A Statistical Framework for Using External Information in Updating Prediction Models with New Biomarker Measures

by

Wenting Cheng

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2017

Doctoral committee:

Professor Jeremy M.G. Taylor, Co-Chair
Professor Bhramar Mukherjee, Co-Chair
Assistant Professor Hui Jiang
Professor Elizaveta Levina

Wenting Cheng

chengwt@umich.edu

ORCID iD: 0000-0003-2199-1467

ACKNOWLEDGEMENTS

I would like to thank my co-advisor, Prof. Jeremy Taylor, for his guidance, encouragement, advice and patience he has provided. I have always been inspired by Jeremy's statistical knowledge, diligence and reliability. Choosing Jeremy as my co-advisor is one of the best decisions I have made in the past years.

I would like to thank my co-advisor, Prof. Bhramar Mukherjee, for her guidance, advice and many many patience for me. I have been extremely lucky to have such a co-advisor who is so knowledgeable, diligent and energetic. Being Bhramar's student has greatly improved my academic writing strategy and presentation skill.

Many thanks to my dissertation committee member, Prof. Elizaveta Levina and Prof. Hui Jiang. Liza was my academic advisor when I was a master student in Statistics Department and she was my role model both as a faculty and a female statistical researcher though I have never told her so in person. Hui, in addition to serving on my committee, gave me helpful suggestions in the algorithms and codes I used in my dissertation research.

I would like to thank my supervisors and collaborators in my GSRA projects, including Prof. Matthew Schipper, Prof. Philip Boonstra from Biostatistics Department, Prof. Richard Neitzel and Ph.D. student Ben Roberts from Environmental Health Sciences Department, Dr. Dawn Misra and Dr. Vinod Misra from Wayne State University.

I also like to thank all the faculty members in the Biostatistics Department, especially Prof. Min Zhang, Prof. Timothy Johnson, Prof. Alexander Tsodikov, Prof. Lu Wang, Prof. Yi Li and Prof. Thomas Braun, for their many helpful suggestions and support in my study and my GSI

work at University of Michigan. I would like to thank my fellows and friends in the Biostatistics Department, especially Yumeng Li, Fan Wu, Sheng Qiu, Tingting Zhou, Krithika Shresh, Lauren Beesley, Allison Furgal, Aaron Chen, Cui Guo, Zihuai He, Zhe Fei, Lu Xia, Lu Tang, Xin Wang and Daniel Muenz.

I must express my gratitude to my parents, for their constant support, understanding and encouragement. Without them, I may never have the resolution and the opportunity to be a graduate student at University of Michigan and eventually enter into the Ph.D. program in Biostatistics. I would like to thank Chang Gong, my boyfriend, who witnesses all my ups and downs in my life and study at University of Michigan. He has always been the cheerleader for my research.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Prediction models are abundant in the clinical and epidemiologic literature. There are established risk prediction models for cancer, cardiovascular diseases and many other chronic diseases. The information from an existing prediction model can be available in the form of coefficient estimates (with or without measures of standard error) or individual prediction probabilities (with or without standard errors). This dissertation poses a principled framework to incorporate such varying types of information while building a new prediction model that adds new candidate biomarkers to the existing model.

In the first chapter, we consider a situation where there is rich historical data available for the coefficients and their standard errors in a linear regression model describing the association between a continuous outcome variable $\mathbf{Y}$ and a set of predicting factors $\mathbf{X}$, from a large study. We would like to utilize this summary information for improving inference in an expanded model of interest, $\mathbf{Y}$ given $\mathbf{X}, \mathbf{B}$. The additional variable $\mathbf{B}$ is a new biomarker, measured on a small number of subjects in a new dataset. We formulate the problem in an inferential framework where the historical information is translated in terms of nonlinear constraints on the parameter space and propose both frequentist and Bayes solutions to this problem. We show that a Bayesian transformation approach proposed by Gunn and Dunson is a simple and effective computational method to conduct approximate Bayesian inference for this constrained parameter problem. The simulation results comparing these methods indicate that historical information on $\mathrm{E}(\mathbf{Y}|\mathbf{X})$ can improve the efficiency of estimation and enhance the predictive power in the regression model of interest $\mathrm{E}(\mathbf{Y}|\mathbf{X}, \mathbf{B})$. We illustrate our methodology by enhancing a published prediction model for bone

lead levels in terms of blood lead and other covariates, with a new biomarker defined through a genetic risk score.

In the second chapter, we further develop and evaluate the strategy of translating the external information into constraints on regression coefficients in the setting of a binary response variable $\mathbf{Y}$ and a logistic regression model. Borrowing from the measurement error literature we establish an approximate relationship between the regression coefficients in the models $\Pr(\mathbf{Y} = 1|\mathbf{X}, \boldsymbol{\beta})$, $\Pr(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B}, \boldsymbol{\gamma})$ and $\mathrm{E}(\mathbf{B}|\mathbf{X}, \boldsymbol{\theta})$ for a Gaussian distribution of $\mathbf{B}$. For binary $\mathbf{B}$ we propose an alternate expression. We illustrate our methodology through simulations and by enhancing the High-grade Prostate Cancer Prevention Trial Risk Calculator, with two new biomarkers prostate cancer antigen 3 and TMPRSS2:ERG.

In the third chapter, the goal is to improve the prediction ability of a risk assessment model, $\Pr(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B})$ constructed from a small dataset by incorporating external information that comes in the form of predicted outcomes from an existing model for $\Pr(\mathbf{Y} = 1|\mathbf{X})$. For example, the existing well-known risk prediction models are often converted into a publicly available online tool or risk calculator to yield a predicted probability of developing the disease for an individual based on a set of risk factors $\mathbf{X}$, but the exact form of the model/algorithm to construct the predictions may not be known. We propose a constrained maximum likelihood method and an approach based on synthetic data and multiple imputation to utilize this information while constructing a model for $\Pr(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B})$.

CHAPTER I

Introduction

Prediction models for an outcome based on a set of input covariates are used in many scientific fields. For example, in biomedicine and public health risk prediction models are often constructed to predict a disease state over a predefined period. The models can be constructed from various perspectives from using simple regression based framework with highly parametric structure to a data-adaptive algorithm-based approach driven by machine learning tools. The stability and accuracy of these models depend on the number of covariates and the sample size (Knofczynski and Mundfrom, 2008). The sample size is desirable to be large in order to provide accurate predictions of the outcome when the primary purpose of the regression model is prediction of response for a new subject.

While increasingly very large datasets are available from which prediction models can be built, in the situation of new or experimental biomarkers the sample sizes of datasets that contain these remain small. In practice, making the best use of available data usually means finding and combining data or information from different sources measuring the same outcome. Often findings from previous similar studies are available in published literature. How to appropriately synthesize all previous study findings on the same outcome of interest has become an increasingly popular topic in statistical research. For example, in evidence-based healthcare, data synthesis has been seen as the key to more coherent and efficient research (Sutton et al., 2009).

For the case that individual-level data is available both in previous studies and current study, and these studies have the same outcome ($\mathbf{Y}$) and share some overlapping set of covariates, there exist statistical methods on aggregating individual-level data from multiple datasets. For example, some methods assume that two datasets with individual-level data are available (Chen and Qin, 2014; Zhan and Ghosh, 2015; Boonstra et al., 2013). Complete data on $(\mathbf{Y}, \mathbf{X}, \mathbf{B})$ was observed in one dataset while data on $(\mathbf{Y}, \mathbf{X})$ was observed in the other dataset where $\mathbf{X}$ and $\mathbf{B}$ are the set of covariates. Chen and Qin (2014) assumed the model of interest is $\mathbf{Y}|\mathbf{X}, \mathbf{B}$ and imputed missing values of $\mathbf{B}$ using data imputation techniques and then the inference on $\mathbf{Y}|\mathbf{X}, \mathbf{B}$ is based on the augmented data. Zhan and Ghosh (2015) proposed a kernel machine-based method to improve the prediction in the model of interest $\mathbf{Y}|\mathbf{B}$ by utilizing the information in $\mathbf{X}$. Boonstra et al. (2013) proposed a shrinkage approach for the inference of $\mathbf{Y}|\mathbf{B}$ model using information derived from a model for $\mathbf{Y}|\mathbf{X}$. All these papers made use of individual-level data.

In many applications, however, the information from historical studies is not individual-level data, but rather summary-level information about estimated models. Such summary-level information can include knowledge about regression parameters in the estimated models. This information can be applied to calibrate the coefficient estimates in the model of interest when fit using the current available dataset (Newcombe et al., 2012) or specified as informative priors for the coefficients of these variables in a Bayesian regression model (Steyerberg et al., 2000). The external information can be used to carry out constrained maximum likelihood estimation or implement estimating equation techniques using the constraints (Chatterjee et al., 2016; Imbens and Lancaster, 1994; Qin, 2000; Qin et al., 2015). Incorporating the summary-level information, such as estimated coefficients from previous studies, into the current study may produce more accurate regression coefficient estimates in the model of interest and improved predictive ability, than an analysis solely based on the individual-level data in the current study.

In this dissertation, the overarching objective is to develop a principled framework for using

external summary-level information when constructing a specific kind of prediction model, a risk prediction model, from a small dataset. Risk prediction models are abundant in the clinical and epidemiologic literature. There are established risk prediction models for cancer, cardiovascular diseases and many other chronic diseases that traditionally incorporate family history, basic socio-demographic factors, lifestyle and behavioral factors, anthropometric measures and alike. With the advent of modern assaying techniques and the omics revolution, new biomarkers are being proposed to add to these existing prediction models for improved personalized predictions.

The information from an existing prediction model can be available in the form of coefficient estimates (with or without measures of standard error) or individual prediction probabilities (with or without standard errors). This dissertation poses a statistical framework to incorporate such varying types of information while building a new prediction model that adds new candidate biomarkers to the existing model. The general premise is that these candidate biomarkers are measured on a small group of subjects while the existing prediction models have been validated in large studies.

In Chapter II, we consider a situation where there is historical information available on the coefficients and their standard errors in a linear regression model describing the association between a continuous outcome variable $\mathbf{Y}$ and a set of predicting factors $\mathbf{X}$, from a large study. We would like to utilize this summary information for improving inference in an expanded model of interest, $\mathbf{Y}$ given $\mathbf{X}, \mathbf{B}$. The additional variable $\mathbf{B}$ is a new biomarker, measured on a small number of subjects in a new dataset. We establish a relationship between $\boldsymbol{\beta}$, the parameters in the model for $\mathrm{E}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta})$ and $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$, the parameters in the models for $\mathrm{E}(\mathbf{Y}|\mathbf{X}, \mathbf{B}, \boldsymbol{\gamma})$ and $\mathrm{E}(\mathbf{B}|\mathbf{X}, \boldsymbol{\theta})$. We formulate the problem in an inferential framework where the historical information is translated in terms of nonlinear constraints on the parameter space and propose both frequentist and Bayes solutions to this problem. We show that a Bayesian transformation approach adapted from the original proposal by Gunn and Dunson (2005) is a simple and effective computational method

to conduct approximate Bayesian inference for this constrained parameter problem. The simulation results comparing these methods indicate that historical information on $\mathrm{E}(\mathbf{Y}|\mathbf{X})$ can improve the efficiency of estimation and enhance the predictive ability in the regression model of interest $\mathrm{E}(\mathbf{Y}|\mathbf{X},\mathbf{B})$. We illustrate our methodology by enhancing a published prediction model for bone lead levels in terms of blood lead and other covariates, with a new biomarker defined through a genetic risk score.

In Chapter III, we further develop and evaluate the strategy of translating the external information into constraints on regression coefficients in the setting of a binary response variable $\mathbf{Y}$ and a logistic regression model. Borrowing from the measurement error literature we establish an approximate relationship between the regression coefficients in the models $\Pr(\mathbf{Y}=1|\mathbf{X},\boldsymbol{\beta})$, $\Pr(\mathbf{Y}=1|\mathbf{X},\mathbf{B},\boldsymbol{\gamma})$ and $\mathrm{E}(\mathbf{B}|\mathbf{X},\boldsymbol{\theta})$ for a Gaussian distribution of $\mathbf{B}$. For binary $\mathbf{B}$ we propose an alternate expression. We illustrate our methodology through simulations and by enhancing the High-grade Prostate Cancer Prevention Trial Risk Calculator (Thompson et al., 2006), with two new biomarkers prostate cancer antigen 3 and TMPRSS2:ERG (Tomlins et al., 2015). The research in Chapter II and Chapter III differs from that in the literature because it directly uses the relationship between $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ in the estimation techniques, whereas other methods incorporate the external information in other ways. The methods we propose also incorporate the standard errors of $\boldsymbol{\beta}$ into the estimation method, which has not been attempted in other literature.

In Chapter IV, the goal is to improve the prediction ability of a risk assessment model, $\Pr(\mathbf{Y}=1|\mathbf{X},\mathbf{B})$ constructed from a small dataset by incorporating external information that comes in the form of predicted outcomes from an existing model for $\Pr(\mathbf{Y}=1|\mathbf{X})$. For example, the well-known risk prediction models are often converted into publicly available online tools or risk calculators to yield a predicted probability of developing the disease for an individual based on a set of risk factors $\mathbf{X}$, but the exact form of the model/algorithm to construct the predictions may not be known. We propose a constrained maximum likelihood method and an approach based

on synthetic data and multiple imputation to utilize this information while constructing a model for $\Pr(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B})$. Simulation results are presented to compare the proposed methods under varying scenarios.

To conclude, aggregation of data from diverse sources that are publicly available is an important problem in the current scientific context in biomedicine and public health. Risk prediction models are being enhanced by newly discovered molecular biomarkers. This dissertation makes a contribution toward principled data integration in this context and leads to multiple new directions of future research in this area.

Bibliography

Boonstra, P. S., Taylor, J. M. and Mukherjee, B. Incorporating auxiliary information for improved prediction in high-dimensional datasets: an ensemble of shrinkage approaches. *Biostatistics*, 14:259–272, 2013.

Chatterjee, N., Chen, Y.-H., Maas, P. and Carroll, R. J. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111(513):107–117, 2016.

Chen, B. and Qin, J. Use of empirical likelihood to calibrate auxiliary information in partly linear monotone regression models. *Statistics in Medicine*, 33(10):1713–1722, 2014.

Gunn, L. H. and Dunson, D. B. A transformation approach for incorporating monotone or unimodal constraints. *Biostatistics*, 6(3):434–449, 2005.

Imbens, G. W. and Lancaster, T. Combining micro and macro data in microeconometric models. *The Review of Economic Studies*, 61(4):655–680, 1994.

Knofczynski, G. T. and Mundfrom, D. Sample sizes when using multiple linear regression for prediction. *Educational and Psychological Measurement*, 68(3):431–442, 2008.

Newcombe, P. J., Reck, B. H., Sun, J., Platek, G. T., Verzilli, C., Kader, A. K., Kim, S.-T., Hsu, F.-C., Zhang, Z., Zheng, S. L., Mooser, V. E., Condreay, L. D., Spraggs, C. F., Whittaker, J. C., Rittmaster, R. S. and Xu, J. A comparison of Bayesian and frequentist approaches to incorporating external information for the prediction of prostate cancer risk. *Genetic Epidemiology*, 36(1):71–83, 2012.

Qin, J. Combining parametric and empirical likelihoods. *Biometrika*, 87(2):484–490, 2000.

Qin, J., Zhang, H., Li, P., Albanes, D. and Yu, K. Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika*, 102(1):169–180, 2015.

Steyerberg, E. W., Eijkemans, M. J. C., Van Houwelingen, J. C., Lee, K. L. and Habbema, J. D. F. Prognostic models based on literature and individual patient data in logistic regression analysis. *Statistics in Medicine*, 19(2):141–160, 2000.

Sutton, A. J., Cooper, N. J. and Jones, D. R. Evidence synthesis as the key to more coherent and efficient research. *BMC Medical Research Methodology*, 9(1):29, 2009.

Thompson, I. M., Ankerst, D. P., Chi, C., Goodman, P. J., Tangen, C. M., Lucia, M. S., Feng, Z., Parnes, H. L. and Coltman, C. A. Assessing prostate cancer risk: results from the prostate cancer prevention trial. *Journal of the National Cancer Institute*, 98(8):529–534, 2006.

Tomlins, S. A., Day, J. R., Lonigro, R. J., Hovelson, D. H., Siddiqui, J., Kunju, L. P., Dunn, R. L., Meyer, S., Hodge, P., Groskopf, J., Wei, J. T. and Chinnaiyan, A. M. Urine TMPRSS2:ERG plus PCA3 for individualized prostate cancer risk assessment. *European Urology*, 70:45–53, 2015.

Zhan, X. and Ghosh, D. Incorporating auxiliary information for improved prediction using combination of kernel machines. *Statistical Methodology*, 22:47–57, 2015.

CHAPTER II

Improving estimation and prediction in linear regression incorporating external information from an established reduced model

2.1 Introduction

In clinical biomedicine, there are many well-known models describing the association between a measure of disease and patient characteristics, treating the measure of disease as the outcome and patient characteristics as predicting variables. Examples include Framingham risk score (D'Agostino et al., 2001), Prostate Cancer Prevention Trial calculator (Thompson et al., 2006) and Gail model (Gail et al., 1989). They can make predictions for future patients, based on their individual characteristics. These models could then be used in the settings of early detection and screening, or help decisions on treatment after diagnosis, or monitor for progression after treatment. While these models are well established, it is possible that including some additional candidate biomarkers and constructing an expanded model will improve the prediction ability.

The challenge of estimating the expanded model is that the additional biomarkers are measured only on a small number of subjects in a new dataset, thus inference in the expanded model tends to give relatively poor coefficient estimates with large standard errors and low prediction accuracy. It is natural to consider incorporating information that is available from an established model into the expanded model to improve the estimates of the parameters and the prediction ability of the newly

developed model. Such external information is often available, however it may not come in a direct or convenient form. We consider a situation where the outcome is a continuous marker of disease risk and the established regression model is described in an article, in which the estimated regression coefficients and their standard errors are presented in tables. The expanded model, however, includes one additional biomarker as a predicting variable. How to incorporate this coefficient information in a principled way is a non-trivial statistical problem.

The use of external information is a popular strategy for improving efficiency in statistical inference. Often the information can be expressed as constraints on the regression coefficients and one can conduct constrained maximum likelihood inference. The problem of inference for regression coefficients from linear regression subject to a set of constraints has been considered from the Bayesian perspective, either by discarding draws violating the constraints (Geweke, 1986) or translating the constraints as informative priors. Geweke uses noninformative priors and an indicator function representing the inequality constraints. The posterior distributions are then computed using importance sampling. Though this idea is easy to implement, it could be extremely slow computationally, especially when the truncation region has a small probability. Dunson and Neelon (2003) and Gunn and Dunson (2005) propose a simple approach to handle constraints by generating sample draws from the unconstrained posterior distribution and mapping these draws to the constrained space. Their interest is primarily in order-restricted inference, and they choose the constrained draw that minimizes the Mahalanobis distance between the unconstrained draws and the ordered draws, across different choices of ordered draws.

There is literature emerging on new frequentist proposals to incorporate external information. Chen et al. (2015) propose a linear regression shrinkage method for predictions in a small dataset calibrated by a larger but biased dataset. Imbens and Lancaster (1994) investigate how aggregate data (e.g., the population average of the response) could be used to improve maximum likelihood estimates in a regression model. They show that the gains from incorporating such information

could be substantial. Qin (2000) propose that the aggregate data can be incorporated into the empirical likelihood and the combination of empirical and parametric likelihood could provide valid inference for the regression coefficients. Qin et al. (2015) consider auxiliary information (e.g., disease prevalence at different levels of risk factors) as constraints on the regression coefficients and the joint covariate distribution. They use empirical likelihood and general estimating equations for estimation. Chatterjee et al. (2016) utilize summary-level information from external data sources of large sample size to calibrate the current regression model.

To introduce notation, let $\mathbf{Y}$ denote the outcome, which is assumed to be continuous, and we have a set of standard risk covariates $\mathbf{X}$ (there is no assumption regarding the distribution of $\mathbf{X}$) and a new continuous covariate $\mathbf{B}$ measured on a small dataset. The model of primary interest is a regression model that describes the joint effect of $\mathbf{X}, \mathbf{B}$ on $\mathbf{Y}$:

$$\mathrm{E}(\mathbf{Y}|\mathbf{X}, \mathbf{B}) = \mathbf{X}\boldsymbol{\gamma_X} + \mathbf{B}\gamma_{\mathrm{B}} \tag{2.1}$$

We could also estimate $\mathrm{E}(\mathbf{B}|\mathbf{X})$ in this small dataset from a model of the form:

$$\mathrm{E}(\mathbf{B}|\mathbf{X}) = \mathbf{X}\boldsymbol{\theta} \tag{2.2}$$

A large, well-characterized previous study describes the association between $\mathbf{X}$ and $\mathbf{Y}$ through a regression model:

$$\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \tag{2.3}$$

The knowledge we obtain from the previous study is summary-level information on model (2.3): estimated regression coefficients and their standard errors. We assume we know the empirical variances (standard error squared) of regression coefficients, but not their covariances, since the estimated covariance matrices are rarely reported in publications. We use $\bar{\boldsymbol{\beta}}$ and $\bar{\mathbf{S}}$ to denote the reported coefficient estimates and their standard errors.

We formulate the problem in an inferential framework where the external information from (2.3) is translated in terms of nonlinear constraints on the regression parameters and propose both

frequentist and Bayes solutions to this problem. The goal is to improve the estimation of linear regression coefficients $\boldsymbol{\gamma}$ and prediction power of model (2.1) incorporating external coefficients information from (2.3), when the sample size is small. When the current dataset is large, the potential gain by incorporating external information from an established reduced model may be limited.

The following is the structure of the remainder of this chapter: in Section 2.2, we discuss how to transform the available external information into constraints on the regression coefficients. We show two unconstrained solutions based on current data, ignoring historical information and propose four constrained solutions that use the historical information: constrained maximum likelihood, partial regression, informative full Bayes and Bayesian transformation approach. We present a simulation study in Section 2.3. Section 2.4 is an application of the approaches to enhance a prediction model for bone lead levels based on data from the Normative Aging Study published in 2009 with new genetic marker information. We discuss the findings and possibilities for future work in Section 2.5.

2.2 Statistical Approaches

2.2.1 *Relationship Equations*

Assume that $\mathbf{X}$ has $p+1$ dimensions (including an intercept). $\mathbf{X_0} = 1$ by notational convention. From models (2.1), (2.2) and (2.3), we find that the regression of $\mathbf{Y}$ on $\mathbf{X}, \mathbf{B}$ is a linear function of $\mathrm{E}(\mathbf{B}|\mathbf{X})$:

$$\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \mathrm{E}(\mathrm{E}(\mathbf{Y}|\mathbf{X},\mathbf{B})|\mathbf{X}) = \mathrm{E}(\mathbf{X}\boldsymbol{\gamma_X} + \mathbf{B}\gamma_{\mathrm{B}}|\mathbf{X}) = \mathbf{X}\boldsymbol{\gamma_X} + \gamma_{\mathrm{B}}\mathrm{E}(\mathbf{B}|\mathbf{X}) \tag{2.4}$$

We are going to estimate model (2.1), $\mathrm{E}(\mathbf{Y}|\mathbf{X},\mathbf{B}) = \gamma_0 + \gamma_1\mathbf{X_1} + \cdots + \gamma_{\mathrm{p}}\mathbf{X_p} + \gamma_{\mathrm{p+1}}\mathbf{B}$ and model (2.2), $\mathrm{E}(\mathbf{B}|\mathbf{X}) = \theta_0 + \theta_1\mathbf{X_1} + \cdots + \theta_{\mathrm{p}}\mathbf{X_p}$ from the current small dataset. Using equation (2.4) and the historical information in model (2.3), we have the following equation:

$$\bar{\beta}_0 + \bar{\beta}_1\mathbf{X_1} + \cdots + \bar{\beta}_{\mathrm{p}}\mathbf{X_p} = \gamma_0 + \gamma_1\mathbf{X_1} + \cdots + \gamma_{\mathrm{p}}\mathbf{X_p} + \gamma_{\mathrm{p+1}}(\theta_0 + \theta_1\mathbf{X_1} + \cdots + \theta_{\mathrm{p}}\mathbf{X_p}) \tag{2.5}$$

which implies the relationship between parameters in models (2.1)-(2.3) as:

$$\beta_j = \gamma_j + \gamma_{p+1}\theta_j, j = 0, \ldots, p \tag{2.6}$$

We summarize the necessary assumptions required for constructing the relationship equations (2.6) in Table 2.1. Essentially, when constructing the relationship equations connecting the parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$, from the models for $\mathrm{E}(\mathbf{Y}|\mathbf{X})$, $\mathrm{E}(\mathbf{Y}|\mathbf{X},\mathbf{B})$ and $\mathrm{E}(\mathbf{B}|\mathbf{X})$ respectively, we assume that $\mathbf{Y}$ and $\mathbf{B}$ are continuous variables. We do assume that though $\mathbf{B}$ is not available in the external historical data, if it were available, $\mathrm{E}(\mathbf{Y}|\mathbf{X},\mathbf{B})$ is linear in $(\mathbf{X},\mathbf{B})$, $\mathrm{E}(\mathbf{Y}|\mathbf{X})$ is linear in $\mathbf{X}$ and $\mathrm{E}(\mathbf{B}|\mathbf{X})$ is linear in $\mathbf{X}$. We further assume that the models for $\mathrm{E}(\mathbf{Y}|\mathbf{X},\mathbf{B})$ and $\mathrm{E}(\mathbf{Y}|\mathbf{X})$ are correctly specified in both the internal and the external study and they are same in both internal and external populations. Moreover, by using the summary information on the regression coefficients from the large dataset, we implicitly assume that the available results provide consistent estimates of $\boldsymbol{\beta}$ and standard errors of $\boldsymbol{\beta}$. No additional distributional assumptions are needed to establish the constraints.

2.2.2 *Unconstrained Solutions*

Direct regression

Without constraints, $\boldsymbol{\gamma}$ can be estimated by ordinary least squares directly:

$$\min_{\gamma} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{Y_i} - \sum_{\mathrm{j}=0}^{\mathrm{p}} \gamma_{\mathrm{j}} \mathrm{X_{ij}} - \gamma_{\mathrm{p+1}} \mathrm{B_i})^2 \tag{2.7}$$

This estimates parameters in model (2.1). To estimate parameters in model (2.2), we obtain least squares estimates of $\boldsymbol{\theta}$ by considering:

$$\min_{\theta} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{B_i} - \sum_{\mathrm{j}=0}^{\mathrm{p}} \theta_{\mathrm{j}} \mathrm{X_{ij}})^2 \tag{2.8}$$

Standard Bayes

Analogous to direct regression, we can perform standard Bayesian linear regression with non-informative conjugate priors (Carlin and Louis, 2009). We use standard Bayes linear regression procedures to fit model (2.1) and model (2.2) separately.

For model (2.1), the likelihood function is derived from $\mathbf{Y}|\mathbf{X},\mathbf{B},\boldsymbol{\gamma},\sigma_1^2 \sim \mathrm{N_n}(\sum_{\mathrm{j=0}}^{\mathrm{p}}\gamma_{\mathrm{j}}\mathbf{X_j} + \gamma_{\mathrm{p+1}}\mathbf{B},\sigma_1^2\mathbf{I_n})$. We specify independent priors for regression coefficient $\boldsymbol{\gamma}$ and residual variance σ_1^2 (Lesaffre and Lawson, 2012): $\pi(\boldsymbol{\gamma},\sigma_1^2) = \pi(\boldsymbol{\gamma})\cdot\pi(\sigma_1^2) = \mathrm{N_{(p+2)}}(\boldsymbol{\mu_0},\boldsymbol{\Sigma_0})\cdot$inverse-gamma$(\frac{\nu_0}{2},\frac{\nu_0\sigma_0^2}{2}) = \mathrm{N}(0,100^2\mathbf{I_{(p+2)(p+2)}})\cdot\mathrm{IG}(0.01,0.01)$. The joint posterior distribution $\mathrm{p}(\boldsymbol{\gamma},\sigma_1^2|\mathbf{Y},\mathbf{X},\mathbf{B})$ is then proportional to $\mathrm{p}(\mathbf{Y}|\mathbf{X},\mathbf{B},\boldsymbol{\gamma},\sigma_1^2)\cdot\pi(\boldsymbol{\gamma})\cdot\pi(\sigma_1^2)$. The full conditional distribution of $\boldsymbol{\gamma}$ (i.e., $\mathrm{p}(\boldsymbol{\gamma}|\mathbf{Y},\mathbf{X},\mathbf{B},\sigma_1^2)$) is then $\mathrm{N}\Big(\big(\boldsymbol{\Sigma_0}^{-1} + \frac{(\mathbf{X},\mathbf{B})^{\mathrm{T}}(\mathbf{X},\mathbf{B})}{\sigma_1^2}\big)^{-1}(\boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0} + \frac{(\mathbf{X},\mathbf{B})^{\mathrm{T}}\mathbf{Y}}{\sigma_1^2}),\big(\boldsymbol{\Sigma_0}^{-1} + \frac{(\mathbf{X},\mathbf{B})^{\mathrm{T}}(\mathbf{X},\mathbf{B})}{\sigma_1^2}\big)^{-1}\Big)$ and the conditional distribution of σ_1^2 is inverse-gamma$\Big(\frac{\mathrm{n}+\nu_0}{2},\frac{1}{2}\big[\big(\mathbf{Y}-(\mathbf{X},\mathbf{B})\boldsymbol{\gamma}\big)^{\mathrm{T}}\big(\mathbf{Y}-(\mathbf{X},\mathbf{B})\boldsymbol{\gamma}\big)+\nu_0\sigma_0^2\big]\Big)$. For model (2.2), the prior specifications and inferences are very similar to that of model (2.1).

Using Markov chain sampling techniques like Gibbs sampling, standard Bayes can be implemented in a fast and easy algorithm to obtain posterior draws of $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$. Direct regression and standard Bayes do not use external information and serve as references for quantifying the amount of efficiency we gain by using external information.

2.2.3 *Constrained Solutions*

Constrained maximum likelihood

The constrained maximum likelihood (constrained ML) method uses optimization of the likelihood under the constraints in (2.6). As we have information on both the point estimate and the standard error of $\boldsymbol{\beta}$, we will require estimates of the parameters such that the new $\boldsymbol{\beta}$ to be within d standard errors of the old point estimate. Our constrained maximum likelihood estimation opti-

mizes the joint log-likelihood, namely:

$$\begin{aligned}\log(\mathrm{L}) = &\sum_{\mathrm{i}=1}^{\mathrm{n}}[-\frac{1}{2}\log(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2}(\mathrm{Y_i} - \sum_{\mathrm{j}=0}^{\mathrm{p}}\gamma_{\mathrm{j}}\mathrm{X_{ij}} - \gamma_{\mathrm{p+1}}\mathrm{B_i})^2]+ \\ &\sum_{\mathrm{i}=1}^{\mathrm{n}}[-\frac{1}{2}\log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}(\mathrm{B_i} - \sum_{\mathrm{j}=0}^{\mathrm{p}}\theta_{\mathrm{j}}\mathrm{X_{ij}})^2]\end{aligned} \tag{2.9}$$

subject to the set of nonlinear constraints: $\gamma_{\mathrm{j}} + \gamma_{\mathrm{p+1}}\theta_{\mathrm{j}} \in [\bar{\beta}_{\mathrm{j}} - \mathrm{d}\bar{\mathrm{S}}_{\mathrm{j}}, \bar{\beta}_{\mathrm{j}} + \mathrm{d}\bar{\mathrm{S}}_{\mathrm{j}}], \mathrm{j} = 0, \ldots, \mathrm{p}$

This method is equivalent to minimizing the weighted sum of squared errors of $\mathbf{Y}|\mathbf{X}, \mathbf{B}$ and the squared errors of $\mathbf{B}|\mathbf{X}$, namely:

$$\min_{\gamma,\theta}\left\{\frac{1}{\sigma_1^2}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{Y_i} - \sum_{\mathrm{j}=0}^{\mathrm{p}}\gamma_{\mathrm{j}}\mathrm{X_{ij}} - \gamma_{\mathrm{p+1}}\mathrm{B_i})^2 + \frac{1}{\sigma_2^2}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{B_i} - \sum_{\mathrm{j}=0}^{\mathrm{p}}\theta_{\mathrm{j}}\mathrm{X_{ij}})^2\right\} \tag{2.10}$$

$s.t. \gamma_{\mathrm{j}} + \gamma_{\mathrm{p+1}}\theta_{\mathrm{j}} \in [\bar{\beta}_{\mathrm{j}} - \mathrm{d}\bar{\mathrm{S}}_{\mathrm{j}}, \bar{\beta}_{\mathrm{j}} + \mathrm{d}\bar{\mathrm{S}}_{\mathrm{j}}], \mathrm{j} = 0, \ldots, \mathrm{p}$

In this optimization problem, instead of treating σ_1^2 and σ_2^2 as unknown parameters, we use $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ as plug-in estimates, which are the OLS residual variances from $\mathrm{E}(\mathbf{Y}|\mathbf{X}, \mathbf{B})$ and $\mathrm{E}(\mathbf{B}|\mathbf{X})$.

The width of the constrained interval is controlled by d, which is a scale parameter representing the strength of external information. From simulations, we find that fixing d as $d = 1$ may be reasonable. This is an optimization problem with nonlinear inequality constraints. To solve it, we use function **solnp** in R package **Rsolnp**, a function that efficiently solves general nonlinear optimization problems using Lagrange multipliers. The starting point is the OLS estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$, namely $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\theta}}$. For computational convenience, we further specify wide lower and upper bounds for each of the parameters: $\gamma_{\mathrm{j}} \in [\hat{\gamma}_{\mathrm{j}} - 5\hat{\mathrm{SE}}(\gamma_{\mathrm{j}}), \hat{\gamma}_{\mathrm{j}} + 5\hat{\mathrm{SE}}(\gamma_{\mathrm{j}})], \mathrm{j} = 0, \ldots, \mathrm{p}+1, \theta_{\mathrm{j}} \in [\hat{\theta}_{\mathrm{j}} - 5\hat{\mathrm{SE}}(\theta_{\mathrm{j}}), \hat{\theta}_{\mathrm{j}} + 5\hat{\mathrm{SE}}(\theta_{\mathrm{j}})], \mathrm{j} = 0, \ldots, \mathrm{p}$.

The standard error of the estimates in the constrained ML solution is hard to derive. Usually the distribution of the constrained maximum likelihood estimate may be derived by expressing the constrained estimate as functions of both the unconstrained estimate and the data and then applying Taylor expansion. However, the fact that the constraints are in the form that is known as "box" constraints in the optimization literature in (2.10) and these box constraints involve nonlinear

functions of the regression parameters makes it impossible to implement this procedure for our solution. Instead we use the bootstrap to estimate the standard error. The bootstrap procedure is described in the Appendix A.

In simulations we find that the constrained maximum likelihood estimate can show substantial bias for small sample sizes. As an alternative we consider the bootstrap bias-corrected estimate $\hat{\gamma}_{bc}$, given by $\hat{\gamma}_{bc} = 2\hat{\gamma} - \tilde{\gamma}$, where $\hat{\gamma}$ is the original estimate, $\tilde{\gamma}$ is the mean of the bootstrap estimates (Efron and Tibshirani, 1986). We use this bootstrap bias-correction procedure to modify the constrained maximum likelihood solution in the simulation studies. For the real data analysis, we also provide a bias-corrected 95% confidence interval: $\left(\mathrm{F}_{\tilde{\gamma}}^{-1}(\Phi(2\mathbf{b}+\mathrm{Z}_{0.025})), \mathrm{F}_{\tilde{\gamma}}^{-1}(\Phi(2\mathbf{b}+\mathrm{Z}_{0.975}))\right)$ where $\mathbf{b}$ is estimated from the bootstrap distribution by $\Phi^{-1}(\Pr(\tilde{\gamma} \leq \hat{\gamma}))$, and Φ is the cumulative distribution function of the normal distribution (Efron, 1981; Carpenter and Bithell, 2000). When the sample size is small, we recommend bootstrap bias-corrected constrained ML instead of the constrained ML as an alternative method for reducing the bias. However this bootstrap bias-corrected constrained ML is not necessary for large sample sizes. Additional simulation studies in the supplementary material demonstrate the performance of this bias-corrected bootstrap procedure in simulation settings of various sample sizes.

Partial regression

Partial regression is an indirect method to estimate the amount by which a dependent variable increases when one of the predicting variables is increased by one unit with all other predicting variables held constant (Abdi, 2004).

Our adaptation of the partial regression method is an attempt to look at the relationship between the response and the new explanatory variable while preserving the effect from the old set of explanatory variables. The following three simple steps describe the proposed partial regression method:

1. Remove the effect of variables $\mathbf{X}$ on the response $\mathbf{Y}$ by computing $\boldsymbol{r_1}$: $\boldsymbol{r_1} = \mathbf{Y} - \sum_{j=0}^{p} \bar{\beta}_j \mathbf{X_j}$.

That is, remove the effect of $\mathbf{X}$ on $\mathbf{Y}$ in the small dataset using historical information, as reflected through plugging in the estimate $\bar{\boldsymbol{\beta}}$

2. Remove from $\mathbf{B}$ the effect of correlation due to variables $\mathbf{X}$: estimate coefficients $\boldsymbol{\theta}$ from $\mathrm{E}(\mathbf{B}|\mathbf{X})$, calculate $\mathbf{r_2} = \mathbf{B} - \sum_{j=0}^{p} \hat{\theta}_j \mathbf{X_j}$. That is, regress $\mathbf{B}$ against $\mathbf{X}$ from our small dataset

3. Regress Y-residuals against B-residuals: estimate coefficients α_0, α_1, in $E(\mathbf{r_1}|\mathbf{r_2}) = \alpha_0 + \alpha_1 \mathbf{r_2}$ by OLS. The estimated coefficients for $\mathrm{E}(\mathbf{Y}|\mathbf{X}, \mathbf{B})$ are then given by:

$$\begin{cases} \hat{\gamma_0} = \hat{\alpha}_0 - \hat{\alpha}_1 \hat{\theta}_0 + \bar{\beta}_0 \\ \hat{\gamma_j} = -\hat{\alpha}_1 \hat{\theta}_j + \bar{\beta}_j, j = 1, \ldots, p \\ \hat{\gamma}_{p+1} = \hat{\alpha}_1 \end{cases} \tag{2.11}$$

The standard error of the partial regression estimate is not easy to derive. The partial regression estimate depends on the estimated $\boldsymbol{\theta}$ from $\mathbf{B}$ regressed on $\mathbf{X}$ and the estimated $\boldsymbol{\beta}$ from historical information. Because of this dependence, the marginal distribution of the partial regression estimate is not in closed form. We use a variance approximation linearization technique to estimate the standard error of the partial regression estimate. This procedure is described in Appendix B.

Informative full Bayes

We suggest a Bayesian approach with informative priors, based on a Markov chain Monte Carlo (MCMC) technique using a Metropolis-Hastings sampling algorithm. The first step is to write down the joint likelihood function, $\mathrm{L}(\mathbf{Y}, \mathbf{B}|\mathbf{X}) = \mathrm{L}(\mathbf{Y}|\mathbf{X}, \mathbf{B})\mathrm{L}(\mathbf{B}|\mathbf{X})$ with prior $\pi(\boldsymbol{\gamma}, \boldsymbol{\theta}, \sigma_1^2, \sigma_2^2)$. This joint likelihood function is a valid likelihood and therefore could be used for Bayesian inference.

$$\mathrm{p}(\boldsymbol{\gamma}, \boldsymbol{\theta}, \sigma_1^2, \sigma_2^2|\mathrm{data}) \propto \left\{ \prod_{\mathrm{i}=1}^{\mathrm{n}} \frac{1}{\sqrt{2\pi\sigma_1^2}} \mathrm{e}^{-\frac{1}{2\sigma_1^2}(\mathrm{Y_i} - \sum_{\mathrm{j}=0}^{\mathrm{p}} \gamma_{\mathrm{j}} \mathrm{X_{ij}} - \gamma_{\mathrm{p}+1} \mathrm{B_i})^2} \cdot \frac{1}{\sqrt{2\pi\sigma_2^2}} \mathrm{e}^{-\frac{1}{2\sigma_2^2}(\mathrm{B_i} - \sum_{\mathrm{j}=0}^{\mathrm{p}} \frac{\beta_{\mathrm{j}} - \gamma_{\mathrm{j}}}{\gamma_{\mathrm{p}+1}} \mathrm{X_{ij}})^2} \right\} \cdot \pi(\boldsymbol{\gamma}, \boldsymbol{\theta}, \sigma_1^2, \sigma_2^2) \tag{2.12}$$

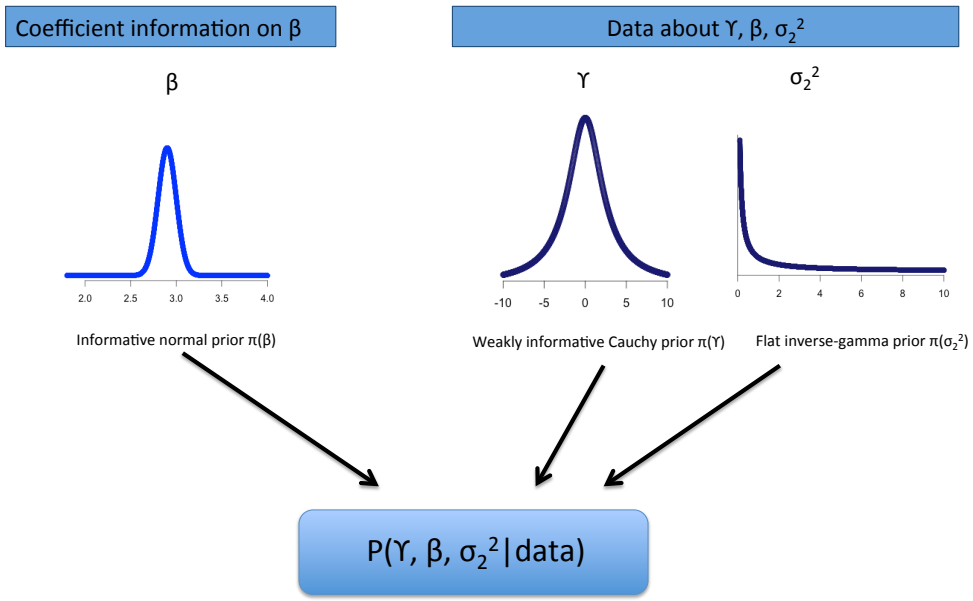


Figure 2.1: Illustration of informative full Bayes

We can re-parameterize (2.12) in terms of variables $\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_1^2, \sigma_2^2$, by a Jacobian transformation using the constraints in (2.6) as the underlying transformation. An illustration of a Jacobian transformation in this informative full Bayes solution is shown in Figure 2.1, for the case that there are three variables, $\mathrm{X}_1, \mathrm{X}_2, \mathrm{B}$. The Jacobian matrix is denoted by $\mathbf{J}$. We further assume independent priors for $\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_1^2$ and σ_2^2. Since we have no information for parameters $\boldsymbol{\gamma}, \sigma_1^2, \sigma_2^2$, we assume non-informative priors $\mathrm{N}(0, 100^2\mathrm{I}_{(\mathrm{p}+2)\times(\mathrm{p}+2)})$, $\mathrm{IG}(0.01, 0.01)$ and $\mathrm{IG}(0.01, 0.01)$; for parameter $\boldsymbol{\beta}$, we use the constraints directly as priors:

$$\beta_{\mathrm{j}} = \gamma_{\mathrm{j}} + \gamma_{\mathrm{p}+1}\theta_{\mathrm{j}} \sim \mathrm{N}(\bar{\beta}_{\mathrm{j}}, \bar{\mathrm{S}}_{\mathrm{j}}^2), \mathrm{j} = 0, \ldots, \mathrm{p} \tag{2.13}$$

Then, we can rewrite the joint posterior distribution of $\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_1^2, \sigma_2^2$ as

$$\mathrm{p}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_1^2, \sigma_2^2|\mathrm{data}) \propto \left\{ \prod_{\mathrm{i}=1}^{\mathrm{n}} \frac{1}{\sqrt{2\pi\sigma_1^2}} \mathrm{e}^{-\frac{1}{2\sigma_1^2}(\mathrm{Y_i} - \sum_{\mathrm{j}=0}^{\mathrm{p}} \gamma_{\mathrm{j}} \mathrm{X_{ij}} - \gamma_{\mathrm{p}+1}\mathrm{B_i})^2} \cdot \frac{1}{\sqrt{2\pi\sigma_2^2}} \mathrm{e}^{-\frac{1}{2\sigma_2^2}(\mathrm{B_i} - \sum_{\mathrm{j}=0}^{\mathrm{p}} \frac{\beta_{\mathrm{j}} - \gamma_{\mathrm{j}}}{\gamma_{\mathrm{p}+1}} \mathrm{X_{ij}})^2} \right\} \cdot$$
$$\pi(\boldsymbol{\beta}) \cdot \pi(\boldsymbol{\gamma}) \cdot \pi(\sigma_1^2) \cdot \pi(\sigma_2^2) \cdot |\mathbf{J}| \tag{2.14}$$

After some algebraic calculations, we find that the conditional distribution of $\beta_0, \ldots, \beta_p$ are normal, each with distribution function $\mathrm{N}(\mu_{\beta_\mathrm{j},\mathrm{n}}, \sigma^2_{\beta_\mathrm{j},\mathrm{n}}), \mathrm{j} = 0, \ldots, \mathrm{p}$; the conditional distribution of $\gamma_0, \ldots, \gamma_p$ are normal, each with distribution function $\mathrm{N}(\mu_{\gamma_\mathrm{j},\mathrm{n}}, \sigma^2_{\gamma_\mathrm{j},\mathrm{n}}), \mathrm{j} = 0, \ldots, \mathrm{p}$; the conditional distribution of σ_1^2 and σ_2^2 are $\mathrm{IG}(\frac{\nu_{1,\mathrm{n}}}{2}, \frac{\nu_{1,\mathrm{n}}\sigma^2_{1,\mathrm{n}}}{2})$ and $\mathrm{IG}(\frac{\nu_{2,\mathrm{n}}}{2}, \frac{\nu_{2,\mathrm{n}}\sigma^2_{2,\mathrm{n}}}{2})$ respectively. The full conditional distribution of γ_{p+1} does not have a closed form. We use a Metropolis-Hastings sampling algorithm to obtain samples from the full conditional of γ_{p+1}. The complete form of the full conditional distributions are presented in Appendix C.

A common drawback of approaches based on vanilla MCMC technique with a constrained parameter space is that the convergence rate is too slow. Roberts (1996) and Gilks and Roberts (1996) suggest that the rate of convergence depends on the posterior correlation between the sample draws of the parameters. We find that in our problem, due to the nonlinear relationship between the parameters, we obtain highly correlated posterior draws in the Markov chain and thus the effective draws are only a small portion of the total draws in the chain. Also, the chain does not move rapidly through the entire support of the posterior distribution and has poor mixing properties. As a consequence, though this informative full Bayes approach provides an exact posterior distribution for all parameters, it is not computationally efficient in our problem.

Bayesian transformation approach

We would like to find an approximate Bayes approach that is computationally inexpensive under constraints when compared to the informative full Bayes described in the previous section. The motivation for our approach stems from the transformation approach incorporating monotone or unimodal constraints proposed in Gunn and Dunson (2005), which we described in Section 2.1. Gunn and Dunson (2005) show that under monotone transformations/order restrictions, the posterior mode with transformed draws can be shown to be consistent estimator for the true posterior mean of the proper constrained Bayes solution. Beyond such monotone constraints, the intuition is that it is desirable that a Bayesian approach could produce posterior draws which are compatible

with the constraints and thus tends to result in values of draws with high density in the constrained space. These constrained posterior draws are based on minimal movement from the unconstrained draws produced by standard Bayes. Inspired by their idea, we first use the unconstrained Bayes method implemented with Gibbs sampling to characterize the posterior distribution and then map the draws from this posterior distribution into the constrained space. We modify their approach in two ways:

1. The constraints: the constraints in our problem are a set of inequality constraints on regression coefficients. The inequality constraints are obtained from historical information and are posed directly on the regression coefficients. These box constraints can be relaxed or strengthened depending on to what extent one wishes to use the historical information, through the choice of the window size "d".

2. The distance measure: the dissimilarity between the unconstrained draws and the constrained draws is measured by normalized Euclidean distance rather than Mahalanobis distance (Gunn and Dunson, 2005). Mahalanobis distance is a dissimilarity measure between two random vectors $\vec{\boldsymbol{v}}_{\mathbf{1}}, \vec{\boldsymbol{v}}_{\mathbf{2}}$ with covariance matrix $\boldsymbol{S}$ as $\mathrm{d}_{\mathrm{MD}}(\vec{\boldsymbol{v}}_{\mathbf{1}}, \vec{\boldsymbol{v}}_{\mathbf{2}}) = \sqrt{(\vec{\boldsymbol{v}}_{\mathbf{1}} - \vec{\boldsymbol{v}}_{\mathbf{2}})^T \boldsymbol{S}^{-1} (\vec{\boldsymbol{v}}_{\mathbf{1}} - \vec{\boldsymbol{v}}_{\mathbf{2}})}$. If the above covariance matrix is diagonal, the Mahalanobis distance reduces to normalized Euclidean distance: $\mathrm{d}_{\mathrm{NED}}(\vec{\boldsymbol{v}}_{\mathbf{1}}, \vec{\boldsymbol{v}}_{\mathbf{2}}) = \sqrt{\sum_{i=1}^{N} \frac{(v_{1i} - v_{2i})^2}{S_i}}$. This normalized Euclidean distance is preferred to the Mahalanobis distance for our problem because it contains only separable functions: the distance measure in each direction is detached from the distance measures in all other directions. This is especially useful for improving the computational efficiency. Also, this distance measure has a natural appeal because it only requires the knowledge of variances instead of the full variance-covariance matrix of the established model, as it is often the case that in literature the established model is presented in a table where standard errors/confidence intervals of the reported estimates are provided.

In general, our transformation approach is defined as follows: assume a vector of parame-

ters $\boldsymbol{\gamma}_{(\mathbf{1}\times \boldsymbol{p})}$ are the coefficients in a regression model and subject to some inequality constraints $\mathbf{C}: \{C_1, C_2, \ldots, C_m\}$. If $\boldsymbol{\gamma}$ are the coefficient estimates that can be easily obtained (i.e., computationally efficient) from standard Bayesian regression ignoring the constraints and $\Omega \subset \mathbb{R}^p$ is a subset of $\mathbb{R}^p$ defined by constraints $\mathbf{C}$ on the elements of $\boldsymbol{\gamma}$, $\Omega = \{\boldsymbol{\gamma} : \boldsymbol{\gamma} \text{ satisfy } \mathbf{C}\}$. Then

$$\forall \boldsymbol{\gamma}, \boldsymbol{\gamma}^\star := \operatorname{argmin}(\mathrm{d}^2_{\mathrm{NED}}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^\star)) \quad \text{s.t.} \boldsymbol{\gamma}^\star \in \Omega \tag{2.15}$$

Figure 2.2 illustrates how a draw of $\boldsymbol{\gamma}$ is transformed to a new $\boldsymbol{\gamma}^\star$. Next, we are going to apply this Bayesian transformation approach to our problem of interest. Suppose the draws from standard Bayesian linear regression on $\mathbf{Y}$ against $\mathbf{X}$, $\mathbf{B}$ are $\gamma_0, \ldots, \gamma_p, \gamma_{p+1}$ and the draws from standard Bayesian linear regression on $\mathbf{B}$ against $\mathbf{X}$ are $\theta_0, \ldots, \theta_p$. We call these draws "raw draws", because these draws are for the unconstrained problem. The corresponding ordinary least squares estimates are $\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}}$ and estimated covariance matrices are $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$. We extract the estimated variances of the coefficients from $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$, denote them by $s^2_{\gamma_0}, \ldots, s^2_{\gamma_p}, s^2_{\gamma_{p+1}}, s^2_{\theta_0}, \ldots, s^2_{\theta_p}$. Then $\boldsymbol{\gamma}^\star, \boldsymbol{\theta}^\star$ are obtained from the unconstrained draws $\gamma_0, \ldots, \gamma_p, \gamma_{p+1}$, $\theta_0, \ldots, \theta_p$ by solving the following optimization problem:

$$\begin{aligned}
\min_{\gamma_0^\star, \ldots, \gamma_p^\star, \gamma_{p+1}^\star, \theta_0^\star, \ldots, \theta_P^\star} \quad & [\mathrm{d}^2_{\mathrm{NED}}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^\star) + \mathrm{d}^2_{\mathrm{NED}}(\boldsymbol{\theta}, \boldsymbol{\theta}^\star)] = \sum_{\mathrm{j}=0}^{\mathrm{p}+1} \frac{(\gamma_{\mathrm{j}} - \gamma_{\mathrm{j}}^\star)^2}{\mathrm{s}^2_{\gamma_{\mathrm{j}}}} + \sum_{\mathrm{k}=0}^{\mathrm{p}} \frac{(\theta_{\mathrm{k}} - \theta_{\mathrm{k}}^\star)^2}{\mathrm{s}^2_{\theta_{\mathrm{k}}}} \\
\text{s.t.} \quad & \gamma_0^\star + \gamma_{p+1}^\star \theta_0^\star \in [\bar{\beta}_0 - d\bar{S}_0, \bar{\beta}_0 + d\bar{S}_0] \\
& \cdots \\
& \gamma_p^\star + \gamma_{p+1}^\star \theta_p^\star \in [\bar{\beta}_p - d\bar{S}_p, \bar{\beta}_p + d\bar{S}_p] \\
& \gamma_j^\star \in [\hat{\gamma}_j - 5s_{\gamma_j}, \hat{\gamma}_j + 5s_{\gamma_j}], j = 0, \ldots, p+1 \\
& \theta_k^\star \in [\hat{\theta}_k - 5s_{\theta_k}, \hat{\theta}_k + 5s_{\theta_k}], k = 0, \ldots, p
\end{aligned} \tag{2.16}$$

where the last two constraints are trivial bounds for each parameter for improving computational efficiency. The scale parameter d controls the degree of trust in the historical information as before. d is drawn from a half normal distribution to reflect the reality that there is uncertainty in the point estimates. Through simulations, we find that the choice of $|\mathrm{N}(0, 1)|$ is reasonable.

Table 2.1: A summary of necessary theoretical assumptions required in constructing the relationship equations

Distribution	Internal dataset	External dataset
$\mathrm{E}(\mathrm{Y}\|\mathrm{X},\mathrm{B})$	Linear in (X, B); Correctly specified	Linear in (X, B); Correctly specified
	$\mathrm{E}(\mathrm{Y}\|\mathrm{X},\mathrm{B})$ is the same in the two datasets	
$\mathrm{E}(\mathrm{Y}\|\mathrm{X})$	Linear in X; Correctly specified	Linear in X; Correctly specified
	$\mathrm{E}(\mathrm{Y}\|\mathrm{X})$ is the same in the two datasets	
$\mathrm{E}(\mathrm{B}\|\mathrm{X})$	Linear in X Correctly specified	Linear in X Correctly specified
	$\mathrm{E}(\mathrm{B}\|\mathrm{X})$ is the same in the two datasets	
$(\mathrm{Y},\mathrm{X},\mathrm{B})$	No assumptions about the joint distribution of (Y, X, B) in either dataset other than the assumption on $\mathrm{E}(\mathrm{Y}\|\mathrm{X},\mathrm{B})$	
(X,B)	No assumptions about the joint distribution of (X, B) in either dataset other than the assumption on $\mathrm{E}(\mathrm{B}\|\mathrm{X})$	
(X)	Distribution of (X) does not have to be the same in the two datasets	

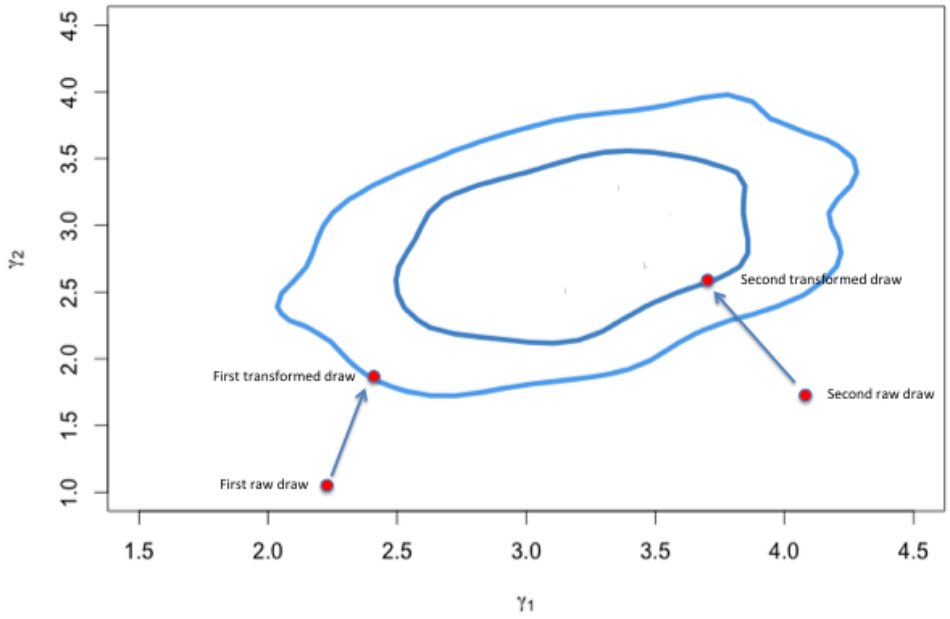


Figure 2.2: An illustration of how two constrained draws are obtained from the raw draws in the Bayesian transformation approach. For each raw draw, we generate a value of d from a half normal distribution $|\mathrm{N}(0,1)|$ thus the two ds are different for these two raw draws

The intuition behind these transformed draws generated by (2.16) is that it will produce values $\boldsymbol{\gamma}^\star, \boldsymbol{\theta}^\star$ subject to the box constraints that are closest to the unconstrained values $\boldsymbol{\gamma}, \boldsymbol{\theta}$ in normalized Euclidean distance. The squared normalized Euclidean distance measure (and its square root version) in the objective function in (2.16) minimizes the weighted Euclidean distance between $(\boldsymbol{\gamma}^\star, \boldsymbol{\theta}^\star)$ and $(\boldsymbol{\gamma}, \boldsymbol{\theta})$. The normalization is to ensure that the distance is relatively small for a particular coefficient if its OLS estimate is more precise (i.e., the estimated variance is relatively small) while the distance is relatively large for those coefficients that have more uncertainty by OLS.

We obtain posterior draws based on the following steps:

- Obtain raw draws: we first obtain draws $\gamma_0, \ldots, \gamma_p, \gamma_{p+1}$, $\theta_0, \ldots, \theta_p$ from standard Bayes. The draws from Standard Bayes are collected after a burn-in period and a thinning procedure
- Transformation of a single draw: for the first draw $\gamma_0^{(1)}, \ldots, \gamma_p^{(1)}, \gamma_{p+1}^{(1)}, \theta_0^{(1)}, \ldots, \theta_p^{(1)}$, generate $d^{(1)}$ from half normal distribution: $\mathrm{d} \sim |\mathrm{N}(0,1)|$. Perform the transformation procedure described in (2.16) and obtain the first posterior draw $\gamma_0^{\star(1)}, \ldots, \gamma_p^{\star(1)}, \gamma_{p+1}^{\star(1)}, \theta_0^{\star(1)}, \ldots, \theta_p^{\star(1)}$
- Iterate: repeat the above step for each draw until all draws are transformed. This means that the transformation procedure is performed 1000 times to obtain 1000 new draws.

Bayesian transformation approach is an ad hoc Bayes-type method that can be easily implemented by using draws from an unconstrained model and making them compatible with constraints derived from the external coefficient information by minimizing the normalized Euclidean distance of the raw draws from the constrained space. It is an approximate and somewhat ad hoc Bayes approach due to the fact that it does not give exact posterior inference as informative full Bayes does. Though this Bayes method does not conduct exact posterior inference, it is a pragmatic choice that was seen to both improve the estimation efficiency of the regression parameters and the predictive power in the regression model of primary interest, compared with standard Bayes without considering constraints. Getting raw draws from a standard Bayes approach is computationally efficient.

The ideal Gibbs procedure with constraints is highly inefficient due to rejection of many draws outside the constrained space. The transformation approach tries to take advantage of making draws from the unconstrained model and mapping them to satisfy the constraints. The proposed Bayesian transformation approach is similar in spirit to constrained MLE in that both of them involve a mapping of the parameters from the unconstrained space to the constrained space. The difference is that the constrained MLE maps a single unconstrained MLE to the constrained space while the Bayesian transformation approach requires a set of unconstrained draws to be converted to the constrained draws, thus providing a natural measure of uncertainty.

The computational efficiency of this approach will be discussed next. The objective function is a convex function and the constraints are a set of nonlinear inequality box constraints. They involve in total $2p + 3$ parameters and $p + 1$ constraints. We could simply rely on R function $\mathbf{solnp}$ in R package $\mathbf{Rsolnp}$ to solve this minimization problem. However, since this minimization procedure needs to be applied multiple times in order to obtain many posterior draws, simplifying the computation is desired. Moreover, with a growing number of predicting variables, this optimization will be of higher dimension and probably harder to solve.

To find an efficient algorithm for this minimization problem, we need to simplify this multiple-parameter minimization problem. We first notice that for these constraints, the nonlinearity is due to the fact that $\gamma^\star_{p+1}$ appears in every constraint. All other parameters appear in pairs $(\gamma^\star_j, \theta^\star_j), j = 0, \ldots, p$ and are only involved in one term within the sum in (2.16). Thus for each pair $(\gamma^\star_j, \theta^\star_j)$, the minimization function reduces to

$$\begin{aligned} \min_{\gamma^\star_j, \theta^\star_j} \quad & \frac{(\gamma_j - \gamma^\star_j)^2}{s^2_{\gamma_j}} + \frac{(\theta_j - \theta^\star_j)^2}{s^2_{\theta_j}} \\ \text{s.t.} \quad & \gamma^\star_j + \gamma^\star_{p+1}\theta^\star_j \in [\bar{\beta}_j - d\bar{S}_j, \bar{\beta}_j + d\bar{S}_j] \end{aligned} \tag{2.17}$$

In a two dimensional space, if we fix $\gamma^\star_{p+1}$, this optimization problem is trying to find coordinates $(\gamma^\star_j, \theta^\star_j)$ between two parallel lines $\gamma^\star_j + \gamma^\star_{p+1}\theta^\star_j = \bar{\beta}_j - d\bar{S}_j$ and $\gamma^\star_j + \gamma^\star_{p+1}\theta^\star_j = \bar{\beta}_j + d\bar{S}_j$

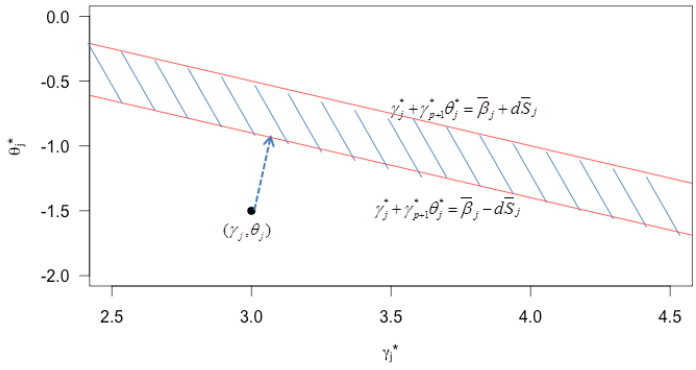


Figure 2.3: Two-dimensional optimization problem

such that it is closest to a point (γ_j, θ_j). We can further translate this problem into a problem that drops a perpendicular from the point with coordinates (γ_j, θ_j) to the line with equation $\gamma_j + \gamma_{p+1}^\star \theta_j = \bar{\beta}_j - d\bar{S}_j$ and another perpendicular to the line with equation $\gamma_j + \gamma_{p+1}^\star \theta_j = \bar{\beta}_j + d\bar{S}_j$ and the foot of each of these two perpendiculars can be easily found. Figure 2.3 illustrates this two-dimensional optimization problem.

As a result, by fixing $\gamma_{p+1}^\star$, the minimization in (2.16) can be divided into $p+1$ two dimensional minimization problems and is solved analytically by re-expressing the solution $\gamma_j^\star, \theta_j^\star$ as functions of $\gamma_{p+1}^\star$. After that, the entire minimization problem is reduced to a simple one-dimensional optimization problem in $\gamma_{p+1}^\star$, which can be easily solved using a one-dimensional optimization method. This iterative conditional optimization procedure is very fast. With $p = 5$ and 500 datasets, on a Mac laptop with 1.6 GHz processor, the computational time of the direct regression is about 2 seconds, 241 seconds for constrained ML, 4 seconds for the partial regression, 235 sec-

onds for standard Bayes, 13865 seconds for informative full Bayes and 775 seconds for Bayesian transformation approach.

2.3 Simulation Study

We present two simulation scenarios. In the first simulation scenario the estimates of $\boldsymbol{\beta}$ and their standard errors are provided from an analysis of a large dataset of size 2000. The current dataset from which to estimate models $\mathrm{E}(\mathbf{Y}|\mathbf{X},\mathbf{B})$ and $\mathrm{E}(\mathbf{B}|\mathbf{X})$ is of size 15 (simulation studies with larger sample sizes are presented in Appendix F). 500 datasets are generated. The first simulation study generates data from a true model of the form $\mathrm{Y_i} = \mu(\mathbf{X_i}, \mathrm{B_i}) + \epsilon_\mathrm{i}$, where $\epsilon_\mathrm{i} \sim \mathrm{N}(0, 6.5^2), \mathrm{i} = 1, \ldots, 15$. $\mu(\mathbf{X},\mathbf{B}) = 4 + 3\mathbf{X_1} + 3\mathbf{X_2} + 2\mathbf{B}$. $\mathbf{X_1}, \mathbf{X_2} \overset{\text{i.i.d}}{\sim} \mathrm{N}(0, 1^2)$ and $\mathbf{B}$ is simulated as $\mathbf{B} = 0.8\mathbf{X_1} + 0.8\mathbf{X_2} + N(0, 1.5^2)$. A linear regression based on the large dataset gives estimates for model $\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \beta_0 + \beta_1\mathbf{X_1} + \beta_2\mathbf{X_2}$. The estimates and standard errors from this fit are $\bar{\beta}_0 = 4, \bar{\mathrm{S}}_0 = 0.16, \bar{\beta}_1 = 4.6, \bar{\mathrm{S}}_1 = 0.16, \bar{\beta}_2 = 4.6, \bar{\mathrm{S}}_2 = 0.17$.

In the second simulation scenario, the dataset from which to estimate models $\mathrm{E}(\mathbf{Y}|\mathbf{X},\mathbf{B})$ and $\mathrm{E}(\mathbf{B}|\mathbf{X})$ is of size 20. 500 datasets are generated. We generate data from a model with a larger number of covariates, $\mathrm{Y_i} = \mu(\mathbf{X_i}, \mathrm{B_i}) + \epsilon_\mathrm{i}$, where $\epsilon_i \sim N(0, 6^2), i = 1, \ldots, 20$. $\mu(\mathbf{X},\mathbf{B}) = 4 + 3\mathbf{X_1} + 3\mathbf{X_2} + 2\mathbf{X_3} + 2\mathbf{X_4} + 2\mathbf{B}$. $\mathbf{X_1}, \mathbf{X_2} \overset{\text{i.i.d}}{\sim} N(0, 1^2)$ and $\mathbf{X_3}, \mathbf{X_4} \overset{\text{i.i.d}}{\sim} N(0, 1.5^2)$ and $\mathbf{B}$ is simulated as $\mathbf{B} = 0.8\mathbf{X_1} + 0.8\mathbf{X_2} + N(0, 1.5^2)$. A linear regression based on a large dataset of 2000 subjects gives estimates for model $\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \beta_0 + \beta_1\mathbf{X_1} + \beta_2\mathbf{X_2} + \beta_3\mathbf{X_3} + \beta_4\mathbf{X_4}$. The estimates and SE's are $\bar{\beta}_0 = 4, \bar{\mathrm{S}}_0 = 0.15, \bar{\beta}_1 = 4.6, \bar{\mathrm{S}}_1 = 0.15, \bar{\beta}_2 = 4.6, \bar{\mathrm{S}}_2 = 0.15, \bar{\beta}_3 = 2.1, \bar{\mathrm{S}}_3 = 0.10, \bar{\beta}_4 = 1.9, \bar{\mathrm{S}}_4 = 0.10$.

For comparing coefficient estimation, we report four quantities: the average of estimated coefficient, relative efficiency of estimated coefficient, mean squared error and the average of the estimated standard error of the coefficient across 500 replicates. The average of estimated coefficient is defined as: $\bar{\gamma}_\mathrm{j} = \frac{1}{500}\sum_{\mathrm{m}=1}^{500}\hat{\gamma}_{\mathrm{m,j}}, \mathrm{j} = 1, \ldots, \mathrm{p}+1$; the relative efficiency of esti-

mated coefficient is defined as: $\mathrm{V}(\hat{\gamma}_{\mathrm{j,direct}})/\mathrm{V}(\hat{\gamma}_{\mathrm{j,method}})$, where $\mathrm{V}(\hat{\gamma}_{\mathrm{j}})$ is the Monte Carlo variance $\frac{1}{500}\sum_{m=1}^{500}(\hat{\gamma}_{m,j}-\bar{\gamma}_j)^2$ and the MSE of an estimated coefficient is defined as: $\frac{1}{500}\sum_{\mathrm{m=1}}^{500}(\hat{\gamma}_{\mathrm{m,j}}-\gamma_{\mathrm{j}})^2$. We report the average of estimated standard errors of the estimates across 500 datasets (i.e., $\frac{1}{500}\sum_{\mathrm{m=1}}^{500}\sqrt{\hat{V}(\gamma_{m,j})}$) for each method: the average of the OLS estimated standard error for the direct regression estimates, the average of the asymptotic standard error for the partial regression estimates, the average of the bootstrap standard error for the constrained ML estimates and the average posterior standard deviation for each of the three Bayes estimates. For the constrained ML solution, we also provide the bootstrap bias-corrected estimate.

For comparing prediction power across different methods, we calculate the average out-of-bag (OOB) R^2 for prediction error in a validation dataset of size 100: OOB $\mathrm{R}^2 = 1 - \sum_{\mathrm{i=1}}^{100}(\mathrm{Y_i} - \sum_{\mathrm{j=0}}^{\mathrm{p}}\hat{\gamma}_{\mathrm{j}}\mathrm{X_{ij}} - \hat{\gamma}_{\mathrm{p+1}}\mathrm{B_i})^2/\sum_{\mathrm{i=1}}^{100}(\mathrm{Y_i}-\overline{\mathrm{Y}})^2$.

Table 2.2 summarizes the simulation results for three-covariate simulation scenario. The bootstrap corrected constrained ML, partial regression and informative full Bayes give estimates of the regression coefficients with low bias. The OOB R^2 of $\mathbf{Y}$ regressed on $\mathbf{X_1}, \mathbf{X_2}$ is low. By looking at relative efficiency of regression parameters γ_1 and γ_2, we find the constrained methods greatly improve the estimation efficiency of coefficients of $\mathbf{X}$. For γ_1 and γ_2, the partial regression, informative full Bayes and Bayesian transformation approach reduce the MSE by more than 50%. As expected, for γ_3, these constrained estimates do not improve systematically from the unconstrained estimates in terms of MSE, as the historical data provides no information on the additional predicting variable **B**. This finding agrees with the conclusion in Qin et al. (2015) that there is large improvement in the coefficients of X but not in the coefficient of B. Among all the methods, the two constrained Bayesian methods, informative full Bayes and Bayesian transformation approach have highest prediction power. They lead to an increase of 41% and 36% in terms of OOB R^2 respectively, compared to direct regression.

Table 2.3 summarizes the simulation results for the five-covariate simulation scenario. The

Table 2.2: Simulation results of three-covariate scenario: Comparison of different methods. OOB R^2 denotes average out-of-bag prediction ability. For each method, the first row includes mean (relative efficiency w.r.t. direct regression) of each regression coefficient and OOB R^2 of this method. The second row shows the MSE of each coefficient and the third row is the average of the standard error across 500 datasets. For constrained ML, we also report a bootstrap bias-corrected constrained ML estimate. A linear regression on $\mathbf{Y}$ on $\mathbf{X_1}, \mathbf{X_2}$ has an OOB R^2 of 0.212

Method	γ_1	γ_2	γ_3	OOBR^2
True value	3	3	2	
Direct regression	3.25(1.00)	3.07(1.00)	1.96(1.00)	0.270
MSE	5.21	5.90	1.92	
Avg.SE	2.20	2.23	1.31	
Constrained ML	2.82(1.59)	2.79(2.20)	2.27(0.80)	0.334
MSE	3.27	2.74	2.46	
Avg.Boot.SE	1.93	2.31	3.06	
Constrained $\mathrm{ML}_{\mathrm{bc}}$	3.01(1.40)	3.01(2.59)	2.00(0.88)	0.346
Partial regression	3.03(2.26)	3.01(2.62)	1.96(1.00)	0.346
MSE	2.29	2.25	1.92	
Avg.Asy.SE	1.58	1.56	1.34	
Standard Bayes	3.24(1.01)	3.06(1.00)	1.97(1.00)	0.270
MSE	5.24	5.93	1.93	
Avg.PSD	2.43	2.46	1.44	
Informative full Bayes	3.06(2.63)	2.99(2.97)	1.98(1.11)	0.382
MSE	1.97	1.98	1.74	
Avg.PSD	1.45	1.48	1.30	
Transformation	3.16(2.14)	3.09(2.31)	1.84(0.88)	0.366
MSE	2.40	2.56	2.22	
Avg.PSD	1.74	1.78	1.65	

OOB R^2 of $\mathbf{Y}$ regressed on $\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_4}$ is moderate. By looking at relative efficiency of regression parameters $\gamma_1, \gamma_2, \gamma_3, \gamma_4$, this table again tells us that, the constrained methods could substantially improve the estimation efficiency of coefficients of $\mathbf{X}$. In fact, the efficiency of coefficient γ_3 and that of γ_4 triple for the partial regression, informative full Bayes and the Bayesian transformation approach compared to direct regression. The informative full Bayes and the Bayesian transformation approach have the highest prediction power, as measured by OOB R^2. They increase by 25% and 23% the OOB R^2 respectively, compared to direct regression. The constrained ML solution also increases by 23% the OOB R^2 when the bootstrap bias-correction procedure is applied.

We also conduct additional simulation studies with different sample sizes for the three-covariate scenario and for the five-covariate scenario. The simulation results are shown in Appendix F. As expected, when the sample size increases, the gain in estimating efficiency and predictive power by incorporating the external information is not as significant as it is in small sample sizes settings.

2.4 Application to the Normative Aging Study

We illustrate our methodology by enhancing a published prediction model for bone lead levels in terms of blood lead and other covariates (Park et al., 2009), with a new biomarker defined through a continuous genetic risk score ($\mathbf{B}$, in terms of the notation used in previous sections). It is known that up to 95% of the total body burden of lead is accumulated in the skeleton (Barry and Mossman, 1970). While blood lead is widely used as a biomarker of recent lead exposure due to the convenience of collecting blood samples, its short half-life ($\sim$30 days) limits its utility in chronic disease epidemiology research. Therefore, bone lead reflects cumulative lead exposure and is considered a better biomarker when examining chronic diseases. Recent development of K X-ray fluorescence instruments makes it possible to take direct measurements of bone lead concentrations (Hu et al., 2007). However, K X-ray fluorescence measurements can only be taken

Table 2.3: Simulation results of five-covariate scenario: Comparison of different methods. OOB R^2 denotes average out-of-bag prediction ability. For each method, the first row includes mean (relative efficiency w.r.t. direct regression) of each regression coefficient and OOB R^2 of this method. The second row shows the MSE of each coefficient and the third row is the average of the standard error across 500 datasets. For constrained ML, we also report a bootstrap bias-corrected constrained ML estimate. A linear regression on $\mathbf{Y}$ on $\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_4}$ has an OOB R^2 of 0.350

Method	γ_1	γ_2	γ_3	γ_4	γ_5	OOB R^2
True value	3	3	2	2	2	
Direct regression	3.11(1.00)	3.02(1.00)	1.93(1.00)	2.04(1.00)	1.92(1.00)	0.421
MSE	3.54	4.28	1.32	1.33	1.22	
Avg.SE	1.84	1.85	1.08	1.08	1.07	
Constrained ML	2.77(1.91)	2.66(1.67)	2.03(2.70)	1.94(2.35)	2.33(0.76)	0.492
MSE	1.91	2.67	0.49	0.56	1.69	
Avg.Boot.SE	1.54	1.59	0.74	0.73	1.76	
Constrained $\mathrm{ML}_{\mathrm{bc}}$	3.07(2.50)	2.97(2.19)	2.04(3.67)	1.94(3.13)	1.95(0.98)	0.519
Partial regression	3.08(2.41)	3.01(2.28)	2.09(3.23)	1.91(3.13)	1.92(1.00)	0.500
MSE	1.48	1.89	0.42	0.43	1.22	
Avg.Asy.SE	1.28	1.29	0.62	0.61	1.11	
Standard Bayes	3.11(1.00)	3.02(1.00)	1.93(1.00)	2.04(1.00)	1.92(1.00)	0.421
MSE	3.55	4.28	1.32	1.33	1.22	
Avg.PSD	1.98	2.00	1.17	1.16	1.16	
Informative full Bayes	3.02(2.92)	2.93(2.54)	2.05(3.93)	1.92(3.33)	2.00(1.26)	0.526
MSE	1.24	1.72	0.34	0.40	0.96	
Avg.PSD	1.14	1.17	0.56	0.56	1.01	
Transformation	3.16(2.34)	3.08(2.07)	2.04(3.44)	1.95(3.23)	1.81(0.81)	0.516
MSE	1.52	2.09	0.38	0.41	1.51	
Avg.PSD	1.42	1.46	0.67	0.66	1.41	

at very few locations in the entire country and thus direct bone lead level measurements are not commonly available. Thus, prediction models of bone lead level were constructed in terms of blood lead levels and other covariates that are more readily available in other studies.

Our study utilizes data from the Normative Aging Study, a longitudinal study established by the Veterans Administration in 1961. The Normative Aging Study enrolled 2280 men, aged 21 to 80, living in the Greater Boston area. Participants were recruited to represent a range of socioeconomic characteristics in terms of education and occupation (Bell et al., 1972). Every 3 to 5 years, participants returned for follow-up visits and information about age, smoking, education level, disease status, medication use, physical activity and dietary intake was recorded. Beginning in 1991, K X-ray fluorescence was used to measure bone lead levels of participants at two sites: tibia (representing cortical bone) and patella (representing trabecular bone).

Park et al. (2009) developed a prediction model for tibia lead level using blood lead levels, age, smoking status, pack-years of cigarette, education and occupation based on 550 participants of the Normative Aging Study. These 6 predictors were selected because they could be routinely collected in epidemiological studies. Table 2.4 shows the estimated tibia lead prediction model presented in their paper.

Park et al. (2009) commented that bone lead levels differ by genetic make-up and including some relevant genetic polymorphisms to the existing model may provide improved prediction accuracy. We want to use the published tibia lead prediction model as external information and see if genes in the lead toxicokinetics and toxicodynamics pathway can enhance the prediction power. We use the data from the Normative Aging study that not only has bone lead levels and these six covariates but also has 19 single-nucleotide polymorphisms (SNPs) relevant to the lead toxicokinetics and toxicodynamics pathway. We would like to include in the model a composite genetic risk score based on the unweighted sum of the risk allele counts of the 19 SNPs.

We exclude those individuals that have missingness in more than 3 SNPs. The remaining

missing values in each genotype are imputed by the average number of risk alleles. The composite genetic risk score is then constructed as the summation of the risk allele counts of these 19 relevant SNPs. The genetic risk score is noted to be roughly normally distributed. We also remove those individuals who have missing values in any of the 6 predictors or the response tibia lead level. Our dataset then consists of 156 observations, including first measurements of 100 participants used as a training dataset and follow-up measurements of another 56 participants used as a testing dataset. These two datasets are independent of each other and our training dataset is independent of the original training data of Park et al. (2009).

Table 2.8 in Appendix D shows the characteristics of the training dataset and the characteristics of the testing dataset. There are no significant differences in variables age, pack-years of cigarette, genetic score, blood lead in these two datasets. There are apparent differences in variables smoking status and education. We would like to build the expanded tibia lead prediction model using our training dataset and the testing dataset will be used for validating our model. The expanded tibia lead prediction model will be estimated by both the unconstrained methods and the constrained methods we described in Section 2.2.

For comparing coefficient estimation across different methods, we report the estimated coefficients and their standard errors. For comparing prediction power, we calculate R^2 in the training dataset and OOB R^2 in the testing dataset. We also estimate a six-predictor model without SNP information as in Park et al. (2009) in our training dataset and find out that the R^2 and OOB R^2 of this tibia lead model without SNP information are 0.42 and 0.16 respectively. The estimated coefficients are shown in Table 2.4.

Table 2.5 presents the expanded tibia lead prediction model fitted to the training dataset. We find that while the R^2 of this model does not increase much comparing to that of the prediction model without SNP information, the OOB R^2 increases 88% if we incorporate external information from Table 2.4 into our model estimating procedure. If we compare the standard errors across

different methods, it is easily seen that the constrained methods will reduce the standard errors of regression coefficients compared to direct regression. For the constrained ML, informative full Bayes and the Bayesian transformation approach, the standard errors of the parameters of the variables blood lead, age, education, white collar, pack-years of cigarette and smoking status decrease at least 50% comparing to the standard errors in direct regression. Meanwhile, partial regression estimates of parameters of variables blood lead, age, education, white collar, pack-years of cigarette and smoking status have more than 80% reduction in standard errors comparing to direct regression. Therefore, it could be easier to identify statistically significant predictors based on these constrained methods. The reason that the partial regression estimates have the smallest estimated standard errors among all constrained solutions is that the standard errors of the regression coefficients from the tibia lead prediction model in Park et al. (2009) have different scales. By using partial regression method, we assume that there is no variation in the estimated coefficients in this tibia lead model (we only plug in the point estimates and do not make use of the standard errors) while the other three constrained solutions take into account the precision in these estimates.

2.5 Discussion

In this study, we demonstrate how to incorporate external information on regression coefficients in linear regression model estimation and prediction. We formulate the problem in an inferential framework in which the historical information is translated into nonlinear inequality constraints for coefficients and propose four constrained solutions: constrained ML, partial regression, informative full Bayes and Bayesian transformation approach. We use simulation studies to assess the performance of these proposed methods and show that incorporating external information can improve the efficiency of model estimation and increase the prediction accuracy. The application to the Normative Aging Study shows that the estimation accuracy of regression coefficients and

the predictive power of a tibia lead prediction model that includes a composite genetic risk score as a new biomarker can be improved, when information from a previously published model with non-genetic data is incorporated.

Among the constrained solutions, our Bayesian transformation approach, motivated by Gunn and Dunson's transformation method, is a simple and effective computation method. The main underlying idea for the Bayesian transformation approach is to first obtain rapid draws from a simple sampling algorithm ignoring the constraints and then transform the draws by minimizing the squared normalized Euclidean distance between unconstrained draws and constrained draws, subject to these constraints.

One point of future consideration is the choice of tuning parameter. We have a quantity labeled d that controls the degree of trust in the historical information and we select it by drawing from a half normal distribution $|\mathrm{N}(0, \sigma_{\mathrm{d}}^2)|$ in Bayesian transformation approach and we fix $d = 1$ in the constrained ML approach. Though we fix $\sigma_d^2 = 1$ or $d = 1$, it can also be considered as a tuning parameter and adaptively selected for a particular dataset. However, how to select this tuning parameter in a principled optimal manner is not determined yet. In the supplementary material we show how the value of d will affect the constrained ML estimates. With bigger d, the standard errors of the regression coefficients will increase, and the predictive power in the validation dataset will decrease. When $d = 10$, these box constraints will be very weak and the estimated model based on the constrained ML is very similar to the estimated model based on direct regression, although the standard errors are smaller.

Another challenge is to demonstrate whether our Bayesian transformation approach provides a good approximation to the informative full Bayes solution. We claim to obtain approximate posterior draws from the Bayesian transformation approach without writing down the posterior distribution functions while informative full Bayes does give the exact posterior distribution. To validate that the Bayesian transformation approach is a good approximation to the true posterior,

we compare the posterior distribution to that obtained from informative full Bayes in simulation studies and find good correspondence (results not shown). But this empirical observation needs further justification.

Among the constrained solutions, the constrained maximum likelihood estimate method and the informative full Bayes approach both depend on the likelihood of $\mathbf{Y}|\mathbf{X},\mathbf{B}$ and $\mathbf{B}|\mathbf{X}$. In this manuscript we primarily discuss the situation when $\mathbf{Y}|\mathbf{X},\mathbf{B}$ and $\mathbf{B}|\mathbf{X}$ are both normal. However if one or both $\mathbf{Y}|\mathbf{X},\mathbf{B}$ or $\mathbf{B}|\mathbf{X}$ is not normal, but still continuous, the joint likelihood function can be modified and these two constrained solutions can be directly extended. In order to perform standard Bayesian inference to produce the initial raw draws, Bayesian transformation approach also depends on the likelihood function of $\mathbf{Y}|\mathbf{X},\mathbf{B}$ and $\mathbf{B}|\mathbf{X}$ which can again be modified to a non-normal likelihood if needed. The simulation study and the data analysis are based on a single continuous variable $\mathbf{B}$. However, these strategies to incorporating the external coefficient information can be extended if $\mathbf{B}$ is multivariate. The exact relationship between parameters shown in equation (2.6) can be extended to the case that $\mathbf{B}$ is multivariate normal with L dimensions:

$$\beta_j = \gamma_j + \sum_{l=1}^{L} \gamma_{p+l}\theta_{lj}, j = 0, \ldots, p \tag{2.18}$$

As a consequence, when $\mathbf{B}$ is multivariate normal, the constrained ML, partial regression, informative full Bayes and Bayesian transformation approach are still applicable.

We consider prediction models to predict a continuous outcome. In future work we will consider predictions of a binary outcome, which are also common, particularly in medical applications, where logistic regression models are frequently used for predicting the risk of a binary disease indicator.

2.6 Software

Software in the form of R code, together with a sample input data set and complete documentation is available on request from the corresponding author (chengwt@umich.edu).

2.7 Supplementary Material

2.7.1 *Appendix A*

Bootstrap estimate of the standard error for the constrained ML estimate

We would like to obtain a bootstrap estimate of the constrained ML estimator's standard error. Regression models can be bootstrapped by (1) treating the design matrix as random and selecting bootstrap samples directly from the observations or (2) treating the design matrix as fixed and resampling from the residuals of the fitted regression models (Fox, 2008). In our study, we implement a residual bootstrap as follows:

- Estimate the regression coefficients $\gamma_0, \ldots, \gamma_{p+1}$ and $\theta_0, \ldots, \theta_p$ by constrained ML method for the original sample. Estimate $\boldsymbol{\beta}$ by $\hat{\beta}_j = \hat{\gamma}_j + \hat{\gamma}_{p+1}\hat{\theta}_j, j = 0, \ldots, p$

- Calculate the fitted outcome pair $(\hat{\mathrm{Y}}_\mathrm{i}, \hat{\mathrm{B}}_\mathrm{i})$ and residual pair $\mathbf{E_i} = (\mathrm{E_{i,Y}}, \mathrm{E_{i,B}})$ for each observation: $\hat{\mathrm{Y}}_\mathrm{i} = \hat{\beta}_0 + \hat{\beta}_1\mathrm{X_{i1}} + \cdots + \hat{\beta}_\mathrm{p}\mathrm{X_{ip}}, \hat{\mathrm{B}}_\mathrm{i} = \hat{\theta}_0 + \hat{\theta}_1\mathrm{X_{i1}} + \cdots + \hat{\theta}_\mathrm{p}\mathrm{X_{ip}}$ and $\mathrm{E_{i,Y}} = \mathrm{Y_i} - \hat{\mathrm{Y}}_\mathrm{i}, \mathrm{E_{i,B}} = \mathrm{B_i} - \hat{\mathrm{B}}_\mathrm{i}$

- Take bootstrap samples of the residual pairs, $\tilde{\mathbf{e}}_\mathbf{b} = [\tilde{\mathbf{E}}_\mathbf{b1}, \ldots, \tilde{\mathbf{E}}_\mathbf{bn}]^\mathrm{T}, \mathrm{b} = 1, \ldots, \mathrm{S}$, calculate bootstrapped $\mathbf{Y}$ values $\tilde{\mathbf{Y}}_\mathbf{b} = [\tilde{Y}_{b1}, \ldots, \tilde{Y}_{bn}]^T$, where $\tilde{\mathrm{Y}}_\mathrm{bi} = \hat{\mathrm{Y}}_\mathrm{i} + \tilde{\mathrm{E}}_\mathrm{bi,Y}$, calculate bootstrapped $\mathbf{B}$ values $\tilde{\mathbf{B}}_\mathbf{b} = [\tilde{\mathrm{B}}_\mathrm{b1}, \ldots, \tilde{\mathrm{B}}_\mathrm{bn}]^\mathrm{T}$, where $\tilde{B}_{bi} = \hat{B}_i + \tilde{E}_{bi,B}$

- Regress $\tilde{\mathbf{Y}}_\mathbf{b}$ on the fixed $\mathbf{X}$ design matrix and bootstrap samples $\tilde{\mathbf{B}}_\mathbf{b}$ to obtain bootstrap estimates of regression coefficients by constrained ML: $\tilde{\gamma}_{b,0}, \ldots, \tilde{\gamma}_{b,p+1}$

- The $\tilde{\boldsymbol{\gamma}}_\boldsymbol{b}$ can be used to construct bootstrap standard error: $\tilde{\sigma}_{.,j} = (\frac{\sum_{b=1}^{S}(\tilde{\gamma}_{b,j} - \bar{\tilde{\gamma}}_{.,j})^2}{S-1})^{1/2}$, $j = 0, \ldots, p+1$, in the usual bootstrap manner as described in Efron and Tibshirani (1986).

For the partial regression estimate, the bootstrap estimate of the standard error can be obtained in a way similar to that described above.

2.7.2 *Appendix B*

Variance approximation for the partial regression estimate

To derive the variance of the partial regression estimate $(\hat{\gamma}_0, \ldots, \hat{\gamma}_{p+1})$, we adapt the variance approximation used in Mukherjee and Chatterjee (2008). Since the partial regression solution only uses the point estimates of $\boldsymbol{\beta}$ but not their standard errors as the external information, we treat $\bar{\boldsymbol{\beta}}$ as constants when deriving the variances of $\hat{\boldsymbol{\gamma}}$ and then the partial regression estimates are functions of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\alpha}}$ only. We can first derive the variance-covariance matrix of $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}$ and then apply delta method to produce the variances of partial regression estimate.

Let ${\mathbf{I}^{\theta}}_{(\mathrm{p+1})\times(\mathrm{p+1})}$ and ${\mathbf{I}^{\alpha}}_{2\times 2}$ denote the observed information matrices for linear regression model $\mathrm{E}(\mathbf{B}|\mathbf{X}) = \mathbf{X}\boldsymbol{\theta}$ and for model $\mathrm{E}(\mathbf{r_1}|\mathbf{r_2}) = \alpha_0 + \alpha_1\mathbf{r_2}$ respectively in the partial regression method. ${\mathbf{I}^{\theta}}_{(\mathrm{p+1})\times(\mathrm{p+1})} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ and ${\mathbf{I}^{\alpha}}_{(2)\times(2)} = \mathbf{V}^{\mathrm{T}}\mathbf{V}$ where $\mathbf{V} = (\mathbf{1}_{(\mathrm{n})\times(1)}, \mathbf{r_2}_{(\mathrm{n})\times(1)})$. Denote the true parameter values of $\boldsymbol{\theta}, \boldsymbol{\alpha}$ by $\boldsymbol{\theta_0}$ and $\boldsymbol{\alpha_0}$. Then asymptotically, $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0} = \sqrt{n}(\mathbf{I}^{\theta})^{-1}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathbf{U}_{\mathbf{i}}^{\theta} + \mathrm{o_p}(\mathrm{n}^{-1/2})$ and $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha_0} = \sqrt{n}(\mathbf{I}^{\alpha})^{-1}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathbf{U}_{\mathbf{i}}^{\alpha} + \mathrm{o_p}(\mathrm{n}^{-1/2})$ where $\mathbf{U}_{\mathbf{i}}^{\theta} = (\mathrm{Y_i} - \mathbf{X_i}\boldsymbol{\theta})\mathbf{X_i}$ and $\mathbf{U}_{\mathbf{i}}^{\alpha} = (\mathrm{r_{i1}} - \mathbf{V_i}\boldsymbol{\alpha})\mathbf{V_i}$ are subject i'th individual score functions for estimate $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\alpha}}$ respectively.

The asymptotic variance-covariance matrix of the vector $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}})^T$ can be represented as

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\alpha}} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} & (\mathbf{I}^{\theta})^{-1}\mathrm{Cov}(\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathbf{U}_{\mathbf{i}}^{\theta}, \sum_{\mathrm{i}=1}^{\mathrm{n}}\mathbf{U}_{\mathbf{i}}^{\alpha})(\mathbf{I}^{\alpha})^{-1\mathrm{T}} \\ (\mathbf{I}^{\alpha})^{-1}\mathrm{Cov}(\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathbf{U}_{\mathbf{i}}^{\alpha}, \sum_{\mathrm{i}=1}^{\mathrm{n}}\mathbf{U}_{\mathbf{i}}^{\theta})(\mathbf{I}^{\theta})^{-1\mathrm{T}} & \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}} \end{pmatrix}$$

where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}$ are the estimated covariance matrices by OLS and $\mathrm{Cov}(\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathbf{U}_{\mathbf{i}}^{\alpha}, \sum_{\mathrm{i}=1}^{\mathrm{n}}\mathbf{U}_{\mathbf{i}}^{\theta}) = \sum_{\mathrm{i}=1}^{\mathrm{n}}\mathbf{U}_{\mathbf{i}}^{\alpha}\mathbf{U}_{\mathbf{i}}^{\boldsymbol{\theta}\mathbf{T}}$.

Since $\hat{\gamma_0} = \hat{\alpha}_0 - \hat{\alpha}_1\hat{\theta}_0 + \bar{\beta}_0 = g(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}})$, by delta method, we have the approximate variance of $\hat{\gamma_0}$: $g'^{T}\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\alpha}}g'$. For other γs, we can obtain the approximate variance in a similar way.

We run a simulation study to evaluate the accuracy of the above two variance estimation methods for the regression coefficients for the three-covariate scenario and for the five-covariate scenario. The results are shown in Table 2.6 and Table 2.7.

2.7.3 *Appendix C*

Posterior distributions in informative full Bayes methods

The conditional distribution of $\beta_0, \ldots, \beta_p$ will be normal, each with distribution function $\mathrm{N}(\mu_{\beta_{\mathrm{j}},\mathrm{n}}, \sigma^2_{\beta_{\mathrm{j}},\mathrm{n}}), \mathrm{j} = 0, \ldots, \mathrm{p}$, where $\mu_{\beta_{\mathrm{j}},\mathrm{n}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{B_i}-\sum_{\mathrm{k}\neq\mathrm{j}}\frac{\beta_{\mathrm{k}}-\gamma_{\mathrm{k}}}{\gamma_{\mathrm{p}+1}}\mathrm{X_{ik}}+\frac{\gamma_{\mathrm{j}}}{\gamma_{\mathrm{p}+1}}\mathrm{X_{ij}})\mathrm{X_{ij}}\frac{\bar{\mathrm{S}}_{\mathrm{j}}^2}{\gamma_{\mathrm{p}+1}}+\bar{\beta}_{\mathrm{j}}\sigma_2^2}{\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{X_{ij}^2}\bar{\mathrm{S}}_{\mathrm{j}}^2}{\gamma_{\mathrm{p}+1}^2}+\sigma_2^2}$, $\sigma^2_{\beta_j,n} = \frac{\sigma_2^2\bar{S}_j^2}{\frac{\sum_{i=1}^{n}X_{ij}^2\bar{S}_j^2}{\gamma_{p+1}^2}+\sigma_2^2}$.

The conditional distribution of $\gamma_0, \ldots, \gamma_p$ will be normal, each with distribution function $\mathrm{N}(\mu_{\gamma_{\mathrm{j}},\mathrm{n}}, \sigma^2_{\gamma_{\mathrm{j}},\mathrm{n}}), \mathrm{j} = 0, \ldots, \mathrm{p}$,

where $\mu_{\gamma_{\mathrm{j}},\mathrm{n}} = \frac{[\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{Y_i}-\sum_{\mathrm{k}\neq\mathrm{j}}\gamma_{\mathrm{k}}\mathrm{X_{ik}}-\gamma_{\mathrm{p}+1}\mathrm{B_i})\mathrm{X_{ij}}\sigma_2^2-\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{B_i}-\sum_{\mathrm{k}\neq\mathrm{j}}\frac{\beta_{\mathrm{k}}-\gamma_{\mathrm{k}}}{\gamma_{\mathrm{p}+1}}\mathrm{X_{ik}}-\frac{\beta_{\mathrm{j}}\mathrm{X_{ij}}}{\gamma_{\mathrm{p}+1}})\mathrm{X_{ij}}\frac{\sigma_1^2}{\gamma_{\mathrm{p}+1}}]\times 100^2}{(\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{X_{ij}^2}\sigma_2^2+\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{x_{ij}^2}}{\gamma_{\mathrm{p}+1}^2}\sigma_1^2)100^2+\sigma_1^2\sigma_2^2}$,

and $\sigma^2_{\gamma_{\mathrm{j}},\mathrm{n}} = \frac{\sigma_1^2\sigma_2^2 100^2}{(\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{X_{ij}^2}\sigma_2^2+\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{X_{i2}^2}}{\gamma_{\mathrm{p}+1}^2}\sigma_1^2)100^2+\sigma_1^2\sigma_2^2}$.

The conditional distribution of σ_1^2, σ_2^2 will be inverse-gamma. The full conditional distribution for σ_1^2 is $\mathrm{inverse}-\mathrm{gamma}(\frac{\nu_{1,\mathrm{n}}}{2}, \frac{\nu_{1,\mathrm{n}}\sigma^2_{1,\mathrm{n}}}{2})$ where $\nu_{1,n} = \nu_0 + n$ and $\sigma^2_{1,\mathrm{n}} = \frac{1}{\nu_{1,\mathrm{n}}}[\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{Y_i} - \sum_{\mathrm{j}=0}^{\mathrm{p}}\gamma_{\mathrm{j}}\mathrm{X_{ij}}-\gamma_{\mathrm{p}+1}\mathrm{B_i})^2+\nu_0\sigma_0^2]$. The full conditional distribution for σ_2^2 is $\mathrm{inverse}-\mathrm{gamma}(\frac{\nu_{2,\mathrm{n}}}{2}, \frac{\nu_{2,\mathrm{n}}\sigma^2_{2,\mathrm{n}}}{2})$ where $\nu_{2,n} = \nu_0 + n$ and $\sigma^2_{2,\mathrm{n}} = \frac{1}{\nu_{2,\mathrm{n}}}[\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{B_i} - \sum_{\mathrm{j}=0}^{\mathrm{p}}\frac{\beta_{\mathrm{j}}-\gamma_{\mathrm{j}}}{\gamma_{\mathrm{p}+1}}\mathrm{X_{ij}})^2 + \nu_0\sigma_0^2]$.

The full conditional distribution of γ_{p+1} is: $\mathrm{p}(\gamma_{\mathrm{p}+1}|\mathbf{Y}, \mathbf{X}, \mathbf{B}, \beta_0, \ldots, \beta_{\mathrm{p}}, \gamma_0, ..., \gamma_{\mathrm{p}}, \sigma_1^2, \sigma_2^2)$
$\propto \left\{\prod_{\mathrm{i}=1}^{\mathrm{n}}\frac{1}{\sqrt{2\pi\sigma_1^2}}\mathrm{e}^{-\frac{1}{2\sigma_1^2}(\mathrm{Y_i}-\sum_{\mathrm{j}=0}^{\mathrm{p}}\gamma_{\mathrm{j}}\mathrm{X_{ij}}-\gamma_{\mathrm{p}+1}\mathrm{B_i})^2} \times \frac{1}{\sqrt{2\pi\sigma_2^2}}\mathrm{e}^{-\frac{1}{2\sigma_2^2}(\mathrm{B_i}-\sum_{\mathrm{j}=0}^{\mathrm{p}}\frac{\beta_{\mathrm{j}}-\gamma_{\mathrm{j}}}{\gamma_{\mathrm{p}+1}}\mathrm{X_{ij}})^2}\right\} \cdot \pi(\gamma_{p+1}) \cdot |\mathbf{J}|$, which does not have a closed form.

2.7.4 *Appendix D*

Characteristics of the training dataset and the characteristics of the testing dataset

The characteristics of the training dataset and the characteristics of the testing dataset are shown in Table 2.8.

2.7.5 *Appendix E*

Sensitivity analysis of the choice of d for the constrained ML in the expanded tibia lead prediction model

In the nonlinear constraints, d plays an important role. d represents our belief in the strength

of the external/historical information. In the manuscript we suggested that fix d as d = 1. The choice $d = 0$ assumes that there is no uncertainty in the regression coefficient estimates in the established historical data model. We evaluate the sensitivity of the choice of d in the bone lead data example. The results are summarized in Table 2.9. It shows that the choice of d will influence both the estimation efficiency of the regression coefficients and the predictive power of the model. With bigger d, the standard errors of the regression coefficients will increase, the predictive power in the validation dataset will decrease. When $d = 10$, these box constraints are observed to be weak and the estimated model based on the constrained methods are more similar to the estimated model based on direct regression on just the internal data. The small differences between the standard errors for $d = 10$ case and those from direct regression are due to the differences in the bootstrap procedure applied when $d = 10$ and information matrix based standard errors for the direct regression method.

2.7.6 *Appendix F*

Additional simulations with different sample sizes

We conducted additional simulations for the case that n = 50; 100; 200; 2000 for the three-covariate simulation scenario and the five-covariate simulation scenario. The results are summarized in Table 2.10 and Table 2.11.

Table 2.4: Regression coefficients externally imported from the tibia lead prediction model (n=550) in Park et al. (2009) and regression coefficients of this tibia prediction model estimated based on our training dataset (n=100)

	Tibia lead prediction model (n=550)			Our training dataset (n=100)		
Variable	β	SE	P	β	SE	P
Intercept	-20.27	5.34	0.0002	-22.20	9.80	0.03
Bone lead	1.03	0.13	< 0.0001	0.93	0.30	0.002
Age	0.59	0.07	< 0.0001	0.55	0.13	< 0.0001
Education						
High school diploma	-3.65	1.75	0.04	3.21	3.42	0.35
≥ 4 yr of college	-7.05	2.09	0.0008	0.02	4.05	1.00
White collar	-3.21	1.18	0.01	-4.46	2.40	0.07
Cumulative cigarette smoking (pack-yr)	0.04	0.03	0.17	0.22	0.06	0.0005
Smoking status						
Former smoker	1.80	1.34	0.18	-1.96	2.52	0.44
Current smoker	0.05	2.48	0.98	-19.22	5.43	0.0006
R^2	0.27			0.42		

Table 2.5: Regression coefficients of the expanded tibia lead prediction model with the genetic score (n = 100). * denotes standard error and ** denotes 95% confidence interval

Variable	Direct regression	Standard Bayes	Constrained ML	Partial regression	Informative full Bayes	Bayesian transformation approach	Constrained $\mathrm{ML}_{\mathrm{bc}}$
Intercept	-23.85(11.27)*	-23.40(11.53)	-27.47(7.27)	-21.77(5.83)	-25.54(5.23)	-20.56(8.82)	-30.98$(-46.71, -16.80$
Blood lead	0.95(0.30)	0.95(0.31)	0.93(0.13)	1.05(0.06)	1.06(0.13)	1.00(0.14)	0.85$(0.71, 1.17)$
Age	0.55(0.13)	0.55(0.14)	0.64(0.06)	0.59(0.01)	0.60(0.05)	0.58(0.06)	0.68$(0.52, 0.72)$
Education							
High school diploma	3.19(3.44)	3.11(3.44)	-1.92(1.42)	-3.68(0.15)	-3.13(1.40)	-2.05(1.24)	-0.47$(-3.47, 0.33)$
$\geq$ 4 yr of college	-0.06(4.08)	-0.11(4.12)	-5.04(1.60)	-7.13(0.33)	-6.62(1.64)	-5.22(1.56)	-3.66$(-7.40, -2.62)$
White collar	-4.48(2.42)	-4.48(2.45)	-2.84(1.02)	-3.24(0.11)	-3.07(1.03)	-3.61(0.98)	-2.77$(-4.98, -1.77)$
Cumulative cigarette smoking	0.22(0.06)	0.22(0.06)	0.06(0.02)	0.03(0.005)	0.04(0.02)	0.06(0.02)	0.08$(0.02, 0.10)$
Smoking status							
Former smoker	-1.92(2.54)	-1.95(2.56)	2.89(0.98)	1.84(0.19)	2.27(1.20)	0.70(1.00)	3.25$(0.29, 3.74)$
Current smoker	-19.07(5.48)	-19.16(5.60)	-2.28(1.90)	0.20(0.61)	-2.44(2.17)	0.19(2.54)	-3.51$(-4.65, 1.90)$
Genetic risk score	0.13(0.44)	0.12(0.45)	0.15(0.48)	0.13(0.45)	0.36(0.31)	-0.10(0.67)	0.13$(-0.76, 1.10)$
R^2	0.42	0.42	0.36	0.32	0.35	0.30	0.37
OOB R^2	0.17	0.17	0.29	0.29	0.30	0.28	0.28

Table 2.6: Simulation results of the three-covariate scenario: for both the constrained ML and the partial regression, we report the ratio of average bootstrap mean and Monte Carlo mean $((\frac{1}{500}\sum_{\mathrm{m}=1}^{500}\bar{\tilde{\gamma}}_{\mathrm{m,j}})/(\frac{1}{500}\sum_{\mathrm{m}=1}^{500}\hat{\gamma}_{\mathrm{m,j}}))$ and the ratio of average bootstrap standard error and Monte Carlo standard deviation $((\frac{1}{500}\sum_{\mathrm{m}=1}^{500}\tilde{\sigma}_{\mathrm{m,j}})/\sqrt{V(\hat{\gamma}_j)})$ of each regression coefficient. For the partial regression solution, we also report the ratio of average asymptotic standard error and Monte Carlo standard deviation $(\frac{1}{500}\sum_{\mathrm{m}=1}^{500}\mathrm{Asy.SE}(\gamma_{\mathrm{m,j}})/\sqrt{V(\hat{\gamma}_j)})$

Sample Size	Method	Ratio	γ_1	γ_2	γ_3
$n = 15$	Constrained ML	Avg.boot.Mean/MC.Mean	0.93	0.92	1.12
		Avg.Boot.SE/MC.SD	1.07	1.41	1.97
	Partial regression	Avg.boot.Mean/MC.Mean	1.00	1.00	1.01
		Avg.Boot.SE/MC.SD	1.14	1.15	1.10
		Avg.Asy.SE/MC.SD	1.05	1.04	0.96
$n = 30$	Constrained ML	Avg.boot.Mean/MC.Mean	0.96	0.96	1.06
		Avg.Boot.SE/MC.SD	0.95	1.03	1.09
	Partial regression	Avg.boot.Mean/MC.Mean	1.00	1.00	1.00
		Avg.Boot.SE/MC.SD	0.94	1.08	1.03
		Avg.Asy.SE/MC.SD	0.91	1.03	1.00
$n = 200$	Constrained ML	Avg.boot.Mean/MC.Mean	1.00	1.00	1.00
		Avg.Boot.SE/MC.SD	1.00	1.00	1.03
	Partial regression	Avg.boot.Mean/MC.Mean	1.00	1.00	1.00
		Avg.Boot.SE/MC.SD	1.00	1.06	1.03
		Avg.Asy.SE/MC.SD	1.00	1.06	1.03

Table 2.7: Simulation results of the five-covariate scenario: for both the constrained ML and the partial regression, we report the ratio of average bootstrap mean and Monte Carlo mean $((\frac{1}{500}\sum_{\mathrm{m=1}}^{500}\bar{\tilde{\gamma}}_{\mathrm{m,j}})/(\frac{1}{500}\sum_{\mathrm{m=1}}^{500}\hat{\gamma}_{\mathrm{m,j}}))$ and the ratio of average bootstrap standard error and Monte Carlo standard deviation $((\frac{1}{500}\sum_{\mathrm{m=1}}^{500}\tilde{\sigma}_{\mathrm{m,j}})/\sqrt{V(\hat{\gamma}_j)})$ of each regression coefficient. For the partial regression solution, we also report the ratio of average asymptotic standard error and Monte Carlo standard deviation $(\frac{1}{500}\sum_{\mathrm{m=1}}^{500}\mathrm{Asy.SE}(\gamma_{\mathrm{m,j}})/\sqrt{V(\hat{\gamma}_j)})$

Sample Size	Method	Ratio	γ_1	γ_2	γ_3	γ_4	γ_5
$n = 20$	Constrained ML	Avg.boot.Mean/MC.Mean	0.89	0.89	1.00	1.00	1.16
		Avg.Boot.SE/MC.SD	1.13	0.99	1.06	0.97	1.40
	Partial regression	Avg.boot.Mean/MC.Mean	1.00	1.00	1.00	1.00	1.00
		Avg.Boot.SE/MC.SD	1.15	1.03	1.02	0.98	1.17
		Avg.Asy.SE/MC.SD	1.06	0.94	0.97	0.94	1.01
$n = 30$	Constrained ML	Avg.boot.Mean/MC.Mean	0.93	0.93	1.00	1.00	1.11
		Avg.Boot.SE/MC.SD	1.08	1.00	0.94	1.07	1.10
	Partial regression	Avg.boot.Mean/MC.Mean	1.00	1.00	1.00	1.00	1.00
		Avg.Boot.SE/MC.SD	1.03	1.01	0.94	1.05	1.05
		Avg.Asy.SE/MC.SD	0.99	0.97	0.96	1.05	0.95
$n = 200$	Constrained ML	Avg.boot.Mean/MC.Mean	0.99	0.99	1.01	0.99	1.01
		Avg.Boot.SE/MC.SD	1.03	1.00	1.00	1.00	1.00
	Partial regression	Avg.boot.Mean/MC.Mean	1.00	1.00	1.00	1.00	1.00
		Avg.Boot.SE/MC.SD	1.03	1.00	1.00	1.07	1.00
		Avg.Asy.SE/MC.SD	1.03	1.00	1.00	1.07	1.00

Table 2.8: Characteristics and lead biomarkers of subjects in training dataset (N = 100) and in testing dataset (N = 56)

	Training dataset		Testing dataset	
	Mean $\pm$ SD	Range	Mean $\pm$ SD	Range
Age (yr)	67.22 $\pm$ 7.44	51.64 - 92.25	68.72 $\pm$ 6.65	53.93 - 82.42
Cumulative cigarette (pack-yr)	18.32 $\pm$ 24.45	0.00 - 136.00	21.50 $\pm$ 26.93	0.00 - 105.50
Genetic risk score	11.45 $\pm$ 2.27	5.76 - 16.97	10.85 $\pm$ 2.36	4.00 - 16.00
Lead biomarkers				
Blood lead (μg/dL)	5.53 $\pm$ 3.29	0.00 - 17.00	5.73 $\pm$ 2.29	1.00 - 11.00
Tibia lead (μg/g)	20.59 $\pm$ 11.86	3.00 - 76.00	20.89 $\pm$ 14.06	-1.00 - 77.00
		N(%)		
Smoking status				
Never		39 (39.00)		16 (28.57)
Former		52 (52.00)		38 (67.86)
Current		9 (9.00)		2 (3.57)
Education				
High school dropout		10 (10.00)		10 (17.86)
High school diploma		58 (58.00)		30 (53.57)
$\geq$ 4 yr of college		32 (32.00)		16 (28.57)
White collar		61 (61.00)		31 (55.36)

Table 2.9: Regression coefficients of the expanded tibia lead prediction model with the genetic score (n = 100)

Variable	Constrained ML				Direct Regressio
	$d = 0$	$d = 1$	$d = 5$	$d = 10$	
Intercept	-21.90(6.56)	-27.47(7.27)	-26.94(10.57)	-23.85(10.72)	-23.85(11.27)
Blood lead	1.05(0.08)	0.93(0.13)	0.91(0.27)	0.95(0.29)	0.95(0.30)
Age	0.59(0.02)	0.64(0.06)	0.60(0.12)	0.55(0.13)	0.55(0.13)
Education					
High school diploma	-3.67(0.44)	-1.92(1.42)	2.34(2.79)	3.19(3.20)	3.19(3.44)
$\geq$ 4 yr of college	-7.12(0.60)	-5.04(1.60)	-0.80(3.41)	-0.06(3.82)	-0.06(4.08)
White collar	-3.24(0.32)	-2.84(1.02)	-3.98(2.29)	-4.48(2.29)	-4.48(2.42)
Cumulative cigarette smoking	0.03(0.01)	0.06(0.02)	0.16(0.04)	0.22(0.05)	0.22(0.06)
Smoking status					
Former smoker	1.83(0.35)	2.89(0.98)	0.05(2.05)	-1.92(2.22)	-1.92(2.54)
Current smoker	0.18(0.85)	-2.28(1.90)	-12.21(3.30)	-19.07(4.59)	-19.07(5.48)
Genetic risk score	0.13(0.51)	0.15(0.48)	0.13(0.44)	0.13(0.43)	0.13(0.44)
R^2	0.32	0.35	0.41	0.42	0.42
OOB R^2	0.29	0.29	0.23	0.17	0.17

Table 2.10: Simulation results of three-covariate scenario: Comparison of different methods. OOB R^2 denotes average out-of-bag prediction ability. For each method, each row includes mean (Monte Carlo standard error) of each regression coefficient and OOB R^2 of this method. A linear regression on $\mathbf{Y}$ on $\mathbf{X_1}, \mathbf{X_2}$ has an OOB R^2 of 0.212

Method	Sample size	γ_1	γ_2	γ_3	OOB R^2
True value		3	3	2	
Direct regression	$n = 15$	3.25(2.27)	3.07(2.43)	1.96(1.39)	0.270
	$n = 50$	2.95(1.08)	2.98(1.07)	2.01(0.58)	0.429
	$n = 100$	2.96(0.75)	3.01(0.78)	2.01(0.44)	0.451
	$n = 200$	2.98(0.52)	2.97(0.54)	2.01(0.30)	0.464
	$n = 2000$	3.01(0.16)	3.00(0.17)	2.00(0.09)	0.474
Constrained ML	$n = 15$	2.82(1.80)	2.79(1.64)	2.27(1.55)	0.334
	$n = 50$	2.92(0.69)	2.93(0.66)	2.08(0.59)	0.453
	$n = 100$	2.95(0.46)	2.98(0.47)	2.05(0.44)	0.462
	$n = 200$	2.97(0.34)	2.98(0.34)	2.03(0.30)	0.469
	$n = 2000$	3.01(0.13)	3.00(0.14)	2.00(0.09)	0.474
Partial regression	$n = 15$	3.03(1.51)	3.01(1.50)	1.96(1.39)	0.346
	$n = 50$	2.99(0.70)	2.99(0.65)	2.01(0.58)	0.446
	$n = 100$	3.00(0.45)	3.01(0.45)	2.01(0.44)	0.459
	$n = 200$	2.99(0.33)	3.00(0.31)	2.01(0.30)	0.468
	$n = 2000$	3.00(0.10)	3.00(0.09)	2.00(0.09)	0.474
Standard Bayes	$n = 15$	3.24(2.28)	3.06(2.44)	1.97(1.39)	0.270
	$n = 50$	2.95(1.08)	2.98(1.07)	2.01(0.58)	0.429
	$n = 100$	2.96(0.75)	3.01(0.78)	2.01(0.44)	0.451
	$n = 200$	2.98(0.52)	2.97(0.54)	2.01(0.30)	0.464
	$n = 2000$	3.01(0.16)	3.00(0.17)	2.00(0.09)	0.474
Informative full Bayes	$n = 15$	3.06(1.42)	2.99(1.43)	1.98(1.33)	0.382
	$n = 50$	2.98(0.65)	2.99(0.62)	2.01(0.56)	0.457
	$n = 100$	2.99(0.44)	3.01(0.44)	2.01(0.43)	0.464
	$n = 200$	2.99(0.32)	2.99(0.30)	2.01(0.30)	0.470
	$n = 2000$	3.01(0.11)	3.00(0.12)	2.00(0.09)	0.475
Transformation	$n = 15$	3.16(1.55)	3.09(1.60)	1.84(1.49)	0.366
	$n = 50$	2.98(0.69)	2.99(0.65)	2.00(0.61)	0.455
	$n = 100$	2.99(0.46)	3.00(0.48)	2.01(0.48)	0.463
	$n = 200$	2.98(0.33)	2.99(0.33)	2.01(0.32)	0.469
	$n = 2000$	3.01(0.11)	3.00(0.12)	2.00(0.10)	0.475

Table 2.11: Simulation results of five-covariate scenario: Comparison of different methods. OOB R^2 denotes average out-of-bag prediction ability. For each method, each row includes mean (Monte Carlo standard error) of each regression coefficient and OOB R^2 of this method. A linear regression on $\mathbf{Y}$ on $\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_4}$ has an OOB R^2 of 0.350

Method	Sample size	γ_1	γ_2	γ_3	γ_4	γ_5	OOB R^2
True value		3	3	2	2	2	
Direct regression	$n = 20$	3.11(1.88)	3.02(2.07)	1.93(1.15)	2.04(1.15)	1.92(1.10)	0.421
	$n = 50$	2.97(1.02)	3.02(1.03)	1.95(0.61)	2.01(0.59)	1.99(0.57)	0.545
	$n = 100$	3.04(0.75)	3.05(0.69)	2.00(0.42)	2.03(0.44)	1.99(0.43)	0.569
	$n = 200$	2.94(0.48)	2.95(0.48)	2.00(0.30)	2.00(0.29)	2.04(0.29)	0.581
	$n = 2000$	3.00(0.15)	3.00(0.16)	2.00(0.09)	2.00(0.09)	1.99(0.09)	0.591
Constrained ML	$n = 20$	2.77(1.36)	2.66(1.60)	2.03(0.70)	1.94(0.75)	2.33(1.26)	0.492
	$n = 50$	2.89(0.67)	2.93(0.67)	2.05(0.30)	1.94(0.31)	2.12(0.59)	0.567
	$n = 100$	2.97(0.49)	2.99(0.46)	2.07(0.21)	1.95(0.22)	2.05(0.43)	0.577
	$n = 200$	2.95(0.31)	2.93(0.33)	2.06(0.15)	1.94(0.15)	2.07(0.29)	0.582
	$n = 2000$	3.00(0.13)	3.00(0.13)	2.03(0.06)	1.97(0.06)	2.00(0.09)	0.589
Partial regression	$n = 20$	3.08(1.21)	3.01(1.37)	2.09(0.64)	1.91(0.65)	1.92(1.10)	0.500
	$n = 50$	3.00(0.65)	3.03(0.67)	2.10(0.33)	1.91(0.31)	1.99(0.57)	0.560
	$n = 100$	3.01(0.48)	3.03(0.46)	2.10(0.21)	1.91(0.22)	1.99(0.43)	0.572
	$n = 200$	2.99(0.31)	2.96(0.32)	2.10(0.15)	1.90(0.14)	2.04(0.29)	0.577
	$n = 2000$	3.00(0.10)	3.01(0.10)	2.10(0.04)	1.90(0.04)	1.99(0.09)	0.583
Standard Bayes	$n = 20$	3.11(1.88)	3.02(2.07)	1.93(1.15)	2.04(1.15)	1.92(1.10)	0.421
	$n = 50$	2.97(1.02)	3.02(1.03)	1.95(0.61)	2.01(0.59)	1.99(0.58)	0.545
	$n = 100$	3.04(0.75)	3.05(0.70)	2.00(0.42)	2.03(0.44)	1.99(0.43)	0.569
	$n = 200$	2.94(0.48)	2.95(0.48)	2.00(0.30)	2.00(0.29)	2.04(0.29)	0.580
	$n = 2000$	3.00(0.15)	3.00(0.16)	2.00(0.09)	2.00(0.09)	1.99(0.09)	0.591
Informative full Bayes	$n = 20$	3.02(1.09)	2.94(1.30)	2.06(0.59)	1.92(0.62)	1.99(0.99)	0.526
	$n = 50$	2.99(0.60)	3.03(0.61)	2.06(0.28)	1.93(0.28)	1.99(0.54)	0.570
	$n = 100$	3.02(0.46)	3.03(0.43)	2.08(0.19)	1.93(0.20)	1.99(0.42)	0.578
	$n = 200$	2.97(0.30)	2.96(0.31)	2.07(0.13)	1.93(0.13)	2.04(0.29)	0.582
	$n = 2000$	3.00(0.11)	3.00(0.11)	2.04(0.05)	1.96(0.06)	1.99(0.09)	0.588
Transformation	$n = 20$	3.16(1.23)	3.08(1.44)	2.04(0.62)	1.95(0.64)	1.81(1.22)	0.516
	$n = 50$	3.01(0.65)	3.05(0.67)	2.06(0.29)	1.93(0.29)	1.96(0.64)	0.568
	$n = 100$	3.03(0.49)	3.04(0.46)	2.07(0.20)	1.94(0.20)	1.97(0.46)	0.578
	$n = 200$	2.97(0.31)	2.95(0.32)	2.07(0.14)	1.93(0.13)	2.05(0.30)	0.582
	$n = 2000$	3.00(0.11)	3.00(0.11)	2.05(0.05)	1.95(0.05)	2.00(0.09)	0.588

Bibliography

Abdi, H. Partial regression coefficients. In Lewis-Beck, M. S., Bryman, A. and Liao, T. F., editors, *The encyclopedia of social science research methods*, pages 796–798. Sage Publications, Inc., CA, 2004.

Barry, P. S. I. and Mossman, D. B. Lead concentrations in human tissues. *British Journal of Industrial Medicine*, 27(4):339–351, 1970.

Bell, B., Rose, C. L. and Damon, A. The normative aging study: an interdisciplinary and longitudinal study of health and aging. *The International Journal of Aging and Human Development*, 3(1):5–17, 1972.

Carlin, B. P. and Louis, T. A. *Bayesian methods for data analysis*. CRC Press, 2009.

Carpenter, J. and Bithell, J. Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in Medicine*, 19(9):1141–1164, 2000.

Chatterjee, N., Chen, Y.-H., Maas, P. and Carroll, R. J. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111(513):107–117, 2016.

Chen, A., Owen, A. B. and Shi, M. Data enriched linear regression. *Electronic Journal of Statistics*, 9(1):1078–1112, 2015.

D'Agostino, R. B., Grundy, S., Sullivan, L. M., Wilson, P. and the CHD Risk Prediction Group. Validation of the framingham coronary heart disease prediction scores: Results of a multiple ethnic groups investigation. *The Journal of the American Medical Association*, 286(2):180–187, 2001.

Dunson, D. B. and Neelon, B. Bayesian inference on order-constrained parameters in generalized linear models. *Biometrics*, 59(2):286–295, 2003.

Efron, B. Nonparametric standard errors and confidence intervals. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 9(2):139–158, 1981.

Efron, B. and Tibshirani, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986.

Fox, J. *Applied regression analysis and generalized linear models*. Sage, 2008.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C. and Mulvihill, J. J. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, 81(24):1879–1886, 1989.

Geweke, J. Exact inference in the inequality constrained normal linear regression model. *Journal of Applied Econometrics*, 1(2):127–141, 1986.

Gilks, W. R. and Roberts, G. O. Strategies for improving mcmc. In Gilks, W. R., Richardson, S. and Spiegelhalter, D., editors, *Markov chain Monte Carlo in practice*, pages 89–110. Chapman and Hall, London, 1996.

Gunn, L. H. and Dunson, D. B. A transformation approach for incorporating monotone or unimodal constraints. *Biostatistics*, 6(3):434–449, 2005.

Hu, H., Shih, R., Rothenberg, S. and Schwartz, B. S. The epidemiology of lead toxicity in adults: Measuring dose and consideration of other methodologic issues. *Environmental Health Perspectives*, 115(3):455–462, 2007.

Imbens, G. W. and Lancaster, T. Combining micro and macro data in microeconometric models. *The Review of Economic Studies*, 61(4):655–680, 1994.

Lesaffre, E. and Lawson, A. B. Choosing the prior distribution. In Lesaffre, E. and Lawson, A. B., editors, *Bayesian Biostatistics*, pages 104–138. John Wiley and Sons, Ltd, West Sussex, 2012.

Mukherjee, B. and Chatterjee, N. Exploiting gene-environment independence for analysis of case-control studies: an empirical bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*, 64(3):685–694, 2008.

Park, S. K., Mukherjee, B., Xia, X., Sparrow, D., Weisskopf, M. G., Nie, H. and Hu, H. Bone lead level prediction models and their application to examining the relationship of lead exposure and hypertension in the third national health and nutrition examination survey (nhanes-iii). *Journal of Occupational and Environmental Medicine / American College of Occupational and Environmental Medicine.*, 51(12):1422–1436, 2009.

Qin, J. Combining parametric and empirical likelihoods. *Biometrika*, 87(2):484–490, 2000.

Qin, J., Zhang, H., Li, P., Albanes, D. and Yu, K. Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika*, 102(1):169–180, 2015.

Roberts, G. O. Markov chain concepts related to sampling algorithms. In Gilks, W. R., Richardson, S. and Spiegelhalter, D., editors, *Markov chain Monte Carlo in practice*, pages 45–54. Chapman and Hall, London, 1996.

Thompson, I. M., Ankerst, D. P., Chi, C., Goodman, P. J., Tangen, C. M., Lucia, M. S., Feng, Z., Parnes, H. L. and Coltman, C. A. Assessing prostate cancer risk: Results from the prostate cancer prevention trial. *Journal of the National Cancer Institute*, 98(8):529–534, 2006.

CHAPTER III

Informing a risk prediction model for binary outcomes with external coefficient information

3.1 Introduction

Risk prediction models for different binary disease endpoints are abundant in the clinical and epidemiological literature. Examples are the breast cancer risk calculator (Gail et al., 1989), the Framingham risk score (D'Agostino et al., 2001) and the Prostate Cancer Prevention Trial Risk Calculator (PCPTrc) (Thompson et al., 2006). Some of these prediction models are included in clinical guidelines for doctors to assess an individual's risk of experiencing a future health event and to make decisions concerning screening and prophylactic prevention. As an example to be used in this paper, the Prostate Cancer Prevention Trial Risk Calculator is an online assessment tool which provides personalized risk estimate of detecting prostate cancer based on risk factors such as age, prostate-specific antigen (PSA) and digital rectal examination (DRE) findings.

While these models are often based on traditional epidemiologic and behavioral risk factors, wider availability of high throughput data and novel assay technologies are generating candidate biomarkers for potential inclusion in existing risk prediction models. It's very likely that the new biomarkers are assessed only on participants in a study of moderate size and cannot be retrospectively measured on the much larger population used for the well-established model. Investigators could directly estimate the expanded model in the new dataset, but results from this expanded pre-

diction model based solely on a limited number of subjects could be highly variable. It is natural to consider using the information from the well-established model to increase the accuracy of the expanded model, which is estimated based on the new dataset.

Substantial research has been done on the problem of enhancing risk prediction models with supplemental external information. The external information from outside the new dataset may be used in an "adaptive" way, which enables us to combine estimates from previous studies with the regression coefficients estimated in the new dataset. Steyerberg et al. (2000) described a method to adjust the multivariate logistic regression model's coefficients estimated in a dataset based on univariate regression models' coefficients in the literature. Newcombe et al. (2012) presented two possible approaches incorporating the effect estimates of a set of predictors: the first one was by adding a composite weighted risk score based on these estimates and the second one was by specifying informative priors for the coefficients of these variables in a Bayesian logistic regression model. Recently Chatterjee et al. (2016) developed a general method for incorporating external coefficients, derived from constrained estimating equations, and they showed their method to be efficient under certain conditions. Other related approaches used constrained maximum likelihood and empirical likelihood (Imbens and Lancaster, 1994; Qin, 2000; Qin et al., 2015).

There are also a number of simple approaches. For the Gail model, Mealiffe et al. (2010) computed a multiplicative risk score based on previously published odds ratios of newly discovered biomarkers. They then multiplied the Gail risk estimate and the multiplicative risk score and the combined risk score was shown to give an improvement in the area under the curve (AUC). Grill et al. (2015) proposed a simple method of incorporating new markers into the Prostate Cancer Prevention Risk Calculator via Bayes Theorem. They updated the posterior odds of getting cancer based on both standard risk factors and new markers by using the likelihood ratio incorporating dependence between the two sets of risk factors to adjust the prior odds of getting cancer based on standard risk alone. Grill et al. (2017) assessed the performance of a set of likelihood ratio

approaches as well as the constrained maximum likelihood estimation approach proposed in Chatterjee et al. (2016) in terms of the ratio of expected to observed cases, AUC and Brier score in a validation dataset.

We consider a situation where the outcome is a binary indicator of disease and the well-established model is described in a published article, in which the estimated regression coefficients and their standard errors are presented in tables. The expanded model includes one additional biomarker as a potential predictor. To introduce notation, let $\mathbf{Y}$ denote the binary outcome, $\mathbf{X}$ is a set of p standard risk factors and $\mathbf{B}$ is a new biomarker. The association between $\mathbf{Y}$ and $\mathbf{X}$ is described through the following logistic model:

$$\text{logit}(\Pr(\mathbf{Y}=1|\mathbf{X})) = \mathbf{X}\boldsymbol{\beta} = \beta_0 + \beta_1\mathbf{X_1} + \cdots + \beta_{\text{p}}\mathbf{X_p} \tag{3.1}$$

We assume we have available summary-level information on the estimated regression coefficients and their standard errors in model (3.1). We use $\bar{\boldsymbol{\beta}}$ and $\bar{\mathbf{S}}$ to denote these known estimated coefficients and their standard errors.

The model of primary interest is an expanded model that describes the joint effect of $\mathbf{X}, \mathbf{B}$ on $\mathbf{Y}$ to be estimated from a small dataset:

$$\text{logit}(\Pr(\mathbf{Y}=1|\mathbf{X},\mathbf{B})) = \mathbf{X}\boldsymbol{\gamma}_{\text{X}} + \mathbf{B}\gamma_{\text{B}} = \gamma_0 + \gamma_1\mathbf{X_1} + \cdots + \gamma_{\text{p}}\mathbf{X_p} + \gamma_{\text{p+1}}\mathbf{B} \tag{3.2}$$

Another model that can be estimated from the current small dataset is:

$$\text{g}(\text{E}(\mathbf{B}|\mathbf{X})) = \mathbf{X}\boldsymbol{\theta} = \theta_0 + \theta_1\mathbf{X_1} + \cdots + \theta_{\text{p}}\mathbf{X_p} \tag{3.3}$$

where g is the link function, which is the identity function $\text{g}(\text{y}) = \text{y}$ for Gaussian $\mathbf{B}$ and the logit function $\text{g}(\text{y}) = \log(\text{y}/(1-\text{y}))$ for binary $\mathbf{B}$. We propose to formulate the problem in an inferential framework where the historical information is translated in terms of non-linear constraints on the regression parameters and propose both frequentist and Bayes solutions to this problem.

The distribution of $\mathbf{B}$ will greatly affect how we translate the historical information into constraints on the regression parameters. In this study, we mainly discuss the cases that $\mathbf{B}$ is Gaussian and that $\mathbf{B}$ is binary. The case that $\mathbf{B}$ is multivariate Gaussian is also described and these strategies for incorporating the external coefficient information can be easily extended.

The following is the structure of the remainder of this chapter: in Section 3.2, we discuss how to establish a relationship equation between the regression coefficients of models (3.1) - (3.3) when B is Gaussian. In Section 3.3, we consider the situation when B is binary and derive the corresponding constrained solutions. We present a simulation study in Section 3.4. In Section 3.5 we demonstrate the proposed approaches by expanding the established High-grade Prostate Cancer Prevention Trial Risk Calculator. Concluding remarks are presented in Section 3.6.

3.2 Statistical Approaches

3.2.1 *Logistic Regression Approximation*

A difficulty in translating the summary information from modeling $\Pr(\mathbf{Y}=1|\mathbf{X})$ to modeling $\Pr(\mathbf{Y}=1|\mathbf{X},\mathbf{B})$ is that a logistic model $\mathrm{logit}(\Pr(\mathbf{Y}=1|\mathbf{X},\mathbf{B}))$ does not reduce to a logistic model $\mathrm{logit}(\Pr(\mathbf{Y}=1|\mathbf{X}))$ when marginalized over the distribution of $\mathbf{B}$. This is the well-known lack of collapsibility property of a logistic regression model. To connect the regression coefficients in models (3.1), (3.2) and (3.3), we need to approximate $\mathrm{logit}(\Pr(\mathbf{Y}=1|\mathbf{X}))$ written in terms of parameters $\boldsymbol{\gamma}, \boldsymbol{\theta}$ and variables $\mathbf{X}$, and equate that to $\mathrm{logit}(\Pr(\mathbf{Y}=1|\mathbf{X}))=\mathbf{X}\boldsymbol{\beta}$. To achieve this, we consider the following integral:

$$\begin{aligned}\Pr(\mathbf{Y}=1|\mathbf{X}) &= \int \Pr(\mathbf{Y}=1|\mathbf{X},\mathbf{B})\mathrm{P}(\mathbf{B}|\mathbf{X})\mathrm{d}\mathbf{B} \\ &= \frac{\int \mathrm{H}(\mathbf{X}\boldsymbol{\gamma}_{\mathrm{x}}+\mathbf{B}\boldsymbol{\gamma}_{\mathrm{B}})\exp\big(-\frac{(\mathbf{B}-\mathbf{X}\boldsymbol{\theta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{B}-\mathbf{X}\boldsymbol{\theta})}{2}\big)\mathrm{d}\mathbf{B}}{(2\pi)^{\mathrm{L}/2}|\mathbf{V}|^{1/2}}\end{aligned} \tag{3.4}$$

where $\mathrm{H}(\mathrm{v})=(1+\exp(-\mathrm{v}))^{-1}$, $\mathbf{B}$ has L dimensions and $\mathbf{B}|\mathbf{X}$ is assumed to follow a multivariate Gaussian distribution $\mathrm{N}(\mathbf{X}\boldsymbol{\theta},\mathbf{V}_{\mathrm{L}\times\mathrm{L}})$. This integral in (3.4) does not have a closed-form solution and approximations are necessary.

By a normal scale mixture representation of the logistic distribution function and computation of the logistic-normal integral (Monahan and Stefanski, 1992), we can find an approximated equation: $\Pr(\mathbf{Y}=1|\mathbf{X}) \approx \mathrm{H}\Big(\frac{\mathbf{X}\boldsymbol{\gamma}_{\mathrm{x}}+(\mathbf{X}\boldsymbol{\theta})\boldsymbol{\gamma}_{\mathrm{B}}}{(1+\boldsymbol{\gamma}_{\mathrm{B}}^{\mathrm{T}}\mathbf{V}\boldsymbol{\gamma}_{\mathrm{B}}/1.7^2)^{\frac{1}{2}}}\Big)$. The derivation of the approximation is given in Supplementary Material Appendix A. In most cases this is a good approximation, unless $\boldsymbol{\gamma}_B^T\mathbf{V}\boldsymbol{\gamma}_B$ is too large (Carroll et al., 2006). Using this approximation, we find an approximate relationship between $\boldsymbol{\gamma}, \boldsymbol{\theta}$ and $\boldsymbol{\beta}$ when $\mathbf{B}$ is multivariate Gaussian:

$$\beta_j \approx \frac{\gamma_j + \sum_{l=1}^{L}\gamma_{p+l}\theta_{lj}}{(1+\boldsymbol{\gamma}_B^T\mathbf{V}\boldsymbol{\gamma}_B/1.7^2)^{\frac{1}{2}}}, j=0,\ldots,p. \tag{3.5}$$

When $\mathbf{B}$ is univariate Gaussian and $\mathbf{B}|\mathbf{X}$ is assumed to follow a Gaussian distribution $\mathrm{N}(\mathbf{X}\boldsymbol{\theta}, \sigma_2^2)$, the approximate relationship between $\boldsymbol{\gamma}, \boldsymbol{\theta}$ and $\boldsymbol{\beta}$ becomes:

$$\beta_j \approx \frac{\gamma_j + \gamma_{p+1}\theta_j}{(1+\gamma_{p+1}^2\sigma_2^2/1.7^2)^{\frac{1}{2}}}, j=0,\ldots,p. \tag{3.6}$$

In this section, we focus on the scenario where $\mathbf{B}$ is univariate Gaussian and propose unconstrained and constrained solutions.

3.2.2 *Firth Correction in Logistic Regression*

Firth correction (Firth, 1993) is a general approach to reduce bias in maximum likelihood estimation by maximizing a penalized log-likelihood function, where the penalty is $\frac{1}{2}|\mathbf{I}|$ and $\mathbf{I}$ is the information matrix. In logistic regression, standard maximum likelihood estimates often experience serious bias or even non-existence due to separability and multicollinearity. Heinze and Schemper (2002) suggested using Firth correction to overcome the separability issue in logistic regression. In our constrained solution, we add Firth correction to stabilize the estimates from standard logistic regression.

3.2.3 *Unconstrained Solutions*

Direct Regression

Without constraints, the unknown parameters $\boldsymbol{\gamma}$ in model (3.2) can be estimated by maximizing the likelihood using an iteratively reweighted least squares approach. Specifically, the estimate solves

$$\max_{\boldsymbol{\gamma}} \left\{ \sum_{\mathrm{i}=1}^{\mathrm{n}} [\mathrm{Y_i}(\sum_{\mathrm{j}=0}^{\mathrm{p}} \gamma_{\mathrm{j}} \mathrm{X_{ij}} + \gamma_{\mathrm{p+1}} \mathrm{B_i}) - \log(1 + \exp(\sum_{\mathrm{j}=0}^{\mathrm{p}} \gamma_{\mathrm{j}} \mathrm{X_{ij}} + \gamma_{\mathrm{p+1}} \mathrm{B_i}))] \right\} \tag{3.7}$$

Since the separability issue could lead to large bias or non-existence of the MLE, we implement Firth's penalized likelihood approach using R package **logistf**.

To estimate parameters in model (3.3), we obtain least squares estimates of $\boldsymbol{\theta}$ by considering:

$$\min_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} (\mathrm{B_i} - \sum_{\mathrm{j}=0}^{\mathrm{p}} \theta_{\mathrm{j}} \mathrm{X_{ij}})^2 \right\} \tag{3.8}$$

Standard Bayes

Analogous to direct linear regression, we can perform standard Bayesian linear regression with flat conjugate priors for parameters in model (3.3). For model (3.2), we can perform standard Bayesian logistic regression. Posterior distributions for Bayesian analysis of a logistic regression model are not available as closed-form expressions based on a conjugate prior and instead standard Bayes can either be implemented by a Metropolis-Hasting sampling technique with flat priors or a Jeffrey's prior. Gelman et al. (2008) suggested weakly informative Cauchy distributions as priors for the regression coefficients in logistic regression to reduce the separability issue. With an approximate EM algorithm, this non-informative Bayes method can be implemented in a fast and easy way to obtain posterior draws. We will use this method with weakly informative Cauchy priors throughout this paper.

3.2.4 *Constrained Solutions*

Constrained Maximum Likelihood

The constrained maximum likelihood (constrained ML) estimation optimizes the joint log-likelihood under the set of constraints generated based on the approximate relationship equations

in (3.6). As we have the point estimates and the standard errors of $\boldsymbol{\beta}$ from the established model, we require the parameter estimates for $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ to result in the desired estimated $\boldsymbol{\beta}$ to be within d standard errors of the old point estimates:

$$\begin{aligned}\min_{\boldsymbol{\gamma},\boldsymbol{\theta}}\Big\{ &\sum_{i=1}^{n}[-\mathrm{Y_i}(\sum_{\mathrm{j=0}}^{\mathrm{p}}\gamma_{\mathrm{j}}\mathrm{X_{ij}}+\gamma_{\mathrm{p+1}}\mathrm{B_i})+\log(1+\exp(\sum_{\mathrm{j=0}}^{\mathrm{p}}\gamma_{\mathrm{j}}\mathrm{X_{ij}}+\gamma_{\mathrm{p+1}}\mathrm{B_i}))] \\ &+\sum_{\mathrm{i=1}}^{\mathrm{n}}\frac{(\mathrm{B_i}-\sum_{\mathrm{j=0}}^{\mathrm{p}}\theta_{\mathrm{j}}\mathrm{X_{ij}})^2}{2\hat{\sigma}_2^2}\Big\} \\ \text{s.t.}\,&\frac{\gamma_j+\gamma_{p+1}\theta_j}{(1+\gamma_{p+1}^2\sigma_2^2/1.7^2)^{\frac{1}{2}}}\in[\bar{\beta}_j-d\bar{S}_j,\bar{\beta}_j+d\bar{S}_j],j=0,\ldots,p\end{aligned} \tag{3.9}$$

In this optimization problem, $\hat{\sigma}_2^2$ is a plug-in estimator and is the OLS residual variance from fitting $\mathrm{E(B|X)}$ and d is a scale parameter representing the strength of external information. From simulations, we find that fixing d as $d=1$ leads to decent properties of the estimates of $\boldsymbol{\gamma}$. To solve this optimization problem, we use function $\mathbf{solnp}$ in R package $\mathbf{Rsolnp}$, a function that efficiently solves a general nonlinear optimization problem using Lagrange multipliers. For computational convenience, we further specify wide lower and upper bounds for each of these parameters: $\gamma_{\mathrm{j}}\in[\hat{\gamma}_{\mathrm{j}}-5\hat{\mathrm{SE}}(\gamma_{\mathrm{j}}),\hat{\gamma}_{\mathrm{j}}+5\hat{\mathrm{SE}}(\gamma_{\mathrm{j}})],\mathrm{j}=0,\cdots,\mathrm{p}+1,\theta_{\mathrm{j}}\in[\hat{\theta}_{\mathrm{j}}-5\hat{\mathrm{SE}}(\theta_{\mathrm{j}}),\hat{\theta}_{\mathrm{j}}+5\hat{\mathrm{SE}}(\theta_{\mathrm{j}})],\mathrm{j}=0,\cdots,\mathrm{p}$ where $\hat{\gamma}_j,j=0,\cdots,p+1$ and $\hat{\theta}_j,j=0,\cdots,p$ are the MLEs and $\hat{\mathrm{SE}}(\gamma_{\mathrm{j}}),\mathrm{j}=0,\cdots,\mathrm{p}+1$ and $\hat{\mathrm{SE}}(\theta_{\mathrm{j}}),\mathrm{j}=0,\cdots,\mathrm{p}$ are the corresponding standard errors.

We also consider a modification to the above constrained ML solution by adding a Firth penalty term to the objective function:

$$\begin{aligned}\min_{\boldsymbol{\gamma},\boldsymbol{\theta}}\Big\{ &\sum_{i=1}^{n}[-\mathrm{Y_i}(\sum_{\mathrm{j=0}}^{\mathrm{p}}\gamma_{\mathrm{j}}\mathrm{X_{ij}}+\gamma_{\mathrm{p+1}}\mathrm{B_i})+\log(1+\exp(\sum_{\mathrm{j=0}}^{\mathrm{p}}\gamma_{\mathrm{j}}\mathrm{X_{ij}}+\gamma_{\mathrm{p+1}}\mathrm{B_i}))] \\ &+\sum_{\mathrm{i=1}}^{\mathrm{n}}\frac{(\mathrm{B_i}-\sum_{\mathrm{j=0}}^{\mathrm{p}}\theta_{\mathrm{j}}\mathrm{X_{ij}})^2}{2\hat{\sigma}_2^2}-0.5\log|\mathbf{I}(\boldsymbol{\gamma})|\Big\} \\ \text{s.t.}\,&\frac{\gamma_j+\gamma_{p+1}\theta_j}{(1+\gamma_{p+1}^2\sigma_2^2/1.7^2)^{\frac{1}{2}}}\in[\bar{\beta}_j-d\bar{S}_j,\bar{\beta}_j+d\bar{S}_j],j=0,\ldots,p\end{aligned} \tag{3.10}$$

where $|\mathbf{I}(\boldsymbol{\gamma})|$ is the determinant of the Fisher information matrix of the likelihood function $\mathrm{L}(\mathbf{Y}|\mathbf{X},\mathbf{B})$. We use the bootstrap as described in Supplementary Material Appendix C to estimate the standard

errors.

Informative Full Bayes

In informative full Bayes, starting with the joint likelihood $\mathrm{L}(\mathbf{Y}|\mathbf{X},\mathbf{B})\mathrm{L}(\mathbf{B}|\mathbf{X})$ we translate the constraints in (3.6) to prior information. The first step is to write down the joint likelihood function with priors on $\boldsymbol{\gamma}, \boldsymbol{\theta}, \sigma_2^2$:

$$\begin{aligned} & p(\boldsymbol{\gamma}, \boldsymbol{\theta}, \sigma_2^2|\mathrm{data}) \propto \mathrm{L}(\mathbf{Y}|\mathbf{X},\mathbf{B},\boldsymbol{\gamma}) \cdot \mathrm{L}(\mathbf{B}|\mathbf{X},\boldsymbol{\theta},\sigma_2^2) \cdot \pi(\boldsymbol{\gamma},\boldsymbol{\theta},\sigma_2^2) \\ & = \Big\{ \prod_{\mathrm{i}=1}^{\mathrm{n}} \frac{\exp((\sum_{\mathrm{j}=0}^{\mathrm{p}} \gamma_{\mathrm{j}} \mathrm{X_{ij}} + \gamma_{\mathrm{p+1}} \mathrm{B_i}) \mathrm{Y_i})}{1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}} \gamma_{\mathrm{j}} \mathrm{X_{ij}} + \gamma_{\mathrm{p+1}} \mathrm{B_i})} \cdot \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\big(-\frac{1}{2\sigma_2^2}(\mathrm{B_i} - \sum_{\mathrm{j}=0}^{\mathrm{p}} \theta_{\mathrm{j}} \mathrm{X_{ij}})^2\big) \Big\} \cdot \pi(\boldsymbol{\gamma},\boldsymbol{\theta},\sigma_2^2) \end{aligned} \tag{3.11}$$

The logistic regression approximation result (3.6) suggests that $\theta_j = \frac{1}{\gamma_{p+1}}(\beta_j \sqrt{1+\frac{\gamma_{p+1}^2\sigma_2^2}{1.7^2}} - \gamma_j), j = 0, \ldots, p$. We can re-parametrize (3.11) in terms of $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and σ_2^2, and include a Jacobian transformation matrix by $\mathbf{J}$, where $|\mathbf{J}| = \frac{1}{|\gamma_{p+1}^{p+1}|}(1+\frac{\gamma_{p+1}^2\sigma_2^2}{1.7^2})^{\frac{p+1}{2}}$. Now the likelihood is represented in terms of $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and σ_2^2.

We specify a non-informative prior inverse-gamma$(0.01, 0.01)$ for σ_2^2 and independent weakly informative Cauchy priors for $\boldsymbol{\gamma}$ (Gelman et al., 2008). For γ_0 we specify a Cauchy prior with location parameter 0, scale parameter 10. For $\gamma_j, j = 1, \ldots, p+1$ we specify a Cauchy prior with location parameter 0, scale parameter 2.5. This is achieved through the hierarchical representation:

$$\begin{aligned} & \gamma_0 \sim \mathrm{N}(0, \mathrm{k_0^2}), \gamma_1 \sim \mathrm{N}(0, \mathrm{k_1^2}), \ldots, \gamma_{\mathrm{p+1}} \sim \mathrm{N}(0, \mathrm{k_{p+1}^2}) \\ & k_0^2 \sim \mathrm{Inv}-\chi^2(1, 10^2), \mathrm{k_1^2} \sim \mathrm{Inv}-\chi^2(1, 2.5^2), \ldots, \mathrm{k_{p+1}^2} \sim \mathrm{Inv}-\chi^2(1, 2.5^2) \end{aligned} \tag{3.12}$$

As a result, the prior distribution for the coefficient $\gamma_j, j = 0, \ldots, p+1$ is a mixture of normals with unknown scale parameter $\mathrm{k_j}$ and $\mathrm{k_j}$ follows an inverse chi-square distribution.

For parameters $\boldsymbol{\beta}$, we use the constraints directly as priors:

$$\beta_{\mathrm{j}} = \frac{\gamma_{\mathrm{j}} + \gamma_{\mathrm{p+1}}\theta_{\mathrm{j}}}{(1+\gamma_{\mathrm{p+1}}^2\sigma_2^2/1.7^2)^{\frac{1}{2}}} \sim \mathrm{N}(\bar{\beta}_{\mathrm{j}}, \bar{\mathrm{S}}_{\mathrm{j}}^2), \mathrm{j} = 0, \ldots, \mathrm{p} \tag{3.13}$$

Then we can rewrite the joint distribution in terms of $\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma_2^2, \mathbf{k}^2$ as $\mathrm{p}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma_2^2, \mathbf{k}^2|\mathbf{Y}, \mathbf{X}, \mathbf{B}) \propto \Big\{\prod_{\mathrm{i}=1}^{\mathrm{n}} \frac{\exp((\sum_{\mathrm{j}=0}^{\mathrm{p}} \gamma_{\mathrm{j}} \mathrm{X}_{\mathrm{ij}} + \gamma_{\mathrm{p}+1} \mathrm{B}_{\mathrm{i}}) \mathrm{Y}_{\mathrm{i}})}{1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}} \gamma_{\mathrm{j}} \mathrm{X}_{\mathrm{ij}} + \gamma_{\mathrm{p}+1} \mathrm{B}_{\mathrm{i}})}\Big\} \cdot \Big\{\prod_{\mathrm{i}=1}^{\mathrm{n}} \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp(-\frac{1}{2\sigma_2^2}(\mathrm{B}_{\mathrm{i}} - \frac{\beta_0\sqrt{1+\frac{\gamma_{p+1}^2\sigma_2^2}{1.7^2}} - \gamma_0}{\gamma_{\mathrm{p}+1}} - \sum_{\mathrm{j}=1}^{\mathrm{p}} \frac{\beta_{\mathrm{j}}\sqrt{1+\frac{\gamma_{p+1}^2\sigma_2^2}{1.7^2}} - \gamma_{\mathrm{j}}}{\gamma_{\mathrm{p}+1}} \mathrm{X}_{\mathrm{ij}})^2)\Big\} \cdot \pi(\boldsymbol{\beta}) \cdot \Big\{\prod_{j=0}^{p+1} \frac{1}{\sqrt{2\pi k_j^2}} \exp(-\frac{\gamma_j^2}{2k_j^2})\Big\} \cdot \Big\{\prod_{j=0}^{p+1} \pi(k_j^2)\Big\} \cdot \pi(\sigma_2^2) \cdot |\mathbf{J}|$

After some algebraic calculations, the full conditional distributions of $\beta_0, \ldots, \beta_p$ turn out to be normal, each with distribution function $\mathrm{N}(\mu_{\beta_{\mathrm{j}},\mathrm{n}}, \sigma^2_{\beta_{\mathrm{j}},\mathrm{n}}), \mathrm{j} = 0, \ldots, \mathrm{p}$. The full conditional distributions of $k_0^2, k_1^2, \ldots, k_{p+1}^2$ are inverse chi-square, each with distribution function $\mathrm{Inv} - \chi^2(2, \frac{1}{2}(\mathrm{s}_{\mathrm{j}}^2 + \gamma_{\mathrm{j}}^2)), \mathrm{s}_0 = 10, \mathrm{s}_1 = \cdots = \mathrm{s}_{\mathrm{p}+1} = 2.5, \mathrm{j} = 0, \ldots, \mathrm{p}+1$. The full conditional distributions of $\gamma_0, \ldots, \gamma_{p+1}$ and the full conditional distribution of σ_2^2 do not have closed form expressions. A Metropolis-Hastings sampling algorithm is needed.

Because of the non-linear relationship between the parameters, the Metropolis-Hasting algorithm cannot be performed efficiently and thus it is computationally slow to obtain uncorrelated draws from the posterior distributions in this informative full Bayes method.

Bayesian Transformation Approach

As the informative full Bayes is computationally intensive, we want to find an approximate Bayesian approach that can produce draws that fall into the constrained space but reduces the computational burden of the informative full Bayes method. The motivation for our approach stems from the transformation approach incorporating monotone or unimodal constraints in posterior inference as proposed in Gunn and Dunson (2005). We modify their approach to the scenario of a regression model with external coefficient information.

Suppose the raw draws from non-informative standard Bayes method as described in Section 3.2 are $\boldsymbol{\gamma}$ and the raw draws from standard Bayesian linear regression are $\boldsymbol{\theta}$. The corresponding posterior covariance matrices are $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. We extract the posterior variances from $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ and denote them by $s^2_{\gamma_0}, \ldots, s^2_{\gamma_p}, s^2_{\gamma_{p+1}}, s^2_{\theta_0}, \ldots, s^2_{\theta_p}$. The OLS residual variance from fitting $\mathrm{E}(\mathbf{B}|\mathbf{X})$ is $\hat{\sigma}_2^2$. Then a set of constrained draws $\boldsymbol{\gamma}^\star, \boldsymbol{\theta}^\star$ are obtained from a set of unconstrained draws by

solving the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\gamma}^\star,\boldsymbol{\theta}^\star} \quad & \left\{\mathrm{d}^2_{\mathrm{NED}}(\boldsymbol{\gamma},\boldsymbol{\gamma}^\star)+\mathrm{d}^2_{\mathrm{NED}}(\boldsymbol{\theta},\boldsymbol{\theta}^\star)\right\} = \sum_{\mathrm{j}=0}^{\mathrm{p}+1}\frac{(\gamma_\mathrm{j}-\gamma_\mathrm{j}^\star)^2}{\mathrm{s}^2_{\gamma_\mathrm{j}}}+\sum_{\mathrm{k}=0}^{\mathrm{p}}\frac{(\theta_\mathrm{k}-\theta_\mathrm{k}^\star)^2}{\mathrm{s}^2_{\theta_\mathrm{k}}} \\ \text{s.t.} \quad & \frac{\gamma_\mathrm{j}^\star+\gamma_{\mathrm{p}+1}^\star\theta_\mathrm{j}^\star}{(1+\gamma_{\mathrm{p}+1}^{\star 2}\hat{\sigma}_2^2/1.7^2)^{\frac{1}{2}}} \in [\bar{\beta}_\mathrm{j}-\mathrm{d}\bar{\mathrm{S}}_\mathrm{j},\bar{\beta}_\mathrm{j}+\mathrm{d}\bar{\mathrm{S}}_\mathrm{j}], \mathrm{j}=0,\ldots,\mathrm{p} \end{aligned} \tag{3.14}$$

where $\mathrm{d}_{\mathrm{NED}}$ stands for normalized Euclidean distance. For the transformation of a single draw, we generate d from half normal distribution: $d \sim |\mathrm{N}(0,1)|$. The intuition behind this transformation procedure is that it will produce values $\boldsymbol{\gamma}^\star, \boldsymbol{\theta}^\star$ subject to the box constraints that are closest to the unconstrained values $\boldsymbol{\gamma}, \boldsymbol{\theta}$ in normalized Euclidean distance. The normalization requires that the distance is relatively small for a particular coefficient if its unrestricted estimate is more precise, while the distance is relatively large for those coefficients that have more uncertainty in unrestricted estimates.

The Bayesian transformation approach can be performed in a computationally efficient way since we have a fast algorithm to solve the optimization problem in (3.14). We fix $\gamma_{p+1}^\star$ and divide the minimization function (3.14) into $p+1$ two-dimensional constrained minimization problems in which the solutions can be re-expressed as functions of $\gamma_{p+1}^\star$. After that, the whole minimization problem is reduced to a one-dimensional optimization problem in $\gamma_{p+1}^\star$, which can be easily solved using a one-dimensional optimization method. The constrained draws produced by the Bayesian transformation approach are not draws from the true posterior distribution, however, they are reasonable approximations that can be generated much faster.

3.3 Statistical Approaches when B is not Univariate Normal

3.3.1 *The Approximate Relationship Equation When B is Binary*

If $\mathbf{B}$ is a binary variable, the logistic regression approximation in Section 3.2 does not hold and the approximate relationship in equation (3.6) is not applicable. However, by Bayes theorem, there is a relationship equation connecting $\Pr(\mathbf{Y}=1|\mathbf{X}), \Pr(\mathbf{Y}=1|\mathbf{X},\mathbf{B})$ and $\mathrm{f}(\mathbf{B}|\mathbf{X},\mathbf{Y})$, regardless

of the nature of $\mathbf{B}$ (continuous/categorical) (Grill et al., 2015; Satten and Kupper, 1993):

$$\frac{\Pr(\mathbf{Y}=1|\mathbf{X},\mathbf{B})}{\Pr(\mathbf{Y}=0|\mathbf{X},\mathbf{B})} = \frac{\mathrm{f}(\mathbf{B}|\mathbf{X},\mathbf{Y}=1)}{\mathrm{f}(\mathrm{B}|\mathbf{X},\mathbf{Y}=0)} \cdot \frac{\Pr(\mathbf{Y}=1|\mathbf{X})}{\Pr(\mathbf{Y}=0|\mathbf{X})} \tag{3.15}$$

In other words, when $\mathbf{B}$ is binary, we need to define a model for $\mathbf{B}|\mathbf{X},\mathbf{Y}$ instead of a model for $\mathbf{B}|\mathbf{X}$. Assume $\mathrm{logit}(\Pr(\mathbf{B}=1|\mathbf{X},\mathbf{Y})) = \sum_{\mathrm{j}=0}^{\mathrm{p}} \phi_{\mathrm{j}}\mathbf{X}_{\mathrm{j}} + \phi_{\mathrm{p}+1}\mathbf{Y}$. Some algebraic simplifications of equation (3.15) followed by a Taylor series expansion (as shown in Supplementary Material Appendix B) result in an approximate relationship equation: $\beta_0 + \beta_1\mathbf{X_1} + \cdots + \beta_{\mathrm{p}}\mathbf{X_p} \approx \gamma_0 + \frac{1}{2}\phi_{\mathrm{p}+1} + \frac{1}{4}\phi_0\phi_{\mathrm{p}+1} + \frac{1}{8}\phi_{\mathrm{p}+1}^2 + \sum_{\mathrm{j}=1}^{\mathrm{p}}(\gamma_{\mathrm{j}} + \frac{1}{4}\phi_{\mathrm{j}}\phi_{\mathrm{p}+1})\mathbf{X_j} + (\gamma_{\mathrm{p}+1} - \phi_{\mathrm{p}+1})\mathbf{B}$. Then the approximate relationship between $\boldsymbol{\gamma}$, $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$ is:

$$\begin{cases} \beta_0 \approx \gamma_0 + \frac{1}{2}\phi_{\mathrm{p}+1} + \frac{1}{4}\phi_0\phi_{\mathrm{p}+1} + \frac{1}{8}\phi_{\mathrm{p}+1}^2 \\ \beta_j \approx \gamma_j + \frac{1}{4}\phi_j\phi_{p+1}, j = 1, \ldots, p \\ \gamma_{p+1} = \phi_{p+1} \end{cases} \tag{3.16}$$

3.3.2 *Unconstrained Solutions*

The two unconstrained solutions, direct regression and standard Bayes can be performed in the same way described in Section 3.2 regardless of the distribution of $\mathbf{B}$.

3.3.3 *Constrained Solutions*

To develop a constrained solution, we need to first define the likelihood function $\mathrm{L}(\mathbf{B}|\mathbf{X})$. It can be shown that $\Pr(\mathrm{B_i}=1|\mathbf{X_i})$ is a weighted average of $\frac{\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}} \mathrm{X_{ij}}\phi_{\mathrm{j}})}{1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}} \mathrm{X_{ij}}\phi_{\mathrm{j}})}$ and $\frac{\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}} \mathrm{X_{ij}}\phi_{\mathrm{j}}+\phi_{\mathrm{p}+1})}{1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}} \mathrm{X_{ij}}\phi_{\mathrm{j}}+\phi_{\mathrm{p}+1})}$. We use estimates of the weights given by $w_{i,\bar{\boldsymbol{\beta}}} = \frac{1}{1+\exp(\mathbf{X}_i\bar{\boldsymbol{\beta}})}$ and $1 - w_{i,\bar{\boldsymbol{\beta}}} = \frac{\exp(\mathbf{X}_i\bar{\boldsymbol{\beta}})}{1+\exp(\mathbf{X}_i\bar{\boldsymbol{\beta}})}$ where $\bar{\boldsymbol{\beta}}$ are the values for $\boldsymbol{\beta}$ from the established model. Then $\mathrm{L}(\mathbf{B}|\mathbf{X})$ can be written as: $\mathrm{L}(\mathbf{B}|\mathbf{X}) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{L}(\mathrm{B_i}|\mathbf{X_i},\boldsymbol{\phi}) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \big(\frac{\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}} \mathrm{X_{ij}}\phi_{\mathrm{j}})}{1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}} \mathrm{X_{ij}}\phi_{\mathrm{j}})} \cdot \mathrm{w}_{\mathrm{i},\bar{\boldsymbol{\beta}}} + \frac{\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}} \mathrm{X_{ij}}\phi_{\mathrm{j}}+\phi_{\mathrm{p}+1})}{1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}} \mathrm{X_{ij}}\phi_{\mathrm{j}}+\phi_{\mathrm{p}+1})} \cdot (1-\mathrm{w}_{\mathrm{i},\bar{\boldsymbol{\beta}}})\big)^{\mathrm{B_i}} \cdot \big(\frac{1}{1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}} \mathrm{X_{ij}}\phi_{\mathrm{j}})} \cdot \mathrm{w}_{\mathrm{i},\bar{\boldsymbol{\beta}}} + \frac{1}{1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}} \mathrm{X_{ij}}\phi_{\mathrm{j}}+\phi_{\mathrm{p}+1})} \cdot (1 - \mathrm{w}_{\mathrm{i},\bar{\boldsymbol{\beta}}})\big)^{(1-\mathrm{B_i})}$.

Constrained Maximum Likelihood

The constrained ML estimation optimizes the following joint log-likelihood $\mathrm{L}(\mathbf{Y}|\mathbf{X},\mathbf{B})\mathrm{L}(\mathbf{B}|\mathbf{X})$ with a set of constraints on $\boldsymbol{\gamma}, \boldsymbol{\phi}$, namely:

$$\begin{aligned}
\min_{\boldsymbol{\gamma},\boldsymbol{\phi}} \Big\{ & \sum_{\mathrm{i}=1}^{\mathrm{n}}[-\mathrm{Y_i}(\sum_{\mathrm{j}=0}^{\mathrm{p}}\gamma_{\mathrm{j}}\mathrm{X_{ij}}+\gamma_{\mathrm{p+1}}\mathrm{B_i})+\log(1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}}\gamma_{\mathrm{j}}\mathrm{X_{ij}}+\gamma_{\mathrm{p+1}}\mathrm{B_i}))] \\
& -\sum_{\mathrm{i}=1}^{\mathrm{n}}\Big[\mathrm{B_i}\log\Big(\frac{\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}}\mathrm{X_{ij}}\phi_{\mathrm{j}})\mathrm{w}_{\mathrm{i},\bar{\boldsymbol{\beta}}}}{1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}}\mathrm{X_{ij}}\phi_{\mathrm{j}})}+\frac{\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}}\mathrm{X_{ij}}\phi_{\mathrm{j}}+\phi_{\mathrm{p+1}})(1-\mathrm{w}_{\mathrm{i},\bar{\boldsymbol{\beta}}})}{1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}}\mathrm{X_{ij}}\phi_{\mathrm{j}}+\phi_{\mathrm{p+1}})}\Big) \\
& +(1-\mathrm{B_i})\log\Big(\frac{\mathrm{w}_{\mathrm{i},\bar{\boldsymbol{\beta}}}}{1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}}\mathrm{X_{ij}}\phi_{\mathrm{j}})}+\frac{(1-\mathrm{w}_{\mathrm{i},\bar{\boldsymbol{\beta}}})}{1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}}\mathrm{X_{ij}}\phi_{\mathrm{j}}+\phi_{\mathrm{p+1}})}\Big)\Big]\Big\} \qquad (3.17)
\end{aligned}$$

$$\text{s.t.}\begin{cases}
\gamma_0+\frac{1}{2}\phi_{p+1}+\frac{1}{4}\phi_0\phi_{p+1}+\frac{1}{8}\phi_{p+1}^2\in[\bar{\beta}_0-d\bar{S}_0,\bar{\beta}_0+d\bar{S}_0] \\
\gamma_j+\frac{1}{4}\phi_j\phi_{p+1}\in[\bar{\beta}_j-d\bar{S}_j,\bar{\beta}_j+d\bar{S}_j], j=1,\ldots,p \\
\gamma_{p+1}=\phi_{p+1}
\end{cases}$$

We also consider a modification to the above constrained ML solution that adds the Firth penalty term to the objective function.

Informative Full Bayes

Analogous to the derivation of the informative full Bayes solution described in Section 3.2, we first write down the product of $\mathrm{L}(\mathbf{Y}|\mathbf{X},\mathbf{B})$ and $\mathrm{L}(\mathbf{B}|\mathbf{X})$ with priors.

$$\begin{aligned}
\mathrm{p}(\boldsymbol{\gamma},\boldsymbol{\phi}|\mathrm{data}) \propto \Big\{ & \prod_{\mathrm{i}=1}^{\mathrm{n}}\frac{\exp((\sum_{\mathrm{j}=0}^{\mathrm{p}}\gamma_{\mathrm{j}}\mathrm{X_{ij}}+\gamma_{\mathrm{p+1}}\mathrm{B_i})\mathrm{Y_i})}{1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}}\gamma_{\mathrm{j}}\mathrm{X_{ij}}+\gamma_{\mathrm{p+1}}\mathrm{B_i})}\cdot \\
& \Big(\frac{\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}}\mathrm{X_{ij}}\phi_{\mathrm{j}})\mathrm{w}_{\mathrm{i},\bar{\boldsymbol{\beta}}}}{1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}}\mathrm{X_{ij}}\phi_{\mathrm{j}})}+\frac{\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}}\mathrm{X_{ij}}\phi_{\mathrm{j}}+\phi_{\mathrm{p+1}})(1-\mathrm{w}_{\mathrm{i},\bar{\boldsymbol{\beta}}})}{1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}}\mathrm{X_{ij}}\phi_{\mathrm{j}}+\phi_{\mathrm{p+1}})}\Big)^{\mathrm{B_i}}\cdot \\
& \Big(\frac{\mathrm{w}_{\mathrm{i},\bar{\boldsymbol{\beta}}}}{1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}}\mathrm{X_{ij}}\phi_{\mathrm{j}})}+\frac{(1-\mathrm{w}_{\mathrm{i},\bar{\boldsymbol{\beta}}})}{1+\exp(\sum_{\mathrm{j}=0}^{\mathrm{p}}\mathrm{X_{ij}}\phi_{\mathrm{j}}+\phi_{\mathrm{p+1}})}\Big)^{(1-\mathrm{B_i})}\Big\}\cdot\pi(\boldsymbol{\gamma},\boldsymbol{\phi})
\end{aligned} \qquad (3.18)$$

We can re-parametrize (3.18) in terms of $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, and include a Jacobian corresponding to this transformation. We denote the Jacobian matrix by $\mathbf{M}$ where $|\mathbf{M}| = |\frac{4}{\gamma_{p+1}}|^{p+1}$. We again specify independent weakly informative Cauchy priors for $\boldsymbol{\gamma}$ and use the constraints directly as priors for $\boldsymbol{\beta}$. Then we can rewrite the joint distribution in terms of $\boldsymbol{\gamma},\boldsymbol{\beta},\mathbf{k}^2$ as $\mathrm{p}(\boldsymbol{\gamma},\boldsymbol{\beta},\mathbf{k}^2|\mathbf{Y},\mathbf{X},\mathbf{B}) \propto$

$$\Bigg\{ \prod_{i=1}^{n} \frac{\exp((\sum_{j=0}^{p} \gamma_j X_{ij} + \gamma_{p+1} B_i) Y_i)}{1+\exp(\sum_{j=0}^{p} \gamma_j X_{ij} + \gamma_{p+1} B_i)} \cdot$$

$$\left[\frac{w_{i,\bar{\boldsymbol{\beta}}}}{1+\exp\left(-\left(\frac{4\beta_0 - 4\gamma_0 - 2\gamma_{p+1} - \frac{1}{2}\gamma_{p+1}^2}{\gamma_{p+1}} + \sum_{j=1}^{p} X_{ij} \frac{4(\beta_j - \gamma_j)}{\gamma_{p+1}}\right)\right)} + \frac{1 - w_{i,\bar{\boldsymbol{\beta}}}}{1+\exp\left(-\left(\frac{4\beta_0 - 4\gamma_0 - 2\gamma_{p+1} - \frac{1}{2}\gamma_{p+1}^2}{\gamma_{p+1}} + \sum_{j=1}^{p} X_{ij} \frac{4(\beta_j - \gamma_j)}{\gamma_{p+1}} + \gamma_{p+1}\right)\right)} \right]^{B_i} \cdot$$

$$\left[\frac{w_{i,\bar{\boldsymbol{\beta}}}}{1+\exp\left(\frac{4\beta_0 - 4\gamma_0 - 2\gamma_{p+1} - \frac{1}{2}\gamma_{p+1}^2}{\gamma_{p+1}} + \sum_{j=1}^{p} X_{ij} \frac{4(\beta_j - \gamma_j)}{\gamma_{p+1}}\right)} + \frac{1 - w_{i,\bar{\boldsymbol{\beta}}}}{1+\exp\left(\frac{4\beta_0 - 4\gamma_0 - 2\gamma_{p+1} - \frac{1}{2}\gamma_{p+1}^2}{\gamma_{p+1}} + \sum_{j=1}^{p} X_{ij} \frac{4(\beta_j - \gamma_j)}{\gamma_{p+1}} + \gamma_{p+1}\right)} \right]^{(1-B_i)} \Bigg\} \cdot$$

$$\pi(\boldsymbol{\beta}) \cdot \left\{ \prod_{j=0}^{p+1} \frac{1}{\sqrt{2\pi k_j^2}} \exp(-\frac{\gamma_j^2}{2k_j^2}) \right\} \cdot \left\{ \prod_{j=0}^{p+1} \pi(k_j^2) \right\} \cdot |\mathbf{M}|$$

After some algebraic calculations, the full conditional distributions of $k_0^2, k_1^2, \ldots, k_{p+1}^2$ are inverse chi-square, each with distribution function $\text{Inv} - \chi^2(2, \frac{1}{2}(s_j^2 + \gamma_j^2)), s_0 = 10, s_1 = \cdots = s_{p+1} = 2.5, j = 0, \ldots, p+1$. The full conditional distributions of $\gamma_0, \ldots, \gamma_{p+1}$ and the full conditional distribution of $\beta_0, \ldots, \beta_p$ do not have closed form expressions. A Metropolis-Hastings sampling algorithm is needed.

Bayesian Transformation Approach

Suppose the raw draws from the non-informative Bayes method for $\mathbf{Y}|\mathbf{X}, \mathbf{B}$ are $\boldsymbol{\gamma}$ and the raw draws from non-informative Bayes method for $\mathbf{B}|\mathbf{X}, \mathbf{Y}$ are $\boldsymbol{\phi}$. The posterior variances are $s_{\gamma_j}^2, j = 0, \ldots, p+1$ and $s_{\phi_k}^2, k = 0, \ldots, p+1$. Then $\boldsymbol{\gamma}^\star, \boldsymbol{\phi}^\star$ are obtained by solving the following optimization problem:

$$\min_{\boldsymbol{\gamma}^\star, \boldsymbol{\phi}^\star} \left\{ d_{\text{NED}}^2(\boldsymbol{\gamma}, \boldsymbol{\gamma}^\star) + d_{\text{NED}}^2(\boldsymbol{\phi}, \boldsymbol{\phi}^\star) \right\} = \sum_{j=0}^{p+1} \frac{(\gamma_j - \gamma_j^\star)^2}{s_{\gamma_j}^2} + \sum_{k=0}^{p+1} \frac{(\phi_k - \phi_k^\star)^2}{s_{\phi_k}^2}$$

$$\text{s.t.} \begin{cases} \gamma_0^\star + \frac{1}{2}\phi_{p+1}^\star + \frac{1}{4}\phi_0^\star \phi_{p+1}^\star + \frac{1}{8}\phi_{p+1}^{\star 2} \in [\bar{\beta}_0 - d\bar{S}_0, \bar{\beta}_0 + d\bar{S}_0] \\ \gamma_j^\star + \frac{1}{4}\phi_j^\star \phi_{p+1}^\star \in [\bar{\beta}_j - d\bar{S}_j, \bar{\beta}_j + d\bar{S}_j], j = 1, \ldots, p \\ \gamma_{p+1}^\star = \phi_{p+1}^\star \end{cases} \tag{3.19}$$

3.3.4 *The Approximate Relationship Equation When* B *is Multivariate Gaussian*

Based on equation (3.5), when $\mathbf{B}|\mathbf{X}$ is multivariate normal with L dimensions, mean $\mathbf{X}\boldsymbol{\theta}$ and covariance matrix $\mathbf{V}_{L\times L}$, the approximate relationship between $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ is:

$$\beta_j \approx \frac{\gamma_j + \sum_{l=1}^{L} \gamma_{p+l}\theta_{lj}}{(1+\boldsymbol{\gamma}_B^T\mathbf{V}\boldsymbol{\gamma}_B/1.7^2)^{\frac{1}{2}}}, j = 0, \ldots, p \tag{3.20}$$

Then the strategies to incorporate the external coefficient information described in Section 3.2 can be easily extended in this case.

3.4 Simulation Study

To evaluate the performance of the various approaches we conduct a simulation study with two main scenarios. The first simulation scenario has three predicting variables, $\mathbf{X_1}, \mathbf{X_2}$ and $\mathbf{B}$ where $\mathbf{B}$ is Gaussian distributed. The dataset from which to estimate model $\Pr(\mathbf{Y}=1|\mathbf{X},\mathbf{B})$ and $\mathrm{E}(\mathbf{B}|\mathbf{X})$ is of size 55. Five hundred replicate datasets are generated. $\mathrm{Y_i}$ is Bernoulli distributed with $\mathrm{logit}(\Pr(\mathrm{Y_i}=1|\mathrm{X_{i1}},\mathrm{X_{i2}},\mathrm{B_i})) = 2+3\mathrm{X_{i1}}+3\mathrm{X_{i2}}+2\mathrm{B_i}$. $\mathrm{X_{i1}}, \mathrm{X_{i2}}$ are independently and identically distributed on $\mathrm{U}(-0.75, 0.25)$ and $\mathrm{B_i}$ is simulated as $\mathrm{B_i} = 0.5\mathrm{X_{i1}}+0.5\mathrm{X_{i2}}+\mathrm{N}(0, 0.75^2)$. A logistic regression based on a large dataset of 10000 subjects gives estimates for the model $\mathrm{logit}(\Pr(\mathbf{Y}=1|\mathbf{X})) = \beta_0+\beta_1\mathbf{X_1}+\beta_2\mathbf{X_2}$. The estimates and standard errors from this fit are $\bar{\beta}_0 = 1.50, \bar{S}_0 = 0.04, \bar{\beta}_1 = 2.95, \bar{S}_1 = 0.09, \bar{\beta}_2 = 3.01, \bar{S}_2 = 0.09$.

The second simulation scenario has three predicting variables $\mathbf{X_1}, \mathbf{X_2}, \mathbf{B}$ where $\mathbf{B}$ is binary. There are 75 observations in each dataset. 500 datasets are generated. $\mathrm{Y_i}$ is Bernoulli distributed with $\mathrm{logit}(\Pr(\mathrm{Y_i}=1|\mathrm{X_{i1}},\mathrm{X_{i2}},\mathrm{B_i})) = 2+4\mathrm{X_{i1}}+4\mathrm{X_{i2}}+2\mathrm{B_i}$. $\mathrm{X_{i1}}, \mathrm{X_{i2}}$ are independently and identically distributed on $\mathrm{U}(-0.75, 0.25)$ and $\mathrm{B_i}$ is simulated from $\mathrm{logit}(\Pr(\mathrm{B_i}=1|\mathrm{X_{i1}},\mathrm{X_{i2}}) = 1+\mathrm{X_{i1}}+\mathrm{X_{i2}}$. A logistic regression based on a large dataset of 10000 subjects gives estimates for the model $\mathrm{logit}(\Pr(\mathbf{Y}=1|\mathbf{X})) = \beta_0+\beta_1\mathbf{X_1}+\beta_2\mathbf{X_2}$. The estimates and standard errors are $\bar{\beta}_0 = 2.97, \bar{S}_0 = 0.06, \bar{\beta}_1 = 3.87, \bar{S}_1 = 0.10, \bar{\beta}_2 = 3.68, \bar{S}_2 = 0.10$.

We report three evaluation metrics related to estimation accuracy: the average of estimated coefficient, relative efficiency of estimated coefficient and mean squared error across 500 replicates. The average of estimated coefficient is defined as: $\bar{\gamma}_j = \frac{1}{500}\sum_{m=1}^{500}\hat{\gamma}_{m,j}$; the relative efficiency of estimated coefficient is defined as: $\frac{\mathrm{V}(\hat{\gamma}_{\mathrm{j,direct}})}{\mathrm{V}(\hat{\gamma}_{\mathrm{j,method}})}$ where $\mathrm{V}(\hat{\gamma}_{\mathrm{j,direct}}) = \frac{1}{500}\sum_{\mathrm{m}=1}^{500}(\hat{\gamma}_{\mathrm{m,j}} - \bar{\gamma}_{\mathrm{j}})^2$ estimated by direct regression; the MSE of estimated coefficient is defined as: $\frac{1}{500}\sum_{\mathrm{m}=1}^{500}(\hat{\gamma}_{\mathrm{m,j}} - \gamma_{\mathrm{j}})^2, \mathrm{j} = 1, \ldots, \mathrm{p}+1$.

The prediction ability of logistic prediction models can be assessed using a variety of methods and metrics, to assess the overall model performance, to quantify the level of agreement between the predicted probability and the observed outcome, or to calculate the goodness-of-fit statistics for calibration (Steyerberg et al., 2010). In this simulation study, we assess the prediction ability of the model by Brier score ($\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{Y_i}-\hat{\mathrm{p}}_{\mathrm{i}})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{Y_i}-0.5)^2}$) and Hosmer-Lemeshow statistic in a validation dataset of size 800. To calculate Hosmer-Lemeshow statistic, the predicted probabilities in the validation dataset are sorted from lowest to highest and then separated into 10 groups of approximately equal size. For each group, the observed numbers of events and non-events and the expected numbers of events and non-events are calculated, where the expected number of events is the sum of the predicted probability in the group. Then Hosmer-Lemeshow statistic is: $\sum_{\mathrm{k}=0}^{1}\sum_{\mathrm{l}=1}^{10}\frac{(\mathrm{O_{kl}}-\mathrm{E_{kl}})^2}{\mathrm{E_{kl}}}$.

Table 3.1 presents the Monte Carlo simulation results for the first simulation scenario. Note the heavy bias in $\hat{\gamma}_1$ and $\hat{\gamma}_2$ in direct regression and non-informative Bayes solutions. By including a Firth penalty, the direct regression successfully reduces the bias inherent in the MLE in direct regression. This reduction in bias in $\hat{\gamma}_1$ and $\hat{\gamma}_2$ is even greater after the constraints on coefficients $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ are applied. By looking at the relative efficiency of regression parameters γ_1 and γ_2, we find the constrained methods greatly improve the estimation efficiency of coefficients of $\mathbf{X}$. The constrained ML can reduce the MSE of $\mathbf{X_1}, \mathbf{X_2}$ by more than 70%. The constrained ML with Firth penalty, the informative full Bayes and the Bayesian transformation approach can reduce the MSE of $\mathbf{X}$ by more than 80%. On the contrary, these constrained solutions can only reduce the MSE of

Table 3.1: Simulation results of the first scenario for Gaussian $\mathbf{B}$: for each method, we report mean (relative efficiency w.r.t. direct regression), MSE, average Brier score, average Hosmer-Lemeshow statistic and computing time for 100 datasets

Method	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	Brier Score	HL	Time
True value	3	3	2			
Direct regression	3.37(1.00)	3.40(1.00)	2.35(1.00)	0.661	253.1	0.68
MSE	3.36	3.48	0.96			
Direct regression + Firth	2.89(1.54)	2.92(1.51)	1.99(1.69)	0.651	78.3	1.25
MSE	2.10	2.18	0.49			
Constrained ML	3.08(3.59)	3.17(3.91)	2.30(1.09)	0.628	93.7	30.70
MSE	0.90	0.88	0.84			
Constrained ML + Firth	2.88(6.08)	2.97(6.22)	1.96(1.84)	0.622	36.9	30.61
MSE	0.55	0.53	0.45			
Non-informative Bayes	2.72(1.80)	2.75(1.76)	2.04(1.74)	0.647	40.5	1.26
MSE	1.88	1.93	0.47			
Informative full Bayes	2.87(4.94)	2.98(5.31)	2.30(1.33)	0.624	35.0	1448.06
MSE	0.66	0.63	0.71			
Transformation	2.89(6.81)	3.00(6.96)	1.93(1.74)	0.622	22.2	339.25
MSE	0.48	0.47	0.48			

γ_3 by 13% (constrained ML) – 53% (constrained ML with Firth penalty). Among all methods, the Bayesian transformation approach has the highest prediction power, as it has the lowest values in both Brier score (0.622) and Hosmer-Lemeshow statistic (22.2). The second best is the informative full Bayes. In terms of computational efficiency, the Bayesian transformation approach takes about 23% the time of the informative full Bayes approach.

Table 3.2 summarizes the simulation results for the second simulation scenario, where $\mathbf{B}$ is binary. The constrained methods in this simulation scenario exhibit greater improvement in estimating efficiency for the coefficients of $\mathbf{X}$ than in the first simulation scenario. The constrained ML with Firth penalty and Bayesian transformation approach can improve the relative efficiency of parameters γ_1 and γ_2 by more than 800%. Among all methods, constrained ML with Firth penalty, informative full Bayes and Bayesian transformation approach have the highest prediction power (their Hosmer-Lemeshow statistic values are 15.7, 14.5 and 15.8). In this simulation scenario, the informative full Bayes method is more computationally intensive since both the conditional distri-

Table 3.2: Simulation results of the second scenario for binary B: for each method, we report mean (relative efficiency w.r.t. direct regression), MSE, average Brier score, average Hosmer-Lemeshow statistic and computing time for 100 datasets

Method	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	Brier Score	HL	Time
True value	4	4	2			
Direct regression	4.49(1.00)	4.40(1.00)	2.22(1.00)	0.560	404.0	1.03
MSE	3.48	3.22	0.82			
Direct regression + Firth	3.98(1.48)	3.90(1.50)	2.00(1.45)	0.554	66.3	1.47
MSE	2.18	2.05	0.53			
Constrained ML	4.08(5.19)	3.94(3.87)	2.13(1.02)	0.537	145.5	23.69
MSE	0.64	0.79	0.77			
Constrained ML + Firth	3.93(11.98)	3.77(11.33)	1.80(1.95)	0.535	15.7	44.70
MSE	0.27	0.32	0.44			
Non-informative Bayes	3.78(1.75)	3.70(1.78)	1.92(1.67)	0.552	31.0	1.39
MSE	1.90	1.81	0.47			
Informative full Bayes	3.91(9.00)	3.75(8.51)	1.95(1.49)	0.534	14.5	8842.90
MSE	0.37	0.42	0.52			
Transformation	3.91(9.31)	3.76(9.43)	1.96(1.63)	0.534	15.8	209.75
MSE	0.35	0.38	0.48			

butions of $\boldsymbol{\gamma}$ and the conditional distributions of $\boldsymbol{\beta}$ do not have closed form expressions. Drawing samples based on a Metropolis-Hasting sampling algorithm in this case is computationally demanding.

3.5 Application to the Prostate Cancer Data

We demonstrate the methods on data from a prostate cancer study in which the samples were collected from multiple clinical sites throughout the United States (Tomlins et al., 2015). The use of Serum prostate-specific antigen to screen for prostate cancer (PCa) is controversial because the test has low specificity and can lead to overtreatment. Therefore, improved tests that use additional information are needed. The Prostate Cancer Prevention Trial Risk Calculator for prostate cancer, and a separate calculator for high-grade (Gleason grade $\geq$ 7) prostate cancer (PCPThg), were the first online prostate cancer risk assessment tools to allow an individual to assess his risk for prostate cancer. These calculators were developed from 5519 men from the placebo arm of the Prostate Cancer Prevention Trial. PCPThg predicts the chance of

detecting high-grade cancer based on PSA, age, DRE findings, prior biopsy history and race (Thompson et al., 2006). The estimated logistic models are available on the PCPTrc website (http://deb.uthscsa.edu/URORiskCalc/Pages/calcs.jsp) and the estimated coefficients and the estimated covariance-variance matrices are also accessible.

Prostate cancer antigen 3 (PCA3) and TMPRSS2:ERG (T2:ERG) gene fusions are two prostate cancer early detection biomarkers which have been shown to have better specificity for PCa than PSA (Truong et al., 2013), (Tomlins et al., 2015). Their transcripts are detectable and quantifiable in urine collected after digital rectal examination. To investigate whether PCA3 and T2:ERG could be combined with the PCPThg calculator to give more accurate risk prediction, Tomlins et al. (2015) undertook a study in 679 men, in whom all the PCPThg calculator variables and both a PCA3 score and a T2:ERG score were measured. An independent validation study of 1218 men was also available.

Tomlins et al. (2015) expanded the PCPThg model by incorporating PCA3 as an additional risk factor. They used the predicted risk score from the PCPThg (i.e., $\hat{\Pr}(\mathrm{Y_i} = 1|\mathbf{X_i}, \bar{\beta}_{\mathrm{PCPThg}}) \times 100$) directly as a predicting variable and estimated the joint effect of the PCPThg risk score and the PCA3 value on the high-grade cancer. They estimated the new model in the training dataset, and found that when applied to the validation dataset the AUC for predicting high-grade PCa increased from 0.707 for the PCPThg model to 0.752 for their model. Tomlins et al. (2015) also constructed another expanded PCPThg model by incorporating T2:ERG as an additional risk factor and showed that the AUC increased from 0.707 to 0.754.

Using the data from Tomlins et al. (2015) we will illustrate the methods described in this paper to develop a logistic model that includes all the PCPThg variables and PCA3. We estimate the new model from the training dataset of 679 men, incorporating the known coefficients and their standard errors from the PCPThg calculator. After a transformation ($\log_2(\mathrm{PCA3} + 1)$) the distribution of PCA3 is roughly normally distributed in both cohorts and thus the approximate relationship

equations (3.6) are applicable. The distribution of T2:ERG looks like a truncated normal whose value is bounded below at zero, with many observations equal to zero. We dichotomized T2:ERG by splitting at the median and develop a logistic model that includes PCPThg variables and a dichotomized T2:ERG. The approximate relationship equations (3.16) would be appropriate in this case.

These two expanded PCPThg models will be estimated by both the unconstrained methods and the constrained methods described in Section 3.2 and Section 3.3. For comparing coefficient estimation across different methods, we report the estimated coefficients and their standard errors calculated from the training dataset. For comparing prediction power, we calculate Brier Score, Hosmer-Lemeshow statistic and AUC based on the validation dataset. We also present the original PCPThg model and the expanded model developed by Tomlins et al. (2015). We give the calibration plots for the original PCPThg model, the expanded model by Tomlins et al. (2015), the expanded PCPThg model estimated without constraints (direct regression) and the expanded PCPThg model estimated with constraints (Bayesian transformation approach). The calibration plot contains the predicted risk of high-grade cancer and the observed risk of high-grade cancer in the 10 equal-sized groups defined in the Hosmer-Lemeshow statistic formula. Perfect predictions should be on the 45° line.

Table 3.3 presents the expanded PCPThg model incorporating these two biomarkers fitted to the training dataset. For the expanded PCPThg model incorporating PCA3 score, if we compare the standard errors across different methods, it is easily seen that the constrained methods can reduce the standard errors of regression coefficients compared to direct regression. For example, the informative full Bayes solution can substantially reduce the standard errors in parameters of variables PSA (0.10 vs 0.19), age (0.008 vs 0.013), DRE findings (0.17 vs 0.27), prior biopsy history (0.17 vs 0.28) and race (0.22 vs 0.31). The constrained ML with Firth penalty can reduce the standard errors of the parameters of variables PSA, age, prior biopsy history and race by at least

Table 3.3: Expanded PCPThg model: for each method, point estimate (standard error) from the training dataset, and Brier score, Hosmer-Lemeshow statistic and AUC from the validation dataset. The sample size of the training dataset is 679. The sample size of the validation dataset is 1218. The methods with the lowest HL and Brier score are bolded

Model	log(PSA)	Age	DRE findings	Prior biopsy history	Race		Brier Score	HL	AUC
Original PCPThg	1.29(0.09)	0.031(0.012)	1.00(0.17)	-0.36(0.18)	0.96(0.27)	–	0.558	43.7	0.707
Estimated PCPThg	1.06(0.18)	0.033(0.012)	1.15(0.26)	-1.44(0.27)	0.44(0.29)	–	0.583	70.7	0.716
Expanded model with $\log_2(\mathrm{PCA3}+1)$						PCA3			
PCPThg score+PCA3	–	–	–	–	–	–	0.568	80.0	0.752
Direct regression	1.00(0.19)	0.009(0.013)	1.07(0.27)	-1.30(0.28)	0.04(0.31)	0.56(0.08)	0.568	91.7	0.767
Non-informative Bayes	0.98(0.18)	0.009(0.013)	1.05(0.27)	-1.27(0.27)	0.04(0.30)	0.56(0.08)	0.568	90.3	0.767
Constrained ML	1.20(0.09)	0.010(0.007)	1.08(0.14)	-0.55(0.13)	0.30(0.19)	0.59(0.08)	0.570	96.5	0.766
Direct regression + Firth	0.97(0.19)	0.009(0.013)	1.06(0.27)	-1.27(0.27)	0.05(0.31)	0.56(0.08)	0.568	91.7	0.767
Constrained ML + Firth	1.19(0.09)	0.012(0.006)	1.08(0.14)	-0.54(0.13)	0.47(0.11)	0.53(0.07)	0.567	87.1	0.764
Informative full Bayes	1.23(0.10)	0.009(0.008)	0.99(0.17)	-0.73(0.17)	0.26(0.22)	0.60(0.08)	0.567	92.3	0.767
Transformation	1.23(0.07)	0.008(0.009)	0.96(0.14)	-0.50(0.13)	0.41(0.19)	0.55(0.08)	**0.528**	**28.1**	0.765
Expanded model with binary T2:ERG						T2:ERG			
PCPThg score + T2:ERG	–	–	–	–	–	–	0.558	50.7	0.732
Direct regression	1.01(0.18)	0.032(0.012)	1.03(0.26)	-1.44(0.28)	0.57(0.29)	0.77(0.20)	0.556	51.3	0.745
Non-informative Bayes	0.99(0.18)	0.032(0.012)	1.01(0.26)	-1.40(0.27)	0.55(0.29)	0.76(0.20)	0.555	52.1	0.745
Constrained ML	1.14(0.07)	0.032(0.004)	1.06(0.14)	-0.52(0.11)	0.81(0.18)	0.74(0.21)	0.555	54.7	0.742
Direct regression + Firth	0.98(0.18)	0.032(0.012)	1.02(0.26)	-1.41(0.27)	0.57(0.29)	0.76(0.20)	0.556	53.3	0.744
Constrained ML + Firth	1.14(0.07)	0.032(0.004)	1.06(0.14)	-0.52(0.11)	0.80(0.17)	0.72(0.20)	0.557	53.6	0.742
Informative full Bayes	1.14(0.09)	0.033(0.007)	0.95(0.14)	-0.76(0.16)	0.77(0.21)	0.73(0.19)	0.552	50.5	0.745
Transformation	1.17(0.07)	0.030(0.007)	0.94(0.12)	-0.50(0.11)	0.89(0.16)	0.74(0.14)	**0.532**	**18.7**	0.742

50%. Therefore, it is easier to identify statistically significant predictors based on these constrained methods.

Among the 1218 validation cohort patients, AUC for PCPThg model and the expanded PCPThg score plus PCA3 model are 0.707 and 0.752. By incorporating PCA3 score in the PCPThg model, the AUC increases to 0.767 in direct regression. However, the constrained methods cannot further increase the AUC. In terms of calibration as measured through HL, there is no benefit in incorporating PCA3 in all solutions except the Bayesian transformation approach. In Figure 3.1 we can see that the expanded PCPThg model incorporating PCA3 tends to overestimate the risk of getting high-grade cancer among those patients with high risk so the calibration is especially imperfect in the high-risk groups. However, the overall calibration ability of the expanded PCPThg model estimated by the Bayesian transformation approach still outperforms that of the original PCPThg model, the expanded PCPCThg score plus PCA3 model or the expanded PCPThg model estimated

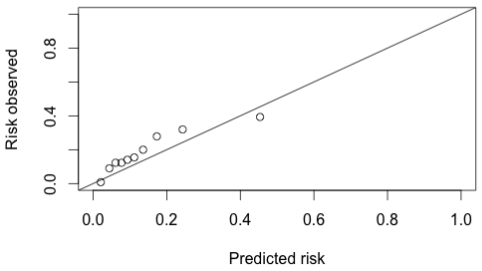
by direct regression.

The expanded PCPThg model incorporating binary T2:ERG fitted to the training dataset again shows that the constrained methods can reduce the standard errors of regression coefficients compared to direct regression. For example, in Table 3.3 the Bayesian transformation approach solution can reduce the standard errors in parameters of variables PSA (0.07 vs 0.18), age (0.007 vs 0.012), DRE findings (0.12 vs 0.26), prior biopsy history (0.11 vs 0.28) and race (0.16 vs 0.29). In Figure 3.1 we can see that the expanded PCPThg model incorporating binary T2:ERG tends to overestimate the risk of getting high-grade cancer among those patients with high risk so the calibration is especially imperfect in the high-risk groups. However, the Bayesian transformation approach predicts the risk well for the high risk groups.

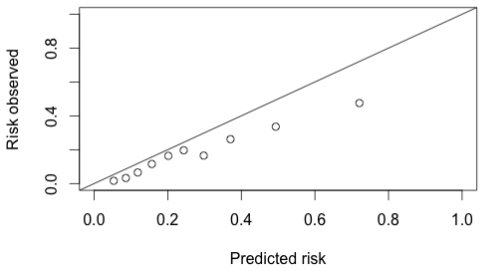
3.6 Discussion

In this study, we propose several strategies for translating the external coefficient information obtained from outside the dataset into constraints on regression coefficients in the setting of a logistic regression model describing $\Pr(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B})$. Simulation studies show that the external coefficient information from the established model can help improve the efficiency of estimation and enhance the predictive power in the expanded model.

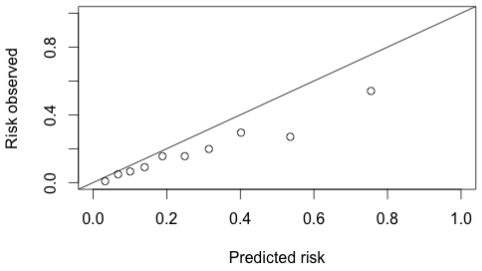
In terms of computational efficiency, in simulation studies the Bayesian transformation approach shows advantage over the informative full Bayes because in the Bayesian transformation approach the raw draws are first obtained in a fast and easy way and then transformed into draws that obey the constraints based on a cost-effective optimization algorithm, while the informative full Bayes solution produces constrained draws inefficiently. When the dimensionality of the predictors $\mathbf{X}$ increases, the computational cost of the Bayesian transformation approach solution will not increase much because the high-dimensional optimization problem that is the key to performing the Bayesian transformation approach will always reduce to a one-dimensional optimization prob-



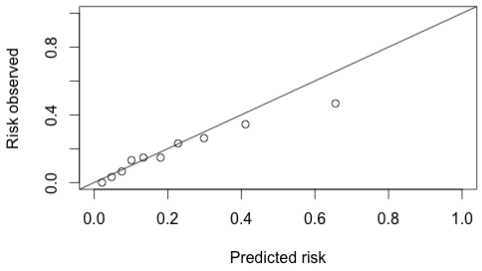
(a) PCPThg model



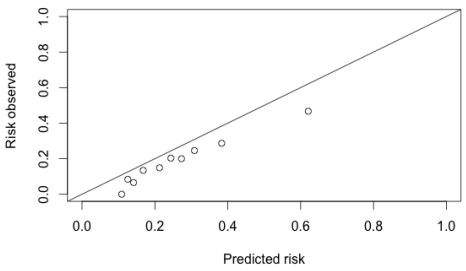
(b) PCPThg score + PCA3, Tomlins et al. (2015)



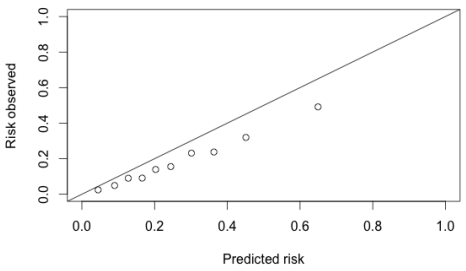
(c) PCPThg covariates + PCA3, direct regression



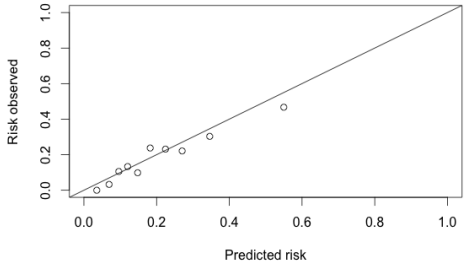
(d) PCPThg covariates + PCA3, Bayesian transformation approach



(e) PCPThg score + dichotomized T2:ERG



(f) PCPThg covariates + dichotomized T2:ERG, direct regression



(g) PCPThg covariates + dichotomized T2:ERG, Bayesian transformation approach

Figure 3.1: Calibration plot of the original high-grade Prostate Cancer Prevention Trial risk calculator (PCPThg) and calibration plots of the expanded PCPThg model by incorporating PCA3 score and dichotomized T2:ERG

lem based on our algorithm regardless of the dimensionality of the predictor space. Furthermore, the correlation among the samples in the Markov chain for the informative full Bayes approach is very high and effective samples are harder to obtain when the dimensionality increases (additional simulation results that validate this finding are not shown). As a consequence, the discrepancy of these two constrained solutions in computational cost will be more apparent in higher dimensions.

The efficiency gain in the expanded model of interest depends on the sample size used to construct the established model and the sample size used to estimate the expanded model of interest. In our simulation studies the established models are based on large datasets with 10000 observations while the current datasets are very small (55 observations in the first simulation scenario and 75 in the second simulation scenario). The relative efficiency gain in the regression coefficient of variables $\mathbf{X}$ by incorporating the external coefficient information is significant and the prediction power in the validation dataset is enhanced. However, when the sample size in the current dataset is large enough to estimate the expanded model, the constrained methods do not lead to much improvement in the prediction ability compared to direct regression, as was the case in the prostate cancer example. However, our numerical results suggest that improved precision of the coefficient estimates, as measured by standard errors, can be achieved even if the current dataset is not small.

It is worth mentioning that the differences in the distributions of $\mathbf{X}$ in the external study and the internal study will not affect the performances of our proposed constrained methods, as found in additional simulation studies (simulation results not shown). This is because the coefficients' approximate relationship equations are constructed based on the conditional distributions $\mathbf{Y}|\mathbf{X},\mathbf{B}$ and $\mathbf{B}|\mathbf{X}$. As long as these two conditional distributions are correctly specified in the internal study, the approximate relationship equations will hold regardless of the differences in the distributions of $\mathbf{X}$ and thus the distributions of $\mathbf{B}$ in these two studies.

One point of future consideration is the distribution of the new biomarker $\mathbf{B}$. We develop the approximate relationship equation for the scenarios that $\mathbf{B}$ is Gaussian and binary. However,

if additional biomarkers follow other distributions, these approximate relationship equations will fail. Therefore, further investigations are needed for the generalization of our proposed constrained solutions to flexibly adapt to other possible distributions of the new biomarker.

3.7 Supplementary Materials

3.7.1 *Appendix A*

Logistic Regression Approximation

The logistic-normal integral of the form $\mathrm{G}(\eta,\tau)=\int_{-\infty}^{+\infty}\mathrm{H(z)}\tau^{-1}\phi(\frac{\mathrm{z}-\eta}{\tau})\mathrm{dz}$ often appears in the studies of logistic regression model calibration where a subset of predicting variables are measured with errors. Monahan and Stefanski (1992) demonstrated a normal scale mixture representation of the logistic cumulative distribution function $\mathrm{H(z)}$, showing that $\mathrm{H(z)}$ can be approximated by a finite location-scale mixture of normal distribution functions: $\mathrm{H(z)}\approx\mathrm{H_k(z)}=\sum_{\mathrm{i=1}}^{\mathrm{k}}\mathrm{p_{k,i}}\Phi(\mathrm{z}\times\mathrm{s_{k,i}}),\mathrm{k}=1,2,\ldots,$ where $p_{k,i}$ is a fixed value and can be considered the weight of each normal CDF. $s_{k,i}$ is also a fixed value and can be considered as the corresponding scale parameter. All values of $p_{k,i}$ and $s_{k,i}$ can be found in their Least Maximum Approximants Table. Numerically studies show that this approximation is remarkably good for k as small as 3. In logistic regression calibration literature it is generally acceptable to take $k=1$. Based on the Least Maximum Approximants Table, the corresponding values of $p_{k,i}, s_{k,i}$ are $p_{1,1}=1$ and $s_{1,1}\approx 0.59$. Then we have the following conclusion:

$$\mathrm{H(z)}\approx\Phi(\mathrm{z}\times\mathrm{s_{1,1}})\approx\Phi(0.59\mathrm{z})\approx\Phi(\mathrm{z}/1.7) \tag{3.21}$$

Sketch Of Proof for Logistic Regression Approximation Equation (3.6)

Assume that $\mathbf{B}|\mathbf{X}$ is univariate normal with mean $\mathrm{m_B}=\mathbf{X}\boldsymbol{\theta}$ and variance σ_2^2.

$\Pr(\mathbf{Y}=1|\mathbf{X})=\int\mathrm{H}(\mathbf{X}\boldsymbol{\gamma}_{\mathrm{x}}+\mathbf{B}\gamma_{\mathrm{B}})\frac{1}{\sqrt{2\pi\sigma_2^2}}\mathrm{e}^{-\frac{(\mathbf{B}-\mathbf{X}\boldsymbol{\theta})^2}{2\sigma_2^2}}\mathrm{d}\mathbf{B}$

$=\int\mathrm{H(z)}\frac{1}{\sqrt{2\pi\sigma_2^2\gamma_B^2}}\mathrm{e}^{-\frac{(\frac{\mathrm{z}-\mathbf{X}\boldsymbol{\gamma_x}}{\gamma_{\mathrm{B}}}-\mathbf{X}\boldsymbol{\theta})^2}{2\sigma_2^2}}\mathrm{dz}$ by changing $\mathbf{X}\boldsymbol{\gamma}_x+\mathbf{B}\gamma_B$ to z

$$\approx \int \Phi(\mathrm{z} \times \mathrm{s}_{1,1}) \frac{1}{\sqrt{2\pi\sigma_2^2\gamma_B^2}} \mathrm{e}^{-\frac{(\frac{\mathrm{z}-\mathbf{X}\boldsymbol{\gamma}_\mathrm{x}}{\gamma_\mathrm{B}}-\mathbf{X}\boldsymbol{\theta})^2}{2\sigma_2^2}} \mathrm{dz}$$

$$= \int \Phi[(\mathbf{X}\boldsymbol{\gamma}_\mathrm{x} + (\mathbf{X}\boldsymbol{\theta})\gamma_\mathrm{B} + \mathrm{C}\gamma_\mathrm{B}\sigma_2)\mathrm{s}_{1,1}] \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\mathrm{C}^2}{2}} \mathrm{dC}$$

$$= \Phi\big(\frac{(\mathbf{X}\boldsymbol{\gamma}_\mathrm{x}+(\mathbf{X}\boldsymbol{\theta})\gamma_\mathrm{B})\mathrm{s}_{1,1}}{\sqrt{1+\gamma_B^2\sigma_2^2 s_{1,1}^2}}\big)$$

$$\approx \mathrm{H}\big(\frac{(\mathbf{X}\boldsymbol{\gamma}_\mathrm{x}+(\mathbf{X}\boldsymbol{\theta})\gamma_\mathrm{B})}{\sqrt{1+\gamma_B^2\sigma_2^2 s_{1,1}^2}}\big)$$

$$\approx \mathrm{H}\big(\frac{(\mathbf{X}\gamma_\mathrm{x}+(\mathbf{X}\boldsymbol{\theta})\gamma_\mathrm{B})}{\sqrt{1+\gamma_B^2\sigma_2^2/1.7^2}}\big)$$

Note that line five holds based on calculating the integral of product of a normal CDF and a standard normal PDF.

Sketch Of Proof for Logistic Regression Approximation Equation (3.5)

For the case that $\mathbf{B}|\mathbf{X}$ is multivariate normal with L dimensions with mean $\mathbf{X}\boldsymbol{\theta}$ and covariance matrix $\mathbf{V}_{L\times L}$, the derivation will be slightly different:

$$\Pr(\mathbf{Y}=1|\mathbf{X}=\mathbf{x}) = \int \Pr(\mathbf{Y}=1|\mathbf{X}=\mathbf{x},\mathbf{B}=\mathbf{b}) \frac{1}{(2\pi)^{\mathrm{L}/2}|\mathbf{V}|^{1/2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{b}-\mathbf{x}\boldsymbol{\theta})^\mathrm{T}\mathbf{V}^{-1}(\mathbf{b}-\mathbf{x}\boldsymbol{\theta})} \mathrm{d}\mathbf{b}$$

$$= \int \mathrm{H}(\mathbf{x}\boldsymbol{\gamma}_\mathrm{x} + \mathbf{b}\boldsymbol{\gamma}_\mathrm{B}) \frac{1}{(2\pi)^{\mathrm{L}/2}|\mathbf{V}|^{1/2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{b}-\mathbf{x}\boldsymbol{\theta})^\mathrm{T}\mathbf{V}^{-1}(\mathbf{b}-\mathbf{x}\boldsymbol{\theta})} \mathrm{d}\mathbf{b}$$

$$\approx \int \Phi(\mathrm{s}_{1,1}(\mathbf{x}\boldsymbol{\gamma}_\mathrm{x} + \mathbf{b}\boldsymbol{\gamma}_\mathrm{B})) \frac{1}{(2\pi)^{\mathrm{L}/2}|\mathbf{V}|^{1/2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{b}-\mathbf{x}\boldsymbol{\theta})^\mathrm{T}\mathbf{V}^{-1}(\mathbf{b}-\mathbf{x}\boldsymbol{\theta})} \mathrm{d}\mathbf{b}$$

$$= \int \Phi(\mathrm{s}_{1,1}(\mathbf{x}\boldsymbol{\gamma}_\mathrm{x} + (\mathbf{x}\boldsymbol{\theta} + \boldsymbol{\Sigma}\mathbf{c})\boldsymbol{\gamma}_\mathrm{B}))\phi(\mathbf{c})\mathrm{d}\mathbf{c} \text{ by changing } \mathbf{b} \text{ to } \mathbf{x}\boldsymbol{\theta} + \boldsymbol{\Sigma}\mathbf{c} \text{ where } \boldsymbol{\Sigma}^T\boldsymbol{\Sigma} = \mathbf{V}$$

$$= \int \Pr(\mathbf{W} \le \mathrm{s}_{1,1}(\mathbf{X}\boldsymbol{\gamma}_\mathrm{x} + (\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\Sigma}\mathbf{C})\boldsymbol{\gamma}_\mathrm{B})|\mathbf{X}=\mathbf{x},\mathbf{C}=\mathbf{c})\phi(\mathbf{c})\mathrm{d}\mathbf{c} \text{ where } \mathbf{W} \text{ is a standard normal random variable and is independent of } \mathbf{C}$$

$$= \Pr(\mathbf{W} \le \mathrm{s}_{1,1}(\mathbf{X}\boldsymbol{\gamma}_\mathrm{x} + (\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\Sigma}\mathbf{C})\boldsymbol{\gamma}_\mathrm{B})|\mathbf{X}=\mathbf{x}) \text{ by the law of total probability}$$

$$= \Pr(-\mathrm{s}_{1,1}(\mathbf{X}\boldsymbol{\gamma}_\mathrm{X} + \mathbf{X}\boldsymbol{\theta}\boldsymbol{\gamma}_\mathrm{B}) \le \mathrm{s}_{1,1}\boldsymbol{\Sigma}\mathbf{C}\boldsymbol{\gamma}_\mathrm{B} - \mathbf{W}|\mathbf{X}=\mathbf{x})$$

Let $\mathbf{Z} = s_{1,1}\boldsymbol{\Sigma}\mathbf{C}\boldsymbol{\gamma}_B - \mathbf{W}$. Then $\mathbf{Z} \sim \mathrm{N}(0, \mathrm{s}_{1,1}^2\boldsymbol{\gamma}_\mathrm{B}^\mathrm{T}\mathbf{V}\boldsymbol{\gamma}_\mathrm{B} + 1)$ by Delta method

Then line seven $= \Pr(-\mathrm{s}_{1,1}(\mathbf{X}\boldsymbol{\gamma}_\mathrm{x} + \mathbf{X}\boldsymbol{\theta}\boldsymbol{\gamma}_\boldsymbol{B}) \le \mathbf{Z}|\mathbf{X}=\mathbf{x})$

$$= \Pr(-\frac{\mathrm{s}_{1,1}(\mathbf{X}\boldsymbol{\gamma}_\mathrm{x}+\mathbf{X}\boldsymbol{\theta}\boldsymbol{\gamma}_\mathrm{B})}{\sqrt{s_{1,1}^2\boldsymbol{\gamma}_B^T\mathbf{V}\boldsymbol{\gamma}_B+1}} \le \frac{\mathbf{Z}}{\sqrt{s_{1,1}^2\boldsymbol{\gamma}_B^T\mathbf{V}\boldsymbol{\gamma}_B+1}}|\mathbf{X}=\mathbf{x})$$

$$= \Phi(\frac{\mathrm{s}_{1,1}(\mathbf{X}\boldsymbol{\gamma}_\mathrm{x}+\mathbf{X}\boldsymbol{\theta}\boldsymbol{\gamma}_\boldsymbol{B})}{\sqrt{s_{1,1}^2\boldsymbol{\gamma}_B^T\mathbf{V}\boldsymbol{\gamma}_B+1}}|\mathbf{X}=\mathbf{x})$$

$$\approx \mathrm{H}(\frac{(\mathbf{X}\boldsymbol{\gamma}_\mathrm{x}+\mathbf{X}\boldsymbol{\theta}\boldsymbol{\gamma}_\mathrm{B})}{\sqrt{s_{1,1}^2\boldsymbol{\gamma}_B^T\mathbf{V}\boldsymbol{\gamma}_B+1}}|\mathbf{X}=\mathbf{x})$$

Throughout this chapter we assume that the established model and the expanded model are

logistic regression models. Therefore, it is necessary to approximate the logistic cumulative function as a normal cumulative distribution function. However, if both the established model and the expanded model are probit regression models, the above logistic regression approximation is no longer applicable. In (3.4) inside the integral will be the product of a normal cumulative distribution function and a normal probability density function and the solution of this integral will be a normal cumulative distribution function. As a result, there exist exact relationship equations connecting parameters $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ if we assume the established model and the expanded model are probit models.

3.7.2 *Appendix B*

When $\mathbf{B}$ is a binary variable, based on Bayes theorem, there is a relationship equation connecting $\Pr(\mathbf{Y}=1|\mathbf{X}), \Pr(\mathbf{Y}=1|\mathbf{X},\mathbf{B})$ and $\mathrm{f}(\mathbf{B}|\mathbf{X},\mathbf{Y})$, regardless of the type of variable $\mathbf{B}$ is (Grill et al., 2015; Satten and Kupper, 1993).

$$\frac{\Pr(\mathbf{Y}=1|\mathbf{X},\mathbf{B})}{\Pr(\mathbf{Y}=0|\mathbf{X},\mathbf{B})} = \frac{\mathrm{f}(\mathbf{B}|\mathbf{X},\mathbf{Y}=1)}{\mathrm{f}(\mathbf{B}|\mathbf{X},\mathbf{Y}=0)} \cdot \frac{\Pr(\mathbf{Y}=1|\mathbf{X})}{\Pr(\mathbf{Y}=0|\mathbf{X})} \tag{3.22}$$

Re-arranging (3.22) and take the log on both sides, we have:

$$\log\left\{\frac{\Pr(\mathbf{Y}=1|\mathbf{X})}{\Pr(\mathbf{Y}=0|\mathbf{X})}\right\} = \log\left\{\frac{\Pr(\mathbf{Y}=1|\mathbf{X},\mathbf{B})}{\Pr(\mathbf{Y}=0|\mathbf{X},\mathbf{B})}\right\} + \log\left\{\frac{\mathrm{f}(\mathbf{B}|\mathbf{X},\mathbf{Y}=0)}{\mathrm{f}(\mathbf{B}|\mathbf{X},\mathbf{Y}=1)}\right\} \tag{3.23}$$

Equation (3.23) indicates that when $\mathbf{B}$ is binary, we need to define a model for $\mathbf{B}|\mathbf{X},\mathbf{Y}$ instead of a model for $\mathbf{B}|\mathbf{X}$. Assume $\mathrm{logit}(\Pr(\mathbf{B}=1|\mathbf{X},\mathbf{Y})) = \sum_{\mathrm{j}=0}^{\mathrm{p}} \phi_{\mathrm{j}}\mathbf{X_j} + \phi_{\mathrm{p}+1}\mathbf{Y}$. So $\mathrm{f}(\mathbf{B}|\mathbf{X},\mathbf{Y}) = \frac{\mathrm{e}^{(\sum_{\mathrm{j}=0}^{\mathrm{p}} \phi_{\mathrm{j}}\mathbf{X_j}+\phi_{\mathrm{p}+1}\mathbf{Y})\mathbf{B}}}{1+\mathrm{e}^{\sum_{\mathrm{j}=0}^{\mathrm{p}} \phi_{\mathrm{j}}\mathbf{X_j}+\phi_{\mathrm{p}+1}\mathbf{Y}}}$. By looking at the log odds ratio in equation (3.3), we find that the left hand side is $\sum_{\mathrm{j}=0}^{\mathrm{p}} \beta_{\mathrm{j}}\mathbf{X_j}$, a linear combination of $\boldsymbol{\beta}$s. So equation (3.23) can be written as:

$$\beta_0+\beta_1\mathrm{X}_1+...+\beta_{\mathrm{p}}\mathrm{X}_{\mathrm{p}} = \gamma_0+\gamma_1\mathrm{X}_1+...+\gamma_{\mathrm{p}}\mathrm{X}_{\mathrm{p}}+\gamma_{\mathrm{p}+1}\mathrm{B}+\log\left\{\frac{\mathrm{e}^{(\sum_{\mathrm{j}=0}^{\mathrm{p}} \phi_{\mathrm{j}}\mathrm{X}_{\mathrm{j}})\mathrm{B}}}{1+\mathrm{e}^{\sum_{\mathrm{j}=0}^{\mathrm{p}} \phi_{\mathrm{j}}\mathrm{X}_{\mathrm{j}}}} \cdot \frac{1+\mathrm{e}^{\sum_{\mathrm{j}=0}^{\mathrm{p}} \phi_{\mathrm{j}}\mathrm{X}_{\mathrm{j}}+\phi_{\mathrm{p}+1}}}{\mathrm{e}^{(\sum_{\mathrm{j}=0}^{\mathrm{p}} \phi_{\mathrm{j}}\mathrm{X}_{\mathrm{j}}+\phi_{\mathrm{p}+1})\mathrm{B}}}\right\} \tag{3.24}$$

The right hand side of this expression can be rewritten as

$$\gamma_0 + ... + \gamma_{\mathrm{p}}\mathrm{X}_{\mathrm{p}} + \gamma_{\mathrm{p}+1}\mathrm{B} + \log\left\{\frac{1+\mathrm{e}^{\sum_{\mathrm{j}=0}^{\mathrm{p}} \phi_{\mathrm{j}}\mathrm{X}_{\mathrm{j}}+\phi_{\mathrm{p}+1}}}{1+\mathrm{e}^{\sum_{\mathrm{j}=0}^{\mathrm{p}} \phi_{\mathrm{j}}\mathrm{X}_{\mathrm{j}}}} \cdot \frac{1}{\mathrm{e}^{\phi_{\mathrm{p}+1}\mathrm{B}}}\right\}$$

$$= \gamma_0 + ... + \gamma_{\mathrm{p}}\mathrm{X_p} + \gamma_{\mathrm{p+1}}\mathrm{B} - \phi_{\mathrm{p+1}}\mathrm{B} + \log\left\{1 + \mathrm{e}^{\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j}+\phi_{\mathrm{p+1}}}\right\} - \log\left\{1 + \mathrm{e}^{\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j}}\right\}$$

This expression can be approximated using Taylor series expansions to obtain an expression which is linear in the X_j's. A number of different ways of approximating it are possible. Below we write out the expressions for expansion about $\phi_0 = \phi_1 = ... = \phi_{p+1} = 0$. Other expansions about $\phi_1 = ... = \phi_{p+1} = 0$ and about the unconstrained MLE's were also considered. The second approximation did lead to slightly improved results in some cases. The third approximation did not lead to a satisfactory linearization. Using the expansion about $\phi_0 = \phi_1 = ... = \phi_{p+1} = 0$ we obtain

$$\begin{aligned}
&\gamma_0 + ... + \gamma_{\mathrm{p}}\mathrm{X_p} + \gamma_{\mathrm{p+1}}\mathrm{B} - \phi_{\mathrm{p+1}}\mathrm{B} + \log\left\{1 + \mathrm{e}^{\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j}+\phi_{\mathrm{p+1}}}\right\} - \log\left\{1 + \mathrm{e}^{\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j}}\right\} \\
&= \gamma_0 + ... + \gamma_{\mathrm{p}}\mathrm{X_p} + \gamma_{\mathrm{p+1}}\mathrm{B} + \log\left\{\frac{1+\mathrm{e}^{\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j}+\phi_{\mathrm{p+1}}}}{1+\mathrm{e}^{\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j}}} \cdot \frac{1}{\mathrm{e}^{\phi_{\mathrm{p+1}}\mathrm{B}}}\right\} \\
&= \gamma_0 + ... + \gamma_{\mathrm{p}}\mathrm{X_p} + \gamma_{\mathrm{p+1}}\mathrm{B} - \phi_{\mathrm{p+1}}\mathrm{B} + \log\left\{1 + \mathrm{e}^{\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j}+\phi_{\mathrm{p+1}}}\right\} - \log\left\{1 + \mathrm{e}^{\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j}}\right\} \\
&\approx \gamma_0 + ... + \gamma_{\mathrm{p}}\mathrm{X_p} + \gamma_{\mathrm{p+1}}\mathrm{B} - \phi_{\mathrm{p+1}}\mathrm{B} + [\log 2 + \tfrac{1}{2}(\textstyle\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j} + \phi_{\mathrm{p+1}}) + \tfrac{1}{8}(\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j} + \phi_{\mathrm{p+1}})^2 \\
&+ \mathrm{O}((\textstyle\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j} + \phi_{\mathrm{p+1}})^3)] - [\log 2 + \tfrac{1}{2}(\textstyle\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j}) + \tfrac{1}{8}(\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j})^2 + \mathrm{O}((\textstyle\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j})^3)] \\
&\approx \gamma_0 + ... + \gamma_{\mathrm{p}}\mathrm{X_p} + \gamma_{\mathrm{p+1}}\mathrm{B} - \phi_{\mathrm{p+1}}\mathrm{B} + \tfrac{1}{2}\phi_{\mathrm{p+1}} + \tfrac{1}{8}(\textstyle\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j} + \phi_{\mathrm{p+1}})^2 - \tfrac{1}{8}(\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j})^2 \\
&= \gamma_0 + \tfrac{1}{2}\phi_{\mathrm{p+1}} + \tfrac{1}{4}\phi_0\phi_{\mathrm{p+1}} + \tfrac{1}{8}\phi_{\mathrm{p+1}}^2 + \textstyle\sum_{\mathrm{j=1}}^{\mathrm{p}}(\gamma_{\mathrm{j}} + \tfrac{1}{4}\phi_{\mathrm{j}}\phi_{\mathrm{p+1}})\mathrm{X_j} + (\gamma_{\mathrm{p+1}} - \phi_{\mathrm{p+1}})\mathrm{B}
\end{aligned}$$

The third last equation is by Taylor series expansions of $\log\left\{1 + \mathrm{e}^{\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j}+\phi_{\mathrm{p+1}}}\right\}$ and $\log\left\{1 + \mathrm{e}^{\sum_{\mathrm{j=0}}^{\mathrm{p}}\phi_{\mathrm{j}}\mathrm{X_j}}\right\}$ at point 0, respectively. By matching the coefficient of each variable on the left hand side and the right hand side of the equation, we find an approximate relationship between $\boldsymbol{\gamma}, \boldsymbol{\phi}$ and $\boldsymbol{\beta}$, when $\mathbf{B}$ is a binary variable:

$$\begin{cases}
\beta_0 \approx \gamma_0 + \frac{1}{2}\phi_{p+1} + \frac{1}{4}\phi_0\phi_{p+1} + \frac{1}{8}\phi_{p+1}^2 \\
\beta_j \approx \gamma_j + \frac{1}{4}\phi_j\phi_{p+1}, j = 1, ..., p \\
\gamma_{p+1} = \phi_{p+1}
\end{cases} \tag{3.25}$$

3.7.3 *Appendix C*

Bootstrap Estimate of the Standard Error for the Constrained ML Estimate When B Is Normal

We would like to obtain a bootstrap estimate of the constrained ML estimator's standard error. In our study, we implement a parametric bootstrap as follows:

- Estimate the regression coefficients $\gamma_0, \ldots, \gamma_{p+1}$ and $\theta_0, \ldots, \theta_p$ by the constrained ML method for the original sample
- Calculate the fitted outcome $\hat{\mathrm{B}}_{\mathrm{i}}$ and residual $\mathrm{E}_{\mathrm{i,B}}$ for each observation: $\hat{\mathrm{B}}_{\mathrm{i}} = \hat{\theta}_0 + \hat{\theta}_1 \mathrm{X}_{\mathrm{i1}} + \cdots + \hat{\theta}_{\mathrm{p}} \mathrm{X}_{\mathrm{ip}}$ and $\mathrm{E}_{\mathrm{i,B}} = \mathrm{B}_{\mathrm{i}} - \hat{\mathrm{B}}_{\mathrm{i}}$
- Take bootstrap samples of the residual (sample with replacement), $\tilde{\mathbf{e}}_{\mathbf{b}} = [\tilde{\mathrm{E}}_{\mathrm{b,1,B}}, \ldots, \tilde{\mathrm{E}}_{\mathrm{b,n,B}}]^{\mathrm{T}}, \mathrm{b} = 1, \ldots, \mathrm{S}$, calculate bootstrapped $\mathbf{B}$ values $\tilde{\mathbf{B}}_{\mathbf{b}} = [\tilde{\mathrm{B}}_{\mathrm{b1}}, \ldots, \tilde{\mathrm{B}}_{\mathrm{bn}}]^{\mathrm{T}}$, where $\tilde{\mathrm{B}}_{\mathrm{bi}} = \hat{\mathrm{B}}_{\mathrm{i}} + \tilde{\mathrm{E}}_{\mathrm{b,i,B}}$
- Calculate bootstrap $\mathbf{Y}$ values: $\tilde{\mathrm{Y}}_{\mathrm{bi}} \sim \mathrm{Bernoulli}(\tilde{\mathrm{P}}_{\mathrm{bi}})$ where $\tilde{\mathrm{P}}_{\mathrm{bi}} = \Pr(\mathrm{Y} = 1 | \mathrm{X}_{\mathrm{i}}, \tilde{\mathrm{B}}_{\mathrm{bi}}, \hat{\boldsymbol{\gamma}})$
- Regress $\tilde{\mathbf{Y}}_{\mathbf{b}}$ on the fixed $\mathbf{X}$ design matrix and bootstrap samples $\tilde{\mathbf{B}}_{\mathbf{b}}$ to obtain bootstrap estimates of regression coefficients by the constrained ML method: $\tilde{\gamma}_{b,0}, \ldots, \tilde{\gamma}_{b,p+1}$
- The $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{b}}$ can be used to construct bootstrap standard error: $\tilde{\sigma}_{.,j} = (\frac{\sum_{b=1}^{S} (\hat{\gamma}_{b,j} - \bar{\tilde{\gamma}}_{.,j})^2}{S-1})^{1/2}$, $j = 0, \ldots, p+1$, in the usual bootstrap manner as described in Efron and Tibshirani (1986).

Bootstrap Estimate of the Standard Error for the Constrained ML Estimate When B Is Binary

- Estimate $\gamma_0, \ldots, \gamma_{p+1}$ and $\phi_0, \ldots, \phi_{p+1}$ by the constrained ML method for the original sample
- Calculate bootstrap $\mathbf{B}$ values: $\tilde{\mathrm{B}}_{\mathrm{bi}} \sim \mathrm{Bernoulli}(\hat{\mathrm{P}}_{\mathrm{B_i}}), \mathrm{b} = 1, \ldots, \mathrm{S}$, where $\hat{\mathrm{P}}_{\mathrm{B_i}} = \Pr(\mathrm{B} = 1 | \mathrm{X}_{\mathrm{i}}, \mathrm{Y}_{\mathrm{i}}, \hat{\boldsymbol{\phi}}, \bar{\boldsymbol{\beta}})$

Table 3.4: Simulation results of parametric bootstrap: we report the ratio of average bootstrap mean and Monte Carlo mean $(\frac{1}{500}\sum_{\mathrm{m}=1}^{500}\tilde{\bar{\gamma}}_{\mathrm{m,j}})/(\frac{1}{500}\sum_{\mathrm{m}=1}^{500}\hat{\gamma}_{\mathrm{m,j}})$ and the ratio of average bootstrap standard error and Monte Carlo standard deviation $(\frac{1}{500}\sum_{\mathrm{m}=1}^{500}\tilde{\sigma}_{\mathrm{m,j}})/\sqrt{V(\hat{\gamma}_j)}$ of each regression coefficient

Method	Ratio	$\hat{\gamma_1}$	$\hat{\gamma_2}$	$\hat{\gamma_3}$
First simulation scenario				
Constrained ML	Avg.Boot.Mean/MC.Mean	1.10	1.10	1.23
	Avg.Boot.SE/MC.SD	1.74	1.81	1.89
Constrained ML + Firth	Avg.Boot.Mean/MC.Mean	0.98	0.99	0.99
	Avg.Boot.SE/MC.SD	1.03	1.05	1.10
Second simulation scenario				
Constrained ML	Avg.Boot.Mean/MC.Mean	1.13	1.12	1.03
	Avg.Boot.SE/MC.SD	1.13	1.00	0.96
Constrained ML + Firth	Avg.Boot.Mean/MC.Mean	1.07	1.06	0.89
	Avg.Boot.SE/MC.SD	0.95	0.94	0.97

- Calculate bootstrap $\mathbf{Y}$ values: $\tilde{\mathrm{Y}}_{\mathrm{bi}} \sim \mathrm{Bernoulli}(\tilde{\mathrm{P}}_{\mathrm{bi}})$ where $\tilde{P}_{bi} = \Pr(\mathrm{Y}=1|\mathrm{X_i}, \tilde{\mathrm{B}}_{\mathrm{bi}}, \hat{\boldsymbol{\gamma}})$

- Regress $\tilde{\mathbf{Y}}_{\mathbf{b}}$ on the fixed $\mathbf{X}$ and bootstrap samples $\tilde{\mathbf{B}}_{\mathbf{b}}$ to obtain bootstrap estimates of regression coefficients by the constrained ML method: $\tilde{\gamma}_{b,0}, \ldots, \tilde{\gamma}_{b,p+1}$

- Construct bootstrap standard error: $\tilde{\sigma}_{.,\mathrm{j}} = (\frac{\sum_{\mathrm{b}=1}^{\mathrm{S}}(\hat{\gamma}_{\mathrm{b,j}}-\tilde{\bar{\gamma}}_{.,\mathrm{j}})^2}{\mathrm{S}-1})^{1/2}$,

  $\mathrm{j} = 0, \ldots, \mathrm{p}+1$.

Comparison of the average bootstrap point estimates and standard errors to the Monte Carlo average and standard deviation are provided in Table 3.4. The bootstrap mean and the Monte Carlo mean appear to be quite similar. The bootstrap estimated standard error is too large in the first scenario for the constrained ML method. However, with the Firth correction the standard errors from the bootstrap method match the empirical standard deviations, in both scenarios.

Bibliography

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. *Measurement Error in Nonlinear Models: a modern perspective, Second edition*. Chapman & Hall /CRC, Boca Raton, Florida, 2006.

Chatterjee, N., Chen, Y.-H., Maas, P. and Carroll, R. J. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111(513):107–117, 2016.

D'Agostino, R. B., Grundy, S., Sullivan, L. M., Wilson, P. and for the CHD Risk Prediction Group. Validation of the framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *The Journal of the American Medical Association*, 286(2):180–187, 2001.

Efron, B. and Tibshirani, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986.

Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C. and Mulvihill, J. J. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, 81(24):1879–1886, 1989.

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y.-S. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

Grill, S., Ankerst, D. P., Gail, M. H., Chatterjee, N. and Pfeiffer, R. M. Comparison of approaches for incorporating new information into existing risk prediction models. *Statistics in Medicine*, 36(7):1134–1156, 2017.

Grill, S., Fallah, M., Leach, R. J., Thompson, I. M., Hemminki, K. and Ankerst, D. P. A simple-to-use method incorporating genomic markers into prostate cancer risk prediction tools facilitated future validation. *Journal of Clinical Epidemiology*, 68:563–573, 2015.

Gunn, L. H. and Dunson, D. B. A transformation approach for incorporating monotone or unimodal constraints. *Biostatistics*, 6(3):434–449, 2005.

Heinze, G. and Schemper, M. A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16):2409–2419, 2002.

Imbens, G. W. and Lancaster, T. Combining micro and macro data in microeconometric models. *The Review of Economic Studies*, 61(4):655–680, 1994.

Mealiffe, M. E., Stokowski, R. P., Rhees, B. K., Prentice, R. L., Pettinger, M. and Hinds, D. A. Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. *Journal of the National Cancer Institute*, 102(21):1618–1627, 2010.

Monahan, J. and Stefanski, L. A. *Normal scale mixture approximations to F*(z) and computation of the logistic-normal integral. in Handbook of the logistic distribution*. CRC Press, New York, 1992.

Newcombe, P. J., Reck, B. H., Sun, J., Platek, G. T., Verzilli, C., Kader, A. K., Kim, S.-T., Hsu, F.-C., Zhang, Z., Zheng, S. L., Mooser, V. E., Condreay, L. D., Spraggs, C. F., Whittaker, J. C., Rittmaster, R. S. and Xu, J. A comparison of Bayesian and frequentist approaches to incorporating external information for the prediction of prostate cancer risk. *Genetic Epidemiology*, 36(1):71–83, 2012.

Qin, J. Combining parametric and empirical likelihoods. *Biometrika*, 87(2):484–490, 2000.

Qin, J., Zhang, H., Li, P., Albanes, D. and Yu, K. Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika*, 102(1):169–180, 2015.

Satten, G. A. and Kupper, L. L. Inferences about exposure-disease associations using probability-of- exposure information. *Journal of the American Statistical Association*, 88(421):200–208, 1993.

Steyerberg, E. W., Eijkemans, M. J. C., Van Houwelingen, J. C., Lee, K. L. and Habbema, J. D. F. Prognostic models based on literature and individual patient data in logistic regression analysis. *Statistics in Medicine*, 19(2):141–160, 2000.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J. and Kattan, M. W. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, 21(1):128–138, 2010.

Thompson, I. M., Ankerst, D. P., Chi, C., Goodman, P. J., Tangen, C. M., Lucia, M. S., Feng, Z., Parnes, H. L. and Coltman, C. A. Assessing prostate cancer risk: results from the prostate cancer prevention trial. *Journal of the National Cancer Institute*, 98(8):529–534, 2006.

Tomlins, S. A., Day, J. R., Lonigro, R. J., Hovelson, D. H., Siddiqui, J., Kunju, L. P., Dunn, R. L., Meyer, S., Hodge, P., Groskopf, J., Wei, J. T. and Chinnaiyan, A. M. Urine TMPRSS2:ERG plus PCA3 for individualized prostate cancer risk assessment. *European Urology*, 70:45–53, 2015.

Truong, M., Yang, B. and Jarrard, D. F. Toward the detection of prostate cancer in urine: a critical analysis. *The Journal of Urology*, 189(2):422 – 429, 2013.

CHAPTER IV

Statistical methods for updating prediction models using individual predicted outcomes from external sources

4.1 Introduction

Constructing a risk prediction model is of interest in many fields. Biomedical researchers and clinical practitioners, in particular, are interested in situations where a set of potential risk factors are related to the occurrence of an adverse health event (e.g, cardiovascular disease or cancer), and want to quantify the possibility of having a specific event occurring to a given patient over a predefined time period. Risk prediction models are fundamental tools to make predictions of such clinical outcomes in terms of estimated risk. They are constructed from large datasets using statistical methods to predict the probability of occurrence of an outcome, treating these risk factors as predicting variables. Many of them are constructed based on a large study and validated externally on another large dataset, so the estimates from these models are considered to be valid and reliable with little estimation uncertainty.

Assuming the predicted outcome is a binary variable $\mathbf{Y}$, for a vector of predicting variables $\mathbf{X}$, the outcome variable and the predicting variables are often connected through a regression model $\mathbf{Y}|\mathbf{X}$. This regression model could be a parametric regression model, such as a logistic regression model. The risk prediction model could also be a complex nonparametric regression model exploiting machine learning techniques. We assume there is no explicit formula and all

that is available is the output about the estimated probability of an outcome based on the input covariates.

In a systematic review of risk prediction models for cardiovascular diseases it is found that for 167 (46%) models, the complete regression formula, including all regression coefficients and intercept are reported while the other 104 (29%) models are presented as online calculators or risk charts that allow individual risk estimation based on the values of a set of risk factors (the remaining 25% models revealed insufficient information to allow calculation of individual risks) (Damen et al., 2016). However the exact form of the model/algorithm behind the online calculator may not be available to the public.

As new risk factors are found to be associated with the disease, it would be ideal to incorporate these new factors into the risk prediction model and construct an expanded risk prediction model of interest, which then may increase prediction accuracy of the current model. We denote the new risk factors as B and specify a logistic regression model of the form $\mathrm{logit}(\Pr(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B})) = \gamma_0 + \mathbf{X}\boldsymbol{\gamma_X} + \mathbf{B}\boldsymbol{\gamma_B}$ and the goal is to estimate γ_0, $\boldsymbol{\gamma_X}$ and $\boldsymbol{\gamma_B}$. The challenge is that the new risk factors $\mathbf{B}$, as well as the set of risk factors $\mathbf{X}$ in the established risk prediction model are only measured in a small number of subjects. A model built from this modest size dataset may not be very reliable. It is natural to consider incorporating the information in the current dataset and the information collected through other sources on the $\Pr(\mathbf{Y} = 1|\mathbf{X})$ model for increasing the efficiency of estimating the coefficients in the model $\Pr(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B})$ and the resultant predictions.

As the established risk prediction model focuses on the same outcome but with a subset of the covariates used in the expanded model, we would like to exploit the information available from the established model. There is substantial literature on how to combine various types of information from different data sources: some studies assume that two datasets of comparable size are available (Chen and Qin, 2014; Zhan and Ghosh, 2015). Individual-level data $(\mathbf{Y}, \mathbf{X}, \mathbf{B})$ was observed in one dataset while individual-level data $(\mathbf{Y}, \mathbf{X})$ was observed in the other dataset and they use missing

data imputation techniques to impute the missing covariates and then the inference is based on the augmented data. In earlier chapters of this dissertation, we reviewed the literature on incorporating summary level information from an external model for $\Pr(\mathbf{Y} = 1|\mathbf{X})$ when the equation for the model was known and the information was available on model coefficients. We also assumed B to be a single binary, continuous or multivariate Gaussian variable.

In this chapter, we consider the situation where the type of information from the external sources comes in the form of the predicted probabilities of binary outcomes from the established model, but the exact algorithm generating the prediction is unknown to us. This reflects the situation of an online calculator which simply returns the output conditional on the input set of predictors, without making the underlying details of the model publicly available.

Denoting the prediction from the "black box" model as $\bar{\mathrm{P}}(\mathbf{X_i})$, one obvious simple way to include the external information is as a covariate in a model, i.e. $\mathrm{logit}(\Pr(\mathrm{Y_i} = 1|\mathbf{X}, \mathbf{B})) = \theta_0 + \theta_1 \mathrm{S}(\bar{\mathrm{P}}(\mathbf{X_i})) + \theta_2 \mathrm{B_i}$ (where $\mathrm{S}(\cdot)$ is a known smooth, flexible function) and then estimate θ_0, θ_1 and θ_2 are obtained from the small dataset (Tomlins et al., 2015).

We approach the problem from two analytic perspectives. The first is a constrained maximum likelihood method. We optimize the likelihood function of Y given $(\mathbf{X}, \mathbf{B})$ with a constraint related to the differences between the predicted outcomes from the established model and the predicted outcomes from the expanded model.

The second method is based on generation of synthetic data and missing data imputation. This is motivated by methods developed in survey methodology literature (Reiter, 2002; Raghunathan et al., 2003; Reiter and Kinney, 2012). Synthetic data are created based on the combination of the information from the established model and the available data on $(\mathbf{Y}, \mathbf{X}, \mathbf{B})$ in the small dataset. Then using imputation, the parameters of the model for $\Pr(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B})$ are estimated from the combination of the original data and the synthetic data.

The remainder of the chapter is organized as follows. In Section 4.2, we present the notation

and definitions needed in this chapter and outline the sets of assumptions that are necessary for implementing these two approaches. In Section 4.3, we propose the constrained maximum likelihood method and the second approach based on synthetic data and missing data imputation. In Section 4.4, we conduct simulation studies to assess the performance of these methods. As both approaches involve the choice of tuning parameters, in Section 4.5 sensitivity analysis for the tuning parameters is performed. Some discussion and possible extensions are presented in Section 4.6.

4.2 Notation, definition and assumptions

We define a p vector of predicting covariates $\mathbf{X}$ that are used in the established model for a binary outcome $\mathbf{Y}$. The effect of $\mathbf{X}$ on the outcome has been well-studied, and an algorithm that exists provides:

$$\bar{\mathrm{P}}(\mathbf{X_i}) = \hat{\mathrm{Pr}}(\mathrm{Y_i} = 1|\mathbf{X_i}) \tag{4.1}$$

However, the exact form of this model is unknown.

We assume that the new covariate is a single variable $\mathbf{B}$ (can be either discrete or continuous). The expanded model of interest is a logistic regression model, describing the joint effect of $\mathbf{X}$ and $\mathbf{B}$ through the following model:

$$\mathrm{logit}(\mathrm{Pr}(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B})) = \gamma_0 + \mathbf{X}\boldsymbol{\gamma_X} + \mathbf{B}\boldsymbol{\gamma_B} \tag{4.2}$$

We assume that equation (4.2) describes the true distribution of $\mathbf{Y}$ given $\mathbf{X}, \mathbf{B}$. The association between $\mathbf{B}$ and $\mathbf{X}$ is defined through the following model:

$$\mathrm{g}(\mathrm{E}(\mathbf{B}|\mathbf{X})) = \mathrm{f}(\mathbf{X}, \boldsymbol{\theta}) \tag{4.3}$$

where g is the link function. If g is the identity link, then $\mathbf{B}|\mathbf{X}$ is a regression model for a Gaussian distributed $\mathbf{B}$ with $f(\mathbf{X}, \boldsymbol{\theta})$ as the mean component. If g is the logit link, then $\mathbf{B}|\mathbf{X}$ is a logistic

regression model for a binary $\mathbf{B}$ with $\mathrm{E}(\mathbf{B}|\mathbf{X}) = \frac{\exp(\mathrm{f}(\mathbf{X},\boldsymbol{\theta}))}{1+\exp(\mathrm{f}(\mathbf{X},\boldsymbol{\theta}))}$ where the function f might includes linear and quadratic terms in $\mathbf{X}$, as well as interactions.

We assume we have an available dataset of size n, $(\mathrm{Y_i}, \mathbf{X_i}, \mathrm{B_i}), \mathrm{i} = 1, ..., \mathrm{n}$. We assume the external population and the current dataset have the same distributions of $\mathbf{Y}|\mathbf{X}, \mathbf{B}$ and $\mathbf{B}|\mathbf{X}$, and thus they have the same distribution for $\mathbf{Y}|\mathbf{X}$. We study robustness of our methods under model misspecification and violation of these assumptions in the simulation section.

4.3 Statistical Approaches

Without any information from external sources on $\Pr(\mathrm{Y_i} = 1|\mathbf{X_i})$, one may perform a standard logistic likelihood analysis, which we refer to as direct regression. For this method $\mathrm{l}(\boldsymbol{\gamma})$ is maximized, where

$$\begin{aligned}\mathrm{l}(\boldsymbol{\gamma}) &= \sum_{\mathrm{i}=1}^{\mathrm{n}} \big[\mathrm{Y_i log}\big(\frac{\exp(\gamma_0 + \mathbf{X_i}\boldsymbol{\gamma_X} + \mathrm{B_i}\gamma_\mathrm{B})}{1+\exp(\gamma_0 + \mathbf{X_i}\boldsymbol{\gamma_X} + \mathrm{B_i}\gamma_\mathrm{B})}\big) + (1-\mathrm{Y_i})\mathrm{log}\big(\frac{1}{1+\exp(\gamma_0 + \mathbf{X_i}\boldsymbol{\gamma_X} + \mathrm{B_i}\gamma_\mathrm{B})}\big)\big] \\ &= \sum_{\mathrm{i}=1}^{\mathrm{n}} \big[\mathrm{Y_i}(\gamma_0 + \mathbf{X_i}\boldsymbol{\gamma_X} + \mathrm{B_i}\gamma_\mathrm{B}) - \log(1+\exp(\gamma_0 + \mathbf{X_i}\boldsymbol{\gamma_X} + \mathrm{B_i}\gamma_\mathrm{B}))\big]\end{aligned} \tag{4.4}$$

where $(\mathrm{Y_i}, \mathbf{X_i}, \mathrm{B_i}), \mathrm{i} = 1, ..., \mathrm{n}$ is the set of available data to fit the expanded model. For small samples this method can be improved by including the Firth correction, which is described in earlier chapters.

4.3.1 *Constrained Maximum Likelihood*

We aim to develop a statistical approach that can use $\bar{\mathrm{P}}(\mathbf{X_i})$, the individual-level predicted outcome from the established model, to help with the inference in the maximum likelihood estimator based on current data and the expanded model for $\mathbf{Y}|\mathbf{X}, \mathbf{B}$. Based on the Law of Total Expectation, we construct an approximate equation to link $\Pr(\mathbf{Y} = 1|\mathbf{X})$ and $\Pr(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B})$:

$$\Pr(\mathbf{Y}=1|\mathbf{X})=\mathrm{E}_{\mathbf{B}|\mathbf{X}}(\Pr(\mathbf{Y}=1|\mathbf{X},\mathbf{B}))=\int\Pr(\mathbf{Y}=1|\mathbf{X},\mathbf{B})\mathrm{p}(\mathbf{B}|\mathbf{X})\mathrm{d}\mathbf{B}\approx\frac{1}{\mathrm{S}}\sum_{\mathrm{s}=1}^{\mathrm{S}}\Pr(\mathbf{Y}=1|\mathbf{X},\mathbf{B}^{(\mathbf{s})}) \tag{4.5}$$

where $\mathbf{B}^{(\mathbf{s})}$ is a draw from the distribution of $\mathbf{B}|\mathbf{X}$ and S is the total number of draws of $\mathbf{B}$. Assuming $\mathbf{Y}|\mathbf{X}$ is the same in the external source and the current small dataset, we have the following approximate equation to connect the predicted outcome from the external source and the predicted outcome from model (4.2), for given values of regression coefficients:

$$\bar{\mathrm{P}}(\mathbf{X_i})\approx\frac{1}{\mathrm{S}}\sum_{\mathrm{s}=1}^{\mathrm{S}}\Pr(\mathrm{Y_i}=1|\mathbf{X_i},\mathrm{B_i^{(s)}})=\frac{1}{\mathrm{S}}\sum_{\mathrm{s}=1}^{\mathrm{S}}\frac{\exp(\gamma_0+\mathbf{X_i}\boldsymbol{\gamma_X}+\mathrm{B_i^{(s)}}\gamma_\mathrm{B})}{1+\exp(\gamma_0+\mathbf{X_i}\boldsymbol{\gamma_X}+\mathrm{B_i^{(s)}}\gamma_\mathrm{B})} \tag{4.6}$$

For subject i, the draws of $\mathrm{B_i}$ are a set of plausible values that represent the uncertainty about the distribution of $\mathrm{B_i}$, conditioning on the values of $\mathbf{X_i}$. To obtain these draws that could reasonably describe the conditional distribution of $\mathrm{B_i}$ we apply a predictive model for $\mathbf{B}$ given $\mathbf{X}$, either a Bayesian parametric regression model or a more flexible nonparametric predictive model for $\mathbf{B}|\mathbf{X}$.

For the first option, we assume the conditional distribution of $\mathbf{B}|\mathbf{X}$ is a well-parametrized distribution. For example, for a continuous $\mathbf{B}$, we assume it is a normal distribution; for a binary $\mathbf{B}$, we assume it is a Bernoulli. We can fit a standard Bayesian regression model according to the type of variable $\mathbf{B}$ is. Denoting the regression parameters as $\boldsymbol{\theta}$, we obtain posterior distribution of $\boldsymbol{\theta}$ and denote it as $\mathrm{p}(\boldsymbol{\theta}|\mathbf{B},\mathbf{X})$. If $\mathbf{B}$ is continuous, we can fit a Bayesian linear regression model and obtain the posterior distribution of $\boldsymbol{\theta},\sigma^2$. Then $\mathrm{B_i^{(s)}},\mathrm{s}=1,...,\mathrm{S}$ are generated based on $\mathrm{B_i^{(s)}}\sim\mathrm{N}(\mathrm{X_i}\boldsymbol{\theta^{(s)}},\sigma^{2(\mathrm{s})})$ where $\boldsymbol{\theta^{(s)}}\sim\mathrm{p}(\boldsymbol{\theta}|\mathbf{B},\mathbf{X})$ is a sample from the posterior distribution of $\boldsymbol{\theta}$ and $\sigma^{2(\mathrm{s})}\sim\mathrm{p}(\sigma^2|\mathbf{B},\mathbf{X})$ is a sample from the posterior distribution of σ^2. If $\mathbf{B}$ is binary, it is a Bayesian logistic regression model and we obtain the posterior distribution of $\boldsymbol{\theta}$ and draw $\mathrm{B_i^{(s)}},\mathrm{s}=1,...,\mathrm{S}$ from $\mathrm{B_i^{(s)}}\sim\mathrm{Bernoulli}(\frac{\exp(\mathbf{X_i}\boldsymbol{\theta^{(s)}})}{1+\exp(\mathbf{X_i}\boldsymbol{\theta^{(s)}})})$ where $\boldsymbol{\theta^{(s)}}\sim\mathrm{p}(\boldsymbol{\theta}|\mathbf{B},\mathbf{X})$ is a sample from the posterior distribution of $\boldsymbol{\theta}$.

For the second option, we first regress $\mathbf{B}$ against $\mathbf{X}$ and obtain $\hat{\mathrm{B}}_\mathrm{i}, \mathrm{i} = 1, ..., \mathrm{n}$ (i.e., an estimate of $\mathrm{E}(\mathbf{B}|\mathbf{X_i})$. Then we find the S nearest neighbors of observation i according to the 1-norm distance (i.e., S number of subjects such that their predicted values of $\mathbf{B}$ are the S closest to $\hat{\mathrm{B}}_\mathrm{i}$). Then $\mathrm{B}_\mathrm{i}^{(\mathrm{s})}, \mathrm{s} = 1, \ldots, \mathrm{S}$ will be the observed values (not the predicted values) corresponding to these nearest neighbors. This is another way to generate multiple draws of $\mathrm{B_i}$, from a distribution that approximates the distribution of $\mathbf{B}|\mathbf{X_i}$ without assuming any parametric form. We call this approach the K-nearest neighbors method.

Since the predicted value of $\mathbf{Y}$ conditioning on $\mathbf{X}$ can be approximated by the average of the predicted outcomes of $\mathbf{Y}$ conditioning on $\mathbf{X}, \mathbf{B}$, given values of $\boldsymbol{\gamma}$, as shown in equation (4.6), we can construct a constraint for the maximum likelihood estimation, requiring the difference between $\bar{\mathrm{P}}(\mathbf{X_i})$ and the average prediction in equation (4.6) given values of $\boldsymbol{\gamma}$ to be small.

We define the constrained maximum likelihood estimator, as the solution of:

$$\min_{\gamma} \left\{ \sum_{i=1}^{n} [-\mathrm{Y_i}(\gamma_0 + \mathbf{X_i}\boldsymbol{\gamma_X} + \gamma_\mathrm{B}\mathrm{B_i}) + \log(1 + \exp(\gamma_0 + \mathbf{X_i}\boldsymbol{\gamma_X} + \gamma_\mathrm{B}\mathrm{B_i}))] \right\} \tag{4.7}$$

$$\text{s.t.} \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\bar{\mathrm{P}}(\mathbf{X_i}) - \frac{1}{\mathrm{S}} \sum_{\mathrm{s}=1}^{\mathrm{S}} \frac{\exp(\gamma_0 + \mathbf{X_i}\boldsymbol{\gamma_X} + \mathrm{B}_\mathrm{i}^{(\mathrm{s})}\gamma_\mathrm{B})}{1 + \exp(\gamma_0 + \mathbf{X_i}\boldsymbol{\gamma_X} + \mathrm{B}_\mathrm{i}^{(\mathrm{s})}\gamma_\mathrm{B})})^2 < \mathrm{C}, \mathrm{C} > 0$$

In the objective function in model (4.7), only $\gamma_0, \boldsymbol{\gamma_X}, \gamma_\mathrm{B}$ are the unknown parameters to be solved. $\bar{\mathrm{P}}(\mathbf{X_i}), \mathrm{i} = 1, ..., \mathrm{n}$ and $\mathrm{B}_\mathrm{i}^{(\mathrm{s})}, \mathrm{i} = 1, ..., \mathrm{n}, \mathrm{s} = 1, ..., \mathrm{S}$ are known values substituted into the constraint.

The implementation of the constrained ML procedure is as follows:

- Based on the type of variable $\mathbf{B}$, choose the appropriate regression model for $\mathbf{B}|\mathbf{X}$. Choose between either a parametric prediction regression model (option one) or the K-nearest neighbor method (option two) to obtain draws of $\mathbf{B}$

- If a parametric prediction model is used, the draws of $\mathrm{B_i}$, $\mathrm{B}_\mathrm{i}^{(\mathrm{s})}, \mathrm{s} = 1, ..., \mathrm{S}$ will be the posterior predictive values of $\mathrm{B_i}$ based on the data and the posterior distribution of the regression

parameters, as described earlier. If a K-nearest neighbor method is used, then S draws of $\mathrm{B_i}$ will be the S closest neighbors' observed values of $\mathbf{B}$ based on K-nearest neighbors methods

- In the constraint in equation (4.7), substitute the values of $\mathrm{B_i^{(s)}}, \mathrm{i} = 1, ..., \mathrm{n}; \mathrm{s} = 1, ..., \mathrm{S}$
- For a fixed C, find the solution to the optimization problem using an optimization algorithm suitable for solving nonlinear constrained optimization.

For solving the optimization problem with nonlinear constraint as in the constrained ML method, we use the *solnp* function in the R package ***Rsolnp*** (Ghalanos and Theussl, 2015; Ye, 1987), which implements a nonlinear optimization using augmented Lagrange method. We use the *solnp* function with its default settings, except that we set minor constraint for each of the regression parameters: $\gamma_\mathrm{j} \in [\hat{\gamma}_\mathrm{j} - 10 * \hat{\mathrm{SE}}_{\gamma_\mathrm{j}}, \hat{\gamma}_\mathrm{j} + 10 * \hat{\mathrm{SE}}_{\gamma_\mathrm{j}}], \forall \mathrm{j}$, where $\hat{\gamma}_\mathrm{j}$ and $\hat{\mathrm{SE}}_{\gamma_\mathrm{j}}$ are the estimates and their standard errors obtained from the unconstrained ML.

We also consider a modification to the above constrained ML solution by adding a Firth penalty term to the objective function:

$$\min_{\boldsymbol{\gamma}} \left\{ \sum_{i=1}^{n} [-\mathrm{Y_i}(\gamma_0 + \mathbf{X_i}\boldsymbol{\gamma_X} + \gamma_\mathrm{B}\mathrm{B_i}) + \log(1 + \exp(\gamma_0 + \mathbf{X_i}\boldsymbol{\gamma_X} + \gamma_\mathrm{B}\mathrm{B_i}))] - 0.5\log|\mathbf{I}(\boldsymbol{\gamma})| \right\} \tag{4.8}$$

$$\mathrm{s.t.} \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\bar{\mathrm{P}}(\mathbf{X_i}) - \frac{1}{\mathrm{S}} \sum_{\mathrm{s}=1}^{\mathrm{S}} \frac{\exp(\gamma_0 + \mathbf{X_i}\boldsymbol{\gamma_X} + \mathrm{B_i^{(s)}}\gamma_\mathrm{B})}{1 + \exp(\gamma_0 + \mathbf{X_i}\boldsymbol{\gamma_X} + \mathrm{B_i^{(s)}}\gamma_\mathrm{B})})^2 < \mathrm{C}, \mathrm{C} > 0$$

where $|\mathbf{I}(\boldsymbol{\gamma})|$ is the determinant of the Fisher information matrix of the likelihood function $\mathrm{L}(\mathbf{Y}|\mathbf{X}, \mathbf{B})$. This Firth penalty term is to correct the biases in the estimates of $\boldsymbol{\gamma}$, and is particularly useful when the sample size is small or when the prevalence of $\mathbf{Y}$ is low.

4.3.2 *Synthetic data method*

We propose a synthetic data approach that can produce synthetic data on $(\mathbf{Y}, \mathbf{X}, \mathbf{B})$, by using the combination of the available information from the established model and the data from the current dataset. The synthetic data would enlarge the sample size and thus help improve the inference

of our maximum likelihood estimation. The information from external sources can be considered as available prior knowledge which can leverage the information from the current data. Our prior information, in this case, is the predicted outcomes $\bar{\mathrm{P}}(\mathbf{X_i}), \mathrm{i} = 1, ..., \mathrm{n}$. This information can be used to generate synthetic data on $(\mathbf{Y}, \mathbf{X})$ and then the value of $\mathbf{B}$ can be considered as missing and it is thus reduced to a missing data problem. We can use standard missing data procedures like multiple imputation to impute the value of $\mathbf{B}$. A key advantage of our method is that it naturally incorporates the prior knowledge into the method by creating large "fake" data compatible with the $\mathbf{Y}|\mathbf{X}$ established model. Inference in the model $\mathbf{Y}|\mathbf{X}, \mathbf{B}$ may be improved with the large sample size of the synthetic data.

We develop two different methods of creating synthetic data, the single synthetic dataset method and the multiple synthetic dataset method.

Single synthetic dataset method

In single synthetic dataset method, the goal is to generate a large amount of data which combines the available information from the external source and the information that is present in the current small dataset. In other words, the generated new data will follow both the conditional distribution of $\mathbf{Y}|\mathbf{X}$ as modeled in the external source and the data structure present in the small dataset.

We can first create synthetic data for $\mathbf{X}$ by taking random replicates of $\mathbf{X}$ from the small dataset. The sample size in the augmented data is $\mathrm{S} * \mathrm{n}$ so that the ratio of the sample size in the synthetic dataset and the sample size in the small dataset is S. We denote the synthetic data on $\mathbf{X}$ as $\mathbf{X}^{\star}_{\mathbf{j}}, \mathrm{j} = 1, ..., \mathrm{S} * \mathrm{n}$. These synthetic data on $\mathbf{X}$ can be considered as draws of $\mathbf{X}$ from the empirical marginal distribution of $\mathbf{X}$.

As the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ is well-studied in the established model, we assume the predicted probability of $\mathbf{Y} = 1$ given $\mathbf{X}$, $\bar{\mathrm{P}}(\mathbf{X})$ is reliable and can well describe the chance that $\mathbf{Y}$ equals one, when the predicting covariates are $\mathbf{X}$ only. Thus we can generate draws

$\mathrm{Y}^{\star}_{\mathrm{j}}, \mathrm{j} = 1, ..., \mathrm{S} * \mathrm{n}$, from a Bernoulli($\bar{\mathrm{P}}(\mathbf{X}^{\star}_{\mathbf{j}})$) distribution, in the synthetic dataset. Since the marginal distribution of $\mathbf{X}$ and the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ are both correct, the joint distribution $(\mathbf{Y}, \mathbf{X})$ will be correct in the synthetic dataset.

By augmenting $\mathbf{X}$ and then obtaining $\mathrm{Y}^{\star}_{\mathrm{j}}, \mathrm{j} = 1, ..., \mathrm{S} * \mathrm{n}$ based on the augmented $\mathbf{X}$, we are essentially generating multiple values of $\mathbf{Y}$. We generate multiple values of $\mathbf{Y}$ based on $\bar{\mathrm{P}}(\mathbf{X})$ instead of a single draw of $\mathbf{Y}$ for three reasons: (1) Multiple draws of $\mathbf{Y}$ based on $\bar{\mathrm{P}}(\mathbf{X})$ can better mimic the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ than a single draw; (2) the differences between the values of these draws can be considered as the variation that is inherent in the prediction function $\bar{\mathrm{P}}(\mathbf{X})$; (3) by producing multiple replicates of $\mathbf{Y}$, we increase the sample size.

Now we have samples of size n with complete data on $(\mathbf{Y}, \mathbf{X}, \mathbf{B})$ in the small dataset and samples of size $\mathrm{S} * \mathrm{n}$ with data on $(\mathbf{Y}, \mathbf{X})$ in the synthetic dataset. Hypothetically the synthetic dataset has data on $\mathbf{B}$ also, however it is missing in this case. So we can impute the value of $\mathbf{B}$ in the synthetic dataset, by using a missing data imputation technique.

Multiple imputation is the classical strategy for analyzing data with missing value. Instead of filling in one value for each missing observation, multiple imputation procedure creates multiple values for each missing observation, because single imputation cannot reflect the uncertainty inherent in the predictions of the unknown missing values. The standard multiple imputation process involves four steps: (1) K imputed values are generated for each missing value based on an imputation algorithm. The choice of imputation algorithm depends on the type of variables the missing data are; (2) after the imputation algorithm is implemented, K complete datasets are generated; (3) these K complete datasets are analyzed separately using standard statistical procedures and thus K sets of estimates are obtained. Denote the parameter of interest as Q and let point estimates from the kth complete dataset be $\hat{\mathrm{Q}}_{\mathrm{k}}$; (4) These estimates $\hat{\mathrm{Q}}_1, ..., \hat{\mathrm{Q}}_{\mathrm{K}}$ are pooled together according to Rubin's rule (Rubin, 1987) to give a final estimate:

$$\bar{\mathrm{Q}} = \frac{1}{\mathrm{K}} \sum_{\mathrm{k}=1}^{\mathrm{K}} \hat{\mathrm{Q}}_{\mathrm{k}} \tag{4.9}$$

In our case we have a single missing covariate $\mathbf{B}$. We use a parametric regression method as our multiple imputation algorithm for such univariate missing data. If $\mathbf{B}$ is continuous, we use a Bayesian linear regression model. If $\mathbf{B}$ is binary, we use a Bayesian logistic regression model. In either case, the predicting variables are $(\mathbf{Y}, \mathbf{X})$ and only the linear term of $\mathbf{Y}$ and $\mathbf{X}$ are used as the predicting covariates. The imputed values are draws from the posterior predictive distribution of $\mathbf{B}$ given $\mathbf{X}$ and $\mathbf{Y}$. To create $\mathrm{S} * \mathrm{n}$ samples of $(\mathbf{Y}, \mathbf{X}, \mathbf{B})$ in the synthetic dataset, we perform the following:

- Generate $\mathrm{S} * \mathrm{n}$ samples of $\mathbf{X}$, $\mathbf{X}^{\star}_{\mathbf{j}}, \mathrm{j} = 1, ..., \mathrm{S} * \mathrm{n}$ from the small dataset by sampling with replacement from the observed values of $\mathbf{X}$

- Generate $\mathrm{Y}^{\star}_{\mathrm{j}}$ from $\mathrm{Y}_{\mathrm{j}} \sim \mathrm{Bernoulli}(\bar{\mathrm{P}}(\mathbf{X}^{\star}_{\mathbf{j}}))$ for every observation of $\mathbf{X}^{\star}$ in the synthetic dataset. Now we have data on $(\mathbf{Y}, \mathbf{X})$ of size $\mathrm{S} * \mathrm{n}$

- Impute values of B_{j} based on multiple imputation: combine the small dataset with complete data on (Y, X, B) and the large dataset with data on (Y, X). By using the parametric regression method as the imputation mechanism, we obtain K such imputed values of $\mathbf{B}$ for those observations in the augmented dataset and denote those values by $\mathrm{B}^{(\mathrm{k})}_{\mathrm{j}}, \mathrm{j} = 1, ..., \mathrm{S} * \mathrm{n}, \mathrm{k} = 1, ..., \mathrm{K}$. Then the kth imputed dataset has $(\mathrm{S} + 1) * \mathrm{n}$ observations and consists of two subsets: $(\mathrm{Y}_{\mathrm{i}}, \mathbf{X}_{\mathbf{i}}, \mathrm{B}_{\mathrm{i}}, \mathrm{i} = 1, ..., \mathrm{n})$ and $(\mathrm{Y}^{\star}_{\mathrm{j}}, \mathbf{X}^{\star}_{\mathbf{j}}, \mathrm{B}^{(\mathrm{k})}_{\mathrm{j}}, \mathrm{j} = 1, ..., \mathrm{S} * \mathrm{n})$

- With these K imputations, we analyze each dataset of size $(\mathrm{S}+1) * \mathrm{n}$ separately and pool the estimates according to Rubin's rule shown in equation (4.9). An illustration of this process is shown in Figure 4.1.

Multiple synthetic dataset method

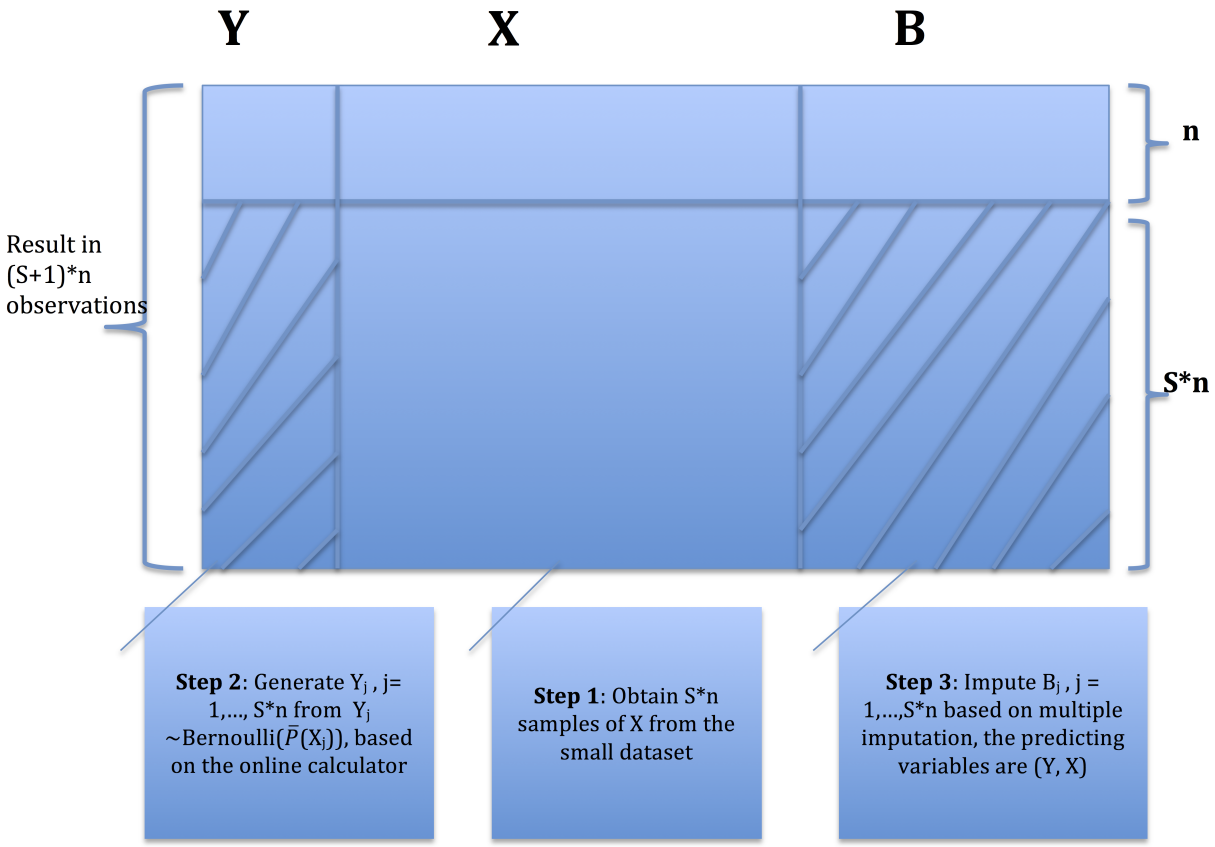


Figure 4.1: Schematic representation of single synthetic dataset method

The multiple synthetic dataset method also creates synthetic dataset using the information from the external sources on $(\mathbf{Y}, \mathbf{X})$ and then creates complete synthetic data $(\mathbf{Y}, \mathbf{X}, \mathbf{B})$ using missing data imputation on $\mathbf{B}$. It is implemented as follows:

- For each subject, generate $\mathrm{Y}_{\mathrm{i}}^{(\mathrm{m})}, \mathrm{m} = 1, ..., \mathrm{M}$ from $\mathrm{Y_i} \sim \mathrm{Bernoulli}(\bar{\mathrm{P}}(\mathbf{X_i})), \mathrm{i} = 1, ..., \mathrm{n}$
- Combine the dataset with data on ($\mathrm{Y}_{\mathrm{i}}^{(\mathrm{m})}$, $\mathbf{X_i}, \mathrm{i} = 1, ..., \mathrm{n}$) with the original data on ($\mathrm{Y_i}, \mathbf{X_i}, \mathrm{B_i}, \mathrm{i} = 1, ..., \mathrm{n}$) to give a dataset with $2\mathrm{n}$ rows. Repeat this M times to give M augmented datasets, each dataset contains complete data on $(\mathbf{Y}, \mathbf{X}, \mathbf{B})$ with sample size n and incomplete data on $(\mathbf{Y}, \mathbf{X})$ with sample size n
- Within each augmented data, using an imputation mechanism to obtain a single imputed value for each missing data of $\mathbf{B}$. That is, we fit a Bayesian regression model of $\mathbf{B}|\mathbf{Y}, \mathbf{X}$ and generate a single draw from the posterior predictive distribution of $\mathbf{B}|\mathbf{Y}, \mathbf{X}$. Then the mth augmented dataset consists of $(\mathrm{Y}_{\mathrm{i}}^{(\mathrm{m})}, \mathbf{X_i}, \mathrm{B}_{\mathrm{i}}^{(\mathrm{m})}, \mathrm{i} = 1, ..., \mathrm{n})$ and $(\mathrm{Y_i}, \mathbf{X_i}, \mathrm{B_i}, \mathrm{i} = 1, ..., \mathrm{n})$
- Within these M datasets, we analyze the data separately using direct regression with Firth correction to obtain estimates of $\gamma_0^{(\mathrm{m})}, \boldsymbol{\gamma}_{\boldsymbol{X}}^{(\boldsymbol{m})}$ and $\gamma_{\mathrm{B}}^{(\mathrm{m})}, \mathrm{m} = 1, ..., \mathrm{M}$
- Pool the estimates. This procedure is schematically shown in Figure 4.2.

Single synthetic dataset method and the multiple synthetic dataset method share some similarities: both of them create synthetic data on $(\mathbf{Y}, \mathbf{X}, \mathbf{B})$ through generating $\mathbf{Y}$ values based on the predicted outcomes from the external source and then imputing $\mathbf{B}$ based on $(\mathbf{Y}, \mathbf{X})$.

The multiple synthetic dataset method is different from the single synthetic dataset method described above in two ways: (1) in the single synthetic dataset method the synthetic data was created as a single long dataset based on random samples of $\mathbf{X}$ while the multiple synthetic dataset method creates multiple datasets, and in each dataset the $\mathbf{X}$ component in the synthetic data is a duplicate of the $\mathbf{X}$ in the original data; (2) the single synthetic dataset method uses multiple imputation to

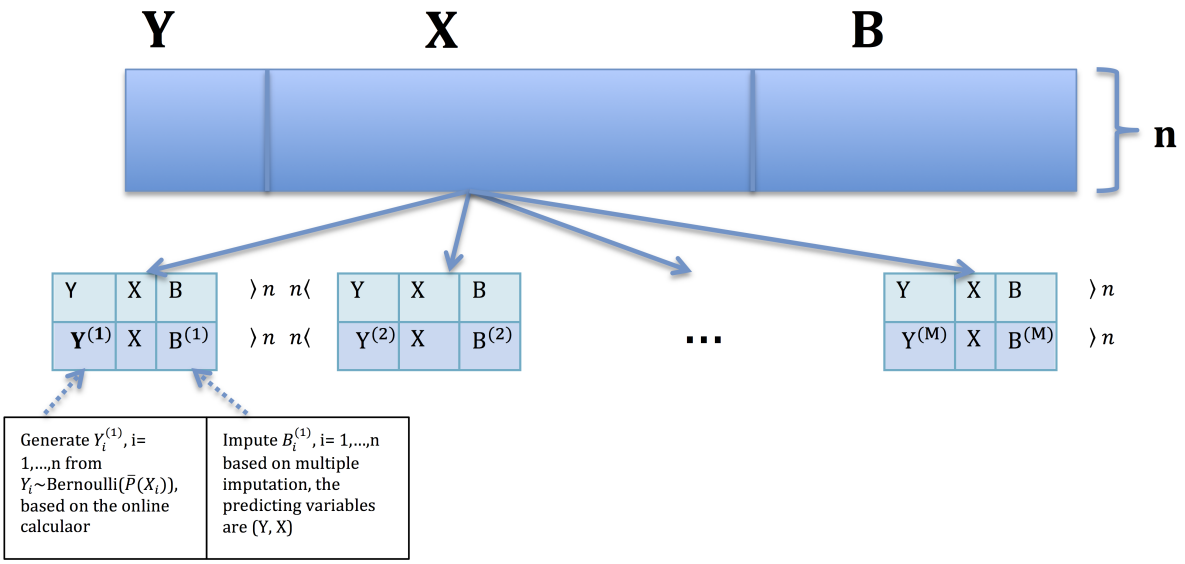


Figure 4.2: Schematic representation of multiple synthetic dataset method

obtain multiple imputed values of B while the multiple synthesis data method imputes a single value of B.

For conducting multiple imputation, we use the R package ***mice*** (Van Buuren and Groothuis-Oudshoorn, 2011). We use the function *mice* with imputation algorithm *logreg* (imputation for binary response variables by the Bayesian logistic regression model) for the imputation of a binary B and the imputation algorithm *norm* (imputation for continuous response variables by the Bayesian linear regression model) for the imputation of a continuous B.

For the single synthetic dataset method, the tuning parameter S (i.e., the ratio of the sample size in the synthetic dataset and that in the small dataset) controls how large the synthetic dataset is. The single synthetic dataset method results in an augmented dataset with a sample size of $(\mathrm{S}+1)*\mathrm{n}$, among which $\mathrm{S}*\mathrm{n}$ are fake data. For the multiple synthetic dataset method, the tuning parameter M (i.e., the number of synthetic datasets) controls the total sample size across all synthetic datasets. The multiple synthetic dataset method results in M augmented datasets, each of size $2\mathrm{n}$, and the

total number of fake data is $\mathrm{M} * \mathrm{n}$. If S equals M, then the sample size of the fake data in the single synthetic dataset method equals the sample size of the fake data in the multiple synthetic dataset method.

4.4 Simulation Study

To assess the performance of the proposed methods, we conduct simulation studies for two types of $\mathbf{B}$ (i.e., a continuous variable or binary variable) and for two settings of the true model for $\mathbf{B}|\mathbf{X}$. In the first setting, the true model of $\mathrm{g}(\mathrm{E}(\mathbf{B}|\mathbf{X}))$ is linear in $\mathbf{X}$ where $\mathrm{g}(\cdot)$ is the link function; in the second setting, the true model of $\mathrm{g}(\mathrm{E}(\mathbf{B}|\mathbf{X}))$ involves a squared term of $\mathbf{X}$.

We compare these models under four given scenarios, all of which have the current data sample size n = 60. For the constrained maximum likelihood method the value of C is 0.005. For each case, we also generate an independent large dataset of $(\mathbf{Y}, \mathbf{X}, \mathbf{B})$ of size 10000 and estimate a logistic regression $\mathrm{logit}(\mathrm{Pr}(\mathrm{Y_i} = 1|\mathbf{X_i}))$. The estimated model $\mathrm{logit}(\hat{\mathrm{Pr}}(\mathrm{Y_i} = 1|\mathbf{X_i}))$ serves as the external source to provide individualized probability of having the disease.

These four simulation scenarios are described as follows:

1. Scenario 1: $\mathbf{B}$ is Gaussian distributed and the true model of $\mathrm{g}(\mathrm{E}(\mathbf{B}|\mathbf{X}))$ is linear in $\mathbf{X}$. The true model of $\mathbf{Y}|\mathbf{X}, \mathbf{B}$ is $\mathrm{logit}(\mathrm{Pr}(\mathrm{Y_i} = 1|\mathrm{X_i}, \mathrm{B_i})) = -2 + 0.5\mathrm{X_i} + 0.5\mathrm{B_i}$. $\mathrm{X_i} \sim \mathrm{N}(0, 1)$ and $\mathrm{B_i}$ is simulated as $\mathrm{B_i} = 1 + 0.5\mathrm{X_i} + \mathrm{N}(0, 1)$. The dataset of size 10000 gives an estimated model of $\mathbf{Y}|\mathbf{X}$: $\mathrm{logit}(\mathrm{Pr}(\mathrm{Y_i} = 1|\mathrm{X_i})) = -1.4626 + 0.7071\mathrm{X_i}$

2. Scenario 2: $\mathbf{B}$ is binary and the true model of $\mathrm{g}(\mathrm{E}(\mathbf{B}|\mathbf{X}))$ is linear in $\mathbf{X}$. The true model of $\mathbf{Y}|\mathbf{X}, \mathbf{B}$ is $\mathrm{logit}(\mathrm{Pr}(\mathrm{Y_i} = 1|\mathrm{X_i}, \mathrm{B_i})) = -1 - 0.5\mathrm{X_i} + 1\mathrm{B_i}$. $\mathrm{X_i} \sim \mathrm{N}(0, 1)$ and $\mathrm{B_i}$ is simulated as $\mathrm{logit}(\mathrm{Pr}(\mathrm{B_i} = 1|\mathrm{X_i})) = 0.5 + 0.3\mathrm{X_i}$. The independent dataset of size 10000 gives an estimated model of $\mathbf{Y}|\mathbf{X}$: $\mathrm{logit}(\mathrm{Pr}(\mathrm{Y_i} = 1|\mathrm{X_i})) = -0.35790 - 0.43698\mathrm{X_i}$

3. Scenario 3: $\mathbf{B}$ is Gaussian distributed and the true model of $\mathrm{g}(\mathrm{E}(\mathbf{B}|\mathbf{X}))$ is not linear in $\mathbf{X}$.

The true model of $\mathbf{Y}|\mathbf{X},\mathbf{B}$ is $\text{logit}(\Pr(\text{Y}_\text{i}=1|\text{X}_\text{i},\text{B}_\text{i}))=-2+0.5\text{X}_\text{i}+0.5\text{B}_\text{i}$. $\text{X}_\text{i}\sim\text{N}(0,1)$ and B_i is simulated as $\text{B}_\text{i}=1+1\text{X}_\text{i}-0.5\text{X}_\text{i}^2+\text{N}(0,1)$. The separate dataset of size 10000 gives an estimated model of $\mathbf{Y}|\mathbf{X}$ as $\text{logit}(\Pr(\text{Y}_\text{i}=1|\text{X}_\text{i}))=-1.4803+0.9743\text{X}_\text{i}-0.2465\text{X}_\text{i}^2$

4. Scenario 4: $\mathbf{B}$ is binary and the true model of $\text{g}(\text{E}(\mathbf{B}|\mathbf{X}))$ is not linear in $\mathbf{X}$. The true model of $\mathbf{Y}|\mathbf{X},\mathbf{B}$ is $\text{logit}(\Pr(\text{Y}_\text{i}=1|\text{X}_\text{i},\text{B}_\text{i}))=-1-0.5\text{X}_\text{i}+1\text{B}_\text{i}$. $\text{X}_\text{i}\sim\text{N}(0,1)$ and B_i is simulated as $\text{logit}(\Pr(\text{B}_\text{i}=1|\text{X}_\text{i}))=0.5+0.3\text{X}_\text{i}+0.2\text{X}_\text{i}^2$. The separate dataset of size 10000 gives an estimated model of $\mathbf{Y}|\mathbf{X}$: $\text{logit}(\Pr(\text{Y}_\text{i}=1|\text{X}_\text{i}))=-0.33909-0.45128\text{X}_\text{i}+0.03229\text{X}_\text{i}^2$

Note that in implementing the constrained ML method, drawing $\mathbf{B}$ given $\mathbf{X}$ involves constructing a regression model for $\mathbf{B}|\mathbf{X}$. In the estimating procedure of $\mathbf{B}|\mathbf{X}$, only linear terms of $\mathbf{X}$ are used as the predicting covariates. In implementing the two synthetic data methods, within the multiple imputation algorithm for imputing missing values in $\mathbf{B}$, we use a Bayesian regression method for $\mathbf{B}|\mathbf{Y},\mathbf{X}$ that is also linear in $\mathbf{Y}$ and $\mathbf{X}$. Since in scenario 3 and in scenario 4, the true model of $\text{g}(\text{E}(\mathbf{B}|\mathbf{X}))$ is not linear in $\mathbf{X}$ but the estimating procedure assumes linearity, mis-specification occurs for the $\mathbf{B}|\mathbf{X}$ model and the $\mathbf{B}|\mathbf{X},\mathbf{Y}$ model. Thus scenario 3 and scenario 4 assess the robustness of our procedures.

We compare these methods in terms of estimation accuracy and prediction ability. For evaluating estimation accuracy of the model, we report the average of estimated coefficients $\boldsymbol{\gamma}$, Monte Carlo standard deviation and mean squared error. For accuracy of the predictions, we evaluate three metrics in a validation dataset of size 800: Brier score ($\frac{\sum_{\text{i}=1}^{800}(\text{Y}_\text{i}-\hat{\text{p}}_\text{i})^2}{\sum_{\text{i}=1}^{800}(\text{Y}_\text{i}-0.5)^2}$), Hosmer-Lemeshow statistic ($\sum_{\text{k}=0}^{1}\sum_{\text{l}=1}^{10}\frac{(\text{O}_\text{kl}-\text{E}_\text{kl})^2}{\text{E}_\text{kl}}$) and the area under the ROC curve (AUC). The Hosmer-Lemeshow statistic calculates the distance between the observed count of $\mathbf{Y}=0$ and $\mathbf{Y}=1$ and the predicted count in group l, for $\text{l}=1,...,10$, where these groups are constructed based on the predicted probabilities.

Table 4.1 shows results of the first simulation scenario based on 500 replications. In this sce-

Table 4.1: Simulation results of the first scenario for Gaussian $\mathbf{B}$, the true model of $\mathrm{g}(\mathrm{E}(\mathbf{B}|\mathbf{X}))$ is linear in X: for each method, we report mean (Monte Carlo standard deviation), MSE, average Brier score, average Hosmer-Lemeshow statistic and average AUC

Method	$\hat{\gamma}_0$	$\hat{\gamma}_{\mathrm{X}}$	$\hat{\gamma}_{\mathrm{B}}$	Brier Score	HL	AUC
True value	-2	0.5	0.5	0.565	21.1	0.751
Online calculator $\bar{\mathrm{P}}(\mathrm{X_i})$				0.612	10.7	0.684
Direct regression	-2.20(0.77)	0.59(0.50)	0.56(0.43)	0.606	100.5	0.731
MSE	0.63	0.26	0.19			
Direct regression + Firth	-2.01(0.65)	0.53(0.44)	0.50(0.38)	0.605	57.2	0.731
MSE	0.43	0.20	0.15			
Plug-in $\hat{\mathrm{P}}_{\mathrm{i}}$				0.608	66.2	0.732
Constrained ML, draw B from $\mathrm{B}|\mathrm{X}$, C = 0.005	-2.15(0.61)	0.56(0.34)	0.53(0.39)	0.591	36.0	0.735
MSE	0.40	0.12	0.15			
Constrained ML, draw B from $\mathrm{B}|\mathrm{X}$ + Firth, C = 0.005	-2.00(0.55)	0.52(0.33)	0.48(0.35)	0.592	32.0	0.735
MSE	0.30	0.11	0.12			
Constrained ML, draw B from KNN, C = 0.005	-2.16(0.67)	0.59(0.35)	0.54(0.41)	0.593	46.2	0.735
MSE	0.47	0.13	0.17			
Constrained ML, draw B from KNN + Firth, C = 0.005	-2.01(0.58)	0.53(0.33)	0.49(0.36)	0.592	35.9	0.735
MSE	0.34	0.11	0.13			
Single synthetic dataset, S = 50, K = 10	-2.12(0.54)	0.51(0.18)	0.51(0.40)	0.581	19.4	0.735
MSE	0.30	0.03	0.16			
Multiple synthetic dataset, M = 50	-2.05(0.53)	0.52(0.26)	0.50(0.37)	0.586	25.1	0.736
MSE	0.28	0.07	0.14			

nario $\mathbf{B}$ is Gaussian distributed and the true model of $\mathrm{g}(\mathrm{E}(\mathbf{B}|\mathbf{X}))$ is linear in $\mathbf{X}$. The prevalence of $\mathbf{Y}$ is 21%. The results show that with a sample size of 60, the unconstrained ML estimates are biased. The Firth correction reduces this bias of the estimates of the regression coefficients compared to the unconstrained ML method, but the MSEs are still high. By using the constrained ML, we are able to reduce the MSE of the estimate of γ_0 from 0.63 to 0.40, the MSE of the estimate of γ_{X} from 0.26 to 0.12, the MSE of the estimate of γ_{B} from 0.19 to 0.15. By applying the Firth correction, the MSEs are further decreased. In terms of the improvement in the prediction, there are only small improvements in Brier score or AUC, compared to the direct regression approach. In terms of HL Statistic, the single synthetic dataset method with S = 50 and K = 10 has the lowest HL statistic (19.4).

Table 4.2 shows results of the second simulation scenario, in which $\mathbf{B}$ is binary and the true model of $\mathrm{g}(\mathrm{E}(\mathbf{B}|\mathbf{X}))$ is linear in $\mathbf{X}$. The prevalence of $\mathbf{Y}$ is 42%. By using the constrained ML, we are able to reduce the MSE of the estimate of γ_0 from 0.33 to 0.18 and the MSE of the estimate

Table 4.2: Simulation results of the second scenario for binary $\mathbf{B}$, the true model of $\mathrm{g}(\mathrm{E}(\mathbf{B}|\mathbf{X}))$ is linear in $\mathbf{X}$: for each method, we report mean (Monte Carlo standard deviation), MSE, average Brier score, average Hosmer-Lemeshow statistic and average AUC

Method	$\hat{\gamma}_0$	$\hat{\gamma}_{\mathrm{X}}$	$\hat{\gamma}_{\mathrm{B}}$	Brier Score	HL	AUC
True value	-1	-0.5	1	0.843	9.6	0.716
Online calculator $\bar{\mathrm{P}}(\mathrm{X_i})$				0.914	4.3	0.644
Direct regression	-1.08(0.57)	-0.56(0.34)	1.11(0.68)	0.888	47.7	0.695
MSE	0.33	0.12	0.47			
Direct regression + Firth	-1.00(0.51)	-0.52(0.31)	1.02(0.62)	0.885	41.3	0.695
MSE	0.26	0.09	0.39			
Plug-in $\hat{\mathrm{P}}_{\mathrm{i}}$				0.888	48.8	0.695
Constrained ML, draw B from $\mathrm{B}|\mathrm{X}$, C = 0.005	-1.03(0.45)	-0.55(0.21)	1.03(0.60)	0.869	23.2	0.702
MSE	0.20	0.05	0.36			
Constrained ML, draw B from $\mathrm{B}|\mathrm{X}$ + Firth, C = 0.005	-0.97(0.42)	-0.52(0.20)	0.96(0.56)	0.869	22.6	0.702
MSE	0.18	0.04	0.32			
Constrained ML, draw B from KNN, C = 0.005	-1.05(0.47)	-0.55(0.22)	1.05(0.62)	0.870	24.8	0.702
MSE	0.22	0.05	0.39			
Constrained ML, draw B from KNN + Firth, C = 0.005	-0.98(0.44)	-0.52(0.21)	0.98(0.58)	0.870	24.1	0.702
MSE	0.19	0.04	0.34			
Single synthetic dataset, S = 50, K = 10	-1.06(0.46)	-0.55(0.11)	1.09(0.68)	0.861	16.9	0.703
MSE	0.22	0.01	0.47			
Multiple synthetic dataset, M = 50	-1.03(0.47)	-0.53(0.18)	1.04(0.64)	0.867	22.7	0.702
MSE	0.22	0.03	0.42			

of γ_{X} from 0.12 to 0.04, when the Firth correction is applied. In terms of HL Statistic, the single synthetic dataset method with S = 50 and K = 10 has the lowest HL statistic (16.9).

In Table 4.3, the simulation results of scenario 3, $\mathbf{B}$ is Gaussian distributed and the true model of $\mathrm{g}(\mathrm{E}(\mathbf{B}|\mathbf{X}))$ is not linear in $\mathbf{X}$. In this case, the model for $\mathbf{B}|\mathbf{X}$ in the constrained ML and the model for $\mathbf{B}|\mathbf{X}, \mathbf{Y}$ are both mis-specified. The prevalence of $\mathbf{Y}$ is calculated to be about 19%. The unconstrained ML has large bias and the predictive ability is low (HL statistic is 559.7). All constrained methods produce biased estimates except constrained ML drawing $\mathbf{B}$ from K nearest neighbors method and Firth correction applied. In terms of MSE, the single synthetic method with S = 50, K = 10 shows the largest decrease in MSE compared to direct regression (0.08 vs 0.60 in $\hat{\gamma}_0$, 0.10 vs 0.34 in $\hat{\gamma}_{\mathrm{X}}$, 0.13 vs 0.18 in $\hat{\gamma}_{\mathrm{B}}$). The constrained methods substantially improve the prediction ability of the model in terms of HL statistic, as we can see that the single synthetic dataset method with S = 50, K = 10 shows a big decrease in terms of HL statistic from 559.7 to 20.2.

Table 4.3: Simulation results of the third scenario for Gaussian $\mathbf{B}$, when the true model of $\mathrm{g}(\mathrm{E}(\mathbf{B}|\mathbf{X}))$ is not linear in $\mathbf{X}$: for each method, we report mean (Monte Carlo standard deviation), MSE, average Brier score, average Hosmer-Lemeshow statistic and average AUC

Method	$\hat{\gamma}_0$	$\hat{\gamma}_{\mathrm{X}}$	$\hat{\gamma}_{\mathrm{B}}$	Brier Score	HL	AUC
True value	-2	0.5	0.5	0.523	17.6	0.765
Online calculator $\mathrm{P}(\mathrm{X_i})$				0.561	5.08	0.696
Direct regression	-2.25(0.74)	0.54(0.58)	0.59(0.42)	0.562	559.7	0.740
MSE	0.60	0.34	0.18			
Direct regression + Firth	-2.01(0.60)	0.50(0.50)	0.51(0.36)	0.561	96.1	0.739
MSE	0.36	0.25	0.13			
Plug-in $\hat{\mathrm{P}}_\mathrm{i}$				0.562	499.0	0.738
Constrained ML, draw B from $\mathrm{B}\|\mathrm{X}$, C = 0.005	-2.05(0.41)	0.43(0.40)	0.49(0.33)	0.546	25.7	0.743
MSE	0.17	0.16	0.11			
Constrained ML, draw B from $\mathrm{B}\|\mathrm{X}$ + Firth, C = 0.005	-1.91(0.38)	0.42(0.37)	0.43(0.30)	0.547	24.8	0.742
MSE	0.15	0.14	0.09			
Constrained ML, draw B from KNN, C = 0.005	-2.21(0.62)	0.51(0.41)	0.57(0.39)	0.548	229.7	0.746
MSE	0.43	0.17	0.16			
Constrained ML, draw B from KNN + Firth, C = 0.005	-2.01(0.52)	0.49(0.37)	0.49(0.34)	0.548	52.8	0.745
MSE	0.27	0.14	0.11			
Single synthetic dataset, S = 50, K = 10	-1.93(0.28)	0.48(0.32)	0.35(0.33)	0.545	20.2	0.737
MSE	0.08	0.10	0.13			
Multiple synthetic dataset, M = 50	-1.93(0.33)	0.47(0.37)	0.41(0.32)	0.547	23.8	0.741
MSE	0.11	0.14	0.11			

Table 4.4 is created based on the simulation result of scenario 4, where $\mathbf{B}$ is binary and the true model of $\mathrm{g}(\mathrm{E}(\mathbf{B}|\mathbf{X}))$ is not linear in $\mathbf{X}$. The prevalence is estimated to be 43%. The constrained ML could alleviate the bias in the estimates of the regression coefficient well, especially after the Firth correction is applied. The multiple synthetic dataset method with M = 50 can also reduce the bias very well. The single synthetic dataset method again has the lowest HL statistic among all methods, which is consistent in all four simulation cases.

Overall the simulation studies under the four scenarios show that: (1) implementing the Firth correction in estimating a logistic regression model always reduces the biases in the estimates, particularly when the prevalence of the outcome is low; (2) the constrained methods can reduce the Monte Carlo standard deviation and MSE of $\hat{\gamma}_X$, using the external information; (3) synthetic data methods reduce the Monte Carlo standard deviation or MSE of $\hat{\gamma}_X$ more than the constrained ML method; (4) the constrained ML or synthetic data methods usually do not reduce the Monte Carlo standard deviation or MSE of $\hat{\gamma}_B$; (4) when the $\mathbf{B}|\mathbf{X}$ model is mis-specified, the K nearest

Table 4.4: Simulation results of the forth scenario for binary $\mathbf{B}$, when $\mathbf{B}|\mathbf{X}$ and the true model of $\mathrm{g}(\mathrm{E}(\mathbf{B}|\mathbf{X}))$ is not linear in $\mathbf{X}$: for each method, we report mean (Monte Carlo standard deviation), MSE, average Brier score, average Hosmer-Lemeshow statistic and average AUC

Method	$\hat{\gamma}_0$	$\hat{\gamma}_{\mathrm{X}}$	$\hat{\gamma}_{\mathrm{B}}$	Brier Score	HL	AUC
True value	-1	-0.5	1	0.855	10.1	0.710
Online calculator $\bar{\mathrm{P}}(\mathrm{X_i})$				0.926	6.2	0.637
Direct regression	-1.09(0.59)	-0.57(0.33)	1.11(0.69)	0.899	47.7	0.688
MSE	0.35	0.11	0.49			
Direct regression + Firth	-1.01(0.53)	-0.52(0.30)	1.02(0.63)	0.897	40.9	0.688
MSE	0.28	0.09	0.40			
Plug-in $\hat{\mathrm{P}}_{\mathrm{i}}$				0.901	51.4	0.688
Constrained ML, draw B from $\mathrm{B}|\mathrm{X}$, C = 0.005	-1.03(0.47)	-0.56(0.21)	1.03(0.61)	0.881	25.1	0.695
MSE	0.22	0.05	0.38			
Constrained ML, draw B from $\mathrm{B}|\mathrm{X}$ + Firth, C = 0.005	-0.97(0.45)	-0.53(0.20)	0.97(0.58)	0.882	24.7	0.695
MSE	0.20	0.04	0.34			
Constrained ML, draw B from KNN, C = 0.005	-1.04(0.49)	-0.56(0.21)	1.05(0.64)	0.882	26.8	0.695
MSE	0.24	0.05	0.41			
Constrained ML, draw B from KNN + Firth, C = 0.005	-0.98(0.45)	-0.53(0.21)	0.98(0.59)	0.882	25.6	0.695
MSE	0.21	0.04	0.35			
Single synthetic dataset, S = 50, K = 10	-1.05(0.50)	-0.55(0.10)	1.08(0.70)	0.873	18.5	0.695
MSE	0.25	0.01	0.50			
Multiple synthetic dataset, M = 50	-1.03(0.49)	-0.54(0.18)	1.05(0.65)	0.879	23.4	0.695
MSE	0.24	0.03	0.43			

neighbors method is a more robust option than the parametric regression model in drawing values of $\mathbf{B}$. (in scenario 3 when the $\mathbf{B}|\mathbf{X}$ model is mis-specified, the constrained ML method drawing $\mathbf{B}$ from K nearest neighbors and Firth correction applied has the smallest biases in the regression coefficient estimates, compared to constrained ML drawing $\mathbf{B}$ from parametric regression model and the two synthetic data methods). For scenario 3 and scenario 4, in simulation results not displayed here we found that if the $\mathbf{B}|\mathbf{X}$ model and the $\mathbf{B}|\mathbf{X}, \mathbf{Y}$ model used for imputation are correctly specified, all constrained methods will have less bias.

4.5 Sensitivity Analysis

In the constrained ML method, there is a fixed parameter C that controls the strength of the constraint. The solution of the constrained optimization problem used in the constrained maximum likelihood estimation is dependent of the value of C. If C is very big, then this constraint would be a very weak constraint and the results from the constrained ML will be very similar to those from

the unconstrained ML. For the single synthetic dataset method, there are two tuning parameters, S, the ratio of the sample size in the synthetic dataset and the original dataset and K, the number of replicates in the multiple imputation procedure. For the multiple synthetic dataset method, there is a tuning parameter M, the number of synthetic datasets.

In practice, the performance of the constrained ML and that of these two synthetic data methods depend on the choice of the corresponding tuning parameters. Here we focus on the tuning parameter selection of C in the constrained ML, the tuning parameter selection of S and K in the single synthetic dataset method and the tuning parameter selection of M in the multiple synthetic dataset method.

We report on simulation studies to investigate the performance of the three constrained methods with different values of the corresponding tuning parameters. We compare their empirical performances under the four simulation scenarios shown in Section 4.4. For each of the three methods, we compare its performance with different values of the tuning parameters within one simulation scenario but also compare the performances across all four simulation scenarios in order to find a consistent conclusion.

For all settings, the results are based on a sample size of 60 and averaged from 500 replicates. To assess the performance of these models, we evaluate bias and Monte Carlo standard deviation of each regression parameter. We also calculate the HL statistic. All results are shown in Table 4.5 and Table 4.6.

Table 4.5 and Table 4.6 show that for the constrained ML, for all scenarios a smaller value of C will result in less biases for the estimates of γ_0. However when C decreases to 0.0025, we don't have a general conclusion about the bias of estimated γ_0 in the constrained ML. In scenario 1, scenario 2 and scenario 4, a smaller value of C will result in less biases for the estimates of γ_{X}. The SD will also be smaller when the C is smaller. On the other hand, when C is very large, like 0.2, it gives a very weak constraint in terms of predicted probabilities from the model and the

results from the constrained ML is about identical to that from the direct regression (unconstrained ML).

For the single synthetic dataset method, with a larger value of S, we observe less bias and smaller SD for γ_0, γ_{X} and γ_{B} in all cases. For example, in scenario 1, when the number of replicates in the multiple imputation procedure is 5, the single synthetic dataset method with S = 50 leads to a bias of -0.10 for γ_0, 0.02 for γ_{X} and 0.00 for γ_{B} where as the single synthetic dataset method with S = 1 has a bias of -0.14 for γ_0, 0.05 for γ_{X} and 0.03 for γ_{B}. If we compare between constrained ML and the single synthetic dataset method, we find that though single synthetic dataset method may lead to less SD in some cases as compared to constrained ML, it is never unbiased.

For the multiple synthetic dataset method, for all scenarios when M increases from 1 to 10, we observe smaller SD for estimates of γ_0, γ_{X} and γ_{B} and a smaller HL statistic. However when M increase from 10 to 50, the decreases in SD for these regression coefficients and the decreases in HL statistic are very limited.

We further investigate the performance of the single synthetic dataset method in scenario 1, with number of replicates K in the multiple imputation procedure being 5. We vary the value of S from 1 to 80 and report the HL statistic and the sum of the MSE of all regression coefficients (i.e. MSE(γ_0) + MSE(γ_{X}) + MSE(γ_{B})). When $\mathrm{S} = 1$, the sample size in the synthetic dataset is n; when $\mathrm{S} = 80$, the sample size in the fake data is $80\mathrm{n}$ and can be considered as sufficiently large. The results are shown in Figure 4.3. We see that with increasing value of S and thus increasing sample size in the synthetic dataset, the HL statistic will decrease at first, but the HL statistic curve will approach a plateau. This phenomenon can also be seen in the sum of the MSE curve, though there are some fluctuations in the values of sum of MSE. Figure 4.3 demonstrates that the improvement in the predictive ability and the improvement in the estimating efficiency of the regression coefficients in the single synthetic dataset method diminish after a certain value of S. If the sample size in the synthetic data is large enough and the performance of the single

Table 4.5: Results for point estimators over 500 replications. The columns correspond to bias and Monte Carlo Standard Deviation for estimating the regression coefficient γ. The last column is HL statistic.

Scenario	Method	$\hat{\gamma}_0$		$\hat{\gamma}_X$		$\hat{\gamma}_B$		HL
		Bias	SD	Bias	SD	Bias	SD	
	Direct regression	-0.20	0.77	0.09	0.50	0.06	0.43	100.5
	CML, C = 0.0025	-0.05	0.46	0.05	0.26	-0.02	0.33	23.6
	CML, C = 0.0050	-0.15	0.61	0.06	0.34	0.03	0.39	36.0
	CML, C = 0.0100	-0.19	0.72	0.07	0.43	0.06	0.42	56.4
	CML, C = 0.0200	-0.20	0.77	0.08	0.48	0.06	0.43	82.9
	CML, C = 0.2000	-0.20	0.77	0.09	0.50	0.06	0.43	100.5
Scenario 1	Single synthetic dataset, S = 1, K = 5	-0.14	0.64	0.05	0.36	0.03	0.42	38.2
	Single synthetic dataset, S = 10, K = 5	-0.11	0.56	0.02	0.22	0.01	0.41	21.5
	Single synthetic dataset, S = 20, K = 5	-0.12	0.54	0.02	0.20	0.02	0.41	20.0
	Single synthetic dataset, S = 50, K = 5	-0.10	0.54	0.02	0.19	0.00	0.41	19.6
	Single synthetic dataset, S = 10, K = 10	-0.12	0.53	0.00	0.21	0.03	0.40	21.1
	Single synthetic dataset, S = 20, K =10	-0.11	0.51	0.01	0.20	0.01	0.38	19.7
	Single synthetic dataset, S = 50, K = 10	-0.12	0.54	0.01	0.18	0.01	0.40	19.4
	Multiple synthetic dataset, M = 1	-0.03	0.63	0.03	0.34	-0.01	0.45	39.2
	Multiple synthetic dataset, M = 10	-0.05	0.55	0.02	0.27	0.01	0.38	26.6
	Multiple synthetic dataset, M = 20	-0.06	0.54	0.02	0.27	0.01	0.38	26.3
	Multiple synthetic dataset, M = 50	-0.05	0.53	0.02	0.26	0.00	0.37	25.1
	Direct regression	-0.08	0.57	-0.06	0.34	0.11	0.68	47.7
	CML, C = 0.0025	0.06	0.35	-0.03	0.14	-0.09	0.49	15.8
	CML, C = 0.0050	-0.03	0.45	-0.05	0.21	0.03	0.60	23.2
	CML, C = 0.0100	-0.07	0.52	-0.06	0.27	0.08	0.66	33.5
	CML, C = 0.0200	-0.08	0.55	-0.06	0.32	0.10	0.67	42.7
	CML, C = 0.2000	-0.08	0.57	-0.06	0.34	0.11	0.68	47.7
Scenario 2	Single synthetic dataset, S = 1, K = 5	-0.08	0.54	-0.05	0.25	0.10	0.71	34.0
	Single synthetic dataset, S= 10, K = 5	-0.06	0.49	-0.04	0.14	0.08	0.72	19.8
	Single synthetic dataset, S = 20, K = 5	-0.05	0.48	-0.05	0.13	0.07	0.71	18.1
	Single synthetic dataset, S = 50, K = 5	-0.05	0.49	-0.04	0.12	0.06	0.71	18.6
	Single synthetic dataset, S = 10, K = 10	-0.06	0.47	-0.04	0.14	0.08	0.68	19.5
	Single synthetic dataset, S = 20, K =10	-0.05	0.46	-0.04	0.12	0.06	0.67	17.4
	Single synthetic dataset, S = 50, K = 10	-0.06	0.46	-0.05	0.11	0.09	0.68	16.9
	Multiple synthetic dataset, M = 1	-0.04	0.57	-0.02	0.25	0.05	0.76	33.8
	Multiple synthetic dataset, M = 10	-0.02	0.47	-0.03	0.19	0.04	0.65	22.7
	Multiple synthetic dataset, M = 20	-0.03	0.47	-0.03	0.18	0.04	0.64	22.6
	Multiple synthetic dataset, M = 50	-0.03	0.47	-0.03	0.18	0.04	0.64	22.7

Table 4.6: Results for point estimators over 500 replications. The columns correspond to bias and Monte Carlo Standard Deviation for estimating the regression coefficient γ. The last column is HL statistic. † In scenario 3 a smaller value of C may lead to non-convergence issue in constrained ML method

Scenario	Method	$\hat{\gamma}_0$		$\hat{\gamma}_X$		$\hat{\gamma}_B$		HL
		Bias	SD	Bias	SD	Bias	SD	
	Direct regression	-0.25	0.74	0.04	0.58	0.09	0.42	559.7
	CML, C = 0.0048†	-0.05	0.41	-0.08	0.39	-0.01	0.34	25.1
	CML, C = 0.0050	-0.05	0.41	-0.07	0.40	-0.01	0.33	25.7
	CML, C = 0.0100	-0.16	0.55	-0.02	0.48	0.05	0.37	41.7
	CML, C = 0.0200	-0.23	0.68	0.03	0.55	0.08	0.40	165.2
	CML, C = 0.2000	-0.25	0.74	0.04	0.58	0.09	0.42	559.7
Scenario 3	Single synthetic dataset, S = 1, K = 5	-0.03	0.45	-0.01	0.45	-0.06	0.37	36.5
	Single synthetic dataset, S= 10, K = 5	0.05	0.33	-0.04	0.35	-0.13	0.35	22.2
	Single synthetic dataset, S= 20, K = 5	0.05	0.32	-0.02	0.35	-0.14	0.36	20.9
	Single synthetic dataset, S = 50, K = 5	0.06	0.29	-0.02	0.33	-0.15	0.34	19.3
	Single synthetic dataset, S = 10, K = 10	0.05	0.30	-0.02	0.35	-0.14	0.33	22.5
	Single synthetic dataset, S = 20, K = 10	0.06	0.29	-0.02	0.34	-0.14	0.34	21.2
	Single synthetic dataset, S = 50, K = 10	0.07	0.28	-0.02	0.32	-0.15	0.33	20.2
	Multiple synthetic dataset, M = 1	0.06	0.47	-0.03	0.46	-0.09	0.39	36.9
	Multiple synthetic dataset, M = 10	0.06	0.34	-0.02	0.38	-0.09	0.33	24.9
	Multiple synthetic dataset, M = 20	0.06	0.34	-0.03	0.37	-0.09	0.32	23.4
	Multiple synthetic dataset, M = 50	0.07	0.34	-0.03	0.37	-0.09	0.32	24.2
	Direct regression	-0.09	0.59	-0.07	0.33	0.11	0.69	47.7
	CML, C = 0.0025	0.07	0.37	-0.04	0.14	-0.10	0.50	17.1
	CML, C = 0.0050	-0.03	0.47	-0.06	0.21	0.03	0.61	25.1
	CML, C = 0.0100	-0.08	0.54	-0.07	0.27	0.09	0.67	36.1
	CML, C = 0.0200	-0.09	0.58	-0.07	0.31	0.11	0.69	44.1
	CML, C = 0.2000	-0.09	0.59	-0.07	0.33	0.11	0.69	47.7
Scenario 4	Single synthetic dataset, S = 1, K = 5	-0.07	0.54	-0.07	0.25	0.10	0.71	33.9
	Single synthetic dataset, S = 10, K= 5	-0.04	0.51	-0.05	0.13	0.07	0.71	20.7
	Single synthetic dataset, S = 20, K = 5	-0.05	0.51	-0.05	0.12	0.08	0.71	19.9
	Single synthetic dataset, S = 50, K = 5	-0.04	0.51	-0.05	0.11	0.06	0.72	19.0
	Single synthetic dataset, S = 10, K = 10	-0.05	0.50	-0.04	0.13	0.08	0.68	20.4
	Single synthetic dataset, S = 20, K =10	-0.04	0.49	-0.05	0.12	0.08	0.69	19.0
	Single synthetic dataset, S = 50, K = 10	-0.05	0.50	-0.05	0.10	0.08	0.70	18.5
	Multiple synthetic dataset, M = 1	-0.03	0.58	-0.05	0.24	0.05	0.78	34.8
	Multiple synthetic dataset, M = 10	-0.02	0.50	-0.04	0.18	0.05	0.66	24.1
	Multiple synthetic dataset, M = 20	-0.03	0.49	-0.04	0.18	0.05	0.65	23.8
	Multiple synthetic dataset, M = 50	-0.03	0.49	-0.04	0.18	0.05	0.65	23.4

synthetic dataset method has reached its best achievement, we will not receive additional benefit from continuing increasing the sample size in the synthetic dataset.

The performance of the multiple synthetic dataset method with varying values of M is shown in Figure 4.4. Similar to the performance of single synthetic dataset method illustrated in Figure 4.4, the improvement in the predictive ability and the improvement in estimating efficiency diminish with increasing M beyond 20, and the performance of the multiple synthetic dataset method will be more stable after 40. This finding also agrees with our previous conclusion that when M increase from 10 to 50, the decreases in HL statistic will be very small. Also, we notice that when the sample size of the synthetic data is very large (i.e., S goes to 80 or M goes to 80), the HL statistic of the single synthetic dataset method will be lower than that of the multiple synthetic dataset method, and the MSE of the single synthetic dataset method will be smaller than that of the multiple synthetic dataset method.

4.6 Discussion

In this article, we have proposed two statistical methods for updating prediction models using information from external sources. The information from the external sources that we consider is the predicted outcomes, as if the results come from a "black-box" in which the input is X and the "black-box" outputs the estimated value of $\Pr(\mathbf{Y} = 1|\mathbf{X})$.

In the constrained ML, we construct a constraint for the regression parameters based on the predicted outcomes. The constraint is constructed by connecting three conditional distributions, $\Pr(\mathbf{Y} = 1|\mathbf{X})$, $\Pr(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B})$ and $\mathrm{f}(\mathbf{B}|\mathbf{X})$. For these three distributions, $\Pr(\mathbf{Y} = 1|\mathbf{X})$ is provided by the "black-box model", $\Pr(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B})$ is the target of interest, and an algorithm is used to draw values of $\mathbf{B}$ from $\mathrm{f}(\mathbf{B}|\mathbf{X})$. The constraint can be thought of as a form of regularization, which requires the parameter estimates for $\Pr(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B})$ be compatible with the external information.

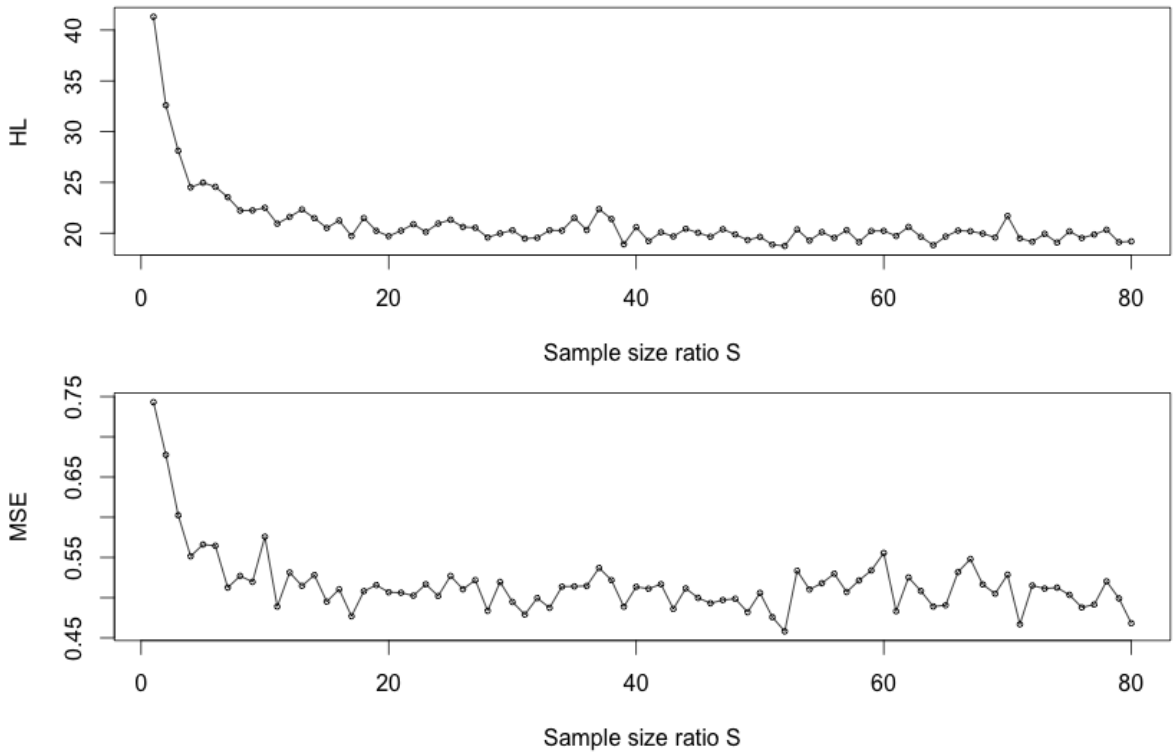


Figure 4.3: Performance of single synthetic dataset method in terms of HL statistic and sum of MSE, with varying values in S and fixed number of replicates in the multiple imputation procedure (K = 5) in scenario 1

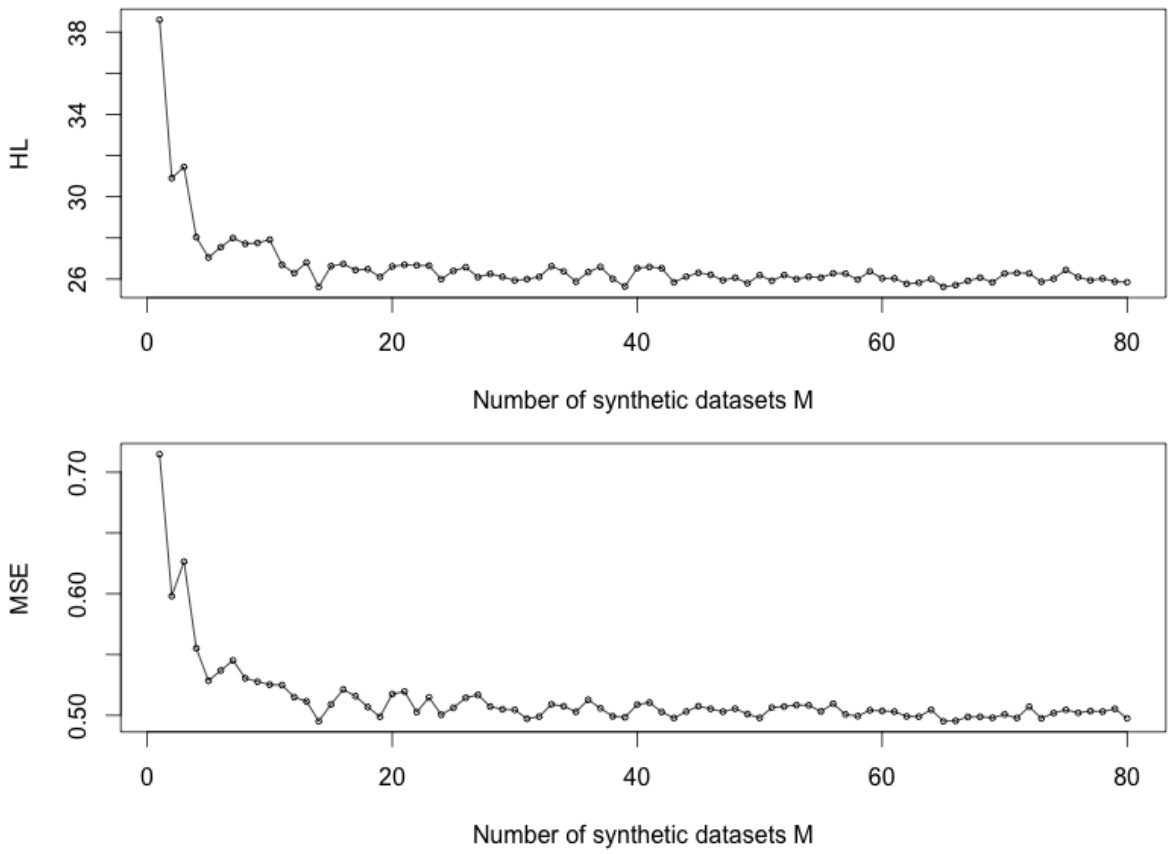


Figure 4.4: Performance of Multiple synthetic dataset method in terms of HL statistic and sum of MSE, with varying values in M and one replicate in the multiple imputation procedure in scenario 1

Though the constraint that we consider in the constrained ML method is to calculate the squared differences between $\bar{\mathrm{P}}(\mathbf{X})$ and the predicted outcomes from the expanded model, other kind of constraints may be possible choices, too. For example, we can calculate the absolute differences between $\bar{\mathrm{P}}(\mathbf{X})$ and the predicted outcomes from the expanded model, or measure the degree of similarity between the two rankings of these two set of probabilities using a rank correlation coefficient.

In the synthetic data methods, we create synthetic data based on the combination of the information from the external sources and the dataset containing complete data on $(\mathbf{Y}, \mathbf{X}, \mathbf{B})$. The synthetic $\mathbf{Y}$ is drawn from a $\mathrm{Bernoulli}(\bar{\mathrm{P}}(\mathbf{X}))$ distribution, where $\bar{\mathrm{P}}(\mathbf{X})$ comes from the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ from the external sources, this is possible even though the exact form of this conditional distribution is unknown. The missing values of $\mathbf{B}$ in the synthetic data are imputed using a multiple imputation algorithm in which the predicting variables include both $\mathbf{Y}$ and $\mathbf{X}$. The synthetic data methods can be thought of as a way to increase the sample size compared to the current small dataset. With the large sample size of the augmented data, the inference based on maximum likelihood estimation is likely to obtain some good large sample properties of the MLE which makes it very appealing for use.

Both methods involve a number of choices by the user. It is possible that these user-defined choices could be optimized to give more efficient and/or more robust estimates for the final model for $\mathrm{Pr}(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B})$.

For the constrained ML method the choices include the choice of C and the form of the constraint. We used a mean of the squared differences in probabilities in the constraint. Other measures of distance, such as absolute difference, or difference on a logit scale could be considered. A choice also has to be made for the method of drawing values of $\mathbf{B}$ given $\mathbf{X_i}$ and how many draws to use in equation (4.6). We present two simple alternatives, one based on a generalized linear model and one based on nearest neighbors. There are evidence in the simulation results (in Table 4.3) that

the nearest neighbors method is more robust when the generalized linear model is mis-specified. More sophisticated methods of drawing $\mathbf{B}$ could be developed, and would be necessary if $\mathbf{B}$ has a non-standard distribution other than the Gaussian and binary cases considered in the simulation study.

For the single synthetic dataset method the choices to be made are about S, the ratio of the sample size of the synthetic data and the sample size of the current small dataset, K the number of imputes of $\mathbf{B}$, and the method for drawing $\mathbf{B}$ given $\mathbf{X}$ and $\mathbf{Y}$. Other than the computational time, there is no reason why S and K should not be very large. However, there are diminishing returns for continually increasing S and K. For the multiple synthetic dataset method, the choices to be made are M, the number of synthetic datasets, and the method for drawing $\mathbf{B}$ given $\mathbf{X}$ and $\mathbf{Y}$. Larger value of M will induce more synthetic datasets. We use a generalized linear model as the basis for drawing values of $\mathbf{B}$ given $\mathbf{X}$ and $\mathbf{Y}$. More sophisticated methods could be developed that provide flexibility to handle complex relationships between $\mathbf{B}$ and $(\mathbf{X}, \mathbf{Y})$ and non-standard distribution of $\mathbf{B}$.

We note that both methods naturally generalize to the situation of multiple new variables $\mathbf{B}$, all that would be needed are algorithms to draw values of $\mathbf{B}$ from the appropriate multivariate distribution. Using the chained equation multiple imputation approach would be a possible solution for this.

We have not investigated how to calculate the standard errors for the parameters under the two methods. For the constrained ML, a possible approach is to use the bootstrap. For the synthetic data method, the variance of these estimates can be obtained using Rubin's multiple imputation rules.

In general most existing calculators that provide $\bar{\mathrm{P}}(\mathbf{X_i})$ do not provide a standard error for $\bar{\mathrm{P}}(\mathbf{X_i})$. In the synthetic data methods we are in fact assuming this standard error is zero, as if the external data was extremely large. This justifies using a very large value of S. If in fact the

external data was of modest size, or there were other reasons to have less confidence in $\bar{\mathrm{P}}(\mathbf{X_i})$, an ad-hoc adjustment would be to use a smaller value of S. The results in Table 4.5 show that a small value of S (specifically, $\mathrm{S} = 1$), lead to some gain in efficiency compared to direct regression, but considerably less gain than for larger values of S.

A finding from this chapter, as well as the previous chapters, is that using external information about the distribution of $\mathbf{Y}|\mathbf{X}$, can lead to considerable gains in efficiency in the coefficients of $\mathbf{X}$ in the model for $\mathbf{Y}|\mathbf{X}, \mathbf{B}$, but minimal gain in efficiency for the coefficient of $\mathbf{B}$. This is to be expected.

In contrast to the regression coefficients, the metrics associated with predictive accuracy are harder to improve. Gains in Brier score, HL statistic and AUC can be seen simply by using $\mathbf{B}$ (in direct regression) compared to not using $\mathbf{B}$. The additional gain by using the external information on $\mathbf{Y}|\mathbf{X}$ is more limited, and only clearly apparent for the HL statistic, which appears to be the most sensitive of the three metrics.

Future work will include a data analysis of prostate cancer data, using information from the prostate cancer prevention trial risk calculator. We also plan on extensions of the model to allow multiple calculators. Because in many cases it is possible that there are multiple calculators focusing on predicting the same outcomes, but with different predicting covariates. In the supplementary material, we show how the constrained ML method can be extended to the case of multiple calculators.

4.7 Supplementary material

4.7.1 *Multiple calculators*

If there are L multiple online calculators, each uses a different subset of $\mathbf{X}$. Assume a subject i is given a risk $\Pr(\mathrm{Y_i} = 1|\mathbf{X_i^{(l)}})$ from online calculator l and $\mathbf{X_i^{(l)}}$ are the predicting variables used in calculator l, $\mathrm{l} = 1, \ldots, \mathrm{L}$. Denote the predicted outcome from the lth calcula-

tor as $\bar{\mathrm{P}}(\mathbf{X_i^{(l)}})$ and denote by $\mathbf{X}^{(-\mathbf{l})}$ the variables in the current dataset but not in the lth calculator. An approximation of the conditional distribution of $\mathbf{Y}$ given $\mathbf{X_i^{(l)}}$ is given by $\Pr(\mathrm{Y_i} = 1|\mathbf{X_i^{(l)}}) = \int \Pr(\mathrm{Y_i} = 1|\mathbf{X_i^{(l)}}, \mathbf{X_i^{(-l)}})\mathrm{p}(\mathbf{X_i^{(-l)}}|\mathbf{X_i^{(l)}})\mathrm{d}\mathbf{X_i^{(-l)}} = \mathrm{E}_{(\mathbf{X_i^{(-l)}}|\mathbf{X_i^{(l)}})}\Pr(\mathrm{Y_i} = 1|\mathbf{X_i^{(l)}}, \mathbf{X_i^{(-l)}}) \approx \frac{1}{\mathrm{S}}\sum_{\mathrm{s}=1}^{\mathrm{S}} \Pr(\mathrm{Y_i} = 1|\mathbf{X_i^{(l)}}, \mathbf{X_i^{(-l),(s)}})$, where $\mathbf{X_i^{(-l),(s)}}$ are draw from the distribution of $\mathbf{X_i^{(-l)}}|\mathbf{X_i^{(l)}}$. As a result, $\bar{\mathrm{P}}(\mathbf{X_i^{(l)}}) \approx \frac{1}{\mathrm{S}}\sum_{\mathrm{s}=1}^{\mathrm{S}} \Pr(\mathrm{Y_i} = 1|\mathbf{X_i^{(l)}}, \mathbf{X_i^{(-l),(s)}}) = \frac{1}{\mathrm{S}}\sum_{\mathrm{s}=1}^{\mathrm{S}} \frac{\exp(\gamma_0+\mathbf{X_i^{(l)}}\boldsymbol{\gamma}_{\boldsymbol{X}(\boldsymbol{l})}+\mathbf{X_i^{(-l),(s)}}\boldsymbol{\gamma}_{\boldsymbol{X}(-\boldsymbol{l})})}{1+\exp(\gamma_0+\mathbf{X_i^{(l)}}\boldsymbol{\gamma}_{\boldsymbol{X}(\boldsymbol{l})}+\mathbf{X_i^{(-l),(s)}}\boldsymbol{\gamma}_{\boldsymbol{X}(-\boldsymbol{l})})}$.

Constrained ML

The constrained ML solution in the case of multiple calculators is:

$$\min_{\gamma}\left\{\sum_{\mathrm{i}=1}^{\mathrm{n}}\big[-\mathrm{Y_i}(\mathbf{X_i}\boldsymbol{\gamma}) + \log(1+\exp(\mathbf{X_i}\boldsymbol{\gamma}))\big]\right\} \tag{4.10}$$

$$\text{s.t. } \frac{1}{\mathrm{n}}\sum_{\mathrm{l}=1}^{\mathrm{L}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\bar{\mathrm{P}}(\mathbf{X_i^{(l)}}) - \frac{1}{\mathrm{S}}\sum_{\mathrm{s}=1}^{\mathrm{S}}\frac{\exp(\gamma_0+\mathbf{X_i^{(l)}}\boldsymbol{\gamma}_{\boldsymbol{X}(\boldsymbol{l})}+\mathbf{X_i^{(-l),(s)}}\boldsymbol{\gamma}_{\boldsymbol{X}(-\boldsymbol{l})})}{1+\exp(\gamma_0+\mathbf{X_i^{(l)}}\boldsymbol{\gamma}_{\boldsymbol{X}(\boldsymbol{l})}+\mathbf{X_i^{(-l),(s)}}\boldsymbol{\gamma}_{\boldsymbol{X}(-\boldsymbol{l})})}\right)^2 < \mathrm{C}$$

A modification of the constrained ML solution is:

$$\min_{\gamma}\left\{\sum_{\mathrm{i}=1}^{\mathrm{n}}\big[-\mathrm{Y_i}(\mathbf{X_i}\boldsymbol{\gamma}) + \log(1+\exp(\mathbf{X_i}\boldsymbol{\gamma}))\big]\right\} \tag{4.11}$$

$$\text{s.t. } \frac{1}{\mathrm{n}}\sum_{\mathrm{l}=1}^{\mathrm{L}}\mathrm{w_l}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\bar{\mathrm{P}}(\mathbf{X_i^{(l)}}) - \frac{1}{\mathrm{S}}\sum_{\mathrm{s}=1}^{\mathrm{S}}\frac{\exp(\gamma_0+\mathbf{X_i^{(l)}}\boldsymbol{\gamma}_{\boldsymbol{X}(\boldsymbol{l})}+\mathbf{X_i^{(-l),(s)}}\boldsymbol{\gamma}_{\boldsymbol{X}(-\boldsymbol{l})})}{1+\exp(\gamma_0+\mathbf{X_i^{(l)}}\boldsymbol{\gamma}_{\boldsymbol{X}(\boldsymbol{l})}+\mathbf{X_i^{(-l),(s)}}\boldsymbol{\gamma}_{\boldsymbol{X}(-\boldsymbol{l})})}\right)^2 < \mathrm{C}$$

Where $\mathrm{w_l} = \left(\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\bar{\mathrm{P}}(\mathbf{X_i^{(l)}}) - \frac{\exp(\mathbf{X_i^{(l)}}\hat{\boldsymbol{\beta}}_{\boldsymbol{X}(\boldsymbol{l})})}{1+\exp(\mathbf{X_i^{(l)}}\hat{\boldsymbol{\beta}}_{\boldsymbol{X}(\boldsymbol{l})})}\right)^2\right)^{-1}/\mathrm{W}$ and $\mathrm{W} = \sum_{\mathrm{l}=1}^{\mathrm{L}}\left(\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\bar{\mathrm{P}}(\mathbf{X_i^{(l)}}) - \frac{\exp(\mathbf{X_i^{(l)}}\hat{\boldsymbol{\beta}}_{\boldsymbol{X}(\boldsymbol{l})})}{1+\exp(\mathbf{X_i^{(l)}}\hat{\boldsymbol{\beta}}_{\boldsymbol{X}(\boldsymbol{l})})}\right)^2\right)^{-1}$ where $\hat{\boldsymbol{\beta}}_{\boldsymbol{X}^{(\boldsymbol{l})}}$ denote the result from direct regression based on covariates $\mathbf{X}^{(\mathbf{l})}$ without any constraints. So this weight gives more emphasis to the calculators whose predictions agree better with our data.

4.7.2 *Additional sensitivity analysis of single synthetic dataset method*

In the single synthetic dataset method, the synthetic data on $\mathbf{X}$, $\mathbf{X_j^\star}, \mathrm{j} = 1, ..., \mathrm{S} * \mathrm{n}$ are random samples of $\mathbf{X}$. Another choice of producing synthetic data on $\mathbf{X}$ is to use exact replicates of $\mathbf{X}$. That is, in step 1 of the implementing procedure of the single synthetic dataset method, we generate

Table 4.7: Results for single synthetic dataset method over 500 replications. The synthetic data on $\mathbf{X}$ are exact replicates of $\mathbf{X}$. The columns correspond to bias and Monte Carlo Standard Deviation for estimating the regression coefficient γ. The last column is HL statistic

Scenario	Method	$\hat{\gamma}_0$		$\hat{\gamma}_X$		$\hat{\gamma}_B$		HL
		Bias	SD	Bias	SD	Bias	SD	
	Single synthetic dataset, S = 1, K = 5	-0.14	0.65	0.05	0.36	0.04	0.42	41.4
	Single synthetic dataset, S = 10, K = 5	-0.09	0.56	0.03	0.22	0.00	0.41	22.1
	Single synthetic dataset, S = 20, K = 5	-0.12	0.54	0.01	0.21	0.02	0.41	19.9
	Single synthetic dataset, S = 50, K = 5	-0.11	0.53	0.01	0.19	0.01	0.40	19.9
	Single synthetic dataset, S = 10, K = 10	-0.13	0.52	0.02	0.21	0.02	0.39	20.4
	Single synthetic dataset, S = 20, K =10	-0.11	0.52	0.02	0.19	0.01	0.39	19.7
Scenario 1	Single synthetic dataset, S = 50, K = 10	-0.12	0.54	0.01	0.19	0.02	0.40	19.3
	Single synthetic dataset, S = 1, K = 5	-0.07	0.54	-0.05	0.24	0.08	0.71	31.5
	Single synthetic dataset, S= 10, K = 5	-0.06	0.49	-0.04	0.14	0.07	0.72	19.7
	Single synthetic dataset, S = 20, K = 5	-0.06	0.48	-0.05	0.12	0.08	0.69	18.0
	Single synthetic dataset, S = 50, K = 5	-0.05	0.47	-0.05	0.11	0.07	0.70	17.3
	Single synthetic dataset, S = 10, K = 10	-0.06	0.47	-0.04	0.14	0.07	0.68	18.3
	Single synthetic dataset, S = 20, K =10	-0.05	0.47	-0.04	0.12	0.07	0.69	18.0
Scenario 2	Single synthetic dataset, S = 50, K = 10	-0.05	0.47	-0.04	0.11	0.07	0.69	17.2

$\mathrm{S} * \mathrm{n}$ samples of $\mathbf{X}$ from the small dataset by duplicating exactly S copies of $\mathbf{X_i}, \mathrm{i} = 1, ..., \mathrm{n}$. The resulting synthetic data on $\mathbf{X}$ still has a sample size of $\mathrm{S} * \mathrm{n}$.

To investigate whether using the exact replicates of $\mathbf{X}$ instead of sampling random samples of $\mathbf{X}$ will influence the estimation results of the single synthetic dataset method, we implement the single synthetic dataset method under each of the four simulation scenarios, with exact replicates of $\mathbf{X}$.

Table 4.7 and Table 4.8 show the simulation results of single synthetic dataset method using exact replicates of $\mathbf{X}$, under the four scenarios described in Section 4.4. Compared with the simulation results of the single synthetic dataset method using random samples of $\mathbf{X}$, we do not observe obvious differences. Therefore, the randomness in choosing $\mathbf{X}$ samples will not likely affect the performance of the single synthetic dataset method.

In Figure 4.5, we see that when the single synthetic dataset method uses exact replicates of $\mathbf{X}$, again with increasing value of S and thus increasing sample size in the synthetic dataset, the HL statistic curve and the MSE curve will decrease at first and then approach a plateau. Compare to

Table 4.8: Results for single synthetic dataset method over 500 replications. The synthetic data on $\mathbf{X}$ are exact replicates of $\mathbf{X}$. The columns correspond to bias and Monte Carlo Standard Deviation for estimating the regression coefficient γ. The last column is HL statistic

Scenario	Method	$\hat{\gamma}_0$		$\hat{\gamma}_{\mathrm{X}}$		$\hat{\gamma}_{\mathrm{B}}$		HL
		Bias	SD	Bias	SD	Bias	SD	
	Single synthetic dataset, S = 1, K = 5	-0.04	0.43	0.01	0.44	-0.06	0.35	32.6
	Single synthetic dataset, S = 10, K = 5	0.07	0.31	-0.03	0.35	-0.14	0.35	22.1
	Single synthetic dataset, S = 20, K = 5	0.05	0.30	-0.01	0.36	-0.15	0.36	21.5
	Single synthetic dataset, S = 50, K = 5	0.06	0.30	-0.03	0.34	-0.15	0.35	19.5
	Single synthetic dataset, S = 10, K = 10	0.05	0.31	-0.02	0.34	-0.13	0.34	21.6
	Single synthetic dataset, S = 20, K =10	0.06	0.29	-0.02	0.33	-0.14	0.33	20.8
Scenario 3	Single synthetic dataset, S = 50, K = 10	0.06	0.29	-0.02	0.33	-0.15	0.34	20.4
	Single synthetic dataset, S = 1, K = 5	-0.07	0.56	-0.06	0.23	0.09	0.70	32.8
	Single synthetic dataset, S= 10, K = 5	-0.06	0.54	-0.05	0.14	0.10	0.74	22.2
	Single synthetic dataset, S = 20, K = 5	-0.04	0.53	-0.05	0.12	0.07	0.74	20.9
	Single synthetic dataset, S = 50, K = 5	-0.05	0.51	-0.05	0.12	0.07	0.72	19.8
	Single synthetic dataset, S = 10, K = 10	-0.04	0.50	-0.05	0.13	0.08	0.69	20.2
	Single synthetic dataset, S = 20, K =10	-0.06	0.50	-0.05	0.12	0.10	0.70	19.3
Scenario 4	Single synthetic dataset, S = 50, K = 10	-0.05	0.50	-0.05	0.11	0.09	0.70	18.6

Figure 4.3 when the single synthetic dataset method uses random samples of $\mathbf{X}$, we do not observe obvious differences.

4.7.3 *Additional simulation studies*

We also investigate the performances of the constrained ML and the two synthetic data methods proposed in this Chapter to the two simulation scenarios discussed in Chapter III. The simulation results are in Table 4.9 and Table 4.10.

For the first simulation scenario, we find that the two synthetic data methods can successfully reduces the bias in the MLE. The constrained ML drawing $\mathbf{B}$ from $\mathbf{B}|\mathbf{X}$ model with Firth correction applied has the lowest MSE for coefficients $\hat{\gamma}_0$, $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $\hat{\gamma}_3$, it also has the lowest Hosmer-Lemeshow statistic among all constraints methods. Compared with the estimates of the proposed constrained ML with Firth correction using summary-level information on regression coefficients in Table 3.1, the constrained ML drawing $\mathbf{B}$ from $\mathbf{B}|\mathbf{X}$ model with Firth correction applied has more biases for the regression coefficient $\hat{\gamma}_B$ but higher prediction power (their Hosmer-Lemeshow

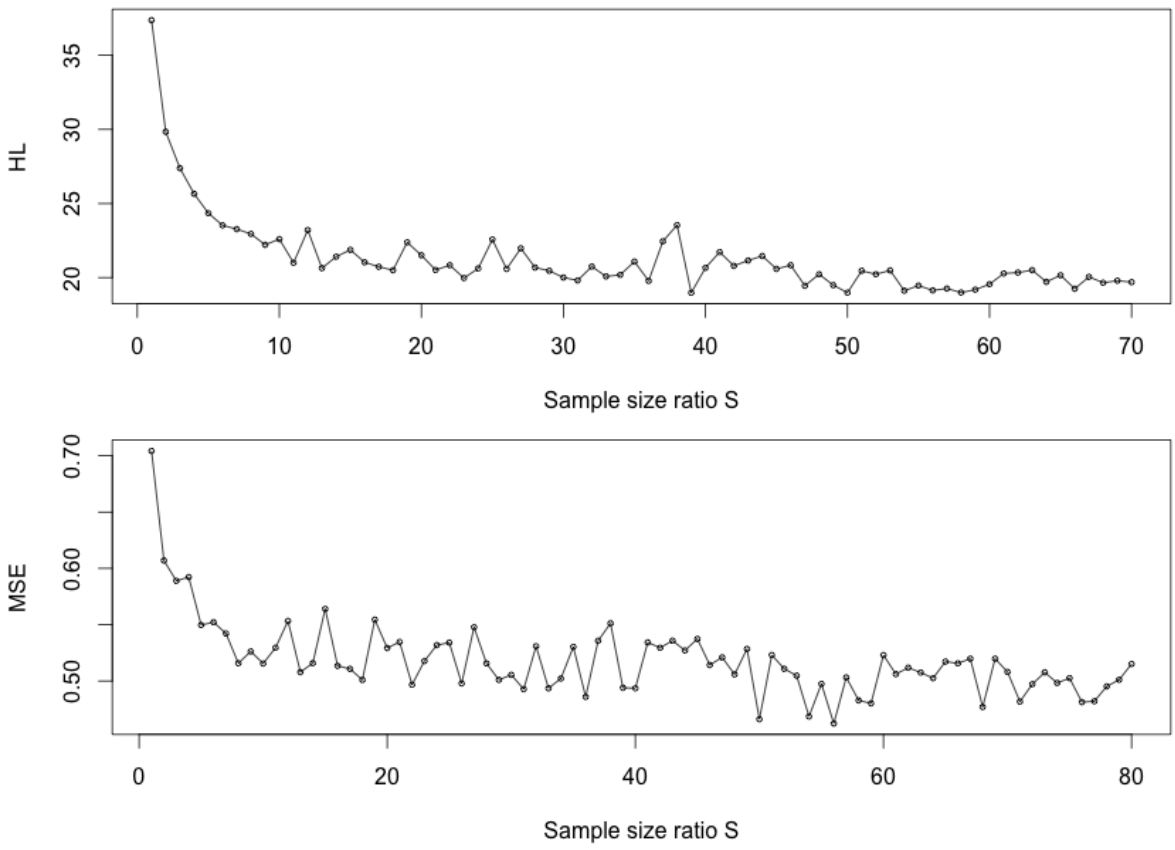


Figure 4.5: Performance of single synthetic dataset method using exact replicates of $\mathbf{X}$ in terms of HL statistic and sum of MSE, with varying values in S and fixed number of replicates in the multiple imputation procedure (K = 5) in scenario 1

Table 4.9: Simulation results of the first simulation scenario in Chapter III for Gaussian B : for each method, we report mean (Monte Carlo standard deviation), MSE, average Brier score, average Hosmer-Lemeshow statistic and average AUC

Method	$\hat{\gamma}_0$	$\hat{\gamma}_{\mathrm{X}_1}$	$\hat{\gamma}_{\mathrm{X}_2}$	$\hat{\gamma}_{\mathrm{B}}$	Brier Score	HL	AUC
True value	2	3	3	2	0.605	11.7	0.864
Online calculator $\bar{\mathrm{P}}(\mathrm{X_i})$					0.798	8.91	0.760
Direct regression	2.27(0.97)	3.37(1.80)	3.40(1.82)	2.35(0.91)	0.661	253.1	0.852
MSE	1.02	3.36	3.48	0.96			
Direct regression + Firth	1.94(0.76)	2.89(1.45)	2.92(1.48)	1.99(0.70)	0.651	78.3	0.852
MSE	0.58	2.10	2.18	0.49			
Plug-in $\hat{\mathrm{P}}_{\mathrm{i}}$					0.900	420.5	0.751
Constrained ML, draw B from $\mathrm{B}\|\mathrm{X}$, C = 0.005	1.96(0.42)	3.05(0.76)	3.13(0.80)	1.64(0.55)	0.623	673.0	0.857
MSE	0.17	0.58	0.66	0.43			
Constrained ML, draw B from $\mathrm{B}\|\mathrm{X}$ + Firth, C = 0.005	1.84(0.35)	2.86(0.66)	2.97(0.66)	1.50(0.38)	0.624	18.4	0.857
MSE	0.15	0.46	0.44	0.39			
Constrained ML, draw B from KNN, C = 0.005	2.07(0.51)	3.13(0.84)	3.25(0.89)	1.90(0.64)	0.627	43.0	0.858
MSE	0.26	0.72	0.85	0.41			
Constrained ML, draw B from KNN + Firth, C = 0.005	1.90(0.44)	2.90(0.77)	3.00(0.78)	1.70(0.54)	0.625	27.3	0.857
MSE	0.20	0.59	0.61	0.38			
Single synthetic dataset, S = 50, K = 10	1.99(0.41)	2.97(0.73)	3.09(0.73)	1.95(0.66)	0.625	35.2	0.858
MSE	0.17	0.53	0.54	0.43			
Multiple synthetic dataset, M = 50	1.96(0.50)	2.92(0.93)	2.99(0.95)	1.95(0.64)	0.630	39.8	0.857
MSE	0.25	0.88	0.90	0.41			

values are 36.9 and 18.4, respectively).

In the second simulation scenario, again these two synthetic data methods can reduce the biases of these regression coefficient estimates remarkably, compared to the these estimates of direct regression. The proposed constrained ML and the two synthetic data methods can reduce the MSE of the regression coefficients very well. In terms of predictive power, the single synthetic dataset method has the lowest Hosmer-Lemeshow statistic (its Hosmer-Lemeshow value is 19.9). However, in this simulation scenario the constrained ML with Firth correction, the informative full Bayes and the Bayesian transformation approach using summary-level information on regression coefficients has lower Hosmer-Lemeshow statistics (15.7, 14.5 and 15.8), as shown in Table 3.2. Compare the prediction ability of these methods shown in Table 4.9, Table 4.10 and the methods shown in Table 3.1 and Table 3.2, we do not have a definite answer about whether using summary-level information on regression coefficients will lead to more improvement in prediction ability than using summary-level information in the form of predicted outcomes from the established model.

Table 4.10: Simulation results of the second simulation scenario in Chapter III for binary B: for each method, we report mean (Monte Carlo standard deviation), MSE, average Brier score, average Hosmer-Lemeshow statistic and average AUC

Method	$\hat{\gamma}_0$	$\hat{\gamma}_{\mathrm{X}_1}$	$\hat{\gamma}_{\mathrm{X}_2}$	$\hat{\gamma}_{\mathrm{B}}$	Brier Score	HL	AUC
True value	2	4	4	2	0.522	39.8	0.874
Online calculator $\mathrm{P}(\mathrm{X_i})$					0.641	10.4	0.800
Direct regression	2.22(0.98)	4.49(1.80)	4.40(1.75)	2.22(0.88)	0.560	404.0	0.861
MSE	1.00	3.48	3.22	0.82			
Direct regression + Firth	1.96(0.82)	3.98(1.48)	3.90(1.43)	2.00(0.73)	0.554	66.3	0.862
MSE	0.67	2.18	2.05	0.53			
Plug-in $\hat{\mathrm{P}}_\mathrm{i}$					1.00	4.2×10^6	0.795
Constrained ML, draw B from $\mathrm{B}\vert\mathrm{X}$, C = 0.005	2.19(0.60)	4.27(0.99)	4.11(0.95)	1.97(0.61)	0.541	27.8	0.830
MSE	0.39	1.05	0.91	0.37			
Constrained ML, draw B from $\mathrm{B}\vert\mathrm{X}$ + Firth, C = 0.005	2.03(0.56)	3.96(0.94)	3.81(0.91)	1.84(0.59)	0.540	23.0	0.864
MSE	0.32	0.89	0.86	0.37			
Constrained ML, draw B from KNN, C = 0.005	2.17(0.63)	4.32(1.08)	4.15(1.04)	2.04(0.69)	0.543	34.3	0.841
MSE	0.43	1.26	1.11	0.47			
Constrained ML, draw B from KNN + Firth, C = 0.005	2.01(0.59)	3.98(0.99)	3.82(0.96)	1.88(0.62)	0.541	25.4	0.864
MSE	0.34	0.98	0.95	0.40			
Single synthetic dataset, S = 50, K = 10	1.99(0.35)	4.07(0.54)	3.90(0.57)	2.09(0.78)	0.534	19.9	0.866
MSE	0.12	0.29	0.34	0.61			
Multiple synthetic dataset, M = 50	1.98(0.48)	4.01(0.84)	3.89(0.82)	2.02(0.71)	0.539	24.8	0.865
MSE	0.23	0.71	0.68	0.51			

Bibliography

Chen, B. and Qin, J. Use of empirical likelihood to calibrate auxiliary information in partly linear monotone regression models. *Statistics in Medicine*, 33(10):1713–1722, 2014.

Damen, J. A. A. G., Hooft, L., Schuit, E., Debray, T. P. A., Collins, G. S., Tzoulaki, I., Lassale, C. M., Siontis, G. C. M., Chiocchia, V., Roberts, C., Schlüssel, M. M., Gerry, S., Black, J. A., Heus, P., van der Schouw, Y. T., Peelen, L. M. and Moons, K. G. M. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*, 353, 2016.

Ghalanos, A. and Theussl, S. *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*, 2015. R package version 1.16.

Raghunathan, T. E., Reiter, J. P. and Rubin, D. B. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1):1, 2003.

Reiter, J. P. Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18(4):531, 2002.

Reiter, J. P. and Kinney, S. K. Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Official Statistics*, 28(4):583–n/a, 2012.

Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc., 1987.

Tomlins, S. A., Day, J. R., Lonigro, R. J., Hovelson, D. H., Siddiqui, J., Kunju, L. P., Dunn, R. L., Meyer, S., Hodge, P., Groskopf, J., Wei, J. T. and Chinnaiyan, A. M. Urine TMPRSS2:ERG plus PCA3 for individualized prostate cancer risk assessment. *European Urology*, 70:45–53, 2015.

Van Buuren, S. and Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011.

Ye, Y. *Interior Algorithms for Linear, Quadratic, and Linearly Constrained Non-Linear Programming*. Ph.D. thesis, Department of ESS, Stanford University, 1987.

Zhan, X. and Ghosh, D. Incorporating auxiliary information for improved prediction using combination of kernel machines. *Statistical Methodology*, 22:47–57, 2015.

CHAPTER V

Discussion

In this dissertation, we consider a variety of statistical approaches to inform a risk prediction model with external summary-level information. In biomedical and clinical research, many times risk prediction models describe the association between a outcome variable $\mathbf{Y}$ and a set of predicting factors $\mathbf{X}$, and the only publicly accessible information about this risk prediction model is some summary-level information. We proposed statistical approaches that involve both frequentist and Bayesian methods to utilize this summary-level information for improving inference in an expanded model of interest, $\mathbf{Y}$ given $\mathbf{X}, \mathbf{B}$. We find that how one incorporates varying types of summary-level information for various types of risk prediction models is a non-trivial statistical problem, especially when the link function is non-linear.

In Chapter II, we demonstrate how to incorporate external summary-level information in a linear regression model $\mathbf{Y}|\mathbf{X}, \mathbf{B}$, when the external summary-level information is in the form of estimated regression coefficients and their standard errors in a reduced model $\mathbf{Y}|\mathbf{X}$. We find that there are exact relationship equations connecting parameters in models $\mathrm{E}(\mathbf{Y}|\mathbf{X}, \mathbf{B})$, $\mathrm{E}(\mathbf{B}|\mathbf{X})$ and $\mathrm{E}(\mathbf{Y}|\mathbf{X})$, and thus the external summary-level information in $\mathbf{Y}|\mathbf{X}$ can be translated into constraints for the regression coefficients in $\mathrm{E}(\mathbf{Y}|\mathbf{X}, \mathbf{B})$. We propose four constrained solutions (i.e., constrained ML, partial regression, informative full Bayes and Bayesian transformation) and these methods are shown to work well in both simulation studies and in the data analysis of a bone lead

prediction model.

In Chapter III, we extend the constrained methods proposed in Chapter II to the setting of a logistic regression model describing $\Pr(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B})$. The relationship equations connecting parameters in $\Pr(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B})$, $\mathrm{E}(\mathbf{B}|\mathbf{X})$ and $\Pr(\mathbf{Y} = 1|\mathbf{X})$ are no longer exact equations but rather dependent on approximations. As such approximations are different depending on the form of the variable $\mathbf{B}$, we considered both continuous and binary $\mathbf{B}$ separately. Three proposed constrained solutions including constrained ML, informative full Bayes and Bayesian transformation approach still work well for using external summary-level information to help improve the efficiency of estimation and enhancing the predictive ability of $\Pr(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B})$.

In general the methods in Chapter II and Chapter III rely on establishing a relationship between the parameters in the model for $\mathbf{Y}|\mathbf{X}$ and the parameters in the model for $\mathbf{Y}|\mathbf{X}, \mathbf{B}$. We demonstrate that this is possible using the Law of Total Expectation and algebraic approximations if $\mathbf{B}$ has a standard parametric distribution. However, this approach may be more problematic if $\mathbf{B}$ has a non-standard distribution, or is a mixed set of continuous and categorical variables.

In Chapter IV, the external summary-level information that we consider is no longer in the form of estimated regression coefficients and their standard errors but comes in the form of predicted outcomes from a "black-box" in which the input is X and the "black-box" outputs the estimated value of $\Pr(\mathbf{Y} = 1|\mathbf{X})$. The two proposed statistical approaches, constrained ML and the synthetic data approach are quite distinct from the constrained methods proposed in Chapter II and Chapter III. Simulation studies show gains in estimating efficiency for the regression parameters in $\Pr(\mathbf{Y} = 1|\mathbf{X}, \mathbf{B})$ by using such kind of external summary-level information, but the metrics associated with predictive accuracy are harder to improve. The methods in this chapter do generalize to the situation of non-standard distributions for $\mathbf{B}$, contingent upon a valid method to draw values of $\mathbf{B}$, either from a model for $\mathbf{B}|\mathbf{X}$, or from a model for $\mathbf{B}|\mathbf{X}, \mathbf{Y}$. Both methods also generalize to the situation of multiple new biomarkers and multiple existing online calculators. Thus we think they

are broadly applicable, and worthy of further research.

One point of future consideration is the choice of tuning parameter. For the constrained ML and the Bayesian transformation approach proposed in Chapter II and Chapter III, we have a quantity labeled d that controls the degree of trust in the historical information and we select it by drawing from a half normal distribution $|\mathrm{N}(0, \sigma_{\mathrm{d}}^2)|$ in Bayesian transformation approach and we fix $d = 1$ in the constrained ML approach. However, σ_d^2 and d can be considered as tuning parameters and chosen adaptively for a dataset. For the proposed methods in Chapter II, we investigate extensively how to select σ_{d}^2 via cross validation. Cross validation is performed for tuning σ_{d}^2 for a large number of possible values of σ_{d}^2 and choosing the optimal σ_{d}^2 that gives the lowest cross validation average error. The simulation results show that the proposed constrained methods with the chosen optimal σ_{d}^2 do not have satisfactory performances (these simulation results are not presented in this thesis). Thus, how to select these tuning parameters in a principled optimal manner needs further investigation.

Though the proposed constrained methods in Chapter II, Chapter III and Chapter IV show improvements in estimating the regression coefficients and predictive ability in the expanded model when the new biomarker $\mathbf{B}$ is a single covariate, the generalization of these proposed methods to the the situation of multiple $\mathbf{B}$ will be critical for future research. $\mathbf{B}$ can be a mixed set of continuous and categorical variables or it can consist of many variables. How to extend the proposed methods to the scenario of multiple $\mathbf{B}$ worths further investigation and will make the proposed constrained methods applicable to more complicated distributions of $\mathbf{B}$.

Another direction for future research is the robustness of these proposed methods when the $\mathbf{Y}|\mathbf{X}$ model, $\mathbf{Y}|\mathbf{X}, \mathbf{B}$ model or the $\mathbf{B}|\mathbf{X}$ model is mis-specified. For example, in Chapter II and Chapter III, we assume that the true model of $\mathbf{Y}|\mathbf{X}, \mathbf{B}$ is linear in $(\mathbf{X}, \mathbf{B})$ and is the same in the external dataset used for estimating the established model and the current small dataset used for fitting the expanded model. However, if the true model of $\mathbf{Y}|\mathbf{X}, \mathbf{B}$ is not linear in $(\mathbf{X}, \mathbf{B})$

(e.g., involve interaction terms between subset of $\mathbf{X}$) but we still fit an expanded model linear in $(\mathbf{X}, \mathbf{B})$, the relationship equations connecting parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ will be less accurate and the performances of the proposed constrained methods may be affected by such mis-specification. Future studies demonstrating the robustness of these proposed methods under mis-specification will make them more appealing.