Exam Econ 4150, spring 2019

1. (15%) The zero conditional mean assumption of the Gauss-Markov conditions is often written as $E[u_i|X_i] = 0$.

   (a) Does $E[u_i|X_i] = 0$ imply that u_i and X_i are uncorrelated? Explain your answer. Yes it does. we can write the covariance between X_i and u_i as $E((X_i - E(X_i))(\varepsilon_i - E(\varepsilon_i)) = E(X_i\varepsilon_i) - E(X_i)E(\varepsilon_i)$. We also know that $E[u_i|X_i] = 0 \implies E[u_iX_i] = 0$ and we know that $E(\varepsilon_i) = 0$. From this it follows that when the expected value of the error term is equal to zero for all values of Xit must be the case that the covariance between $X_i$and u_i is zero.

   (b) Give an example of a regression equation where the $E[u_i|X_i] = 0$ is likely to be violated. The error term contains all variables that affect an outcome Y but that are not included in the regression. If the expected value of this error term depends on X we have an endogeneity problem. This will, for example occur if there is another variable, not included in the regression that is correlated both with the outcome we are interested in (the dependent variable) and the independent variable X. Suppose for example that our goal is to measure the effect of education (X) on wages (Y) but we do not include the ability of a person. It is likely that ability will be positively correlated with both years of education and wages. This then means that the expected value of the error term u increases in years of education and in a wage regression that does not include ability will give a biased estimate of the

   (c) If $E[u_i|X_i] \neq 0$ will more observations of X help? No

2. (7,5%) You draw a sample of N individuals from a population and use income data from the sample to estimate the average income in the population. Explain why $\tilde{y} = \frac{1}{N-1}\sum_{i=1}^{N} y_i$ is a biased but consistent estimator of the mean income in the population. It is biased since the expected value of this estimator is not equal to the average in the population (it is always a bit below). But when sample size goes to infinity this bias vanishes, it is therefore consistent.

3. (7,5%) You don't have access to your computer and you have to estimate the parameters in this simple OLS model by hand:

$$y = \beta_0 + \beta_1 x + u.$$

   Lucky for you there are only five observations, displayed below. Find the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

obs.	y	x
1	2	1
2	5	2
3	6	3
4	7	3
5	5	1

The ols formula is given by $\hat{\beta}_1 = \frac{COV(y,x)}{VAR(x)}$ and $\hat{\beta}_0 = \bar{y} - \beta\bar{x}$. Doing the calculations gives $\hat{\beta}_1 = 1,5$ and $\hat{\beta}_0 = 2$

4. (20%) A researcher wants to estimate how mothers influence the earnings of their daughters and collects data on the wage of 100 women and data on their mother's education, IQ and reading habits. She obtains the following results (the standard error is in parenthesis below the coefficient)

$$\widehat{lwage} = \underset{(0.54)}{1.58} + \underset{(0.23)}{1.24}\,mothedu + \underset{(0.56)}{1.60}\,IQ + \underset{(1.09)}{1.19}\,books$$
$$N = 100,\ R^2 = 0.42,\ F(3.96) = 18,9$$

where *mothedu*, *IQ*, *books* refer to an individual's mother's logged education level, their logged score on a standard IQ test and the logged number of books they have read respectively.

(a) What is the interpretation on the coefficient on books? It captures the elasticity of wages with respect to books read. It is the % increase in wages associated with a 1% increase in the number of books read.

(b) Are each of the variables individually significant at the 95% confidence level? No books has a t-stat below the critical level for 95% significance.

(c) Test whether the coefficient on *mothedu* is significantly different from 1 at a 5% significance level? The test-stat here is $t = \frac{1,24-1}{0,23} = 1,04$ which is below the critical value for a 5% significance.

(d) It turns out that there is a strong positive correlation between *books* and *mothedu*, is this a problem? Books is very correlated with each of the other factors. This means that inclusion of this variable along with the other two may cause large standard errors, and cause instability in the regression results.

(e) What would happen to the coefficient on *mothedu* if *books* was omitted from the above regression? Why would this happen? Since *mothedu* is likely positively correlated with books, we would expect that the removal of books from the regression results will cause an increase in the coefficient on *mothedu*. The intuition here is that books is taking some of the credit away from *mothedu*.

5. (50%) Data from a household survey ($N = 9000$) in Bangladesh contains information on whether or not a household has a member, a migrant, that works abroad. Migrants regularly send money to their household back home. These transfers are called remittances. A researcher wants to use the household survey data to estimate whether receiving remittances from migrants affect the income earned locally at home by the household that receives money from their migrant member.

   A priori one can imagine that receiving money from the migrant can either reduce or increase how much the household earns at home. To investigate this question the researcher uses OLS to regress the log of income earned by the household, excluding the money received from the migrant, ($ln(income)$ on a dummy that indicates whether or not the household receives remittances ($rem = 1$ if a household receives remittences (money from the migrant), 0 if not):

$$ln(income)_i = \beta_0 + \beta_1 rem_i + u_i \tag{1}$$

   (a) She obtains $\widehat{\beta}_1 = 0.11$, give an interpretation of this coefficient. This means that on average having a member of the family that sends remittances increases the income earned at home with 11 %.

   (b) The standard error of $\widehat{\beta}_1$ is 0.03. What is the 95 % confidence interval for $\widehat{\beta}_1$? The interval has a lower bound of $0.11 - 0,03 * 1,96 = 0,051$. The upper bound of this interval is $0.11 + 0,03 * 1,96 = 0,169$

   (c) The researcher is also interested in estimating how remittances affect the poverty status of a household. To this end she creates a indicator variable *poverty* that is equal to 0 for households with an income above the poverty line and equal to 1 for households with an income at or below the poverty line.

      i. Is it problematic to use *poverty* as the dependent variable in an OLS regression? Explain your answer. It is possible to use OLS with a binary outcome, it is called the linear probability model. The advantage with this model is that it is easy to interpret the coefficient, the disadvantage is that the OLS estimator suffers from heteroskedasticity (that can be relatively early fixed) and it also predicts probabilities outside 0,1. Which is of course nonsensical.

      ii. What alternatives to OLS can be used? Can use non-linear models that always predict a probability inside the 0,1 interval. There are two often used models the probit and the logit model. In these models we have $Pr(Yi = 1) = G(Z)$ with $Z = 0 + \beta_1 X_{1i} + .. + \beta_k X_{ki}$ and $0 \leq G(Z) \leq 1$. In the probit case $G(Z) = \phi(Z)$ where $\phi(Z)$ is the cumulative normal distribution. In the logit case we have $G(Z) = \frac{1}{1+e^{-Z}}$.

(d) Return to the regression model $ln(income)_i = \beta_0 + \beta_1 rem_i + u_i$. Discuss this statement: $\widehat{\beta}_1 = 0.11$ captures the average causal effect on earned family income of having a family member abroad that sends remittences? This is probably not true since it is very likely that the decision to send a migrant and to receive remittances is correlated the income that is measured in the regression.

(e) The researcher uses distance from Dhaka (capital in Bangladesh) as an instrument for remittances in an IV-estimation. What criteria must distance fulfill to be a valid instrument. A good instrument must be relevant; that is it must be correlated with the variable that we are interested in, here remittances. This can be checked. A second requirement is exclusion, or exogeneity, the instrument must not affect income in any other way than through the variable of interest; the instrument must not be correlated with the error term. So in this case it must (i) be the case that

(f) Write down the first stage equation of the IV-regression that uses distance from Dhaka as an instrument. What would you look for in this first stage to determine the validity of the instrument? The first stage here would be to regress remittances on distance from Dhaka: $rem_i = \pi Dist - Dhak_i + \gamma_i$. I would check that $Dist - Dhak$ is strongly correlated with receiving remittences. Otherwise Distance to Dhaka is not a relevant instrument.

(g) Someone suggests that another instrument could be used, namely the ownership of non-agricultural land.

   i. Is it possible to use both instruments simultaneously to predict migration? Explain your answer. It sure is. We can then estimate the first stage equation $rem_i = \pi_1 Dist - Dhak_i + \pi_2 Own - land_i + \gamma_i$. We would use the estimates, $\hat{\pi_1}$ and $\hat{\pi_2}$ of this first stage to predict the remittances status from household i.

   i. With two potential instruments, can any tests be performed to check the exogeneity of the instruments? Explain your answer. When we have more than one instrument we can test if at least one of the instruments are endogenous. The basic idea (and I do not expect any formal analysis here) is that with homogenous treatment effects (the effect on income of an exogenous change in remittances is homogenous) then we would expect the two instruments, if they are both exogenous, to produce the same IV estimate. If they are very different we can reject that they are both exogenous. There is a formal test for this and that is called the J-test.