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

Master exam Summer term 2022

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

Preliminary remarks:

Grading:	A maximum of 60 points can be earned. The points for each problem are indicated in parentheses. They correspond to the recommended time to be spent on each problem (in minutes).
Allowed tools:	• Calculator • Dictionary
Important note:	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 points)

You are interested in the determinants of graduating with honors. The data set contains observations on U.S. students.

Variable	Description
honcomp	=1, if a student graduates with honors; =0 otherwise
$female$	=1, if a student is female; =0, if a student is male
read	standardized test score in reading
science	standardized test score in science

The following model is estimated: $honcomp_i = \Lambda(\beta_1 + \beta_2 female_i + \beta_3 read_i + \beta_4 science_i)$

```
Logit estimates                                   Number of obs   =        200
                                                  LR chi2(3)      =      71.05
                                                  Prob > chi2     =     0.0000
Log likelihood =  -80.11818                       Pseudo R2       =     0.3072

------------------------------------------------------------------------------
     honcomp |      Coef.   Std. Err.      z    P>|z|     [95\% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   1.482498   .447399     3.31   0.001         .605611   2.359384
        read |    .103536   .025766     4.02   0.000         .053035    .154037
     science |    .094790   .030453     3.11   0.002         .035102    .154478
       _cons |  -2.467000   .500000    -4.93   0.000      -3.447000  -1.487000
------------------------------------------------------------------------------
```

Note: Round all results to the third digit.

1.1 Interpret the coefficient of female in terms of its direction and statistical significance. (2 points)

1.2 The Stata output above shows the test statistic of a likelihood ratio test. State the null and alternative hypothesis of the test for the case at hand and briefly describe verbally the basic idea and the decision logic of the test procedure. (3 points)

1.3 Calculate the marginal effect of the variable $read_i$ for women with a standardized test score in science of 7 and in reading of 5. Interpret the result. Note: $\text{ME}(\mathbf{x}) = p \cdot (1-p) \cdot \beta_x$ (6 points)

1.4 Describe the idea of the maximum likelihood (ML) method. In what situation may the ML estimator show a local optimum? (3 points)

1.5 The ML estimator for β in the linear regression model is equivalent to the OLS estimator under the assumption of normally distributed error terms. If $\epsilon_i \sim N(0, \sigma^2 I)$, the distribution of y_i with $y_i = x_i'\beta + \epsilon_i$ is given by

$$f(y_i|x_i'\beta, \sigma_\epsilon^2) = \tfrac{1}{\sqrt{2\pi\sigma^2}} exp\{-\tfrac{1}{2\sigma^2}(y_i - x_i'\beta)^2\}.$$

Assume independence of the observations $i = 1, 2, ..., N$. Establish the likelihood function and derive the log-likelihood function for the normally distributed y. Explain verbally how coefficient estimates are obtained. (6 points)

Problem 2 (10 points)

You would like to investigate the determinants of the length of family holidays. For this purpose, you have information from 7,543 families.

Variable	Description
Numtraveld	number of travel days
Holiday	=1, if the family has gone on holiday; =0, otherwise
Agem	age of the mother in years
Numchild	number of children
lninc	logarithmised monthly household income

```
Heckman selection model -- two-step estimates   Number of obs     =      7,543
(regression model with sample selection)              Selected    =      6,086
                                                      Nonselected =      1,457

                                                Wald chi2(3)      =      72.50
                                                Prob > chi2       =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95\% Conf. Interval]
-------------+----------------------------------------------------------------
Numtraveld   |
        Agem |    .65074     .09137     7.12   0.000        .47166      .82983
    Numchild |   -.15269     .96854    -0.16   0.875      -2.05099     1.74560
       _cons |   1.95216   17.95557     0.11   0.913     -33.24011    37.14443
-------------+----------------------------------------------------------------
Holiday      |
       lninc |    .007136    .00218     3.27   0.001       .002859     .011413
        Agem |   -.000149    .00138    -0.11   0.914      -.002860     .002563
    Numchild |   -.006056    .01441    -0.42   0.674      -.034303     .022191
       _cons |    .468362    .14261     3.28   0.001       .188862     .747869
-------------+----------------------------------------------------------------
             |
      lambda | -14.19619    5.3220    -2.67   0.008     -24.62714   -3.765243
-------------+----------------------------------------------------------------
         rho |  -1.00000
       sigma |  14.19619
------------------------------------------------------------------------------
```

Note: Round all results to the third digit.

2.1 Explain in general and with reference to the analysis of trip length the meaning of endogenous sample selection. What consequences follow from endogenous sample selection for least squares estimation? Give an example why this can be the case in the current application. (3 points)

2.2 State the properties which the variable *lninc* must have in order to be an appropriate exclusion restriction. Explain whether the variable is suitable as an exclusion restriction in the present case. (3 points)

2.3 Which statistic in the Stata output above can be used to assess the selectivity of the sample? What does the statistic measure in general? Explain whether selectivity is present in the example. (3 points)

2.4 Which alternative estimation approaches could be used for this dependent variable ignoring sample selection? (1 point)

Problem 3 (14 points)

You want to examine the effect of a health care reform on the number of doctor visits. You have a data set with 2,227 observations and the following variables:

Variable	Description
numvisit	number of doctor visits in the past year
reform	=1, if the observation took place after the reform; = 0, otherwise
age	age in years
badh	=1, if poor health condition indicated; =0, otherwise
loginc	log household income

You estimate a Negbin II model with *numvisit* as dependent variable and obtain the following output:

```
Negative binomial regression                    Number of obs     =      2,227
                                                LR chi2(5)        =     300.46
Dispersion     =       mean                     Prob > chi2       =     0.0000
Log likelihood = -4563.3905                     Pseudo R2         =     0.0319

------------------------------------------------------------------------------
    numvisit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    reform   |   -.137408    .051133    -2.69   0.007     -.237627   - .037188
    age      |    .005570    .002399     2.32   0.020      .000866     .010273
    educ     |   -.005319    .011477    -0.46   0.643     -.027814     .017174
    badh     |   1.131247    .074780    15.13   0.000      .984679    1.277814
    loginc   |    .037129    .007119     5.22   0.004     0.023175     .051083
    _cons    |   -.407475    .533606    -0.76   0.445    -1.453323     .638374
-------------+----------------------------------------------------------------
    /lnalpha |    .002129    .047539                     -.0910466    .0953047
-------------+----------------------------------------------------------------
       alpha |   1.002131   .0476408                      .9129752    1.099994
------------------------------------------------------------------------------
LR test of alpha=0: chibar2(01) = 2760.88              Prob >= chibar = 0.000
```

Note: Round all results to the third digit.

3.1 Interpret the magnitude and statistical significance of the estimated coefficients of the variables *reform* and *loginc*. (5 points)

3.2 Briefly explain the meaning of overdispersion and its consequence for Poisson estimation results. (2 points)

3.3 Explain two differences between a Poisson and a Negbin II model. (2 points)

3.4 Describe how you can test for overdispersion in this example with the given output. Also, state the hypotheses, the test statistic, and your test decision. Does the test confirm overdispersion? (5 points)

Note: *For the Negbin II model we have:* $Var(y_i|x_i) = \lambda + \alpha\lambda^2$.

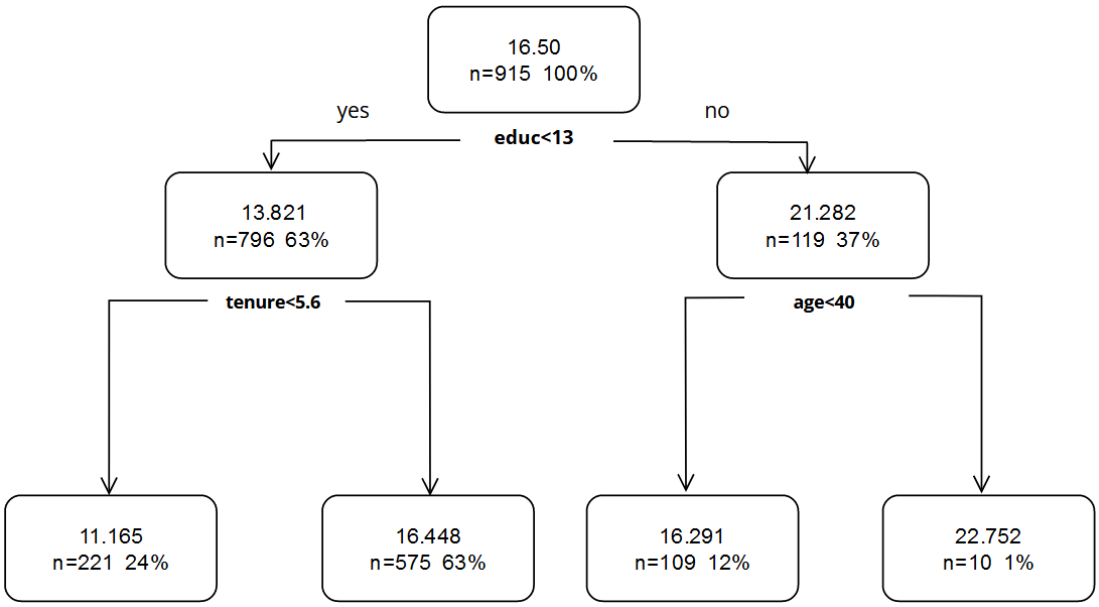
Problem 4 (16 points)

You would like to fit a model to predict hourly wages based on the following data:

Variable	Description
hwage	hourly wage in Euro
age	age in years
educ	education in years
tenure	tenure with current employer in years

$n = 915$

You have fitted a regression tree using hourly wage as dependent variable and three features (age, educ, tenure) as explanatory variables.



4.1 Is the fitted tree a classification tree? Explain your answer. (2 points)

4.2 Interpret the right leave (22.752, n= 10, 1%) of the tree. (4 points)

4.3 Mention two weaknesses of trees. (2 points)

4.4 Explain the purpose of bagging and briefly describe the approach. (3 points)

4.5 Describe the procedure of k-Fold Cross Validation with $k = 10$ in general and the decision that is based on it. (5 points)