1352.0.55.078

Australian Bureau of Statistics

Research Paper

Assessing the Quality of Modelled Estimates

www.abs.gov.au

New Issue

Research Paper

Assessing the Quality of Modelled Estimates

Laurie Nitschke, Shiji Zhao and Lewis Conn

Analytical Services Branch

Methodology Advisory Committee

23 June 2006, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THURS 6 SEP 2007

ABS Catalogue no. 1352.0.55.078
ISBN 978 0 64248 368 3

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Dr Shiji Zhao, Analytical Services Branch on Canberra (02) 6252 6053 or email <analytical.services@abs.gov.au>.

CONTENTS

ABSTRACT 1

1. INTRODUCTION 1

2. WHAT ARE MODELLED ESTIMATES AND WHY DO WE NEED THEM? 3
   - 2.1 What are 'modelled' statistics? 3
   - 2.2 Why 'modelled' statistics? 4

3. AN EXAMPLE – ESTIMATING EMPLOYEE NUMBERS 5
   - 3.1 A conceptual framework 5
   - 3.2 Data description and analysis 7
   - 3.3 Data treatment 10
   - 3.4 Model specification 12
   - 3.5 Estimation results 16
   - 3.6 Subsequent analysis 19

4. ASSESSING THE QUALITY OF THE EMPLOYEE NUMBERS MODEL AND ITS OUTPUTS 21
   - 4.1 A quality assessment framework 21
   - 4.2 Consistency 23
   - 4.3 Plausibility 25
   - 4.4 Robustness 28

5. CONCLUDING REMARKS 30

ACKNOWLEDGEMENTS 31

REFERENCES 32

APPENDIXES

A. DESCRIPTION OF ANZSIC SUBDIVISIONS 33

B. PARAMETER ESTIMATES, 2003 34

ASSESSING THE QUALITY OF MODELLED ESTIMATES

Laurie Nitschke, Shiji Zhao and Lewis Conn
Analytical Services Branch

ABSTRACT

In recent years, official statistical agencies have increasingly used analytical methods to compile statistics. For example, econometric models are sometimes applied to statistical processes or used to solve certain statistical problems. Modelling techniques have enabled statisticians to meet the demand for statistics that otherwise would be too costly or difficult to produce.

This paper proposes a framework that may be used to assess the quality of statistics that are generated using econometric models. We applied the framework to a recent project that is intended to estimate business employee numbers based on the Economic Activity Survey (EAS) and Business Income Tax (BIT) data.

An earlier version of this paper was presented to the Methodology Advisory Committee (MAC) in June 2006. Subsequent to the MAC meeting, further analysis of the model specifications and underlying assumptions was undertaken. The revised model has been used in the compilation of the employment statistics published in *Australian Industry* (ABS cat. no. 8155.0), *Manufacturing Industry, Australia* (ABS cat. no. 8221.0) and *Mining Industry, Australia* (ABS cat. no. 8415.0).

1. INTRODUCTION

Modelled estimates are becoming an increasing part of ABS statistics. Uses of these modelled estimates include adjusting data to determine variables, estimating values for missing variables and data for missing time periods.

The focus of this paper is how to assess the quality of these modelled estimates. To illustrate the issue, the estimation of employee numbers from the Economic Activity Survey (EAS) and Business Income Tax (BIT) data will be used as an example. The quality assessment of the employee estimates will then be used as a starting point to quality assess other modelled estimates.

We believe this is an interesting topic for the members of the Methodology Advisory Committee (MAC) as modelled statistics involve diverse requirements (e.g. filling gaps in variables, observations, frequency of collections, smaller geographic and other domains), data (ABS survey data, administrative or business data etc.) and techniques

(i.e. models used, etc.), yet we do not have a standard approved framework to assess the quality of the outputs.

Therefore we would like to explore ways for the ABS to develop a 'framework' to assess the quality of these estimates.

We are keen to hear the comments of the MAC members on the work presented in this paper. We are particularly interested in the views on two broadly defined areas. The first is specifically about the quality of our estimation of employee numbers from the EAS and BIT data. For example, MAC members may wish to comment on

- the model specifications and estimation procedure;
- the economic and statistical assumptions;
- the method we use to clean and analyse the data; and
- estimation results and findings.
- How do we package the quality assessment information so that it is relevant and useful for users? The information supplied should inform users of the quality and the limitations of the results.

The second area concerns the framework that we have applied to assess the quality of the estimation. For example, we are interested in

- how to improve the quality assessment framework;
- whether we have appropriately applied the framework to this project; and
- what do we need to take into account, if we are to generalise and apply it to assess other modelled statistics?

Section 2 introduces modelled estimates and why we need them. Section 3 outlines the issues and findings from the project of estimating employee numbers. Section 4 applies a quality assessment framework to the work on employee numbers and Section 5 concludes.

2. WHAT ARE MODELLED ESTIMATES AND WHY DO WE NEED THEM?

2.1 What are 'modelled' statistics?

'Model' is a frequently used word among many professionals including mathematicians, statisticians and economists. However, this word can be interpreted very differently even within the same group of professionals. According to *A Dictionary of Statistical Terms*,

> "A model is a formalised expression of a theory or the causal situation which is regarded as having generated observed data." (Marriott, 1991, p. 132)

Based on this definition, it may be argued that almost all statistics involve a certain degree of modelling at various stages of data generation process (e.g. sample design, data collection and cleaning, aggregation and even presentation). Therefore, nearly all statistics are 'model-based'.

This paper focuses on statistics that are generated through a particular kind of modelling processes or techniques. We do not intend to provide a definition of modelled statistics as such and, indeed, it is not the purpose of this paper. However, a description of their broad characteristics will be helpful.

Generally speaking, modelled statistics are usually derived from one or more processes of mathematical or statistical transformations. The transformations may vary from sophisticated methods, such as regression or other statistical techniques, to simple techniques such as prorating. In practice, survey, business or administrative data are used as inputs in the process of generating this kind of statistics.

Modelled statistics are increasing in importance among official statistical offices. For example, many national statistical agencies have applied hedonic methods to the compilation of price indexes of computer hardware. Hedonic methods are a regression-based technique.

In recent years, developing methods for modelled statistics has been an important theme of analysis in the Analytical Services Branch (ASB). For example, various modelling techniques have been used in a number of projects that include, for example,

- small area estimation;
- measuring quality-adjusted labour inputs;
- hedonic methods for house price index;
- allocating social transfers-in-kind in the fiscal incidence study; and
- price indexes for computer software.

The seasonally adjusted and trend statistics and the project to be discussed in this paper are also good examples of modelled statistics.

2.2 Why 'modelled' statistics?

There are several reasons for modelled statistics to become more popular among official statistical agencies. A main driver for using models to produce statistics is the increased availability of administrative data. Because these data are usually collected through administrative processes and varying degrees of modification and transformation are necessary before they can be used to produce meaningful statistics.

Modelled statistics can reduce costs and the load on providers when they are used in place of sample surveys. However the use of model-dependent methods to replace surveys is reliant on the availability of suitable input data to produce the estimates.

Modelled estimates are often an efficient and accurate way to quality adjust the price of high technology goods. Many high technology goods have frequently changing attributes which makes pricing them difficult. Traditional pricing methods require the quality of the product to remain constant and thus are not suitable for these type of products. Model-dependent methods such as hedonics allow for changing quality and provide a systematic way of producing 'pure' price change.

Model-dependent estimation is able to assist in the measurement of products that are not clearly defined or standardised, this would include computer software and bundled goods and services. Model-dependent methods could be useful in defining mobile phone prices when they are bundled with other services.

Another use of model-dependent methods is to adjust data when the data do not meet statistical compilation requirements. An example of this would be adjustments made to the sales data for use in the house price indexes. In many circumstances, hedonic methods provide a powerful tool to control for the heterogeneity in the characteristics of houses and increase the comparability of samples collected at different points in time.

3. AN EXAMPLE – ESTIMATING EMPLOYEE NUMBERS

To illustrate the issues surrounding modelled estimates, this paper will examine a project which involved estimating employee numbers. The aim of this project is to produce an estimate of employment for each ANZSIC industry subdivision.

The need for these estimates is a result of a reduction in the amount of survey data and the lack of a employee variable on the administrative data. In previous years employee numbers have been estimated through the use of sample surveys (i.e. The ABS Economic Activity Survey). However the amount of data produced through direct collection has been reduced and supplemented with administrative data from the Australian Taxation Office (ATO). The administrative data used is the Business Income Tax (BIT) data and does not contain direct information regarding employee numbers. This project explores methods of using the two datasets jointly to produce estimates for the number of employees.

The process is to look at the survey data still collected to form a relationship between employee numbers and possible explanatory variables. We then apply this relationship to the appropriate variables on the BIT data set to produce an estimate of employee numbers.

3.1 A conceptual framework

Broadly speaking, it involves two steps to derive estimates of the number of employees using the BIT data. First, we run a regression based on the data from the Economic Activity Survey (EAS),

$$N_{EAS} = f(W_{EAS}, Z_{EAS}) \tag{1}$$

where N is for number of employees, W for wages and salaries and Z for characteristics of the firm such as size, the industry the firm belongs to and legal identifies etc..

In the second step, the coefficients obtained from equation (1) are used to 'predict' the numbers of employees based on the BIT data.

$$\hat{N}_{BIT} = f(w_{BIT}, z_{BIT}) \tag{2}$$

where w and z are wages and salaries and the firm characteristics from the BIT data. The $\widehat{N}$ are the predicted values for the number of employees.

The key assumption underlying this method is that there exists a relationship between the number of employees and total wages and salaries among business firms and this relationship is consistently represented in both the EAS and BIT data. In reality, the

relationship between the number of employees and wages and salaries does vary between firms. Then the main question is how to reconcile the model and the reality, and how to control for the variations between firms in order to produce a set of meaningful estimates for the number of employees at firm level data that can then be aggregated to industry level.

In theory, it may be argued that two identical firms (that produce identical products, have the same management and organisational structure, use the same technology and operate in the same markets for inputs and outputs), should employ a similar labour force and pay similar wages and salaries. In this hypothetical situation, the model is expected to work well. However, in reality, the relationship between the number of employees and wages and salaries differs for many reasons. Here are some examples.

Firms usually produce different products and, as a result, they use different technologies – in the sense of both the engineering technology and the ratio of labour and capital – and operate in different product markets and legislative environments. Because of these reasons, their labour employment are expected to differ in, for example, the composition of skilled/non-skilled or part-time/full-time labour. It is unrealistic to perfectly control for these kinds of differences between firms. However, we may be able to control for this type of 'heterogeneity' among the firms to a certain degree, by including industry classification (such as ANZSIC) as an independent variable (i.e. the Z) in the regression model.

Even when two firms produce identical products and employ equal numbers of employees, they may still have very different production processes and organisational structures and use different technologies. As a result, they may require a different composition of labour force (e.g. skilled and unskilled workers) and they are unlikely to pay exactly the same wage and salaries. It is very difficult to capture these kinds of differences.

Like human beings, business firms also have their life cycles. At different stages of the life cycle, firms that produce identical products and operate in similar markets and legislative environments may differ considerably in size and organisational structure. These differences may also be reflected in the composition of labour and result in different relationships between the number of employees and wages and salaries.

Both the EAS and BIT data cover firms from all eight Australian States and Territories. Although it may be argued that the labour market in Australia is reasonably competitive (meaning that wages and salaries are similar for identical jobs or persons with similar skills and experience), it is not realistic to assume that the pay and conditions are exactly the same between the States and Territories or that they will remain constant over time. This means that the relationship between employee

numbers and wages and salaries may vary between States and Territories and over time.

These are just a few examples and, in reality, there are almost countless factors – economic, demographic, social, locational and physical – that may cause a difference in the relationship between employee numbers and wages and salaries for two business firms. In this project, we use a set of variables (i.e. Z in equations (1) and (2)) trying to control for these differences and minimise their impact on the resulting estimates.

However, it is unrealistic for any econometric techniques to control for all the differences among the firms and fit a 'perfect' relation. This is because there are too many factors that have an influence and their impacts are too complex to be modelled using a simple equation. In addition, imperfections in the data, from both EAS and BIT, will have an impact on the results. More details will be given in Sections 3.2 and 3.3 about the data problems and how we address these issues. In this project we hope that, by using econometric techniques, we will be able to obtain a set of estimates of reasonable quality. An assessment of the quality of our estimates is presented in Section 4.

3.2 Data description and analysis

This section provides an overview of the data including data analysis and treatments applied to the data.

To form a model between employee numbers and possible explanatory variables, data from the EAS were used. The EAS data is an annual survey and years 2002 to 2004 were used in this investigation. This survey includes all the large and complex businesses and a sample of smaller businesses and it was supplemented with Income Tax Survey data.

The EAS data set contains over 300 variables, including an employment variable. For each year there are approximately 20,000 observations, with each observation referring to a single business/entity. EAS data are available for the years pre-2002, but due to framework changes, the data are not easily comparable.

The variables in the EAS data set were examined to determine as to which variables could be used as Z in equations (1 and 2) to provide explanatory information regarding employee numbers. The variables considered need to exist in both the EAS and BIT data sets. The variables must be in the EAS data set to form the model, and they must exist in the BIT data set so that the coefficients of the model can be applied to the prediction of employee numbers. Table 3.1 lists variables that exist in both data sets and could have some possible explanatory power.

Previous investigations into the modelling of employee numbers have found that *Wages & salaries* provide the most explanatory power out of the available variables. The *ANZSIC*, *TOLO* and *Total income* variables will be considered as they may provide additional information.

3.1 Explanatory variables available in both the EAS and BIT data sets

Explanatory variables	*Description*
Wages & salaries	Total wages paid in that year
ANZSIC	Industry classification
TOLO	Type of legal entity
Total income	Income in that year
Not for profit	Variable indicating if the business is 'not for profit'
Total assets	Total assets held by the business
Total expenses	Total expenses in that year

Preliminary data analysis

Initial analysis of the data examined the quality of the *Wages & salaries* and *Employee numbers* variables. The *Wages & salaries* variable should have the strongest relationship with *Employee numbers*, so an understanding of the quality of these variables will be useful to the analysis.

A limitation of the EAS data is that the employee numbers variable does not distinguish between full and part time employment. This limitation makes it impossible for us to standardise the variable of employee numbers and may potentially 'weaken' the relationships between wages and employees or make it difficult to interpret the results.

Table 3.2 is provided to give an indication of the number and size of the *Wages & salaries* and *Employee numbers* variables in each ANZSIC subdivision of the 2004 EAS data.

As a result of the assumption that the relationship between *Employee numbers* and *Wages & salaries* may be different for different industries, the data have been split by ANZSIC sub-division. This will help us to identify any industries where the data may not be reliable.

Unreliable data may be the result of

- numbers of observations in that subdivision being too small for our purpose;
- a high proportion of zero observations; and/or
- a *Wages & salaries* / *Employee numbers* ratio that is not plausible.

3.2 Data from the 2004 EAS data set

	Total obs.	*Wages >$10m*	*Wages >$30m*	*Employ & wages zero*	*Wages only zero*	*Employ only zero*	*Average wage >$250k*	*Average wage <$2k*
01 Agriculture	275	6	#	82	33	7	#	7
02 Services to Agriculture	184	#	#	75	19	7	#	6
03 Forestry and Logging	269	4	#	92	50	4	#	4
04 Commercial Fishing	300	#	#	80	54	#	#	#
11 Coal Mining	147	28	11	61	#	#	#	#
12 Oil and Gas Extraction	99	9	5	42	#	#	#	#
13 Metal Ore Mining	193	29	16	57	4	7	4	#
14 Other Mining	90	7	#	27	#	#	#	#
15 Services to Mining	130	22	10	34	#	4	4	#
21 Food, Beverages and Tobacco	917	143	50	55	#	26	#	#
22 Textile, Clothing, Footwear and Leather	757	22	#	29	18	22	#	#
23 Wood and Paper Products	748	34	13	16	8	14	#	#
24 Printing, Publishing and Recorded Media	519	50	16	19	7	5	#	#
25 Petroleum, Coal, Chemicals	835	88	35	31	#	16	#	#
26 Non-Metallic Mineral Products	360	33	14	17	#	10	#	#
27 Metal Products	920	79	38	38	12	11	#	#
28 Machinery and Equipment	1,572	141	42	94	11	34	#	#
29 Other Manufacturing	707	11	#	20	4	19	#	#
36 Electricity and Gas Supply	381	38	24	191	36	9	#	#
37 Water Supply, Sewerage, Drainage Services	137	15	5	17	#	#	#	#
41 General Construction	248	64	32	59	18	5	#	#
42 Construction Trade Services	504	35	12	147	73	14	6	6
45 Basic Material Wholesaling	135	34	13	29	11	#	#	#
46 Machinery and Motor Vehicle Wholesaling	239	97	43	26	4	#	#	#
47 Personal and Household Good Wholesaling	241	70	23	44	15	6	#	#
51 Food Retailing	1,148	31	14	129	78	29	#	37
52 Personal and Household Good Retailing	366	83	28	49	37	10	#	5
53 Motor Vehicle Retailing and Services	689	28	7	100	46	20	#	8
57 Accommodation, Cafes and Restaurants	485	52	14	120	28	12	#	#
61 Road Transport	808	40	16	167	105	19	#	11
62 Rail Transport	36	11	8	9	#	#	#	#
63 Water Transport	156	8	#	69	4	4	#	#
64 Air and Space Transport	124	16	5	27	#	5	#	#
65 Other Transport	71	#	#	23	9	#	#	#
66 Services to Transport	132	38	19	28	6	#	#	#
67 Storage	60	9	6	11	#	#	#	#
71 Communication Services	150	16	9	58	4	8	#	#
75 Services to Finance and Insurance	869	64	32	557	48	9	9	#
77 Property Services	1,458	37	15	751	156	18	#	7
78 Business Services	1,085	276	128	311	82	32	24	7
84 Education	589	65	13	172	53	10	57	7
86 Health Services	364	119	43	74	25	#	#	#
87 Community Services	456	65	11	146	9	9	#	#
91 Motion Picture, Radio, Television Services	162	15	8	58	7	4	#	#
92 Libraries, Museums and the Arts	187	4	#	46	42	5	#	11
93 Sport and Recreation	158	27	7	35	11	#	#	#
95 Personal Services	199	12	#	48	27	11	#	#
96 Other Services	534	24	8	170	36	11	#	4
	21,193	2,105	806	4,540	1,212	460	134	150

Note:

1 Table 3.2 has been confidentialised. Observations of three or less have been replaced with a #.

2 The first column shows the subdivision codes defined in the ANZSIC 1993 classification.

Table 3.2 also includes information on *Average wages* – that is, reported *Wages & salaries* divided the reported *Employee numbers* for that year. The purpose of this is to check the plausibility of the data provided.

The *Wages & salaries* variable and the *Employee numbers* variable are measured on a different basis which may cause some misleading results when a ratio of these variables is taken (to produce *Average wages*, for example). Readers should interpret the numbers with caution. The *Wages & salaries* variable is the amount paid over a year, while the *Employee numbers* variable is a 'point in time' count of the number of persons employed. Consider a business employing ten staff and paying each staff member at a rate of $60,000 per annum. Ten months into the year the business restructures and reduces its employees to a single staff member on $120,000 per annum. The total wages paid in a year for this business will be (10 months × $50,000/month) + (2 months × $10,000/month) = $520,000. When surveyed, the business reports only one employee. The average wage for this business is then $520,000 per annum, which is misleading.

Table 3.2 also includes information on businesses reporting a value of zero for either *Wages & salaries* or the number of employees. Overall, approximately 30% of the observations have a zero for either *Wages & salaries* or *Employee numbers*. Several industries have a high proportion of reported zeros, including ANZSIC subdivisions 75, 77 and 36. ANZSIC subdivision 75 has a zero value for approximately 70% of its observations. Many of the zero observations will be from small businesses that do not employ staff.

The number of observations in a particular ANZSIC subdivision is important for modelling. *ANZSIC 62: Rail transport* is an issue with only 24 non-zero observations. At the other end of the scale, *ANZSIC 28: Machinery and equipment manufacturing* is well surveyed with 1,433 observations.

The table also includes information on the size of the *Wages & salaries* variable. Approximately 90% of observations have a total value of less than $10 million, and less than 4% report wages over $30 million.

3.3 Data treatment

As a result of our data analysis, observations that were considered implausible or did not reflect 'economic' activity (i.e. no wages or employees) were removed. A description of which observations were removed and the reason for their removal are listed below.

Removal of zero observations

Observations that reported a zero for either *Wages & salaries* or the number of employees were removed from the modelling process. These observation provide no information on the relationship between *Employee numbers* and *Wages & salaries*.

Removal of observations based on average wages [1]

An *Average wages* variable was calculated by dividing reported wages by the reported number of employees for all observations in the EAS data set. This *Average wages* variable was used to remove observations where the wages/employee ratio was not plausible. Initially the upper and lower limits were set at $250,000 and $2,000. However after consultation with industry experts the lower limit was revised to $500. A lower limit of $500 was considered relevant for industries that have part-time workers on low pay rates – an example would be *Food retailing*.

Removal of observations based on large wages

The model formed from the EAS data is to be applied to the BIT dataset. However, many of the large units in the BIT dataset are ABS maintained and, because the *Employee numbers* for those units will be available from the direct collection, the model will only be applied to the non-ABS maintained subset of the BIT data. Table 3.3 provides a list of upper limits by industry. All observations in the non-ABS subset of BIT have values for *Wages & salaries* under these limits. These upper limits were determined after investigating the wage ranges of non-ABS maintained businesses.

3.3 Upper limits by ANZSIC subdivision

ANZSIC	*Upper limit*
01 , 02 , 03 , 04	$10,000,000
11 – 37	$5,000,000
41 , 42	$20,000,000
45, 46 , 47	$30,000,000
51 – 75	$20,000,000
77 , 78	$30,000,000
81 , 82 , 84	$5,000,000
86 , 87	$20,000,000
91 – 99	$10,000,000

The model will only be applied to units with *Wages & salaries* less than the amounts listed in table 3.3. Thus for consistency the units with wages greater than those shown in table 3.3 will be removed from the EAS data set.

1 In this paper, 'wages' and 'wages and salaries' are used interchangeably.

3.4 Model specification

The explanatory variables available are limited to variables that exist in both the EAS and BIT data sets (see table 3.1). The following variables were initially considered for use in a model as independent variables.

Wages & salaries (W)

The focus of the modelling is the relationship between *Wages & salaries* and the *Employee numbers*. As expected plots produced in the original data analysis indicated that the two variables were positively correlated in the EAS data set. That is, an increase in *Wages & salaries* is associated with an increase in employment which makes economic sense.

ANZSIC

This variable may influence the relationship between *Employee numbers* and *Wages & salaries* indirectly. ANZSIC is an industry classification and businesses classified to different ANZSIC subdivisions are considered to produce different products. Furthermore, the workforce employed in different industries may differ in the composition of skilled and unskilled labour and in the proportion of part-time and full time arrangements. Therefore, ANZSIC classification is expected to capture some of these attributes between industries and, of course, it is also assumed that firms within a particular industry will produce similar products and have a similar workforce composition (e.g. proportions of skilled vs unskilled and part-time and full-time workers).

Type of legal organisation (TOLO)

The classification system is broadly arranged on private sector and public sector lines. TOLO includes categories such as, Proprietary, Sole Proprietor, Family Partnership, Australian Government Department, State Government Department and Local Government Authority. This variable is likely to capture the impact of different management and organisational structures on the relationship between employment and *Wages & salaries*.

Total income (TI)

Total income is used as an indication of the size of firms. Firms of different size may differ in terms of management and organisational structure and the composition of the labour force (e.g. skilled vs unskilled and part-time vs full-time workers) – which will have implications to the relationship between the *Employee numbers* and *Wages & salaries*.

It should be noted that the relationships between firm size (measured by *Total income*) and total labour employment may not necessarily be consistent between firms, even within the same ANZSIC subdivision. Larger businesses usually employ more labour, however a large business in terms of *Total income* may also employ a small amount of labour.

Total expenses (TE)

A direct relationship may exist between *Total expenses* and the *Employee numbers* but it is not expected to provide strong additional explanatory power because *Wages & salaries* account for a significant proportion of *Total expenses*.

Not for profit (NPI)

The variable of *NPI* is another variable that may explain some differences in the relationship between the *Employee numbers* and *Wages & salaries* between business firms.

Ratio of total assets to total income (TATI)

We also derived a few variables and tested their usability in the regression. One such variable is the *Ratio of total assets to total income*. This variable was considered to possibly identify the capital intensity of firms.

Model specification and variable selection

A previous ABS investigation into modelling employee numbers used a linear model and a regression was run for each ANZSIC subdivision. However, as the observations were split by ANZSIC subdivisions, the numbers of observations in some ANZSIC subdivisions was small, and as a result, the estimates (i.e. the coefficients and the 'predicted' values for the number of employees) became unreasonably volatile.

For this investigation we ran a regression on the complete data set and used the ANZSIC classification (as dummy variables) to distinguish firms that belong to different industries.

While *Wages & salaries* is expected to be the most powerful variable, we also tested the usability of other variables mentioned in this section. In our modelling process, we started with equation (3):

$$N = f(W, ANZSIC, TOLO, TI, TE, NPI, TATI) \qquad (3)$$

and we tested various functional forms (such as log, semi-log and quadratic etc.).

The variable selection was carried out based on both statistical diagnosis and the plausibility of the results. In the process, we found the *ANZSIC* variable (i.e. used as dummy variables) was statistically significant consistently across various model specifications.

The *TOLO* variable was statistically significant in some model specifications that we tested but the inclusion of this variable caused some of the estimates to be inconsistent over time. As a result, this variable was not included. Other variables were found to be either insignificant or their coefficients (and the 'predicted' values of *Employee numbers*) were unbelievable.

Functional form is another important consideration. We tried linear, log, semi-log and quadratic forms for the continuous variables. [2]

In the end, we came up with a very simple equation:

$$N = \alpha_0 W + \alpha_1 log(W) + \sum_{i=1}^{M} \beta_i (ANZSIC) \times W + \varepsilon \qquad (4)$$

where N is for *Employee numbers*, W for *Wages & salaries* and i for ANZSIC industry subdivision I. The coefficients from this model appear to be the most robust and the 'predicted' numbers of employees most plausible.

We would like to make three observations regarding equation (4). First, we restricted the intercept of the equation and forced it to be zero. This restriction was imposed because a model without this restriction would suggest that businesses that paid zero dollars in wages had employed a negative number of employees. This does not make sense. Forcing the intercept to equal zero ensures that that the firms reporting zero wages will show zero value in our predicted value of the number of employees.

Second, the dummy variables (represented by *ANZSIC*) are used interactively with the variable of *Wages & salaries*. This implies that the relationships between the *Employee numbers* and *Wages & salaries* are assumed to be different in the slope (rather than intercept) between firms belonging to different industries.

Third, the logarithm of *Wages & salaries* proved to be a useful term in the model. We found that firms with small reported *Wages & salaries* employed more than the expected number of staff. For these organisations the log(W) term made the 'predicted' number of employees more believable over a number of industries. We are not exactly sure why this phenomenon was observed. However, we suspect a greater use of part-time staff is a possible reason for the larger than expected employee number with the small organisations.

2 We did not use Box–Cox transformation to explore other functional forms mainly because the production system is believed to be unable to support more complex functional forms for the purpose of regular statistical compilations and publications.

3.4 Regression estimates, 2004

	Coefficient	*Standard error*	*t-statistic*	*Prob > |t|*
Wages	0.0276	0.0010	27.2490	0.0001
log(Wages)	0.4835	0.0272	17.7470	0.0001
ANZSIC				
01	0.0050	0.0024	2.0870	0.0369
02	0.0172	0.0041	4.2360	0.0001
03	–0.0025	0.0029	–0.8710	0.3840
04	0.0048	0.0042	1.1610	0.2456
11	–0.0150	0.0059	–2.5500	0.0108
12	–0.0170	0.0060	–2.8460	0.0044
13	–0.0122	0.0038	–3.1860	0.0014
14	–0.0117	0.0036	–3.2480	0.0012
15	–0.0132	0.0029	–4.5260	0.0001
21	–0.0021	0.0016	–1.3360	0.1816
22	–0.0018	0.0020	–0.9140	0.3609
23	–0.0032	0.0022	–1.4890	0.1365
24	–0.0069	0.0017	–4.1240	0.0001
25	–0.0066	0.0016	–4.0810	0.0001
26	–0.0085	0.0023	–3.7540	0.0002
27	–0.0062	0.0014	–4.3610	0.0001
28	–0.0069	0.0013	–5.1410	0.0001
29	–0.0042	0.0018	–2.2810	0.0226
36	–0.0146	0.0044	–3.3480	0.0008
37	–0.0064	0.0030	–2.1600	0.0308
41	–0.0077	0.0012	–6.6360	0.0001
42	–0.0111	0.0011	–9.7970	0.0001
45	–0.0081	0.0011	–7.0920	0.0001
46	–0.0091	0.0011	–8.4490	0.0001
47	–0.0080	0.0011	–7.3910	0.0001
51	0.0371	0.0014	26.9220	0.0001
52	0.0088	0.0011	7.8270	0.0001
53	–0.0025	0.0012	–2.0350	0.0418
57	0.0109	0.0011	9.7850	0.0001
61	–0.0046	0.0012	–3.8920	0.0001
62	–0.0040	0.0027	–1.5110	0.1308
63	–0.0079	0.0019	–4.2520	0.0001
64	–0.0102	0.0016	–6.4970	0.0001
65	–0.0113	0.0025	–4.5470	0.0001
66	–0.0086	0.0012	–7.1260	0.0001
67	–0.0058	0.0022	–2.6760	0.0074
71	–0.0118	0.0014	–8.2730	0.0001
75	–0.0136	0.0013	–10.7730	0.0001
77	–0.0078	0.0011	–6.8490	0.0001
78	–0.0053	0.0010	–5.1160	0.0001
84	–0.0002	0.0013	–0.1380	0.8903
86	0.0047	0.0011	4.2910	0.0001
87	0.0124	0.0011	11.1250	0.0001
91	–0.0060	0.0018	–3.2810	0.0010
92	0.0015	0.0025	0.6020	0.5472
93	–0.0001	0.0013	–0.1110	0.9116
95	0.0012	0.0020	0.6030	0.5462
R-square	0.7410			
Adjusted R-square	0.7405			

Note: The first column shows the subdivision codes defined in the ANZSIC 1993 classification.

3.5 Estimation results

The estimated coefficients, t-statistics and R-square for the data of 2004 are provided in table 3.4. The results for the data of 2003 are provided in Appendix B.

$$N = (0.02762 + \alpha_{2004,i})W + 0.48347\, log(W) \qquad (5)$$

where $\alpha_{2004,i}$ equals the parameter estimate of the relevant ANZSIC dummy term listed in table 3.4.

Table 3.4 shows that the coefficient for wages and salaries is 0.0276 and the coefficient for the log term is 0.483471. Because the measurement unit of wages and salaries is in thousands of dollars, these coefficients seem plausible. For example, our model predicts that a firm paying a wage bill of $100,000 will hire 3.7 workers – at an average wage of $26,819.

Of course, this value varies significantly between industries and the differences are captured by the coefficients of the ANZSIC dummy variables. Most of the signs and values are reasonably consistent with our expectations. For example, for *ANZSIC 75: Services to Finance and Insurance*, *ANZSIC 77: Property Services* and *ANZSIC 78: Business Services*, it is expected that *Wages & salaries* will be higher than for most other industries and, as a result, the coefficients of the industry dummy variables are expected to be negative. The signs of the coefficients are consistent with our expectation and are statistically significant. For the industry *ANZSIC 12: Oil and Gas Extraction*, the coefficient is –0.020671. This implies that a firm in this industry that pays a wage bill of $100,000 is expected to hire 2.03 workers. In other words, the average wage for this industry is $49,367 – which is significantly higher than the average and appears to be reasonable for this industry.

The t-statistics are significant for most ANZSIC dummy variables suggesting that the wage levels do differ across industries. This means that prediction of the number of employees would have been misleading, if we had not included the ANZSIC classification in the model.

The R-square for the 2004 data is 0.7410. This statistic provides a measure of the goodness-of-fit of the relationship between the dependent and independent variables in the regression analysis. In this instance, it is the percentage of variation in *Employee numbers* that is explained by the *Wages & salaries* and *ANZSIC* variables. We consider the R-square value of 0.7410 to be reasonable. The EAS data have many limitations (as discussed in Sections 3.2 and 3.3). One particular constraint is the limited number of variables that are available for us to use as independent variables in our regression. The inadequate number of explanatory variables severely limited our ability to control for differences in the relationship between *Employee numbers* and *Wages & salaries* between the firms.

3.5 Reported vs predicted employees, 2003 and 2004 – EAS Data

ANZSIC	*Reported employees 2003*	*Predicted employees 2003*	*Reported employees 2004*	*Predicted employees 2004*	*Reported change*	*Predicted change*
01	23,068	19,142	48,985	44,088	112.3%	130.3%
02	30,381	20,498	61,500	37,365	102.4%	82.3%
03	10,415	8,278	20,971	18,338	101.4%	121.5%
04	11,096	14,751	22,149	30,502	99.6%	106.8%
11	9,727	9,571	11,370	11,386	16.9%	19.0%
12	2,921	2,863	2,664	2,607	–8.8%	–8.9%
13	16,765	16,723	16,895	16,996	0.8%	1.6%
14	11,676	11,460	12,888	12,190	10.4%	6.4%
15	21,652	21,872	24,937	26,163	15.2%	19.6%
21	208,688	193,264	209,773	193,048	0.5%	–0.1%
22	91,620	87,428	90,056	85,123	–1.7%	–2.6%
23	75,986	69,704	81,842	74,000	7.7%	6.2%
24	138,690	137,240	133,870	130,652	–3.5%	–4.8%
25	135,582	123,750	138,593	124,207	2.2%	0.4%
26	52,119	48,674	53,351	48,423	2.4%	–0.5%
27	206,824	198,437	210,577	201,652	1.8%	1.6%
28	283,327	283,066	296,732	292,946	4.7%	3.5%
29	117,334	110,134	117,311	112,207	0.0%	1.9%
36	16,580	17,567	16,136	16,910	–2.7%	–3.7%
37	26,177	25,166	25,333	24,633	–3.2%	–2.1%
41	169,835	168,670	180,004	176,911	6.0%	4.9%
42	436,657	447,224	501,029	528,722	14.7%	18.2%
45	100,043	93,194	92,149	88,329	–7.9%	–5.2%
46	175,858	177,051	175,119	170,284	–0.4%	–3.8%
47	223,366	208,505	223,440	202,836	0.0%	–2.7%
51	317,308	323,894	387,255	410,138	22.0%	26.6%
52	570,379	597,154	603,828	708,769	5.9%	18.7%
53	238,611	253,825	207,360	206,752	–13.1%	–18.5%
57	610,587	603,331	675,883	613,739	10.7%	1.7%
61	205,996	220,751	192,329	199,534	–6.6%	–9.6%
62	3,175	3,275	3,835	3,618	20.8%	10.5%
63	11,020	10,241	12,947	12,784	17.5%	24.8%
64	10,872	9,636	12,847	11,979	18.2%	24.3%
65	7,084	5,788	5,323	4,775	–24.9%	–17.5%
66	77,643	73,152	70,051	68,888	–9.8%	–5.8%
67	11,928	11,785	11,953	12,185	0.2%	3.4%
71	20,881	18,894	28,429	25,325	36.1%	34.0%
75	35,143	36,432	36,514	36,168	3.9%	–0.7%
77	272,369	260,233	258,027	268,974	–5.3%	3.4%
78	1,217,345	1,226,136	1,285,370	1,299,673	5.6%	6.0%
84	345,790	336,002	333,464	327,699	–3.6%	–2.5%
86	537,681	616,861	513,664	569,128	–4.5%	–7.7%
87	379,071	380,321	383,205	386,699	1.1%	1.7%
91	40,287	43,988	33,613	30,220	–16.6%	–31.3%
92	38,936	39,440	42,607	39,702	9.4%	0.7%
93	144,303	126,206	175,279	138,190	21.5%	9.5%
95	133,188	154,331	152,125	169,435	14.2%	9.8%
96	172,920	172,802	170,728	164,547	–1.3%	–4.8%

Note: The first column shows the subdivision codes defined in the ANZSIC 1993 classification.

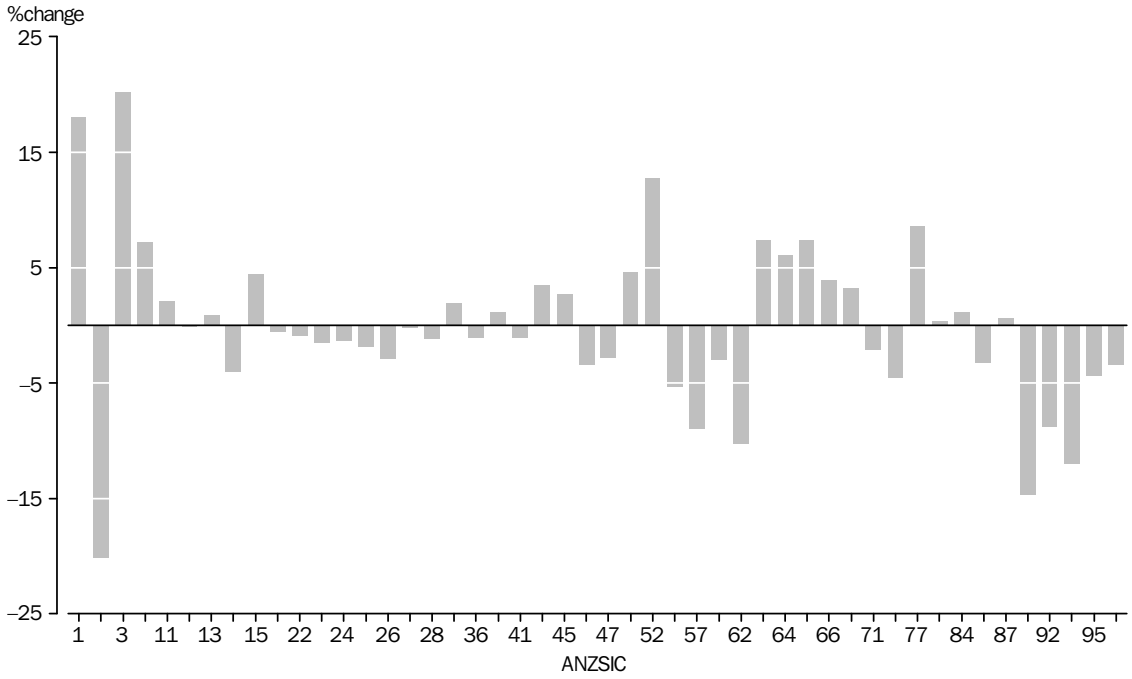
It is worth noting that the coefficients estimated from the data for the two years (i.e. 2003 and 2004) appear to be reasonably stable. (See table 3.4 and the table in Appendix B.) It is reassuring that the model appears to be reasonably robust to data conditions.

Table 3.5 is the result of applying the model to the EAS data and comparing the reported and predicted numbers of employees.

The last two columns compare the 'predicted' *Employee numbers* with those numbers aggregated from the reported EAS (i.e. the 'clean' data that were used as inputs in the modelling process) in terms of percentage movement between 2003 and 2004. At the aggregate level, the model appears to perform reasonably well.

However, the differences between the two sets of figure are more obvious at the industry level and for some industries they are quite significant. We calculate the differences between the predicted and reported values in percentage points and present them in figure 3.6. Out of 48 ANZSIC subdivisions, seven are within the bounds of ±1 percentage points, 16 within ±2 percentage points and 33 within ±5 percentage points. There are seven subdivisions where the differences are outside the bounds of ±10 percentage points, among which the predicted and reported value differs by more than 20 percentage points for industries *ANZSIC 02: Services to Agriculture* and *ANZSIC 03: Forestry and Logging*.

3.6 Differences between predicted and reported changes, 2003–04



These differences are significant but not totally unexpected. There may be several reasons for the differences. First, the sample sizes are small for certain industries. If the firms in those industries are not very homogeneous in the composition of labour employed (e.g. skilled vs unskilled and full time vs part time), then the estimates may differ depending on the methods of aggregation. In such circumstances, an econometric model is likely to generate a different estimate from those derived from straightforward aggregation methods. Our results seem to suggest that a further analysis of the 'diversity' of the firms within certain industries is required in order to test the hypothesis (of heterogenieity) and to understand the problem and improve the estimates. If our hypothesis is confirmed, then we will recommend to increase the sample sizes of the industries which appear to be highly diverse in labour employment.

Second, the ANZSIC classification covers the whole economy and we know that certain industries could be very different from others in terms of the composition of employment (e.g. skilled vs unskilled and part time vs full time workers). As mentioned earlier, one particular problem with EAS data is that we are unable to distinguish full-time and part-time employees. If an industry is too unique (in the employment composition), then the model is unlikely to perfectly pick up the differences between this and the rest of the industries, even if the sample sizes are adequate. It appears to us that the only way to reduce the impact of this problem is to further improve the modelling method.

3.6 Subsequent analysis

Since June 2006, further analysis was undertaken to address key issues identified by the Methodology Advisory Committee.

The committee noted certain inconsistencies in the model specification. For example, some MAC members questioned the inclusion of the log term in Equation 4 and suggested that an intercept term also be included. A further investigation suggested that the final estimates had been overstating employment level and growth of several industries and for the whole economy overall by a significant margin. In response to these concerns, we examined the impact of the log term and tested the model with an intercept term.

We found that the inclusion of the log term tended to overestimate employment of small businesses with total wage bills less than $150,000. These businesses accounted for a relatively small proportion of the EAS survey sample, but they made up over 80% of the business population in BIT. When we use the coefficients from equation (4) to predict the business employment (using BIT data), the log term had a significant impact on the results. Therefore, we decided to remove this variable from the model.

We also found that the inclusion of the intercept term would cause an even larger over-estimation of employment. Consequently, we decided not to include the intercept term. The revised model is presented in equation (6).

$$N = \alpha_0 W + \sum_{i=1}^{M} \beta_i (ANZSIC) \times W + \varepsilon \qquad (6)$$

This model assumes, for a specific industry, the number of employees and wages have a linear relationship. Our tests suggested that the relationship is a reasonably good approximation for businesses within certain size ranges. In this study, we estimated the model and used the coefficients to 'predict' the employee numbers for small businesses (with no more than 200 employees). We also applied industry-specific restrictions on the size of total wages as outlined in table 3.3. This model has been used to estimate the final 2004–05 employment estimates for industries, published in *Australian Industry* (ABS cat. no. 8155.0), *Manufacturing Industry, Australia* (ABS cat. no. 8221.0) and *Mining Industry, Australia* (ABS cat. no. 8415.0).

4. ASSESSING THE QUALITY OF THE EMPLOYEE NUMBERS MODEL AND ITS OUTPUTS

Section 3 provided a few diagnoses of the estimates from the regression which indicate that the model works reasonably well. In this section, we go a step further by examining the quality of the estimates from multiple perspectives.

4.1 A quality assessment framework

A major difference between the statistics presented in the previous section and those derived directly from survey data is the fact that, in this project, an econometric model is used in the statistical compilation. This means that we need to examine the quality of the statistics from at least three perspectives: statistical results (including the regression estimates and the 'predicted' values for the number of employees), the assumptions underlying the model, and the model specifications.

Table 4.1 shows the aspects that we consider important in examining the qualities of modelled statistics. They are used as the 'criteria' in the quality assessment. We are still at the early stage of building a quality assessment framework for modelled statistics and, in this paper, we do not intend to provide a 'definition' as such for each of the criteria. However, a brief interpretation of these criteria will be useful for readers to understand and evaluate the quality of the statistics presented in this paper.

4.1 Quality assessment criteria

Criteria	*Statistical results*	*Model assumptions*	*Model specifications*
Relevance	✔		
Accuracy	✔		
Plausibility	✔	✔	
Consistency	✔	✔	
Robustness	✔	✔	✔
Interpretability	✔	✔	✔
Transparency		✔	✔
Free of 'endogeneity'		✔	
Minimal Data Requirements			✔
Sustainability			✔
Cost Effectiveness			✔

Note: The tick in the box indicates that the criterion is relevant to each of the three perspectives from which we assess the quality of the statistical results.

Relevance and *accuracy* are basic requirements for any official statistics. Generally speaking, relevance means that the statistics fit the purpose of users. Users will find the statistics useless, if they are not relevant to their purposes, and misleading, if they are not sufficiently accurate. These two criteria have the same interpretation as their counterparts in a quality assessment framework described by Allen (2002).

Plausibility is a criterion relevant to both statistical results and model assumptions. As far as statistical results are concerned, plausibility requires the statistics to closely measure the concept (that it is intended to measure) and the numbers are believable. In the same time, model assumptions are plausible if they are 'realistic' and consistent with existing knowledge.

Consistency is also required for both statistical results and model assumptions. Generally speaking this requirement means that the results and assumptions must not be in disagreement with relevant theories in statistics, economics and social science.

Robustness is an important consideration in all the three perspectives. In terms of statistical results, it means that the estimates from the model (such as coefficients from a regression) and the derived statistics are not heavily influenced by a small number of unusual or unrepresentative observations. It requires the model assumptions to be reasonably general and applicable to a wide range of economic, social, demographic or policy conditions. If the assumptions are too specific and unique to a particular condition, then we will find the statistics quickly out of date and will have to update the models frequently. *Robustness* in model specification becomes an important consideration when we have opportunities to choose from two or more modelling techniques. In such circumstances, we should use the technique that is most general in the model assumptions and 'robust' to the data conditions.

Interpretability is another consideration applicable to the results, assumptions and model specifications. This criterion is reasonably self-explanatory. Literally, it means that the statistics and the model assumptions can be explained by the theories, data and economic, social and other conditions relevant to the statistics. If they are in conflict with existing knowledge, the disagreements should be explainable. Some modelling techniques are easy to explain and others are less so. If we have a choice, the former should be the preferred ones.

Transparency is a very important consideration in communicating with the users of the statistics. Because the mathematical or statistical transformations often make it harder for users to 'connect' the input data and the statistical results, it becomes important for us to help users establish the connection by properly explaining the estimation process and assumptions underlying the model. To achieve that, it requires model assumptions readily understood and clearly explainable. It also means that we should choose a technique (or model specification) that is less complex and has fewer steps in the transformation, if we have a choice.

Endogeneity is a term that is used in this paper to refer to a particular situation where the model assumptions defeats the purpose of users. For example, in this project, we assumed and estimated a particular relation between the number of employees and wages and salaries. However, if the main users of the statistics (of the employee numbers) want to explore the relation or test whether such a relation exists or not,

then our results are not going to be very helpful because they are derived based on the assumption. In a sense, our modelling process has already 'endogenised' the application of the statistics (or the purpose of the users). 'Free of endogeneity' requires us to examine the model assumptions to make sure that they do not defeat the purposes of the statistics.

Some models are more data demanding than others. Smaller or simpler models are preferred mainly because they are usually more manageable and often 'economical' in the production process. Furthermore, use of data from multiple sources may complicate the analysis and make it difficult to interpret the results. So, in practice, a 'smaller' model (i.e. less data-demanding model) is preferred to 'bigger' models.

Sustainability and *Cost efficiency* are required for the model to be able to support the ongoing production of the statistics at the lowest possible costs.

We have applied the 'quality assessment framework' to our work and found it useful. We are inclined to believe that this framework may be useful as a general guidance for assuring the quality of other modelled statistics. However, one should use it with caution. First, the criteria set out in this framework may not be exhaustive and, depending on specific circumstances, other factors may need to be taken into account in the quality assessment. Second, in the applications of this framework, all the criteria may not have equal importance and they need to be prioritised on a case-by-case basis.

The quality of the estimates of the employment numbers was assessed using this framework and the following 3 sections present some of the findings. Section 4.2 focuses on *consistency*, Section 4.3 on *plausibility* and Section 4.4 on *robustness* of the regression estimates.

4.2 Consistency

As explained in Section 3.1, to establish a meaningful relationship between *Employee numbers* and *Wages & salaries* (i.e. the coefficient of W in Equation (1)), it is critically important to adequately control for 'heterogeneity' (or the differences in the attributes) among the business firms. We attempted to achieve this by incorporating a set of variables in the regression (i.e. Z in equation (1)) that represent the characteristics of the firms. However, this approach is constrained by at least two factors.

First, EAS was not designed for our purpose and, as a result, the data set does not include all the variables required for Z in equation (1). Furthermore, in order to use BIT data to predict the *Employee numbers* based on equation (2), the variables chosen must be available in both the EAS and BIT datasets and they must be similar in

concept and definition. The limitations of the two datasets constrained our ability to control for the differences in firm characteristics.

In the analysis, we explored a range of variables and statistically tested their suitability for inclusion in the *Z*. The following variables are just a few examples.

- *ANZSIC classification* (included in *Z* as dummy variables) is used to reflect the differences in the firms' products and the conditions of the product markets;
- *Total assets* is used to control for differences in the size of firms;
- *Type of legal organisation* is used to control for potentially different management and organisational structures;
- *Total assets* and *Total expenses* are used to capture the effects of different sizes and technologies.

Second, even if we are able to obtain a perfect set of variables for *Z*, we may still be unable to perfectly capture (and adjust for) their impact on the relationship between the number of employees and wages and salaries. This is because the factors described in Section 3.1 – differences in products, technologies, sizes, management and organisational structure, product and labour markets – may influence the relationship in a very complex way. It is unrealistic to expect a simple equation (such as equation (4) or (6)) to be able to capture all the impacts perfectly.

In this project we explored a wide range of functional forms and analysed the impact of the variables from both economic and statistical perspectives. Unfortunately most of them were found to be statistically insignificant or have little impact on the estimates. The findings from these tests (presented in Sections 3.5, 4.3 and 4.4) led us to believe that the simple equation (5) may have been an adequate (and possibly the best) representation of the relationship between the number of employees and wages and salaries.

Another important issue is about the direction of causality in the relationship between the employee numbers and wages and salaries. Statistically, equation (1) assumes that it is *Wages & salaries* that determines the level of *Employee numbers*. In theory, this issue is about how business firms make employment decisions. We can imagine two broad (and extreme) scenarios.

In the first scenario, it may be argued that the firms determine their employment (e.g. the size and composition) based on their (planned) budgets. In this circumstance, it is reasonable to assume that *Employee numbers* is a dependent variable and *Wages & salaries* an independent variable. In other words, the causality runs from the latter to the former and therefore equation (1) is appropriate.

Alternatively, it is perhaps more likely that firms determine their employment (and other inputs to maximise profits) based on the needs in the production process (subject to budget constraints, as suggested by the standard economic theory). The *Wages & salaries* will follow, once the decisions on the size and composition of the labour inputs are made. If this scenario is true, then equation (1) is not a correct representation of the relationship between *Employee numbers* and *Wages & salaries*.

In reality, most firms' employment decisions may be somewhere on the spectrum between the two extreme scenarios. A technical issue is whether equation (1) is a reasonable representation of the employment decision of a so-called 'representative' firm. We are unable answer this question definitively at this stage and it may become a research topic in the future.

In fact, concerns over the causality between dependent and independent variables are not confined to this project. In recent years, we came across this problem in several projects where we attempted to develop modelled statistics using regression techniques. In most cases, we did not have a 'theory' as a guidance to help us determine the direction of causality. This appears to be a quite general issue. We imagine that, in the end, we may have to resort to a kind of statistical framework to guide our analysis.

4.3 Plausibility

This section assesses the plausibility of the estimates generated from the model. Two methods are used in the assessment. First, we compare the predicted average wages of the top and bottom five industries for 2003 and 2004. Second, we compare the 'predicted' number of employees with data from the ABS Labour Force Survey.

Predicted Average wages

Tables 4.2 and 4.3 below list the ANZSIC subdivisions with the lowest and highest predicted wages for 2003.

4.2 Lowest average wages for 2003

	Predicted average wage
Food Retailing	$12,385
Commercial Fishing	$18,236
Personal and Household Retailing	$21,020
Community Services	$21,070
Agriculture	$21,893

4.3 Highest average wages for 2003

	Predicted average wage
Oil and Gas Extraction	$82,104
Coal Mining	$67,656
Services to Finance and Insurance	$64,414
Electricity and Gas Supply	$60,469
Communication Services	$60,209

The prediction of the lowest five industries by average wage fits with expectations. For example, *Food Retailing* is an industry where many employees are on a minimum wage and work only a small number of hours per week. Also, *Commercial Fishing* is a seasonal industry – thus many of its employees are only employed for a few months of the year.

The prediction of the high earning industries also fits with expectations. *Oil and Gas Extraction* and *Coal Mining* are known to be industries with high employee earnings.

Tables 4.4 and 4.5 below list the ANZSIC subdivisions with the lowest and highest predicted wages for 2004. These tables have produced similar results to the 2003 tables and again these industries fit with expectations of industries that have employees with high and low wage levels.

It is worth noting that five of the average wages for different industries listed in tables 4.2–4.5 have declined between 2003 and 2004. A number a possible explanations for this to happen. A decline in average wage could result from a compositional shift where employee numbers have increased at a greater rate than the wages paid due to changes in the part time/full time composition. The decline in the wages may also be caused in the volatility of the data or other economic conditions.

4.4 Lowest average wages for 2004

	Predicted average wage
Food Retailing	$13,155
Services to Agriculture	$18,319
Commercial Fishing	$20,967
Personal Services	$21,539
Agriculture	$22,369

4.5 Highest average wages for 2004

	Predicted average wage
Oil and Gas Extraction	$76,018
Services to Finance and Insurance	$65,331
Coal Mining	$62,407
Electricity and Gas Supply	$62,335
Communication Services	$57,854

How the predictions compare with other ABS data

We also examined the plausibility of our restuls by comparing the year-to-year movements of the predicted numbers of employees with the estimates from other sources.

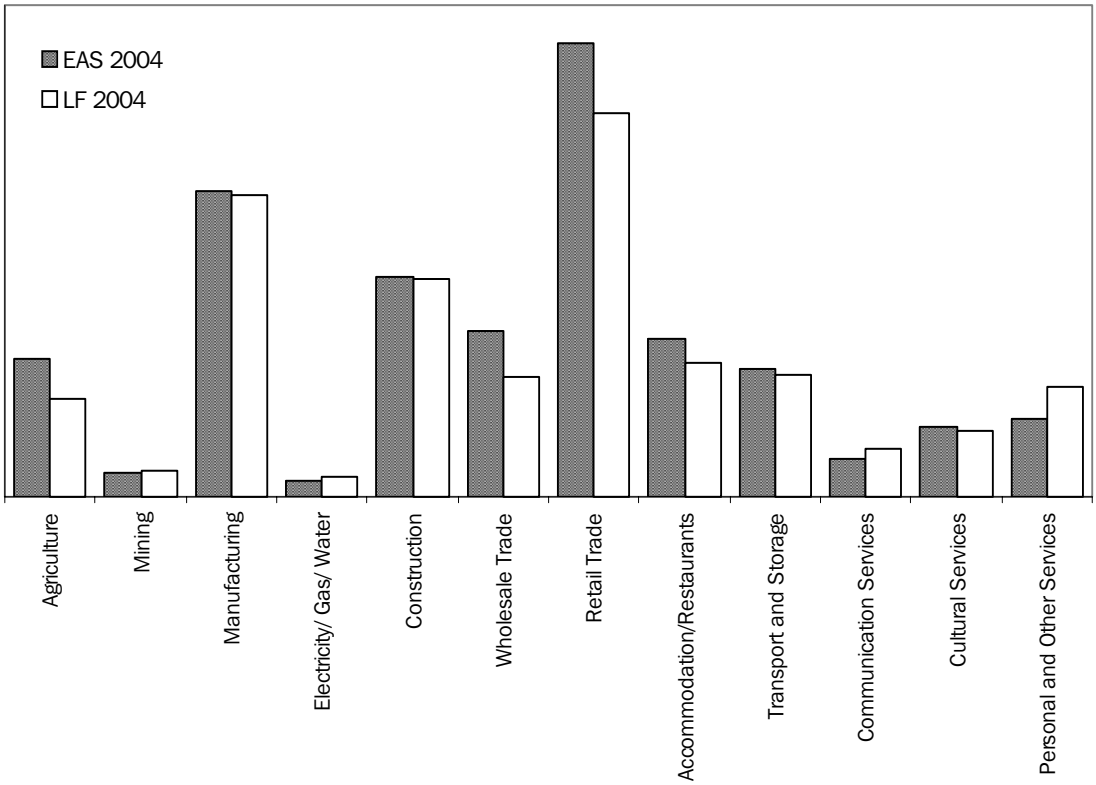
Figure 4.6 compares the predicted levels by industry with estimates from Labour Force. The comparison is made at the ANZSIC Division level.

The predicted number of employees for 8 (out of 12) industries are very close to the estimates from the Labour Force Survey suggesting that our results are believable. The estimates for the remaining 4 industries differ by more than 18%. However, the numbers in figure 4.6 must be interpreted with caution. A possible reason for some of the differences is that a person with multiple positions will be counted once in the labour force survey, but multiple times in the EAS data. This is likely to occur in industries such as food retailing.

Although the numbers in figure 4.6 do not perfectly conform with the estimates from the labour force survey, we found that the differences are explainable and they are mainly attributable to the imperfection of the data (rather than modelling techniques). So we are satisfied that the results are reasonably plausible and acceptable. This will continue to be monitored as more data become available.

In figure 4.6, four industries have been removed as scope differences mean the comparison is not meaningful. These industries are *Finance and Insurance*, *Property and Business Services*, *Education* and *Health and Community Services*.

4.6 Comparison: EAS model vs Labour Force statistics



4.4 Robustness

Initially, we did assess the goodness-of-fit of the employee numbers model by looking at the estimated R-squared, which was found to be reasonably high for the case study. Moreover, the analysis of variance table revealed a significant F-value, at 0.05 level of significance, indicating the overall significance of the chosen model. Though the t-test statistics for both the parameters of *Wages & salaries* and the logarithm of *Wages & salaries* already showed significance at 0.01 level of significance, restricted least squares estimations further supported that both of these variables significantly affect *Employee numbers* and should be included in the model.

The consistency and plausibility of the model have also been taken into account and were discussed in the previous subsections of the paper. Next, the standard assumptions behind the multiple regression modelling need to be examined for possible violations. To do this, further diagnostics using the residual values from the regression analysis were conducted to assess the robustness of the estimated model chosen in Section 3, most importantly the presence of heteroskedastic errors.

As a starting point, we investigated the plot of the residuals versus the predicted values as well as the plot of the residuals versus the independent variables. The plots of the residuals revealed increasing patterns as we moved from left to right, that is, the variances tend to increase. Such behaviour was much clearer in the plot of the

residuals against wages and log of wages. This suggested that heteroskedasticy might be a problem in the chosen model. If this was the case, then our least squares estimates of the parameter coefficients are still linear and unbiased but may no longer the the best. Least squares estimation will lead to incorrect standard errors and any hypothesis tests that use these standard errors might lead to wrong conclusion.

While the plots of the residuals suggest the likely existence of heteroskedastic errors, a much formal test using Goldfeld–Quandt (GQ) procedure was used to provide statistical evidence against the null hypothesis of homoskedasticity. The GQ test revealed that, at 0.05 level of significance, heteroskedasticity does exist. To remedy the problem, a generalised least square procedure was initially conducted and results showed lower estimates of standard errors as well as the estimated variance of the regression. Another alternative to remedy the heteroskedasticity problem is to use White's approximation procedure for the variances of the least square estimates. The latter procedure has not yet been implemented in the case study.

Though there has been some initial treatment of outliers in the data, the studentized residuals were also obtained and examined. The plot of studentized residuals revealed three to four outliers, and these may influence the results of the estimation. These should be investigated further.

We also tested whether the residuals are normally distributed. All four tests, namely, Jarque–Bera, Kolmogorov–Smirnov, Cramer–von Mises and Anderson–Darling, indicated non-normality in the residuals. This might suggests that the chosen functional form of the model can still be improved upon.

The regression modelling utilised cross-sectional firm level data. We would expect that the randomness of the sample data for these analysis implies that the error terms for different observations (firms) are uncorrelated, hence autocorrelation will not be a problem.

5. CONCLUDING REMARKS

This paper presents the main findings from our project of estimating numbers of employees as well as a framework that we proposed and used to assess the quality of the estimates. The results of the numbers of employees were intended be used in the official publication if the estimates were considered to have achieved sufficient quality. The quality assessment framework is at an embryonic stage and it is applied to a real situation for the first time. We plan to continue our efforts to further improve the framework and, if it turns out to be practically useful, we will use it to assess the quality of other modelled statistics in the future.

In the project of estimating numbers of employees, we used regression as the key technique in the data transformation. In the assessment of the quality of the estimates, we have focused on their consistency, plausibility and robustness.

Consistency

Our model assumed that there exists a statistical relationship between the number of employees and wages and salaries. This appears to be a reasonable assumption and we have shown that this relationship can be estimated and is reasonably consistent with our expectation.

However we tried but found it unrealistic to perfectly control for the heterogeniety of the firms within an industry and between industries. The limitations of the data is a main barrier to achieve our goal.

One unresolved issue is the difficulty in determining the direction of causality between the independent and dependent variables. This requires further investigations.

Plausibility

The coefficients from the regression appear to be reasonable in size and their signs are consistent with the *a priori* expectation. The coefficients of the key variables (i.e. wage and log wage) are statistically significant and so are most dummy variables used to control for the divergence in the firms' goods and services (i.e. ANZSIC classification).

We compared our estimates of the employee numbers with data from other sources and found the two sets of estimates are reasonably consistent. However, because the two data sources have different scopes, the estimates are not strictly comparable. Therefore the test of plausibility (based on this approach) does not necessarily lead a definitive conclusion and we wish to run more tests on the plausibility once an improved dataset is available.

Robustness

We ran a number of diagnostic tests focusing on the existence of heteroschadasticity and non-normality of the residuals and possible outliers in the data.

We found evidence of heteroschadasticity. Our investigation suggested that a general least square procedure may be able to adequately address the problem. The studentized results indicate that there may be a small number of outliers (i.e. three to four) in the data used for estimation. The tests of normality also suggests that chosen functional forms may also have room for further improvement.

ACKNOWLEDGEMENTS

The authors would like to thank Franklin Soriano and Peter Rossiter for their valuable contributions to this paper.

REFERENCES

Allen B. (2002) *Qualifying Quality – A Framework for Supporting Quality-Informed Decisions*, Discussion Paper, Australian Bureau of Statistics, June 2002.

Marriott, F.H.C. (ed.) (1991) *A Dictionary of Statistical Terms*, Fifth edition, Published for the International Statistical Institute by Longman Scientific and Technical, New York.

APPENDIXES

A. DESCRIPTION OF ANZSIC SUBDIVISIONS

A.1 Description of ANZSIC subdivisions

ANZSIC subdivision	*Description*
01	Agriculture
02	Services to Agriculture
03	Forestry and Logging
04	Commercial Fishing
11	Coal Mining
12	Oil and Gas Extraction
13	Metal Ore Mining
14	Other Mining
15	Services to Mining
21	Food and Beverage Manufacturing
22	Textile and Clothing Manufacturing
23	Wood and Paper Product Manufacturing
24	Printing, Publishing and Recorded Media
25	Petroleum, Coal and Chemical Manufacturing
26	Non-Metallic Mineral Product Manufacturing
27	Metal Product Manufacturing
28	Machinery and Equipment Manufacturing
29	Other Manufacturing
36	Electricity and Gas Supply
37	Water Supply and Sewerage Services
41	General Construction
42	Construction Trade Services
45	Basic Material Wholesaling
46	Machinery and Motor Vehicle Wholesaling
47	Personal and Household Good Wholesaling
51	Food Retailing
52	Personal and Household Good Retailing
53	Motor Vehicle Retailing and Services
57	Accommodation, Cafes and Restaurants
61	Road Transport
62	Rail Transport
63	Water Transport
64	Air and Space Transport
65	Other Transport
66	Services to Transport
67	Storage
71	Communication Services
75	Services to Finance and Insurance
77	Property Services
78	Business Services
84	Education
86	Health Services
87	Community Services
91	Motion Picture, Radio and Television Services
92	Libraries, Museums and the Arts
93	Sport and Recreation
95	Personal Services

B. PARAMETER ESTIMATES, 2003

B.1 Parameter estimates, 2003

	Parameter estimate	*Standard error*	*t-statistic for H_0: parameter=0*	*Prob > |t|*
Wages	0.0304	0.0010	30.2490	0.0001
log(Wages)	0.4582	0.0262	17.4860	0.0001
ANZSIC				
01	0.0033	0.0035	0.9400	0.3474
02	0.0054	0.0051	1.0580	0.2902
03	–0.0011	0.0046	–0.2470	0.8053
04	0.0098	0.0052	1.8790	0.0602
11	–0.0185	0.0053	–3.5110	0.0004
12	–0.0207	0.0056	–3.7150	0.0002
13	–0.0157	0.0038	–4.1600	0.0001
14	–0.0140	0.0037	–3.7980	0.0001
15	–0.0134	0.0031	–4.3950	0.0001
21	–0.0037	0.0016	–2.4090	0.0160
22	–0.0042	0.0019	–2.1380	0.0325
23	–0.0046	0.0022	–2.1140	0.0345
24	–0.0095	0.0016	–6.0440	0.0001
25	–0.0085	0.0016	–5.2850	0.0001
26	–0.0100	0.0022	–4.4830	0.0001
27	–0.0079	0.0014	–5.5920	0.0001
28	–0.0088	0.0013	–6.6570	0.0001
29	–0.0051	0.0018	–2.8620	0.0042
36	–0.0170	0.0043	–3.9610	0.0001
37	–0.0089	0.0029	–3.1030	0.0019
41	–0.0106	0.0012	–9.0460	0.0001
42	–0.0115	0.0012	–9.7830	0.0001
45	–0.0099	0.0011	–8.6350	0.0001
46	–0.0105	0.0011	–9.8520	0.0001
47	–0.0096	0.0011	–8.8930	0.0001
51	0.0400	0.0015	27.1610	0.0001
52	0.0090	0.0011	8.0050	0.0001
53	–0.0024	0.0012	–2.0410	0.0412
57	0.0093	0.0011	8.4330	0.0001
61	–0.0073	0.0012	–6.1720	0.0001
62	–0.0055	0.0025	–2.1930	0.0283
63	–0.0084	0.0020	–4.1960	0.0001
64	–0.0123	0.0017	–7.3270	0.0001
65	–0.0136	0.0025	–5.4650	0.0001
66	–0.0108	0.0012	–9.1720	0.0001
67	–0.0072	0.0019	–3.7020	0.0002
71	–0.0158	0.0014	–11.0010	0.0001
75	–0.0163	0.0013	–12.9360	0.0001
77	–0.0105	0.0011	–9.3330	0.0001
78	–0.0092	0.0010	–9.0080	0.0001
84	–0.0027	0.0012	–2.1770	0.0295
86	0.0032	0.0011	2.9890	0.0028
87	0.0128	0.0011	11.4800	0.0001
91	–0.0094	0.0018	–5.1950	0.0001
92	0.0016	0.0025	0.6280	0.5300
93	–0.0001	0.0014	–0.0810	0.9354
95	–0.0021	0.0020	–1.0430	0.2969
R-square =	0.7523			
Adjusted R-square =	0.7518			

FOR MORE INFORMATION . . .

INTERNET	**www.abs.gov.au** the ABS website 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 website 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 website, 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 statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS	www.abs.gov.au

2000001568514
ISBN 9780642483683

RRP $11.00