1352.0.55.051

Australian Bureau of Statistics

Research Paper

Winsorization for Generalised Regression Estimation

www.abs.gov.au

Research Paper

Winsorization for Generalised Regression Estimation

John Preston and Carl Mackin

Statistical Services Branch

Methodology Advisory Committee

22 November 2002, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) MON 20 MAY 2002

ABS Catalogue no. 1352.0.55.051
ISBN 0 642 48155 5

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr John Preston, Statistical Services Branch on Brisbane (07) 3222 6229 or email <john.preston@abs.gov.au>.

WINSORIZATION FOR GENERALISED REGRESSION ESTIMATION

John Preston and Carl Mackin
Statistical Services Branch

EXECUTIVE SUMMARY

The availability of Business Activity Statement (BAS) data collected by the Australian Taxation Office (ATO) has provided the Australian Bureau of Statistics (ABS) with opportunities to improve the efficiency of sample design and estimation for its business surveys. The ABS business surveys currently use two methods of estimation; number-raised estimation and ratio estimation. While ratio estimation allows the use of one auxiliary variable to improve the precision of the estimates, generalised regression (GREG) estimation allows the use of more than one auxiliary variable, and hence has the potential to be more efficient (i.e. reduce the current sample sizes for ABS business surveys with no reduction in the precision of the estimates) than number-raised and ratio estimation.

The generalised regression estimator is unbiased with respect to the assumed model. However, if by chance there are several units in the sample with unusually large residuals under the generalised regression model, then the generalised regression estimator may grossly underestimate or overestimate the population totals. One solution to this problem is to modify values outside preset cutoff values to values closer to these cutoff values. This estimator is called the 'winsorized' estimator. Although the winsorized estimator is biased, it may have a considerably smaller mean squared error than the generalised regression estimator.

In order to minimise the mean squared error of the winsorized estimator, the choice of the cutoff values can be written as functions of the proposed regression model and the bias of the winsorized estimator. The suitability of the winsorized estimator will ultimately depend on the choice of the cutoff values, and hence the methods used to estimate the bias parameters and regression parameters used to calculate these cutoff values.

The estimate of bias parameters can be calculated using the approach outlined in Kokic and Bell (1994). However, there are many solutions to the problem of estimating robust regression parameters. It is worth noting that the task is not to find the best robust regression model, but rather to find a robust regression fitting procedure which results in the best performing winsorized estimator. Furthermore, since this procedure will have to be used for many ABS business surveys and understood by a range of people, it is desirable that the procedure is simple, flexible

and easily traceable. An investigation was performed using a number of different robust regression fitting techniques to determine which techniques resulted in the best performing winsorized estimator.

A simulation study was undertaken to assess the performance of these various methods to estimate robust regression parameters, and hence estimate the cutoff values used in the winsorized estimator. The simulation study examined:

- the 'best' robust regression fitting technique;
- the data used to calculate cutoff values under winsorization;
- the level to calculate bias parameters under winsorization; and
- the sample weights to calculate the cutoff values under winsorization.

One of the key findings of the study was that diagnostics should be incorporated into the regression fitting procedure before using the regression parameters to generate cutoff values. In particular, checking that the regression model fits the data well, checking that units with large influences are removed from the regression model, and checking that the regression model fits to current data to be winsorized.

There often exists linear relationships between the various data items collected and derived in ABS business surveys, and it is important that these linear relationships still hold after winsorization. The current ABS estimation system allows the linear relationships to be maintained by two methods. Unfortunately, there are some situations where these two methods perform quite poorly. An alternative method which attempts to overcome the shortcomings of the two methods is suggested, which requires the specification of a distance function between the original and final winsorized values. Although any one of a number of distance functions could be used, the one examined in this paper is the generalised least squares distance function.

DISCUSSION POINTS FOR MAC

The questions for MAC members in relation to winsorization for generalised regression estimation are:

- Is the solution under linear interpolation to estimate the bias parameters better than taking the last positive breakpoint?
- Is it logical to apply the same regression model used for the generalised regression estimator, to generate the regression parameters for the winsorized cutoff values? Should this same model be used to form winsorized cutoff values for all variables collected in the survey? If a regression model is known which fits another variable better, then should this regression model be used to form winsorization cutoff values for this other variable, even though it was not used for the generalised regression estimator?
- Is it logical to fit regression model at different levels to the generalised regression model?
- What is the best way to ensure the regression model fitted to the historical data is still applicable to the current data to be winsorized? What action should be taken when the regression model does not fit the current data to be winsorized?
- What is the best way to deal with units with large historical values which have an adverse influence on the estimate of the bias parameter?
- Is there any reason why the design weights should not be used in the calculation of cutoff values used in the winsorized estimator?
- Is the concept of minimising a distance function to ensure linear relationships between variables still hold after winsorization appropriate? Is the generalised least squares distance function appropriate?

CONTENTS

WINSORIZATION FOR GENERALISED REGRESSION ESTIMATION

John Preston and Carl Mackin
Statistical Services Branch

1. INTRODUCTION

The availability of Business Activity Statement (BAS) data collected by the Australian Taxation Office (ATO) has provided the Australian Bureau of Statistics (ABS) with opportunities to improve the efficiency of sample design and estimation for its business surveys. The ABS business surveys currently use two methods of estimation; number-raised estimation and ratio estimation. While ratio estimation allows the use of one auxiliary variable to improve the precision of the estimates, generalised regression (GREG) estimation allows the use of more than one auxiliary variable, and hence has the potential to be more efficient (i.e. reduce the current sample sizes for ABS business surveys with no reduction in the precision of the estimates) than number-raised and ratio estimation. The BAS data will potentially be a rich source of auxiliary variables for use in GREG estimation.

1.1 Generalised regression estimator

Consider a finite population $U = \{1,...,i,...,N\}$, from which a probability sample $s\,(s \subseteq U)$ is drawn according to a sample design with selection probabilities $\pi_i = \Pr(i \in s)$. The sampling weights $w_i = 1/\pi_i$ are those used in the Horvitz–Thompson estimator, $\hat{t}_{y\pi} = \sum_{i \in s} w_i y_i$, for variable of interest y. The objective is to estimate the population total $Y = \sum_{i \in U} y_i$, where y_i is the value of the variable of interest y for unit i. Assume there exists a set of auxiliary variables $\underset{\sim}{x}_i = (x_{1i},\ldots,x_{ki},\ldots,x_{Ki})^T$ for which the population totals $\underset{\sim}{t}_x = \sum_{i \in U} \underset{\sim}{x}_i$ are known. The generalised regression estimator is given by (Sarndal, Swensson and Wretman, 1992):

$$\hat{t}_{yreg} = \sum_{i \in s} w_i y_i + \left(\underset{\sim}{t}_x - \sum_{i \in s} w_i \underset{\sim}{x}_i \right)^T \underset{\sim}{\hat{\beta}}$$

where

$$\underset{\sim}{\hat{\beta}} = \left(\sum_{i \in s} w_i \underset{\sim}{x}_i \underset{\sim}{x}_i^T \right)^{-1} \left(\sum_{i \in s} w_i \underset{\sim}{x}_i y_i \right).$$

The generalised regression estimator is often written as:

$$\hat{t}_{yreg} = \sum_{i \in s} w_i g_i y_i = \sum_{i \in s} \tilde{w}_i y_i$$

where g_i the g-weight for unit i, defined as:

$$g_i = \left(1 + \underset{\sim}{x}_i^T \left(\sum_{i \in s} w_i \underset{\sim}{x}_i \underset{\sim}{x}_i^T \right)^{-1} \left(\underset{\sim}{t}_x - \sum_{i \in s} w_i \underset{\sim}{x}_i \right)^T \right)$$

1.2 Winsorization

The generalised regression estimator is unbiased with respect to the assumed model. However, if by chance there are several units in the sample with unusually large residuals under the generalised regression model, then the generalised regression estimator may grossly underestimate or overestimate the population totals. It is desirable to robustify the estimator against such unusually large residuals. There are two distinct solutions to this problem. The first is to modify the weights associated with these units (Hidiroglou and Srinath, 1981), while the second is to modify the values of the variables of interest for these units. One approach to this second solution is to modify values outside preset cutoff values to values closer to these cutoff values. This estimator is called the 'winsorized' estimator (Searls, 1966). Although the winsorized estimator is biased, it may have a considerably smaller mean squared error than the generalised regression estimator.

Let $\hat{t}_y = \sum_{i \in s} \tilde{w}_i y_i$ be an unbiased estimator of the population total, under the model:

$$E(Y_i) = \mu_i$$

$$Var(Y_i) = \sigma_i^2$$

and suppose the winsorized estimator of the population total is given by:

$$\hat{t}_{ywin} = \sum_{i \in s} \tilde{w}_i y_i^*$$

where the winsorized value, y_i^*, is calculated using the Type II winsorized estimator (Gross, Bode, Taylor and Lloyd–Smith 1986), modified for two-sided winsorization:

$$y_i^* = \begin{cases} \left(\frac{1}{\tilde{w}_i} \right) y_i + \left(1 - \frac{1}{\tilde{w}_i} \right) K_{Ui} & \text{,if } y_i > K_{Ui} \\ y_i \quad \text{,if } K_{Ui} \geq y_i \geq K_{Li} & \\ \left(\frac{1}{\tilde{w}_i} \right) y_i + \left(1 - \frac{1}{\tilde{w}_i} \right) K_{Li} & \text{,if } y_i < K_{Li} \end{cases}$$

(i.e. the outlier contributes its unweighted values, while the non-sampled units, represented by the remainder of the weight, $\tilde{w}_i - 1$, contribute preset upper or lower cutoff values, K_{Ui} and K_{Li}, to the estimate of the population).

In order to minimise the mean squared error of the winsorized estimator under the model, the choice of the cutoff value (Clark 1995) is given by:

$$K_{Ui} = \mu_i^* - \frac{B_U}{(\tilde{w}_i - 1)}$$

$$K_{Li} = \mu_i^* - \frac{B_L}{(\tilde{w}_i - 1)}$$

where $\mu_i^* = E\left(Y_i^*\right)$, and B_U and B_L are the bias of $\hat{t}_{ywinU}$ and $\hat{t}_{ywinL}$:

$$B_U = E[\hat{t}_{ywinU} - t_y]$$

$$B_L = E[\hat{t}_{ywinL} - t_y]$$

where $\hat{t}_{ywinU}$ is the estimate of population total when only upper winsorization is performed and $\hat{t}_{ywinL}$ is the estimate of population total when only lower winsorization is performed.

In practice μ_i^* is difficult to estimate. Under the assumptions that winsorization is mild and reasonably symmetric μ_i^* is replaced with μ_i to give approximately optimal cutoffs:

$$K_{Ui} = \mu_i - \frac{B_U}{(\tilde{w}_i - 1)}$$

$$K_{Li} = \mu_i - \frac{B_L}{(\tilde{w}_i - 1)}$$

2. CHOICE OF CUTOFF VALUES

The suitability of the winsorized estimator will ultimately depend on the choice of the cutoff values, and hence the methods used to estimate the bias parameters and regression parameters used to calculate these cutoff values.

2.1 Estimation of bias parameters

The estimate of bias parameter B_U under winsorization depends on the values of the upper cutoffs, and can be calculated using the approach outlined in Kokic and Bell (1994). Firstly, define weighted residuals:

$$D_i = (Y_i - \mu_i)(\tilde{w}_i - 1)$$

and let $U = -B_U$ such that $K_{Ui} = \mu_i + \dfrac{U}{(\tilde{w}_i - 1)}$, then the upper bias parameter can be written as:

$$\begin{aligned}
B_U(U) &= E[\hat{t}_{ywinU} - t_y] \\
&= \sum_{i \in s} (\tilde{w}_i - 1)\{E[\min(Y_i, K_{Ui})] - \mu_i\} \\
&= \sum_{i \in s} E[\min\{(Y_i - \mu_i)(\tilde{w}_i - 1), (K_{Ui} - \mu_i)(\tilde{w}_i - 1)\}] \\
&= \sum_{i \in s} E[\min\{D_i, U\}] \\
&= \sum_{i \in s} E[\min\{0, U - D_i\} + D_i] \\
&= -E[\sum_{i \in s} \max\{D_i - U, 0\}]
\end{aligned}$$

The value of B_U can be found by solving the equation:

$$U - E[\sum_{i \in s} \max\{D_i - U, 0\}] = 0$$

Let $\hat{\mu}_i$ be a robust estimate of μ_i and define $\hat{D}_i = (Y_i - \hat{\mu}_i)(\tilde{w}_i - 1)$, then the previous equation is piecewise linear with discontinuities at $U = \hat{D}_i$. By setting $\hat{D}_{(1)} \geq \hat{D}_{(2)} \geq \ldots \geq 0 \geq \ldots$ as the ordered values of $\hat{D}_i$, the distinct breakpoints of the equation can be expressed as:

$$\begin{aligned}
\psi_U\left(\hat{D}_{(k)}\right) &= \hat{D}_{(k)} - \sum_{i \in s} \max\{\hat{D}_{(i)} - \hat{D}_{(k)}, 0\} \\
&= (k+1)\hat{D}_{(k)} - \sum_{j=1}^{k} \hat{D}_{(j)}
\end{aligned}$$

Therefore, the optimal value of U can be found by solving the equation $\psi_U\left(\hat{U}\right)=0$. In general there will be no exact solution to this equation. The solution given by Chambers, Kokic, Smith and Cruddas (2000) is:

$$\hat{U}=\frac{1}{\left(k^*+1\right)}\sum_{j=1}^{k^*}\hat{D}_{(j)}$$

where k^* is the last value of k for which $\psi_U\left(\hat{D}_{(k)}\right)$ is non-negative.

An alternative solution is to use linear interpolation between $\frac{1}{(k^*+1)}\sum_{j=1}^{k^*}\hat{D}_{(j)}$ and $\frac{1}{(k^*+2)}\sum_{j=1}^{k^*+1}\hat{D}_{(j)}$:

$$\hat{U}=\frac{\psi_U\left(\hat{D}_{(k^*+1)}\right)\left[\frac{1}{(k^*+1)}\sum_{j=1}^{k^*}\hat{D}_{(j)}\right]-\psi_U\left(\hat{D}_{(k^*)}\right)\left[\frac{1}{(k^*+2)}\sum_{j=1}^{k^*+1}\hat{D}_{(j)}\right]}{\left(\psi_U\left(\hat{D}_{(k^*+1)}\right)-\psi_U\left(\hat{D}_{(k^*)}\right)\right)}$$

If there is limited data or the extreme weighted residuals differ significantly, then the solution under linear interpolation will produce a lower value of $\hat{U}$ than taking to the last positive $\psi_U\left(\hat{D}_{(k)}\right)$. Hence, the solution under linear interpolation will reduce, to some extent, the influence of individual units on the value of $\hat{U}$.

QUESTION 1: Is the solution under linear interpolation to estimate the bias parameters better than taking the last positive breakpoint?

The estimate of the lower bias parameter B_L under winsorization can be found in the same way. Let $L=-B_L$ and then the lower bias parameter can be written as:

$$\begin{aligned}B_L(L)&=E[\hat{t}_{ywinL}-t_y]\\&=-E[\sum_{i\in s}\min\{D_i-L,0\}]\end{aligned}$$

By setting $\hat{E}_{(1)}\le\hat{E}_{(2)}\le\ldots\le 0\le\ldots$ as the ordered values of $\hat{D}_i$ then the distinct breakpoints of the equation are:

$$\psi_L\left(\hat{E}_{(k)}\right) = \hat{E}_{(k)} - \sum_s \min\{\hat{E}_{(i)} - \hat{E}_{(k)}, 0\}$$

$$= (k+1)\hat{E}_{(k)} - \sum_{j=1}^{k} \hat{E}_{(j)}$$

The linear interpolation solution to the equation $\psi_L\left(\hat{L}\right) = 0$ is:

$$\hat{L} = \frac{\psi_L\left(\hat{E}_{(k^{**}+1)}\right)\left[\frac{1}{(k^{**}+1)}\sum_{j=1}^{k^{**}}\hat{E}_{(j)}\right] - \psi_L\left(\hat{E}_{(k^{**})}\right)\left[\frac{1}{(k^{**}+2)}\sum_{j=1}^{k^{**}+1}\hat{E}_{(j)}\right]}{\left(\psi_L\left(\hat{E}_{(k^{**}+1)}\right) - \psi_L\left(\hat{E}_{(k^{**})}\right)\right)}$$

where k^{**} is the last value of k for which $\psi_L\left(\hat{E}_{(k)}\right)$ is non-positive.

The upper and lower bias parameter estimates, $\hat{B}_U$ and $\hat{B}_L$, depend on the values of the top and bottom weighted residuals $\hat{D}_{(1)}, \hat{D}_{(2)}, \ldots, \hat{D}_{(k^*)}$ and $\hat{E}_{(1)}, \hat{E}_{(2)}, \ldots, \hat{E}_{(k^{**})}$. If the data used to generate the cutoff values is the same as the data for which the winsorized cutoff values are to be applied, then it is those units with the values of the top and bottom weighted residuals that will be winsorized. In this case the estimated bias, $\hat{B}_U + \hat{B}_L$, will be realised. However, if the data used to generate the cutoff values is different then it is assumed that data for which the winsorized cutoff values are to be applied fits the same model as the data used to generate the cutoff values. If this assumption holds then the realised bias should be approximately $\hat{B}_U + \hat{B}_L$.

2.2 Estimate of robust regression parameters

Suppose the generalised regression estimator is based on the model:

$$E\left(Y_i\right) = \mu_i = \underset{\sim}{\beta}^T \underset{\sim}{x}_i$$

$$Var\left(Y_i\right) = \sigma_i^2$$

where $\underset{\sim}{x}_i = \left(x_{1i}, \ldots, x_{ki}, \ldots, x_{Ki}\right)^T$ is a set of auxiliary variables for which the population totals are known. It would appear logical to apply this same model to generate $\hat{\mu}_i$ to be used as a robust estimate of the parameter μ_i, as well as in the estimation of the upper and lower bias parameters, B_U and B_L.

QUESTION 2: Is it logical to apply the same regression model used for the generalised regression estimator, to generate the regression parameters for the winsorized cutoff values? Should this same model be used to form winsorized cutoff values for all variables collected in the survey? If a regression model is known which fits another variable better, then should this regression model be used to form winsorization cutoff values for this other variable, even though it was not used for the generalised regression estimator?

There are many solutions to the problem of estimating robust regression parameters. It is worth noting that the task is not to find the best robust regression model, but rather to find a robust regression fitting procedure which results in the best performing winsorized estimator. Furthermore, since this procedure will have to be used for many ABS business surveys and understood by a range of people, it is desirable that the procedure is simple, flexible and easily traceable. An investigation was performed using a number of different robust regression fitting techniques to determine which method resulted in the best performing winsorized estimator. The techniques are listed in the following paragraphs with the results of the investigation presented in Section 2.3.

2.2.1 Trimmed least squares

The Trimmed Least Squares (TLS) technique consisted of fitting an Ordinary Least Squares (OLS) regression model to minimise the function:

$$F = \sum_{i \in s} \left(y_i - \beta^T x_i \right)^2$$

The residuals were calculated by applying the regression model back to the data used to fit the model. A percentage of the units with the largest positive and negative residuals were then removed from the data. A second regression model was then fitted to the reduced data, to estimate the robust regression parameters. The percentage of units removed and actual method of removing these units was varied in the investigation. The TLS technique has the advantage that it is extremely quick to run, simple to understand and easy to trace. Standard regression diagnostics can be generated from the fit of the regression model.

2.2.2 Trimmed least absolute value or L^1 regression technique

The Trimmed Least Absolute Value (LAV) or L^1 Regression Technique consisted of fitting a regression model to minimise the function:

$$F = \sum_{i \in s} \left| y_i - \beta^T x_i \right|$$

The residuals were calculated by applying the regression model back to the data used to fit the model. A percentage of the units with the largest positive and negative residuals were then removed from the data. A second regression model was then fitted to the reduced data, to estimate the robust regression parameters. The percentage of units removed and actual method of removing these units was varied in the investigation.

This technique is very similar to the current method used to perform winsorization in ABS business surveys using ratio estimation. The difference is that the current method truncates the data to a percentile value (10% and 90%) rather than removing the data. The LAV technique should result in a more robust regression model than the TLS technique because large residuals have less influence on the regression parameters, since the residuals are not squared.

2.2.3 Sample splitting technique

The Sample Splitting (SS) Technique consists of fitting an OLS regression model after the data has been randomly split into two halves. The 'residuals' were calculated by applying the regression model back to the half of the data not used to fit the model. The units with the largest positive and negative residuals were then removed from the data after the two halves were then merged back together. This process was repeated until a percentage of the units had been removed. The SS technique should result in a more robust regression model than the TLS technique because the residuals used to remove the 'outlier' units are not calculated from a regression model that has been generated using these 'outlier' units.

2.2.4 Least median of squares

The Least Median of Squares (LMS) technique, described by Rousseeuw and Leroy (1987), consisted of minimising the median of all sample squared residuals. The LMS regression parameters cannot be found analytically, so a resampling technique similar to the bootstrap is applied to find an approximate solution. The LMS technique is approximated by calculating the median of squared residuals of many trial regression parameters, and then selecting the regression parameters with the smallest median of squared residuals. The LMS technique should result in a more robust regression model than the TLS technique because it has the effect of fitting an OLS regression model in the absence of "outlier" units, without totally removing these 'outlier' units.

2.3 Simulation study

A simulation study was undertaken to assess the performance of these various techniques to estimate robust regression parameters, and hence estimate the cutoff values used in the winsorized estimator. The simulation study examined:

- the 'best' robust regression fitting technique (Section 3.3.1)
- the data used to calculate cutoff values under winsorization (Section 3.3.2)
- the level to calculate bias parameters under winsorization (Section 3.3.3)
- the sample weights to calculate the cutoff values under winsorization (Section 3.3.4)

The simulation study was performed using a survey population of approximately 700,000 units, based on the survey frame used for the Quarterly Economic Activity Survey (QEAS). QEAS uses a stratified random sample design with strata defined by the variables state, industry and employment size. The total sample size is approximately 16,400. The reported QEAS sales variable was used as the response variable for the study. BAS wages and BAS turnover values were merged to the frame, to be used as the auxiliary variables. For the non sampled units on the frame a QEAS sales value was generated using a regression model involving BAS wages, BAS turnover, frame employment and an error term:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{\varepsilon}_i$$

where

$\hat{y}_i$ is the predicted QEAS Sales for unit *i*

x_{1i} is the BAS Wages for unit *i*

x_{2i} is the BAS Turnover for unit *i*

x_{3i} is the Business Register Employment for unit *i*

$\hat{\alpha}$ is the intercept parameter from fitting the model on the sampled units

$\hat{\beta}_1$ is the BAS Wages parameter from fitting the model on the sampled units

$\hat{\beta}_2$ is the BAS Turnover parameter from fitting the model on the sampled units

$\hat{\beta}_3$ is the Business Register Employment parameter from fitting the model on the sampled units

$\hat{\varepsilon}_i$ is random noise for unit *i* from $N(0,\sigma_i^2)$

σ_i^2 is the variance of the predicted value for unit *i*

The regression model was fit at stratum level wherever there were sufficient sampled units. Where there were less than five responding units in a stratum the QEAS sales value was generated from a model fit at employment size level.

2.3.1 'Best' robust regression fitting technique

The simulation study consisted of selecting three independent stratified random samples to generate cutoff values under the various robust regression fitting techniques, and then applying these cutoff values to another independent stratified random sample to calculate the winsorized estimator, $\hat{t}_{ywin}$. This process was repeated a large number of times, *R*. The measures used to assess the performance of the various robust regression fitting techniques were the Mean Squared Error (MSE) and the bias:

$$MSE(\hat{t}_{ywin}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{t}_{ywin,r} - t_y)^2$$

$$Bias(\hat{t}_{ywin}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{t}_{ywin,r} - t_y)$$

where $\hat{t}_{ywin,r}$ is the winsorized estimator for the *r*-th simulated sample selected from the population, and t_y is the known population total.

The MSE and bias under the various robust regression fitting techniques were compared with the MSE and bias of the 'unwinsorized' estimator:

$$MSE(\hat{t}_y) = \frac{1}{R}\sum_{r=1}^{R}(\hat{t}_{y,r} - t_y)^2$$

$$Bias(\hat{t}_y) = \frac{1}{R}\sum_{r=1}^{R}(\hat{t}_{y,r} - t_y)$$

The percentage reduction in MSE for the various robust regression fitting techniques was calculated as:

$$MSE\ Reduction = \frac{MSE(\hat{t}_{ywin}) - MSE(\hat{t}_y)}{MSE(\hat{t}_y)} \times 100\%$$

$$Bias\ Reduction = \frac{Bias(\hat{t}_{ywin}) - Bias(\hat{t}_y)}{Bias(\hat{t}_y)} \times 100\%$$

and the methods with the largest percentage reduction in MSE and smallest percentage bias were considered the 'best'.

The regression parameters were calculated at the industry level, using an intercept term and two auxiliary variables, BAS wages and BAS turnover. For the TLS and SS methods 5% of units were removed. The regression models were fit at the industry level since it appeared logical to fit the regression model at the same level as the generalised regression model. There were major problems associated with the LAV technique when several auxiliary variables were attempted, because the matrices in the minimisation problems were frequently ill conditioned. Therefore, the results for the LAV technique are based on a single auxiliary variable, BAS Turnover. The percentage reduction in MSE and percentage bias for the various robust regression fitting techniques are presented in Table 1. 200 Repetitions (R = 200) were used to generate these results.

Table 1: Percentage reduction in MSE and bias of Winsorized estimates for various techniques

Robust Regression Fitting Techniques	*MSE Reduction (%)*	*Bias (%)*
TLS	–57.93	–0.31
LAV	–29.62	–1.14
SS	–58.46	–0.34
LMS	17.37	–1.81

The LMS technique performed quite poorly, with an increase in MSE, due to a larger bias. The explanation for this poor performance is that while this technique results in a very good model for the core of the data, it can result in a very poor model for the tails of the distribution. Since most distributions in ABS business surveys are positively skewed (i.e. large upper tails), this method is more likely to result in very large values of $\hat{D}_i$ for units in the upper tails of the distribution, where the majority of 'outlier' units are located. Hence this technique has the potential to result in very high bias parameters.

While the LAV technique did not perform as well as the TLS and SS techniques, this was primarily due to the fact that the LAV technique was based on a single auxiliary variable. Indeed, the performance of the LAV technique was similar to the TLS and SS techniques based on single auxiliary variables. The SS and TLS techniques performed very well, with the exception of several industries. The percentage reduction in MSE and percentage bias for the various robust regression fitting techniques at the industry level are presented in the Attachment.

In light of the poor performance in industries 18, 30 and 35 some further investigation was undertaken into the level of fit of the regression model, the handling of influential units and the percentage of units to be removed. While these three industries performed much better, when the regression parameters were calculated at the

stratum level, rather than the industry level, most other industries performed slightly worse. This suggests that there might be a need to fit of the regression model at different levels to the generalised regression model.

QUESTION 3: Is it logical to fit a regression model at different levels to the generalised regression model?

Generally speaking, the greater percentage of units removed from the fit of the regression model, the more robust the regression parameters. However, caution should be taken not to remove too many units, as this can lead to excessive large bias parameters. An investigation was undertaken into the impact on the MSE and bias of the various methods when the percentage of units to be removed was varied. This investigation found that the optimal percentage of units to be removed varied across industries. Furthermore, there was no conclusive evidence to suggest whether it was better to remove units based on weighted or unweighted residuals.

The investigations also found that there were some industries where the regression model fitted to the three historical samples was very different from the regression model fit to the current data. This indicates that diagnostics should be incorporated into the regression fitting procedure before using the regression parameters to generate cutoff values. In particular, checking that the regression model fits the data well, checking that units with large influences are removed from the regression model, and checking that the regression model fits to current data to be winsorized.

QUESTION 4: What is the best way to ensure the regression model fitted to the historical data is still applicable to the current data to be winsorized? What action should be taken when the regression model does not fit to the current data to be winsorized?

2.3.2 Data used to calculate cutoff values

The current practice for ABS business surveys is to use several cycles of historical data to estimate the regression parameters and bias parameter, and assume that the same regression model holds for the current data. This practice was based on work by Clark (1995) who suggested that the quality of parameters can be improved by using more data. The results of the simulation study support these findings, showing a

much tighter distribution of the bias parameter when more cycles of the data are used.

An exception to this practice is the monthly Retail Trade Survey which is affected by seasonal variation. In this situation it is more effective to use a single cycle of historical data, the corresponding month in the previous year, to estimate the parameters. There is a need for further investigation into weighting the various cycles of historical data to maximise the stability of parameters over time, in order to minimise the impact on the movement estimates.

Several ABS business surveys have experienced problems with individual units with large historical values that adversely influence the estimate of the bias parameter. It can be seen that if the largest weighted residual, $\hat{D}_{(1)}$, is more than double the second largest weighted residual, $\hat{D}_{(2)}$, then the bias parameter calculation algorithm will stop after the second breakpoint, $\psi_U\left(\hat{D}_{(2)}\right) = 2\hat{D}_{(2)} - \hat{D}_{(1)} < 0$. The estimate of the bias parameter generated will be an interpolation between $\frac{\hat{D}_{(1)}}{2}$ and $\frac{\hat{D}_{(1)} + \hat{D}_{(2)}}{3}$. If the unit with the largest weighted residual was not present or was smaller then estimate may be quite different.

Another problem with these large historical values is that they can make the bias parameters unstable over time, and hence result in large impacts on movement estimates. The current practice is to remove units with large historical values which have an adverse influence on the estimate of the bias parameter. This is justified by assuming the units come from a different population entirely and so including them does not add any information about the tail of the distribution of interest. These units should be removed with caution however, otherwise it could lead to cutoff values that are too small.

QUESTION 5: What is the best way to deal with units with large historical values which have an adverse influence on the estimate of the bias parameter?

2.3.3 Level of calculate bias parameters

The level at which the bias parameters, B_U and B_L, are calculated will determine the performance of estimates at the various levels. If the bias parameters are calculated at a broad level (e.g. Australia or industry levels), then these broad level estimates should perform well, but the finer level estimates may have large variances, as 'outlier' units may be undetected at these levels. On the other hand, if the bias parameters are

calculated at finer levels, then these fine level estimates should perform well, but the broad level estimates may exhibit large biases, as too many units may be winsorized at these levels.

The solution suggested by Chambers, Kokic, Smith and Cruddas (2000), and current implemented for ABS business surveys, is to calculate the bias parameters at the most important level. This treatment could well produce poor quality estimates of at the finer levels. In practice any further units that adversely affect the estimates at the finer levels have usually been made surprise outliers (i.e. had their weight set to one) to overcome this problem. It is expected that some surprise outliering will always be required regardless of the winsorisation methodology, although a compromise solution, to calculate the bias parameters at an artificial intermediate level, has promise for reducing the extent of surprise outliering required.

An alternative approach suggested to this problem is to calculate bias parameters at broad and fine levels. The fine level estimates would then be modified using rescaling factors such that they are consistent with the broad level estimates. Although this method has its merits, it has several disadvantages. Firstly, under this approach either the unit record data will no longer add to published estimates; or the rescaling factors will need to be applied to all units in the survey. Secondly, this approach will become very complex where there are large number of data items or relationships between the data items.

2.3.4 Sample weights used to calculate cutoff values

The estimates of regression parameters, $\hat{\mu}_i$, and bias parameters, $\hat{B}_U$ and $\hat{B}_L$, are usually generated from historical data and hence are treated as independent from the current data. However, the cutoff values do depend on the current data through the sample weights, $\tilde{w}_i$ (i.e. generalised regression weights under generalised regression estimation). The use of generalised regression weights to calculate the cutoff values means that the generalised regression weights need to be available to perform winsorization. On the other hand, the use of design weights to calculate the cutoff values has the advantage that the cutoff values can be generated in advance of the generalised regression weights to allow sufficient time for quality checking. Furthermore, the use of design weights also simplifies variance estimation under the bootstrap methodology, since the same winsorized values can be used in all replicate samples, rather than being calculated separately for each replicate sample. Therefore, an investigation was undertaken into the performance of the winsorized estimator using the design weight, $w_i = 1/\pi_i$, in place of the generalised regression weight, μ_i, to generate the cutoff values.

The investigation used three independent samples to generate cutoff values based on TLS technique, with 5% of units removed. The regression parameters were calculated

at the industry level, using two auxiliary variables, BAS wages and BAS turnover. These cutoff values were then applied to a fourth independent sample. This process was repeated 200 times. The relative differences between the winsorized estimates when cutoff values are calculated using the generalised regression weights and design weights are presented in Table 2.

Table 2: Relative difference between Winsorized estimates using generalised regression weight and design weight

Difference between Winsorized Estimates	*Percentage of Units*
0–0.5 %	88.5
0.5–1 %	9.0
1–3 %	2.5
3–5 %	0.0
5–10 %	0.0
> 10 %	0.0

At the Australia level, 88.5% of independent samples resulted in less than 0.5% difference between the winsorized estimates using the generalised regression weights and design weights. At the industry level, most estimates differed by less than 3.0%. Those industries with the larger differences (i.e. Industries 18, 26, 30 and 35) have already been identified as having questionable regression models throughout the simulation study. The relative differences between the winsorized estimates using the generalised regression weights and design weights at the industry level are presented in Attachment 1.

QUESTION 6: Is there any reason why the design weights should not be used in the calculation of cutoff values used in the winsorized estimator?

3. LINEAR RELATED ITEMS

Most ABS business surveys collect and derive a wide range of data items. Furthermore, there often exists linear relationships between these data items. For example, suppose a survey collects a set of variables, $y_1, y_2, ..., y_K$, and a derived variable, y_0, is calculated as a linear combination of these variables, $y_0 = \sum_{k=1}^{K} a_k y_k$. In this situation there exists the following linear relationship between the variables, $\sum_{k=0}^{K} a_k y_{ki} = 0$, where $a_0 = -1$. It is important that these linear relationships still hold after winsorization. Therefore, an important issue for winsorization is to develop a method to maintain the linear relationships between the variables.

In theory, winsorization can be applied to all the survey variables. However, in most cases the linear relationships between these variables will no longer hold after winsorization. The current ABS estimation system allows the linear relationships to be maintained by either:

- winsorizing the set of component variables, $y_1^*, y_2^*, ..., y_K^*$, and then calculating the 'winsorized' derived variable based on these winsorized component variables, $y_{0i}^* = \sum_{k=1}^{K} a_k y_{ki}^*$; or
- winsorizing the derived variable, y_{0i}^*, and then calculating the 'winsorized' set of component variables by applying the same proportional adjustment, $y_{ki}^* = \frac{y_{0i}^*}{y_{0i}} y_{ki}$.

Unfortunately, there are some situations where these two methods perform quite poorly. The major problem with the first of these methods is that it could result in very poor values for the 'winsorized' derived variables, in particular where some of the a_k are negative. For example, suppose profits y_0 is derived based on total income y_1 minus total expenses y_2 plus opening stocks y_3 minus closing stocks y_4 (i.e. $y_0 = y_1 - y_2 + y_3 - y_4$). Suppose a unit has an unusually large total expenses, but total income, opening stocks and closing stocks are not unusual. Furthermore, suppose the derived profit for the unit is negative. Using the first of the methods under the current ABS estimation system, the 'winsorized' derived profit for the unit could easily end up positive, since only total expenses is winsorized. This treatment could well have a detrimental impact on the sign of the estimates of profit from the survey.

The major problem with the second of these methods is that it could result in very poor values for the 'winsorized' set of component variables, in particular where some of the components of a derived variable are usually much smaller than other

components. For example, suppose total income is derived based on a number of variables, including sales income (generally a large component of total income) and royalties income (generally a small component of total income). However, suppose a unit has an unusually large royalties income, but total income is not unusual. Using the second of the methods under the current ABS system, the royalties for the unit will not be winsorized, since total income is not winsorized. This treatment could well produce poor quality estimates of royalties from the survey. In practice these units have usually been made surprise outliers (i.e. had their weight set to one) to overcome this problem. Another problem with this method is that it can be quite cumbersome to maintain multiple linear relationship between the variables.

Chambers, Kokic, Smith and Cruddas (2000) suggested an alternative to the second method, which distributes the difference between the original and winsorized derived variable amongst the largest component variables. This method is based on the principle that an outlier on the derived variables will generally be due to one or several of the component variables being unusually large, rather than all the component variables. Let $y_{(1)} \geq y_{(2)} \geq \ldots \geq y_{(K)}$ denote the ordered set of component variables with coefficients $a_{(1)}, a_{(2)}, \ldots, a_{(K)}$, then the 'winsorized' set of component variables are computed using the equation:

$$y^*_{ki} = \frac{\sum_{k=1}^{j^*} a_{(k)} y_{(k)i} + y^*_{0i} - y_{0i}}{\sum_{k=1}^{j^*} a_{(k)}}$$

where j^* is the largest value of j for which

$$\psi_{ji} = y_{0i} - y^*_{0i} - \sum_{k=1}^{j} a_{(k)} y_{(k)i} + y_{(k)i} \sum_{k=1}^{j} a_{(k)} \geq 0$$

It should be noted this method cannot be applied in its current form in the situation where some of the a_k are negative. However, it is relatively simple to modify this method to be appropriate for this situation, and where there are two-sided winsorized cutoff values. While this method has its merits, it suffers from the same problems as the second of the methods under the current ABS estimation system.

Another alternative method which attempts to overcome the shortcomings of the two methods under the current ABS estimation system is to winsorize all the survey variables and then modify these winsorized values, so that the linear relationships between these variables still hold, by a process known as calibration. A new set of winsorized values for variable k for unit i, y^{**}_{ki}, are sought which lie as close as possible

to the set of original winsorized values, y^{*}_{ki}. The calibration requires the specification of distance function between the original and final winsorized values.

Although any one of a number of distance functions could be used, one of the most commonly used is the generalised least squares distance function:

$$D_i = \sum_{k=0}^{K} c_k \frac{\left(y^{**}_{ki} - y^{*}_{ki}\right)^2}{\left|y^{*}_{ki}\right|}$$

where c_k are specified positive factors that control the relative importance of the variables.

QUESTION 7: Is the concept of minimising a distance function to ensure linear relationships between variables still holds after winsorization appropriate? Is the generalised least squares distance function appropriate?

Minimisation of the generalised least squares distance function using Lagrange multipliers, subject to satisfying the linear relationship constraint, $\sum_{k=0}^{K} a_k y^{**}_{ki} = 0$, leads to the final winsorized values:

$$y^{**}_{ki} = \begin{cases} y^{*}_{ki}\left(1 - \dfrac{\left(\dfrac{a_k}{c_k}\right)\sum_{k=0}^{K} a_k y^{*}_{ki}}{\sum_{k=0}^{K} \dfrac{a_k^2 \left|y^{*}_{ki}\right|}{c_k}}\right) & \text{,if } y^{*}_{ki} \geq 0 \\ y^{*}_{ki}\left(1 + \dfrac{\left(\dfrac{a_k}{c_k}\right)\sum_{k=0}^{K} a_k y^{*}_{ki}}{\sum_{k=0}^{K} \dfrac{a_k^2 \left|y^{*}_{ki}\right|}{c_k}}\right) & \text{,if } y^{*}_{ki} < 0 \end{cases}$$

This method can easily be extended to multiple linear relationship constraints (i.e. $\sum_{k=0}^{K} a_k y^{**}_{ki} = 0$). One disadvantage of this method is that some of the final winsorized values can be negative for variables which should always be positive (and vice versa). This problem can be overcome by imposing range restrictions on the final winsorized values, $L \leq y^{**}_{ki} \leq U$, where L and U are suitable lower and upper bounds. In order to

satisfy the linear relationship constraints and the range restrictions, the calculation of the final winsorized values needs to be undertaken using an iterative method.

The first of the methods under the current ABS estimation system is a special case of this alternative method. If the factors for the set of component variables are all set to infinity (i.e. $c_1 = c_2 = \ldots = c_K = \infty$) and $a_0 = -1$ then the final winsorized values simplify to:

$$y_{ki}^{**} = \begin{cases} y_{ki}^{*} & \text{,for } k = 1, 2, \ldots, K \\ \sum_{k=1}^{K} a_k y_{ki}^{*} & \text{,for } k = 0 \end{cases}$$

which is equivalent to the first method under the current ABS estimation system. On the other hand, the second of the methods under the current ABS estimation system is not a special case of this alternative method.

4. CONCLUSION

The effectiveness of winsorization at robustifying the GREG estimator against unusually large residuals will ultimately depend on the choice of the cutoff values, and hence the methods used to estimate the bias parameters and regression parameters. This paper has investigated several techniques for fitting robust regression models and found that the winsorized estimator performs best under models that are only moderately robust. Some conceptual questions have been raised about the link that should exist between the GREG model and the model used to estimate regression parameters for cutoffs. Current thinking is that the models should involve the same auxillary variables and be fitted at the same level, however the simulation study found cases where different models improved performance.

One of the key findings of the study was that diagnostics should be incorporated into the regression fitting procedure before using the regression parameters to generate cutoff values. In particular, checking that the regression model fits the data well, checking that units with large influences are removed from the regression model, and checking that the regression model fits to current data to be winsorized.

The winsorization methodology described in this paper works well for point in time estimates. Further work to assess the performance of winsorization on GREG estimates of movement between time points may prove useful. Movement estimates are a key output for many ABS business surveys. It has been noted that large historical values can make bias parameters unstable over time and hence impact on movement estimates. Further work to determine the best way of dealing with these values and the best way of weighting various cycles of historical data to maximise stability of parameters over time is recommended.

Linear relationships between data items is another area warranting further investigation. Shortcomings of the current ABS estimation system's ability to handle linear relationships between survey variables have been discussed and an alternative method presented. The alternative method requires specification of a distance function between original and final winsorized values. Investigation is planned into the performance of this method and the suitability of the generalised least squares distance function.

The simulation study presented here independently replicated the process of selecting historical samples, estimating cutoffs and applying cutoffs to independent samples. A large number of replicates were generated to produce estimates of the MSE and bias of the winsorized estimator. In practice only a single set of cutoffs is generated, based on historical data, and used to winsorize the present sample. This introduces a source of variability which is not reflected in current variance estimates and which would be difficult to incorporate. The simulation study found cutoffs to be

more stable when several cycles of historical data was used to estimate parameters and this is the approach that will be implemented for GREG estimation. Further work could involve using the data from the simulation study to quantify the significance of the variability introduced through estimating cutoffs.

BIBLIOGRAPHY

Chambers, R., Kokic, P., Smith, P. and Cruddas, M. (2000) "Winsorization for Identifying and Treating Outliers in Business Surveys", *Proceedings of the Second International Conference on Establishment Surveys* (ICES II), pp. 687–696.

Clark, R.G. (1995) *Winsorization Methods in Sample Surveys*, Masters thesis, Department of Statistics, Australian National University.

Gross, W.F., Bode, G., Taylor, J.M. and Lloyd–Smith, C.W. (1986) "Some finite population estimators which reduce the contribution of outliers", *Proceedings of the Pacific Statistical Conference*, Auckland, New Zealand, 20–24 May 1985.

Hidiroglou, M.H. and Srinath, K.P. (1981) "Some estimators of population total from simple random samples containing large units", *Journal of the American Statistical Association*, 76, pp. 690–695.

Kokic, P.N. and Bell, P.A. (1994) "Optimal winsorizing cutoffs for a stratified finite population estimator", *Journal of Official Statistics*, 10, pp. 419–435.

Rousseeuw, P.J., and Leroy, P.M. (1987) *Robust Regression and Outlier Detection*, John Wiley & Sons.

Sarndal, C.–E., Swensson, B. and Wretman, J.H. (1992) *Model Assisted Survey Sampling*, Springer–Verlag.

Searls, D.T. (1966) "An estimator which reduces large true observations", *Journal of the American Statistical Association*, 61, pp. 1200–1204.

ATTACHMENT

Table 1: Percentage reduction in MSE and bias of Winsorized estimates for various methods

	MSE Reduction (%)				*Bias (%)*			
Industry	*TLS Method*	*LAV Method*	*SS Method*	*LMS Method*	*TLS Method*	*LAV Method*	*SS Method*	*LMS Method*
01	–19.32	–17.65	–19.33	–24.78	–0.06	0.11	–0.07	–0.46
03	–49.77	–49.08	–50.33	–36.57	0.52	0.61	0.49	–0.80
04	–57.41	–45.50	–58.62	–34.04	0.69	–1.63	0.60	–2.63
05	–52.86	–44.71	–57.15	–22.35	–2.21	–3.04	–1.96	–4.23
06	–50.97	–39.52	–50.82	–38.69	–0.03	–1.99	–0.04	–2.03
07	–42.57	–15.18	–39.81	–41.43	0.43	1.31	0.56	–0.46
08	–40.60	–30.51	–40.78	–24.60	–1.97	–2.76	–2.06	–3.18
10	–25.09	–22.78	–25.73	–17.14	0.39	0.62	0.54	–0.11
11	–65.66	–54.16	–65.54	–58.78	–0.60	–2.21	–0.66	–1.91
12	–18.13	–13.92	–18.94	–9.93	–1.15	–3.00	–1.31	–4.32
13	–66.00	–66.60	–65.61	–66.13	1.61	0.92	1.88	0.42
14	–27.38	–20.14	–26.85	–17.50	–0.01	–1.14	–0.44	–1.32
15	–27.75	–25.68	–27.69	–19.81	–0.24	–0.57	–0.26	–1.32
16	–39.63	–38.51	–41.01	–36.18	0.21	–0.40	0.11	–0.88
17	–51.01	–49.59	–50.19	–42.16	–0.22	–0.74	–0.05	–1.51
18	–50.01	–37.46	–53.97	–35.17	–8.13	–9.98	–7.53	–10.30
19	–44.12	–44.97	–44.77	–41.79	–0.14	–0.36	–0.15	–0.66
22	–32.63	–33.70	–32.25	–35.42	1.12	1.03	1.70	0.67
25	–43.36	74.68	–42.89	–34.17	–0.32	0.97	–0.32	–0.68
26	–39.89	–31.24	–40.04	–34.28	–0.79	–4.75	–0.70	–5.78
27	–49.20	–52.68	–49.54	–53.14	2.41	–0.16	2.61	–0.93
28	–46.59	–51.61	–43.87	–55.91	2.39	1.71	2.71	0.43
30	–19.45	–41.04	–22.01	–39.80	15.16	9.96	15.46	10.24
31	0	0	0	0	0	0	0	0
33	–46.81	–45.08	–47.37	–45.02	–0.18	–0.53	–0.16	–0.78
35	–37.43	–48.07	–34.63	–47.37	8.34	4.75	9.04	3.73
36	–56.52	–46.90	–56.67	–37.20	–6.73	–9.42	–6.83	–11.49
37	–46.08	–45.40	–46.33	–41.76	0.79	–0.59	0.58	–1.40
38	–32.17	–28.74	–33.81	–23.74	–2.63	–3.20	–2.55	–3.95
39	–29.47	–26.41	–31.06	–22.09	0.74	–0.02	0.60	–0.77
40	–77.02	–75.52	–77.40	–75.96	–0.78	–2.49	–0.67	–2.44
41	–67.76	–65.51	–67.93	–61.23	1.86	–0.49	2.40	–1.85
42	–41.23	–61.01	–67.65	–53.78	2.27	–4.32	–1.53	–5.42
43	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0
76	–19.08	–19.96	–19.05	–19.16	–0.43	–0.48	–0.45	–0.80
Total	–57.93	–29.62	–58.46	17.37	–0.31	–1.14	–0.34	–1.81

Table 2: Relative difference between Winsorized estimates using generalised regression weight and design weight

	Difference between Winsorized estimates					
Industry	*0–0.5 %*	*0.5–1 %*	*1–3 %*	*3–5 %*	*5–10 %*	*> 10 %*
01	83.5	14.5	2.0	0	0	0
03	91.0	5.5	3.5	0	0	0
04	64.0	16.0	17.5	2.5	0	0
05	70.5	18.5	10.5	0.5	0	0
06	44.5	23.5	23.0	8.0	1.0	0
07	74.0	11.5	13.5	1.0	0	0
08	94.0	5.5	0.5	0	0	0
10	88.0	10.0	2.0	0	0	0
11	77.0	14.5	8.0	0.5	0	0
12	20.0	22.0	39.5	14.5	3.5	0.5
13	50.5	19.5	27.5	2.5	0	0
14	70.0	18.0	11.5	0.5	0	0
15	81.5	15.5	3.0	0	0	0
16	77.5	16.0	6.5	0	0	0
17	74.5	16.0	9.0	0.5	0	0
18	34.5	12.0	30.5	10.5	8.5	4.0
19	84.5	12.5	3.0	0	0	0
22	76.5	17.0	6.0	0.5	0	0
25	93.0	6.0	1.0	0	0	0
26	23.5	13.0	33.0	12.0	16.5	2.0
27	30.0	19.5	30.5	11.0	6.0	3.0
28	36.5	20.0	34.5	5.5	2.0	1.5
30	9.5	6.5	35.0	16.5	21.0	11.5
31	100.0	0	0	0	0	0
33	91.0	7.0	2.0	0	0	0
35	20.5	13.0	31.0	18.5	12.0	5.0
36	38.5	24.5	27.5	7.0	2.0	0.5
37	64.5	14.5	17.5	3.5	0	0
38	77.0	16.0	6.0	1.0	0	0
39	67.5	15.5	15.0	2.0	0	0
40	44.0	18.5	29.5	3.5	2.5	2.0
41	45.5	17.5	26.5	8.5	2.0	0
42	35.5	12.0	39.0	10.5	2.5	0.5
43	100.0	0	0	0	0	0
44	100.0	0	0	0	0	0
45	100.0	0	0	0	0	0
46	100.0	0	0	0	0	0
47	100.0	0	0	0	0	0
48	100.0	0	0	0	0	0
76	84.0	12.0	4.0	0	0	0
Total	88.5	9.0	2.5	0	0	0

FOR MORE INFORMATION . . .

INTERNET	**www.abs.gov.au** the ABS web site is the best place for data from our publications and information about the ABS.
LIBRARY	A range of ABS publications are available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our web site for a list of libraries.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our web site, or purchase a hard copy publication. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE	1300 135 070
EMAIL	client.services@abs.gov.au
FAX	1300 135 211
POST	Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All ABS statistics can be downloaded free of charge from the ABS web site.

WEB ADDRESS	www.abs.gov.au

2000001524381
ISBN 0 642 48158 X

RRP $11.00