Postponed Exam ECON4150: Introductory Econometrics.
June 19; 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

A researcher wants to investigate the wage returns to a job training program. He has set up an experiment where 400 individuals were randomly assigned to a treatment group (the job training program) and to a control group. After the experiment the researcher collected data for the 400 individuals on wages ($Wage$) (in NOK) and on whether the individual has participated in the job training program ($Training$=1) or not ($Training$=0). The researcher decides to estimate the following regression model by OLS

$$Wage_i = \beta_0 + \beta_1 \cdot Training_i + u_i \tag{1}$$

and obtains the following regression results

```
. regress Wage Training, robust

Linear regression                                Number of obs =       400
                                                    F(  1,   398) =    243.25
                                                    Prob > F      =    0.0000
                                                    R-squared     =    0.3794
                                                    Root MSE      =    52.024
```

Wage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf. Interval]	
Training	81.18112	5.205143	15.60	0.000	70.94811	91.41413
_cons	162.1307	3.741819	43.33	0.000	154.7745	169.4869

a) Give an interpretation, in words, of the two estimated coefficients.

Solution (5 points): $\widehat{\beta}_0 = 162.13$ *NOK is the mean wage of the individuals who did not participate in the job training program and* $\widehat{\beta}_1 = 81.18$ *NOK is the difference between the mean wage of the individuals who participated in the job training program and the mean wage of those who did not participate. The mean wage of the individuals who participated in a job training program is equal to* $\widehat{\beta}_0 + \widehat{\beta}_1 = 243.31$ *NOK.*

b) The researcher wants to analyze whether there is a difference in the returns to participating in a job training program between men and women. Describe in detail how you would extend model (1), such that you can test the null hypothesis that the wage returns to a job training program do not dependent on gender.

Solution (10 points): *The regression should be augmented to include an interaction term as follows:*

$$Wage_i = \delta_0 + \delta_1 Training_i + \delta_2 F_i + \delta_3 (Training_i \times F_i) + \epsilon_i$$

whereby F_i equals one if female and zero if male. The hypothesis can be tested by using a t or F test testing H_0: $\delta_3 = 0$.

c) Some individuals who were assigned to the treatment group did not participate in the job training program while some individuals who were assigned to the control group did participate in the job training program Do you think that the OLS estimate of β_1 in model (1) is a consistent estimate of the effect of participating in the job training program on wages? Explain why or why not.

Solution (10 points): *This is an example of failure to follow the randomized treatment protocol or partial compliance. Although assignment to the treatment or control group is random, participating in the job training program is not random and likely related to unobserved variables that affect wages. An example of such an unobserved variable is motivation. If motivated individuals decide to participate although they are assigned to the control group and unmotivated individuals do not participate although they are assigned to the treatment group and motivated individuals earn higher wages than unmotivated individuals, the OLS estimator will be inconsistent (omitted variable bias).*

d) The researcher decides to use an instrumental variable approach to estimate the returns to participating in a job training program. He uses the assignment to the treatment ($Z_i = 1$) or control group ($Z_i = 0$) as an instrument for whether or not an individual participated in the job training program. The researcher obtains the following first stage OLS estimates.

```
. regress Training Z, robust noheader
```

Training	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
Z	.155	.0495038	3.13	0.002	.0576783	.2523217
_cons	.435	.0351433	12.38	0.000	.3659103	.5040897

Do you think that the instrument relevance condition holds? Is Z a weak instrument?

Solution (10 points): *Instrument relevance, $Corr(Training_i, Z_i) \neq 0$ can be investigated using the first stage regression. The first stage F-statistic equals $F = (t)^2 = (3.13)^2 = 9.8$, which is smaller than the rule-of-thumb value of 10. The instrument relevance condition holds but Z is a (relatively) weak instrument.*

e) Next to investigating the effect of participating in a job training program on wages the researcher also wants to know whether participating in the program increases the likelihood of being promoted to a higher position within a company. The data set contains an additional variable $Promotion_i$ which is equal to one if an individual has been promoted to a higher position and is zero otherwise. The researcher estimates a probit model and obtains the following estimation results

```
Probit regression                               Number of obs   =          400
                                                LR chi2(    1)      =      15.15
                                                Prob > chi2     =       0.0001
Log likelihood =  -269.18127                      Pseudo R2       =       0.0274
```

Promotion	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Training	.4916828	.1268232	3.88	0.000	.2431139	.7402516
_cons	-.1874829	.0903272	-2.08	0.038	-.364521	-.0104447

On the basis of these estimation results, what is the change in the probability that an individual is promoted to a higher position that is associated with participating in the job training program?

Solution (10 points): *On the basis of these estimation results the change in the probability that an individual is promoted to a higher position that is associated with participating in a job training program is equal to*

$$\widehat{\Pr(Prom_i = 1|Train_i = 1)} - \widehat{\Pr(Prom_i = 1|Train_i = 0)} = \Phi(-0.187 + 0.491) - \Phi(-0.187)$$

$$= 0.6179 - 0.4286$$

$$= 0.1893$$

Question 2

An economist wants to build a forecasting model for the annualized rate of inflation. He has quarterly data on the inflation rate (`inflation`). Let `d_inflation` be the change in the inflation rate from period $t-1$ to period t.

a) The economist estimates the following AR(1) model $\triangle inflation_t = \beta_0 + \beta_1 \triangle inflation_{t-1} + u_t$ and obtains the following estimation results

```
> regress d_inflation L1.d_inflation if tin(1973q1,2004q4), noheader
```

d_inflation	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
d_inflation						
L1.	**-.2227082**	**.0871244**	[redacted]	[redacted]	[redacted]	[redacted]
_cons	**-.0074748**	**.16348**	**-0.05**	**0.964**	**-.330997**	**.3160474**

Compute a 95% confidence interval for β_1.

Solution (5 points): 95% confidence interval for

$$\beta_1 = \left[\widehat{\beta}_1 - 1.96 \times SE\left(\widehat{\beta}_1\right), \widehat{\beta}_1 + 1.96 \times SE\left(\widehat{\beta}_1\right)\right]$$

Using the results in the stata output gives:

$$[-0.223 - 1.96 \times 0.087\ ,\ -0.223 + 1.96 \times 0.087]$$

$$[-0.39\ ,\ -0.05]$$

b) The economist wants to know how many lags of $\triangle inflation_t$ to include in the autoregression. He estimates $AR(p)$ models for $p = 1, 2$ and 3 over the sample period 1973:1 to 2004:4 (first quarter of 1973 through the fourth quarter of 2004) and obtains the following sum of squared residuals (SSR) for each of the estimated models.

p	**1**	**2**	**3**
SSR	431.014	364.635	352.8687

Use the Bayes Information Criterion (BIC) to estimate the number of lags that should be included in the autoregression.

Solution (10 points): *Choose the model with the smallest value of the Bayes Information Criterion (BIC)*

$$BIC(p) = ln\left[\frac{SSR(p)}{T}\right] + (p+1)\frac{ln(T)}{T}$$

*T is the number of time periods which equals 128 (32*4).*

p	**1**	**2**	**3**
SSR	**431.014**	**364.635**	**352.8687**
$ln\left[\frac{SSR(p)}{T}\right]$	**1.214**	**1.0469**	**1.0141**
$(p+1)\frac{ln(T)}{T}$	**0.076**	**0.1137**	**0.1516**
BIC	**1.2899**	**1.1606**	**1.1657**

The model with $p = 2$ has the smallest BIC, so on the basis of the BIC the economist should include 2 lags in the model.

c) The economist augments the AR(1) model of part (a) with three lagged values of the annualized unemployment rate. The economist computes the Granger-causality F-statistic on the three lags of the unemployment rate and obtains the following results.

```
. test L1.unemployment=L2.unemployment=L3.unemployment=0

 ( 1)  L.unemployment - L2.unemployment = 0
 ( 2)  L.unemployment - L3.unemployment = 0
 ( 3)  L.unemployment = 0

       F( [redacted],   123) =    11.78
            Prob > F =     0.0000
```

Do the unemployment rates help to predict the inflation rate (at a 5% significance level)?

Solution (10 points): *The claim that a variable has no predictive content corresponds to the null hypothesis that the coefficients on all lags of the variable are zero.*

$$H_0 : \beta_{L1unemployment} = \beta_{L2unemployment} = \beta_{L3unemployment} = 0$$

The critical value of the F-statistic at a 5% significance level equals $F_{3,123} = 2.60$. Since F=11.78 it is bigger than the critical value we conclude that Unemployment is a useful predictor for the change in the inflation rate.

d) The previous regressions were based on $\triangle inflation_t$, because the economist is worried that $inflation_t$ has a stochastic trend. Use the estimation results below to test for the presence of a stochastic trend in $inflation$, use a 5% significance level.

```
. regress d_inflation L1.inflation L1.d_inflation L2.d_inflation if tin(1973q1,2004q4)
```

d_inflation	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
inflation						
L1.	-.0933371	.0499211	-1.87	0.064	-.1921449	.0054707
d_inflation						
L1.	-.2491508	.0869439	-2.87	0.005	-.4212371	-.0770645
L2.	-.3516491	.0848594	-4.14	0.000	-.5196097	-.1836886
_cons	.4367815	.2789657	1.57	0.120	-.1153697	.9889326

Solution (10 points): *We have to perform the augmented Dickey-Fuller test*

$$H_0 : \beta_{L1.inflation} = 0 \quad vs \quad H_1 : \beta_{L1.inflation} < 0$$

The DF-statistic is the t-statistic on `L1.inflation` *in the estimation output:* $DF = -1.87$
The critical value of the augmented Dickey-Fuller statistic at a 5% significance level is -2.86. Since -1.87 is less negative than the critical value, we do not reject the null hypothesis of a stochastic trend.

Question 3

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

a) The R^2 can never be equal to zero.

Solution (5 points) *Incorrect. 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 explained sum of squares is zero and the* $R^2 = 0$.

b) Omitted variable bias can be solved by computing heteroskedasticity robust standard errors.

Solution (5 points) *Incorrect. Omitted variable bias leads to a correlation between a regressor and the error term making the OLS estimate of the coefficient of the regressor inconsistent. Computing heteroskedasticity robust standard errors solves the problem of heteroskedastic error terms* ($Var(u_t) \neq \sigma_u^2$) *but does not solve for omitted variable bias.*

c) By including entity fixed effects in a panel data model you control for omitted variables that vary across entities but not over time.

Solution (5 points) *Correct. In the model* $Y_{it} = \beta X_{it} + \alpha_i + u_{it}$, α_i *are entitity fixed effect which only have a subscript* i*. These* α_i *control for all (un)observed variables that vary between entities but which are constant over time.*

d) A forecast error is the same as an OLS residual.

Solution (5 points) *Incorrect. OLS residuals* $\widehat{u}_t = Y_t - \widehat{Y}_t$ *for* $t \leq T$ *are "in-sample". They are calculated for the observations in the sample used to estimate the regression. A forecast error* $Y_{T+j} - \widehat{Y}_{T+j|T}$ *for* $j \geq 1$ *is "out-of-sample". It is calculated for some date beyond the data set used to estimate the regression.*

e) In a regression without explanatory variables the OLS estimate of the constant term is equal to the mean of the dependent variable.

Solution (5 points) *Correct. In a regression without explanatory variables* $Y_i = \beta_0 + u_i$ *. Minimizing the sum of squared residuals gives* $\widehat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}$.

f) If the first stage F-statistic is larger than 10 the instrument exogeneity condition is satisfied.

Solution (5 points) *Incorrect. If the first stage F-statistic is larger than 10 this indicates that the instrument(s) is (are) not weak and that the instrument relevance condition is satisfied. Instrument exogeneity cannot be tested by investigating the first stage F-statistic.*

Question 4

Consider the following population regression model $W_i = \beta_0 + \beta_1 E_i + u_i$ with $E[u_i|E_i] = 0$. The researcher observes E_i but does not observe W_i, instead he observed a noisy measure $W_i^* = W_i + \varepsilon_i$ where $Cov(W_i, \varepsilon_i) = Cov(E_i, \varepsilon_i) = 0$, $E[\varepsilon_i] = 0$ and $Var(\varepsilon_i) = \sigma_\varepsilon^2$. The researcher estimates the following equation by OLS

$$W_i^* = \beta_0 + \beta_1 E_i + v_i$$

b) Is the OLS estimator of β_1 consistent?

Solution (10 points):

$$\widehat{\beta}_1 = \frac{s_{W^*E}}{s_E^2} \xrightarrow{p} \frac{Cov(W_i^*, E_i)}{Var(E_i)} = \frac{Cov(E_i, \beta_0 + \beta_1 E_i + v_i)}{Var(E_i)} = \beta_1 + \frac{Cov(E_i, v_i)}{Var(E_i)}$$

Substituting for $W_i = W_i^* - \varepsilon_i$ in $W_i = \beta_0 + \beta_1 E_i + u_i$ gives:

$$\begin{aligned} W_i^* - \varepsilon_i &= \beta_0 + \beta_1 E_i + u_i \\ W_i^* &= \beta_0 + \beta_1 E_i + \underbrace{u_i + \varepsilon_i}_{v_i} \end{aligned}$$

Substituting $v_i = u_i + \varepsilon_i$ in $Cov(E_i, v_i)$ gives:

$$\begin{aligned} \widehat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{Cov(E_i, v_i)}{Var(E_i)} &= \beta_1 + \frac{Cov(E_i, u_i + \varepsilon_i)}{Var(E_i)} \\ &= \beta_1 + \frac{Cov(E_i, u_i) + Cov(E_i, \varepsilon_i)}{Var(E_i)} \\ &= \beta_1 \end{aligned}$$

$Cov(E_i, u_i) = 0$ because $E[u_i|E_i] = 0$ and $Cov(E_i, \varepsilon_i) = 0$ as given in the exercise. This means that the OLS estimator of β_1 is consistent.