Working Paper No. 108A
Industrial Relations Section
Princeton University
June 1979

Estimating the Union/Non-Union Wage Differential:

A Statistical Issue

by

Charles Mulvey and John M. Abowd*

* University of Glasgow and University of Chicago, respectively. An earlier version of this paper was circulated as Princeton University, Industrial Relations Section Working Paper No. 108 and was written while the first author was visiting Princeton at the Industrial Relations Section. The current version was written while the second author was visiting the Centre for Labour Economics (C.L.E.) at the London School of Economics. Helpful comments from Daniel Hamermesh, Gregg Lewis, Sherwin Rosen and especially Orley Ashenfelter are gratefully acknowledged. Financial support was provided by the S.S.R.C. and the C.L.E.

ESTIMATING THE UNION/NON-UNION WAGE DIFFERENTIAL: A STATISTICAL ISSUE

by

Charles Mulvey and John M. Abowd*

Prior to the availability of micro-data sets for the U.S. which indicated the union or non-union status of individuals, almost all estimates of the conditional average union/non-union wage differential, given union membership, were made using mean average wages from a sample of industries or occupations. The statistical and economic methodology employed in such studies is basically the same. It involves (a) the construction of a statistical relationship between the average wage and the extent of unionization and (b) the use of a maintained hypothesis about the economic structure of union/non-union wages under which the resulting estimate of the average conditional differential is unbiased or, at least, consistent. We show below that this incomplete data method has produced biassed estimates of the union/non-union wage differential when applied to data from the U.K.

This is an important matter. In Britain a number of estimates of the differential have been made using this incomplete data method. A concensus has emerged that the differential is around 25%.[1] We show below that this estimate is quite likely to be biassed by as much as 50 per cent. It is worthwhile stressing at this point that we are not referring to (potentially) biassed estimates of the structural effect of unionism on wages[2] but rather to (observably) biassed estimates of the conditional effect of unionism on wages.[3] Before U.S. estimates based on individual data became available, there was some concensus that the average differential was between 20 and 30 per cent.[4] Micro-data estimates now put the U.S. figure at around 12%.[5] There are no comparable micro-data estimates for Britain. Hence, it is important to know if the incomplete data estimates for Britain may be subject to the same biasses as apparently arise when the

procedure is used in the U.S. In the next section we outline the differences between the (conventional) incomplete data method and some straightforward complete data methods. In the last section we apply both methods to industry average data from the New Earnings Survey (1974) and examine the resulting biasses.

The Estimation Methods

The incomplete data method is characterised by the fact that only an average wage W_i and the extent of unionism U_i are observed for each observation i. The categories might be industries or occupations or some other grouping. We have used industries in our empirical analysis so we will refer to the observation unit as an industry. Let W_{ui} be the average union wage in industry i and W_{ni} be the average non-union wage in industry i. A complete data method is characterised by the fact that W_{ui}, W_{ni}, and U_i are all observable. The conventional incomplete data model assumes that the regression function for ℓnW_i is given by

$$(1) \quad E\left[\ell nW_i \middle| U_i\right] = U_i E\left[\ell nW_{ui} \middle| U_i\right] + (1 - U_i) E\left[\ell nW_{ni} \middle| U_i\right]$$

where $E\left[\tilde{y} \middle| x\right]$ denotes the conditional expectation of (random) y given x, and ℓn denotes the natural logarithm. Equation (1) is often transformed into

$$(2) \quad E\left[\ell nW_i \middle| U_i\right] = E\left[\ell nW_{ni} \middle| U_i\right] + \phi_i U_i$$

where ϕ_i is, by definition, the conditional differential between ℓnW_{ui} and ℓnW_{ni} in industry i. By contrast, a complete data method makes use of $E\left[\ell nW_{ui} \middle| U_i\right]$ and $E\left[\ell nW_{ni} \middle| U_i\right]$ directly, since W_{ui} and W_{ni} are both observed.

In order to actually use equation (2), assumptions must be made about the functional form of $E\left[\ell nW_{ni} \middle| U_i\right]$ and the incidental parameter ϕ_i must be eliminated. In practice, the non-union wage is usually taken to depend on a vector of exogenous characteristics of the workers and firms in industry i. The incidental parameter ϕ_i is eliminated by assuming either independence as

(3a) $\phi_i = \phi$

or linear dependence as

(3b) $\phi_i = \phi + \gamma(U_i - \bar{U})$

where ϕ represents the average differential and $\bar{U}$ is the average extent of unionism. Any assumption about the dependence of $E\left[\ell nW_{ni} | U_i\right]$ on exogenous characteristics is, of course, implicitly an assumption about the structural relationship between ℓnW_{ni} and U_i. Biasses arising from errors of this type of assumption, while important, are not the subject of our analysis. Biasses arising from counterfactual assumptions used to eliminate the incidental parameter ϕ_i -- that is, errors in equations (3a) or (3b) -- are the main subject of this analysis. The parameter of interest, however, is ϕ, the conditional mean of the industry-by-industry union/non-union wage differential, given the inter-industry extent of unionism. In practice, ϕ corresponds to the average wage differential across industries and may always be defined as

$$(4)\quad \phi = E\left[\sum_{i=1}^{N} (\ell nW_{ui} - \ell nW_{ni})/N \,\middle|\, U_1, \ldots, U_N\right]$$

where N is the number of industries used in the analysis. In addition, the possibility of linear or higher order dependence of ϕ_i on U_i can always be expressed using the regression function of $\ell nW_{ui} - \ell nW_{ni}$ on powers of U_i. In the complete data model, therefore, ϕ can always be estimated without bias using equation (4). In addition, no potentially counterfactual assumptions need to be made about $E\left[\ell nW_{ni} | U_i\right]$ or the dependence of ϕ_i on U_i.

Our analysis is designed to isolate the potential biasses in the incomplete data method by comparing estimates of ϕ based on the regression functions

$$(5a)\quad E\left[\ell nW_i | U_i\right] = E\left[\ell nW_{ni} | U_i\right] + \phi U_i + (\phi_i - \phi)U_i$$

and

$$(5b)\quad E\left[\ell nW_i | U_i\right] = E\left[\ell nW_{ni} | U_i\right] + \phi U_i + \gamma(U_i - \bar{U})U_i + (\phi_i - \phi - \gamma(U_i - \bar{U}))U_i.$$

Equations (5a) and (5b) are derived from equations (2), (3a) and (3b). Our complete data analysis is based on the regression function

$$(6)\quad E\left[\ell nW_{ui} - \ell nW_{ni} | U_i\right] = \phi + \gamma(U_i - \bar{U}) + \gamma_2(U_i - \bar{U})^2 + \gamma_3(U_i - \bar{U})^3$$

where ϕ and γ are the same parameters as arise in the incomplete data analysis, and γ_2 and γ_3 allow for higher order dependence on U_i.[6] The statistical origins of the potential biasses in estimates of the parameters of equations (5a) and (5b) are, formally, (a) the ommission of the last term in either equation (5a) or (5b); (b) incorrect specification of $E\left[\ell nW_{ni} | U_i\right]$; (c) measurement error if ℓnW_{ni} is used for $E\left[\ell nW_{ni} | U_i\right]$; and (d) incorrect functional form in (5a) or (5b) arising from non-zero γ_2 or γ_3 in equation (6). Since ϕ is the conditional mean differential, simultaneous determination of U_i and the wages ℓnW_{ui} and ℓnW_{ni} is not a potential source of bias.

The Empirical Analysis

Our data are drawn from industry average salaries in the New Earnings Survey with averages taken at the two digit classification level.[7] All data required to compute our regressions are contained in Appendix A. The variable definitions are:

W_{ui}: average hourly wage of all workers in industry i covered by a collective bargaining agreement;

W_{ni}: average hourly wage of all workers in industry i not covered by a collective bargaining agreement;

U_i : percentage of all workers in industry i covered by a collective bargaining agreement;

L_i : total employment in industry i.

The complete data analysis is reported in Table 1 while the incomplete data analysis is given in Table 2.

Since the complete data method is free of the potential biasses discussed above, we discuss these results first. Columns (1) and (5) of Table 1 report

the unweighted and weighted estimates of ϕ implied by the observed industry differentials.[8] The 90% confidence interval for the unweighted estimate of ϕ is $.0574 \pm .0207$. This implies a union/non-union wage differential in the 90% confidence interval of 3.75% to 8.13%. The 90% confidence interval for the wage differential implied by the weighted estimate of ϕ is 4.32% to 9.14%. Columns (2), (3) and (4) of Table 1 report the unweighted estimate of the linear, quadratic and cubic versions of equation (6). The reported F statistic tests the joint hypothesis that all the powers of U_i estimated in a given column are zero. All these F tests fail to reject the null hypothesis at conventional significance levels. We conclude that these data show no relationship of the differential with any reasonable power of U_i. The results in columns (6), (7) and (8) show exactly the same result for the weighted regressions.

Table 2 gives the results of the incomplete data analysis when the specifications in equations (5a) and (5b) are used. The first row shows the coefficient estimate on U_i which is always the implied estimate of ϕ. Columns (1) and (3) are the ordinary least squares estimates of the two incomplete data specifications. The estimates can be seen to be very different from each other and from the unbiassed estimate in Table 1. Column (1) implies a 90% confidence interval for ϕ of $.1282 \pm .1283$. The implicit estimate of the union/non-union wage differential is between zero and 29.60%. This is much more imprecise than the interval implied by the unbiassed estimate. In addition, the biassed point estimate of 14.00% is outside the 90% confidence interval of the unbiassed estimate reported in Table 1. The bias is 8.07 percentage points or over 57% of the reported point estimate. The point and interval estimates of ϕ implied by column (3) of Table 2 are wholly unreasonable. As Table 1 shows, there is no linear relationship between the differential and U_i in these data. We should not, therefore, expect the counterfactual assumption (3b) embodied in equation (5b) to give a sensible result and it doesn't. Columns (2) and (4) use an instrumental variable procedure to check for the possibility of measurement error in ℓnW_{ni} as a potential

problem in columns (1) and (3), respectively.[9] Our conclusions would be unchanged using these instrumental variable results instead of the ordinary least squares results. Columns (5) - (8) repeat the analysis of columns (1) - (4), respectively using appropriate weighted procedures. Once again, our conclusions would not be affected by using these results in comparison with the weighted results of Table 1. In fact, every column other than column (1) of Table 2 implies a more serious bias than the bias discussed in relation to the estimate in column (1).

Conclusions

Our major conclusion is that the possibility of bias as large as 50% or more in the estimate of the union/non-union wage differential in Britain is quite likely. All estimates based on the incomplete data methods we have discussed are suspect. This includes many of the estimates reviewed by Metcalf (1977). Further analysis using the New Earnings Survey is clearly in order.

Footnotes

1 See Metcalf (1977) for a review of this literature.

2 We mean here the effect of an increase in unionism on union/non-union wages in a model where unionism is also endogenous.

3 We mean here the effect of an increase in unionism on union/non-union wages when unionism is exogenous.

4 See, especially, Lewis (1963); and Weiss (1966), Throop (1968), and Rosen (1969).

5 Every survey or panel data wage regression including a union member dummy variable or using separate equations for union and non-union members generates an estimate of this parameter. In this context hundreds of such estimates have been produced based on the Panel Study of Income Dynamics, the National Longitudinal Study, the Current Population Survey, the public use U.S. census data and other micro-data sets. We are not aware of any complete survey of these estimates, although Lewis has privately tabulated many such results. Rees (1976) noticed the same point.

6 Note that when γ_2 or $\gamma_3 \neq 0$ the parameters ϕ and γ are not exactly as in the incomplete data analysis. Specifically, they now refer to the value of equation (6) at $\bar{U}$ rather than to the value at $\bar{U}$, $\overline{(U - \bar{U})^2}$ and $\overline{(U - \bar{U})^3}$.

7 The data were extracted with the generous cooperation of the Department of Employment with financial support from a grant by the S.S.R.C. to the Department of Social and Economic Research, University of Glasgow.

8 All weighted estimates have been computed using L_i (total industry employment) as the weight.

9 The implied reduced form regression of ℓnW_{ni} on $U_i - \bar{U}$ and its square and cube fits well: $R^2 = .38$, $F = 4.48$ with (3,22) degrees of freedom, P-value = .013.

References

Department of Employment (1974). New Earnings Survey 1973, H.M.S.O.

Johnson, G.E. (1977). 'The Determination of Wages in the Union and Non-Union Sectors', British Journal of Industrial Relations, Vol. XV (July) pp.211-225.

Lewis, H.G. (1963). Unionism and Relative Wages in the United States, Chicago University Press.

Metcalf, D. (1977). 'Unions, Incomes Policy and Relative Wages in Britain', British Journal of Industrial Relations, Vol. XV (July) pp.157-175.

Rees, A. (1976). 'H. Gregg Lewis and the Development of Analytical Labor Economics', Journal of Political Economy, Vol. 84 (August) pp.53-58.

Rosen, S. (1969). 'Trade Union Power, Threat Effects and the Extent of Organization', Review of Economic Studies, Vol. 36 (April) pp.185-196.

Throop, A.W. (1968). 'The Union/Non-Union Wage Differential and Cost-Push Inflation', American Economic Review, Vol. 58 (March) pp.79-99.

Weiss, L. (1966). 'Concentration and Labor Earnings', American Economic Review, Vol. 56 (March) pp.96-117.

TABLE 1

Regression Estimates of the Conditional Union Wage Differential using the Dependent Variable $\ell nW_u - \ell nW_n$. (Standard Errors in Parentheses)

	(1)	(2)	(3)	(4)	(5)[a]	(6)[a]	(7)[a]	(8)[a]
Constant [ϕ]	.0574 (.0121)	.0574 (.0124)	.0457 (.0171)	.0426 (.0182)	.0648 (.0132)	.0648 (.0137)	.0482 (.0188)	.0450 (.0208)
$U - \bar{U}$		.0193 (.0780)	.0786 (.0984)	-.0034 (.1750)		.0207 (.0968)	.1647 (.1486)	.1216 (.1867)
$(U - \bar{U})^2$			.4667 (.4722)	.8261 (.7915)			.8253 (.6523)	1.2803 (1.3316)
$(U - \bar{U})^3$				1.8622 (3.2638)				1.7008 (4.3136)
S.E.E.	n.a.	.0629	.0630	.0640	n.a.	.0671	.0650	.0647
R^2	n.a.	.0025	.0432	.0571	n.a.	.0019	.0669	.0734
F	n.a.	.0610	.5190	.4444	n.a.	.00008	.8239	.5809
d.f.	n.a.	(1,24)	(2,23)	(3,22)	n.a.	(1,24)	(2,23)	(3,22)

[a]Computed using GLS with total employees per industry as the weights. R^2 and F have been corrected.

TABLE 2

Regression Estimates of the Conditional Union Wage Differential Using the Dependent Variable $\ell n(U \cdot W_u + (1-U) \cdot W_n)$. (Standard Errors in Parentheses)

	(1)	(2)[a]	(3)	(4)[a]	(5)[b]	(6)[c]	(7)[b]	(8)[c]
U [ϕ]	.1282 (.0746)	.1745 (.1210)	-.0606 (.2845)	-.0404 (.4727)	.1787 (.0940)	.2355 (.1450)	-.0517 (.4241)	-.0337 (1.5937)
$U(U - \bar{U})$			.2799 (.4067)	.2598 (.5532)			.3382 (.6064)	.3175 (1.8717)
$\ell n\ W_n$	.8758 (.1026)	.7525 (.2702)	.8964 (.1080)	.8772 (.3752)	.7844 (.1257)	.6397 (.3045)	.8159 (.1397)	.8042 (1.0152)
Constant	-.0882 (.0770)	-.1582 (.1619)	.0565 (.2244)	.0361 (.4432)	-.1516 (.1002)	-.2378 (.1938)	.0391 (.3567)	.0215 (1.5405)
S.E.E.	.0516	.0532	.0522	.0600	.0544	.0559	.0540	.0540
R^2	.8437	.8339	.8470	.8468	.7717	.7586	.7749	.7748
F	62.09	57.74	40.60	40.53	38.88	36.14	25.25	25.23
d.f.	(2,23)	(2,23)	(3,22)	(3,22)	(2,23)	(2,23)	(3,22)	(3,22)

[a]Computed using 2SLS with instruments: $U - \bar{U}$, $(U - \bar{U})^2$, and $(U - \bar{U})^3$. F is asymptotically chi-square.

[b]Computed using GLS as in fn. a, Table 1. R^2 and F have been corrected.

[c]Computed using 2SLS as in fn. a and GLS as in fn. a, Table 1.

Appendix A

SIC(1968)	W_u	W_n	U	L
I	.59	.57	.440	1056
II	.90	.80	.962	2149
III	.78	.81	.701	2497
IV	.95	.95	.834	177
V	.87	.87	.803	1612
VI	.89	.83	.938	2562
VII	.83	.81	.840	3831
VIII	.79	.79	.684	390
IX	.83	.82	.834	2140
X	.87	.82	.951	806
XI	.99	.86	.953	3360
XII	.83	.79	.749	1968
XIII	.76	.76	.799	1690
XIV	.69	.70	.707	122
XV	.80	.73	.574	511
XVI	.82	.79	.822	1305
XVII	.83	.72	.772	1024
XVIII	.96	.80	.869	1901
XIX	.85	.77	.740	1072
XX	.81	.82	.890	6757
...[a]	...	...	...	...
XXII	.83	.71	.909	6601
XXIII	.72	.66	.541	2993
XXIV	.78	.74	.409	511
XXV	.67	.62	.882	1738
XXVI	.70	.61	.536	2417
XXVII	.68	.69	.983	2376

[a]XXI not provided by the D.E.