SORT Special issue: Privacy in statistical databases, 2011, 7-24

Remote data access and the risk of disclosure from linear regression

P. Bleninger[1], J. Drechsler[1,*], and G. Ronning[2]

[1]*Institute for Employment Research,* [2]*Tübingen University*

Abstract

In the endeavor of finding ways for easy data access for external researchers remote data access seems to be an attractive alternative to the current standard of data perturbation or restricted access only at designated data archives or research data centers. However, even if the microdata are not available directly, disclosure of sensitive information is still possible. We illustrate that an ill-intentioned user could use some commonly available background information to reveal sensitive information using simple linear regression. We demonstrate the real risks from this approach with an empirical evaluation based on a German establishment survey, the IAB Establishment Panel.

MSC: 62J68, 62P68

Keywords: Remote Data Access, Data Privacy, Disclosure, IAB Establishment Panel, Linear Regression, Artificial Outlier, Strategic Dummy.

1. Introduction

Data collecting agencies generally have two options if they are willing to provide access to their data for external researchers. They can release data sets to the public if they can guarantee that the dissemination will not harm the privacy of any survey respondent or they can allow external researchers on-site access to the data in research data centers (RDC) or data enclaves. Since most data have to be altered in some way to allow data dissemination, many researchers prefer the direct access to the unaltered data at the RDC, especially if the data dissemination requires perturbation of the microdata. For this reason more and more agencies deposit their data at data enclaves or set up their own research data centers. However, the use of these facilities comes at a high price both for

* Corresponding author: Jörg Drechsler, Institute for Employment Research, Regensburger Str. 104, 90478 Nuremberg; e-mail: joerg.drechsler@iab.de
Received: November 2010
Accepted: March 2011

the researcher and the providing agency. Researchers have to travel to the agency before they ever get in touch with the original data. Although some agencies provide dummy data sets to give the researcher an idea of the real data, these dummy data sets often are of very low quality and the researcher might not realize that the data collected by the agency is not suitable for her analysis before traveling to the agency. Furthermore, researchers can request a certain time slot at the RDC in which they expect to finish their research. It is very difficult for the researcher to anticipate how long the data preparation will take without access to the data, and unexpected problems might require more days than the admitted time slot will allow. Besides, if the researcher wants to extend her research maybe using more variables than she asked for in the original proposal, she might have to go through the complete reviewing process again before she can actually add the variables to her analysis. On the other hand, the agency has to check every output from the analysis for potential disclosure violations. Only cleared outputs may leave the RDC and may be used by the researcher for publication. At present, this output checkin is still carried out manually. With the growing popularity of the RDCs the capacity of handling all this output checking is at the limit.

Given these drawbacks remote data access seems to be the panacea for data access for external researchers. In an ideal world full remote access would enable the external researcher to connect to a host server from her desktop machine. She would see the microdata on the screen and would be allowed to manipulate them in any way but the actual data would never leave the server and it would not be possible to store the microdata on the desktop computer. Requested queries would be automatically scanned for possible confidentiality violations and only those queries that pass the confidentiality check would be answered by the server. Remote access would free the researcher from the burden of traveling to the RDC and it would render the cost intensive and time consuming manual output checking unnecessary. However, there are many obstacles with this approach making the full implementation of a remote data access more than questionable. Apart from the technical issues of guaranteeing a safe connection between the desktop computer of the external user and the microdata server at the agency, direct access to the unchanged microdata is prohibited by law in many countries. For example in Germany, the data accessible for external researchers is required to be *de facto* anonymised which means that the effort that is necessary to identify a single unit in the data set is higher than the actual benefit the potential intruder would achieve by this identification. This is still a privilege compared to the *absolute* anonymity that is required for all published results. One solution in this context could be that the researcher would only see an anonymised version of the microdata on her screen but the queries she submits to the server would actually be run on the original data. However, this would still require the server to identify all queries that might lead to a breach of confidentiality.

Some of these queries are easy to identify. For example queries that ask for the maximum or minimum of a variable should never be allowed. For tabulation queries potentially identifying small cells could be suppressed using standard rules from the

cell suppression literature.[1] However, there are other analyses for which it is not that obvious that they actually might impose an increased risk of disclosure and illustrating this for a specific set of queries is the main aim of this paper.

We focus on the risks from simple linear regression analysis under the assumption that the user will never see the true microdata. Given the legal restrictions in many countries (see discussion above), we believe that even under remote access the user will only see an anonymised version of the true microdata. In this sense our notion of remote access is located somewhere in the middle between the dream of a full remote access and the idea of a remote analysis server that can only answer specified queries without providing access to any microdata at all. We note that our findings are also relevant in the context of a plain remote analysis server.

Often regression analysis is considered as safe in the sense that it is assumed that no output checking is required. Following the discussion in Gomatam *et al.* (2005) we illustrate that an intruder with background knowledge on some of the variables contained in the data set can get accurate estimates for any sensitive variable she is interested in using only the results from a linear regression analysis. We use the IAB Establishment Panel to demonstrate empirically that at least for business data very limited and easily available background information can be sufficient to allow the intruder to obtain sensitive information with this approach.

The remainder of the paper is organized as follows. Section 2 recapitulates the basic concept that allows the intruder to retrieve sensitive information for a single respondent based on the background information she has about that respondent. In this section we follow the outline described in Gomatam *et al.* (2005). In Section 3 we briefly introduce the data set we used for the empirical simulations: the IAB Establishment Panel. This data set is used in Section 4 to illustrate that only very limited background information is required to learn sensitive information about a survey respondent in this setting. The paper concludes with some final remarks.

2. The formal approach

In the following we assume that the intruder has at least approximate knowledge about some of the variables contained in the survey for a certain survey respondent *m*. It is important to note that this knowledge may refer to any set of variables in the data set, no matter if the variables are sensitive or not. For example in a business survey, the external information available to the intruder might be the energy consumption or the total production time. The intruder would then use these variables for obtaining information on sensitive variables such as investment, sales, or research expenditures.

1. Even cell suppression can quickly become problematic, if we allow dynamic queries. In this case, the server would have to keep track of all earlier queries and would have to guarantee that requests submitted at a later point in time would not allow the calculation of cell entries that are being suppressed now.

In the following we denote the variable for which information is at hand by x and the true value for this variable provided by the survey respondent m by x_m^0. Let $\hat{x}_m$ be the external information the intruder obtained about the survey respondent m for this variable. Finally, let y_m be the reported value for respondent m for the sensitive variable of interest y.

Gomatam *et al.* (2005) pointed out that the knowledge of $\hat{x}_m$ may be used to obtain information for any other variable contained in the microdata set for this respondent by making the variable of interest the dependent variable in a simple linear regression analysis. The authors propose two approaches: (i) The intruder could generate an "artificial outlier" obtained by transformation. (ii) Alternatively, the intruder could employ a "strategic dummy variable" which uses the background information for identifying the respondent m.

2.1. Artificial outliers

For the artificial outlier approach we assume the intruder knows the exact reported value for x_m, that is $\hat{x}_m = x_m^0$. She defines a new regressor variable

$$z = \frac{1}{|x - \hat{x}_m| + \varepsilon} \tag{1}$$

where ε is arbitrarily small. If we include this regressor variable in a linear regression with the variable of interest specified as the dependent variable, the regressor z will become extremely large for the respondent m and therefore generates a leverage point such that the predicted value of the dependent variable tends towards the true value y_m^0 for this respondent. A formal proof that

$$\lim_{z_m \to \infty} \hat{y}_m \quad = \quad y_m$$

holds, is given in Appendix A1. It is important to note that this is true only if no other respondent reports a value for x that is equal to x_m^0. If other respondents report the same value, y_m will generally not be predicted exactly (see Appendix A1 for details).

2.2. Strategic dummies

Alternatively, the intruder could define a dummy that exploits the knowledge regarding the variable x. In case of exact knowledge of the reported value the dummy would be given by

$$\Im_{x=x_m} = \begin{cases} 1 & \text{if } x = \hat{x}_m \\ 0 & \text{else.} \end{cases} \tag{2}$$

In other situations only vague information might be available represented by an interval in which the true value x_m^0 must fall. This range might be formulated in additive or multiplicative terms, that is

$$x_m^0 - \gamma < \hat{x}_m < x_m^0 + \gamma \quad \text{or} \quad (1-\delta)x_m^0 < \hat{x}_m < (1+\delta)x_m^0.$$

Thus, assuming only approximate knowledge one would create a strategic dummy according to

$$\Im_{x \simeq x_m} = \begin{cases} 1 & \text{if } x-\gamma < \hat{x}_m < x+\gamma \\ 0 & \text{else} \end{cases} \tag{3}$$

or the corresponding multiplicative specification mentioned above.
It is shown in Appendix A.2 that a simple regression which uses just this dummy variable and any variable of interest as the dependent variable will result in

$$\hat{y}_m = y_m^0.$$

The result remains valid if other regressors are added to the model (see Appendix A.2).

However, the proof again is based on the assumption that only a single respondent is identified using the knowledge regarding x. If x is a categorical variable, this is an unrealistic assumption and even for continuous variables more than one respondent may report the same value. Still, with the dummy variable approach the constructed dummy can easily be based on more than one variable exploiting all the information the intruder has about the survey respondent. In our business survey example this could mean that the intruder uses her information about the industry, an approximate number of employees, and regional information about the establishment she is looking for. In this case we could define an indicator dummy for each variable for which the intruder has background information.

Let $x_1, \ldots, x_p$ be the variables for which background information is available and let $\Im_1, \ldots \Im_p$ be the corresponding indicators defined as in (2) or (3). Now the final indicator can be defined as follows:

$$\Im = \begin{cases} 1 & \text{if } \Im_1 = 1 \wedge \Im_2 = 1 \wedge \cdots \wedge \Im_p = 1 \\ 0 & \text{else.} \end{cases} \tag{4}$$

It is important to note that both the artificial and the strategic dummy approach critically rely on the assumption that a single record can be identified with the external information the intruder has about m. However, the artificial outlier approach requires that the intruder knows x_m^0 exactly. This is often unrealistic in reality. With the dummy variable approach it can be sufficient to have a rough estimate of x_m^0.

3. The IAB Establishment Panel

Since our empirical evaluations in the next section are based on the wave 2007 of the IAB Establishment Panel a short introduction of the data set should prelude our illustrations. The IAB Establishment Panel is based on the German employment register aggregated via the establishment number as of 30 June of each year. The basis of the register, the German Social Security Data (GSSD) is the integrated notification procedure for the health, pension and unemployment insurances, which was introduced in January 1973. This procedure requires employers to notify the social security agencies about all employees covered by social security. As by definition the German Social Security Data only include employees covered by social security – civil servants and unpaid family workers for example are not included – approx. 80% of the German workforce are represented. However, the degree of coverage varies considerably across the occupations and the industries.

Since the register only contains information on employees covered by social security, the panel includes establishments with at least one employee covered by social security. The sample is drawn using a stratified sampling design. The stratification cells are defined by ten classes for the size of the establishment, 16 classes for the region, and 17 classes for the industry.

These cells are also used for weighting and extrapolation of the sample. The survey is conducted by interviewers from TNS Infratest Sozialforschung. For the first wave, 4,265 establishments were interviewed in West Germany in the third quarter of 1993. Since then the Establishment Panel has been conducted annually – since 1996 with over 4,700 establishments in East Germany in addition. In the wave 2007 more than 15,000 establishments participated in the survey. Each year, the panel is accompanied by supplementary samples and follow-up samples to include new or reviving establishments and to compensate for panel mortality. The list of questions contains detailed information about the firms' personnel structure, development and personnel policy. For a detailed description of the data set we refer to Fischer *et al.* (2008) or Kölling (2000). For the simulations we use one data set with all missing values imputed. We treat all imputed values like originally observed values for simplicity. See Drechsler (2010) for a description of the multiple imputation of the missing values in the survey.

4. Empirical evidence

For our empirical evaluations, we use the wave 2007 of the establishment survey and treat the turnover of an establishment as the sensitive variable to be disclosed. Thus, we exclude all entities from the survey that do not report turnover such as non-industrial organizations, regional and local authorities and administrations, financial institutions, and insurance companies. The remaining data set includes 12,814 completely observed establishments. We analyze different subsets of the dataset defined by quantiles of the establishment size to illustrate the increased risk for larger establishments.

Table 1: *Disclosure risk evaluations using an artificial outlier generated from establishment size.*

	quantile	N	prop. uniqu. identified	Δ all identified	Δ uniqu. identified
	all	12814	0.034	13773.795	0.001
size	0.5	6516	0.066	26936.156	0.001
	0.75	3217	0.134	51952.053	0.001
	0.9	1282	0.335	1.957	0.001
	0.99	129	0.969	0.011	0.0001

4.1. Empirical evidence for artificial outliers

Using the number of employees as the available background information we construct a variable z according to (1) setting $\epsilon = 0.0001$. To evaluate the risks for the complete data set we successively treat each record in the data set as the target m for which background information is available. Table 1 summarizes the results of the artificial outlier regressions for different subsets of the data. The first column defines the subset of the data. For example, the results for the 90% quantile represent only the largest 10% of establishments. The second column provides the number of records that are contained in the subset. Column 3 contains the percentage of records that are uniquely identified based on an artificial outlier derived from the establishment size, i. e. it contains the percentage of unique high leverage points regarding the number of employees. If there is more than one high leverage point, additional establishments reduce the prediction accuracy for the target's turnover (see the proof in Appendix A.1). Column 4 presents the average absolute relative error between the predicted and the observed value for turnover for the target record m, i.e.

$$\Delta = \frac{1}{N} \sum_{j=1}^{N} \frac{|\hat{y}_{m=j} - y_{m=j}|}{y_{m=j}} \tag{5}$$

for all records in the subset. Finally, column 5 presents the same quantity only for the records that are uniquely identified and therefore generate a unique high leverage point for z.

As expected the disclosure risk clearly increases with establishment size. Under the assumption that the intruder would know the exact reported establishment size, we observe a substantial increase in the risk when going from the largest 10% of the establishments (33.5% correctly identified) to the largest 1% of establishments (96.9% correctly identified). Below these thresholds identification risks are relatively low since establishment size alone will not uniquely identify a single record. The results in column 4 illustrate that generally risks are low as long as a unique identification is not possible. The average absolute relative error is very large (often far more than 100%) indicating that the predicted value on average differs substantially from the reported value. Finally,

all the values close to zero in the last column are by no means surprising. This is a direct result of the proof given in Appendix A.1. We only include these results to emphasize that once a record is uniquely identified, the intruder does not have to have direct access to the microdata. Instead she can use the artificial outlier approach (or the dummy variable approach discussed below) to exactly reveal any sensitive information about the identified record.

Often the intruder will have more background information on the target than just one variable. Generally she can use this information to generate more artificial outliers and also include them in the regression. For brevity we omit the proof that an exact prediction is possible with more than one outlier variable. A detailed proof can be found in Ronning *et al.* (2010).

However, simply using two outlier variables in the regression will not necessarily increase the number of uniquely identified records. The proof only holds if both outliers individually identify the same single record uniquely. This means that in general there is no benefit from adding a second artificial outlier to the regression since the dependent variable will only be predicted correctly for those units for which one of the background variables alone already uniquely identifies the target. The same results would be achievable if the intruder would run two separate regressions using one outlier at a time. To fully utilize the additional background knowledge the intruder should interact the background variables and apply the artificial outlier approach to the interaction term. If the joint background information identifies a record uniquely, the value of the interaction term will also be a unique value in the data set.

We illustrate the increased risks if the intruder has information on more than one variable in Table 2. We assume the intruder knows the exact number of employees and the German Federal State in which the establishment is located and uses the interaction of the two variables to generate the artificial outlier.

As expected the disclosure risks increase considerably. For example 15.9% (87.2%) of the establishments in the complete data set (of the largest 10% of the establishments) are identified uniquely compared to only 3.4% (33.5%) if the establishment size is used alone to identify the target. In theory the intruder could further improve her results if more background information is available. The more variables are interacted to generate the artificial outlier the higher is the chance of a unique identification and thus a perfect

Table 2: *Disclosure risk evaluations using artificial outliers generated from the interaction term of establishment size and German Federal State.*

	quantile	N	prop. uniqu. identified	Δ all records	Δ uniqu. identified
size*fed. state	all	12814	0.159	17820.512	0.035
	0.5	6516	0.312	33908.516	0.036
	0.75	3217	0.596	63907.225	0.006
	0.9	1282	0.872	0.338	0.0003
	0.99	129	1	$1.86 * 10^{-5}$	$1.86 * 10^{-5}$

prediction. However, a regression using three-way, four-way or even higher interaction terms will look very suspicious or might not be allowed in a remote access setting.

4.2. Empirical evidence for strategic dummies

For the strategic dummy approach we evaluate for each record if a unique identification is possible using a varying amount of background information. For the background information we chose four variables that we believe are easy to obtain for an intruder from public records, namely the (approximate) size of the establishment, i.e. its (approximate) total number of employees, the German Federal State the establishment is located in, its legal form and its industrial sector (recorded in 40 categories). We evaluate the increase in risk if these variables are added successively to the strategic dummy. The results are summarized in the Table 3. Not surprisingly the same percentage of records as in Ta-

Table 3: *Disclosure risk evaluations using the strategic dummy approach.*

quantile	N	indicators $\Im_k$	prop. uniqu. identified	Δ all records	Δ uniqu. identified
all	12814	exact size	0.034	13801.825	0
		approx. size	0.0009	11025.450	0
		+ federal state	0.023	11739.574	0
		+ legal form	0.116	13633.345	0
		+ branch	0.658	1.478	0
0.5	6516	exact size	0.066	26985.871	0
		approx. size	0.002	21526.008	0
		+ federal state	0.046	21945.190	0
		+ legal form	0.200	26774.728	0
		+ branch	0.846	0.323	0
0.75	3217	exact size	0.134	52023.417	0
		approx. size	0.003	40965.983	0
		+ federal state	0.085	39651.427	0
		+ legal form	0.228	48390.483	0
		+ branch	0.868	0.147	0
0.9	1282	exact size	0.335	1.956	0
		approx. size	0.009	4.296	0
		+ federal state	0.186	1.944	0
		+ legal form	0.352	1.499	0
		+ branch	0.895	0.070	0
0.99	129	exact size	0.969	0.011	0
		approx. size	0.085	1.311	0
		+ federal state	0.682	0.136	0
		+ legal form	0.806	0.055	0
		+ branch	0.953	0.021	0

ble 1 are identified, if the exact establishment size is used as a dummy. Relaxing the unrealistic assumption of exactly knowing the size of the establishment we use an indicator for the approximate total number of employees that identifies all records that lie within $\pm 2.5\%$ of the reported establishment size. This information alone almost never uniquely identifies a record in the data set. Even for the top 0.1% of establishments only 31% are uniquely identified. However, adding more information significantly increases the risk. When all four background variables are used, more than 65% of the establishments are identified uniquely in the entire data set. Since arguably intruders will only be interested in the larger establishments and not in small family businesses, the fact that almost 90% of the records can be uniquely identified for the largest 10% of the establishments based on very little background information is an alarming result. Again, we only include the results in the last column of the table to emphasize that once a record is uniquely identified all information in the data set for that record can be revealed easily without access to the actual microdata.

This leads to the question how the intruder will know that she has indeed uniquely identified the mth respondent. Of course, the natural way would be to check the residuals of the regression for zeroes. However, residuals usually are not reported in remote access. Alternatively, for the dummy variable approach the intruder could check the mean of the generated dummy variable which should be $1/n$ in case of unique identification. If the agency decides to suppress means for binary variables with few positive (or negative) outcomes, the intruder could compute the variance of the dummy variable. Given a unique identification it should be equal to $Var(\Im) = 1/n + 1/n^2$. Both approaches are of course not possible when generating an artificial outlier since z would just be a new continuous variable with unknown mean and variance. In this case, the intruder might check, if a unique maximum exists for z. Only if the maximum is unique, a single record has been identified. However, such requests will likely be suppressed by the remote server. This can be seen as an additional argument in favor of the strategic dummy approach.

5. Conclusion

It is obvious that agencies – once they are aware of the risks described in the previous sections – can easily prevent this type of disclosure, e.g. by prohibiting regressions that contain dichotomous regressors with less than say 3 positive outcomes or by allowing only certain transformations for the variables. But it is important that the agency must be aware of the problem to prevent it. The point that we are trying to make is that there are many constellations that might lead to a risk of disclosure. Some are obvious whereas others are more difficult to detect in advance. Full remote access without any intervention of the agency would require that all possible constellations are considered and ruled out before data access is provided. The risk from linear regressions that is the main topic of this paper is only one example of a disclosure risk that might not

be obvious at first glance. We believe there are many other situations that might be equally harmful. For example it is well known that saturated models can reveal the exact information for small cell table entries that would have been protected by cell suppression or any other statistical disclosure limitation technique if the table would have been requested directly. We believe that more research in the area is needed to detect other user queries that might impose a risk of disclosure. Whether it will be possible to rule out all potential disclosure risks in advance remains an open question.

Acknowledgments

This research was supported by a grant from the German Federal Ministry of Education and Research.

References

Drechsler, J. (2010). Multiple imputation in practice – a case study using a complex German establishment survey. *Advances in Statistical Analysis (online first)*.

Fischer, G., Janik, F., Müller, D. and Schmucker, A. (2008). The IAB Establishment Panel – from sample to survey to projection. Technical report, FDZ-Methodenreport, No. 1.

Gomatam, S., Karr, A. F., Reiter, J. P. and Sanil, A. P. (2005). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers. *Statistical Science*, 20, 163–177.

Hoaglin, D. and Welsh, R. (1978). The hat matrix in regression and anova. *The American Statistician*, 32, 17–22.

Kölling, A. (2000). The IAB-Establishment Panel. *Journal of Applied Social Science Studies*, 120, 291–300.

Ronning, G., Bleninger, P., Drechsler, J. and Gürke, C. (2010). Remote Access - Eine Welt ohne Mikrodaten? (in German). *IAW Discussion Papers*, 66.

A. Artificial outliers and strategic dummies

We consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{6}$$

where $\mathbf{y}$ and $\mathbf{u}$ are n-dimensional vectors, $\boldsymbol{\beta}$ is a K-dimensional vector and $\mathbf{X}$ a $(n \times K)$ matrix with $\boldsymbol{\iota}' = (1, 1, \ldots, 1)$ as the first column. The vector of predicted values is given by

$$\hat{\mathbf{y}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y} \tag{7}$$

where **H**, called the hat matrix, measures the "leverage" of a certain regressor (see, e.g. Hoaglin and Welsh (1978)).

A.1 Artificial outliers

In the following we assume that the observations are ordered such that observations for survey respondent m are in the first row of the data matrix. Therefore z_1 contains the artificial outlier which tends towards infinity; compare the definition (1) of artificial outliers in the main text.

Unique identification

In the special case of a simple regression ($K = 2$) with

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{\iota} & \mathbf{z} \end{pmatrix}$$

the elements of the hat matrix are given by

$$h_{jk} = \frac{1}{n\sum z_i^2 - \left(\sum z_i\right)^2}\left(\sum_{i=1}^n z_i^2 - z_j\sum_{i=1}^n z_i - z_k\sum_{i=1}^n z_i + nz_jz_k\right),$$

with $j = 1,\ldots,n$ and $k = 1,\ldots,n$. Therefore the jth element of the vector of predicted values $\hat{\mathbf{y}}$ is given by

$$\hat{y}_j = \sum_{k=1}^n h_{jk}y_k = \frac{1}{n\sum z_i^2 - \left(\sum z_i\right)^2}\sum_{k=1}^n\left(\sum_{i=1}^n z_i^2 - z_j\sum_{i=1}^n z_i - z_k\sum_{i=1}^n z_i + nz_jz_k\right)y_k$$

and in particular for $j = 1$ we have

$$\begin{aligned}
\hat{y}_1 &= \frac{1}{n\sum z_i^2 - \left(\sum z_i\right)^2}\sum_{k=1}^n\left[\sum_{i=1}^n z_i^2 - z_1\sum_{i=1}^n z_i - z_k\sum_{i=1}^n z_i + nz_1z_k\right]y_k \\
&= \frac{1}{n\sum z_i^2 - \left(\sum z_i\right)^2}\left[\left(\sum_{i=1}^n z_i^2 - z_1\sum_{i=1}^n z_i\right)\sum_{k=1}^n y_k - \left(\sum_{i=1}^n z_i - nz_1\right)\sum_{k=1}^n z_ky_k\right] \\
&= \frac{\left(\sum_{i=1}^n z_i^2 - z_1\sum_{i=1}^n z_i\right)\sum_{k=1}^n y_k}{n\sum z_i^2 - \left(\sum z_i\right)^2} - \frac{\left(\sum_{i=1}^n z_i - nz_1\right)\sum_{k=1}^n z_ky_k}{n\sum z_i^2 - \left(\sum z_i\right)^2} \\
&= \frac{\left(z_1^2 + \sum_{i>1} z_i^2 - z_1\left(z_1 + \sum_{i>1} z_i\right)\right)\sum_{k=1}^n y_k}{n\left(z_1^2 + \sum_{i>1} z_i^2\right) - \left(z_1 + \sum_{i>1} z_i\right)^2} - \frac{\left(z_1 + \sum_{i>1} z_i - nz_1\right)\left(z_1y_1 + \sum_{k>1} z_ky_k\right)}{n\left(z_1^2 + \sum_{i>1} z_i^2\right) - \left(z_1 + \sum_{i>1} z_i\right)^2} \\
&= A - B.
\end{aligned}$$

In order to obtain results for $z_1 \to \infty$ we write the two terms as follows:

$$A = \frac{\left[\left(1+\frac{\sum_{i>1} z_i^2}{z_1^2}\right) - \left(1+\frac{\sum_{i>1} z_i}{z_1}\right)\right] \sum_{k=1}^{n} y_k}{n\left(1+\frac{\sum_{i>1} z_i^2}{z_1^2}\right) - \left(1+\frac{\sum_{i>1} z_i}{z_1}\right)^2}$$

and

$$B = \frac{\left(1+\frac{\sum_{i>1} z_i}{z_1} - n\right)\left(y_1 + \frac{\sum_{k>1} z_k y_k}{z_1}\right)}{n\left(1+\frac{\sum_{i>1} z_i^2}{z_1^2}\right) - \left(1+\frac{\sum_{i>1} z_i}{z_1}\right)^2}$$

from which we obtain

$$\lim_{z_1\to\infty} \hat{y}_1 \;=\; \lim_{z_1\to\infty} (A-B) \;=\; \frac{0}{n-1} - \frac{(1-n)\,y_1}{n-1} \;=\; y_1. \tag{8}$$

Therefore for a sufficiently large z_1 we can approximate y_1 by its predicted value $\hat{y}_1$.

Non-unique identification

To this point we assumed that the target is uniquely identified by the background information resp. the transformed outlier generating variable (see (1) in the main text). Now consider the case where more than a single subject is identified by x_m resp. z. In this case the matrix containing the outlier is given by

$$\mathbf{X}_2 = \begin{pmatrix} z_1 \boldsymbol{\iota}_q \\ \mathbf{z}_2 \end{pmatrix}.$$

We assume q subjects are identified, i.e. have the exact same value for the background variable as the target record. These q subjects are transformed to artificial outliers. Without loss of generality let them be the first q observations in the dataset. $\boldsymbol{\iota}_q$ is a q-vector of ones and $\mathbf{0}$ a $(n-q)$-vector of zeros so that

$$\mathbf{X}_2 \left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2' = \frac{1}{q z_1^2 + \sum_{i>q} z_i^2} \begin{pmatrix} z_1^2 \boldsymbol{\iota}_q \boldsymbol{\iota}_q' & z_1 \boldsymbol{\iota}_q \mathbf{z}_2' \\ z_1 \mathbf{z}_2 \boldsymbol{\iota}_q' & \mathbf{z}_2 \mathbf{z}_2' \end{pmatrix}$$

and

$$\mathbf{X}_2 \left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2' \left(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right) = \frac{1}{q z_1^2 + \sum_{i>q} z_i^2} \begin{pmatrix} z_1^2 \boldsymbol{\iota}_q \boldsymbol{\iota}_q' & z_1 \boldsymbol{\iota}_q \mathbf{z}_2' \\ z_1 \mathbf{z}_2 \boldsymbol{\iota}_q' & \mathbf{z}_2 \mathbf{z}_2' \end{pmatrix} \left(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right).$$

The predicted values are given by

$$\hat{\mathbf{y}} = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \frac{1}{q z_1^2 + \sum_{i>q} z_i^2} \begin{pmatrix} z_1^2 \boldsymbol{\iota}_q \boldsymbol{\iota}_q' & z_1 \boldsymbol{\iota}_q \mathbf{z}_2' \\ z_1 \mathbf{z}_2 \boldsymbol{\iota}_q' & \mathbf{z}_2 \mathbf{z}_2' \end{pmatrix} \left(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right)$$

respectively

$$\begin{pmatrix} \hat{\mathbf{y}}_q \\ \hat{\mathbf{y}}_{n-q} \end{pmatrix} =$$

$$= \begin{pmatrix} \mathbf{X}_{1q} \\ \mathbf{X}_{1,n-q} \end{pmatrix} \hat{\boldsymbol{\beta}}_1 + \frac{1}{q z_1^2 + \sum_{i>q} z_i^2} \begin{pmatrix} z_1^2 \boldsymbol{\iota}_q \boldsymbol{\iota}_q' & z_1 \boldsymbol{\iota}_q \mathbf{z}_2' \\ z_1 \mathbf{z}_2 \boldsymbol{\iota}_q' & \mathbf{z}_2 \mathbf{z}_2' \end{pmatrix} \left\{ \begin{pmatrix} \mathbf{y}_q \\ \mathbf{y}_{n-q} \end{pmatrix} - \begin{pmatrix} \mathbf{X}_{1q} \\ \mathbf{X}_{1,n-q} \end{pmatrix} \hat{\boldsymbol{\beta}}_1 \right\}$$

For the first q elements of the vector $\hat{\mathbf{y}}$ of predicted values we get

$$\begin{aligned} \hat{\mathbf{y}}_q &= \mathbf{X}_{1q} \hat{\boldsymbol{\beta}}_1 + \frac{1}{q z_1^2 + \sum_{i>q} z_i^2} \begin{pmatrix} z_1^2 \boldsymbol{\iota}_q \boldsymbol{\iota}_q' & z_1 \boldsymbol{\iota}_q \mathbf{z}_2' \end{pmatrix} \left\{ \begin{pmatrix} \mathbf{y}_q \\ \mathbf{y}_{n-q} \end{pmatrix} - \begin{pmatrix} \mathbf{X}_{1q} \\ \mathbf{X}_{1,n-q} \end{pmatrix} \hat{\boldsymbol{\beta}}_1 \right\} \\ &= \mathbf{X}_{1q} \hat{\boldsymbol{\beta}}_1 \\ &+ \frac{1}{q z_1^2 + \sum_{i>q} z_i^2} z_1^2 \boldsymbol{\iota}_q \boldsymbol{\iota}_q' (\mathbf{y}_q - \mathbf{X}_{1q} \hat{\boldsymbol{\beta}}_1) \\ &+ \frac{1}{q z_1^2 + \sum_{i>q} z_i^2} z_1 \boldsymbol{\iota}_q \mathbf{z}_2' (\mathbf{y}_{n-q} - \mathbf{X}_{1,n-q} \hat{\boldsymbol{\beta}}_1) \\ &= \mathbf{X}_{1q} \hat{\boldsymbol{\beta}}_1 \\ &+ \frac{1}{q + \frac{\sum_{i>q} z_i^2}{z_1^2}} \boldsymbol{\iota}_q \boldsymbol{\iota}_q' (\mathbf{y}_q - \mathbf{X}_{1q} \hat{\boldsymbol{\beta}}_1) \\ &+ \frac{1}{q + \frac{\sum_{i>q} z_i^2}{z_1^2}} \frac{1}{z_1} \boldsymbol{\iota}_q \mathbf{z}_2' (\mathbf{y}_{n-q} - \mathbf{X}_{1,n-q} \hat{\boldsymbol{\beta}}_1) \end{aligned} \tag{9}$$

If z_1 becomes infinitely large the limit of the predicted values is

$$\lim_{z_1 \to \infty} \hat{\mathbf{y}}_q = \frac{1}{q} \boldsymbol{\iota}_q \boldsymbol{\iota}_q' \, \mathbf{y}_q + \mathbf{X}_{1q} \hat{\boldsymbol{\beta}}_1 - \frac{1}{q} \boldsymbol{\iota}_q \boldsymbol{\iota}_q' \mathbf{X}_{1q} \hat{\boldsymbol{\beta}}_1 \tag{10}$$

resulting in

$$\hat{y}_i = \bar{y}_q + \begin{pmatrix} 1, & x_{i2} - \bar{x}_q^{(2)}, & \ldots, & x_{iK} - \bar{x}_q^{(K)} \end{pmatrix} \hat{\boldsymbol{\beta}}_1 \, , i = 1, 2, \ldots, q. \tag{11}$$

Here we use

$$\bar{y}_q = \frac{1}{q} \sum_{i=1}^{q} y_i \quad \text{and} \quad \bar{x}_q^{(k)} = \frac{1}{q} \sum_{i=1}^{q} x_{ik} \, , k = 2, \ldots, K \, .$$

Both (10) and (11) show that

- if $q = 1$, i.e. unique identification, the result reduces to (8) because $\bar{y}_q = y_1$ and $\bar{x}_q^{(k)} = x_{1k}$ for all regressors.
- If only the artificial outlier generating z is used in a simple linear regression it holds that

$$\lim_{z_1 \to \infty} \hat{y}_i \quad = \quad \bar{y}_q \,, \quad i = 1, \ldots, q,$$

  for all q subjects selected.
- In general however, under non-unique identification no clear-cut statement regarding the difference between $\hat{y}_i$ and y_i , $i = 1, 2, \ldots, q$, can be made.

A.2 Strategic dummy variables

Simple regression

In case of unique identification by (2), (3) or (4) in the main text the regressor matrix is given by

$$\mathbf{X} = \left(\begin{array}{cc} \boldsymbol{\iota} & \mathbf{e}_1 \end{array} \right),$$

where $\mathbf{e}_1$ is an n-dimensional vector with 1 as the first element and 0 for the remaining $n-1$ elements. Therefore

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{(n-1)} \left(\begin{array}{cc} 1 & -1 \\ -1 & n \end{array} \right)$$

and

$$\mathbf{H} = \frac{1}{n-1} \left(\begin{array}{cc} n-1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\iota}_{n-1}\boldsymbol{\iota}_{n-1}' \end{array} \right),$$

where $\mathbf{0}$ is the $(n-1)$-dimensional null vector and $\boldsymbol{\iota}_{n-1}$ a $(n-1)$-dimensional vector of ones. Note that $h_{11} = 1$ and $h_{1j} = 0$, $j > 1$, so that the predicted value for y_1 is given by

$$\hat{y}_1 = \sum_{k=1}^{n} h_{1k} y_k = \frac{1}{n-1} \left((n-1) y_1 + \sum_{k>1} 0 \cdot y_k \right) = y_1.$$

The case of additional regressors

We now consider the case that other regressors are added to the regression which might be motivated by the idea that the use of a strategic dummy is not so easily detected by the agency if other regressors are also included in the model. We write the model in partitioned form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} = \begin{pmatrix} \mathbf{X}_1 \, \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \mathbf{u} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}.$$

with

$$\mathbf{X}_2 = \mathbf{e}_1$$

so that this submatrix contains only the information regarding the strategic dummy. Then the vector of predicted values can be written as

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2\hat{\boldsymbol{\beta}}_2 \\ &= \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\left(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right) \end{aligned} \tag{12}$$

Since

$$\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2' = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

and

$$\begin{aligned} \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\left(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right) &= \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\left(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right) \\ &= \begin{pmatrix} y_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} 1 & x_{12} & \dots & x_{1K} \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix}\hat{\boldsymbol{\beta}}_1, \end{aligned}$$

we obtain for the vector of predicted values in (12):

$$\hat{\mathbf{y}} = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \begin{pmatrix} y_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} 1 & x_{12} & \dots & x_{1K} \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix}\hat{\boldsymbol{\beta}}_1 \tag{13}$$

and in particular for the first element we get

$$\hat{y}_1 = \begin{pmatrix} 1 & x_{12} & \dots & x_{1K} \end{pmatrix} \hat{\boldsymbol{\beta}}_1 + y_1 - \begin{pmatrix} 1 & x_{12} & \dots & x_{1K} \end{pmatrix} \hat{\boldsymbol{\beta}}_1 = y_1. \qquad (14)$$

Non-unique identification

The empirical example in Section 4 shows that y_m and $\hat{y}_m$ may differ substantially if more than one respondent is identified using the background information available for x_m. In this section we evaluate the fitted value $\hat{y}_m$ in this case.

If more than one respondent is picked by the strategic dummy the submatrix $\mathbf{X}_2$ (which actually is a vector) has the form

$$\mathbf{X}_2 = \begin{pmatrix} \boldsymbol{\iota}_q \\ \mathbf{0} \end{pmatrix}$$

where we assume that q units in the data set have the same reported value for the available background information as the target record x_m and that they are placed in the first q rows of the data matrix. $\boldsymbol{\iota}_q$ is a vector of ones and $\mathbf{0}$ denotes a $n-q$ dimensional vector of zeroes. Moreover, we have

$$\mathbf{X}_2 (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2' = \frac{1}{q} \begin{pmatrix} \boldsymbol{\iota}_q\boldsymbol{\iota}_q' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

and

$$\mathbf{X}_2 (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2' \left(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right) = \frac{1}{q} \begin{pmatrix} \boldsymbol{\iota}_q\boldsymbol{\iota}_q' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \left(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right)$$

$$= \begin{pmatrix} \bar{y}_q \\ \bar{y}_q \\ \vdots \\ \bar{y}_q \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} 1 & \bar{x}_q^{(2)} & \dots & \bar{x}_q^{(K)} \\ 1 & \bar{x}_q^{(2)} & \dots & \bar{x}_q^{(K)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \bar{x}_q^{(2)} & \dots & \bar{x}_q^{(K)} \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix} \hat{\boldsymbol{\beta}}_1,$$

where we use

$$\bar{y}_q = \frac{1}{q}\sum_{i=1}^{q} y_i \quad \text{and} \quad \bar{x}_q^{(k)} = \frac{1}{q}\sum_{i=1}^{q} x_{ik}\,, k = 2,\dots,K.$$

Comparing this with (9) we note that for the first q elements of the vector $\hat{\mathbf{y}}$ we obtain

$$\hat{y}_i = \bar{y}_q + \left(1, \quad x_{i2} - \bar{x}_q^{(2)}, \quad \ldots, \quad x_{iK} - \bar{x}_q^{(K)} \right) \hat{\boldsymbol{\beta}}_1 \, , i = 1, 2, \ldots, q. \tag{15}$$

which implies the following: (i) If $q = 1$ and therefore a single unit is identified, the above result is equivalent with (14) because then $\bar{y}_q = y_1$ and for all regressors $\bar{x}_q^{(k)} = x_{1k}$. (ii) If the strategic dummy is used as a single regressor then for all q units

$$\hat{y}_i = \bar{y}_q$$

holds, that is, the estimated value of y equals the arithmetic mean of all q units. (iii) If more regressors are added to the model, no clear-cut statement regarding the difference between y_m and $\hat{y}_m$ can be made.