PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 31, Fasc. 2 (2011), pp. 177–182

UNBIASED ESTIMATES FOR LINEAR REGRESSION WITH ROUNDOFF ERROR

BY

CHRISTOPHER S. WITHERS (LOWER HUTT)
AND SARALEES NADARAJAH (MANCHESTER)

Abstract. We consider the linear regression model, where the residuals have zero mean and an otherwise unspecified distribution F. Suppose that least squares estimates are formed by using *rounded* values of the dependent variables. We show that these are still unbiased, and that unbiased estimates for the moments and cumulants of F are given by applying Sheppard's corrections to their estimates.

2000 AMS Mathematics Subject Classification: Primary 62J05; Secondary 62G05.

Key words and phrases: Grouping, least squares, regression, roundoff error, Sheppard's corrections, unbiased.

1. INTRODUCTION

The linear regression model is one of the most popular models in statistics. It is also one of the simplest models in statistics. It has received applications in almost every area of science, engineering and medicine.

The data on the dependent variables are often *rounded off*. There may be many reasons that the data are rounded off, for example: recording standard, faulty equipments, human error or the lack of technology to measure variables accurately.

Many authors have considered the important problem: the estimates of linear regression with the rounded values used instead of the actual ones. For most excellent reviews of the literature, see Heitjan [5] and Schneeweiss et al. [8]. However, much of the work has focused on correcting for bias of the estimates under suitable conditions.

In this short note, we show that the estimates of linear regression can, in fact, be unbiased if the rounded values are used instead of the actual ones; see Theorem 2.1 in Section 2. We also give unbiased estimates for the rth moment and the rth cumulant of the residuals of linear regression for any r; see Theorem 2.2 in Section 2. The related proofs are given in Section 4. We assume throughout that Sheppard's corrections hold exactly.

We believe it is the first time that a general result like Theorem 2.2 has been proven. This result could have potentially wide spread applications. The first four moments and the first two cumulants have obvious physical interpretations and use in statistical inference. There are often situations, where one also needs higher moments and higher cumulants. For example, there are many situations in insurance and economics that require moments of orders higher than four. We mention:

• Taleb [9] suggests using moments of order higher than four to measure the risk of an option. For example, the fifth moment is presumed as the asymmetry sensitivity of the fourth one. The seventh moment is suggested as the sign of the convexity change as the underlying asset moves up or down.

• Avramidis and Matzinger [1] show that an estimator for pricing American options can be improved using moments of order higher than four. An example in the area of Cornish–Fisher expansions (Cornish and Fisher [2]; Fisher and Cornish [3]) is described in Section 3.

2. MAIN RESULTS

Suppose we observe $Y_N = \mathbf{x}_N^{'}\boldsymbol{\beta} + e_N$, $1 \leqslant N \leqslant n$, where $\mathbf{x}_N$ and $\boldsymbol{\beta}$ are known and unknown q-vectors, respectively, and $\{e_N\}$ are independently and identically distributed according to F, an unknown distribution with zero mean and unknown central moments and cumulants $\{\mu_r, \kappa_r, r \geqslant 2\}$.

Suppose we record not $\mathbf{Y}^{'} = (Y_1, \ldots, Y_n)$ but $\mathbf{Z}^{'} = (Z_1, \ldots, Z_n)$, where $Z_N = T_h(Y_N)$, say, is Y_N rounded to the nearest integral multiple of a given $h < 0$, with mid-values rounded down (or F continuous). Let $\widehat{\boldsymbol{\beta}}_h$ denote the least squares estimate about $\{Z_N\}$:

$$\widehat{\boldsymbol{\beta}}_h = (\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{Z}, \quad \text{where } \mathbf{X}^{'} = (x_1, \ldots, x_n).$$

THEOREM 2.1. *We have* $\mathbb{E}[\widehat{\boldsymbol{\beta}}_h] = \boldsymbol{\beta}$.

In Withers and Nadarajah [12], unbiased estimates (UEs) were given for $\{\mu_r, \kappa_r, r \geqslant 2\}$, say $\widehat{\mu}_r = \widehat{\mu}_r(Y), \widehat{\kappa}_r = \widehat{\kappa}_r(Y)$. In particular,

$$\widehat{\mu}_2 = |\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}|^2/(n-q),$$

as is well known.

THEOREM 2.2. *UEs of* $\{\mu_r, \kappa_r, r \geqslant 2\}$ *are given by applying Sheppard's corrections to* $\{\widehat{\mu}_{r,h}, \widehat{\kappa}_{r,h}, r \geqslant 2\}$, *where* $\widehat{\mu}_{r,h} = \widehat{\mu}_r(\mathbf{Z})$, $\widehat{\kappa}_{r,h} = \widehat{\kappa}_r(\mathbf{Z})$. *That is*, *UEs are given by*

$$\begin{aligned}
\widetilde{\mu}_2 &= \widehat{\mu}_{2,h} - H/3, \\
\widetilde{\mu}_3 &= \widehat{\mu}_{3,h}, \\
\widetilde{\mu}_4 &= \widehat{\mu}_{4,h} - 2H\widehat{\mu}_{2,h} + 7H^2/5, \\
\widetilde{\mu}_5 &= \widehat{\mu}_{5,h} - 10H\widehat{\mu}_{3,h}/3,
\end{aligned}$$

where $H = h^2/4$. *In general, for* $r \geqslant 2$

$$\widetilde{\mu}_r = \sum_{0 \leqslant m \leqslant r/2} b_m H^m \binom{r}{2m} \widehat{\mu}_{r-2m,h},$$

$$\widetilde{\kappa}_r = \widehat{\kappa}_{r,h} - h^r B_r / r,$$

where

$$\widehat{\mu}_{r,h} = \sum_{0 \leqslant m \leqslant r/2} (2m+1)^{-1} H^m \binom{r}{2m} \widetilde{\mu}_r,$$

$$b_m = (2 - 2^{2-2m}) B_{2m},$$

and B_r *is the* r*th Bernoulli number.*

The conditions on F in these theorems are as for Sheppard's corrections with $r = 1$ for Theorem 2.1. Basically, this requires sufficient regularity of F in its tails.

3. EXAMPLE

Here, we describe an example, where the results of Theorem 2.2 can be applied.

Suppose we wish to make inferences about a parameter θ. Let $\widehat{\theta}$ denote an estimator of θ. Suppose the cumulants of $\widehat{\theta}$ satisfy the standard expansion

$$\kappa_r(\widehat{\theta}) = \sum_{j=r-1}^{\infty} a_{r,j} n^{-j} \tag{3.1}$$

for $r \geqslant 1$ with the coefficients $\{a_{r,j}\}$ bounded as $n \to \infty$, $a_{2,1}$ bounded away from zero. This assumption is key to obtaining Cornish–Fisher expansions for $\widehat{\theta}$ and is known to be true for a wide variety of estimates, including smooth functions of sample means.

By (3.1), the mean and variance of $\widehat{\theta}$ are asymptotic to $a_{1,0}$ and $a_{2,1}/n$, respectively. So, a central limit approximation is to assume that

$$Y_n = (n/a_{2,1})^{1/2} (\widehat{\theta} - a_{1,0}) \tag{3.2}$$

has the standard normal distribution for n sufficiently large. But this approximation can be crude.

It would be better to use corrections for the central limit version. Withers [10] (see also Kolassa and McCullagh [7]) showed that under the assumption (3.1) the asymptotic expansions of Cornish and Fisher [2] and Fisher and Cornish [3] for

the distribution and quantiles of $\widehat{\theta}$ reduce to the form

$$P_n(x) = P\left(Y_n \leqslant x\right) = \Phi(x) - \phi(x) \sum_{r=1}^{\infty} n^{-r/2} h_{0,r}(x), \tag{3.3}$$

$$\Phi^{-1}\big(P_n(x)\big) = x - \sum_{r=1}^{\infty} n^{-r/2} f_r(x), \tag{3.4}$$

$$P_n^{-1}\big(\Phi(x)\big) = x + \sum_{r=1}^{\infty} n^{-r/2} g_r(x), \tag{3.5}$$

where $\Phi(x)$ and $\phi(x)$ are the distribution and density, respectively, of a standard normal random variable. The functions $h_{0,r}(x)$, $f_r(x)$, $g_r(x)$ are certain polynomials of degree $3r-1$, $r+1$, $r+1$ in x, respectively, and $A_{r,i} = a_{r,i}/a_{2,1}^{r/2}$.

The infinite sums on the right-hand sides of (3.3)–(3.5) can be truncated to provide better inferences for θ than the central limit version given by (3.2). The polynomials $h_{0,r}(x)$, $f_r(x)$, $g_r(x)$ involve the rth order cumulants of $\widehat{\theta}$. Consequently, higher order inferences for θ would require knowing estimates for higher order cumulants of $\widehat{\theta}$.

Withers and Nadarajah [11] have demonstrated how (3.1) and (3.3)–(3.5) can be applied to construct improved confidence intervals for parameters of interest when the data are rounded off.

4. PROOFS

P r o o f o f T h e o r e m 2.1. Sheppard showed (see Kendall and Stuart [6], p. 76) that if $X \sim F$ is rounded to produce $X_h = T_h \sim F_h$, say, then their noncentral moments and their cumulants are connected by

$$\mu'_{r,h} = \mu'_r\left(F_h\right) = \sum_{0 \leqslant m} (2m+1)^{-1} H^m \binom{r}{2m} \mu'_r, \tag{4.1}$$

$$\mu'_r = \mu'_r(F) = \sum_{0 \leqslant m} b_m H^m \binom{r}{2m} \mu'_{r,h} \tag{4.2}$$

and, for $r \neq 1$,

$$\kappa_r = \kappa_{r,h} - h^r B_r / r. \tag{4.3}$$

In particular,

$$\mathbb{E}\left(X_h\right) = \mathbb{E}(X). \tag{4.4}$$

So, $\mathbb{E}(Z) = \mathbb{E}(Y)$ and Theorem 2.1 holds. ∎

To prove Theorem 2.2 we need the following lemma.

LEMMA 4.1. *For any constant c, $T_h(X-c)$ and $T_h(X)-c$ have the same moments.*

P r o o f. By (4.4), we have $\mathbb{E}[T_h(X-c)] = \mathbb{E}(X) - c = \mathbb{E}[T_h(X)] - c$. Let $G(y) = P(X - c \leqslant y) = F(y+c)$ and G_h be the distribution of $T_h(X-c)$. Note that (4.1) and (4.2) hold for central moments, i.e. with the dashes dropped. So, for $r \geqslant 2$,

$$\begin{aligned}\mu_r(G_h) &= \sum (2m+1)^{-1} H^m \binom{r}{2m} \mu_r(G) \\ &= \sum (2m+1)^{-1} H^m \binom{r}{2m} \mu_r(F) \\ &= \mu_r(F_h) = \mu_r[T_h(X) - c],\end{aligned}$$

where F, F_h are the distributions of X, $T_h(X)$, respectively, and

$$\mu_r[X] = \mu_r(F). \quad \blacksquare$$

P r o o f o f T h e o r e m 2.2. Set $e_{h,N} = Z_N - \mathbf{x}_N' \boldsymbol{\beta}$. Then, by equations (1.2), (2.2), (3.1) of Withers and Nadarajah [12], $\widehat{\mu}_{r,h}$ and $\widehat{\kappa}_{r,h}$ each have the form

$$\sum R_{i_1,\ldots,i_r}(X) e_{h,i_1} \ldots e_{h,i_r}.$$

Note that $\mathbb{E}[e_{h,i}] = \mathbb{E}[e_i] = 0$ by (4.4), so

$$\mathbb{E}\left[e_{h,i_1} \ldots e_{h,i_r}\right] = \prod_j \mathbb{E}[e_{h,L_j}^{a_j}],$$

where $\{L_j\}$ are the distinct values of $\{i_j\}$ and a_j is the number of i's equal to L_j. By Theorem 2.1, $\mathbb{E}[e_{h,L}^a] = \mu_a(F_h)$, where F_h is the distribution of $T_h(e_N)$, $N \geqslant 1$. So,

$$\mathbb{E}\left[e_{h,i_1} \ldots e_{h,i_r}\right] = \prod_j \mu_{a_j}(F_h).$$

But in the case of no round-off,

$$\mathbb{E}\left[e_{i_1} \ldots e_{i_r}\right] = \prod_j \mu_{a_j}(F)$$

and $\mu_r(F) = \mathbb{E}[\widehat{\mu}_r] = \mu_r(F)$, $\mathbb{E}[\widehat{\kappa}_r] = \mu_r(F)$. So, $\mathbb{E}[\widehat{\mu}_{r,h}] = \mu_r(F_h)$ and $\mathbb{E}[\widehat{\kappa}_{r,h}]$ $= \kappa_r(F_h)$. Thus, the result for moments follows by the fact that Sheppard's corrections apply to central moments, and the result for cumulants is implied by (4.3). $\blacksquare$

Acknowledgments. The authors would like to thank the editor and the referee for carefully reading and for their comments which greatly improved the paper.

REFERENCES

[1] A. N. Avramidis and H. Matzinger, *Convergence of the stochastic mesh estimator for pricing American options*, in: *Proceedings of the 2002 Winter Simulation Conference*, E. Yücesan, C.-H. Chen, J. L. Snowdon and J. M. Charnes (Eds.), 2002, pp. 1560–1567.

[2] E. A. Cornish and R. A. Fisher, *Moments and cumulants in the specification of distributions*, Revue de l'Institut International de Statistics 5 (1937), pp. 307–322. Reproduced in: *The Collected Papers of R. A. Fisher*, Vol. 4.

[3] R. A. Fisher and E. A. Cornish, *The percentile points of distributions having known cumulants*, Technometrics 2 (1960), pp. 209–225.

[4] P. Hall, *The Bootstrap and Edgeworth Expansion*, Springer, New York 1992.

[5] D. F. Heitjan, *Inference from grouped continuous data: A review*, Statist. Sci. 4 (1989), pp. 164–179.

[6] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. 1, second edition. Griffin, London 1963.

[7] J. E. Kolassa and P. McCullagh, *Edgeworth expansions for lattice distributions*, Ann. Statist. 18 (1990), pp. 981–985.

[8] H. Schneeweiss, J. Komlos and A. S. Ahmad, *Symmetric and asymmetric rounding*, Discussion paper 479, Sonderforschungsbereich 386, University of Munich, 2006.

[9] N. Taleb, *Dynamic Hedging*, Wiley, New York 1997.

[10] C. S. Withers, *Asymptotic expansions for distributions and quantiles with power series cumulants*, J. Roy. Statist. Soc., Ser. B 46 (1984), pp. 389–396.

[11] C. S. Withers and S. Nadarajah, *Adjusting Cornish–Fisher expansions and confidence intervals for the effect of roundoff*, Statistics (2011), doi: 10.1080/02331888.2010.539691.

[12] C. S. Withers and S. Nadarajah, *Unbiased estimates for moments and cumulants in linear regression*, J. Statist. Plann. Inference (2011), to appear.

Applied Mathematics Group
Industrial Research Limited
Lower Hutt, New Zealand
E-mail: c.withers@cri.irl.nz

School of Mathematics
University of Manchester
Manchester M13 9PL
United Kingdom
E-mail: mbbsssn2@manchester.ac.uk

Received on 13.4.2010;
revised version on 2.2.2011