Problem Set 3

1 Pencil and Paper Questions

1. Omitted Variable Bias

(a)

True model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \equiv \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \epsilon_i$
CLRM assumptions hold, especially $E(\epsilon_i | X_{1i}, X_{2i}) = 0$
Our estimation: $Y_i = \alpha_0 + \alpha_1 X_{1i} + \epsilon_i$

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

Calculate $E(\widehat{\alpha}_1)$

$$\mathbf{E}(\widehat{\boldsymbol{\alpha}}_{\mathbf{1}}) = \frac{cov(X_{1i}, \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i)}{var(X_{1i})} = \frac{cov(X_{1i}, \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \epsilon_i)}{var(X_{1i})}$$
$$= \alpha_1 + \alpha_2 \frac{cov(X_{1i}, X_{2i})}{var(X_{1i})} = \boldsymbol{\beta_1} + \boldsymbol{\beta_2} \frac{\boldsymbol{cov(X_{1i}, X_{2i})}}{\boldsymbol{var(X_{1i})}}$$

(b)

Is $\hat{\alpha}_1$ likely to be unbiased?

So, when we look at the formula above, $E(\widehat{\alpha}_1) = \alpha_1 + \alpha_2 \frac{cov(X_{1i}, X_{2i})}{var(X_{1i})}$, where $\alpha_2 = \beta_2$, we see that there will be no bias if either $\alpha_2 = 0$ or $cov(X_{1i}, X_{2i}) = 0$. It is not likely that α_2 equals zero as it is likely that the subject of graduation does impact Y. So, we pray that $cov(X_{1i}, X_{2i}) = 0$. This seems to be more realistic, as it is not clear for us why your chosen field of study should impact your GPA. It is possible for both finance and economics students to get higher or lower marks, dependent on their success in exams. Therefore, these two variables can be seen as independent, which results in the conclusion that $cov(X_{1i}, X_{2i}) = 0$. So it seems like our prayers have been fulfilled and there is no omitted variable bias in our estimated regression function.

(c)

Sign of the bias:

	$cov(X_{1i}, X_{2i})$		
$cov(X_{2i}, Y_i)$		+	-
	+	Upward Bias	Downward Bias
	-	Downward Bias	Upward Bias

In b) we came to the conclusion that there exists no bias. However, to be able to answer this question we assume that there exists a bias. It is conceivable that $cov(X_{2i}, Y_i) > 0$ because students with knowledge in economics or finance are more useful for hiring firms. As a consequence, they will pay higher wages to these students. In our opinion X_{1i} (Grade Point Average) and X_{2i} (economics or finance-dummy) have no covariance as we see the grade distribution in the respective field of study as independent. So, a good grade in the master studies is reachable for a history or literature student as well as for a economics or finance student (as described in b)). However, it is imaginable that in history or literature studies the GPA's are higher than in economics or finance as in these subjects the liaison between students and professors may be closer and therefore the grading better for these students. So, we could argue that $cov(X_{1i}, X_{2i}) < 0$. So, with this assumption in mind we could mock that the bias is downward, but we don't feel confident about this assumption. Given our other justification holds (no bias), then the sign of this bias is irrelevant.

(d)

Included another variable:
Truth: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i \equiv \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \epsilon_i$
You estimate: $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_3 X_{3i} + \epsilon_i$

Mathematically shown:

$$\mathrm{E}(\widehat{\alpha}_1) = \mathrm{E}((\mathrm{X}'_{1\mathrm{i}} M_{[-1]} X_{1i})^{-1} \mathrm{X}'_{1\mathrm{i}} M_{[-1]} y)$$

Plug in the true model:

$$\begin{aligned}
\mathrm{E}(\widehat{\alpha}_1) = {} & E((\mathrm{X}'_{1\mathrm{i}} M_{[-1i]} X_{1i})^{-1} \mathrm{X}'_{1\mathrm{i}} M_{[-1]} (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i)) \\
& = 0 + \beta_1 E((\mathrm{X}'_{1\mathrm{i}} M_{[-1i]} X_{1i})^{-1} \mathrm{X}'_{1\mathrm{i}} M_{[-1]} X_{1i}) \\
& + \beta_2 E((\mathrm{X}'_{1\mathrm{i}} M_{[-1i]} X_{1i})^{-1} \mathrm{X}'_{1\mathrm{i}} M_{[-1]} X_{2i} + \beta_3 E((\mathrm{X}'_{1\mathrm{i}} M_{[-1i]} X_{1i})^{-1} \mathrm{X}'_{1\mathrm{i}} M_{[-1]} X_{3i}) \\
& + E((\mathrm{X}'_{1\mathrm{i}} M_{[-1i]} X_{1i})^{-1} \mathrm{X}'_{1\mathrm{i}} M_{[-1]} \epsilon_i \\
& = \beta_1 E((\mathrm{X}'_{1\mathrm{i}} M_{[-1i]} X_{1i})^{-1} \mathrm{X}'_{1\mathrm{i}} M_{[-1]} X_{1i}) + \beta_2 E((\mathrm{X}'_{1\mathrm{i}} M_{[-1i]} X_{1i})^{-1} \mathrm{X}'_{1\mathrm{i}} M_{[-1]} X_{2i} \\
& + \beta_3 E((\mathrm{X}'_{1\mathrm{i}} M_{[-1i]} X_{1i})^{-1} \mathrm{X}'_{1\mathrm{i}} M_{[-1]} X_{3i}) + 0 \\
& = \beta_1 + \beta_2 E((\mathrm{X}'_{1\mathrm{i}} M_{[-1i]} X_{1i})^{-1} \mathrm{X}'_{1\mathrm{i}} M_{[-1]} X_{2i} \\
& + \beta_3 E((\mathrm{X}'_{1\mathrm{i}} M_{[-1i]} X_{1i})^{-1} \mathrm{X}'_{1\mathrm{i}} M_{[-1]} X_{3i}) \\
& = \beta_1 + \beta_2 \hat{\beta}_{X_{2i}\ on\ X_{1i}\ after\ controlling\ for\ the\ intercept, X_{2i}\ and\ X_{3i}} \\
& + \beta_3 \hat{\beta}_{X_{3i}\ on\ X_{1i}\ after\ controlling\ for\ the\ intercept, X_{2i}\ and\ X_{3i}}
\end{aligned}$$

Difficult to say in this case, as the bias would be driven by both the conditional (on X_{3i}) correlation between X_{2i} and X_{1i}. And further the bias would be driven by both the conditional (on X_{2i}) correlation between X_{3i} and X_{1i}. So a statement is rather impossible and wouldn't leed to a sophisticated answer.

2. Measurement Error in y

(a)

True model: $y_i^* = \alpha + \beta x_i + \epsilon_i^*$
Estimated model: $y_i = \alpha + \beta x_i + \epsilon_i$, where $y_i = y_i^* + \eta$
$\rightarrow y_i^* + \eta = \alpha + \beta x_i + \epsilon_i^* + \eta$
The measurement error is $\epsilon_i = \epsilon_i^* + \eta_i$.

Mean

The mean of ϵ_i is:

$$E(\epsilon_i|x_i) = E(\epsilon_i^* + \eta_i) = E(\epsilon_i^*) + E(\eta_i)$$

From the (conditional-)mean-zero-error assumption, which states that $E(\epsilon_i^*|x_i) = 0$, it follows that $E(\epsilon_i^*) = E_{x_i}E(\epsilon_i^*|x_i) = E_{x_i} \cdot 0 = 0$. And from the assumption that $\eta_i \sim (0, \sigma_\eta^2)$, it follows that $E(\eta_i) = 0$.
Therefore:

$$E(\epsilon_i|x_i) = E(\epsilon_i^* + \eta_i) = E(\epsilon_i^*) + E(\eta_i) = 0 + 0 = 0$$

In summary, the mean of the error ϵ_i is zero.

Variance

The variance of ϵ_i is:

$$\begin{aligned} V(\epsilon_i) &= V(\epsilon_i^* + \eta_i) = V(\epsilon_i^*) + V(\eta_i) + 2Cov(\epsilon_i^*, \eta_i) \\ &= E\left[\left(\epsilon_i^* - \mu_{\epsilon_i^*}\right)^2\right] + E\left[\left(\eta_i - \mu_{\eta_i}\right)^2\right] + 2Cov(\epsilon_i^*, \eta_i) \end{aligned}$$

As shown above, the mean of ϵ_i^*, $\mu_{\epsilon_i^*}$, and the mean of η_i, μ_{η_i}, are zero. And from the assumption that $E(\eta_i|\epsilon_i^*) = 0$, it follows that $Cov(\epsilon_i^*, \eta_i) = 0$.
Therefore:

$$\begin{aligned} V(\epsilon_i) &= E[(\epsilon_i^* - 0)^2] + E[(\eta_i - 0)^2] + 2 \cdot 0 \\ &= E[(\epsilon_i^*)^2] + E[(\eta_i)^2] \\ &= \sigma_{\epsilon_i^*}^2 + \sigma_{\eta_i}^2 \end{aligned}$$

In summary, the variance of ϵ_i is $V(\epsilon_i) = \sigma_{\epsilon_i^*}^2 + \sigma_{\eta_i}^2$.

(b)

The estimator of β, $\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon$, is unbiased if $E(\hat{\beta}) = \beta$.

$$\begin{aligned} E(\hat{\beta}) &= E(\beta + (X'X)^{-1}X'\epsilon) = E(\beta) + E((X'X)^{-1}X'\epsilon) \\ &= \beta + E((X'X)^{-1}X'\epsilon) \end{aligned}$$

As shown in a), $E(\epsilon|X) = (\epsilon_i^*) + E(\eta_i) = 0$.
Therefore:

$$E(\hat{\beta}) = \beta + E((X'X)^{-1}X'0) = \beta + 0 = \beta$$

In summary, $E(\hat{\beta}) = \beta$ and therefore, the estimator $\hat{\beta}$ is unbiased.

(c)

The variance of $\hat{\beta}$ without measurement error is:

$$V(\hat{\beta}) = V(\epsilon^*|X)(X'X)^{-1} = \sigma_*^2(X'X)^{-1}$$

where σ_*^2 refers to the variance of the error term of the true model, ϵ^*, thus is equal to $\sigma_{\epsilon_i^*}^2$. The variance of the error term of the model with measurement is $\sigma^2 = V(\epsilon) = \sigma_{\epsilon_i^*}^2 + \sigma_{\eta_i}^2 = \sigma_*^2 + \sigma_{\eta_i}^2$

Therefore:

$$V(\hat{\beta}) = V(\epsilon|X)(X'X)^{-1} = (\sigma_*^2 + \sigma_{\eta_i}^2)(X'X)^{-1}$$

The variance of the estimator $\hat{\beta}$, $V(\hat{\beta})$, is larger for the model with the measurement error in y than for the model without measurement error.

(d)

A measurement in the dependent variable y does not bias the estimator $\hat{\beta}$, but it increases the variance of the estimator, $V(\hat{\beta})$. Therefore, a measurement error in the dependent variable is not a big problem, whereas a measurement in an independent variable biases the estimator and therefore is a big deal.

2 Empirical Application

1. Dealing with Measurement Error

We will analyse 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.

Statistic	N	Mean	St. Dev.	Min	Max
corruptionun	90	-0.00000	1.000	-2.213	1.511
mortalityun	90	0.000	1.000	-0.921	3.183
ruleoflaw	90	-0.000	1.000	-3.042	1.944
govmort	90	-0.000	1.000	-1.794	3.270
hospital_deaths	90	-0.000	1.000	-1.789	4.035

The analysed countries are:

```
Albania                  Argentina            Australia     Austria         Burundi
Belgium                  Benin                Burkina Faso  Bangladesh      Bulgaria
Bolivia                  Brazil               Barbados      Bhutan          Botswana
Central African Republic Canada               Switzerland   Chile           China
Cameroon                 Colombia             Comoros       Costa Rica      Germany
Denmark                  Dominican Republic   Algeria       Ecuador         Spain
Finland                  Fiji                 Gabon         United Kingdom  Ghana
Guinea-Bissau            Greece               Guatemala     Guyana          Honduras
Indonesia                India                Ireland       Iceland         Israel
Italy                    Jamaica              Jordan        Japan           Kenya
Sri Lanka                Lesotho              Morocco       Madagascar      Mexico
Mali                     Mongolia             Mozambique    Mauritius       Malawi
Niger                    Netherlands          Norway        Nepal           New Zealand
Panama                   Peru                 Philippines   Portugal        Paraguay
Rwanda                   Saudi Arabia         Senegal       Sierra Leone    El Salvador
Sweden                   Swaziland            Seychelles    Chad            Togo
Thailand                 Trinidad and Tobago  Tunisia       Turkey          Uganda
Uruguay                  United States        Vietnam       South Africa    Zambia
```

1. (a)

The variable **corruption** is very likely to be subject to measurement error, due to the fact that corruption is illegal and therefore often not reported. So there is a high number of unreported cases. So the numbers from UN are probably too low. Further the more corrupt a county is the more difficult it gets for the UN to estimate an appropriate number.

Suppose the true model is:

$$mortalityun_i = \alpha + \beta corruption_i + \epsilon_i^*$$

But there is a measurement error:

$$corruptionun_i = corruption_i + \eta_i$$

where

$$\eta_i \sim (0, \sigma_\eta^2)$$

So:

$$\begin{aligned} mortalityun_i &= \alpha + \beta corruption_i + \beta\eta_i + (\epsilon_{it}^* - \beta\eta_i) \\ &= \alpha + \beta corruptionun_i + \epsilon_{it} \end{aligned}$$

where

$$\epsilon_{it} = (\epsilon_{it}^* - \beta\eta_i)$$

We now have a problem however:

$$Cov(corruptionun_i, \epsilon_i) = Cov(corruption_i + \eta_i, \epsilon_{it}^* - \beta\eta_i) = -\beta\sigma_\eta^2$$

and:

$$Var(corruptionun_i) = Var(corruption_i + \eta_i) = \sigma_{corruption}^2 + \sigma_\eta^2$$

So:

$$\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon = \beta + \frac{\frac{1}{N}X'\epsilon}{\frac{1}{N}X'X} \xrightarrow{p} \beta\frac{\sigma_{corruption}^2}{\sigma_{corruption}^2 + \sigma_\eta^2} \xrightarrow{p} \lambda\beta$$

If we assume that the true $\beta > 0$, then the $\hat{\beta}$ has a negative bias, what is known as attenuation bias.

The unrecorded **mortality** cases may be higher in developing countries due to the fact the access to healthcare sector is limited and therefore these children do not occur in any statistics. If we assume that the expected error term = 0, we are perfectly fine, the beta estimator is unbiased.

(b) i.

If we estimate a model $mortalityun_i \sim corruptionun_i$ we get the following output:

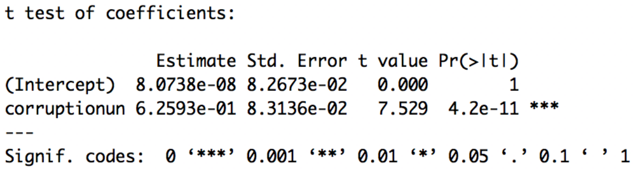
	Dependent variable:
	mortalityun
corruptionun	0.626^{***}
	(0.083)
Constant	0.00000
	(0.083)
Observations	90
R^2	0.392
Adjusted R^2	0.385
Residual Std. Error	0.784 (df = 88)
F Statistic	56.685^{***} (df = 1; 88)
Note:	*p<0.1; **p<0.05; ***p<0.01

The estimator for corruptionun_i is 0.626. This value is with a P-Value lower than 1% highly significant. The adjusted R^2 is relatively high with a value of 0.392.

In a next step we calculate a **one-sided hypothesis test**:

$$H_0: \beta_{corruptionun} \leq 0 \; H_1: \beta_{corruptionun} > 0$$

Input data (under the assumption of homoscedasticity):



```
t test of coefficients:

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.0738e-08 8.2673e-02   0.000        1
corruptionun 6.2593e-01 8.3136e-02   7.529  4.2e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

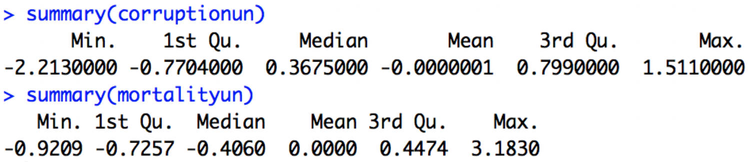
$$\hat{\beta}_{corruptionun} = 0.62593$$
$$Std.Error(\hat{\beta}_{corruptionun}) = 0.083136$$
$$d.o.f. = (N - K - 1) = (90 - 1 - 1) = 88$$
$$a = 5\%\ (one\ sided) \rightarrow Critical\ Value: 1.662354$$

Calculation of t-Value:

$$t_{Value} = \frac{\hat{\beta}_{corruptionun} - 0}{Std.Error(\hat{\beta}_{corruptionun})} = \frac{0.62593}{0.083136} = 7.528989$$

So we already can reject the H_0 hypothesis.
The resulting P-Value is 4.2e-11 and therefore nearly 0.

(b) ii.

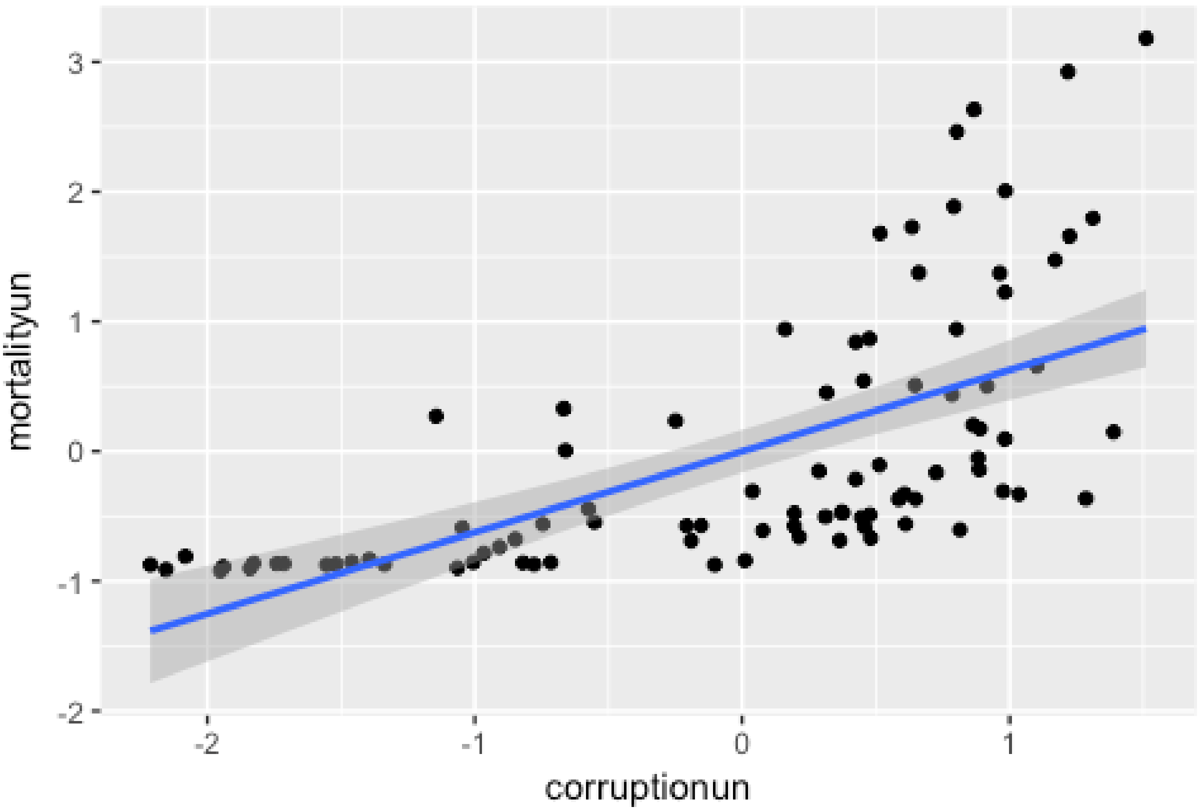


```
> summary(corruptionun)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max. 
-2.2130000 -0.7704000  0.3675000 -0.0000001  0.7990000  1.5110000 
> summary(mortalityun)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-0.9209 -0.7257 -0.4060  0.0000  0.4474  3.1830 
```

The effect says that if we increase the $corruptionun$ by 1 unit we expect an increase in $mortalityun$ by 0.626 unit. The corruption index by UN observers is a standardized index with a scale between -2.213 and 1.511 and the index for child mortality rate under age 5 also reported by the UN is also a standardized index with a scale between -0.9209 and 3.1830. If we have these numbers as a proxy the impact is relatively high.

(b) iii.

In the following scatter plot you can see the data points as well as the sample regression line (blue) and its confidence intervals (grey area: $a = 5\%$).



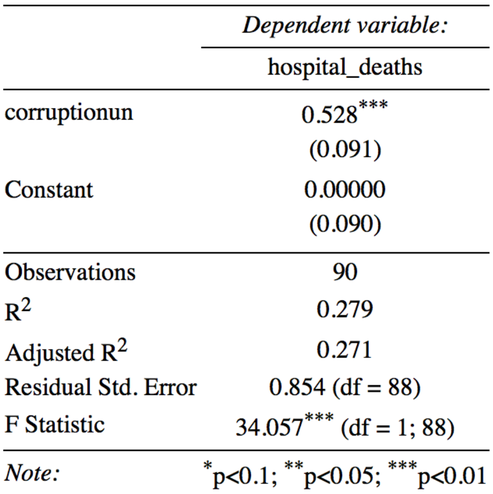
(c) i.

If we assume that the error term are random as given in the task and as we measure the dependent variable we conclude as we have proven above that the beta is unbiased and therefore we are fine. In this case we expect the beta to be the same and an increase in variance. $\hat{\beta} \rightarrow$
BUT:

Compared to the section Pencil and Paper it is in our opinion not the same because we measure different data. With the hospital death we only focus on well developed countries with good health infrastructure and therefore we underestimate the true death rate. As we measure something else we get a complete different result, which is difficult to compare with the original model. Under this assumption we expect the beta to be lower because in our opinion the excess ability shows a certain level of development which could mean the impact of corruption is not as big as to people with limited access to healthcare. $\hat{\beta} \downarrow$

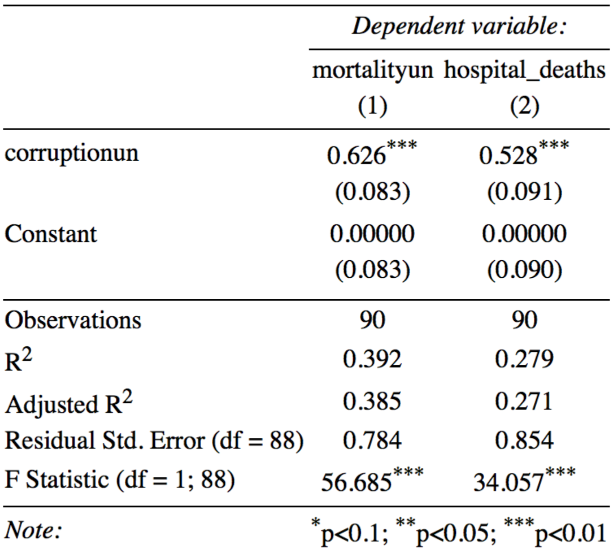
(c) ii.

If we do not have the official mortality data for child mortality rate under age 5 ($mortalityun$) and we estimate a model instead with the variable of mortality generated by number of deaths in hospitals ($hospital\_deaths$), we get the output below:



	Dependent variable:
	hospital_deaths
corruptionun	0.528***
	(0.091)
Constant	0.00000
	(0.090)
Observations	90
R2	0.279
Adjusted R2	0.271
Residual Std. Error	0.854 (df = 88)
F Statistic	34.057*** (df = 1; 88)
Note:	*p<0.1; **p<0.05; ***p<0.01

To compare with the “true” relationship between the two variables, the output of both estimations are presented below:



	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
R2	0.392	0.279
Adjusted R2	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	

The impact of the corruptionun on $hospital\_deaths$ (0.528) is smaller compared to the impact of the corruptionun on $mortalityun$ (0.626). As we can see in the table the given assumption that the errors are randomly distributed is wrong, because the beta differs. But it seems like our second proposition may hold as the $\hat{\beta}_{hospital\_death}$ is lower than $\hat{\beta}_{mortalityun}$.

Analytical explanation:

$$E\left(\epsilon_i \middle| X_{corruptionun_i}\right) = E(\epsilon_i^* + \eta_i) = E(\epsilon_i^*) + E(\eta_i)$$

The variance of $E(\epsilon_i^*) = 0$ due to assumption 2 but in this case $E(\eta_i) < 0$

$$E(\epsilon_i|x_i) = E(\epsilon_i^* + \eta_i) = E(\epsilon_i^*) + E(\eta_i) \rightarrow E(\eta_i) < 0$$

Now we built the expectations:

$$E(\hat{\beta}) = E(\beta + (X'X)^{-1}X'\epsilon) = E(\beta) + E((X'X)^{-1}X'\epsilon)$$
$$= \beta + E((X'X)^{-1}X'(\epsilon^* + \eta))$$

Therefore:

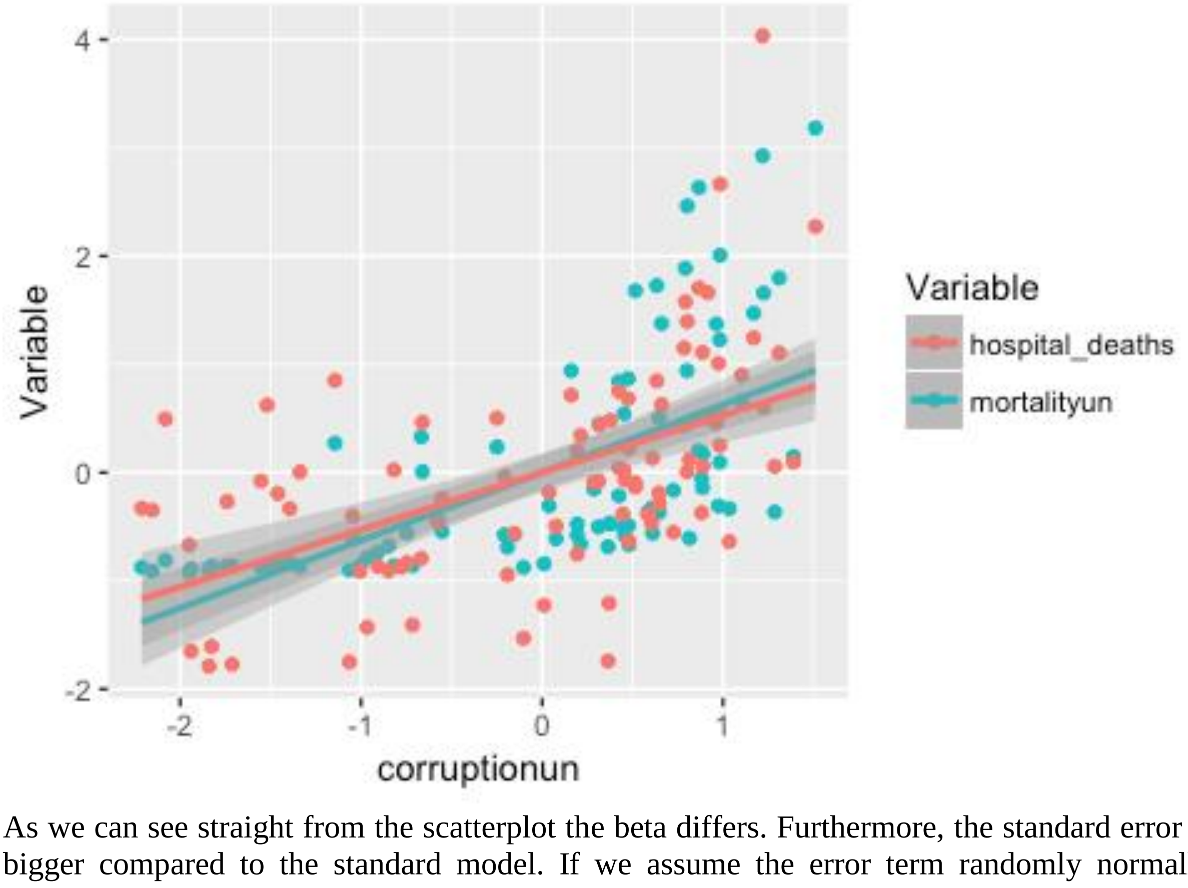
$$E(\hat{\beta}) = \beta + E((X'X)^{-1}X'\eta) = \beta + E(\eta_i)$$

So:

$$E(\widehat{\beta_{1c}}) < \beta_{1b}$$

(c) iii.

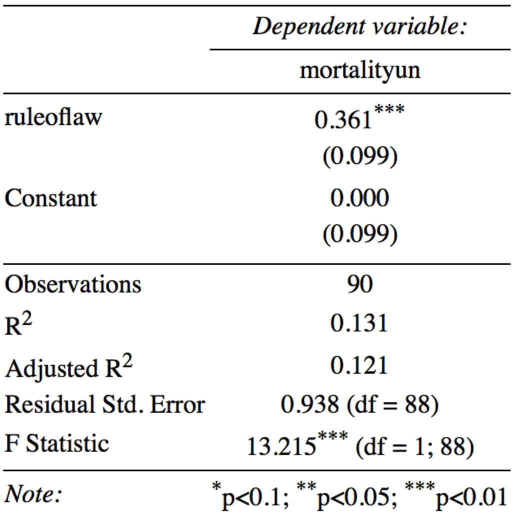
In the following scatter plot you can see the data points of the model (1) in green and of the model (2) in red. Furthermore, both sample regression lines ((1) green, (2) red) and its confidence intervals (((1) green, (2) red): $a = 5\%$).



As we can see straight from the scatterplot the beta differs. Furthermore, the standard error is bigger compared to the standard model. If we assume the error term randomly normally distributed, then this outcome holds with our expectation due to the fact that the estimator is unbiased but has a bigger variance.

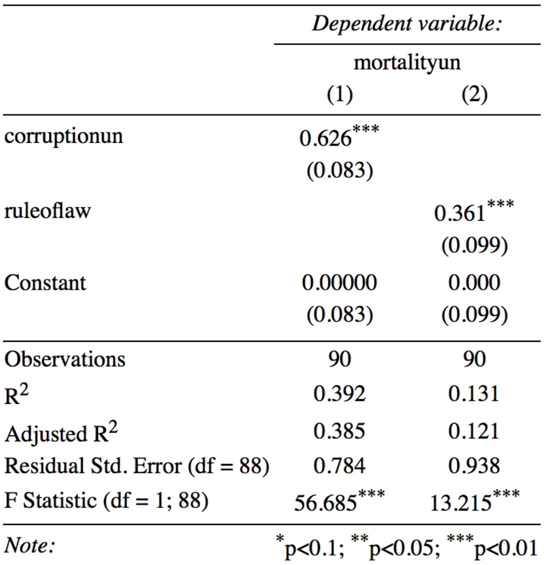
(d)

If we do have the official mortality data for child mortality rate under age 5 ($mortalityun$) but this time not the corruption index by UN observers ($\text{corruption}un$). Instead we use a proxy for corruption based of the degree to which laws and regulations are actually enforceable in the country (ruleoflaw), we get the output below:



	Dependent variable:
	mortalityun
ruleoflaw	0.361^{***}
	(0.099)
Constant	0.000
	(0.099)
Observations	90
R^2	0.131
Adjusted R^2	0.121
Residual Std. Error	0.938 (df = 88)
F Statistic	13.215^{***} (df = 1; 88)
Note:	*p<0.1; **p<0.05; ***p<0.01

To compare with the "true" relationship between the two variables, the outputs of both estimations are presented below:



	Dependent variable:	
	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	

The impact of the corruptionun on $mortalityun$ (0.626) is bigger compared to the impact of the $ruleoflaw$ on $mortalityun$ (0.361). The beta of the second model seems to be biased due to measurement error arising from taking an alternative X with random distributed error term between $mortalityun$ and $ruleoflaw$.

Suppose the true model is:

$$mortalityun_i = \alpha + \beta corruptionun_i + \epsilon_i^*$$

But there is a measurement error:

$$ruleoflaw_i = corruptionun_i + \eta_i$$

where

$$\eta_i \sim (0, \sigma_\eta^2)$$

So:

$$\begin{aligned} mortalityun_i &= \alpha + \beta corruptionun_i + \beta\eta_i + (\epsilon_{it}^* - \beta\eta_i) \\ &= \alpha + \beta ruleoflaw_i + \epsilon_{it} \end{aligned}$$

where

$$\epsilon_{it} = (\epsilon^*_{it} - \beta\eta_i)$$

move on:

$$Cov(ruleoflaw_i, \epsilon_i) = Cov(corruptionun_i + \eta_i, \epsilon^*_{it} - \beta\eta_i) = -\beta\sigma^2_\eta$$

and:

$$Var(ruleoflaw_i) = Var(corruptionun_i + \eta_i) = \sigma^2_{corruptionun} + \sigma^2_\eta$$

So:

$$\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon = \beta + \frac{\frac{1}{N}X'\epsilon}{\frac{1}{N}X'X} \xrightarrow{p} \beta\frac{\sigma^2_{corruptionun}}{\sigma^2_{corruptionun} + \sigma^2_\eta} \xrightarrow{p} \lambda\beta$$

If we assume that the true $\beta > 0$, then the $\hat{\beta}$ has a negative bias, what is known as attenuation bias. So the σ^2_η in this case has to be positive and $\neq 0$. That is why the estimator is biased towards 0.

(e) i.

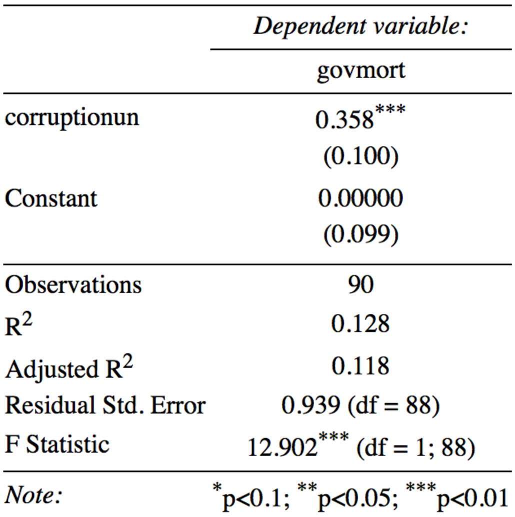
The biggest problem with this estimation is the fact that the governments can report the number by themselves. Therefore, the data suffer under a systematical measurement error. This results in correlation between the error term and the X's. So, the more developed and richer countries have no incentive to fake the data whereas poor and/or corrupt country can do. So the measurement error might not be random, what is very critical/bad. To conclude here we don't have a classical measurement error as seen in the lecture.

(e) ii.

High corruption states probably reduce their child mortality rate to make a better impression. So the bias could be negative.

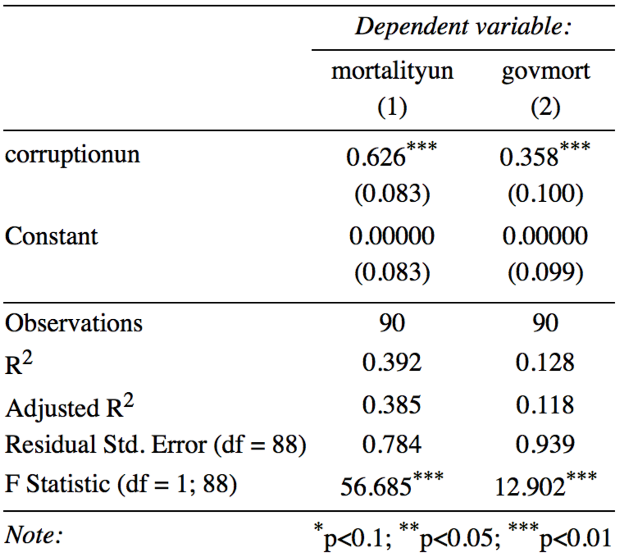
(e) iii.

If we do not have the official mortality data for child mortality rate under age 5 ($mortalityun$) and we use a mortality rate self-reported by each country ($govmort$), we get the output below:



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

To compare with the "true" relationship between the two variables, the outputs of both estimations are presented below:

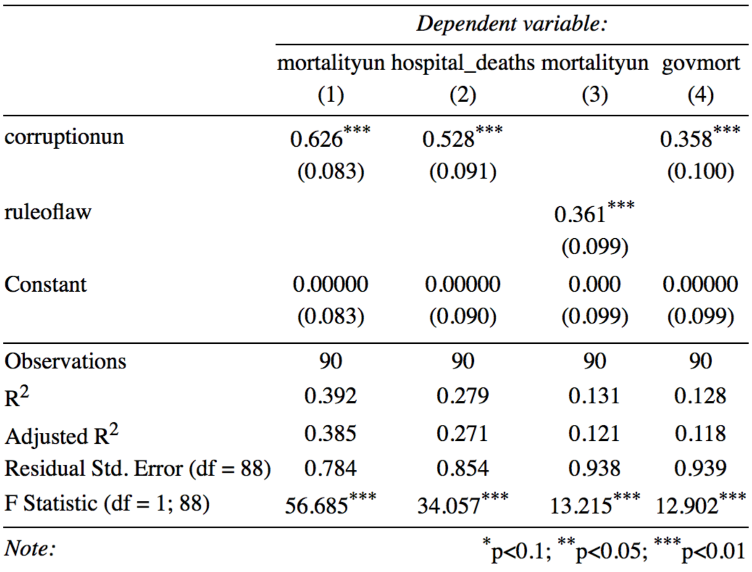


	Dependent variable:	
	mortalityun	govmort
	(1)	(2)
corruptionun	0.626^{***}	0.358^{***}
	(0.083)	(0.100)
Constant	0.00000	0.00000
	(0.083)	(0.099)
Observations	90	90
R^2	0.392	0.128
Adjusted R^2	0.385	0.118
Residual Std. Error (df = 88)	0.784	0.939
F Statistic (df = 1; 88)	56.685^{***}	12.902^{***}
Note: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

The impact of the corruptionu$n$ on $mortalityun$ (0.626) is much bigger compared to the impact of the corruptionu$n$ on $govmort$ (0.358). First of all, the second model is not comparable with the original model as explaned above. In our opinion it is lower because the reported govmort rate is only a subset of the true model (mortalityun) as in corrupt countries not every death is part of official statistics.

(f)

To compare with the "true" relationship between the two variables, the outputs of all estimations are presented below:



	Dependent variable:			
	mortalityun	hospital_deaths	mortalityun	govmort
	(1)	(2)	(3)	(4)
corruptionun	0.626^{***}	0.528^{***}		0.358^{***}
	(0.083)	(0.091)		(0.100)
ruleoflaw			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

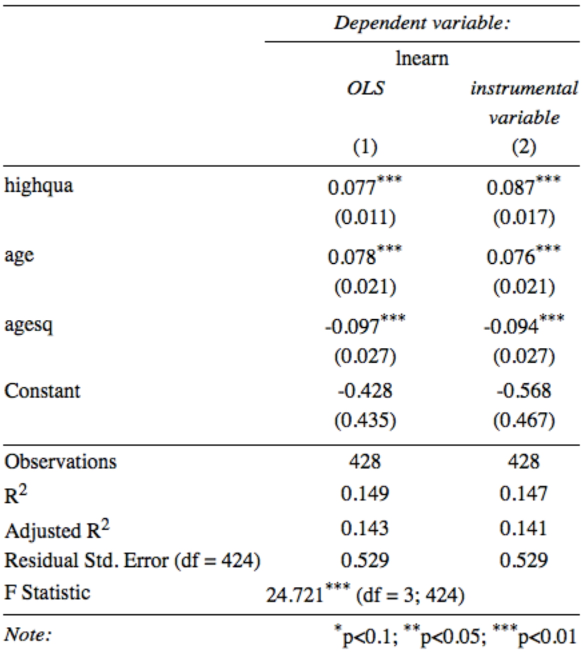
If we define model (1) as the true model, we can rank the models as follows:

1. Model (2): There we have a change in Y but which should not affect the beta and the data generating process is clear and well defined.
2. Model (3): We change the explanatory variable and as we have proven above the estimator beta is biased toward zero.
3. Model (4): Here we change again the Y but this time the governments reports the number instead of collecting the data by ourselves. Here we have systematical deviation between the two Ys. So that we estimate a different model which is and should not be compared with the other models.

2. IV Regression

a) [1]

If we regress these relations, we get the following table:



	Dependent variable:	
	lnearn	
	OLS	*instrumental variable*
	(1)	(2)
highqua	0.077***	0.087***
	(0.011)	(0.017)
age	0.078***	0.076***
	(0.021)	(0.021)
agesq	-0.097***	-0.094***
	(0.027)	(0.027)
Constant	-0.428	-0.568
	(0.435)	(0.467)
Observations	428	428
R^2	0.149	0.147
Adjusted R^2	0.143	0.141
Residual Std. Error (df = 424)	0.529	0.529
F Statistic	24.721*** (df = 3; 424)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

a) i.

For the OLS regression, we get exactly the same results for all coefficient as the ones reported in BCHHS. For the IV regression, our results are slightly different. More precisely, the coefficient for *highqua* is slightly higher (0.087, compared to 0.085 in BCHHS) with a slightly higher standard error (0.017, compared to 0.012 in BCHHS). The coefficients for *age* and *agesq* are slightly lower (0.076, compared to 0.077 in BCHHS, and -0.094 compared to -0.095 in BCCHS, respectively), but have the same standard errors. In summary, both the coefficients and the standard errors are all exactly the same or very similar. Therefore, neither the magnitude of the effects nor their significance is influenced. Thus, the discrepancy is not "serious".

a) ii.

The authors run the regression with a constant term, which is important, but they do not report the coefficient on the constant term. This coefficient contains the information (predicted) log earnings for *highqha* = 0. As it can be assumed that *highqua* is only zero for children and children do not earn any money, this information does not seem to be important. However, the coefficient is also needed to predict log earnings for given values of *highqua*, *age* and *agesq*. Thus, the coefficient on the constant is important if one intends to use the regression results for prediction.

a) iii.

If the assumptions for the IV estimation hold, i.e. if the instrument is relevant and exogenous, the coefficient of education can be interpreted as the causal effect on earnings. More precisely, the coefficient of education of 0.087 suggests that each year of additional education increases earnings by 8.7%.

[1] To get comparable results we also divided agesqu by 100 as done in the Paper.

b) i.

There are mainly three sources of dangers for the assumption 2 $(E(\varepsilon_i|x_i) = 0$: (1) correlated unobservables, (2) classical measurement error and (3) simultaneous causality.

(1) Correlated unobservables refers to factors which are correlated with both the dependent variable, in our case earnings, and a dependent variable, in our case education or age, but not observed. Such a factor is for example someone's ability. Ability could be positively correlated with education, i.e. more able people get higher education (for example because it might be less costly for someone with high ability to get another year of school than for someone with lower ability). Further, ability could be positively correlated with earnings, i.e. more able people get have higher earnings (for example because they perform better in their job). Thus, *highqua* might be endogenous because of correlated unobservables, such as for example ability.

(2) Measurement error is a situation in which either the dependent variable or one of the independent variables are measured with. All three independent variables, *highqua*, *age* and *agesq*, are easy to measure, thus we expect no measurement error due to difficulty in measurement. However, it is possible that people overstate their education. Therefore, *highqua* is might be endogenous due to measurement error.

(3) Simultaneous causality refers to a situation in which the dependent variable has an effect on one of the independent variables. In our case, this means that earnings have an effect on *highqua*, *age* or *agesq*. An effect of income on age is impossible, thus *age* and *agesq* are not endogenous due to simultaneous causality. An effect of income on education is possible if also education during adult life is considered. However, as we assume *highqua* to refer to education at young age, income is not likely to have an influence on that. Thus, simultaneous causality seems not to be a relevant threat.

b) ii.

(1) We assume the likely sign of the bias due to correlated unobservable (such as for example ability) to be positive. The factor ability is expected to have a positive effect on *highqua* and *lnearn*. Therefore, when ignoring it, we overestimate the effect of *highqua* on *lnearn* and have a positive bias.

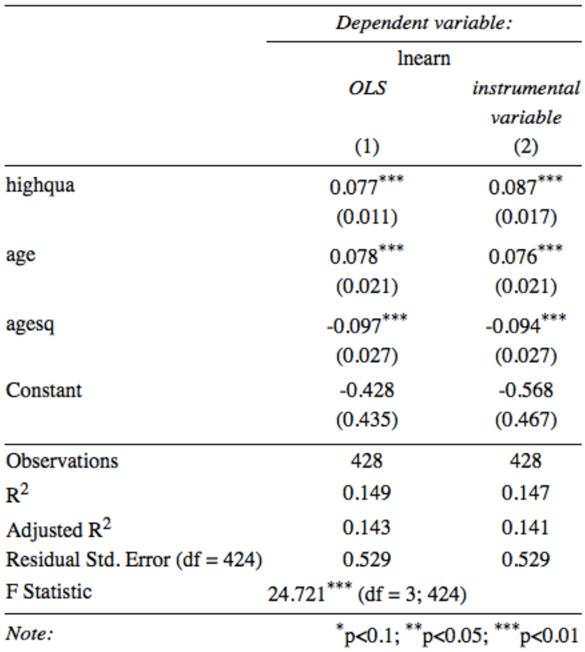
(2) We assume that people do rather over- than understate their education. Therefore, we assume the likely sign of the bias due to measurement error in *highqua* to be negative.

(3) We do not expect any bias due to simultaneous causality.

b) iii.

Source of Endogeneity	**Relevance**	**Exogeneity**
Correlated Unobservables (e.g. Ability)	The twin's report of someone's education is most likely correlated with the actual education.	Someone's earnings do most likely not influence their twin's education.
Measurement Error		

b) iv.

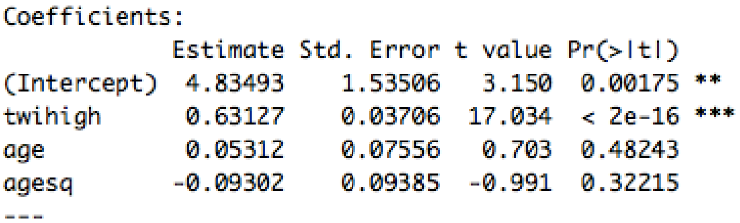


	Dependent variable:	
	lnearn	
	OLS	*instrumental variable*
	(1)	(2)
highqua	0.077***	0.087***
	(0.011)	(0.017)
age	0.078***	0.076***
	(0.021)	(0.021)
agesq	-0.097***	-0.094***
	(0.027)	(0.027)
Constant	-0.428	-0.568
	(0.435)	(0.467)
Observations	428	428
R^2	0.149	0.147
Adjusted R^2	0.143	0.141
Residual Std. Error (df = 424)	0.529	0.529
F Statistic	24.721*** (df = 3; 424)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

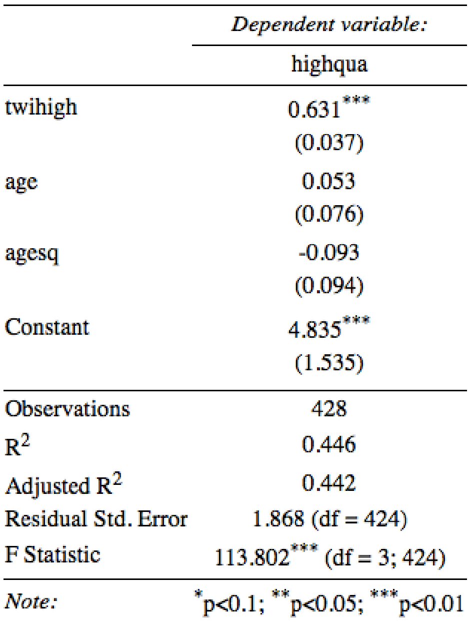
The OLS regression results suggest a coefficient of *highqua* of 0.077, while the IV regression results suggest a coefficient of 0.087. The IV coefficient is slightly higher, which suggests that the OLS regression suffered from a negative bias. This matches with our expectations of a bias due to measurement error, i.e. due to the fact that people are likely to overstate their education.

b) v.

First of all, we have a look on the estimated numbers:



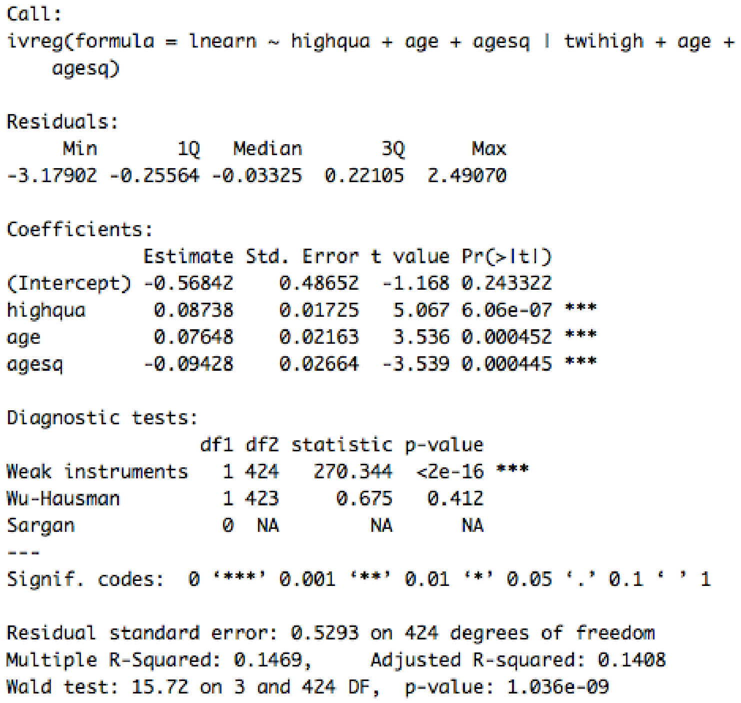
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.83493    1.53506   3.150  0.00175 ** 
twihigh      0.63127    0.03706  17.034  < 2e-16 ***
age          0.05312    0.07556   0.703  0.48243    
agesq       -0.09302    0.09385  -0.991  0.32215    
---
```



	Dependent variable:
	highqua
twihigh	0.631***
	(0.037)
age	0.053
	(0.076)
agesq	-0.093
	(0.094)
Constant	4.835***
	(1.535)
Observations	428
R^2	0.446
Adjusted R^2	0.442
Residual Std. Error	1.868 (df = 424)
F Statistic	113.802*** (df = 3; 424)
Note:	*p<0.1; **p<0.05; ***p<0.01

To test whether the instrument is weak, we use an F test. If we get an F value lower than 10, we say that the instrument is weak.

The F value of the variable *twihigh* is equal to the squared t value: $F = 17.034^2 \cong 290$. Thus, the F value is much higher than 10 and therefore, the instrument is not weak.



```
Call:
ivreg(formula = lnearn ~ highqua + age + agesq | twihigh + age +
    agesq)

Residuals:
     Min       1Q   Median       3Q      Max 
-3.17902 -0.25564 -0.03325  0.22105  2.49070 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.56842    0.48652  -1.168 0.243322    
highqua      0.08738    0.01725   5.067 6.06e-07 ***
age          0.07648    0.02163   3.536 0.000452 ***
agesq       -0.09428    0.02664  -3.539 0.000445 ***

Diagnostic tests:
                 df1 df2 statistic p-value    
Weak instruments   1 424   270.344  <2e-16 ***
Wu-Hausman         1 423     0.675   0.412    
Sargan             0  NA        NA      NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5293 on 424 degrees of freedom
Multiple R-Squared: 0.1469,	Adjusted R-squared: 0.1408 
Wald test: 15.72 on 3 and 424 DF,  p-value: 1.036e-09
```

b) vi.

The first stage regression suggests that *twihigh* has a statistically significant effect of 0.631 on *highqua*. Further, the F statistic supports the significance of *twihigh* as an instrument for *highqua*. Additionally, direction of bias which is suggested by the IV regression is intuitively reasonable. Therefore, we believe these results, based on our assumption of validity of the instrument used.

R Code

```
setwd("~/Dropbox/Empirical Methods/Problemset 3/Working Progress/Marco")
library(lmtest)
library(zoo)
library(stargazer)
library(dplyr)
library(sandwich)
library(ggplot2)

###Empirical Application
### 1. Dealing with Measurement Error
## import data
data <- read.csv("indicators.csv", header = TRUE, sep = ",")
head(data)
summary(data)
stargazer(data, type = "text", out = "summary.htm")
data$country
attach(data)

## (a)
## texttexttext

## (b)
model1 <- lm(mortalityun~corruptionun)
stargazer(model1, type = "text", out = "Regression_1b).htm")

## (b) i)
model1_fittet <- model1$fitted.values
model1_fittet
t.test(model1_fittet,y=NULL,alternative = c("greater"), mu=0)
# controll
model.str.elpct <- model1
model.coeftest <- coeftest(model.str.elpct, vcov=vcovHC(model.str.elpct,"HC1"))
model.coeftest

## (b) ii)
summary(corruptionun)
summary(mortalityun)
## texttexttext

## (b) iii)
plot(corruptionun,mortalityun,  ylab="index for child mortality rate under age 5 also reported
by the UN", xlab="index of corruption reported by UN observers", main="Scatter plot",
xlim=c(-2,2),ylim=c(-1,3))
abline(lm(mortalityun~corruptionun), col="red")
newx<-seq(-2,2)
prd<-predict(model1,newdata=data.frame(x=newx),interval = c("confidence"),level =
0.9,type="response")
abline(newx,prd[,"upr"],col="blue",lty=2)
abline(newx,prd[,"lwr"],col="blue",lty=2)
```

```
## doesn't work! next try:
corruptionunmortalityun <- subset(data, select=c(corruptionun, mortalityun) )
modeltest<- ggplot(corruptionunmortalityun, aes(x=corruptionun, y=
mortalityun))+geom_point()
print(modeltest)
g <- modeltest + geom_smooth( method= "lm")
print(g)

## (c) i)
## texttexttext

## (c) ii)
model2 <- lm(hospital_deaths~corruptionun)
stargazer(model2, type = "text", out = "Regression_1c).htm")
## compare
stargazer(model1, model2, type = "text", out = "Regression_1c)2.htm")
## texttexttext

## (c) iii)
plot(corruptionun,mortalityun,  ylab="mortality variables", xlab="index of corruption reported
by UN observers", main="Scatter plot", xlim=c(-2,2),ylim=c(-1,3), col="green")
points(corruptionun,hospital_deaths, col="red")
abline(lm(mortalityun~corruptionun), col="green")
abline(lm(hospital_deaths~corruptionun), col="red")
newx<-seq(-2,2)
prd<-predict(model1,newdata=data.frame(x=newx),interval = c("confidence"),level =
0.9,type="response")
abline(newx,prd[,"upr"],col="green",lty=2)
abline(newx,prd[,"lwr"],col="green",lty=2)
newx2<-seq(-2,2)
prd<-predict(model2,newdata=data.frame(x=newx2),interval = c("confidence"),level =
0.9,type="response")
abline(newx,prd[,"upr"],col="red",lty=2)
abline(newx,prd[,"lwr"],col="red",lty=2)
## doesn't work! next try:
testtest <- ggplot(data = data, aes(Corruption, Mortality)) +
  geom_point(aes(x = corruptionun, y = mortalityun), legend=  TRUE,  xlab="X", ylab="Y",
colour=alpha('red')) +
  geom_point(aes(x = corruptionun, y = hospital_deaths), legend = TRUE, colour=alpha('blue')
        + geom_smooth(method= 'lm',aes(x=corruptionun, y = mortalityun))
        + geom_smooth(method= 'lm',aes(x=corruptionun, y = hospital_deaths)))
testtest
g <- testtest + geom_smooth( method= "lm")
print(g)
#woopwoop
ggplot(data, aes(x=corruptionun, y = Variable, color = Variable)) +
geom_point(aes(x=corruptionun, y = mortalityun, col = "mortalityun")) +
geom_point(aes(x=corruptionun, y = hospital_deaths, col =
"hospital_deaths"))+geom_smooth(method= 'lm',aes(x=corruptionun, y = mortalityun, col =
"mortalityun"))+ geom_smooth(method= 'lm',aes(x=corruptionun, y = hospital_deaths, col =
"hospital_deaths"))
```

```
## (d)
model3 <- lm(mortalityun~ruleoflaw)
stargazer(model3, type = "text", out = "Regression_1d).htm")
## compare
stargazer(model1, model3, type = "text", out = "Regression_1c)2.htm")

## (e) i)
## texttexttext

## (e) ii)
## texttexttext

## (e) iii)
model4 <- lm(govmort~corruptionun)
stargazer(model3, type = "text", out = "Regression_1e).htm")
## compare
stargazer(model1, model4, type = "text", out = "Regression_1e)2.htm")

## (f)
stargazer(model1, model2, model3, model4, type = "text", out = "Regression_1f).htm")
## texttexttext

##IV

setwd("~/Dropbox/Empirical Methods/Problemset 3/Working Progress/Marco/IV")
data <- read.csv("IVdata.csv", sep = ",", header = TRUE)
head(data)
attach(data)
library(car)
library(lmtest)
library(zoo)
library(dplyr)
library(sandwich)
library(ivmodel)
library(AER)
library(stargazer)


data2 <- mutate(data, lnearn=log(earning), agesq=((age^2)/100))
head(data2)
attach(data2)

## a)
model1 <- lm(lnearn~highqua+age+agesq)
summary(model1)
#iv
model2 <- ivreg(lnearn~highqua+age+agesq |twihigh+age+agesq)
summary(model2)
summary(model2, vcov = sandwich, diagnostics = TRUE)
```

```
stargazer(model1, model2, type = "text", out = "Model12.htm")

model3 <- lm(highqua~twihigh+age+agesq)
summary(model3)
stargazer(model3, type = "text", out = "bv.htm")
jtest(model1, model2)
```