Problem Set 3 Empirical Methods

Patrick Glettig (13-252-143), Carmine Ragone (14-995-021)

19.11.2018

Paper and Pencil Questions

1 Coefficient Interpretation

a)

$$consumption_i = \beta_1 + \beta_2 income_i + \delta_i$$

$$consumption_i = 5800.441 + 0.267 income_i + \delta_i$$

A 1-unit increase in income (1 USD) would effect in an increased consumption of 0.267 USD.

b)

$$consumption_i = \beta_1 + \beta_2 income_i + \beta_3 famsize + \upsilon_i$$

$$consumption_i = 4429.22 + 0.254 income_i + 625.431 famsize + \upsilon_i$$

A 1-unit increase in income (1 USD) would effect in an increased consumption of 0.254 USD, holding family size constant. Every additional family member increases consumption by 625 USD.

c)

The coefficient on income is still 0.254, so the effect size did not change from the model in b). But we can now interpret it as: a 1-unit increase in income (1 USD) would effect in an increased consumption of 0.254 USD, controlling for family size and house owners.

d)

The coefficients change from model (1) to model (2) as part of the coefficient in *income* also reflected in $famsize$, a intuitive explanation would be that people with larger families have to work (& hence earn) more to feed the family which biases the effect of income. The coefficients β_2 and β_3 barely change from model (2) to model (3), so the fact that someone is owning a house did not explain any part of the coefficients on *income* and $famsize$. As this makes model (3) inefficent, model (2) is the right one since it allows causal statements on the effect of income and family size on consumption. Furthermore, the standard error in house is almost as large as the coefficient itself, which hints to a bias or noise in the model.

2. Omitted Variable Bias

a)

$\hat{\alpha}_1$ is defined as:

$$\hat{\alpha}_1 = \frac{cov(X_{1i}, Y_i)}{var(X_{1i})}$$

Plug in Y_i of the true model to find the conditional expectation:

$$E(\hat{\alpha}_1|X) = \frac{\overbrace{cov(X_{1i}, \beta_0)}^{=0} + \overbrace{cov(X_{1i}, \beta_1 X_{1i})}^{=\beta_1 var(X_{1i})} + \overbrace{cov(X_{1i}, \beta_2 X_{2i})}^{=\beta_2 cov(X_{1i}, x_{2i})} + \overbrace{cov(X_{2i}, \epsilon_i)}^{=0, \text{ under A2}}}{var(X_{1i})}$$

$$= \beta_1 + \beta_2 \frac{cov(X_{1i}, X_{2i})}{var(X_{1i})} = \beta_1 + \beta_2 \hat{\beta}_{X_{2i} on X_{1i}} \quad \text{(a)}$$

Where $\hat{\beta}_{X_{2i} on X_{1i}}$ is the OLS coefficient from the regression of X_{2i} on X_{1i}.

b)

From slide 30 in slidedeck 2a we can find two conditions for no omitted variable bias:

1) The variable omitted, $\beta_2 = 0$ or
2) X_{1i} is uncorrelated with X_{2i}

For 1): the true data generation process includes β_2, so by definition it is not equal to zero (otherwise it would not be part of the true data generating process.)

For 2): field of study is likely to be correlated with the GPA, in other words X_{1i} will most likely correlate with X_{2i}. From a statistical point of view, the probability that any two covariates are orthogonal is zero. We would assume that studying econ or finance is much harder because you have to take empirical methods so you will have a lower GPA.

By using $Y_i = \alpha_0 + \alpha_1 X_{1i} + \epsilon_i$, with $fieldofstudy$ a relevant variable is omitted, therefore $\hat{\alpha}_1$ is likely to be biased.

c)

$$\hat{\alpha}_1 = \beta_1 + \underbrace{\beta_2}_{\text{impact of} \beta_2 \text{on} Y} \underbrace{\frac{cov(X_{1i}, X_{2i})}{var(X_{1i})}}_{Cor(GPA, EcoFin)}$$

$$E(\hat{\alpha}_1|X) = \beta_1 + \beta_2 \hat{\beta}_{X_{2i} on X_{1i}} \qquad = \beta_1 + \overbrace{\beta_2}^{\text{impact of} \beta_2 \text{on} Y} \overbrace{(X_1' M_{-1} X_1)^{-} 1 X_1' M_{-1} X_2}^{Cor(GPA, EcoFin)}$$

To identify the sign of the bias, we need to make assumptions on the two parts influencing the bias. *Impact of β_2 on Y* is assumed to be positive as the valuable skills you learned in Empirical Methods (which you had to take as part of *EcoFin*) pay off in the job market. *Correlation between GPA and EcoFin* is assumed to be negative because of the reasons we stated in b). As we multiply a positive number with a negative, the bias will be negative.

d)

With β_3, the OVS bias changes to:

$$E(\hat{\alpha}_1|X) = \beta_1 + \beta_2 (X_1' M_{-1} X_1)^{-1} X_1' M_{-1} X_2$$

where $(X_1' M_{-1} X_1)^{-1} X_1' M_{-1} X_2$ would be the coefficient of an OLS regression of ϵ_{x_2} on ϵ_{x_1} after controlling for x_3.

Therefore, as opposed to questions 2c and 2d, there is now a conditional correlation between x_1 and x_2 as x_1 is controlled for x_3. The interpretation of the unconditional correlation in 1c would still hold, however we

cannot be sure if the conditional correlation can be interpreted as the uncoditional one. Therefore the sign of the bias cannot be confidently be identified because there is less intuition for conditional correlation.

If we had to sign the conditional correlation, we would assume that the interpretation of the conditional correlation is similar to the one from the unconditional one. Under this assumption, we would conclude that the conditional covariance is negative and therefore, the addition of x_3 would not change our answer (i.e. the bias is still negative).

The bias will be less than before as with the addition of x_3, fewer variance in Y is left for x_1.

3. Measurement Error in y

a)

First of all, we derive the relationship between ϵ_i, ϵ_i^* and η_i. Thus, we have:

$$\epsilon_i = y_i - \alpha_i - \beta x_i'$$

Plugging in the value of y_i in the equation above we have:

$$\begin{aligned}\epsilon_i &= y_i^* + \eta_i - \alpha_i - \beta x_i' \\ &= \underbrace{y_i^* - \alpha_i - \beta x_i'}_{\epsilon_i^*} + \eta_i = \epsilon_i^* + \eta_i\end{aligned}$$

So it is possible to calculate the mean and the variance of ϵ_i

Mean:

$$E[\epsilon_i] = E[\epsilon_i^* + \eta_i] = E[\epsilon_i^*] + E[\eta_i] = 0 + 0 = 0$$

Variance:

$$Var[\epsilon_i] = Var[\epsilon_i^* + \eta_i] = Var[\epsilon_i^*] + Var[\eta_i] + 2Cov(\epsilon_i^*, \eta_i) = \sigma_*^2 + \sigma_\eta^2$$

Note that the $Cov(\epsilon_i^*, \eta_i)$ is 0 because of the conditional expectaion of ϵ_i^* and η_i is also 0.

b)

We know that an estimator is unbiased only if the expected value of $\hat{\beta}$ is β itself. So, we calculate $E[\hat{\beta}]$:

$$\begin{aligned}E[\widehat{\beta}] &= E[\beta + (X'X)^{-1}X'\epsilon_i] \\ &= \beta + (X'X)^{-1}X'E[\epsilon_i] \\ &= \beta + (X'X)^{-1}X'E[\epsilon_i^* + \eta_i] \\ &= \beta + (X'X)^{-1}X' \underbrace{E[\epsilon_i]}_{=0, \text{ under A2}} + (X'X)^{-1}X' \underbrace{E[\eta_i]}_{=0, \text{ given}} = \beta\end{aligned}$$

Therefore, we can conclude that the estimator is not biased.

Empirical Part

Read data for empirical part:

```r
indicators <- fread("indicators.csv")
head(indicators)  # Check the format
dim(indicators)  #Check length
summary(indicators)  #Explore the data
```

We created one table for all of the regressions needed in the following questions:

```r
Model1 = mortalityun ~ corruptionun
Model2 = hospital_deaths ~ corruptionun
Model3 = mortalityun ~ ruleoflaw
Model4 = govmort ~ corruptionun

lm1 <- lm(Model1, data = indicators, x = TRUE)
lm2 <- lm(Model2, data = indicators, x = TRUE)
lm3 <- lm(Model3, data = indicators, x = TRUE)
lm4 <- lm(Model4, data = indicators, x = TRUE)
```

Table 1: Regression Results Empirical Exercise

	Dependent variable:			
	Mortality UN	Hospital Deaths	Mortality UN	Mortality Gov
	(1)	(2)	(3)	(4)
Corruption UN	0.626***	0.528***		0.358***
	(0.083)	(0.091)		(0.100)
Rule of Law			0.361***	
			(0.099)	
Constant	0.00000	0.00000	0.000	0.00000
	(0.083)	(0.090)	(0.099)	(0.099)
Observations	90	90	90	90
R^2	0.392	0.279	0.131	0.128
Adjusted R^2	0.385	0.271	0.121	0.118
Residual Std. Error (df = 88)	0.784	0.854	0.938	0.939
F Statistic (df = 1; 88)	56.685***	34.057***	13.215***	12.902***
Note:				*p<0.1; **p<0.05; ***p<0.01

a)

Child Mortality: Every country might have a different definition of child mortality, e.g. until which age is a human defined as a child. Therefore, the UN reported measure might suffer from less bias since it proposes a unified measurement. If child mortality is self-reported, countries might have a need to present them better as they actually are, so there might be a downwards bias. Even the UN index cannot be guaranteed to be measured exactly the same way in each country, so it is likely that *child mortality* includes some measurement error.

Corruption: the variable for corruption, *corruptionun* is a measured by the UN and therefore uses a standardized framework. However, corruption might appear in different forms and depending on the type of measurement countries may have different scores although the effective corruption is the same. Also, even with the assumption that corruption appears in the same way in every country (and can thus be measured with a standardized index), the measurement must also be enforced the same way in every country, assuming we had neutral observators in every country. A standardized index by the UN is probably the best way to measure corruption, nevertheless it is not free from a possible measurement error.

b)

i)

The OLS estimate ($\hat{\beta}$) of the relationship between corruption and child mortality is 0.626 (see Table 1, column 1):

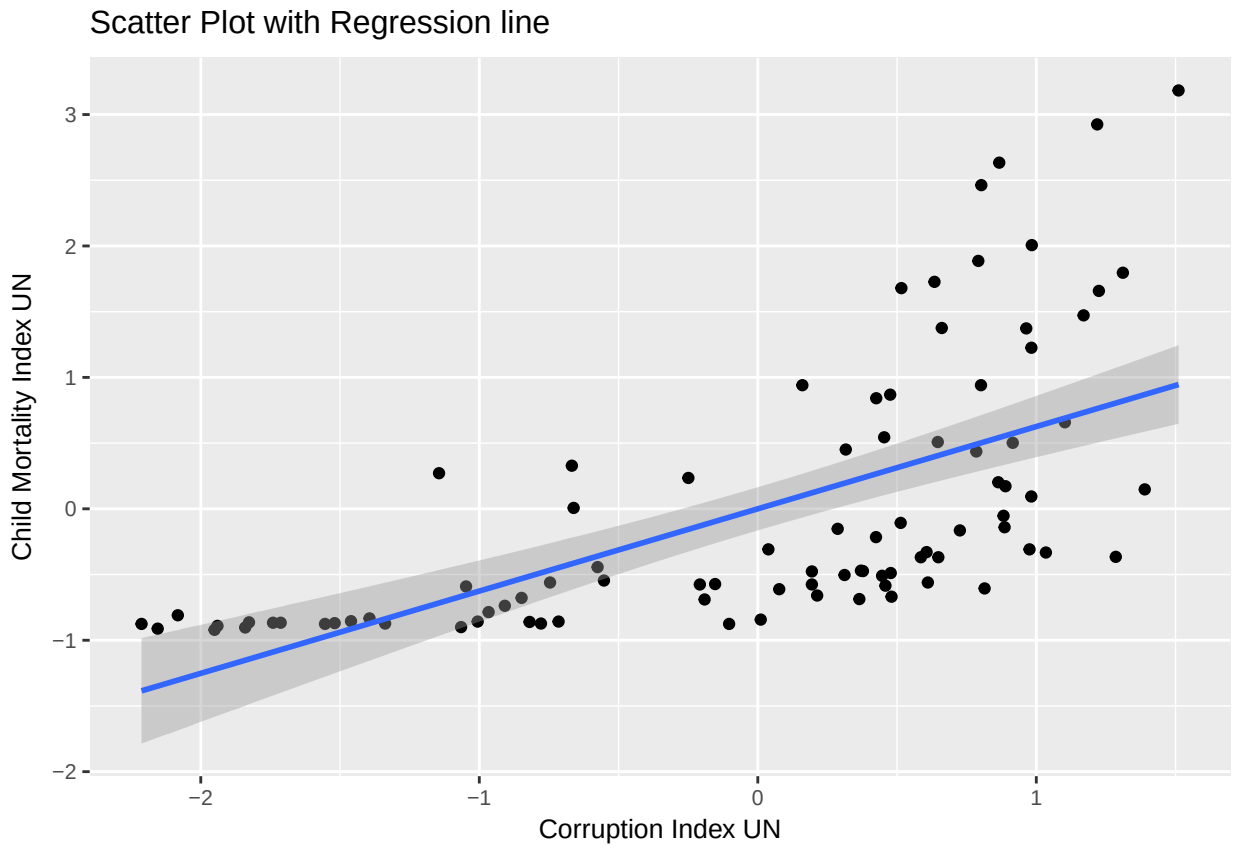
```
lm1$coefficients
##  (Intercept) corruptionun
## 8.073839e-08 6.259262e-01
```

To test whether $\hat{\beta} > 0$, we can perform a one side test with $H_0 : \hat{\beta} \leq 0$ and $H_1 : \hat{\beta} > 0$. Given the t-value of 7.5289673 and 89 degrees of freedom, we can look up the probability to find a t-value higher than 7.5289673 in a percentage t-disribution table. In the table, we find that this probability is equal to 0 (i.e. the p-value is 0). Hence the likelihood that we fail to reject the null hypothesis is 0% and we can reject that $\hat{\beta} < 0$.

ii)

A one-unit increase in *corruptionun* increases *mortalityun* by $\sim$ 0.63. Given that both variables are standardized with mean zero and standard deviation one, this also equals an elasticity of 63%, a large effect.

iii)



c)

i)

The conditions for classical measurement error are $E(\eta_i|x_i^*) = 0$ and $E(\eta_i|\epsilon_i^*) = 0$. It seems reasonable to assume that the later is fulfilled, a correlation between the measurement error η_i and the true error term ϵ_i^* seems unlikely, at least there is no obvious reason for it. There might be the possibility that $E(\eta_i|x_i^*) \neq 0$ as the η_i could increase (decrease) with x_i (and hence $Cov(x_i, \epsilon_i + \eta_i) \neq 0$), however as we are willing to assume that the measurement error is random, we can also state that this condition is fulfilled.

Based on this, the setting is similar from the one in question 3 of the paper and pencil part. From regressing hospital death on corruption index UN, we expect the OLS estimator to be unbiased but inefficient (i.e. the variance will be larger due to the possible measurement error) since $V(x) = \sigma_{x^*}^2 + \sigma_\eta^2$ (by slide 85 of topic 2a).

ii)

We compare column (1) to column (2) in Table 1 and see that the variance has gone up from 0.083^2 to 0.091^2, so our expectations were fully confirmed. Explanation given above.

iii)

The standard error of the estimate on *mortalityun* (0.083) is lower than the one for *hospitaldeaths* (0.091). Larger standard errors result in larger confidence intervals, which is also confirmed by the figure. This aligns with our expectations. Also, the red scatters (variable with measurement error) are more spread out than the blue ones (true variable), which again confirms our expectation that variance in *hospitaldeaths* will be higher than in *mortalityun*.

d)

The coefficient *rule of law* is 0.361 (see column 3) and significantly smaller if compared to the coefficient of *corruption* which is 0.626 (see column 1). This result, however, is not a surprise. In the case of a classical measurement error in the independent variable, the estimate is biased towards zero which results in an attenuation bias of the indepentend variable. In this specific circumstances, it is the case because we are assuming that any error between the *UN corruption index* and *UN index for Rule of Law* is random.

e)

i)

The setting is similar to question 3 in the paper and pencil part because there is as well a measurement error in the dependent variable. Although, as we assume that the classical measurement error conditions do not hold, it is not possible to consider it equivalent. In this case, we assume that the measurement error correlates to the independent variable corruption UN. Countries with a higher corruption index are more likely to report falsified figures regarding child mortality.

ii)

The sign of the bias will probably be negative since the covariance between corruption and the measurement error in govmort is negative. A plausible story for this results could be that the countries with higher corruption index have incentives to falsify their statistics to show better performances given a lower child mortality rate than in reality.

$$E[\widehat{\beta}] = \beta + (X'X)^{-1}X' \underbrace{E[\epsilon_i]}_{=0} + (X'X)^{-1}X' \underbrace{E[\eta_i]}_{<0} = (X'X)^{-1}X'E[\eta_i] < \beta$$

iii)

The estimate of corruption UN is 0.358 (see column 4) which is again lower if compared to the question 1.b. As suspected, the coefficient is downwards biased due to the negative correlation between corruption and measurement error.

f)

The worst case possible is when you mistakenly assume that an estimator is unbiased whereby it is biased. In this example, the worst case is (c) because we assumed that the classical measurement error conditions are holding. This assumption is rather "strong" since there is the possibility that our measurement error correlates with the true error term. Especially in poorer countries, children have less access to hospitals than in richer one. Thus, the important variable wealth causes a bias, because we don't control for it. As a consequence, we wrongly assume that the estimator is unbiased when it is biased.