# Problemt Set 1

Global Poverty and Economic Development

Wenjie Tu

Fall Semester 2021

## 1 Education & Poverty

$$\ln(inc_i) = \theta_0 + \theta_1 educ_i + \epsilon_i \tag{1}$$

### 1.1

In equation (1), the education in level terms while the dependent variable (income) in log terms.

Take the differential of both sides:

$$d\ln(inc_i) = \theta_1 d(educ_i)$$
$$\frac{d(inc_i)}{inc_i} = \theta_1 d(educ_i)$$

Multiply both sides by 100 and rearrange:

$$\frac{100 \times d(inc_i)}{inc_i} = 100 \times \theta_1 d(educ_i)$$

$$100 \times \theta_1 = \frac{\frac{100 \times d(inc_i)}{inc_i}}{d(educ_i)} = \frac{\%\Delta inc_i}{\text{unit}\Delta educ_i}$$

Therefore, $100 \times \theta_1$ can be interpreted as the percentage change in $inc_i$ for a unit increase in $educ_i$. In other words, holding other independent variable constant (only one independent variable in equation (1)), an additional year of schooling is associated with a 5% increase in income per year.

Mechanism: since the relationship is positive, it suggests that more years of schooling imply a higher income level. This could be because people who spend more time in school are able to access jobs with higher payment.

### 1.2

Null and alternative hypotheses:

- $H_0 : \theta_1 = 0$
- $H_1 : \theta_1 \neq 0$

$t$-statistic:

$$t = \frac{\hat{\theta}_1 - \theta_1}{\text{SE}(\hat{\theta})} = \frac{0.07}{0.4} = 0.175$$

Construct confidence interval:

$$\begin{aligned} CI &= \hat{\theta} \pm t_{\frac{\alpha}{2}} \times \text{SE}(\hat{\theta}) \\ &= 0.07 \pm 2.048 \times 0.4 \\ &= (-0.7492, 0.8892) \end{aligned}$$

Note: we can use the command `qt(1-alpha/2, df=30-2)` in R console to get the corresponding $t_{\frac{\alpha}{2}} = 2.048$.

```
## [1] 2.048407
```

Clearly, 95% confidence interval covers 0. We cannot reject the null hypothesis at $\alpha = 0.05$ level.

### 1.3

In specification (1), there is only one independent variable - years of schooling. The estimate for $\theta_1$ is more likely to be biased due to the following reasons:

- Measurement error: when asked in the survey, subjects might overstate their incomes out of vanity. There might be a measurement error in dependent variable.
- Omitted variables bias: some variables such as family, innate ability also have an impact on both income and education but they are not included in the model. The error term absorbs the omitted variables, which leads to a violation of mean-zero-error assumption (i.e., $\mathbb{E}[\epsilon_i | educ_i] \neq 0$).
- Reverse causality: individuals with higher income would also pursue higher further education.

### 1.4

There is a measurement error in income data. People tend to overstate their incomes out of vanity. The estimate would **NOT** be biased in expectations. But a possible concern is that the standard errors could be much larger. A measurement error in $Y$ is not a big problem since we can still get unbiased estimates (with larger standard errors). It is only an issue if the error is correlated with education.

## 2 Data Analysis

### 2.1

### 2.2

Regression output is formatted in a table (see *Table 4*).

Table 1: Summary Statistics for Immunizations

| Statistic | N | Mean | St. Dev. | Min | Pctl(25) | Pctl(75) | Max |
|---|---|---|---|---|---|---|---|
| BCG | 5,199 | 83.319 | 20.627 | 1.000 | 77.000 | 98.000 | 99.000 |
| DPT | 6,352 | 78.410 | 23.101 | 1.000 | 69.000 | 95.000 | 99.000 |
| HepB3 | 3,070 | 83.718 | 19.660 | 1.000 | 79.000 | 96.000 | 99.000 |
| Hib3 | 2,358 | 86.055 | 17.458 | 1.000 | 83.000 | 97.000 | 99.000 |
| Pol3 | 6,358 | 79.138 | 22.983 | 1.000 | 71.000 | 96.000 | 99.000 |
| measles | 6,234 | 77.320 | 22.536 | 1.000 | 66.000 | 95.000 | 99.000 |

Table 2: Summary Statistics for Life Expectatncy

| Statistic | N | Mean | St. Dev. | Min | Pctl(25) | Pctl(75) | Max |
|---|---|---|---|---|---|---|---|
| female | 10,723 | 66.093 | 12.142 | 22.394 | 57.183 | 75.500 | 87.300 |
| male | 10,723 | 61.426 | 10.974 | 16.286 | 53.878 | 69.561 | 81.600 |
| total | 10,723 | 63.703 | 11.502 | 19.266 | 55.585 | 72.412 | 84.278 |

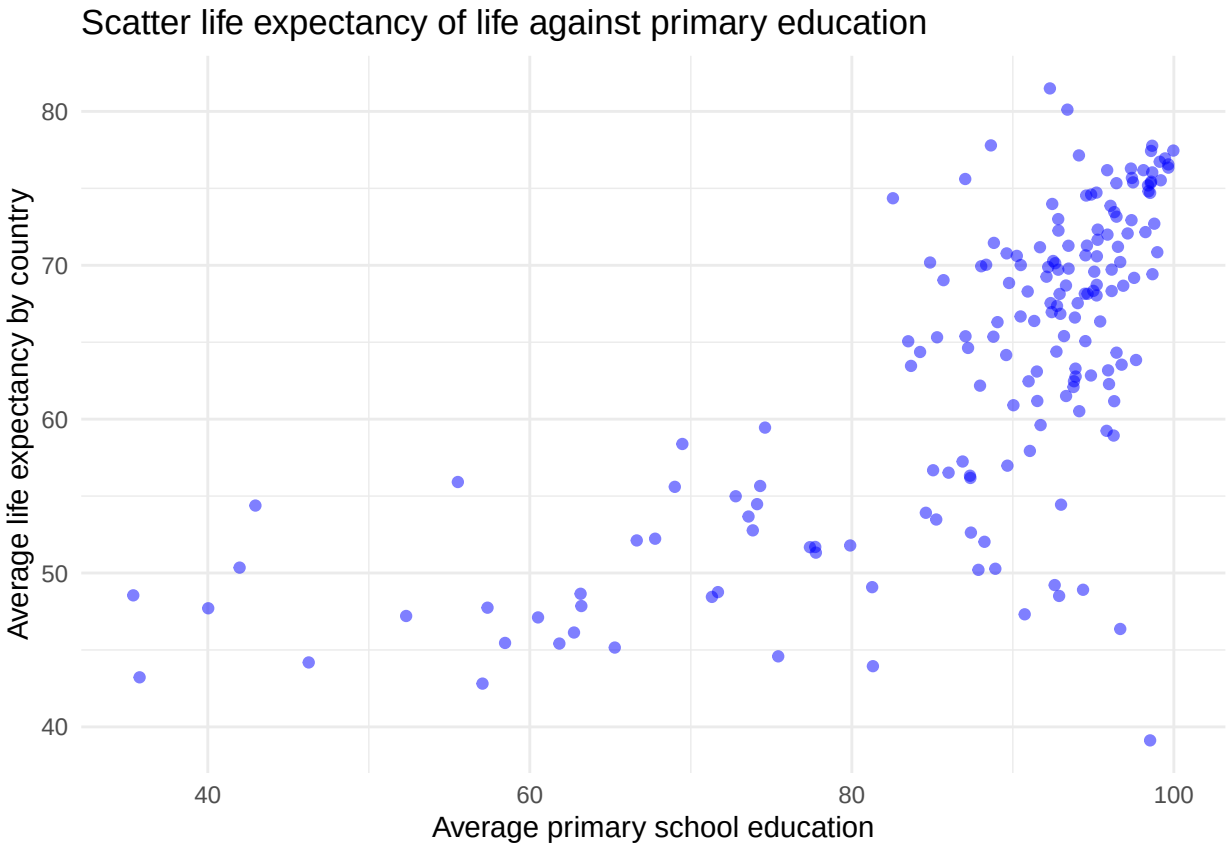
Table 3: Summary Statistics for School Enrollment

| Statistic | N | Mean | St. Dev. | Min | Pctl(25) | Pctl(75) | Max |
|---|---|---|---|---|---|---|---|
| net primary | 2,115 | 88.422 | 13.589 | 25.618 | 86.624 | 96.815 | 99.999 |
| net secondary | 1,577 | 69.451 | 25.008 | 2.684 | 53.617 | 89.037 | 100.000 |

Table 4: Regression output 2.2

| | Life expectancy at birth (total) | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Private expenditure | $-0.843^{***}$ (0.094) | | |
| Public expenditure | $2.154^{***}$ (0.067) | | |
| Sanitation access | | $0.184^{***}$ (0.004) | |
| Water access | | $0.164^{***}$ (0.007) | |
| Log of primary | | | 1.605 (1.135) |
| Log of secondary | | | $12.118^{***}$ (0.364) |
| Constant | $62.690^{***}$ (0.391) | $41.400^{***}$ (0.383) | $14.262^{***}$ (4.069) |
| Observations | 3,528 | 4,574 | 1,377 |
| $R^2$ | 0.258 | 0.753 | 0.670 |
| *Note:* | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

### 2.3




## 3 Fixed Effects

### 3.1

The mean of $x_1$ is 0.720 with a standard deviation of 0.299

### 3.2

From table 5, we see that the coefficient on $x_1$ is 0.415

Table 5: Regression table for 3.2

| | *Dependent variable:* |
|---|---|
| | y |
| x1 | 0.415 (0.326) |
| Constant | 6.516*** (0.254) |
| Observations | 1,000 |
| $R^2$ | 0.002 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

### 3.3

From table 6, we see that the coefficient on $x_1$ is 0.475 if we use household fixed effects.

Table 6: Regression table for 3.3

| | *Dependent variable:* |
|---|---|
| | y |
| x1 | 0.475*** (0.031) |
| Observations | 1,000 |
| $R^2$ | 0.212 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

### 3.4

The mean of $\hat{x}_1$ is 40.720 with a standard deviation of 49.013

### 3.5

Table 7: OLS regressions comparison in 3.5

| | *Dependent variable:* | |
|---|---|---|
| | y | |
| | (1) | (2) |
| x1 | 0.415 (0.326) | |
| x1hat | | 0.052*** (0.001) |
| Constant | 6.516*** (0.254) | 4.687*** (0.070) |
| Observations | 1,000 | 1,000 |
| $R^2$ | 0.002 | 0.691 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

$$\text{OLS}: \quad y_{it} = \theta_0 + \theta_1 x_{it} + \eta_{it} \tag{2}$$

$$\text{FE}: \quad y_{it} = \alpha_i + \beta x_{it} + \epsilon_{it} \tag{3}$$

In table 7, we see the coefficients of interest are different in OLS regression.

Intuition: in OLS regression (2), $\theta_0$ is the overall intercept and it is the same across all households. We cannot control for any household-specific characteristics in OLS specification. Hence, the coefficient of $\hat{x}_1$ differs from the coefficient of $x_1$.

### 3.6

In table 7, we see the coefficients of interest are the same in fixed-effects model.

Table 8: Fixed effects regressions comparison in 3.6

| | *Dependent variable:* | |
|---|---|---|
| | y | |
| | (1) | (2) |
| x1 | 0.475*** (0.031) | |
| x1hat | | 0.475*** (0.031) |
| Observations | 1,000 | 1,000 |
| $R^2$ | 0.212 | 0.212 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

Intuition: in fixed-effects model (3), $\alpha_i$ is the household-specific intercept. It allows for any household-specific effect. We can control for household-specific effect in an indirect way simply by demeaning the variables. $\beta$ is also called within estimator. We can difference out the household-specific effect (e.g. intercept). Therefore, the coefficient of $\hat{x}_1$ does not differ from the coefficient of $x_1$.

Within transformation:

$$y_{it} - \bar{y}_i = \underbrace{(\alpha_i - \bar{\alpha}_i)}_{0} + \beta(x_{it} - \bar{x}_{it}) + \epsilon_{it} - \bar{\epsilon}_i$$

# 4 Omitted Variable Bias

## 4.1

True model:

$$Y_i = \theta_0 + \theta_1 T_i + \theta_2 X_i + \epsilon_i$$

We estimate:

$$Y_i = \beta_0 + \beta_1 T_i + \eta_i$$

Take covariance w.r.t. $T_i$:

$$\begin{aligned} Cov(T_i, Y_i) &= Cov(T_i, \beta_0 + \beta_1 T_i + \eta_i) \\ &= \beta_1 Var(T_i) + Cov(T_i, \eta_i) \end{aligned}$$

Rearrange:

$$\hat{\beta}_1 = \frac{Cov(T_i, Y_i)}{Var(T_i)}$$

Substitute $Y_i$ with the true model:

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{Cov(T_i, Y_i)}{Var(T_i)} \\
&= \frac{Cov(T_i, \theta_0 + \theta_1 T_i + \theta_2 X_i + \epsilon_i)}{Var(T_i)} \\
&= \frac{\theta_1 Var(T_i) + \theta_2 Cov(T_i, X_i) + Cov(T_i, \epsilon_i)}{Var(T_i)} \\
&= \theta_1 + \theta_2 \frac{Cov(T_i, X_i)}{Var(T_i)} + \frac{Cov(T_i, \epsilon_i)}{Var(T_i)}
\end{aligned}
$$

$$
\mathbb{E}[\hat{\beta}_1] = \theta_1 + \theta_2 \frac{Cov(T_i, X_i)}{Var(T_i)} \tag{4}
$$

Given that:

$$
\mathbb{E}[X_i | T_i] = \alpha_0 + \alpha_1 T_i
$$

Rewrite the above equation:

$$
X_i = \alpha_0 + \alpha_1 T_i + \nu_i
$$

$$
\begin{aligned}
Cov(T_i, X_i) &= Cov(T_i, \alpha_0 + \alpha_1 T_i + \nu_i) \\
&= \alpha_1 Var(T_i) + Cov(T_i, \nu_i)
\end{aligned}
$$

Rearrange:

$$
\hat{\alpha}_1 = \frac{Cov(T_i, X_i)}{Var(T_i)}
$$

Substitute above into equation (4):

$$
\mathbb{E}[\hat{\beta}_1] = \theta_1 + \theta_2 \hat{\alpha}_1
$$

## 4.2

Because people are randomized to treatment and control groups, on average there is no difference between these two groups on any characteristics other than their treatment.

This means that before the treatment is given, on average the two groups (T and C) are equivalent to one another on every observed and unobserved variable. For example, the two groups should be similar in all pre-treatment variables: age, gender, motivation levels, heart disease, math ability, etc. When the treatment is assigned and implemented, any differences between outcomes can be attributed to the treatment. Thus, OVB can be addressed through randomization.

## 4.3

### 4.3(a)

I would expect the sign of $\theta_1$ to be negative. If there is a conflict in a geographical unit, the income for the geographical unit will decrease. Hence, $\theta_1 < 0$.

**4.3(b)**

I would expect the sign of $\theta_2$ to be positive and the sign of $\alpha_1$ to be positive.

- An increase in mineral prices would lead to a rise in people's income since they can sell minerals at higher prices. Therefore, $Cov(X_i, Y_i) > 0$ and $\theta_2 > 0$.
- More conflicts between groups or villages would increase the mineral prices. The reasoning is that higher selling prices must compensate for the losses in conflicts. Thus, $Cov(T_i, X_i) > 0$ and $\alpha_1 > 0$.

**4.3(c)**

$$\mathbb{E}[\hat{\beta}_1] = \theta_1 + \theta_2 \hat{\alpha}_1$$

It is argued in previous question that $\theta_1 > 0$ and $\hat{\alpha}_1 > 0$. Hence, the estimated effect of conflict on income would be biased upward (i.e., $\mathbb{E}[\hat{\beta}_1] > \theta_1$).

Regarding examples below, equation (4) is used for analysis.

**Example 1:**

- $Y$ is the hourly wage in the first job.
- $T$ is whether the individual receives higher education.
- $X$ is unobserved innate ability.

In this setting, individuals with higher innate ability **on average** will earn higher wages (i.e., $\theta_2 > 0$), and individuals with higher innate ability **on average** will receive higher education (i.e., $Cov(T, X) > 0$). Therefore, there is a **upward bias** in the estimated effect.

**Example 2:**

- $Y$ is the final score for a course.
- $T$ is the time spent on the course per week.
- $X$ is unobserved innate ability.

In this setting, individuals with higher innate ability **on average** will achieve higher score (i.e., $\theta_2 > 0$) while individuals with higher innate ability **on average** will spend a bit less time on the course (i.e., $Cov(T, X) < 0$). Hence, there is a **downward bias** in the estimated effect.