Research Article

Statistics in Medicine

Received 17 December 2012, Accepted 15 July 2013 Published online 12 August 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5940

Propensity score-based diagnostics for categorical response regression models

Philip S. Boonstra,[a*†] Irina Bondarenko,[a] Sung Kyun Park,[b] Pantel S. Vokonas[c,d] and Bhramar Mukherjee[a]

For binary or categorical response models, most goodness-of-fit statistics are based on the notion of partitioning the subjects into groups or regions and comparing the observed and predicted responses in these regions by a suitable chi-squared distribution. Existing strategies create this partition based on the predicted response probabilities, or propensity scores, from the fitted model. In this paper, we follow a retrospective approach, borrowing the notion of balancing scores used in causal inference to inspect the conditional distribution of the predictors, given the propensity scores, in each category of the response to assess model adequacy. We can use this diagnostic under both prospective and retrospective sampling designs, and it may ascertain general forms of misspecification. We first present simple graphical and numerical summaries that can be used in a binary logistic model. We then generalize the tools to propose model diagnostics for the proportional odds model. We illustrate the methods with simulation studies and two data examples: (i) a case-control study of the association between cumulative lead exposure and Parkinson's disease in the Boston, Massachusetts, area and (ii) and a cohort study of biomarkers possibly associated with diabetes, from the VA Normative Aging Study.

Keywords: balancing score; multinomial logistic; proportional odds; residual diagnostic; score test

1. Introduction

To check model fit and adequacy for categorical response regression models, it is common to compare observed frequencies and estimated expected frequencies within a given partition of the covariate space. The expected frequencies are calculated under the assumed model and compared with the observed frequencies via a χ^2-type goodness-of-fit statistic, as in the Hosmer and Lemeshow (HL) statistics [1, 2]. These statistics use a grouping strategy based on the values of estimated response probabilities given the predictors and compare the observed and the expected responses within the groups. Two methods of grouping based on the estimated probabilities are discussed in [2]: (i) fixed cutpoints in the $[0, 1]$ interval or (ii) sample quantiles/deciles with equal sized groups. A disadvantage of both approaches is a dependence on the choice of cutpoints. Tsiatis [3] suggests a score test that circumvents the problem of HL statistics lacking discriminatory power in the region of the predictor space that gives rise to the same estimated probability. However, this test is also dependent on the choice of partition in the exposure space. Stukel [4] adopts a generalized logistic model framework to test the adequacy of a fitted logistic model. le Cessie and van Houwelingen [5] address these problems by proposing a class of tests based on smoothed residuals, while Royston [6] uses the partial sums of residuals. Hosmer *et al.* [7] present an excellent overview.

Lipsitz *et al.* [8] generalizes the popular HL statistic proposed for a logistic regression model with binary data to regression with c-category ordinal responses. Toledano and Gatsonis [9] gives a generalization of a receiver operating characteristic curve that plots sensitivity against (1 - specificity) for

[a]*Department of Biostatistics, University of Michigan, Ann Arbor, MI, 48109, U.S.A.*
[b]*Department of Epidemiology, University of Michigan, Ann Arbor, MI, 48109, U.S.A.*
[c]*VA Normative Aging Study, Veterans Affairs Boston Healthcare System, Boston, MA, U.S.A.*
[d]*Department of Epidemiology, Boston University School of Medicine and School of Public Health, Boston, MA, U.S.A.*
**Correspondence to: Philip S. Boonstra, Department of Biostatistics, University of Michigan, Ann Arbor, MI, 48109, U.S.A.*
[†]*E-mail: philb@umich.edu*

every possible collapsing of the c categories. Kim [10] proposes a graphical method for assessing the proportional odds assumption. All of the aforementioned methods check the overall adequacy of the proportional odds model, but they do not give a close view of model misspecification corresponding to specific covariates.

Lin *et al.* [11, 12] and Arbogast and Lin [13] develop graphical and numerical methods for assessing the adequacy of the functional form of a covariate in Cox regression and the logistic regression model using the cumulative sums of residuals. In standard linear regression models, the plot of residuals against the explanatory variable $\boldsymbol{x}$ is often viewed as a diagnostic tool to examine model misspecification in $\boldsymbol{x}$. The residuals for a binary logistic model are typically defined as the difference between observed response and the estimated probability of the response, conditional on the covariates. The plot of the residuals versus $\boldsymbol{x}$ is difficult to interpret in such cases. Arbogast and Lin [13] recommend using cumulative sums of the residuals over the covariate of interest to check for functional misspecification in $\boldsymbol{x}$. When the model is correctly specified, they show that the cumulative residual process converges weakly to a zero-mean Gaussian process. Thus, they compare the observed pattern of the cumulative residuals with simulated realizations based on the limiting Gaussian process under the null hypothesis that the model is correctly specified. Liu *et al.* [14] generalize this graphical diagnostic idea to the proportional odds model. Beyond this limited work, graphical diagnostics for ordinal data models have not gained wider acceptance. The procedures that involve cumulative-sum-based residuals are computationally hard to simulate from a limiting Gaussian process distribution, which restricts their use.

Simple graphical diagnostics for overall model assessment are still lacking for even binary regression models because of the discreteness of data. Landwehr *et al.* [15] propose graphical tools like local mean deviance plots, empirical probability plots, and smoothed partial residual plots for detecting model inadequacies for binary data. In a discussion of [15], Rubin [16] proposes that for model assessment, one should take a retrospective view, examining the implied distribution of the covariate $\boldsymbol{x}$ in the $y = 0$ and $y = 1$ groups rather than the prospective distribution $p(y = 1|\boldsymbol{x})$. The fundamental idea lies in a powerful result established by Rosenbaum and Rubin [17]: Given the true $\pi(x) = p(y = 1|\boldsymbol{x})$, $\boldsymbol{x}$ and y are conditionally independent. This result suggests that if $\hat{\pi}$ is an adequate estimate of $p(y = 1|\boldsymbol{x})$, then the observed differences between $p(\boldsymbol{x}|\hat{\pi}, y = 1)$ and $p(\boldsymbol{x}|\hat{\pi}, y = 0)$ may point to elements of $\boldsymbol{x}$ that need to be adjusted in the model and also help to identify outliers. The basic idea is to enrich the model for $\hat{\pi}$ such that conditional on $\hat{\pi}$, the covariates $\boldsymbol{x}$ are independent of the outcome y. Numerical tests for model adequacy can also be constructed, but easy-to-understand graphical tools are the principal attraction of such an approach. Extending this idea of Rosenbaum and Rubin [17], which is primarily used to balance covariates between *treatment groups*, we develop graphical diagnostics for the logistic model, balancing covariates between *response groups*, and also extend the idea to outcomes with more than two categories. We present graphical diagnostics for the proportional odds model, one of the most popular models for ordinal data. The extension of propensity scores to the proportional odds model is indicated in Joffe and Rosenbaum [18], who point out that a scalar balancing score is sufficient under the proportional odds structure. The aforementioned 'balancing score' result of Rosenbaum and Rubin [17] is technically extended to generalized treatment regimes, beyond the binary case, in Imbens [19]. Lu *et al.* [20] apply matching in terms of the scalar balancing score to an observational study of drug abuse. Imai and van Dyk [21] provide a general theory that covers continuous, semi-continuous, and multivariate treatment (y in our notation) regimes and effectively balance a high-dimensional covariate, $\boldsymbol{x}$, by a low-dimensional function of the propensity scores, perhaps not scalar. None of this work has been applied to the development of diagnostic tools. Note that unlike the causal inference literature, we do not have any 'treatment' groups or a randomized trial. Rather, we are simply exploiting the mathematical result of conditional independence as justification for our model diagnostics.

Our proposed graphical diagnostics, which follow Rubin's idea, are computationally simple, as is the motivating rationale. Further, owing to the retrospective formulation, they may be readily used under outcome dependent sampling, for example, in a case-control study. There have been concerns regarding the use of propensity scores in case-control or case-cohort studies for adjusting confounders [22] because of artifactual effect modification and reduced ability to control for potential confounding factors, when estimating treatment effects. Because treatment-effect estimation is not our goal here, those concerns do not apply. The conditional independence result between y and $\boldsymbol{x}$ conditional on $\boldsymbol{\pi}(\boldsymbol{x})$ holds in a case-control study because the propensity score model from prospective and retrospective likelihoods differs only through the intercept term, which does not involve $\boldsymbol{x}$. Our proposals for model diagnostics are based solely on this independence result and do not change for cohort versus case-control sampling.

We organize the rest of the paper as follows. In Section 2, we review the result of Rosenbaum and Rubin [17] for the binary logistic model and its extension to the proportional odds model. We describe how these results lead to some proposed graphical and numerical diagnostics and present simulation results investigating the efficacy of these diagnostics for the binary logistic and proportional odds models. In Section 3, we analyze two datasets to assess model adequacy: A logistic regression example comes from a case-control study of Parkinson's disease (PD) and its association with lead exposure, and the proportional odds example originates from a cohort study of diabetes. Section 4 ends with a discussion.

2. The balancing score and its extension

Rosenbaum and Rubin [17] prove a simple but powerful result: a binary outcome y (which is treatment in their context) and a set of covariates $\boldsymbol{x}$ are conditionally independent given the propensity score $\pi = \pi(\boldsymbol{x}) = p(y = 1|\boldsymbol{x})$. This implies the following:

(1) $p(\boldsymbol{x}, y|\pi) = p(\boldsymbol{x}|\pi)p(y|\pi)$.
(2) $p(\boldsymbol{x}|\pi, y) = p(\boldsymbol{x}|\pi)$.
(3) $p(\boldsymbol{x}|\pi, y = 1) = p(\boldsymbol{x}|\pi, y = 0)$.

In fact, they prove a more general result. Our presentation is simpler, and our argument differs from the original proof so as to provide insight into the subsequent extension and construction of the diagnostics.

Theorem 2 from Rosenbaum and Rubin [17]
For a binary outcome y, a set of length-p covariates $\boldsymbol{x}$, and a function $b(\boldsymbol{x})$, which is the balancing score,

$$\boldsymbol{x} \perp\!\!\!\perp y|b(\boldsymbol{x}) \Leftrightarrow \pi(\boldsymbol{x}) = h(b(\boldsymbol{x})),$$

for some function h.

Proof

(i) Left-hand side (LHS) ⇒ Right-hand side (RHS) We note that, in general,

$$f(y, \boldsymbol{x}|b(\boldsymbol{x})) = f_Y(y|\boldsymbol{x}, b(\boldsymbol{x}))\, f_X(\boldsymbol{x}|b(\boldsymbol{x})).$$

Thus, the conditional independence in the LHS of the theorem, namely $f(y, \boldsymbol{x}|b(x)) = f_Y(y|b(\boldsymbol{x}))\, f_X(\boldsymbol{x}|b(\boldsymbol{x}))$, is equivalent to stating that

$$f_Y(y|b(\boldsymbol{x})) = f_Y(y|x, b(\boldsymbol{x})).$$

Hence,

$$\begin{aligned}\pi(\boldsymbol{x}) = \mathbb{E}_Y[\, y|\boldsymbol{x}\,] &= \int y f_Y(y|\boldsymbol{x})\, \mathrm{d}y \\ &= \int y f_Y(y|b(\boldsymbol{x}))\, \mathrm{d}y,\end{aligned}$$

which, by construction, must be a function $h(b(\boldsymbol{x}))$ for some $h(\cdot)$.

(ii) RHS ⇒ LHS: Consider the conditional mean of $y|b(\boldsymbol{x})$:

$$\begin{aligned}\mathbb{E}[\, y|b(\boldsymbol{x})\,] &= \mathbb{E}_{\boldsymbol{x}|b(\boldsymbol{x})}[\,\mathbb{E}[\, y|b(\boldsymbol{x}), \boldsymbol{x}\,]\,] \\ &= \mathbb{E}_{\boldsymbol{x}|b(\boldsymbol{x})}[\,\pi(\boldsymbol{x})\,], \text{ by definition.}\end{aligned}$$

By the condition in RHS, $\pi(\boldsymbol{x}) = h(b(\boldsymbol{x}))$ is a deterministic function of $b(\boldsymbol{x})$. Hence, $\mathbb{E}_{\boldsymbol{x}|b(\boldsymbol{x})}[\,\pi(x)\,] = \pi(x)$, and

$$\mathbb{E}[\, y|b(\boldsymbol{x})\,] = \pi(\boldsymbol{x}).$$

Now, as the mean of $y|\boldsymbol{x}$ determines its (conditional) distribution, we see that

$$f_Y(y|b(\boldsymbol{x})) = f_Y(y|\boldsymbol{x}, b(\boldsymbol{x})) = \pi(\boldsymbol{x}),$$

and, thus, the separation property required must hold.

□

Now consider a general categorical outcome $y \in 0, 1, \cdots, K$. There is, in general, no scalar function of the covariates that can satisfy the aforementioned balancing score property as $\mathbb{E}(y|\boldsymbol{x})$ does not fully describe the distribution of y. The proof presented previously is very much contingent upon this property. However, for the proportional odds or ordinal-logit model,

$$\text{logit Pr}(y \leq k|\boldsymbol{x}) = \alpha_k + \boldsymbol{\beta}^\top \boldsymbol{x}, \qquad k = 0, 1, \cdots, K-2,$$

where $\text{logit}(x) = \log(x/[1-x])$. Thus, the distribution of $y|\boldsymbol{x}$ depends on $\boldsymbol{x}$ only through $b(\boldsymbol{x}) = \boldsymbol{\beta}^\top \boldsymbol{x}$, that is, $\Pr(y = k|\boldsymbol{x}) = \Pr(y = k, b(\boldsymbol{x}) = \boldsymbol{\beta}^\top \boldsymbol{x})$, and a single balancing score determines the entire distribution of y. It also follows that

(1) $p(\boldsymbol{x}, y = k|b(\boldsymbol{x})) = p(\boldsymbol{x}|b(\boldsymbol{x}))p(y|b(\boldsymbol{x}))$.
(2) $p(\boldsymbol{x}|b(\boldsymbol{x}), y = k) = p(\boldsymbol{x}|b(\boldsymbol{x}))$.
(3) $p(\boldsymbol{x}|b(\boldsymbol{x}), y = k) = p(\boldsymbol{x}|b(\boldsymbol{x}), y = k')$, for all $k' \neq k$.

Remark 1
For general polytomous models, Imbens [19] defines a concept of weak dependence through generalized balancing scores $b(k, x)$ where $b(k, \boldsymbol{x}) = p(y = k|\boldsymbol{x})$ and shows that $\mathbb{E}(\boldsymbol{x}|b(k, \boldsymbol{x}), y = k) = \mathbb{E}(\boldsymbol{x}|b(k, \boldsymbol{x})$. This result may be used to extend the aforementioned three equalities to the multinomial logistic model, a potentially more flexible alternative to the proportional odds model. However, no distributional independence of y and $\boldsymbol{x}$ can be deduced with a general structure of $b(k, \boldsymbol{x})$, and we do not further develop this extension.

2.1. *Graphical summaries*

To present our proposed graphical diagnostics for model misspecification in a concrete fashion, consider the following generating model for binary outcomes:

$$\text{logit Pr}(y = 1|\boldsymbol{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_1 x_2. \tag{2.1.1}$$

The linear predictor, $\boldsymbol{\beta}^\top \boldsymbol{x}$, is the balancing score, which shares a 1-1 correspondence with the propensity score, $\pi = \Pr(y = 1|\boldsymbol{x})$. Theorem 2 implies that y is independent of $\boldsymbol{x}$ given $\boldsymbol{\beta}^\top \boldsymbol{x}$. Suppose we estimate the balancing score of (2.1.1) assuming $\beta_4 = \beta_5 = 0$ in the fitted model. Theorem 2 is satisfied only when this assumption is true in the generating model. If $\beta_4 \neq 0$ or $\beta_5 \neq 0$, then $p(\boldsymbol{x}|\pi, y = 1) \neq p(\boldsymbol{x}|\pi, y = 0)$, meaning there will be residual association between y and $\boldsymbol{x}$. Consider next two generating models for ordinal outcomes

$$\text{logit Pr}(y \leq k|\boldsymbol{x}) = \alpha_k + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \quad k = 0, \ldots, K-2, \tag{2.1.2}$$

$$\Pr(y = k|\boldsymbol{x}) \propto \exp\{\alpha_k + \beta_{1k} x_1 + \beta_2 x_2 + \beta_3 x_3\}, \quad k = 0, \ldots, K-2. \tag{2.1.3}$$

Here, only for (2.1.2) is $\boldsymbol{\beta}^\top \boldsymbol{x}$ the balancing score, and, in this case, the extension of Theorem 2 to ordinal outcomes implies that y is independent of $\boldsymbol{x}$ given $\boldsymbol{\beta}^\top \boldsymbol{x}$. The linear predictor is not a balancing score in (2.1.3) for two reasons: (i) the data are generated from a multinomial logit model, which automatically violates the proportional odds assumption, and (ii) β_{1k} depends on k. Note that for each model, the probability of falling in category $K-1$ is constrained for identifiability purposes. If we estimate the balancing score assuming (2.1.2) but the data are generated according to (2.1.3), there will be residual association between each y and $\boldsymbol{x}$ conditional on $\hat{\boldsymbol{\beta}}^\top \boldsymbol{x}$. Based on these conclusions, we propose two graphical diagnostics for model misspecification. Figures 1 and 2 present representative plots for binary and ordinal outcomes, respectively, which we now describe:

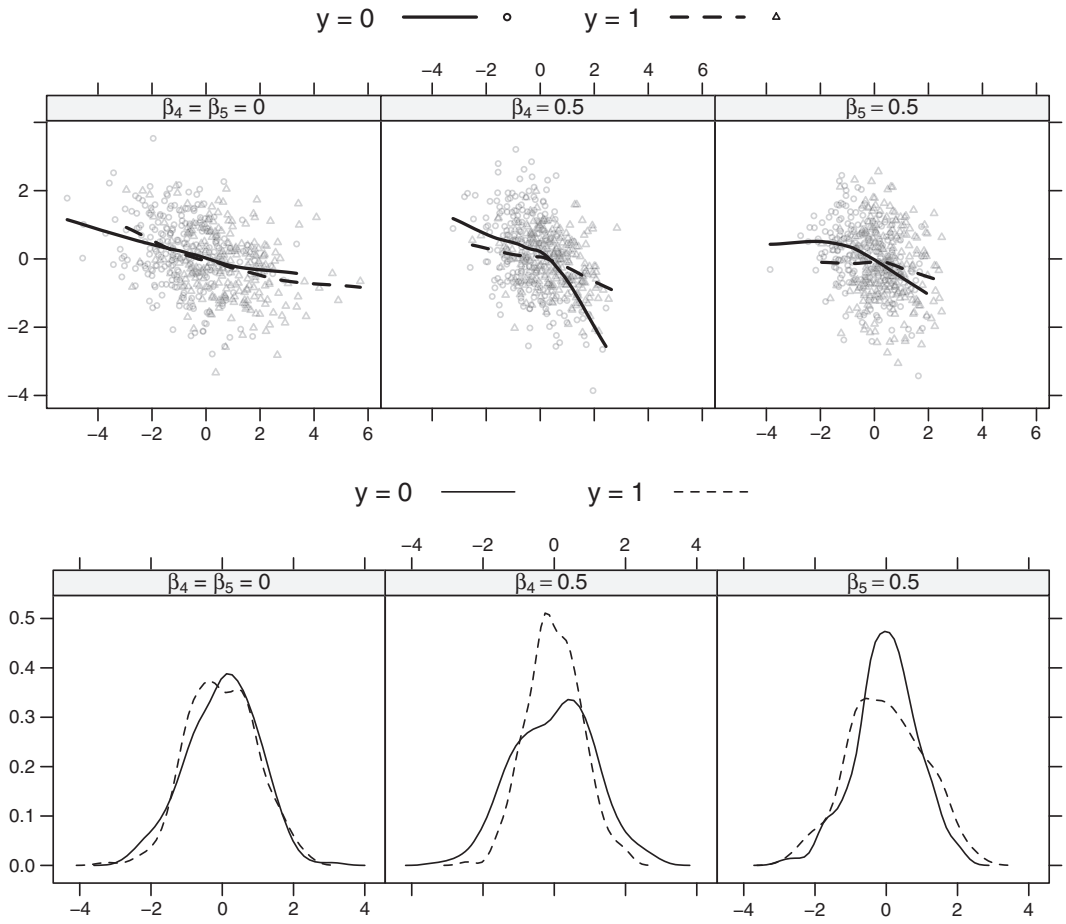


Figure 1. The *Scatterplot Diagnostic* (top row) and *Residual Density Diagnostic* (bottom row) for three simulated datasets, one for each column, of size $n = 500$ generated from (2.1.1). For all three columns, $\beta_1 = 0.5$, $\beta_2 = 1.0$, $\beta_3 = -0.5$, α is such that $\Pr(y = 1) = 0.5$, and the fitted model assumes $\beta_4 = \beta_5 = 0$. In the first column, $\beta_4 = \beta_5 = 0$, and the curves corresponding to $y = 0$ and $y = 1$ are approximately collinear because the fitted model matches the generating model. In the second and third columns, $\beta_4 = 0.5$ and $\beta_5 = 0.5$, respectively, and so the functional form is violated in the fitted model.

Scatterplot Diagnostic. *Grouping by the response y, plot the individual components of $\boldsymbol{x}$ versus $\hat{\boldsymbol{\beta}}^\top \boldsymbol{x}$ and annotate with a locally weighted scatterplot smoothing (LOWESS) curve for each group.* In a model with sufficiently good fit, the conditional independence result gives that the distribution of $\boldsymbol{x}$ at each value of $\hat{\boldsymbol{\beta}}^\top \boldsymbol{x}$ should be 'similar'. The *LOWESS* curve summarizes the degree of similarity between groups and allows for simple comparisons between models to assess relative improvement of model fit. The first row of Figure 1 gives examples of this for three simulated datasets of size 500. In each, we plot x_1 against $\hat{\boldsymbol{\beta}}^\top \boldsymbol{x}$, coming from fitting (2.1.1) and assuming $\beta_4 = \beta_5 = 0$. In all cases, $\beta_1 = 0.5$, $\beta_2 = 1.0$, and $\beta_3 = -0.5$, and α is such that $\Pr(y = 1) = 0.5$. In the first column, $\beta_4 = \beta_5 = 0$, and the two lines corresponding to $y = 0$ or $y = 1$ are approximately collinear. As the effect of x_1 changes in the quadratic term ($\beta_4 = 0.5$, the second column) or upon inclusion of an interaction term ($\beta_5 = 0.5$, the third column), the *LOWESS* curves separate. The first row of Figure 2 presents this same diagnostic for the case of ordinal outcomes. We generated two datasets of size 500, one each from (2.1.2) and (2.1.3), estimated $\boldsymbol{\beta}^\top \boldsymbol{x}$ in both, assuming (2.1.2), and compared *LOWESS* curves of x_1 by $\hat{\boldsymbol{\beta}}^\top \boldsymbol{x}$ for each level of y. The figure's caption gives the true coefficient values. In the second column, where the generating model is (2.1.3), the curve for the $y = 1$ group is shifted down compared to the other groups.

Residual Density Diagnostic. *Plot the kernel density estimate (KDE) of the residuals corresponding to the regression of each component of $\boldsymbol{x}$ on $\hat{\boldsymbol{\beta}}^\top \boldsymbol{x}$, grouping by the response y.* Models with good fit should result in plots in which the *KDE* curves for different values of y are similar in shape and location.

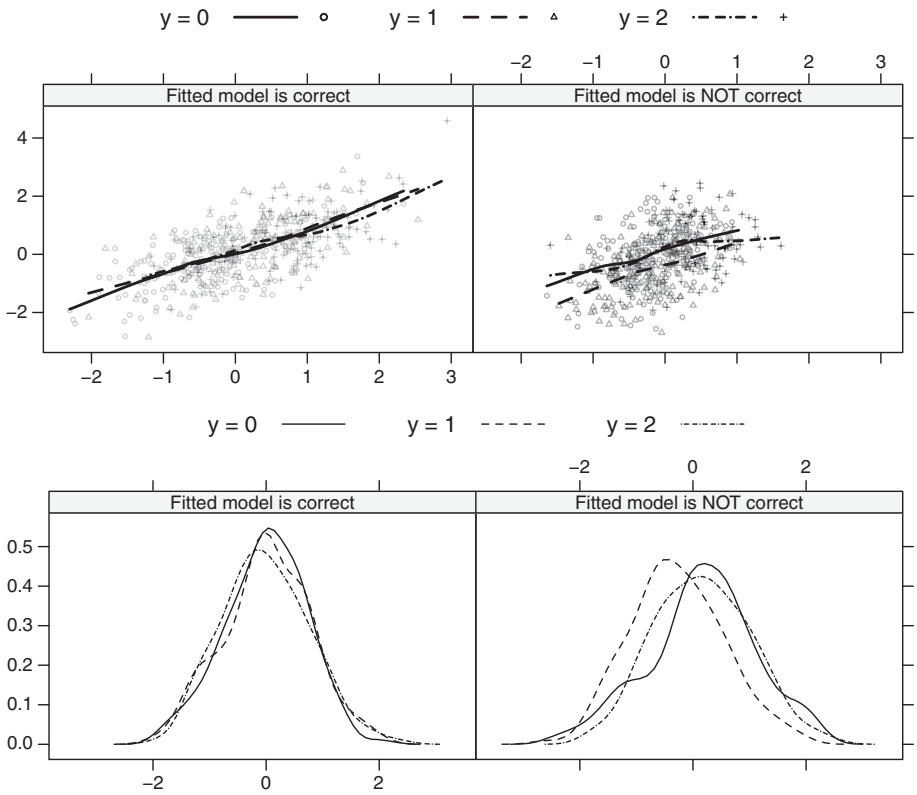


Figure 2. The *Scatterplot Diagnostic* (top row) and *Residual Density Diagnostic* (bottom row) for two simulated datasets, one for each column, of size $n = 500$ generated from (2.1.2) and (2.1.3), respectively. In both columns, the fitted model is (2.1.2). The true parameter values are as follows: $\beta_2 = -0.5$, $\beta_3 = -0.5$, $\beta_3 = -0.5$, and α_0 and α_1 are such that $\Pr(y = 0) = \Pr(y = 1)$. In the first column, $\beta_1 = 0.5$, and the fitted model matches the generating model, so the curves corresponding to each group are approximately collinear. In the second column, $\beta_{10} = 0.5$, $\beta_{11} = 1.0$, and the fitted model does not match the generating model, so $\hat{\boldsymbol{\beta}}^\top \boldsymbol{x}$ is *not* the balancing score.

The second row of Figure 1 presents these plots for the binary outcome model. When $\beta_4 = 0.5$ or $\beta_5 = 0.5$, one of the groups has a taller mode, and the other group has heavier tails. The second row of Figure 2 gives the same diagnostic for the case of ordinal outcomes, where we see a location shift in the *KDE* between groups.

These diagnostics are heuristic. Data-specific idiosyncracies, such as sample size or the bandwidth of the curves, will affect whether the plots indicate model misspecification. Each column in Figures 1 and 2 represents a single dataset, so specific qualities of these plots are not representative. Additionally, interpretation of results is up to the analyst, and these plots should be used in conjunction with other model-checking techniques, in particular the numerical tests we propose next. Rather than linear regression, a more flexible non-linear model, for example, a generalized additive model [23], may be used to create these residual plots, as long as the resulting estimated propensity score is conditioned upon. Such less-parametric approaches may offer statistically better transformations at a cost of interpretability. We restrict ourselves to normal linear regression to illustrate how the proposed graphical diagnostics may indicate a need for improvement of the proposed model and to preserve interpretability. The aforementioned simulated examples simply show that the graphical diagnostics exhibit differences between response groups when the fitted propensity score model differs from the true generating model. We further explain the graphical diagnostics in the real data examples, for which the true generating models are unknown.

2.2. *Numerical tests*

The concept of a balancing score also suggests numerical tests for detecting model misspecification. Let $\hat{\boldsymbol{\pi}} = \{\hat{\boldsymbol{\pi}}_1, \ldots, \hat{\boldsymbol{\pi}}_n\}$ be the vector of estimated propensity scores from an initial fitted model, for example,

one with only main effects for each component of $\boldsymbol{x}$. Let R_ℓ denote the interval $\left(\hat{\boldsymbol{\pi}}_{[20(\ell-1)]}, \hat{\boldsymbol{\pi}}_{[20\ell]}\right)$, $\ell = 1, \ldots, 5$, where $\hat{\boldsymbol{\pi}}_{[p]}$ is the pth sample percentile of $\hat{\boldsymbol{\pi}}$. Also define $\boldsymbol{Z}_i = \left[\boldsymbol{Z}_{i1}^\top, \boldsymbol{Z}_{i2}^\top, \boldsymbol{Z}_{i3}^\top\right]^\top$, where $\boldsymbol{Z}_{i1} = [1, I\{\hat{\boldsymbol{\pi}}_i \in R_2\}, I\{\hat{\boldsymbol{\pi}}_i \in R_3\}, I\{\hat{\boldsymbol{\pi}}_i \in R_4\}, I\{\hat{\boldsymbol{\pi}}_i \in R_5\}]^\top$, $\boldsymbol{Z}_{i2} = I\{y_i = 0\}$ and $\boldsymbol{Z}_{i3} = \boldsymbol{Z}_{i1} \times \boldsymbol{Z}_{i2}$, $i = 1, \ldots, n$. Rosenbaum and Rubin [24] fit a two-way ANOVA (two outcomes $\times$ five quintiles) on the jth covariate, $j = 1, \ldots, p$, to look for differences in covariate distributions across response categories, conditional on the propensity score quintile. The mean of the jth covariate in the ith observation from the ANOVA model is given by

$$\mathbb{E}[\boldsymbol{x}_{ij}|\boldsymbol{Z}_i] = \boldsymbol{\delta}_j^\top \boldsymbol{Z}_i, \tag{2.2.1}$$

where the coefficients $\boldsymbol{\delta}_j = \left[\boldsymbol{\delta}_{j1}^\top, \delta_{j2}^\top, \boldsymbol{\delta}_{j3}^\top\right]^\top$ correspond to the elements of $\boldsymbol{Z}_i$. Rosenbaum and Rubin [24] compute the F-statistics for the null hypotheses testing the main or interaction effects, respectively, $H_{0j} : \boldsymbol{\delta}_{j2} = 0$ or $H_{0j} : \boldsymbol{\delta}_{j3} = \boldsymbol{0}_4$, $j = 1, \ldots, p$, and inspect the five-number summary (minimum, three quartiles, maximum) of the size-p set of F-statistics. The propensity score model is then enriched, for example, with interaction and quadratic effects, thereby modifying $\boldsymbol{Z}_i$ and yielding smaller F-statistics. This process is repeated, each time plotting the new F-statistics until they are small enough, as determined by visual inspection. Our proposed diagnostic builds on this idea: For a given propensity score $\hat{\boldsymbol{\pi}}$, formally test the coefficients in (2.2.1). Because inference on $\boldsymbol{\delta}_{j2}$ alone in the presence of interaction effects is difficult to interpret, we propose to test either the interaction effects alone, $H_{0j} : \boldsymbol{\delta}_{j3} = \boldsymbol{0}_4$, or the main and interaction effects simultaneously, $H_{0j} : \{\boldsymbol{\delta}_{j2} = 0\} \cap \{\boldsymbol{\delta}_{j3} = \boldsymbol{0}_4\}$. There are p such null hypotheses, corresponding to each of p covariates. A simple testing strategy goes as follows. First, conduct a multiple response test of every null hypothesis as an overall diagnostic. That is, add the p log-likelihoods from individually regressing the jth covariate $\{\boldsymbol{x}_{ij}\}_i$ on $\{\boldsymbol{Z}_i\}_i$, and conduct a single composite test. If this test is rejected, then follow up with covariate-specific tests to find the likely source of misspecification. We summarize the two numerical tests of functional misspecification of the covariates as follows:

INT $H_0 : \boldsymbol{\delta}_{13} = \boldsymbol{\delta}_{23} = \cdots = \boldsymbol{\delta}_{p3} = \boldsymbol{0}_4$ (test interactions only)

JOINT $H_0 : \{\boldsymbol{\delta}_{12} = \boldsymbol{\delta}_{22} = \cdots = \boldsymbol{\delta}_{p2} = 0\} \cap \{\boldsymbol{\delta}_{13} = \boldsymbol{\delta}_{23} = \cdots = \boldsymbol{\delta}_{p3} = \boldsymbol{0}_4\}$ (test interactions and main effects).

Remark 2

We derive these parametric tests, which may be viewed as numerical analogs to the graphical Scatterplot Diagnostic, under the assumption of a normal linear model, which may be inadequate, even after an appropriate transformation of the covariate. An alternative would be a nonparametric testing approach using residuals, which is a numerical analog to the Residual Density Diagnostic. Define $\boldsymbol{r}_{ij} = \boldsymbol{x}_{ij} - \hat{\mathbb{E}}[\boldsymbol{x}_{ij}|\boldsymbol{Z}_{i1}]$, that is, the residuals from fitting a regression of the jth covariate to the categorical linear predictor. If balance is achieved, then the residual vector $\boldsymbol{r}_i = \{\boldsymbol{r}_{i1}, \boldsymbol{r}_{i2}, \ldots, \boldsymbol{r}_{ip}\}^\top$ should not depend on y_i. A nonparametric multiple response ANOVA would formally test for equality of the location, and the Kolmogorov–Smirnov or Kuiper [25] tests would test for equality of the full residual distribution. We found these to have little power to detect functional misspecification and do not consider them further. For completeness, we describe the competing approaches to our proposed methods.

HL [1] This is a standard goodness-of-fit test for models with binary outcomes. Given the vector of propensity scores, $\hat{\boldsymbol{\pi}}$, construct the quantile intervals R_ℓ as discussed in the beginning of this section. The number of quantiles need not be 5; we use $G = 10$ for all analyses, and any G satisfying $G > \dim\left(\hat{\boldsymbol{\beta}}\right) + 1$ may be used, where $\dim\left(\hat{\boldsymbol{\beta}}\right)$ is the number of parameters in the model. Define $o_{1\ell} = \sum_{i \in R_\ell} y_i$, $o_{0\ell} = \sum_{i \in R_\ell} 1 - y_i$, $e_{1\ell} = \sum_{i \in R_\ell} \hat{\boldsymbol{\pi}}_i$, and $e_{0\ell} = \sum_{i \in R_\ell} 1 - \hat{\boldsymbol{\pi}}_i$. Compare the test statistic $C_G = \sum_{k=0}^{1} \sum_{\ell=1}^{G} (o_{k\ell} - e_{k\ell})^2 / e_{k\ell}$ to a χ^2 distribution with degrees of freedom $G-2$ to find the appropriate p-value.

LIPSITZ [8] This extends HL for models with $K \geqslant 2$ ordinal outcomes. First, derive a score, $\hat{\mu}_i$, for each observation. Following the authors' suggestion, we use $\hat{\mu}_i = \sum_{k=0}^{K-1} k \hat{\Pr}(y_i = k)$. Again, similar to the aforementioned construction of the R_ℓ regions, find the G quantiles of the $\hat{\mu}_i$'s, and define indicators $I_{ig} = 1\{\hat{\mu}_i \in \text{quantile } g\}$, $g = 2, \ldots, G$. Finally, fit the model, logit $\Pr(y_i \leqslant k) = \alpha_k + \boldsymbol{\beta}^\top \boldsymbol{x}_i + \sum_{g=2}^{G} I_{ig} \gamma_g$, and test $H_0 : \gamma_2 = \cdots = \gamma_G = 0$ using a likelihood

Table I. *Logistic regression – functional misspecification of covariates.* Rejection rates from 10,000 simulated datasets of size 500 generated by (2.1.1).

Setting		Main effects			Rejection rates			
Quadratic, interaction effects	$\Pr(Y=1)$	β_1	β_2	β_3	INT	JOINT	LIPSITZ	HL
	0.50	0.50	0.00	0.00	0.12	0.12	0.05	0.05
	0.10	0.50	0.00	0.00	0.09	0.07	0.07	0.05
Type 1 error	0.50	0.50	0.50	−0.50	0.10	0.05	0.06	0.05
$\beta_4=0, \beta_5=0$	0.10	0.50	0.50	−0.50	0.09	0.05	0.08	0.05
	0.50	0.50	1.00	−0.50	0.10	0.06	0.06	0.05
	0.10	0.50	1.00	−0.50	0.09	0.09	0.06	0.05
	0.50	0.50	0.00	0.00	0.94	0.89	0.92	0.92
	0.10	0.50	0.00	0.00	0.23	0.14	0.25	0.18
Power	0.50	0.50	0.50	−0.50	0.64	0.53	0.13	0.12
$\beta_4=0.5, \beta_5=0$	0.10	0.50	0.50	−0.50	0.08	0.04	0.08	0.05
	0.50	0.50	1.00	−0.50	0.39	0.30	0.07	0.06
	0.10	0.50	1.00	−0.50	0.07	0.08	0.06	0.04
	0.50	0.50	0.00	0.00	1.00	0.99	0.95	0.95
	0.10	0.50	0.00	0.00	0.39	0.27	0.33	0.27
Power	0.50	0.50	0.50	−0.50	0.81	0.75	0.20	0.18
$\beta_4=1, \beta_5=0$	0.10	0.50	0.50	−0.50	0.08	0.04	0.08	0.05
	0.50	0.50	1.00	−0.50	0.57	0.49	0.11	0.09
	0.10	0.50	1.00	−0.50	0.06	0.06	0.06	0.04
	0.50	0.50	0.50	−0.50	0.34	0.22	0.15	0.14
Power	0.10	0.50	0.50	−0.50	0.27	0.18	0.09	0.05
$\beta_4=0, \beta_5=0.5$	0.50	0.50	1.00	−0.50	0.35	0.26	0.13	0.11
	0.10	0.50	1.00	−0.50	0.36	0.31	0.08	0.05
	0.50	0.50	0.50	−0.50	0.83	0.72	0.38	0.36
Power	0.10	0.50	0.50	−0.50	0.64	0.53	0.11	0.08
$\beta_4=0, \beta_5=1$	0.50	0.50	1.00	−0.50	0.85	0.77	0.42	0.38
	0.10	0.50	1.00	−0.50	0.84	0.80	0.17	0.17

ratio test statistic. Under H_0, the test statistic will asymptotically follow a χ^2 distribution with $G-1$ degrees of freedom. As with HL, we use $G=10$ for all analyses in this paper. Related to LIPSITZ is the Score test [26], which tests whether the parameters corresponding to different cumulative logits are the same. This is provided by `PROC LOGISTIC` in SAS but is liberal in its rejection rates [27].

We evaluated the operating characteristics of these tests via a small simulation study, described in three parts as follows.

2.3. *Simulation design*

2.3.1. Logistic regression – functional misspecification of covariates. For binary outcomes, we generated datasets of size 500 from (2.1.1). Each covariate was independently drawn from a standard normal distribution. We chose values of $\boldsymbol{\beta}$ to satisfy either null, that is, $\beta_4=\beta_5=0$, or non-null scenarios. The intercept α was such that $\Pr(y=1)=0.5$ or $\Pr(y=1)=0.1$, marginalized over the covariates. For each α and $\boldsymbol{\beta}$, empirical rejection rates from 10,000 datasets were recorded to simulate the type I error or power of each test. We give the results in Table I. JOINT has slightly inflated type I error (0.05 to 0.12), and INT more so (0.09–0.12). LIPSITZ, while designed for ordinal outcomes, may be applied to logistic regression as a special case; the LIPSITZ test statistic is approximately equal in distribution to the HL test statistic, although HL appears to be the only test with exactly nominal type I error. JOINT has favorable power properties compared to LIPSITZ and HL, despite only slightly inflated type I error.

Table II. *Ordinal regression – functional misspecification of covariates.* Rejection rates from 10,000 simulated datasets of size 500 generated by (2.1.2).

Setting			Main effects			Rejection rates		
Quadratic, interaction effects	$\Pr(Y=1)$	$\Pr(Y=2)$	β_1	β_2	β_3	INT	JOINT	LIPSITZ
	0.33	0.33	0.50	0.00	0.00	0.12	0.13	0.06
	0.30	0.60	0.50	0.00	0.00	0.11	0.12	0.05
Type 1 error	0.33	0.33	0.50	0.50	−0.50	0.10	0.07	0.05
$\beta_4=0, \beta_5=0$	0.30	0.60	0.50	0.50	−0.50	0.10	0.07	0.06
	0.33	0.33	0.50	1.00	−0.50	0.13	0.11	0.06
	0.30	0.60	0.50	1.00	−0.50	0.10	0.11	0.06
	0.33	0.33	0.50	0.00	0.00	0.96	0.93	0.98
	0.30	0.60	0.50	0.00	0.00	0.98	0.99	0.99
Power	0.33	0.33	0.50	0.50	−0.50	0.68	0.65	0.17
$\beta_4=0.5, \beta_5=0$	0.30	0.60	0.50	0.50	−0.50	0.89	0.90	0.31
	0.33	0.33	0.50	1.00	−0.50	0.42	0.42	0.08
	0.30	0.60	0.50	1.00	−0.50	0.71	0.75	0.11
	0.33	0.33	0.50	0.00	0.00	1.00	1.00	0.98
	0.30	0.60	0.50	0.00	0.00	1.00	1.00	0.99
Power	0.33	0.33	0.50	0.50	−0.50	0.89	0.86	0.29
$\beta_4=1, \beta_5=0$	0.30	0.60	0.50	0.50	−0.50	0.99	0.98	0.63
	0.34	0.33	0.50	1.00	−0.50	0.68	0.66	0.11
	0.30	0.60	0.50	1.00	−0.50	0.95	0.96	0.28
	0.33	0.33	0.50	0.50	−0.50	0.33	0.28	0.21
Power	0.30	0.60	0.50	0.50	−0.50	0.25	0.23	0.24
$\beta_4=0, \beta_5=0.5$	0.33	0.33	0.50	1.00	−0.50	0.36	0.35	0.18
	0.30	0.60	0.50	1.00	−0.50	0.19	0.24	0.21
	0.33	0.33	0.50	0.50	−0.50	0.82	0.78	0.54
Power	0.30	0.60	0.50	0.50	−0.50	0.75	0.77	0.74
$\beta_4=0, \beta_5=1$	0.33	0.33	0.50	1.00	−0.50	0.85	0.85	0.58
	0.30	0.60	0.50	1.00	−0.50	0.56	0.67	0.77

2.3.2. Ordinal regression – functional misspecification of covariates. Here, we used the generating model given by

$$\text{logit } \Pr(y \leqslant k|\boldsymbol{x}) = \alpha_k + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_1 x_2, \quad k = 0, 1, \tag{2.3.1}$$

which is a more general version of (2.1.2). We used the same choices of $\boldsymbol{\beta}$ as in the binary case and chose α_0 and α_1 to achieve $\Pr(y=0) = \Pr(y=1) = 1/3$ or $\Pr(y=0) = 0.1$ and $\Pr(y=1) = 0.3$. We give the results in Table II. All methods have some type I error inflation, INT having the most (0.10–0.13), followed by JOINT (0.07–0.13), and finally LIPSITZ. Although the type I error inflation makes a direct power comparison difficult, in some cases; for example, when $\beta_4 = 0.5$ and $\beta_5 = 0$, JOINT has a considerable power advantage over LIPSITZ.

2.3.3. Ordinal regression – violation of the proportional odds assumption. The final simulation we conducted looked at the tests' behavior in the presence of a violation to the proportional odds assumption of an ordinal regression. To achieve this, we generated the data from an underlying model given by (2.1.3). We give the results in Table III. To be clear, the proportional odds assumption is violated regardless of whether $\beta_{10} = \beta_{11}$, and we refer back to Table II for type I error properties. We chose α_0 and α_1 so that $\Pr(y=0) = \Pr(y=1) = 1/3$ or $\Pr(y=0) = 0.1$ and $\Pr(y=1) = 0.3$. We also emphasize that this is a scenario for which our proposed numerical tests are ill suited, namely a structural violation to the underlying probability of falling within each category rather than a violation to the functional form of the covariates, and a more general goodness-of-fit approach, like LIPSITZ, may be better suited here. In fact, none of the methods has acceptable power properties

Table III. *Ordinal regression – violation of proportional odds assumption.* Rejection rates from 10,000 simulated datasets of size 500 generated by (2.1.3).

			Main effects				Rejection rates		
Setting	$\Pr(Y=1)$	$\Pr(Y=2)$	β_{10}	β_{11}	β_2	β_3	INT	JOINT	LIPSITZ
	0.33	0.33	0.50	0.50	0.00	0.00	0.10	0.12	0.06
	0.30	0.60	0.50	0.50	0.00	0.00	0.10	0.11	0.06
Power	0.33	0.33	0.50	0.50	−0.50	−0.50	0.10	0.09	0.16
	0.30	0.60	0.50	0.50	−0.50	−0.50	0.10	0.07	0.07
	0.33	0.34	0.50	0.50	−1.50	−0.50	0.12	0.16	0.82
	0.30	0.60	0.50	0.50	−1.50	−0.50	0.10	0.10	0.37
	0.33	0.33	0.50	0.75	0.00	0.00	0.11	0.26	0.11
	0.30	0.60	0.50	0.75	0.00	0.00	0.13	0.17	0.06
Power	0.33	0.33	0.50	0.75	−0.50	−0.50	0.10	0.28	0.25
	0.30	0.60	0.50	0.75	−0.50	−0.50	0.10	0.13	0.10
	0.33	0.33	0.50	0.75	−1.50	−0.50	0.13	0.38	0.86
	0.30	0.60	0.50	0.75	−1.50	−0.50	0.09	0.15	0.47
	0.33	0.33	0.50	1.00	0.00	0.00	0.13	0.50	0.17
	0.30	0.60	0.50	1.00	0.00	0.00	0.15	0.28	0.08
Power	0.33	0.33	0.50	1.00	−0.50	−0.50	0.10	0.77	0.32
	0.30	0.60	0.50	1.00	−0.50	−0.50	0.10	0.29	0.14
	0.33	0.34	0.50	1.00	−1.50	−0.50	0.13	0.87	0.87
	0.30	0.60	0.50	1.00	−1.50	−0.50	0.09	0.39	0.56

All rejection rates in this table correspond to power because the data are generated from a multinomial logistic model, which automatically violates the proportional odds assumption.

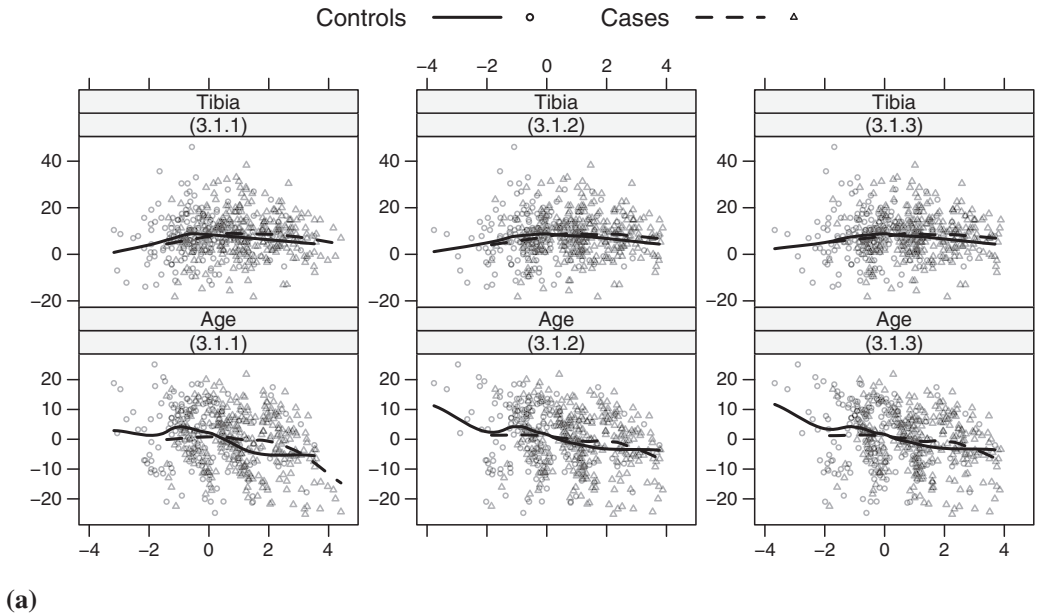
when $\beta_{10} = \beta_{11}$. INT has small power in all cases, but JOINT has more power than LIPSITZ when $\beta_{11} - \beta_{10} = 0.5$

3. Data analysis

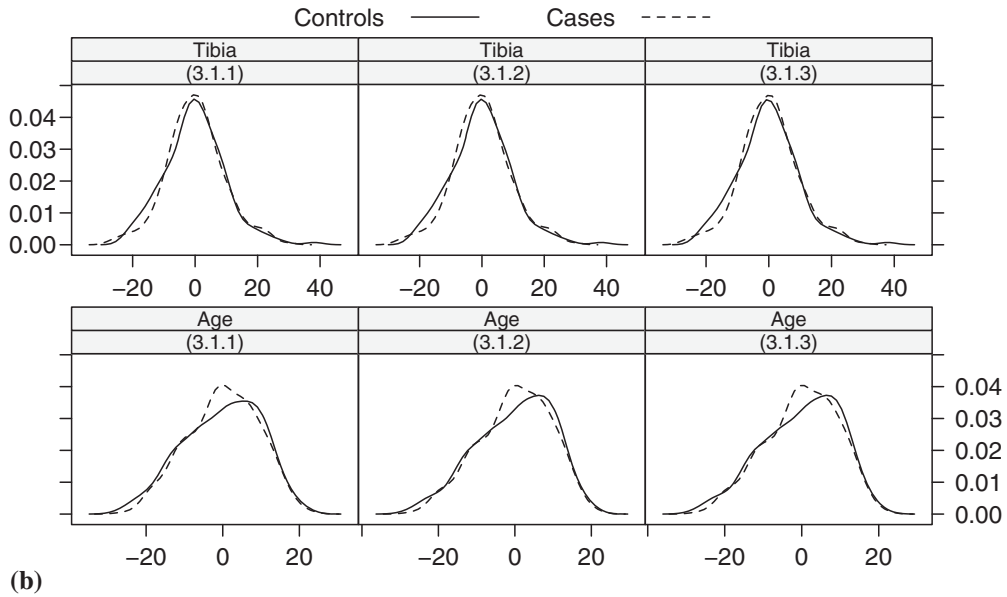
3.1. A case-control study of Parkinson's disease and cumulative lead exposure

We analyze data from a case-control study in the Boston, Massachusetts, area conducted from 2003–2007 that explores association between lifetime cumulative lead exposure and PD [28]. PD is a complex disease for which environmental determinants have been hypothesized to be particularly important. The primary hypothesis of the study is that the risk of PD increases with increased lifetime exposure to heavy metals, including lead. As a biomarker for cumulative lead exposure, the study used measurements of lead levels in the tibia bone assessed with a K-shell X-ray fluorescence (KXRF) technique. Cases were recruited from four movement disorder clinics in the Boston area. Controls included family, in-laws, and friends of cases, as well as responders to community advertisements and eligible subjects in the Harvard Cooperative Program on Aging, a registry of elderly volunteer research subjects. In the original study, controls and self-reported cases from the Normative Aging Study were also included; here, we do not include those subjects. After removing individuals with missing covariates or KXRF measurements that had large uncertainty, 300 cases and 194 controls were included in the present analysis. We illustrate how the graphical and numerical diagnostic tests proposed in previous sections incrementally change between models we consider, beginning with a logistic model that uses continuous values of tibia lead measurements and ending with the actual model presented in the paper by Weisskopf *et al.* [28], which uses categorized tibia lead measurements.

Apart from normalized continuous tibia lead measurements (in μg/g), denoted as 'Tibia' in the model, the other covariates used were 'Age' (centered, in years), smoking status (in pack-years of smoking, 'Pack-Yr'), education ('Educ'; high school graduate or less, some college or college degree, or post-graduate degree), 'Race' (white/non-white), 'Sex', and ascertainment location ('Loc'; BU = Boston University Medical Center, BW = Brigham and Women's Hospital, BI = Beth Israel Deaconess Medical Center, or HV = Harvard Vanguard Medical Associates). Following the original paper, separate



(a)



(b)

Figure 3. (a) The *Scatterplot Diagnostic* for two models applied to the Parkinson's disease case-control study. Covariate values are on the y-axes, the linear predictor $\hat{\boldsymbol{\beta}}^\top \boldsymbol{x}$ is on each x-axis, and the data are grouped by case status. (b) The *Residual Density Diagnostic* for two models applied to the Parkinson's disease case-control study. Each panel is a *kernel density estimate* of the residuals from regressing the continuous covariates on the linear predictor $\hat{\boldsymbol{\beta}}^\top \boldsymbol{x}$, grouped by case status.

fixed effects for each of the four clinics account for potential heterogeneity between locations, and the controls ascertained from the general community are treated as coming from BU. A logistic model with linear terms is

$$\begin{aligned}
&\text{logit Pr(PD} = 1|\text{covariates)} \\
&\quad = \alpha + \beta_{\text{TIB}}\text{Tibia} + \beta_{\text{AGE}}\text{Age} + \beta_{\text{PKYR}}\text{Pack-Yr} + \beta_{\text{ED1}}\{\text{Educ} = \text{College}\} + \beta_{\text{ED2}}\{\text{Educ} = \text{Postgrad.}\} \\
&\qquad + \beta_{\text{RAC}}\{\text{Race} = \text{White}\} + \beta_{\text{SEX}}\{\text{Sex} = \text{Female}\} + \beta_{\text{LOC.BU}}\{\text{Loc} = \text{BU}\} \\
&\qquad + \beta_{\text{LOC.BW}}\{\text{Loc} = \text{BW}\} + \beta_{\text{LOC.HV}}\{\text{Loc} = \text{HV}\} \\
&\quad \equiv \alpha + \beta_{\text{TIB}}\text{Tibia} + \beta_{\text{AGE}}\text{Age} + \boldsymbol{\gamma}^\top \boldsymbol{W}. \qquad (3.1.1)
\end{aligned}$$

Table IV. Regression coefficients with tests of significance (top) and numerical tests of functional misspecification (bottom) for three models fit to the Parkinson's disease case-control data.

	Parameter estimate (p-value)					
Model	$\beta_{\text{TIB}}(\mu g/g)$	$\beta_{\text{TIB.Q2}}$	$\beta_{\text{TIB.Q3}}$	$\beta_{\text{TIB.Q4}}$	β_{AGE}(yr)	β_{AGESQ}(yr^2)
(3.1.1)	0.024 (0.049)				-0.049 ($< 10^{-4}$)	
(3.1.2)	0.025 (0.047)				-0.051 ($< 10^{-4}$)	-0.002 (0.045)
(3.1.3)		0.179 (0.576)	0.299 (0.343)	0.529 (0.107)	-0.049 ($< 10^{-4}$)	-0.002 (0.060)

	Test statistic (degrees of freedom, p-value) for testing functional misspecification			
Model	HL	LIPSITZ	INT	JOINT
(3.1.1)	5.9 (8, 0.659)	9.1 (9, 0.431)	19.4 (8, 0.013)	19.7 (10, 0.032)
(3.1.2)	4.5 (8, 0.808)	7.2 (9, 0.616)	5.4 (8, 0.716)	5.5 (10, 0.853)
(3.1.3)	7.7 (8, 0.464)	7.9 (9, 0.543)	4.8 (8, 0.781)	5.1 (10, 0.885)

The coefficients $\beta_{\text{TIB.Q2}}$, $\beta_{\text{TIB.Q3}}$, and $\beta_{\text{TIB.Q4}}$ correspond to the effects of 2nd, 3rd, and 4th quartiles of the categorized measurements of 'Tibia' relative to the 1st quartile.

We focus our analysis on the continuous covariates 'Tibia' and 'Age', and the remaining covariates are hereafter collectively called $\boldsymbol{W}$. We give the graphical diagnostics corresponding to the models we consider in Figure 3. Table IV lists estimates of β_{TIB} and β_{AGE} with p-values corresponding to the Wald test statistic. In the initial model as described previously, $\hat{\beta}_{\text{TIB}} = 0.024$ ($p = 0.049$), and $\hat{\beta}_{\text{AGE}} = -0.049$ ($p < 10^{-4}$). Thus, controlling for age and $\boldsymbol{W}$, increased lead exposure seems to increase the odds of PD.

Also in Table IV are the numerical diagnostic tests for this initial model with linear terms of age and tibia. HL and LIPSITZ have non-significant p-values, respectively, $p = 0.659$ and $p = 0.431$, while the INT and JOINT statistics are significant, with $p = 0.013$ and $p = 0.032$. Covariate specific tests from INT and JOINT, which are not given in the table, suggest a lack of fit for 'Age'. This is somewhat supported by the Scatterplot Diagnostic, given in the first column of Figure 3(a), in which younger cases tend to have larger values of the linear predictor, $\hat{\boldsymbol{\beta}}^\top \boldsymbol{x}$, causing the dashed line to drop at the right of the panel. To a lesser degree, the Residual Density Diagnostic (first column, Figure 3(b)) also suggests misspecification of 'Age'. Because $\hat{\beta}_{\text{AGE}} = -0.049$ implausibly suggests a monotone decrease in risk of PD with age, we add a quadratic term, yielding the following model:

$$\text{logit Pr(PD} = 1|\text{covariates)} = \alpha + \beta_{\text{TIB}}\text{Tibia} + \beta_{\text{AGE}}\text{Age} + \beta_{\text{AGESQ}}\text{Age}^2 + \boldsymbol{\gamma}^\top \boldsymbol{W}. \qquad (3.1.2)$$

From Table IV, the effect of 'Tibia' is about the same, as is the first-order term for 'Age'. However, the quadratic term is negative, -0.002 ($p = 0.045$). Thus, fixing all other covariates, the risk of PD is estimated to increase until about age 55. The p-values for HL, LIPSITZ, INT, and JOINT are, respectively, 0.808, 0.616, 0.716, and 0.853, and the results from INT and JOINT suggest improved overall fit of the covariates. From the second column of Figure 3(a), younger cases no longer have disproportionately large values of $\hat{\boldsymbol{\beta}}^\top \boldsymbol{x}$, although there may be some remaining lack of fit suggested by several older controls having small values of $\hat{\beta}^\top \boldsymbol{x}$. The Residual Density Diagnostic (second column, Figure 3(b)) is insensitive to the added quadratic term and shows little change from model (3.1.1).

The actual model presented in [28] differs from (3.1.2) in representing the effect of 'Tibia'; we now modify (3.1.2) to match with the model for 'Tibia' as presented in [28]. Specifically, we replace the continuous measurements with a four-level categorical variable, where the levels are determined by the empirical quartiles. The model is

$$\begin{aligned}\text{logit Pr(PD} = 1|\text{covariates)} =& \alpha + \beta_{\text{TIB.Q2}}\text{TibiaQ2} + \beta_{\text{TIB.Q3}}\text{TibiaQ3} \\ &+ \beta_{\text{TIB.Q4}}\text{TibiaQ4} + \beta_{\text{AGE}}\text{Age} + \beta_{\text{AGESQ}}\text{Age}^2 + \boldsymbol{\gamma}^\top \boldsymbol{W}. \qquad (3.1.3)\end{aligned}$$

From the corresponding rows in Table IV, this modification yields results very similar to those of (3.1.2), suggesting that the categorization is not needed from a model-fit perspective. All numerical diagnostic tests are still non-significant (Table IV), but the differences in p-values between models should not be used for model selection purposes. The graphical diagnostics in the third columns of Figure 3(a and b) are

minimally changed from the second column. This is consistent with INT and JOINT previously pointing to 'Age', rather than 'Tibia', as the source of covariate misspecification.

Remark 3

An alternative to the aforementioned numerical tests would be to look at numerical summaries that are often used to check covariate balance via descriptive statistics. One can compute standardized bias estimates within the five propensity score quintiles. For each covariate and each quintile, calculate the difference in mean values for that covariate between cases and controls falling in that stratum, and divide by the estimated standard deviation of the mean difference, exactly as in a studentized two-sample pooled t-statistic. The main advantage of this approach over the aforementioned testing procedures is that it is less dependent on sample size. But it is difficult to integrate into a single test or numeric summary. Table S1 in the Supporting Information[‡] gives standardized biases for 'Tibia' and 'Age'. In agreement with our analysis, there is no clear change between models for 'Tibia', but there is a significant drop in bias for 'Age' between models (3.1.1) and (3.1.2).

3.2. *Normative Aging Study*

The Normative Aging Study (NAS) is a multidisciplinary longitudinal study of aging in men established by the Veteran's Administration in 1963. NAS subjects have reported for medical examination every 3–5 years. Although the study records data on a wide spectrum of variables, including several health-related measures, dietary and behavioral exposures, exposure to certain metals in their environment, and psychosocial events, our analysis focuses on exploring the relationship of fasting blood glucose (FBG) with two markers of systemic inflammation, namely white blood cell count (WBC, $10^3/\text{mm}^3$) and blood levels of C-reactive protein (CRP, mg/L), after controlling for age (y). The measurements were taken between January 2000 and December 2004; in cases where multiple measurements were available on the same subject, we consider only the last complete observation available. The data contain observations on 682 men in the age range of 48–93 years. FBG was categorized into three categories according to established diagnostic criteria for diabetes [29], with values less than 110 mg/dl defined as normal (FBG = 1), those between 110 and 126 mg/dl, inclusive, defined as impaired fasting glucose (FBG = 2), and those exceeding 126 mg/dl defined as diabetes (FBG = 3). It has been suggested that oxidative stress-induced inflammatory response increases insulin resistance, resulting in hyperglycemia or elevated levels of FBG, which in turn causes further oxidative stress [30]. Inflammation is known to be a risk factor for diabetes [31]. WBC count and CRP levels may be viewed as biomarkers of systemic inflammation and thus could potentially be associated with FBG levels, leading to this analysis. We posited the model

$$\text{logit Pr}(\text{FBG} \leqslant k|\text{covariates}) = \alpha_k + \beta_{\text{AGE}}\text{Age} + \beta_{\ell_{\text{WBC}}}\log(\text{WBC}) + \beta_{\ell_{\text{CRP}}}\log(\text{CRP}), \quad k = 0, 1. \tag{3.2.1}$$

Note that both WBC and CRP are log-transformed. Fitting the model, we have $\hat{\beta}_{\ell_{\text{WBC}}} = -0.615$ ($p = 0.024$) and $\hat{\beta}_{\ell_{\text{CRP}}} = -0.081$ ($p = 0.195$). Thus, WBC and CRP are associated with increased levels of FBG, and $\hat{\beta}_{\ell_{\text{WBC}}}$ is significant using size 0.05. Age was not found to be significantly associated with FBG levels. LIPSITZ, INT, and JOINT had p-values, respectively, of 0.294, 0.727, and 0.906. Based on recommendations by Agresti [27], we also evaluated the proportional odds assumption by collapsing the FBG categories and inspecting the coefficients from the logistic models corresponding to the two possible collapsings. The signs and effect sizes of the coefficient estimates were consistent, which would be expected under the proportional odds assumption.

The graphical diagnostics are in Figure 4. From the Scatterplot Diagnostic, there does appear to be a lack of collinearity between *LOWESS* curves, particularly in log(CRP); however, this is driven by a few influential observations with extreme values of the linear predictor, and we did not view this as indicative of systemic misspecification. Similarly, the Residual Density Diagnostic indicates agreement between the distributions of residuals. Several outliers in the data may warrant further investigation. Based on these and our interpretation of the graphical diagnostics, we did not fit additional models.

These case studies highlight the utility of our graphical and numerical diagnostics as a supplement to, rather than replacement of, existing strategies for evaluating model fit and the subjectivity of their application. These tests provide an initial assessment of model misspecification. In these analyses, the

[‡]*Supporting information may be found in the online version of this article.*

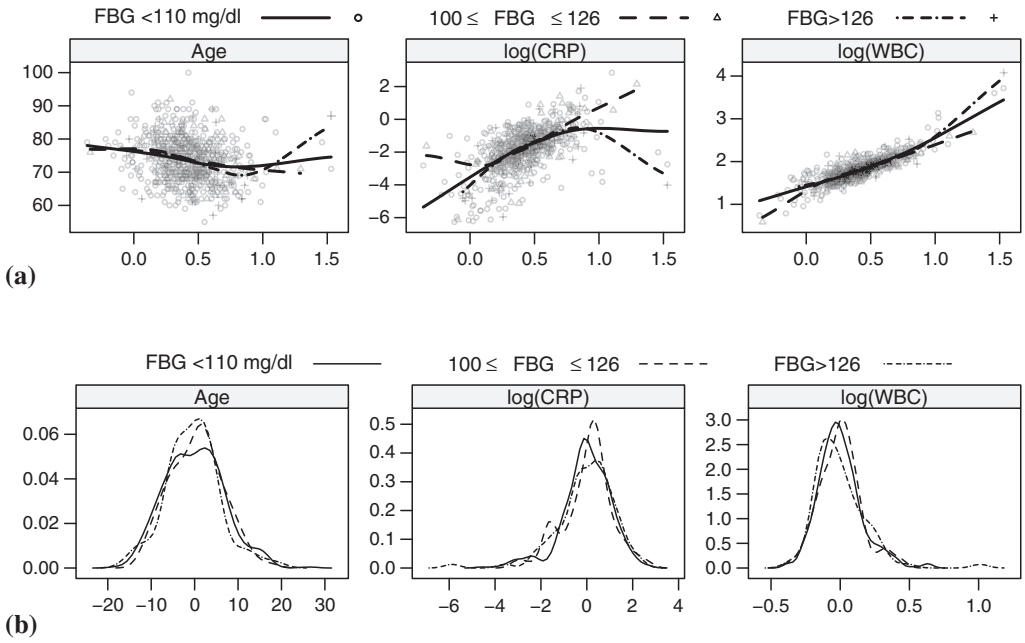


Figure 4. (a) The *Scatterplot Diagnostic* applied to the Normative Aging Study (NAS) cohort. Covariate values are on the y-axes, the linear predictor $\hat{\boldsymbol{\beta}}^\top \boldsymbol{x}$ is on each x-axis, and the data are grouped by fasting blood glucose (FBG) categories. The lack of collinearity of the *LOWESS* curves on the edges of the plots is driven by a few influential observations. (b) The *Residual Density Diagnostic* applied to the NAS cohort. Each panel is a *kernel density estimate* of the residuals from regressing the continuous covariates on the linear predictor $\hat{\boldsymbol{\beta}}^\top \boldsymbol{x}$, grouped by FBG categories.

goal is not to obtain large p-values but rather p-values that are not small, and this should additionally be supported by the graphical diagnostics, as INT and JOINT do not always maintain nominal type I error. In the PD case-control study, the p-values for INT and JOINT improved after introducing a quadratic term for 'Age', as did the collinearity of the LOWESS curves in the Scatterplot Diagnostic. In contrast, the p-values from HL and LIPSITZ never suggested model misspecification. The added quadratic term was motivated by scientific rationale rather than a particular aspect of the diagnostic tests; in other words, our diagnostics are useful for detecting general model misspecification but, when used in isolation, not necessarily remedies thereof. In the NAS example, the first model we fit was considered adequate. However, the diagnostics could have been used to compare (3.2.1) to a model with un-transformed CRP and WBC levels. For brevity, we do not include this comparison.

4. Discussion

Simple graphical tools for categorical response regression models are lacking. Following the suggestion of Rubin [16], in this paper, we use the balancing score to develop visual model diagnostics for categorical data models. The visual summaries are informative about a global model misspecification, not just covariate misspecification. Thus, this breadth is both a strength and weakness because it also means the visual summaries cannot point to the specific nature of the misspecification, for example, a functional or link misspecification. As a by-product, we also examine certain tests that have been used to check covariate balance in treatment groups for the purpose of identifying the source of model misspecification. The tests and diagnostics developed in the paper may serve as simple tools to discern misspecification in ordinal models. Further research is warranted to provide insight into the form of misspecification and derive targeted solutions to alleviate the poor fit.

Acknowledgements

The NSF grant DMS 1007494 and NIH grants CA156608 and ES020811 partially supported the research of Philip S. Boonstra and Bhramar Mukherjee. The authors thank the Robert E. Feldman Memorial Genetics and Environmental Metals Parkinson's Disease Study, supported by NIH grant ES 0010798, for sharing the

Parkinson's disease data. The Cooperative Studies Program/Epidemiology Research and Information Center of the US Department of Veterans Affairs supports the VA NAS, and it is a component of the Massachusetts Veterans Epidemiology Research and Information Center, Boston, MA. The first two authors contributed equally to this work. All analyses were conducted using R [32]. Code for the diagnostics is available at http://www-personal.umich.edu/~philb.

References

1. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics, Series A* 1980; **9**:1043–1069.
2. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*, 3rd edn. John Wiley & Sons, Inc.: Hoboken, NJ, 2013.
3. Tsiatis AA. A note on a goodness-of-fit test for the logistic regression model. *Biometrika* 1980; **67**:250–251.
4. Stukel TA. Generalized logistic models. *Journal of the American Statistical Association* 1988; **83**:426–431.
5. le Cessie S, van Houwelingen HC. Testing the fit of a regression model via score tests in random effects models. *Biometrics* 1995; **51**:600–614.
6. Royston P. The use of cusums and other techniques in modelling continuous covariates in logistic regression. *Statistics in Medicine* 1992; **11**:1115–1129.
7. Hosmer DW, Hosmer T, le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine* 1997; **16**:965–980.
8. Lipsitz SR, Fitzmaurice GM, Molenberghs G. Goodness-of-fit tests for ordinal response regression models. *Journal of the Royal Statistical Society, Series C* 1996; **45**:175–190.
9. Toledano AY, Gatsonis C. Ordinal regression methodology for ROC curves derived from correlated data. *Statistics in Medicine* 1996; **15**:1807–1826.
10. Kim JH. Assessing practical significance of the proportional odds assumption. *Statistics & Probability Letters* 2003; **65**:233–239.
11. Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* 1993; **80**:557–572.
12. Lin DY, Wei LJ, Ying Z. Model-checking techniques based on cumulative residuals. *Biometrics* 2002; **58**:1–12.
13. Arbogast PG, Lin DY. Model-checking techniques for stratified case-control studies. *Statistics in Medicine* 2005; **24**:229–247.
14. Liu I, Mukherjee B, Suesse T, Sparrow D, Park SK. Graphical diagnostics to check model misspecification for the proportional odds regression model. *Statistics in Medicine* 2009; **28**:412–429.
15. Landwehr JM, Pregibon D, Shoemaker AC. Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association* 1984; **79**:61–71.
16. Rubin DB. Graphical methods for assessing logistic regression models: comment. *Journal of the American Statistical Association* 1984; **79**:79–80.
17. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
18. Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. *American Journal of Epidemiology* 1999; **150**:327–333.
19. Imbens G. The role of the propensity score in estimating dose-response functions. *Biometrika* 2000; **87**:706–710.
20. Lu B, Zanutto E, Hornik R, Rosenbaum PR. Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association* 2001; **96**:1245–1253.
21. Imai K, van Dyk D. Causal inference with general treatment regimes: generalizing the propensity score. *Journal of the American Statistical Association* 2004; **99**:854–866.
22. Månsson R, Joffe MM, Sun W, Hennessy S. On the estimation and use of propensity scores in case-control and case-cohort studies. *American Journal of Epidemiology* 2007; **166**:332–339.
23. Hastie T, Tibshirani R. *Generalized Additive Models*, Vol. 43. Chapman & Hall/CRC: Boca Raton, FL, 1990.
24. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; **79**:516–524.
25. Kuiper N. Tests concerning random points on a circle. *Koninklijke Nederlandse Akademie van Wetenschappen, Series A* 1962; **63**:38–47.
26. Peterson B, Harrell FE, Jr. Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society, Series C* 1990; **39**:205–217.
27. Agresti A. *Analysis of Ordinal Categorical Data*. John Wiley & Sons, Inc.: Hoboken, NJ, 2010.
28. Weisskopf MG, Weuve J, Nie H, Saint-Hilaire MH, Sudarsky L, Simon DK, Hersh B, Schwartz J, Wright RO, Hu H. Association of cumulative lead exposure with Parkinson's disease. *Environmental Health Perspectives* 2010; **118**:1609–1613.
29. Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Report of the expert committee on the diagnosis and classification of diabetes mellitus. *Diabetes Care* 2002; **25**:s5–s20.
30. Pliquett R, Fasshauer M, Bluher M, Paschke R. Neurohumoral stimulation in type-2-diabetes as an emerging disease concept. *Cardiovascular Diabetology* 2004; **3**:1–8.
31. Nakanishi S, Yamane K, Kamei N, Okubo M, Kohno N. Elevated C-reactive protein is a risk factor for the development of type 2 diabetes in Japanese Americans. *Diabetes Care* 2003; **26**:2754–2757.
32. R Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013. http://www.R-project.org/.