RESEARCH SUPPORT
UNIVERSITY OF MICHIGAN BUSINESS SCHOOL

APRIL 1997

ON SCALE MIXTURES OF UNIFORM DISTRIBUTIONS AND THE LATENT WEIGHTED LEAST SQUARES METHOD

WORKING PAPER #9712-09

BY
STEPHEN WALKER
IMPERIAL COLLEGE, LONDON
PAUL DAMIEN
UNIVERSITY OF MICHIGAN
AND
MARY MEYER
UNIVERSITY OF MICHIGAN

On Scale Mixtures of Uniform Distributions and the Latent Weighted Least Squares Method

Stephen Walker[1]
Paul Damien[2]
and
Mary Meyer[3]

[1] Department of Mathematics, Imperial College, 180 Queen's Gate, London SW7 2BZ.

[2] Business School, University of Michigan, Ann Arbor, 48109-1234, USA.

[3] Department of Statistics, University of Michigan, Ann Arbor, MI, 48109.

Summary

In this paper we introduce a new estimator for the coefficients of a linear regression model. The estimator is based on the characterisation of a normal distribution as a scale mixture of a uniform, the mixing distribution being a particular gamma distribution. The Student t and exponential power distributions are also characterised as scale mixtures of uniforms; these, in turn, are seen to be a special case of a new (and general) family of distributions. It is shown that this new family of distributions coincides with the class of unimodal, symmetric distributions.

Key words: EM algorithm, Latent variables, Weighted least squares, Robustness.

1 Introduction

This paper considers the familiar linear regression model

$$y_i = \sum_{j=1}^{k} x_{ij}\beta_j + e_i, \quad i = 1, \cdots, n, \tag{1}$$

where the y_i are observable dependent variables, the x_{ij} are known covariates, the β_j unknown regression coeffecients and the parameter of interest, and the e_i are independent and identically distributed (iid) error terms. We make the following assumptions about these errors:

1) $Ee_i = 0$ and the e_i are symmetric about 0, and

2) $\text{var}(e_i) = \sigma^2$.

The most popular and widely used estimate for the regression parameter $\beta = (\beta_1, \cdots, \beta_k)$ is the ordinary least squares (OLS) estimator:

$$\hat{\beta}_{OLS} = [\sum_i X_i X_i']^{-1} \sum_i y_i X_i,$$

where X_i is the column vector with entries $(x_{i1}, \cdots, x_{ik})$. The OLS is the maximum likelihood (ML) estimator under the assumption that the e_i are normally distributed.

It is well known that the OLS estimator is likely to be unsatisfactory in a number of possible sampling scenarios. For example, if the errors have a heavier tailed distribution to that of the normal, or if there are outlier observations: in the latter case, the OLS estimator gives too much influence to outliers; that is, it is not robust. 'Moving' a response observation to infinity would drag the estimator to infinity as well. It is desirable to have an estimator that 'downweights' the effect of outliers. A number of alternative (ML) estimators are available by considering non-normal error models; for example, the Student t distribution; see Meng and Van Dyk (1997) for a recent likelihood based approach and O'Hagan (1988) for a Bayesian treatment.

In this paper we propose a new estimator, Latent Weighted Least Squares (LWLS), for the parameters in a linear regression model in cases of outlier and/or non-normal error models, using an EM type algorithm that is both simple to code and fast to execute. The development of the LWLS estimator starts from a characterisation of the normal distribution as a scale mixture of a uniform, the mixing distribution being a particular gamma distribution. We

are therefore able to characterise any scale mixture of a normal distribution (Andrews and Mallows, 1974; West, 1987) as a scale mixture of a uniform distribution. In Section 2 we characterise the Student t and exponential power distributions as scale mixtures of uniforms. These characterisations lead to a new family of distributions and a characterisation of the Weibull distribution.

In Section 3 we introduce the LWLS estimation method which is based on the characterisation of the normal as a scale mixture of a uniform. Section 4 contains a simulation study and real data analyses to investigate the LWLS estimator by, comparing with the OLS estimator.

2 Scale mixtures of uniform distributions

We can write the model (1) in a different way: first, introduce the latent variable $u = (u_1, \cdots, u_n)$, with each u_i defined on $(0, \infty)$. Consider the model

$$y_i|u_i = \sum_{j=1}^{k} x_{ij}\beta_j + \tau_i\sqrt{u_i}, \;\; i = 1, \cdots, n \tag{2}$$

where the τ_i are iid from the uniform distribution on $(-1, +1)$, and the u_i are iid from some distribution f_U defined on $(0, \infty)$. The τ_i and the u_i are also independent of each other.

Theorem 1. For the model given in (2)
(i) $Ey_i = \sum_{j=1}^{k} x_{ij}\beta_j$ and y_i is symmetric about the mean; and
(ii) if $Eu_i = 3\sigma^2$ then $\text{var}(y_i) = \sigma^2$.

The proof is straightforward, and is omitted.

Therefore provided $Eu_i = 3\sigma^2$ then the conditions 1) and 2) will be satisfied. One possibility is the normal distribution which arises when f_U is a particular gamma distribution:

Theorem 2. (The Normal error regression model). For the model in (2), if f_U is the gamma distribution with parameters $(3/2, \lambda/2)$, and mean value $3/\lambda$, where $\lambda = \sigma^{-2}$, then marginally each e_i is normally distributed with mean 0 and variance σ^2.

Proof. This follows from the result

$$\int_{u>y^2} \exp(-u)du = \exp(-y^2)$$

and noting that model (2) is equivalent to $y_i|u_i$ being uniformly distributed on the interval $(\mu_i - \sqrt{u_i}, \mu_i + \sqrt{u_i})$, where $\mu_i = \sum_{j=1}^{k} x_{ij}\beta_j$.

We have noted that the normal distribution can be represented as a scale mixture of a uniform distribution. Here we note that two other distributions, the Student t and exponential power distributions also arise as scale mixtures of uniform distributions. Both of these distributions have a scale mixture of normal representation (West, 1987); for example, the Student t, with v degrees of freedom, has the representation

$$y|\xi \sim N(\mu, \sigma^2/\xi),$$

$$\xi \sim \text{ga}(v/2, v/2).$$

We can write the normal first stage as the scale mixture of uniform distribution:

$$y|u \sim U(\mu - \sigma\sqrt{u}, \mu + \sigma\sqrt{u}),$$

$$u|\xi \sim \text{ga}(3/2, \xi/2).$$

We can then combine $f(u|\xi)$ and $f(\xi)$, integrate over ξ, to obtain the marginal distribution of u and hence obtain a scale mixture of uniform representation:

Theorem 3 (The Student t distribution). If f_U has density given up to proportionality by

$$f(u) \propto \frac{\sqrt{u}}{(v+u)^{(v+3)/2}}$$

and

$$y|u \sim U(\mu - \sigma\sqrt{u}, \mu + \sigma\sqrt{u}),$$

then y has a Student t distribution with mean μ, scale parameter σ and v degrees of freedom.

We state the result for the exponential power distribution:

Theorem 4 (The exponential power distribution). If f_U has density proportional to

$$f(u) \propto u^{1/\tau - 1/2} \exp\left(-u^{1/\tau}\right)$$

and

$$y|u \sim U(\mu - \sigma\sqrt{u}, \mu + \sigma\sqrt{u}),$$

then

$$f(y) \propto \exp\left(-\left|\frac{y-\mu}{\sigma}\right|^{2/\tau}\right),$$

where $\tau \in (0, 2]$.

This characterisation of the exponential power distribution appears to be more tractable than the alternative scale mixture of a normal characterisation (West, 1987), which is only valid for $\tau \in (1, 2]$. We can obtain an interesting result by combining the result of West with ours:

Theorem 4 (West, 1987). If $y|\lambda \sim N(0, \lambda)$, here λ denotes 1/(variance), and $f(\lambda) \propto \lambda^{-1/2} p_{1/\tau}(\lambda)$, $(1 < \tau < 2)$, where $p_a(.)$ denotes the density of the positive stable distribution with index a $(0 < a < 1)$, then $f(y) \propto \exp(-|y|^{2/\tau})$.

We can now insert the uniform and gamma mixture to replace the normal, leading to the following 3 stage mixture:

$$y|u \sim U(-\sqrt{u}, +\sqrt{u})$$

$$u|\lambda \sim \text{ga}(3/2, \lambda/2)$$

and

$$f(\lambda) \propto \lambda^{-1/2} p_{1/\tau}(\lambda).$$

Combining the last two stages implies:

$$u^{1/\tau - 1/2} \exp(-u^{1/\tau}) \propto u^{1/2} \int_{\lambda=0}^{\infty} \lambda^{3/2} \exp(-0.5\lambda u) \lambda^{-1/2} p_{1/\tau}(\lambda) d\lambda.$$

Therefore,

Theorem 5. If $u|\lambda$ has the exponential distribution with mean $2/\lambda$ and

$f(\lambda) \propto p_{1/\tau}(\lambda)$, $(1 < \tau < 2)$, then $f(u) \propto u^{1/\tau - 1}\exp(-u^{1/\tau})$, a Weibull distribution.

In fact any distribution which has a scale mixture of normal representation also has a scale mixture of uniform representation. This is easy to see; if $x|\lambda \sim N(0, \lambda)$ and $\lambda \sim g$ then $x|u \sim U(-\sqrt{u}, +\sqrt{u})$ with $u \sim f$ where

$$f(u) \propto \sqrt{u} \int_{\lambda=0}^{\infty} \lambda^{3/2} \exp(-0.5\lambda u) g(\lambda) d\lambda.$$

There does not seem to be any reason why we should just consider the Student t and exponential power families.

Consider the general family of distributions:

$$y|u \sim U(\mu - \sigma\sqrt{u}, \mu + \sigma\sqrt{u})$$

$$u \sim f_U.$$

Then the following hold: $Ey = \mu$, $\text{var}(y) = \sigma^2 Eu/3$ and

$$\text{kurtosis}(y) = \frac{9E[u^2]}{5[Eu]^2} - 3.$$

So the mean and variance of u determine the variance and kurtosis of y. To obtain $\text{var}(y) = \sigma^2$ and $\text{kurtosis}(y) = \tau$ we require $Eu = 3$ and $\text{var}(u) = 5\tau + 6$. Note that we must have $\tau > -6/5$ which is the kurtosis for the uniform density. A particular distribution which satisfies these requirements is given by

$$f_U = \text{ga}(9\alpha, 3\alpha),$$

where $\alpha = (5\tau + 6)^{-1}$. This new family of distributions has parameters (μ, σ, τ), with mean μ, variance σ^2 and kurtosis τ. We recover the normal distribution when $\tau = 0$ ($\alpha = 1/6$).

In fact the scale mixture of uniform family coincides with the class of unimodal, symmetric distributions:

Theorem 6. If f_X is a unimodal, symmetric density about 0 and $f'_X(x)$ exists for all x then

$$f_X(x) = 1/2 \int_{u>x^2} f_U(u) du/\sqrt{u},$$

where $f_U(u) = -f_X'(\sqrt{u})$.

Therefore, we can write $x|u \sim U(-\sqrt{u}, +\sqrt{u})$ with $u \sim f_U$, provided f_U is a density on $(0, \infty)$. Note that $-\int_{u=0}^{\infty} f_X'(\sqrt{u})du = 1$ which follows from

$$1 = \int_{-\infty}^{+\infty} f_X(x)dx = 1/2 \int_{u=0}^{\infty} -f_X'(\sqrt{u})/\sqrt{u} \left[\int_{x=-\sqrt{u}}^{+\sqrt{u}} dx \right] du,$$

and $f_U(u) \geq 0$ iff $f_X(x)$ is unimodal.

Theorem 6 is a consequence of a theorem of Feller (1971, pp. 155); see, also, Brunner and Lo (1989). We note that our approach appears to be more general and simpler than the one provided by Feller. Feller (see, also, Brunner and Lo) considers the different scale mixture of uniform model, given by

$$x|u \sim (\mu - u, \mu + u),$$
$$u \sim G,$$

for some distribution G with support on $(0, \infty)$. Brunner and Lo then assign G a Dirichlet process prior. However, this model only provides the unimodal, symmetric distribution for X and the first *four* moments of U are all required to specify the first four moments of X. With our model

$$x|u \sim (\mu - \sigma\sqrt{u}, \mu + \sigma\sqrt{u}),$$
$$u \sim f_U;$$

it is clear that since we are explicitly modelling the variance of X, only the first *two* moments of U are required to define the first four moments of X. We can 'improve' on the Brunner and Lo (1990) model by considering the following model:

$$x|u \sim (\mu - \sigma\sqrt{u}, \mu + \sigma\sqrt{u}),$$
$$u \sim G,$$

with G taken from a Polya tree prior, which happens to generalise the Dirichlet prior. For this we would need to fix the location of G, which would otherwise be confounded with σ. We can achieve this by fixing the median of G, an easy task, and this development will be reported elsewhere.

3 Latent weighted least squares (LWLS)

The advantage of introducing the latent model (2) is the availability of the 'natural' weights $(u_1, \cdots, u_n)$ for a weighted least squares method, based on $\text{var}(y_i|u_i) = u_i/3$. Given the u_i, the standard estimator for the regression parameter is the weighted least squares (WLS) estimate;

$$\hat{\beta}_{WLS}(u) = [\sum_i w_i X_i X_i']^{-1} \sum_i w_i y_i X_i \tag{3}$$

where $w_i = 1/u_i$. This will 'downweight' extreme observations since it is clear from (2) that for large y_i the larger u_i will need to be.

In principle, we have a linear random effects model not unlike the original linear random effects models considered by Laird and Ware (1982). A slightly modified version of their hybrid EM algorithm is used to obtain a LWLS estimate. We treat the u_i as missing data so that (y, u) represents the complete dataset. With this complete dataset we have, with f_U as ga$(3/2, \lambda/2)$,

$$E(y_i|u_i) = \sum_{j=1}^{k} x_{ij}\beta_j \text{ and } \text{var}(y_i|u_i) = u_i/3$$

and we take $\hat{\beta} = \hat{\beta}_{WLS}(u)$ given by (3). Following Laird and Ware we then obtain the expectations of the sufficient statistics for the missing data; that is, $\hat{u}_i = E(u_i|\beta, \sigma, y)$, so

$$\hat{u}_i = (y_i - \sum_{j=1}^{k} x_{ij}\beta_j)^2 + 2\sigma^2. \tag{4}$$

Finally, we maximise the complete data likelihood $l(\sigma^2|y, u)$ to update σ:

$$\hat{\sigma}^2 = \sum_i u_i/(3n). \tag{5}$$

Combining (3), (4) and (5) together gives our algorithm for obtaining the LWLS estimator for the linear regression model:

1) $\hat{\beta} = [\sum_i w_i X_i X_i']^{-1} \sum_i w_i y_i X_i$, where $w_i = 1/\hat{u}_i$,
2) $\hat{\sigma}^2 = \sum_i \hat{u}_i/(3n)$ and,
3) $\hat{u}_i = (y_i - \sum_{j=1}^{k} x_{ij}\hat{\beta}_j)^2 + 2\hat{\sigma}^2$.

The algorithm iterates over 1), 2) and 3) until convergence. Suitable starting values should be the OLS estimator for β and the sample variance for σ^2. A simulation study and real data analyses, comparing LWLS with OLS, is presented in Section 4.

4 Examples

To begin our investigation of the LWLS estimator we simulated 50 standard normal random variates and computed the OLS estimate and the LWLS estimate. We repeated this exercise 1000 times and computed the mean and variance of the sample estimates: for the OLS these are $(5.9e-04, 0.02)$, and for the LWLS these are $(-1.1e-03, 4.4e-06)$.

At the other end of the scale, we simulated 50 standard Student t random variates with 2 degrees of freedom and, again, repeated the exercise 1000 times. The corresponding mean and variance for the sample estimates are $(-5.4e-03, 0.22)$ and $(7.2e-03, 5.5e-05)$ for the OLS and LWLS estimators, respectively.

It is quite obvious from these results that the LWLS estimator improves quite remarkably on the OLS estimator in these two fundamental, yet common, sampling scenarios.

How does the LWLS estimator cope with contaminated data and outliers? To investigate this point we sampled 50 random variates from the contaminated standard normal, obtained by adding 5 to any of the standard normal random variates with probability 1/10. The corresponding mean and variance for the sample estimates are $(0.49, 0.06)$ and $(0.11, 1.1e-04)$ for the OLS and LWLS estimators, respectively.

For an analysis involving a real dataset we turn to Box and Tiao (1973, Table 3.4.1). The data consists of 20 experiments relating the rate of a chemical reaction to the temparature at which the experiment was conducted. The linear model

$$y_i = \beta_1 + \beta_2 x_i + e_i,$$

is used where y_i and x_i denote the log reaction rate and a a measure of the temperature for the ith experiment, respectively. Box and Tiao, within a robust Bayesian framework, use the exponential power family for modelling the error distribution, which includes the normal distribution.

They conclude that the normality assumption, by consideration of the power parameter of the exponential power distribution, is valid and under this asumption obtain an estimate for (β_1, β_2) as $(-4.013, -0.203)$. Our LWLS estimator turns out to be $(-4.008, -0.202)$.

In the same chapter of their book, Box and Tiao reanalyse the Darwin dataset (Table 3.2.1) consisting of 15 iid observations. They again make use

of the exponential power family to model the observations. They conclude in this case that the normality assumption is not valid and the true distribution shows signs of leptokurticity.

The mean of the data is 20.933 and the LWLS estimate is given by 26.469. This higher estimate is in keeping with the results of Box and Tiao (see, Figures 3.2.3 and 3.2.6).

5 Discussion

In this paper, we study scale mixtures of uniform distributions: which coincides with the family of unimodal, symmetric densities on the real line. We introduce a new (mixture) family of distributions which is ideally suited to the modelling of location, scale and kurtosis.

A new estimator for the linear regression model, the Latent Weighted Least Squares (LWLS) estimator, is developed which is simple to code and fast to execute. Illustrative analyses exemplifying the method are provided and demonstrate vast improvement on the OLS estimator in a number of sampling scenarios. We encourage its widespread use.

References

Andrews, D.F. and Mallows, C.L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B* **36**, 99-102.

Box,G.E.P. and Tiao,G.C. (1973). *Bayesian inference in statistical analysis.* Addison-Wesley, Massachusetts.

Brunner, L.J. and Lo, A.Y. (1989). Bayes methods for a symmetric unimodal density and its mode. *The Annals of Statistics* **17**, 1550-1566.

Feller, W. (1971). *An introduction to probability theory and its applications* II, Wiley, New York.

Laird, N.M. and Ware, J.H. (1982). Random effects model for longitudinal data. *Biometrics* **38**, 963-974.

Meng, X. and Van Dyk, D. (1997). The EM algorithm: an old folk-song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society, Series B.* To appear.

O'Hagan, A. (1988). Modelling with heavy tails. In *Bayesian Statistics 3*, pp. 345-359. J.M.Bernardo, M.H.DeGroot, D.V.Lindley and A.F.M.Smith (eds.) Oxford University Press.

West, M. (1987). Scale mixtures of normal distributions. *Biometrika* **74**, 646-648.