Prof. Dr. Joachim Winter

Department of Economics
LMU Munich

Summer Semester 2015

Microeconometrics

Final Exam

Tuesday, July 14, 2015, 9.00–11.00

- The exam is set for 120 minutes and a total of 120 credits.
- All questions must be answered. Thus, you should not spend more than 1 minute for 1 credit, on average.
- This is an "open notes" exam. You may use your class and lecture notes. Books or print-outs of scientific papers are NOT allowed. Furthermore you may use a non-programmable calculator, but we anticipate that you won't need it.
- There are no nasty pitfalls in these questions. If you nevertheless think that a question misses a piece of information which you need to come up with an answer, you may make an appropriate assumption yourself.
- Good luck!

Question 1 **$5 + 4 + 3 + 15 + 2 + 4 + 4 + 3 = 40$ credits**

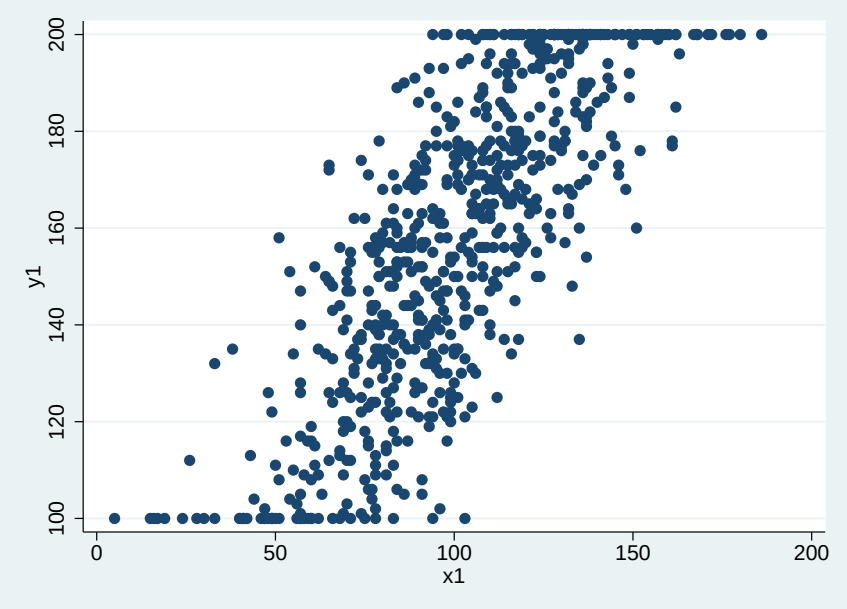
Part 1

A friend of yours shows you the following `Stata` output (graph and code) and asks you what it is all about.

```
. sum y1 x1 x2
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          y1 |       700    157.4914    30.87468       100        200
          x1 |       700      100.71    29.14652         5        186
          x2 |       700    .4828571    .5000634         0          1

. tab x2
         x2 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        362       51.71       51.71
          1 |        338       48.29      100.00
------------+-----------------------------------
      Total |        700      100.00

. scatter y1 x1
. graph export "Graph.pdf", as(pdf) replace
(file Graph.pdf written in PDF format)
```



(a) Explain the nature of all variables. State an observation rule.

(b) Describe an (economic) application in which the type of data you described in (a) could occur. Give examples of all variables.

(c) What specific model is implied if you consider the nature of the data (and if y_1 is the outcome variable and x_1 and x_2 are the explanatory variables)? Provide the latent variable model specification and explain the underlying assumptions. (*Hint*: We have discussed a similar, but not identical model in the lecture and the class.)

(d) How does the conditional density of the observed outcome look like? State it and derive the log-likelihood for this model. If you have to define new variables, give a short explanation of what they stand for.

Part 2

The `Stata` program named `xyz` models the likelihood function from above. When you run it, you get the following Stata output:

```
. ml model lf xyz (y1 = x1 x2 ) (), vce(opg)

. ml maximize, nolog noheader
initial:       log likelihood =     -<inf>  (could not be evaluated)
feasible:      log likelihood = -80081.429
rescale:       log likelihood = -3522.2039
rescale eq:    log likelihood = -3003.4133
------------------------------------------------------------------------------
             |                 OPG
          y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
eq1          |
          x1 |   1.026432   .0309502    33.16   0.000     .9657708    1.087093
          x2 |   15.53259   1.632343     9.52   0.000     12.33325    18.73192
       _cons |   48.05065   3.312752    14.50   0.000     41.55778    54.54353
-------------+----------------------------------------------------------------
eq2          |
       _cons |   20.53029   .6273257    32.73   0.000     19.30076    21.75983
------------------------------------------------------------------------------
```

(e) What is estimated in the last row (`eq2 _cons`) of the above regression table?

(f) Interpret the coefficients of x_1 and x_2.

(g) What does the variance-covariance estimation option `vce(opg)` mean? How does it relate to the default `vce(oim)`? Explain the underlying principle and state formulas where appropriate. (*Hint*: You do not need to derive any quantities yourself here.)

(h) Your friend suggests to use the `heckman` command in Stata to estimate an alternative model. What are the differences to the model described above? Do you think it is suitable for the DGP described above? Explain your answer briefly.

Question 2 **$2 + 3 + 5 + 10 = 20$ credits**

You would like to analyze the factors that influence the expenditures spent on cultural activities (like visiting a museum or the opera).

(a) Your first approach is to ask people in a museum how much money they spend on cultural activities. What is wrong with this sampling design?

Suppose you now have a random sample of the population. In this sample, you observe expenditures for cultural activities. Unfortunately, the dataset contains only two explanatory variables – age and gender. After a quick look at the data, you come to the conclusion that the outcome variable is highly skewed, with many people who do not spend anything on cultural activities.

(b) Due to the many zeros in your outcome variable, you decide to use a standard two-part model – a binary outcome model in the first part and a linear regression with the logarithm of expenditure as the dependent variable in the second part. What advantage does this model have compared to the censored regression (Tobit) model?

(c) Now, compare the standard two-part model with the bivariate sample selection model (the so-called Heckman model). What is the main advantage of the sample selection model? Do you think that this model extension is helpful in this case?

(d) Briefly explain how you can implement Heckman's two-step estimator. What is an advantage of this approach compared to the one-step maximum likelihood estimation of the sample selection model? Under what circumstances would you prefer maximum likelihood estimation over the two-step approach?

Question 3 **$2 + 3 + 3 + 4 + 8 = 20$ credits**

This question is based on the "fishing mode" example in the chapter on multinomial models in the textbook by Cameron and Trivedi (2005). (All relevant information is provided below.)

The dependent variable takes the value 1, 2, 3, or 4 depending on which of the mutually exclusive modes of fishing is chosen:

1 beach

2 pier

3 private

4 charter

The regressors that are used to predict fishing mode choice are:

- individual income (constant across alternatives)
- price of fishing (alternative-specific)
- catch rate (alternative-specific)

On the next page, the Stata output for two models of fishing mode choice is provided. Both models were estimated using the same dataset.

Please indicate for each of the following statements whether it is true or false and provide a very brief explanation (i. e., explain what number(s) in the Stata output you use to answer each question).

(a) The number of fishing choices that have been used for estimation of both models is 4728.

(b) A fishing mode is less likely to be chosen if its price increases.

(c) Compared to beach fishing, a higher income leads to reduced likelihood of private boat fishing.

(d) An individual's income does not predict fishing mode choice as well as the alternative-specific variables.

And finally, an open-ended question:

(e) Can you think of a better model to predict fishing mode choices? Briefly describe that model and explain why you would prefer it over the two models whose results are reported here.

Stata output for fishing mode choice: Conditional Logit (CL)

```
Iteration 0:   log likelihood = -1581.9099
Iteration 1:   log likelihood = -1363.5718
Iteration 2:   log likelihood = -1317.8453
Iteration 3:   log likelihood = -1312.1013
Iteration 4:   log likelihood = -1311.9797
Iteration 5:   log likelihood = -1311.9796

Conditional (fixed-effects) logistic regression   Number of obs   =       4728
                                                  LR chi2(2)      =     653.24
                                                  Prob > chi2     =     0.0000
Log likelihood = -1311.9796                       Pseudo R2       =     0.1993

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
price        |  -.0204765   .0012231   -16.74   0.000    -.0228737   -.0180794
catch rate   |   .9530985   .0894134    10.66   0.000     .7778514    1.128346
------------------------------------------------------------------------------
```

Stata output for fishing mode choice: Multinomial Logit (MNL)

```
Iteration 0:   log likelihood = -1497.7229
Iteration 1:   log likelihood = -1477.5265
Iteration 2:   log likelihood = -1477.1514
Iteration 3:   log likelihood = -1477.1506

Multinomial logistic regression                   Number of obs   =       1182
                                                  LR chi2(3)      =      41.14
                                                  Prob > chi2     =     0.0000
Log likelihood = -1477.1506                       Pseudo R2       =     0.0137

------------------------------------------------------------------------------
        mode |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pier         |
      income |  -.1434029   .0532882    -2.69   0.007    -.2478459     -.03896
       _cons |   .8141503   .2286316     3.56   0.000     .3660405     1.26226
-------------+----------------------------------------------------------------
private      |
      income |   .0919064   .0406638     2.26   0.024     .0122069    .1716059
       _cons |   .7389208   .1967309     3.76   0.000     .3533352    1.124506
-------------+----------------------------------------------------------------
charter      |
      income |  -.0316399   .0418463    -0.76   0.450    -.1136571    .0503774
       _cons |   1.341291   .1945167     6.90   0.000     .9600457    1.722537
------------------------------------------------------------------------------
(Outcome mode==beach is the comparison group)
```

Question 4 – Lindeboom and van Doorslaer (2004) **5 + 10 = 15 credits**

(a) The authors begin the paper by stating that "[t]here is some concern that ordered responses on health questions may differ across populations or even across subgroups of a population." Briefly describe the reasons for the reporting heterogeneity they refer to.

(b) Explain the differences between a standard Ordered Probit model and the Cut-Point Shift model formally (ideally, using the key model equations).

Question 5 – Spindler *et al.* (2014) **2 + 10 + 3 = 15 credits**

(a) What is the basic prediction of theories of asymmetric information on insurance markets that the authors test?

(b) State and explain (ideally, using the key model equations) *two* of the various models that are estimated in this paper.

(c) Explain why data on regional variation in risk exposure (say, weather conditions) that is not reflected in premiums would be useful in the empirical analysis of asymmetric information on insurance markets.

Question 6 – Aghion *et al.* (2009) **3 + 7 = 10 credits**

(a) One of the models the authors estimate is a zero-inflated Poisson model. What is the dependent variable? Which feature of this variable leads them to estimate a zero-inflated model?

(b) Which general econometric approach do the authors use to address endogeneity in the zero-inflated Poisson model? (You do not need to state which specific variables they use.)