Health Outcomes in West Africa

Aileen Gui

Fall 2019

Purpose of project

Data

The 'GD' dataset: GiveDirectly (GD) is an NGO that started distributing large cash transfers (roughly $1000 USD) to thousands of households in rural western Kenya in 2014. These transfers were targeted to relatively poor households and grants were unconditional. The amount is equivalent to more than 50% of total annual income for many recipient households. Approximately 2 years after the intial cash transfers, the households were surveyed on a range of important economic life outcomes. I used this dataset for its measures on health and nutrition. The villages in the GD cash transfer program were randomly chosen. This dataset is part of an ongoing study currently being conducted by D. Egger, J. Haushofer, E. Miguel, P. Niehaus and M. Walker. Each observation (row) in the dataset represents one household.

- cash: indicator for treatment households
- male: indicator for if the household respondent is male
- age35: indicator for if the household respondent is at least 35 years old
- schooling: indicator for if the household respondent has completed secondary school

The 'DD' dataset: Restricts the GD dataset to only those villages where households were chosen to participate in the study. Each observation (row) in the dataset represents one household in one time period ("time").

- time: indicator for endline survey round
- eligible: indicator for household treatment eligibility
- male: indicator for if the household respondent is male
- age35: indicator for if the household respondent is at least 35 years old
- schooling: indicator for if the household respondent has completed secondary school

The 'MSS' dataset: Provided by the Miguel Satyanath and Sergenti (2004) article on rainfall as an instrumental variable for gdp which was claimed to cause greater civil conflict. This is a partial extract of data from the actual article where each observation in the dataset represents one country in one year. This is panel data.

- country_id: numerical country identifier
- year: year of the observation
- gdp_growth: annual GDP growth in that country and year
- green_index_growth: measure of land greenness (NDVI)
- democracy: measure of democratic institutions in that country and year (based on Polity dataset where higher values denote more democratic institutions and lower values denote less democratic institutions)

Section 1: Household Income and Child Health (Treatment Randomly Assigned)

In this section, we will be examining household income's effect on child health by using the "gd" dataset. It's important to note that the project's randomized design helps address ommitted variable bias, and its randomized design affects the methods chosen to to estimate treatment effects of the cash transfer intervention.

The cash transfer intervention described above was randomized across villages as opposed to within villages in order to estimate a treatment effect taking into consideration gains due to spillovers or local treatment externalities. In particular, randomization at the village level ensures that control households (those households in villages where no households were treated) were not impacted by the intervention, and so remain a valid control group. If the cash transfer intervention had been randomized within villages, then some of the benefits experienced by treated households may have spilled over or exerted a positive externality on untreated households within the same village; as a result, the estimated treatment effect would have *understated* the effectiveness of cash transfers. Randomizing across villages instead of within villages ensure that the estimated treatment effect truly captures the impact of the cash transfer program.

The randomized design implies that if externality effects are localized within villages, then treatment effects can be estimated by simply computing the difference in outcomes between households in treatment villages and households in control villages. The randomized design helped to ensure that treatment and control villages had very similar characteristics in the absence of treatment, so that the only difference between treatment and control villages was their treatment status itself. As such, any difference in average outcomes between treatment and control villages can be attributed to the treatment itself, rather than due to some underlying pre-existing observable or unobservable differences. Put differently, the randomization procedure removed any potential omitted variable bias (OVB), so that estimating the treatment effect by computing the difference in outcomes between the treatment and control villages will not lead to a biased estimate of the treatment effect.

Let's first look at summary statistics for the gd dataset that we will be using in this section.

```
knitr::opts_chunk$set(echo = TRUE)
install.packages("ivpack")
library(stargazer)
library(ivpack)
```

```
gd<- read.csv("gd.csv")
```

```
dim(gd)
```

```
## [1] 2000    6
```

```
summary(gd)
```

```
##       hhid             cash            age35         schooling
##  Min.   :   1.0   Min.   :0.0000   Min.   :0.000   Min.   :0.000
##  1st Qu.: 500.8   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.000
##  Median :1000.5   Median :1.0000   Median :1.000   Median :0.000
##  Mean   :1000.5   Mean   :0.5025   Mean   :0.506   Mean   :0.055
##  3rd Qu.:1500.2   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:0.000
##  Max.   :2000.0   Max.   :1.0000   Max.   :1.000   Max.   :1.000
##       male          nutrition
##  Min.   :0.0000   Min.   :-4.71500
##  1st Qu.:0.0000   1st Qu.:-0.35900
##  Median :0.0000   Median : 0.09100
##  Mean   :0.3095   Mean   : 0.06619
```

```
##  3rd Qu.:1.0000   3rd Qu.: 0.89500
##  Max.   :1.0000   Max.   : 3.79300
```

```
sd(gd$nutrition)
```

```
## [1] 0.9568018
```

Based on a quick look at the summary statistics, we can see that this dataset contains 2000 observations (households), where half of the households received the cash treatment, roughly 31% of households' primary respondents are male, 51% are over the age of 35, and 5.5% have completed secondary schooling. The mean nutritional index for respondents is around 0.06 with a standard deviation of 0.96.

We would like to estimate the following equation:

$$Y_i = a + bCASH_i + e_i$$

where Y is the characteristic of interest (`male, age35, schooling`), and `CASH` takes on value 1 if the village received the cash transfers. In this equation, b describes the average difference between treatment and control village households, and a gives us the average value of the variable in control village households.

Below is the code to run the five regressions, where the first three regressions determine average differences between treatment and control households for each each of the three characteristics within the full sample, and the final two regressions determine the average differences between treatment and control households in terms of percent female and percent over age 35, restricting to respondents who have not completed secondary schooling.

```
reg_male <- lm(male ~ cash, data=gd)
reg_age35 <- lm(age35 ~ cash, data=gd)
reg_schooling <- lm(schooling ~ cash, data=gd)
reg_male_noschooling <- lm(male ~ cash, data=subset(gd, schooling==0))
reg_age35_noschooling <- lm(age35 ~ cash, data=subset(gd,schooling==0))
```

These are the results displayed in table format:

```
stargazer(reg_male, reg_age35, reg_schooling,
          reg_male_noschooling, reg_age35_noschooling,
          out="Table 1",
          type = "latex",
          title="Baseline Covariate Analysis",
          dep.var.labels=c("Male", "Age 35+", "Sec School", "Male", "Age 35+"),
          covariate.labels=c("Cash Treatment"),
          align=TRUE,
          table.placement = "!h",
          header=FALSE,
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),
          no.space=TRUE)
```

The first column indicates that the fraction of male respondents is 0.309 among households in the control group and 0.02 higher among households in the treatment group. The second column indicates that the fraction of respondents above age 35 is 0.502 in control households, while this fraction is on average 0.009 higher in treatment households. The third column indicates that the fraction of respondents who have completed secondary school is 0.0503 among control households and 0.003 higher among treatment households.

Restricting attention to those households in which the primary respondent has not completed secondary school, treatment households are slightly more likely to have a primary respondent who is male (where the fraction male is 0.296 among control households and 0.002 higher in among treatment households) and slightly more likely to have a primary respondent over the age of 35 (where the fraction above age 35 is 0.5 among control households and 0.014 higher among treatment households).

Table 1: Baseline Covariate Analysis

	Dependent variable:				
	Male	Age 35+	Sec School	Male	Age 35+
	(1)	(2)	(3)	(4)	(5)
Cash Treatment	0.002	0.009	0.003	0.002	0.014
	(0.021)	(0.022)	(0.010)	(0.021)	(0.023)
Constant	0.309***	0.502***	0.053***	0.296***	0.500***
	(0.015)	(0.016)	(0.007)	(0.015)	(0.016)
Observations	2,000	2,000	2,000	1,890	1,890
Note:					*p<0.1; **p<0.05; ***p<0.01

Standard errors are reported in parentheses below the appropriate estimated coefficients. Recall that standard errors help to answer the question of how precisely estimated each coefficient is, and whether that coefficient is or is not statistically distinguishable from zero. In this context, if a coefficient is statistically indistinguishable from zero, that would indicate that the treatment and control groups are equally balanced in terms of the characteristic included as the outcome of each regression.

Comparison of the coefficient estimate and the standard error allows us to compute a t-statistic for each coefficient, which in turns allows us to determine whether or not the coefficient is statistically distiguishable from zero at a particular level of significance, usually 1%, 5%, or 10% (corresponding to 99%, 95%, and 90% confidence intervals). Let's say we are interested in determining whether coefficients are statistically indistinguishable from zero at the 5% level, as is a common standard in economics. If the absolute value of the t-statistic we compute is less than 1.96, then we cannot reject that the coefficient is statistically indistinguishable from zero at the 5% level. In other words, if the absolute value of the t-statistic we compute is less than 1.96, then we cannot reject the null hypothesis of no difference across the two groups (that is, the null hypothesis that $b = 0$), but instead, we can conclude that treatment and control households are "balanced" for that particular variable.

The formula we will use is: t=b/se(b). For the variable `male`, the t-statistic is 0.002/0.021=0.095, which is quite a bit less than 1.96. For the variable `age35`, the t-statistic is 0.009/0.022=0.409, which is again less than 1.96. For the variable `schooling`, the t-statistic is 0.003/0.010=0.3 which is less than 1.96 in absolute value. For the variable `male`, restricted to households where the respondent has not completed secondary school, the t-statistic is 0.002/0.021=0.095, which is less than 1.96. For the variable `age35`, restricted to households where the respondent has not completed secondary school, the t-statistic is 0.014/0.023=0.609, which is again less than 1.96.

In summary, it appears that the randomization succeeded in creating comparable treatment and control groups, both for the full sample and when we restrict attention to only those households where the respondent has not completed secondary school.

Now that we've established that cash dispersal was indeed random, since there seems to be no significance between certain attributes and whether their household was more likely to receive cash transfers, we want to determine the difference between treatment and control households in terms of their nutritional status in the endline survey. Nutritional status measure is an index of food security, and is created by combining survey questions on the respondents and their household members' numbers of meals eaten, number of days that they had to cut back on or skip meals in the last week, and number of days they went to bed hungry. More positive values denote better outcomes. Taken together, the index can be thought of as a summary measure of the household's regular access to adequate food. This index is standardized, and can take on positive or negative values.

The regression of interest this time is:

$$NUTRITION_i = a + bCASH_i + e_i$$

Since cash treatment was randomly assigned, it appears to have successfully created comparable treatment and control groups, and we can interpret estimates of b as causal estimates of the impact of cash transfers on nutritional status. We will run 2 regressions: the first on the cash transfer's effect on household nutrition, and the second will run the same regression but restrict the dataset to only those households with no education past secondary.

```
reg_nutrition <- lm(nutrition ~ cash, data=gd)
reg_s0_nutrition <- lm(nutrition ~ cash, data = subset(gd,schooling == 0))

stargazer(reg_nutrition, reg_s0_nutrition,
          out="Table 2",
          title="Treatment Effect of Cash on Nutrition",
          dep.var.labels="Nutrition",
          covariate.labels=c("Cash Treatment"),
          type="latex",
          table.placement="!h",
          header=FALSE,
          align=TRUE,
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),
          no.space=TRUE)
```

Table 2: Treatment Effect of Cash on Nutrition

	Dependent variable:	
	Nutrition	
	(1)	(2)
Cash Treatment	0.102**	0.117***
	(0.043)	(0.044)
Constant	0.015	−0.007
	(0.030)	(0.031)
Observations	2,000	1,890
Note:	*p<0.1; **p<0.05; ***p<0.01	

Based on the regression output, we estimate that receiving a large cash transfer approximately two years earlier *improves* respondents' nutritional status by 0.102 units of the nutritional status index with a standard error of 0.043. (Note that since the nutritional status index is standardized to have mean equal to 0 with standard deviation equal to 1 within the control group, a 0.102 unit increase corresponds to an increase of 0.102 of a standard deviation.) The t-statistic associated with this coefficient is b/se(b)=0.102/0.043=2.37. Since this is greater than 1.96 (the appropriate cutoff for the 5% significance/95% confidence level), we can say that the estimated treatment effect is significantly different from zero at 5% significance/with 95% confidence (in other words, we reject the null hypothesis of no treatment effect, since zero doesn't lie in the 95% confidence interval).

When we restrict attention only to those respondents who have not completed secondary school, receiving a large cash transfer approxiately two years earlier again *improves* recipients' nutritional status by 0.117 units of the nutritional status index (or 0.117 of a standard deviation), with a standard error of 0.044. The t-statistic associated with this coefficient is b/se(b)=0.117/0.044=2.66. Again, since this is greater than 1.96 (the appropriate cutoff for the 5% significance/95% confidence level), we can say that this positive treatment effect estimate is significantly different from zero at 5% significance/with 95% confidence.

Ex ante, we might expect the cash transfer to have different impacts among the general population (represented by the full, unrestricted sample) compared to the subset that has not completed secondary school. On the one hand, if someone has completed secondary school, they may be better able to make use of

a large cash transfer. For example, if those with secondary education are more likely to operate a small business, receiving a cash transfer might provide extra resources that could be invested in capital or other business inputs, in turn increasing incomes and ability to consume more or higher quality nutrition. Alternatively, it could be that those who have completed secondary school may be more financially literate or more well-informed about the importance of nutrition for improving health and other outcomes, and so make decisions after receiving the cash that are directed towards improving their nutritional outcomes. On the other hand, if someone has not completed secondary school, they might have a lower income, be more cash-constrained, and have a lower nutritional index to start with. If there are higher returns to additional investments/consumption decisions that would improve nutritional outcomes at relatively low levels of nutrition, there might be higher returns to initially low-education/low-nutrition households. (Think back to earlier in the course when we discussed that low-income countries could potentially have higher returns to capital given initially lower levels of capital relative to high-income counterparts; in the same way, individuals with relatively lower nutritional status to start may have higher returns to any interventions (such as a cash transfer) that enable them invest/consume in such a way as to boost their nutritional outcomes.)

The results from our regressions indicate that the cash transfer treatment did have a slightly greater impact among those households where the primary respondent has not completed secondary school (comparing 1.117 to 0.102), though this represents a relatively small increase. That said, notice that restricting attention to those households that have not completed secondary school only reduces the number of observations from 2000 to 1890. In other words, the difference in the estimates we see is driven largely by exclusion of only 110 households, so we can expect that the effect would have therefore been substantially smaller among those 110 households.

Now, we re-run the two regressions above but include `male` and `age35` as additional explanatory variables. The regression of interest follows the form:

$$NUTRITION_i = a + bCASH_i + cMALE_i + dAGE35_i + e_i$$

```
reg_nutrition_more <- lm(nutrition ~ cash + male + age35, data=gd)
reg_s0_nutrition_more <- lm(nutrition ~ cash + male + age35, data = subset(gd,schooling == 0))

stargazer(reg_nutrition_more, reg_s0_nutrition_more,
          out="Table 3",
          title="Treatment Effect of Cash on Nutrition",
          dep.var.labels="Nutrition",
          covariate.labels=c("Cash Treatment", "Male", "Age 35+"),
          type="latex",
          table.placement="!h",
          header=FALSE,
          align=TRUE,
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),
          no.space=TRUE)
```

In the first column, we see that when adding controls for gender and age (in other words, adding `male` and `age35` as covariates), the estimated treatment effect on the health outcomes among the general population decreases slightly from 0.102 to 0.099. The standard error on the estimated coefficient of 0.099 is 0.042, which indicates that the t-statistic is 0.099/0.042=2.36. Since this is greater than the cutoff of 1.96, we can say that this impact is significantly different from zero at the 5% significance/95% confidence level.

Comparing the case when we restrict attention to those households in which the respondent has not completed secondary schooling, we see that the treatment effect of the cash transfers slightly decreases from 0.117 to 0.113 when we add in controls for the respondent being male or over age 35. The standard error on the estimated coefficient is 0.044, so the associated t-statistic is 0.113/0.044=2.57, so again significantly different from zero at the 5% significance/95% confidence level.

What do we learn from this exercise? First of all, the inclusion of controls does very little to our estimated treatment effects. In the full sample case, the estimated treatment effect declines from 0.102 to 0.099 and

Table 3: Treatment Effect of Cash on Nutrition

	Dependent variable:	
	Nutrition	
	(1)	(2)
Cash Treatment	0.099**	0.113**
	(0.042)	(0.044)
Male	0.119***	0.104**
	(0.046)	(0.048)
Age 35+	0.303***	0.301***
	(0.043)	(0.044)
Constant	−0.174***	−0.189***
	(0.040)	(0.041)
Observations	2,000	1,890
Note:	*p<0.1; **p<0.05; ***p<0.01	

remains significant; similarly, when we look at those households where respondents have not completed secondary school, the estimated treatment effect declines from 0.117 to 0.113 and remains significant.

The results related to the treatment effect after controlling for gender and age *are* what we would expect: Since the research design (randomization across villages) ensured that treatment and control households had similar baseline values for male heads of households and heads of households over age 35, we should not expect that adding these variables as covariates would impact the results. In other words, it is not surprising that adding these controls to the regressions doesn't modify our conclusions tremendously, because we showed in part (b) that the treatment and control groups were well balanced along these observable dimensions.

Note that in both regressions, the coefficient on `male` is positive and significant (0.119 in the full sample and 0.104 in the subset of the sample that has not completed secondary school). Similarly in both regressions, the coefficient on `age35` is positive and significant (0.303 in the full sample and 0.301 in the subset of the sample that completed secondary school). What does this tell us? this simply tells us that households headed by a male respondent and/or those with respondent above age 35 have better nutritional outcomes on average. While these findings are interesting in their own right, these coefficients describe **correlational** relationships, not **causal** relationships.

Section 2: Household Income and Child Health (Treatment Unrandomly Assigned)

As a thought experiment, let's supposed how we would go about this same analysis if treatment was not randomly assigned per household. Imagine, instead, that we only have data from the treatment villages and not any of the control villages. In the treatment villages, roughly one third of households were eligible for transfers. Since the cash transfers were not randomized between households within treatment villages, and were targeted at relatively poor households who are in all likelihood different from the relatively rich in many ways besides income. Relatively poor households may have better or worse nutrition for many reasons unrelated to the cash treatment, which invites potential ommitted variable bias. Having baselin data for both groups allows us to get a sense of how different the treatment and control groups were initially. In addition, we assume that, in the absence of treatment, the average difference in nutritional outcomes between treatment and control group would have stayed constant (otherwise known as the "parallel trends" assumption), then the change in the difference between treatment and control groups from baseline to endline would give us a valid estimate of the causal effect of treatment. This differencing within differencing mitigates the issue of potential ommitted variable bias. We will estimate the imapct of the GD cash transfer on nutrition using data only from eligible and in eligible households in the treatment villages, from both the baseline and the endline surveys. This is also known as the "diff-in-diff" approach.

Let's first load and look at the data for this section:

Based on a quick look at the summary statistics, we can see that this dataset contains 4000 observations from treatment villages (`cash` == 1), where as expected, 50% of observations are at endline (`time` == 1) and 50% are at baseline (`time` == 0). Moreover, 50% of households are eligible and were treated at endline (`time` == 1). Roughly 25% of household respondents are male (`male`==1), 67% are over age 35 (`age35`==1), and only 6% have completed secondary school (`schooling`==1).

Assessing balance of covariates at baseline

To start, we want to estimate the following equation:

$$Y_{i0} = a + b \cdot ELIGIBLE_i + e_{i0}$$

where Y_{i0} is the characteristic of interest at baseline (`male, age25, schooling`), and $ELIGIBLE_i$ takes on value 1 if the household was eligible for cash transfers. In this equation, b describes the average difference in Y between eligible and ineligible households at baseline, and a gives us the average value of Y for ineligible households at baseline.

Below is the code to run the three regressions. Using stargazer, we can combine the output from the three regressions run so far into one table.

```
dd_male <- lm(male ~ eligible, data = subset(dd, time==0))
dd_age35 <- lm(age35 ~ eligible, data = subset(dd, time==0))
dd_schooling <- lm(schooling ~ eligible, data = subset(dd, time==0))
```

```
stargazer(dd_male,dd_age35,dd_schooling,
          out="Table 4",type="latex",header=FALSE,
          title="Baseline Characteristics: Eligible vs Ineligible",
          dep.var.labels=c("Male", "Age 35+", "Sec School"),
          covariate.labels=c("Eligible"),align=TRUE,
          report = "vc*st",
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),no.space=TRUE)
```

The first column indicates that 20.0% of ineligible household respondents are male, and among eligible households, the value is 10.1% higher. The second column indicates that 81.7% of respondents in ineligible households are above age 35, while on average, the percent of eligible households with respondents above

Table 4: Baseline Characteristics: Eligible vs Ineligible

	Dependent variable:		
	Male	Age 35+	Sec School
	(1)	(2)	(3)
Eligible	0.101***	−0.300***	0.000
	(0.019)	(0.020)	(0.011)
	$t = 5.245$	$t = -15.007$	$t = 0.000$
Constant	0.200***	0.817***	0.061***
	(0.014)	(0.014)	(0.008)
	$t = 14.689$	$t = 57.798$	$t = 8.056$
Observations	2,000	2,000	2,000
Note:			*p<0.1; **p<0.05; ***p<0.01

age 35 is 30.0 percentage points lower. The third column indicates that 6.1% of respondents in ineligible households have completed secondary school, and that this value does not change for eligible households.

Standard errors are reported in parentheses below the estimated coefficients. Recall that standard errors help to answer the question of how precisely estimated each coefficient is, and whether that coefficient is or is not statistically distinguishable from zero. In this context, if a coefficient is statistically indistinguishable from zero, that would indicate that eligible and ineligible households are similar at baseline in terms of the characteristic included as the outcome of each regression.

A comparison of the coefficient estimate and the standard error allows us to compute a t-statistic for each coefficient, which in turns allows us to determine whether or not the coefficient is statistically distiguishable from zero at a particular level of significance, usually 1%, 5%, or 10% (corresponding to 99%, 95%, and 90% confidence intervals). Let's say we are interested in determining whether coefficients are statistically indistinguishable from zero at the 5% level, as is a common standard in economics. If the absolute value of the t-statistic we compute is less than 1.96, then we cannot reject that the coefficients is statistically indistinguishable from zero at the 5% level. In other words, if the absolute value of the t-statistic we compute is less than 1.96, then we can be reasonably certain that the coefficient is not different than zero, but instead, we can conclude that eligible and ineligible households "balanced" for that particular variable. For the 10% and 1% significance levels, the critical values for the t-statistics are 1.65 and 2.58 respectively.

The formula we will use is: t=b/se(b). For the variable `male`, the t-statistic is 0.101/0.019= 5.32, which is larger than 2.58 in absolute value. That is, we can reject the hypothesis that the share of respondents that are male is the same in eligible and ineligible households at baseline at the 1%, 5% and 10% significance levels. For the variable `age35`, the t-statistic is -0.300/0.020=-15.0, which bigger than 2.58 in absolute value, so we can reject that the hypothesis that eligible and ineligible households are the same at the 1%, 5% and 10% significance levels. For the variable `schooling`, the t-statistic is 0.000/0.011=0.00, which smaller than 1.65 in absolute value, so we cannot reject that eligible and ineligible households have the same proportion of respondents who completed secondary school at the 10% significance level. (Note: You don't always have to calculate the t-statistics or significance tests by yourself. By using the "report = "c*st"" option in stargazer, we told stargazer to report coefficients (with significance stars), standard errors and t-statistics.)

In summary, eligible households are statistically significantly younger, and more likely to be male. However, they are similar in average proportion of respondents that have completed secondary school. It is not surprising that those two groups are not balanced, since being eligible for a cash transfer within a village was not randomly chosen. Instead, cash transfers were targeted towards the relatively poor. While the patterns may appear surprising at first, it does make sense that households headed by relatively younger individuals would be relatively poor (think of the earnings profile going up over time). Moreover, young male-headed households may represent young males who have not yet married, and who are just starting out in their careers, and haven't had the time or experience yet to build up a more stable economic existence and

therefore more likely to be eligible – although this is more speculative, and we would need more evidence to investigate whether this holds true.

Next, we will determine the average difference between the eligible and ineligible households in terms of their nutritional status ("nutrition") in both the baseline survey and the endline survey round.

The regressions of interest for baseline and endline are:

$$Y_{i0} = a + b \cdot ELIGIBLE_i + c \cdot MALE_i + d \cdot AGE35_i + f \cdot SCHOOLING_i + e_{i0}$$
$$Y_{i1} = a + b \cdot ELIGIBLE_i + c \cdot MALE_i + d \cdot AGE35_i + f \cdot SCHOOLING_i + e_{i1}$$

The code for running the relevant regressions and reporting the results in a table follows:

```
dd_baseline <- lm(nutrition ~ eligible + male + age35 + schooling, data = subset(dd,time == 0))
dd_endline <- lm(nutrition ~ eligible + male + age35 + schooling, data = subset(dd, time == 1))

stargazer(dd_baseline,dd_endline,
          out="Table 5",type="latex",header=FALSE, table.placement = "h!",
          dep.var.labels=c("Nutrition"),
          covariate.labels=c("Eligible", "Male", "Age 35+", "Sec School"),
          column.labels = c("Baseline", "Endline"),
          title="Nutritional Outcomes: Eligible vs Ineligible",align=TRUE,
          report = "vc*st",
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),no.space=TRUE)
```

Table 5: Nutritional Outcomes: Eligible vs Ineligible

	Dependent variable:	
	Nutrition	
	Baseline	Endline
	(1)	(2)
Eligible	−0.100**	0.008
	(0.046)	(0.043)
	$t = -2.189$	$t = 0.193$
Male	0.078	0.032
	(0.051)	(0.047)
	$t = 1.548$	$t = 0.682$
Age 35+	−0.237***	−0.172***
	(0.048)	(0.045)
	$t = -4.906$	$t = -3.838$
Sec School	0.148	0.260***
	(0.091)	(0.084)
	$t = 1.637$	$t = 3.093$
Constant	0.213***	0.170***
	(0.050)	(0.047)
	$t = 4.222$	$t = 3.635$
Observations	2,000	2,000
Note:	*p<0.1; **p<0.05; ***p<0.01	

In column (1), we see that, controlling for the respondent's gender, age and schooling, eligible households have statistically significantly lower nutritional outcomes than ineligible households at baseline. The difference is -0.1 points of the normalized nutrition index (corresponding to a 0.1 standard deviations). In addition, we find that, at baseline, male-headed households have slightly higher nutrition scores, though the difference is

not statistically significant. Being over 35 years of age is associated with respondents' having a statistically significantly lower nutrition status by 0.24 standard deviations on average. Having completed secondary school is associated with a statistically insignificant 0.15 standard deviation higher nutrition status index on average.

In column (2), we see that, controlling for the respondent's gender, age and schooling, eligible households now have *better* nutrition than ineligible households on average, but the difference is not statistically significant. The patterns for the other covariates remain similiar: Male households have statistically insignificantly higher nutritional status than female households. Younger households have statistically significantly better nutritional outcomes. And, completing secondary school is associated with better nutrition, though that coefficient now becomes statistically significant at the 1% level.

Overall, it makes intuitive sense that eligible households would have lower nutrition scores at baseline, since cash transfers targeted poorer households. Moreover, it is not surprising that, controlling for the other included variables, older, less educated and female households tend to have lower nutritional outcomes on average.

Next, we will consider a difference-in-differences analysis that uses data from both time periods simultaneously. In order to carry out this analysis, we will need to construct a new treatment variable that is the interaction of the eligible indicator variable and the time variable. Since this variable does not exist in the data, we will create the new variable.

```
dd$treat <- dd$eligible * dd$time
```

We will carry out the standard diff-in-diff regression, including "male", "age35" and "schooling" as additional explanatory variables. The regression of interest is as follows:

$$Y_{it} = a + \alpha \cdot ELIGIBLE_i + \beta \cdot TIME_t + \gamma \cdot (ELIGIBLE_i \cdot TIME_T) + c \cdot MALE_i + d \cdot AGE35_i + f \cdot SCHOOLING_i + e_i$$

```
dd_diffindiff <- lm(nutrition ~ eligible + time + treat
                    + male + age35 + schooling,
                    data=dd)
stargazer(dd_diffindiff,
          out="Table 6",type="latex",header=FALSE, table.placement = "h!",
          title="Nutritional Outcomes: Diff in Diff",
          dep.var.labels=c("Nutrition"),
          covariate.labels=c("Eligible", "Time", "Treat", "Male", "Age 35+", "Sec School"),
          align=TRUE,report = "vc*st",
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),no.space=TRUE)
```

Before we jump to our interpretation: There may be reasons why we think the difference-in-difference estimator is not a good estimator of the true causal effect in this case. For our causal interpretation to be valid, we need to be confident in our parallel trends assumption. Would average differences in nutritional status for eligible and ineligible households have remained stable between baseline and endline in the absence of treatment? There are many ways to answer this question, and we can never know for sure. However, we know that eligibles at baseline are, on average, younger than ineligibles (see question 2a): If average nutritional values rise faster for young households than for older households, maybe because their earnings rise faster as they become more experienced, we would expect the difference or gap in nutrition between eligibles and ineligibles to decline over time in the absence of treatment. In other words, eligible households might have caught up with ineligibles over time even in the absence of receiving a cash transfer. Our DD estimator would therefore *overestimate* the causal effect of treatment (draw a graph).

With this caveat in mind, our difference-in-differences estimator of the causal effect of cash transfers on nutrition is 0.084, implying that getting a cash transfer increases the nutrition index by 0.08 standard deviations on average. This estimate, however, is not statistically different from zero at conventional confidence levels. Thus, we cannot reject the hypothesis that receiving a cash transfer has no effect on nutritional outcomes after about 2 years.

Table 6: Nutritional Outcomes: Diff in Diff

	Dependent variable:
	Nutrition
Eligible	-0.088^{**}
	(0.043)
	$t = -2.059$
Time	0.008
	(0.042)
	$t = 0.190$
Treat	0.084
	(0.059)
	$t = 1.439$
Male	0.055
	(0.034)
	$t = 1.599$
Age 35+	-0.205^{***}
	(0.033)
	$t = -6.207$
Sec School	0.204^{***}
	(0.062)
	$t = 3.305$
Constant	0.188^{***}
	(0.040)
	$t = 4.670$
Observations	4,000
Note:	*p<0.1; **p<0.05; ***p<0.01

Section 3: GDP's Effect on Civil Conflict (Instrumental Variable Approach)

Briefly describe the motivation, the data, econometric approach, and main empirical findings of the MSS (2004) article.

Motivation: Miguel, Satyanath, and Sergenti (2004) build on the body of literature that seeks to understand the link between economic outcomes and armed civil conflict, focusing in particular on Sub-Saharan Africa in the late 20th century. The presence of civil conflict in this region during this period was substantial. MSS (2004) cite that in Sub-Saharan Africa "29 of 43 countries suffered from civil conflict during the 1980s and 1990s" and "in the median sub-Saharan African country, hundreds of thousands of people were displaced from their homes as a consequence of civil war during this period." Answering the question of how economic outcomes causally impact civil conflict, however, is challenging as a result of endogeneity issues: there could be reverse causality (where the presence of armed conflict influences economic outcomes directly) or there could be many omitted factors (for example, the quality of institutions) that influence both economic outcomes and the likelihood of armed civil conflict.

Econometric Approach: MSS (2004) make a unique contribution to the literature by using an instrumental variables strategy that allows them to address the endogeneity issue present in other research on the topic. Their instrumental variables strategy relies on using exogenous variation in rainfall as an instrument for GDP growth to estimate the *causal* effect of GDP growth on the incidence of armed conflict. For this strategy to be valid, there first needs to be a strong link

between changes in rainfall patterns and GDP growth. In this context, this is the case, as many SSA economies are predominantly agricultural, and irrigation methods are less widely applied than in developed countries (first stage). Second, the rainfall should not be directly affected by conflict (feedback loop), and should be uncorrelated with factors other than GDP growth which also affect conflict (exogeneity). Third, the only channel through which rainfall patterns impact armed conflict is through GDP growth; that is, rainfall does not impact armed conflict directly or through other channels (exclusion restriction).

Data: (1) Armed Conflict: Data on the presence of armed conflict comes from the Armed Conflict Data database produced by the International Peace Research Institute of Oslo, Norway, and the University of Uppsala, Sweden. This dataset focuses on politically-motivated violence (notably not capturing other forms of violence) at the national and annual level. The data captures all conflicts that result in at least 25 deaths and an indicator for those resulting in at least 1000 deaths. MSS focus on within-country conflict. (2) GDP Growth: GDP Growth measures come from the Penn World Tables and the World Bank. (3) Rainfall: Rainfall data comes from the Global Precipitation Climatology Project (GPCP), which records rainfall data at 2.5-degree intervals of latitude and longitude. MSS aggregate these up to country-year averages, and compute for each country a measure of the proportional change in rainfall relative to the prior year as $CHANGEinR_t = \frac{R_t - R_{t-1}}{R_{t-1}}$. (4) Other Country-Level Data: Covariates such as ethnic and religious fractionalization, poulation, democracy, terrain, etc. come from a variety of World Bank and other data bases. Excellent answers would discuss the reliability of those data, and likely sources of potential measurement error.

Main Findings: Using an Instrumental Variables Two Stage Least Squares (IV-2SLS) estimation strategy including country fixed effects and country-specific time trends (as per column (6) of Table 4), MSS find that a one percentage point increase in GDP decreases the probability of civil conflict by 2.55 percentage points. Put differently, a five percentage point decline in GDP leads to an over 12 percentage point increase in the likelihood of armed conflict. Excellent answers discussed whether those estimates appear big / small and why.

```
## Load in the data
mss <- read.csv("mss.csv")

## Create summary statistics and distributions
stargazer(mss,out="Table 7",
          title="Summary Statistics",
          type="text",header=FALSE)
```

Summary Statistics

Statistic N Mean St. Dev. Min Pctl(25) Pctl(75) Max

country_id 639 497.757 53.536 404 450 546 625 year 639 1,990.798 4.863 1,983 1,987 1,995 1,999 green_index_growth 639 0.009 0.094 -0.468 -0.037 0.047 0.655 gdp_growth 639 -0.006 0.065 -0.474 -0.035 0.024 0.319 democracy 639 -3.354 5.498 -10 -7 0 9

```
## Create histogram of NDVI measure (green index growth)
png("Histogram Green.png")
hist(mss$green_index_growth,breaks=20,
     xlab="Proportional change in the green index relative to last year",
     main="Distribution of Green Index Growth Measure")
dev.off()
```

pdf 2

```
## Create histogram of GDP growth
png("Histogram GDP.png")
hist(mss$gdp,breaks=20,
     xlab="Log GDP growth since last year",
     main="Distribution of GDP growth rates")
dev.off()
```

pdf 2

From this exercise, we are able to confirm the following about the variables at hand: (1) democracy is an integer variable taking on values between -10 and +9 in our data, (2) NDVI (green index growth) captures the proportional change in the average value of green vegetation for a particular country and year relative to that of the previous year, and (3) gdp growth is measured in log changes relative to the previous year (i.e. a value of 0.05 corresponds roughly to a growth rate of 5%). There are 639 observations for 40 countries over the course of 17 years (from 1983 to 1999).

The first stage, second stage, and reduced form regressions with year fixed effects are as follows:

The first stage regression is given by:

$$GDPGROWTH_{ct} = \pi_0 + \pi_1 NDVI_{ct} + \pi_2 YEAR_t + u_{ct}$$

The second stage regression (or structural equation) is given by:

$$DEMOCRACY_{ct} = \beta_0 + \beta_1 GDPGROWTH_{ct} + \beta_2 YEAR_t + \varepsilon_{ct}$$

The reduced form regression is given by:

$$DEMOCRACY_{ct} = \gamma_0 + \gamma_1 NDVI_{ct} + \gamma_2 YEAR_t + \nu_{ct}$$

In the above regressions, GDPGROWTH, DEMOCRACY, and NDVI are as defined above, where YEAR is included as a linear control, and the unit of observation is at the country-year level (represented by the ct indices).

Optional material: You may have also added in year fixed effects instead of adding year as a linear control, in which case your set of regressions would be:

- First Stage: $GDPGROWTH_{ct} = \pi_0 + \pi_1 NDVI_{ct} + \delta_{1t} + u_{ct}$
- Second Stage: $DEMOCRACY_{ct} = \beta_0 + \beta_1 GDPGROWTH_{ct} + \delta_{2t} + \varepsilon_{ct}$
- Reduced Form: $DEMOCRACY_{ct} = \gamma_0 + \gamma_1 NDVI_{ct} + \delta_{3t} + \nu_{ct}$

where δ_{1t}, δ_{2t}, and δ_{3t} represent year fixed effects. Note: The linear control assumes a model where the outcome variable changes linearly over time. For example, in the first stage, GDP growth would increase or decrease linearly over time: If π_2 is 0.01, this implies that the growth rate of GDP increases by 1% each year. This seems an unrealistic model for GDP growth. Year fixed effects, on the other hand, allow the average of the outcome variable across all countries to vary for each specific year. In other words, adding year fixed effects controls for factors that affect the outcome variable in the same way across all countries in a given year. In the first stage equation, for example, GDP growth is allowed to have a different average value across all countries in each year, so factors such as global recessions are absorbed. This seems much more reasonable. How would you interpret the coefficients on year fixed effects? For example, in the first stage, $\pi_0 + \delta_{1,1990}$ identifies the average growth rate across all countries in 1990 conditional on NDVI not changing since last year, $\pi_0 + \delta_{1,1991}$ identifies the average growth rate across all countries in 1991 conditional on NDVI not changing since last year, and so on.

We can plug the first stage into the second stage regression (or structural equation) to get:

$$DEMOCRACY_{ct} = \beta_0 + \beta_1 GDPGROWTH_{ct} + \beta_2 YEAR_t + \varepsilon_{ct}$$
$$DEMOCRACY_{ct} = \beta_0 + \beta_1\Big(\pi_0 + \pi_1 NDVI_{ct} + \pi_2 YEAR_t + u_{ct}\Big) + \beta_2 YEAR_t + \varepsilon_{ct}$$
$$DEMOCRACY_{ct} = \beta_0 + \beta_1\pi_0 + \beta_1\pi_1 NDVI_{ct} + \beta_1\pi_2 YEAR_t + \beta_2 YEAR_t + \beta_1 u_{ct} + \varepsilon_{ct}$$
$$DEMOCRACY_{ct} = \underbrace{\beta_0 + \beta_1\pi_0}_{\gamma_0} + \underbrace{\beta_1\pi_1}_{\gamma_1} NDVI_{ct} + \underbrace{(\beta_1\pi_2 + \beta_2)}_{\gamma_2} YEAR_t + \underbrace{\beta_1 u_{ct} + \varepsilon_{ct}}_{\nu_{ct}}$$

Since $\gamma_1 = \beta_1\pi_1$, the coefficient of interest β_1 can be estimated using an indirect least squares (ILS) procedure as $\hat{\beta}_1^{ILS} = \frac{\hat{\gamma}_1}{\hat{\pi}_1}$ (where we write $\hat{\beta}_1^{ILS}$ to emphasize that this is an ILS estimate). Note that this since we have one endogenous regressor (GDP growth) and one instrument (NDVI measure), the coefficient of interest can be estimated in this way. More generally, a two-stage least squares (2SLS) estimation method can be used to compute IV estimates, even in the case of more than one endogenous variable and one or more instruments for each of those endogenous variables. In the case of one endogenous regressor and one instrument, indirect least squares and two stage least squares estimates are equivalent, that is, $\beta_1^{ILS} = \beta_1^{2SLS}$.

First, we produce the second stage regression (structural equation) by regressing our democracy measure on GDP growth.

```
reg1 <- lm(democracy ~ gdp_growth + year, data=mss)
mss$year.f <- factor(mss$year)
reg1_yearfe <- lm(democracy ~ gdp_growth + year.f, data=mss)

stargazer(reg1, reg1_yearfe,
          out="Table 8", type="latex",
          header=FALSE,title="Structural Equation",
          dep.var.labels=c("Democracy"),
          covariate.labels=c("Log GDP Growth","Year"),
          align=TRUE,
          omit="year.f",
          report="vc*st"
          ,omit.stat=c("LL","ser","f","rsq","adj.rsq"),
          no.space=TRUE,
          notes = "Column (2) includes year fixed effects; coefficients not reported",
          notes.append = TRUE)
```

Based on the regression output, we can see that using a simple OLS estimation strategy including year as a linear control (as in column 1), annual GDP growth is negatively correlated with more democratic institutions. Specifically, the coefficient on annual GDP growth (β_1^{OLS}, where we include the superscript OLS to emphasize that β_1 was estimated using OLS) is -1.697 with an associated standard error of 2.995. The magnitude of this estimate implies that a 1 percentage point increase in GDP growth (corresponding to a increase in log GDP growth by 0.01) is associated with a decline in the democracy score of 0.017. The t-statistic is $t = \frac{\beta_1^{OLS}}{SE(\beta_1^{OLS})} = \frac{-1.697}{2.995} = -0.567$. Since $|-0.567| < 1.96$, we conclude that the OLS estimate of β_1^{OLS} is not statistically significantly different from zero. The estimated cofficient on the year variable positive, at 0.519 and statistically significant, indicating that the democracy score was increasing by 0.519 units on average each year across all countries. This makes sense, as Sub-Saharan African countries were democratising in the 1990s afer a period of more autocratic rule.

In the specification including year fixed effects (column 2), β_1^{OLS} is estimated to be -0.465 with a a standard error of 3.030, indicating significance at 95% confidence. Hereafter, in the interest of simplicity, we focus our attention and interpretation on the specification with the inclusion of year as a linear control rather than the specification including year fixed effects.

Table 7: Structural Equation

	Dependent variable:	
	Democracy	
	(1)	(2)
Log GDP Growth	−1.697	−0.465
	(2.995)	(3.030)
	$t = -0.567$	$t = -0.153$
Year	0.519***	
	(0.040)	
	$t = 12.929$	
Constant	−1,036.066***	−5.632***
	(79.880)	(0.777)
	$t = -12.970$	$t = -7.251$
Observations	639	639
Note:		*p<0.1; **p<0.05; ***p<0.01
	Column (2) includes year fixed effects; coefficients not reported	

Why don't we stop here? Why do we want to proceed with an instrumental variables analysis? We are interested in getting at the causal impact of GDP growth on democratization. Estimating the relationship using OLS, however, likely delivers a *biased* estimate of the effect. Why? First, reverse causality: it could be that democratization influences a variety of economic outcomes, among them annual GDP growth. Second, omitted variables: there are likely many other factors that simultaneously influence both economic conditions (such as annual GDP growth) and political outcomes (such as the extent of democracy). For example, movements towards democratization may be associated with political instability, or even violent conflict which also affect the economy. Either of these scenarios would lead to correlaton or covariance between GDP growth and the error term ε, leading OLS to deliver a biased estimate of β_1.

Next, we regress the GDP growth measure on the NDVI growth measure (`green_index_growth`) and the year variable. In an IV framework where NDVI growth will serve as an instrument for GDP growth, this regression is known as the first stage regression.

```
## First Stage: Regress GDP growth on the NDVI measure and year
reg2 <- lm(gdp_growth ~ green_index_growth + year, data=mss)
reg2_yearfe <- lm(gdp_growth ~ green_index_growth + year.f, data=mss)

stargazer(reg2, reg2_yearfe,
          out="Table 9", type="latex", header=FALSE,
          title="First Stage",dep.var.labels=c("Log GDP Growth"),
          covariate.labels=c("NDVI", "Year"),
          omit="year.f",
          align=TRUE,
          report="vc*st",
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),
          notes = "Column (2) includes year fixed effects; coefficients not reported",
          notes.append=TRUE,
          no.space=TRUE)
```

The regression output indicates that the NDVI measure is indeed strongly correlated with GDP growth. The coefficient on the NDVI measure (π_1) is 0.109, which indicates that a one-unit increase in the green growth index is associated with a 10.9 percentage point increase in GDP

Table 8: First Stage

	Dependent variable:	
	Log GDP Growth	
	(1)	(2)
NDVI	0.109***	0.077***
	(0.027)	(0.029)
	$t = 4.016$	$t = 2.639$
Year	0.001***	
	(0.001)	
	$t = 2.685$	
Constant	−2.796***	−0.033***
	(1.039)	(0.010)
	$t = -2.692$	$t = -3.290$
Observations	639	639
Note:		*p<0.1; **p<0.05; ***p<0.01
	Column (2) includes year fixed effects; coefficients not reported	

growth. (Note: Since the NDVI growth rate is proportional relative to last year, a one unit increase in NDVI growth implies a doubling of green coverage.) The associated standard error is 0.027, and so the t-statistic is $t = \frac{\pi_1}{SE(\pi_1)} = \frac{0.109}{0.027} = 4$. Since $|4| > 1.96$, we know that this estimate for π_1 is significant at 95%. The coefficient on year is statistically significantly positive. Each year, GDP growth increased by 0.1 percentage points on average across all countries. Over the period, growth in Sub-Saharan Africa was therefore accelerating, corresponding to what we saw in one of the first lectures.

Now we regress the democracy measure on the NDVI measure and the year variable. This is known as the reduced form regression.

```
## Reduced Form: Regress democracy score on the NDVI growth measure and year
reg3 <- lm(democracy ~ green_index_growth + year, data=mss)
reg3_yearfe <- lm(democracy ~ green_index_growth + year.f, data=mss)

stargazer(reg3, reg3_yearfe,
          out="Table 10", type="latex", header=FALSE,
          title="Reduced Form",dep.var.labels=c("Democracy"),
          covariate.labels=c("NDVI", "Year"),
          align=TRUE,
          omit="year.f",
          report="vc*st",
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),
          no.space=TRUE,
          notes = "Column (2) includes year fixed effects; coefficients not reported",
          notes.append=TRUE)
```

We can see from this reduced form analysis that the coefficient on the NDVI measure (γ_1) is 0.060. This indicates that a one-unit increase in NDVI measure is associated with a 0.06 unit increase in the democracy score. The associated standard error is 2.073 and the t-statistic is 0.029, indicating that this result not statistically significant at the 90% confidence level. Once again, the coefficient on the year control is statistically significantly positive (0.516), indicating that Sub-Saharan African countries were becoming more democratic over the study period.

Table 9: Reduced Form

	Dependent variable:	
	Democracy	
	(1)	(2)
NDVI	0.060	0.016
	(2.073)	(2.215)
	$t = 0.029$	$t = 0.007$
Year	0.516***	
	(0.040)	
	$t = 12.932$	
Constant	−1,030.986***	−5.615***
	(79.462)	(0.771)
	$t = -12.975$	$t = -7.283$
Observations	639	639
Note:		*p<0.1; **p<0.05; ***p<0.01
	Column (2) includes year fixed effects; coefficients not reported	

Finally, carry out the estimation of the instrumental variables (IV) model, in which the NDVI measure serves as an instrument for GDP growth. Carry out and report the IV estimation you laid out in part b., using the "ivreg" command from the "ivpack" package in R . Please include the year variable as a control. Remember that ivreg works using `ivreg(y ~ x + w1 | z + w1, x=TRUE, data=mydata)`, where y is the outcome of interest, x is the endogenous regressor, z is the instrument and w1 is another covariate. What is the implied effect of economic growth on democracy based on your analysis? Interpret both the magnitude of the effect and the statistical significance.

```
# IV Procedure: Regress democracy score on GDP growth, using the
#               NDVI measure as an instrument for GDP growth.
reg4 <- ivreg(democracy ~ gdp_growth + year |
                  green_index_growth + year, x=TRUE, data=mss)
reg4_yearfe <- ivreg(democracy ~ gdp_growth + year.f |
                         green_index_growth + year.f, x=TRUE, data=mss)

## Comparing OLS and IV Results side-by-side
stargazer(reg1, reg4, reg1_yearfe, reg4_yearfe,
          out="Table 11", type="latex", header=FALSE,
          title="OLS and IV Estimates",
          dep.var.labels=c("Democracy"),
          covariate.labels=c("Log GDP Growth", "Year"),
          column.labels = c("OLS", "IV","OLS","IV"),
          model.names = FALSE,align=TRUE,omit="year.f",report="vc*st",
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),no.space=TRUE,
          notes = "Columns (3) and (4) include year fixed effects; coefficients not reported",
          notes.append=TRUE)
```

Column 1 reports estimation results of the second stage or structural equation regression using OLS (reproduced from question 1c above), and column 2 reports estimation results of the second stage regression or structural equation using IV. We can see that while $\beta_1^{OLS} = -1.697$, $\beta_1^{IV} = 0.553$. The IV estimate implies that a 1 percentage point increase in GDP growth is associated with a 0.006 unit increase in the democracy score. Neither the OLS nor IV coefficient are statistically significantly different from zero. We cannot reject the hypothesis that economic growth has no impact on democratization.

Table 10: OLS and IV Estimates

	Dependent variable:			
	Democracy			
	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
Log GDP Growth	−1.697	0.553	−0.465	0.208
	(2.995)	(19.054)	(3.030)	(28.770)
	$t = -0.567$	$t = 0.029$	$t = -0.153$	$t = 0.007$
Year	0.519***	0.515***		
	(0.040)	(0.049)		
	$t = 12.929$	$t = 10.559$		
Constant	−1,036.066***	−1,029.440***	−5.632***	−5.608***
	(79.880)	(97.249)	(0.777)	(1.272)
	$t = -12.970$	$t = -10.586$	$t = -7.251$	$t = -4.410$
Observations	639	639	639	639

Note: *p<0.1; **p<0.05; ***p<0.01
Columns (3) and (4) include year fixed effects; coefficients not reported

An important caveat to for our analysis is that our data is yearly. Thus, we find no effect of GDP growth on democratization in the same year.

Lastly, we re-evaluate the validity of the instruments. Instrument validity relies on the existence of a strong first stage, satisfaction of exogeneity, and satisfaction of the exclusion restriction. Answers to this question may vary. We list below some considerations related to each of these assumptions in turn.

- First Stage: There seems to be a strong first stage, where the NDVI measure strongly predicts GDP growth changes. The first stage results support that when green index growth is high (indicative of good rainfall and good agricultural conditions), GDP growth is high. In contrast, when green index growth is low (indicative of poor rainfall and poor agricultural conditions), GDP growth is low. This is precisely what we would expect to find for economies with a large share of production in agriculture and low prevalence of artificial irrigation methods, which reflects a strong first stage. This is also supported by our results. Including diagnostics in the summary of our IV regression, we can see that the F-stat associated with a test for weak instruments is 16, which exceeds the benchmark of 10.

```
## Test for weak instruments
reg4_diagnostics <- summary(reg4, diagnostics = TRUE)$diagnostics
reg4_diagnostics
```

```
##                  df1 df2   statistic      p-value
## Weak instruments   1 636 16.12846841 6.625619e-05
## Wu-Hausman         1 635  0.01428832 9.048903e-01
## Sargan             0  NA          NA           NA
```

- Exogeneity: First, there is likely no feedback loop from the democracy score to the green index growth measure. Democratization most likely has no direct impact on rainfall and the green index measure. One violation to this would be if democratization leads directly to a change in vegetated areas (e.g. democratic leaders decide to plant trees against climate change or redistribute the land to small-scale farmers who plant more crops), which would make the green index measure not quite exogenous. Second, we need that changes in the green index growth measure are uncorrelated with omitted variables or factors other than GDP growth that influence democratization in Sub-Saharan African countries. This

correlation could be causal (as with violations of the exlusion restriction below), or merely correlational. For example, rainfall may be correlated with other climate-related phenomena associated with higher/lower green index growth. We have seen in lecture, for instance, that higher temperatures are associated with more aggressive behavior of individuals in laboratory conditions. So, if less rainfall is correlated with higher temperatures, and higher temperatures lead to more violence, which in turn affects democratic movements (positively or negatively) other than through GDP growth, then temperature could be an omitted variable that confounds the effect of rainfall on conflict.

- Exclusion Restriction: The exclusion restriction requires that GDP growth is the *only* channel through which green index growth influences democratization. As usual, we cannot directly show that the exclusion restriction holds. One potential violation of the exclusion restriction would be if green index growth directly influences democratization, independent of GDP growth. For instance, more rainfall (and hence, higher green index growth) may lead people to stay inside more, protest less, and therefore reduce the power of democratic movements.

Part h) [1 point]

In your opinion, which estimates are more informative, those from part e. or those from part f., and why?

Answers to this question will vary. We will look for an answer that indicates you understand the distinction between the reduced form regression from part (e) and the second stage IV regression from part (f) and includes some discussion of what parameter estimates from each regression mean, what might be the strengths or limitations of each regression, etc.

The regression estimates from part (e) show the *reduced form* relationship between green index growth and democracy. Even if you believe the exclusion restriction could potentially be violated, this regression (and the estimated parameter values) can still prove informative in that it depicts the *reduced form* relationship between green index growth and democracy without necessarily taking a stand on the particular causal channels underlying that relationship. Put differently, γ_1 captures the relationship between green index growth and democracy which may be of interest in and of itself, but stops short of asserting that GDP growth is the only specific causal channel driving this relationship or estimating the strength of GDP growth as a causal channel.

The regression estimates from part (f) show the *causal* impact of GDP growth on democratization, provided that the exclusion restriction is satisfied. If your primary relationship of interest is to pin down the causal relationship between GDP growth and democratization *and* you believe the exclusion restriction holds, then you may consider the regression estimates from part (f) (and the estimated parameter values) to be more informative. Put differently, if the green index growth is a valid instrument (that is, satisfies the three conditions for a valid instrument), then β_1^{IV} gives an unbiased estimate of the causal impact of GDP growth on democratization, allowing you to estimate the nature and strength of the causal relationship of interest.

Full Code

```
################################
## Health Outcomes in West Africa
################################

## Install packages: only run once to install
#install.packages("stargazer")
#install.packages("ivpack")

## Set working directory and load packages
setwd("/home/rstudio/westafrica")
library(stargazer)
library(ivpack)

## Load in data in csv form
gd <- read.csv("gd.csv") #Section 1

#Section 1
## Take a look at the data and present summary statistics for selected variables
dim(gd)
summary(gd)
sd(gd$nutrition)

## b) Balance check: Determine average difference across cash and control
## households in terms of the following characteristics:
##  (1) Male (2) Age 35 or older (3) Secondary schooling completed
# univariate regression of maleness measure on cash infusion
reg_male <- lm(male ~ cash, data=gd)
# univariate regression of age on cash infusion
reg_age35 <- lm(age35 ~ cash, data=gd)
# univariate regression of secondary school completion on cash infusion
reg_schooling <- lm(schooling ~ cash, data=gd)

summary(reg_male)
summary(reg_age35)
summary(reg_schooling)

## Repeat balance checks, restricting attention to those who have not
## completed secondary schooling
# univariate regression of maleness measure on cash infusion where the
# respondent has not completed secondary school
reg_male_noschooling <- lm(male ~ cash, data=subset(gd, schooling==0))
# univariate regression of maleness measure on cash infusion where the
# respondent has not completed secondary school
reg_age35_noschooling <- lm(age35 ~ cash, data=subset(gd,schooling==0))
summary(reg_male_noschooling)
summary(reg_age35_noschooling)

# Display results
stargazer(reg_male, reg_age35, reg_schooling,
          reg_male_noschooling, reg_age35_noschooling,
          out="Table 1",
```

```
          title="Baseline Covariate Analysis",
          dep.var.labels=c("Male", "Age 35+", "Sec School", "Male", "Age 35+"),
          covariate.labels=c("Cash Treatment"),
          align=TRUE,
          type="text",
          table.placement = "!h",
          header=FALSE,
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),
          no.space=TRUE)

## c) Treatment Effect: Determine average difference between treatment
## and control households in nutritional status at endline survey
# univariate regression of nutrition on cash infusion
reg_nutrition <- lm(nutrition ~ cash, data=gd)
#univariate regression of nutrition on cash infusion where the respondent
#has not completed secondary school
reg_s0_nutrition <- lm(nutrition ~ cash, data = subset(gd,schooling == 0))
## Display results
stargazer(reg_nutrition, reg_s0_nutrition,
          out="Table 2",
          title="Treatment Effect of Cash on Nutrition",
          dep.var.labels="Nutrition",
          covariate.labels=c("Cash Treatment"),
          type="text",
          table.placement="!h",
          header=FALSE,
          align=TRUE,
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),
          no.space=TRUE)

## d) Treatment Effect: Repeat treatment effect analysis, including
## controls for male and age 35 or older
#multivariate regression of nutrition on cash infusion, maleness, and age
reg_nutrition_more <- lm(nutrition ~ cash + male + age35, data=gd)
#multivariate regression of nutrition on cash infusion, maleness, and age
#where the respondent has not completed secondary school
reg_s0_nutrition_more <- lm(nutrition ~ cash + male + age35, data = subset(gd,schooling == 0))

## Display results
stargazer(reg_nutrition_more, reg_s0_nutrition_more,
          out="Table 3",
          title="Treatment Effect of Cash on Nutrition",
          dep.var.labels="Nutrition",
          covariate.labels=c("Cash Treatment", "Male", "Age 35+"),
          type="text",
          table.placement="!h",
          header=FALSE,
          align=TRUE,
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),
          no.space=TRUE)

#Section 2
```

```r
# Load data in csv form
# In section 1, we know the villages in the GD cash transfer program were randomly chosen.
# Another experiment was run where instead, we only have data from the treatment villages and not any o
# In the treatment villages, roughly one third of households - typically relatively poor households
# Were eligible for cash transfers
# Estimate the impact of GD cash transfers on nutrition using data on eligible and ineligible household
dd <- read.csv("dd.csv")


## Average difference between eligible and
## ineligible households in terms of female, over age 25,
## and having completed primary schooling at baseline.
#univariate regression of maleness measure on cash infusion
dd_male <- lm(male ~ eligible, data = subset(dd, time==0))
#univariate regression of age on cash infusion
dd_age35 <- lm(age35 ~ eligible, data = subset(dd, time==0))
#univariate regression of secondary school completion on cash infusion
dd_schooling <- lm(schooling ~ eligible, data = subset(dd, time==0))

stargazer(dd_male,dd_age35,dd_schooling,
          out="Table 2",type="text",header=FALSE,
          title="Baseline Characteristics: Eligible vs Ineligible",
          dep.var.labels=c("Male", "Age 35+", "Sec School"),
          covariate.labels=c("Eligible"),align=TRUE,
          report = "vc*st",
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),no.space=TRUE)

## Average difference in health between eligible and ineligible
## households at baseline (q2c_baseline) and at endline (q2c_endline)
dd_baseline <- lm(nutrition ~ eligible + male + age35 + schooling, data = subset(dataset,time == 0))
dd_endline <- lm(nutrition ~ eligible + male + age35 + schooling, data = subset(dataset, time == 1))

stargazer(dd_baseline,dd_endline,
          out="Table 3",type="text",header=FALSE,
          dep.var.labels=c("Nutrition"),
          covariate.labels=c("Eligible", "Male", "Age 35+", "Sec School"),
          column.labels = c("Baseline", "Endline"),
          title="Nutritional Outcomes: Eligible vs Ineligible",align=TRUE,
          report = "vc*st",
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),no.space=TRUE)

## Create interaction term (eligible X endline)
dd$treat <- dd$eligible * dd$time

## Run regression and report results
dd_diffindiff <- lm(nutrition ~ eligible + time + treat
                    + male + age35 + schooling,
                    data=dd)

stargazer(dd_diffindiff,
          out="Table 4",type="text",header=FALSE, table.placement = "h!",
          title="Nutritional Outcomes: Diff in Diff",
          dep.var.labels=c("Nutrition"),
```

```
          covariate.labels=c("Eligible", "Time", "Treat", "Male", "Age 35+", "Sec School"),
          align=TRUE,report = "vc*st",
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),no.space=TRUE)

## Section 3
mss <- read.csv("mss.csv")

## Create summary statistics and distributions
stargazer(mss,out="Table 1",
          title="Summary Statistics",
          type="text",header=FALSE)

## Create histogram of NDVI measure (green index growth)
png("Histogram Green.png")
hist(mss$green_index_growth,breaks=20,
     xlab="Proportional change in the green index relative to last year",
     main="Distribution of Green Index Growth Measure")
dev.off()

## Create histogram of GDP growth
png("Histogram GDP.png")
hist(mss$gdp,breaks=20,
     xlab="Log GDP growth since last year",
     main="Distribution of GDP growth rates")
dev.off()

## Second Stage (Structural Equation): Regress democracy score on GDP growth and year
## Structural Equation: Regress democracy score on GDP growth and year
reg1 <- lm(democracy ~ gdp_growth + year, data=mss)
mss$year.f <- factor(mss$year)
reg1_yearfe <- lm(democracy ~ gdp_growth + year.f, data=mss)

stargazer(reg1, reg1_yearfe,
          out="Table 2", type="latex",
          header=FALSE,title="Structural Equation",
          dep.var.labels=c("Democracy"),
          covariate.labels=c("Log GDP Growth","Year"),
          align=TRUE,
          omit="year.f",
          report="vc*st"
          ,omit.stat=c("LL","ser","f","rsq","adj.rsq"),
          no.space=TRUE,
          notes = "Column (2) includes year fixed effects; coefficients not reported",
          notes.append = TRUE)

## First Stage: Regress GDP growth on the NDVI measure and year
reg2 <- lm(gdp_growth ~ green_index_growth + year, data=mss)
reg2_yearfe <- lm(gdp_growth ~ green_index_growth + year.f, data=mss)

stargazer(reg2, reg2_yearfe,
          out="Table 3", type="latex", header=FALSE,
          title="First Stage",dep.var.labels=c("Log GDP Growth"),
          covariate.labels=c("NDVI", "Year"),
```

```
          omit="year.f",
          align=TRUE,
          report="vc*st",
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),
          notes = "Column (2) includes year fixed effects; coefficients not reported",
          notes.append=TRUE,
          no.space=TRUE)

## Reduced Form: Regress democracy score on the NDVI growth measure and year
reg3 <- lm(democracy ~ green_index_growth + year, data=mss)
reg3_yearfe <- lm(democracy ~ green_index_growth + year.f, data=mss)

stargazer(reg3, reg3_yearfe,
          out="Table 4", type="latex", header=FALSE,
          title="Reduced Form",dep.var.labels=c("Democracy"),
          covariate.labels=c("NDVI", "Year"),
          align=TRUE,
          omit="year.f",
          report="vc*st",
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),
          no.space=TRUE,
          notes = "Column (2) includes year fixed effects; coefficients not reported",
          notes.append=TRUE)

# IV Procedure: Regress democracy score on GDP growth, using the
#               NDVI measure as an instrument for GDP growth.
reg4 <- ivreg(democracy ~ gdp_growth + year |
                  green_index_growth + year, x=TRUE, data=mss)
reg4_yearfe <- ivreg(democracy ~ gdp_growth + year.f |
                         green_index_growth + year.f, x=TRUE, data=mss)

## Comparing OLS and IV Results side-by-side
stargazer(reg1, reg4, reg1_yearfe, reg4_yearfe,
          out="Table 5", type="latex", header=FALSE,
          title="OLS and IV Estimates",
          dep.var.labels=c("Democracy"),
          covariate.labels=c("Log GDP Growth", "Year"),
          column.labels = c("OLS", "IV","OLS","IV"),
          model.names = FALSE,align=TRUE,omit="year.f",report="vc*st",
          omit.stat=c("LL","ser","f","rsq","adj.rsq"),no.space=TRUE,
          notes = "Columns (3) and (4) include year fixed effects; coefficients not reported",
          notes.append=TRUE)

## Test for weak instruments
reg4_diagnostics <- summary(reg4, diagnostics = TRUE)$diagnostics
reg4_diagnostics
```