IAW DISCUSSION PAPERS

iaw@iaw.edu | www.iaw.edu

IAW Diskussionspapiere | Nr. 72 Juni 2011 | No. 72 June 2011

Disclosure Risk from Interactions and Saturated Models in Remote Access

Gerd Ronning

Institut für Angewandte Wirtschaftsforschung e.V.
Ob dem Himmelreich 1 | 72074 Tübingen | Germany
Tel.: +49 7071 98960 | Fax: +49 7071 989699

ISSN: 1617-5654

IAW-Diskussionspapiere

Das Institut für Angewandte Wirtschaftsforschung (IAW) Tübingen ist ein unabhängiges außeruniversitäres Forschungsinstitut, das am 17. Juli 1957 auf Initiative von Professor Dr. Hans Peter gegründet wurde. Es hat die Aufgabe, Forschungsergebnisse aus dem Gebiet der Wirtschafts- und Sozialwissenschaften auf Fragen der Wirtschaft anzuwenden. Die Tätigkeit des Instituts konzentriert sich auf empirische Wirtschaftsforschung und Politikberatung.

Dieses IAW-Diskussionspapier können Sie auch von unserer IAW-Homepage als pdf-Datei herunterladen:

http://www.iaw.edu/Publikationen/IAW-Diskussionspapiere

ISSN 1617-5654

Weitere Publikationen des IAW:

- IAW-News (erscheinen 4x jährlich)
- IAW-Forschungsberichte

Möchten Sie regelmäßig eine unserer Publikationen erhalten, dann wenden Sie sich bitte an uns:

IAW Tübingen, Ob dem Himmelreich 1, 72074 Tübingen,
Telefon 07071 / 98 96-0
Fax 07071 / 98 96-99
E-Mail: iaw@iaw.edu

Aktuelle Informationen finden Sie auch im Internet unter:

http://www.iaw.edu

Disclosure Risk From Interactions and Saturated Models in Remote Access

GERD RONNING[1]

(This version: June 20, 2011

Abstract

Empirical research using micro data via remote access has been advocated in recent time by statistical offices since confidentiality is easier warranted for this approach. However, disclosure of single values and units cannot be completely avoided. Binary regressors (dummy variables) bear a high risk of disclosure, especially if their interactions are considered as it is done by definition in saturated models. However, contrary to views expressed in earlier publications the risk is only existing if besides parameter estimates also predicted values are reported to the researcher. The paper considers saturated specifications of the most popular linear and nonlinear microeconometric models and shows that in all cases the disclosure risk is high if some design points are represented by a (very) small number of observations. For two of the models not belonging to the exponential family (probit model and negative binomial regression model) we show that the same estimates of the conditional expectations arise here although the parameter estimates are defined by a modified equation. In the last section we draw attention to the fact that interaction of binary regressors can be used to construct "strategic dummy variables" which lead to hight disclosure risk as shown, for example, in Bleninger et al. (2010) for the linear model. In this paper we extend the analysis to the set of established nonlinear models, in particular logit, probit and count data models.

Keywords: logit model , probit model , poisson regression , negative binomial regression model , strategic dummy variable , tabular data.

[1]Wirtschaftswissenschaftliche Fakultät, Universität Tübingen, D-72074 Tübingen. email: gerd.ronning@uni-tuebingen.de. Research is related to project "Eine informationelle Infrastruktur für das ‚E-Science Age' Auf dem Weg zum ‚Remote-Access' – Verbesserung der kontrollierten Datenfernverarbeitung bei wirtschaftsstatistischen Daten durch Datenstrukturfiles und automatisierte Ergebniskontrolle" which is financially supported by the German Ministry of Education and Research.

Contents

1 Introduction

Many national official statistical agencies nowadays consider remote access to micro data as the best compromise in balancing needs for confidentiality against the users' demand for original data. The fact that there is also a disclosure risk from this way of providing the data, has been pointed out by many authors. See, for example, Gomatam et al. (2005) and O'Keefe and Good (2009) who consider various scenarios where the output sent back to the user may contain confidential details. By far the best known example is the simultaneous release of predicted values and residuals from which the values of the dependent variable can be obtained by mere addition. Another example is the exploitation of the knowledge regarding a single observational unit from which the value of any other variable for this unit can be obtained by a "strategic dummy variable" or an "artificial outlier". See Gomatam et al. (2005) and also Bleninger et al. (2011).

Reznek (2003) seems to be the first who draw attention to the disclosure risk of including interactions of binary variable into the set of explanatory variables. He also pointed out that even without interactions a disclosure risk exists if for some binary explanatory variable the number of cases is small for one or both categories since the regression coefficients provide the conditional arithmetic means of the dependent variable: The estimate of the corresponding effect equals the value of the dependent variable if this estimate is based on a single observation. Therefore O'Keefe and Good (2009), p. 1178 suggest with regard to output control under the heading 'Restricted Access': " Do not return any results if the model has: (a) An explanatory factor variable with a level with few values. (b) Interactions between factors with few values in their levels." Furthermore under the heading 'Restricted Analyses' they postulate that " at most two-way interactions between variables can be included. (a) Factor interactions with a small number of values in any cell are not permitted."

Reznek (2003), pp. 3446 - 3448, has also considered "models involving binary dependent variables", e.g. logit and probit models, and showed by means of numerical examples that the same kind of disclosure risk may exist in case of binary regressors (dummy variables). He also points out that the probit model seems to behave differently with regard to the "recovery of frequency cross-classifications" (Reznek (2003), p. 3446) . Without making an explicit statement, by this remark he has drawn attention to the fact that binary regressors may also imply a *disclosure risk for confidential tables*. Ronning et al. (2010), section 3.4 have given a formal analysis for both linear and nonlinear models which will be generalized in this paper.

The use of interactions is well motivated from a statistical point of view; a separate interpretation of main effects when interaction effects are significant is meaningless . See, for example, Fahrmeir et al. (1996), chapter 5.1 . Therefore, the general exclusion of interactions in statistical models would not be acceptable for users of micro data. On the other hand, the recommendations

cited above draw attention to interactions of order greater than two and point in particular to cases where such interactions identify single observations. In fact, already interactions of low order may lead to single observations especially for small sample size. The corresponding column in the regressor matrix has only a single "1" and zeroes elsewhere.

Since the user is not allowed to see the micro data, disclosure risk only arises if "the intruder" starts to look for such information.[2] The natural way would be to check the residuals of the regression for zeroes. However, residuals usually are not reported in remote access. Alternatively, he 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 $1/n$, too.[3]

In this paper we want to show that interactions allow the retrieval of tabular information if predicted values are provided by the server and certain manipulations are not suppressed. This would be of special concern if some cell contains only a single observation. By doing so we also want to make clear that the specification of interactions per se is not a disclosure risk.

In the next section we present an empirical example which will be used to illustrate the theoretical results. Section 3 contains the formal results restricting however the analysis to the case of only two binary regressors. Modifications for the general case of an arbitrary number of binary regressors and the construction of "strategic dummy variables" from their interactions are discussed in section 4. A short summary is given in section 5.

2 An example

Let us assume that the statistical office has micro data for $n = 10$ enterprises. For each enterprise the variables sales (million Euro) (SAL), employment (EMP) and the existence/nonexistence of a works council (WOC) are available from the data set. Moreover, regional information ("north (N)" or "south (S)") and the retail status ("retailer (R)" or "wholesaler (W)") are given. The matrix of micro data may be as follows:

[2] For the following see Bleninger et al. (2011).

[3] This result assumes that the empirical variance is computed from $s^2 = \frac{1}{n-1}\sum_i (x_i - \overline{x})^2$. Otherwise we obtain $\frac{n-1}{n} \cdot \frac{1}{n} = \frac{n-1}{n^2}$.

$$\mathbf{A} = \left(\begin{array}{ccc|cc|cc} SAL & EMP & WOC & \text{retail status} & & \text{region} & \\ & & & W & R & N & S \\ \hline 1 & 31 & 0 & 1 & 0 & 1 & 0 \\ 2 & 22 & 1 & 1 & 0 & 1 & 0 \\ 3 & 73 & 1 & 0 & 1 & 1 & 0 \\ 4 & 24 & 0 & 0 & 1 & 1 & 0 \\ 17 & 17 & 0 & 1 & 0 & 0 & 1 \\ 5 & 35 & 0 & 1 & 0 & 0 & 1 \\ 8 & 18 & 1 & 1 & 0 & 0 & 1 \\ 7 & 97 & 0 & 0 & 1 & 0 & 1 \\ 12 & 124 & 1 & 0 & 1 & 0 & 1 \\ 6 & 67 & 0 & 1 & 0 & 0 & 1 \end{array}\right)$$

For example, the first observation refers to an wholesale enterprise with headquarters in the north. Its sales amount to 1 million Euro; it has 31 employees and there is no works council in this enterprise.

Table 2.1: Number of enterprises by region and retail status

	W	R	$\sum$
N	2	2	4
S	4	2	6
$\sum$	6	4	10

From this matrix we obtain the frequencies of enterprises in a certain cell. See table 2.1. For example, there are 2 wholesale enterprises in the northern region. Please note that all cells contain more than one observation.

Table 2.2: Sales by region and retail status

	W	R	$\sum$
N	3	7	10
S	36	19	55
$\sum$	39	26	65

If we cross-classify the three other variables sales, employment and existence of works council by region and retail status, we obtain tables which we now will describe in detail. Consider first Table 2.2 which is called "magnitude table". It reports aggregated sales by region and retail status. In section 3.2.2 we will show that such table can be estimated by means of a saturated linear model (analysis of variance).

Table 2.3 considers the discrete variable employment which is cross-classified by region and retail status. Note that in this table we consider count data which later on are analyzed by count data models. Such models can be used

Table 2.3: Employment by region and retail status

	W	R	$\sum$
N	53	137	190
S	73	185	258
$\sum$	126	322	448

to estimate the entries of this table when the binary regressors region and retail status are used. See sections 3.2.3 and 3.3.2 below.

Table 2.4: Existence of works council by region and retail status

	W	R	$\sum$
N	1 (2)	1 (2)	2 (4)
S	1 (4)	1 (2)	2 (6)
$\sum$	2 (6)	2 (4)	4 (10)

Number of enterprises in parentheses.
See table 2.1.

Finally, from table 2.4 one can obtain the information regarding the binary variable 'existence of works council'. Each cell reports the frequency of enterprises where a council has been established. This is a frequency table which is based on a binary variable.[4] In sections 3.2.4 and 3.3.3 below we will use logit and probit models to estimate the entries of this table.

3 Models with two binary regressors

In the first subsection we consider a simple example which uses the data from subsection 2 above and then give a general formulation of the the result in the subsections to follow.

3.1 Notation and terminology

In this section we consider the effect of the two binary regressors 'retail status' (denoted by d_1) and 'region' (denoted by d_2) on the three variables (sales, employment and existence of works council) which will be denoted in all three cases by y. In case of the continuous variable sales we use the analysis of variance, in case of the discrete variable employment count data models are

[4] In practice relative frequencies will be preferred which could be deduced from the figures in parentheses.

appropriate and in case of the binary variable existence of works council we consider logit and probit models.

Of course, we could add another binary regressor by defining the interaction term

$$d_{12,i} = d_{1i} \cdot d_{2i} \quad .$$

For the case of two binary regressors this would imply the specification of a saturated model.[5] The resulting regressor matrix $\mathbf{X}$ is given by

$$\mathbf{X} \quad = \quad \left(\begin{array}{cccc} \text{const} & d_1 = W & d_2 = N & d_{12} = d_1 \cdot d_2 \\ \hline 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{array} \right) \quad .$$

Please note that the columns of this matrix act as summation operators. For example, in case of the vector of sales from the data matrix $\mathbf{A}$ (see section 2)

$$\mathbf{y}' = \left(\begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 17 & 5 & 8 & 7 & 12 & 6 \end{array} \right)$$

we obtain

$$\mathbf{X}'\mathbf{y} \quad = \quad \left(\begin{array}{c} 1+2+3+4+17+5+8+7+12+6 \\ 1+2+17+5++8+6 \\ 1+2+3+4 \\ 1+2 \end{array} \right) = \left(\begin{array}{c} 65 \\ 39 \\ 10 \\ 3 \end{array} \right)$$

$$= \quad \left(\begin{array}{c} \sum_{i=1}^{n} y_i \\ \sum_{i \epsilon W} y_i \\ \sum_{i \epsilon N} y_i \\ \sum_{i \epsilon} y_i \end{array} \right) = \left(\begin{array}{c} \text{total sales} \\ \text{sales in wholesale sector} \\ \text{sales in northern region} \\ \text{sales of wholesale sector in northern region} \end{array} \right) .$$

If we compare this vector with table 2.2, we see that the four figures are sufficient to reproduce this table! Clearly the last column of $\mathbf{X}$ representing the interaction term is responsible for identifying not only the margins but also the four inner cells. Corresponding statements apply to the other two data vectors concerning employment and existence of works councils.

We definine the conditional expectation of y by

$$\theta_i \quad \equiv \quad E[y_i \,|\mathbf{x}_i, \boldsymbol{\beta}] \tag{3-1}$$

[5] For a more general definition see section 4.1.

for observation i where $\mathbf{x}_i$ is the i-th row from the matrix $\mathbf{X}$ and $\boldsymbol{\beta}$ denotes the vector of coefficients. In the following we consider an estimate of θ_i which we denote by

$$\hat{\theta}_i = E[y_i \,|\mathbf{x}_i, \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}] \qquad .$$

In case of linear models we usually write $\hat{y}_i$ for this estimate; however in order to include also the case of discrete or binary y's we prefer the more general notation of (3-1).

3.2 Models satisfying $\mathbf{X}'\hat{\boldsymbol{\theta}} = \mathbf{X}'\mathbf{y}$

We now consider models for which manipulation of predicted values of the kind $\mathbf{X}'\hat{\boldsymbol{\theta}}$ leads to disclosure risk since entries of the table can be read from this expression.

3.2.1 A general remark

Let us first consider implications of this result. Please note that in case of our simple example with regressor matrix $\mathbf{X}$ from section 3.1 the (set of) equation(s)

$$\mathbf{X}'\hat{\boldsymbol{\theta}} = \mathbf{X}'\mathbf{y} \tag{3-2}$$

can be written as follows:

$$\begin{aligned}
\sum_{i\epsilon W\cap N} y_i + \sum_{i\epsilon \overline{W}\cap N} y_i + \sum_{i\epsilon W\cap \overline{N}} y_i + \sum_{i\epsilon \overline{W}\cap \overline{N}} y_i &= \sum_{i\epsilon W\cap N} \hat{\theta}_i + \sum_{i\epsilon \overline{W}\cap N} \hat{\theta}_i + \sum_{i\epsilon W\cap \overline{N}} \hat{\theta}_i + \sum_{i\epsilon \overline{W}\cap \overline{N}} \hat{\theta}_i \\
\sum_{i\epsilon W\cap N} y_i + \sum_{i\epsilon \overline{W}\cap N} y_i &= \sum_{i\epsilon W\cap N} \hat{\theta}_i + \sum_{i\epsilon \overline{W}\cap N} \hat{\theta}_i \\
\sum_{i\epsilon W\cap N} y_i + \sum_{i\epsilon W\cap \overline{N}} y_i &= \sum_{i\epsilon W\cap N} \hat{\theta}_i + \sum_{i\epsilon W\cap \overline{N}} \hat{\theta}_i \\
\sum_{i\epsilon W\cap N} y_i &= \sum_{i\epsilon W\cap N} \hat{\theta}_i
\end{aligned} \tag{3-3}$$

As we see later on more clearly, estimate $\hat{\theta}_i$ for a certain cell is a constant so that, for example,

$$\hat{\theta}_i \quad = \quad \hat{\theta}_{W\cap N}\,,\, \forall i\,\epsilon\, W \cap N \qquad .$$

for the cell containing all wholesale firm from the northern region. Therefore the last equation from (3-3) can be written as

$$n_{W\cap N}\,\hat{\theta}_{W\cap N} \quad = \quad \sum_{i\epsilon W\cap N} y_i \qquad ,$$

where $n_{W\cap N}$ is the number of observations in cell $W\cap N$. In other words, the estimate of this conditional expectation is given by the corresponding conditional

arithmetic mean. With the help of the last equation we find corresponding estimates for the remaining three cells from the first three equations of (3-3). Summarizing this, we write

$$\hat{\theta}_i \quad = \quad \begin{cases} \dfrac{1}{n_{W\cap N}} \displaystyle\sum_{i\epsilon W\cap N} y_i \ , & i\,\epsilon\, W\cap N \\ \dfrac{1}{n_{\overline{W}\cap N}} \displaystyle\sum_{i\epsilon \overline{W}\cap N} y_i \ , & i\,\epsilon\, \overline{W}\cap N \\ \dfrac{1}{n_{W\cap \overline{N}}} \displaystyle\sum_{i\epsilon W\cap \overline{N}} y_i \ , & i\,\epsilon\, W\cap \overline{N} \\ \dfrac{1}{n_{\overline{W}\cap \overline{N}}} \displaystyle\sum_{i\epsilon \overline{W}\cap \overline{N}} y_i \ , & i\,\epsilon\, \overline{W}\cap \overline{N} \end{cases} \quad . \tag{3-4}$$

3.2.2 Analysis of variance

If we would employ an ANOVA to estimate the effect of retail status and region on sales, we would obtain the vector of predicted values $\hat{\mathbf{y}}$ or $\hat{\boldsymbol{\theta}}$ given by

$$\hat{\boldsymbol{\theta}} \ = \ \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \ \mathbf{X}'\mathbf{y}$$

and applying the regressor matrix to this vector we obtain

$$\mathbf{X}'\hat{\boldsymbol{\theta}} = \mathbf{X}'\,\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \ \mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{y} \qquad .$$

Note that this equation has the structure of (3-2). Therefore manipulation of predicted values $\hat{\theta}_i$ in the described manner leads to disclosure risk and should be interdicted if some figures in the table are confidential. This is of special concern if some cell contains information for a single enterprise: Then the sales figure of this enterprise would be revealed.

Please note that already the vector $\hat{\boldsymbol{\theta}}$ could be seen as a disclosure risk since each $\hat{\theta}_i$ equals the empirical mean for one of the four cells in one of the empirical tables given above. For the sales example of table 2.2 the vector of predicted values is given by

$$\hat{\boldsymbol{\theta}}' = \left(\ 1.5 \quad 1.5 \quad 3.5 \quad 3.5 \quad 9.0 \quad 9.0 \quad 9.0 \quad 9.5 \quad 9.5 \quad 9.0 \ \right)$$

or more generally by

$$\hat{\boldsymbol{\theta}}' = \left(\ \overline{y}_{W\cap N} \quad \overline{y}_{W\cap N} \quad \overline{y}_{\overline{W}\cap N} \quad \overline{y}_{\overline{W}\cap N} \quad \overline{y}_{W\cap \overline{N}} \quad \overline{y}_{W\cap \overline{N}} \quad \overline{y}_{W\cap \overline{N}} \quad \overline{y}_{\overline{W}\cap \overline{N}} \quad \overline{y}_{\overline{W}\cap \overline{N}} \quad \overline{y}_{W\cap \overline{N}} \ \right) .$$

For example, the first two elements of the vector each display the mean of the two wholesale enterprises in the northern region. Therefore total sales for this cell ($W \cap N$) would be determinded from multyplying this mean by the corresponding frequency: $n_{W\cap N} = 2$ (see table 2.1). Usually the number of firms will not be revealed; however as in the above example frequencies may be determined from counting the number of times a certain value appears. See

the above example where "1.5" appears twice. Therefore, it may be wise (and in case of single observations absolutely essential) not to disclose the predicted values towards the user in any case!

On the other hand this would imply that even the estimated coefficients cannot be provided since from these estimates the predicted values can be easily determined. In case of the ANOVA the model is given by

$$y_i \quad = \quad \mu \, + \, \beta_1 \, d_{1i} \, + \, \beta_2 \, d_{2i} \, + \, \beta_3 \, d_{12i} \, + \, \varepsilon_{ijk}$$

so that for example expected value for wholesale ($d_1 = 1$) in northern region ($d_2 = 1$) is given by

$$E[y|d_1 = 1, d_2 = 1] = \mu + \beta_1 + \beta_2 + \beta_3 \qquad ,$$

whereas in case of wholesale in the south the expected value is given by

$$E[y|d_1 = 1, d_2 = 0] = \mu + \beta_1 \qquad .$$

For the sales example above the vector of estimated coefficients is given by

$$\begin{pmatrix} \hat{\mu} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} \quad = \quad \begin{pmatrix} 9.5 \\ -0.5 \\ -6.0 \\ -1.5 \end{pmatrix}$$

so that

$$E[y|d_1 = 1, d_2 = 1] = 9.5 - 0.5 - 6.0 - 1.5 = 1.5$$

and

$$E[y|d_1 = 1, d_2 = 0] = 9.5 - 0.5 = 9.0 \quad .$$

This shows that a saturated version of an analysis of variance bears a general disclosure risk even if predicted values are not provided since they can be determined from the estimated coefficients. However, there is one important difference: The frequency of the several means is *not* accessible (since the regressor matrix is not known to the user!) and therefore the entries of the table cannot be reconstructed! The argument of Reznek (2003) that the estimated coefficients bear a risk, therefore is only valid if already the means bear a disclosure risk! This usually would not be the case if the number of observations for each cell are large enough, at least greater than 2. However, in section 4.2 we consider interactions of higher order which are constructed with the aim to identify single observations.

3.2.3 Poisson regression model

For the Poisson regression model with two interacting binary regressor the conditional expectation is given by

$$\theta(d_1, d_2) = \exp(\mu + \beta_1 \, d_1 + \beta_2 \, d_2 \, + \, \beta_3 \, d_{12}) \qquad ,$$

and the first order condition of ML estimation of the coefficients leads to[6]

$$\mathbf{X}'\mathbf{y} \quad = \quad \mathbf{X}' \begin{pmatrix} \hat{\theta}(\mathbf{x}_1;\hat{\boldsymbol{\beta}}) \\ \hat{\theta}(\mathbf{x}_2;\hat{\boldsymbol{\beta}}) \\ \vdots \\ \hat{\theta}(\mathbf{x}_n;\hat{\boldsymbol{\beta}}) \end{pmatrix} \qquad ,$$

where $\mathbf{x}_i$ denotes the i-th row of the regressor matrix $\mathbf{X}$. Note that this equation has the structure of (3-2) and therefore tells us that if we estimate the effect of retail status and region on employment, premultiplication of the vector of predicted values by the regressor matrix $\mathbf{X}$ would reveal the complete information from table 2.3. This corresponds to the results for the sales example in case of the ANOVA. And of course the remarks concerning disclosure risk of the vector of predicted values itself apply here, too. In particular, again predicted values are given by the appropriate arithmetic means. In case of the employment example above we obtain

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} 26.5 & 26.5 & 48.5 & 48.5 & 34.25 & 34.25 & 34.25 & 110.5 & 110.5 & 34.25 \end{pmatrix}$$

or more generally

$$\hat{\boldsymbol{\theta}}' = \begin{pmatrix} \overline{y}_{W\cap N} & \overline{y}_{W\cap N} & \overline{y}_{\overline{W}\cap N} & \overline{y}_{\overline{W}\cap N} & \overline{y}_{W\cap\overline{N}} & \overline{y}_{W\cap\overline{N}} & \overline{y}_{W\cap\overline{N}} & \overline{y}_{\overline{W}\cap\overline{N}} & \overline{y}_{\overline{W}\cap\overline{N}} & \overline{y}_{W\cap\overline{N}} \end{pmatrix}$$

where y now denotes employment. So clearly again the predicted values would identify observations from a single enterprise if some cell contains only information from this particular unit. And again predicted values for the different "design points" can be determined from the parameter estimates but not their frequencies from which the tabular entries could be inferred.

3.2.4 The logit model

Both (binary) logit and probit in case of two interaction binary regressors consider the conditional probabilities

$$P(Y_i = 1|d_1, d_2) = F(\mu + \beta_1\, d_1 + \beta_2\, d_2 \,+\, \beta_3\, d_{12}) \qquad ,$$

which equal the corresponding conditional expectations θ_i. In this subsection we concentrate on the logit case for which first order conditions of ML estimation results in[7]

$$\mathbf{X}'\mathbf{y} \quad = \quad \mathbf{X}' \begin{pmatrix} \hat{\theta}(\mathbf{x}_1;\hat{\boldsymbol{\beta}}) \\ \hat{\theta}(\mathbf{x}_2;\hat{\boldsymbol{\beta}}) \\ \vdots \\ \hat{\theta}(\mathbf{x}_n;\hat{\boldsymbol{\beta}}) \end{pmatrix} \qquad (3\text{-}5)$$

[6] See, for example, Greene(2000), chapter 19.9.
[7] See, for example, Greene(2000), chapter 19.4 .

with

$$\hat{\theta}(\mathbf{x}_i;\hat{\boldsymbol{\beta}}) = \frac{1}{1+\exp\left\{\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right\}} \quad .$$

Therefore remarks concerning the disclosure risk of predicted values , in this case predicted probabilities, apply here, too! For completeness we report the results of estimating the effect of retail status and region on the existence of works councils for the above example. We obtain the vector[8]

$$\hat{\theta} = \begin{pmatrix} 0.5 & 0.5 & 0.5 & 0.5 & 0.25 & 0.25 & 0.25 & 0.5 & 0.5 & 0.25 \end{pmatrix}$$

For example, the first two elements of the vector are given by the mean of the two elements of the cell "$W \cap N$" , that is $(0 + 1)/2 = 0.5$, and the number of works councils therefore is $n_{W\cap N} \cdot \overline{y}_{W\cap N} = 2 \cdot 0.5 = 1$. And for the four elements of the cell "$W \cap \overline{N}$" we obtain $n_{W\cap\overline{N}} \cdot \overline{y}_{W\cap\overline{N}} = 4 \cdot 0.25 = 1$.

We would like to add that in case of the multinomial logit model the same remarks apply, i.e. predicted probabilities can be used to retrieve tabular information since first order conditions for this model have the same structure as in (3-5) for each of the categories of the polytomous dependent variable except for the reference category.

3.3 Models satisfying $\mathbf{X'D}\hat{\theta} = \mathbf{X'Dy}$

We now turn to two other important microeconometric models for which estimates are defined by a more general equation (as given in the title of this subsection). This will be discussed further below in subsections 3.3.2 for the negative binomial regression model and in 3.3.3 for the probit model. However, we first want to show that the estimated vector $\hat{\theta}$ defined by the (set of) equation(s)

$$\mathbf{X'D}\hat{\theta} = \mathbf{X'Dy} \tag{3-6}$$

satisfies also the equation $\mathbf{X}'\hat{\theta} = \mathbf{X}'\mathbf{y}$ in (3-2) so that the same disclosure risk exists as in case of the models discussed above.

3.3.1 Proof that both equations hold

We start by writing the $(n \times r)$ regressor matrix as

$$\mathbf{X} \quad = \quad \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_{n-1}' \\ \mathbf{x}_n' \end{pmatrix}$$

[8] The small sample size of 10 does not allow much variation of predicted values. For three out of four cells the mean is 0.5.

so that each r-dimensional vector $\mathbf{x}_i', i = 1, \ldots, n,$ denotes a certain row of this matrix. Furthermore the Matrix $\mathbf{D}$ in (3-6) is diagonal and each element d_i depends on row i of the regressor matrix $\mathbf{X}$:

$$\mathbf{D} = \begin{pmatrix} d_1(\mathbf{x}_1) & & & & \\ & d_2(\mathbf{x}_2) & & & \\ & & \ddots & & \\ & & & d_{n-1}(\mathbf{x}_{n-1}) & \\ & & & & d_n(\mathbf{x}_n) \end{pmatrix} \quad .$$

In case of two binary regressors with interaction we have four distinct rows of $\mathbf{X}$ with multiplicity equalling the number of observations for this cell i.e. design point. Therefore the regressor matrix can - after reordering of rows - be written as

$$\mathbf{X} \quad = \quad \begin{pmatrix} \iota_{\mathbf{W}\cap\mathbf{N}} \otimes \mathbf{x}'_{\mathbf{W}\cap\mathbf{N}} \\ \iota_{\overline{\mathbf{W}}\cap\mathbf{N}} \otimes \mathbf{x}'_{\overline{\mathbf{W}}\cap\mathbf{N}} \\ \iota_{\mathbf{W}\cap\overline{\mathbf{N}}} \otimes \mathbf{x}'_{\mathbf{W}\cap\overline{\mathbf{N}}} \\ \iota_{\overline{\mathbf{W}}\cap\overline{\mathbf{N}}} \otimes \mathbf{x}'_{\overline{\mathbf{W}}\cap\overline{\mathbf{N}}} \end{pmatrix}$$

where $\otimes$ denotes the Kronecker product and $\iota_{\mathbf{W}\cap\mathbf{N}}$, for example, is a vector of ones of dimensionality $n_{W\cap N}$. Compare the matrix $\mathbf{X}$ in section 3.1. Moreover, the matrix $\mathbf{D}$ then may be written as

$$\mathbf{D} = \begin{pmatrix} d_{W\cap N} \otimes \mathbf{I}_{\overline{W}\cap\overline{N}} & & & \\ & d_{\overline{W}\cap N} \otimes \mathbf{I}_{\overline{W}\cap N} & & \\ & & d_{W\cap\overline{N}} \otimes \mathbf{I}_{W\cap\overline{N}} & \\ & & & d_{\overline{W}\cap\overline{N}} \otimes \mathbf{I}_{\overline{W}\cap\overline{N}} \end{pmatrix} \quad .$$

If we now proceed as in section 3.2.1, we can write (3-6) as follows:

$$\begin{aligned} \sum_{i\epsilon W\cap N} d_i y_i + \sum_{i\epsilon \overline{W}\cap N} d_i y_i + \sum_{i\epsilon W\cap\overline{N}} d_i y_i + \sum_{i\epsilon \overline{W}\cap\overline{N}} d_i y_i &= \sum_{i\epsilon W\cap N} d_i\hat{\theta}_i + \sum_{i\epsilon \overline{W}\cap N} d_i\hat{\theta}_i + \sum_{i\epsilon W\cap\overline{N}} d_i\hat{\theta}_i + \sum_{i\epsilon \overline{W}\cap\overline{N}} d_i\hat{\theta}_i \\ \sum_{i\epsilon W\cap N} d_i y_i + \sum_{i\epsilon \overline{W}\cap N} d_i y_i &= \sum_{i\epsilon W\cap N} d_i\hat{\theta}_i + \sum_{i\epsilon \overline{W}\cap N} d_i\hat{\theta}_i \\ \sum_{i\epsilon W\cap N} d_i y_i + \sum_{i\epsilon W\cap\overline{N}} d_i y_i &= \sum_{i\epsilon W\cap N} d_i\hat{\theta}_i + \sum_{i\epsilon W\cap\overline{N}} d_i\hat{\theta}_i \\ \sum_{i\epsilon W\cap N} d_i y_i &= \sum_{i\epsilon W\cap N} d_i\hat{\theta}_i \end{aligned} \tag{3-7}$$

Again we make use of the fact that the estimate $\hat{\theta}_i$ for a certain cell is a constant so that, for example,

$$\hat{\theta}_i \quad = \quad \hat{\theta}_{W\cap N}\,,\; i\,\epsilon\, W\cap N \quad .$$

Additionally we know from the matrix $\mathbf{D}$ above that d_i for a certain cell is a constant, too:

$$d_i \quad = \quad d_{W\cap N}\,,\; i\,\epsilon\, W\cap N \quad .$$

Therefore the last equation from (3-3) can be written as

$$n_{W\cap N}\, d_{W\cap N}\, \hat{\theta}_{W\cap N} \quad = \quad d_{W\cap N} \sum_{i\epsilon W\cap N} y_i \quad .$$

Since $d_{W\cap N}$ cancels out, we arrive at the same result as for the last equation from (3-6). The same is true for the three other equations above so that the estimated conditional expectations again equal the corresponding arithmetic means! In other words, the result (3-4) also applies here!

3.3.2 The NEGBIN model

Let us assume that the Poisson Parameter λ is a random variable satisfying

$$\lambda = \bar{\lambda}\,\varepsilon$$

where $\bar{\lambda}$ denotes some "average' which is deterministic and ε is a nonnegative random variable with expectation 1 so that

$$E\left(\lambda\right) = \bar{\lambda} \qquad .$$

One calls ε the heterogeneity component. If we insert the above specification into the Poisson distribution , we get the conditional distribution

$$P\left(Y = y|\,\varepsilon\right) = \frac{\exp(-\bar{\lambda}\,\varepsilon)\;(\bar{\lambda}\,\varepsilon)^{y}}{y\,!} \qquad .$$

We now assume that ε follows a gamma distribution with $E\left(\varepsilon\right) = 1$ or formally

$$f\left(\varepsilon\right) = \frac{\kappa^{\kappa}}{\Gamma\left(\kappa\right)}\,\varepsilon^{\kappa-1}\,\exp(-\kappa\,\varepsilon) \quad .$$

We derive the unconditional distribution of Y by "integrating out" the heterogeneity component:[9]

$$\begin{array}{rcl} P\left(Y = y\right) & = & \int_0^{\infty} \frac{\exp(-\bar{\lambda}\,\varepsilon)\;(\bar{\lambda}\,\varepsilon)^{y}}{y\,!}\;\frac{\kappa^{\kappa}}{\Gamma\left(\kappa\right)}\,\varepsilon^{\kappa-1}\,\exp(-\kappa\,\varepsilon)\,d\,\varepsilon \\ & = & \frac{\Gamma\left(\kappa+y\right)}{\Gamma\left(\kappa\right)\,y\,!}\,\left(\frac{\kappa}{\bar{\lambda}+\kappa}\right)^{\kappa}\,\left(\frac{\bar{\lambda}}{\bar{\lambda}+\kappa}\right)^{y} \quad . \end{array}$$

Comparing this result with the general formula of the negative binomial distribution[10] we see that

$$E[Y] = \bar{\lambda}$$

and

$$V[Y] = \bar{\lambda}\left(1 + \frac{\bar{\lambda}}{\kappa}\right) > E[Y]$$

the latter result indicating "overdispersion". If we now set

$$\bar{\lambda} = \bar{\lambda}(\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta}) \tag{3-8}$$

we arrive at the likelihood function

$$\mathcal{L}(\boldsymbol{\beta}, \kappa|\mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n} \frac{\Gamma\left(\kappa + y_i\right)}{\Gamma\left(\kappa\right)\,y_i\,!}\;\frac{\kappa^{\kappa}\;\exp\left(\mathbf{x}_i\boldsymbol{\beta}\right)^{y_i}}{\left[\kappa + \exp\left(\mathbf{x}_i\boldsymbol{\beta}\right)\right]^{\kappa+y_i}}$$

[9] See, e.g., Ronning (1991), section 4.2.4.

[10] See, e.g., Ronning (1991), appendix A11 .

which has to be maximized with respect to $\boldsymbol{\beta}$ and κ.

We consider first partial derivatives of the log-likelihood function

$$\begin{aligned} L(\boldsymbol{\beta}, \kappa|\mathbf{y}, \mathbf{X}) \;&=\; \sum_{i=1}^{n} \{\log(\Gamma(\kappa + y_i)) \;-\; \log(\Gamma(\kappa)) \;-\; \log(y_i\,!) \\ &\qquad + \kappa \log(\kappa) \;+\; y_i\,\mathbf{x}_i\boldsymbol{\beta} \;-\; (\kappa + y_i)\,\log([\kappa + \exp(\mathbf{x}_i\boldsymbol{\beta})])\} \end{aligned}$$

from which the first order conditions are given by

$$\frac{\partial}{\partial \boldsymbol{\beta}} L \;=\; \sum_{i=1}^{n} \left\{ y_i\,\mathbf{x}_i \;-\; \frac{(\kappa + y_i)\,\exp(\mathbf{x}_i\boldsymbol{\beta})}{\kappa + \exp(\mathbf{x}_i\boldsymbol{\beta})}\,\mathbf{x}_i \right\} \quad = \quad \mathbf{0} \qquad (3\text{-}9)$$

and

$$\begin{aligned} &\frac{\partial}{\partial \kappa} L \\ &= \sum_{i=1}^{n} \left\{ \psi(\kappa + y_i) \;-\; \psi(\kappa) + \;(\log(\kappa) + 1) \;-\; y_i \log([\kappa + \exp(\mathbf{x}_i\boldsymbol{\beta})]) \;-\; \frac{\kappa + y_i}{\kappa + \exp(\mathbf{x}_i\boldsymbol{\beta})} \right\} \\ &= \qquad 0 \qquad . \end{aligned} \qquad (3\text{-}10)$$

For the first equation we may also write

$$\frac{\partial}{\partial \boldsymbol{\beta}} L \;=\; \sum_{i=1}^{n} \left\{ \frac{1}{1 + \kappa^{-1}\exp(\mathbf{x}_i\boldsymbol{\beta})}\,(y_i - \exp(\mathbf{x}_i\boldsymbol{\beta}))\,\mathbf{x}_i \right\} = \mathbf{0}$$

which equals results given, for example , by Winkelmann(2008), section 4.2.2.

Using the $(n \times n)$ diagonal matrix

$$\mathbf{D}_{\hat{\theta}} \quad = \quad \begin{pmatrix} \frac{1}{1+\hat{\kappa}^{-1}\exp(\mathbf{x}_1\hat{\boldsymbol{\beta}})} & & & & \\ & \frac{1}{1+\hat{\kappa}^{-1}\exp(\mathbf{x}_2\hat{\boldsymbol{\beta}})} & & & \\ & & \ddots & & \\ & & & \frac{1}{1+\hat{\kappa}^{-1}\exp(\mathbf{x}_{n-1}\hat{\boldsymbol{\beta}})} & \\ & & & & \frac{1}{1+\hat{\kappa}^{-1}\exp(\mathbf{x}_n\hat{\boldsymbol{\beta}})} \end{pmatrix}$$

equation (3-9) may be written in the form

$$\mathbf{X}'\,\hat{\mathbf{D}}\,\mathbf{y} \quad = \quad \mathbf{X}'\,\hat{\mathbf{D}} \begin{pmatrix} \bar{\lambda}(\mathbf{x}_1;\hat{\boldsymbol{\beta}}) \\ \bar{\lambda}(\mathbf{x}_2;\hat{\boldsymbol{\beta}}) \\ \vdots \\ \bar{\lambda}(\mathbf{x}_{n-1};\hat{\boldsymbol{\beta}}) \\ \bar{\lambda}(\mathbf{x}_n;\hat{\boldsymbol{\beta}}) \end{pmatrix} \qquad (3\text{-}11)$$

with $\bar{\lambda}$ from (3-8). This equation has the structure of (3-6). Please note that additionally the condition (3-10) must be satisfied.

3.3.3 Probit model

Maximum likelihood estimation of the binary probit model leads to [11]:

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \sum_{t=1}^{n} \frac{\phi_t}{\Phi_t(1-\Phi_t)} (y_t - \Phi_t) \mathbf{x}_t = \mathbf{0} \quad . \tag{3-12}$$

where

$$\Phi_t = \Phi(\mathbf{x}_t\boldsymbol{\beta})$$

and

$$\phi_t = \phi(\mathbf{x}_t\boldsymbol{\beta})$$

denote the cdf and density, respectively, of the standard normal distribution. Using the $(n \times n)$ diagonal matrix

$$\mathbf{D} = \begin{pmatrix} \frac{\phi_1}{\Phi_1(1-\Phi_1)} & & & & \\ & \frac{\phi_2}{\Phi_2(1-\Phi_2)} & & & \\ & & \ddots & & \\ & & & \frac{\phi_{n-1}}{\Phi_{n-1}(1-\Phi_{n-1})} & \\ & & & & \frac{\phi_n}{\Phi_n(1-\Phi_n)} \end{pmatrix}$$

the first order conditions (3-12) can also be written as

$$\mathbf{X}'\hat{\mathbf{D}}\,\mathbf{y} = \mathbf{X}'\hat{\mathbf{D}} \begin{pmatrix} \hat{\Phi}_1 \\ \hat{\Phi}_2 \\ \vdots \\ \hat{\Phi}_{n-1} \\ \hat{\Phi}_n \end{pmatrix} \tag{3-13}$$

which has the structure of (3-6).

4 Additional remarks

4.1 Arbitrary number of binary regressors

So far we have considered the special case of just two binary regressors. For the generalization to the case of an arbitrary number of binary variables we refer the reader to any textbook on experimental design. See, for example, Hocking (2003). For $p = 2$ we have two main effects and one interaction, i.e. two regressors plus the constant term so that $r = 4$. In case of $p = 3$ we have three main effects, three interaction of first order and one interaction of second

[11] See, for example, Ronning (1991), section 2.2.1

order, plus constant term so that $r = 8$. More generally, for p different binary variables the number r of regressors equals

$$r \quad = \quad \sum_{j=0}^{p} \left(\begin{array}{c} p \\ j \end{array} \right) \quad = \quad 2^p \qquad .$$

From the exposition given for the case of $q = 2$ binary regressors in section 3 it is clear that for all models considered there which satisfy either (3-2) or (3-6) the estimated conditional expectations are given by

$$\hat{\theta}_i \quad = \quad \frac{1}{n(\mathcal{A}_j)} \sum_{i \epsilon \mathcal{A}_j} y_i \qquad , i \, \epsilon \, \mathcal{A}_j \quad , \tag{4-1}$$

where $\mathcal{A}_j$, $j = 1, \ldots, r,$ defines the r sets of $n(\mathcal{A}_j)$ identical rows from $\mathbf{X}$.

Of course, interactions of order higher than two will not be considered very often in applications. However, in the next subsection we present a situation where such interactions play an important rule with regard to disclosure risk although they need not to arise in (fully) saturated models.

4.2 Interactions and strategic dummy variables

Bleninger et al. (2010) consider disclosure risk of so-called "strategic" dummies. These are specified in such a way that exact exact knowledge of a variable for some observational unit will identify it in the data set. For example, if an intruder knows the exact employment of some enterprise, he could specify a dummy variable

$$\Im_{x=x_m} \quad = \left\{ \begin{array}{ll} 1 & \text{if } x = \hat{x}_m \\ 0 & \text{else} \end{array} \right. \tag{4-2}$$

where $\hat{x}_m$ is the guessed employment figure. Assuming only approximate knowledge one would rather use

$$\Im_{x \simeq x_m} \quad = \left\{ \begin{array}{ll} 1 & \text{if } x - \gamma < \hat{x}_m < x + \gamma \\ 0 & \text{else} \end{array} \right. \tag{4-3}$$

The authors show that if a *single unit is identified* by one of these this dummy, then the the estimated value of y, denoted here by $\hat{\theta}_m$ will equal y_m. So by this procedure any sensitive variable such as sales or investment for enterprise m could be revealed if predicted values are reported to the user of remote access.

The situation with only a single observational unit may not happen frequently, especially if x is a categorical variable, and even for continuous variables especially in case of (4-3) many units may report the same value because of rounding. 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 also uses information concerning industry classification and the region. In this case one could define an indicator dummy for each variable for which the intruder has background information. Let $x_1, ..., x_q$ be the variables for which background information is available and let $\Im_1, ...\Im_q$ be the q indicators defined as in (4-2) . The strategic dummy variable is then defined as the interaction of all these indicators:

$$\Im \quad = \begin{cases} 1 & \text{if } \Im_1 = 1 \wedge \Im_2 = 1 \wedge \cdots \wedge \Im_p = 1 \\ 0 & \text{else} \end{cases} \quad . \qquad (4\text{-}4)$$

If such interaction leads to a single observation in the corresponding cell, then it is clear from the discussion in sections 3.2.1 and 3.3 above that the corresponding predicted value $\hat{\theta}_m$ will equal y_m, the value of the dependent variable for enterprise m. Bleninger et al. (2011), table 2 report results based on the German IAB Establishment Panel: They considered the interaction of a dummy for a certain employment interval with information on the site (federal state), legal form of enterprise and industry classification: For larger firms more than 90 % of enterprises were uniquely identified by such an interaction term.

The important thing to note is that in Bleninger et al. (2011) a *linear regression model* was assumed with a strategic dummy included in the set of regressors (binary and/or continuous). In their paper it is proved that $\hat{y}_m = y_m$ if a single unit is uniquely identified by the strategic dummy. On the other hand, in our paper we assume that for a set of q binary regressors a saturated model is specified. This means that if we would consider an analysis of variance with a dummy of the type (4-2) or (4-3) together with binary regressors for site, legal status, region and industry and specify interactions of all orders for these $q = 5$ effects, then in case of a single observation i identified by the interaction of order 5 would imply $\hat{\theta}_i = y_i$ for this observation. But in this case the result *holds not only for the linear regression model, but also for all nonlinear models* (count data models and choice models) which we discussed in section 3.

Let us shortly consider the case that any continuous regressor x is added to the saturated specification. In case of $q = 2$ binary regressors for the Poisson regression model or the logit model we would obtain from (3-2) the set of equations given by (3-3) supplemented by a fifth equation:

$$\sum_{i\epsilon W\cap N} y_i + \sum_{i\epsilon \overline{W}\cap N} y_i + \sum_{i\epsilon W\cap \overline{N}} y_i + \sum_{i\epsilon \overline{W}\cap \overline{N}} y_i$$

$$= \sum_{i\epsilon W\cap N} \hat{\theta}_i + \sum_{i\epsilon \overline{W}\cap N} \hat{\theta}_i + \sum_{i\epsilon W\cap \overline{N}} \hat{\theta}_i + \sum_{i\epsilon \overline{W}\cap \overline{N}} \hat{\theta}_i$$

$$\sum_{i\epsilon W\cap N} y_i + \sum_{i\epsilon \overline{W}\cap N} y_i = \sum_{i\epsilon W\cap N} \hat{\theta}_i + \sum_{i\epsilon \overline{W}\cap N} \hat{\theta}_i$$

$$\sum_{i\epsilon W\cap N} y_i + \sum_{i\epsilon W\cap \overline{N}} y_i = \sum_{i\epsilon W\cap N} \hat{\theta}_i + \sum_{i\epsilon W\cap \overline{N}} \hat{\theta}_i$$

$$\sum_{i\epsilon W\cap N} y_i = \sum_{i\epsilon W\cap N} \hat{\theta}_i$$

$$\sum_{i\epsilon W\cap N} x_i y_i + \sum_{i\epsilon \overline{W}\cap N} x_i y_i + \sum_{i\epsilon W\cap \overline{N}} x_i y_i + \sum_{i\epsilon \overline{W}\cap \overline{N}} x_i y_i$$

$$= \sum_{i\epsilon W\cap N} x_i\hat{\theta}_i + \sum_{i\epsilon \overline{W}\cap N} x_i\hat{\theta}_i + \sum_{i\epsilon W\cap \overline{N}} x_i\hat{\theta}_i + \sum_{i\epsilon \overline{W}\cap \overline{N}} x_i\hat{\theta}_i$$

It is obvious that from the fourth equation for $n_{W\cap N} = 1$ we would obtain $\hat{\theta}_{W\cap N} = y_{W\cap N}$. Therefore the result still holds if continuous regressors are added, and of course it is not necessary that the main effects are specified. It is only necessary that the interaction is included.

However we still have to check whether for the NEGBIN model and for the probit model which satisfy the more general equation (3-6) the same result is true. In this case from adding a continuous regressor we obtain

$$\sum_{i\epsilon W\cap N} d_i y_i + \sum_{i\epsilon \overline{W}\cap N} d_i y_i + \sum_{i\epsilon W\cap \overline{N}} d_i y_i + \sum_{i\epsilon \overline{W}\cap \overline{N}} d_i y_i$$

$$= \sum_{i\epsilon W\cap N} d_i\hat{\theta}_i + \sum_{i\epsilon \overline{W}\cap N} d_i\hat{\theta}_i + \sum_{i\epsilon W\cap \overline{N}} d_i\hat{\theta}_i + \sum_{i\epsilon \overline{W}\cap \overline{N}} d_i\hat{\theta}_i$$

$$\sum_{i\epsilon W\cap N} d_i y_i + \sum_{i\epsilon \overline{W}\cap N} d_i y_i = \sum_{i\epsilon W\cap N} d_i\hat{\theta}_i + \sum_{i\epsilon \overline{W}\cap N} d_i\hat{\theta}_i$$

$$\sum_{i\epsilon W\cap N} d_i y_i + \sum_{i\epsilon W\cap \overline{N}} d_i y_i = \sum_{i\epsilon W\cap N} d_i\hat{\theta}_i + \sum_{i\epsilon W\cap \overline{N}} d_i\hat{\theta}_i$$

$$\sum_{i\epsilon W\cap N} d_i y_i = \sum_{i\epsilon W\cap N} d_i\hat{\theta}_i$$

$$\sum_{i\epsilon W\cap N} d_i x_i y_i + \sum_{i\epsilon \overline{W}\cap N} d_i x_i y_i + \sum_{i\epsilon W\cap \overline{N}} d_i x_i y_i + \sum_{i\epsilon \overline{W}\cap \overline{N}} d_i d_i x_i y_i$$

$$= \sum_{i\epsilon W\cap N} d_i x_i\hat{\theta}_i + \sum_{i\epsilon \overline{W}\cap N} d_i x_i\hat{\theta}_i + \sum_{i\epsilon W\cap \overline{N}} d_i x_i\hat{\theta}_i + \sum_{i\epsilon \overline{W}\cap \overline{N}} d_i x_i\hat{\theta}_i$$

where the fifth equation has been added. Although d_i now varies also with the continuous regressor x, in case of $n_{W\cap N} = 1$ we would obtain from the fourth equation $d_{W\cap N}\,\hat{\theta}_{W\cap N} = d_{W\cap N}\,y_{W\cap N}$ or $\hat{\theta}_{W\cap N} = y_{W\cap N}$ as above. Therefore the result also holds for NEGBIN and probit. And of course it is not necessary that the main effects are specified. It is only necessary that the interaction is included which identifies a single unit.

5 Résumé

We show in this paper that predicted values, i.e. estimated conditional expectations, in saturated models bear high disclosure risk if they are based on few observations only. In case of single observations the corresponding estimate is identical to the corresponding value of the dependent variable. This is not only true for the linear model. We show in detail that this pertains also all important nonlinear microeconometric models: Poisson and NEGBIN regression models as well as logit and probit models. In the last section we draw attention to the fact that not the main effects but the interactions are of special concern since they can be used as "strategic dummy variables" exploiting external knowledge regarding a set of variables of some observational units. Therefore interactions should be checked carefully with regard to disclosure risk. On the other hand, principally interdicting any interactions cannot be accepted from a statistical point of view since interaction terms may be important in interpreting estimated main effects in saturated models.

References

[Bleninger et al. (2011)] Bleninger, P,. Drechsler, J., Ronning, G. . Remote data access and the risk of disclosure from linear regression: An empirical study. Statistics and Operations Research Transactions (SORT)**, **–** (2011)

[Fahrmeir et al. (1996)] Fahrmeir, L., Hamerle, A., Tutz, G. (editors) . Multivariate Statistische Verfahren. De Gruyter: Berlin, second edition. (1996)

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

[Greene(2000)] Greene, W.H., Econometric Analysis, Prentice Hall, Upper Saddle River (NJ), 4th edition (2000)

[Hocking (2003)] Hocking, R.R. . Methods and Applications of Linear Models: Regression and the Analysis of Variance, Wiley, New York, 2nd edition. (2003)

[O'Keefe and Good (2009)] O'Keefe, C.O., Good, N.M. , Regression output from a remote analysis server. Data & Knowledge Engineering 68, 1175-1186. (2009)

[Reznek (2003)] Reznek, A.P. , Disclosure risks in cross-section regression models. Proceedings of the American Statistical Association, Government Statistics Section, [CD-ROM], Alexandria, VA, American Statistical Association, 3444 - 3451. (2003)

[Ronning (1991)] Ronning, G. , Mikroökonometrie, Springer, Berlin (1991)

[Ronning et al. (2010)] Ronning, G., , Bleninger, Ph., Drechsler, J., Gürke , Ch., Remote Access – Eine Welt ohne Mikrodaten?? IAW Discussion Paper 66 (June 2010). (2010). http://www.iaw.edu/RePEc/iaw/pdf/iaw_dp_66.pdf. Accessed 11 June 2011.

[Winkelmann(2008)] Winkelmann, R. (2003). *Econometric Analysis of Count Data*. Springer: Berlin, 5th edition.

IAW-Diskussionspapiere

Die IAW-Diskussionspapiere erscheinen seit September 2001. Die vollständige Liste der IAW-Diskussionspapiere von 2001 bis 2008 (Nr. 1-44) finden Sie auf der IAW-Internetseite www.iaw.edu/publikationene/iaw-diskussionspapiere.

IAW-Diskussionspapiere seit Juli 2008:

Nr. 45 (Oktober 2008)
Effects of Dismissal Protection Legislation on Individual Employment Stability in Germany
Bernhard Boockmann / Daniel Gutknecht / Susanne Steffes

Nr. 46 (November 2008)
Trade's Impact on the Labor Share: Evidence from German and Italian Regions
Claudia M. Buch / Paola Monti / Farid Toubal

Nr. 47 (März 2009)
Network and Border Effects: Where Do Foreign Multinationals Locate in Germany?
Julia Spies

Nr. 48 (März 2009)
Stochastische Überlagerung mit Hilfe der Mischungsverteilung (Stand: 18. März 2009 – Version 49)
Gerd Ronning

Nr. 49 (April 2009)
Außenwirtschaftliche Verbindungen der deutschen Bundesländer zur Republik Österreich
Anselm Mattes /Julia Spies

Nr. 50 (Juli 2009)
New Firms – Different Jobs? An Inquiry into the Quality of Employment in Start-ups and Incumbents (Stand: 28. Juli 2009 – Version 1.3)
Andreas Koch / Jochen Späth

Nr. 51 (Juli 2009)
Poverty and Wealth Reporting of the German Government: Approach, Lessons and Critique
Christian Arndt / Jürgen Volkert

Nr. 52 (August/September 2009)
Barriers to Internationalization: Firm-Level Evidence from Germany
Christian Arndt / Claudia M. Buch / Anselm Mattes

Nr. 53 (September 2009)
IV-Schätzung eines linearen Panelmodells mit stochastisch überlagerten Betriebs- und Unternehmensdaten
Elena Biewen / Gerd Ronning / Martin Rosemann

Nr. 54 (November 2009)
Financial Constraints and the Margins of FDI
Claudia M. Buch / Iris Kesternich / Alexander Lipponer / Monika Schnitzer

Nr. 55 (November 2009)
Offshoring and the Onshore Composition of Tasks and Skills
Sascha O. Becker / Karolina Ekholm / Marc-Andreas Muendler

Nr. 56 (November 2009)
Intensifying the Use of Benefit Sanctions – An Effective Tool to Shorten Welfare Receipt and Speed up Transitions to Employment?
Bernhard Boockmann / Stephan L. Thomsen / Thomas Walter

Nr. 57 (November 2009)
The Responses of Taxable Income Induced by Tax Cuts – Empirical Evidence from the German Taxpayer Panel
Peter Gottfried / Daniela Witczak

Nr. 58 (November 2009)
Reformoption Duale Einkommensteuer – Aufkommens- und Verteilungseffekte
Peter Gottfried / Daniela Witczak

IAW-Diskussionspapiere

Nr. 59 (Februar 2010)
The Impact of Horizontal and Vertical FDI on Labor Demand for Different Skill Groups
Anselm Mattes

Nr. 60 (Februar 2010)
International M & A: Evidence on Effects of Foreign Takeovers
Anselm Mattes

Nr. 61
The Impact of Regional Supply and Demand Conditions on Job Creation and Destruction (Februar 2010)
Raimund Krumm / Harald Strotmann

Nr. 62 (April 2010)
The Effects of Foreign Ownership Change on the Performance of German Multinational Firms
Christian Arndt / Anselm Mattes

Nr. 63 (April 2010)
The Export Magnification Effect of Offshoring
Jörn Kleinert / Nico Zorell

Nr. 64 (März 2010)
Kundenbetreuung aus einer Hand im SGB II? – Integration versus Spezialisierung von Fallmanagement, Vermittlung und materiellen Leistungen
Harald Strotmann / Martin Rosemann / Sabine Dann / Christine Hamacher

Nr. 65 (Mai 2010)
The Combined Employment Effects of Minimum Wages and Labor Market Regulation – A Meta-analysis
Bernhard Boockmann

Nr. 66
Remote Access – Eine Welt ohne Mikrodaten ?? (Stand: 20.06.2010, Version 18) (Juni 2010)
Gerd Ronning / Philipp Bleninger / Jörg Drechsler / Christopher Gürke

Nr. 67
Opening Clauses in Collective Bargaining Agreements: More Flexibility to Save Jobs? (Oktober 2010)
Tobias Brändle / Wolf Dieter Heinbach

Nr. 68
Interest Rate Policy and Supply-side Adjustment Dynamics (Dezember 2010)
Daniel Kienzler / Kai Schmid

Nr. 69
Should Welfare Administration be Centralized or Decentralized? Evidence from a Policy Experiment (Dezember 2010)
Bernhard Boockmann / Stephan L. Thomsen / Thomas Walter / Christian Göbel / Martin Huber

Nr. 70
Banks in Space: Does Distance Really Affect Cross-Border-Banking? (Februar 2011)
Katja Neugebauer

Nr. 71
An Almost Ideal Wage Database Harmonizing the ILO Database October Inquiry (Februar 2011)
Daniela Harsch / Jörn Kleinert

Nr. 72
Disclosure Risk from Interactions and Saturated Models in Remote Access (Juni 2011)
Gerd Ronning