Postponed Exam ECON4150: Introductory Econometrics
Spring 2015

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, each sub-question will count for 1/12th of the total grade.

Guideline for correctors: *In this exam a total of 120 points can be obtained. With each sub-question a maximum of 10 points can be obtained.*

Question 1

An economist wants to investigate the effect of family size on the educational attainment of children. He performs a regression of years of education (when the child has finished his or her education) $Education_i$ on the variable $More2kids$ which equals 1 if a child has at least two siblings (his mother had more than 2 children) and zero if the child has one sibling (his mother had 2 children). The economist estimates the following regression by OLS

$$Education_i = \beta_0 + \beta_1 More2kids_i + u_i$$

and obtains the following OLS estimates

```
. regress Education More2kids, r

Linear regression                               Number of obs =      30000
                                                F(  1, 29998) =    6759.41
                                                Prob > F      =     0.0000
                                                R-squared     =     0.1834
                                                Root MSE      =     1.0208
```

Education	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
More2kids	**-.9871441**	**.0120068**	[redacted]	[redacted]	[redacted]	[redacted]
_cons	**12.69732**	**.0076388**	**1662.20**	**0.000**	**12.68235**	**12.71229**

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

Solution (10 points): $\widehat{\beta}_0 = 12.70$ *is the average number of years of schooling of the individuals who have one sibling and* $\widehat{\beta}_1 = -0.99$ *is the difference between the average years of schooling of the individuals who have more than one sibling and the average years of schooling of those who have one sibling. The average years of schooling of the individuals with more than two siblings is equal to* $\widehat{\beta}_0 + \widehat{\beta}_1 = 11.71$

b) Test the null hypothesis that the coefficient on $More2kids_i$ is equal to zero using a 5 percent significance level.

Solution (10 points): *Test* $H_0: \quad \beta_1 = 0$ *versus* $H_1: \quad \beta_1 \neq 0$.
Compute the t-statistic:

$$t = \frac{-0.987}{0.012} = -82.22$$

The critical value of the t-statistic at a 5% significance level is 1.96. Since $|-82.22|$ *is bigger than 1.96 we reject the null hypothesis that* $\beta_1 = 0$ *at a 5% significance level.*

c) Describe one potential threat to the internal validity of the current regression results.

Solution (10 points): *A potential threat to the internal validity is omitted variable bias. Parents that have more than 2 children might differ in unobserved characteristics from parents that decide to have 2 children. Often parents that have more than two children are lower educated compared to parents with 2 children and parent's level of education might have a direct effect on the education of their children.*

d) The economist thinks that he can obtain a consistent estimate of the effect family size on the educational attainment of children by performing 2SLS. He decides to use the incidence of twins at second birth as instrument for family size. He estimates the following first stage regression

$$More2kids_i = \pi_0 + \pi_1 Twins2_i + v_i$$

where $Twins2$ equals one if the mother had twins at second birth and zero otherwise. He obtains the following OLS estimates.

```
. regress More2kids Twins2, r

Linear regression                               Number of obs =      30000
                                                F(  1, 29998) =      79.40
                                                Prob > F      =     0.0000
                                                R-squared     =     0.0027
                                                Root MSE      =     .48947
```

More2kids	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
Twins2	.1974412	.0221584	8.91	0.000	.1540097	.2408726
_cons	.3977492	.0028496	139.58	0.000	.3921638	.4033346

Do you think that the instrument relevance condition holds? Is $Twins2_i$ a weak instrument?

Solution (10 points): *Instrument relevance, $Cov(More2kids_i, Twins2_i) \neq 0$ can be investigated using the first stage regression. The first stage F-statistic equals $F = (t)^2 = (8.91)^2 = 79.4$, which is bigger than the rule-of-thumb value of 10. The instrument relevance condition holds and $Twins2_i$ is a not a weak instrument.*

e) Do you think that $Twins2_i$ satisfies the instrument exogeneity condition? Explain why or why not.

Solution (10 points): *Instrument exogeneity states that $Cov(Twins2_i, u_i) = 0$ which implies that $Twins2_i$ should be uncorrelated with unobserved characteristics that affect the education of children (independence), in addition $Twins2_i$ should not have a direct effect on the education of children (exclusion restriction). One potential reason why the instrument exogeneity condition might not hold is that older mother's are more likely to have twins and mother's age (at birth) might have a direct effect on the education of children.*

f) The following table shows the averages of $Education_i$ and $More2kids_i$ for the individuals whose mother had twins at second birth ($Twins2_i = 1$) and for the individuals whose mother did not have twins at second birth ($Twins2_i = 0$). Use the results in the table below to obtain the instrumental variable estimate of the effect of $More2kids_i$ on years of education ($Education_i$).

	$Twins2_i = 1$	$Twins2_i = 0$
$\widehat{E}\left[Education_i \vert Twins_i = x\right]$	12.20	12.30
$\widehat{E}\left[More2kids_i \vert Twins_i = x\right]$	0.60	0.40

Solution: *(10 points) The instrument $Twins2_i$ is binary, We therefore have that the IV estimator equals the so called Wald estimator:*

$$\hat{\beta}_{IV} = \frac{S_{ZY}/S_Z^2}{S_{ZX}/S_Z^2} = \frac{\widehat{E}\left[Education_i \vert Twins_i = 1\right] - \widehat{E}\left[Education_i \vert Twins_i = 0\right]}{\widehat{E}\left[More2kids_i \vert Twins_i = 1\right] - \widehat{E}\left[More2kids_i \vert Twins_i = 0\right]}$$

The instrumental variable estimate of the effect of $More2kids_i$ on years of education ($Education_i$) equals

$$\hat{\beta}_{IV} = \frac{12.2 - 12.3}{0.6 - 0.4} = -0.5$$

Question 2

A researcher decides to build a forecasting model for the annualized rate of inflation. She 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 researcher estimates the following AR(2) model

$$\triangle inflation_t = \beta_0 + \beta_1 \triangle inflation_{t-1} + \beta_2 \triangle inflation_{t-2} + u_t$$

and obtains the following estimation results

```
. regress d_inflation L1.d_inflation L2.d_inflation if tin(1962q1,1994q4)
```

Source	SS	df	MS		
Model	**73.3612493**	**2**	**36.6806246**	Number of obs =	**132**
Residual	**347.325089**	**129**	**2.69244255**	F(2, 129) =	**13.62**
				Prob > F =	**0.0000**
				R-squared =	**0.1744**
				Adj R-squared =	**0.1616**
Total	**420.686339**	**131**	**3.2113461**	Root MSE =	**1.6409**

d_inflation	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
d_inflation						
L1.	**-.2873615**	**.0818998**	**-3.51**	**0.001**	**-.4494022**	**-.1253208**
L2.	**-.3692611**	**.0818178**	**-4.51**	**0.000**	**-.5311396**	**-.2073826**
_cons	**.0245487**	**.1428358**	**0.17**	**0.864**	**-.2580554**	**.3071529**

Compute a 99% confidence interval for β_1.

Solution (10 points): 99% confidence interval for β_1 is

$$\left[\widehat{\beta}_1 - 2.58 \times SE\left(\widehat{\beta}_1\right), \widehat{\beta}_1 + 2.58 \times SE\left(\widehat{\beta}_1\right)\right]$$

Using the results in the stata output gives:

$$[-0.287 - 2.58 \times 0.082\ ,\ -0.287 + 2.58 \times 0.082]$$

$$[-0.499\ ,\ -0.075]$$

b) The previous regression is based on $\triangle inflation_t$, because the researcher is worried that $inflation_t$ has a stochastic trend. The researcher estimates the following regression model

$$\triangle inflation_t = \gamma_0 + \delta \cdot inflation_{t-1} + \gamma_1 \triangle inflation_{t-1} + \gamma_2 \triangle inflation_{t-2} + u_t$$

and obtains the following estimation results.

```
. regress d_inflation L1.inflation L1.d_inflation L2.d_inflation if tin(1962q1,1994q4),
```

d_inflation	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
inflation						
L1.	-.0993034	.0474993	-2.09	0.039	-.1932889	-.0053179
d_inflation						
L1.	-.2307487	.0852646	-2.71	0.008	-.3994594	-.062038
L2.	-.3306665	.0828522	-3.99	0.000	-.4946038	-.1667293
_cons	.5055666	.2698531	1.87	0.063	-.0283838	1.039517

Use the estimation results to test for the presence of a stochastic trend in $inflation_t$. Use a 5% significance level.

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

$$H_0 : \delta = 0 \quad vs \quad H_1 : \delta < 0$$

The DF-statistic is the t-statistic on `L1.inflation` *in the estimation output:* $DF = -2.09$
The critical value of the augmented Dickey-Fuller statistic at a 5% significance level is -2.86. Since -2.09 is less negative than the critical value, we do not reject the null hypothesis of a stochastic trend.

c) The researcher wants to know how many lags of $\triangle inflation_t$ to include in the autoregression. She estimates $AR(p)$ models for $p = 1, 2, 3$ and 4 over the sample period 1962:1 to 1994:4 (first quarter of 1962 through the fourth quarter of 1994) and obtains the following sum of squared residuals (SSR) for each of the estimated models.

p	**1**	**2**	**3**	**4**
SSR	402.168	347.3251	331.970	331.383

Use the Akaike Information Criterion (AIC) 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 Akaike Information Criterion (AIC)*

$$AIC(p) = ln\left[\frac{SSR(p)}{T}\right] + (p+1)\frac{2}{T}$$

*T is the number of time periods which equals 132 (33*4).*

p	**1**	**2**	**3**	**4**
SSR	**402.168**	**347.3251**	**331.970**	**331.383**
$ln\left[\frac{SSR(p)}{T}\right]$	**1.114**	**0.967**	**0.922**	**0.920**
$(p+1)\frac{2}{T}$	**0.030**	**0.045**	**0.061**	**0.076**
AIC	**1.144**	**1.013**	**0.983**	**0.996**

The model with $p = 3$ has the smallest AIC, so on the basis of the AIC the researcher should include 3 lags in the model.

d) The researcher augments the AR(2) model of part (a) with four lagged values of the annualized unemployment rate. The researcher computes the Granger-causality F-statistic on the four lags of the unemployment rate and obtains the following results.

```
. test L1.unemployment=L2.unemployment=L3.unemployment=L4.unemploymen=0

 ( 1)  L.unemployment - L2.unemployment = 0
 ( 2)  L.unemployment - L3.unemployment = 0
 ( 3)  L.unemployment - L4.unemployment = 0
 ( 4)  L.unemployment = 0

       F(  [redacted],   125) =    11.07
                  [redacted]
```

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.*

$$H_0 : \beta_{L1unemployment} = \beta_{L2unemployment} = \beta_{L3unemployment} = \beta_{L4unemployment} = 0$$

The critical value of the F-statistic at a 5% significance level equals $F_{4,125} \approx F_{4,\infty} = 2.37$. Since F=11.07 it is bigger than the critical value we conclude that Unemployment is a useful predictor for the change in the inflation rate.

Question 3

Consider the following population regression model $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^3 + u_i$ with $E\left[u_i | X_i, X_i^3\right] = 0$. A researcher has a large sample with i.i.d observations on Y_i and X_i and estimates the following equation by OLS

$$Y_i = \beta_0 + \beta_1 X_i + v_i$$

a) What is $Cov\left(X_i, v_i\right)$?

Solution (10 points):

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 X_i^3 + u_i \\ &= \beta_0 + \beta_1 X_i + v_i \end{aligned}$$

This implies that $v_i = u_i + \beta_2 X_i^3$

$$\begin{aligned} Cov\left(X_i, v_i\right) &= Cov\left(X_i,\ u_i + \beta_2 X_i^3\right) \\ &= Cov(X_i, u_i) + Cov\left(X_i, \beta_2 X_i^3\right) \\ &= Cov(X_i, u_i) + \beta_2 Cov\left(X_i, X_i^3\right) \\ &= 0 + \beta_2 Cov\left(X_i, X_i^3\right) \\ &= \beta_2 Cov\left(X_i, X_i^3\right) \end{aligned}$$

$Cov(u_i, X_i) = 0$ because $E\left[u_i | X_i, X_i^3\right] = 0$, but $Cov\left(X_i, X_i^3\right) \neq 0$.

b) Is the OLS estimator of β_1 consistent?

Solution (10 points):

$$\widehat{\beta_1} = \frac{s_{XY}}{s_X^2} \xrightarrow{p} \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

$$\begin{aligned} \widehat{\beta_1} \xrightarrow{p} \frac{Cov(X_i, Y_i)}{Var(X_i)} &= \frac{Cov(X_i, \beta_0 + \beta_1 X_i + v_i)}{Var(X_i)} \\ &= \frac{Cov(X_i, \beta_0) + \beta_1 Cov(X_i, X_i) + Cov(X_i, v_i)}{Var(X_i)} \\ &= \beta_1 + \frac{Cov(X_i, v_i)}{Var(X_i)} \\ &= \beta_1 + \beta_2 \frac{Cov\left(X_i, X_i^3\right)}{Var(X_i)} \end{aligned}$$

Since $\widehat{\beta_1} \longrightarrow \beta_1 + \beta_2 \frac{Cov\left(X_i, X_i^3\right)}{Var(X_i)}$, the OLS estimator of β_1 is only consistent if $\beta_2 = 0$.