REFERENCES

Goldberger, A.S., Econometric Theory, John Wiley & Sons, Inc., New York, 1964.

Hancock, K. & B. Hughes, "Relative wages, institutions and Australian labour markets", Working Paper No. 1, Institute of Labour Studies, Flinders University of South Australia, May 1973.

Hancock, K. & K. Moore, "Occupation wage structure in Australia since 1914", British Journal of Industrial Relations, Vol. 10 (1), March 1972, pp. 107-22 and reprinted in J.R. Niland & J.E. Isaac, (eds.) Australian Labour Economics Readings, 2nd ed., Sun Books, Melbourne, 1975, pp. 207-225.

Hughes, B., "The wages of the strong and the weak", Journal of Industrial Relations, Vol. 15 (1), March 1973.

Kmenta, J., "Inter-industry wage differentials in Australia, 1947-1954" Australian Economic Papers, Vol. 2 (1), June 1963, pp. 85-106.

Parham, D.J. & G.J. Ryland, "Models of Skill Substitution and Transformation in an Occupationally Disaggregated Labour Market", Impact of Demographic Change on Industry Structure in Australia, Preliminary Working Paper No. OP-18, Industries Assistance Commission, Melbourne, March 1978 (mimeo).

IMPACT OF DEMOGRAPHIC CHANGE ON INDUSTRY STRUCTURE IN AUSTRALIA

A joint study by the Australian Bureau of Statistics, the Department of Employment and Industrial Relations, the Department of Environment, Housing and Community Development, the Department of Industry and Commerce and the Industries Assistance Commission

CONFIDENTIAL : Not for quotation without prior clearance from the authors; comments welcome

ABS LABOUR FORCE SURVEY AND INCOME DISTRIBUTION SURVEY DATA : PRELIMINARY ANALYSIS

by

Dean Parham and G. J. Ryland
Industries Assistance Commission

Preliminary Working Paper No. IP-05 Melbourne September 1978

IMPACT Project, Industries Assistance Commission, 608 St. Kilda Road, Melbourne, Victoria 3004 Australia
Area Code 03 Telephone 51 8611

In our case, the variables a and b are respectively income per person and hours per person. Various values were selected for ρ_{ab}, but the regression results were found to be relatively insensitive. The value of ρ_{ab} used in Tables 12 and 13 was 0.5 . To indicate sensitivity of parameter estimates, we provide a table of results for 1973/74 income per hour using various values of ρ_{ab} .

TABLE 14 : WEIGHTED REGRESSION RESULTS - EARNED INCOME PER HOUR (1973/74) - FOR SELECTED VALUES OF ρ_{ab}

ρ_{ab}	CONST	OCC	IND	$\bar{R}^2$	S.E.
	-0.133 (-0.45)	0.9794 (9.602)		0.9695	1.393
0.1	0.4918 (1.60)		0.7628 (7.195)	0.962	1.56
	-1.812 (-6.30)	0.8980 (11.60)	0.6680 (9.267)	0.9826	1.051
	-0.1124 (-0.39)	0.9799 (9.871)		0.9679	1.722
0.5	0.4973 (1.551)		0.7635 (6.977)	0.9581	1.966
	-1.825 (-6.319)	0.9042 (11.92)	0.6680 (9.123)	0.9815	1.308
	(-0.029) (-0.1083)	0.984 (10.75)		0.9633	2.771
0.9	0.4399 (1.194)		0.8019 (6.589)	0.9463	3.352
	-1.928 (-6.576)	0.9196 (13.15)	0.6998 (9.108)	0.9788	2.105

APPENDIX 3 : CALCULATION OF VARIANCES IN INCOME PER HOUR

The asymptotic variance of a ratio of two random variables is developed as follows (Goldberger, p. 124).

If $y = f(a,b)$,

then

$$\text{as.var}(y) = \sigma_y^2 = j'\Sigma j$$

where

$$j' = \begin{bmatrix} \frac{\partial y}{\partial a} & \frac{\partial y}{\partial b} \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{bmatrix}$$

In the case $y = a/b$,

$$\sigma_y^2 = \left(\frac{\partial y}{\partial a}\ \sigma_a\right)^2 + \left(\frac{\partial y}{\partial b} \cdot \sigma_b\right)^2 + 2\,\frac{\partial y}{\partial a} \cdot \frac{\partial y}{\partial b} \cdot \sigma_{ab} \ ;$$

and, since

$$\sigma_{ab} = \rho_{ab} \cdot \sigma_a\, \sigma_b$$

(where ρ_{ab} = correlation coefficient between a & b) ,

then

$$\sigma_y^2 = \left(\frac{\partial y}{\partial a} \cdot \sigma_a\right)^2 + \left(\frac{\partial y}{\partial b} \cdot \sigma_b\right)^2 + 2\rho_{ab}\left(\frac{\partial y}{\partial a} \cdot \sigma_a\right)\left(\frac{\partial y}{\partial b} \cdot \sigma_b\right) .$$

CONTENTS

APPENDIX 2 : IMPACT INDUSTRY CLASSIFICATIONS

IMPACT Industry	CLI	ASIC
1. Agriculture	02-06	01-02
2. Forestry and Fishing	00-01, 07	03-04
3. Mining-Metallic and Non-metallic Minerals	08-09	11,14,15,16
4. Mining-Coal and Crude Petroleum	Nil (included in 3)	12-13
5. Food, Beverages and Tobacco	26-34	21-22
6. Textiles, Clothing and Footwear	21-25	23-24
7. Wood and Wood Products; Paper and Paper Products	35-38	25-26
8. Basic Chemicals; Other Chemicals and Related Products	16,39,40	271-272
9. Petroleum Refining : Petroleum and Coal Products	11	273-274
10. Glass, Clay and Other Non-metallic Mineral Products	10	28
11. Basic and Fabricated Metal Products	12-15	29, 31
12. Transport Equipment	20	32
13. Industrial Machinery and Household Appliances	17-19, 44	33
14. Rubber and Plastic Products Manufacturing	43, 45-46	342, 343
15. Other Manufacturing Industries	41,42,47	341, 344
16. Electricity, Gas, Water, Sewerage & Drainage	48-49	36-37
17. Building and Construction	50-51	41-42
18. Wholesale and Retail Trade	61-64	46-48
19. Transport and Storage	52-56	51-55
20. Communication	57	56
21. Finance, Insurance, Real Estate & Business Services	58-60, 74	61-63
22. Public Administration and Defence	65-67	71-72
23. Community Services and Health	68-73	81-84
24. Entertainment and Recreational Services	75-79	91-94
25. Other - Non Classifiable Establishments	80-99	99

Appendix 1 (contd)

	IMPACT Minor Groups	IMPACT Major Groups	Parham/Ryland Groups
15.	Semi-skilled Metal and Electrical	8. Semi-skilled and Unskilled Blue Collar	5. Unskilled Blue Collar
16.	Building		
17.	Miners		
18.	Drivers		
19.	Protective Services		
20.	Production & Process		
21.	Services		
22.	Labourers		
23.	Farmers	9. Rural Workers	
24.	Farm Workers		
25.	Officers	10. Armed Services	6. Other
26.	Other Ranks		
27.	Other (n.e.c.)	11. Other	

ABS LABOUR FORCE SURVEY AND INCOME DISTRIBUTION SURVEY DATA : PRELIMINARY ANALYSIS

by Dean Parham & G.J. Ryland*

1. INTRODUCTION

There are very few Australian studies of the occupation and industry specific components in the variability of labour market characteristics such as hours worked and earned income.[1] A major factor in explaining this scant attention is the lack of a data base suitably disaggregated by industry and occupation. However, as part of its participation in the IMPACT Project, the Australian Bureau of Statistics (ABS) has processed additional unpublished data from its Labour Force Surveys in a form suitable for subsequent analysis. These data provide information on, inter alia, an industry by occupation disaggregation of earned income, hours worked and numbers of persons employed. It is on this subset of the data that the investigation in this paper is focused.

It should be emphasized from the outset that we are using the results of the Labour Force Surveys for a purpose for which they were not primarily designed. Rather than focusing mainly on labour force characteristics at the industry by occupation level of disaggregation, the surveys are intended to serve multiple purposes. Among these objectives are the analyses of the distribution of income, workforce participation rates and hours worked according to age,

* With the usual caveat regarding responsibility for omissions and errors, we thank Alan Powell for his encouragement and advice and Bill Pattinson, Andrew Bertie and Ashok Tulpulé for providing information on the ABS Surveys.

1. The only studies of which we are aware are analyses of occupational award wage rates (Hancock and Moore (1972)) and of industrial earnings (Kmenta (1963), Hancock & Hughes (1973)).

sex, marital status, educational status, and finally industry and occupation.

It is largely due to these considerations of design, that the data, when processed at an industry by occupation level of disaggregation, do not meet the normal standards of reliability required by the ABS for publication. However, they do constitute the only available basis for estimation of certain parameters in the IMPACT model.

Our task in this paper is to subject the data to preliminary analysis. Our prime concern is to assess their reliability at the industry by occupation level of disaggregation and thus their suitability for subsequent use in parameter estimation. We are particularly interested in the reliability of the hours worked, earned income, and persons employed data which will be used to estimate substitution and transformation parameters in a model of occupationally disaggregated labour demands and supplies (Parham and Ryland (1978)).

The outline of the paper is as follows. In the next section, we describe, in broad terms, the design of the Labour Force Survey and the Income Distribution Survey, the nature of the data collected, and more particularly the unpublished data provided by the ABS. In Section 3 we discuss factors which cause concern for the reliability of the data. These factors explain, at least in part, the deficiencies apparent from scrutiny of the data in Section 4. In Section 5, we use further information to calculate variances in labour force characteristics, which form an important feature of subsequent regression analysis. A preliminary analysis of the variability of labour force characteristics using coefficients of variation and simple regression analysis is conducted in Section 6. The final section summarizes the discussion.

APPENDIX 1 : IMPACT OCCUPATION CLASSIFICATIONS

	IMPACT Minor Groups	IMPACT Major Groups	Parham/Ryland Groups
1.	Scientists Engineers Medical Societal	1. Professional White Collar	1. Professional
2.	Tertiary Teachers Secondary Teachers Technical Teachers Primary Teachers	2. Lecturers and Teachers	
3.	Technicians	3. Skilled White Collar	2. Skilled White Collar
4.	Para-medical Creative		
5.	Government		
6.	Employers		
7.	Clerical	4. Semi-skilled and Unskilled White Collar	3. Unskilled White Collar
8.	Sales		
9.	Semi-skilled Medical Audio-visual		
10.	Metal Trades	Skilled Blue Collar Metal and Electrical	4. Skilled Blue Collar
11.	Electrical Trades		
12.	Instrument Trades		
13.	Wood Trades Brick, Stone, Glass Trades Painter	6. Skilled Blue Collar Building	
14.	Food Trades Textile Trades Printing Trades	7. Skilled Blue Collar Other	

Variances were also used in weighted regression analysis, conducted to identify the relative contribution of occupation and industry effects in explaining variability in labour force characteristics. Occupation and industry effects accounted for little variation in hours per person, but, to the extent that variation was explained, the results attributed a greater importance to the industry effect. However, the occupation effect was clearly dominant in analyses of income per person and income per hour. This last result offers support to the IMPACT research strategy in which wages are treated as occupation (rather than as industry) specific.

2. DESCRIPTION OF THE SURVEY DATA

Sample Design

Data on hours worked and income earned were generated from ABS survey information. As part of the Population Survey, the ABS conducts a regular Labour Force Survey (LFS), which is designed primarily to extract information on the employment status (employed, unemployed, not in the workforce) of the population. In addition, the ABS has appended irregularly an Income Distribution Survey (IDS), as a supplement to the LFS. The primary focus of the IDS is on the levels and sources of income. The LFS provides distributions of the workforce according to age, sex, marital status, full-time/part-time employment status, birthplace, hours worked, industry and occupation. The IDS collects additional information on income distributions according to educational attainment, age, sex, marital status, etc..

The surveys relate to the civilian population of Australia, aged fifteen years or over. The LFS, with a sample which covers approximately one per cent of the population,[1] is conducted at quarterly intervals.[2] The IDS is a sub-set of the LFS and has a sample size approximately half that of the LFS. The IDS is an infrequent survey, and, to date, has been conducted only in conjunction with the November LFS's of 1969 and 1974.

The primary sampling unit of the surveys is the household. In the selection of households, the sample is first grouped into private dwellings and special dwellings (e.g., hotels, hospitals, etc.). The sample is further stratified according to the area (State and region), and age and sex of occupants. In general, sampling is random up to the stage of selecting Census Collectors' Districts (CDs), but, within CDs, selections of households are clustered.

---

1. A major redesign of the LFS was phased in over 1972. Amongst changes made was the reduction of the sample size to two thirds of one per cent.

2. From February, 1978, the LFS has been conducted monthly.

All survey information is collected by personal interview at the household level. This contrasts with the Census, in which information is collected in accordance with the respondent's compliance with and interpretation of a questionnaire.

The sample of the LFS is rotated one eighth each quarter. Thus approximately seven eighths[1] of the respondents in one survey will be present in the next survey. Further, since the two existing IDSs are more than two years apart, the results of each of these surveys are based on an entirely different sample.

Definitions

In general, information contained in the LFS relates to the working week preceding the survey week. On the other hand, the additional information on income contained in the IDS refers to the previous financial year. The principal definitions of the main variables of interest are listed below.

Employed persons : employees, employers or self-employed persons who, during the survey week, did any work for pay, profit, commission or payment in kind.

Unemployed persons : persons who did no work, had no job and who were actively seeking work.

Full-time/part-time workers : persons who usually (or, during the survey week, actually did) work 35 hours or more are classed as full-time workers.

Earned Income : income in the previous year from wages or salary, own business, trade, profession or share in a partnership.

1. Some respondents may change address and therefore will be removed from the sample.

7. SUMMARY

There have been two purposes in writing this paper. The first has been to assess the reliability and compatibility of the Labour Force Survey and Income Distribution Survey in providing data bases of hours worked and income earned (respectively). This precedes the use of the data for estimation of parameters of the IMPACT model. The second has been to conduct preliminary analyses of the data, and in particular to assess the relative contributions of industry effects and occupation effects in explaining variations in labour force characteristics.

The surveys are primarily designed to provide information at an aggregate level and not at a level disaggregated by industry and occupation. Consequently, the small sample size and the stratification scheme used imply that sampling errors of estimates at the disaggregated level are, in some cases, prohibitively large. Given that no other source of data for the analysis of detailed labour force characteristics exists, this is particularly unfortunate. Additional errors in the special tabulations available to IMPACT result from deconfidentialization, truncated expansion factors and imprecise classification concordances. Further, there are problems with matching earned income with hours worked.

We have extracted information on the variances of characteristics at each occupation by industry level of aggregation. This enabled the construction of a relative index of variability, the coefficient of variation. According to this measure, the dispersion in earned income per person narrowed considerably between 1968/69 and 1973/74, especially for professionals.

We also analysed occupation and industry effects on hourly income. The income per hour variable was constructed as the ratio of wages per person to hours per person. Thus the previously discussed mismatching of income and hours and the mismatching of numbers of persons from the two data sources, are inherent in this variable.

For the sake of completeness we include weighted regression results for income per hour. However, we are more tentative in presenting these results as the calculation of the variances in income per hour, detailed in Appendix 3, involves further approximations.

The results of the income per hour analysis (Tables 12 and 13) clearly attribute dominance to the occupation effect. This is in spite of the fact that only five occupations (versus 23 industries) are distinguished in the analysis. The $\bar{R}^2$ for the occupation only regressions are 0.42 (and 0.48) in 1968/69 and 0.39 (and 0.43) in 1973/74, whereas the corresponding values for the industry only regressions are 0.03 (0.13) and 0.07 (0.12).

Income from wages or salary is defined as gross income (including overtime, bonuses, gratuities, etc.) before taxation and other deductions are made.

Other sources of income differentiated in the IDS include government social service benefit; superannuation; interest, dividends, rents, etc.; other (e.g., trust or will, maintenance or alimony).

Data supplied to IMPACT

The ABS has prepared special tabulations of the two IDSs, relating respectively to earned income received in 1968/69 and in 1973/74. In the case of the LFS, special tabulations of the May survey in each calendar year from 1968 to 1974 were prepared, with two exceptions. First, the 1968 tabulation refers to the August survey, and second, no 1969 tabulation is available.

The tabulations provide information on the number of persons employed, number of hours worked and number of people in each of 13 income ranges. The distribution of these variables according to age, sex, marital status, birthplace, and industry and occupation of employment is also provided. The variables are grouped into 25 IMPACT industry classifications and 27 IMPACT occupation classifications.[1]

Because of the fine level of disaggregation, population estimates in many individual cells are small. The ABS has determined that any (non-zero) population estimate of 500 persons or less is to be considered

1. Occupation and industry classifications are given in Appendices 1 and 2 respectively.

confidential. To circumvent this, the ABS has substituted a hybrid estimate in confidential cells, the estimate being the average estimate of population size across all confidential cells.

TABLE 12 : REGRESSION RESULTS : EARNED INCOME PER HOUR (1968/69)

	CONST	OCC	IND	$\bar{R}^2$	S.E.
UNWEIGHTED	0.2712 (1.314)	0.9257 (9.058)		0.4198	0.6627
	1.264 (3.440)		0.4562 (2.204)	0.033	0.8554
	-0.6177 (-1.834)	0.9357 (9.541)	0.5021 (3.263)	0.4662	0.6356
WEIGHTED[1]	-0.0385 (-0.212)	0.9454 (8.222)		0.480	1.712
	0.3072 (1.60)		0.7095 (5.95)	0.128	1.895
	-0.9550 (-4.812)	0.8848 (9.246)	0.6353 (7.084)	0.515	1.419

1. The three observations deleted in the 1968/69 income per person weighted regressions were also deleted in the income per hour weighted regressions.

TABLE 13 : REGRESSION RESULTS : EARNED INCOME PER HOUR (1973/74)

	CONST	OCC	IND	$\bar{R}^2$	S.E.
UNWEIGHTED	1.199 (3.959)	0.7134 (8.637)		0.3944	1.006
	2.219 (4.497)		0.4545 (3.054)	0.0686	1.248
	-0.2642 (-0.5735)	0.7132 (9.206)	0.4541 (4.034)	0.4671	0.9437
WEIGHTED[1]	-0.1150 (-0.398)	0.9799 (9.869)		0.428	1.722
	0.5036 (1.564)		0.7608 (6.921)	0.116	1.972
	-1.819 (-6.288)	0.9046 (11.91)	0.6650 (9.059)	0.497	1.309

1. The one observation deleted in the 1973/74 hours per person weighted regressions was also deleted in the income per hour weighted regressions.

TABLE 10 : REGRESSION RESULTS : WEEKLY HOURS PER PERSON (1968)

	CONST	OCC	IND	$\bar{R}^2$	S.E.
UNWEIGHTED	32.52 (4.893)	0.1562 (0.8693)		-0.0021	6.881
	25.81 (4.694)		0.3368 (2.281)	0.0356	6.75
	20.06 (2.356)	0.1562 (0.885)	0.3368 (2.279)	0.0337	6.757
WEIGHTED	31.92 (14.76	0.1804 (3.088)		-0.002	92.41
	3.949 (1.343)		0.9035 (11.78)	0.036	64.46
	-5.965 (-2.024)	0.2273 (6.483)	0.9434 (14.3)	0.032	55.21

TABLE 11 : REGRESSION RESULTS : WEEKLY HOURS PER PERSON (1973/74)

	CONST	OCC	IND	$\bar{R}^2$	S.E.
UNWEIGHTED	19.77 (4.459)	0.4799 (3.924)		0.1121	5.316
	18.50 (4.862)		0.5093 (4.914)	0.1688	5.144
	1.217 (0.2293)	0.4799 (4.368)	0.5093 (5.292)	0.2834	4.776
WEIGHTED[1]	21.80 (8.167)	0.4515 (6.257)		0.352	78.45
	3.042 (0.889)		0.9281 (10.36)	0.360	65.12
	-6.142 (1.924)	0.3489 (6.778)	0.8320 (10.81)	0.438	55.01

1. One observation with a very small calculated variance was deleted for the weighted regressions.

3. COMMENTS ON THE SURVEY DESIGN

In this section and the next, we examine the reliability of the LFS and IDS in providing a data base of labour force characteristics at the occupation by industry (O:I) level of disaggregation. Here, we present a series of factors which, from considerations of survey design and processing of results, cause concern for the reliability of the survey as an O:I data base. These factors presumably contribute significantly to an explanation of the deficiencies in the data, apparent from the analysis of Section 4.

We consider relevant factors of the survey method and design, as well as factors specific to the tabulations provided by ABS.

Survey Method

The personal interview approach to survey reduces the extent of bias due to non-response considerably. Further, it narrows the scope for misinterpretation and imprecise replies on the part of respondents. However, it does suffer from the possibility that the respondent is not the household's major participant in the workforce, but perhaps a mal-informed dependant.

This last factor is likely to be a significant source of error in estimates of wages, hours worked, industry and occupation. The positive aspect of the personal interview approach (reduced non-response etc.) would be a contributing factor in explaining differences in, for example, labour force participation rates estimated by the LFS as opposed to the Census.

The data base will be used to examine the relationships between hours and wages by using the results of comparable LFSs and IDSs. However,

doubt must be cast upon the precision with which hours worked (from the LFS) can be matched with wages received (from the IDS). First, hours worked refers to a particular working week identified by the LFS, whereas wages information relates to the financial year preceding the IDS. The working week may include an abnormal level of overtime hours which would not be reflected in annual income. Similarly income may reflect unemployment or indisposition during the financial year which is not present in the LFS survey week. Second, full information on actual hours worked is collected, whereas income is identified only by selection of a broad income range. Third, there is no provision for matching part-time or second-job income with (extra) hours worked.

Survey Design

The LFS and IDS have not been designed to provide a reliable occupation by industry disaggregation of labour force characteristics. Even with a relatively small number of industries and occupations, the population estimates in many O:I cells are small. As can be seen from the following table, the standard errors of estimate in cells containing small numbers are so large that the information content cannot be considered statistically reliable. Indeed, the ABS cautions against the use of data based on population estimates of less than 4000.

To gain some idea of the number of cells with population size estimates below the levels provided in Table 1, refer to Table 2. The figures in this table were calculated using a 5 occupation and 24 industry split (120 cells).

TABLE 8 : REGRESSION RESULTS : WEEKLY EARNED INCOME PER PERSON (1968/69)

	CONST	OCC	IND	$\bar{R}^2$	S.E.
UNWEIGHTED	-10.55 (-1.49)	1.252 (13.0)		0.5999	20.74
	19.83 (1.18)		0.9136 (3.51)	0.0919	31.25
	-63.98 (-5.6)	1.24 (14.52)	0.8562 (5.593)	0.6857	18.39
WEIGHTED[1]	2.40 (0.429)	0.9219 (9.37)		0.643	1.61
	-6.66 (-0.862)		1.013 (7.933)	0.189	1.72
	-46.69 (-7.78)	0.839 (12.17)	0.8993 (10.76)	0.700	1.12

1. Three observations, for which the calculated variances were zero, were omitted in the weighted regressions.

TABLE 9 : REGRESSION RESULTS : WEEKLY EARNED INCOME PER PERSON (1973/74)

	CONST	OCC	IND	$\bar{R}^2$	S.E.
UNWEIGHTED	-2.808 (-0.294)	1.112 (14.91)		0.6619	26.98
	37.08 (1.414)		0.8502 (3.763)	0.1043	43.92
	-99.44 (-6.53)	1.111 (18.12)	0.8446 (7.405)	0.7717	22.17
WEIGHTED	-2.905 (-0.311)	1.007 (11.23)		0.662	1.70
	-8.825 (-0.604)		1.02 (7.53)	0.104	2.02
	-91.48 (-9.268)	0.935 (15.4)	0.8994 (11.65)	0.765	1.14

of the weighted regressions to the original untransformed data.[1]

The general presence of multicollinearity among the explanatory variables, when both the industry and occupation means are included, makes difficult any meaningful partition of the explained sum of squares into the separate industry and occupation contributions. Consequently, the analysis of the relative contributions of industry and occupation effects must rest with a comparison of the degree of explanation provided by the occupation only regression as opposed to the industry only regressions.

On this basis, the results of the earned income per person regressions (Tables 8 and 9) suggest that the occupation effect, has a much greater explanatory power than the industry effect. In the 1968/69 income per person unweighted regressions, the occupation only regression accounts for 0.6 of total variation whereas the industry only regression accounts for 0.1 . The corresponding fractions for weighted regressions, based on three fewer observations, are 0.64 and 0.19 . The occupation effect is even more dominant in the 1973/74 regressions, where the occupation only regressions accounted for 0.66, whereas the industry only regressions accounted for 0.1 .

The degree of explantion provided by the various regressions is lower for weekly hours per person (Tables 10 and 11). This is particularly true for 1968 hours per person. However, to the extent that variation is explained, the industry effect appears dominant.

1. The quasi-R^2 is calculated by taking firstly the square of the correlation coefficient between the original untransformed variable (y_{ij}) and the predicted variable $(\hat{y}_{ij})$ using the estimated coefficients of weighted regressions. This value is then adjusted for degrees of freedom.

TABLE 1 : STANDARD ERRORS OF ESTIMATE FROM LFS AND IDS

Size of Estimate (Persons)	Approximate Standard Error of Estimates (% of Estimate)	
	LFS	IDS
500	65.7	90.6
1000	43.5	58.9
2000	28.8	38.2
4000	19.0	24.8
10000	11.0	14.0
20000	7.3	9.1
50000	4.2	5.2
100000	2.8	3.4

Source : Published information (ABS The Labour Force (Ref. 6.22) and Income Distribution (Ref. 17.6) and extrapolation.

A power curve of the form $y = ax^{-b}$, where y = percentage standard error and x = size of estimate of population, was used for extrapolation (and interpolation). The parameters a and b were estimated by fitting a power curve to the published data. For the IDS a = 4331.9, b = 0.62 and for the LFS a = 2669.2, b = 0.6. The goodness of fit in both cases was extremely high.

TABLE 2 : NUMBER OF CELLS[1] BELOW VARIOUS POPULATION SIZE ESTIMATES[2]

Survey		Number of Cells Which Have Population Estimates Less Than or Equal To							
		100000	50000	20000	10000	4000	2000	1000	500
1968/69	IDS	99	89	60	47	28	16	8	5
1973/74	IDS	104	91	63	51	27	12	6	2
1968	LFS	106	95	68	52	33	16	9	3
1973	LFS	103	92	65	52	23	14	8	5
1974	LFS	103	91	65	50	27	11	8	5

1. Total number of cells is 120 (5 occupations x 24 industries) for all surveys except the 1968/69 IDS, which has a total of 115 (industries 3 and 4 not differentiated).

2. The number of cells with population estimates of 500 or less in Table 2 should not be taken as an indication of the pervasiveness of confidentiality. First, the numbers of cells includes null cells, which are not confidential. Second, deconfidentialization takes place at the 27 occupation split and aggregation into 5 occupations may, for example, take a confidential cell, when summed with a larger cell, into the greater than 500 category.

These two tables together imply that the statistical reliability of a significant proportion of our occupation by industry information is doubtful. This results largely from the fact that the survey is not designed to provide reliable estimates at the disaggregated level. Specifically, the statistical reliability would only be increased at the occupation by industry level if the sample size were increased and if emphasis were placed on stratification of the sample according to industry and occupation rather than area, age and sex.

The small number problem is quite likely intensified by the one eighth rotation of survey after each quarterly LFS. In cells with small population sizes, estimates may be based on a sample of only a few people. The effect of inclusion or exclusion of households through rotation could be quite drastic, meaning that standard errors on quarter to quarter changes would be very high. This is affirmed by ABS published standard errors, which indicate a much higher standard error on quarter-to-quarter changes for cells with small size population estimates.[1] The reliability of comparisons of data relating to surveys five years apart is thus somewhat uncertain.

Changes in the sample design contribute further to the unreliability of estimates of quarter-to-quarter changes. A major design change was phased in over 1972 leading to higher than usual standard errors in quarter-to-quarter changes over that period and to some discontinuity in the series.

1. See, for example, the technical note in ABS, The Labour Force 1972, Canberra, 1974 (Ref. 6.22).

In equation (4) ,

$$(4b) \qquad \bar{Y}_{.*} = W^{-1}\bar{y}_{.*} \; ; \; \bar{Y}_{*.} = W^{-1}\bar{y}_{*.} \; ; \; \bar{Y}_{*j} = W^{-1}\bar{y}_{*j} \; ;$$

$$\bar{J} = W^{-1}\bar{1} \; ; \; \bar{E} = W^{-1}\bar{\varepsilon}$$

Fitting (4a) by ordinary least squares (OLS) is equivalent to fitting (1) by generalized least squares. This procedure can be expected to be relatively more efficient than fitting (1) by OLS, because the allowance made for heteroskedasticity ensures that the observations are weighted according to their information content.

We present regression results for weekly[1] earned income per person (Tables 8 and 9), hours worked per person (Tables 10 and 11) and earned income per hour (Tables 12 and 13). In each table, results of both the unweighted model (as in equation (1)) and weighted model (as in equation (4)) are presented. For each model, estimated coefficients are presented for equations including (i) the occupation means (ii) the industry means and (iii) both the occupation and industry means. Calculated "t" statistics are given in parentheses below associated regression coefficients. We present also values of the adjusted coefficient of multiple determination ($\bar{R}^2$) for unweighted regressions and the quasi-$\bar{R}^2$ for weighted regressions. The quasi-$\bar{R}^2$ reflects the goodness of fit

1. Income per person is divided in all cases by 48 to provide an estimate of weekly earned income.

$$W^{-1} = \begin{bmatrix} \sqrt{p_{11}}/\sigma_{11} & 0 & \cdots & 0 \\ 0 & \sqrt{p_{12}}/\sigma_{12} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sqrt{p_{nm}}/\sigma_{nm} \end{bmatrix}$$

where

σ_{ij} = standard deviation of the per capita characteristic in O:I cell ij ,

p_{ij} = estimate of the number of persons in O:I cell ij ,

and

$$WW' = V$$

Thus

(3) $$W^{-1} V(W^{-1})' = I .$$

If we now rewrite (1) in regression format as,

$$\bar{y} = \alpha\bar{y}._* + \beta\bar{y}_{*}. + u\bar{1} + \bar{\varepsilon}$$

and then premultiply by W^{-1} , we obtain the transformed equation,

(4) $$\bar{Y} = \alpha\bar{Y}_{*}. + \beta\bar{Y}._* + u\bar{J} + \bar{E} \ ;$$

or, in scalar notation,

(4a) $$Y_{ij} = \alpha Y_{i*} + \beta Y_{*j} + U + E_{ij} \ .$$

Time Conformability

In order to match hours worked with wages received, we have had to assume, firstly, that the LFS data available to IMPACT, corresponding to the May survey of each year, are representative of the entire calendar year. Secondly, to ensure an improved conformability with the financial year information of the IDS, we have averaged the LFS data from the appropriate consecutive calendar years. However, as no 1969 LFS information could be provided in the special tabulations required, we assume that the August 1968 LFS data is representative of the financial year 1968/69.

Deconfidentialization

Before supplying the labour force data to IMPACT, the ABS processes information in cells which are considered confidential; i.e., those cells which contain a non-zero size of population estimate of less than 500. The average value in confidential cells in a particular tabulation is substituted for the actual estimate.

The deconfidentialization process introduces a serious loss of information in small sized cells (although the standard errors on the actual estimates would be very large in any case). Further, it introduces error into the row and column totals in a tabulation, since the totals are calculated after deconfidentialization has taken place.

An indication of the extent of confidentiality is given in the following table.[1]

TABLE 3 : PERCENTAGE OF PERSONS IN CONFIDENTIAL CELLS

	LFS	IDS
1968 (/69)	6.3	10.0
1973 (/74)	6.6	9.1
1974	5.9	

Truncated Expansion Factors

In essence, the published population estimates are generated from the survey results by the following procedure.[2] Population estimates from some other source are classified by State, region, age and sex to form population benchmarks.[3] The population estimate from the survey is calculated as the ratio of the number of persons with the characteristic being measured in the benchmark cell to the total number of persons in the sample in that benchmark cell, multiplied by

1. The calculations in Table 3 were made for a 27 occupation by 25 industry disaggregation of the workforce.
2. For a more detailed outline of the estimation procedure, see, for example, the explanatory notes in ABS, The Labour Force 1972, Canberra, 1974 (Ref. 6.22), p. 4 and p. 54.
3. There are 8 State, 2 region, 12 age and 2 sex classifications, implying 384 benchmarks, and benchmark cells.

where

y_{ij} = element in the ith row (occupation) and jth column (industry) in the body of the O:I table,

y_{i*} = aggregate mean value for occupation i (margin value for row i),

y_{*j} = aggregate mean value for industry j (margin value for column j),

u = a constant,

ε_{ij} = error term.

However, as was mentioned previously, the characteristics, y_{ij}, do not exhibit constant variance, and thus introduce heteroskedasticity in the disturbances, ε_{ij}. We can write

$$E\left[\bar{\varepsilon}\ \bar{\varepsilon}'\right] = \lambda V \qquad (2)$$

where

ε = vector of disturbance terms of order $(nm \times 1)$,

λ = unknown (constant) scalar,

V = known symmetric positive definite matrix of order $(nm \times nm)$.

The assumption contained in (2) is that we know the variances and covariances of the dusturbance term up to some scale factor. In fact, we have insufficient information to estimate covariances and consequently specialize V to be a diagonal matrix.

We assume that the variance of the disturbance term is proportional to some index of the variance in the dependent variable, y_{ij}. We define a matrix, W^{-1}, in terms of the standard errors in y_{ij}

(4) Although the professionals had the greatest dispersion in earned income in 1968/69, the distribution of their incomes narrowed markedly by 1973/74. The dispersion in the earned incomes of unskilled workers also narrowed over this interval, but by a much smaller margin.

(5) The hours worked of unskilled blue collar workers shows greatest variation among the occupations, principally due to the large proportion of rural workers in this occupation (20%). Apart from this, professionals show the greatest variation, followed by unskilled white collar workers. Note also that the dispersion for professionals increased markedly by 1973/74.

Regression Analysis

In order to capture the occupation and industry influences on variability in hours worked and earned income in a systematic way, we conduct simple regression analysis. We seek to explain variations in characteristics at the O:I level in terms of the occupation (mean) effect and the industry (mean) effect, using the model[1] :

$$(1) \qquad y_{ij} = \alpha y_{i*} + \beta y_{*j} + u + \varepsilon_{ij} \quad , \qquad \begin{matrix} i=1,\ldots,n \\ j=1,\ldots,m \end{matrix}$$

1. We stress that none of the models in this section constitutes a model of wage determination from the point of view of economic theory. Numerous studies (see Rosen (1977)) examine age earning profiles usually with cross-section data and variables such as family background, status and educational attainment are used to explain wage differences. However, these studies tend to avoid the influence of specific occupation and industry effects on earnings.

the population benchmark estimates (or expansion factor). This ensures that the survey estimates conform to the previously estimated age and sex distributions, and tends to offset any age and sex stratification errors in the sample.

The procedure presents problems when the independent population estimates and benchmarks become revised, indicating that previous survey estimates have been biased by less reliable benchmarks.

In the special tabulations, the problem is exacerbated by the use of a single overall average expansion factor, rather than the area, sex and age specific expansion factors.

The average expansion factor error in LFS tabulations is given in the following table. The same order of error would apply to the IDS estimates.

TABLE 4 : AVERAGE LFS EXPANSION FACTOR ERROR[1]

LFS	Estimates of Persons Employed		
	Published ('000)	Impact ('000)	% Difference
1968	5003.5	4960.4	-0.9
1973	5575.8	5546.4	-0.5
1974	5750.1	5730.4	-0.3

1. The average expansion factor used to generate the special tabulations would reflect the original expansion factors used for published data and not factors revised in the light of updated population estimates.

Classification Concordance

The survey estimates are transferred from the original occupation and industry classifications to IMPACT classifications of 27 occupations and 25 industries. However, the original set of classifications used in the surveys has changed over time with varying degrees of concordance with the IMPACT classifications. For example, the 1968/69 IDS estimates have been stored on a 2 digit (Classified List of Industries) (CLI) basis, the LFS of 1968 on a 3 digit CLI, and both surveys after 1972 on an Australian Standard Industrial Classification (ASIC) basis.

TABLE 7 : COEFFICIENTS OF VARIATION FOR WEEKLY HOURS OF WORK PER PERSON BY INDUSTRY AND OCCUPATION FOR 1968/69 AND 1973/74

INDUSTRY	OCCUPATION: Professional 1968	Professional 1973/74	Skilled White Collar 1968	Skilled White Collar 1973/74	Unskilled White Collar 1968	Unskilled White Collar 1973/74	Skilled Blue Collar 1968	Skilled Blue Collar 1973/74	Unskilled Blue Collar 1968	Unskilled Blue Collar 1973/74	All Occupations 1968	All Occupations 1973/74
1	1.98	.72	.66	.68	.23	.90	.71	.52	2.25	2.51	2.17	2.42
2	.82	.35	1.91	.30	.66	1.39	.94	1.57	.32	.27	.41	.35
3	.14	.26	.41	.33	.17	.35	.19	.18	.21	.38	.21	.34
4	.49	.53	.79	.25	.55	.94	.14	.48	.40	.33	.39	.36
5	.30	.36	.20	.16	.26	.20	.30	.15	.14	.27	.11	.23
6	.73	.40	.12	.27	.31	.31	.02	.10	.13	.11	.15	.03
7	.60	.08	.23	.15	.24	.33	.12	.17	.04	.21	.10	.22
8	.50	.13	.93	.24	.46	.16	.21	.42	.16	.07	.28	.20
9	.50	.31	.32	.45	.22	.40	.07	.26	.27	.26	.28	.36
10	.79	.19	.40	.10	.36	.37	.20	.31	.08	.18	.17	.23
11	.17	.18	.24	.25	.26	.26	.11	.18	.06	.14	.13	.16
12	.22	.16	.40	.24	.03	.15	.09	.07	.07	.12	.06	.11
13	.14	.26	.29	.29	.21	.37	.04	.11	.15	.15	.10	.20
14	.59	.10	.33	.11	.32	.21	.44	.30	.22	.18	.09	.19
15	.49	1.11	.86	.13	.26	.34	.37	.16	.35	.07	.39	.14
16	.02	.23	.35	.22	.13	.33	.20	.30	.07	.20	.15	.27
17	.27	.11	.20	.15	.45	.53	.14	.08	.09	.16	.06	.14
18	.26	.27	.28	.28	.26	.31	.13	.03	.20	.24	.19	.23
19	.02	.52	.19	.22	.32	.33	.06	.38	.23	.21	.22	.26
20	.11	.57	.51	.28	.27	.30	.13	.22	.23	.34	.22	.30
21	.22	.26	.27	.30	.27	.31	.29	.24	.37	.48	.27	.32
22	.42	.36	.21	.27	.30	.38	.34	.42	.28	.30	.31	.35
23	.41	.74	.18	.30	.24	.38	.17	.21	.27	.38	.31	.49
24	.19	.66	.36	.38	.21	.35	.25	.19	.38	.49	.27	.37
All Industries	.37	.58	.03	.08	.26	.34	.05	.14	.64	.64	.42	.44

TABLE 6 : COEFFICIENTS OF VARIATION FOR ANNUAL EARNED INCOME PER PERSON BY INDUSTRY AND OCCUPATION FOR 1968/69 AND 1973/74

INDUSTRY	OCCUPATION: Professional 1968/69	Professional 1973/74	Skilled White Collar 1968/69	Skilled White Collar 1973/74	Unskilled White Collar 1968/69	Unskilled White Collar 1973/74	Skilled Blue Collar 1968/69	Skilled Blue Collar 1973/74	Unskilled Blue Collar 1968/69	Unskilled Blue Collar 1973/74	All Occupations 1968/69	All Occupations 1973/74
1	0	.47785	.41802	.36895	.83426	.74394	.66213	.81079	.96134	.80558	.95721	.80349
2	.31857	.49902	.82975	0	0	.31754	0	.23999	.61415	.52028	.61164	.51844
3	.47530	.36789	.69180	.56886	.55140	.59264	.34744	.41134	.50062	.46801	.54112	.49295
4	0	.19096	0	.13793	0	0	0	.48941	0	.35167	0	.41835
5	.39329	.25267	.51191	.48601	.52115	.56178	.53549	.45068	.47301	.51553	.55318	.53315
6	.43276	.18270	.57104	.43538	.56191	.54034	.70854	.44529	.58463	.47116	.74661	.63854
7	.47071	.41871	.58884	.48102	.68845	.53378	.48906	.48296	.60085	.54954	.64364	.57595
8	.38942	.36150	.54869	.39113	.39198	.41208	.36928	.21133	.44551	.37094	.52551	.44412
9	0	.26735	.14943	.47586	.25260	.57112	.47465	.22878	.41305	.57424	.55665	.53960
10	.60000	.42291	.47975	.43038	.56896	.46530	.33042	.38782	.45361	.38871	.50263	.45122
11	.53415	.31528	.52413	.45829	.41344	.54400	.46787	.43974	.45315	.49581	.53333	.51298
12	.73806	.29268	.63499	.39100	.50045	.42158	.47966	.44355	.47527	.44338	.60276	.46966
13	.35252	.47664	.58694	.40978	.59385	.52077	.46482	.45699	.53247	.47643	.59624	.53923
14	.56250	.60124	.56887	.35693	.56270	.32863	.48152	.27086	.61536	.45647	.77714	.45455
15	0	.58703	.38178	.64237	.51642	.64986	.62519	.50435	.86563	.63955	.84645	.64589
16	.32277	.26669	.40071	.42470	.47148	.46070	.41321	.37276	.38189	.42082	.47580	.45819
17	.55265	.30404	.57398	.51953	.69939	.68994	.49555	.46723	.50996	.47011	.58197	.52585
18	.60501	.53025	.61661	.53126	.73861	.76923	.60107	.48220	.56481	.65182	.79442	.73458
19	.34633	.40459	.60220	.39902	.50182	.54607	.35294	.42428	.47855	.46688	.58266	.51366
20	.20886	.28508	.48714	.42163	.57486	.51795	.35747	.32349	.53566	.48850	.52765	.46616
21	.63534	.39268	.62204	.52797	.74058	.66992	.63304	.32871	.65639	.74012	.83784	.71817
22	.48001	.45811	.53288	.44531	.57226	.54163	.37778	.40635	.52102	.57624	.63493	.57646
23	.84585	.58428	.72495	.67611	.63351	.60442	.80304	.50840	.63198	.65787	.92563	.70925
24	.61981	.79100	.78034	.62411	.96250	.87158	.51918	.68004	.88831	.83462	.97300	.86190
All Industries	.73730	.53045	.69402	.58316	.70227	.68572	.50063	.46614	.70883	.64113	.74840	.65425

4. ASSESSMENT OF DATA RELIABILITY

Some comparisons of estimates from different sources were conducted in an attempt to assess the extent to which the factors mentioned in the last section have biased the information obtained from Labour Force Surveys. Detailed comparisons with published data are not possible because of (i) the general lack of occupation and industry disaggregation in published data and (ii) the imprecise concordance between industry and occupation classifications used in published estimates and the IMPACT classifications used in the tabulations supplied. Consequently, emphasis was placed on internal comparisons, i.e., analysis of changes in variables from survey to survey and comparison of estimates of the same variable arising from tabulations corresponding to different surveys.

One test of the data is to compare estimates of the numbers of persons employed arising from different sources. For example, a comparison of industry totals from the 1971 LFS and the 1971 Census exposed differences within acceptable limits. However, a comparison between the LFS and IDS at the O:I level indicated widespread deficiencies.

The LFS/IDS comparison was conducted with a 5 way occupation split. The five occupations chosen reflected the major types of occupation of interest (in the light of the need to preserve homogeneity within occupations and heterogeneity between occupations) and avoided analysis of large numbers of zeros in O:I cells. The five occupations used were[1] :

1. The relationship between these groupings and standard IMPACT groupings is demonstrated in Appendix 1.

Professional
Skilled White Collar
Unskilled White Collar
Skilled Blue Collar
Unskilled Blue Collar

Preliminary analysis showed major discrepancies in some industry classifications. To combat these major deficiencies, the following industries were aggregated.

Industries Aggregated	Reasons for Aggregation
Agriculture Forestry & Fishing	Imprecise concordance. Confidentiality[1] High standard errors of estimate[2]
Mining - Metallic & Non-metallic Minerals Mining - Coal and Crude Petroleum	Not differentiated in 1968/69 IDS
Basic Chemicals Petroleum Refining	Imprecise Concordance

This left a disaggregation of the survey estimates by 5 occupations and 21 industries (105 cells).

With a more systematic analysis, further deficiencies became apparent. Comparisons of estimates of persons employed were made between contemporaneous IDS and LFS tabulations. Specifically, we determined whether the difference between estimates from the two sources was significantly

1. The small numbers in occupations other than rural workers (included in Unskilled Blue Collar) makes a distinction between these two industries pointless.

2. The ABS states that standard errors of estimate for agricultural industries are higher than for other industries.

CV is a unit free index of relative variability and can be used for comparative purposes.

It should be stressed that the calculated CV may, in some cases, be a poor representation of the variation in the population, for two reasons. First, the use of midpoints (representative points) in income (hours) ranges may be a poor approximation. Second, the calculated variance ignores the influence of sampling error. This latter factor is likely to be particularly serious in cells with small estimated populations and at the "tails" of the distribution across measurement intervals.[1]

CV's for each O:I for earned income in 1968/69 and 1973/74 are presented in Table 6, and CV's for hours worked for 1968 and 1973/74 in Table 7. The main features of these tables are summarized below.

(1) CV's for weekly hours of work have a wider range of values at a low level of aggregation than those for annual income earned. However, in general, variation is higher in earned income than in hours worked.

(2) Although average wages have increased over the five year period, variation in earned incomes has, in general, decreased.

(3) At the highest level of aggregation the variation in hours worked has remained stable. The overall average hours worked have declined by about half an hour per week.

1. Because of larger numbers, the row and column total CV's are likely to be more accurate than individual cell CV's.

6. ANALYSIS OF VARIABILITY IN LABOUR FORCE CHARACTERISTICS

Using the various sets of information on labour force characteristics, variances, and standard errors we conduct a preliminary analysis of variability in labour force characteristics. We are not concerned here with a determination of factors affecting characteristics such as the relationship between hours worked and income earned in the context of a labour demand-supply model.[1] Our concern is with the extent of variability in labour force characteristics at the O:I level, and, in particular, with the relative influence of the industry and the occupation components in explaining variability.

We proceed in two ways. First, we present coefficients of variation as an index of relative variability in each O:I cell. Second, we use weighted and unweighted regression analysis to test hypotheses regarding the extent to which variability of each O:I characteristic may be explained in terms of the industry component and the occupation component.

Coefficients of Variation

The coefficient of variation (CV) is defined as the standard deviation of a characteristic divided by its mean value. The standard deviation is simply the square root of the variance of the characteristic across individuals at the O:I level, which was calculated in the last section as one element of the standard error. Since the standard deviation is corrected for order of magnitude by division by the mean value, the

1. This task is the concern of Parham & Ryland (1978).

different from zero at various levels of confidence. The numbers of cells (out of a total of 105) in which the difference in population estimated in the LFS and the IDS was significantly different from zero are presented in Table 5.[1]

TABLE 5 : NUMBER OF CELLS IN WHICH ESTIMATES OF POPULATION FROM IDS AND LFS ARE SIGNIFICANTLY DIFFERENT

Comparison	Level of Confidence (Percent)		
	90	95	99
1968 LFS / 1968/69 IDS	24	18	11
1973/74 LFS[a] / 1973/74 IDS	27	19	7

a Average of the 1973 and 1974 LFSs.

The number of failures is sufficient to suggest serious deficiencies in the data base. We do not report other comparisons conducted (e.g., year-to-year changes in the LFS and LFS versus Census), since they do not provide further information on the degree of unreliability in the data base.

1. The standard error of the difference between the estimates of population from the two sources is calculated assuming the distribution of the results of each survey to be independent of each other. The standard errors of estimate of each survey were generated from power curves fitted to published data (see note to Table 1 on p. 9).

5. DISPERSION OF A CHARACTERISTIC AND POPULATION SIZE : STANDARD ERROR

There are significant deficiencies in the data base of labour force characteristics, arising largely from sampling error. However, even apart from this factor, there are different degrees of variability in a characteristic about the mean value for each O:I cell observation. It is important, particularly for any subsequent regression analysis, that each O:I cell observation on a left hand variable should exhibit constant variance. Put another way, in any estimation process each observation should be given a weight which varies inversely with its variance.

We are able to calculate variances of variables for each O:I cell. Each observation on a characteristic at the O:I level is an average over all individuals in that cell. In the ij^{th} cell, the variance of this (mean) value is equal to the variance (σ_{ij}^2) across individuals divided by the total number p_{ij} of individuals in the cell. In subsequent discussion, we refer to the square root of the variance in a characteristic, viz. the standard error, $SE = \sigma_{ij}/\sqrt{p_{ij}}$.

Population variances in earned income were calculated as follows. Numbers of people in each O:I cell from the IDS tabulations are distributed across 13 income ranges. Neglecting variation within ranges (which is tantamount to assuming that the representative person earns the midpoint of each income range),[1] it is a simple exercise to calculate the mean and variance of income per person.

1. To use the midpoint, it is necessary to assume that the distribution of persons in an income range is uniform about the midpoint. The likely imprecision of this assumption is recognised, but we maintain it in absence of further information.

The calculation of variance in hours worked was more problematical. The ABS tabulations provided estimates of total hours worked, calculated from the number of hours worked reported by survey respondents, as well as numbers of people distributed across 5 hours worked ranges. However, midpoints were found to be unsatisfactory as representative points in the hours ranges since they could not be used to replicate the total numbers of hours supplied by the ABS.[1] Consequently, to obtain the approximate representative points for each hours range so as to minimize variance, we solved a quadratic programming problem subject to the restrictions that the representative point of each class interval was bounded by the upper and lower end-points of each class interval.[2]

The SE can be included in any estimation process (see, for example, the next section) to induce constant variance across observations at the O:I mean level. It incorporates further information into the estimation process on the variance in individuals' characteristics and the population size of each O:I cell. The effect of the incorporation of the standard error in the estimation process is to give greater weight to observations with smaller variances.

1. Note that this exercise was impossible for the income data, since precise income figures were not supplied by survey respondents and therefore an estimate of total actual income is not available.

2. The problem is to find values of X_i such that $\sum_i (X_i - \bar{X})^2$ is minimized where $\bar{X}$ is mean hours worked over all O:I classifications subject to $\sum_i A_{ij} X_i \leq b_j$ where A_{ij} is number of persons in each class interval i and O:I classification j and b_j is number of hours worked for a particular O:I classification. Bounds were imposed on X_i.