Problem Set 2

ECONOMICS 172: Issues in African Economic Development

Oscar Chaix

03/06/19

1 - Impacts of a Malaria Vaccine for Africa

Life expectancy & other health measures are lower in Sub-Saharan Africa (from now on Africa) than anywhere else, even before accounting for the HIV epidemic. Widespread tropical diseases, including sleeping sickness (trypanosomiasis), yellow fever and malaria have plagued the region for centuries, not only largely reducing African population until the 1870s but also affecting colonisation and slave trade and inducing the formation of extractive institutions in highly affected areas (see Acemoglu et al). Malaria is the most important tropical disease, killing about 400 thousand Africans per year, with children and pregnant women being the most vulnerable sub-groups as adults gain partial immunity with time. Malaria prevalence however thankfully decreased by about 40-50% since 2000, in large part due to expanded use of insecticide-treated bednets.

Based on the article, the vaccine has several limitations. First, according to the World Health Organization, it is only effective for 30 to 50% of patients. Moreover, its effectiveness diminishes with time and it is apparently less successful for the most exposed patients. Finally, the pilot will only be complete by 2021, and "researchers will continue monitoring feasibility and safety until 2022" according to the article, so it will take some time before we can see the effects of the vaccine, if it is actually deployed on a national scale. Nevertheless, the vaccine seems promising in the current trial in Kenya and it should make a difference in the long run to help eradicate malaria. In fact, since the vaccine will potentially be administered to as much as 120 thousand children from 6 to 24 month year old, it targets the most susceptible population at a very early age, thereby potentially significantly reducing malaria prevalence in their early childhood. Combined with the continued use of insecticide-treated bednets, the vaccine should thus be effective in the long run to significantly reduce malaria prevalence and malaria-induced mortality in Kenya.

If the treatment is successful, a direct demographic short term consequence should be lower infant mortality in Kenya and an even higher dependency ratio, as long as fertility rates remain high. This would lead to less savings and investment, since more children will pose a higher financial burden on parents, and thus to lower growth. However, another mechanic effect of the

vaccine if it is successful is an increase in the (healthy) population, which should increase economic growth in the long run through a more abundant productive workforce.

With regards to economic outcomes, there are other positive potential consequences to consider. According to Bloom and Sachs in their paper "Geography, Demography, and Economic Growth in Africa", published in 1998, tropical diseases like malaria lead to low labor productivity an low per capita income, as well as to lower foreign investment and technological transfers. For example, diseases like malaria limit the visit of foreign businessmen and tourists in Africa and constrain the ability of African businessmen to leave for a significant time, since their partial immunity to malaria would be at risk. These investment and labor productivity channels, according to Bloom and Sachs, directly slow long run economic growth in Africa. If their theory and regression prove to be correct in the Kenyan case, the vaccine could potentially lift up Kenya and African countries' economic growth in a significant way.

A channel linking health and income that has been shown to be potentially very significant is the education channel. More focused approaches, like Miguel and Kremer's 2004 study, have in fact demonstrated a significant link between health, education and labor productivity. Their randomised control trial on school-deworming's impact shows no short term effects on education measures (like test gains), but shows that in the long run deworming programs is an efficient human capital investment, as treated individuals achieve more schooling, profit form higher 'quality' peer groups, and end up having higher adult earnings and occupations. Their 2007-2008 follow-up survey in fact demonstrates overall better health and working performance among those previously treated in school, with for example hours worked 17% higher, which shows these individuals can better keep up with heavy workdays, and earnings about 20% higher. Miguel and Kremer hence show that the long run benefits of school-deworming to the economy are very large, and a similar reasoning can be applied to the malaria vaccine: if young individuals are treated in mass, it should lead to significant long run benefits for the economy.

2 - Household Income and Child Health, Part 2

The main econometric assumption that needs to be met in order to have an accurate difference-in-difference (DD) approach is the stability of omitted variable bias over time. If the nature of the differences between the two non-randomly selected groups is stable over time, we can use the DD method to remove omitted variable bias.

In other words, if the treatment and control outcomes move in parallel in absence of treatment (at t = 0), the divergence of a post-treatment path from the trend established by a comparison group may signal an unbiased treatment effect. The DD estimator subtracts the post-treatment difference between both groups from the pre-treatment difference, thereby adjusting for the fact that the groups were not the same initially.

As an illustration, if we assign $t = 0$ to the pre-program period (the "baseline") and $t = 1$ to the post-program period (the "follow-up"), we have:

(1) $Yit = a + bTit + dXit + eit$

In our specific situation, the program (or "event") corresponds to an household receiving the cash transfer, and the outcome corresponds to their health outcome in the baseline and endline surveys. In period $t = 1$, the treatment estimate is as we calculated in problem set 1:

(2) $E(Yi1 \mid Ti1 = 1) - E(Yi1 \mid Ti1 = 0)$

$= [a + b + dE(Xi1 \mid Ti1 = 1) + E(ei1 \mid Ti1=1)] - [a + 0 + dE(Xi1 \mid Ti1 = 0) + E(ei1 \mid Ti1 =0)]$

$= b + d[E(Xi1 \mid Ti1 = 1) - E(Xi1 \mid Ti1 = 0)]$

Where b corresponds to the true effect term and the right part of the equation corresponds to the omitted variable bias (OVB) term for t=1.

However, in period $t = 0$, the treatment estimate is different, because we did not apply the

treatment yet. Before the event (i.e. no one has received the cash transfer yet), the difference between the two groups is the following, where we already know which households will receive the cash transfer in period $t = 1$ ($Ti1 = 1$):

(3) $E(Yi0 \mid Ti1 = 1) - E(Yi0 \mid Ti1 = 0)$

$= [\, a + \mathbf{0} + dE(Xi0 \mid Ti1 = 1) + E(ei0 \mid Ti1=1)] - [a + 0 + dE(Xi0 \mid Ti1 = 0) + E(ei0 \mid Ti1 =0)]$

$= d[E(Xi0 \mid Ti1 = 1) - E(Xi0 \mid Ti1 = 0)]$

Where the final expression underlines only the OVB term for t=0, which makes sense since the there hasn't been any treatment yet.

Finally, to calculate the difference-in-difference estimator, we take the difference between equation (2) and (3) to adjust for initial difference between the two groups:

(4) $[E(Yi1 \mid Ti1 = 1) - E(Yi1 \mid Ti1 = 0)] - [E(Yi0 \mid Ti1 = 1) - E(Yi0 \mid Ti1 = 0)]$

$= b + d[E(Xi1 \mid Ti1 = 1) - E(Xi1 \mid Ti1 = 0)] - d[E(Xi0 \mid Ti1 = 1) - E(Xi0 \mid Ti1 = 0)]$

<=> True effect + [(OVB in $t = 1$) - (OVB in $t = 0$)]

If our assumption holds and the OVB between both groups is stable over time, then this second term disappears and cancels out, and the DD estimator delivers only the true effect. However, if the OVB between both group is *not* stable over time, then the second term does *not* cancel out and there is still omitted variable bias in the regression estimates.

Therefore, it is important to have access to data from both the baseline and endline survey rounds to be able to compare the relative trajectories between both groups before the cash transfer and after the cash transfer, to accurately capture the evolution of the omitted variable bias over time. With more data, the assumption of stable OVB can be probed, tested, and relaxed, as we can for

example check that the trend is actually in parallel for both household groups before the cash transfer.

The lack of randomisation in the allocation of cash transfers within treatment villages affect the estimation of the treatment effect because it creates omitted variable bias since the cash transfer is allocated to households with different characteristics: the transfers were targeted to relatively poor households. However, if we take into account this original difference and assume that it does not change, we can remove this bias in the estimated treatment effect through the DD method. To consider the required conditions underlying our assumption of parallel trend, we need to ask ourselves the counterfactual: would health outcomes between the two groups have evolved similarly over time in the absence of the cash transfer? We therefore need to consider other unobserved factors that could have affected the outcome over the study period, for example changes in local policy, in sickness prevalence, or environmental factors over time and throughout villages. We also need to consider how the eligible households and ineligible households would potentially differ in their health habits and hygiene in the absence of the cash transfer.

In my opinion, the assumption of parallel health level trend before and after the cash transfer is likely to be met in the case of GD cash transfers in Kenya, since we saw in problem set 1 that income did not significantly affect health outcomes. Hence, even though the cash transfer targeted relatively poor households, the DD approach is still a reasonable approach to evaluate the policy because we did not see any significant effects of income on health in our last study. In addition, the eligible and ineligible groups come from the same villages, so even though external factors like new plagues, natural disasters or new local policies can emerge, they should affect the eligible and ineligible households similarly. In other words, if there is a confounding factor that emerges at any point, the trends of both groups should vary in a similar way and thus stay parallel. Finally, as there was surveying both at baseline and after the cash transfer, the panel data should be sufficiently rich to thoroughly assess the relative trends and health outcomes of both groups over time.

(b)

Summary statistics

First let's have a look at the summary statistics of the baseline survey data:

Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
hhid	1,000	500.500	288.819	1	250.8	750.2	1,000
eligible	1,000	0.500	0.500	0	0	1	1
time	1,000	0.000	0.000	0	0	0	0
cash	1,000	1.000	0.000	1	1	1	1
female	1,000	0.742	0.438	0	0	1	1
age25	1,000	0.895	0.307	0	1	1	1
schooling	1,000	0.363	0.481	0	0	1	1
health	1,000	-0.170	1.014	-2.410	-0.570	1.270	1.270

We can see that these summary statistics concern only the baseline survey data of the treatment group, since the variable time has a mean of 0 and the variable cash has a mean of 1. Looking at the three explanatory variables (female, age25, schooling), we see that this population has high proportions of females and respondents older than 25, equal to about 74.2% and 89.5% respectively, and a low proportion of people that completed primary school, at only 36.3%. Concerning the health variable, it looks like at baseline treated households generally felt not healthy, with an index of about -0.170, but with a high standard deviation of about 1 index point.

Average differences between the eligible and non-eligible groups

Now let's look at the covariation of the "eligible" dummy variable with the control characteristics:

Baseline Covariate Results

	Dependent variable:		
	female	age25	schooling
	(1)	(2)	(3)
eligible	-0.048*	-0.126***	0.062**
	(0.028)	(0.019)	(0.030)
Constant	0.766***	0.958***	0.332***
	(0.020)	(0.013)	(0.021)
Observations	1,000	1,000	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01		

We can first note that the constant coefficients, corresponding to how much female, age25, and schooling participation there is overall regardless of eligible or non-eligible groups, are similar to the proportions we found in the summary statistics, at 0.766 (or 76.6%), 0.958 (95.8%) and 0.332 (33.2%) respectively. These estimates furthermore have high statistical significance, with p-values that are less than 0.01: in other words, we are 99% confident that these estimates are not simply due

to chance variation. Our regressions therefore confirm our summary statistics results, as they show that there are a majority of females and age 25 and higher households and a minority of primary-educated households in both groups.

First, our regression of the female variable on the eligible variable (corresponding to column 1) gives us an estimated coefficient of -0.048 with a standard error of 0.028 and t-value of -1.735. The slope coefficient is relatively small but appreciable, which shows that being in the eligible or ineligible group slightly explain why a household respondent is a male or female. In fact, an increase in the eligible variable of one unit (meaning that we only consider treated households) is associated with an average decrease in female proportion of -4.8% (since 1 is where all persons are female). In other words, the coefficient does not highlight an important difference between eligible and ineligible groups in terms of the proportion of females both have, but a considerable 5% difference.

However, although this estimated parameter is statistically significant at the 10% level, it has a lower statistical significance than the estimated slopes of the other two regressions. As the standard error is equal to 0.028, a 90% confidence interval around our parameter is (β - 1.645 • SE , β + 1.645 • SE) <=> (-0.048 - 1.645*0.028 , -0.048 + 1.645*0.028) <=> (-0.094 , -0.002). Since the confidence interval does not includes zero, we can reject the null hypothesis of no correlation at the 90% level . In other words, we are 90% sure that the real coefficient β is not 0. The t-statistic gives us the same information, however if we test the significance at the 5% level, since $t = -1.735$, we have $t < 1.96$, and we cannot reject the null with 95% confidence.

Thus, eligible and ineligible groups have somewhat significantly different proportions of women, but this difference is small and not highly significant. Being in one group or another is only slightly correlated with being a man or a woman, with statistical significance only when we consider the 10% level.

Second, our regression of the age25 variable on the eligible variable (corresponding to column 2) gives us an estimated coefficient of -0.126 with a standard error of 0.019 and t-value of -6.634. The slope coefficient is more consequent here, which shows that being in the eligible or ineligible group does explain why a household respondent is 25 years old and older or not. In fact, an increase in the eligible variable of one unit (meaning that we only consider treated households) is associated with an average decrease in age25 proportion of -12.6% (since 1 is where all persons are 25 or older). In other words, the estimated coefficient highlights a substantial difference between the eligible and ineligible groups in terms of the proportion of age 25 and older both have, a 12.6% difference.

Furthermore, this estimated parameter is statistically highly significant, as the standard error is equal to 0.019. In fact, a 95% confidence interval around our parameter is (β - 1.96 • SE , β + 1.96 • SE) <=> (-0.126 - 1.96*0.019 , -0.126 + 1.96*0.019) <=> (-0.163 , -0.088). Since the confidence interval does not includes zero, we can reject the null hypothesis of no correlation at the 95% level. In other words, we are 95% sure that the real coefficient β is not 0. The t-statistic gives us the same information: since $t = -6.634$, we have $t > 1.96$, and we can reject the null with 95% confidence.

Thus, eligible and ineligible groups have substantially and significantly different proportions of people that are at least 25 and being in one group or another is correlated with being at least 25 years old.

Third, our regression of the schooling variable on the eligible variable (corresponding to column 3) gives us an estimated coefficient of 0.062 with a standard error of 0.030 and t-value of

2.041. The slope coefficient is less consequent than for the age25 variable and relatively small, but again appreciable enough, which shows that being in the eligible or ineligible group slightly explains why a household respondent has completed primary schooling or not. In fact, an increase in the eligible variable of one unit (meaning that we only consider treated households) is associated with an average increase in primary-educated proportion of 6.2%. In other words, the coefficient highlights a slight difference between eligible and ineligible groups in terms of the proportion of females both have, a 6% difference.

Furthermore, this estimated parameter is statistically significant, as the standard error is equal to 0.030. In fact, a 95% confidence interval around our parameter is (0.003 , 0.121). Since the confidence interval does not includes zero, we can reject the null hypothesis of no correlation at the 95% level. In other words, we are 95% sure that the real coefficient β is not 0. The t-statistic gives us the same information: since $t = 2.041$, we have $t > 1.96$, and we can reject the null with 95% confidence.

Thus, eligible and ineligible groups have slightly but significantly different proportions of people that have completed primary school and being in one group or another is slightly correlated with being primary-educated.

All in all, even when the regression estimates point to slight correlations, we can observe significant covariation between the eligible variable and all three characteristics. Taken together, the eligible and ineligible households appear to be different along these three demographic characteristics at baseline. The targeted selection of the eligible population thus appears, as expected, to have created significant selection bias between the eligible and non-eligible groups at baseline.

(c)

Let's now look at the average health status differences between the eligible and non-eligible groups in both the baseline survey and endline survey:

Baseline and Endline Regressions of Health Outcomes

	Dependent variable:	
	health	
	(1)	(2)
eligible	0.201***	0.078
	(0.065)	(0.060)
female	-0.116	-0.101
	(0.074)	(0.068)
age25	-0.341***	-0.475***
	(0.107)	(0.099)
schooling	0.162**	0.217***
	(0.068)	(0.062)
Constant	0.063	0.186
	(0.139)	(0.128)
Observations	1,000	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01	

The first column corresponds to the regression of the health index on the eligible variable and the other there explanatory variables (female, age25 and schooling) using the baseline survey, while the second column corresponds to the same regression but using the endline survey data.

At baseline, it appears that there are large health differences between the eligible and ineligible groups. In fact, the slope coefficient is substantially large, at about 0.2, which shows that being in the eligible or ineligible group does explain why a household is healthy or not at baseline. In fact, an increase in the eligible variable of one unit (meaning that we only consider treated households) is associated with an average increase in the health index of 0.201. In other words, the estimated slope coefficient highlights a substantial difference between the eligible and ineligible groups in terms of their health at baseline, with eligible households being on average healthier.

Furthermore, this estimated parameter is statistically highly significant, as the standard error is equal to 0.065. In fact, a 95% confidence interval around our parameter is (0.074, 0.329). Since the confidence interval does not includes zero, we can reject the null hypothesis of no correlation at the 95% level. In other words, we are 95% sure that the real coefficient β is not 0. The t-statistic gives us the same information: since $t = 3.106$, we have $t > 1.96$, and we can reject the null with 95% confidence. Finally, the three stars associated to the estimated coefficient indicate that the p-value is lower than 0.01, showing that in fact we are 99% confident that the real coefficient β is not 0. Thus, eligible and ineligible groups have substantially and significantly different health levels at baseline, at least if we trust the health index to accurately measure the health of an individual.

The other parameters in this baseline regression also indicate substantial differences at baseline in health status between households of different age group and different primary-school

background. The estimated coefficients for both explanatory variables are in fact substantial, at -0.341 for the age25 variable and 0.162 for the schooling variable. Both are also statistically significant, with p-value of less than 0.01 and 0.05 respectively. The female explanatory variable however is not statistically significant, so we cannot infer from this regression any substantial differences in health status between different genders at baseline. The constant also cannot be interpreted here, since it is not statistically significant.

However, looking at the second regression, corresponding to the endline data, there are no significant differences between the eligible and ineligible groups in terms of health status. In fact, the slope coefficient is much lower, at about 0.078, which shows that being in the eligible or ineligible group does not explain why a household is healthy or not in the endline survey.

Furthermore, this estimated parameter is not statistically significant, as the standard error is equal to 0.060. In fact, a 95% confidence interval around our parameter is (-0.04, 0.20). Since the confidence interval includes zero, we cannot reject the null hypothesis of no correlation at the 95% level. In other words, we are not 95% sure that the real coefficient β is not 0. The t-statistic gives us the same information: since $t = 1.314$, we have $t < 1.96$, and we cannot reject the null with 95% confidence. Thus, eligible and ineligible groups do not have substantially and significantly different health levels in the endline survey.

The other parameters, like in the first regression, indicate substantial differences in the endline survey in health status between households of different age group and different primary-school background. The coefficients for both explanatory variables are in fact substantial, at -0.475 for the age25 variable and 0.217 for the schooling variable. Both are also highly significant, with p-values of less than 0.01. The female explanatory variable however is again not statistically significant, so we cannot infer from our regression any substantial differences in health status between different genders in the endline survey. The constant also cannot be interpreted here, since it is not statistically significant.

Taken together, neither of these analyses on their own have a causal interpretation. Even though the first regression shows statistically significant and substantial differences between eligible and ineligible groups in terms of health status, we cannot make any causal interpretation from it since there hasn't been any cash transfer to the eligible group yet. Differences in health status between both groups at baseline are thus due to selection bias, and not to a treatment effect. Then, looking solely at the second regression of health status on all other variables in the endline survey, we do not find any statistically significant health index differences between eligible groups and ineligible groups. Therefore, the second regression does not demonstrate any causal impact of the treatment on health status either. This was expected, since in problem set 1 we saw that there wasn't any significant effect of cash transfers on health outcomes, even looking at households from different treatment and control villages.

(d) Finally, let's consider a difference-in-difference (DD) analysis that uses data from both time periods simultaneously. Our regression equation needs to take the form:

$Yit = \alpha + \gamma Ni + \lambda dt + \delta(Ni \bullet dt) + \varepsilon it$

with:

*Y*it: health index

Ni: eligible indicator (or "treatment indicator")

dt : time indicator (or "pre-post indicator")

Ni•dt : interaction term (or "DD term")

To set up this regression, we thus need to create an interaction variable "treat" that takes the product of the eligible and time variables. This term is an indicator variable which takes the value 1 when the eligible and time indicator equal to 1, and which takes the value 0 otherwise (which makes sense, since we multiply two indicator variables). The regression estimate associated with this variable is the estimator of interest in our analysis. δ in fact represents the effect of the cash transfer on the eligible group's health status while taking in account differences (or selection bias) between the two groups over time. In other words, it underscores the difference between the health outcome of the eligible group with the treatment and what would have been the health outcome of the eligible group without the treatment (the counterfactual), assuming the trend of the counterfactual to be the same as the trend of the ineligible group. Adding this variable to the dataset, and including the other three explanatory variables in our regression, we have:

DD Regression of Health on Eligibility, Time, the Intersection Term and Controls

	Dependent variable:
	health
eligible	0.191***
	(0.061)
time	0.026
	(0.061)
treat	-0.103
	(0.086)
female	-0.108**
	(0.050)
age25	-0.408***
	(0.073)
schooling	0.189***
	(0.046)
Constant	0.112
	(0.099)
Observations	2,000
Note:	*p<0.1; **p<0.05; ***p<0.01

The impact of receiving a large cash transfer approximately two years earlier on respondents' health status, using a difference-in-difference approach, is neither substantial nor statistically significant. Actually, if there is an effect, it is more negative than positive, as we can see that the interaction term "treat" estimate decreases by -0.103. However, we cannot infer any effect or impact from this estimate coefficient, since it is not statistically significant. The standard error being equal to 0.086, a 95% confidence interval around our parameter is (-0.27, 0.066). Since the confidence interval includes zero, we cannot reject the null hypothesis of no correlation at the 95% level. In other words, we are not 95% sure that the real coefficient β is not 0, and this estimator is not significantly different from 0 at 95% confidence. The t-statistic gives us the same information: since $t = -1.203$, we have $t < 1.96$, and we cannot reject the null with 95% confidence. Thus, eligible and ineligible groups do not have substantially and significantly different health status levels after the cash transfers.

The implication of this result is that there is no causal impact of higher household income on health outcomes. In fact, like in problem set 1, we find no statistically significant effect of higher household income on health outcomes, even controlling for age, gender and schooling. The eligible and ineligible groups have similar health outcomes two years after the cash transfer and being in one group or another is not correlated with being healthier, even when we do a thorough difference-in-difference analysis that takes in account time and household eligibility. This could be due to various phenomenas that need to be investigated more thoroughly in further research. For example, maybe there is a transition period (or lag time) before respondents spend extra income on health care and/or before results manifest; maybe are health outcomes not easily ameliorated with more spending in health in rural Kenya; or maybe do people rarely spend extra-income on health and prefer consuming other goods and services. In any case, our DD regression results do not support a causal impact of income on health, and even less a positive one.

A final note is that the estimated coefficients for the three explanatory variable are high and have high statistical significances in this last regression. We can interpret these coefficients as underscoring a generally negative effect of being a female or being old on the health index survey, which intuitively makes sense and is in line with what we found in problem set 1. These results are what we would expect since, intuitively, different genders and age levels should have different effects on one's health, especially in the context of a developing country. Concerning the age variable, persons are usually less healthy as they get older, or at least they feel like they are less healthy. Concerning the female variable, in the context of rural western Kenya, this could be because women generally make less income and/or are treated with less care by their families, which could provoke a negative effect on their health. Finally, schooling seems to have a positive effect on health outcome with high statistical significance ($p<0.01$), which also intuitively makes sense since educated people should have better hygiene and value more their health.

R-History

```
#PB Set 2
data=read.csv("~/Desktop/Econ172_S19_ProblemSet2_data.csv")
library(stargazer)
library(dplyr)
######b######
summary(data$eligible)
baseline_data = filter(data,time==0)
summary_stats = summary(baseline_data)
summary_stats
###Average differences between households eligible and ineligible for
treatment###
##Gender differences
reg1= lm(female ~ eligible, baseline_data)
reg1
summary(reg1) #coeff = -0.048 , SE = 0.028, t-value = -1.735
##Age differences
reg2= lm(age25 ~ eligible, baseline_data)
reg2
summary(reg2) #coeff = -0.126, SE = 0.0190 , t-value = -6.634 (***)
##Education differences
reg3= lm(schooling ~ eligible, baseline_data)
reg3
summary(reg3) #coeff = 0.062 , SE = 0.0304, t-value = 2.041 (*)
##Result Outputs
stargazer(baseline_data,
out="Table 1.html",type="html",header=FALSE,
titles="Baseline Summary Statistics",align=TRUE,no.space=TRUE,stats =
c("mean", "sd", "min", "med", "max", "n.valid"))
stargazer(reg1,reg2,reg3,
out="Table 2.html", type="html",header=FALSE,
title="Baseline Covariate Results",align=TRUE,
omit.stat=c("LL","ser","f","rsq","adj.rsq"),no.space=TRUE)
######c######
##Regression of health on eligbility, female, age25 and shcooling for the
baseline data
reg4 = lm(health ~ eligible + female + age25 + schooling, baseline_data)
reg4
summary(reg4)
##Regression of health on eligbility, female, age25 and shcooling for the
endline data
endline_data = filter(data,time==1)
reg5 = lm(health ~ eligible + female + age25 + schooling, endline_data)
reg5
summary(reg5)
##Result Outputs
stargazer(reg4,reg5,
out="Table 3.html", type="html",header=FALSE,
title="Baseline and Endline Regressions of Health Outcomes",align=TRUE,
omit.stat=c("LL","ser","f","rsq","adj.rsq"),no.space=TRUE)
######d######
#Creation of the interaction variable "treat" and inclusion in data
data = mutate(data, treat = data$eligible*data$time)
##DD regression of health on eligbility, time, the intereaction term, and
controls
reg6 = lm(health ~ eligible + time + treat + female + age25 + schooling,
data)
reg6
summary(reg6)
```

```
##Result Outputs
stargazer(reg6,
out="Table 4.html", type="html",header=FALSE,
title="DD Regression of Health on Eligibility, Time, the Intersection Term
and Controls",align=TRUE,
omit.stat=c("LL","ser","f","rsq","adj.rsq"),no.space=TRUE)
timestamp(stamp=date())
##------ Wed Mar  6 17:22:19 2019 ------##
savehistory(file="PB_Set_2_Oscar_CHAIX.Rhistory")
```

R-Script Code

```
#PB Set 2

data=read.csv("~/Desktop/Econ172_S19_ProblemSet2_data.csv")
library(stargazer)
library(dplyr)

######b######

summary(data$eligible)
baseline_data = filter(data,time==0)
summary_stats = summary(baseline_data)
summary_stats

###Average differences between households eligible and ineligible for treatment###

##Gender differences

reg1= lm(female ~ eligible, baseline_data)
reg1
summary(reg1) #coeff = -0.048 , SE = 0.028, t-value = -1.735

##Age differences

reg2= lm(age25 ~ eligible, baseline_data)
reg2
summary(reg2) #coeff = -0.126, SE = 0.0190 , t-value = -6.634 (***)

##Education differences

reg3= lm(schooling ~ eligible, baseline_data)
reg3
summary(reg3) #coeff = 0.062 , SE = 0.0304, t-value = 2.041 (*)

##Result Outputs

stargazer(baseline_data,
          out="Table 1.html",type="html",header=FALSE,
          titles="Baseline Summary Statistics",align=TRUE,no.space=TRUE,stats = c("mean", "sd", "min",
"med", "max", "n.valid"))

stargazer(reg1,reg2,reg3,
          out="Table 2.html", type="html",header=FALSE,
          title="Baseline Covariate Results",align=TRUE,
omit.stat=c("LL","ser","f","rsq","adj.rsq"),no.space=TRUE)

######c######

##Regression of health on eligbility, female, age25 and shcooling for the baseline data

reg4 = lm(health ~ eligible + female + age25 + schooling, baseline_data)
reg4
summary(reg4)

##Regression of health on eligbility, female, age25 and shcooling for the endline data

endline_data = filter(data,time==1)
reg5 = lm(health ~ eligible + female + age25 + schooling, endline_data)
```

```
reg5
summary(reg5)

##Result Outputs

stargazer(reg4,reg5,
          out=“Table 3.html", type="html",header=FALSE,
          title="Baseline and Endline Regressions of Health Outcomes",align=TRUE,
omit.stat=c("LL","ser","f","rsq","adj.rsq"),no.space=TRUE)

######d######

#Creation of the interaction variable "treat" and inclusion in data

data = mutate(data, treat = data$eligible*data$time)

##DD regression of health on eligbility, time, the intereaction term, and controls

reg6 = lm(health ~ eligible + time + treat + female + age25 + schooling, data)
reg6
summary(reg6)

##Result Outputs

stargazer(reg6,
        out="Table 4.html", type="html",header=FALSE,
        title="DD Regression of Health on Eligibility, Time, the Intersection Term and
Controls",align=TRUE, omit.stat=c("LL","ser","f","rsq","adj.rsq"),no.space=TRUE)
```