Fertig, Michael; Görlitz, Katja

Missing wages: How to test for biased estimates in wage functions?

Economics Letters

Provided in Cooperation with:
RWI – Leibniz-Institut für Wirtschaftsforschung, Essen

Suggested Citation: Fertig, Michael; Görlitz, Katja (2013) : Missing wages: How to test for biased estimates in wage functions?, Economics Letters, ISSN 0165-1765, Elsevier, Amsterdam, Vol. 118, Iss. 2, pp. 269-271,
https://doi.org/10.1016/j.econlet.2012.10.036 ,
http://www.sciencedirect.com/science/article/pii/S0165176512005897

This Version is available at:
https://hdl.handle.net/10419/148299

Missing wages: How to test for biased estimates in wage functions?

Michael Fertig [1]

Katja Görlitz [2, 1]

Abstract: This paper investigates how to test for nonresponse selection bias in wage functions induced by missing income information. We suggest an "easy-to-implement" approach which requires information on interviewer IDs and the interview date rather than hard to get interviewer characteristics.

Keywords: Item nonresponse, wages

[1] ISG - Institut für Sozialforschung und Gesellschaftspolitik GmbH, Barbarossaplatz 2, 50674 Köln, Germany

[2] RWI Essen - Rheinisch-Westfälisches Institut für Wirtschaftsforschung, Hohenzollernstraße 1-3, 45128 Essen, Germany

[1] All correspondence to Katja Görlitz, RWI Essen, Hohenzollernstraße 1-3, 45128 Essen, Germany, Phone: +49 201 8149268, Fax: +49 201 8149200, goerlitz@rwi-essen.de. All remaining errors are our own.

1. Introduction

It is well known that item nonresponse is particularly high when income and wage information is surveyed. If the response inclination is systematically related to wages, the estimates of wage equations could suffer from serious biases (Zweimüller 1992). Missing wages are still a rather neglected problem in empirical studies. A common way to deal with nonresponse is to eliminate missing cases. However, such a procedure assumes implicitly that wages are missing at random which seems to be at odds with the finding that item nonresponse on wage questions is more common in the tails of the income distribution (see e.g. Biewen 2001, Lillard et al. 1986).

The Heckman model is an appropriate framework to test for selection bias induced by missing information. One crucial requirement when applying this model is to find a valid exclusion restriction. A possible candidate could be taken from the interview situation, i.e. from the characteristics of the interview. For instance, Bollinger and Hirsch (2010) present two exclusion restrictions based on information on whether oral or telephone interviews were conducted and when interview performance was evaluated that both varied by calendar month of the survey. However, such variation in the interview situation by survey month might be a rather seldom case.

Another possibility is using interviewer characteristics because they are related to the response inclination (see e.g. Riphahn and Serfling 2005, Sousa-Poza and Henneberger 2000). Unfortunately, interviewer characteristics are often unobservable due to data protection regulations. Therefore, we suggest an exclusion restriction that only requires information on interviewer IDs and the date of the interview. Another advantage of our approach is that interviewer IDs and the date of the interview are observable for all observations, while interviewer characteristics are unavailable if some interviewers have refused to provide this information.

2. Data and empirical strategy

The empirical investigation is based on the German data set "WeLL" that was designed to analyze continuous training activities of individuals. The first wave covers 6,404 employees who were interviewed by telephone between October 2007 and January 2008.[3] In addition to information on continuous training activities, the data covers socio-demographic characteristics, education and job characteristics. For the analysis, individuals with no job and with no information on core variables were excluded (reducing the sample size by 3% and by 2%, respectively). The final sample consists of 6,054 observations.

A specific feature of the data is that additional information can be merged from administrative records of the social security system (which covers approximately 80% of the German workforce). The administrative data contains, amongst others, exact information on wages. However, German data protection regulations do not allow merging data from different sources without the respondent's approval. Therefore, in the WeLL questionnaire, respondents were asked to declare their agreement to link administrative data to their survey information. In the WeLL data, wages are missing for those 9% of respondents who did not agree.[4]

[3] For more information on the data set, see Bender et al. (2009).

[4] In many countries (including the UK and US), record linkage became a common tool in survey data, not only in health surveys but also in individual or household surveys (Jenkins et al. 2006, Sala et al. 2012).

To investigate whether missing wages induce a bias in a wage regression, a selection model is estimated (Heckman 1979). The outcome equation is a Mincer earnings function (Mincer 1974). The Probit selection equation indicates whether individuals agreed to merge data. The logarithm of the gross monthly wage is not observed for individuals who denied merging. The set of explanatory variables contains individual and job characteristics. In addition, the selection equation needs to include at least one variable that is related to the decision to declare agreement but unrelated to wages. To construct such an exclusion restriction, we suggest exploiting information on interviewer IDs (presented in more detail in the next paragraphs). The error terms are assumed to follow a bivariate normal distribution. If they are correlated with each other, missing wage information cannot simply be ignored and a sample selection correction needs to be incorporated in the wage regression. Estimation is carried out by Maximum likelihood (ML) as well as by the two step procedure.

As an exclusion restriction, a possible choice would be to use interviewer fixed effects. In our case, however, the selection equation cannot be estimated properly since 107 interviewers had an agreement rate of 100%. Thus, the 762 corresponding respondents had to be omitted from the Probit regression. In other settings, such an approach might be applicable, in particular, when missing wages occur more frequently or when the number of interviews per interviewer is large. Instead we construct an exclusion restriction based on information on interviewer IDs in addition to the date of the interview. The time period between the first and the last interview was 102 days. For each interviewer, we observe the day of each interview. These days are coded 1 for interviews on the first day, 2 for the second day […] and 102 for the last day. The exclusion restriction is generated by calculating the standard deviation of the interviewer-specific days.

The idea behind this measure is that interviewers being more intensely engaged in the survey (indicated by a low standard deviation) are more able to concentrate on the specific issues of the questionnaire. The question on record linkage is by no means standard to interviewers and it can induce further requests from respondents. More concentrated and focused interviewers might have higher agreement rates as they respond more adequately to queries or as they are perceived as being more sensitive or trustworthy. Those 18 interviewers having conducted only one interview were set to zero for the analysis. Since the assignment of interviewers to respondents is random and since it is the interviewer's choice to conduct the survey more or less intensely, we consider this exclusion restriction to be valid. In addition, when inserting this exclusion restriction in the wage equation, the coefficient becomes statistically insignificant which is interpreted as descriptive evidence of a valid exclusion restriction. Table 1 contains the description of all variables and sample means.

Table 1: Variable description and summary statistics

Variable	Description	Mean
ln(wage)	Logarithm of gross monthly wages (in Euro)	7.92
Male	Dummy: 1 for males, 0 otherwise	0.63
Married	Dummy: 1 for married employees, 0 otherwise	0.73
Children	Dummy: 1 for having children aged <=18 years, 0 otherwise	0.38
Male × Children	Interaction term between male and child	0.25
Years of schooling	Years of schooling	12.98
Potential experience	Age-years of schooling-6	26.17
Training incidence	Dummy: 1 for training participation in last two year, 0 otherwise	0.65
Tenure	Tenure in current job (months)	207.48
White collar employee	Dummy: 1 for white collar workers, 0 otherwise	0.65
Full time job	Dummy: 1 for full-time job, 0 otherwise	0.84
Temporary contract	Dummy: 1 for temporary contract, 0 otherwise	0.06
Agreement to merge wages	Dummy: 1 for agreement to merge data, 0 otherwise	0.91
Standard deviation of days conducting the interviews	Standard deviation of the days when interviewers have conducted their interviews	13.48

Notes: 6,054 observations (wages: 5,538 observations).

3. Results

Table 2 documents the main results showing that the coefficient of the exclusion restriction is statistically significant in the selection equation.[5] A higher standard deviation of days conducting the interviews is negatively related to the individual's likelihood to declare their agreement. This holds regardless of using the ML approach or the two-step procedure. The insignificant ρ indicates that there is no statistically significant correlation between the error terms of the wage and the selection equation. Thus, in our data, ignoring observations with missing wage information yields unbiased results (see also the coefficients of an OLS model estimated on the non-missing observations; Table 2, column 3).

Our result contrasts with the findings of Zweimüller (1992) who identifies a serious bias from ignoring missing cases. It is, however, similar to the conclusion drawn by Sousa-Poza and Henneberger (2000). Even though these studies directly investigate refusing wages which is different to our case of not declaring agreement to merge data, we still suggest that differences in the missing-wage rate could explain different results. Zweimüller (1992) is confronted with a missing-wage rate of almost 40%, Sousa-Poza and Henneberger (2000) face 14% and we have 9%.

[5] Among the other covariates, children, training and working full time are positively associated with declaring agreement on a statistically significant level. The coefficient of the interaction between child and male has a negative sign.

Table 2: Estimation results

	Heckman Selection Model				OLS Model,	
	Maximum Likelihood		Two-step Model		Missing Wages Deleted	
	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.
Selection equation						
Standard deviation of days conducting the interviews	-0.012 ***	0.004	-0.012 ***	0.004		
Covariates from wage equation	Yes		Yes			
ρ	-0.001		-0.04			
Wald test (p-value)	0.97					
Bootstrapped std. errors (p-value)			0.96			
Wage equation						
Male	0.206 ***	0.014	0.206 ***	0.014	0.206 ***	0.014
Married	-0.008	0.011	-0.008	0.011	-0.008	0.011
Children (y/n)	-0.081 ***	0.025	-0.082 ***	0.027	-0.081 ***	0.025
Male × Children	0.149 ***	0.028	0.150 ***	0.032	0.149 ***	0.028
Years of schooling	0.046 ***	0.003	0.046 ***	0.003	0.046 ***	0.003
Potential experience	0.014 ***	0.003	0.014 ***	0.003	0.014 ***	0.003
Potential experience squared	-0.0003 ***	0.000	-0.0003 ***	0.000	-0.0003 ***	0.000
Training incidence	0.121 ***	0.012	0.120 ***	0.020	0.121 ***	0.012
Tenure	0.001 ***	0.000	0.001 ***	0.000	0.001 ***	0.000
White collar employee	0.181 ***	0.014	0.182 ***	0.016	0.181 ***	0.014
Full time contract	0.642 ***	0.030	0.641 ***	0.033	0.642 ***	0.030
Temporary contract	-0.123 ***	0.026	-0.123 ***	0.026	-0.123 ***	0.026
Observations	6,054		6,054		5,538	
Censored Observations	516		516			
Uncensored Observations	5,538		5,538			

Notes: Standard errors are clustered at the interviewer level (287 clusters). Significance level: *** 1%, ** 5%.

4. Conclusion

This study shows that deleting missing wages is a valid way to deal with item nonresponse when using the WeLL data. Even though this result is not directly transferable to other data sets, our approach to test for selectivity is widely applicable. Especially when the number of missing cases is high, testing for selection bias (and if necessary correcting for it) is important and can be implemented by our approach. Furthermore, we suppose that its applicability is not only limited to the case of selectivity induced by merging data but also to cases in which wage questions are directly refused. This is because the necessary assumptions for the exclusion restriction are similar in either case.

Having access to interviewer IDs and information on the date of the interview enlarges the set of methods to test for selection bias due to missing information and, hence, helps to improve the quality of empirical work. Furthermore, with respect to data protection regulations both pieces of information are by far less problematic than interviewer characteristics. Survey administrators should, therefore, provide interviewer IDs in survey data.

Acknowledgements

The authors are grateful to Manfred Antoni, Alfredo Paloyo, Sandra Schaffner, Joel Stiebale, Marcus Tamm and an anonymous referee for helpful comments and suggestions. Financial support from the “Leibniz Gemeinschaft” is gratefully acknowledged.

References

Bender, S., Fertig, M., Görlitz, K., Huber, M., Schmucker, A., 2009. WeLL - Unique Linked Employer-Employee Data on Further Training in Germany. Journal of Applied Social Science Studies 129 (4), 637-643.

Biewen, M., 2001. Item non-response and inequality measurement: Evidence from the German earnings distribution. Allgemeines Statistisches Archiv 85, 409-425.

Bollinger, C. R., Hirsch, B. T., 2010. Is Earnings Nonresponse Ignorable? *Review of Economics and Statistics*, forthcoming.

Heckman, J. J., 1979. Sample Selection Bias as a Specification Error. Econometrica 47 (1), 153-161.

Jenkins, S. P., Cappellari, L., Lynn, P., Jäckle, A., Sala, E., 2006. Patterns of consent: evidence from a general household survey. Journal of the Royal Statistical Society. Series A (Statistics in Society) 169 (4), 701-722.

Mincer, J., 1974. Schooling, Experience, and Earnings. New York: National Bureau of Economic Research.

Riphahn, R. T., Serfling, O., 2005. Item non-response on income and wealth questions. Empirical Economics (30), 521-538.

Sala, E., Burton, J., Knies, G., 2012. Correlates of Obtaining Informed Consent to Data Linkage: Respondent, Interview, and Interviewer Characteristics. Sociological Methods & Research 41, 414-439.

Sousa-Poza, A., Henneberger, F., 2000. Wage data collected by telephone interviews: an empirical analysis of the item nonresponse problem and its implications for the estimation of wage functions. Zeitschrift für Volkswirtschaft und Statistik 136 (1), 79-98.

Zweimüller, J. 1992. Survey non-response and biases in wage regressions. Economics Letters 39, 105-109.