Development Economics Assignment #1

George Tyler

working with Han Zhou, Nicholas Somogyi, Melanie Lautrup

August 23, 2019

1 World Bank Development Report Data

1. The mean and standard deviation GNP per capita are $5340 and $9307 respectively; the mean infant mortality is 41 per thousand and with a standard deviation of 38.4. The illiteracy rate among females is higher on average (mean 25.37% female versus 16.43% male); hence the gender gap for illiteracy is on average 8.94%. Assuming that literacy is a proxy for basic education, we can surmise some reasons why a literacy gap exists. Poor families have constrained resources to invest in their children's human capital, and they might favour educating a boy if a boy represents a higher expected future income (King & Hill, 1993). An important factor is the existence of gendered expectations and norms against girls' education. Other obstacles specific to girls' secondary education - for example, early marriage and inadequate sanitation facilities at school - might exacerbate the gender disparity in adult literacy if literacy is not achieved by the end of primary education.

Table 1: World Bank Development Report

	(1)	
	mean	sd
GNP per capita (US$1995)	5340.887	9307.99
% illiterate, female aged 15+	25.37681	26.20977
% illiterate, male aged 15+	16.43478	17.89261
infant mortality rate, per 1000	41.16923	38.40462
Observations	196	

2. In the 50 richest countries, the mean illiteracy rate is 5.21%, with a range between 0% and 27%. In 36 of the 50 poorest countries[1], the mean illiteracy rate is 45.69%, with a range between 1 and 86. This suggests a negative relationship between illiteracy and income: a richer country is likely to have less

[1] Illiteracy data is missing for the other 14.

illiteracy. Infant mortality in the 50 richest countries has a mean of 13.3 per 1000, with range between 4 and 87 per 1000; in the 50 poorest, a mean of 85.16 per 1000, with range between 20 and 170. Under 5 mortality in the 50 richest countries has a mean of 18.04, with range between 5 and 136; in the 50 poorest, a mean of 134.5 with range 23 to 286 per 1000. These mortality statistics also show a clear correlation of higher income with lower rates of infant and under-5 mortality.

Table 2: Top 50

	(1)		
	mean	min	max
country name	.	.	.
% illiterate, female aged 15+	55.05556	2	93
% illiterate, male aged 15+	36.02778	1	78
% illiterate, total aged 15+	45.69444	1	86
infant mortality rate, per 1000	85.16	20	170
¡ age 5 mortality rate, per 100	134.5	23	286
GNP per capita (US$1995)	335.64	112	656
GNP per capita (PPP)	1193.191	410	2240
1977 infant mort rate, per 1000	132.3256	62	263
1977 GNP per capita (US$1995)	363.5625	156	906
Observations	50		

Table 3: Bottom 50

	(1)		
	mean	min	max
country name	.	.	.
% illiterate, female aged 15+	6.119048	0	38
% illiterate, male aged 15+	4.261905	0	28
% illiterate, total aged 15+	5.214286	0	27
infant mortality rate, per 1000	13.3	4	87
¡ age 5 mortality rate, per 100	18.04348	5	136
GNP per capita (US$1995)	14885.22	3304	46448
GNP per capita (PPP)	14627.55	4930	29230
1977 infant mort rate, per 1000	27.97826	8	122
1977 GNP per capita (US$1995)	11311.65	1092	38314
Observations	50		

3. The median GNP per capita is $1430.5, which is much less than the mean GNP of $5340. The distribution of income thus has a strong rightward skew; that is, a large amount of global income is concentrated among a small set of rich countries. Hence, there is a large amount of global income inequality.

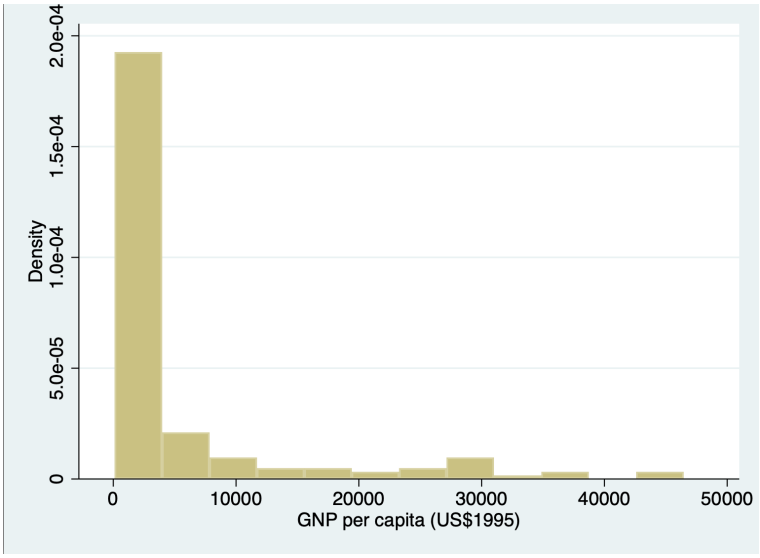

Figure 1: Histogram of GNP distribution.

4. Regressing illiteracy as dependent variable upon GNP per capita yields the coefficient of GNP/k as $\beta = -.0010328$ with standard error .0001743. This implies that an increase of $1000 in per-capita GDP results in a decrease in illiteracy of %1.0328; however due to the skewness of the distribution a log transformation should likely be applied. That said, the sign is as expected - more income per capita should be correlated with a decrease in total illiteracy, since an increase in national wealth allows for more expenditure on education. The t-statistic for the coefficient is -5.92, which satisfies our critical value, so we can reject the null hypothesis that the coefficient is equal to 0. We can also say that in 95% of samples, the true coefficient lies between -0.0013781 and -0.0006875.

Table 4: Regression 1.4

	(1)
	% illiterate, total aged 15+
GNP per capita (US$1995)	-0.00103***
	(-5.92)
Constant	28.40***
	(13.17)
Observations	118

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5. Regressing infant mortality against GNP per capita yields the coefficient as $\beta = -0.0019967$; a $1000 increase in GNP/k leads to the infant mortality per 1000 people dropping by nearly 2. This is statistically significant from 0; we know this, as it passes the t-test. The critical value for the t-distribution at 157

df is -1.646; since our calculated t-value is -7.15, well beyond this, we can reject the null hypothesis that $\beta = 0$.

Table 5: Regression 1.5

	(1) infant mortality rate, per 1000
GNP per capita (US$1995)	-0.00200***
	(-7.15)
Constant	54.44***
	(18.14)
Observations	159

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

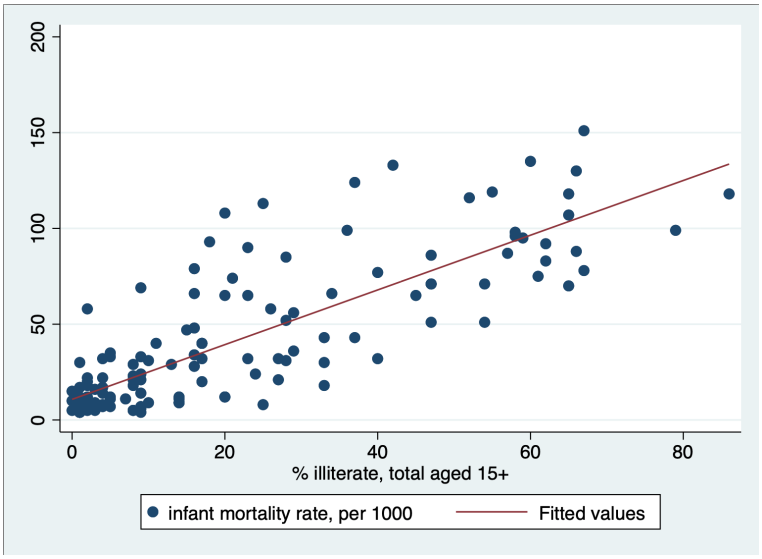

Figure 2: Scatterplot of illiteracy rate against infant mortality rate.

6. We see that infant mortality rate is positively correlated with illiteracy rate, due among other things to the GNP/k confounder. Note also the clustering around the origin.

7. From the three regressions run, we learn that GNP is negatively correlated with illiteracy and infant mortality, and that infant mortality and illiteracy are positively correlated to a relatively high extent ($R^2 = 0.6736$). These two variables coincide, but do not cause each other. We can imagine that illiterate mothers face obstacles in obtaining child healthcare, but it is more likely that these variables are simultaneously caused by the level of GNP. With this in place, we only need to ascertain whether GNP *causes* higher rates of illiteracy and infant mortality, or whether the reverse is true. It is likely

that *both* causal pathways exist - illiteracy is an obstacle to higher GNP, much like low GNP prevents expenditure on reducing illiteracy.

2 Cross-country growth regressions

1. In estimating the growth regression, J. L. Gallup, Sachs, and Mellinger (1999, p. 33) discuss the potential endogeneity of Malaria cases with GDP growth. There may be an instance of reverse causation, in that an increase in GDP allows for more resources to control malaria, and that the more effective institutions coincident with high GDP allow for better malaria control (J. Gallup & Sachs, 2001). In this case, an instrument is needed to isolate the effect of malaria on GDP; Sachs chooses 'malaria ecology' as an instrument: different malaria vectors are geographically distributed, and have differing efffectiveness in passing on the disease (and are thus highly correlated with incidence). However, unlike malaria incidence, malaria ecology is independent of GDP, and can affect it only *through* incidence.

2. Sachs' finding that countries with malaria grow slower is a causal finding. However, malaria incidence is highly correlated with the 'tropicar' variable, and may also act as a proxy for other geographically associated diseases. Despite this, Sachs is able to construct a valid instrument, address reverse causation, and so successfully isolates the effect of malaria independent of geographic region (although these remain tightly linked). Although it is difficult to support this finding with a robust story backed up by microeconomic research, there are various channels. Namely, these are a reduction in foreign direct investment and tourism, limitation of internal movement (underexposed city-dwellers are reluctant to visit high-risk rural areas), and a reduction in child health and early cognitive development (J. Gallup & Sachs, 2001). A possible alternative to the hypothesis that malaria matters is the institutional hypothesis,

3. In order to isolate the effect of geography, and specifically the effect of malaria within geography, Sachs must include other determinants of GDP growth. The control variables included in table 3, specification

7 (J. L. Gallup et al., 1999, p. 201) are:

$$\begin{aligned}
\text{GDP per capita 1965} &= -2.4 \\
\text{Years of secondary schooling 1965} &= 0.1 \\
\text{Log life expectancy 1965} &= 3.4 \\
\text{Openness (0-1)} &= 1.8 \\
\text{Public institutions (0-10)} &= 0.4 \\
\text{Tropicar (percentage)} &= -0.5 \\
\text{Malaria index 1966} &= -1.6 \\
\text{dMal6694} &= -1.9 \\
\text{Log coastal density} &= 0.2 \\
\text{Log interior density} &= -0.1
\end{aligned}$$

Generally, the coefficients correspond to our expectations, and allow us to isolate the effect of the malaria index and its change on GDP growth. Because Sachs has constructed his index taking ecology into account, he claims that this instrument is independent of GDP affecting malaria cases. So, he has resolved the joint determination issue by using inherent area-level ecology, which does not correlate to GDP.

4. Another set of diseases correlated with area-level ecology are the various forms of *Filiarisis*, which is caused by an infection of *Filiaroidea* roundworms transmitted by vectors including flies and mosquitoes. Diseases in this family include river blindness and lymphatic filiarisis. These diseases are endemic to the tropics, allowing for a similar analysis: 120 million people are infected, with 94% living in South-East Asia or Sub-Saharan Africa (WHO, 2019). Therefore, a similar argument can be made for an instrumental variable identification strategy: while incidence of *Filiarisis* is endogenous with respect to GDP growth, using area-level ecology to create an instrument that is highly correlated with *Filiarisis* incidence yet orthogonal to GDP growth. We might like to focus, for example, on the different vectors of the worms; according to the WHO, Lymphatic filariasis is transmitted the urban Culex mosquito, rural Anopheles mosquito, and Aedes mosquito, mainly in endemic islands in the Pacific. Alternatively we could separate our sample into areas similar to the paper. These would provide us with a GDP-orthogonal geographic instrument.

3 Poverty and nutrition

1. We set income as the dependent variable, and kcal as our independent variable, obtaining:

$$inc_i = \alpha + \beta_1 kcal_i + \epsilon_i \tag{1}$$

as our estimating equation. If we had sufficient data, we could include Mincerian controls and demographic information in our estimating equation as a vector.

2. The regression cannot establish a one-way causal link between kcal and income: there remains the possibility of simultaneity, which is quite likely (wealthier people are able to eat more, and eating more enables more money to be earned).

3. The coefficient of the above regression should be positive: *ceteris paribus*, a higher caloric intake should correspond to higher earnings. However, the rate of change is likely to decrease above a subsistence caloric intake, so a squared term might also be appropriate.

4.

$$inc_i = \alpha + \beta_1 educ_i + \epsilon_i \tag{2}$$

$$kcal_i = \gamma + \delta_1 educ_i + \nu_i \tag{3}$$

β and δ should be positive: education should correlate positively to both income and calorie intake, though I would surmise that education would affect calorie intake indirectly, *via* income.

5. We are estimating income as our dependent variable; the short regression will suffer from omitted variable bias from equation 2, as we have left this out. We will also see, due to the positive interaction between *kcal* and *educ*, a postive sign to this bias. We derive this as follows.

$$\begin{aligned} \text{Short equation:} \quad & inc_i = \alpha + \beta_1 kcal_i + \epsilon_i \\ \text{Long equation:} \quad & inc_i = \alpha + \beta_1 kcal_i + \beta_2 educ_i + \epsilon_i \end{aligned}$$

Our variable of interest is β_1. We can re-arrange equation 3 as $educ_i = \gamma + \delta_1 kcal_i + \nu_i$ and so substitute into the long equation, rearranging:

$$\begin{aligned} inc_i &= \alpha + \beta_1 kcal_i + \beta_2(\gamma + \delta_1 kcal_i + \nu_i) + \epsilon_i \\ inc_i &= (\alpha + \beta_2\gamma) + (\beta_1 + \beta_2\delta_1)kcal_i + (\beta_2\nu_i + \epsilon_i) \end{aligned}$$

The bias on β_1 is therefore expressed as the difference between the 'true' coefficient in the long equation, β_1, and that derived from our auxiliary regression, $\beta_2\delta$. Since we have hypothesised in equations 2 and

3 that β_2 and δ are positive, we observe that their product is positive, and conclude that the sign of the omitted variable bias is positive. Equivalently, this result is backed up by evaluating the 'formula' for omitted variable bias (Greene, 2003, p. 148):

$$E(b_1) = \beta_1 + \beta_2 \frac{Cov(kcal, educ)}{Var(kcal)}$$

which we can see is positive, since Cov(kcal, educ) and Var(kcal) are positive.

6. There are broadly two categories of measurement error: attenuation bias, and non-classical measurement error.

   Attenuation bias comes from errors ν_i in each measurement of $kcal_i + \nu_i$ being uncorrelated with the regression error ϵ_i. In our context, this would mean that the variance of $kcal_i$ would be larger than is actually the case. This would entail that the probability limit of the estimator would take the form

$$\text{plim}\hat{\beta} = \beta \frac{var(kcal_i)}{\text{var}(kcal_i) + \text{var}(\nu_i)}$$

   . We see that the estimate of β will be scaled by a factor less than 1 and so will *underestimate* our coefficient.

   A non-classical measurement error that would upwardly bias our coefficient would come from a systemic bias in measurement between richer and poorer people. If rich people underestimated their caloric intake, and poor people overestimated it, this would mean that the residuals for poor and rich would both approach the middle of the distribution of calories. Assuming a positive relationship, this would reduce the amount of variance compared to the true variance, and *over-estimate* the correlation between nutritional intake and income. In turn, this would cause the regression to over-estimate the effect of caloric intake, overestimating the coefficient.

4 Estimating the Returns to Schooling

a. The parameter b in the Mincerian regression equation $\ln w_i = a_i + bS_i$ can be interpreted as the percentage increase in wage for an extra year in school. It is therefore the rate of return to schooling. This coefficient may not be a good estimate of the population parameter because of omitted variable bias, which include family background variables such as income, parental education, and school quality (Griffin & Ganderton, 1996); experience, and ability.

b. **i.** We can calculate the first stage difference-in-differences (DD) estimate of INPRES on schooling by using 'young' as the time dummy and 'high' as the treatment dummy, and calculating the coefficient

of their interaction as the DD estimate. The corresponding regression equation is therefore

$$yeduc_i = \alpha + \beta_1 high_i + \beta_2 young_i + \gamma high * young_i + \epsilon_i$$

The first stage DD estimate is 0.117: the students living in the high-intensity areas gained 0.117 *more* years of schooling after the treatment than those living in the low-intensity areas. As the program was targeted at under-performing areas, this is a 'catch-up' effect: ultimately, the students who had the treatment did better than those who did not, after controlling for time effects. We can therefore attribute this increase as the result of the INPRES program.

ii. The effect of INPRES on log hourly wages is 0.00656: the students living in the high-intensity areas earned 0.6% more per hour after the treatment than those living in the low-intensity areas. This is subject to the parallell trends assumption: we must assume that if the program had not gone ahead, the two regions would have had a similar rate of increase in earnings. If this is the case, the jump of 0.6% in earnings from the treatment group stands out as a causal effect. We can test this assumption by comparing the 'old' and 'veryold' groups, both before the treatment happened (this is the 'control experiment' Duflo undertakes to test for mean reversion, (Duflo, 2001, p. 799)). If the overall pre-treatment trend is similar for both high and low intensity regions, we can extrapolate and say that they would have grown in the same fashion if not for the treatment.

Table 6: Regression 4b

	(1) log hourly wage
dummy high program intensity region	-0.124***
	(-11.28)
dummy born 1968-1972	-0.299***
	(-30.66)
high_young	0.00656
	(0.43)
Constant	7.065***
	(1003.74)
Observations	30720

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

iii. The additional assumption for unbiasedness of the Wald estimate is that the potential outcomes of wages are not directly affected by the treatment; in other words, the program must affect wages only

through education. We could perhaps test this by using a cohort-by-cohort analysis. Each cohort will have had a different exposure time to the treatment; therefore, if the effect has an upward trend depending on the length of exposure to the treatment, then we can say that the program affected education (Figure 1 in Duflo 2001). However, we cannot say for certain whether general equilibrium effects of the program affected overall wages (since it was such a large government expenditure, this is a valid concern).

iv. We divide 0.00656 by 0.117 = 0.056068376, to obtain our Wald estimate of 5.6. This is the IV estimate of the coefficient of education on wages in our reduced-form equation; hence, due to the DD identification strategy we can say that an extra year of schooling *causes* hourly wage to increase by 5.6%.

c. **i.** The estimated coefficient of program intensity on mean difference in education between old and young cohorts is 0.278; the p-value is 0.00 indicating high significance.

Table 7: Regression 4ci

	(1)
	educ_dif
(mean) prog_int	0.278***
	(5.48)
(mean) ch71	0.00000397***
	(6.53)
Constant	-1.010***
	(-5.60)
Observations	280

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

ii. The estimated coefficient of program intensity on mean difference in log hourly wage between cohorts is 0.026; the p-value is 0.00 indicating high significance. .

iii. Dividing equations (1) and (2) by alpha and gamma respectively we obtain:

$$\frac{1}{\alpha}(S_{Y_j} - S_{O_j}) = P_j + \frac{\nu_j}{\alpha}$$
$$\frac{1}{\gamma}(y_{Y_j} - y_{O_j}) = P_j + \frac{\epsilon_j}{\gamma}$$

Table 8: Regression 4cii

	(1) wage_dif
(mean) prog_int	0.0262**
	(2.60)
(mean) ch71	0.000000883***
	(7.52)
Constant	-0.525***
	(-14.50)
Observations	280

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

We can now subtract and rearrange the equations to obtain:

$$\frac{1}{\gamma}(y_{Y_j} - y_{O_j}) - \frac{1}{\alpha}(S_{Y_j} - S_{O_j}) = P_j + \frac{\epsilon_j}{\gamma} - P_j - \frac{\nu_j}{\alpha}$$

$$y_{Y_j} - y_{O_j} = \gamma(P_j + \frac{\epsilon_j}{\gamma}) - \gamma(P_j + \frac{\nu_j}{\alpha}) + \frac{\gamma}{\alpha}(S_{Y_j} - S_{O_j})$$

This is in the same form as equation (3) above, so we can conclude that $b = \frac{\gamma}{\alpha}$.

iv. $\frac{\gamma}{\alpha} = 0.0935 \implies 9.4\%$: compared to our Wald estimate of 5.6%, we see that it is nearly double as an estimate of the rate of return to schooling. We have now more granularity of program intensity across regions, as we are including them all in our analysis rather than separating into 'high' and 'low'. Also, we now can control for population effects by region, as we include the number of children in 1971.

v. The new estimate of b is indeed 0.0941058; this estimation of edif-pred matches part iv.

d. **i.** If we were to run the above regressions, α_k would be interpreted as the impact of program intensity on the change in years of education compared to the oldest cohort. We would expect that, if a cohort k in region i had a high intensity program, then they would improve education against the 1950 baseline by more than the same age cohort in a low-intensity region j. γ_k can be interpreted similarly; for any given cohort, the improvement in hourly wage over the 1950 baseline should be higher for those countries which experienced a more intensive INPRES program.

ii. The requirements for these to be good instruments are relevance, and the exclusion restriction. The dummy variables are relevant (that is, highly correlated with the variable of interest, years of education); this is likely to be the case, as it is a basic premise that exposure to the program (the meaning of the dummy) increases years of schooling. In fact, this much is shown in the regression

Table 9: Regression 4cv

	(1) wage_dif
Fitted values	0.0941**
	(2.60)
(mean) ch71	0.000000510***
	(3.81)
Constant	-0.430***
	(-22.73)
Observations	280

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

in question **i**. We must also state that exposure to the program affects hourly wage only through the increase in *amount* of schooling - note, not the *quality* of schooling, which must be assumed constant for the exclusion restriction (Duflo, 2001). Assuming that the program did not affect quality is reasonable, as Duflo tests for this and finds no relation between the quantity of schools constructed and the student/teacher ratio, which is a good proxy for school quality. Also, as mentioned above, we should be aware of possible general equilibrium effects arising from spillovers from intensive construction investment in a given region. This may kick-start a process of mean reversion, thereby increasing future wages for all in the region without having to actually being exposed to the program; also, the INPRES program could be correlated with other government investment in a given region. However, Duflo tests for this when doing her DD and discards this possibility.

iii. **1.** I regressed years of education on the instruments and control variables, forming the predicted values of education which I use in the next question.

2. The 2SLS estimate of b is 0.0962; this is our final estimate of returns to education at 9.6% increase in wages for a 1 year increase in schooling. It is unsurprisingly very similar to the previous Wald estimator.

3. Duflo mentions the important caveat of school *quality*; this is assumed to be constant. If the INPRES program does indeed improve school quality, this violates the exclusion restriction. We only have data on the number of schools built, so if there is a systematic relationship between this and the quality of education, this effect will fail to be captured by our 'program intensity' dummy. Hence, we will not have constructed a valid instrument, as program intensity will affect wages not only through years of education but also through quality of education. However, if this effect were to exist, we would see it in the 2SLS results as an impact of the

Table 10: Regression 4diii2

	(1)
	log hourly wage
Fitted values	0.0962***
	(8.52)
Observations	59938

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

program directly upon wages of those who have completed 9 or more years of education (i.e., the 'full-graduates' from program schools would earn more than those from other schools). No such relationship exists, so we can conclude that the program has an effect only through increasing the quantity of years of schooling, rather than the quality. This satisfies the exclusion restriction and verifies our analysis.

References

Duflo, E. (2001). Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment. *The American Economic Review*, *91*(4), 795–813. Retrieved August 22, 2019, from https://www.jstor.org/stable/2677813

Gallup, J., & Sachs, J. (2001). The economic burden of malaria. *The American Journal of Tropical Medicine and Hygiene*, *64*, 85–96. Retrieved August 21, 2019, from http://www.ajtmh.org/content/journals/10.4269/ajtmh.2001.64.85

Gallup, J. L., Sachs, J. D., & Mellinger, A. D. (1999). Geography and economic development. *International regional science review*, *22*(2), 179–232.

Greene, W. H. (2003). *Econometric analysis* (5th ed). Upper Saddle River, N.J: Prentice Hall.

Griffin, P., & Ganderton, P. T. (1996). Evidence on omitted variable bias in earnings equations. *Economics of Education Review*, *15*(2), 139–148. Retrieved August 22, 2019, from http://www.sciencedirect.com/science/article/pii/0272775796000015

King, E. M., & Hill, M. A. (1993). *Women's education in developing countries: Barriers, benefits, and policies*. The World Bank. Retrieved August 8, 2019, from http://elibrary.worldbank.org/doi/book/10.1596/0-8018-4534-3

WHO. (2019). Lymphatic filariasis factsheet. Retrieved August 22, 2019, from http://www.who.int/lymphatic_filariasis/epidemiology/en/

5 .do file

```
** author: George Tyler
** project: Development Economics Assignment 1
**EXTERNAL PACKAGES: estout, univar
**Section 1: wbdr.dta
use wbdr.dta
**Question 1.1
estpost sum gnppc illit_f illit_m mort_inf
esttab ., cells("mean␣sd"), using a1tables.tex, label  title (World Bank Development
   Report) append
**Question 1.2 and 1.3
preserve
        drop if missing(gnppc)
        sum mort_inf illit_t, detail
        gsort gnppc
        estpost sum in 1/50
        esttab ., cells("mean␣min␣max"), using a1tables.tex, label  title(Top 50) append
        estpost sum in -50/l
        esttab ., cells("mean␣min␣max"), using a1tables.tex, label  title(Bottom 50)
           append
        univar gnppc
        hist gnppc
restore

**Question 1.4-1.6
reg illit_t gnppc
esttab using a1tables.tex, label  title(Regression 1.4\label{Reg1.4}) append
reg mort_inf gnppc
esttab using a1tables.tex, label  title(Regression 1.4\label{Reg1.4}) append
reg mort_inf illit_t
esttab using a1tables.tex, label  title(Regression 1.5\label{Reg1.5}) append
twoway scatter mort_inf illit_t || lfit mort_inf illit_t

**Question 4
use supas.dta
**Question 4.a: DD estimation yeduc
preserve
        drop if intermed
        drop if veryold
        gen high_young = high*young
        reg yeduc high young high_young, r
**Question 4.b: DD estimation lwage
        reg lhwage high young high_young, r
```

```
        esttab using a1tables.tex, label  title(Regression 4b\label{Reg4b}) append
restore
**Question 4.c: Indirect Least Squares
preserve
        gen educ_yng = yeduc if young==1
        gen educ_old = yeduc if old==1
        gen lhwage_yng = lhwage if young==1
        gen lhwage_old = lhwage if old==1
        collapse (mean) educ_old educ_yng lhwage_old lhwage_yng ch71 prog_int, by(ROB)
        gen educ_dif = educ_yng - educ_old
        gen wage_dif = lhwage_yng - lhwage_old
        reg educ_dif prog_int ch71, r
        esttab using a1tables.tex, label  title(Regression 4ci\label{Reg4ci}) append
        reg wage_dif prog_int ch71, r
        esttab using a1tables.tex, label  title(Regression 4cii\label{Reg4cii}) append
        reg educ_dif prog_int ch71, r
        predict edif_pred
        reg wage_dif edif_pred ch71, r
        esttab using a1tables.tex, label  title(Regression 4cv\label{Reg4cv}) append
restore

**Question 4.d: 2 Stage Least Squares
gen d62=(YOB==62)
gen d63=(YOB==63)
gen d64=(YOB==64)
gen d65=(YOB==65)
gen d66=(YOB==66)
gen d67=(YOB==67)
gen d68=(YOB==68)
gen d69=(YOB==69)
gen d70=(YOB==70)
gen d71=(YOB==71)
gen d72=(YOB==72)

gen z62 = d62*prog_int
gen z63 = d63*prog_int
gen z64 = d64*prog_int
gen z65 = d65*prog_int
gen z66 = d66*prog_int
gen z67 = d67*prog_int
gen z68 = d68*prog_int
gen z69 = d69*prog_int
gen z70 = d70*prog_int
gen z71 = d71*prog_int
```

```
gen z72 = d72*prog_int

xi: reg yeduc z62 z63 z64 z65 z66 z67 z68 z69 z70 z71 z72 i.YOB*ch71, r

predict yeduc_fitted

xi: reg lhwage yeduc_fitted i.YOB*ch71, r
esttab, keep(yeduc_fitted), using a1tables.tex, label  title(Regression 4diii2 \
   label{Reg4diii2}) append
```