WORKING
PAPER

Measurement Error and Misclassification

A Comparison of Survey and Register Data

ARIE KAPTEYN
JELMER YEB YPMA

WR-283-1

November 2006

RAND LABOR AND POPULATION

Measurement Error and Misclassification: A Comparison of Survey and Register Data*

Arie Kapteyn

RAND, 1776 Main Street, Santa Monica, CA90407, U.S.A.

Email: Kapteyn@rand.org

Jelmer Yeb Ypma

RAND and University of Groningen

Email: Ypma@rand.org

October 9, 2006

Abstract

We provide both a theoretical and empirical analysis of the relation between administrative and survey data. By distinguishing between different sources of deviations between survey and administrative data we are able to reproduce several stylized facts in the literature. In doing so, we deviate from the almost universal assumption that the administrative data represent the truth. We illustrate the implications of different error sources for estimation in (simple) econometric models and find potentially very substantial biases, both when using survey data and when using administrative data. The analysis is applied to Swedish data that have been collected for a validation study as part of a larger European health and retirement study (SHARE: Survey of Health, Ageing, and Retirement in Europe). Thus this paper makes two contributions: (1) it adds to the limited number of empirical validation studies of earnings measurement in surveys and (2) it shows the sensitivity of some findings in the literature for the assumption that administrative data represent the truth. We find in particular that the common finding of substantial mean reversion in survey data largely goes away once we allow for a richer error structure.

*This paper is part of an NIA funded project (Comparison of Survey and Register Data: The Swedish Case, R03AG21780) in collaboration with Anders Klevmarken (Uppsala University) and Susann Rohwedder (RAND). We thank Susann Rohwedder, Anders Klevmarken and Fredrik Johansson for their helpful comments.

1 Introduction

Micro-data are an essential ingredient of research in economics and other social sciences. Such data, where information is available for each micro-unit (individual, firm, etc.) separately, is usually obtained through a survey or from registers.[1] Both methods of data collection have their advantages and disadvantages.

One problem in surveys is nonresponse, both to single questions (item nonresponse) and to the entire questionnaire (unit nonresponse). Another problem is measurement error. Furthermore, surveys are generally costly. In registers, like the Social Security Administration in the U.S., information is typically available for large numbers of individuals. These data are generally assumed to be reliable, but may not measure exactly the concept a researcher is interested in.

In this paper we concentrate on sources of measurement error in survey and administrative data by comparing individual survey information and administrative information on the same variables. By doing so, we both replicate a number of earlier studies using new data and propose extensions of models in the literature that may help us better understand the nature of measurement error in both survey data and administrative data.

The number of datasets allowing for validation studies of survey information appears to be quite limited. We discuss three of them here.[2]

A first example of a study comparing survey data and data from administrative sources is the Panel Study of Income Dynamics Validation Study (PSIDVS). In 1983 and 1987 a questionnaire, based on the PSID questionnaire, but shorter, was administered to employees of a manufacturing company in the Detroit-area to measure their earnings in the preceding years. At the same time payroll records of the employees were collected from this firm. The data is assumed to be very accurate, since the firm was highly cooperative (Pischke, 1995).

Duncan and Hill (1985) use the PSIDVS from 1983, which includes questions about annual earnings in the two preceding years (1981 and 1982). They find that means of ln(earnings) in the survey data and the validation data do not differ significantly. However, at the individual level there may still be substantial differences. Assuming the administrative data to contain the 'true' values, measurement error can be defined as the difference between survey data and administrative data. They find a reliability ratio between .64 and .84, depending on which year they look at, and whether outliers are removed or not. As expected, the reliability ratio is lower for the question with

[1] We will refer to such data as register or administrative data.

[2] Bound, Brown and Mathiowetz (2001) provide a more extensive overview of validation studies dealing with earnings measures.

a longer recall period, i.e. earnings in 1981 versus earnings in 1982. Pischke (1995) finds, when using data from 1982 and 1986, that administrative data and survey data do not differ much in either mean or variance. However, measurement error is found to be weakly negatively correlated with 'true' ln(earnings). In cross-section data of 1982 this correlation is stronger than in cross-section data of 1986. When restricting data to respondents who are present in both years, no significant correlation is found between 'true' ln(earnings) and measurement error. The negative correlation is not significant when leaving out hourly workers, i.e. looking only at salaried employees. Rodgers, Brown and Duncan (1993) and Bound et al. (1994) also report a negative correlation between measurement error and the true value. In contrast to Pischke (1995) they do not distinguish between hourly workers and salaried employees.

Another dataset suitable for comparison of survey and administrative data is a match constructed between the Current Population Survey (CPS) and data from the Social Security Administration (SSA) for the years 1976 and 1977 (e.g. Bound and Krueger, 1991). Once again the maintained hypothesis for most of their paper is that the administrative data are error-free. Using cross-section data, the reliability ratio for ln(earnings) for men is .844 in 1976 and .819 in 1977. For women this ratio is higher, .939 and .924. For men they find large negative correlations between measurement error and 'true' ln(earnings), $-.46$ in 1976 and $-.42$ in 1977. This correlation is small for women. Bollinger (1998) finds that the negative relationship between measurement error and earnings is mainly driven by overreporting among low earners.

When comparing their results obtained with the PSIDVS with the CPS-SSA data, Bound et al. (1994) find that qualitatively, the results are similar. However, they notice a large difference in the standard deviation of the measurement error (.13 in the PSIDVS data versus .32 in the CPS-SSA data). The difference seems to lie in the tails of the measurement error distribution, with very large outliers in the CPS-SSA data. Bound et al. (1994) suggest that these very large measurement errors are not necessarily due to misreporting in the survey, but rather due to errors in the SSA data.

A number of studies have addressed the possibility of measurement error in the earnings data collected in the Survey of Income and Program Participation (SIPP). We briefly discuss two studies relevant to our analysis. Pedace and Bates (2001) use a match between the 1992 SIPP longitudinal file and the Social Security Summary Earnings Records. These authors also assume the administrative data to represent the truth. It appears that respondents with low SSA earnings tend to overreport their earnings, whereas respondents with high earnings underreport. In contrast to the studies mentioned so far, Stinson (2002) does not make the assumption that the administrative data represent the truth. She instead estimates an earnings function allowing for (mutually uncorrelated) measurement error in both the survey and administrative data. She finds that both measures have similar magnitudes of measurement error, with the error in the administrative data being slightly larger than in the survey data.

The studies cited here reach somewhat contrasting conclusions regarding whether there is negative correlation between the 'true' value and measurement error or not. While this correlation seems evident in the CPS-SSA data and the SIPP data, in the PSIDVS data no such correlation is found for the salaried workers, but it is found when including hourly workers. This *mean reversion* in the measurement error has gradually attained the status of a stylized fact. Kim and Solon (2003) discuss its implications for the modeling of earnings dynamics.

In the remainder of the paper we will provide both a theoretical and empirical analysis of the relation between administrative and survey data. By distinguishing between different sources of deviations between survey and administrative data we will be able to reproduce several stylized facts in the literature. In doing so, we deviate from the almost universal assumption that the administrative data represent the truth.[3] We will illustrate the implications of these error sources for estimation in (simple) econometric models. The analysis is applied to Swedish data that have been collected for a validation study as part of a larger European health and retirement study (SHARE: Survey of Health, Ageing, and Retirement in Europe). Thus this paper makes two contributions: (1) it adds to the limited number of empirical validation studies of earnings measurement in surveys and (2) it shows the sensitivity of some findings in the literature for the assumption that administrative data represent the truth.

In Section 2 we describe the data. In Section 3 a relatively straightforward model of different error sources is proposed and their empirical implications are explored, both for the observed relation between survey and administrative data and for some econometric models incorporating survey and/or administrative data. Among other things we address the question when it is preferable to use administrative data and when survey data are to be preferred. Section 4 estimates the model of Section 3 using our Swedish dataset. We are able to identify various sources of error in the data and actually flag observations that suffer from different types of error. Section 5 concludes.

2 Data and Project Description

The Scandinavian countries have thirty years of experience with using administrative data for statistical purposes. The statistical offices in Denmark, Norway, Sweden, and Finland have made important progress in making information from various registers compatible and to ensure the link among various sources. The wealth of information

[3] As in most of the papers discussed above. The same assumption is made in more formal analyses of the use of validations samples, including Lee and Sepanski (1995) and Chen, Hong, and Tamer (2005).

contained in these registers is considerable. In our empirical work we will use Swedish data.

Every Swede has a unique social security number, which is also available in every register. In principle this allows interviewers to ask respondents for their social security number and permission to link them to the information available in the registers. If the respondent agrees this may substantially shorten interviews, because questions, for instance about income, can be skipped. Interviews can then focus on information *not* available in registers, while combining the interview information with the (presumably) more reliable register information.

The experiment generating our data is motivated exactly by this consideration. As part of SHARE (Survey of Health, Ageing, and Retirement), a pan-European data collection effort among individuals 50 and over, an experiment was devised in Sweden to assess the usefulness of combining survey and administrative data. The purpose of the experiment was to inform SHARE about the possibilities of combining such data and potentially implement it in other countries than Sweden. Aspects to be investigated include: (1) selectivity in survey responses in a number of important domains; (2) reliability of selected survey measures including income and pension entitlements by comparison with administrative data; (3) estimating biases in a number of important empirical relationships (e.g. health and socio-economic status (SES)) when using survey data rather than administrative data; (4) generalizing from these findings to the limitations of international comparisons if administrative data are available in some countries, but not in others. In the current paper we mainly concentrate on (2) and pay some attention to (3).

2.1 Administrative Data

In our empirical analysis we will use a sample of individuals from LINDA (Longitudinal Individual Data for Sweden) over age 50. LINDA is a register-based longitudinal database that is representative of the Swedish population since 1960. LINDA has two subsamples. The first subsample is the population sample, representative of the entire population with a coverage rate of 3.35%. The second subsample is the immigrant sample, covering 19.5% of the immigrant population. The samples are kept both cross-sectionally and longitudinally representative of their respective populations. There are two principal data sources: (1) the Income Registers, available annually since 1968 and (2) the Population Censuses, available every fifth year from 1960 to 1990 (no census has been taken after 1990). Other registers have been added during the nineties (see Edin and Frederiksson (2000) for more details).

By drawing from LINDA we have a wealth of information available about our respondents from the administrative data.[4] The information in the different registers is linked by the social security number of individuals (See Figure 1). Since the information comes from several registers an incorrect social security number in a register may lead to a wrong link.

As Abowd and Vilhuber (2005) notice, while looking at unemployment spells, errors in linking can be a real problem. Biases arise, which are the result of errors in period-to-period linking of records. When using a different probabilistic matching algorithm they achieve smaller error rates than the overall error rate of 7.8%, the Bureau of Labor Statistics found in an SSN validation project. Errors in database-linking can be caused by similar errors in period-to-period linking. These would be most likely the result of recording a wrong or mistyped SSN. Due to little human intervention in the process of data collection for LINDA, we expect the number of mismatches to be small.

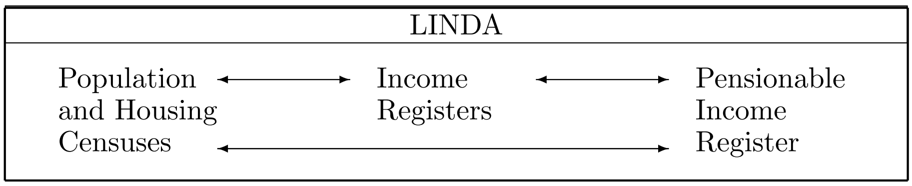


Figure 1 – Some of the registers used in LINDA, linked via social security numbers

We will be using mostly demographic variables (e.g. education, age and gender) and income variables. Since the survey contained mostly questions about 2002, we will be using data from that year.

2.2 Survey Data

In the beginning of every year, just before tax returns need to be filed, people in Sweden receive pre-printed tax-statements from the tax authorities. Around this time in 2003 a survey was conducted of 1431 individuals aged fifty or over. Out of the 1431 individuals who were contacted, 881 responded; 469 of them are women and 412 are men. The timing of the survey was chosen so as to optimize the information available to respondents when answering the questions in the survey. The questionnaire contains several questions about household income and expenses, partner income and assets. Besides the

[4] Among others, LINDA includes annual cash earnings, annual taxable benefits, social security sickness compensation (only if not old-age pension), if single family home (owner occupied), if condominium, if secondary house, tax assessed value on house, market value of house, stocks and shares, bank holdings, bonds, mutual funds, mortgages, other loans, schooling by number of years and category, pensions (social security, old age and disability pension, group (occupational) pensions, Private pension insurance (annuity)), capital incomes (interest and dividends received, realized capital gains, interest paid), total tax (income tax on earnings, capital income tax, real estate tax, wealth tax).

financial questions there are some smaller sections about household composition, health, retirement, and education of the respondent.

2.3 Descriptive comparisons of administrative and survey data

Our analysis will concentrate on three monetary variables: earnings, pensions, and taxes. Table 1 shows the evolution of the sample if we move from the gross sample of 1431 respondents drawn from the registers to the sample of respondents who answered at least some questions over the phone. The survey has 881 respondents, a response rate of 61.6%. In view of the age distribution of the sample, it is not surprising that many respondents report zero earnings. The answer to the survey question about taxes could be given either as an amount or as a percentage of income. We don't consider the percentage answers as the calculated amount of taxes paid would most likely exhibit a different error structure as the other responses. Roughly speaking, the implied error is a result of taking the ratio of two variables (cf. Duncan and Hill (1985)). Thus the number of respondents in the survey with positive survey taxes is much lower than in the administrative data.

Table 1 – Evolution of the sample

	earnings	pensions	taxes
Total number of persons in sample	1431	1431	1431
Number of respondents	881	881	881
Number of respondents with positive administrative ...	511	492	845
Number of respondents with positive survey ...	414	376	495
Both values positive	400	369	487

Table 2 compares a number of administrative variables across different subsamples. Comparing the sample of respondents (column 2) to the original (gross) sample (column 1) shows that the age and gender composition is essentially the same. The respondent sample is slightly better educated. Perhaps related to that, the respondent sample exhibits somewhat higher (administrative) earnings (7%). If we only consider observations with positive earnings the difference disappears: respondents earn 1.5% less on average than the overall mean of the gross sample. Pensions among the respondents are 5% higher on average in the respondent sample than in the gross sample, and 9% higher if we only consider positive pensions. For taxes the difference is 8.5%; and 6% if we only consider respondents with positive (administrative) tax payments.

Our empirical analysis will be primarily concerned with a comparison of non-zero

Table 2 – Comparison of population and sample characteristics

	(1)		(2)		(3)		(4)		(5)	
	Orig. sample[b]		Resp. sample[c]		Earnings sample		Pension sample		Taxes sample	
Variable (adm.)	Mean	Sd.	Mean	Sd.	Mean	Sd.	Mean	Sd.	Mean	Sd.
female	53.0%		53.2%		52.5%		53.7%		54.4%	
age	63.6	(9.6)	63.5	(9.5)	57.1	(5.3)	70.8	(7.2)	64.3	(9.4)
low education	28.1%		26.1%		20.0%		28.2%		24.8%	
middle education	35.8%		37.8%		48.3%		27.4%		39.4%	
high education	17.5%		19.1%		31.3%		11.9%		18.1%	
education missing	18.6%		17.0%		0.5%		32.5%		17.7%	
earnings[a]	114	(153)	122	(153)	–		–		–	
earnings > 0	214	(149)	211	(147)	246	(138)	–		–	
pensions[a]	74	(129)	78	(154)	–		–		–	
pensions > 0	128	(147)	140	(184)	–		141	(74)	–	
taxes[a]	59	(85)	64	(100)	–		–		–	
taxes > 0	64	(87)	67	(101)	–		–		67	(121)
N	1431		881		400		369		487	

Earnings, pensions and taxes are given in 1000 SEKs per year

[a] 1430 observations are used, since register values are missing for 1 individual.

[b] In the original sample the number of observations with a positive amount is 765 for earnings, 828 for pensions and 1328 for taxes.

[c] In the respondents sample the number of observations with a positive amount is 511 for earnings, 492 for pensions and 845 for taxes.

observations in both the administrative and survey data. Columns 3, 4, and 5 present administrative values for the subsamples that have positive values for both the administrative and survey variables. Now age and education vary considerably across subsamples for obvious reasons. Respondents with earners are typically younger and respondents with pensions are typically older. Education levels are also different, reflecting cohort differences in educational achievement. Comparing column 2 with columns 4 and 5 shows essentially no difference in means for pensions and taxes respectively. A comparison of columns 2 and 3 shows that mean earnings are substantially higher in the latter column than in the former. It is worthwhile therefore to investigate the differences between administrative data and survey data in some more detail.

The survey contained a number of questions on income and income related variables. A number of measures were taken to improve survey quality and to improve immediate comparison to the administrative variables. According to Hurd et al. (2004) one way to increase the quality of report, is by giving respondents the opportunity to report income in a timespan consistent with how they receive their income. For example, instead of forcing respondents to provide a yearly amount for earnings and pensions, they were given the possibility to report these amounts either per month or per year.

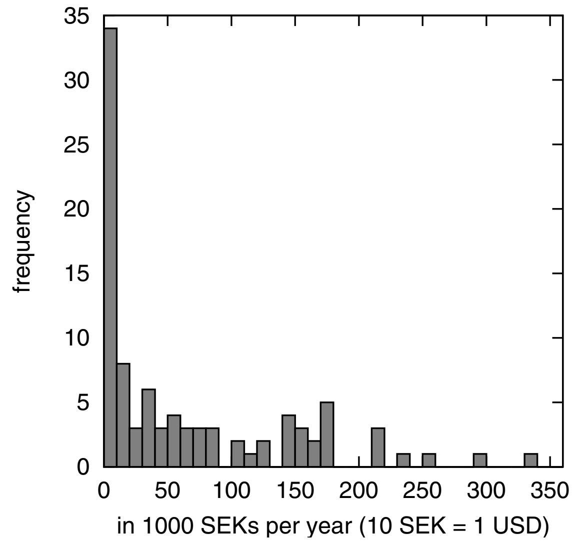


Figure 2 – Histogram of administrative earnings for respondents reporting no earnings

The question 'Did you have any income from employment in 2002?' was answered affirmatively by 435 respondents. When comparing these answers to the register, some discrepancies are found (see Table 3). Seventeen respondents claimed to have earnings, when according to the register they have zero earnings. Ninety-three respondents reported to have no earnings, when they should have earnings according to the register. As can be seen from Figure 2, two groups can be distinguished. One large group has low earnings (for instance 34 respondents earn less than 10,000 SEK[5] per year), which are easy to forget or perhaps thought not to be worth mentioning. Since this group is included in the mean earnings in column 2 of Table 2, but not in column 3, the mean in column 3 is substantially higher than in column 2. For the other group, respondents with substantial amounts of earnings, the explanation of rounding to zero is less plausible. Errors can be made by both the respondent and the interviewer, e.g. an interviewer may simply type in the wrong code. A different explanation could be mismatching, where a non-earning respondent is incorrectly matched with a register individual with positive earnings.

It is interesting to compare these findings to those of Pedace and Bates (2001) in their analysis of the SIPP. They find that of those who had earnings according to the administrative data, 5.5% said not to have had any earnings in the survey. In our survey that percentage is 18.2. On the other hand, Pedace and Bates (2001) find that of those who had no earnings according to the registers, 18% reported earnings in the SIPP. In our data that percentage is 4.6. Thus, in percentage terms, the discrepancy between administrative data and survey data appears similar across the two datasets, but in

[5]SEK represents the Swedish Kronor. In 2002, 10 SEK approximately equals 1 U.S. Dollar.

the SIPP zero administrative earnings are reported to be positive in the survey data at about the same rate as positive administrative earnings in Sweden are reported to be zero in the survey. Not too much should be read into this, if only because of the different age compositions of the Swedish and U.S. samples.

For the pension data we find a similar pattern of false responses. The percentage of respondents, with positive administrative pensions, but stating they do not have any pensions is 12.6. A smaller percentage of respondents (1.8%) report having a pension, while this pension is not found in the administrative data. 410 respondents gave an answer to the monthly pension question, including 2 zeros, 6 refusals and 43 don't knows. Only 26 respondents chose to give a yearly amount. For 369 respondents we have both a positive administrative value and a positive survey value.

From now on we concentrate on positive values of the monetary variables. We will not pay attention to sources of selectivity such as refusals, zero responses and don't knows. See Johansson and Klevmarken (2005) for a detailed analysis of various sources of non-response in the data. We will generally use logarithms of the monetary variables we are considering, to achieve distributions that are approximately symmetric.

In Table 4 some statistics for the monetary variables of interest are given. The difference between ln(earnings) measured in the survey and the administrative data is on the order of .02 on average. Strikingly, the variance of the survey data is smaller than the variance of the administrative data. Under the assumption of classical measurement error for the survey data and no error in the administrative data, this would not be possible. Looking at ln(pensions), we again observe modest, though somewhat larger, differences between the log-means of the administrative values and the survey values. Just as in the earnings data we see a substantial negative correlation between the difference between survey and administrative data, $s - r$, and the administrative data, r. Finally, mean ln(taxes) are about .08 higher in the survey data than in the administrative data. The correlation between $s - r$ and r is still negative, but closer to zero than for earnings and pensions.

Ten people (2.5%) gave an earnings amount exactly equal to the administrative value, while an additional 49 (12.3%) respondents provided amounts within 1000 SEK of the administrative value. For the pension and taxes data we find respectively 40 (13.3%) and 25 (5.1%) answers exactly equal to the administrative value, with an additional 59 (16.0%) and 72 (14.8%) answers within the 1000 SEK bound.

Figure 3 presents histograms of the difference between survey data and administrative data for earnings, pensions and taxes. Most of the values are close to zero, but we also see a large positive difference for earnings and large negative differences for pensions and taxes. The large positive and negative values force the scales in Figure 3 to be rather coarse. Hence, in Figure 4 the same histograms are given, but now truncated at a

Table 3 – Correspondence between administrative and survey 'answers' on relevant questions

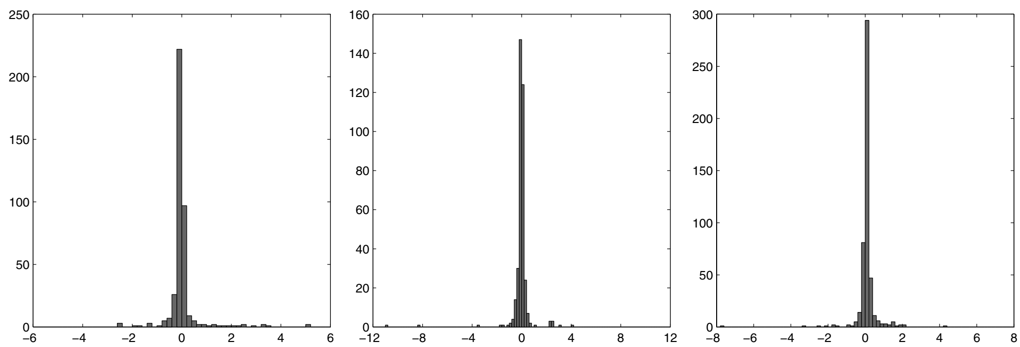
Question			survey: Yes		No		DK		RF	
Did you have any income from employment in 2002?		adm. > 0	418	(81.8%)	93	(18.2%)	0	(0%)	0	(0%)
		adm. $= 0$	17	(4.6%)	352	(95.1%)	0	(0%)	1	(0.3%)
Did you receive any type of old-age pension in 2002, such as (...)?		adm. > 0	430	(87.4%)	62	(12.6%)	0	(0%)	0	(0%)
		adm. $= 0$	7	(1.8%)	381	(97.9%)	1	(0.3%)	0	(0%)
			> 0		$= 0$		DK		RF	
How much did you earn per month in 2002, before taxes?		adm. > 0	327		1		6		5	
		adm. $= 0$	7		1		1		0	
Altogether, about how much did you earn from your main job in 2002, before taxes		adm. > 0	73		0		6		0	
		adm. $= 0$	7		0		1		0	
How much did you receive in pension payments per month in 2002, before taxes?		adm. > 0	353		2		43		6	
		adm. $= 0$	6		0		0		0	
Altogether, about how much did you receive in old-age pension payments (before taxes) in 2002?		adm. > 0	16		1		8		0	
		adm. $= 0$	1		0		0		0	
Think about the total income you received in 2002 from employment, pensions and taxable benefits. About how much did you pay (will you pay) in income tax on that amount?	amount	adm. > 0	487		1		12		0	
		adm. $= 0$	8		8		0		0	
	percentage	adm. > 0	230		0		10		0	
		adm. $= 0$	3		0		0		0	

Table 4 – Summary of administrative and survey variables

	N	survey	administrative	$s-r$	reliability*	corr(m,r)
ln(earnings)	400	12.196	12.172	.0243	.6821	–.5395
std.dev.		(.821)	(.961)	(.656)		
ln(pension)	369	11.649	11.694	–.0451	.3595	–.2699
std.dev.		(.945)	(.658)	(.878)		
ln(taxes)	487	10.869	10.786	.0828	.6734	–.1875
std.dev.		(.917)	(.829)	(.577)		

* $\frac{\sigma_r^2}{\sigma_r^2+\sigma_m^2}$, where r is the administrative variable and m is the difference between survey and administrative variable

value of $\pm.6$. Although the mean difference between survey and administrative earnings is positive (Table 4), we see from the truncated histogram that for most observations administrative earnings are larger than survey earnings. If the administrative data are assumed correct, this would indicate that most people underestimate their earnings. The difference in pensions data seems to centered around zero, whereas the difference in taxes data is mostly positive.

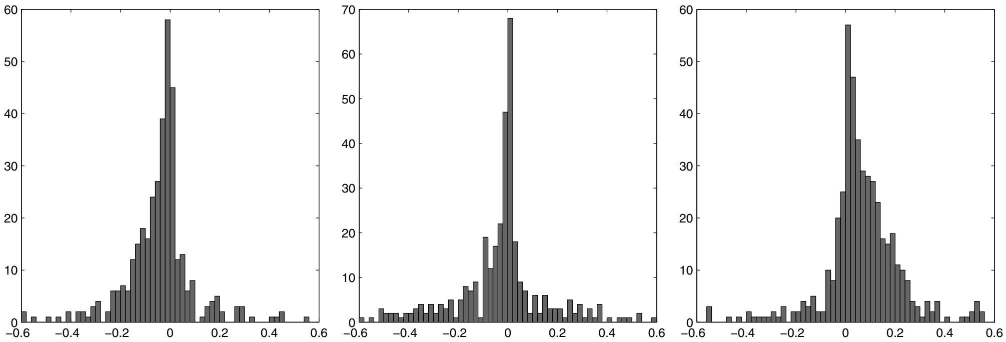


(a) Earnings (b) Pensions (c) Taxes

Figure 3 – Histograms of $m_i = s_i - r_i$ for earnings, pensions and taxes

3 Modeling Different Errors

Let ξ_i be the logarithm of the true value of the variable or interest (e.g. log-income) for individual i. This true value is not measured directly, but instead two sources of data, capturing the same concept, are available. In contrast to the assumptions typically made in the literature, we will assume that both sources of information, administrative



(a) Earnings (b) Pensions (c) Taxes

Figure 4 – Histograms of $m_i = s_i - r_i$ for earnings, pensions and taxes, truncated at $\pm.6$

and survey data, may contain error. The structure of the error, however, is different per source.

Define four i.i.d. and mutually independent normal variables: $\xi_i \sim N(\mu_\xi, \sigma_\xi^2)$, $\zeta_i \sim N(\mu_\zeta, \sigma_\zeta^2)$, $\eta_i \sim N(\mu_\eta, \sigma_\eta^2)$ and $\omega_i \sim N(\mu_\omega, \sigma_\omega^2)$, where i indexes the unit of observation. Denote the value elicited in the survey by s_i and the corresponding administrative value by r_i. For the derivations in this section the normality assumptions we are making are not necessary, but will be exploited later in ML estimation.

Errors in the administrative data are assumed to be due only to mismatching. With probability π_r the observed administrative value, r_i, is equal to the true value of individual i, ξ_i. In the case of a mismatch, which occurs with probability $1 - \pi_r$, the administrative value r_i corresponds to the true value of someone else.[6] This mismatched value will be denoted by ζ_i, where no correlation exists between ξ_i and ζ_i. Note that our sample comes from a subset of the population, containing only individuals of age 50 and older, whereas a mismatch may come from the complete sample available in LINDA. The distributions of ξ and ζ can therefore be different. The observed values r_i are a mixture of correct matches and mismatches

$$r_i = \begin{cases} \xi_i & \text{with probability } \ \pi_r \\ \zeta_i & \text{with probability } \ (1 - \pi_r). \end{cases}$$

For the survey data we distinguish three cases. The observed survey value is correct with probability π_s. With probability $1 - \pi_s$ the survey data contains response error, part of which is mean-reverting. Some of these observations, a proportion π_ω, are contaminated,

[6]Although it is attractive to think of mismatch as administrative data being linked to the wrong individual, other cases may be covered as well. For instance in some cases administrative data may be formally correct, but measure something that that is conceptually different. An example would be an individual who writes off a heavy capital loss in a given year. This may lead to low or even negative taxable income, while for most modeling purposes income would probably be defined differently.

modeled by adding an extra error-term, ω_i

$$s_i = \begin{cases} \xi_i & \text{with probability } \pi_s \\ \xi_i + \rho(\xi_i - \mu_\xi) + \eta_i & \text{with probability } (1-\pi_s)(1-\pi_\omega) \\ \xi_i + \rho(\xi_i - \mu_\xi) + \eta_i + \omega_i & \text{with probability } (1-\pi_s)\pi_\omega. \end{cases}$$

Contamination can for instance be the result of erroneously reporting income as annual, whereas the amount is a monthly amount, or vice versa, omitting a second job or working only part of the year. Each of these cases may result in large differences between survey value and true value. A value of ρ smaller than zero implies mean-reverting response error in the sense of Bound and Krueger(1991).

Next define the difference between survey data and administrative data, m_i, as

$$m_i \equiv s_i - r_i. \tag{1}$$

Note that the difference between survey data and administrative data no longer equals the response error, since in this model the administrative data may contain error. It is useful to calculate a number of moments of r_i, s_i and m_i to gain insight in the behavior of the model and its differences with a model without administrative error. We have (for derivations see Appendix A):

$$\mu_r \equiv E(r_i) = \pi_r\mu_\xi + (1-\pi_r)\mu_\zeta \tag{2}$$

$$\mu_s \equiv E(s_i) = \mu_\xi + (1-\pi_s)\left[\mu_\eta + \pi_\omega\mu_\omega\right] \tag{3}$$

$$\mu_m \equiv E(s_i) - E(r_i) = (1-\pi_r)\left[\mu_\xi - \mu_\zeta\right] + (1-\pi_s)\left[\mu_\eta + \pi_\omega\mu_\omega\right] \tag{4}$$

$$\sigma_r^2 \equiv Var(r_i) = \pi_r\sigma_\xi^2 + (1-\pi_r)\sigma_\zeta^2 + \pi_r(1-\pi_r)\left[\mu_\xi - \mu_\zeta\right]^2 \tag{5}$$

$$\begin{aligned} \sigma_s^2 &\equiv Var(s_i) \\ &= \pi_s\sigma_\xi^2 + (1-\pi_s)\{(1+\rho)^2\sigma_\xi^2 + \sigma_\eta^2 + \pi_\omega\sigma_\omega^2 + \pi_s\left[\mu_\eta + \pi_\omega\mu_\omega\right]^2 + \pi_\omega(1-\pi_\omega)\mu_\omega^2\} \end{aligned} \tag{6}$$

$$\begin{aligned} \sigma_m^2 &\equiv Var(m_i) \\ &= \Big[(1-\pi_r)\pi_s + \pi_r(1-\pi_s)\rho^2 + (1-\pi_r)(1-\pi_s)(1+\rho)^2\Big]\sigma_\xi^2 + \\ &\quad (1-\pi_r)\sigma_\zeta^2 + (1-\pi_s)\left[\sigma_\eta^2 + \pi_\omega\sigma_\omega^2\right] + \pi_r(1-\pi_r)\left[\mu_\xi - \mu_\zeta\right]^2 + \\ &\quad \pi_s(1-\pi_s)\left[\mu_\eta + \pi_\omega\mu_\omega\right]^2 + (1-\pi_s)\pi_\omega(1-\pi_\omega)\mu_\omega^2 \end{aligned} \tag{7}$$

$$\begin{aligned} \sigma_{mr} &\equiv E\left[(m_i - \mu_m)(r_i)\right] \\ &= \rho\pi_r(1-\pi_s)\sigma_\xi^2 - (1-\pi_r)\sigma_\zeta^2 - \pi_r(1-\pi_r)\left[\mu_\xi - \mu_\zeta\right]^2 \end{aligned} \tag{8}$$

The last expression can be seen as a measure of mean reversion we expect to see in the data under the (possibly incorrect) assumption that the administrative data are measured without error. We note that σ_{mr} is, for negative ρ, unambiguously negative, implying indeed mean reversion. Observe however that σ_{mr} is still negative if $\rho = 0$. That is, even without "true" mean reversion the data will still suggest that mean reversion is present as long as $\pi_r \neq 1$, i.e. if the administrative data suffer from at least some mismatch. This is not unique to the current set-up and is essentially an example of regression towards the mean. As soon as r_i suffers from measurement error, we expect $s_i - r_i$ to be negatively correlated with r_i. As a second observation we consider

$$\begin{aligned}\sigma_s^2 - \sigma_r^2 &= \sigma_\xi^2\left[\pi_s - \pi_r + (1-\pi_s)(1+\rho)^2\right] + (1-\pi_s)\left[\sigma_\eta^2 + \pi_\omega\sigma_\omega^2\right] \\ &\quad -(1-\pi_r)\sigma_\zeta^2 - \pi_r(1-\pi_r)[\mu_\xi - \mu_\zeta]^2 \end{aligned} \tag{9}$$

For $\pi_r = 1$, i.e. no mismatch in the administrative data, (9) reduces to

$$\sigma_s^2 - \sigma_r^2 = (1-\pi_s)\left[\left((1+\rho)^2 - 1\right)\sigma_\xi^2 + \sigma_\eta^2 + \pi_\omega\sigma_\omega^2\right] \tag{10}$$

which shows that if the administrative data are assumed to be measured perfectly, the variance of the survey data can only be smaller than the variance of the administrative data if the survey data exhibit mean reversion ($\rho < 0$). As a matter of fact, (10) will be negative if

$$(1+\rho)^2 < 1 - \frac{\sigma_\eta^2 + \pi_\omega\sigma_\omega^2}{\sigma_\xi^2} \tag{11}$$

Thus, the bigger the measurement errors in the survey data are assumed to be, the more mean reversion one needs to rationalize the data.

Under the scenario that neither the survey data nor the administrative data are flawless the question arises, which of the two should one use in modeling, and under what circumstances. An alternative, related, question would be: Given that survey data are usually more easily available than administrative data, how much do we lose by using survey data rather than administrative data. Below we illustrate the answers to these questions for a very simple linear model. One could also ask, what is the best way of combining survey and administrative data if both are available, but that question is beyond the scope of the current paper.[7]

3.1 Implications if ξ_i is a dependent variable

We consider a very simple linear regression model of the form:

$$\xi_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{12}$$

where we make the conventional assumption that ε_i is uncorrelated with x_i (and with any of the other random variables we have defined so far). If we have only survey data available, we would replace ξ_i by s_i on the left hand side of (12). Denote the ensuing estimate of β_1 by $\widehat{\beta}_1$. Then we have

$$p\lim\widehat{\beta}_1 = \beta_1\left[1 + \rho(1-\pi_s)\right] \tag{13}$$

[7] Our ML estimation using both administrative and survey data is one answer to that question.

Thus the estimator is consistent if $\rho = 0$, i.e. if there is no mean reverting error. The biasing effect of the mean reversion is mitigated by the observations that are exactly correct.

If we replace ξ_i by r_i we obtain for the probability limit of the estimator (say $\widetilde{\beta}_1$):

$$p\lim \widetilde{\beta}_1 = \pi_r \beta_1 \tag{14}$$

so that the percent bias is equal to the percentage of mismatched administrative observations. Comparing (13) to (14) shows that, for $\rho < 0$, using survey data will lead to less bias if:

$$1 + \rho(1 - \pi_s) > \pi_r \tag{15}$$

Clearly, this always holds if $\rho = 0$ and the administrative data are not perfect.

3.2 Implications if ξ_i is an independent variable

Now consider a model of the form:

$$z_i = \gamma_0 + \gamma_1 \xi_i + \nu_i \tag{16}$$

Let $\widehat{\gamma}_1$ be the OLS estimator of γ_1 if we replace ξ_i by s_i. It is straightforward to show that

$$p\lim \widehat{\gamma}_1 = \gamma_1 \frac{[1 + \rho(1 - \pi_s)]\, \sigma_\xi^2}{\pi_s \sigma_\xi^2 + (1 - \pi_s)\{(1+\rho)^2 \sigma_\xi^2 + \sigma_\eta^2 + \pi_\omega \sigma_\omega^2 + \pi_s \left[\mu_\eta + \pi_\omega \mu_\omega\right]^2 + \pi_\omega (1 - \pi_\omega) \mu_\omega^2\}} \tag{17}$$

where the denominator is equal to (6). Clearly if $\rho = 0$, $\pi_s = 0$ and $\pi_\omega = 0$, we are back in the case of classical measurement error. In that case (17) reduces to $p\lim \widehat{\gamma}_1 = \gamma_1[\sigma_\xi^2/(\sigma_\xi^2 + \sigma_\eta^2)]$. For $\rho = 0$ and $\pi_s, \pi_\omega \neq 0$ (17) reduces to

$$p\lim \widehat{\gamma}_1 = \gamma_1 \frac{\sigma_\xi^2}{\sigma_\xi^2 + (1 - \pi_s)\{\sigma_\eta^2 + \pi_\omega \sigma_\omega^2 + \pi_s \left[\mu_\eta + \pi_\omega \mu_\omega\right]^2 + \pi_\omega (1 - \pi_\omega) \mu_\omega^2\}} \tag{18}$$

As one would expect, having a proportion π_s of exactly measured variables mitigates the attenuation due to measurement error in the explanatory variable, while on the other hand a proportion π_ω of contaminated samples worsens the effect. Having measurement errors or contaminated samples with mean unequal to zero also increases the bias.

Let $\widetilde{\gamma}_1$ be the OLS estimator of γ_1 if we replace ξ_i by r_i. Now we obtain:

$$p\lim \widetilde{\gamma}_1 = \gamma_1 \frac{\pi_r \sigma_\xi^2}{\pi_r \sigma_\xi^2 + (1 - \pi_r)\sigma_\zeta^2 + \pi_r (1 - \pi_r)(\mu_\xi - \mu_\zeta)^2} \tag{19}$$

where the denominator is equal to (5). Clearly $\widetilde{\gamma}_1$ is consistent for $\pi_r = 1$, as it should be. A special case of interest is where a mismatch is a drawing from the same distribution, i.e. $\mu_\xi = \mu_\zeta$, $\sigma^2_\xi = \sigma^2_\zeta$. In that case (19) reduces to $p \lim \widetilde{\gamma}_1 = \gamma_1 \pi_r$. This is exactly the same shrinkage as in (14).

Comparison of biases introduced by either using less than perfect survey data or partly mismatched administrative data shows that survey data are to be preferred if

$$\begin{aligned} & \frac{[1+\rho(1-\pi_s)]\,\sigma^2_\xi}{\pi_s\sigma^2_\xi + (1-\pi_s)\{(1+\rho)^2\sigma^2_\xi + \sigma^2_\eta + \pi_\omega\sigma^2_\omega + \pi_s\left[\mu_\eta + \pi_\omega\mu_\omega\right]^2 + \pi_\omega(1-\pi_\omega)\mu^2_\omega\}} \\ > \; & \frac{\pi_r\sigma^2_\xi}{\pi_r\sigma^2_\xi + (1-\pi_r)\sigma^2_\zeta + \pi_r(1-\pi_r)(\mu_\xi - \mu_\zeta)^2} \end{aligned} \tag{20}$$

which is not particularly informative. More insight can be gained for the case where $\rho = 0$ (which is close to the empirically relevant case as we shall see for our data), $\mu_\omega = \mu_\eta = 0$ and $\mu_\xi = \mu_\zeta$, $\sigma^2_\xi = \sigma^2_\zeta$. In that case (20) reduces to

$$\frac{\sigma^2_\xi}{\sigma^2_\xi + (1-\pi_s)[\sigma^2_\eta + \pi_\omega\sigma^2_\omega]} > \pi_r \tag{21}$$

Thus survey data exhibit less bias if the reliability ratio of the survey data is greater than the proportion of perfect administrative data. In the more general case where $\mu_\xi \neq \mu_\zeta$, and $\sigma^2_\zeta \geqslant \sigma^2_\xi$ the balance tips a little more in favor of the survey data.

4 Estimation and results

We extend the model, discussed in the last section, by including covariates. We parameterize μ_ξ as a function of individual characteristics. The 'true' variable, ξ_i, is assumed to be dependent on variables such as gender, age and education. In this case we have

$$\xi_i = x_i\beta + \varepsilon_i, \tag{22}$$

where ε_i is Gaussian noise. Note, that when covariates are included, mean-reversion gets a slightly different interpretation. Values don't get adjusted toward the overall mean, but toward the mean within a group of people with the same values of x_i. Both the model with covariates, where $\xi_i \sim N(x_i\beta, \sigma^2_\xi)$, and the model without covariates, where $\xi_i \sim N(\mu_\xi, \sigma^2_\xi)$, are estimated. The covariates[8] used are age, age^2 and dummies for gender, degree of education and self-assessed retirement status. These are all survey values in order to avoid the mismatching problems accompanying administrative values. One drawback of this method, is possible error in survey data and a lot of don't know values for education (coded as *dkedu*).

[8]The LINDA variable used to define age is *bald*, which is the age of the individual at the end of the tax year. We define age = (*bald*-40) and age^2=(*bald*-40)^2/100.

For estimation we write the model as a mixture of one univariate normal distribution, when $r_i = \xi_i$ and $s_i = \xi_i$, and five bivariate normal distributions with different means and covariance matrices. Maximum likelihood is used to obtain estimates. For a detailed description of the estimation procedure we refer to Appendix B. It is perhaps somewhat remarkable that such a rich error structure can identified. This is the direct result of the non-normality of the error structure. It has been known for quite a while that normality is a very unfavorable assumption for identifiability of parameters in linear errors in variables models, e.g. Aigner et al. (1984) and Bekker (1986). Kane, Rouse and Staiger (1999) provide an example of how exploiting non-normality aids identifiability. Meijer and Ypma (2006) provide a simple proof of identification of the mixture of two normal distributions case, of which the current model is a generalization. A different strand of literature examines how bounds on measurement error can be exploited to bound parameter estimates, e.g. Klepper and Leamer (1984), and Bekker, Kapteyn, and Wansbeek (1987). In the current context such bounds don't seem necessary, although conceivably this would further narrow the range of plausible parameter estimates.

Appendix C presents all estimation results. For each of the variables of interest, earnings, pensions and taxes, two tables are presented: One table with covariates included and one without. Besides the full model, three other models are estimated, with certain constraints imposed on the full model. These include a model without contamination of the survey data, $\pi_\omega = 0$, a model where no mismatching occurs, $\pi_r = 1$ and a model where both are left out. This last model can be seen as the model used in most previous studies, with only the addition that survey observations are equal to the truth with positive probability.[9]

Table 7 presents the estimation results for earnings when no covariates are included. The most striking observation is probably that allowing for mismatches or contaminated samples in the model, leads to a dramatic fall in the estimated value of the mean-reversion parameter, ρ. Only when we do not allow for either contamination or mismatches, do we reproduce the "stylized fact" of substantial mean reversion. Table 6, which includes covariates, leads to qualitatively similar conclusions. The fact that allowing for mismatch may lead to a sharp drop in the estimate of ρ is consistent with equation (8), which shows that a negative covariance between the "measurement error" and the administrative values can be generated by mismatches even if $\rho = 0$. The fact that contamination of the survey data can also rationalize a negative covariance between m and r is a little harder to grasp intuitively.

[9]Since it is sometimes found that women provide more accurate answers than men (e.g. Bound and Krueger, 1991) we have also estimated a model where μ_ε and σ_ε are allowed to differ by gender, but the differences are negligible. For instance, for earnings the estimate of σ_ε is .100 for men and .104 for women; the t-value of the difference is .33. The t-value for the difference in μ_ε is even lower: .10.

With respect to the pattern of estimated mean reversion, Tables 9 and 8 (for pensions) provide a qualitatively similar picture, although mean reversion in the full model is somewhat higher than for earnings. For taxes, mean reversion in the full model and in the models with only mismatches or only contaminated samples is essentially zero (Tables 11 and 10).

The estimated percentage of correct survey data, π_s, ranges from 15 percent in the earnings data to a little over 25 percent in the pension data. The fraction of contaminated survey values, $\pi_\omega(1-\pi_s)$, lies between .04 and .13.

A parameter of particular interest is π_r, or rather $1-\pi_r$ the proportion of mismatched administrative values. The estimate of $1-\pi_r$ varies from 2% in the pension data to about 8% in the tax data. We can use equations (13), (14), (17), and (19) to asses the biases that would arise in the estimate of a slope parameter if ξ_i were either a dependent (LHS) or an independent (RHS) variable in a univariate regression. Table 5 presents the proportional biases for both cases and for the three variables we are considering in this paper.[10] When ξ_i is a dependent (LHS) variable, biases are modest, with the administrative data leading to slightly smaller biases than the survey data for pensions, whereas for earnings and taxes survey data yield less bias. Biases are on the order of 5%. When ξ_i is a RHS variable, the picture is dramatically different. Biases are much larger, up to 65%. When using administrative data, we find that the bias is largest for earnings, about 50%. Inspection of formula (19) shows that this is due to the fact that the 4% mismatched data appear to be drawn from a distribution that has a much lower log-mean ($\mu_\zeta = 9.187$, while $\mu_\xi = 12.283$) and substantially higher dispersion ($\sigma_\zeta = 1.807$, while $\sigma_\xi = .717$). For the survey data the bias is particularly large in the case of pensions, about 65%. Inspection of formula (17) suggests that the main cause for the big bias lies in the low mean and high variance of the contaminated data ($\mu_\omega = -1.632$, $\sigma_\omega = 3.801$).

Table 5 – Proportional biases resulting from using administrative or survey data

	Earnings		Pensions		Taxes	
	Admin.	Survey	Admin.	Survey	Admin.	Survey
ξ_i LHS	0.959	0.989	0.981	0.904	0.935	0.991
ξ_i RHS	0.491	0.701	0.719	0.363	0.789	0.735

Note: cells present proportional asymptotic biases in OLS estimates if we replace the true variable by administrative or survey data; 1 means no bias.

[10] We use the estimates for the models without covariates

A different way to gauge the importance of a proper treatment of the various error sources, is to compare different conventional estimation methods and how their results differ from the full model. Tables 12, 13, and 14 present estimates of the parameters of economic interest for a number of different estimation methods: ML on the full model; OLS, robust regression, and median regression. The latter three estimation methods are applied twice, once with administrative data as dependent variable and once with survey data as dependent variable.

Considering earnings (Table 12) we note that the estimates of the effect of education on earnings may vary by at least a factor of 2, depending on the estimation method chosen. Table 12 suggests that running OLS of the administrative variables on the explanatory variables provides estimates close to those of the full model. However when we consider pensions that conclusion changes quite a bit. Robust regression and median regression yield estimates that may be quite far removed from the estimates obtained with ML on the full model. Another noteworthy phenomenon is the wide variation in the estimates of the effect of the semi-retired dummy on pensions (Table 13). Estimates vary from significantly positive to significantly negative depending on the estimation method used and the choice of dependent variable. For taxes the different estimation methods appear to provide roughly comparable estimates of parameters of economic interest.

Since the full model is a mixture of a number of different regimes it is of interest to assign observations to different regimes. To do so we have used the fact that the likelihood for each observation is a weighted sum of densities corresponding to the different regimes. We have assigned observations to the regime that produces the highest density for that observation. Figures 5-7 present the results. For the earnings data there is some suggestion that respondents with low earnings tend to give high survey values, as also found by Bollinger (1998). Of particular interest are the observations that are classified as mismatched administrative variables. Most of these points lie above the 45 degree line, whereas the points classified as contaminated in the survey data lie below the 45 degree line. Naturally, the assignment procedure used here is merely indicative and probabilistic; and hence no great importance should be attached to the classification of each observation

It is of interest to see if some of the classifications may be externally validated in some way. For about 70% of the LINDA sample a third source of earnings information is available from employer records. It consists of two variables. One variable containing the fulltime equivalent monthly earnings and one containing the percentage of fulltime equivalent employed. Although there are some problems with this information, we can still compare these values with the survey value and the administrative value. A simple way to look for possible mismatches is to consider observations where survey and employer earnings are fairly close, while differing substantially from the administrative data. For instance, if we select observations for which survey earnings and administrative earnings are at least 50,000 SEK apart, while survey earnings and employer earnings differ by less than 10,000 SEK we find twelve cases where this condition holds true. For these 12 observations the ratio of survey and administrative earnings varies between .5 and 4.5.

5 Conclusion

In comparison with most studies in the literature we have allowed for a richer specification of possible error sources in survey data and administrative data. Our results suggest that some conclusions in the literature may be quite sensitive to the assumption that administrative data are flawless. In a sense, the question if administrative data represent the truth, is almost a philosophical question. For instance, the examples of detected mismatches given above do not necessarily imply there is true mismatching going on in the administrative data. Rather, it appears that sometimes the survey data (and the alternative source of administrative data) measure a different concept than the administrative data. Be this as it may, also under the latter interpretation one would be hard pressed to maintain that the difference between survey data and administrative data exhibits strong mean reversion.

Our results also suggest that substantive conclusions may be affected quite a bit by changes in assumptions on the nature of error in survey and administrative data. Application of robust methods, like median or robust regression, yields results quite far removed from ML on the full model. Thus these methods do not appear to provide a solution for dealing with different sources of error in survey or administrative data.

There are many good reasons for wanting to use administrative data, including sample sizes, cost of surveys, and data quality. However, unless administrative data measure exactly the concept that one is interested in and do so without error, these data are not a panacea. Our illustrative calculations in Table 5 suggest that biases resulting of using administrative data as right hand side variables may be very substantial. As always, one has to be careful in modeling the sources and nature of errors and take that into account when investigating hypotheses of substantive interest.

References

[1] John M. Abowd and Lars Vilhuber. The sensitivity of economic statistics to coding errors in personal identifiers. *Journal of Business & Economic Statistics*, 23(2):133–152, April 2005.

[2] Dennis Aigner, Cheng Hsiao, Arie Kapteyn, and Tom Wansbeek. Latent variable models in econometrics. In Zvi Griliches and Michael Intriligator, editors, *Handbook of Econometrics*, volume II, pages 1321–1393. Elsevier, Amsterdam, 1984.

[3] Paul Bekker, Arie Kapteyn, and Tom Wansbeek. Consistent sets of estimates for regressions with correlated or uncorrelated measurement error in arbitrary subsets of all variables. *Econometrica*, 55:1223–1230, 1987.

[4] Paul A. Bekker. Comment on identification in the linear errors in variables model. *Econometrica*, 54:215–217, 1986.

[5] Christopher R. Bollinger. Measurement error in the current population survey: A nonparametric look. *Journal of Labor Economics*, 16(3):576–594, July 1998.

[6] J. Bound, C. Brown, and N. Mathiowetz. Measurement error in survey data. In J. Heckman and E. Leamer, editors, *Handbook of Labor Economics*, volume 5, pages 3707–3843. Elsevier, Amsterdam, 2001.

[7] John Bound, Charles Brown, Greg J. Duncan, and Willard L. Rodgers. Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics*, 12(3):345–368, July 1994.

[8] John Bound and Alan B. Krueger. The extent of measurement error in longitudinal earnings data: Do two wrongs make a right? *Journal of Labor Economics*, 9(1):1–24, January 1991.

[9] Xiaohong Chen, Han Hong, and Elie Tamer. Measurement error models with auxiliary data. *The Review of Economic Studies*, 72(2):343–366, April 2005.

[10] Greg J. Duncan and Daniel H. Hill. An investigation of the extent and consequences of measurement error in labor-economic survey data. *Journal of Labor Economics*, 3(4):508–532, October 1985.

[11] Per-Anders Edin and Peter Frederiksson. LINDA - Longitudinal INdividual DAta for Sweden. *working paper*, November 2000.

[12] Michael Hurd, F. Thomas Juster, and James P. Smith. Enhancing the quality of data on income: Recent developments in survey methodology. *Working paper*, 2004.

[13] Fredrik Johansson and Anders Klevmarken. Explaining the size and nature of response behavior in a survey on health status and economic standard. *Working paper*, Uppsala University, May 2005.

[14] Thomas Kane, Cecilia Rouse, and Douglas Staiger. Estimating returns to schooling when schooling is misreported. *NBER Working paper*7235, 1999.

[15] Bonggeun Kim and Gary Solon. Implications of mean-reverting measurement error for longitudinal studies of wages and employment. *Working paper*, December 2003.

[16] Steven Klepper and Edward Leamer. Consistent sets of estimates for regressions with errors in all variables. *Econometrica*, 52:163–184, 1984.

[17] LungFei Lee and Jungsywan H. Sepanski. Estimation of linear and nonlinear errors-in-variables models using validation data. *Journal of the American Statistical Association*, 90(429):130–140, March 1995.

[18] Erik Meijer and Jelmer Ypma. A simple identification proof for a mixture of two univariate normal distributions. *Working paper* University of Groningen, March 2006.

[19] Roberto Pedace and Nancy Bates. Using administrative records to assess earnings reporting error in the survey of income and program participation. *Journal of Economic and Social Measurement*, 26(3-4):173–192, 2001.

[20] Jörn-Steffen Pischke. Measurement error and earnings dynamics: Some estimates from the PSID validation study. *Journal of Business & Economic Statistics*, 13(3):305–314, July 1995.

[21] Richard A. Redner and Homer F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, April 1984.

[22] Willard L. Rodgers, Charles Brown, and Greg J. Duncan. Errors in survey reports of earnings, hours worked, and hourly wages. *Journal of the American Statistical Association*, 88(424):1208–1218, December 1993.

[23] Martha Stinson. Estimating measurement error in SIPP annual job earnings: A comparison of census survey and SSA administrative data. *Technical Paper* No. TP-2002-24, U.S. Census Bureau, Suitland, MD, September 2002.

A Derivation of moments

The expectation of r_i

$$\begin{aligned}\mu_r \equiv E[r_i] &= \pi_r E[r_i|r_i = \xi_i] + (1-\pi_r)E[r_i|r_i = \zeta_i] \\ &= \pi_r E[\xi_i] + (1-\pi_r)E[\zeta_i] \\ &= \pi_r \mu_\xi + (1-\pi_r)\mu_\zeta.\end{aligned}$$

The expectation of s_i

$$\begin{aligned}\mu_s &\equiv E[s_i] \\ &= \pi_s E[s_i|s_i = \xi_i] + (1-\pi_s)(1-\pi_\omega)E[s_i|s_i = \xi_i + \rho(\xi_i - \mu_\xi) + \eta_i] + \\ &\quad (1-\pi_s)\pi_\omega E[s_i|s_i = \xi_i + \rho(\xi_i - \mu_\xi) + \eta_i + \omega_i] \\ &= \pi_s E[\xi_i] + (1-\pi_s)(1-\pi_\omega)E[\xi_i + \rho(\xi_i - \mu_\xi) + \eta_i] + \\ &\quad (1-\pi_s)\pi_\omega E[\xi_i + \rho(\xi_i - \mu_\xi) + \eta_i + \omega_i] \\ &= \pi_s \mu_\xi + (1-\pi_s)(1-\pi_\omega)[\mu_\xi + \mu_\eta] + (1-\pi_s)\pi_\omega[\mu_\xi + \mu_\eta + \mu_\omega] \\ &= \mu_\xi + (1-\pi_s)[\mu_\eta + \pi_\omega \mu_\omega].\end{aligned}$$

Expectation of m_i

$$\mu_m \equiv E[m_i] = E[s_i - r_i] = E[s_i] - E[r_i] = (1-\pi_r)[\mu_\xi - \mu_\zeta] + (1-\pi_s)[\mu_\eta + \pi_\omega \mu_\omega].$$

The variance of r_i is

$$\begin{aligned}\sigma_r^2 &\equiv E[r_i - \mu_r]^2 \\ &= \pi_r E[r_i - \mu_r | r_i = \xi_i]^2 + (1-\pi_r)E[r_i - \mu_r | r_i = \zeta_i]^2 \\ &= \pi_r E[\xi_i - \pi_r\mu_\xi - (1-\pi_r)\mu_\zeta]^2 + (1-\pi_r)E[\zeta_i - \pi_r\mu_\xi - (1-\pi_r)\mu_\zeta]^2 \\ &= \pi_r E[(\xi_i - \mu_\xi) + (1-\pi_r)(\mu_\xi - \mu_\zeta)]^2 + (1-\pi_r)E[(\zeta_i - \mu_\zeta) - \pi_r(\mu_\xi - \mu_\zeta)]^2 \\ &= \pi_r[\sigma_\xi^2 + (1-\pi_r)^2(\mu_\xi - \mu_\zeta)^2] + (1-\pi_r)[\sigma_\zeta^2 + \pi_r^2(\mu_\xi - \mu_\zeta)^2] \\ &= \pi_r\sigma_\xi^2 + (1-\pi_r)\sigma_\zeta^2 + \pi_r(1-\pi_r)(\mu_\xi - \mu_\zeta)^2.\end{aligned}$$

The variance of s_i is

$$\begin{aligned}\sigma_s^2 &= E[s_i - \mu_s]^2 \\ &= \pi_s E[s_i - \mu_s | s_i = \xi_i]^2 + (1-\pi_s)(1-\pi_\omega)E[s_i - \mu_s | s_i = \xi_i + \rho(\xi_i - \mu_\xi) + \eta_i]^2 + \\ &\quad (1-\pi_s)\pi_\omega E[s_i - \mu_s | s_i = \xi_i + \rho(\xi_i - \mu_\xi) + \eta_i + \omega_i]^2. \end{aligned} \tag{23}$$

First calculate the three variances separately. We have

$$\begin{aligned}E[s_i - \mu_s | s_i = \xi_i]^2 &= E[\xi_i - \mu_\xi - (1-\pi_s)(\mu_\eta + \pi_\omega\mu_\omega)]^2 \\ &= E[\xi_i - \mu_\xi - (1-\pi_s)(\mu_\eta + \pi_\omega\mu_\omega)]^2 \\ &= \sigma_\xi^2 + (1-\pi_s)^2(\mu_\eta + \pi_\omega\mu_\omega)^2,\end{aligned} \tag{24}$$

and

$$
\begin{aligned}
& E[s_i - \mu_s | s_i = \xi_i + \rho(\xi_i - \mu_\xi) + \eta_i]^2 \\
= \; & E[\xi_i + \rho(\xi_i - \mu_\xi) + \eta_i - \mu_\xi - (1 - \pi_s)(\mu_\eta + \pi_\omega \mu_\omega)]^2 \\
= \; & E[(1+\rho)(\xi_i - \mu_\xi) + (\eta_i - \mu_\eta) + \pi_s(\mu_\eta + \pi_\omega \mu_\omega) - \pi_\omega \mu_\omega]^2 \\
= \; & (1+\rho)^2 \sigma_\xi^2 + \sigma_\eta^2 + [\pi_s(\mu_\eta + \pi_\omega \mu_\omega) - \pi_\omega \mu_\omega]^2, \qquad (25)
\end{aligned}
$$

and

$$
\begin{aligned}
& E[s_i - \mu_s | s_i = \xi_i + \rho(\xi_i - \mu_\xi) + \eta_i + \omega_i]^2 \\
= \; & E[\xi_i + \rho(\xi_i - \mu_\xi) + \eta_i + \omega_i - \mu_\xi - (1 - \pi_s)(\mu_\eta + \pi_\omega \mu_\omega)]^2 \\
= \; & E[(1+\rho)(\xi_i - \mu_\xi) + (\eta_i - \mu_\eta) + (\omega_i - \mu_\omega) + \pi_s(\mu_\eta + \pi_\omega \mu_\omega) + (1 - \pi_\omega)\mu_\omega]^2 \\
= \; & (1+\rho)^2 \sigma_\xi^2 + \sigma_\eta^2 + \sigma_\omega^2 + [\pi_s(\mu_\eta + \pi_\omega \mu_\omega) + (1 - \pi_\omega)\mu_\omega]^2. \qquad (26)
\end{aligned}
$$

Define

$$
\delta = \mu_\xi - \mu_\zeta, \quad \text{and} \quad \Delta = \mu_\eta + \pi_\omega \mu_\omega.
$$

Substituting (24), (25) and (26) in (23) we have

$$
\begin{aligned}
\sigma_s^2 \; = \; & E[s_i - \mu_s]^2 \\
= \; & \pi_s[\sigma_\xi^2 + (1 - \pi_s)^2 \Delta^2] + \\
& (1 - \pi_s)(1 - \pi_\omega)[(1+\rho)^2 \sigma_\xi^2 + \sigma_\eta^2 + [\pi_s \Delta - \pi_\omega \mu_\omega]^2] + \\
& (1 - \pi_s)\pi_\omega[(1+\rho)^2 \sigma_\xi^2 + \sigma_\eta^2 + \sigma_\omega^2 + [\pi_s \Delta + (1 - \pi_\omega)\mu_\omega]^2] \\
= \; & [\pi_s + (1 - \pi_s)(1+\rho)^2]\sigma_\xi^2 + (1 - \pi_s)[\sigma_\eta^2 + \pi_\omega \sigma_\omega^2] \; + \; \Omega, \qquad (27)
\end{aligned}
$$

where Ω is

$$
\begin{aligned}
\Omega \; = \; & \pi_s(1 - \pi_s)^2 \Delta^2 + (1 - \pi_s)(1 - \pi_\omega)[\pi_s \Delta - \pi_\omega \mu_\omega]^2 + (1 - \pi_s)\pi_\omega[\pi_s \Delta + (1 - \pi_\omega)\mu_\omega]^2 \\
= \; & \pi_s(1 - \pi_s)^2 \Delta^2 + (1 - \pi_s)\pi_s^2 \Delta^2 + (1 - \pi_s)(1 - \pi_\omega)\pi_\omega^2 \mu_\omega^2 + (1 - \pi_s)\pi_\omega(1 - \pi_\omega)^2 \mu_\omega^2 \\
& -2(1 - \pi_s)(1 - \pi_\omega)\pi_s \Delta \pi_\omega \mu_\omega + 2(1 - \pi_s)\pi_\omega \pi_s \Delta(1 - \pi_\omega)\mu_\omega \\
= \; & [\pi_s(1 - \pi_s)^2 + (1 - \pi_s)\pi_s^2]\Delta^2 + (1 - \pi_s)[(1 - \pi_\omega)\pi_\omega^2 + \pi_\omega(1 - \pi_\omega)^2]\mu_\omega^2 \\
= \; & \pi_s(1 - \pi_s)\Delta^2 + (1 - \pi_s)\pi_\omega(1 - \pi_\omega)\mu_\omega^2. \qquad (28)
\end{aligned}
$$

We can now calculate the variance using (27), (28) and the definition of Δ

$$
\begin{aligned}
\sigma_s^2 \; = \; & [\pi_s + (1 - \pi_s)(1+\rho)^2]\sigma_\xi^2 + (1 - \pi_s)[\sigma_\eta^2 + \pi_\omega \sigma_\omega^2] + \\
& \pi_s(1 - \pi_s)(\mu_\eta + \pi_\omega \mu_\omega)^2 + (1 - \pi_s)\pi_\omega(1 - \pi_\omega)\mu_\omega^2.
\end{aligned}
$$

Using the same procedure as above, the covariance between r and s is

$$\begin{aligned}
\sigma_{rs} &\equiv E\big[r_i - \mu_r\big]\big[s_i - \mu_s\big] \\
&= \pi_r\pi_s\big(\sigma_\xi^2 - (1-\pi_r)(1-\pi_s)\delta\Delta\big) + \\
&\quad \pi_r(1-\pi_s)(1-\pi_\omega)\big((1+\rho)\sigma_\xi^2 + (1-\pi_r)\pi_s\delta\Delta - (1-\pi_r)\delta\pi_\omega\mu_\omega\big) + \\
&\quad \pi_r(1-\pi_s)\pi_\omega\big((1+\rho)\sigma_\xi^2 + (1-\pi_r)\pi_s\delta\Delta + (1-\pi_r)\delta(1-\pi_\omega)\mu_\omega\big) + \\
&\quad (1-\pi_r)\pi_s\big(\pi_r(1-\pi_s)\delta\Delta\big) + \\
&\quad (1-\pi_r)(1-\pi_s)(1-\pi_\omega)\big(-\pi_r\pi_s\delta\Delta + \pi_r\delta\pi_\omega\mu_\omega\big) + \\
&\quad (1-\pi_r)(1-\pi_s)\pi_\omega\big(-\pi_r\pi_s\delta\Delta - \pi_r\delta(1-\pi_\omega)\mu_\omega\big) \\
&= \pi_r\sigma_\xi^2 + \pi_r(1-\pi_s)\rho\sigma_\xi^2.
\end{aligned}$$

The variance of m_i is then easily obtained as

$$\begin{aligned}
\sigma_m^2 &\equiv E[m_i - \mu_m]^2 = E[(s_i - \mu_s) - (r_i - \mu_r)]^2 = \sigma_s^2 + \sigma_r^2 - 2\sigma_{rs} \\
&= \Big[\pi_s + (1-\pi_s)(1+\rho)^2 + \pi_r - 2\pi_r - 2\pi_r(1-\pi_s)\rho\Big]\sigma_\xi^2 + \\
&\quad (1-\pi_r)\sigma_\zeta^2 + (1-\pi_s)\big[\sigma_\eta^2 + \pi_\omega\sigma_\omega^2\big] + \pi_r(1-\pi_r)\big[\mu_\xi - \mu_\zeta\big]^2 + \\
&\quad \pi_s(1-\pi_s)\big[\mu_\eta + \pi_\omega\mu_\omega\big]^2 + (1-\pi_s)\pi_\omega(1-\pi_\omega)\mu_\omega^2,
\end{aligned}$$

and the covariance between m and r is

$$\begin{aligned}
\sigma_{mr} &\equiv \sigma_{sr} - \sigma_r^2 \\
&= \pi_r\sigma_\xi^2 + \pi_r(1-\pi_s)\rho\sigma_\xi^2 - [\pi_r\sigma_\xi^2 + (1-\pi_r)\sigma_\zeta^2 + \pi_r(1-\pi_r)(\mu_\xi - \mu_\zeta)^2] \\
&= \rho\pi_r(1-\pi_s)\sigma_\xi^2 - (1-\pi_r)\sigma_\zeta^2 - \pi_r(1-\pi_r)(\mu_\xi - \mu_\zeta)^2.
\end{aligned}$$

B Maximum likelihood

As described in Section 4, we assume that the administrative data are a mixture of two normal distributions, while the survey data are a mixture of three different normal distributions. Since we assume the processes underlying the administrative data and the survey data to be independent, the combined set of observations (r_i, s_i) follow a mixture of six distributions.

The general shape of the log-likelihood function of a mixture of M distributions and N observations is the following (Redner and Walker (1984))

$$l(\theta) = \sum_{i=1}^{N} \log\left(\sum_{m=1}^{M} \pi_m f_m(x_i|\theta)\right),$$

where θ is the vector of parameters, including mixing proportions π_m and parameters describing the distributions. Redner and Walker (1984) mention some special cases of mixtures, for instance when some of the observations are labeled. In our case, the observations where $r_i = s_i$ can be seen as labeled observations. We assume that these observations come from the distribution, where both administrative and survey data are correct, referred to as group 1. Since all of the observations from group 1 can be labeled, we have a completely labeled group.

Let's assume that observations $i = 1, \ldots, n_1$ are observations from the completely labeled group 1, and the other observations, $i = n_1 + 1, \ldots, N$ are a mixture of the remaining five distributions. The following log-likelihood can then be derived

$$l(\theta) = \sum_{i=1}^{n_1} \log\left(\pi_1 f_1(x_i|\theta)\right) + \sum_{i=n_1+1}^{N} \log\left(\sum_{m=2}^{5} \pi_m f_m(x_i|\theta)\right),$$

where

$$\begin{aligned}
\pi_1 &= \pi_r \pi_s \\
\pi_2 &= \pi_r(1-\pi_s)(1-\pi_\omega) \\
\pi_3 &= \pi_r(1-\pi_s)\pi_\omega \\
\pi_4 &= (1-\pi_r)\pi_s \\
\pi_5 &= (1-\pi_r)(1-\pi_s)(1-\pi_\omega) \\
\pi_6 &= (1-\pi_r)(1-\pi_s)\pi_\omega.
\end{aligned}$$

The density function f_1 is the probability density function of a $N(\mu_\xi, \sigma_\xi^2)$-distribution, since both r_i and s_i are equal to ξ_i in this case. In the five other cases we have the bivariate normal distributions listed below

$$f_2(r_i, s_i) \sim N\left[\begin{pmatrix} \mu_\xi \\ \mu_\xi + \mu_\eta \end{pmatrix}, \begin{pmatrix} \sigma_\xi^2 & \frac{(1+\rho)\sigma_\xi^2}{\sigma_\xi\sqrt{(1+\rho)^2\sigma_\xi^2+\sigma_\eta^2}} \\ \frac{(1+\rho)\sigma_\xi^2}{\sigma_\xi\sqrt{(1+\rho)^2\sigma_\xi^2+\sigma_\eta^2}} & (1+\rho)^2\sigma_\xi^2 + \sigma_\eta^2 \end{pmatrix}\right],$$

when $r_i = \xi_i$ and $s_i = \xi_i + \rho(\xi_i - \mu_\xi) + \eta_i$,

$$f_3(r_i, s_i) \sim N\left[\begin{pmatrix} \mu_\xi \\ \mu_\xi + \mu_\eta + \mu_\omega \end{pmatrix}, \begin{pmatrix} \sigma_\xi^2 & \frac{(1+\rho)\sigma_\xi^2}{\sigma_\xi\sqrt{(1+\rho)^2\sigma_\xi^2+\sigma_\eta^2+\sigma_\omega^2}} \\ \frac{(1+\rho)\sigma_\xi^2}{\sigma_\xi\sqrt{(1+\rho)^2\sigma_\xi^2+\sigma_\eta^2+\sigma_\omega^2}} & (1+\rho)^2\sigma_\xi^2 + \sigma_\eta^2 + \sigma_\omega^2 \end{pmatrix}\right],$$

when $r_i = \xi_i$ and $s_i = \xi_i + \rho(\xi_i - \mu_\xi) + \eta_i + \omega_i$,

$$f_4(r_i, s_i) \sim N\left[\begin{pmatrix} \mu_\zeta \\ \mu_\xi \end{pmatrix}, \begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & \sigma_\xi^2 \end{pmatrix}\right],$$

when $r_i = \zeta_i$ and $s_i = \xi_i$,

$$f_5(r_i, s_i) \sim N\left[\begin{pmatrix} \mu_\zeta \\ \mu_\xi + \mu_\eta \end{pmatrix}, \begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & (1+\rho)^2\sigma_\xi^2 + \sigma_\eta^2 \end{pmatrix}\right],$$

when $r_i = \zeta_i$ and $s_i = \xi_i + \rho(\xi_i - \mu_\xi) + \eta_i$ and

$$f_6(r_i, s_i) \sim N\left[\begin{pmatrix} \mu_\zeta \\ \mu_\xi + \mu_\eta + \mu_\omega \end{pmatrix}, \begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & (1+\rho)^2\sigma_\xi^2 + \sigma_\eta^2 + \sigma_\omega^2 \end{pmatrix}\right],$$

when $r_i = \zeta_i$ and $s_i = \xi_i + \rho(\xi_i - \mu_\xi) + \eta_i + \omega_i$. One final note has to be made about the labeling of group 1. Observations are labeled as a member of group 1 if the difference between r_i and s_i is smaller than 1000 SEK. The proportion π_s then is the proportion of survey observations that differ less than 1000 SEK from the administrative data. In principle, this broader definition of "equal" observations affects the consistency of our estimates. However, we expect these effects to be a minor.

C Estimation results

Table 6 – Estimates using Log(Earnings)

	Full model		No cont.		No mismatch		Basic model	
	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.
Log likelihood	–503.2		–558.8		–623.4		–762.8	
female	–.268	(.052)	–.253	(.062)	–.163	(.100)	–.225	(.071)
age	.087	(.024)	.095	(.031)	.136	(.025)	.111	(.034)
age^2	–.275	(.062)	–.275	(.082)	–.421	(.059)	–.330	(.095)
fullret	–1.015	(.125)	–1.152	(.091)	–1.253	(.127)	–1.210	(.114)
semiret	–.421	(.114)	–.273	(.112)	–.694	(.276)	–.378	(.106)
midedu	.244	(.102)	.370	(.117)	.266	(.128)	.361	(.228)
highedu	.345	(.101)	.438	(.122)	.364	(.148)	.435	(.211)
dkedu	.568	(.104)	.694	(.119)	.603	(.120)	.684	(.197)
μ_ξ	11.592	(.262)	11.335	(.315)	11.092	(.301)	11.181	(.287)
σ_ξ	.543	(.021)	.653	(.021)	.840	(.032)	.754	(.031)
μ_ζ	9.843	(.646)	11.385	(.243)	–	–	–	–
σ_ζ	2.026	(.334)	1.738	(.170)	–	–	–	–
μ_ω	–.259	(.211)	–	–	.442	(.234)	–	–
σ_ω	1.313	(.152)	–	–	1.791	(.174)	–	–
μ_η	–.046	(.007)	–.053	(.008)	–.040	(.008)	–	–
σ_η	.102	(.008)	.110	(.007)	.104	(.006)	.523	(.020)
π_r	.948	(.015)	.864	(.020)	–	–	–	–
π_s	.155	(.019)	.170	(.020)	.148	(.018)	–	–
π_ω	.123	(.028)	–	–	.168	(.024)	–	–
ρ	–.064	(.020)	–.002	(.024)	–.072	(.021)	–.525	(.036)

Table 7 – Estimates using Log(Earnings)

	Full model		No cont.		No mismatch		Basic model	
	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.
Log likelihood	–607.5		–646.5		–708.5		–881.2	
μ_ξ	12.283	(.032)	12.246	(.041)	12.178	(.046)	12.191	(.038)
σ_ξ	.717	(.021)	.843	(.026)	1.116	(.035)	.960	(.030)
μ_ζ	9.187	(.691)	11.387	(.254)	–	–	–	–
σ_ζ	1.807	(.424)	1.751	(.178)	–	–	–	–
μ_ω	–.304	(.174)	–	–	.432	(.187)	–	–
σ_ω	1.239	(.129)	–	–	1.407	(.151)	–	–
μ_η	–.048	(.007)	–.056	(.007)	–.047	(.008)	–	–
σ_η	.099	(.007)	.112	(.006)	.100	(.007)	.552	(.019)
π_r	.959	(.013)	.867	(.020)	–	–	–	–
π_s	.152	(.018)	.169	(.020)	.148	(.018)	–	–
π_ω	.156	(.028)	–	–	.187	(.026)	–	–
ρ	–.013	(.014)	.022	(.012)	–.013	(.014)	–.369	(.029)

Table 8 – Estimates using Log(Pensions)

	Full model		No cont.		No mismatch		Basic model	
	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.
Log likelihood	–589.5		–755.7		–653.6		–775.8	
female	–.422	(.048)	–.423	(.106)	–.375	(.063)	–.410	(.059)
age	.058	(.025)	.020	(.061)	.077	(.029)	.072	(.028)
age^2	–.085	(.041)	–.011	(.100)	–.111	(.048)	–.100	(.047)
fullret	.527	(.154)	.276	(.268)	.575	(.163)	.531	(.159)
semiret	.034	(.187)	–.487	(.554)	–.219	(.253)	–.150	(.273)
midedu	.224	(.055)	.167	(.124)	.201	(.075)	.165	(.069)
highedu	.156	(.089)	.352	(.167)	.111	(.090)	.088	(.099)
dkedu	.487	(.068)	.520	(.155)	.489	(.099)	.407	(.091)
μ_ξ	10.379	(.429)	10.964	(.814)	9.948	(.379)	10.059	(.390)
σ_ξ	.501	(.021)	1.060	(.071)	.669	(.027)	.572	(.021)
μ_ζ	8.957	(.307)	11.234	(.226)	–	–	–	–
σ_ζ	.764	(.235)	1.206	(.160)	–	–	–	–
μ_ω	–1.472	(.983)	–	–	–.305	(.809)	–	–
σ_ω	3.676	(.713)	–	–	3.707	(.610)	–	–
μ_η	–.049	(.013)	–.037	(.022)	–.040	(.015)	–	–
σ_η	.212	(.012)	.187	(.012)	.209	(.012)	.838	(.031)
π_r	.981	(.007)	.907	(.021)	–	–	–	–
π_s	.267	(.023)	.295	(.025)	.263	(.023)	–	–
π_ω	.056	(.018)	–	–	.077	(.019)	–	–
ρ	–.199	(.032)	–.161	(.027)	–.186	(.031)	–.459	(.073)

Table 9 – Estimates using Log(Pensions)

	Full model		No cont.		No mismatch		Basic model	
	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.
Log likelihood	–650.5		–777.3		–696.4		–830.3	
μ_ξ	11.742	(.032)	11.679	(.056)	11.693	(.038)	11.685	(.025)
σ_ξ	.628	(.030)	1.123	(.033)	.769	(.033)	.657	(.023)
μ_ζ	9.023	(.409)	11.256	(.222)	–	–	–	–
σ_ζ	.843	(.344)	1.202	(.159)	–	–	–	–
μ_ω	–1.632	(1.077)	–	–	–.302	(.773)	–	–
σ_ω	3.801	(.775)	–	–	3.608	(.584)	–	–
μ_η	–.044	(.014)	–.036	(.015)	–.038	(.015)	–	–
σ_η	.217	(.012)	.190	(.012)	.211	(.012)	.845	(.022)
π_r	.981	(.008)	.905	(.022)	–	–	–	–
π_s	.268	(.023)	.295	(.025)	.263	(.023)	–	–
π_ω	.050	(.017)	–	–	.080	(.020)	–	–
ρ	–.131	(.023)	–.117	(.022)	–.128	(.023)	–.361	(.067)

Table 10 – Estimates using Log(Taxes)

	Full model		No cont.		No mismatch		Basic model	
	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.
Log likelihood	–759.2		–818.5		–788.0		–924.4	
female	–.453	(.064)	–.412	(.080)	–.471	(.076)	–.509	(.059)
age	.030	(.023)	.028	(.022)	.039	(.020)	.035	(.020)
age^2	–.070	(.040)	–.076	(.042)	–.097	(.035)	–.087	(.036)
fullret	–.409	(.129)	–.368	(.117)	–.381	(.132)	–.320	(.100)
semiret	.153	(.120)	.234	(.129)	.156	(.157)	.198	(.183)
midedu	.332	(.083)	.290	(.095)	.291	(.091)	.284	(.077)
highedu	.381	(.103)	.406	(.110)	.389	(.107)	.394	(.099)
dkedu	.648	(.110)	.664	(.099)	.648	(.109)	.641	(.097)
μ_ξ	10.752	(.293)	10.741	(.269)	10.647	(.258)	10.709	(.256)
σ_ξ	.686	(.026)	.890	(.028)	.790	(.028)	.689	(.022)
μ_ζ	9.846	(.300)	10.235	(.137)	–	–	–	–
σ_ζ	1.197	(.145)	1.097	(.094)	–	–	–	–
μ_ω	–.591	(.360)	–	–	.025	(.188)	–	–
σ_ω	1.847	(.267)	–	–	1.639	(.136)	–	–
μ_η	.095	(.007)	.097	(.007)	.101	(.005)	–	–
σ_η	.105	(.006)	.103	(.006)	.106	(.006)	.567	(.018)
π_r	.921	(.023)	.848	(.020)	–	–	–	–
π_s	.216	(.020)	.239	(.021)	.199	(.018)	–	–
π_ω	.094	(.028)	–	–	.180	(.024)	–	–
ρ	–.043	(.013)	–.043	(.015)	–.056	(.014)	–.192	(.037)

Table 11 – Estimates using Log(Taxes)

	Full model		No cont.		No mismatch		Basic model	
	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.
Log likelihood	–840.0		–876.0		–862.8		–1018.2	
μ_ξ	10.839	(.089)	10.801	(.043)	10.784	(.042)	10.809	(.034)
σ_ξ	.846	(.117)	1.006	(.037)	.958	(.019)	.828	(.023)
μ_ζ	9.645	(.448)	10.220	(.140)	–	–	–	–
σ_ζ	1.189	(.167)	1.109	(.095)	–	–	–	–
μ_ω	–.474	(.442)	–	–	.051	(.174)	–	–
σ_ω	1.640	(.471)	–	–	1.500	(.130)	–	–
μ_η	.092	(.010)	.088	(.010)	.095	(.007)	–	–
σ_η	.107	(.008)	.108	(.007)	.105	(.005)	.572	(.015)
π_r	.935	(.031)	.853	(.020)	–	–	–	–
π_s	.213	(.020)	.235	(.021)	.199	(.018)	–	–
π_ω	.111	(.045)	–	–	.189	(.024)	–	–
ρ	–.011	.021	–.0004	(.019)	–.015	(.012)	–.133	(.030)

Table 12 – Different regressions for earnings

	Both sources		Administrative earnings						Survey earnings					
	Full model		OLS		Robust Regr.		Median Regr.		OLS		Robust Regr.		Median Regr.	
	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sds.
female	−.268	(.052)	−.142	(.076)	−.256	(.036)	−.226	(.037)	−.262	(.065)	−.244	(.036)	−.249	(.039)
age	.087	(.024)	.133	(.038)	.049	(.018)	.126	(.019)	.101	(.032)	.068	(.018)	.081	(.020)
age2	−.275	(.062)	−.416	(.102)	−.152	(.048)	−.380	(.049)	−.292	(.086)	−.203	(.048)	−.233	(.052)
fullret	−1.015	(.125)	−1.260	(.176)	−.663	(.084)	−.826	(.085)	−1.189	(.149)	−.591	(.084)	−.857	(.089)
semiret	−.421	(.114)	−.520	(.175)	−.433	(.084)	−.544	(.084)	−.316	(.148)	−.594	(.083)	−.597	(.089)
midedu	.244	(.102)	.219	(.136)	.145	(.065)	.120	(.065)	.423	(.115)	.192	(.064)	.168	(.069)
highedu	.345	(.101)	.334	(.141)	.283	(.067)	.313	(.068)	.479	(.119)	.317	(.067)	.302	(.072)
dkedu	.568	(.104)	.511	(.138)	.461	(.066)	.404	(.067)	.759	(.117)	.504	(.066)	.476	(.071)
cons	11.592	(.262)	11.158	(.385)	11.943	(.184)	11.326	(.188)	11.192	(.326)	11.708	(.183)	11.590	(.197)

Table 13 – Different regressions for pensions

	Both sources		Administrative pensions						Survey pensions					
	Full model		OLS		Robust Regr.		Median Regr.		OLS		Robust Regr.		Median Regr.	
	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sds.
female	–.422	(.048)	–.395	(.062)	–.432	(.042)	–.410	(.054)	–.476	(.096)	–.405	(.041)	–.381	(.053)
age	.058	(.025)	.083	(.030)	.064	(.020)	.057	(.025)	.028	(.046)	.063	(.020)	.034	(.025)
age^2	–.085	(.041)	–.120	(.049)	–.099	(.033)	–.083	(.041)	–.018	(.075)	–.097	(.032)	–.044	(.041)
fullret	.527	(.154)	.605	(.172)	.649	(.116)	.679	(.145)	.223	(.266)	.616	(.113)	.538	(.143)
semiret	.034	(.187)	–.066	(.257)	.675	(.174)	–.465	(.214)	–.501	(.398)	–.232	(.169)	–.695	(.211)
midedu	.224	(.055)	.180	(.072)	.200	(.049)	.169	(.062)	.099	(.111)	.221	(.047)	.244	(.061)
highedu	.156	(.089)	.036	(.105)	.256	(.071)	.196	(.091)	.304	(.162)	.289	(.069)	.299	(.088)
dkedu	.487	(.068)	.401	(.096)	.440	(.065)	.401	(.082)	.428	(.148)	.496	(.063)	.505	(.081)
cons	10.379	(.429)	9.861	(.419)	10.246	(.284)	10.294	(.354)	10.881	(.648)	10.223	(.276)	10.647	(.357)

Table 14 – Different regressions for taxes

	Both sources		Administrative taxes						Survey taxes					
	Full model		OLS		Robust Regr.		Median Regr.		OLS		Robust Regr.		Median Regr.	
	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sd.	Coef.	Sds.
female	–.453	(.064)	–.528	(.063)	–.394	(.041)	–.391	(.041)	–.456	(.073)	–.406	(.046)	–.379	(.053)
age	.030	(.023)	.037	(.022)	.025	(.014)	.042	(.014)	.027	(.025)	.024	(.016)	.058	(.018)
age^2	–.070	(.040)	–.092	(.039)	–.045	(.026)	–.071	(.025)	–.073	(.045)	–.041	(.028)	–.101	(.033)
fullret	–.409	(.129)	–.314	(.109)	–.405	(.071)	–.496	(.070)	–.338	(.126)	–.496	(.079)	–.575	(.091)
semiret	.153	(.120)	.182	(.192)	–.207	(.126)	–.118	(.121)	.242	(.221)	–.158	(.139)	–.174	(.156)
midedu	.332	(.083)	.286	(.084)	.218	(.055)	.203	(.054)	.282	(.097)	.292	(.061)	.277	(.070)
highedu	.381	(.103)	.387	(.102)	.324	(.066)	.337	(.065)	.412	(.117)	.393	(.074)	.379	(.085)
dkedu	.648	(.110)	.632	(.101)	.465	(.066)	.458	(.064)	.666	(.116)	.557	(.073)	.530	(.083)
cons	10.752	(.293)	10.663	(.271)	10.789	(.177)	10.578	(.174)	10.834	(.312)	10.864	(.196)	10.459	(.227)

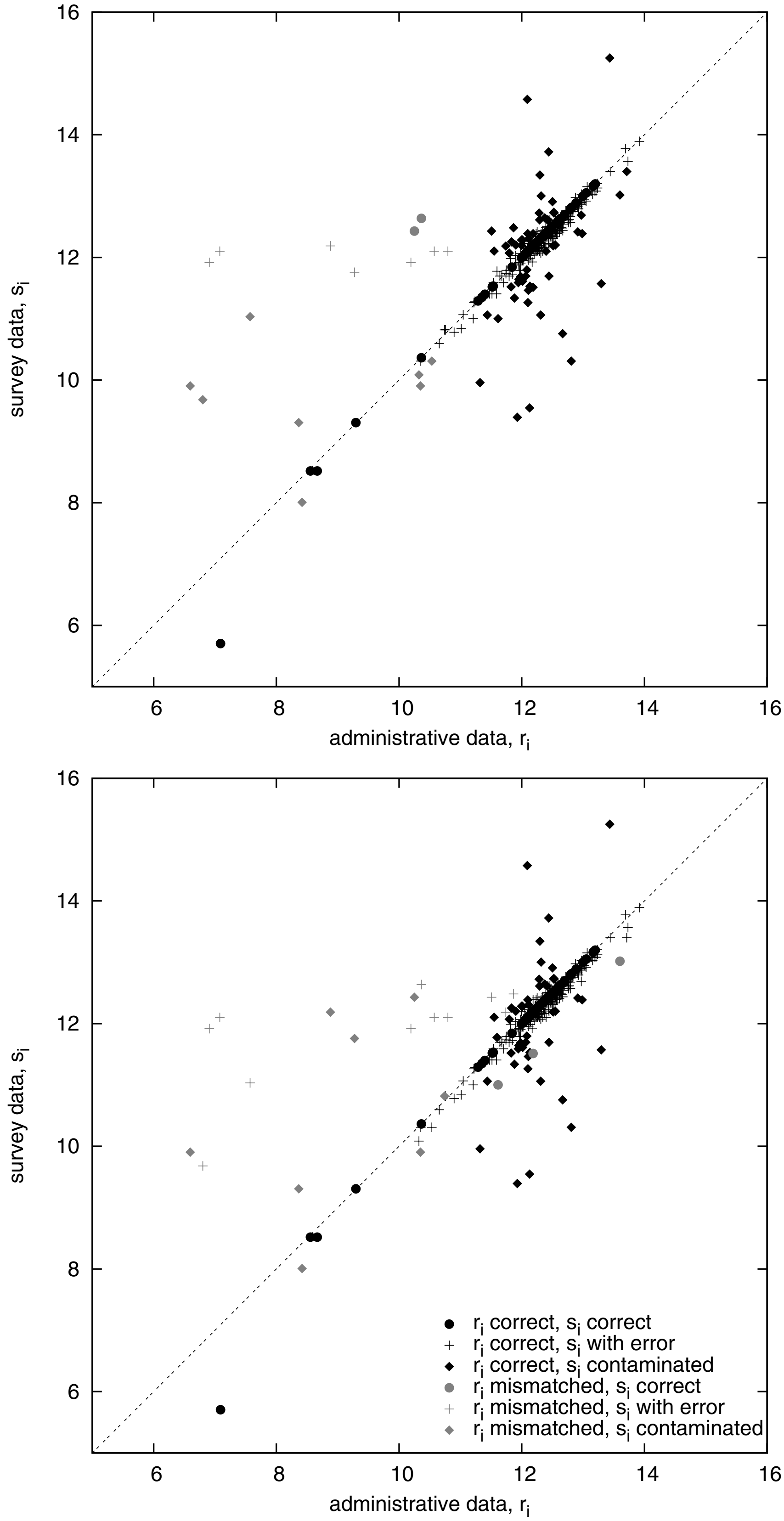


Figure 5 – ln(earnings) without (upper) and with (lower) covariates

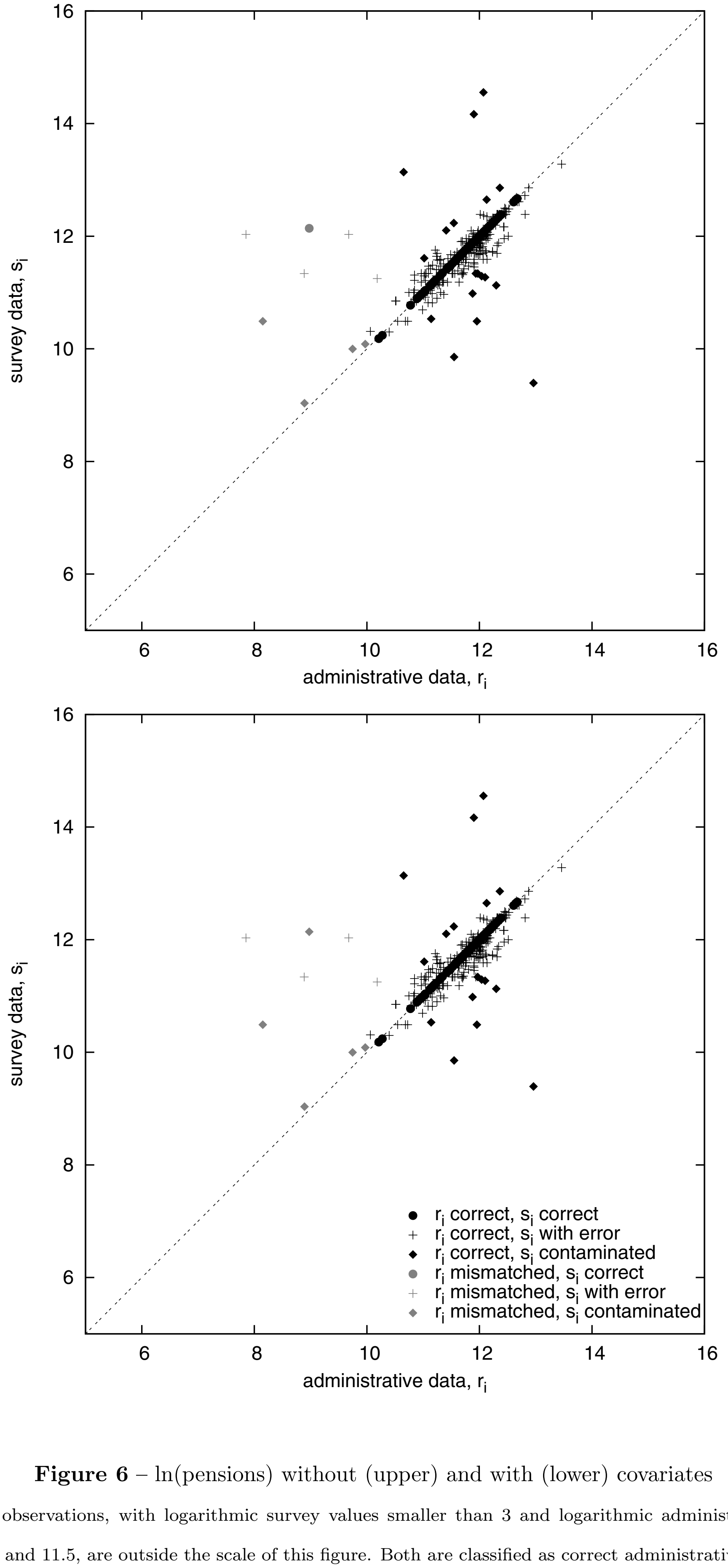
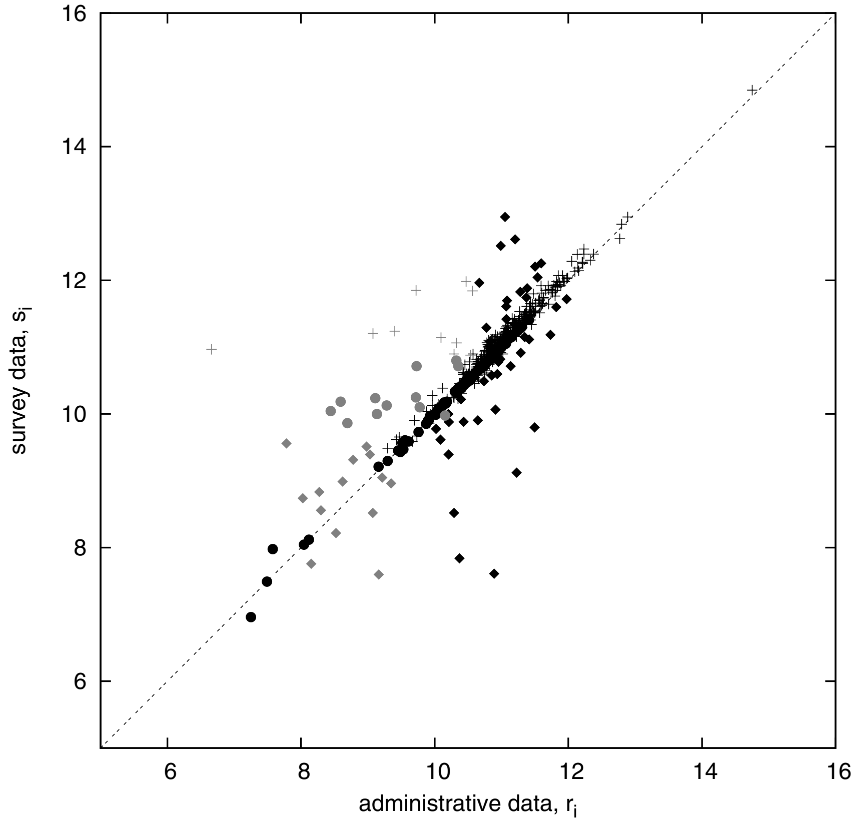


Figure 6 – ln(pensions) without (upper) and with (lower) covariates

Note: two observations, with logarithmic survey values smaller than 3 and logarithmic administrative values between 11 and 11.5, are outside the scale of this figure. Both are classified as correct administrative data and a contaminated survey value.



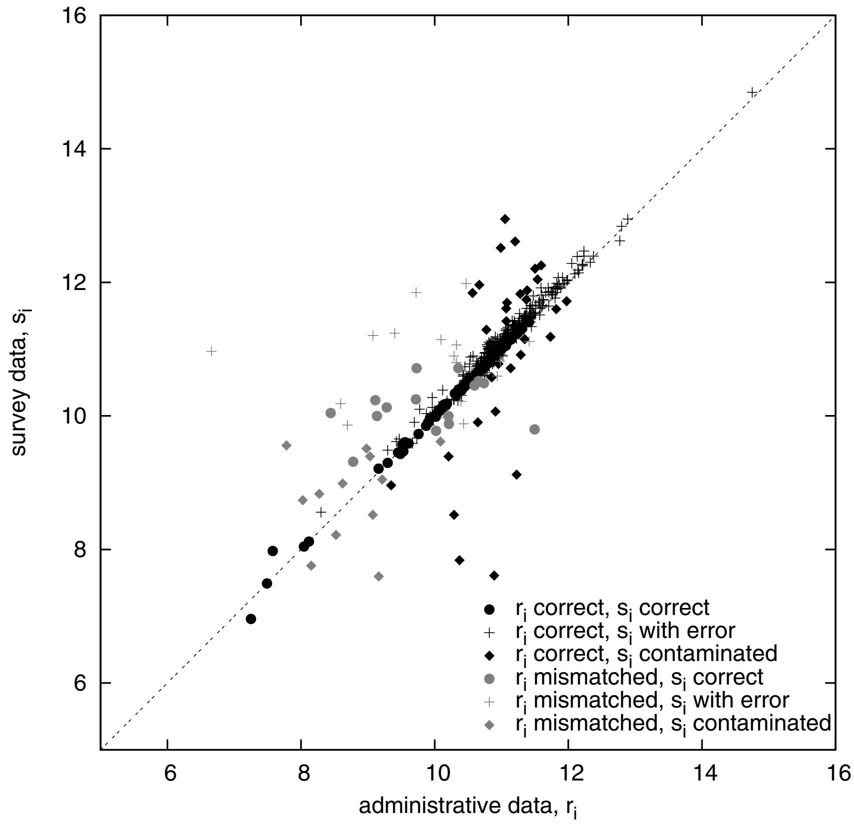


Figure 7 – ln(taxes) without (upper) and with (lower) covariates

Note: one observation, with logarithmic survey value smaller than 4 and logarithmic register value between 11 and 11.5, is outside the scale of this figure. The survey value is classified as contaminated and the register value is classified as correct (with covariates) or as a mismatch (without covariates).