Econometrics: Problem Set #2

Due on April 5, 2024 at 4:00pm

Professor Ben Faber Section 101

Zachary Brandt

Question 1

The Ministry of Commerce in a large country wants to know the causal effect of membership in a local Chamber of Commerce on firm revenues and profits. Firms pay for their membership and they supposedly benefit from the network of information and contacts that the local Chambers of Commerce offer them. But usually, it is only a small minority of all firms that end up paying for the membership.

The Ministry plans to estimate the causal effect of being a member in a local Chamber of Commerce by letting the Ministry's staff estimate the average percentage change in annual firm sales between firms that are members and the rest of the firms that are non-members of their local Chamber of Commerce.

A) Write down the OLS regression specification that the Ministry's staff could use to implement their analysis described above. Interpret what the intercept and slope coefficients would capture in such a specification.

B) Using notation from the Potential Outcomes Framework, briefly explain the concept of the Average Treatment Effect (ATE) to the Minister, and how what they plan to estimate in A) relates to this definition.

C) Referring to the expressions you use in your answer to B), explain why a randomized control trial (RCT) could be useful, and very briefly describe the basics of the RCT design for how the Ministry could set this up.

D) The Ministry mentions that it has no legal authority to force firms to become members in their local Chambers of Commerce. Using notation from the Potential Outcomes Framework, explain why this information could be important for the interpretation of the results from the RCT relative to the ATE, and how the Ministry should address this concern in the RCT analysis?

E) The Ministry talked to other economists, and now it is worried about spillover effects on the control group. The staff don't fully understand what the concern is, however. Briefly explain to them the intuition behind this concern, and explain how they could potentially address it when designing the RCT.

Question 1 continued on next page...

Part A

Write down the OLS regression specification that the Ministry's staff could use to implement their analysis described above. Interpret what the intercept and slope coefficients would capture in such a specification.

Solution

The following OLS regression specification is one that the Ministry could implement:

$$ln(Y_i) = \beta_0 + \beta_1 D_i + u_i$$

where

the subscript i runs over the observations, $i = 1, \ldots, n$;

Y_i is the *dependent variable*, annual firm sales

D_i is the *dummy variable*, $D_i = 1$ if the firm is a local Chamber of Commerce member and 0 otherwise

β_0 is the *intercept* of the regression, the population mean value of annual, non-member firm sales

β_1 is the coefficient on D_i, associating a change in D_i by one unit with a $100 \times \beta_1\%$ change in Y_i

$\beta_0 + \beta_1$ is the population mean value of annual, member firm sales

u_i is the *error* term, all the factors responsiblze for the difference between predicted and observed values

Question 1 continued on next page...

Part B

Using notation from the Potential Outcomes Framework, briefly explain the concept of the Average Treatment Effect (ATE) to the Minister, and how what they plan to estimate in A) relates to this definition.

Solution

One core challenge with evaluating the causal effect of membership in a local Chamber of Commerce on firm revenues and profits is that we cannot observe a firm i in two different states of the world: one where the firm is a member, and one where it is not. Instead, we can compare the mean outcomes of two firm groups (members vs. non-members) to learn about the true, average treatment effect (ATE) of membership using our regression specification in part A).

The ATE is defined as:

$$ATE = E(Y_i(1) - Y_i(0))$$

where $Y_i(1)$ represents the potential outcome (the annual sales) of a firm i if part of a local Chamber of Commerce and $Y_i(0)$ represents the annual sales for the **same** firm if not.

To compare the mean outcomes of the two groups of firms (members vs. non-members of a local Chamber of Commerce), we instead estimate:

$$E(Y_i(1) \mid X_i = 1) - E(Y_i(0) \mid X_i = 0)$$

where we condition our observation of Y_i on X_i, with $X_i = 1$ representing if firm i is part of a local Chamber of Commerce, and $X_i = 0$ if it is not. The above is not necessarily equal to the ATE defined earlier because X_i, or membership status, is *not independent* of potential outcomes (annual firm sales). We would have to show first that, in the absence of a local Chamber of Commerce, member and non-member firms would have to be on average identical (there would have to be no selection bias).

Unpacking the above expression we can see the effect of selection bias:

$$\begin{aligned} E(Y_i(1) \mid D_i = 1) - E(Y_i(0) \mid D_i = 1) = &\underbrace{E(Y_i(1) \mid D_i = 1) - E(Y_i(0) \mid D_i = 1)}_{\text{Average Treatment Effect on the Treated}} \\ &+ \underbrace{E(Y_i(0) \mid D_i = 1) - E(Y_i(0) \mid D_i = 0)}_{\text{Selection Bias}} \end{aligned}$$

The first term is the average treatment effect on the treated (ATT), the difference between $Y_i(1)$ and $Y_i(0)$ for the group of firms that are members of a local Chamber of Commerce. The second term estimates the selection bias by comparing $Y_i(0)$, annual sales before being a member, between members, $X_i = 1$, and non-members, $X_i = 0$. $E(Y_i(0) \mid X_i = 1)$ cannot be observed directly for both of these terms, because we cannot rerun the experiment a second time where the treatment is not applied. The bias term appears because firms that are members of a local Chamber of Commerce may have different outcomes than firms that are not members, *even in the absence of a local Chamber of Commerce.*

Question 1 continued on next page...

Part C

Referring to the expressions you use in your answer to B), explain why a randomized control trial (RCT) could be useful, and very briefly describe the basics of the RCT design for how the Ministry could set this up.

Solution

In a RCT, treatment and control groups are randomly selected, and the average outcomes of these two groups are compared after the treatment. By randomly allocating treatment status, we can use the simple difference in outcomes between treatment and control groups to estimate the ATE. This is because the randomization removes selection bias by ensuring that member firms and non-member firms are comparable in terms of both observable and unobservable characteristics and that any difference between the two groups is not due to systematic differences. This implies that the ATE = ATT.

Connecting back to the potential outcomes framework in part B) and our regression from part A), $D_i = 1$ indicates if firm i is a member of a local Chamber of Commerce (the treatment), and $D_i = 0$ if the firm is not (D_i indicates treatment status). $Y_i(1)$ are the annual sales (the potential outcome) for a firm i if treated (part of a local Chamber or Commerce), and $Y_i(0)$ are the annual sales for the same firm if not.

The ATE is then defined as:

$$ATE = E(Y_i(1) - Y_i(0))$$

If the treatment assignment D_i is assigned randomly and independent of the potential outcomes of the firms ($D_i \perp (Y_i(1), Y_i(0))$), we can compare the mean outcomes of the two groups (members vs. non-members of a local Chamber of Commerce):

$$\begin{aligned} E(Y_i(1) \mid D_i = 1) - E(Y_i(0) \mid D_i = 1) &= \underbrace{E(Y_i(1) \mid D_i = 1) - E(Y_i(0) \mid D_i = 1)}_{\text{Average Treatment Effect on the Treated}} \\ &\quad + \underbrace{E(Y_i(0) \mid D_i = 1) - E(Y_i(0) \mid D_i = 0)}_{\text{Selection Bias}} \\ &= E(Y_i(1) \mid D_i = 1) - E(Y_i(0) \mid D_i = 1) \quad \text{(selection bias is zero)} \\ &= ATT \end{aligned}$$

because the treatment status is independent of outcomes, we can also show:

$$ATT = E(Y_i(1) - Y_i(0)) = ATE$$

This means that if we randomly assign treatment status (membership in a local Chamber of Commerce) and control status (non-membership), we can estimate the ATE by comparing mean sales outcomes across the two groups of observation. An experimenter would have to randomly select from a sample of firms which ones to assign to the control group and become or remain non-members of a local Chamber of Commerce and assign the others to remain or become members of a local Chamber of Commerce. After a certain period, the experimenter will conclude the experiment and record sales data and assignment status for each firm in the study.

We can report the estimated treatment effect of the RCT design from the OLS regression, $ln(Y_i) = \beta_0 + \beta_1 D_i + u_i$. The β_1 coefficient is the difference in means between the treatment group relative to the control group. Dividing this difference by its standard error produces a t-statistic that allows us to determine whether the observed difference in annual sales between firms part of a local Chamber of Commerce and firms that are not is statistically significant.

Question 1 continued on next page...

Part D

The Ministry mentions that it has no legal authority to force firms to become members in their local Chambers of Commerce. Using notation from the Potential Outcomes Framework, explain why this information could be important for the interpretation of the results from the RCT relative to the ATE, and how the Ministry should address this concern in the RCT analysis?

Solution

In the case where assigned treatment status, D_i, cannot be enforced on firms, T_i represents the effective treatment status of a firm i. Some firms may not comply with their assigned treatment status: firms assigned to become members of a local Chamber of Commerce may opt not to, and firms assigned to be non-members might become members. However, firms' choices to deviate from their assigned treatment are not random, and T_i may be correlated with annual sales (potential outcomes). $T_i \perp (Y_i(0), Y_i(1))$ is no longer true, which means the difference between the mean outcomes of the control and treatment groups may be influenced by selection bias.

However, the initial assignment status, D_i, is still randomly assigned and may have strong predictive power on actual treatment status, T_i. So, to address our original concern, we can use D_i as an instrumental variable for T_i to estimate the causal effect. Using an instrumental variable will limit the variation in the effective treatment status, T_i, to only the part that can be predicted by the randomly assigned D_i. This will pick out only the exogneous variation of the independent variable T_i to deal with concerns of bias in the case where there is only partial compliance. To be a valid instrumental variable, D_i must satisfy three conditions: instrument relevance, instrument exogeneity, and the instrument exclusion restriction.

Instrument relevance. The instrument needs to be related strongly enough with the endogenous, explantory variables of interest T_i. So, as long as the assignment protocol is partially followed, then the actual treatment status will be partially determined by the assigned treatment status, making D_i a relevant instrumental variable.

Instrument exogenity. The instrument must be uncorrelated with any observed variables that also affect Y_i in the error term. That is, $Cor(D_i, u_i) = 0$. Since initial treatment assignment is random, then D_i is distributed independently of u_i, so the instrument is exogenous.

Instrument exclusion restriction. The instrument must not have a direct effect on Y_i except through its relationship with T_i. That is, conditional on T_i, D_i has no effect on Y_i. Again, since the initial treatment assignment is random, D_i is distributed independenty of any other variable than T_i that can have an effect on Y_i, satisfying the exclusion restriction.

Therefore, the original random treatment assignment is a valid instrumental variable, and the following is the two-stage least squares regression:

$$\begin{aligned} T_i &= \beta_0' + \beta_1' D_i + u_i' \quad \text{(first stage)} \\ ln(Y_i) &= \beta_0 + \beta_1 \widehat{T_i} + u_i \quad \text{(second stage)} \\ ln(Y_i) &= \beta_0'' + \beta_1'' D_i + u_i'' \text{ (reduced form)} \end{aligned}$$

In contrast to our original regression, our point estimate from the second stage is a weighted average of the treatment effect β_{1i}, where the weights increase with the degree to which the instrument D_i influences T_i in the first stage. This is because the treatment effect from the instrumental variables regression gives greater weights to units that respond to the instrument in the first place:

$$\widehat{\beta_{IV}} = \frac{\beta_1''}{\beta_1'} = \frac{E(Y_i|D_i = 1 - Y_i|D_i = 0)}{E(T_i|D_i = 1 - T_i|D_i = 0)} = \frac{\text{Intent-to-treat effect (ITT)}}{\text{Difference in fraction treated between groups}} = LATE$$

That is, instead of estimating the average treatment effect, or ATE, we instead estimate the local average treatment effect (LATE) after isolating the variation in T_i to the group of compliers. This is because we don't know if the firms that complied are representative to the whole sample of firms chosen for the experiment and is why the causal effect in the instrumental variable case is a local average. So, when we adjust for less than full compliance in our RCT by instrumenting the effect treatment status with assigned treatment status, our point estimate, $\widehat{\beta_{IV}}$, reflects a ratio between an "intent-to-treat" effect (ITT), the difference in mean outcomes between members and non-members of a local Chamber of Commerce, and the proportion of compliers, producing the LATE.

The average treatment effect on the treat (ATT) is then the same thing as the LATE: the causal effect of being a member of a local Chamber of Commerce estimated on the group of firms that chose to take up the experiment when offered to them. This is different from the ATE because it was not randomized which firms will actually follow adhere to their assigned treatment status. It's instead an independent decision made by each firm. In a rational world, the ATT will usually demonstrate a stronger effect than the ATE because the firms that take up the treatment probably have more to gain.

Question 1 continued on next page...

Part E

The Ministry talked to other economists, and now it is worried about spillover effects on the control group. The staff don't fully understand what the concern is, however. Briefly explain to them the intuition behind this concern, and explain how they could potentially address it when designing the RCT.

Solution

The concern behind spillover effects is that, even if we were to isolate the treatment (membership in a local Chamber of Commerce) to firms only in the treatment group and restrict membership for those in the control group, we still have to worry about the effects that the treatment has on the control group. For example, in the case where membership in a local Chamber of Commerce improves firm sales for members, increased sales may contribute to overall economic surplus and increas sales for firms in the control group. This means that the control group is no longer a reference in which nothing has changed.

One solution would be to seperate treatment and control group in some way so that the effect of nearby firms assigned to the treatment group affecting the outcomes of firms assigned to the control group by the end of the period. For example, if the true causal effect of being a member of a local Chamber of Commerce is positive, that is β_1 reflects a positive slope for D_i, these firms may experience increased annual sales. If that is the case, a better business environment may in some way stimulate demand so that non-member, control-group firms may now also see higher sales. Or the opposite: firms part of a local Chamber of Commerce might corner the market and non-member firms might experience greater losses than they otherwise would have. I have identified two different ways below an experimenter may seperate treatment and control group firms so as to mitigate spillover effects.

Geographical seperation. Assuming the treatment (assignment to a local Chamber of Commerce) has a non-zero effect on economic activity, segregating treatment and control groups to different locations may isolate the "ripple effects" of the different economic activity of members of a local Chamber of Commerce to a confined area, or at least make these effects neglible to the mean outcomes of non-member firms. For example, firms in the city of Portland, Oregon are assigned to join a local Chamber of Commerce, and the firms of Atlanta, Georgia constitute the control group. The economic consequences of firms in Portland now members of a local Chamber of Commerce may not be felt all the way in Atlanta. There is the problem of confounding variables however, in that, something about the treatment group's location may affect how they interact with the treatment. In this case, an experimenter may want to increase the sample size to reduce the effects of selection bias.

Industry stratification. Instead of, or in conjunction with, geographical seperation of treatment and control groups, it may be prudent to assign firms to treatment and control groups on an industry-wide level. For example, firms engaged in the fishing industry might be assigned to become members of a local Chamber of Commerce, whereas home furniture firms may not. In the case where the treatment increases fishing firm sales, this might have no effect on a very inelastic market for home furnishings. The same type of problem identified in the above method is relevant with this stratification method. There might be something inherent to fishing industry firms that affects the way they interact with the treatment that might not be representative of all firms. That is, there might be selection bias. In this case, an experimenter would need to select a wide variety of industries in both the treatment and control groups to "average out" the effects of this bias.

Question 2

To answer this question, we will use R and the dataset "Mexico_PS2.csv" that you can download from bCourses. The dataset contains 1153 Mexican municipalities that reported some amount of local tourism activity (measured by local hotel sales) in the year 2000. Write up the answers below in the same document as above. In addition, also attach the complete code that you used to answer the questions. You are not required to export results in LaTeX, it is enough to show your output in R (using either knitR/RMarkdown or just showing the output, e.g. with screenshots).

A) Visualize a table that lists the number of observations, the mean, the standard deviation, the minimum value and the maximum value for each of the variables in the dataset (if you had difficulties with this in the previous problem set, use the summary() command). Briefly describe what we learn from the table about the sample of Mexican municipalities.

B) Use the data to obtain an OLS point estimate of the effect of the average years of education on the logarithm of local average monthly household incomes. Show your result in a regression table, using the command stargazer(). Comment on the interpretation and statistical significance of your results.

C) List three plausible arguments why the point estimate in b) could be biased upwards or downwards relative to the true causal effect of education on monthly household incomes. Be specific in signing the bias for each argument.

D) Now your GSI suggests that the natural logarithm of kilometer distance between the center of the municipality and the nearest segment of the US-Mexico border could be a valid instrumental variable for the average years of education in a municipality (because incentives for education are higher in regions exporting to the US). List the assumptions that need to hold true for this to be correct.

E) Verify if the assumption of instrument relevance is satisfied, and export the results into the same regression table that you used above (again using the stargazer() command). Comment on the interpretation and statistical significance of your result.

F) Now estimate the 2^{nd} stage IV (TSLS) point estimate as suggested by your GSI, and show your result in the same regression table you used before (again using the stargazer() command). Comment on the interpretation and statistical significance of your result. In reference to your answer to c), is the difference between the OLS and IV point estimates as you expected or rather not?

G) Now one of your friends suggests that the distance to the US border is likely correlated with other local characteristics that affect local incomes, such as the logarithm of local tourism sales, the logarithm of the average temperature, the logarithm of the average precipitation, and the proportion of indigenous population. Estimate the first-stage regression and the reduced-form regression both before and after you add those variables as additional controls in the two regressions (estimating 4 regression specifications in total). Export your regression results in a new regression table. Comment on the results and what they imply about the validity of the instrumental variable strategy.

Question 2 continued on next page...

Part A

Visualize a table that lists the number of observations, the mean, the standard deviation, the minimum value and the maximum value for each of the variables in the dataset (if you had difficulties with this in the previous problem set, use the summary() command). Briefly describe what we learn from the table about the sample of Mexican municipalities.

Solution

Statistic	N	Mean	St. Dev.	Min	Max
year	1,153	2,000.000	0.000	2,000	2,000
municode	1,153	18,168.720	8,045.480	1,001	32,056
inc_m	1,153	2,583.081	1,257.101	299.757	14,792.910
ind_lang	1,153	0.114	0.221	0.000	0.984
educ_years	1,153	7.031	1.354	2.488	11.870
sales_hotel	1,153	27,120.140	201,033.400	1	5,229,616
logtemp	1,153	5.283	0.212	4.658	5.671
logprecip	1,153	4.304	0.562	1.790	5.746
dist_us_km	1,153	696.171	278.301	6.609	1,348.003

From the table, the sample of Mexican municipalities includes 1,153 units of observation. All units were recorded in the year 2000, as the year statistic has no variation across observations in the sample. The units of observation are Mexican municipalities, which, besides the year statistic, also record the municipality code (municode), average monthly income (inc_m), the proportion of the indigenous population (ind_lang), average years of education (educ_years), local hotel sales (sales_hotel), the logarith of the average temperature (logtemp), the logarithm of the average precipitation (logprecip), and the distance to the US border (dist_us_km).

Average monthly income. Average monthly incomes have a low mean value and a relatively large standard devation. The distribution could be right skewed as the mean is far closer to the minimum value than the maximum value.

Proportion of the indigenous population. The mean value for the indigenous proportion of Mexican municipalities population stands at 11.4%. However, the standard devation is very high, suggesting a broad distribution of proportions. Some municipalities have effectively 0% and some are near 100%.

Average years of education. In contrast to the aformentioned statistics, the variation in average years of educational attainment is smaller, suggesting a tighter distribution. Most observations have statistics close to the sample mean, with few outliers.

Local hotel sales. Local hotel sales vary massively. Many municipalities record neglible sales. The distribution appears to be right skewed with a long tail towards higher sales.

Question 2 continued on next page...

Part B

Use the data to obtain an OLS point estimate of the effect of the average years of education on the logarithm of local average monthly household incomes. Show your result in a regression table, using the command stargazer(). Comment on the interpretation and statistical significance of your results.

Solution

	Dependent variable:
	log(inc_m)
educ_years	0.183***
	(0.007)
Constant	6.485***
	(0.047)
Observations	1,153
R^2	0.400
Adjusted R^2	0.399
Residual Std. Error	0.304 (df = 1151)
F Statistic	767.224*** (df = 1; 1151)
Note:	*p<0.1; **p<0.05; ***p<0.01

The following OLS regression specification was applied to produce the results above:

$$\widehat{ln(inc\_m_i)} = \widehat{\beta_0} + \widehat{\beta_1} \times educ\_years_i$$

where

$\widehat{\beta_0}$ is the *intercept* of the regression. This sample mean value of the logarithm of local average monthly household incomes is 6.485 when average years of education is 0. This result is statistically significant at the 0.01 level (*** corresponds to a p-val below the 1% threshold). With a t-stat of $\frac{6.485}{0.047} = 137.979$ in magnitude, greater than the t-stat of 1.96 for a 5% threshold to reject $H_0 : \beta_0 = 0$, a confidence interval built with this standard error at the 5% level does not contain zero.

$\widehat{\beta_1}$ is the coefficient on $educ\_years_i$, associating a one year increase in the average years of education with an 18.3% increase in local average monthly household incomes. This result is statistically significant at the 0.01 level (*** corresponds to a p-val below the 1% threshold). With a t-stat of $\frac{0.183}{0.007} = 26.143$ in magnitude, greater than the t-stat of 1.96 for a 5% threshold to reject $H_0 : \beta_1 = 0$, a confidence interval built with this standard error at the 5% level does not contain zero.

The R^2 is 0.4, meaning 44% of the variation in the logarithm of local average monthly incomes can be explained by this model.

Question 2 continued on next page...

Part C

List three plausible arguments why the point estimate in b) could be biased upwards or downwards relative to the true causal effect of education on monthly household incomes. Be specific in signing the bias for each argument.

Solution

Below are three reasons why the point estimate $\widehat{\beta_1}$ from part B) may be biased relative to the true causal effect of education on incomes, β_1.

Omitted variable bias. Below is the formula to identify the effects of omitted variable bias on our point estimate with $educ\_years$ as our variable of interest and X as another variable currently in the error term:

$$\underbrace{\widehat{\beta_1}}_{\text{Our point estimate}} = \underbrace{\beta_1}_{\text{the true causal effect}} \underbrace{+ \beta_2 \times \frac{Cov(educ\_years, X)}{Var(educ\_years)}}_{\text{OVB bias}}$$

Monthly household incomes could be impacted by a host of other factors that could be correlated with education, such as ability or age. Ability might have a direct effect on household incomes and might also be positively correlated with education, with more able people tending to invest more into education and driving up the average years of education. As such the covariance between education and age would be positive, leading to a positive bias. Age might have a similar effect, with older people tending to have more years in education. Age might have an effect on incomes too, with more senior workers having higher incomes. Again, the covariance between these two variables would be positive. By not including these factors, age and ability, as variables in our regression, there is more undue weight given to education, overestimating the true causal effect β_1.

Measurement error. Below is the formula to identify the effects of measurement error on the explanator variable, $educ\_years$, on our point estimate:

$$\underbrace{\widehat{\beta_1}}_{\text{Our point estimate}} = \underbrace{\beta_1'}_{\text{the true causal effect}} \underbrace{- \beta_1' \times \frac{Var(e))}{Var(educ\_years)}}_{\text{ME bias}}$$

is "noise" in measurements of the explanator variable of average years of education, $educ_years$ will be correlated with the error term. The effect of measurement error on the explanatory variable is to exert negative bias on the point estimate when the true effect of average years of education on local average monthly household incomes is positive and vice versa. That is, the $bias > 0$ when $\beta_1' < 0$ and $bias < 0$ when $\beta_1' > 0$. This make sense if the "noise" is not biased in any particular direction (that is, average years of education are generally overreported as much as they are underreported). In our case then, with a positive point estimate, the measurement error bias is negative.

Reverse causality bias.

$$educ\_years \longleftrightarrow ln(inc\_m)$$

One reason that there might be reverse causality bias at play is because wealthier people (measured in terms of local average monthly household incomes) might already pursue more education than those who don't. Thus there may be a positive bias from our dependent variable now explaining some of the variation in the explanatory variable.

Question 2 continued on next page...

Part D

Now your GSI suggests that the natural logarithm of kilometer distance between the center of the municipality and the nearest segment of the US-Mexico border could be a valid instrumental variable for the average years of education in a municipality (because incentives for education are higher in regions exporting to the US). List the assumptions that need to hold true for this to be correct.

Solution

There are three assumptions that need to be true for a variable to be a valid instrument:

1. **Instrument relevance**: The potential instrument needs to be related strongly enough with the endogenous, explanatory variable.
2. **Instrument exogeneity**: The instrument must not be correlated with any other unobserved variables (in the error term) that affect the dependent variable. That is, $Cor(ln(dist\_us\_km), u) = 0$.
3. **Instrument exclusion restriction**: The instrument must not have a direct effect on the dependent variable except throught its influence on the endogenous explanatory variable. That is, conditioning for $educ\_years$, $ln(dist\_us\_km)$ has no effect on $ln(inc\_m)$, so again $Cor(ln(dist\_us\_km), u) = 0$.

Question 2 continued on next page...

Part E

Verify if the assumption of instrument relevance is satisfied, and export the results into the same regression table that you used above (again using the stargazer() command). Comment on the interpretation and statistical significance of your result.

Solution

	Dependent variable:	
	log(inc_m)	educ_years
	(1)	(2)
educ_years	0.183***	
	(0.007)	
log(dist_us_km)		−0.424***
		(0.052)
Constant	6.485***	9.738***
	(0.047)	(0.332)
Observations	1,153	1,153
R^2	0.400	0.055
Adjusted R^2	0.399	0.055
Residual Std. Error (df = 1151)	0.304	1.317
F Statistic (df = 1; 1151)	767.224***	67.597***
Note:	*p<0.1; **p<0.05; ***p<0.01	

The following first-stage regression specification was applied to produce the results for (2) above:

$$\widehat{educ\_years}_i = \widehat{\beta_0} + \widehat{\beta_1} \times ln(dist\_us\_km_i)$$

where

$\widehat{\beta_0}$ is the *intercept* of the regression. This sample mean value of the average years of education is 9.738 when the distance to the US border is 0. This result is statistically significant at the 0.01 level (*** corresponds to a p-val below the 1% threshold). With a t-stat of $\frac{9.738}{0.332} = 29.331d$ in magnitude, greater than the t-stat of 1.96 for a 5% threshold to reject $H_0 : \beta_0 = 0$, a confidence interval built with this standard error at the 5% level does not contain zero.

$\widehat{\beta_1}$ is the coefficient on $ln(dist\_us\_km)$, associating a one percent increase in the distance to the US border with 0.004 decrease in the average years of education. This result is statistically significant at the 0.01 level (*** corresponds to a p-val below the 1% threshold). With a t-stat of $c-0.4240.052 = -8.154$ in magnitude, greater than the t-stat of 1.96 for a 5% threshold to reject $H_0 : \beta_0 = 0$, a confidence interval built with this standard error at the 5% level does not contain zero.

The F-stat is 67.597, a result statistically significant at the 0.01 level and above 10, meaning $log(dist\_us\_km)$ is related strongly enough with $educ\_years$ to pass the instrument relevance test.

Question 2 continued on next page...

Part F

Now estimate the 2nd stage IV (TSLS) point estimate as suggested by your GSI, and show your result in the same regression table you used before (again using the stargazer() command). Comment on the interpretation and statistical significance of your result. In reference to your answer to c), is the difference between the OLS and IV point estimates as you expected or rather not?

Solution

	Dependent variable:		
	log(inc_m)	educ_years	log(inc_m)
	OLS	*OLS*	*IV*
	(1)	(2)	(3)
educ_years	0.183***		0.453***
	(0.007)		(0.044)
log(dist_us_km)		−0.424***	
		(0.052)	
Constant	6.485***	9.738***	4.589***
	(0.047)	(0.332)	(0.309)
Observations	1,153	1,153	1,153
R^2	0.400	0.055	−0.465
Adjusted R^2	0.399	0.055	−0.466
Residual Std. Error (df = 1151)	0.304	1.317	0.475
F Statistic (df = 1; 1151)	767.224***	67.597***	
Note:			*p<0.1; **p<0.05; ***p<0.01

The following second-stage regression specification was applied to produce the results for (3) above:

$$\widehat{ln(inc\_m_i)} = \widehat{\beta_0} + \widehat{\beta_1} \times \widehat{educ\_years}_i$$

where

$\widehat{educ\_years}_i$ are the predicted values for the endogenous explanatory variable, average years of education, instrumented on distance to the US border in the first-stage model from (1).

$\widehat{\beta_0}$ is the *intercept* of the regression. This sample mean value of the logarithm of local average household income is 4.589 when the average years of education is 0. This result is statistically significant at the 0.01 level.

$\widehat{\beta_1}$ is the coefficient on $educ\_years_i$, associating a one year increase in the average years of education with a 45.3% increase in local average monthly household incomes. This result is statistically significant at the 0.01 level.

The difference between the OLS and IV point estimates for β_1 was different that what I expected because in part C I identified three reasons why the OLS estimate might be the result of positive bias. However, the new IV point estimate suggests there was negative bias deflating the true value of β_1, or the effect of an extra year of average years of education on local average monthly incomes.

Question 2 continued on next page...

Part G

Now one of your friends suggests that the distance to the US border is likely correlated with other local characteristics that affect local incomes, such as the logarithm of local tourism sales, the logarithm of the average temperature, the logarithm of the average precipitation, and the proportion of indigenous population. Estimate the first-stage regression and the reduced-form regression both before and after you add those variables as additional controls in the two regressions (estimating 4 regression specifications in total). Export your regression results in a new regression table. Comment on the results and what they imply about the validity of the instrumental variable strategy.

Solution

	Dependent variable:			
	educ_years		log(inc_m)	
	(1)	(2)	(3)	(4)
log(dist_us_km)	-0.424^{***}	-0.034	-0.192^{***}	-0.091^{***}
	(0.052)	(0.048)	(0.014)	(0.015)
log(sales_hotel)		0.239^{***}		0.042^{***}
		(0.012)		(0.004)
logtemp		-0.714^{***}		-0.105^{**}
		(0.153)		(0.047)
logprecip		-0.432^{***}		-0.124^{***}
		(0.072)		(0.022)
ind_lang		-1.246^{***}		-0.439^{***}
		(0.150)		(0.046)
Constant	9.738^{***}	11.537^{***}	9.000^{***}	9.233^{***}
	(0.332)	(0.787)	(0.092)	(0.241)
Observations	1,153	1,153	1,153	1,153
R^2	0.055	0.435	0.135	0.367
Adjusted R^2	0.055	0.433	0.135	0.365
Residual Std. Error	1.317 (df = 1151)	1.020 (df = 1147)	0.365 (df = 1151)	0.313 (df = 1147)
F Statistic	67.597^{***} (df = 1)	176.598^{***} (df = 5)	180.291^{***} (df = 1)	133.152^{***} (df = 5)

Note: *p<0.1; **p<0.05; ***p<0.01

The model that produced results for (1) was already described in part E). The following first-stage and reduced-form regression specifications were applied to produce the results for (2), (3), and (4):

(2) $\widehat{educ\_years}_i = \widehat{\alpha_0} + \widehat{\alpha_1}\, ln(dist\_us\_km_i) + \widehat{\alpha_2}\, ln(sales\_hotel_i) + \widehat{\alpha_3}\, logtemp_i + \widehat{\alpha_4}\, logprecip_i + \widehat{\alpha_5}\, ind\_lang_i$

(3) $\widehat{ln(inc\_m_i)} = \widehat{\beta_0} + \widehat{\beta_1}\, ln(dist\_us\_km_i)$

(4) $\widehat{ln(inc\_m_i)} = \widehat{\gamma_0} + \widehat{\gamma_1}\, ln(dist\_us\_km_i) + \widehat{\gamma_2}\, ln(sales\_hotel_i) + \widehat{\gamma_3}\, logtemp_i + \widehat{\gamma_4}\, logprecip_i + \widehat{\gamma_5}\, ind\_lang_i$

Question 2 continued on next page...

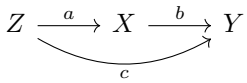
From the (2) first-stage regression of average years of education on the logarithm of distance to the US border alongside the additional control variables, the coefficent on $log(dist\_us\_km_i)$ is no longer statistically significant at either the 0.01, 0.05, or 0.1 levels in contrast to (1), the model without the control variables. This suggests that our instrumental variables strategy has a weak first stage, and our IV point estimate is unreliable. This unreliability in our IV point estimate might be validated for a skeptic when computing $\widehat{\beta_{IV}}$ as a ratio between the controlled first-stage point estimate and the controlled reduced-form point estimate:

$$\frac{\widehat{\gamma_1}}{\widehat{\alpha_1}} = \frac{-0.091}{-0.034} = 2.676$$

This is a relatively huge IV value and might be a symptom of having a weak instrument (it suggests that, conditioning for the other control variables, a one percent change in the distance to the US border corresponds to a 267.6% change in local average monthly household incomes). Economic theory and real world experience point to this conclusion not being reliable.

Even for regression (1) the instrument effect on education is quite miniscule (from part E), the point estimate relates a one percent change in the distance to the US border with a 0.004 unit decrease in the average years of education) and now with the controls in place it doesn't have a statistically significant effect at all.

It's possible that our instrument also does not pass the instrument exclusion restriction. While the effect of the logarithm of distance to the US border might not be statistically significant, its effect on the logarithm of local average monthly household incomes is. In this case, our instrument does not have an affect on the dependent variable only through our endogenous explanatory variable but rather through other ways.



For a potential instrumental variable to be a valid candidate, the effect it has on the dependent variable must only act through the explanatory one (through 'a' and then 'b' in the diagram above). If it impacts Y through 'c' for instance, it is not valid as the explanatory variable is no longer exogenous. In regression (4) we can see how, in contrast to (2), the point estimate for the causal effect of $ln(dist\_us\_km)$ on $ln(inc\_m)$ is statistically significant at the 0.01 level. That is, our point estimate for γ_1 is statistically significant.

Code

```
    #install.packages(c("dplyr", "stargazer", "knitr", "ivreg"))
    library(dplyr)
    library(stargazer)
    library(knitr)
    library(ivreg)

    setwd(paste("/home/zachary/Desktop/ECON 140/Problem Set 2/code", sep=""))
    mexico <- read.csv("Mexico_PS2.csv")

    # part a
    stargazer(mexico, type = "text")
```

```
##
## ============================================================
## Statistic     N      Mean     St. Dev.     Min      Max
## ------------------------------------------------------------
## year        1,153 2,000.000     0.000     2,000    2,000
## municode    1,153 18,168.720  8,045.480   1,001    32,056
## inc_m       1,153 2,583.081   1,257.101  299.757 14,792.910
## ind_lang    1,153   0.114       0.221     0.000    0.984
## educ_years  1,153   7.031       1.354     2.488    11.870
## sales_hotel 1,153 27,120.140 201,033.400    1    5,229,616
## logtemp     1,153   5.283       0.212     4.658    5.671
## logprecip   1,153   4.304       0.562     1.790    5.746
## dist_us_km  1,153  696.171     278.301    6.609  1,348.003
## ------------------------------------------------------------
```

```
    # part b
    ols <- lm(log(inc_m) ~ educ_years, mexico)
    stargazer(ols, type = "text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                              log(inc_m)
## -----------------------------------------------
## educ_years                    0.183***
##                                (0.007)
##
## Constant                      6.485***
##                                (0.047)
##
## -----------------------------------------------
## Observations                    1,153
## R2                              0.400
## Adjusted R2                     0.399
```

```
## Residual Std. Error      0.304 (df = 1151)
## F Statistic          767.224*** (df = 1; 1151)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
    # part e
    first_stage <- lm(educ_years ~ log(dist_us_km), mexico)
    stargazer(ols, first_stage, type = "text")
```

```
##
## ===============================================================
##                                    Dependent variable:
##                                ---------------------------
##                                  log(inc_m)    educ_years
##                                     (1)            (2)
## ---------------------------------------------------------------
## educ_years                         0.183***
##                                    (0.007)
##
## log(dist_us_km)                                   -0.424***
##                                                    (0.052)
##
## Constant                           6.485***       9.738***
##                                    (0.047)        (0.332)
##
## ---------------------------------------------------------------
## Observations                        1,153          1,153
## R2                                  0.400          0.055
## Adjusted R2                         0.399          0.055
## Residual Std. Error (df = 1151)     0.304          1.317
## F Statistic (df = 1; 1151)        767.224***     67.597***
## ===============================================================
## Note:                            *p<0.1; **p<0.05; ***p<0.01
```

```
    # part f
    second_stage <- ivreg(log(inc_m) ~ educ_years | log(dist_us_km), data = mexico)
    stargazer(ols, first_stage, second_stage, type = "text")
```

```
##
## ===================================================================
##                                        Dependent variable:
##                                    ---------------------------------
##                                    log(inc_m) educ_years  log(inc_m)
##                                       OLS        OLS     instrumental
##                                                            variable
##                                       (1)        (2)         (3)
## -------------------------------------------------------------------
## educ_years                          0.183***               0.453***
##                                     (0.007)                (0.044)
##
```

```
## log(dist_us_km)                              -0.424***
##                                               (0.052)
##
## Constant                          6.485***   9.738***    4.589***
##                                   (0.047)    (0.332)     (0.309)
##
## ---------------------------------------------------------------------
## Observations                       1,153      1,153       1,153
## R2                                 0.400      0.055       -0.465
## Adjusted R2                        0.399      0.055       -0.466
## Residual Std. Error (df = 1151)    0.304      1.317       0.475
## F Statistic (df = 1; 1151)      767.224*** 67.597***
## =====================================================================
## Note:                                    *p<0.1; **p<0.05; ***p<0.01
```

```
    # part g
    first_stage_controls <- lm(educ_years ~ log(dist_us_km) + log(sales_hotel) +
                                            logtemp + logprecip + ind_lang, mexico)
    reduced_form <- lm(log(inc_m) ~ log(dist_us_km), mexico)
    reduced_form_controls <- lm(log(inc_m) ~ log(dist_us_km) + log(sales_hotel) +
                                            logtemp + logprecip + ind_lang, mexico)
    #stargazer(first_stage, first_stage_controls, reduced_form,
    #            reduced_form_controls, type = "text")
```