Empirical Methods - Exercise 3

Solutions

November 20-21, 2019

Theory Q1 - Coefficient Interpretation

Table 1:

	Dependent variable:		
	consumption		
	(1)	(2)	(3)
income	0.267^{***}	0.254^{***}	0.254^{***}
	(0.006)	(0.006)	(0.006)
fam_size		625.431^{***}	625.445^{***}
		(75.731)	(75.726)
house			1,395.781
			(1,021.504)
Constant	$5,800.441^{***}$	$4,429.220^{***}$	$4,413.947^{***}$
	(132.779)	(212.165)	(212.445)
Observations	6,371	6,371	6,371
R^2	0.221	0.229	0.229
Note:			$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

(a) *Write down the regression model estimated in column (1). What is the value of the coefficient on income? What is its interpretation?*

Answer:

The regression model is:

$$C_i = \beta_0 + \beta_1 Y_i + \epsilon_i \tag{1}$$

where the subscript i identifies a family.

This model tells us that a 1 unit increase of income is associated to an increase in consumption of β_1 units. If we assume that both consumption and income are measured in dollars, then our estimated model implies that a \$100 increase in income is associated with \$26.7 increase in family consumption, on average. In other words US families spend on average between $\frac{1}{3}$ and $\frac{1}{4}$ of their income on consumption, according to our estimated model.

(b) *Now write down the regression model in column (2). What is the interpretation of the coefficient on income? What is the interpretation of the coefficient on family size?*

Answer:

The regression model is:

$$C_i = \beta_0 + \beta_1 Y_i + \beta_2 N_i + \epsilon_i \tag{2}$$

where the subscript i identifies a family. This model tells us that a 1 unit increase of income is related with an increase in consumption of β_1 units for families of the same size. Also an extra family member icreases consuption by β_2 USD, all else equal. According to our estimated equation, this means that 1 extra dollar of income increases consumption by 0.25 dollars, on average and keeping the size of the family fixed. Moreover, adding an extra family member is associated with an increase in family *yearly* consumption of $2,500, on average, all else equal.

(c) *In the last column you add whether i owns a home. What is now the value of the coefficient on income? How do you interpret it?*

Answer:

This means that $1 extra increases family consuption by 0.25 US dollars, ceteris paribus. As we can observe, the estimated coefficient for the dummy variable "house" is not statistically significant, despite its large magnitude.

(d) *Why are your answers changing (if they are)? Which is "right", i.e. which interpretation should you be trying to get?*

Answer:

Our answers are changing because by including additional variables into the model we acknowledge that family's level of consumption can change not only with changes in income, but also with changes in other variables. For example, when a child is born (i.e., family size changes), a family is usually forced to consume more than they did before, even though parents' income might not change. Including other relevant factors that are likely to influence income is the "right" way to go, since we are likely to try to use the model's results either for prediction or for policy purposes (for example, to assess impact of a tax cut - a policy change that increases households' income - on consumers' spending behavior), and we will be interested in knowing how different types of families are likely to be affected.

Theory Q2 - Omitted Variable Bias

Suppose the true data generation process for a student's salary in their first job after they graduate with a Master's degree (their "starting salary") is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

where Y_i = the starting salary of individual i, X_{1i} = i's Grade Point Average (GPA) in their Master's coursework, and X_{2i} = a dummy for whether their Master's degree was in economics or finance (versus, e.g., history or literature), and the standard CLRM assumptions hold, especially that $E(\epsilon_i | X_{1i}, X_{2i}) = 0$

Suppose instead that you estimate the model

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + \epsilon_i$$

(a) *Write down the formula for $\hat{\alpha}_1$ and calculate $E(\hat{\alpha}_1|X)$.*

Answer:

$$
\begin{aligned}
\hat{\alpha_1} &= \frac{Cov(x_{i1}, y_i)}{Var(x_{i1})} = \frac{\sum_i (x_{i1} - \bar{x}_1)(y_i - \bar{y})}{\sum_i (x_{i1} - \bar{x}_1)^2} \\
&= \frac{\sum_i (x_{i1} - \bar{x}_1)y_i - \bar{y}\sum_i (x_{i1} - \bar{x}_1)}{\sum_i (x_{i1} - \bar{x}_1)^2} \\
&= \frac{\sum_i (x_{i1} - \bar{x}_1)y_i}{\sum_i (x_{i1} - \bar{x}_1)^2} \\
&= \frac{\sum_i (x_{i1} - \bar{x}_1)(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i)}{\sum_i (x_{i1} - \bar{x}_1)^2} \\
&= \beta_0 \frac{\sum_i (x_{i1} - \bar{x}_1)}{\sum_i (x_{i1} - \bar{x}_1)^2} + \beta_1 \frac{\sum_i (x_{i1} - \bar{x}_1)x_{i1}}{\sum_i (x_{i1} - \bar{x}_1)^2} + \beta_2 \frac{\sum_i (x_{i1} - \bar{x}_1)x_{i2}}{\sum_i (x_{i1} - \bar{x}_1)^2} \\
&\quad + \frac{\sum_i (x_{i1} - \bar{x}_1)\epsilon_i}{\sum_i (x_{i1} - \bar{x}_1)^2} \\
&= \ldots
\end{aligned}
$$

(continues)

$$\hat{\alpha_1} = 0 + \beta_1 + \beta_2 \frac{\sum_i (x_{i1} - \bar{x}_1) x_{i2}}{\sum_i (x_{i1} - \bar{x}_1)^2} + \frac{\sum_i (x_{i1} - \bar{x}_1) \epsilon_i}{\sum_i (x_{i1} - \bar{x}_1)^2}$$

So that:

$$\begin{aligned} E(\hat{\alpha_1} | x_1, x_2) &= \beta_1 + \beta_2 \frac{\sum_i (x_{i1} - \bar{x}_1) x_{i2}}{\sum_i (x_{i1} - \bar{x}_1)^2} + E\left[\frac{\sum_i (x_{i1} - \bar{x}_1) \epsilon_i}{\sum_i (x_{i1} - \bar{x}_1)^2}\right] \\ &= \beta_1 + \beta_2 \frac{\sum_i (x_{i1} - \bar{x}_1) x_{i2}}{\sum_i (x_{i1} - \bar{x}_1)^2} \end{aligned}$$

As a result,$\hat{\alpha_1}$ is unbiased if:

(i) $\beta_2 = 0$; or

(ii) $\sum_i (x_{i1} - \bar{x}_1) x_{i2} = 0$.

(b) *Is $\hat{\alpha}_1$ likely to be unbiased? Why or why not?*

Answer:

No because on one side it is very likely that $\beta_2 \neq 0$ since the field of degree is presumably correlated with the first salary, and on the other side $\frac{Cov(x_1,x_2)}{Var(x_1)} \neq 0$ since the type of degree is very likely correlated with GPA score.

(c) *Given your answer to question (b), do you think any bias will be positive or negative? Explain. If you need to make an assumption in order to answer the question, state your assumption clearly and give the reasons that you made it.*

Answer:

To answer this question, we need to make assumptions about the correlation between GPA and a major in either Economics or Finance. This correlation is probably negative - average grades are generally lower in more science-y subjects. Assuming that $\beta_2 > 0$, i.e. a major in Economics or Finance rewards in the labor market, while $Cov(x_1, x_2) < 0$, we expect our estimate to be biased downward (i.e., we have *negative bias*).

(d) *Suppose the true data generating process included another variable, X_{3i}, which measured the time spent per week by student i on extracurricular activities (e.g. sports, travel, etc.), and that you were able to include this in your model, i.e.:*

$$\begin{array}{rcl} \text{Truth:} & Y_i & = & \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i \\ \text{You estimate:} & Y_i & = & \alpha_0 + \alpha_1 X_{1i} + \alpha_3 X_{3i} + \varepsilon_i \end{array}$$

(Note that you are still omitting X_{2i}, the dummy for the Master's degree.)
How does the addition of this new variable change your answer to question (c), if at all? In particular, do you think any bias on $\hat{\alpha}_1$ is likely to be greater or less than your answer in part (c)?

Answer:

When we add a new variable, the thing that determines bias on $\hat{\alpha}_1$ is the correlation between the variable in question (GPA) and the omitted variable (Economics Major) after "netting out" the effect of the other variables in the model (number of hours spent on extracurricular activities). That is, we have to think about what is $Corr(GPA, EconMajor)$ after controlling for time on extracurricular activities.

First, we may think that, assuming sciency degrees are more demanding, sciency majors do less in extracurricular activities. As a result, students enrolled in sciency degrees may have more time for studying, relative to others. If this was the case, the gap in GPA between the two groups will shrink a bit, before we included *ExtraAct* in our regression model.

Once we control for extracurricular activities however, we may imagine that there is an even greater advantage in GPA for non-sciency majors, implying an even larger bias relative to before. At the same time however, one may also argue that, conditional on spending lots of time in extracurricular activities, $Corr(GPA, EconMajor) > 0$: maybe those that spend lots of time on extracurricular activities are also those that can actually handle both things (extremely good/motivated students).

The thought process should be something like: (a) high extracurriculars explains some of people???s low GPAs, (b) high extracur- riculars is probably uncorrelated with majors, (c) what then is the correlation between extracurricular-adjusted GPA and being an econ major? Probably still negative? As you may see, what is important here is not the final answer, but the thought process!

Empirical Application Question - Dealing with Measurement Error

Dataset *indicators.csv* contains values for the following 8 variables in 2015: country, countrycode, mortality rate reported by UN, # deaths in hospitals, mortality rate reported by goverments, corruption index reported by UN and a variable proxing the rule of law of the country (i.e. how laws and regulations are actually enforceable in the country).

In this Exercise we will try to understand what are the different types of measurement error and what their consequences can be when estimating a model. To do so, we will analyze the relationship between corruption and child mortality. The corruption indexes are constructed such that higher values of the index indicate that the country is *more* corrupt. Furthermore, to ensure comparability, all indexes are standardized with a mean of zero and standard deviation of one.

(a) *Do you think these two variables are likely to be subject to measurement error? Explain.*

Answer:

Those two variables are very likely to be subject to measurement error. Corruption is usually an index that is based on observing specific situations like fraud, slow justice, bribery and so on. So it usually aggregates different things. It is also possible that those components are misreported in a non-random way (eg. corrupted governments who bribe international observers). Mortality as well is subject to ME. It may be hard in some countries to keep track of all the deaths when not all births are registered. In this respect, measurement error for mortality rate may well be correlated with the economic development of the country (poorer countries may have less resources to be able to properly register this type of data).

(b) *Suppose you believe that the corruption and mortality scores reported by the UN are the most reliable. Regress mortality on corruption using these measures.*

(b.i) *What is your OLS estimate of the relationship between them? Call your estimate* $\hat{\beta}$*. What is the p-value from the one-sided hypothesis test that* $\hat{\beta} > 0$*?*

Table 2: Answer b.i

	Dependent variable:
	mortalityun
corruptionun	0.626^{***}
	(0.083)
Constant	0.00000
	(0.083)
Observations	90
R^2	0.392
Adjusted R^2	0.385
Residual Std. Error	0.784 (df = 88)
F Statistic	56.685^{***} (df = 1; 88)
Note:	*p<0.1; **p<0.05; ***p<0.01

The relationship between corruption and mortality is positive: more corrupted countries on average tend to have higher levels of under 5 child mortality.

$H_0 : \hat{\beta} \leq 0$ and the alternative $H_1 : \hat{\beta} > 0$. Our t statistic for the one sided test is:

$t = \frac{\beta_{corruption}}{se(\beta_{corruption})} = \frac{0.626}{0.083} = 7.54$

By looking at the table for the t-statistic, with $N - K = 90 - 2 = 88$ degrees of freedom, the probability that we get a t-stat of 7.54 when the null hypothesis is true is smaller than 0.01, i.e. $P(t_{88} \geq 7.54) = 1 - F(7.54) < 0.01$. As a result, we can reject the null hypothesis at any standard significance level.

(b.ii) *Suppose the CLRM assumptions are satisfied: how do you interpret $\hat{\beta}$? Is this a large or small effect in your opinion?*

Answer:

We should keep in mind that both *corruptionun* and *mortalityun* are standardized variables. By construction, this implies that they have mean zero and s.d. one. It is standard to interpret their estimated coefficients in terms of standard deviations (also note that they indeed do not have a unit of measurement).

As such, a 1 standard deviation increase in the corruption score increases mortality rate by 0.63 standard deviations, on average - which arguably is a large effect.

Let's briefly revise together a model with standardized variables. Consider the simple following model: $y = \beta_0 + \beta_1 x_1 + \epsilon$.

Subtract $\bar{y}$ from both sides: $y - \bar{y} = \beta_0 + \beta_1 x_1 + \epsilon - \bar{y}$

Now, remember that the constant β_0 is equivalent to: $\beta_0 = \bar{y} - \beta_1 \bar{x}_1$. Plug this into the previous equation and we get:

$$\begin{aligned} y - \bar{y} &= \bar{y} - \beta_1 \bar{x}_1 + \beta_1 x_1 + \epsilon - \bar{y} \\ &= \beta_1 (x_1 - \bar{x}_1) + \epsilon \\ &= \beta_1 \frac{s_1}{s_1} (x_1 - \bar{x}_1) + \epsilon \\ &= \beta_1 s_1 \tilde{x}_1 + \epsilon \end{aligned}$$

where $\tilde{x}_1$ is indeed a standardized measure. We now have to standardize also the dependent variable. So, we divide both sides of the previous equation by s_y i.e., the standard deviation of y, and we get:

$$\frac{y - \bar{y}}{s_y} = \beta_1 \frac{s_1}{s_y} \tilde{x}_1 + \frac{\epsilon}{s_y}$$

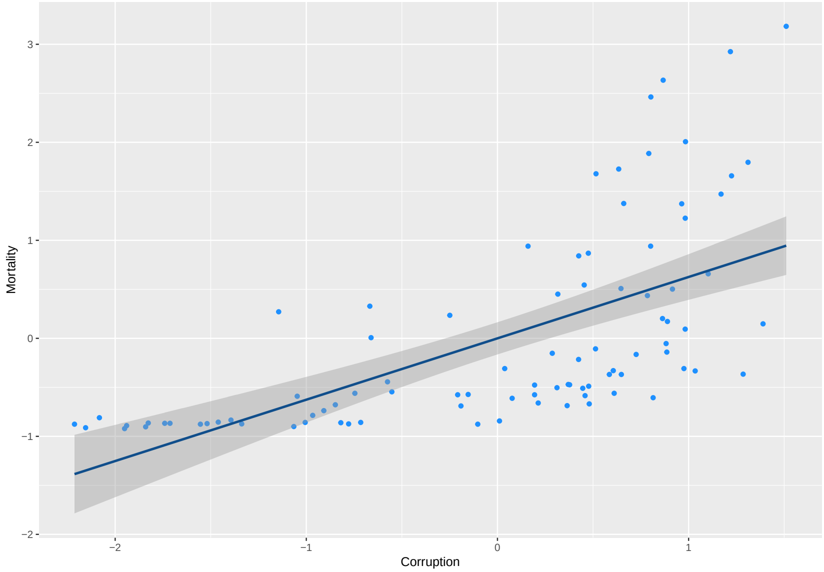
So that we finally get:

$$\tilde{y} = \tilde{\beta}_1 \tilde{x}_1 + \tilde{\epsilon}$$

(Note that indeed with a standardized dependent variable, the model has no constant!)

(b.iii) *Make a graph with the scatter plot of mortality and corruption together with the fitted regression line and confidence intervals. For the rest of the exercise, suppose this is the “true” relationship between the two variables.*

```r
g1 <- ggplot(indicators, mapping =
      aes(indicators$corruptionun,indicators$mortalityun)) +
      geom_point(colour = "dodgerblue") +
      geom_smooth(method = "lm", se=TRUE, colour = "dodgerblue4") +
      labs(x = "Corruption", y = "Mortality")
```


Mortality
3
2
1
0
−1
−2
−2
−1
0
1
Corruption

(c) *Suppose now that official mortality data (i.e. mortalityun) are not available. However, you have access to hospitals records in each country from which you - with much time and effort - manually extracted the number of deaths of infants under the age of 5 to build your mortality index. Call this index "hospital_deaths". It's possible that you made mistakes doing this, but you are willing to assume that any such mistakes were probably random.*

(c.i) *Do you think this new variable is likely to satisfy the conditions of classical measurement error? What findings do you expect from regressing hospital_deaths on corruptionun? Explain.*

Answer:

This variable likely suffers from classical measurement error, which implies that we expect no bias but larger standard errors. As long as any mistakes that were done in data collection are random (e.g., are not idiosyncratic to a specific hospital or country), they only increase the noise, and hence the standard errors.

We may hence still expect a positive estimated coefficient for corruption.

(c.ii) *Regress hospital_deaths on corruptionun. How does your coefficient estimate compare to that you estimated in question (b)? Is this consistent with your expectations? Explain.*

Answer:

The results are consistent with what we expected: the effect is a bit smaller and standard errors are larger when we compare column (2) to column (1) below.

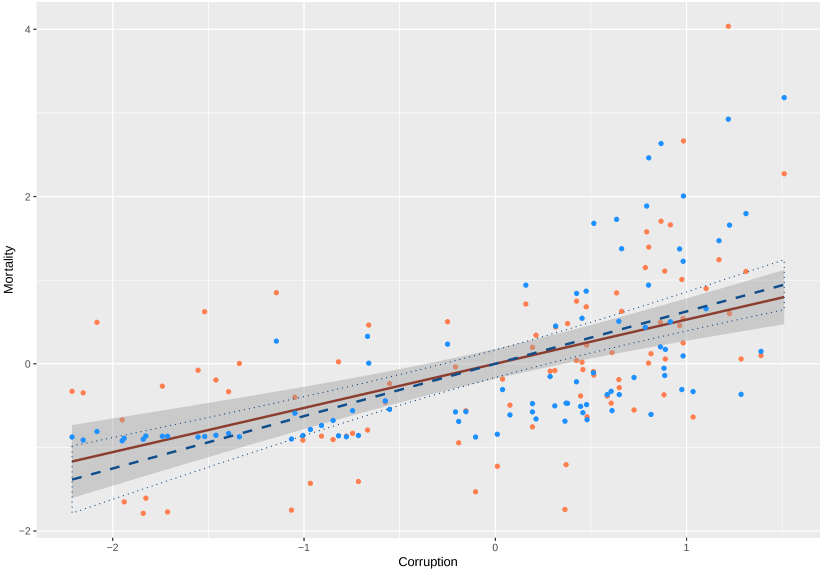
Table 3: Answer c.ii

	Dependent variable:	
	mortalityun	hospital_deaths
	(1)	(2)
corruptionun	0.626***	0.528***
	(0.083)	(0.091)
Constant	0.00000	0.00000
	(0.083)	(0.090)
Observations	90	90
R^2	0.392	0.279
Adjusted R^2	0.385	0.271
Residual Std. Error (df = 88)	0.784	0.854
F Statistic (df = 1; 88)	56.685***	34.057***
Note:	*p<0.1; **p<0.05; ***p<0.01	

(c.iii) *Plot in a single figure the scatterplot of both of your mortality variables against corruptionun as well as each of your regression lines. How do they differ in terms of standard errors and confidence intervals? Is this consistent with your expectations? Explain.*

Answer:

```r
g2 <- ggplot(indicators) +
  geom_point(aes(indicators$corruptionun,indicators$hospital_deaths), colour = "coral") +
    geom_smooth(mapping = aes(indicators$corruptionun,indicators$hospital_deaths)
                , method = "lm", se=TRUE,  colour = "coral4") +
  geom_point(aes(indicators$corruptionun,indicators$mortalityun), colour = "dodgerblue") +
    geom_smooth(mapping = aes(indicators$corruptionun,indicators$mortalityun)
                , method = "lm", se=FALSE, colour = "dodgerblue4"
                , linetype = "dashed", fill = "NA") +
      labs(x = "Corruption", y = "Mortality") +
  stat_smooth(mapping = aes(indicators$corruptionun,indicators$mortalityun)
              , method="lm",fill=NA,colour="dodgerblue4",linetype=3,geom="ribbon")
```


4
2
0
−2
Mortality
−2
−1
0
1
Corruption

As before blue dots are official UN mortality data; orange ones are # hospital deaths by country.

If the measurement error on the dependent variable (mortality) is random, the estimates will be different but the regression line will not diverge too much from the true one. However, standard error will be bigger, leading to less precise estimates. As we can see, the red distribution is more “spread around” vertically than the original blue dots.

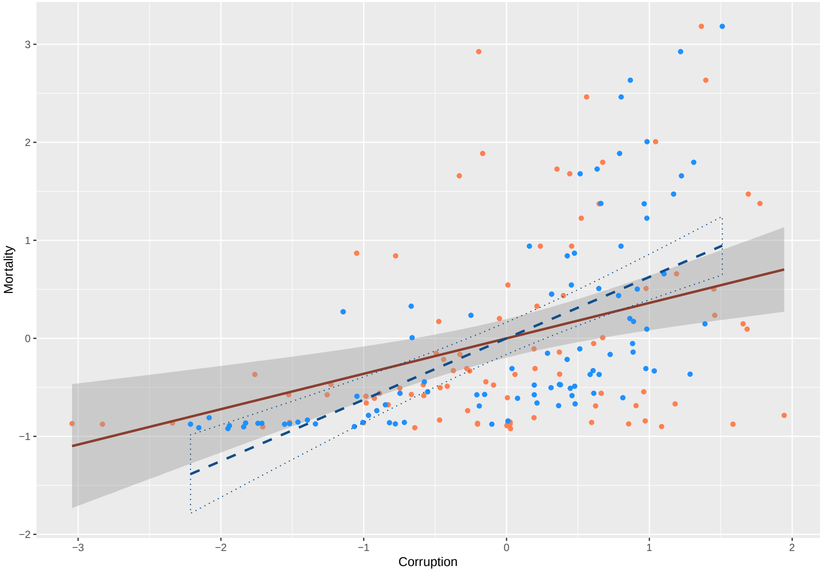
(d) *Suppose now that your UN mortality index is available but your UN corruption index (corruptionun) is not. Instead, you have the UN index for Rule of Law. This index is based on different measures of corruption and is highly correlated with the UN corruption index. You can safely assume that any error between the two is random.*

(d.i) *Regress mortalityun on ruleoflaw. How does the coefficient compare to that from question (b)? Is this consistent with your expectations? Explain.*

Table 4: Answer d.i

	Dependent variable:	
	mortalityun	mortalityun
	(1)	(2)
corruptionun	0.626^{***}	
	(0.083)	
ruleoflaw		0.361^{***}
		(0.099)
Constant	0.00000	0.000
	(0.083)	(0.099)
Observations	90	90
R^2	0.392	0.131
Adjusted R^2	0.385	0.121
Residual Std. Error (df = 88)	0.784	0.938
F Statistic (df = 1; 88)	56.685^{***}	13.215^{***}

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


Mortality
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
Corruption

As seen in the lecture, when the measurement error is on the independent variable (corruption), even if it is random, it is very likely to bias the estimates. As we can see, the red dots (which show the relationship between rule of law and mortality rate) are much more spread over the x axis, which will bias *downwards* the coefficient of interest. This is called *attenuation bias*. The estimated coefficient in this model is indeed (almost 50%) lower than before.

(e) *As in question (1c), suppose that mortalityun is not available. Nor were you able to collect yourself the raw data. What is available is a mortality rate self-reported by each country in the data called govmort.*

(e.i) *Do you think this variable is likely to satisfy the conditions of classical measurement error? Explain.*

Answer:

Measurement error is on the dependent variable again. However, if more corrupted governments are more likely to report lower levels of mortality, this measurement error will not be random but rather systematic. This means that the assumptions behind the "classical" measurement error will not hold. In particular, $E(\eta|X) \neq 0$.

(e.ii) *If yes, leave this question blank. If not, what is the likely sign of any bias in the coefficient on corruptionun from a regression of govmort on corruptionun?*

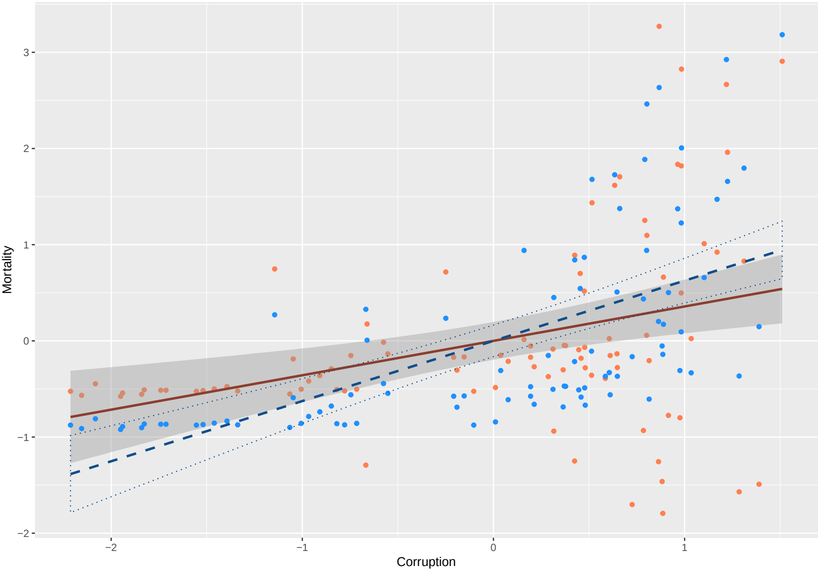
Answer:

One can show (see PS3) that:
$E(\hat{\beta}) = \beta + (X'X)^{-1}X'E_X[E(\epsilon^*|X)] + (X'X)^{-1}X'E_X[E(\eta|X)]$. If measurement error is systematic, the last term does not disappear, implying $E(\hat{\beta}) = \beta + (X'X)^{-1}X'E_X[E(\eta|X)]$. The sign of the bias depends on $(X'X)^{-1}X'E(\eta|X)$, i.e. on $Cov(corruptionun_i, \eta_i)$. If we assume that more corrupted governments report lower levels of mortality, this correlation is negative, and hence we would expect to find a downward biased coefficient.

(e.iii) *Regress govmort on corruptionun. How does the coefficient compare to (b)? Is this result consistent with your expectations? Why or why not?*

Table 5: Answer e.iii

	Dependent variable:	
	mortalityun	govmort
	(1)	(2)
corruptionun	0.626^{***}	0.358^{***}
	(0.083)	(0.100)
Constant	0.00000	0.00000
	(0.083)	(0.099)
Observations	90	90
R^2	0.392	0.128
Adjusted R^2	0.385	0.118
Residual Std. Error (df = 88)	0.784	0.939
F Statistic (df = 1; 88)	56.685^{***}	12.902^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	


Mortality
Corruption
3
2
1
0
−1
−2
−2
−1
0
1

Looking at the scatterplot, highly corrupted countries tend to report lower than true levels of mortality (assuming UN values are indeed the true ones): the vertical gap between “true” values and reported ones is indeed way larger among more corrupted countries. Overall, estimates from the regression with self-reported mortality values are biased and the relationship between the two variables appears to be less strong (precisely because more corrupted governments under-report mortality values).

(f) *Which of the above four cases - (c) to (e) - do you believe to be more dangerous in terms of identification? Explain.*

Answer:

(1) Classical measurement error is worse if arises in the x than in y: hence, d worse than c;

(2) Systematic measurement error is worse than Classical measurement error: e worse than c;

$\longrightarrow$ c is the "least-worst" scenario (classical error in y)

(3) Is classical measurement error in x worse than systematic measurement error in y? If the measurement error in x is not too severe, and hence the attenuation bias is not too bad, then systematic error in y is worse than classical error in x.

To sum up: e worse than d worse than c.