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

01/2015
EN

Using longitudinal wage information in linked data sets

The example of ALWA-ADIAB

Malte Reichelt

Bundesagentur für Arbeit

Using longitudinal wage information in linked data sets

The example of ALWA-ADIAB

Malte Reichelt (IAB)

Die FDZ-Methodenreporte befassen sich mit den methodischen Aspekten der Daten des FDZ und helfen somit Nutzerinnen und Nutzern bei der Analyse der Daten. Nutzerinnen und Nutzer können hierzu in dieser Reihe zitationsfähig publizieren und stellen sich der öffentlichen Diskussion.

FDZ-Methodenreporte (FDZ method reports) deal with methodical aspects of FDZ data and help users in the analysis of these data. In addition, users can publish their results in a citable manner and present them for public discussion.

Contents

Abstract

Using longitudinal wage information in linked datasets such as ALWA-ADIAB is not straightforward. First, due to the upper earnings limit for social security contributions, parts of the wages have to be imputed. Second, although information from the ALWA survey is linked to administrative data from the Federal Employment Agency (BA) on the personal/individual level, matching single employment spells may not always be possible. I propose a polynomial time function to smoothen wage information over time and close gaps between spells in the administrative data. The information can then be matched to the respondents in ALWA and adjusted to meet the episode structure given in the survey. Moreover, survey information on contract hours allows for the calculation of hourly wages. The result is the supplement of ALWA with longitudinal wage information for all main employment episodes in West Germany, starting after January 1975, and for episodes in East Germany starting after January 1993. Although the estimation of wages obtained through the proposed method is connected to some assumptions and thus to a certain degree of uncertainty it produces more complete wage information than is provided through simple matching on a monthly basis. Moreover, a comparison with the net income that respondents reported for their interview date shows that predicted daily and hourly wages are highly reliable.

Zusammenfassung

Lohnangaben im Längsschnitt in gelinkten Datensätzen wie dem ALWA-ADIAB sind meist nicht ohne weiteres nutzbar. Erstens sind Löhne aufgrund der Beitragsbemessungsgrenze zensiert und müssen imputiert werden. Zweitens wurden Survey- und administrative Daten der Bundesagentur für Arbeit (BA) zwar auf Personenebene verknüpft; die jeweiligen Beschäftigungsepisoden einander zuzuordnen ist aber nicht immer möglich. Ich schlage vor, mithilfe einer Polynom-Funktion Lohnangaben über die Zeit zu glätten und Lücken in den administrativen Daten zu schließen. Die so erhaltenen Informationen können dann zu den Befragungsdaten in ALWA hinzugespielt und auf die Episodenangaben angepasst werden. Darüber hinaus können Survey-Informationen zu vertraglich vereinbarten Arbeitszeiten verwendet werden, um Stundenlöhne zu berechnen. Das Ergebnis ist die Erweiterung von ALWA mit Lohnangaben im Längsschnitt für die Haupterwerbsepisoden in Westdeutschland seit Januar 1975 und in Ostdeutschland seit Januar 1993. Obwohl die so geschätzten Angaben bestimmten Annahmen und so einer gewissen Unsicherheit unterliegen, generiert die Methode vollständigere Lohninformationen als durch simples monatliches Matching erreichbar wäre. Außerdem zeigt ein Vergleich der so gewonnenen Tagesentgelte und Stundenlöhne mit den erfragten Lohnangaben zum Interviewzeitpunkt, dass die Informationen zuverlässig geschätzt werden.

Keywords: ALWA-ADIAB, wage imputation, hourly wages, episode matching

I would like to thank Britta Matthes, Corinna Kleinert, Florian Lehmer and Manfred Antoni for helpful comments and Johannes Ludsteck for advice concerning the wage imputation.

1 Introduction

The "ALWA survey data linked to administrative data of the IAB" (ALWA-ADIAB) provides a dataset that links the retrospective survey "Working and Learning in a Changing World" (ALWA) (Antoni et al., 2010) to administrative data on the person and firm level (Antoni, Jacobebbinghaus, & Seth, 2011; Antoni & Seth, 2012). The survey was conducted in 2007/2008 and includes 10,177 computer assisted telephone interviews with German speaking respondents. The data encompass retrospective residential, educational, employment and partnership histories in Germany and provides complete life-course data on a monthly basis (Kleinert et al., 2011).

The data thus offers a rich source for longitudinal analyses in the context of employment histories. However, income information—which often is of central importance in such analyses—is only available for those employment relationships that were given at the time of the interview. Consequently, the only option to use income information is in cross-sectional analyses. Linking the ALWA survey to administrative data introduces the possibility of using longitudinal wage information for most respondents. ALWA-ADIAB[1] is a unique dataset that connects retrospective survey information to precise daily employment and unemployment data, provided through employers' mandatory reports on social security contributions.

Using the wage information from administrative data, however, imposes several problems that need to be addressed. *First*, wages are subject to the upper earnings limit for social security contributions and thus partly have to be imputed. *Second*, ALWA and the administrative data contain employment spells, which are not congruent and cannot be readily matched. *Third*, the administrative data provides daily wages, calculated from the yearly remuneration or the total of an employment episode. The wages are provided without corresponding information on actual working days or contract hours. *Fourth*, the administrative data are left censored and provide wages for West Germany, starting from 1975 and for East Germany, starting from 1993.

I propose a method that tackles the first three problems and allows for matching wage information and calculating hourly wages for main employment episodes in ALWA. Some uncertainty in the data remains, but I am able to show that the wage information is reliable at least for the most recent employment spells, where information is provided by both sources: survey and administrative data.

2 Challenges when matching employment spells

The linked ALWA-ADIAB data is based on the survey population of ALWA. Those respondents that agreed could be matched to their administrative data using the names, sex, birth dates and addresses. Linkage was achieved using error-tolerant record linkage techniques

[1] Access to the dataset is provided via the Research Data Center (FDZ) of German Federal Employment Agency at the IAB and is given through on-site-use and subsequent remote data access. See http://fdz.iab.de/en.aspx for more information.

(Antoni & Seth, 2012), which allowed for matching the majority of respondents. While 91.61 percent have agreed to link their data, 86.49 percent of these respondents could be connected to the administrative data. Thus, about 80 percent of all respondents were successfully matched. More detailed information on the matching process can be found in Antoni et al. (2011).

For these individuals, the administrative part of ALWA-ADIAB provides rich longitudinal data. However, the wage data cannot be used immediately and without further data manipulation. The main problem is that even though survey and administrative data of the respondents may be matched successfully, their employment episodes are not. For illustration, Figure 1 shows an exemplary person with employment spells in administrative data and the ALWA survey. As the administrative data provides spell data and contains information on a daily basis, I aggregated the information to monthly data. The administrative data that is used for wage information mostly consists of yearly reports from employers as well as starting and ending notices of employment relations. In order to obtain complete employment episodes, I combined directly connected spells at the same establishment.

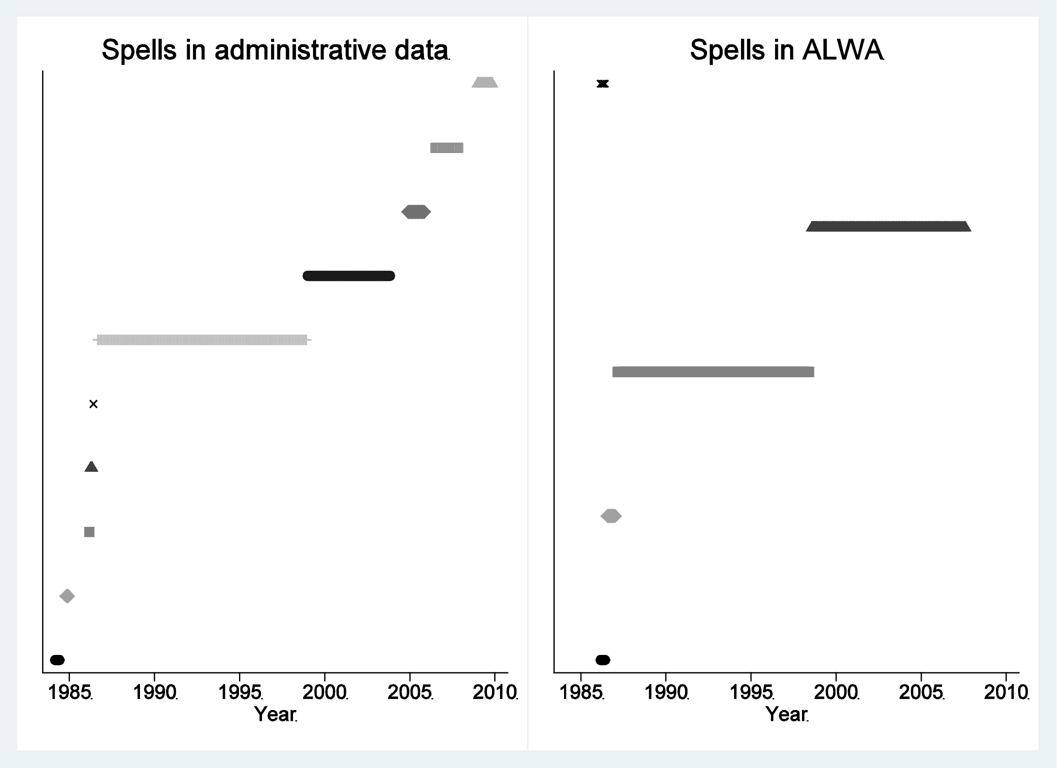

Figure 1 Exemplary spell structure in administrative data and ALWA

The left hand graph shows employment episodes, created from the administrative data, whereas the right hand graph shows the employment episodes that are reported in ALWA. When comparing both graphs and trying to match the employment episodes, several challenges become obvious. First, both data sources provide a different number of spells. There may be multiple reasons, as contracts within a firm that belongs to one corporation are counted as multiple episodes in the administrative data but may be counted as single episodes in ALWA. Moreover, subsequent contracts with short interruptions may be defined differently in both sources. Whereas respondents may report one continuous episode, the administrative data would indicate two episodes. One further reason is provided by the type of employment. Whereas episodes of marginally employment before 1999, self-employment, informal labor and employment relations of civil servants or freelancers are reported in ALWA, they cannot be found in the administrative data.

A second problem is provided by parallel spells. ALWA differentiates between main and second employment spells, but even for main employment spells, a number of respondents reported parallel and overlapping spells in the survey. Moreover, multiple part-time contracts may result in parallel employment episodes in the survey as well as parallel spells in the administrative data, which makes it difficult to assign the employment relations to each other.

The greatest challenge, however, results from different starting and ending dates of the episodes. As Figure 1 illustrates, these dates are seldom identical, resulting in differing lengths and different positions of the episodes. ALWA uses a combination of modularized self-reports and event history calendars (EHC), which has been shown to improve completeness and dating accuracy (Drasch & Matthes, 2013). However, episode dates that lie further in the past may be inaccurate to some degree. Although they are consistent in the timeline reported in the survey, they must not be congruent with the administrative reports. Moreover, the quality of the administrative data is dependent on the quality of the employers' reporting behavior. Mistakes or delays may as well lead to differences in both episode structures.

I propose a method that circumvents these problems and reduces uncertainty in matching wage information from administrative data to employment spells in ALWA. The following sections explain multiple single steps that tackle one of the above described problems, while the online appendix provides the do-file for implementing or adjusting the method in Stata.

3 Identifying and matching employment spells

When matching longitudinal data, parallel spells may pose a challenge. In the case of multiple employment spells, identifying those episodes in the administrative data and the survey data that belong together may not be possible. Moreover, the probability that starting and ending dates of parallel secondary employment episodes are remembered correctly should be lower than for main employment spells, as these mostly represent short episodes that may be less important in terms of wage and career options. Therefore, I focus on main employment spells and try to identify the respective episodes in both data sources.

3.1 Issues when selecting and matching main employment spells

In ALWA, respondents had to differentiate between main and secondary employment relationships. However, even these spells may be overlapping or parallel. Moreover, Figure 1 shows that even for main employment episodes, it is difficult to match episodes. For example, the last ALWA-employment episode on the right hand side could represent multiple spell combinations in the administrative data.

A possible solution to this issue is allowing episodes in the administrative data to vary in position, starting and ending date. Applying such a procedure raises the number of episodes that can be linked, but still leaves the majority of spells in ALWA-ADIAB unmatched. Therefore, I propose manipulating main employment spells to obtain non-overlapping spells, matching wage information on a monthly basis and adjusting the information to fit the newly generated episodes.

3.2 Manipulating the spell structure

In the administrative data, I identify main employment episodes using the spells with the highest daily wage. Should spells be parallel, it is most likely that secondary employment spells offer lower wages. Moreover, I only select episodes subject to social security contributions without specific characteristics. I thus delete wage information for certain groups, such as marginally employed, employees performing community service and apprenticeships. Turning to the survey data, I apply the following rules to create employment episodes that are neither overlapping nor completely parallel:

- I select employment episodes that by definition can be found in the administrative data, thus excluding marginally employed, employees performing community service and self-employed.

- In the case of overlapping episodes, the starting date of the second episode is recoded to fit the ending date of the previous one. There may be several reasons for overlapping episodes, for instance the start of a new employment contract before the formal termination of the previous one or imprecisely reported ending dates. In such a case, a contract ending in December could denote the beginning just as well as the ending of the month. In any case, it is reasonable to assume that the starting date of a new employment relation denotes the beginning of a new main employment spell, which is accompanied by higher wages.

- If an employment episode is completely enclosed by another, it is ignored. Again, I suspect the longer episode to be the main employment episode, which is accompanied by higher wages.

- If two episodes are completely parallel, the one with more working hours is chosen.

Figure 2 illustratively shows how employment spells are manipulated when applying the rules described above. All four lines represent employment episodes in ALWA. In order to identify starting and ending dates of these episodes in all three manipulation states, all lines are provided with a different texture.

If episode structure A were given, applying the manipulation rules results in structure B with one main employment spell and shortened overlapping episodes as well as one completely enclosed episode. Further simplifying the information results in structure C, where parallel spells are subsumed into a dummy variable. This reduction allows for controlling effects of parallel spells in longitudinal analyses.

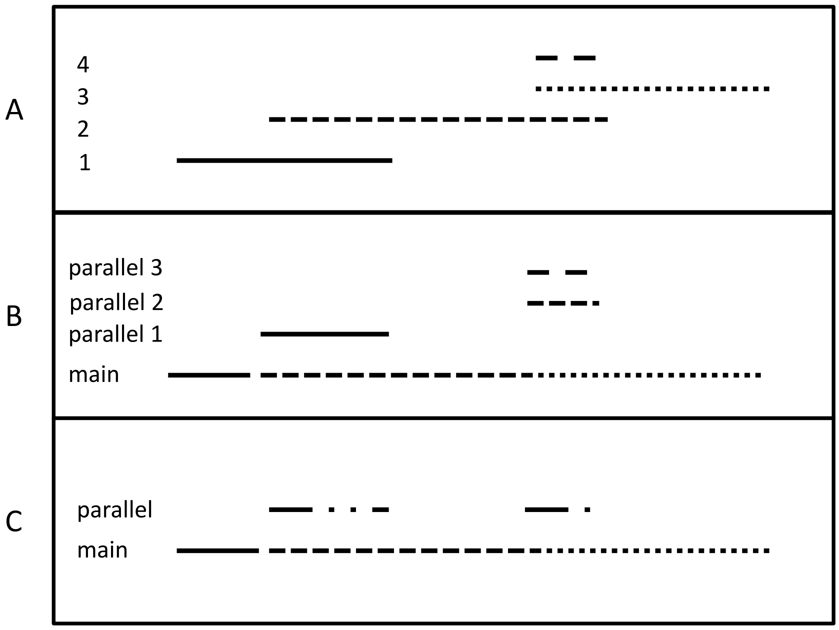

Figure 2 Exemplary manipulation of employment spells in ALWA

Having created single employment episodes in both the administrative and the survey data, the information is brought into the same format in order to be matched. Starting or ending dates before or on the 15th of a month are recoded to the beginning of the month and dates after the 15th are recoded to the beginning of the next month. Technically, income information could now be matched; however, episodes may still have different lengths and different positions in the timeline.

4 Imputing and matching wages in ALWA-ADIAB

In order to solve the problem of differing lengths and positions of main employment episodes, I propose regressing the daily wage on time polynomials for every respondent. Such a procedure results in individual wage trajectories instead of single wage measurements. As a consequence, matching employment episodes from both sources becomes redundant. The predicted wages can easily be cut to fit the employment spells given in the ALWA survey. To calculate such a function accurately, however, censoring of the wage information has to be accounted for.

Moreover, deflating wages may be sensible for some analyses. Therefore, after imputing and matching wages, I deflate the wages according to the consumer price index, which results in gross daily wages adjusted for price changes over the years. Both daily and hourly wages will be provided in their original as well as in their deflated form.

4.1 Imputing censored wages in the administrative data

The administrative data provides wage information that is subject to the upper earnings limit for social security contributions. Ignoring the censoring may result in biased results. Therefore, it is sensible to calculate a truncated regression, predict the right censored values and add an error term (Dustmann, Ludsteck, & Schönberg, 2009; Gartner, 2005).

First, I delete wages below the limit for marginally employed, because such wage values should not be observed for standard employment relations that are subject to social security contributions. Moreover, I set wages above the right censoring limit to missing as these wages rather represent bonus payments instead of regular wages. Both limits are reported in Drews (2007)[2]. I calculate separate interval regressions for each year in the data and calculate the expected value of the logarithm of the daily wage. The interval regression is a generalization of the tobit regression, which can account for any kind of censoring or truncation. The general form is a linear model

$$\mathrm{lnw} = \mathrm{x}'\hat{\upbeta} + \epsilon,$$

where the likelihood contribution is $\Pr(Y_i = y_i)$ for all point data from individual i and is $\Pr(Y_i \geq y_i)$ for all right censored data. y_i is the observed (censored) value and Y_i represents the random dependent variable in the model. Here, the prediction is based on schooling, age, sex, 3-digit occupations, job position including information on part- and full-time[3], size of the firm and an indicator for East Germany. Of course, further variables can easily be implemented.

In order to prevent the predicted log wages $E(lnw_i) = x'_i\hat{\beta}$ from correlating overly high with the covariates, it is useful to add an error term. I draw a random value lnw^{imp} from a normal distribution $N(x'\hat{\beta}, \sigma^2)$, as log wages usually follow such a distribution. The value can then be added as an error term η to the predicted value, resulting in the prediction:

$$E(lnw_i) = x'_i\hat{\beta} + \eta_i.$$

As the true value of lnw_i must be larger than the censoring limit, one can draw the random variable from a truncated distribution. Gartner (2005) shows that the randomly drawn error term can be written as

$$\eta = \Phi^{-1}\left(Y\left(1 - \Phi(\alpha)\right) + \Phi(\alpha)\right).$$

The censored wages can then be replaced by predicted values and the error term. Figure 3 shows the distribution of the logarithm of daily wage before and after the imputation. The imputation results in an approximately normal distribution for log daily wages.

[2] Censoring limits can as well be found online at http://fdz.iab.de/en/FDZ_Individual_Data/IAB_Employment_Samples/IAB_Employment_Samples_Working_Tools.aspx

[3] I use both, full- and part-time employees for the wage imputation. As most wage measurements above the contribution limit can be ascribed to full-time employees, potential bias should be low.

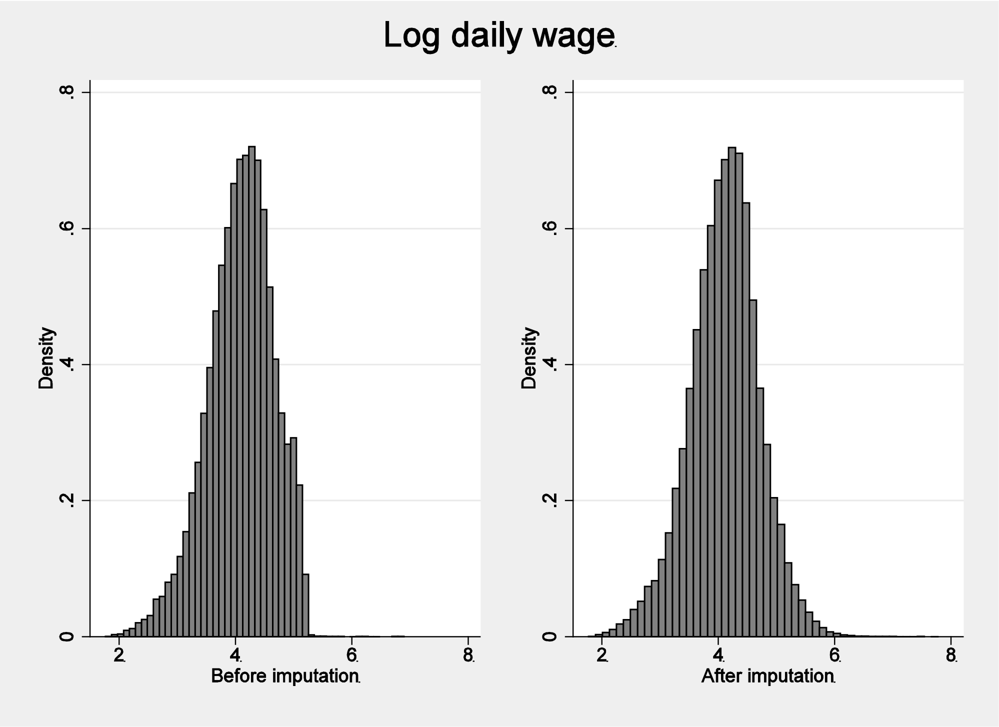

Figure 3 Logarithm of daily wage and imputed logarithm of daily wage

4.2 Calculating a wage function

Having imputed the wages, the next step involves matching the wage data to their respective episodes in ALWA. Until now, the wage data matches the episodes in the administrative data, which may not be identical to the episodes in the survey. Therefore, I propose regressing the daily wage on time polynomials to predict the wages for every individual i:

$$\ln(wage)_i = \beta_{0i} + \sum_{k=1}^{6} \beta_{ki} t^k \; + \epsilon,$$

where t represents the time, measured in months since 1960 and β_{ki} represents the predicted parameter for the k^{th} degree of time for individual i.

This procedure results in a wage trajectory for every individual with employment episodes in both data sources. Thus, gaps between employment episodes are closed and outliers, for instance resulting from bonus payments, are smoothed. As a consequence, the wage information does not encompass the exact daily wage for every spell anymore but rather depicts a wage trajectory.

Figure 4 shows the logarithm of the daily wage and the predicted logarithm of the daily wage from the polynomial function as well as the ALWA employment spells for the same example respondent used above. It becomes obvious that for the employment spells in ALWA, we now obtain a smoothed wage measure.

However, smoothing the wage trajectory results in some limitations when using the predicted measure. First, predictions outside the time range of the administrative data are hardly possible. Calculating such a function thus imposes uncertainty at both ends of the employment history. If wage measures in the administrative data end before the last employment spell in ALWA ends, the functional form may result in the prediction of implausibly high or low wage values. Put differently, if ALWA employment episodes were positioned before the first or after the last wage measurement in the administrative data, the prediction would not be useful at all. Another limitation of this kind of wage-smoothing affects the type of statistical analyses that can be conducted. Regarding the predicted log daily wage, it should become obvious that direct jumps in wage cannot be identified anymore. Comparing immediate returns to occupational mobility for example may thus not be possible.

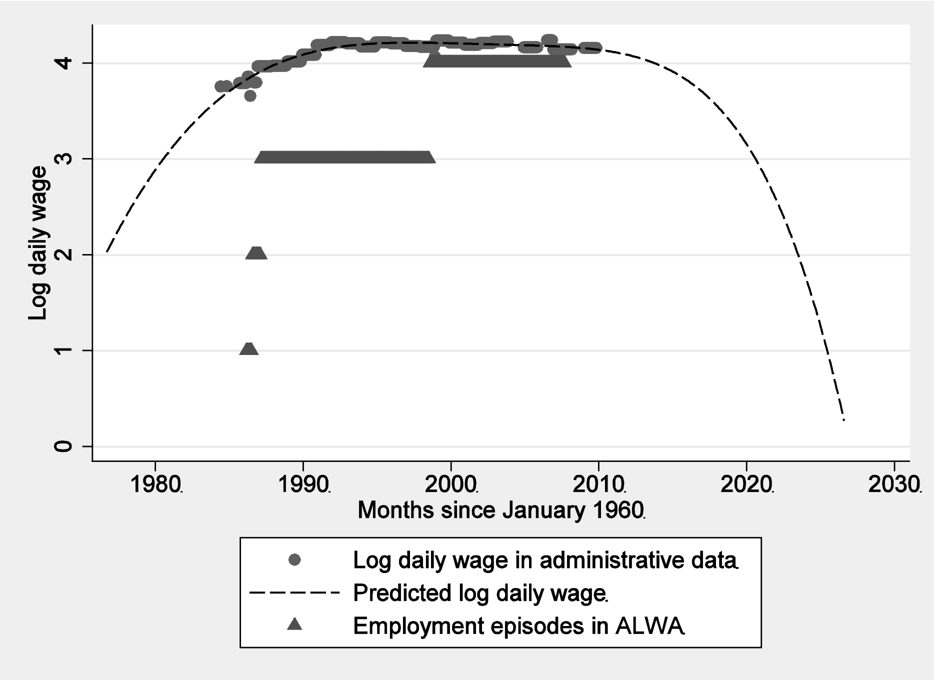

Figure 4 Predicted log daily wage

As a consequence, the first measure in the first episode in ALWA is written back to six months before the beginning of the episode and the last measure of the last episode is extrapolated to six months after the end. If the first or last episodes in ALWA cannot be ascribed to any wage measure in the administrative data, the episode is ignored.

The newly generated wage trajectory can then be cut to fit the employment spells—eliminating gaps and different episode lengths. For all employment episodes that are subject to social security contributions in ALWA, about 13 percent of the monthly observations did not have a wage measure in the administrative data. Filling the gaps, using the polynomial function, thus significantly improves the quantity of the data, which can be used for analyses. Moreover, the correlation of the predicted daily wage and the wage reported in the administrative data is 0.97, indicating a good fit.

4.3 Calculating hourly wages

Having obtained daily wages, the employment information from ALWA can be used to calculate hourly wages. In ALWA, respondents were asked to report their contract hours at the beginning of each employment spell. While contract hours may be subject to change throughout an employment episode, ignoring contract hours may lead to a bias if daily wages of both full- and part-time workers are being used. The daily wage reported in the administrative data lacks information on working days or exact contract hours. However, using information from ALWA allows for calculating hourly wages under the assumption that contract hours are constant during an employment episode.[4]

Figure 5 shows the trajectory for actually measured log daily wages as well as the predicted logarithms of daily wage and hourly wages.

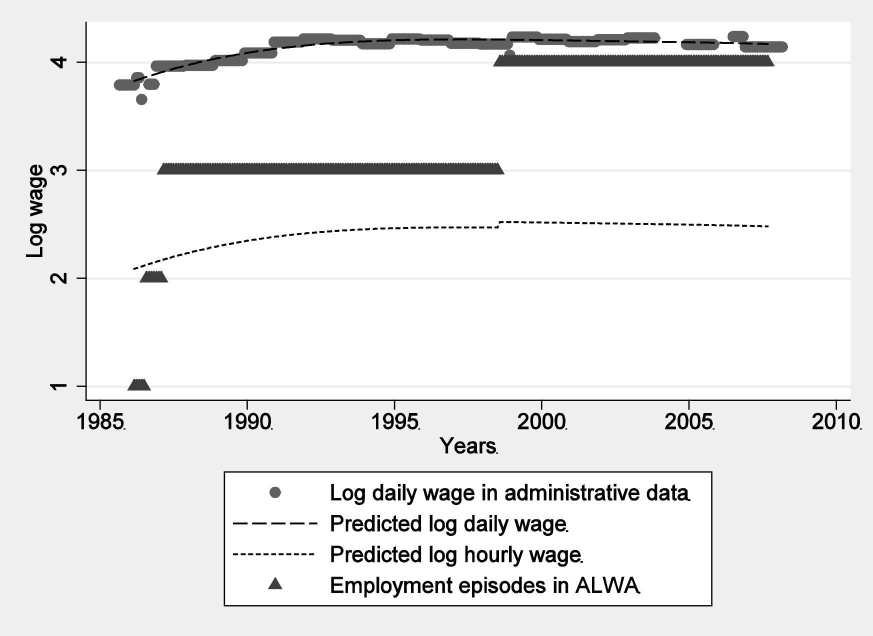

Figure 5 Predicted hourly wage

5 A matching indicator: Comparing reported income in ALWA with predicted daily wages

Predicting daily and hourly wages provides the possibility to use longitudinal wage information in ALWA. However, as outlined above, several assumptions have to be made and some sources for potential bias remain.

[4] Especially for women, this assumption may be problematic. Childbirth may lead to maternity leave and a subsequent reduction in working hours. Moreover, contract hours may later on be increased again. One should be aware of this limitation when using hourly wages in statistical analyses. However, the information if contract hours are altered throughout an employment relationship is included in ALWA and may serve as an additional control variable.

To check the validity of the data and retrieve a measure on how good the predicted values fit the actual income, it is possible to compare the predicted wages and the income that respondents reported at their interview date, as well as hourly wages calculated from both sources. All respondents were asked for their net income they obtained the month before the interview date (Matthes & Trahms, 2010). Comparing the logarithm of the predicted daily wage for the respective month with the logarithm of net income for full-time employees with at least 35 contract hours a week in ALWA gives an indication on how valid the prediction is. However, one has to bear in mind that respondents were asked for the sum of their income, whereas the prediction from the administrative data only provides information on the daily wages. Moreover, the administrative data provides gross wages, whereas respondents were asked to report net of their income. Third, social desirability and rounding of income might lead to slightly different values.

Despite all the potential sources for bias, the correlations show that indeed the predicted measures seem to be a good approximation to the actual earnings. First, the predicted values correlate highly (0.97) with the administrative wage data, indicating a good fit of the wage trajectories. Second, the gross daily wages for full-time employees correlate fairly high (0.84) with the net wages from the survey. Third, hourly wages for all employees including part-time workers correlate highly (0.83).

6 Conclusion

Using longitudinal wage information in linked datasets such as ALWA-ADIAB is not straightforward. First, due to the upper earnings limit for social security contributions, parts of the wages have to be imputed. Second, although information from the ALWA survey is linked to administrative data from the Federal Employment Agency (BA) on the personal/individual level, matching single employment spells may not always be possible. To solve these problems, I proposed to impute wage information that is censored due to the upper earnings contribution and to calculate a polynomial time function to smoothen wage information over time. The predicted wage can afterwards readily be matched to the respondents in ALWA. The result is the supplement of ALWA with longitudinal wage information for all main employment episodes in West Germany, starting after January 1975, and for episodes in East Germany starting after January 1993.

The prediction of daily wages from administrative data and the calculation of hourly wages introduce the possibility of conducting statistical analyses in ALWA using longitudinal wage information. The method may also be of interest for other datasets that are linked to administrative data sources. Using the method I have proposed in this report, however, limits the kind of analyses that can be conducted. As the wage measure is already slightly smoothened, comparing two values that follow each other closely in time may not be expedient. When using such wage information, one should be aware that the predicted variable rather contains trajectories and is not able to show direct jumps in wages through job changes.

References

Antoni, M., Drasch, K., Kleinert, C., Matthes, B., Ruland, M., & Trahms, A. (2010). Arbeiten und Lernen im Wandel * Teil I: Überblick über die Studie - März 2011 (2. aktualisierte Fassung des Berichtes vom Mai 2010) *FDZ-Methodenreport* (Vol. 5). Nürnberg: IAB.

Antoni, M., Jacobebbinghaus, P., & Seth, S. (2011). ALWA-Befragungsdaten verknüpft mit administrativen Daten des IAB (ALWA-ADIAB) 1975-2009. *FDZ-Datenreport, 05/2011*.

Antoni, M., & Seth, S. (2012). ALWA-ADIAB – Linked Individual Survey and Administrative Data for Substantive and Methodological Research. *Schmollers Jahrbuch, 132*, 141-146.

Drasch, K., & Matthes, B. (2013). Improving retrospective life course data by combining modularized self-reports and event history calendars. *Quality & Quantity. International Journal of Methodology, 47*(2), 817-838.

Drews, N. (2007). Variablen der schwach anonymisierten Version der IAB-Beschäftigten-Stichprobe 1975-2004. *FDZ-Datenreport, 3*.

Dustmann, C., Ludsteck, J., & Schönberg, U. (2009). Revisiting the German Wage Structure. *The Quarterly Journal of Economics, 124*(2), 843-881.

Gartner, H. (2005). The imputation of wages above the contribution limit with the German IAB employment sample. *FDZ-Methodenreport, 2*.

Kleinert, C., Matthes, B., Antoni, M., Drasch, K., Ruland, M., & Trahms, A. (2011). ALWA – New Life Course Data for Germany. *Schmollers Jahrbuch, 131*, 625-634.

Matthes, B., & Trahms, A. (2010). Arbeiten und Lernen im Wandel Teil II – Codebuch. *FDZ-Datenreport, 2*.

Imprint

FDZ-Methodenreport 1/2015

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
Stefan Bender, Dagmar Theune

Technical production
Dagmar Theune

All rights reserved
Reproduction and distribution in any form, also in parts,
requires the permission of FDZ

Download
http://doku.iab.de/fdz/reporte/2015/MR_01-15.pdf

Internet
http://fdz.iab.de/

Corresponding author:
Malte Reichelt
Institute for Employment Research (IAB)
Regensburger Str. 104
D-90478 Nürnberg
Phone: +49-911-179 3349
E-Mail: Malte.Reichelt@iab.de