FDZ

Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

FDZ·Methodenreport

Methodological aspects of labour market data

06/2018
EN

On Omitted Variables, Proxies and Unobserved Effects in Analysis of Administrative Labour Market Data

Documentation version: DOI: 10.5164/IAB.FDZM.1806.en.v1

Shihan Du,
Pia Homrighausen,
Ralf A. Wilke

Bundesagentur für Arbeit

On Omitted Variables, Proxies and Unobserved Effects in Analysis of Administrative Labour Market Data[*]

Shihan Du[†], Pia Homrighausen[‡], Ralf A. Wilke[§]

Abstract: Empirical research addresses omitted variable bias in regression analysis by means of various approaches. We present a framework that nests some of them and put it to German linked administrative labour market data. We find evidence for sizable omitted variable bias in a wage regression, while a labour market transition model appears to be less affected. Additional survey variables contribute only to the wage model, while the use of work history variables and panel models lead to changes in coefficients in the two models. Overall, unobserved effects panel data models with a restricted regressor set are found to control for more information than cross sectional analysis with an extended variable set.

Keywords: linked survey-administrative data, statistical regularisation

[*] We thank the DIM unit of the IAB for providing the data and Arne Bethmann for his support with the PASS data.

[†] Copenhagen Business School, Department of Economics, Porcelaenshaven 16A, 2000 Frederiksberg, Denmark, E–mail: sd.eco@cbs.dk

[‡] Institute for Employment Research (IAB), E–mail: pia.homrighausen@iab.de.

[§] Corresponding author: Copenhagen Business School, Department of Economics, Porcelaenshaven 16A, 2000 Frederiksberg, Denmark, Phone: +4538155648, E–mail:rw.eco@cbs.dk

1 Introduction

Linked administrative data are increasingly used for empirical research in economics, social sciences and related disciplines. Their main advantages over survey based data sources are bigger sample sizes and higher precision of key variables. Administrative data cover the population and therefore its availability is not restricted to smaller and possibly non-random samples. Key variables are generated through operations in firms and public services and should be less prone to misclassification due to recall errors of respondents. However, administrative data have also disadvantages over survey data. The variable set is restricted to information generated through operations. Thus, there is often a systematic lack of information on everything that exceeds the operational processes. This includes for example the motivation of individuals, their personality traits, the size of social network and working climate in firms among many other things. Indeed, a number of studies based on survey data has shown that such additional variables contribute to the model. Beside that their availability enables the researcher to analyse problems which couldn't be analysed with administrative data. Examples include Nyos and Pons (2005), Mueller and Plug (2006), and Heineck and Anger (2010) who use survey data with information about personality traits to analyse individual labour market outcomes.

The existence of administrative data does not directly imply that all information collected is indeed accessible to the researcher. In particular, not all variables may be available due to a lack of data linkage between administrative registers. Moreover, usually data providers only give access to a random sample of the population data and only to a restricted set of variables due data confidentiality restrictions. Therefore, typical research based on administrative data is far away from using complete information about the population with all variables collected in administrative processes. A sizeable random sample should not raise too many concerns for making inference with these data, however, the unavailability of important variables casts concerns for the consistency of estimates. There is extensive literature that considers the problem of omission of variables in regression analysis. For example Gelbach (2016) suggests a variable selection approach that takes into account how much the omission of an available variable induces a bias for the coefficients on the other still included variables. Oster (2017) presents a comprehensive treatment of omission of variables because they are unavailable. She suggests approaches how the size of the resulting omitted variable bias can be approximated under restrictions.

In this paper we focus on common empirical strategies for reducing omitted variable bias in labour market research. First, empirical research often uses constructed variables from the individual work history. Examples include Kauhanen and Napari (2012) who use linked employer-employee data to study career and wage dynamics within and between firms in Finland. Fernández-Kranz and Rodríguez-Planas (2011) investigate the earnings effect of women who switch to part-time work under different types of contracts in Spain. Their study is based on longitudinal Spanish data from social security records. Baptista et al. (2012) obtain new insights in career mobility using Portuguese longitudinal matched employer-employee data. Using German administrative data, Biewen et al. (2014) conduct an analysis of treatment effects of labour market programmes. Work history variables may directly belong to the population model or they might be proxies for otherwise unobserved variables such as performance. A prominent example in labour economics is that human capital is difficult to measure and usually unobserved. However, human capital is supposed to be an important variable in wage regression models. Thus, researchers use test scores, such as the IQ, as proxies for human capital (compare Neal and Johnson, 1996; Bollinger, 2003). While the use of proxies is practically appealing, there is no guarantee that their use leads to a bias reduction or consistent estimation.

Second, another approach to mitigate the omission of variables is adding survey based variables to the administrative data, especially information on personal traits. While adding variables is appealing, the generation of survey data is typically costly and time consuming. Moreover, the question arises to what extent these variables indeed contribute to the model. Third, the availability of panel data makes it possible to control for correlated unobserved time invariant effects, reducing the need to control for as many variables as possible compared to cross sectional analysis.

Despite the widespread use of work history and survey based variables, little systematic research has been conducted to assess how they contribute to the estimation of the models. Attempts to investigate their role are so far restricted to sensitivity analysis, which is how their inclusion into the model affects estimation results. For example Lechner and Wunsch (2013), Arni et al. (2014) and Caliendo et al. (2014) investigate whether estimated treatment effects of labour market programmes on labour market outcomes are sensitive with respect to the inclusion of additional variables. Our analysis exceeds a sensitivity analysis as it tests a number of relationships and restrictions that can be partly derived from a panel model. This is helpful in obtaining a deeper understanding on the viability of the different approaches.

We use a sample of linked administrative data which are linked to extensive survey data from Germany. In particular, we use the Integrated Employment Biographies (IEB) of the Institute for Employment Research (IAB), which is linked with the Panel Study "Labour Market and

Social Security" (PASS). The PASS survey was funded by the German government to provide a more comprehensive data base for the evaluation of the effects of the so-called Hartz reforms during the 2000s. Our data therefore contains many non-operations based variables which are not available in administrative data. Centered around this scenario we provide a formal framework for estimation bias due to omission of important variables and estimation bias due to imperfect proxy variables. Our starting point is a widely used administrative data product with only a limited number of variables. We then assess to what extent additional survey based non operations related variables and work history variables contribute to the model and change the results. We suggest a statistical framework that allows us to test the conditions for the work history variables to be feasible proxy variables. Moreover, we relate the results of cross sectional analysis with those of panel analysis to investigate to what extent additional cross sectional variables explain the variation in unobserved individual time invariant effects. In our analysis we do exemplary wage regressions and an analysis of labour market transitions. Our results suggest that additional cross sectional variables control for considerably less relevant information than fixed effects in panel analysis. Panel data analysis is found to give significantly different results, in particular in the wage regression model. The endogeneity of a number of regressors in the cross sectional models is confirmed.

In Section 2 the econometric problem is outlined. Section 3 describes the data and section 4 presents the empirical findings. The last section summarises.

2 The model

We consider the situation when a researcher has access to some standard administrative data product, which means that only a smaller number of administrative registers is linked. Thus, the set of available variables is restricted to some core variables. We restrict ourself to the multiple linear regression model and the population model is

$$y = X\beta + W\gamma + v, \tag{1}$$

where β $(J \times 1)$ and γ $(L \times 1)$ are unknown parameters, X $(1 \times J)$ are observable regressors (including the first element being a constant) and W $(1 \times L)$ are unobserved regressors. We will later relax this to some of the components of W being observed. We assume that the components of X and W are not perfectly multicollinear. y is observed and v is unobserved. We assume $E(v|X, W) = 0$. Because W is unobserved, the model in (1) cannot be directly estimated. Instead, one could omit the unobserved variables and use OLS to estimate the model

$$y = X\beta + u, \tag{2}$$

where $u = W\gamma + v$. This is what is typically estimated in applications. It is well known that if $\text{cov}(x_j, u) \neq 0$ for some j causes $\hat{\beta}$, the OLS estimator for β, to be inconsistent. We focus here on a model with an unknown number of omitted variables as this is the most realistic scenario in applications. When there are more than one omitted variables, the L linear projections of W onto the observable regressors are

$$W = X\delta + R,$$

with δ is $J \times L$ and R is $1 \times L$. Let r_l be the l'th component of R. By definition $E(r_l) = 0$ and $\text{cov}(x_j, r_l) = 0$ for $j = 1, ..., J$ and $l = 1, ..., L$. When plugging W into (1) we obtain

$$y = X(\beta + \delta\gamma) + R\gamma + v.$$

In this model, all regressors are uncorrelated with the composite error and therefore the probability limit of the OLS estimator $\hat{\beta}$ for model (2) is

$$\text{plim}\hat{\beta} = \beta + \delta\gamma. \tag{3}$$

This is the well known omitted variables bias and its size depends on the strength of the partial correlation between W and X and the size of the elements of γ, i.e. the relevance of the omitted variables in the population model (1). Since W is not observed, the size and direction of the bias are unknown in an application. This is in contrast to the approach in Gelbach (2016) that focuses on variable selection. Oster (2017) provides an in-depth analysis of omitted variable bias and derives expressions that can be estimated under several restrictions without observing one or multiple W. The restrictions are on the relationship between X and the omitted factors (proportional selection relationship) and knowledge of the R^2 of the population model. In particular, her model considers the case of one component of X being related to W and requires that the components of W to be uncorrelated. We have applied her method to our problem using information on some of the components of W and found that the sign and magnitude of the estimated proportional selection relationship jumped strongly across variables. Given this instability and that the restrictions on her model exceed what we assume in our model, we focus on alternative approaches aiming at reducing the omitted variable bias. However, non of these approaches is able to entirely remove bias or reveal the size of the bias in absence of additional restrictions.

One approach to mitigate omitted variable bias is to plug-in in generated variables from the observable history of cross section units. In labour market research these are for example variables that characterise the work history of an individual and not simply lagged observable variables. These are denoted as Z $(1 \times P)$. We assume that none of the components of X and Z are highly correlated or perfectly multicollinear in an application. In most applications P is a small integer and $P \leq L$. This means there are fewer constructed variables than omitted variables. The role of Z requires some discussion. For the reasons provided in the introduction, a special case is attained if a z_j is a proxy variable for one unobserved w_l, i.e. $z_j = w_l + error$ with $E(error) = 0$. However, more generally z_j can be related to any W, i.e. $z_j = \theta_0 + W\theta_j + m_j$ with $E(m_j|W) = 0$ for all j. θ_0 (1×1) and θ_j $(L \times 1)$ are unknown to the researcher. If z_j is a proxy for w_l, then only the l'th element of θ_j is nonzero. This is the case that is typically considered by the proxy variable literature (Lubotsky and Wittenberg, 2006, Bollinger and Minier, 2015). Using Z instead of W can be also interpreted as a measurement error problem. Here any deviation from the linear combination $W\theta_j$, which is m_j, is the measurement error. Alternatively, one could think of $z_j \in W$. In this case the constructed variable would directly belong to the population model. Then $m_j = 0$, one component of θ_j is 1 and the others are 0. Lastly, z_j may not be correlated with any component of W. In this case $\theta_j = 0$ and z_j should not be included at all. A researcher normally faces the problem of not knowing the exact role of the components of Z. In any case it depends on the statistical relationship between X, W and the m_js, whether the inclusion of Z mitigates or increases the omitted variable bias. Given that W and L are unknown, it is more convenient to write the linear projection on the linear combination of Ws, i.e, $W\gamma = \alpha + Z\lambda + e$ with $E(e|Z) = 0$) and parameters α (1×1) and λ $(P \times 1)$. e can be interpreted as the measurement or approximation error between $W\gamma$ and $Z\lambda$, which is the variation in the linear combination of unobserved variables that is not explained by the linear combination of constructed and included variables in Z. Therefore

$$\begin{aligned} y &= X\beta + W\gamma + v \\ &= X\beta + Z\lambda + \alpha + e + v. \end{aligned} \tag{4}$$

For β in model (4) to be consistently estimated by OLS, it is additionally required that e is uncorrelated with X and v with Z. The former is not the case if X plays a role in the linear projection of Z and X on $W\gamma$, so it is required $E(W\gamma|X, Z) = E(W\gamma|Z)$. The latter requires $E(y|X, W, Z) = E(y|X, W)$, i.e. the redundancy of Z in the population model. Whether the bias in $\hat{\beta}$ in model (4) is smaller or greater than in model (2) is an empirical question. This depends on whether the correlations between the components of X and $W\gamma$ are greater or smaller than the correlations between the components of X and e, respectively. If for example the size of the components of δ are zero or very small, the inclusion of Z will increase the bias in $\hat{\beta}$ if there is correlation between Z and both X and v. Evidently, the better the fit of the model for $W\gamma$ on Z, the more likely plugging in Z leads to bias reduction. This is because e becomes smaller in magnitude which reduces its covariance with X. It is remarked that λ has the interpretation of parameters of the linear projection on $W\gamma$ and we ignore the identifiability of α and the first component of β because the intercept is assumed to be not of interest.

Another approach to mitigate omitted variable bias is to enhance the regressor set by conducting a survey or by using additional administrative variables that are normally not accessible. Suppose that a subset W_1 of W, by assumption the first L1 variables of W, is observable for some random sample of the population. The idea is to do an analysis with a richer variable set. For direct comparability of the results across models we always restrict the analysis to the cross section units for which we have information on W_1. Thus, we ignore the potential loss in precision and focus on asymptotic bias only. We consider the case, where the researcher is primarily interested in estimating the partial relationship between y and elements of X, rather than between y and elements W_1, although the latter will be typically also of interest. W_2 is $1 \times L2$ and comprises of the last $L2$ elements of W with $L1 + L2 = L$. W_2, the remaining unobservable variables, may be correlated with X and W_1. Therefore, their omission induces a bias for estimated β and $\gamma_{(1)}$ in the regression of y on X and W_1:

$$y = X\beta + W_1\gamma_{(1)} + u_2, \tag{5}$$

where $\gamma_{(1)}$ contains the first $L1$ elements of γ and $u_2 = W_2\gamma_{(2)} + v$, where $\gamma_{(2)}$ consists of the last $L2$ elements of γ. Unfortunately, there is no guarantee that including more variables indeed reduces the bias but in practice one should expect this. The reason is that the number of summands in the bias term in equation (3) decreases from L to $L2$, when reducing the number of omitted variables. However, this may not lead to a reduction in the bias as the magnitude and sign of the various components of δ and γ are not restricted.

Instead of enhancing the set of observable variables, one can exploit the availability of longitudinal information, i.e. panel data, to mitigate the bias from the omission of W. y, X and Z are observed in periods $t = 1, ..., T$ with $T \geq 2$ and observations are denoted as y_{it}, X_{it} and Z_{it}, respectively, for units $i = 1, ..., N$. W_1 is assumed to be observed in one period only and W_2 is never observed, thus, W has to be omitted from the model. In order to relax the exogeneity restrictions on X, we consider a fixed effects model:

$$y_{it} = X_{it}\beta + a_i + q_{it}$$

with $a_i + q_{it} = u_{it}$. a_i is assumed to be time invariant (the so called fixed effect) and q_{it} is a time varying error. Though, X is allowed to be correlated with a, the fixed effects estimator will only consistently estimate β if $E(q_{it}|X_i, a_i) = 0$ with $X_i = (X'_{i1}, ..., X'_{iT})'$. However, this depends on the relationship between W and X because

$$\begin{aligned} y_{it} &= X_{it}\beta + W_{it}\gamma + v_{it} \\ &= X_{it}\beta + (\bar{W}_i + C_{it})\gamma + v_{it} \\ &= X_{it}\beta + a_i + q_{it} \end{aligned} \tag{6}$$

with $\bar{W}_i = \sum_{t=1}^{T} W_{it}/T$ and $q_{it} = C_{it}\gamma + v_{it}$. a_i therefore corresponds to the time constant part of $W_{it}\gamma$, which is not only the time constant variables in W but also the time average of the time varying components of W. $E(C_{it}\gamma|X_i, \bar{W}_i\gamma) = 0$ is required for consistent estimation by means of a fixed effects panel data model provided that v is idiosyncratic. It is also insightful to consider the role of Z when used in the fixed effects model. As discussed above, $W\gamma$ can be expressed as a linear combination of the Z plus a measurement error. In terms of the panel model this is $W_{it}\gamma = Z_{it}\lambda + b_i + s_{it}$. This linear projection decomposes the measurement error into a time constant part (b_i) and a time varying part (s_{it}). Then, for the main model we have

$$\begin{aligned} y_{it} &= X_{it}\beta + W_{it}\gamma + v_{it} \\ &= X_{it}\beta + Z_{it}\lambda + b_i + s_{it} + v_{it}. \end{aligned} \tag{7}$$

In order to consistently estimate β by means of a fixed effects model, b_i is allowed to be correlated with X_{it} and Z_{it}, but we need $E(s_{it}|X_i, Z_i, b_i) = 0$ and $E(v_{it}|X_i, Z_i, b_i) = 0$ with $Z_i = (Z'_{i1}, ..., Z_{iT})'$. The latter is again satisfied if Z does not play a role in the population model. The former, however, requires some discussion. b_i captures all time constant features of W which are not being absorbed by Z. The more of the time varying information of W is captured by Z, the smaller is s_{it}. If the time varying information in Z_{it} is related to the time varying part of W_{it}, s_{it} is smaller in size than $C_{it}\gamma$. Then the inconsistency of the estimated β compared to model (6) is smaller. If the measurement error is time constant, i.e. $s_{it} = 0$, the fixed effects estimator for model (7) is consistent (Wooldridge, 2010). A roughly time constant measurement error (i.e. $s_{it} \approx 0$) may not be implausible in applications if Z_{it} has the interpretation of containing proxies.

In our empirical analysis we do a comparative estimation of the various approaches in order to see to what extent the results are sensitive. Our analysis exceeds a sensitivity analysis by relating the result patterns to the theoretical considerations outlined in this section. The underlying theory also provides a starting point for testing and checking restrictions, to obtain a better understanding of the role of Z, how the considered approaches relate in terms of the ability to control for parts of $W\gamma$ and to provide insights which of the X and Z show evidence of endogeneity.

The availability of W_1 makes it possible to get some ideas of how usually omitted variables are related to Z. In particular, one can estimate the strength of the relationship between $W_1\gamma_{(1)}$ and the Z. This shows which of the Z variables are related with unobservables and how much the variation in Z is able to explain the variation in $W_1\gamma_{(1)}$. A high R^2 would point to small measurement error. One can also test restrictions required for Z being a set of valid proxy variables, however, valid inference requires that a model without the omitted W_2 can be consistently estimated, i.e. W_2 is uncorrelated with all included variables. Testable restrictions are $E(W_1\gamma_{(1)}|X,Z) = E(W_1\gamma_{(1)}|Z)$ and $E(y|X,W_1,Z) = E(y|X,W_1)$, which have been motivated above. However, any correlations between (X,Z) and W_2 invalidate the inference.

Once panel models (6) and (7) have been estimated, one can relate the estimated fixed effects to W_1 and Z in a cross sectional model. It had been shown that $a = \bar{W}\gamma$ in model (6) and $b = W\gamma - Z\lambda - s$ in model (7). However, given that only W_1 is observed in one period, the following linear projections are suggested:

$$\hat{a} = W_1\rho + d \tag{8}$$

$$\hat{b} + Z\hat{\lambda} = W_1\varrho + f, \tag{9}$$

with d and f being unobserved and uncorrelated with W_1 and $E(d) = E(f) = 0$. The dependent variables in these models are the estimated components of the panel models (6) and (7) that are supposed to control for the omitted W. These regressions tell us two things: First, whether there is a linear partial relationship between the components of W_1 and the dependent variables. This shows which components of W_1 are indeed at least to some extent controlled for. Second, the R^2 of these models shows us how much the variation in W_1 explains the variation of the components that control for W. A low R^2 would point to that the panel models mainly control for information that is not in W_1 and Z, thus information in W_2. This would suggest that a panel analysis using a reduced regressor set is expected to be the more fruitful empirical approach than a cross sectional analysis with an expanded regressor set. In contrast, if the R^2 was high, the reverse applies. This suggests that the fixed effects capture only little time constant information of W_2, meaning a fixed effects panel analysis does not control for much more than what is in W_1. It is remarked that the R^2 of models (8) and (9) increases with $L1$ and approaches 1 if the entire W was used. Moreover, the models use W_1 at one time point and not the time constant part of W_1 which is expected to result in a lower R^2. However, the more important the cross sectional variation in W_1 than the longitudinal variation, the smaller the expected effect on the R^2.

Finally, simple regression based tests of the endogeneity of X and Z can be conducted once fixed effects have been estimated. The idea is here to regress $\hat{a}$ or $\hat{b}$ on X or (X, Z), respectively. Any significant relationship points to that the fixed effects are partially correlated with the observables, thus leading to inconsistencies of OLS estimates for β for models (2) or (4). These tests will also reveal which variables or groups of variables possess these patterns. As an extension we outline in Appendix II how the validity of an instrumental variable can be checked for cross sectional models (2) and (4), when W_1, a or b were available.

3 German Administrative Data linked with Survey Data

For our analyses we use the Integrated Employment Biographies (IEB) of the IAB. These administrative registers contain information for every German once employed in a job subject to social insurance contributions since 1973. This information includes socio-demographic characteristics as well as daily records on employment and job seeking periods, receipt of unemployment benefits and information about participation in active labour market policy programs.

Usually, access to these data is restricted to random samples and a subset of variables due to data confidentiality reasons. In our application we mimic the situation of a researcher working with a standard administrative data set, which is accessible to a wider group of data users. In particular, we focus on the widely used scientific use file version of the "Sample of Integrated Labour Market Biographies" (SIAB, cf. vom Berge et al. 2013). The SIAB is a 2 percent random sample drawn from the IEB (approximately 1.6M individuals) and provides restricted access to variables available in the IEB records. The SIAB is available as a standard data set through the Research Data Center (FDZ) of IAB (http://fdz.iab.de/).

We enrich the administrative data by linking it with comprehensive survey data on individual level, with the household panel study "Labour Market and Social Security" (PASS, cf. Berg et al. 2012). The PASS survey was implemented in 2006 to gain more insights into the living conditions of (means-tested) unemployment benefit recipients in the household context. Since then, the PASS survey in general provides several waves of survey data from household and individual interviews on a wide variety of issues relating to the socioeconomic situation. About

80 percent of the individuals interviewed in the PASS survey agreed to linking the PASS survey data to the administrative records (approximately 22,000 individuals). A very similar linked dataset is the ”PASS survey data linked to administrative data of the IAB” (PASS-ADIAB) that is also available through the Research Data Center (FDZ) of IAB. For more information on these data see Antoni and Bethmann (2014).

Table 1: Data sources

	Size	IEB Variables	SIAB Variables (X)	PASS survey variables (W_1)
Integrated Employment Biographies (IEB)	100% of the population	x	x	
Sample of Integrated Labour Market Biographies (SIAB)	2% of IEB		x	
Panel Study "Labour Market and Social Security" linked with IEB (PASS-ADIAB)	0.03% of IEB		x	x

For our comparative analysis we restrict the sample to individuals aged 16 to 64 of different households who have participated in the 5th wave of the PASS survey in 2011. This leaves us with approximately 9,700 individuals. Since it is a common situation that survey data is only available for one period, we do not use further waves of the PASS survey. We restrict the analysis to the 5th wave to have information on personality traits that are not available in prior waves. Using both, the restricted IEB data as well as information from the PASS data, our sample contains variables from administrative registers available in the SIAB (X), generated work history variables (Z) as well as additional survey-based variables from PASS (W_1).

With these data, we perform two exemplary applications: one wage regression and one labour market transition analysis. Focusing only on individuals who are observed at least three years in the administrative data, the sample of the wage regression consists of 2,435 persons employed during the interview months. The sample of the transition regression consists of 1,484 persons who once have been registered as unemployed during the interview year and are observed at least for three years in the administrative data. The dependent variable y of the wage regression is the logarithmized average daily gross wage at the time of the interview. X includes socio-demographic and employment related variables such as gender, age, trainee status, education, nationality, and industrial sector. The dependent variable y in the transition analysis is a dummy variable indicating whether an unemployed individual left unemployment within 12 months ($y = 1$) or not ($y = 0$). As regressors we use a subset of the variables of the wage regression as well as dummies of unemployment related registers such as the receipt of unemployment insurance benefits (German: Arbeitslosengeld, ALG I) and means-tested unemployment benefits (German: Arbeitslosengeld II, ALG II). Table 10 in Appendix III presents the full set of regressors used in the wage and transition regression as well as their descriptive statistics.

The survey based variables constituting W_1 are linked PASS data. Among the survey variables those are of particular interest that are supposed to have an impact on wage levels and/or labour market transitions. While the survey incorporates a wide array of topics, we mainly focus on labour market related information. This includes information on personality traits and attitudes (Big Five), job search, working hours and other social factors. Table 10 in Appendix III presents the full set of survey variables used as well as their descriptive statistics.[1]Despite that we use a rich set W_1 variables, there may well be further important variables in the population models that are unobservable to us (and thus in W_2).

Variables Z are constructed from individual (un-) employment histories. Thus, they are computed from past administrative records on employment and unemployment among other past labour market outcomes. We construct four variables for the wage regression: length of job tenure, share of time employed over total length of recorded labour market history, past unemployment history, and working experience. For the transition analysis we construct five variables: past unemployment history at time of transition, duration of current unemployment episode, recall history, past long-term unemployment (i.e. last unemployment episode longer than 12 months), and participation in active labour market programmes within the last three years.

4 Empirical Analysis

Suppose a researcher has access to some standard administrative data product as described in Section 3. This contains y and X for multiple periods and can be enhanced with constructed work history variables (Z) and additional variables that are collected through a survey (W_1). The latter are available for one period and for a subset of the population only. Due to the limited size of the survey population we restrict ourselves to two exemplary linear regression models: A wage regression and a linear probability transition model. For both models we do the analysis steps as outlined in Section 2. From our findings we derive some general guidance for empirical researchers who work with these or similar data.

The idea behind using work history variables Z in wage or transition models is twofold: These

[1] See www.fdz.iab.de for a full list of variables available in the PASS survey data.

variables capture otherwise unobserved individual features related to past labour market performance and therefore they can be interpreted as proxy variables. In our application Z include among others the unemployment history and tenure with the current employer. While past unemployment experiences should be related to work motivation and performance, the tenure in a job should reflect job specific skills. Thus, these variables are correlated with something that is typically not observable.

However, work history variables may actually belong to the population model. This is for example if past unemployment experiences play a direct role in hiring decisions and therefore for the probability of starting a new job. Similarly job safety or the collective wage bargaining process can be direct functions of tenure due to legal restrictions. In many countries, dismissal protection is stronger for long-time employees and recently hired employees are normally not entitled to wage increases. However, if a component of Z belonged to the model, it is correlated with unobservables for the reasons mentioned above. Thus, it is endogenous. This is why adding additional variables W_1 to the model is expected to not only uncover endogeneity of X but in particular of Z. The PASS data provide a large number of additional variables. We apply the LASSO and an elastic net (see Appendix I) as tools for selection of relevant variables in the two models. While for the wage regression 35 variables are selected as set of relevant W_1 variables (see Table in Appendix III), none of the survey variables appears to be relevant in the transition model.

4.1 Wage Regression

We consider a standard Mincer type wage equation with y is the log of the average daily daily gross wage. As regressors X we use individual level and firm level data such as age, gender, education and industry. As Z we use work history variables related to previous working experience, tenure and previous unemployment experiences. For a complete list of variables in this model see Table 10 in Appendix III.

As the first step we apply ordinary least squares to estimate linear models for $E(y|X)$, $E(y|X,Z)$, $E(y|X,W_1)$ and $E(y|X,Z,W_1)$. Table 2 contains the main estimation results for these models, denoted by W.A-W.D. The coefficients on W_1 are for completeness reported in Table 11 in Appendix III. The R^2 increases from 0.32 in Model W.A to 0.41 in Model W.B and to 0.50 in Model W.C. It increases further to 0.57 in Model W.D, pointing to that the set of variables individually and jointly contribute to explaining variation in the dependent variable. It is found that a number of coefficients on X differ considerably across the models, pointing to omitted variable bias. For example the coefficient on gender decreases from 0.499 in Model W.A to 0.148 in Model W.D. This suggests that the estimated gender wage gap is much smaller (only around 14% compared to 39%) when non operations based variables are included. Although still highly significant, this is an economically relevant reduction. In contrast, other coefficients such as nationality and several business sectors are invariant across models W.A-W.D.

Table 2: Wage regression: Dependent variable log(wage)

	W.A	W.B	W.C	W.D	W.E	W.F
	$E(y\|X)$	$E(y\|X,Z)$	$E(y\|X,W_1)$	$E(y\|X,Z,W_1)$	$E(y_{it}\|X_{it})$	$E(y_{it}\|X_{it},Z_{it})$
	coef. / (SE)	coef. / (SE)	coef. / (SE)	coef. / (SE)	coef. / (SE)	coef. / (SE)
Gender (male=1)	0.499***	0.443***	0.174***	0.148***	6.537*	6.199*
	(0.026)	(0.024)	(0.027)	(0.026)	(3.564)	(3.498)
Age	0.006***	-0.001	0.011***	0.002	-0.055	-0.047
	(0.001)	(0.001)	(0.001)	(0.001)	(0.084)	(0.082)
Dummy: trainee	-0.437	-0.342	-0.628**	-0.518**	-0.494***	-0.515***
	(0.369)	(0.305)	(0.285)	(0.246)	(0.166)	(0.164)
Missing information on education	-0.528***	-0.438***	-0.394***	-0.300*	0.060	0.055
	(0.163)	(0.154)	(0.141)	(0.154)	(0.180)	(0.175)
No formal degree	-0.264**	-0.215**	-0.146	-0.112	-0.053	-0.040
	(0.117)	(0.107)	(0.090)	(0.081)	(0.146)	(0.141)
Vocational training	0.030	-0.008	0.045	0.022	0.062	0.080
	(0.113)	(0.103)	(0.085)	(0.076)	(0.133)	(0.127)
Higher Education	0.522***	0.483***	0.434***	0.417***	0.050	0.056
	(0.117)	(0.107)	(0.088)	(0.079)	(0.130)	(0.125)
Dummy: German nationality	0.030	-0.057	0.001	-0.072	-0.048	-0.030
	(0.058)	(0.055)	(0.048)	(0.045)	(0.130)	(0.127)
Agriculture	-0.627***	-0.443***	-0.558***	-0.404***	-0.189	-0.174
	(0.097)	(0.093)	(0.077)	(0.080)	(0.561)	(0.561)
Hotel and restaurant	-0.543***	-0.368***	-0.571***	-0.436***	-0.309	-0.328*
	(0.076)	(0.074)	(0.063)	(0.059)	(0.200)	(0.197)
Construction	-0.310***	-0.211***	-0.284***	-0.200***	0.139	0.130
	(0.055)	(0.053)	(0.049)	(0.047)	(0.155)	(0.155)
Trade	-0.249***	-0.175***	-0.193***	-0.135***	-0.025	-0.022
	(0.038)	(0.035)	(0.034)	(0.031)	(0.087)	(0.086)
Services	-0.232***	-0.124***	-0.198***	-0.115***	-0.108*	-0.115*
	(0.034)	(0.032)	(0.032)	(0.030)	(0.064)	(0.063)
Education and social health	-0.136***	-0.054	-0.092***	-0.033	-0.066	-0.073
	(0.037)	(0.035)	(0.032)	(0.030)	(0.101)	(0.099)
Public institutions	0.082*	0.070	0.086**	0.076**	0.158	0.153
	(0.045)	(0.044)	(0.037)	(0.036)	(0.179)	(0.177)
Other sectors	-0.083	-0.010	-0.015	0.038	-0.514**	-0.502**

Continued on next page

... table 2 continued

	W.A	W.B	W.C	W.D	W.E	W.F
	(0.074)	(0.061)	(0.062)	(0.050)	(0.226)	(0.218)
Tenure (in years)		0.019***		0.018***		0.005
		(0.002)		(0.002)		(0.006)
Share of working experience over total observation time		0.217***		0.118***		-0.123
		(0.047)		(0.043)		(0.106)
Additional working experience (in years)		0.011***		0.013***		0.008
		(0.002)		(0.002)		(0.007)
Dummy: unemployment history in the past		-0.492***		-0.427***		0.101
		(0.032)		(0.029)		(0.145)
Constant	3.770***	4.200***	2.506***	2.996***		
	(0.131)	(0.125)	(0.189)	(0.178)		
N	2435	2435	2435	2435	3× 2435	3×2435
R^2	0.319	0.412	0.502	0.570	0.997	0.997

Robust standard errors of model W.A-W.D and clustered standard errors of model W.E-W.F in parentheses.
* p<0.10, ** p<0.05, *** p<0.010

The coefficients on several components of X, such as gender and higher education, change monotonically from Models W.A to W.D. This could be interpreted as an improvement of the estimates and a reduction in the omitted variable bias as the model R^2 increases. As outlined in Section 2, however, there is no theoretical foundation that this is always true. For some X, such as vocational training and nationality, the change is small and not statistically significant. For other variables in X, such as trainee, the coefficients do not change monotonically (although not significantly) but they gain in precision and become statistically significant. As all Z variables are individually significant in Model W.D, the restriction $E(y|X, W_1, Z) = E(y|X, W_1)$ is violated. It can be seen from Table 3 that all but one component of Z are individually significant in the linear projection on $W_1\hat{\gamma}_{(1)}$. This suggests that there is a statistical partial relationship between the linear combination of Z and the linear combination of W. However, the R^2 of only 0.04 points to that the variation in Z only very little explains the variation in $W_1\hat{\gamma}$ and therefore Z are poor proxies for W_1. This is also confirmed by a rejection of the restriction $E(W_1\gamma_{(1)}|X, Z) = E(W_1\gamma_{(1)}|Z)$ with a P-value of virtually 0. Moreover, the coefficients on Z are mainly unchanged between Models W.B and W.D., which also suggest that the endogeneity of Z is not removed by adding W_1. If anything, these observations suggest that the Z variables are either components of W_2 or they proxy for components in W_2. This would be in line with the increase in the R^2 when we go from Model W.C to W.D.

Table 3: Wage regression: Test restrictions for Z being feasible proxy variables

	$E(W_1\hat{\gamma}_{(1)}\|Z)$ coef. / (SE)
Tenure (in years)	-0.000 (0.001)
Share of working experience over total observation time	0.241*** (0.028)
Additional working experience (in years)	-0.006*** (0.001)
Dummy: unemployment history in the past	-0.175*** (0.019)
N	2435
R^2	0.042

Robust standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.010

In order to shed more light on the role of W_1 and Z in the previous models we estimate panel data regression (6) and (7) with 3 periods for the same individuals as for the other models. We include period interactions for all regressors and only report the coefficients for the period that is used in the cross sectional models. In order to obtain coefficients on the time constant variables, we estimate a dummy variable regression model with 2,435 individual specific dummy variables. The results - without the estimated a - are displayed in Table 2 as Models W.E and W.F, respectively. It is evident that the coefficients on several of the X and Z variables change considerably when using a panel model that allows for correlation between (X, Z) and the time constant part of the error. This points to violations of the stronger exogeneity restrictions in cross sectional analysis. For example, the coefficient on higher education drops sharply from 0.483 in Model W.B to 0.056 in Model W.F. A similar pattern can be observed for several of the business sectors, while other previously strongly significant coefficients become weakly or insignificant in the panel analysis (e.g. gender). The multicollinearity pattern driving this result is briefly discussed at the end of this subsection. But there are also variables, such as trainee, for which precision increases. The coefficients on the Z variables decrease in magnitude and these variables become considerably less individually significant. A robust test whether the components of Z are jointly significant in Model W.F has a p-value of 0.704. This observation and given that the R^2 of Model W.F is not higher than that of Model W.E suggest that Z does not additionally contribute to the model. The relevance of Z in Models W.B and W.D is therefore more likely due to correlation with W_2 rather than because Z directly belongs to the population model.

In the following we shed light on two more questions: First, to what extent do the variables in W_1 explain the variation of the estimated part of the panel model that is supposed to capture the omitted W? Second, to what extent are the estimated fixed effects statistically related to the included X and Z? Any relationship suggests endogeneity of the latter in a cross sectional regression.

Table 4: The statistical relationship between the estimated component of the panel model that controls for omitted W and the observable W_1

	$E(\hat{a}\|W_1)$ coef. / (SE)	$E(\hat{b} + Z\hat{\lambda}\|W_1)$ coef. / (SE)
Big Five: I am rather cautious, reserved	0.036 (0.047)	0.029 (0.057)
Big Five: I tend to criticise people	0.000 (0.041)	-0.013 (0.051)
Big Five: I attend to all my assignments with precision	0.044 (0.066)	0.045 (0.080)
Big Five: I have versatile interests	-0.132** (0.060)	-0.171** (0.073)
Big Five: I am inspirable and can inspire other people	0.027 (0.052)	0.032 (0.064)
Big Five: I easily trust in people and believe in the good in humans	0.070* (0.041)	0.094* (0.050)
Big Five: I tend to be lazy	-0.203*** (0.043)	-0.250*** (0.053)

Continued on next page...

... table 4 continued

	$E(\hat{a}\|W_1)$ coef. / (SE)	$E(\hat{b}+Z\hat{\lambda}\|W_1)$ coef. / (SE)
Big Five: I am profound and like to think about things	-0.115** (0.045)	-0.139** (0.055)
Big Five: I am rather quiet, introverted	-0.292*** (0.046)	-0.375*** (0.056)
Big Five: I can act cold and distant	0.004 (0.040)	0.017 (0.049)
Big Five: I am industrious and work hard	0.200*** (0.076)	0.282*** (0.092)
Big Five: I worry a lot	0.225*** (0.041)	0.291*** (0.051)
Big Five: I have a vivid imagination and have a lot of phantasy	-0.188*** (0.052)	-0.225*** (0.063)
Big Five: I am outgoing and like company	-0.005 (0.054)	0.021 (0.066)
Big Five: I can be gruff and repellend towards other people	-0.117*** (0.043)	-0.140*** (0.053)
Big Five: I make plans and carry them out	-0.025 (0.057)	-0.038 (0.070)
Big Five: I easily get nervous and insecure	0.141*** (0.048)	0.200*** (0.058)
Big Five: I treasure artistic and aesthetic impressions	0.219*** (0.047)	0.248*** (0.057)
Big Five: I am not very interested in art	-0.152*** (0.044)	-0.176*** (0.053)
Dummy: satisfied with one?s life in general	0.389*** (0.142)	0.410** (0.173)
Dummy: was looking for a new job	-0.458*** (0.165)	-0.402** (0.202)
Dummy: was looking for an additional job	-0.708* (0.379)	-0.808* (0.463)
Dummy: was looking for a new and an additional job	0.170 (0.881)	0.402 (1.078)
strength of connection to place of residence	-0.031 (0.046)	-0.033 (0.057)
Frequency of misunderstandings, tensions or conflicts	-0.108** (0.049)	-0.168*** (0.059)
Number of children in total (within and outside the household)	0.212*** (0.056)	0.090 (0.068)
Number of children in household	0.295*** (0.084)	0.415*** (0.103)
Dummy: none of parents has a HE degree	0.056 (0.092)	0.019 (0.112)
Dummy: one parent has a HE degree	0.032 (0.173)	-0.016 (0.212)
Current contract working time,total, without mini-job	-0.042*** (0.008)	-0.055*** (0.010)
Current actual working time, main occupation, without mini-job	-0.074*** (0.014)	-0.095*** (0.017)

Continued on next page...

... table 4 continued

	$E(\hat{a}\|W_1)$ coef. / (SE)	$E(\hat{b}+Z\hat{\lambda}\|W_1)$ coef. / (SE)
Current actual working time,total, without mini-job	0.024* (0.014)	0.029* (0.017)
Dummy: none of parents with migrational background	-0.369** (0.182)	-0.314 (0.223)
Size of household	-0.753*** (0.064)	-0.844*** (0.078)
Constant	9.482*** (0.639)	9.667*** (0.782)
N	2435	2435
R^2	0.327	0.334

Robust standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.010

Table 4 displays the results of the linear projections of W_1 on the estimated components of the panel models that capture the unobserved W as given by (8) and (9) for the cross sectional data. In the case of Model (6) this is simply the estimated fixed effects $\hat{a}$. In the case of Model (7), this is the estimated fixed effect plus the estimated component related to Z, i.e. $\hat{b} + Z\hat{\lambda}$. The estimated coefficients are from the panel regressions. Given that the two regressions in Table 4 have different dependent variables with different variation, the estimated coefficients and the R^2 are not directly comparable. However, they show that the variation in W_1 explains around one third of the variation of the dependent variables. They also show that a number of W_1 variables is partially related with the dependent variables. This is evidence that for the panel models effectively controlling for information in W_1 without directly using it. However, the remaining 2/3 of the variation must be due to W_2. This suggests that the panel models also effectively control for additional unobservables.

Table 5: Wage regression: Regression based endogeneity test for components of X and Z

	$E(\hat{a}\|X)$ coef. / (SE)	$E(\hat{b}\|X,Z)$ coef. / (SE)
Gender (male=1)	-4.867*** (0.026)	-6.249*** (0.024)
Age	0.060*** (0.001)	0.034*** (0.001)
Dummy: trainee	-0.150 (0.375)	-0.033 (0.299)
Missing information on education	-0.582*** (0.158)	-0.483*** (0.155)
No formal degree	-0.188* (0.107)	-0.126 (0.098)
Vocational training	-0.050	-0.082

Continued on next page...

... table 5 continued

	$E(\hat{a}\|X)$ coef. / (SE)	$E(\hat{b}\|X,Z)$ coef. / (SE)
	(0.103)	(0.094)
Higher Education	0.329***	0.304***
	(0.106)	(0.097)
Dummy: German nationality	0.066	-0.037
	(0.059)	(0.055)
Agriculture	-0.473***	-0.285***
	(0.094)	(0.095)
Hotel and restaurant	-0.132	0.056
	(0.080)	(0.075)
Construction	-0.285***	-0.179***
	(0.059)	(0.058)
Trade	-0.147***	-0.076**
	(0.039)	(0.036)
Services	-0.145***	-0.030
	(0.035)	(0.033)
Education and social health	-0.185***	-0.102***
	(0.037)	(0.034)
Public institutions	-0.192***	-0.196***
	(0.045)	(0.044)
Other sectors	0.223***	0.293***
	(0.080)	(0.066)
Tenure (in years)		0.017***
		(0.002)
Share of working experience over total observation time		0.327***
		(0.048)
Additional working experience (in years)		0.004*
		(0.002)
Dummy: unemployment history in the past		-0.479***
		(0.034)
Constant	3.735***	4.172***
	(0.124)	(0.120)
N	2435	2435
R^2	0.954	0.974

Robust standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.010

Table 5 reports the results for regressions of the estimated fixed effects from the panel analysis on the included regressors in the two models using the cross sectional data. It is apparent that a large number of the coefficients differ significantly from 0. This points to partial correlation between fixed effects and regressors and thus to endogeneity of the latter in the cross sectional models of Table 2. This means there is significant bias in many of the estimated coefficients of the cross sectional models W.A and W.B in Table 2. The large values of the R^2 for the two models in Table 5 reveal that the included regressors nearly entirely explain the variation in estimated fixed effects. This causes a strong multicollinearity pattern between some of the variables in the panel models of Table 2, which is reflected by the partly huge standard errors in Models W.E and W.F., e.g. for the coefficient on gender. A solution to mitigate this pattern would be to use information from additional periods but this would then lead to an unbalanced panel.

4.2 Transition Analysis

In this subsection we repeat the analysis by considering a probability model for leaving unemployment. In particular, the dependent variable is binary and takes on the value one if the individual has left unemployment within 12 months since the time of the interview. We apply the linear probability model to estimate the partial relationship between various observables and the transition probability. As this analysis is restricted to unemployed job seekers, the sample conditions on those being in the job seekers register. For this reason, firm level variables are no longer available but other variables such as claiming unemployment benefits. The full set of variables is again provided in Table 10 in Appendix III. Also, the set of work history variable Z changes and becomes related to past unemployment experiences and participation in active labour market policy programs. As mentioned above, none of the survey variables have been selected by the Lasso and elastic net, thus the set W_1 is empty. This suggests that the survey does not contribute relevant information to the analysis.

Table 6: Transition analysis: Binary dependent variable (Dummy: left unemployment within 12 months)

	T.A $E(y\|X)$ coef. / (SE)	T.B $E(y\|X,Z)$ coef. / (SE)	T.C $E(y_{it}\|X_{it})$ coef. / (SE)	T.D $E(y_{it}\|X_{it},Z_{it})$ coef. / (SE)
Gender (male=1)	0.043* (0.022)	0.034 (0.021)	0.111 (0.879)	-0.833 (0.966)
Age	-0.003*** (0.001)	-0.002* (0.001)	0.016 (0.023)	0.033 (0.026)
Missing information on education	-0.007 (0.104)	0.006 (0.099)	0.048 (0.183)	0.013 (0.143)
No formal degree	-0.010 (0.095)	-0.030 (0.090)	-0.003 (0.172)	-0.033 (0.133)
Vocational training	0.025 (0.094)	0.005 (0.089)	0.030 (0.168)	-0.003 (0.127)
Higher Education	0.043 (0.128)	-0.021 (0.127)	0.094 (0.186)	0.053 (0.151)
Dummy: German nationality	0.002 (0.037)	0.003 (0.034)	-0.077 (0.074)	-0.058 (0.068)

Continued on next page...

... table 6 continued

	T.A coef. / (SE)	T.B coef. / (SE)	T.C coef. / (SE)	T.D coef. / (SE)
Dummy: receiving unemployment insurance benefits	0.121*** (0.026)	0.034 (0.025)	0.322*** (0.097)	0.201*** (0.073)
Dummy: receiving mean-tested unemployment benefits	-0.163*** (0.031)	-0.088** (0.036)	0.003 (0.144)	0.092 (0.121)
Dummy: West Germany	-0.006 (0.023)	-0.023 (0.023)	0.027 (0.097)	0.024 (0.095)
Dummy: left unemployment in the past		0.510*** (0.044)		0.537*** (0.063)
Unemployment duration (in months)		-0.003*** (0.000)		-0.020*** (0.002)
Dummy: left long-term unemployment (>12 months) in the past		-0.009 (0.032)		0.293*** (0.043)
Dummy: be recalled in the past		0.022 (0.025)		-0.004 (0.039)
Dummy: participation in active labour market programmes in the past 3 years		0.035 (0.022)		-0.000 (0.026)
Constant	0.949*** (0.119)	0.555*** (0.118)		
N	1484	1484	3×1484	3×1484
R^2	0.036	0.151	0.925	0.936
Percent correctly predicted (PCP)	0.763	0.787		

Robust standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.010

Table 6 shows the estimation results for the models $P(y = 1|X)$, $P(y = 1|X, Z)$, $P(y_{it} = 1|X_{it})$ and $P(y_{it} = 1|X_{it}, Z_{it})$, which are denoted as T.A - T.D, respectively. It is apparent that the estimated coefficients on the X variables are often similar and statistically not different across the regressions T.A and T.B, except for the benefit claim related variables, which both decrease in magnitude. The R^2 increases from 0.036 to 0.151, which shows that the work history variables contribute to the model, though, the Percent Correctly Predicted (PCP) only very marginally increases from 0.763 to 0.787 due to the inclusion of Z. The results for the panel models T.C and T.D in Table 6 show also only evidence for a small number of coefficients (X, Z) to be sizably different in comparison to Models T.A and T.B. The coefficient on receiving unemployment insurance benefits increases considerably, while the coefficient on receiving mean-tested unemployment benefits changes sign but looses statistical significance. Among the Z variables, only the coefficients on unemployment duration and on having left long-term unemployment change. In particular, they increase strongly in magnitude in model T.D. As the latter is the interaction of having been long term unemployed and having left unemployment in the past, these results suggest that past successes play an important role in explaining future successes. While the W_1 variables turned out to be irrelevant, the Z variables appear to be the most important variables in the transition model and they remain relevant after controlling for unobserved fixed effects. Thus, they may be relevant in the population model or may be related to time varying information in W_2 that has not been captured by the panel models.

Table 7: Transition sample: Regression based endogeneity test for components of X and Z

	(1)	(2)
	$E(a\|X)$	$E(b\|X,Z)$
	β / (SE)	β / (SE)
Gender (male=1)	-2.844***	-2.020***
	(0.018)	(0.018)
Age	0.151***	0.110***
	(0.001)	(0.001)
Missing information on education	-0.024	0.024
	(0.073)	(0.077)
No formal degree	-0.005	-0.013
	(0.067)	(0.070)
Vocational training	0.008	0.014
	(0.066)	(0.069)
Higher Education	-0.027	-0.032
	(0.098)	(0.097)
Dummy: German nationality	0.054*	0.051*
	(0.030)	(0.027)
Dummy: receiving unemployment insurance benefits	-0.232***	-0.196***
	(0.021)	(0.020)
Dummy: receiving mean-tested unemployment benefits	-0.117***	-0.117***
	(0.028)	(0.034)
Dummy: West Germany	-0.062***	-0.077***
	(0.019)	(0.019)
Dummy: left unemployment in the past		0.040
		(0.039)
Unemployment duration (in months)		0.017***
		(0.000)
Dummy: left long-term unemployment (>12 months) in the past		-0.274***
		(0.026)
Dummy: be recalled in the past		0.031
		(0.020)
Dummy: participation in active labour market programmes in the past 3 years		0.041**
		(0.017)
Constant	0.953***	0.565***
	(0.089)	(0.096)
N	1484	1484
R^2	0.976	0.968

Robust standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.010

Table 7 presents the results for the linear projection of X and X, Z on the estimated fixed effects. Similar to the wage regressions, there is evidence for a number of variables being endogenous in the cross sectional analysis of Models T.A and T.B. However, as already discussed, our results suggest that the omitted variable bias is limited for most variables in the transition model. Table 7 also shows that the included regressors almost perfectly explain the variation in estimated fixed effects, which suggests again a multicollinearity pattern for a subset of the regressors in Table 6.

Before finishing this section we pay some special focus on the variable "participation in an active labour market policy program" as participation in an active labour market policy program, such as training, is a policy relevant variable that has received a lot of attention in empirical labour market research. We do not find economically, nor statistically relevant changes when comparing Models T.B and T.D. Lechner and Wunsch (2013) and Caliendo et al. (2014), who focus among other things on the estimation of treatment effects on labour market transitions, add more an more operations based administrative or interview based survey variables to their models to check sensitivity of results. Our findings confirm their findings that the estimated treatment effects are stable and thus may not be affected by omitted variables. Despite these findings the endogeneity tests in Table 7 provide some evidence for this variable being endogenous. In order to tackle endogeneity of labour market treatment variables, the academic literature typically applies instrumental variable techniques. Similar to Frölich and Lechner (2010) and Bookmann et al. (2014) we construct an additional variable, the regional treatment intensity as a candidate for an instrument. In Appendix II we outline how the models of Section 2 can be used to test for validity of candidates for instrumental variables. When applying this to the transition sample we find some evidence for the instrument to be correlated with unobserved model components, although the estimated correlations are small (around 0.05 in magnitude). Despite that there is some evidence for instrument invalidity, it is an empirical question whether an IV procedure is more or less biased than an OLS estimate and will depend on the instrument strength. It is remarked that instrument endogeneity can arise through correlation with omitted model components despite that the instrument is not a direct component of the population model and not related to other observable regressors.

5 Summary and Discussion

In virtually any empirical regression analysis there is only limited availability of observed variables and limited prior knowledge which variables belong to the model. This paper provides a unified framework that nests various approaches aiming at reducing omitted variable bias in linear regression analysis. We work out the mechanisms driving the size of the bias and how various models with different regressor sets or unobserved effects relate. Without imposing restrictions on the relationship and role of the variables, it is, however, not possible to derive model rankings that are valid in every application.

In our applications we find evidence for sizable omitted variable bias for a number of variables in the wage regression, while only a small number of coefficients is systematically affected in the transition analysis. While the use of work history and survey variables in the transition analysis hardly changes the results, they seem to contribute to a reduction in omitted variable bias in the wage regression as by including more and more variables the coefficients often converge to their values in the most comprehensive panel data model. In particular, key socio-demographic variables appear to move closer to the results of a panel analysis. When exploiting the availability of panel data, we obtain evidence for cross sectional results being biased due to correlations with unobserved effects. Our results suggest that panel analysis is expected to capture more relevant unobservable model components than an expanded regressor set at one point of time. Beside asymptotic bias considerations, an analysis based on administrative data only should also benefit from a higher precision due to the larger sample size, if for example survey based variables were only available for a small subset of the population.

Our results are not only important for empirical researchers but also for data providers. Due to cost and data confidentiality constraints, data providers aim at supplying a maximum amount of relevant information but a minimum of irrelevant information. Given our findings, the availability of longitudinal information for key variables appears to add more to the analysis than a greatly but possibly unfocused set of additional (survey) variables at one time point.

References

[1] Antoni, M., and Bethmann, A. (2014), PASS-Befragungsdaten verknüpft mit administrativen Daten des IAB (PASS-ADIAB) 1975–2011. *FDZ-Datenreport*, 03/2014, Nürnberg.

[2] Arni, P., Caliendo, M., Künn, S. and Mahlstedt, R. (2014), Predicting the Risk of Long-Term Unemployment: What can we learn from Personality Traits, Beliefs and other Behavioral Variables?, *Working Paper*, Potsdam.

[3] Baptista, R., Lima, F., and Preto, M. T. (2012), How former business owners fare in the labor market? Job assignment and earnings. *European Economic Review*, 56(2), 263-276.

[4] Berg, M., Cramer, R., Dickmann, C., Gilberg, R., Jesske, B., Kleudgen, M., Bethmann, A., Fuchs, B., Trappmann, M., and Wurdack, A. (2013), Codebook and documentation of the panel study 'Labour Market and Social Security' (PASS). Datenreport wave 5. *FDZ-Datenreport*, 06/2012 (en), Nürnberg.

[5] Biewen, M., Fitzenberger, B., Osikominu, R. and Paul, M. (2014), The Effectiveness of Public Sponsored Training Revisited: The Importance of Data and Methodological Choices, *Journal of Labor Economics*, 32(4), 837–897.

[6] Bollinger, C.R. (2003), Measurement Error in Human Capital and the Black-White Gap, *Review of Economics and Statistics*, 85, 578–585.

[7] Bollinger, C.R. and Minier, J. (2015), On the Robustness of Coefficient Estimates to the Inclusion of Proxy Variables, *Journal of Econometric Methods*, 4, 101–122.

[8] Bookmann, B., Thomsen, S.L. and Walter, T. (2014), Intensifying the use of benefit sanctions: an effective tool to increase employment? *IZA Journal of Labor Policy*, 3, 21.

[9] Caliendo, M., Mahlstedt, R. and Mitknik, O.A. (2014), Unobservable, but Unimportant? The Influence of Personality Traits (and Other Usually Unobserved Variables for the Evaluation of Labor Market Policies, *IZA Discussion Paper* No. 8337, IZA Bonn.

[10] Fernández-Kranz, D., and Rodríguez-Planas, N. (2011). The part-time pay penalty in a segmented labor market. *Labour Economics*, 18(5), 591-606.

[11] Friedmann, J, Hastie, T. and Tbishirani, R. (2010), Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, 33.

[12] Frölich, M. and Lechner, M. (2010), Exploiting Regional Treatment Intensity for the Evaluation of Labor Market Policies, *Journal of the American Statistical Association*, 105, 1014–1029.

[13] Gelbach, J. (2016), When Do Covariates Matter? And Which Ones, and How Much? *Journal of Labor Economics*, 34, 509–543.

[14] Heineck, G. and Anger, S. (2010), The returns to cognitive abilities and personality traits in Germany, *Labour Economics*, 17, 535–546.

[15] Kauhanen, A., and Napari, S. (2012), Career and wage dynamics: Evidence from linked employer-employee data. *Research in Labor Economics*, 36, 35-76.

[16] Lechner, M. and Wunsch, C. (2013), Sensitivity of matching-based program evaluations to the availability of control variables, *Labour Economics*, 21, 111–121.

[17] Lubotsky, D. and Wittenberg, M. (2006), Interpretation of Regressions with Multiple Proxies, *Review of Economics and Statistics*, 88, 549–562.

[18] Mueller, G. and Plug, E.J.S. (2006), Estimating the Effect of Personality on Male and Female Earnings, *Estimating the Effect of Personality on Male and Female Earnings*, *Industrial & Labor Relations Review*, 60, Art 1, 1–20.

[19] Neal, D.A. and Johnson, W.R. (1996), The Role of Premarket Factors in Black-White Wage Differences. *Journal of Political Economy*, 104, 869–895.

[20] Nyos, E. and Pons, E. (2005), The effects of personality on earnings, *Journal of Economic Psychology*, 26, 363–384.

[21] Oster, E. (2017), Unobservable Selection and Coefficient Stability ,*Journal of Business and Economic Statistics*, published online.

[22] Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso, *Journal of the*

Royal Statistical Society B, 58, 267–288.

[23] Townsend, W. (2017), "ELASTICREGRESS: Stata module to perform elastic net regression, lasso regression, ridge regression," *Statistical Software Components* S458397, Boston College Department of Economics, revised 01 Feb 2018.

[24] vom Berge, P., König, M., and Seth, S. (2013), Sample of Integrated Labour Market Biographies (SIAB) 1975-2010. *FDZ-Datenreport*, 01/2013 (en), Nürnberg.

[25] Wooldridge, J.M. (2010), *Econometric Analysis of Cross Section and Panel Data*, 2nd ed., MIT Press, Cambridge.

Appendix I: Statistical regularisation and variable selection

The survey data contains a large number of variables (around 40). But it unknown which of them actually belong to the regression model. In an overfitted model with many variables, estimated coefficients may become implausibly large, while not contributing much to the precision of the model fit. Moreover, the reporting of results is more convenient if irrelevant variables are excluded. We apply the Lasso (least absolute shrinkage and selection operator) as a numerical procedure to eliminate variables that do not or very little contribute to the model. Beside their elimination, the model constraints the sum of the parameters on the regressors, making their interpretation easier. The objective is minimising the usual sum of squared residuals subject to a linear inequality constraint

$$(\hat{\beta}, \hat{\gamma}) \text{ arg min} \sum_{i=1}^{N} (y_i - X_i\beta - W_i\gamma) \text{ subject to } \sum_j \gamma_j \leq \lambda$$

with λ being the regularisation parameter (Tibshirani, 1996) and W is a regressor set that contains the eventually chosen W_1 but does not include W_2. The linear inequality constraint leads the Lasso to un-select variables, i.e. $\hat{\gamma}_j = 0$, if their coefficient is small in magnitude. It also leads to the selection of one variable in the case of a group of highly correlated regressors. In order to find $(\hat{\beta}, \hat{\gamma})$ we apply the algorithm suggested by Friedmann et al. (2010). λ is determined by cross-validation such that it minimises the mean squared error. We use the STATA package `elasticregress` (Townsend, 2017). As for the transition model the Lasso does not select any of the variables, we have also applied an Elastic Net which combines the Lasso with a Ridge regression. The Elastic Net did not select additional variables for the transition model.

Appendix II: Using additional information to test the validity of an instrument in cross sectional analysis

Instrumental variable analysis is popular in applied economic and social sciences research but in order for a candidate for an instrumental variable to be valid it has to pass certain restrictions. In particular, let m be for simplicity one instrumental variable for a component of X, say x_j that is assumed to be endogenous, i.e. $cov(x_j, u) \neq 0$, where u is the error of a regression model , e.g. Model (2). For m being a valid instrument, we need that it is partially correlated with x_j, i.e. playing a role in the reduced form for x_j. This is something one can easily test for. The second requirement is $cov(m, u) = cov(m, W\gamma + v) = cov(m, a + q) = 0$ or $cov(m, b + s + v) = 0$ with a and q as in Model (6) and b, s and v as in Model (7). These conditions cannot easily be tested for in applications, because of unavailability of u, W, a and b. By exploiting the availability of W_1 and the panel dimension of the data as outlined in Section 2, we suggest the following tests for instrument validity. Once $W_1\gamma_1$, a and b are (consistently) estimated, one can use them to empirically check if m is partially correlated with $W_1\hat{\gamma}_1$, $\hat{a}$ or $\hat{b}$. Any non zero covariance would point to endogeneity of m. Even including Z and or W_1 in the model will not render m exogenous if time constant unobservables are correlated with fixed effects.

We apply this to the data of the transition analysis. As candidate for an instrumental variable we use the regional treatment intensity. This is the share of unemployed job seekers that has been assigned into a treatment measure by an employment agency district. Similar variables have been used to address possible endogeneity of the individual level variable "participation in an active labour market policy program in the past 3 years" (or subsequently denoted as z_j) that is of high policy relevance. Table 8 reports the resulting covariances and pairwise correlations between estimated fixed effects $\hat{b}$, and the candidate instrument m and the regressor z_j, respectively. We do not consider $W_1\hat{\gamma}_{(1)}$ as there is no W_1 in this model and we do not consider $\hat{a}$ as the endogeneous regressor is an element of Z. It is apparent that the relationships are rather weak but not strictly zero. The correlation of m with $\hat{b}$ is -0.05 with a p-value of 0.04. There is therefore some evidence for the instrument not being valid. It is an empirical question whether a 2SLS estimator has a smaller or larger bias than an OLS estimator. But given that $\rho_{\hat{b},z_j}$ is only marginally larger in magnitude than $\rho_{\hat{b},m}$, the instrument needs to be strong to make the IV estimator less biased and sufficiently precise in the application.

Table 8: Transition Analysis: Testing for endogeneity

	$Cov(\hat{b}, .)$	$\rho_{\hat{b},.}$
Regional treatment intensity	-0.0106	-0.0522**
Participation in active labour market programmes in the past 3 years	0.064038	0.0738***

ρ: pairwise correlation
* p<0.10, ** p<0.05, *** p<0.010

Appendix III: Tables

Table 9: Variable selection by LASSO and elastic net

Variable Names
Wage sample:
Big Five: I am rather cautious, reserved
Big Five: I tend to criticise people
Big Five: I attend to all my assignments with precision
Big Five: I have versatile interests
Big Five: I am inspirable and can inspire other people
Big Five: I easily trust in people and believe in the good in humans
Big Five: I tend to be lazy
Big Five: I am profound and like to think about things
Big Five: I am rather quiet, introverted
Big Five: I can act cold and distant
Big Five: I am industrious and work hard
Big Five: I worry a lot
Big Five: I have a vivid imagination and have a lot of phantasy
Big Five: I am outgoing and like company
Big Five: I can be gruff and repellend towards other people
Big Five: I make plans and carry them out
Big Five: I easily get nervous and insecure
Big Five: I treasure artistic and aesthetic impressions
Big Five: I am not very interested in art
Dummy: satisfied with one?s life in general
Dummy: was looking for a new job
Dummy: was looking for an additional job
Dummy: was looking for a new and an additional job
strength of connection to place of residence
Frequency of misunderstandings, tensions or conflicts
Number of children in total (within and outside the household)
Number of children in household
Dummy: none of parents has a HE degree
Dummy: one parent has a HE degree

Continued on next page...

… table 9 continued

Variable Names
Current contract working time,total, without mini-job
Current actual working time, main occupation, without mini-job
Current actual working time,total, without mini-job
Dummy: none of parents with migrational background
Size of household
Not Selected by LASSO:
Big Five: I tend to be depressed, crestfallen
Big Five: I am relaxed and don?t let stress get to me
Dummy: satisfied with health
Working even without being dependent on wage
Number of real close friends/family members outside the household
Transition sample:
No variable is selected by LASSO and elastic net

Table 10: Descriptive statistics

Variable Names	Wage				Transition			
	mean	sd	min	max	mean	sd	min	max
y variables:								
log(average daily gross wage)	4.15	0.66	1.21	5.44	-	-	-	-
Dummy: left unemployment within 12 months	-	-	-	-	0.76	0.42	0.00	1.00
X variables:								
Gender (male=1)	0.47	0.50	0.00	1.00	0.40	0.49	0.00	1.00
Age	42.87	10.19	18.00	64.00	39.74	11.30	17.00	63.00
Dummy: trainee	0.00	0.04	0.00	1.00	-	-	-	-
Missing information on education	0.00	0.07	0.00	1.00	0.06	0.25	0.00	1.00
No formal degree	0.12	0.33	0.00	1.00	0.31	0.46	0.00	1.00
Vocational training	0.75	0.43	0.00	1.00	0.60	0.49	0.00	1.00
Higher Education	0.11	0.31	0.00	1.00	0.01	0.12	0.00	1.00
Dummy: German nationality	0.96	0.19	0.00	1.00	0.89	0.32	0.00	1.00
Dummy: West Germany	-	-	-	-	0.63	0.48	0.00	1.00
Agriculture	0.01	0.09	0.00	1.00	-	-	-	-
Hotel and restaurant	0.03	0.18	0.00	1.00	-	-	-	-
Construction	0.05	0.22	0.00	1.00	-	-	-	-
Trade	0.14	0.35	0.00	1.00	-	-	-	-
Services	0.29	0.45	0.00	1.00	-	-	-	-
Education and social health	0.20	0.40	0.00	1.00	-	-	-	-
Public institutions	0.07	0.25	0.00	1.00	-	-	-	-
Other sectors	0.02	0.14	0.00	1.00	-	-	-	-
Dummy: receiving unemployment insurance benefits	-	-	-	-	0.67	0.47	0.00	1.00
Dummy: receiving mean-tested unemployment assistance	-	-	-	-	0.37	0.48	0.00	1.00
Z variables:								
Tenure (in years)	4.85	5.99	0.00	36.59	-	-	-	-
Share of working experience over total observation time	0.53	0.34	0.00	1.00	-	-	-	-
Additional working experience (in years)	8.31	8.03	0.00	36.14	-	-	-	-
Dummy: unemployment history in the past	0.74	0.44	0.00	1.00	-	-	-	-
Dummy: left unemployment in the past	-	-	-	-	0.89	0.31	0.00	1.00
Unemployment duration (in months)	-	-	-	-	40.19	30.20	0.00	133.13
Dummy: left long-term unemployment(>12 months) in the past	-	-	-	-	0.22	0.41	0.00	1.00
Dummy: be recalled in the past	-	-	-	-	0.73	0.44	0.00	1.00
Dummy: participation in an active labour market policy program in the past 3 years	-	-	-	-	0.39	0.49	0.00	1.00
$W1$ variables:								
Big Five: I am rather cautious, reserved	2.82	1.16	1.00	5.00	-	-	-	-
Big Five: I tend to criticise people	2.73	1.12	1.00	5.00	-	-	-	-
Big Five: I attend to all my assignments with precision	4.41	0.71	1.00	5.00	-	-	-	-
Big Five: I have versatile interests	4.22	0.83	1.00	5.00	-	-	-	-
Big Five: I am inspirable and can inspire other people	3.80	1.02	1.00	5.00	-	-	-	-

Continued on next page...

... table 10 continued

Variable Names	Wage				Transition			
	mean	sd	min	max	mean	sd	min	max
Big Five: I easily trust in people and believe in the good in humans	3.57	1.12	1.00	5.00	-	-	-	-
Big Five: I tend to be lazy	2.21	1.10	1.00	5.00	-	-	-	-
Big Five: I am profound and like to think about things	3.70	1.05	1.00	5.00	-	-	-	-
Big Five: I am rather quiet, introverted	2.52	1.21	1.00	5.00	-	-	-	-
Big Five: I can act cold and distant	3.10	1.24	1.00	5.00	-	-	-	-
Big Five: I am industrious and work hard	4.31	0.65	1.00	5.00	-	-	-	-
Big Five: I worry a lot	3.23	1.18	1.00	5.00	-	-	-	-
Big Five: I have a vivid imagination and have a lot of phantasy	3.81	0.96	1.00	5.00	-	-	-	-
Big Five: I am outgoing and like company	3.71	1.01	1.00	5.00	-	-	-	-
Big Five: I can be gruff and repellend towards other people	3.01	1.18	1.00	5.00	-	-	-	-
Big Five: I make plans and carry them out	3.98	0.85	1.00	5.00	-	-	-	-
Big Five: I easily get nervous and insecure	2.43	1.03	1.00	5.00	-	-	-	-
Big Five: I treasure artistic and aesthetic impressions	3.34	1.19	1.00	5.00	-	-	-	-
Big Five: I am not very interested in art	2.76	1.26	1.00	5.00	-	-	-	-
Dummy: satisfied with one's life in general	0.88	0.32	0.00	1.00	-	-	-	-
Dummy: was looking for a new job	0.08	0.27	0.00	1.00	-	-	-	-
Dummy: was looking for an additional job	0.01	0.12	0.00	1.00	-	-	-	-
Dummy: was not looking for a new job	0.91	0.29	0.00	1.00	-	-	-	-
Dummy: was looking for a new and an additional job	0.00	0.05	0.00	1.00	-	-	-	-
strength of connection to place of residence	1.98	0.96	1.00	5.00	-	-	-	-
Frequency of misunderstandings, tensions or conflicts	3.55	0.95	1.00	5.00	-	-	-	-
Number of children in total (within and outside the household)	1.53	1.13	0.00	7.00	-	-	-	-
Number of children in household	1.07	1.02	0.00	7.00	-	-	-	-
Dummy: none of parents has a HE degree	0.53	0.50	0.00	1.00	-	-	-	-
Dummy: one parent has a HE degree	0.08	0.27	0.00	1.00	-	-	-	-
Current contract working time,total, without mini-job	33.83	9.17	0.00	80.00	-	-	-	-
Current actual working time, main occupation, without mini-job	37.49	11.37	0.00	80.00	-	-	-	-
Current actual working time,total, without mini-job	37.85	11.87	0.00	120.00	-	-	-	-
Dummy: none of parents with migrational background	0.06	0.24	0.00	1.00	-	-	-	-
Size of household	3.11	1.09	2.00	10.00	-	-	-	-

Table 11: Appendix: Estimated coefficients on W_1 variables in Models W.C and W.D (continued from Table 2)

	$E(y\|X,W1)$ coef. / (SE)	$E(y\|X,Z,W1)$ coef. / (SE)
Big Five: I am rather cautious, reserved	-0.017 (0.011)	-0.017* (0.010)
Big Five: I tend to criticise people	0.025** (0.010)	0.026*** (0.009)
Big Five: I attend to all my assignments with precision	-0.019 (0.014)	-0.016 (0.013)
Big Five: I have versatile interests	0.001 (0.013)	0.006 (0.012)
Big Five: I am inspirable and can inspire other people	-0.009 (0.011)	-0.009 (0.011)
Big Five: I easily trust in people and believe in the good in humans	0.002 (0.009)	0.007 (0.008)
Big Five: I tend to be lazy	0.036*** (0.010)	0.036*** (0.009)
Big Five: I am profound and like to think about things	0.018* (0.010)	0.018* (0.009)
Big Five: I am rather quiet, introverted	-0.009 (0.010)	-0.010 (0.009)
Big Five: I can act cold and distant	-0.005 (0.008)	-0.005 (0.008)
Big Five: I am industrious and work hard	-0.008 (0.018)	-0.002 (0.017)
Big Five: I worry a lot	-0.023** (0.009)	-0.018** (0.009)
Big Five: I have a vivid imagination and have a lot of phantasy	0.004 (0.011)	-0.000 (0.010)
Big Five: I am outgoing and like company	-0.028** (0.011)	-0.029*** (0.010)
Big Five: I can be gruff and repellend towards other people	-0.007 (0.010)	-0.000 (0.009)
Big Five: I make plans and carry them out	0.059*** (0.013)	0.050*** (0.012)
Big Five: I easily get nervous and insecure	-0.023** (0.011)	-0.019* (0.011)
Big Five: I treasure artistic and aesthetic impressions	0.005 (0.010)	0.009 (0.009)
Big Five: I am not very interested in art	0.000 (0.010)	-0.004 (0.009)
Dummy: satisfied with one?s life in general	0.164*** (0.031)	0.122*** (0.029)
Dummy: was looking for a new job	-0.180*** (0.035)	-0.128*** (0.033)
Dummy: was looking for an additional job	-0.028 (0.087)	-0.001 (0.079)
Dummy: was looking for a new and an additional job	-0.093 (0.171)	-0.030 (0.179)

Continued on next page...

… table 11 continued

	$E(y\|X, W1)$ coef. / (SE)	$E(y\|X, Z, W1)$ coef. / (SE)
strength of connection to place of residence	0.017*	0.024***
	(0.010)	(0.009)
Frequency of misunderstandings, tensions or conflicts	-0.024**	-0.023**
	(0.011)	(0.011)
Number of children in total (within and outside the household)	-0.060***	-0.027**
	(0.014)	(0.013)
Number of children in household	0.050***	0.049***
	(0.017)	(0.017)
Dummy: none of parents has a HE degree	0.033	0.013
	(0.020)	(0.019)
Dummy: one parent has a HE degree	0.105***	0.100***
	(0.036)	(0.034)
Current contract working time,total, without mini-job	0.021***	0.020***
	(0.002)	(0.002)
Current actual working time, main occupation, without mini-job	0.022***	0.020***
	(0.003)	(0.003)
Current actual working time,total, without mini-job	-0.011***	-0.010***
	(0.004)	(0.003)
Dummy: none of parents with migrational background	0.089**	0.108***
	(0.040)	(0.039)
Size of household	0.011	-0.015
	(0.013)	(0.013)
N	2435	2435
R^2	0.502	0.570

* p<0.10, ** p<0.05, *** p<0.010

Imprint

FDZ-Methodenreport 6/2018

Publisher
The Research Data Centre (FDZ)
of the Federal Employment Agency
in the Institute for Employment Research
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Dana Müller, Dagmar Theune

Technical production
Dagmar Theune

Download
http://doku.iab.de/fdz/reporte/2018/MR_06-18_EN.pdf

Internet
http://fdz.iab.de/

Corresponding author:
Pia Homrighausen
Institute for Employment Research (IAB)
Regensburger Str. 104
D-90478 Nürnberg
Email: pia.homrighausen@iab.de