# Exact inference on poverty predictors based on logistic regression approach

**Ottó Hajdu,**
Associate professor
Budapest University of Technology
E-mail: hajdu@finance.bme.hu

The paper deals with the exact inference on poverty indicators included in a predictive logistic regression model as predictor variables. Based on a multiple stratification applied in a household survey, small size or unbalanced subgroups are likely to occur in practice with regard to the number of poor and hence the standard unconditional maximum likelihood estimation of a regression parameter may fail to exist. Focus is brought on exact inference which is still possible to make even at this case. The paper gives a brief overview of problems of exact *p*-value and confidence interval calculation in small samples for the case when the unconditional maximum likelihood estimate does not exist or the large sample asymptotic properties are violated. Besides, some empirical examples are presented based on a survey of Hungarian households.

KEYWORDS:
Logistic and loglinear models.
Estimation. Poverty and social deprivation.

Logistic regression is a straightforward method to evaluate the risk of being poor based on an appropriate poverty line.[1] This paper focuses mainly on the problem of the relevancy of poverty indicators when the so-called *p*-value criterion is applied to select a set of predictor variables. Since several types of homogenous households can be defined by their socio-economic and demographic characteristics, small size or unbalanced subgroups are likely to be formed for measuring poverty in a stratified survey. Calculating *p*-values for inference accurately as possible is clearly a key stage of the model building process but the standard unconditional maximum likelihood approach exhibits appealing asymptotic properties only in the large sample case. Otherwise, considering small or unbalanced samples the exact approach of inference gives the correct *p*-values and confidence intervals. The main purpose of this paper is to give a brief guide to apply exact methods and to interpret their results correctly.

## 1. The predictive model

Consider a set of independent binary random variables, $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ where $i = 1,\ 2,\ ....,\ n$ stands for an individual household. The response variable $Y_i$ takes the value of 1 in the case of households falling below the poverty line and 0 otherwise. Corresponding to each response variable, $Y_i$, there is a $(p \times 1)$ covariate vector $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})'$ of predictor variables. Let $\pi_x$ denote the conditional probability $Pr(Y = 1|\mathbf{x})$ i.e. the probability that $Y = 1$ conditioned on the levels of the covariates. Using the so-called odds-measure defined as $\pi_x / (1 - \pi_x)$ the conditional probability of the event „1" can be written as follows:

$$\pi_x = \frac{\pi_x / (1 - \pi_x)}{1 + \pi_x / (1 - \pi_x)} = \frac{\text{odds}_x}{1 + \text{odds}_x}. \qquad /1/$$

When $\pi_x$ exceeds a critical value of $C$ the prediction is $\hat{Y} = 1$ otherwise the predicted value is 0. In the case of logistic regression models the dependency of $\pi_x$ on the predictor variables is expressed through the relationship

[1] For a recent work in the field see for instance *Havasi* [2005].

$$\log\left(\text{odds}_x\right)=\boldsymbol{\beta}\mathbf{x}\,, \qquad /2/$$

where $\boldsymbol{\beta}$ is a $(1\times p)$ vector of unknown parameters. An *unstratified* model is defined when there is only one single constant term included among the parameters. This distinguishes it from the *stratified* model when data come from several strata and a different stratum specific constant term is introduced to each stratum provided common slope parameters.

Based on a random sample of size $n$ included in the $(n\times 1)$ vector $\mathbf{y}=(y_1,y_2,...,y_n)'$ the unconditional likelihood of the sample to be maximized with respect to $\boldsymbol{\beta}$ is

$$L(\boldsymbol{\beta})=\prod_{i=1}^{n}\frac{\left(\text{odds}_i\right)^{y_i}}{1+\text{odds}_i}=\frac{e^{\sum_{i=1}^{n}y_i\sum_{j=1}^{p}\beta_j x_{ij}}}{\prod_{i=1}^{n}\left(1+e^{\sum_{j=1}^{p}\beta_j x_{ij}}\right)}=\frac{e^{\sum_{j=1}^{p}\beta_j t_j}}{\prod_{i=1}^{n}\left(1+e^{\sum_{j=1}^{p}\beta_j x_{ij}}\right)}\to\max\,, \qquad /3/$$

where the

$$t_j=\sum_{i=1}^{n}y_i x_{ij}\quad (j=1,2,...,p) \qquad /4/$$

sample statistic is termed as *sufficient statistic* for the parameter $\beta_j$ in the literature of the exact logistic regression.[2] Recalling now that $y_i=\{0,1\}$, $t_j$ is simply the sum of the predictor variable $x_j$ for those who fall below the poverty line. Hence, if a constant term is included in the model, its sufficient statistic means the number of poor households. Obviously, $t_j$ is the sample outcome of the random variable $T_j$.

In order to make inference on parameters of our interest consisting of the vector $\boldsymbol{\beta}_1$, partition the parameter vector $\boldsymbol{\beta}$ into two components $\boldsymbol{\beta}=(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)$. Suppose we are primarily interested in inference about $\boldsymbol{\beta}_1$, and regard $\boldsymbol{\beta}_2$ as „nuisance" parameter(s). The partitioned vector of the sufficient statistics corresponding to $\boldsymbol{\beta}=(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)$ is $\mathbf{t}=(\mathbf{t}_1,\mathbf{t}_2)$ or in matrix notation

$$\mathbf{X}_1'\mathbf{y}=\mathbf{t}_1,\quad \mathbf{X}_2'\mathbf{y}=\mathbf{t}_2\text{, where} \qquad /5/$$

$$\mathbf{X}=\left[\mathbf{X}_1,\mathbf{X}_2\right] \qquad /6/$$

[2] For the definition of a sufficient statistic see *Garthwaite–Jolliffe–Jones* [1995].

is the partitioned form of the $(n \times p)$ covariate matrix $\mathbf{X}$.

The unconditional likelihood can be written in an equivalent form of

$$f_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}(\mathbf{t}_1,\mathbf{t}_2) = \frac{c(\mathbf{t}_1,\mathbf{t}_2)e^{\boldsymbol{\beta}_1\mathbf{t}_1+\boldsymbol{\beta}_2\mathbf{t}_2}}{\sum_{\mathbf{u}} c(\mathbf{u}_1,\mathbf{u}_2)e^{\boldsymbol{\beta}_1\mathbf{u}_1+\boldsymbol{\beta}_2\mathbf{u}_2}}, \qquad /7/$$

where $c(\mathbf{t})$ is the number of ways of selecting the binary sequence $\mathbf{y}$ so as to satisfy the $\mathbf{X}'\mathbf{y} = \mathbf{t}$ sample condition. The summation in the denominator is over all $\mathbf{u} = \mathbf{X}'\mathbf{Y}$ generated by any $(n \times 1)$ binary sequence $\mathbf{Y}$. Maximizing /7/ by choice of $\boldsymbol{\beta}$ is equivalent to maximizing /3/ by choice of $\boldsymbol{\beta}$ (see *LogXact 7 User Manual* [2005] p. 503).

## 2. Hypothesis testing

Three methods of inference are available for testing: *1.* unconditional likelihood inference, *2.* conditional likelihood inference, and *3.* conditional exact inference. Unconditional likelihood inference is based on estimating the entire parameter vector $\boldsymbol{\beta}$ by maximizing the unconditional likelihood function. This approach assumes the asymptotic normality of the maximum likelihood estimates to make inferences about $\boldsymbol{\beta}_1$. Conditional likelihood inference, in contrast, is based on maximizing the likelihood function derived by conditioning on the sufficient statistics for $\boldsymbol{\beta}_2$ in /7/. By definition this likelihood function is *free* of the nuisance parameters $\boldsymbol{\beta}_2$. Finally, conditional exact inference is based on deriving the exact permutational distribution $c(\mathbf{u}_1,\mathbf{t}_2)$ of the sufficient statistics for $\boldsymbol{\beta}_1$ conditional on the sufficient statistics for $\boldsymbol{\beta}_2$.

It is apparent, that conditioning on $\mathbf{t}_2$ plays a key role in inference because thereby parameter $\boldsymbol{\beta}_2$ can entirely be eliminated from the likelihood:

$$f_{\boldsymbol{\beta}_1}(\mathbf{t}_1 \mid \mathbf{t}_2) = L(\boldsymbol{\beta}_1 \mid \mathbf{t}_2) = \frac{c(\mathbf{t}_1,\mathbf{t}_2)e^{\boldsymbol{\beta}_1\mathbf{t}_1}}{\sum_{\mathbf{u}_1} c(\mathbf{u}_1,\mathbf{t}_2)e^{\boldsymbol{\beta}_1\mathbf{u}_1}} \to \max . \qquad /8/$$

Maximizing /8/ with respect to $\boldsymbol{\beta}_1$ yields conditional maximum likelihood estimates (CMLE). The fundamental difference between unconditional maximum likelihood estimates (MLE) and conditional CMLE inference is that MLE needs to estimate $\boldsymbol{\beta}_2$ even if it were only a nuisance parameter, interest being focused mainly

on $\boldsymbol{\beta}_1$ while based on the CMLE approach $\boldsymbol{\beta}_2$ is conditioned out from the estimation.

To test the null hypothesis

$$H_0 : \boldsymbol{\beta}_1 = \mathbf{0}, \qquad /9/$$

based on the MLE approach three basic methods are available: *1.* the *scores* statistic (also known as Lagrange-Multiplier test statistic), *2.* the *likelihood ratio* statistic and *3.* the *Wald* statistic. All of these statistics are asymptotically chi-squared distributed on *d* degrees of freedom under $H_0$ where *d* is the number of restrictions imposed. It must be emphasized that the scores statistic does not depend on the full model MLE. It is derived based on the MLE of the restricted model only. This means that the scores statistic may exist even when the MLE of the full model does not exist.

When CMLE is applied the corresponding conditional versions of these test statistics are in hand to be applied. Further, when data are from several strata and number of the stratum specific "nuisance constants" is large relative to the number of observations the MLE estimates for the slope parameters may be inconsistent. In this case the CMLE estimates for the slope parameters $\boldsymbol{\beta}_1$ (eliminating the stratum specific constant terms included in $\boldsymbol{\beta}_2$ from the computation) are consistent. Since we have eliminated the stratum specific constant terms from the likelihood function it is no longer possible to test hypothesis about these parameters.

For small or highly imbalanced data sets the asymptotic chi-squared distributions for the scores, likelihood ratio and Wald tests might not hold. This is the situation when it is appropriate to generate the true permutation distribution of $\mathbf{T}_1$ given $\mathbf{t}_2$. Armed with this permutation distribution one can perform exact hypothesis tests and generate exact confidence intervals for the parameters of interest. Based on /8/ under $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$ the exact conditional probability that $\mathbf{T}_1 = \mathbf{t}_1$ given $\mathbf{t}_2$ that is the null-distribution of $f_{\boldsymbol{\beta}_1}(\mathbf{t}_1 \mid \mathbf{t}_2)$ reduces to

$$f_0(\mathbf{t}_1 \mid \mathbf{t}_2) = \frac{c(\mathbf{t}_1, \mathbf{t}_2)}{\sum_{\mathbf{u}_1} c(\mathbf{u}_1, \mathbf{t}_2)}. \qquad /10/$$

An exact *p*-value for testing $H_0$ is then defined as follows:

$$p = \sum_{\mathbf{u}_1 \in R} f_0(\mathbf{u}_1 \mid \mathbf{t}_2), \qquad /11/$$

where *R* is the region of the conditional sample space of $\mathbf{T}_1$ given $\mathbf{t}_2$ in which the values $\mathbf{T}_1 = \mathbf{u}_1$ are all considered to be more extreme under $H_0$ than the observed value $\mathbf{T}_1 = \mathbf{t}_1$.

An adjusted version of $p$, the so-called *mid-p-value* can be applied to correct the discrete test without compromising on its significance level:

$$p_{mid} = p - 0.5 f_0\left(\mathbf{t}_1 \mid \mathbf{t}_2\right). \qquad /12/$$

The choice of $R$ depends on the type of exact test selected. We provide three methods for selecting an exact test: *1.* exact conditional scores test (based on either asymptotic or exact variance); *2.* exact conditional probability test; and *3.* exact likelihood ratio test.

For the *exact conditional scores test*, $R$, the extreme region of the sample space over which the $p$-value is calculated is defined to be all values of the test statistic having a value greater than or equal to the *observed* test statistic:

$$q = \left(\mathbf{t}_1 - E\left(\mathbf{T}_1\right)\right)' Cov_{\mathbf{T}_1}^{-1}\left(\mathbf{t}_1 - E\left(\mathbf{T}_1\right)\right). \qquad /13/$$

For the *exact conditional probability test*, $R$, the extreme region of the sample space over which the $p$-value is calculated is defined to be all values of the test statistic having a probability smaller than or equal to the probability of the observed test statistic. Finally, for the *exact likelihood ratio test*, $R$, the extreme region of the sample space over which the $p$-value is calculated is defined to be all values of the test statistic having a likelihood ratio value greater than or equal to that of the observed data.

Although the exact tests are guaranteed to protect from type-1 error at any specified level, the conditional scores test may be the preferred one. The main reason is that characterizing the rejection region $R$ in terms of larger conditional scores rather than smaller conditional probabilities is intuitively appealing. Considering the exact conditional probability approach in the univariate case (when $\mathbf{t}_1$ is a scalar $t_1$) it may happen that the conditional probability distribution $f_0\left(t_1 \middle| \mathbf{t}_2\right)$ has multiple modes so that the rejection region $R$ is not a contiguous interval. This problem does not occur when exact conditional scores test is applied. Moreover, the conditional maximum likelihood estimate fails to exist in certain situations hence it is impossible to carry out a likelihood ratio test. Notice, that for the conditional scores test, parameter estimates are not required.

## 3. Exact parameter estimation

Estimation of the parameters by maximizing /8/ with respect to $\boldsymbol{\beta}_1$ yields conditional CMLE point estimates. However, since the marginal interpretation of the

slope parameters is meaningful only in a partial sense it is reasonable to perform exact inference on each individual scalar parameter $\beta_j$ separately, step-by-step. This is carried out by taking the special partitioning $\boldsymbol{\beta}_1 = \beta_1$ which is done successively for each parameter of interest:

$$f_{\beta_1}\left(t_1 \mid \mathbf{t}_2\right) = \frac{c\left(t_1, \mathbf{t}_2\right) e^{\beta_1 t_1}}{\sum_{u_1 = t_{1\min}}^{t_{1\max}} c\left(u_1, \mathbf{t}_2\right) e^{\beta_1 u_1}} \to \max , \qquad /14/$$

where $t_{1\min}$ and $t_{1\max}$ are the minimum and maximum values respectively in the range of the random variable $T_1$ conditional on $\mathbf{t}_2$. Notice that /14/ depends only on $\beta_1$.

If the observed value of the sufficient statistic $t_1$ is at one extreme of its range (when either $t_1 = t_{1\min}$ or $t_1 = t_{1\max}$) it is no longer possible to maximize /14/ with respect to $\beta_1$. This is the case when the sample is *separated* with respect to the predictor $x_1$ (see *Christmann–Rousseeuw* [2001]). In this case the likelihood function increases strictly monotonically as $|\beta_1|$ goes towards $\infty$. When the CMLE fails to exist the so-called "median unbiased estimate" (MUE) is available for exact conditional estimation by solving the following equation:

$$f_{\hat{\beta}_1}\left(t_1 \mid \mathbf{t}_2\right) = 0.5 . \qquad /15/$$

Considering the two-sided $\alpha$ level confidence interval (CI) the upper and lower bounds $\beta_+$ and $\beta_-$ are defined respectively based on the left and right tails of the distribution of $T_1 \mid \mathbf{t}_2$ at the observed value $T_1 = t_1$ in the following manner:

$$F_{\beta+}\left(t_1\right) = \sum_{u_1 = t_{1\min}}^{t_1} f_{\beta_+}\left(u_1 \mid \mathbf{t}_2\right) = \alpha / 2 , \qquad /16/$$

$$G_{\beta_-}\left(t_1\right) = \sum_{u_1 = t_1}^{t_{1\max}} f_{\beta_-}\left(u_1 \mid \mathbf{t}_2\right) = \alpha / 2 . \qquad /17/$$

Solving /16/ and /17/ gives interval as desired. A one-sided CI can also be obtained spending the entire $\alpha$ error using only the left tail distribution $F_{\beta_+}\left(t_1\right) = \alpha$, when $F_0\left(t_1\right) \le G_0\left(t_1\right)$ and the right tail distribution $G_{\beta_-}\left(t_1\right) = \alpha$, when $F_0\left(t_1\right) > G_0\left(t_1\right)$. The resulted intervals are $\left(-\infty, \beta_+\right)$ and $\left(\beta_-, \infty\right)$ respectively.

Also, regardless the type of CI requested only an open interval is available when either $t_1 = t_{1\min}$ or $t_1 = t_{1\max}$ because the cumulative probability on the entire range of

the random variable $T_1 | \mathbf{t}_2$ always equals 1 and hence it is independent at the value of $\beta_1$. Then the CI bounds are constructed as follows:

$$F_{\beta+}(t_1) = \sum_{u_1=t_{1\min}}^{t_{1\min}} f_{\beta_+}(u_1 | \mathbf{t}_2) = \alpha/2, \quad \beta_- = -\infty, \qquad /18/$$

$$G_{\beta_-}(t_1) = \sum_{u_1=t_{1\max}}^{t_{1\max}} f_{\beta_-}(u_1 | \mathbf{t}_2) = \alpha/2, \quad \beta_+ = +\infty. \qquad /19/$$

Finally, whenever a CI reported at $\alpha$ level it is necessary to compute a *p*-value that preserves the consistency between the conclusions derived from the CI and from the *p*-value. Based on this particular method it is ensured that the exact *p*-value is less than $\alpha$ if and only if the exact $(1-\alpha)$ CI excludes the corresponding model parameter. According to this alternative definition the exact two-sided *p*-value is double the one-sided *p*-value: $p_2 = 2p_1$, where the one-sided *p*-value is the minimum of the left and right tail probabilities:

$$p_1 = \min\{F_0(t_1), G_0(t_1)\}. \qquad /20/$$

## 4. Some empirical examples

Let us consider the subgroup of *households* in *Budapest* with *6 persons or more*, in the year of 2003. The source of data is the HCSO, Household Expenditure Survey 2003. Those households with per capita income below the poverty line (0.6 median income) are considered to be poor.[3] Their status is recorded in the binary poverty variable by the value of Poverty = 1 which plays the role of the response variable. The examples of our interest are as follows.

### 4.1. The unconditional, asymptotic maximum likelihood estimation fails to exist

The estimation results are shown in Table 1 and the frequency distributions of the predictor variables considered are included in Table 2. The computations are carried out by the program LogXact 7 (www.cytel.com).

[3] The per capita median income level in the year of 2003 is HUF754.000 and per *capita* here stands for per *consumption unit* as defined by the OECD scheme using 1, .7 and .5 units representing first and further adults and children respectively.

Table 1

*Parameter estimates when MLE does not exist*

| Model | Type | Beta | SE(Beta) | Type | 95 percent CI lower | 95 percent CI upper | 2*1-sided=$p_2$ |
|---|---|---|---|---|---|---|---|
| Model term 1 | | | | | | | |
| Constant | MLE | ? | ? | asymptotic | ? | ? | ? |
| Gender | MLE | ? | ? | asymptotic | ? | ? | ? |
| | MUE | 4.481 | NA | exact | 2.804 | +INF | 1.094e-024 |
| Model term 2 | | | | | | | |
| Constant | MLE | ? | ? | asymptotic | ? | ? | ? |
| Permanently sick | MLE | ? | ? | asymptotic | ? | ? | ? |
| | MUE | –5.29 | NA | exact | –INF | –3.614 | 5.809e-052 |
| Model term 3 | | | | | | | |
| Constant | MLE | –8.522 | 0.3566 | asymptotic | –9.221 | –7.823 | 2.493e-051 |
| Education score | MLE | 0.5927 | 0.03053 | asymptotic | 0.5328 | 0.6525 | 2.327e-043 |
| | CMLE | 0.5926 | 0.03053 | exact | 0.534 | 0.6547 | 3.763e-202 |
| Model term 4 | | | | | | | |
| Constant | MLE | ? | ? | asymptotic | ? | ? | ? |
| Education level | MLE | ? | ? | asymptotic | ? | ? | ? |
| | MUE | 7.092 | NA | exact | 5.418 | +INF | 6.977e-257 |

*Note.* NA for not applicable, ? means does not exist, INF is infinite, e is exponent.

Table 2

*Frequency distributions of the predictor variables*

| Denomination | Poverty = 0 | Poverty = 1 | Total |
|---|---|---|---|
| Gender | | | |
| 0: Female | 601 | 0 | 601 |
| 1: Male | 6294 | 642 | 6936 |
| Permanently sick persons | | | |
| 0: not present | 5678 | 642 | 6320 |
| 1: present | 1217 | 0 | 1217 |
| Education score | | | |
| 3 | 1009 | 0 | 1009 |
| 5 | 1573 | 0 | 1573 |
| 7 | 370 | 0 | 370 |
| 8 | 1383 | 0 | 1383 |
| 11 | 545 | 355 | 900 |
| 12 | 1809 | 126 | 1935 |
| 13 | 206 | 161 | 367 |

*(Continued on the next page.)*

*(Continuation.)*

| Denomination | Poverty = 0 | Poverty = 1 | Total |
|---|---|---|---|
| Education level | | | |
| 1 | 1009 | 0 | 1009 |
| 2 | 3326 | 0 | 3326 |
| 3 | 2560 | 642 | 3202 |
| Unemployed persons | | | |
| 0 | 6459 | 516 | 6975 |
| 1 | 436 | 0 | 436 |
| 2 | 0 | 126 | 126 |
| Economic activity | | | |
| 111 type | 681 | 0 | 681 |
| 112 type | 986 | 161 | 1147 |
| 113 type | 410 | 0 | 410 |
| 114 type | 656 | 0 | 656 |
| 115 type | 1520 | 0 | 1520 |
| 117 type | 996 | 0 | 996 |
| 121 type | 609 | 481 | 1090 |
| 122 type | 601 | 0 | 601 |
| 232 type | 436 | 0 | 436 |
| *Total* | *6895* | *642* | *7537* |

Firstly, taking the gender of the head of household as a single predictor variable (Model term 1) "Female" is a perfect predictor hence the MLE does not exist (this is indicated by a ? mark) while the MUE point estimate and a one-sided confidence interval are available. The upper bound of the CI is +INF because the zero frequency occurs at the lower extreme value on the range of gender i.e. at female when Gender = 0 in Table 2.

In contrast, let us consider an another single binary predictor variable namely whether a permanently sick person is present in the household "1" or not "0" (Model term 2). The conclusions are similar to those made earlier with the only exception that the lower bound of the CI is -INF since the zero frequency in Table 2 occurs at the upper extreme value of 1 in the range i.e. when a permanently sick person is present in the household.

Merging categories can also influence the existence of the MLE. Considering the education score of the head of household as a single predictor variable (see Model term 3)[4] it is apparent that both MLE and CMLE exist *despite* the fact that zero

[4] The entire ordinal range of the level of education: [1,2,...,13] with 13 indicating a PhD degree.

frequencies appear *only* at the lower *tail* of the distribution published in Table 2. However, merging the scores into only three levels as shown in Table 2 the MLE does not exist any longer as it can be seen in Table 1 under Model Term 4.

## 4.2. Differences between exact and asymptotic *p*-values

Despite the relatively large sample size – given that the sample is unbalanced (i.e. the poor/non-poor ratio is 642/6895) – one can expect that the asymptotic and exact *p*-values differ significantly. Let us take the number of unemployed persons in the household as a single predictor variable (Model term 5). Results of Table 3 show that this is not the case. The number of unemployed persons is significant at any level and the point and interval estimations are quite similar in magnitude.

Selecting now the number of the dependent persons in the household as predictor variable (Model term 6), Table 3 shows that the exact *p*-value can considerably differ from the unconditional one. Apparently, the number of dependent persons is not significant at usual error levels but if the extreme 40 percent level were applied as a critical cut-off-value the different methods would yield different conclusions. It might be the case of several predictors in several strata not investigated in this paper.

Table 3

*Parameter estimates when the MLE does exist*

| Model | Type | Beta | SE(Beta) | Type | 95 percent CI lower | 95 percent CI upper | 2*1-sided=$p_2$ |
|---|---|---|---|---|---|---|---|
| Model term 5 | | | | | | | |
| Constant | MLE | –2.6420 | 0.04741 | asymptotic | –2.7350 | –2.5490 | 3.92e-085 |
| Unemployed | MLE | 1.4910 | 0.07773 | asymptotic | 1.3390 | 1.6440 | 6.443e-043 |
| | CMLE | 1.4910 | 0.07772 | exact | 1.3360 | 1.6470 | 3.333e-073 |
| Model term 6 | | | | | | | |
| Constant | MLE | –1.4590 | 0.41440 | asymptotic | –2.2710 | –0.6464 | 0.000432 |
| Dependents | MLE | –0.0877 | 0.10120 | asymptotic | –0.2860 | 0.1106 | 0.386 |
| | CMLE | –0.0876 | 0.10110 | exact | –0.2912 | 0.1158 | 0.4143 |

## 4.3. Strata specific constant terms are conditioned out

Let us discuss again the impact of the number of unemployed persons in the household as the single predictor but controlling on the type of households regarding

the economic activity of the households as a stratum variable (Model term 7 in Table 4). Several strata can be defined based on the classification whether an active person is present in the household or not and besides regarding the activity of the head of the household as well. The codes of these types are presented in Table 2 but their exact meaning is not relevant from our methodological point of view. The fact must be highlighted that after this stratification the MLE is not available but exact MUE exists and exact *p*-value in Table 4 tells us that variable "Unemployed" is still significant at any level. Notice, that both the intercept term and the nuisance stratum specific constant terms are *cancelled out* from the estimation.

Table 4

*Stratified estimates by the type of household's economic activity*

| Model | Type | Beta | SE(Beta) | Type | 95 percent CI lower | 95 percent CI upper | 2*1-sided=$p_2$ |
|---|---|---|---|---|---|---|---|
| Model term 7 | | | | | | | |
| Unemployed | MLE | ? | ? | asymptotic | ? | ? | ? |
| | MUE | 2.8680 | NA | exact | 2.0230 | +INF | 9.471e-050 |
| Model term 8 | | | | | | | |
| Education score | MLE | –0.2139 | 0.09588 | asymptotic | –0.4018 | –0.02596 | 0.0257 |
| | CMLE | –0.2139 | 0.09588 | exact | –0.4065 | –0.02154 | 0.02889 |

Finally, Table 4 reconsiders the model with the education score of the household's head as the predictor variable but using the stratified sample described previously. Although both the MLE and the CMLE exist the predictor "score of education" at 2 percent level is no longer significant, moreover, the *signs* of both parameters have changed. Notice again that both the intercept term and the nuisance stratum specific constant terms are cancelled out from the estimation.

## 4.4. Conflicting test results

So far, in our study only the 2*1-sided type $p_2$-value was applied to ensure the consistency with the 95 percent CI published. However, the exact *p*-values may vary depending on the special test statistic type applied, such as *scores*, *likelihood ratio* and *Wald*, especially when the sample size is extremely small. This problem is illustrated as follows.

Table 5 gives the exact test results for two predictors. Firstly, for the age of the head of the household (Age) then subsequently for the response (yes or no) whether the household has suffered poverty *ever before* (Suffered). The sample has been restricted to the *types* of the households in *Budapest* those with *more than 6 persons*

exhibiting gender of the household's head. Table 5 shows that in the case of Age only the score test exists among the asymptotic tests but its *p*-values yield different decisions at 5 percent error level depending on the type of test considered. Moreover, exact tests give the same $p = 0.07143$ value in this case (this is not necessary in general) but the *p-mid* value for the exact likelihood ratio test accepts the null-hypothesis at 5 percent level while the other exact tests reject it.

Finally, considering the question of "Suffering poverty ever before" based on the *p*-values at 5 percent error level the conclusion of the exact probability test differs from that of the other exact types and both the *p*-values and the *p-mid* values vary substantially.

Table 5

*Exact test results*

| Type of test | Statistics | DF | *p*-value | *p*-mid |
|---|---|---|---|---|
| | $H_0$ :Beta_Age = 0 | | | |
| Score | 4.317 | 1 | 0.03774 | NA |
| Likelihood ratio | ? | ? | ? | ? |
| Wald | ? | ? | ? | ? |
| Exact score_asy | 4.317 | NA | 0.07143 | 0.05357 |
| Exact score | 3.777 | NA | 0.07143 | 0.05357 |
| Exact probability | 0.03571 | NA | 0.07143 | 0.05357 |
| Exact likelihood ratio | 8.997 | NA | 0.07143 | 0.03571 |
| | $H_0$ : Beta_Suffered = 0 | | | |
| Score | 6.107 | 1 | 0.01347 | NA |
| Likelihood ratio | ? | ? | ? | ? |
| Wald | ? | ? | ? | ? |
| Exact score | 5.343 | NA | 0.03571 | 0.01786 |
| Exact probability | 0.03571 | NA | 0.07143 | 0.05357 |
| Exact likelihood ratio | 8.997 | NA | 0.03571 | 0 |

Finally, we can conclude that when a stratification of the data is required in many dimensions in order to control some factors, other covariates of the response variable are likely to become perfect predictors. In this case the exact approach of inference is appropriate providing reliable results if they exist.

## References

AGRESTI, A. [2002]: *Categorical data analysis*. Wiley. New York.

ALBERT, A. – ANDERSON, J. A.[1984]: On the existence of maximum likelihood estimates in logistic models. *Biometrica.* 71. p. 1–10.

BULL, S B. – MAK, C. – GREENWOOD, C. M. T. [2002]: A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis*. 39. p. 57–74.

CHRISTMANN, A. – ROUSSEEUW, P. J. [2001]: Measuring overlap in logistic regression. *Computational Statistics and Data Analysis*. 37. p. 65–75.

COX, D. R. – SNELL, E. J. [1989]: *Analysis of binary data*. Chapman and Hall. London.

CRAMER, J. S. [1999]: Predictive performance of the binary logit model in unbalanced samples. *The Statistician*. 48. p. 85–94.

GARTHWAITE, P. H. – JOLLIFFE, I. T. – JONES, B. [1995]: *Statistical inference*. Prentice Hall. New York.

HAVASI, É. [2005]: *A jövedelem mint az anyagi jólét és a szegénység mérőszáma*. KSH. Budapest.

HIRJI, K. F. [1992]: Exact distributions for polytomous data. *JASA*. 87. p. 487–492.

HIRJI, K. F. – MEHTA, C. R. – PATEL, N. R. [1987]: Computing distributions for exact logistic regression. *JASA*. 82. p. 1110–1117.

HIRJI, K. F. – TSIATIS, A. A. – MEHTA, C. R. [1989]: Median unbiased estimation for binary data. *The American Statistician.* 43. p. 7–11.

HUNYADI, L. [2001]: *Statisztikai következtetéselmélet közgazdászoknak* . KSH. Budapest.

KING, G. – ZENG, L. [2001]: Logistic regression in rare events data. *Political Analysis*. 9. p. 137–163.

KING, E. N. – RYAN, T. P. [2002]: A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. *The American Statistician*. Vol. 56. No. 3. p. 163–170.

*LogXact 7 User Manual, Cytel Statistical Software & Services* [2005]. Cytel. Cambridge.

MEHTA, C.R. – PATEL, N. R. [1995]: Exact logistic regression: Theory and examples. *Statistics in Medicine*. Vol. 14. p. 2143–2160.

MEHTA, C. R. – PATEL, N. R. – SENCHAUDHURI, P. [2000]: Efficient Monte Carlo methods for conditional logistic regression. *Journal of the American Statistical Association*. Vol. 95. No. 449. p. 99–108.

SANTNER, T. J.– DUFFY, D. E. [1989]: The statistical analysis of discrete data. Springer-Verlag. New York.

TRICHLER, D. [1984]: An algorithm for exact logistic regression. *Journal of the American Statistical Association*. 79. p. 709–711.