Chair of Empirical Economics
Prof. Regina T. Riphahn, Ph.D.

Master exam Summer term 2019

Subject: Microeconometrics and machine learning
Term: Summer 2019
Examiner: Prof. Regina T. Riphahn, Ph.D.

Preliminary remarks:

Number of problems: The exam consists of 4 problems.

Grading: A maximum of 60 points can be earned. The points for each task are indicated in parentheses. They correspond to the recommended time to be spent on each task (in minutes).

Allowed tools::

- Statistical distribution tables (attached)
- Calculator
- Dictionary

Important notes:

- If the statistical distribution tables included with the exam do not show the value of degrees of freedom you are looking for, note it, and use the closest value.
- If a piece of information or a necessary assumption for the calculation is missing, note it, and make a plausible assumption for the missing value.

Problem 1 (20.5 points)

Your task is to analyse the determinants of starting an extramarital affair. The following information about married individuals is are available to you:

$affair_i$	=1, if person i has had at least one affair; otherwise =0
$male_i$	=1, if person i is male; otherwise =0
$yrsmarr_i$	years of marriage of person i
age_i	age of person i in years
$naffairs_i$	Number of affairs of person i

You estimate the following model: $P(affair_i = 1) = \Lambda(\beta_1 + \beta_2 male_i + \beta_3 yrsmarr_i + \beta_4 age_i)$

```
Logistic regression                               Number of obs     =        601
                                                  LR chi2(3)        =      19.02
                                                  Prob > chi2       =     0.0003
Log likelihood = -328.17599                       Pseudo R2         =      ?????

------------------------------------------------------------------------------
affair       |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
male         |    .345624   .1983282     1.74   0.081    -.0430922    .7343402
yrsmarr      |   .1108364   .0281062     3.94   0.000     .0557492    .1659235
age          |  -.0404532   .0172147    -2.35   0.019    -.0741934    -.006713
_cons        |  -.8997599   .4029636    -2.23   0.026    -1.689554   -.1099657
------------------------------------------------------------------------------
```

Note: Round all results to the third digit.

1.1 Interpret the coefficient of *male* regarding its sign and statistical significance (2 points)

1.2 Interpret the coefficient of *male* using the odds ratio. (2 points)

1.3 Name two alternative ML test methods besides the Likelihood Ratio test shown in the output. In which situation do you expect identical test results for the three test methods? (2 points)

1.4 Calculate and interpret the marginal effect of years of marriage on the probability of starting an affair for a 40-year-old man who has been married for 10 years. (4.5 points)

1.5 What is the McFadden R^2 used for? Define the McFadden R^2 and explain its components. Calculate it for the given estimated model. (3.5 points)
Note: The value of the log likelihood function of the model above
(1) without the variable *male* is -329.700.
(2) without the constant is -330.595.
(3) only with constant value -337.688.

1.6 In the following, assume that you have 7 independent observations for the number of affairs. The number of affairs, denoted with X, follows a discrete distribution. This distribution is known to you and only depends on the parameter Θ. Determine $\hat{\theta}$ based on the observations of a subsample in the table below, using the maximum likelihood method. Use $\mathcal{L} = \prod_{i=1}^{7} P(X_i)$ as the likelihood function. (6.5 points)

Realization of X	0	1	2	3
probability of X: P(X)	$\frac{1}{\theta}$	2θ	$\frac{2}{\theta}$	$(1-\theta)$
observed frequencies	3	1	2	1

Problem 2 (8.5 points)

You are working with the data from problem 1. You are now using a Poisson regression instead of a logistic regression for the analysis of the determinants of number of extramarital affairs and obtain following results:

```
Poisson regression                                Number of obs     =        601
                                                  LR chi2(3)        =      14.08
                                                  Prob > chi2       =     0.0028
Log likelihood = -351.15167                       Pseudo R2         =     0.0197

------------------------------------------------------------------------------
naffairs     |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
male         |   .2506842   .1686504     1.49   0.137    -.0798644    .5812329
yrsmarr      |   .0809766   .0235737     3.44   0.001      .034773    .1271803
age          |  -.0292303   .0146721    -1.99   0.046    -.0579871   -.0004735
_cons        |  -1.266648   .3482522    -3.64   0.000     -1.94921    -.584086
------------------------------------------------------------------------------
```

2.1 Interpret the coefficient of *age* and determine the exact effect. (2 points)

2.2 The Poisson regression requires a deterministic relationship between the Poisson distribution parameters λ_i and $e^{\beta' x_i}$ and no autocorrelation in y. What is the consequence of violating these assumptions for the relation between variance and expected value of y? Name an alternative method of estimation for count data with less restrictive assumptions. Name two differences to the Poisson model. (3 points)

2.3 Describe the steps of a procedure for checking whether there is over-dispersion. (3.5 points)

Problem 3 (14 points)

You intend to analyse the determinants of health satisfaction based on the survey data of 9,000 respondents aged between 17 and 98 years. The following variables are available:

sat	health satisfaction (1 = low, 2 = medium, 3 = high)
age	age
agesq	age squared
inc	monthly income in thousand euros
educ	education in years
disease	= 1 if a person has a chronic disease, = 0 otherwise.

The estimates of an ordered probit model are as follows:

```
Iteration 0:   log likelihood = -5467.8914

Ordered probit regression                           Number of obs   =       9000
                                                    LR chi2(5)      =     427.16
                                                    Prob > chi2     =     0.0000
Log likelihood = -5265.9881                         Pseudo R2       =          ?

------------------------------------------------------------------------------
         sat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.02552     .00560      -4.56   0.000    -.03650      -.01454
       agesq |   .00035     .00003      11.67   0.000     .00029       .00041
         inc |   .10637     .00912      11.66   0.000     .08849       .12426
        educ |   .02188     .00581       3.76   0.000     .01049       .03328
     disease |  -.62530     .03999     -15.64   0.000    -.70368      -.54692
-------------+----------------------------------------------------------------
       /cut1 |  -1.456627   .1312109                     -1.713796   -1.199459
       /cut2 |   .0861512   .1298953                      -.168439    .3407414
------------------------------------------------------------------------------
```

3.1 Explain why the estimates do not contain a constant. Briefly describe the necessary adjustment of the model if it contained a constant. (2 points)

3.2 Interpret the coefficient *educ* in relation to the observed health satisfaction. (3 points)

3.3 Determine the age at which the latent satisfaction with health is minimized. (2 points)

3.4 Calculate the compensating variation for income for a chronically ill person. Interpret your findings. (4 points)

3.5 For the given case, you could also estimate a multinomial logit model. Explain one advantage and one disadvantage in comparison to the ordered Probit model. (2 points)

Problem 4 (17 points)

Using a cross-sectional dataset for the first half of 2019, you estimate the determinants of vacation expenditures. 150 of the 790 individuals in the sample have not yet had any vacation expenses in 2019 (i.e. *urlaub*=0). The variables of the dataset are defined as follows:

urlaub	vacation expenditures in 2019 in €
alter	age in years
bil1	=1, if basic school qualification (Hauptschulabschluss); = 0 otherwise
bil2	=1, if secondary school diploma (Realschulabschluss); =0 otherwise
bil3	=1, if A-levels (Abitur); =0 otherwise

A Tobit model provides the following results:

```
Tobit regression                          Number of obs   =      790
                                          LR chi2(?)      =      ???
                                          Prob > chi2     =      ???
Log likelihood = -745.947                 Pseudo R2       =   0.7087

---------------------------------------
      urlaub |      Coef.   Std. Err.
-------------+-------------------------
       alter |   77.85490   23.80924
        bil1 |   -370.676   1.046733
        bil2 |   -129.863   23.42910
       _cons |   6.979211   7.60e-03
-------------+-------------------------
      /sigma |   .9952454   .0110585
---------------------------------------
```

4.1 Briefly explain the difference between truncation and censoring in the context of the variable *urlaub*. What is the case for the *urlaub*? Give reasons for your response. (2 points)

4.2 Name two problems of a least-squares-estimation based on the truncated data. Formally denote the expected value of the dependent variable in such an OLS estimation and explain your notation. How many observations would you use in this example? (3 points)

4.3 Conduct a Wald-Test at the 1% significance level to test the hypothesis that vacation expenditures differ in at least one of the lower education groups (this means at most basic und secondary school diploma) from the expenditures in the highest education group (i. e. the A-levels). Denote the null and alternative hypotheses, degrees of freedom, critical value, test statistics and test decision. (6 points)

Note:

(1) The test statistic of the Wald-Test is: $W = \hat{\boldsymbol{\beta}}'\widehat{Var}(\hat{\boldsymbol{\beta}})^{-1}\hat{\boldsymbol{\beta}} \sim \chi^2_k$

(2) For the estimates vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_{bil1}\hat{\beta}_{bil2})'$ use the following estimated inverse variance-covariance matrix:
$\widehat{Var}(\hat{\boldsymbol{\beta}})^{-1} = \begin{pmatrix} 0.5 & 0 \\ 0 & -1.5 \end{pmatrix}$

4.4 Conduct a likelihood-ratio test for an overall significance of the model at the 10% significance level. Provide the null and alternative hypotheses, degrees of freedom, critical value, test statistics and test decision.
Note: The log likelihood value of a model that contains only a constant is -805.447. (4 points)

4.5 Name two weaknesses of the Tobit estimator. (2 points)