EKT 713/714
Semester 1 2020 Exam (3 hours)

By submitting your Exam, you are confirming that you have read the EMS Statement.

EMS Statement:

"The University of Pretoria commits itself to produce academic work of integrity. I affirm that I am aware of and have read the Rules and Policies of the University, more specifically the Disciplinary Procedure and the Tests and Examinations Rules, which prohibit any unethical, dishonest or improper conduct during tests, assignments, examinations and/or any other forms of assessment. I am aware that no student or any other person may assist or attempt to assist another student, or obtain help, or attempt to obtain help from another student or any other person during tests, assessments, assignments, examinations and/or any other forms of assessment."

Instruction: Please show your work and/or explain your answer or you will receive zero credit. By show your work, I mean enough to indicate to me that you credibly completed the work on your own. Make sure to use the actual variable names instead of generic labels like x, y, and h. If you use generic information copied from the book or another source, you will receive 0 credit. Points are shown at the end of questions. The acronym TFU stands for True, False, or Uncertain question that you must provide adequate support for your answer. Each number should be considered an independent question unless stated otherwise.

Part I

You are interested in the determinants of working women having children under the age of 6. The descriptive statistics by age group are in Table 1. The variables *kidslt6*, *educ*, and *age* are count variables representing number of kids under 6, education, and age. The variables *hours* and *hushrs* are continuous variables representing hours worked and hours worked by husband. If OLS was used to estimate the model, the estimating equation would be

$$kidslt6_i = \beta_0 + \beta_1 * educ_i + \beta_2 * hours_i + \beta_3 * age_i + \beta_4 * hushrs_i + u_i \quad (1)$$

Each number should be considered an independent question unless stated otherwise.

Table 1: Descriptive Statistics by Age Group

Age Groups	30 to 39	40 to 49	50 to 60	Total
N (percentage of sample)	268 (39.5)	257 (37.9)	153 (22.6)	678 (100.0)
** continuous variables	***	***	***	***
Hours Worked/Year, mean (sd)	766.090 (880.861)	792.759 (867.830)	609.954 (865.095)	740.965 (874.056)
Hours Worked/Year Husband, mean (sd)	2309.989 (576.516)	2212.860 (587.830)	2226.941 (625.627)	2254.431 (593.021)
** discrete variables	***	***	***	***
Number of Kids<6, mean (sd)	0.489 (0.684)	0.089 (0.299)	0.013 (0.114)	0.230 (0.516)
Educational Attainment, mean (sd)	12.519 (2.201)	12.261 (2.220)	11.830 (2.528)	12.265 (2.297)
Age of Woman, mean (sd)	34.269 (2.914)	44.708 (2.734)	53.778 (2.870)	42.628 (8.089)

Stand deviation is in parenthesis except in the first row. In the first row the number in parenthesis is percentage of sample.

1. (TFU) A researcher notes that mother's education (*meduc*) of each individual could be a good instrument for *educ* although it could be correlated with error term in equation (1). You do show that *educ* and *meduc* are strongly correlated. Since $E(meduc, u) \neq 0$ is likely true, you should use OLS.

2. We want to estimate a Probit model instead of OLS. Write down the latent variable model for equation (1). Be specific to the information given.

3. We have estimated the model described in 2 and the results are Calculate the partial effect at the averages for

Figure 1: Results

```
Probit
Iteration 0:   log likelihood = -328.46693
Iteration 1:   log likelihood = -228.78726
Iteration 2:   log likelihood = -222.20262
Iteration 3:   log likelihood = -222.15762
Iteration 4:   log likelihood = -222.15762

Probit regression                                 Number of obs     =        678

Log likelihood = -222.15762

------------------------------------------------------------------------------
     kidslt6 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0273265   .0307288     0.89   0.374    -.0329009    .0875538
       hours |  -.0005953   .0000917    -6.49   0.000     -.000775   -.0004157
         age |  -.1205088   .0113689   -10.60   0.000    -.1427915   -.0982261
      hushrs |   -.000222   .0001189    -1.87   0.062     -.000455    .0000111
       _cons |   4.411773   .6810511     6.48   0.000     3.076937    5.746608
------------------------------------------------------------------------------
```

hours and *hushrs* for the groups 30 to 39 and 50 to 60. Show work.

4. Using the information in 2 and 3, what is the partial effect at the averages for a woman in the sample that has a high school education ($educ = 12$) vs someone with 11 years of education ($educ = 11$)?

5. Using the information in 2 and 3, test the hypothesis that $\beta_1 = \beta_2 = \beta_3 = \beta_4$. Show work and explain.

6. Using the information in 2 and 3, calculate the psuedo R-squared. Show work and explain.

7. You have decided to revert back to using $kidslt6$ as a discrete variable. Setup a Heckman Selection Model assuming that age meets the exclusion restriction. Provide an adequate explanation with steps and be specific to the information given. How would you determine whether selection is a problem?

Part II

Suppose you have data on fatality rates ($frate_{it}$)from traffic accidents for 100 districts in the years 2012, 2014, and 2018. The determinants (regressors) are alcohol tax revenue (tax_{it}), average high school graduation rate ($grad_{it}$), and average unemployment rate ($unemploy_{it}$). Note that we also need binary variables representing the year in the model.

$$frate_{it} = \beta_0 + \beta_1 Tax_{it} + \beta_2 grad_{it} + \beta_3 unemploy_{it} + D_{2014} + D_{2018} + dis_i + u_{it} \tag{2}$$

The variable α_i is the unobserved district heterogeneity. Treat number as an individual questions unless otherwise specified.
Each number should be considered an independent question unless stated otherwise.

1. (TFU) Since we can not account for dis_i in the model, the unobserved district will cause bias in the estimates if we estimate equation (2).

2. A researcher suggest that the unobserved district heterogeneity (dis_i) is correlated with u_{it}. Setup a model that would consistently estimate the effect tax_{it} on $frate_{it}$ (hint: First Difference). Be sure to state required assumptions and show procedures to get estimating equation.

3. The same researcher notes that heteroscedasticity may also be a problem in the model you setup in 2. Setup a test that uses nonlinearites to show whether the researcher could possibly be correct. Be specific and explain.

4. Suppose the researcher was correct in 3 and you must adjust for heteroscedasticity (this should not affect your answer in 3). Set up a model that corrects for heteroscedasticity. Be specific to the information given and explain.

5. In 2016, a new program was initiated by some districts controlled by the AA political party. The AA party decided restrict the time individual 21 years old or younger can drive. The AA party believes young drivers are the primary cause of accidents so they restricted the time they can legally drive to between 06:00-22:00. Setup a model to analyze the new policy using equation 2 as the base (hint: Diff-in-Diff). Explain and be specific.

6. (TFU) Without using the method in 5, you will over estimate the policy effect of the new restriction if you used First Difference instead.