# Empirical Methods, Autumn Semester 17, University of Zurich

# Problem Set 3

Team AliDaJo*

12th December 2017

## 1 Pencil and Paper

### 1.1 Omitted Variable Bias

#### a)

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

Now, we plug in $Y_i$ of the true model and take the conditional expectation:

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

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

#### b)

$\hat{\alpha}_1$ is unbiased only if:

- $\beta_2 = 0$, but in this case there would be no omitted variable, *or*
- $X_{1i}$ is uncorrelated with $X_{2i}$

*Danagul Adilbayeva (10-611-416), Joel Hampton(11-747-318) and Alice Horner (12-211-769)

First, we know that the true data generating process includes the $\beta_2$ (field of study). Therefore, by definition, it is unequal zero, because otherwise it would not be part of the true data generating process.
Second, it is also very unlikely that the correlation between $X_{1i}$ and $X_{2i}$ is zero. In general, from a statistical perspective, the probability that any two covariates are orthogonal is zero. In our example, we assume that if you study econ or finance, it is harder to get a high GPA compared to if you study history or literature. Thus, *field of study* and *GPA* are correlated. So, we actually omit a relevant variable *field of study* in our sample regression. Thus, we conclude that $\hat{\alpha}_1$ is likely to be biased.

## c)

We have to assess the sign of the bias given the formula in (1.1/a). In order to define the direction of the bias, we have to make the following two assumptions:

1. If you studied econ or finance, we assume that your starting salary will be higher compared to students in history or literature (i.e. $Cov(field, salary) > 0$). The reason is that you are more likely to start a career in a bank, insurance, consulting firm and alike. These industries are usually paying a higher starting salary in order to attract the best students.
2. If you study econ or finance, it is harder to get a high GPA (i.e. $Cov(field, GPA) < 0$), because the content is more abstract and the exams are more difficult.

Bottom line: Omitting the variable *field of study* causes a negative bias for $\hat{\alpha}_1$.

## d)

The formula for omitted variable bias changes slightly from (1.1/a):

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

where $(X_1' M_{-1} X_1)^{-1} X_1' M_{-1} X_2$ is equivalent to the coefficient in a simple OLS regression when we regress the left-over variation in $X_2$ on the left-over variation $X_1$ after we controlled for $X_3$.

The difference between question (1c) and question (1d) is that we have now a case of a conditional correlation between $X_1$ and $X_2$ (i.e. we are controlling for $X_3$). The unconditional correlation from question (1c) would still hold. However, we don't know if the unconditional correlation has a similar interpretation as the conditional one. This being said we cannot determine the direction of the bias for sure, since it is hard to have an intuition for the 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).

## 1.2 Measurement Error in y

### a)

First of all, we derive the relationship between $\epsilon_i$, $\epsilon_i^*$ and $\eta_i$.
After rearranging our estimated model, we have:

$$\epsilon_i = y_i - \alpha - \beta x_i$$

We plug in the given relationship between $y_i$ and $y_i^*$:

$$\epsilon_i = y_i^* + \eta_i - \alpha - \beta x_i = \overbrace{y_i^* - \alpha - \beta x_i}^{=\epsilon_i^*} + \eta_i = \epsilon_i^* + \eta_i$$

Now, we are calculating the mean and variance of $\epsilon_i$:

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

$$Var[\epsilon_i] = Var[\epsilon_i^* + \eta_i] = Var[\epsilon_i^*] + Var[\eta_i] + 2 \overbrace{Cov(\epsilon_i^*, \eta_i)}^{=0,\, b/c\, E[\eta_i|\epsilon_i^*]=0} = \sigma_*^2 + \sigma_\eta^2$$

### b)

$$\begin{aligned} E(\hat{\beta}|X) &= E(\beta + (X'X)^{-1}X'\epsilon|X) = \beta + (X'X)^{-1}X'E(\epsilon|X) \\ &= \beta + (X'X)^{-1}X'E(\epsilon^* + \eta|X) \\ &= \beta + (X'X)^{-1}X' \overbrace{E(\epsilon^*|X)}^{=0,\text{ under A2}} + (X'X)^{-1}X' \overbrace{E(\eta|X)}^{=0,\text{ because given}} = \beta \end{aligned}$$

$\hat{\beta}$ is unbiased in this case.

### c)

$$\begin{aligned} V(\hat{\beta}|X) &= V(\beta + (X'X)^{-1}X'\epsilon|X) \\ &= E\left[(\beta + (X'X)^{-1}X'\epsilon - \beta|X)(\beta + (X'X)^{-1}X'\epsilon - \beta|X)'\right] \\ &= E((X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}|X) \\ &= (X'X)^{-1}X'X(X'X)^{-1}E(\epsilon\epsilon'|X) \\ &= (X'X)^{-1}E((\epsilon^* + \eta)(\epsilon^* + \eta)'|X) \\ &= (X'X)^{-1}V(\epsilon^* + \eta|X) \\ &= (X'X)^{-1}(\sigma_*^2 + \sigma_\eta^2) \end{aligned}$$

When there is a measurement error, the variance is increased by $\sigma_\eta^2$. This is the case because adding an additional variation to the dependent variable increases the variance of the estimate.

#### d)

In the presence of the measurement error in the dependent variable, it is only a big problem when it creates bias, because then the estimated coefficient is wrong in expectation. But this is only the case, if the classical measurement error conditions do not hold. Given our answer in (1.2/b), we do agree that the measurement error is not such a big deal, because we have an unbiased estimator (i.e. the classical measurement error conditions hold). However, as explained in (1.2/c), the variance of our estimate is increased by $\sigma_\eta^2$ and thus, it is inefficient. As a consequence, we are less likely to reject the null hypothesis of $\beta = 0$ due to larger standard errors of our estimate (i.e. we might have a less significant estimate and it might be further away from the "true" value).

## 2 Empirical Application

### 2.1 Dealing with Measurement Error

#### 2.1/a)

- **Corruption:** We believe that it is impossible to measure corruption with complete accuracy. We base our argument on the following three reasons: First, due to its illegal nature, econometricians cannot observe corruption easily. Second, corruption is hard to define precisely, because it can manifest itself in many different forms and sometimes it is not so clear, if a certain behaviour can be described as corruptive. Third, many measures of corruption are based on surveys from individuals, which might be subjective to a certain degree. So, we believe that any measure of corruption is likely to include measurement errors.
- **Child mortality:** Compared to corruption, child mortality is less subject to measurement error. If we considered developed countries (e.g. USA, European countries, Japan, Singapore) then we believe that the measurement error in child mortality is extremely small due to the fact that women give birth in the hospitals and children are then supervised by doctors which are subject to reporting child mortality. If we consider developing countries (e.g. India, China, Indonesia, Ethiopia) where there are women (esp. in rural areas) who are not registered (the state does not know about their existence) and their children die, it will not appear in any statistics. There are also women who give birth at home (not at the hospital) and if their children die this will not be reported either. In sum, we think that measurement error in child mortality is likely to happen.

#### 2.1/b)

**i)** The OLS estimate ($\hat{\beta}$) of the relationship between corruption and child mortality is 0.626 (as depicted in table 1 column 1).

We have to test for the null hypothesis of $\hat{\beta} < 0$ (according to a post from Matteo in the forum on the 1st December). This means we have to find out what the p-value is of finding a

Table 1: Regression table for the whole question 1 of computer exercise part

| | *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 |

t-value higher than the value in our regression output. Our t-value is 7.529. According to a percentage t-distribution table, we will find a t-value higher than 7.529 with a probability of 0 (i.e. our p-value is equal 0). The likelihood that we fail to reject the null hypothesis is 0%. Therefore, we reject the null-hypothesis of $\hat{\beta} < 0$.

Note: We would have found the same result if we had taken the given p-value in the regression output, since we are testing for the null hypothesis of $\hat{\beta} < 0$.

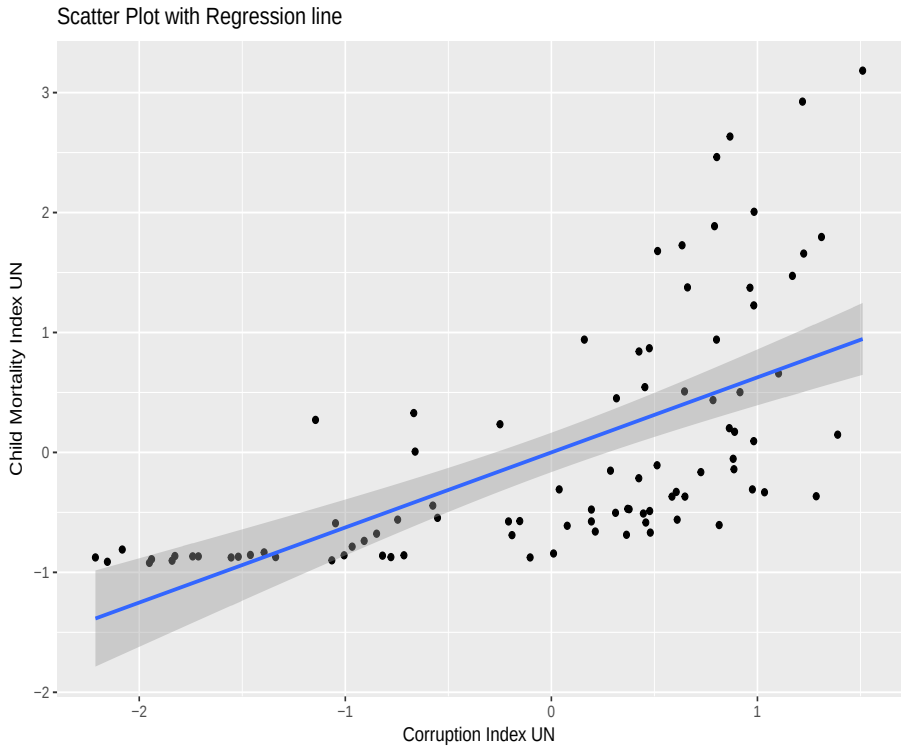
**ii)** Assuming that the CLRM assumptions are satisfied, an increase of corruption by one full index point increases on average the mortality index by 0.626.
In order to estimate the magnitude: A one standard deviation increase in the corruption index increases the mortality index by 0.626 standard deviations. This is a large effect. In a country with mean values in both indices, an one-standard increase in corruption raises child mortality above the 3rd quartile.

**iii)** Please see figure 1 on the next page.

## 2.1/c)

**i)** Under the assumption that the recorded documentation in the hospitals are correct and not influenced by the level of corruption, we come to the conclusion that the setting is the same as in Question 2 of the Pencil-and-Paper section. This means the classical measurement error conditions hold and the measurement error in y (i.e. $\eta$) is uncorrelated to x and the residual. 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).

Figure 1: Scatter plot with one OLS regression line



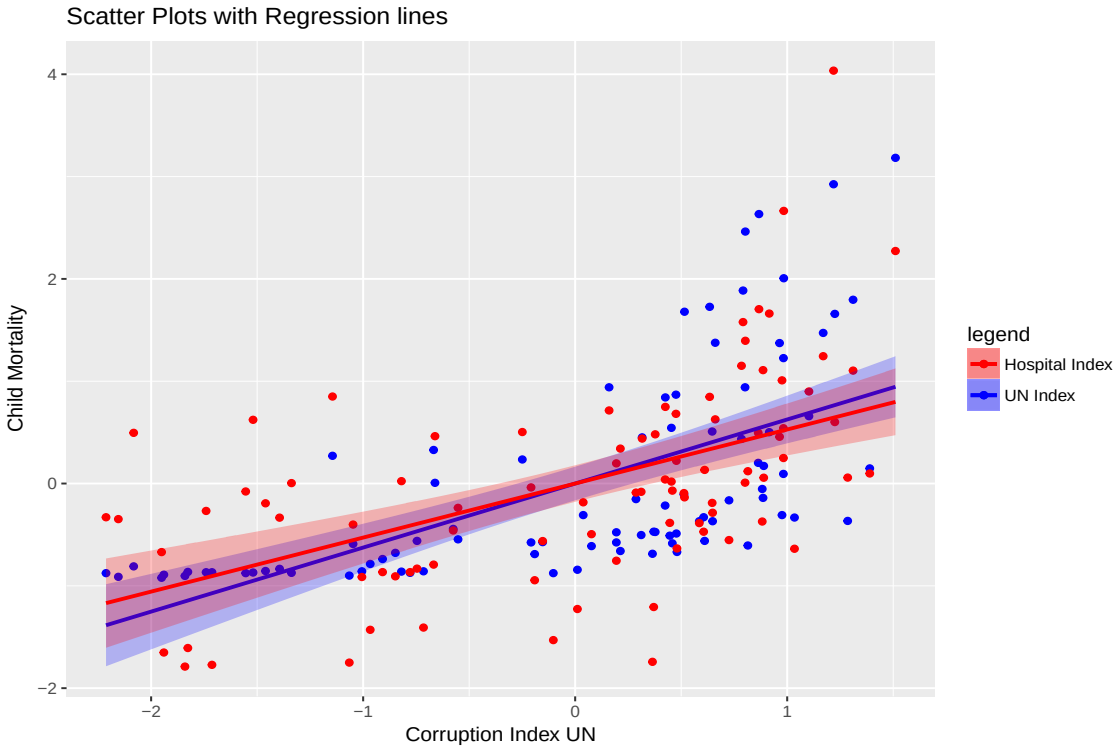

**ii)** The estimate on corruption is 0.528 which is about 0.1 lower than in (2.1/b)(see table 1 column 1).
It is consistent with our expectations. First, the coefficient is still unbiased because it is still concentrated around the true value (true value is 0.626 from 2.1/b - see table 1 column 1). Second, the variance has gone up from 0.006889 ($=0.083^2$) to 0.008281 ($=0.091^2$), which allows for further deviation from the true value.

**iii)** The standard error of the estimate on *mortalityun* (0.083) is lower than the one for *hospital deaths* (0.091). Thus, the confidence interval will also be smaller for *mortalityun* compared to *hospital deaths*.
Yes, this is consistent with our expectation, because when we add a classical measurement error into the regression, the variance of the estimate becomes larger (also the standard error).

## 2.1/d)

Our coefficient on *rule of law* is 0.361 (see table 1 column 3). Compared to the estimate of *corruption UN* in (2.1/b), the estimate of rule of law is lower by 0.265. This finding is consistent with our expectation. In the case of a classical measurement error in the independent variable, the estimate is biased towards zero (i.e. attenuation bias). We assume the classical measurement error is fulfilled, because any error between *UN corruption index* and *UN Rule*

Figure 2: Scatter plot with both OLS regression lines




*of Law index* is random. So, this explains why our coefficient on *rule of law* is smaller than the coefficient on *corruption UN* (in (2.1/b) - see table 1 column 1).

## 2.1/e)

**i)** The setting is similar to the question 2 in the paper and pencil part, because there is also a measurement error in the dependent variable. But, as we assume that the classical measurement error conditions do not hold, it is definitely not the same. We expect that the measurement error is correlated to the independent variable *corruption UN*. Countries with a higher index in corruption are more likely to report faked child mortality numbers.

**ii)** We assume that the sign of the bias will be negative, because the covariance between *corruption* and the measurement error in *govmort* is negative. We think that the reason for this is that the countries with higher corruption index might misreport their statistics in order to have a lower child mortality rate (i.e. better performance). From (1.2/b) we know:

$$E(\hat{\beta}|X) = \beta + (X'X)^{-1}X' \overbrace{E(\epsilon^*|X)}^{=0, \text{ under A2}} + (X'X)^{-1}X' \overbrace{E(\eta|X)}^{<0} = \beta + (X'X)^{-1}X'E(\eta|X) < \beta$$

**iii)** The estimate of corruption UN is 0.358 (see table 1 column 4) and it is lower by 0.268 compared to the question (2.1/b) (see table 1 column 1). As expected, we see that the coefficient is downward biased due to the negative correlation between corruption and measurement error.

### 2.1/f)

In order to rank the cases we care more about the bias than the efficiency because in all three cases the estimators are inefficient due to the measurement error. Within bias we label a case as being dangerous when we mistakenly assume that the estimator is unbiased whereby it is actually biased.
Based on this logic, we rank the cases in the following order (from most dangerous to less dangerous): (c) > (e) > (d). Our argumentation goes as follows:
The only case, in which we could mistakenly interpret an estimator as unbiased, but in fact it is biased, is when we have a measurement error in Y. Particularly, this is the case if we wrongly assumed that the conditions of classical measurement error hold. Based on these argumentation, (d) should be the least dangerous case, because it will be biased regardless of whether the conditions of classical measurement error hold or not.
In case (e), we assumed that more corruptive governments misreported their numbers of child mortality. Thus, as econometricians we are aware that there might be a bias. But since we are unsure if the governments truly fake their numbers, there is a slight chance that there is no bias.
However, in case (c) we assumed that the classical measurement error conditions hold. We know that this is a "strong" assumption from our side, since it might be the case that the measurement error is correlated with the true error term. Particularly, in poorer countries, children under 5 might have less access to hospitals than in richer countries. In this situation, the relevant variable *wealth* causes a bias, because we don't control for it. As a consequence, we would mistakenly assume that the estimator is unbiased, but in fact it is biased. Thus, the case (c) is more severe than the case (e).

## 2.2 IV Regression

### 2.2/a)

**i)** Yes, there is a discrepancy in the IV coefficient of *highqua*. In the paper the reported value is 0.085 and in our regression the value of the coefficient is 0.087. The coefficients of *age* and *age squared* are almost equivalent to those in the report, i.e. the discrepancies are negligible. The reported coefficient of *age* is 0.077 and our coefficient is 0.076. The reported coefficient of *age* squared is $-0.095$ and our coefficient is $-0.094$. At this point we want to stress that for comparability reasons, we divided the covariate of *age squared* by 100 in our regression because BCHHS did the same in their report.
In order to judge if the discrepancy in the *highqua* coefficient is "serious" we look at the impact on *log hourly wage*. The difference accounts for 0.2%. For a 10£ hourly wage, this would make a difference of 2 penny. Therefore, we interpret this discrepancy as "non-serious".

**ii)** We think that there is no important information in the intercept. In general, the intercept can be interpreted as a value of the dependent variable for the group for which all the independent variables are equal zero. In our case, the intercept would be the log hourly wage for a person with no years of education and with an age of zero (i.e. a baby born today).

**iii)** Since we assume that our IV results are consistent and IV1 as well as IV2 hold, we can interpret the coefficient as follows: holding all other independent variables constant, an additional year of education is associated with a 8.7% higher hourly wage.

## 2.2/b)

**i)** We could think of the following two reasons why *highqua* can be endogenous.

- First, the error term might contain "ability" which is correlated with years of schooling and with hourly wage.
- Second, an individual might misreport deliberately its own years of schooling in order to be conform with social expectations (i.e. individuals with years of education below the average might have a tendency to report a higher number in order to avoid to be perceived as "stupid" or "uneducated"). This is a measurement error which does not fulfil the conditions of a classical measurement error, since it is correlated with the number of years of schooling (i.e. negative correlation).

**ii)** The formula to determine the sign on the bias when omitting a relevant variable is:

$$E[\hat{\beta_{highqua}}] = \beta_{highqua} + \beta_{abil}\hat{\beta}_{abil\_on\_highqua}$$

Thus, we expect the overall sign of the bias to be positive. The reasons are: first, the cov(ability, hourly wage) is positive, because more able individuals are more productive and thus, are receiving a higher hourly wage. Second, the cov(years of schooling, ability) is also positive, because more capable individuals are more likely to acquire a higher level of education due to their ability. This leads to more years of schooling.

The formula to determine the sign of the bias for a measurement error is:

$$\begin{aligned}\hat{\beta_{highqua}} &= \beta_{highqua} + (X'_{highqua}X_{highqua})^{-1}X'_{highqua}\epsilon \\ &= \beta_{highqua} + \frac{\frac{1}{N}X'_{highqua}\epsilon}{\frac{1}{N}X'_{highqua}X_{highqua}} \\ &\xrightarrow{\text{p}} \lambda\beta_{highqua}\end{aligned}$$

Based on the formula above, we assume that our estimate will be biased towards zero even more than in the classical measurement error situation, because we have a non-zero covariance between years of schooling and measurement error (note: in a classical measurement error case the bias only consisted of the variance of the measurement times the coefficient).

**iii)** For the first reason ("omitted variable"):

- **relevance:** We believe that *twihigh* and *highqua* are highly correlated, since your twin should know how many years of education you had. We would even assume that it is close to 1.
- **exogeneity:** We assume that *twihigh* is as much correlated with *ability* as *highqua* is. Therefore, with *twihigh* we cannot pull out the non-problematic variation in *highqua*. Consequently, IV1 property is not fulfilled.

For the second reason, ("measurement error"):

- **relevance:** We believe that *twihigh* and *highqua* are still highly correlated for the same reason.
- **exogeneity:** Since your twin should not have an incentive to deliberately misreport (i.e. overstate the amount of years), we believe that*twihigh* is able to pull out the clean variation in *highqua*. However, if your sibling is randomly misreporting your years of schooling, there would still be correlation between the instrument (*twihigh*) and the error term due to the classical measurement error. We believe that a classical measurement error is likely in our case.

**iv)**

Table 2: Regression table for the IV-Problem

| | *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 |

We assume that the chosen instrumental variable tries to deal with the measurement error and thus, the IV estimate went into the right direction, because we had an attenuation bias in the OLS estimate (i.e. OLS estimate is lower than IV estimate). This is true because *twihigh* is a more precise measure for years of schooling (i.e. there is no measurement error from social expectation in it).

**v)** According to our regression output in table 3, we find a t-value for *twihigh* of 17.034 ($= \frac{0.63127}{0.03706}$). So, our t-value is higher than our rule of thumb (3.16) - our instrument is not weak (this is a relevance test on statistical grounds. The theoretical one we did in (2.2/b/iii)).

Table 3: First-Step Regression

| | *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 |

**vi)** With the IV regression, the coefficient is moving up as we expected. Consequently, we can conclude that our instrumental variable allows us to receive more reasonable results. However, we believe that the IV regression still does not measure the causal effect between *years of schooling* and *hourly wage*, because of two reasons: First, we still don't control for *ability*. Second, we believe that there is a classical measurement error in the instrument *twihigh*.

# 3 R-Code

```
#######################################################
# Problem Set 3/ Empirical Methods/ HS17
#######################################################
# Group: AliDaJo
# Datum: 12. December 2017
# Version: Final Version
#######################################################

# Load libraries
library(ggplot2)
library(stargazer)
library(dummies)
library(foreign)
# install.packages("AER")
library(AER)
```

```
#------------------------------
# Exercise 2.1: Dealing with Measurement Error
#------------------------------
# Load the Data
d.indicators <- read.csv("indicators.csv", sep=',', header=TRUE)
head(d.indicators); tail(d.indicators); summary(d.indicators);
levels(d.indicators$country)

colnames(d.indicators)

summ <- subset(d.indicators, select = c(corruptionun, mortalityun,
ruleoflaw, govmort, hospital_deaths))
stargazer(summ, header = TRUE, type = "text")
summary(summ)

# 2.1/b): regress mortality on corruption
#------------------------------
r.mod1 <- lm(mortalityun~1+corruptionun, data = d.indicators)
stargazer(r.mod1, header = FALSE, type = "text")
summary(r.mod1) # p-value from a t-distribution table

g.1 <- ggplot(d.indicators, aes(x=corruptionun, y=mortalityun)) +
geom_point(lwd = 1.5) +
   geom_smooth(method = "lm", se = TRUE) +
   ggtitle("Scatter Plot with Regression line") +
   xlab("Corruption Index UN") +
   ylab("Child Mortality Index UN")

# 2.1/c): regress hospital_deaths on corruption
#------------------------------
# ii
r.mod2 <- lm(hospital_deaths~1+corruptionun, data = d.indicators)
stargazer(r.mod2, header = FALSE, type = "text")
summary(r.mod2) # p-value from a t-distribution table

# iii
g.2 <- ggplot(d.indicators) + geom_point(aes(x=corruptionun,
y=mortalityun, colour="UN Index")) +
  geom_point(aes(x=corruptionun, y=hospital_deaths, colour="Hospital Index")) +
  geom_smooth(aes(x=corruptionun, y=mortalityun, color = "UN Index", fill="UN Index"),
   method = "lm", se = TRUE, alpha = 0.25) +
  geom_smooth(aes(x=corruptionun, y=hospital_deaths, color = "Hospital Index",
   fill="Hospital Index"),
   method = "lm", se = TRUE, alpha = 0.25) +
  labs(title = "Scatter Plots with Regression lines", x= "Corruption Index UN",
   y="Child Mortality") +
  scale_colour_manual(name="legend", values=c("red", "blue"))+
  scale_fill_manual(name="legend", values=c("red", "blue"))
stargazer(r.mod1, r.mod2, type="text")
```

```
conf.mod1 <- confint(object=r.mod1, parm="corruptionun", level = 0.95)
conf.mod2 <- confint(object=r.mod2, parm="corruptionun", level = 0.95)
stargazer(conf.mod1, conf.mod2, type = "text")

# 2.1/d): regress mortality on RuleOfLaw
#-------------------------------
r.mod3 <- lm(mortalityun~1+ruleoflaw, data = d.indicators)
stargazer(r.mod3, header = FALSE, type = "text")
stargazer(r.mod1,r.mod2,r.mod3, header = FALSE, type = "text")
summary(r.mod3)

# 2.1/e): regress mortality on RuleOfLaw
#-------------------------------
r.mod4 <- lm(govmort~1+corruptionun, data = d.indicators)
stargazer(r.mod1, r.mod2, r.mod3, r.mod4, header = FALSE, type = "text")

#-------------------------------
# Exercise 2.2: IV Regression
#-------------------------------
# Load the Data
d.educ_twins <- read.csv("https://raw.githubusercontent.com/lachlandeer/
moec415_data_archive/master/ps01/BCHHS_data.csv")
head(d.educ_twins); summary(d.educ_twins); names(d.educ_twins)

# Generate lnearn and agesq
lnearn <- log(d.educ_twins$earning)
agesq <- (d.educ_twins$age)^2/100
d.educ_twins <- cbind(d.educ_twins, lnearn, agesq)

# 2.2/a)
#-------------------------------
r2.mod1 <- lm(lnearn~1+highqua+age+agesq, data = d.educ_twins)
stargazer(r2.mod1, type = "text")
summary(r2.mod1)

r2.mod2 <- ivreg(lnearn~1+highqua+age+agesq | 1+twihigh+age+agesq, data = d.educ_twins)
stargazer(r2.mod1, r2.mod2, type="latex")
summary(r2.mod2)

# 2.2/b)
#-------------------------------
stage1_r2.mod.2 <- lm(highqua~twihigh+age+agesq, data = d.educ_twins)
stargazer(stage1_r2.mod.2, header = TRUE, type = "latex")
summary(stage1_r2.mod.2)
```