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

Master exam Summer term 2023

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:	• Answers in German will be graded as well. • 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 (12 points)

The amount of credit (*credit*, measured in euro) granted to a firm is modeled as a function of log sales (*ln_sales*, measured in millions of euro) and firm size (*size*, measured as number of employees). In the given dataset, 1,845 of 5,658 firms have been granted a credit. Estimation of a Tobit model yields the following output for the dependent variable *credit*:

	Tobit	
Variable	coefficient	std.error
ln_sales	7.213	(0.027)
size	-0.058	(0.000)
constant	-37.676	(0.105)
σ	0.995	(0.011)
$\Phi(\mathbf{x}'\boldsymbol{\beta}/\sigma)$	0.193	
observations (N)	5,658	

Note: Round all results to the third digit.

1.1 Using the variable *credit* as an example, briefly explain the difference between truncation and censoring. For this example, state the number of observations used in the truncated regression. (3 points)

1.2 What is the direct interpretation of Tobit coefficients? Interpret the coefficient estimate of the variable *ln_sales*. (2 points)

1.3 Calculate and interpret the marginal effect of firm size on the censored amount of credit (i.e. for all firms). (3 points)

1.4 The Tobit model was estimated using the following log-likelihood function. Briefly explain what the two sums $\sum_{y_i=0}$ and $\sum_{y_i>0}$ each stand for. What assumption regarding $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ is made in the Tobit model? (3 points)

$$\ln L = \sum_{y_i=0} \ln \Phi\left(\frac{0 - \mathbf{x}_i'\boldsymbol{\beta}_1}{\sigma}\right) + \sum_{y_i>0} \ln \frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}_2}{\sigma}\right)$$

1.5 Explain one weakness of the Tobit estimator. (1 point)

Problem 2 (17 points)

Determinants of the decision to apply to graduate school are analysed with an ordered logistic model. The data set contains observations on U.S. college juniors. The variables and regression results are presented below:

Variable		**Description**
application	=	application to graduate school (= 0 unlikely; = 1 somewhat likely; = 2 very likely)
pedu	=	1, if at least one parent has a graduate degree; 0, else
public	=	1, if undergraduate institution is public; 0, if private
gpa	=	student's grade point average

The following model is estimated: $application_i = f(\beta_1 pedu_i + \beta_2 public_i + \beta_3 gpa_i)$

```
Ordered logistic regression                       Number of obs   =        400
                                                  LR chi2(3)      =      24.18
                                                  Prob > chi2     =     0.0000
Log likelihood = -358.51244                       Pseudo R2       =     0.0326

------------------------------------------------------------------------------
 application |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pedu |   1.047664   .2657891     3.94   0.000     .5267266    1.568601
      public |  -.0586828   .2978588    -0.20   0.844    -.6424754    .5251098
         gpa |   .6157458   .2606311     2.36   0.018     .1049183    1.126573
-------------+----------------------------------------------------------------
       /cut1 |   2.203323   .7795353                      .6754621    3.731184
       /cut2 |   4.298767   .8043147                       2.72234    5.875195
------------------------------------------------------------------------------
```

Note: Round all results to the third digit.

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

2.2 Formally explain the relationship between the observed application intention y_i and the latent application intention y_i^*. Define the quantities used. (3 points)

2.3 How does the estimation change if another category (3 = certain) is added to the dependent variable *application*? Explain the changes in terms of the parameters and interpretation. (3 points)

2.4 A student attends a public undergraduate institution. By how much would the student's GPA have to increase to compensate the resulting decline in the intention to apply to graduate school? (3 points)

2.5 Alternatively, you estimate a multinomial logit model.

i. How many parameters are estimated in total? Explain your solution briefly. (2 points)

ii. In this multinomial logit model, how would you test the hypothesis that parental education does not contribute to the decision to apply to graduate school? Suggest the test procedure, provide the respective degress of freedom of the test statistic, the null and alternative hypothesis. (4 points)

Problem 3 (7 points)

You estimate a Poisson model in which you want to explain the number of children a woman has. You have the following information for 14,786 women:

kids	=	number of children
age	=	age in years
educ	=	education in years

You get the following output:

```
Iteration 0:   log pseudolikelihood = -38999.8
Iteration 1:   log pseudolikelihood = -38795.8

Poisson regression                                Number of obs   =      14786
                                                  Wald chi2(2)    =     100.67
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -38795.8                   Pseudo R2       =     0.0058

------------------------------------------------------------------------------
             |               Robust
        kids |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0943417   .0270359     3.49   0.000     .0413523     .147331
        educ |  -.0106121   .0027562    -3.81   0.000    -.0159142     -.00511
       _cons |   14.56069    1.87507     7.77   0.000     10.88562    18.23576
------------------------------------------------------------------------------
```

Note: Round all results to the third digit.

3.1 Interpret the coefficient of education in terms of its magnitude and statistical significance. (2 points)

3.2 Briefly explain the concept of underdispersion. What does this mean for the estimation results? (2 points)

3.3 You want to validate the goodness of fit of your model using the Akaike Information Criterion (AIC). Calculate the value of the AIC for the estimated model. State an alternative goodness of fit measure and its advantage over the AIC. (*hint*: AIC = 2k - 2lnL.) (3 points)

Problem 4 (10 points)

You analyze the duration of unemployment (measured in weeks) with a Weibull regression. Your dataset contains information on 3,674 unemployed individuals. The following explanatory variables are available:

$assist_i$	=1, if individual is assisted by the employment agency in job search; 0 otherwise
age_i	age in years i

```
Weibull regression -- log relative-hazard form

No. of subjects =         3674                     Number of obs =   3674
No. of failures =         2571
Time at risk    =          557
LR chi2(1)      =       243.86
Log likelihood  =   -1045.4234                     Prob > chi2   =  0.0000

------------------------------------------------------------------------------
_t           | Coef.        Std. Err.      z     P>|z|   [95% Conf. Interval]
-------------+----------------------------------------------------------------
assist       |  2.317153     .4855492    4.77    0.000   1.365494    3.268812
age          | -.2874122     .0391988   -3.41    0.001  -.2106459   -.0569896
_cons        | -.2610042     1.479498   -0.18    0.860  -3.160767    2.638759
-------------+----------------------------------------------------------------
/ln_p        |  .4570259     .1665073    2.74    0.006   .1306776    .7833742
-------------+----------------------------------------------------------------
p            |   1.57937     .2629766                      1.1396    2.188845
1/p          |  .6331639     .1054264                     .4568619    .8775007
------------------------------------------------------------------------------
```

Note: Round all results to the third digit.

4.1 Explain the terms *flow sample* and *stock sample*. Name one problem that may arise with *stock samples*. (3 points)

4.2 In another Weibull estimation, the patient's gender is included in the model as an additional explanatory variable. The estimation yields a log-likelihood value of -1042.3425. Test whether the explanatory power of the model has improved significantly. Report the test statistic, degrees of freedom, and critical value at the $\alpha = 0.05$ significance level. Calculate the empirical test statistic and make a test decision. (7 points)

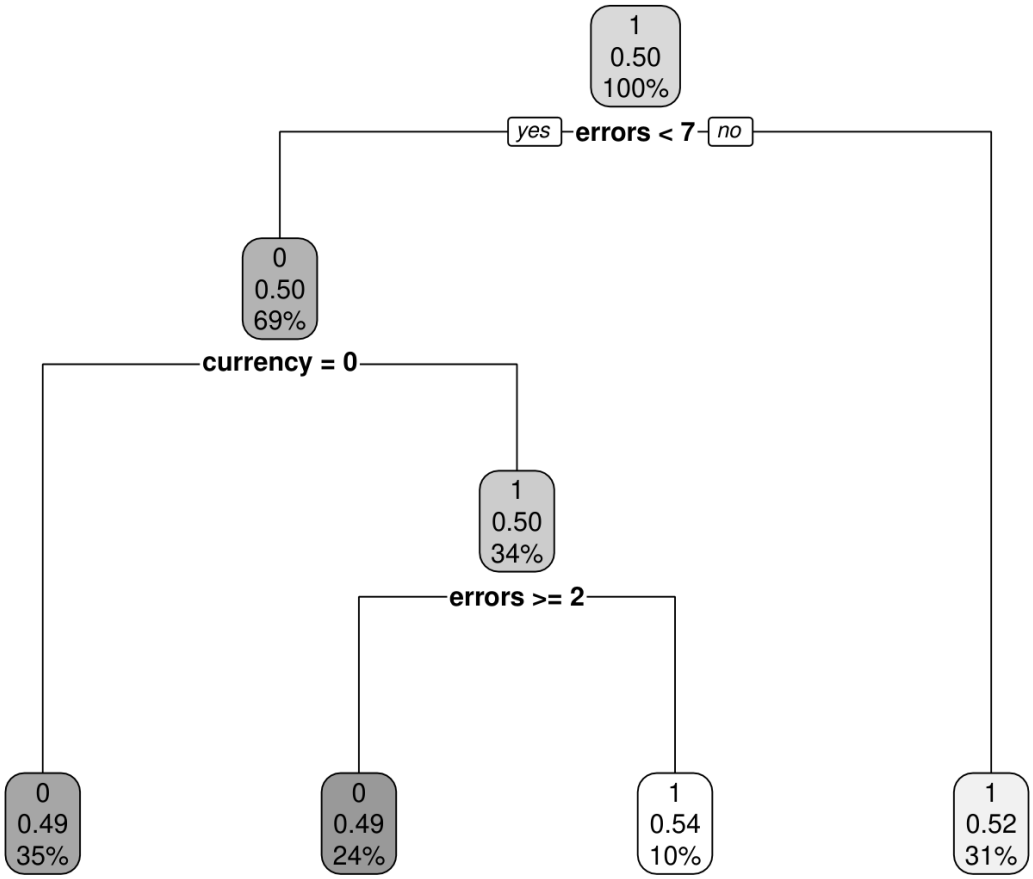
Problem 5 (14 points)

5.1 Explain whether and why the regression tree analysis is a supervised or unsupervised learning method. Discuss the idea of regression trees intuitively and also mention the criterion or function to be minimized. (6 points)

You are given a dataset of 300 corporate emails. Your job is to predict whether an email is spam or not. The following variables are available:

$spam_i$	=1, if an email i is spam; 0, otherwise
$errors_i$	number of grammatical errors in the text i
$currency_i$	=1, if an email i contains words for money; 0, otherwise

You have fitted a tree using *spam* as dependent variable and two features (*errors*, *currency*) as explanatory variables.



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

5.3 Interpret the left leaf (0, 0.49, 35%) of the tree. How many emails are in this leaf? (4 points)

5.4 Mention two weaknesses of trees. (2 points)

Table 1: Percentiles of χ^2 distribution
Cell entry: c, so that $P[\chi^2_n \leq c] = P$, with n degrees of freedom

n \ P	0.005	0.01	0.025	0.05	0.1	0.25	0.5	0.75	0.9	0.95	0.975	0.99	0.995
1	0.00004	0.0002	0.001	0.004	0.016	0.102	0.455	1.323	2.706	3.842	5.024	6.635	7.879
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
35	17.19	18.51	20.57	22.47	24.80	29.05	34.34	40.22	46.06	49.80	53.20	57.34	60.27
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77
45	24.31	25.90	28.37	30.61	33.35	38.29	44.34	50.98	57.51	61.66	65.41	69.96	73.17
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49