Lehrstuhl für Statistik und empirische Wirtschaftsforschung
Prof. Regina T. Riphahn, Ph.D.

Masterprüfung Sommersemester 2021

Fach: Mikroökonometrie und Maschinelles Lernen
Semester: Sommersemester 2021
Prüfer: Prof. Regina T. Riphahn, Ph.D.

Vorbemerkungen:

Anzahl der Aufgaben:	Die Klausur besteht aus 5 Aufgaben, die alle bearbeitet werden müssen. **Es wird nur der Lösungsbogen eingesammelt.**
Bewertung:	Es können maximal 60 Punkte erworben werden. Die maximale Punktzahl für jede Aufgabe ist in Klammern angegeben. Sie entspricht der für die Aufgabe empfohlenen Bearbeitungszeit in Minuten.
Erlaubte Hilfsmittel:	• Tabellen der statistischen Verteilungen (sind der Klausur beigefügt) • Taschenrechner • Fremdwörterbuch
Wichtige Hinweise:	• Sollte es vorkommen, dass die statistischen Tabellen, die dieser Klausur beiliegen, den gesuchten Wert der Freiheitsgrade nicht ausweisen, machen Sie dies kenntlich und verwenden Sie den nächstgelegenen Wert. • Sollte es vorkommen, dass bei einer Berechnung eine erforderliche Information fehlt, machen Sie dies kenntlich und treffen Sie für den fehlenden Wert eine plausible Annahme.

Aufgabe 1 (19 Punkte)

Determinanten der Häufigkeit des Bierkonsums werden mit einem geordneten Probit Modell analysiert. Es liegen folgende Variablen und Regressionsergebnisse vor:

Variable	Beschreibung
bier	Bierkonsum (=1, nie; =2, selten; =3, manchmal; =4, oft)
educ	Jahre der Schul- und Berufsausbildung
alter	Alter in Jahren
*alter*2	Alter quadriert
mann	= 1, falls Mann; =0, falls Frau

```
Ordered probit regression                         Number of obs   =       8793
                                                  LR chi2(?)      =        ?
                                                  Prob > chi2     =        ?
Log likelihood = -10671.792                       Pseudo R2       =     0.0846

------------------------------------------------------------------------------
        bier |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mann |   1.035621   .0244472    42.36   0.000     .9877053    1.083537
        educ |   .0253862   .0043604     5.82   0.000       .01684    .0339324
       alter |   .0149494   .0039597     3.78   0.000     .0071885    .0227102
      alter2 |  -.0001927   .0000379    -5.09   0.000     -.000267   -.0001185
-------------+----------------------------------------------------------------
       /cut1 |   .5305678   .1073922                      .3200829    .7410526
       /cut2 |   1.363246   .1078928                       1.15178    1.574712
       /cut3 |   2.371798   .1091321                      2.157904    2.585693
------------------------------------------------------------------------------
```

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

1.1 Stellen Sie formal mit Hilfe des Schwellenwertkonzepts den Zusammenhang zwischen dem latenten Bierkonsum y_i^* und dem beobachteten Bierkonsum y_i für die Antwortkategorien *nie* und *selten* dar. (2 Punkte)

1.2 Interpretieren Sie das Vorzeichen des Koeffizientenschätzers der Variable *mann* in Bezug auf die Wahrscheinlichkeit nie bzw. selten Bier zu konsumieren. (2 Punkte)

1.3 Erklären Sie kurz, warum das Modell ohne Konstante geschätzt wurde. Bestimmen Sie die geschätzten Werte der Konstante ($\hat{\beta}_{cons}$) und des Parameters von *mann* ($\hat{\beta}_{mann}$), wenn der erste Schwellenwert auf 0 normalisiert und das Modell mit Konstante geschätzt wird. (3 Punkte)

1.4 Führen Sie einen Likelihood-Ratio-Test auf Gesamtsignifikanz am 1%-Signifikanzniveau durch. Dabei sind Null- und Alternativhypothese, Teststatistik, Freiheitsgrade, kritischer Wert und Schlusslogik anzugeben. Interpretieren Sie Ihre Testentscheidung. *Hinweis*: Nutzen Sie die Angabe des Pseudo (=McFadden) $R^2 = 1 - \frac{\ln L_u}{\ln L_r}$. (7 Punkte)

1.5 Sie schätzen alternativ ein multinomiales Logit Modell.

i. Wie viele Parameter werden insgesamt geschätzt? Erklären Sie Ihren Lösungsweg kurz. (2 Punkte)

ii. Zeigen Sie allgemein, wie der Koeffizientenschätzer für die Kategorie *selten* ($j = 2$) aus einem Modell mit der Basiskategorie *nie* ($j = 1$) auf ein Modell mit der Basiskategorie *oft* ($j = 4$) umgerechnet werden kann. (3 Punkte)

Aufgabe 2 (9 Punkte)

Der Umfang sportlicher Aktivität wird mit einem Tobit-Modell analysiert. Es liegen folgende Variablen und Regressionsergebnisse vor:

Variable	Beschreibung
sport_hrs	Umfang sportlicher Aktivität (in Stunden pro Woche)
sport_yes	=1, wenn *sport_hrs* >0; =0 sonst
educ	Schulausbildung in Jahren
kids	Anzahl der eigenen Kinder

```
Tobit regression                                  Number of obs   =       3412
                                                  LR chi2(2)      =    3404.59
                                                  Prob > chi2     =     0.0000
Log likelihood = -2303.7544                       Pseudo R2       =     0.4249

------------------------------------------------------------------------------
   sport_hrs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1982764   .0029339    67.58   0.000      .192524    .2040288
        kids |  -.0050942   .0064673    -0.79   0.431    -.0177744     .007586
       _cons |  -1.967411   .0422941   -46.52   0.000    -2.050335   -1.884486
-------------+----------------------------------------------------------------
      /sigma |   .4996532   .0074094                      .4851259    .5141806
------------------------------------------------------------------------------
Obs. summary:       1025  left-censored observations at activity<=0
                    2387     uncensored observations
                       0 right-censored observations
```

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

2.1 Erläutern Sie am Beispiel der Variable *sport_hrs* knapp den Unterschied zwischen Stutzung und Zensierung. (2 Punkte)

2.2 Berechnen und interpretieren Sie den marginalen Effekt der Bildung auf den Umfang sportlicher Aktivität (i) für Sportler und (ii) für die gesamte Bevölkerung. *Hinweise*: In der vorliegenden Aufgabe ist der Korrekturfaktor für das gestutzte Modell $(1 - \delta(\alpha)) = 0,618$, $\Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) = 0,832$ und der Anteil der unzensierten Beobachtungen beträgt 70 Prozent. (3 Punkte)

2.3 Nennen Sie zwei Schwächen des Tobit-Schätzers. (2 Punkte)

2.4 Skizzieren Sie grob, wie ein Test auf Heteroskedastie im Fall des Tobit-Modells durchgeführt werden kann. Wie modellieren Sie hierfür Heteroskedastie? (2 Punkte)

Aufgabe 3 (12 Punkte)

In einer Impfstudie werden 231 zufällig ausgewählte Personen gemäß ihrem Impfstatus den Treatment- und Kontrollgruppen zugeteilt. Sie möchten die Verweildauer von Beginn der Studie bis zum Eintreten der Krankheit untersuchen. Folgende Variablen sind in dem Datensatz enthalten:

Variable	Beschreibung
treated	=1 wenn Studienteilnehmer geimpft wurde; =0 sonst
t_healthy	Wochen bis zum Auftreten der Krankheit falls erkrankt, sonst Zeit seit Beginn der Studie

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

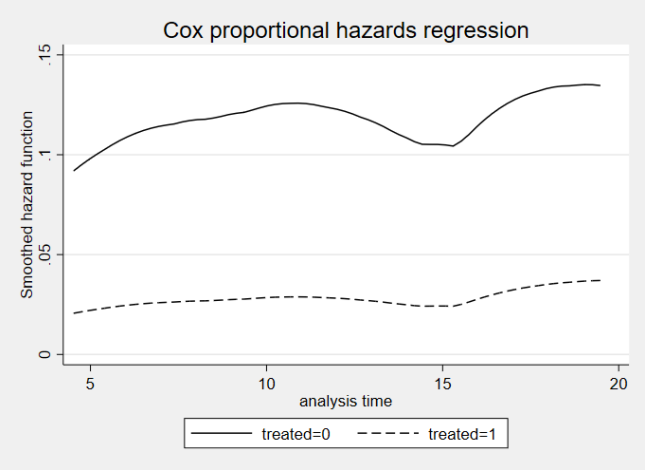
3.1 Diskutieren Sie kurz, ob es in dem gegebenen Kontext zu Rechtszensierung der Episoden kommen könnte. (2 Punkte)

3.2 Sie schätzen ein Cox Proportional Hazard Modell mit *t_healthy* als abhängige und *treated* als unabhängige Variable und erhalten folgenden Output. Interpretieren Sie das Ergebnis inhaltlich und statistisch. (4 Punkte)

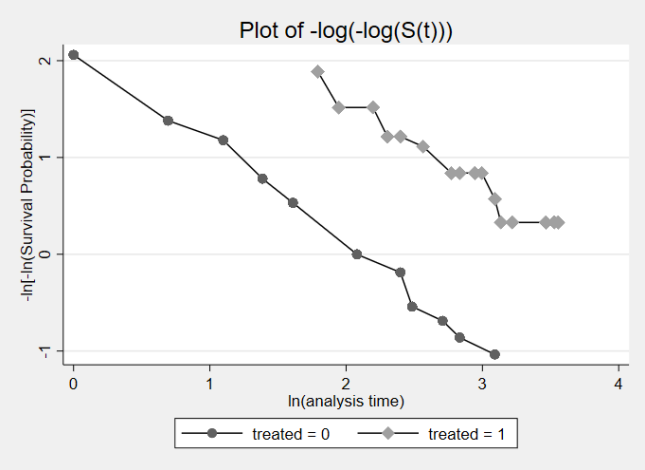
```
------------------------------------------------------------------------------
             |               Robust
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     treated |  -1.557952   .1612415    -9.66   0.000     -1.87398   -1.241925
------------------------------------------------------------------------------
```

3.3 Zeigen Sie, dass in einem Cox Modell mit nur einer binären unabhängigen Variable (*treated*) das Hazard Ratio zwischen zwei Beobachtungen i (mit $X_i = treated = 1$) und j ($X_j = treated = 0$) konstant ist. (Hinweis: $\lambda_k(t) = \lambda_0(t)exp(\beta' X_k)$ für $k = 1, \ldots, n$.) (4 Punkte)

3.4 Folgend betrachten Sie die durch das Modell geschätzten Hazard (siehe a) und die entsprechenden transformierten Survival Funktionen (siehe b) getrennt für die beiden zugrundeliegenden Gruppen. Würden sie anhand einer der Grafiken die Proportionalitätsannahme bestätigen? Erläutern Sie Ihre Antwort. (2 Punkte)



(a) Prognostizierte Hazards



(b) Transformation von S(t)

Aufgabe 4 (7 Punkte)

1025 vollzeitbeschäftigte Personen wurden nach der Anzahl während der vergangenen Woche in Büro verbrachten Arbeitstage befragt. Folgende Variablen stehen in dem Datensatz zur Verfügung:

Variable	**Beschreibung**
d_in_office	Anzahl Tage im Büro in der vergangenen Woche
$female$	=1 wenn die Person weiblich ist; =0 sonst
$married$	=1 wenn die Person verheiratet ist; =0 sonst
$under15$	Anzahl Kinder unter 15 Jahre
car	=1 wenn die Person ein Auto besitzt; =0 sonst
$internet$	=1 wenn die Person schnelles Internet zu hause hat; =0 sonst

Sie untersuchen die Determinanten der Anzahl der Büroarbeitstage. Dazu schätzen Sie ein Poisson Modell mit folgendem Output:

```
Poisson regression                              Number of obs     =      1,025
                                                Wald chi2(5)      =      64.64
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -769.21732               Pseudo R2         =     0.0558

------------------------------------------------------------------------------
             |               Robust
d_in_office  |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
female       |    .086736   .1202713     0.72   0.471    -.1489914    .3224634
married      |   .2154893   .1186576     1.82   0.069    -.0170753    .4480538
under15      |   .1990644   .0622003     3.20   0.001     .0771541    .3209747
car          |   .5255471   .3140864     1.67   0.094    -.0900509    1.141145
internet     |  -.8825217   .1823894    -4.84   0.000    -1.239998   -.5250451
_cons        |  -1.493201   .3450456    -4.33   0.000    -2.169478   -.8169236
------------------------------------------------------------------------------
```

4.1 Interpretieren Sie den geschätzten Koeffizienten der Variable *internet* inhaltlich und statistisch. Runden Sie auf die dritte Nachkommastelle. (3 Punkte)

4.2 Erläutern Sie kurz, was unter Überstreuung zu verstehen ist. Was bedeutet das für die Schätzergebnisse? (2 Punkte)

4.3 Erläutern Sie zwei Unterschiede zwischen einem Poisson und einem Negbin Modell. (2 Punkte)

Aufgabe 5 (13 Punkte)

Ihnen liegen Daten aus einer Befragung zu Löhnen und persönlichen Charakteristika von 914 Personen vor. Sie führen eine Lasso-Regression mit der abhängigen Variable *ln_wage* durch und erhalten folgenden Output:

```
Lasso linear model                          No. of obs        =        914
                                            No. of covariates =        277
Selection: Cross-validation                 No. of CV folds   =         10

------------------------------------------------------------------------------
         |                              No. of      Out-of-      CV mean
         |                             nonzero       sample   prediction
      ID |     Description      lambda     coef.    R-squared        error
---------+--------------------------------------------------------------------
       1 |    first lambda    .9090511         0       0.0010     18.33331
      23 |   lambda before    .1174085        58       0.3543     11.82553
    * 24 | selected lambda    .1069782        64       0.3547     11.81814
      25 |    lambda after    .0974746        66       0.3545      11.8222
      28 |     last lambda    .0737359        80       0.3487     11.92887
------------------------------------------------------------------------------
* lambda selected by cross-validation.
```

5.1 Basierend auf dem oberen Output, erläutern Sie kurz das Ergebnis der Lasso-Schätzung bezüglich des gewählten Wertes von Parameter λ und der Anzahl von eliminierten Variablen. (2 Punkte)

5.2 Kann man die Koeffizienten der Lasso-Schätzung direkt interpretieren? Erläurtern Sie kurz. Welche Voraussetzungen sollten die gültigen Standardfehler für die statistische Inferenz erfüllen? (2 Punkte)

5.3 Nennen Sie eine Schwäche der Lasso Schätzung. (1 Punkt)

5.4 Erklären Sie den Ansatz von Post-Lasso-KQ-Schätzung und einen Vorteil gegenüber der Lasso-Schätzung. (2 Punkte)

5.5 Alternativ führen Sie eine Ridge Regression durch. Welche Bedeutung hat der Parameter λ? Welche Werte erwarten Sie für die Schätzer $\hat{\beta}^{ridge}$ in einer Ridge Regression, wenn $\lambda \approx 0$? (2 Punkte)

5.6 Erläutern Sie in Ihren Worten das Vorgehen bei k-Fold Cross Validation mit $k = 10$ im Allgemeinen. (4 Punkte)

Tabelle 3: Perzentile der χ^2-Verteilung

Zelleneintrag: c, sodass Prob$\left[\chi_n^2 \leq c\right] = P$, mit n Freiheitsgraden

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

Quelle: In R generiert