# Health Outcomes in West Africa

*Aileen Gui*

*Fall 2019*

## Purpose of project

I've always wondered exactly what kind of difference money makes. What kinds of problems will it seem to "solve"? A policy question that has always interested me was the effect of income on health. There have been several unconditional cash transfer programs in recent years in developing economies since the dollar goes further outside of the US. For relatively few dollars, a household in certain districts of Kenya can afford much more food, visits to the doctor, parental attention, etc than without. One such cash transfer was done through GiveDirectly (GD). The NGO started distributing large cash transfers (roughly $1000 USD) to thousands of households in rural western Kenya in 2014. The amount was equivalent to more than 50% of total annual income for many recipient households. Approximately 2 years after the initial cash transfers, the households were surveyed on a range of life outcomes including but not limited to: how often they went hungry, how many hours did children spend in school versus working, how many days did they miss school for illnesses, etc. Our endline outcome of interest in this case is nutrition. 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 first part of this analysis focuses on establishing the link between household income and health. The second part will further examine what problems money "solves", particularly conflict in developing households. What difference does GDP make on the level of civil conflict?

## Data

The 'GD' dataset: GiveDirectly (GD) kept records of their cash transfer program and survey results. 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
- nutrition: Nutritional status measure

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
- nutrition: Nutritional status measure

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. The dataset itself is comprised of multiple components:

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 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.
2. GDP Growth: Data comes from Penn World Tables and the World Bank
3. Rainfall: Data comes from the Global Precipitation Climatology Project (GPCP), which records rainfall data at 2.5-degree intervals of latitude and longitude.
4. Other Country-Level Data: Comprised of covariates including ethnic and religious fractionalization, democracy, terrain, etc sourced from World Bank and other databases

- 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 intervention was randomized across villages as opposed to within villages. This was done in order to estimate a treatment effect that would already take into account spillovers and local treatment externalities. Randomization between villages ensures that control households (households in villages where no households were treated) were not impacted by the intervention, keeping the control group valid.

Randomized design ensures that treatment and control villages have similar characteristics in the absence of treatment, so that the only difference between the two gropus is the treatment itself. Therefore, any difference in average outcomes between the two groups can be attributed to the treatment, rather than other pre-existing observable or unobservable differences, which removes the omitted variable bias.

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
```

From the summary statistics, we can see that this dataset contains 2000 households where half 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.

### Assessing balance of covariates at baseline

We would like to estimate the following equation:

$$Y_i = a + bCASH_i + e_i$$

where $Y$ is the characteristic of interest, and `CASH` takes on value 1 if the village received the cash transfers. Here, $b$ denotes the average difference between treatment and control households, and $a$ denotes the average value of the variable in control households.

We will need to run 5 separate regression to view the results: * average difference between treatment and control households for `male` * average difference between treatment and control households for `age35` * average difference between treatment and control households for `schooling` * average difference between treatment and control households in terms of percent female, restricting respondents to those who have not completed secondary schooling * average difference between treatment and control households in terms of percent over age35, restricting respondents to those 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))
```

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 |

The fraction of male respondents is 0.0309 among control households and 0.02 higher among treatment households. The fraction of respondents above age 35 is 0.502 in control households and is on average 0.009 higher in treatment households. The fraction of respondents who have completed secondary school is 0.0503 among control households and on average 0.003 higher among treatment households.

For those households which do not have primary respondents that completed secondary school, treatment households are slightly more likely to have a primary respondent who is male, and slightly more likely to have a primary respondent over the age of 35. Standard errors are reported in the parentheses below the estimated coefficients. Standard errors help to address how precisely estimated each coefficient is, and whether or not that coefficient is statistically distinguishable from zero, often done use t-tests. Here, if a coefficient is indistinguishable from zero, that implies the treatment and control groups are balanced for the characteristic of interest its respective regression.

We will use the 5% significance cutoff, as is standard practice in economics research. 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. Instead, we conclude that treatment and control households are balanced for that particular variable.

The formula we will use is: t=b/se(b). The t-statistic for `male` is 0.095, for `age35` is 0.409, for `schooling` is 0.3, for `male` restricting households to ones where respondents have not completed secondary school is 0.095, and for `age35` with that same restriction is 0.609. All of these values are less than 1.96, indicating that

treatment and control households are balanced for these chracteristics, regardless of restriction to secondary school completion.

This establishes that the randomness intended with initial cash disperal was achieved! There seems to be no significance between certain attributes and whether their household was more likely to receive cash transfers. Now, we want to determine the difference between treatment and control households in terms of their nutritional status in the endline survey.

The regression of interest this time is:

$$NUTRITION_i = a + bCASH_i + e_i$$

Since we've established that randomization of the cash transfers was successful, we can interpret estimates of $b$ as causal estimates of the impact of cash transfers on nutritional status.

We will run 2 regressions: * cash transfer's effect on household nutrition * cash transfer's effect on household nutrition, restricting the dataset to only those households with no schooling past secondary

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

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 can see that receiving a large cash transfer approximately two years earlier improves respondents' nutritional status by an average of 0.102 units of the nutritional status index with a standard error of 0.043. The t-statistic associated with this coefficient is b/se(b)=0.102/0.043=2.37. Since this is greater than 1.96, we can say that the estimated treatment effect is significantly different from zero at 5% significance.

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

Prior to running this analysis, we might expect the cash transfer to have different impacts among the general population versus the subset of the population that has not completed secondary school. This helps to address potential policy concerns about the ways education in a household will impact money usage, and ultimately desired nutrition outcomes. Perhaps people who have completed secondary school are able to make better use of a large cash transfer. Maybe those who've completed secondary education are more likely to operate a small business and the additional cash transfer would provide resources to invest in business inputs, which would generate even more income for the household than the initial cash transfer, increasing incomes and ability to consume higher quality nutrition. Alternatively, those who've completed secondary education might be better informed about the effects of healthy eating, be more financially literate, and

so make decisions in regards to the cash transfer that would be aimed towards increasing long term health rather than restrict themselves to short-term outcomes.

An alternative narrative might be that perhaps those who have not completed secondary school are lower income, more cash-constrained, and thus the lump $1000 sum cash transfer might help them even more than households where primary respondents have completed secondary education. Perhaps households who have not completed secondary schooling possess lower levels of nutrition to begin with, and so much like we see in graphs of diminishing marginal returns, their nutritional gain is higher than those households who start at a higher nutritional index.

The regressions run above indicate that the cash transfer treatment did have a slightly higher impact among those households where the primary respondent has *not* completed secondary school, though this represents a relatively small increase. However, it should be noted that restricting households only eliminated 110 from the original dataset. We should expect that the effect would have been substantially smaller along 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))
```

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 | |

In the first column, we see that when adding controls for gender and age the estimated treatment effect on the health outcomes among the general population decreases from 0.102 to 0.099. The standard error on the estimated coefficient of 0.099 is 0.042, so the t-statistic is 0.099/0.042=2.36. Since this is greater than the cutoff of 1.96, this impact is significantly different from zero at the 5% significance.

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. From these two regressions, we see that inclusion of controls does very little to affect our estimated treatment effects. The results from congrolling gender and age are what we would expect! Since the design of the study was to randomize across villages, that ensured that treatment and control households had similar baseline values for male respondents and respondents over the age of 35. We should not expect that adding these variables as covariates should impact results.

## Household Income and Child Health (Treatment Unrandomly Assigned)

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.

GiveDirectly's program was special in that they were able to collect information from a wide spread of villages, including those in which they did not distribute cash. Many other NGO's run transfer programs similar to this one, but are only able to collect data from villages where they've distributed cash. Let's do another analysis with the data from GiveDirectly, this time limiting ourselves to only data from 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 baseline 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:

Table 4: Descriptive Statistics

| | Mean | Std.Dev | Min | Median | Max | N.Valid |
|---|---|---|---|---|---|---|
| cash | 1.00 | 0.00 | 1.00 | 1.00 | 1.00 | 4000 |
| eligible | 0.50 | 0.50 | 0.00 | 0.50 | 1.00 | 4000 |
| time | 0.50 | 0.50 | 0.00 | 0.50 | 1.00 | 4000 |
| male | 0.25 | 0.43 | 0.00 | 0.00 | 1.00 | 4000 |
| age35 | 0.67 | 0.47 | 0.00 | 1.00 | 1.00 | 4000 |
| schooling | 0.06 | 0.24 | 0.00 | 0.00 | 1.00 | 4000 |
| nutrition | 0.06 | 0.93 | -8.51 | 0.09 | 3.78 | 4000 |

Looking at the summary statistics, we see that this dataset contains 4000 observations from treatment villages where 50% of observations are at endline, and 50% of observations are at baseline. This is to be expected! 50% of households are eligible and were treated at endline. Roughly 25% of household respondents are male, 67% are over age 35, and 6% have completed secondary school.

### 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, and $ELIGIBLE_i$ is an indicator of whether the household was eligible for cash transfers. $b$ denotes the average difference in $Y$ between eligible and ineligible households at baseline. $a$ denotes the average value of $Y$ for ineligible households at baseline.

Below is the code to run the three regressions:

```
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))
```

Table 5: 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 |

The first column indicates that 20% of ineligible household respondents are male, and among eligible households, the value is on average 10.1% higher. The second column indicates that 81.7% of respondents in eligible households are above age 35, while this figure is 30 percentage points lower in eligible households. The third column indicates that 6.1% of respondents in ineligible households have completed secondary schooling, while this figure doesn't change for eligible househlds. Standard errors are reported in parentheses below the estimated coefficients.

We will again use t-tests for each coefficient with a 5% significance level/95% confidence level to determine if the results are statistically different from zero. If the t-statistic computed is less than 1.96, then we cannot reject that the coefficient is statistically different from zero at the 5% significance level. For the variable `male`, the t-statistic is 0.101/0.019= 5.32, for the variable `age35` is -0.300/0.020=-15.0, for the variable `schooling` it is 0.000/0.011=0.00. The t-statistics calculated for `male` and `age35` are both absolutely greater than 1.96, so we conclude that we can reject the hypothesis that the share of respondents that are male and over age35 is the same in eligible and ineligible households at baseline for the 5% significance level. The t-statistic for `schooling` is less than 1.96 in absolute terms, and so we cannot reject the null hypothesis that eligible and ineligible households have the same proportion of respondents who completed secondary school at the 5% significance level.

Eligible households are statistically significantly younger and more likely to be male. However, eligible and ineligible households are similar in average portion of respondents that have completed secondary school. These results are not surprising. While in the original study, villages were randomized, selection within villages was not. Being eligible for a cash transfer within a village meant the household was relatively poor. We should expect some characteristics to vary between the relatively poor and more well-off households.

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

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 below runs the relevant regressions.

```
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))
```

Table 6: 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 | |

The first column indicates that when we control for the respondents' gender, age, and schooling, eligible households have statistically significantly lower nutritional outcomes than ineligible households at baseline. At baseline, households where a male is the primary respondent have slightly higher nutritional scores, though the difference is not statistically significant. Primary respondents over the age of 35 is associated with respondents' having a statistically significantly lower nutrition status (0.24 standard deviations on average). Completing secondary schooling is associated with a statistically insignificant 0.15 average standard deviation higher nutrition status index. The results of the second column indicate that after the controls are added, eligible households now have better nutrition than ineligible households on average, but that difference is not statistically significant. This pattern holds for each of the individual covariates as well. Male respondent households have statistically insignificantly higher nutritional status than female respondent households. Younger households have statistically significantly better nutrition than their 35+ counterparts. Completing secondary school is associated with better nutritional outcomes even at the 1% significance level.

These results are what we would expect! Eligible households have lower nutrition scores at baseline, since cash transfers targeted poorer households. It is also expected that older, less educated, and female-headed households would tend to have lower nutrition on average.

In order to estimate the treatment effect, we will use a difference-in-differences analysis. In order to carry out this method, we will construct an interaction variable of the eligible indicator and the time indicator.

```
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)
```

Table 7: 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 |

I mentioned before that in order for diff-in-diff to truly capture the treatment effect, the parallel trends assumption must hold. There are several reasons why this assumption might not hold in this context, although it is hard to know for sure. If average nutritional values rise faster for young households than for older households, perhaps because their earnings rise faster in youth, we would expect the difference between eligible 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 the cash infusion. Our estimator would therefore overestimate the causal effect of the treamtent.

With this in mind, let us continue our diff-in-diff estimate of the treatment effect, knowing that accuracy rests on the validity of the parallel trends assumption. The estimator of the causal effect of cash transfers on nutrition is 0.084, indicating that getting a cash transfer increases the nutritional index on average by 0.08 points. This estimate is not statistically significant from zero at the 5% confidence level. Thus, we cannot reject the hypothesis that receiving a cash transfer has no effect on nutritional outcomes after about 2 years.

# Section 2

## GDP's Effect on Civil Conflict (Instrumental Variable Approach)

We saw the effects of income on a household level. Now let's expand that lens to see the effects of income on a macro level. How does GDP influence civil conflict? This question suffers from endogeneity issues: perhaps there is reverse causality between civil conflict and national income, or there could be many ommitted variables that affect both income and civil conflict level simultaneously. We will use the instrumental variable strategy to untangle this potential link.

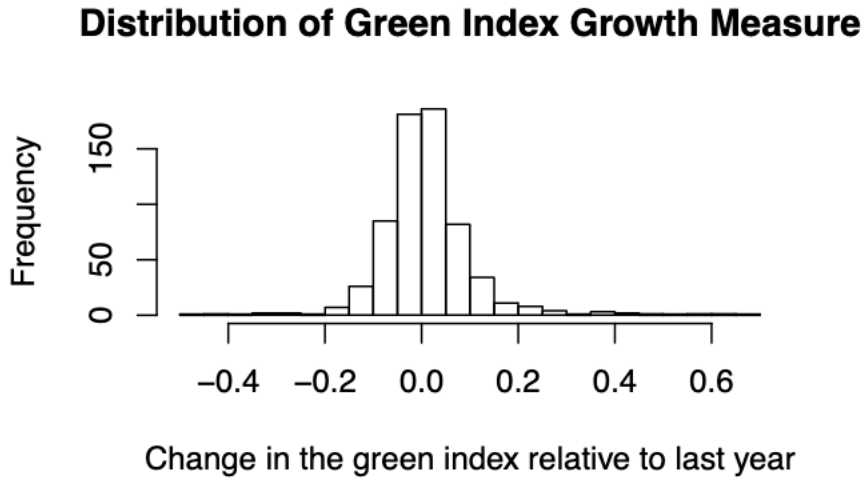
Research from Miguel Satyanath and Sergenti (2004) used the rainfall as the instrumental variable that affect income but not civil conflict level, because many Sub-Suharan African countries' economies are primarily agricultural. For this strategy to be valid, we first need to prove a strong link exists between changes in rainfall and GDP growth. Next, the rainfall should not directly affect conflict (feedback loop) and should be uncorrelated with factors other than GDP growth which also affect conflict (exogeneity). Lastly, the only channel through which rainfall can affect conflict must be through GDP growth (exlusion restriction).

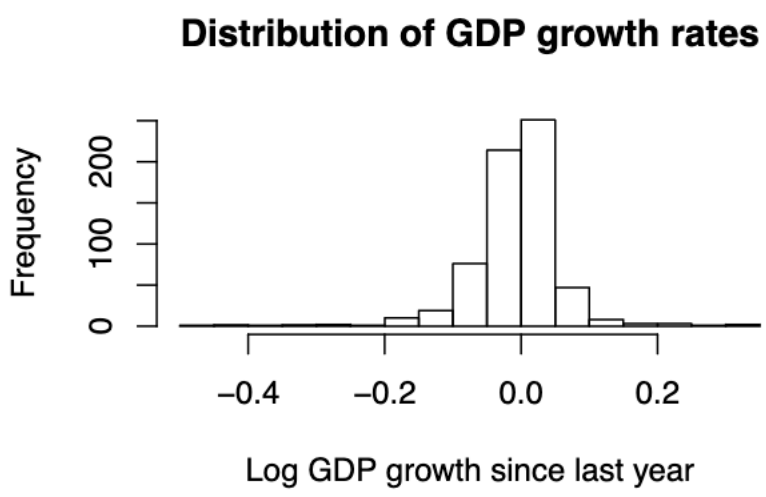
Let's first prove a strong link exists between rainfall and GDP growth:

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

Table 8: 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 |






The results above visualize our variables: democracy is an integer variable taking on values from -10 to +9, NDVI 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 gdp growht is measured in log changes (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 added as a linear control are as follows:

First stage regression establishes the strength of the instrument:

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

Second stage regression (structural equation):

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

Reduced form regression:

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

The same equations with year fixed effects instead of year as a linear control:

$$GDPGROWTH_{ct} = \pi_0 + \pi_1 NDVI_{ct} + \delta_{1t} + u_{ct}$$

$$DEMOCRACY_{ct} = \beta_0 + \beta_1 GDPGROWTH_{ct} + \delta_{2t} + \varepsilon_{ct}$$

$$DEMOCRACY_{ct} = \gamma_0 + \gamma_1 NDVI_{ct} + \delta_{3t} + \nu_{ct}$$

Using an Instrumental Variables Two Stage Least Squares (IV-2SLS) estimation strategy including country fixed effects and country-specific time trends, we find that a one percentage point increase in GDP decreases the probability of civil conflict by 2.55 percentage points. This is major. A five percentage point decline in GDP would lead to an over 12 percentage point increase in the likelihood of civil conflict.

Year fixed effects allow the average of the outcome variable across all countries to vary for each specific year. It allows us to control 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.

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 $\beta_1$ can be estimated using an indirect least squares procedure as $\hat{\beta_1}^{ILS} = \frac{\hat{\gamma_1}}{\hat{\pi_1}}$. This is because we have one endogenous regressor (GDP growth) and one instrument (NDVI measure). 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: $\beta_1^{ILS} = \beta_1^{2SLS}$.

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)
```

Table 9: 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 | |

From the results above, we see that from the linear control method, annual GDP growth is negatively associated with more democratic institutions. The coefficient on annual GDP growth ($\beta_1^{OLS}$ = -1.697 (with a standard error of 2.995). This means a 1 percentage point increase in GDP growth is associated with a 0.017 decline in deocracy score. The associated t-statistic = -0.567. Since |-0.567|<1.96, we can conclude that the OLS estimate is not statistically significant at the 5% significance level. The coefficient on the year variable is statistically significant at 0.519, indicating that the democracy score increases 0.519 units on average each year across all countries. Historically, this makes sense. Sub-Saharan African countries have been democratizing since the 1990s. When we instead run the regression including year fixed effects, the coefficient estimated is statistically significant at -0.465 (with a standard error of 3.030) and 95% confidence. You could pick either method to go with, but to simplify things, I'll keep going with only the models that include year as a linear control.

Next, we regress the GDP growth measure on the NDVI growth measure and the year variable. This is 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)
```

Table 10: 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 | |

The results above indicate that the NDVI measure is a strong instrument for GDP growth because of its strong correlation. The coefficient on the NDVI measure is 0.109, which means that a 1-unit increase in the index measure is associated with a 10.9 percentage point increase in GDP growth (with a standard error of 0.027). The associated t-statistic is is 4>1.96, meaning the estimate is significant at the 95% confidence level. The coefficient for years is also statistically significant at the 95% confidence level, indicating that GDP growth was accelerating in Sub-Saharan Africa during the years of data collection.

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)
```

From the results above, we see that the coefficient on the NDVI measure is 0.060. This means that a 1-unit increase in NDVI measure is associated with a 0.06 unit increase in democracy score (with a standard error of 2.073 and corresponding t-statistics of 0.029<1.96). This result is not statistically significant at the 95% confidence level. The coefficient on the year control was significantly positive, indicating that Sub-Saharan African coutnries were becoming more democratic over the period of data collection.

Finally, we carry out the estimation of the instrumental variable model, including the year variable as a control:

```
# 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)
```

Table 11: 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

Table 12: 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

The results in column 1 represent the second stage (structural equation) regression using OLS. Column 2 results represent results of the second stage regression using IV. We see that the beta for OLS is -1.697, and the beta for IF is 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 estimate is statistically significant from zero at the 95% confidence level. Therefore, we cannot reject the null hypothesis that GDP growth has no impact on democratization within the same year (since we have yearly data).

Lastly, let's re-evaluate instrument validity which relies on 3 main factors: existence of strong first stage, satisfaction of exogeneity, and satisfaction of the exclusion restriction.

First Stage: There seems to be a strong connection between NDVI measures and GDP growth changes. First stage results support that when the NDVI measure is high, GDP growth is high. When that same measure is low, GDP growth is low. In summarizing our IV regression, we see that the F-stat associated with a test for weak instruments 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: There is likely no feedback loop from democracy score to rainfall measure. A potential violation scenario, however, might be if democratization leads to a change of vegetated areas, since NDVI is based on satellite imagery of the amount of greenery in the country. This would cause NDVI measures to not be exogenous. We also need the changes in the green index growth measure to be uncorrelated with ommitted variables or other factors besides GDP growth that influence democratization in Sub-Saharan African countries. On example of violation might be if other climate-related phenomena are associated with green index growth. For instance, perhaps heat is associated with aggressive behavior, and if less rainfall is associated with higher temperatures, and higher temperatures lead to more violence that affect democratic movements, this would introduce temperature as an omitted variable which confounds the effect of rainfall on conflict.

Exclusion Restriction: This requires that GDP growth is the only channel through which rainfall influences civil conflict. We can not directly show (through regression or trends) that the exclusion restriction holds. One potential violation might be if rainfall influences democratization, independent of GDP growth. Perhaps more rainfall influences people to stay inside, away from protests, and thereby reduces the power of democratic movements.

The regression estimates from 2SLS are still useful in that even if you are skeptical of exclusion restriction satisfaction, the regression still proves a relationship between green index growth and democracy (civil conflict), without taking a stance on causality. The regression estimates from IV rest on the assumption that all 3 instrument requirements hold, and shows the impact of GDP growth on civil conflict. If the goal is to pin down the causal relationship between GDP growth and civil conflict, and the exclusion restriction is held, then the most accurate regression estimates are the IV estimates.

## 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
```