# Problem Set 5

This problem set will be due at 6:30pm on Tuesday, April 24th. Hand in either in-class or in my mailbox (6th floor, 19 West 4th Street). Show your work (required for full and part marks).

# 1 Theory Questions

## Question 1

In this problem we will understand why economists began to think about instrumental variables in the first place. We will explore simultaneity bias. Suppose that you are an economist trying to predict the effect that raising tariffs will have on the price and output of coffee. To do this, you need to understand the supply and demand functions of coffee. For now, suppose we just want to understand demand. We observe quantities demanded and prices over $m = 1, ..., M$ market. Market demand has the following simple form:

$$Q_m^D = \beta_0 + \beta_1 P_m + u_m$$

Similarly, suppose that market supply has the following form:

$$Q_m^S = \alpha_0 + \alpha_1 P_m + \epsilon_m$$

We suspect that $\alpha_1 > 0$ since as price is higher, the market is more willing to supply coffee since more producers will enter. We further assume that $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ and $u_i \sim \mathcal{N}(0, \sigma_u^2)$ and $u$ and $\epsilon$ are uncorrelated.

Simultaneity bias arises because in equilibrium we have that quantity supplied must equal quantity demanded so we have that: $Q_m^S = Q_m^D$. Equilibrium price and quantity is thus given by:[1]

$$P_m = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{u_m - \epsilon_m}{\alpha_1 - \beta_1}$$
$$Q_m = \frac{\alpha_1\beta_0 - \beta_1\alpha_0}{\alpha_1 - \beta_1} + \frac{\alpha_1 u_m - \beta_1\epsilon_m}{\alpha_1 - \beta_1}$$

**Part (a):** Solve for $Cov(P, Q)$.

**Part (b):** Solve for $Var(P)$.

[1] Some good practice: solve out for $P_m$ and $Q_m$ yourself.

**Part (c):** Suppose your friend does not know too much about economics and so he runs the following regression:

$$Q_m = \beta_0 + \beta_1 P_m + u_m$$

What is the expected value of $\hat{\beta}_1^{OLS}$? Is the OLS estimator unbiased? (hint: part (a) and part(b) should help you a lot here...)

**Part (d):** Suppose that you will use weather as an instrument for price. Define weather as the instrumental variable $Z_m$. Our instrumental variable is valid, and so we have that $Cov(Z, P) \neq 0$ and $Cov(Z, U) = 0$. Our IV estimator here is given by:

$$\hat{\beta}_1^{IV} = \frac{\sum_{m=1}^{M} (z_i - \bar{z})(q_m - \bar{q})}{\sum_{m=1}^{M} (z_i - \bar{z})(p_m - \bar{p})}$$

Show that this IV estimator is unbiased. (**Note:** As I stated in class, you are welcome to the fact that $\mathbb{E}\left[\sum_{m=1}^{M} (z_i - \bar{z})(q_m - \bar{q})\right] = Cov(Z, Q)$ to avoid using summation notation when doing these proofs; but it is up to you.)

**Part (e):** Discuss the validity of weather as an instrument for the price of coffee.

# 2 Practical Questions

## 2.1 Practical Question 1

Instrumental variables have been used by economists to try to answer some very wide-ranging questions, including on the effects of political institutions on economic outcomes. In Acemoglu, Johnson and Robsion (2001, AER) the authors use different colonial institutions as an instrument for economic outcomes today. The dataset "ps5_iv.csv" has been posted online. Please download this data and load it into R. There are a total of 163 observations in the data set (although note many observations have missing data). The data structure is $c$, where $c$ indexes countries. The variables in the data set are:

$shortnam_c$ = Abbreviation of country name
$lat\_abst_c$ = lattitude of capital (divided by 90)
$euro1900_c$ = Number of european settlements in 1900
$avexpr_c$ = Average protection against appropriation (measure of property rights)
$loggpgp95_c$ = log GDP per capita in 1995
$logem4_c$= log settler mortality

The authors want to investigate the effect of property rights on economic outcomes and so are interested in the following regression:

$$loggpgp95_c = \beta_0 + \beta_1 avexpr_c + u_c$$

**Part (a):** Please find a variable that is omitted in the above regression that is likely to cause omitted variable bias and sign the bias resulting from the omission of that variable in the above regression.

**Suggested Instrument:** The authors suggest using settler mortality as an instrument for property rights, the idea being that in places where Europeans faced high mortality rates, they were less likely to settle and were more likely to set up extractive institutions which persist to the present. The authors thus argue that the instrument is relevant.

**Part (b):** Assess instrument relevance. To do so, clearly state the regression that you are running in `R` to assess relevance and perform the relevant hypothesis test. Is this instrument a weak instrument?

**Part (c):** Find the IV estimate when you investigate the effect of property rights on GDP in 1995 using settler mortality as an instrument for property rights. Include lattitude of capital as an additional control in your model. Clearly state the regression(s) that you are running in `R` to estimate this 2SLS model. Perform a hypothesis test to test if property rights causally affect GDP (assuming the instrument is valid).

**Part (d):** Assess instrument validity.

**Part (e):** In the paper, the authors suggest using both settler mortality and number of European settlements in 1900 as instruments. With these two instruments in hand, perform a J-test to test instrument validity. Clearly describe what your results imply about instrument validity.

## 2.2 Practical Question 2

In this problem we look at one of the (arguably) most famous use of regression discontinuity: estimating the effect of class size on student test scores:[2] we will do so in the context of New York City.[3] In New York City elementary schools in 2009-2013, there was a class size rule in place that class sizes could not exceed 32 students.[4] Therefore, a school-grade with enrollment of 32 students could form one class of 32 without violating the rule, while a school-grade with enrollment of 33 had to form two classes (one of 16 and one of 17) to abide by the rule. The dataset "ps5_rd.csv" which is derived from New York City publicly available data has been posted online. Please download this data and load it into `R`. There are a total of 4,053 observations in the data set. The data structure is *s*-*g*-*t*, where *s* indexes schools, *g* indexes grades, and *t* indexes year. The variables in the data set are:

$schoolname_s$= Name of school
$averageclasssize_{sgt}$ = Average class size
$year_t$ = Year (data is organized so 2009 represents school year 2009-10)
$grade_g$ = grade
$enroll_{sgt}$ = school-grade-year enrollment
$schoolid_s$= school identifier
$mathscore_{sgt}$ = standardized math test scores (these are thus in standard deviation units)
$x_{sgt}$ = Normalized running variable[5]
$pctdisability_{sgt}$ = proportion of students with learning disability
$pctblack_{sgt}$ = proportion of students that are African American
$pcthisp_{sgt}$ = proportion of students that are Hispanic
$pctwhite_{sgt}$ = proportion of students that are white (note: omitted category is proportion Asian)

**Part (a):** Using a bandwidth of 10 and a linear functional form, show that the class size rule generates a discontinuity in class size. What is the exact change in class size we observe when school-grades cross the threshold?

**Part (b):** Using a bandwidth of 10 and a linear functional form, use the class size rule to estimate the causal effect of a **one student** increase in class size on student achievement. Please also include the following variables as controls in your regression(s): $enroll_{sgt}$, $pctblack_{sgt}$, $pctwhite_{sgt}$, $pcthisp_{sgt}$, $pctdisability_{sgt}$, and $pctEL_{sgt}$.

[2]See well-know papers such as Angrist and Lavy (1999) as well as Hoxby (2000) for examples.

[3]Data for this question actually comes from one of my own research papers if you are interested.

[4]Actually this rule has been in place in other years too. However, it was only really followed during this time period because the union was fighting with Mayor Bloomberg which led to the union stringently enforcing the class size rule through the courts.

[5]I was nice and normalized this for you. Basically, if your school-grade enrollment is 32 then you get assigned a value of -0.5 and if your enrollment is 33 you get 0.5. In addition, since each multiple of 32 is also a class size cap (going from 64 students to 65 you must make three classes) a school-grade enrollment of 64 is assigned a value of -0.5 while 65 is assigned 0.5 (and similarly for enrollment of 96).

**Part (c):** Formally assess the validity of the regression discontinuity design (to be clear: I expect you to check whether each covariate is balanced and to visually describe whether schools precisely manipulate their enrollment around the class size cap).

**Part (d):** Describe which type of school-grade that identifies our RD estimate. How does this affect external validity?