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|2019 EN

A simple algorithm to link "last hires" from the Job Vacancy Survey to administrative records

Benjamin Lochner

Bundesagentur für Arbeit

A simple algorithm to link "last hires" from the Job Vacancy Survey to administrative records

Benjamin Lochner (IAB)

Contents

List of figures

List of tables

Zusammenfassung

Die IAB Stellenerhebung befragt Betriebe in Deutschland unter anderem zu ihrem letzten Fall einer Neueininstellung. In den Daten existiert dabei keine Möglichkeit, die Informationen zu der eingestellten Person, über deren Neueininstellung berichtet wurde, mit den administrativen Daten des IAB direkt zu verknüpfen. Dieser Bericht beschreibt einen Algorithmus, der versucht diese Lücke zu füllen. Er benutzt dazu beobachtbare Charakteristika der eingestellten Beschäftigten. Durch eine Vielzahl von Plausibilitätsprüfungen soll eine hohe Verknüpfungsqualität gewährleistet werden. Durch den Algorithmus ist eine Verknüpfung von ca. 70 Prozent der Neueininstellungen möglich, die grundsätzlich als verknüpfbar gelten. Für diese Fälle ist eine Identifizierung der in der IAB Stellenerhebung erfassten Neueininstellungen in den Integrierten Erwerbsbiographien des IAB möglich.

Abstract

The IAB Job Vacancy Survey asks German establishments, among other things, about their most recent hire. Unfortunately, a worker identifier that would allow the direct linking to administrative records is not available. This report describes an algorithm that allows to find reported hires in the administrative employment histories. Based on observable characteristics, the algorithm runs several plausibility checks that make sure that a valid and unique linkage is performed. With its default parameterization the algorithm finds around 70 percent of hires that were mergeable in the first place. The result is the identification of the most recent hire reported in the IAB Job Vacancy Survey in the Integrated Employment Biographies of the IAB.

Keywords

Job Vacancy Survey, administrative records, record linkage, algorithm

1 Introduction

The IAB Job Vacancy Survey (hereafter JVS) collects data on a variety of topics with regard to the hiring process of German establishments (see Bossler et al. 2019 for a detailed data description). It identifies the number of vacancies on the German labor market, including those vacancies that are not reported to the Federal Employment Agency (FEA), Germany's public employment service. The main questionnaire which is conducted in every fourth quarter of a year collects information on the number and structure of vacancies, future labor demand, the current economic situation, and the expected economic development of participating establishments. A major part of the survey collects information on the most recent hire of an establishment. In particular, establishments are asked whether or not they have filled a position during the last 12 months. If they did, they are further asked about certain job characteristics such as the exact job requirements, the hiring channel, the search duration and the exact hiring date. Furthermore, establishments report certain individual attributes of the most recent hire such as gender, age, as well as match-specific characteristics like educational qualification, wage bargaining, and (in some waves) the hourly wage.

In some research contexts the employment history of the last hire, both prior to and after the hiring process, is of primary interest. Unfortunately, any worker identifier that would allow the linkage to existing administrative data is not available.

This report describes an algorithm that aims to link the reported last hires in the JVS to the administrative records even in the absence of an individual identifier in the JVS. More precisely, it tries to identify the job spell in the Integrated Employment Biographies (IEB) record that is related to the hire that is reported in the JVS. Once a link to the IEB has been established, it is possible to uncover the worker's identification number in the IEB. For a given worker identifier one can link the worker's full (un)employment history from the admin data to the JVS.

Please note that for validation purposes, the availability of the record linkage is currently confined to internal use at the Institute for Employment Research.

2 Data

The algorithm relies on two main data sources. The JVS and the IEB. In order to enhance the accuracy of the record linkage, additionally the initials of workers' last names, which were provided by the Data and IT Management of the IAB, were merged to the IEB.

Since 2010, surveyed establishments implicitly declared that researchers are allowed to link the survey to existing IAB data. In 2009, a subset of establishments directly consented to linking their information. Therefore, this algorithm uses the main JVS that is executed in every fourth quarter from 2009 to 2016. Some questions in the questionnaire are retrospective, so for each establishment that has been surveyed between 2008 and 2016, the full IEB records are used.

3 The algorithm

3.1 Key assumptions

Since the IEB contains each job spell that is subject to social security contributions, the key assumption the algorithm relies on is that the reported last hire from the JVS must be observable at least once in the IEB. Therefore, all of the first spells of newly hired workers in a given establishment are needed. This assumption seems trivial at first glance. However, we will see later in this report that due to hires that are very similar in some observable attributes, this assumption gets crucial.

3.2 Initial data preparation

3.2.1 JVS

The JVS initially has no establishment identifier. However, there is a one-to-one mapping from surveyed establishments to a unique, IAB internal (system free) establishment identifier. Hence, the first step is to merge these identifiers to the raw survey data.

Although establishments are asked to report their most recent hire during the last 12 months, it happens that they do not. In some rare cases they report a hire with the same date in two successive survey waves. Since this constellation would cause issues when merging the survey data to the administrative data, the reported hire is selected that is closest to the survey date.

3.2.2 IEB

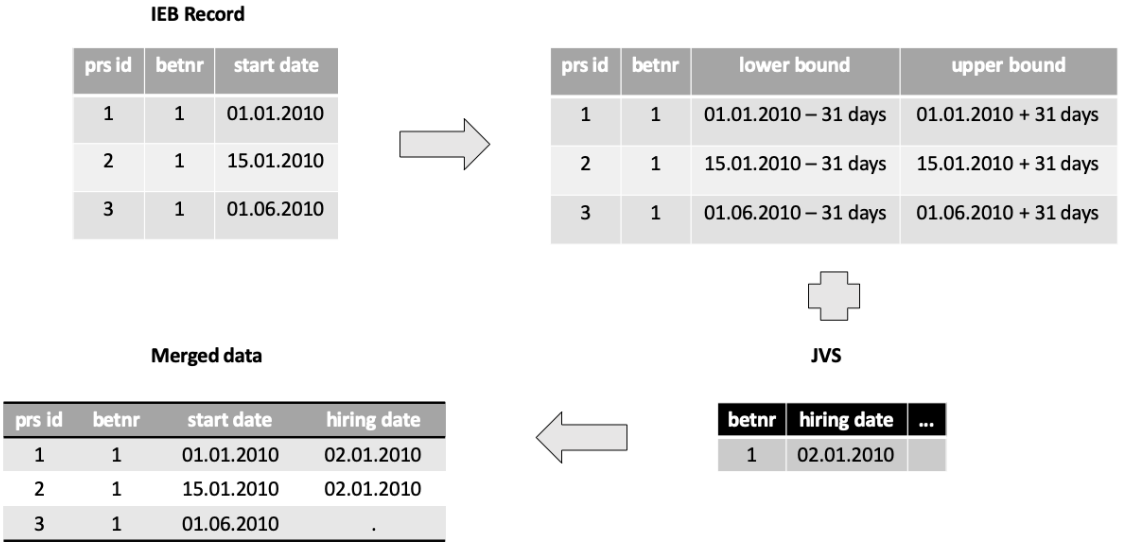
It happens that there are duplicate job spells with the same start date for a given establishment-worker pair in the IEB. For instance, this can be the case when a contract was supposed to end but is extended spontaneously or there is a bonus payment at the end of the year. If those duplicates are observed, the longest spell and/ or the one with the highest average daily wage is kept.

3.3 Merging strategy

The first step is to connect the survey data with the administrative data. To link both data sets, one could join both data sets by the establishment identifier and ignore the time dimension. However, this technique would be computationally complex because each wave of the JVS (and therefore potentially each case of last hire) would be combined with all the job spells of each establishment in the IEB. Many of them would obviously be wrong in terms of the hiring date and would never lead to a “true” match. If one could be sure that there was no noise in the information on the date of a hiring, one could even merge by establishment and the exact hiring date, because these information are present in both data sets. Unfortunately, it turned out that the information on the hiring date is noisy, so this procedure would only expose a fraction of last hires.

It turned out that a mixture of both merging strategies appears to be sufficient to solve both merging issues. Figure 1 shows a simple representation of the merging procedure. The logic is as follows: first, a time band around all starting dates in the IEB is constructed. Then, all last hires from the JVS that have a starting date within that band at the same establishment are merged. All observations that could not be merged are dropped from the data sets, i.e. those outside the time band. The default time band is 31 days, symmetrical before and after the starting date of a job in the IEB. However, the researcher can change this time band easily if required. The merging procedure results in pairs of observations (or job spells) that potentially match, i.e. the reported hiring date lies in the constructed time range.

Figure 1: Merging procedure



IEB Record

prs id	betnr	start date
1	1	01.01.2010
2	1	15.01.2010
3	1	01.06.2010

prs id	betnr	lower bound	upper bound
1	1	01.01.2010 – 31 days	01.01.2010 + 31 days
2	1	15.01.2010 – 31 days	15.01.2010 + 31 days
3	1	01.06.2010 – 31 days	01.06.2010 + 31 days

JVS

betnr	hiring date	...
1	02.01.2010	

Merged data

prs id	betnr	start date	hiring date
1	1	01.01.2010	02.01.2010
2	1	15.01.2010	02.01.2010
3	1	01.06.2010	.

Note: Figure shows a representation of the initial merging procedure that links last hires from the JVS to the IEB.

3.4 Identifying last hires

In order to identify the "true" matches (or weed out false matches) a simple but effective algorithm is conducted. In principle, the idea of this algorithm is simple: after merging the two data sets, the worker-specific attributes that are present in both data sets are compared and then the algorithm picks the most "reliable" match (according to a few definitions that are explained below). The hard part is finding out what "reliable" means. Since these definitions are obviously debatable, these definitions are partly up to the researcher who wants to use this algorithm. However, some default values are provided.

The main variables that are useful to find the hire in the IEB are

- the establishment identifier
- the date of hiring
- the gender
- the age, and
- the hire's occupation.

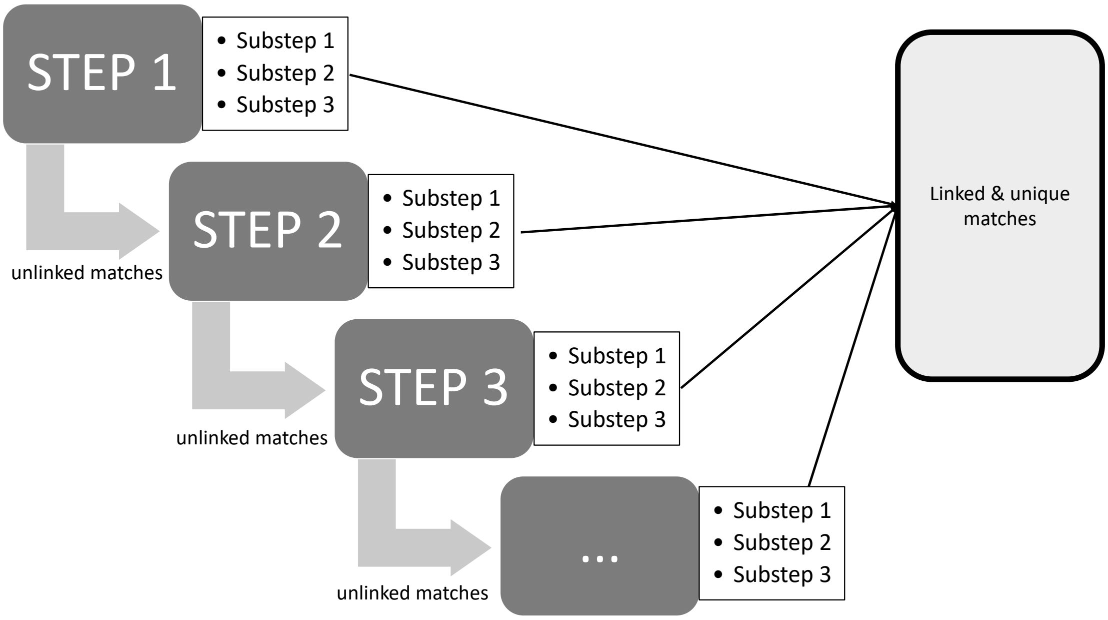
In addition, the method uses a few more variables such as the first letter of workers' last names, working hours, and the wage information for a few plausibility checks which are described in more detail below.

The algorithm consists of six main steps.[1] Each of these steps has potentially three substeps.
The aim of each step is to find combinations of the two data sets that match according to certain step-specific definitions. Each step ends whenever there are no more valid and unique matches to find. A match is valid and unique if and only if i) in a given time period, in a given establishment, the step-specific conditions are met AND importantly ii) there is no other hire in the administrative data that has the same attributes.
If a match is valid and unique, the algorithm does not proceed to the next step. Whenever a match is valid but not unique, the algorithm continues with the three substeps. First, the working hours are compared to the information about part- and full-time employment in the IEB. Second, the algorithm checks if conditional on working hours, the wage information in the administrative data is "reliable". "Reliable" in this step means i) not zero, and ii) above the social security threshold of marginal employment.

[1] The order of steps can influence the outcome, however the impact is minor.

In the third substep, an instruction to the person who is asked to answer the JVS is used. When the establishments are asked about their last hire, they were told that in case they hired two workers at the same date, they should report the hire whose name comes first in the alphabet. This information can be used, however it is necessary to check for plausibility. If and only if the potential match passes substep one and two, the match with the name first in the alphabet is chosen. Figure 2 shows the procedure in greater detail.

Figure 2: Implementation



The definition of the main steps are:

Step 1:

The first step is the most restrictive one, which means that all of the four main variables (hiring date, gender, age, occupation code) must be equal without any deviations between the two data sets.

Step 2:

The second step does not allow for any deviations in age, gender, and the occupation code, but allows for minor deviations in the hiring date between the two data sets. The reason for this is that it might be the case that establishments report this information imprecisely. The default deviation the algorithm takes into account is 31 days, symmetrically before and after the reported date in the survey.[2]

Step 3:

The third step takes into account missing values in the occupation codes. More precisely, it allows for constellations where in either in the administrative data or in the survey data the occupation code is missing while in the other data set this information is present. Note that while this step loosens the occupational criteria, it tightens the hiring date restriction as it goes back to the exact date constraint as in step 1 (no deviation in the hiring date).

Step 4:

The fourth step allows for absolute deviations in age. Since in the JVS establishments do not report the exact date of birth, but the absolute age at the hiring date, this variable is potentially noisy. The default value allows age deviations of up to two years. Note that this step goes back to the stricter condition for the hiring date (no deviation between the two data sets).

Step 5:

The fifth step combines Step 3 and 4 as it allows deviations in age of up to two years as well as missing occupational information. Note that this step conditions on the exact hiring date as in step 1 (no deviation between the two data sets).

Step 6:

All matches that did not pass steps one to five go to the last step. This step allows for implausible occupation code combinations as well as deviations in the hiring date up to again 31 days.

4 Results

From 2009 to 2016 there are 67,028 reported cases of last hires from 102,281 establishment-year observations from 79,002 unique establishments in the JVS.

Among those, there are 55,336 observations that are i) not missing in the main variables (hiring date, gender, age) and ii) can potentially be linked within the default time range of 31 days.[3]

[2] It is up to the researcher to change this value. However, please note that this deviation must be smaller or equal to the time range in the first merge command.

[3] Note that this number might change if the default time range changes. Due to missing information in the main variables the sample reduces to 57,864 cases of last hires. For the remainder (57,864 – 55,336 = 2,528), there are no hires reported in the IEB within the default time range.

Table 1 shows the number of potential matches, initial duplicates, duplicates that could be solved, and finally the number of valid and unique matches.

Duplicates are matches that are valid but not unique, i.e. duplicates are reported hires from the JVS that fit the definition of a step (valid), however in the IEB there are many (more than one) of them (non-unique). As duplicates inflate the number of (potential) matches, it is informative to distinguish between the number of hires and the number of observations these hires produce in the merged data. Note that by construction these numbers add up, e.g. in Step 1 there are 9,673 initially matched observations from 9,289 JVS hires. These are inflated by 554 duplicate observations from 170 JVS hires. 66 of the duplicates could be solved by the algorithm. This leads to 9,185 valid and unique matched hires (9,673 -554 +66=9,185 or 9,289 -170+66=9,185).

Table 1: Matches and duplicates

Step	potentially matched observations/ number of JVS hires	initial duplicate observations /number of JVS hires	duplicates solved (number of JVS hires)	valid & unique matches (number of JVS hires)
1	9,673/ 9,289	554/ 170	66	9,185
2	3,736/ 2,229	1,932/ 425	142	1,946
3	17,751/ 14,682	3,903/ 834	377	14,225
4	4,484/ 3,820	874/ 210	83	3,693
5	13,126/ 7,282	6,887/ 1,043	415	6,654
6	9,916/ 3,976	7,174/ 1,234	389	3,131

Note: The table shows the number of potential matches, intital duplicates, solved duplicates, and final matches.

In total, around 70 percent (38,834 out of 55,336) of mergeable cases of hires (i.e. those without missing information on the main variables) can be identified under the underlying assumptions. Table 2 in the Appendix shows a comparison between some of the hires‘ characteristics in the initial JVS and in the linked data.

5 Discussion

It is plausible to think that the probability of the occurrence of duplicates and hence the number of matched hires is related to certain characteristics. Larger establishments are likely to exhibit a higher probability that relatively similar people, in terms of the characteristics the algorithm conditions on, are hired simultaneously. Figure 3 to Figure 6 in the Appendix deal with this issue. Figure

3 shows that the majority of reported hires in the JVS stems from small establishments with less than 25 employees. Figure 4 shows the number of hires by steps of the algorithm that are not unique in the first run. The results with respect to establishment size are mixed: In step 1 and 3, there is a positive relationship between the number of non-unique hires and size, while it is negative for step 4. For the remaining steps, there is no clear-cut pattern. Figure 5 shows the distribution of duplicates per hire conditional on a positive number of duplicates. Intuitively, these graphs show that, if there are hires which are not unique, these produce more duplicates in large establishments as compared to small establishments. This pattern seems to be true for all steps of the algorithm and particularly driven by the very largest establishment size category. Figure 6 shows the number of hires that the algorithm could recover, per number of initial non-unique hires. Intuitively, this statistic is a measure for the success of the algorithm and how it is distributed across size categories. The results are mixed: For steps 1, 2, and 6, there is a tendency that the algorithm can recover more non-unique hires in larger establishments, whereas there seems to be no clear-cut relationship in the other steps.

6 Conclusion

The report describes an algorithm that aims to link hires from the JVS to administrative records. The algorithm performs several steps that make sure that a valid and unique linkage is guaranteed. With its default parameterization the algorithm finds around 70 percent of hires that were mergeable in the first place. The result is the identification of the IAB internal worker identification number that allows linking the workers' full (un)employment history from the administrative records to the JVS.

7 Literature

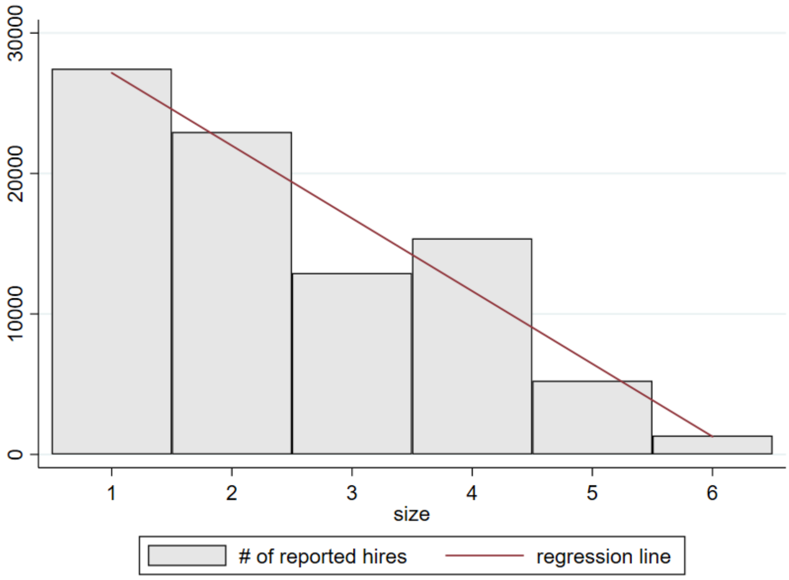
Bossler, Mario; Gartner, Hermann; Kubis, Alexander; Küfner, Benjamin; Rothe, Thomas (2019): The IAB Job Vacancy Survey: Establishment survey on labour demand and recruitment processes, Waves 2000 to 2016 and subsequent quarters 2006 to 2017. (FDZ-Datenreport, 03/2019 (en)), Nürnberg, 18 S.

8 Appendix

Table 2: Sample comparison

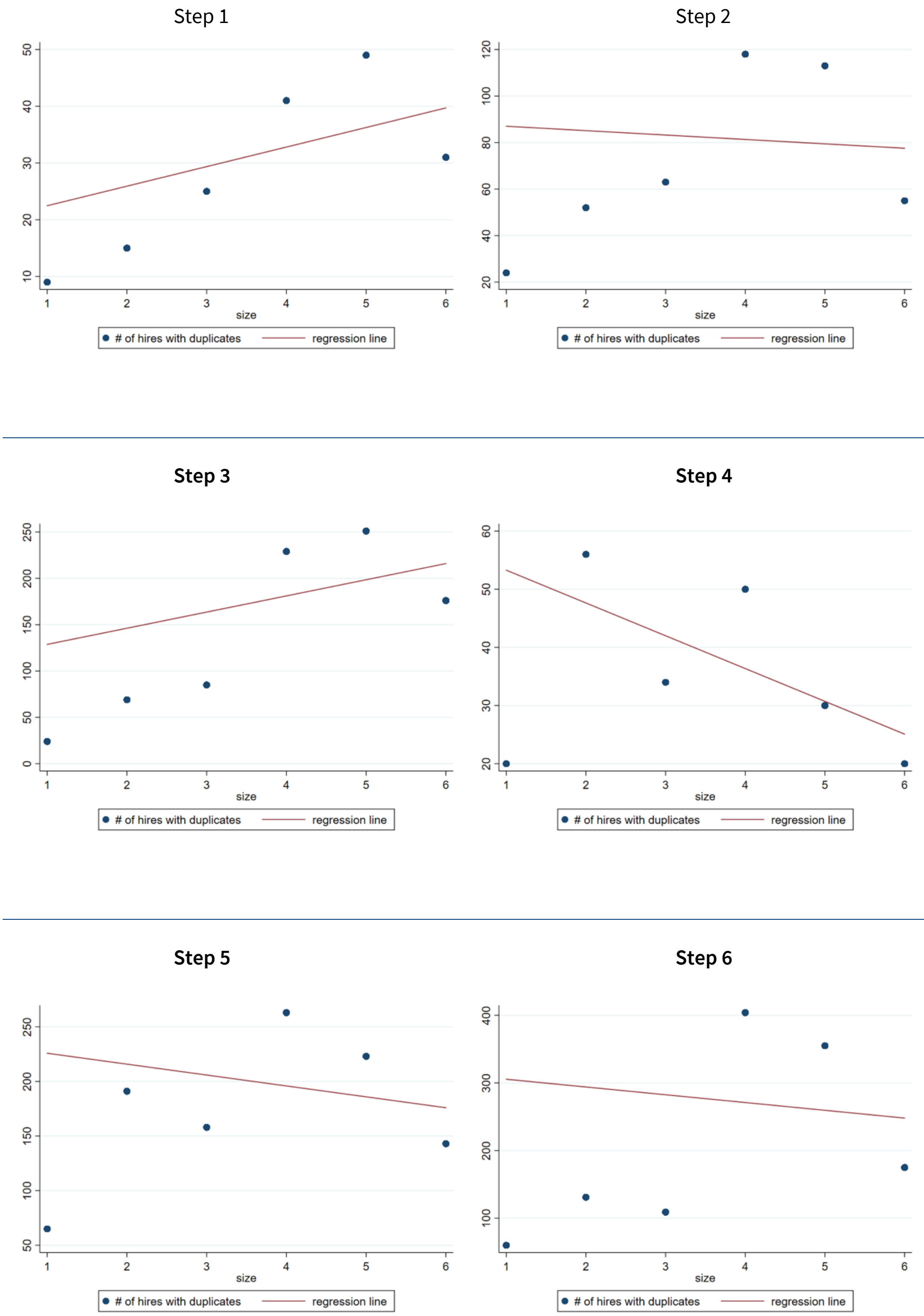
	Number Obs		Mean		Standard Dev.		Min/Max	
Attribute	JVS Hires	Matched Hires	JVS Hi-res	Matched Hires	JVS Hires	Matched Hires	JVS Hires	Matched Hires
Gender (1=Female)	58,770	38,834	1.55	1.55	0.50	0.50	1.00/ 2.00	1.00/ 2.00
Age	57,864	38,834	36.05	35.94	10.91	10.96	15.00/ 74.00	15.00/ 73.00
Working Hours	57,525	38,053	36.54	36.58	6.92	6.83	1.00/ 60.00	1.00/ 60.00
Establish-ment Size	58,770	38,834	137.33	146.82	1,091	1,152	1.00 167,447	1.00/ 167,447

Figure 3: Number of reported hires in the JVS across establishment size categories



Note: establishment size is binned into 6 size bins: 1 = less than 10 employees, 2 = 11- 25 employees, 3 = 26 - 50 employees, 4 = 51 – 250 employees, 5 = 251 - 1000 employees, 6 = more than 1000 employees.

Figure 4: Number of non-unique hires



Step 1

Step 2

Step 3

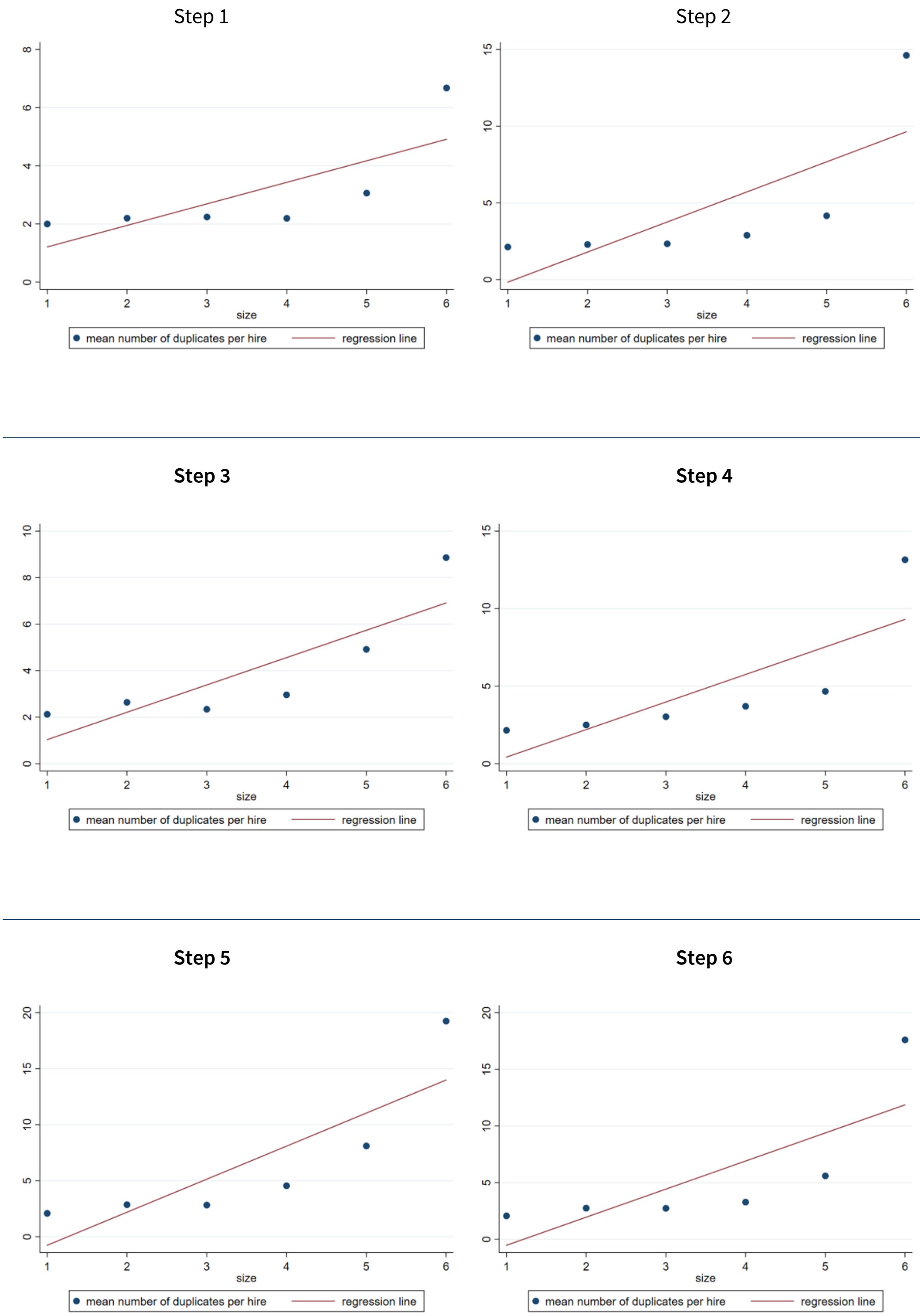
Step 4

Step 5

Step 6

Note: establishment size is binned into 6 size bins: 1 = less than 10 employees, 2 = 11- 25 employees, 3 = 26 - 50 employees, 4 = 51 – 250 employees, 5 = 251 - 1000 employees, 6 = more than 1000 employees.

Figure 5: Number of duplicates per non-unique hires



Step 1

Step 2

Step 3

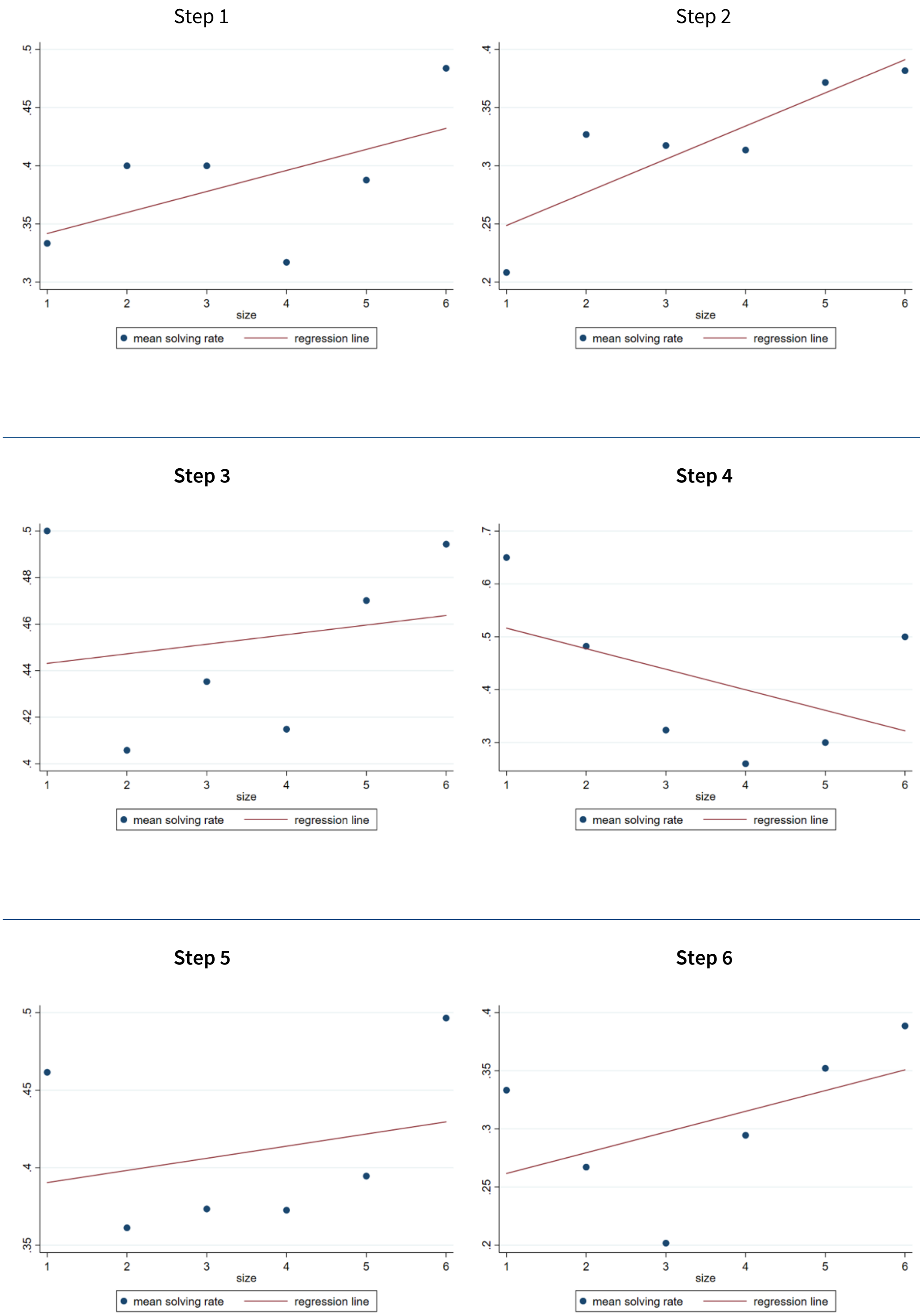
Step 4

Step 5

Step 6

Note: establishment size is binned into 6 size bins: 1 = less than 10 employees, 2 = 11- 25 employees, 3 = 26 - 50 employees, 4 = 51 – 250 employees, 5 = 251 - 1000 employees, 6 = more than 1000 employees.

Figure 6: Number of recovered hires per non-unique hires ("solving rate")



Step 1

Step 2

Step 3

Step 4

Step 5

Step 6

Note: establishment size is binned into 6 size bins: 1 = less than 10 employees, 2 = 11- 25 employees, 3 = 26 - 50 employees, 4 = 51 – 250 employees, 5 = 251 - 1000 employees, 6 = more than 1000 employees.

Imprint

FDZ-Methodenreport 06|2019 EN

Date of publication
10. December 2019

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

Download
http://doku.iab.de/fdz/reporte/2019/MR_06-19_EN.pdf

Documentation version

DOI: 10.5164/IAB.FDZM.1906.en.v1

All publications in the series "FDZ-Methodenreport" can be downloaded from
https://fdz.iab.de/en/FDZ_Publications/FDZ_Publication_Series/FDZ-Methodenreporte.aspx

Website
http://fdz.iab.de

Corresponding author
Dr. Benjamin Lochner
Phone: +49 911 179-6564
Email Benjamin.Lochner@iab.de