Exam ECON4150: Introductory Econometrics.
May 13; 09:00h-12.00h.

This is an open book examination where all printed and written resources, in addition to a calculator, are allowed. If you are asked to derive something, give all intermediate steps. Do not answer questions with a "yes" or "no" only, but carefully motivate your answer. In the grading, questions 1 and 2 will together count for 2/3 and questions 3 and 4 will together count for 1/3.

Guideline for correctors: *In this exam a total of 120 points can be obtained. The number of points that can be obtained by answering a question correctly are indicated in the solution box below the question.*

Question 1

The government of a developing country wants to implement a program where poor families receive food stamps that can be used to purchase prepackaged foods with high nutritional value. The government decides to set up an experiment where 500 families (each with 1 child) are randomly assigned to a treatment group (eligible for food stamps, $T_i = 1$) and to a control group (ineligible for food stamps, $T_i = 0$). The government has hired a researcher to investigate the effect of food stamps on the probability that a child has poor health.
After the experiment the researcher performs a regression of H_i (a binary variable that equals 1 if a child has poor health) on F_i (a binary variable that equals 1 if a family received food stamps). She obtains the following OLS estimation results.

```
. regress H F, robust

Linear regression                               Number of obs =        500
                                                   F(  1,   498) =    20.08
                                                   Prob > F      =   0.0000
                                                   R-squared     =   0.0375
                                                   Root MSE      =   .49141
```

H	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf. Interval]	
F	-.2090787	.0466629	-4.48	0.000	-.3007592	-.1173983
_cons	.6538462	.0381663	17.13	0.000	.5788593	.728833

a) Interpret the two estimated coefficients.

Solution: *(5 points). The estimated constant term can be interpreted as the mean probability that a child has poor health in families that did not receive food stamps (which equals 0.654, or 65%). The estimated coefficient on F is the difference in the mean probability that a child has poor health between families with and without food stamps and indicates that children in families with food stamps are on average 21 percent less likely to have poor health.*

b) The researcher finds out that some of the families in the control group received food stamps. Explain whether we can interpret the estimated OLS coefficient on F as the causal effect of food stamps on child health?

Solution: *(10 points) No we cannot interpret the coefficient on F as the causal effect of food stamps. Although assignment to the treatment group is random, actual receipt of food stamps is not random and can be related to (unobserved) family characteristics that affect child health (leading to omitted variable bias). This is an example of "failure to follow the treatment protocol".*

The researcher decides to estimate the effect of food stamps using an instrumental variable approach. She uses assignment to the treatment group as instrument for the actual receipt of food stamps. She obtains the following first stage estimation results.

```
. regress F T, robust noheader
```

F	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
T	.624	.0306963	20.33	0.000	.5636897	.6843103
_cons	.376	.0306963	12.25	0.000	.3156897	.4363103

c) Do you think that the instrument relevance condition holds? Is T a weak instrument?

Solution: *(10 points) Instrument relevance, $Corr(T_i, F_i) \neq 0$ can be investigated using the first stage regression. The first stage F-statistic equals $F = (t)^2 = (20.33)^2 = 413.3089$, which is larger than the rule-of-thumb value of 10 so the instrument relevance condition holds and T is not a weak instrument.*

d) Do you think that the instrument exogeneity condition holds?

Solution: *(10 points) The instrument should be uncorrelated with the error term, $Corr(T_i, u_i) = 0$. This assumption cannot be tested but in this case the assumption very likely holds because treatment assignment was random (and there are no clear reasons for the pressence of a direct effect of treatment assignment on child health).*

e) The following table shows the averages of H_i and F_i for those assigned to treatment group ($T_i = 1$) and for those assigned to the control group ($T_i = 1$). Use the results in the table below to obtain the instrumental variable estimate of the effect of food stamps on the probability that a child has poor health.

	$T_i = 1$	$T_i = 0$
$\widehat{E}\left[H_i \mid T_i = x\right]$	0.476	0.544
$\widehat{E}\left[F_i \mid T_i = x\right]$	1	0.376

Solution: *(10 points) The instrument T_i is binary, We therefore have that the IV estimator equals the so called Wald estimator:*

$$\hat{\beta}_{IV} = \frac{S_{ZY}/S_Z^2}{S_{ZX}/S_Z^2} = \frac{\widehat{E}\left[H_i \mid T_i = 1\right] - \widehat{E}\left[H_i \mid T_i = 0\right]}{\widehat{E}\left[F_i \mid T_i = 1\right] - \widehat{E}\left[F_i \mid T_i = 0\right]}$$

the instrumental variable estimate of the effect of food stamps on the probability that a child has poor health therefore equals

$$\hat{\beta}_{IV} = \frac{0.476 - 0.544}{1 - 0.376} = -0.109$$

Question 2

The Norwegian government wants to know whether restricting the opening hours of liquor stores reduces alcohol consumption. Holger, an employee of Statistics Norway, is asked to investigate this research question. He uses panel data for $n = 60$ municipalities observed in $T = 10$ time periods. The data set contains information on per capita alcohol consumption (in liters per year) in municipality i in year t ($alcohol_{it}$) and on the number of hours that liquor stores were open during year t in municipality i ($hours_{it}$). Holger estimates

$$alcohol_{it} = \beta_0 + \beta_1 ln(hours_{it}) + u_{it}$$

by OLS and obtains the following estimation results.

```
. regress alcohol lnhours, robust

Linear regression                               Number of obs =        600
                                                [redacted]
                                                R-squared     =     0.0745
                                                Root MSE      =      .9568
```

alcohol	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnhours	**5.288235**	**.7486464**	[redacted]	[redacted]	[redacted]	[redacted]
_cons	**2.768777**	**5.686286**	**0.49**	**0.626**	**-8.398742**	**13.9363**

a) Test the null hypothesis that $\beta_1 = 0$ at a 1% significance level.

Solution: (5 points)

$$H_0 : \beta_1 = 0 \quad vs \quad H_0 : \beta_1 \neq 0$$

Compute the t-statistic:

$$t = \frac{5.288}{0.748} = 7.07$$

The critical value at a 1% significance level is 2.58. Since 7.07 is bigger than 2.58 we reject the null hypothesis that $\beta_1 = 0$ *at a 1% significance level.*

b) Use the above estimation results to predict the change in alcohol consumption if the opening hours of liquor stores are reduced by 20 percent.

Solution: *(10 points) On the basis of the estimation results, the predicted change in alcohol consumption if the opening hours of liquor stores are reduced by 20 percent equals a reduction of* $20 \times 0.01 \times 5.28 \approx 1.06$ *liters per capita per year.*

c) Marit, Holger's colleague, suggests to augment the model with municipality fixed effects (α_i)

$$alcohol_{it} = \beta_0 + \beta_1 ln(hours_{it}) + \alpha_i + u_{it} \tag{1}$$

Explain how you could estimate model (1).

Solution: *(10 points) There are two ways they only have to discuss one.*

1. *Least Squares with dummy variables: Create 60 dummy variables for the municipalities $D1_i, D2_i, ..., D60_i$ with $D1_i = 1$ if $i = 1$ and zero otherwise, $D2_i = 1$ if $i = 2$ and zero otherwise etc. Augment the model by including $60 - 1$ dummy variables, or include 60 dummy variables and exclude the constant term.*

   (a) *Within estimation. First step: demean $alcohol_{it}$ and $ln(hours_{it})$. Second step: regress $(alcohol_{it} - \overline{alcohol_i})$ on $(ln(hours_{it}) - \overline{ln(hours_{it})})$.*

d) Holger decides to estimate a model that includes both municipality and year fixed effects. Both Holger and Marit are confident that by including municipality and year fixed effects the estimated coefficient on $ln(hours_{it})$ cannot suffer from omitted variable bias problems. Do you agree with Holger and Marit?

Solution: *(10 points) No, there can still be omitted variable bias. Municipality fixed effects control for all (unobserved) variables that are constant over time and year fixed effects control for all (unobserved) variables that vary over time but that are constant across municipalities. However, variables that vary over time and across municipalities that are omitted from the regression can still cause omitted variable bias in a model with municipality and year fixed effects.*

Question 3

Discuss whether each of the following statements is correct or not.

a) A high R^2 implies no omitted variables bias.

Solution: *(5 points) Incorrect, a high R^2 means that the regressors produce good predictions of the dependent variable but it is not informative about the presence of omitted variable bias in any of the estimated coefficients due to correlation between a regressor and omitted variables in the error term.*

b) A high p-value indicates that we cannot reject the null hypothesis.

Solution: *(5 points) Correct, if the p-value is for example above 0.10 this indicates that the null hypothesis cannot be rejected at a 10 percent significance level.*

c) In a regression model with no explanatory variables the R^2 is equal to 0.

Solution: *(5 points) Correct. The R^2 is the ratio of the explained sum of squares to the total sum of squares. If a regression model does not include explanatory variables the model does not explain anything and the explained sum of squares is zero.*

d) If a variable follows a random walk with drift, the best forecast of the variable tomorrow is the value of the variable today.

Solution: (5 points) *Incorrect. In the random walk with drift model ($Y_t = \beta_0 + Y_{t-1} + u_t$) the best forecast of the variable tomorrow is the value of the variable today plus the drift β_0.*

Question 4

Consider a labour market with a labour supply function $L_i^s = \beta_0 + \beta_1 W_i + \varepsilon_i^s$ and a labour demand fuction $L_i^d = \gamma_0 + \varepsilon_i^d$ and a market equilibrium condition $L_i^s = L_i^d$. In addition $E[\varepsilon_i^s] = E[\varepsilon_i^d] = 0$ and $Cov(\varepsilon_i^s, \varepsilon_i^d) = 0$.

a) Show that $Cov(W_i, \varepsilon_i^s) \neq 0$.

Solution: *(10 points) In equilibirum we have*

$$\beta_0 + \beta_1 W_i + \varepsilon_i^s = \gamma_0 + \varepsilon_i^d$$

hence solving for W_i yields

$$W_i = \frac{1}{\beta_1}(\gamma_0 - \beta_0 + \varepsilon_i^d - \varepsilon_i^s)$$

from which (using $Cov(\varepsilon_i^s, \varepsilon_i^d) = 0$) we see that

$$Cov(W_i, \varepsilon_i^s) = Cov(\frac{1}{\beta_1}(\gamma_0 - \beta_0 + \varepsilon_i^d - \varepsilon_i^s), \varepsilon_i^s) = -\frac{1}{\beta_1} Var(\varepsilon_i^s) \neq 0$$

b) Show that the OLS estimator of β_1 is inconsistent.

Solution: *(10 points) The OLS estimator of the supply equation is*

$$\widehat{\beta_1} = \frac{s_{WL}}{s_W^2} \rightarrow \frac{Cov(W_i, L_i)}{Var(W_i)} = \frac{Cov(W_i, \beta_0 + \beta_1 W_i + \varepsilon_i^s)}{Var(W_i)} = \beta_1 + \frac{Cov(W_i, \varepsilon_i^s)}{Var(W_i)} \neq \beta_1$$

from a) we know that $Cov(W_i, \varepsilon_i^s) \neq 0$. The OLS estimator is not a consistent estimator of β_1.