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

Master exam summer term 2024

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 (21 points)

Determinants of health satisfaction are analyzed using survey data on 7,000 individuals. The following variables are available:

sat	=	Health satisfaction (1 = low, 2 = medium, 3 = high)
age	=	Age in years
agesq	=	Age squared in years
inc	=	Monthly income in thousands of Euros
educ	=	Education in years
disease	=	Chronic illness (=1, otherwise=0).

The estimation results of an ordered probit model are as follows:

```
Iteration 0:   log likelihood = -5265.9881
Iteration 1:   log likelihood = -5265.9881

Ordered probit regression                         Number of obs   =       7000
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        |  -1.4566   .1312109                      -1.713796   -1.199459
/cut2        |   .08615   .1298953                       -.168439    .3407414
------------------------------------------------------------------------------
```

Note: Round all results to the third digit.

1.1 Formally explain the relationship between the observed health satisfaction y_i and the latent health satisfaction y_i^*. Define the terms used. (3 points)

1.2 Explain in detail why the estimate does not contain a constant. Briefly describe the necessary change to the model if it is to be estimated with a constant. (3 points)

1.3 Interpret the coefficient of *educ* statistically and in terms of its meaning. (4 points)

1.4 Calculate the compensating variation in income for a chronically ill person. Interpret the result. (3 points)

1.5 What is the purpose of the McFadden R^2? Define and explain the measure. Calculate it for the given estimation. Assume that the log-likelihood in the model with only the constant is -5467.8914. (4 points)

1.6 In the given case, you could also estimate a multinomial logit model (MNL). Describe and explain one advantage of MNL and one disadvantage compared to the ordered probit model. (4 points)

Problem 2 (7 points)

For a sample of 11,874 students, labor market participation while studying is analyzed using a linear probability model as well as probit and logit models. The binary dependent variable is y=1 if a person is working and y=0 otherwise.

2.1 Name two disadvantages of a linear probability model. (2 points)

2.2 You first consider the age of the students as an explanatory variable and estimate a logit model (Logit 1). Then you also include gender in the model and estimate it again (Logit 2). The following table compares both models. *Note*: AIC denotes the Akaike information criterion and BIC the Schwarz information criterion.

Model	Obs	df	AIC	BIC
Logit 1	11874	2	9710.496	9725.26
Logit 2	11874	3	9601.815	9623.962

Based on the AIC and BIC, which model is preferable? Explain your answer. Explain which criteria are taken into account when calculating the two measures of fit. (3 points)

2.3 What is the main difference between logit and probit models? (2 points)

Problem 3 (11 points)

You estimate a Poisson model in which you want to analyze the relationship between the days a student is absent during the school year and the characteristics of the student. You have the following information for 316 students:

daysabs	=	Number of days absent during the school year
mathnce	=	Math standardized test score, 1...100 points
langnce	=	Language standardized test score, 1...100 points
female	=	Female (=1, otherwise=0)

You get the following output:

```
Iteration 0:   log likelihood = -1547.9709
Iteration 1:   log likelihood = -1547.9709

Poisson regression                              Number of obs   =        316
                                                LR chi2(3)      =     175.27
                                                Prob > chi2     =     0.0000
Log likelihood = -1547.9709                     Pseudo R2       =     0.0536

------------------------------------------------------------------------------
     daysabs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mathnce |  -.0035232   .0018213    -1.93   0.053     -.007093    .0000466
     langnce |  -.0121521   .0018348    -6.62   0.000    -.0157483   -.0085559
      female |   .4009209   .0484122     8.28   0.000     .3060348     .495807
       _cons |   2.286745   .0699539    32.69   0.000     2.149638    2.423852
------------------------------------------------------------------------------
```

Note: Round all results to the third digit.

3.1 Name two advantages of the Poisson model over the linear model. (2 points)

3.2 Interpret the coefficient of the language standardized test score in terms of its magnitude and statistical significance. (2 points)

3.3 Briefly explain the concept of overdispersion. What does this mean for the estimation results? (3 points)

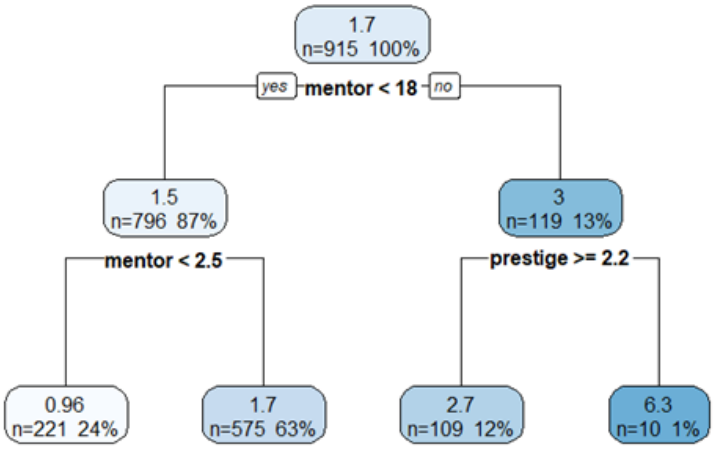
3.4 Name two differences between a Poisson and a Negbin model. (4 points)

Problem 4 (9 points)

You would like to fit a model to predict the number of graduate students' publications based on the following data:

articles	Number of articles published by a graduate student.
mentor	Number of articles published by student's mentor.
prestige	Prestige of the graduate program.

You have fitted a tree using *articles* as dependent variable and two features *(mentor, prestige)* as explanatory variables.



4.1 Interpret the right leaf (6.3, n=10, 1%) of the tree. (4 points)

4.2 Explain the method of Random Forest. Mention one advantage and one disadvantage of Random Forest compared to a tree. (5 points)

Problem 5 (12 points)

5.1 Describe the procedure of Leave-one-out Cross Validation (LOOCV). (4 points)

5.2 Explain the motivation behind splitting the original dataset in a training and a test dataset. Which task is performed in the training dataset? Which one in the test dataset? (4 points)

5.3 Discuss the similarities and differences between Ridge Regression and the Least Absolute Shrinkage and Selection Operator (LASSO). How is the elastic net related to LASSO and Ridge? (4 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