Introduction to Econometrics

Final exam

The time limit is 90 minutes and the exam is worth a total of 30 points. You are NOT allowed to look up for solutions in books, notes or internet and not allowed to consult the problems with your classmates or anyone knowledgeable. Any violation of academic honesty will be punished to the fullest extent possible.

Date: 15.01.2021

Instructor: Dali Laxton

Multiple choice questions

30 min, 2 points each

1) Consider the following simple regression model y=β0 + β1x1 + u. The variable *z* is a poor instrument for *x* if _____.
a. there is a high correlation between z and x
b. there is a low correlation between z and x
c. there is a high correlation between z and u
d. there is a low correlation between z and u

2) In the following regression equation, y is a binary variable:

$$y= \beta_0+\beta_1x_1+\ldots\beta_k x_k+ u$$

In this case, the estimated slope coefficient,$\widehat{\beta_1}$ measures _____.
a. the predicted change in the value of y when x_1 increases by one unit, everything else remaining constant
b. the predicted change in the value of y when x_1 decreases by one unit, everything else remaining constant
c. the predicted change in the probability of success when x_1 increases by one unit, everything else remaining constant
d. the predicted change in the probability of success when x_1 decreases by one unit, everything else remaining constant

3) Which of the following assumptions is required to obtain a first-differenced

estimator in a two-period panel data analysis?
a. The idiosyncratic error at each time period is uncorrelated with the explanatory variables in both time periods.

b. The explanatory variable does not change over time for any cross-sectional unit.

c. The explanatory variable changes by the same amount in each time period.
d. The variance of the error term in the regression model is not constant.

4) Consider the following regression model: log(y) = β0 + β1x1 + β2x1^2 + β3x3 + u. This model will suffer from functional form misspecification if _____.
a. β0 is omitted from the model
b. u is heteroskedastic
c. x1^2 is omitted from the model
d. x3 is a binary variable

5) Consider the following regression equation: y = β0+β1x1+…βk xk+ u
In which of the following cases, the dependent variable is binary?
a. y indicates the gross domestic product of a country
b. y indicates whether an adult is a college dropout
c. y indicates household consumption expenditure
d. y indicates the number of children in a family

Problem 1. (**30 min**) We want to measure the impact of holding a health insurance (*healthin*) on the medical expenses (*medexp)*. The following is the simple model expressing the relationship:

$$\log(medexp) = \beta_0 + \beta_1 healthin + \varepsilon$$

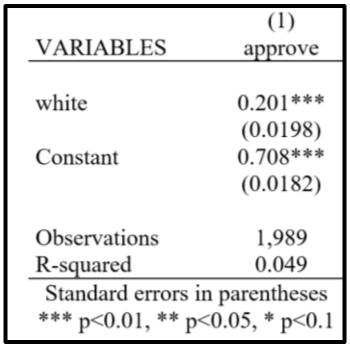
- **a)** (**2pt**) Why might $healthin$ be correlated with ε? **Some variables omitted like illnesses, age, could also be some reverse causality**
- **b)** (**2pt**) Explain why $healthin$ is likely to be related to the *age* and *illnesses* of the insured. Does this mean *age* and *illnesses* are good IV for $healthin$? Why or why not? **The sicker the person, more likely he will purchase an insurance to avoid tremendous medical expenses, also, older a person – more likely to have the illnesses. It is not a good IV, it is a good variable to control in the regression but as IV it will fail the instrument exogeneity requirement.**
- **c)** (**2pt**) After controlling for *age* and *illnesses*, you still believe that $healthin$ suffers from endogeneity issue. In particular, you believe that the risk-aversion of individuals drives both variables $healthin$ and $medexp$. Justify why social security income replacement rate[1] may be a good instrument. **The income replacement rate is correlated with *healthin* since higher the proportion of after retirement pension with the income before retirement, the more likely the person can afford the same standard of living as before retirement. This rate is however not correlated with the medical expenses, people do not change their replacement ratio based on medical expenses. It is just a ratio, it does not say anything about the amount of income itself**
- **d)** (**2pt**) How would you proceed with the estimation using the IV? (describe the 2SLS technique in this particular example). **First regress *healthin* on income replacement rate for each individual, find predicted values. In the secnd stage, regress the predicted values on the logarithmic expenses**
- **e)** (**2pt**) Propose an alternative instrument in order to solve the endogeneity issue. **Any natural or field experiment which ensures that the endogeneous variable is randomly assigned would do the trick. For example, if government subsidizes 30% of the population with health insurance randomly, we will define an instrument weather an individual received the subsidy or not, this variable will be strongly correlated with healthin, after this we can regress predicted values on log(medexp).**

[1] A social security income replacement rate is the percentage of a worker's pre-retirement income that is paid out by a pension program after retirement.

Problem2. (**30 min**) Suppose you want to assess the impact of a race of an individual on the likelihood of approving a mortgage loan. In the example below the key explanatory variable is white, a dummy variable equal to one if the applicant was white. The other applicants in the data set are black and Hispanic. To test for the discrimination in the mortgage loan market, a linear probability model (LPM) can be used:

$$approve = \alpha_0 + \alpha_1 white + u$$

a) (**2pt**) Suppose you obtain the following output from the regression above. Interpret the coefficient on *white.*



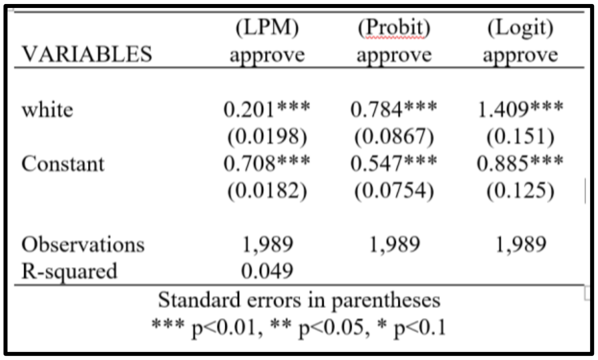
	(1)
VARIABLES	approve
white	0.201***
	(0.0198)
Constant	0.708***
	(0.0182)
Observations	1,989
R-squared	0.049

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

White people are 20% more likely to obtain approved mortgage than blacks and Hispanics.

b) (**2pt**) Name at least one pro and one con of using an LPM. **It is easy to interpret, predicted values goes out of [0,1] interval.**

c) (**2pt**) Suppose now that you run probit and logit models as well, interpret the coefficients on *white* for probit and logit models and compare them with the LPM model.



	(LPM)	(Probit)	(Logit)
VARIABLES	approve	approve	approve
white	0.201***	0.784***	1.409***
	(0.0198)	(0.0867)	(0.151)
Constant	0.708***	0.547***	0.885***
	(0.0182)	(0.0754)	(0.125)
Observations	1,989	1,989	1,989
R-squared	0.049		

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In probit and logit models you cannot directly interpret the coefficients, what matters is only signs – if a person is white, he is more likely to get mortgage approved.

d) (**2pt**) By how much is it more likely for white people to obtain mortgage loan in comparison to minorities according to probit model? How different is this result from LPM result?

$$\boldsymbol{P(Y = 1 \middle| X) = \phi(0.547 + 0.784 * 1) = \phi(1.331) = 0.908}$$
$$\boldsymbol{P(Y = 0|X) = \phi(0.547) = 0.705}$$

Comparison: 0.908-0.705=0.203 -> by 20% more likely just like in the LPM.

e) (**2pt**) By how much is it more likely for white people to obtain mortgage loan in comparison to minorities according to logit model? Note that the functional form of the logit model is $\Lambda(\cdot) = \frac{\exp(\cdot)}{1+\exp(\cdot)}$.

$\boldsymbol{\Lambda(2.294) = \frac{\exp(2.294)}{1+\exp(2.294)} = 0.908.}$

$\boldsymbol{\Lambda(0.885) = \frac{\exp(0.885)}{1+\exp(0.885)} = 0.708.}$

Difference is about 20% meaning that white people are 20% more likely to get mortgage approved.