MSD 2019 Final Project

A replication and extension of Housing, Health, and Happiness by Matias D. Cattaneo, Sebastian Galiani, Paul J. Gertler, Sebastian Martinez, and Roccio Titiunik, American Economic Journal: Economic Policy 2009

Nancy Thomas (nkt2111), Patrick Alrassy (pa2492), Brigid Lynch (bml2133)
2019-05-12 21:16:47

Contents

Background 1

Data 2

Import Data 2

Reproduction: 2

Model 1: no controls 3

Model 2: age, demographic, and health-habit controls 4

Model 3: age, demographic, health-habit and public social programs controls 6

Control Group Means and Standard Deviations 8

Compile Results into Tables 4, 5, 6 11

Compare Clustered and Non Clustered SE: 13

Extension: Logistic Regression 15

Logistic Regression 16

Extension: R Squared 21

R Squared 21

R Squared Plots 24

Issues 33

Works Cited: Original Paper 33

Works Cited: Cited Paper 34

Background

For our final project, we chose to reproduce and extend the main results in the paper "Housing, Health, and Happiness" published in the American Economic Journal: Economic Policy in 2009 by Matias D.Cattaneo, Sebastian Galiani, Paul J. Gertler, Sebastian Martinez, and Rocio Titiunik. In their paper, they explore the impact of Piso Firme, a program in Mexico that replaced dirt floors with cement floor, on child health and adult happiness. They found significant improvements in child health in homes affected by Piso Firme, as measured by parasite counts, diarrhea, anemia, the MacArthur Communicative Development Test score,

the Picture Peabody Vocabulary Test percentile score, height, and weight. Furthermore, they also found significant improvements in adult happiness as measured by satisfaction with floor quality, satisfaction with house quality, satisfaction with quality of life, depression scale, and perceived stress scale. The goal of Piso Firme is to provide better living standards to people in Mexico through cheaper means in relation to other social programs. This paper showed that Piso Firme does meet its intended goal.

Data

The data for this paper comes from surveys of control and treatment groups that the team conducted with the Mexican National Institute of Public Health in 2005 as well as the 2000 Mexican census, vital statistic mortality files, and the 1994-2000 household surveys. Since this was a natural experiment and the assignment of the homes into treatment or control groups was not done experimentally, the authors had to give special care to make sure that the treatment and control groups did not differ besides with regards to whether or not they received the treatment. The authors noticed that there is an urban area that is shared between two states, one of which implemented Piso Firme before the study was conducted and the other which did not. Since this was one urban space, however, the demographic and socioeconomic status of its residents seemed to be relatively uniform regardless of which state the individual home belonged to. The researchers took advantage of this characteristic to conduct a natural experiment, in which the homes in the urban space that received Piso Firme were part of the treatment group, and those that did not were part of the control. In order to further confirm the pre-treatment health, socioeconomic status, and demographic makeup of the groups, the researchers specifically looked at mean difference between numerous variables such as: average number of rooms per household, proportion of houses with no gas heater, average overcrowding index, and age. The clustered standard errors were also calculated at the census block level. Using these values they tested for significance in difference between the variables. They found that very few of these variables differed significantly, and confirmed that there were in fact no significant differences between the groups before the treatment.

Import Data

Read in the two data files used for replication of the main results. The houshold dataset has information at the houshold level and includes data from both the 2000 Mexican Census and the 2005 Survey. The individual dataset has information at the individual level and includes data from the 2005 Survey.

```
household_dat <- read_dta(file = "data/PisoFirme_AEJPol-20070024_household.dta")
individual_dat <- read_dta(file = "data/PisoFirme_AEJPol-20070024_individual.dta")
```

Divides the data into treatment and control groups.

```
household_treatment <- household_dat %>% filter(dpisofirme == 1)
household_control <- household_dat %>% filter(dpisofirme == 0)
individual_treatment <- individual_dat %>% filter(dpisofirme == 1)
individual_control <- individual_dat %>% filter(dpisofirme == 0)
```

Reproduction:

The main part of the paper, as mentioned earlier, was to show that Piso Firme had significant impacts on children's health and on adult happiness as measured by a few specific variables present in the data set. In order to do this, the researchers conducted linear regressions with three different models, one where the only independent variable was a dummy variable indicating whether or not the house received the treatment, another where the independent variable also included age, demographic, and health habits, and a third that also included social programs. The dependent variables varied between all of the indicator variables for child health and adult happiness. The researchers also included a table, Table 4, where they

conducted this same regression model, but the dependent variables indicated the share of rooms with cement floors, and dummy variables indicating whether or not there is a cement floor in the kitchen, dining room, bathroom, and bedroom. The point of this was to show that Piso Firme did in fact install cement floors, since those receiving the treatment did not have to accept it and the researchers conducted an intent to treat analysis. They found that Piso Firme did lead to significant increases in all of these measures. For all of the regressions, the researchers used three models. The first had no controls, the second had age, demographic, and health-habit controls, and the third had age, demographic, health-habit, and public social programs controls. It is important to note that here, the researchers use regression to determine statistical significance. For each dependent variable, they report its coefficient as well as its clustered standard error at census-block level, which was the measure used to determine significance. Finally, they also reported one hundred times the coefficient divided by the control mean, a measure roughly representing the percentage change in the measure of the dependent variable as a result of the mean. In Table 5, which showed children's health, they found significant results for all variables except height, weight, and Picture Peabody Vocabulary Test for model 1. In Table 6, which shows adult happiness, they found significant results for all variables. Thus, overall, the researchers were able to show, since the control and treatment groups appeared to be similar besides the treatment, that Piso Firme does improve child health and adult happiness in a significant way.

Model 1: no controls

Here, we fit linear models, varying the dependent variable and extracting the correlation coefficient as well as both the clustered and non-clustered standard errors.

```
# function for individual data set
model_1_i <- function(dependent,cluster=T) {
  data_updated<-individual_dat%>%filter(!is.na(dependent) & !is.na(individual_dat$idcluster))
  dependent_updated <- dependent[!is.na(dependent)& !is.na(individual_dat$idcluster)]
  if(cluster==T){
  return(tidy(lm_robust(dependent_updated ~ dpisofirme,data_updated,clusters=idcluster)))}
  else{
    return(tidy(lm_robust(dependent_updated ~ dpisofirme,data_updated)))
  }
}
# function for household data set
 model_1_hh <- function(dependent,cluster=T) {
  data_updated<-household_dat%>%filter(!is.na(dependent) & !is.na(household_dat$idcluster))
dependent_updated <- dependent[!is.na(dependent)& !is.na(household_dat$idcluster)]
  if(cluster==T){
  return(tidy(lm_robust(dependent_updated ~ dpisofirme,data_updated,clusters=idcluster)))}
  else{
    return(tidy(lm_robust(dependent_updated ~ dpisofirme,data_updated)))
  }
 }

 #coefficient:  $estimate[2] for standard error: $std.error[2]
 #for non clustered std error make argument false:$std.error[2]

# caluclates coefficients for each dependent variable
model_1_coeff <- c(model_1_hh(household_dat$S_shcementfloor)$estimate[2],model_1_hh(household_dat$S_ceme
# calculates clustered standard errors for each dependent variable
model_1_std_error_clustered<- c(model_1_hh(household_dat$S_shcementfloor)$std.error[2],model_1_hh(housel
# calculates non-clusted standard errors for each dependent variable
model_1_std_error<- c(model_1_hh(household_dat$S_shcementfloor,cluster=F)$std.error[2],model_1_hh(housel
variables <- c("share_cement_floors", "kitchen", "dining_room", "bathroom", "bedroom", "parasite","diar
```

```
Model_1 <- data.frame(var = variables,coeff_1 = model_1_coeff,sce_1 = model_1_std_error_clustered,se_1=r
```

Model 2: age, demographic, and health-habit controls

Here, we fit linear models with age, demographic, and health-habit controls, varying the dependent variable and extracting the correlation coefficient as well as both the clustered and non-clustered standard errors.

```
# control variables, set na to 0
individual_dat$S_HHpeople[is.na(individual_dat$S_HHpeople)]<- 0
individual_dat$S_rooms[is.na(individual_dat$S_rooms)]<- 0
individual_dat$S_age[is.na(individual_dat$S_age)]<- 0
individual_dat$S_gender[is.na(individual_dat$S_gender)]<- 0
individual_dat$S_childma[is.na(individual_dat$S_childma)]<- 0
individual_dat$S_childmaage[is.na(individual_dat$S_childmaage)]<- 0
individual_dat$S_childmaeduc[is.na(individual_dat$S_childmaeduc)]<- 0
individual_dat$S_childpa[is.na(individual_dat$S_childpa)]<- 0
individual_dat$S_childpaage[is.na(individual_dat$S_childpaage)]<- 0
individual_dat$S_childpaeduc[is.na(individual_dat$S_childpaeduc)]<- 0
individual_dat$S_waterland[is.na(individual_dat$S_waterland)]<- 0
individual_dat$S_waterhouse[is.na(individual_dat$S_waterhouse)]<- 0
individual_dat$S_electricity[is.na(individual_dat$S_electricity)]<- 0
individual_dat$S_hasanimals[is.na(individual_dat$S_hasanimals)]<- 0
individual_dat$S_animalsinside[is.na(individual_dat$S_animalsinside)]<- 0
individual_dat$S_garbage[is.na(individual_dat$S_garbage)]<- 0
individual_dat$S_washhands[is.na(individual_dat$S_washhands)]<- 0
# function for individual data set
model_2_i <- function(dependent,cluster=T) {
  # removes entries with na values
  data_updated<-individual_dat%>%filter(!is.na(dependent) & !is.na(individual_dat$idcluster))
  # control variables
  x1<- data_updated$S_HHpeople
  x2<-data_updated$S_rooms
  x3<-data_updated$S_age
  x4<-data_updated$S_gender
  x5<-data_updated$S_childma
  x6<-data_updated$S_childmaage
  x7<-data_updated$S_childmaeduc
  x8<-data_updated$S_childpa
  x9<-data_updated$S_childpaage
  x10<-data_updated$S_childpaeduc
  x11<-data_updated$S_waterland
  x12<-data_updated$S_waterhouse
  x13<-data_updated$S_electricity
  x14<-data_updated$S_hasanimals
  x15<-data_updated$S_animalsinside
  x16<-data_updated$S_garbage
  x17<-data_updated$S_washhands
  x18<- data_updated$dpisofirme
  updated_dependent<- dependent[!is.na(dependent)& !is.na(individual_dat$idcluster)]
  if(cluster==T)
  {
    return(tidy(lm_robust(updated_dependent ~ x18 + x1 + x2 + x3  + x4+ x5+ x6+ x7+ x8+ x9+ x10+ x11+ x
  }else{
```

```
    return(tidy(lm_robust(updated_dependent ~ x18 + x1 + x2 + x3  + x4+ x5+ x6+ x7+ x8+ x9+ x10+ x11+ x
  }
}
# control variables, set na to 0
household_dat$S_HHpeople[is.na(household_dat$S_HHpeople)]<-0
household_dat$S_headage[is.na(household_dat$S_headage)]<-0
household_dat$S_spouseage[is.na(household_dat$S_spouseage)]<-0
household_dat$S_headeduc[is.na(household_dat$S_headeduc)]<-0
household_dat$S_spouseeduc[is.na(household_dat$S_spouseeduc)]<-0
household_dat$S_dem1[is.na(household_dat$S_dem1)]<-0
household_dat$S_dem2[is.na(household_dat$S_dem2)] <-0
household_dat$S_dem3[is.na(household_dat$S_dem3)]<-0
household_dat$S_dem4[is.na(household_dat$S_dem4)] <-0
household_dat$S_dem5[is.na(household_dat$S_dem5)]<-0
household_dat$S_dem6[is.na(household_dat$S_dem6)]<-0
household_dat$S_dem7[is.na(household_dat$S_dem7)] <-0
household_dat$S_dem8[is.na(household_dat$S_dem8)]<-0
household_dat$S_waterland[is.na(household_dat$S_waterland)]<-0
household_dat$S_waterhouse[is.na(household_dat$S_waterhouse)]<-0
household_dat$S_electricity[is.na(household_dat$S_electricity)]<-0
household_dat$S_hasanimals[is.na(household_dat$S_hasanimals)]<-0
household_dat$S_animalsinside[is.na(household_dat$S_animalsinside)]<-0
household_dat$S_garbage[is.na(household_dat$S_garbage)]<-0
household_dat$S_washhands[is.na(household_dat$S_washhands)]<-0
# function for household data set
model_2_hh <- function(dependent,cluster=T) {
  # removes entries with na values
  data_updated<-household_dat%>%filter(!is.na(dependent)&!is.na(idcluster))
  # control variables
  x1<- data_updated$S_HHpeople
  x2<-data_updated$S_headage
  x3<-data_updated$S_spouseage
  x4<-data_updated$S_headeduc
  x5<-data_updated$S_spouseeduc
  x6<-data_updated$S_dem1
  x7<-data_updated$S_dem2
  x8<-data_updated$S_dem3
  x9<-data_updated$S_dem4
  x10<-data_updated$S_dem5
  x11<-data_updated$S_dem6
  x12<-data_updated$S_dem7
  x13<-data_updated$S_dem8
  x14<-data_updated$S_waterland
  x15<-data_updated$S_waterhouse
  x16<-data_updated$S_electricity
  x17<-data_updated$S_hasanimals
  x18<-data_updated$S_animalsinside
  x19<-data_updated$S_garbage
  x20<-data_updated$S_washhands
  x21<-data_updated$dpisofirme
  updated_dependent<- dependent[!is.na(dependent)& !is.na(household_dat$idcluster)]
  if(cluster==T)
  {
```

```
    return(tidy(lm_robust(updated_dependent ~ x21 + x1 + x2 + x3  + x4+ x5+ x6+ x7+ x8+ x9+ x10+ x11+ x
  }else{
    return(tidy(lm_robust(updated_dependent ~ x21 + x1 + x2 + x3  + x4+ x5+ x6+ x7+ x8+ x9+ x10+ x11+ x
  }
}
# caluclates coefficients for each dependent variable
model_2_coeff <- c(model_2_hh(household_dat$S_shcementfloor)$estimate[2],model_2_hh(household_dat$S_ceme
# caluclates clustered standard errors for each dependent variable
model_2_std_error_clustered<- c(model_2_hh(household_dat$S_shcementfloor)$std.error[2],model_2_hh(housel
# caluclates non-clustered standard errors for each dependent variable
model_2_std_error<- c(model_2_hh(household_dat$S_shcementfloor,cluster=F)$std.error[2],model_2_hh(housel
Model_2 <- data.frame(var = variables,coeff_2 = model_2_coeff,sce_2 = model_2_std_error_clustered,se_2=r
```

Model 3: age, demographic, health-habit and public social programs controls

Here, we fit linear models with age, demographic, healh-habit, and public social programs controls, varying the dependent variable and extracting the correlation coefficient as well as both the clustered and non-clustered standard errors.

```
# additional control variables, set na to 0
individual_dat$S_cashtransfers[is.na(individual_dat$S_cashtransfers)]<- 0
individual_dat$S_milkprogram[is.na(individual_dat$S_milkprogram)]<- 0
individual_dat$S_foodprogram[is.na(individual_dat$S_foodprogram)]<- 0
individual_dat$S_seguropopular[is.na(individual_dat$S_seguropopular)]<- 0
# function for individual data set
model_3_i <- function(dependent,cluster=T) {
  # removes entries with na values
   data_updated<-individual_dat%>%filter(!is.na(dependent) & !is.na(individual_dat$idcluster))
   # control variables
   x1<- data_updated$S_HHpeople
   x2<-data_updated$S_rooms
   x3<-data_updated$S_age
   x4<-data_updated$S_gender
   x5<-data_updated$S_childma
   x6<-data_updated$S_childmaage
   x7<-data_updated$S_childmaeduc
   x8<-data_updated$S_childpa
   x9<-data_updated$S_childpaage
   x10<-data_updated$S_childpaeduc
   x11<-data_updated$S_waterland
   x12<-data_updated$S_waterhouse
   x13<-data_updated$S_electricity
   x14<-data_updated$S_hasanimals
   x15<-data_updated$S_animalsinside
   x16<-data_updated$S_garbage
   x17<-data_updated$S_washhands
   x18<-data_updated$S_cashtransfers
   x19<-data_updated$S_milkprogram
   x20<-data_updated$S_foodprogram
   x21<-data_updated$S_seguropopular
   x22<- data_updated$dpisofirme
```

```
   updated_dependent<- dependent[!is.na(dependent)& !is.na(individual_dat$idcluster)]
  if(cluster==T)
  {
    return(tidy(lm_robust(updated_dependent ~ x22 + x1 + x2 + x3  + x4+ x5+ x6+ x7+ x8+ x9+ x10+ x11+ x
  }else{
    return(tidy(lm_robust(updated_dependent ~ x22 + x1 + x2 + x3  + x4+ x5+ x6+ x7+ x8+ x9+ x10+ x11+ x
  }
}

# additional control variables, set na to 0
household_dat$S_cashtransfers[is.na(household_dat$S_cashtransfers)]<- 0
household_dat$S_milkprogram[is.na(household_dat$S_milkprogram)]<- 0
household_dat$S_foodprogram[is.na(household_dat$S_foodprogram)]<- 0
household_dat$S_seguropopular[is.na(household_dat$S_seguropopular)]<- 0
# function for household data set
model_3_hh <- function(dependent,cluster=T) {
  data_updated <- household_dat%>%filter(!is.na(dependent) & !is.na(household_dat$idcluster))
  x1<- data_updated$S_HHpeople
  x2<-data_updated$S_headage
  x3<-data_updated$S_spouseage
  x4<-data_updated$S_headeduc
  x5<-data_updated$S_spouseeduc
  x6<-data_updated$S_dem1
  x7<-data_updated$S_dem2
  x8<-data_updated$S_dem3
  x9<-data_updated$S_dem4
  x10<-data_updated$S_dem5
  x11<-data_updated$S_dem6
  x12<-data_updated$S_dem7
  x13<-data_updated$S_dem8
  x14<-data_updated$S_waterland
  x15<-data_updated$S_waterhouse
  x16<-data_updated$S_electricity
  x17<-data_updated$S_hasanimals
  x18<-data_updated$S_animalsinside
  x19<-data_updated$S_garbage
  x20<-data_updated$S_washhands
  x21<- data_updated$dpisofirme
  x22<-data_updated$S_cashtransfers
  x23<-data_updated$S_milkprogram
  x24<-data_updated$S_foodprogram
  x25<-data_updated$S_seguropopular
  updated_dependent<- dependent[!is.na(dependent)& !is.na(household_dat$idcluster)]
  if(cluster==T)
  {
    return(tidy(lm_robust(updated_dependent ~ x21 + x1 + x2 + x3  + x4+ x5+ x6+ x7+ x8+ x9+ x10+ x11+ x
  }else{
    return(tidy(lm_robust(updated_dependent ~ x21 + x1 + x2 + x3  + x4+ x5+ x6+ x7+ x8+ x9+ x10+ x11+ x
  }
}

# caluclates coefficients for each dependent variable
model_3_coeff <- c(model_3_hh(household_dat$S_shcementfloor)$estimate[2],model_3_hh(household_dat$S_ceme
```

```
# caluclates clustered standard errors for each dependent variable
model_3_std_error_clustered<- c(model_3_hh(household_dat$S_shcementfloor)$std.error[2],model_3_hh(housel
# caluclates non-clustered standard errors for each dependent variable
model_3_std_error<- c(model_3_hh(household_dat$S_shcementfloor,cluster=F)$std.error[2],model_3_hh(housel
Model_3 <- data.frame(var = variables,coeff_3 = model_3_coeff,sce_3 = model_3_std_error_clustered,se_3=r
```

Control Group Means and Standard Deviations

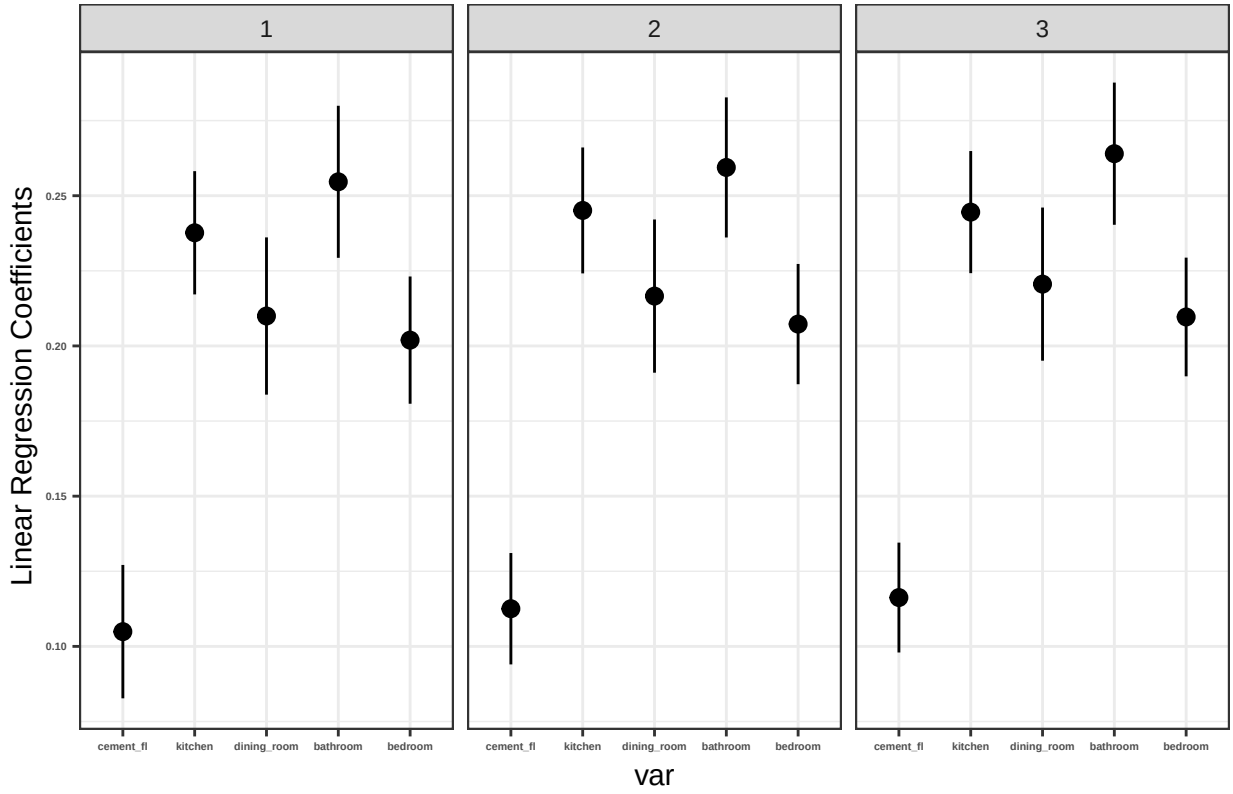
Calculates control group means and standard deviations, which are used as to understand the proportional impact of the dependent variable.

```
# function to calculate control mean
control_mean <- function(dependent) {
  updated_dependent<- dependent[!is.na(dependent)]
  return(mean(updated_dependent,na.rm=T))
}
# function to calculate control standard deviation
control_sd <- function(dependent) {
  updated_dependent<- dependent[!is.na(dependent)]
  return(sd(updated_dependent,na.rm=T))
}
# computes control mean for each dependent variable
control_mean <- c(control_mean(household_control$S_shcementfloor), control_mean(household_control$S_ceme
# computes control standard deviation for each dependent variable
control_sd <- c(control_sd(household_control$S_shcementfloor), control_sd(household_control$S_cementflo
Mean_SD <- data.frame(var = variables, control_group_mean = control_mean, control_group_sd = control_sd)
```

Here we have graphical representations of the regression coefficients for each dependent variables separated by model and table. We used standard clustered error for the error bars.

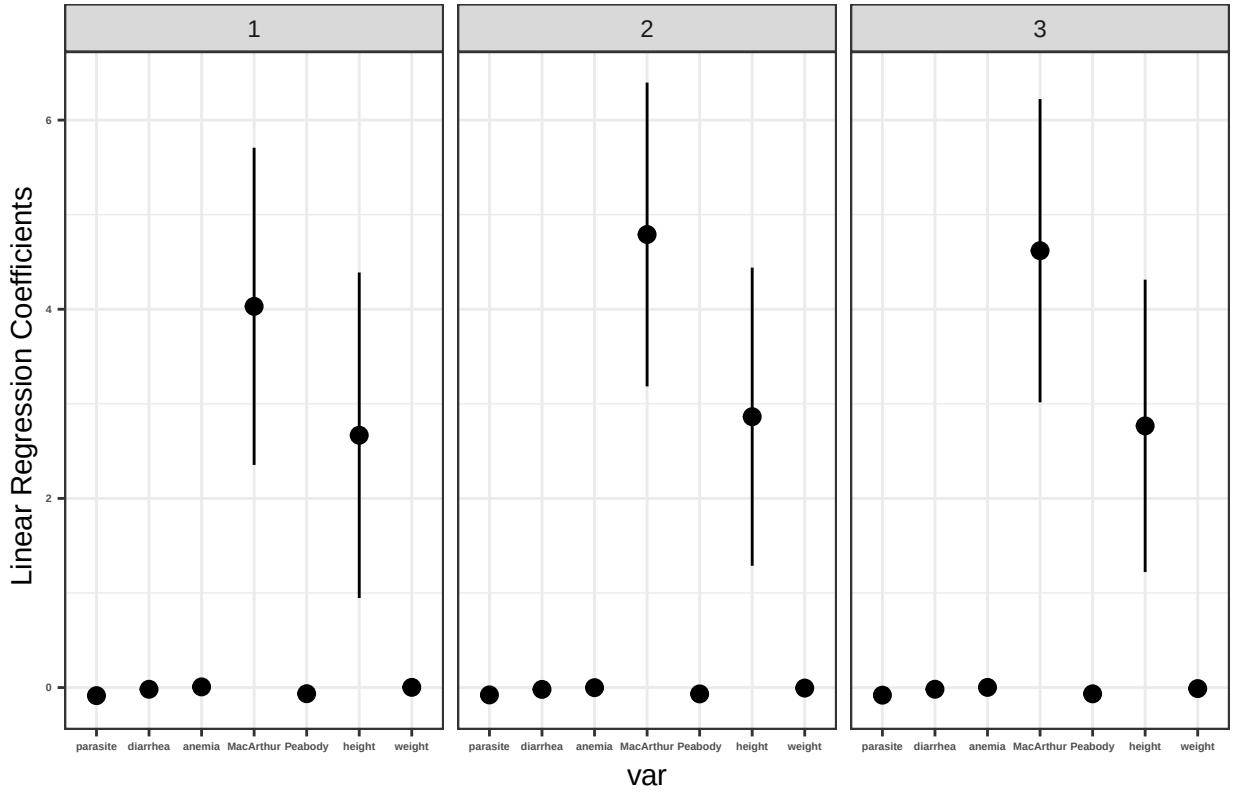
```
#create a df so we can facet_grid by model type
graphing<-rbind(Model_1%>%mutate(model_type=1)%>%select(coeff_1,sce_1,var,model_type)%>%rename(coeff=coe
#visualizations of coefficients for each model, grouped by dependent variable type (effectiveness, adul
#corresponds to tables 4-6

ggplot(graphing[c(1:5,18:22,35:39),],aes(x=var,y=coeff))+
  geom_pointrange(aes(ymin=coeff-sce,ymax=coeff+sce))+scale_x_discrete(labels=c("cement_fl",variables[2
```

Table 4 Regression Coefficents by Model

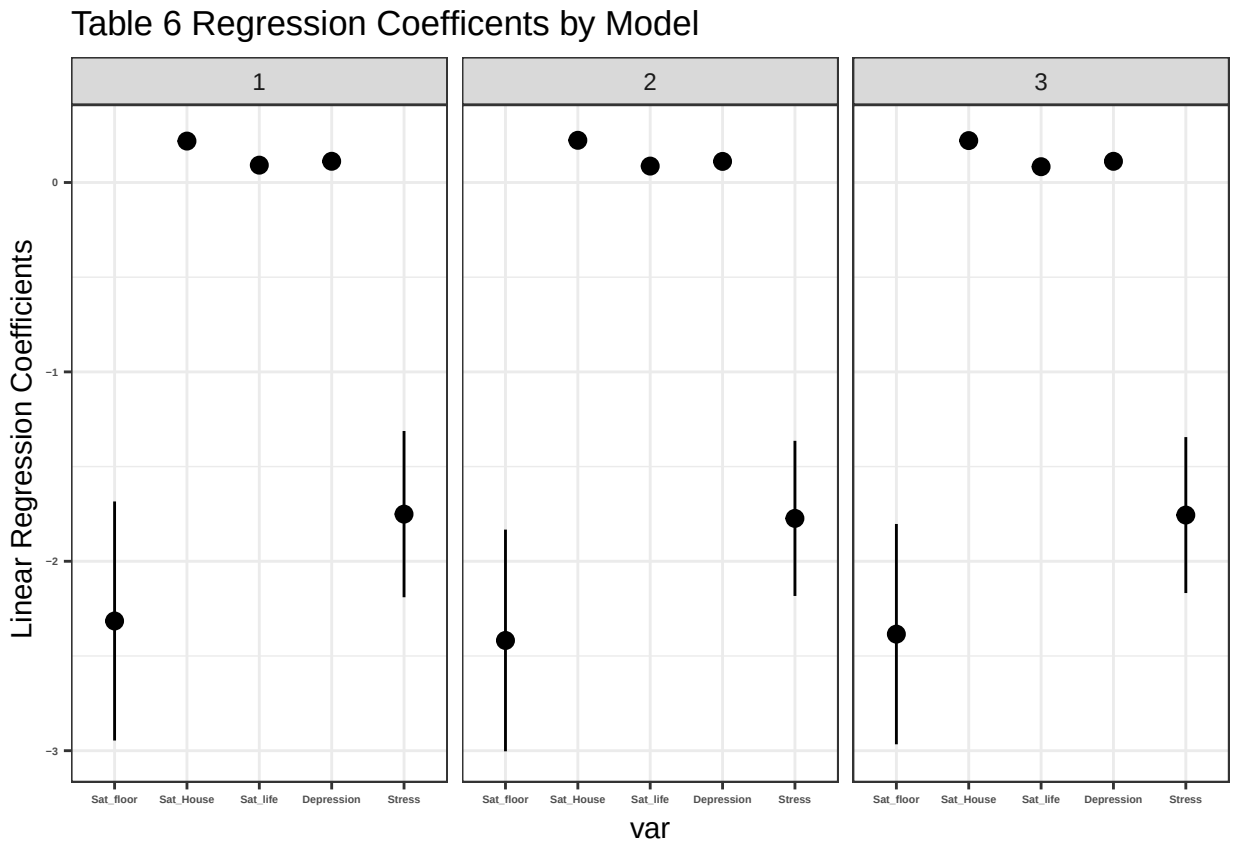



```
ggplot(graphing[c(6:12,23:29,40:46),],aes(x=var,y=coeff))+
  geom_pointrange(aes(ymin=coeff-sce,ymax=coeff+sce))+scale_x_discrete(labels=c(variables[6:12]))+facet
```

Table 5 Regression Coefficents by Model




```
ggplot(graphing[c(13:17,30:34,47:51),],aes(x=var,y=coeff))+
  geom_pointrange(aes(ymin=coeff-sce,ymax=coeff+sce))+scale_x_discrete(labels=c(variables[13:17]))+face
```



Compile Results into Tables 4, 5, 6

Organizes above results into Tables 4, 5, and 6 as in the paper.

```
Model <- Model_1 %>% left_join(Model_2, by = "var") %>% left_join(Model_3, by = "var") %>% left_join(Mea



Table_4 <- Model %>% filter(var == "share_cement_floors" | var == "kitchen" | var == "dining_room" | var



Table_5 <- Model %>% filter(var == "parasite" | var == "diarrhea" | var == "anemia" | var == "MacArthur

Table_6 <- Model %>% filter(var == "Sat_floor" | var == "Sat_house" | var == "Sat_life" | var == "Depres

Table_4
```

```
##            Dependent control_group_mean control_group_sd   coeff_1
## 1 share_cement_floors          0.7277989        0.3628952 0.2019351
## 2             kitchen          0.6712132        0.4699410 0.2546311
## 3         dining_room          0.7085427        0.4545968 0.2099595
## 4            bathroom          0.8025844        0.3981916 0.1049046
## 5             bedroom          0.6676238        0.4712342 0.2376625
##        sce_1 coef_mean_1   coeff_2      sce_2 coef_mean_2   coeff_3
```

```
## 1 0.02117636   0.2774600 0.2072762 0.02001508   0.2847988 0.2096413
## 2 0.02530671   0.3793595 0.2594093 0.02331204   0.3864782 0.2639879
## 3 0.02616822   0.2963258 0.2165915 0.02549685   0.3056859 0.2205735
## 4 0.02219845   0.1307085 0.1125345 0.01853916   0.1402151 0.1162511
## 5 0.02051527   0.3559827 0.2450988 0.02095768   0.3671211 0.2445425
##        sce_3 coef_mean_3
## 1 0.01975315   0.2880484
## 2 0.02366543   0.3932996
## 3 0.02548561   0.3113059
## 4 0.01828329   0.1448460
## 5 0.02032116   0.3662878
```

```
Table_5
```

```
##   Dependent control_group_mean control_group_sd      coeff_1       sce_1
## 1  parasite          0.3326948        0.6733949 -0.065024607 0.032974232
## 2  diarrhea          0.1420428        0.3491769 -0.018208558 0.009472691
## 3    anemia          0.4259354        0.4946108 -0.085437680 0.028603484
## 4 MacArthur         13.3543046       18.9523874  4.030575089 1.676726842
## 5   Peabody         30.6560588       24.8641493  2.667587219 1.720902576
## 6    height         -0.6047102        1.1041380  0.007021171 0.043429621
## 7    weight          0.1250194        1.1325997  0.002237342 0.034795771
##   coef_mean_1      coeff_2       sce_2 coef_mean_2     coeff_3       sce_3
## 1 -0.19544824 -0.067313764 0.031319469 -0.20232890 -0.06638076 0.031732065
## 2 -0.12819069 -0.019470531 0.009529533 -0.13707514 -0.01797303 0.009762238
## 3 -0.20058834 -0.077757313 0.027259949 -0.18255658 -0.08035310 0.027255742
## 4  0.30181842  4.790327482 1.606184878  0.35871036  4.61881271 1.603381128
## 5  0.08701664  2.863684903 1.576153745  0.09341334  2.76698293 1.545689983
## 6 -0.01161080 -0.001151036 0.041476572  0.00190345  0.00163365 0.042119849
## 7  0.01789595 -0.005480864 0.035778246 -0.04384009 -0.01093887 0.036315855
##    coef_mean_3
## 1 -0.199524501
## 2 -0.126532513
## 3 -0.188650909
## 4  0.345866957
## 5  0.090258926
## 6 -0.002701541
## 7 -0.087497332
```

```
Table_6
```

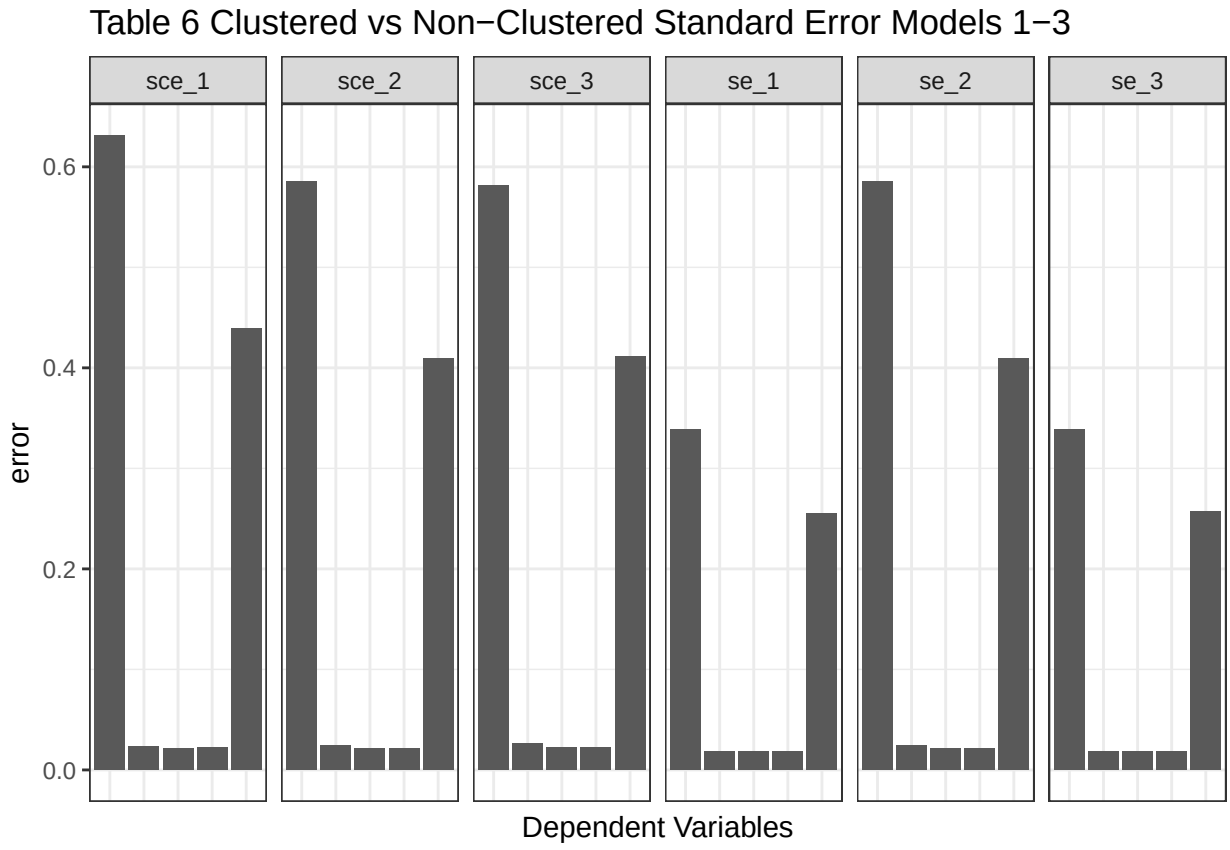
```
##    Dependent control_group_mean control_group_sd    coeff_1      sce_1
## 1  Sat_floor          0.5111271        0.5000557  0.2186820 0.02364742
## 2   Sat_life          0.6008615        0.4898972  0.1120607 0.02213988
## 3 Depression         18.5324207        9.4020791 -2.3152863 0.63116967
## 4     Stress         16.5140591        6.9140506 -1.7509980 0.43885790
##   coef_mean_1    coeff_2      sce_2 coef_mean_2    coeff_3      sce_3
## 1   0.4278428  0.2228228 0.02486332   0.4359440  0.2213322 0.02629294
## 2   0.1865001  0.1113385 0.02159705   0.1852981  0.1118668 0.02268118
## 3  -0.1249317 -2.4180563 0.58543276  -0.1304771 -2.3846136 0.58158648
## 4  -0.1060307 -1.7735720 0.40959517  -0.1073977 -1.7559568 0.41165289
##   coef_mean_3
## 1   0.4330278
## 2   0.1861774
```

```
## 3  -0.1286725
## 4  -0.1063310
```

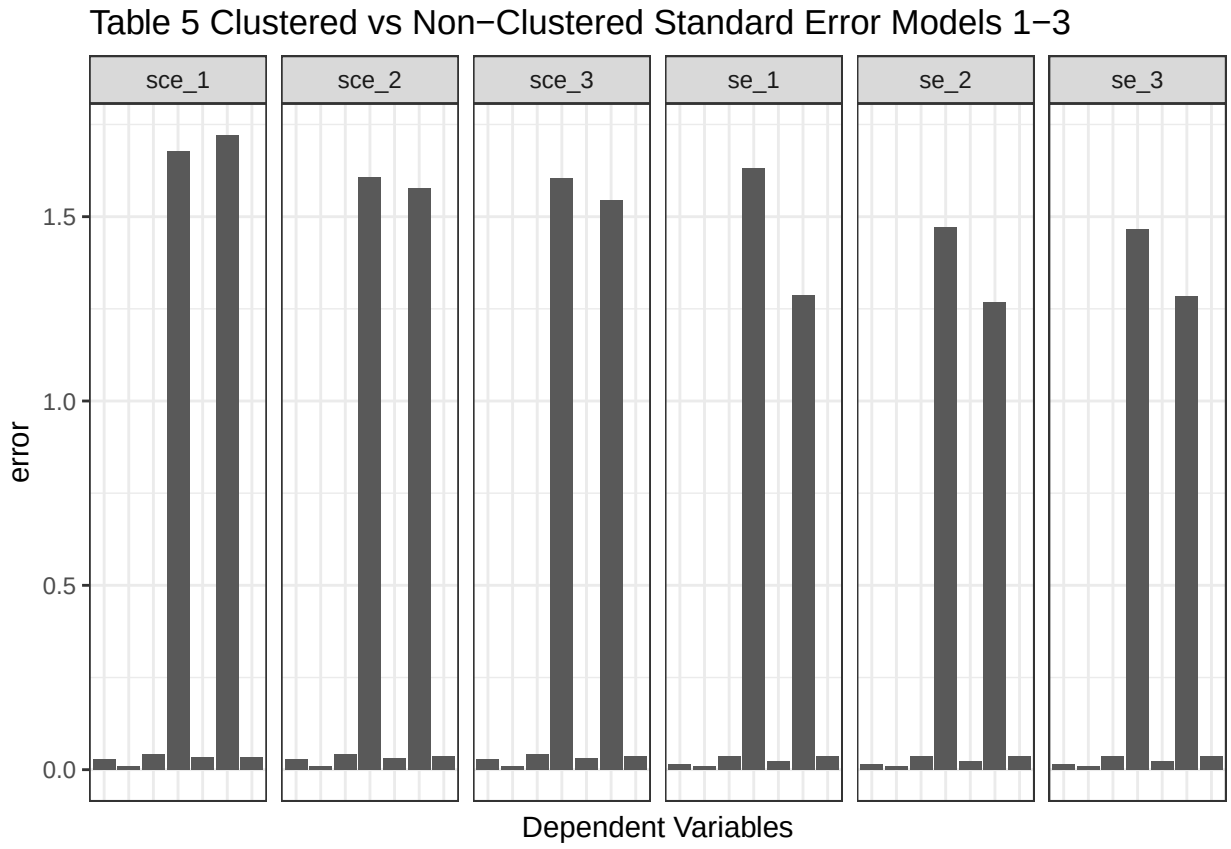
Compare Clustered and Non Clustered SE:

Housing and Happiness made use of clustered standard errors, clustering at the census block level. Miller explains the need for clustering by region, "Then model errors for individuals in the same region may be correlated, while model errors for individuals in different regions are assumed to be uncorrelated" (Cameron and Miller 2). He also notes that a failure to control for any cluster correlations would result in misleadingly small standard errors. As part of our extension we provide both the clustered standard error and non-clustered standard error, which as Miller predicted was much smaller than the clustered standard error. This indicates that there is a correlation among individuals at the census block level, and thus Cattaneo et al's use of standard clustered error is justified. Reporting a misleadingly small standard error would have made the regression performance appear better than it is. It is important to note that we obtained the same clustered standard errors at census block level as determined in the paper, and thus found statistical significance at the same levels for the same dependent variables as in the paper.

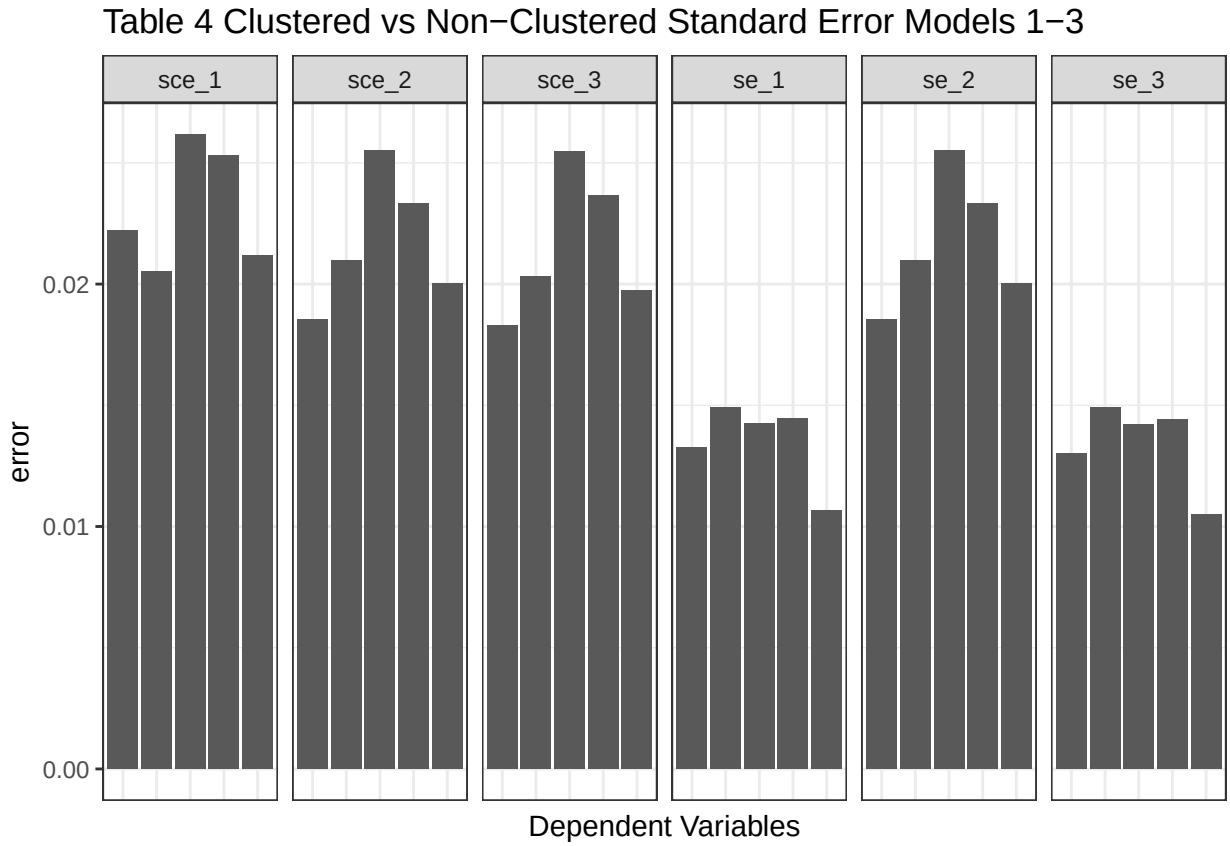
```
errors<-Model%>%select(var,sce_1,se_1,sce_2,se_2,sce_3,se_3)
errors<-errors%>%gather("type","error",2:7)
tb6_err<-errors%>%filter(var == "Sat_floor" | var == "Sat_House" | var == "Sat_life" | var == "Depressi
tb5_err<-errors%>% filter(var == "parasite" | var == "diarrhea" | var == "anemia" | var == "MacArthur"
tb4_err<-errors%>% filter(var == "share_cement_floors" | var == "kitchen" | var == "dining_room" | var =

ggplot(tb6_err)+geom_col(aes(x=var,y=error))+facet_grid(~type)+ ggtitle("Table 6 Clustered vs Non-Clust
```



```
ggplot(tb5_err)+geom_col(aes(x=var,y=error))+facet_grid(~type)+ ggtitle("Table 5 Clustered vs Non-Cluste
        axis.ticks.x=element_blank())+xlab("Dependent Variables")
```




```
ggplot(tb4_err)+geom_col(aes(x=var,y=error))+facet_grid(~type)+ ggtitle("Table 4 Clustered vs Non-Cluste
        axis.ticks.x=element_blank())+xlab("Dependent Variables")
```



Extension: Logistic Regression

Housing and Happiness by Cattaneo et al is an example of a natural experiment. In this case the government decided to give the treatment to one town and not another although they are geographically close and socio-economically similar. Although this was not necessarily a naturally random experiment (households were not randomly selected to retrieve the treatment) but instead the towns differ only by an "administrative split" (different states). Additionally, they also used a smallest distance algorithm to only select the control census blocks that closest match the treatment census blocks (using pre Piso Firme 2000 census data). In order to extend our analysis we wanted to use classification see how similar/dissimilar the control census blocks and treatment census blocks were. We used a logistic regression model in order to predict whether or not a census block received Piso Firme, with features from the 2000 census data. In theory, if the two towns were as similar as Cattaneo et al. found in their extensive analysis then we would expect to see an accuracy of about 50% (same as random selecting a census block to receive treatment). Before applying the logistic regression we made sure that every census block was unique because we found that every data entry that had the same census block also contained the exact same 2000 census information, so if we kept all of the raw data for the analysis it gave a false sense of high accuracy. Since this lowered the sample size for the training and testing set we also implemented a 5-fold cross validation. We found the average accuracy: .59 , precision:.66 , and recall: .72 . In addition, we found the variables with the lowest and highest logistic regression coefficients. Typically we saw proportion of household with illiterate members, proportion of household with no water connection outside the house as the highest indicators and proportion of households who lacked a refrigerator and proportion of households with dirt floors as the most negative. The model thus found that areas with a greater percent of illiteracy and no water collection as indicators of receiving piso Firme and greater proportions of households without refrigerators or with dirt floors as the greatest indicators of not getting PisoFirme. It is interesting that having more dirt floors indicates a shift towards not getting the treatment. We also made visuals of those indicators to see how their distributions differed from

treatment and control.

Logistic Regression

Here, we try to see if we can predict whether or not a house recieved the treatment based on the pre-treatment variables. If we can predict whether or not the house received the treatment, we would have evidence to suggest that the treatment and control groups are not relatively equal, as claimed in the paper.

```
set.seed(42)
household_dat$dpisofirme <- factor(household_dat$dpisofirme)
# selects pre-treatment variables
controlled_household <- household_dat %>%group_by(idcluster)%>%select(dpisofirme,C_blocksdirtfloor,C_HH
```

```
## Adding missing grouping variables: `idcluster`
```

```
# Set up 5 fold cross validation using household data
num_folds <- 5
num_rows <- nrow(controlled_household)
frac_train <- 0.8
num_train <- floor(num_rows * frac_train)
ndx <- sample(1:num_rows, num_train, replace=F)
classify<- controlled_household[ndx, ] %>%
  mutate(fold = (row_number() %% num_folds) + 1)
  # do 5-fold cross-validation within each value of
#initiate result lists to average final results
accuracy<-c()
#topterm and bottomterm are the variables with most pos/neg log regression
#coefficients
topterm<-c()
bottomterm<-c()
recall<-c()
precision<-c()
counter<-1
for (f in 1:num_folds) {
    # fit on the training data
    training <- filter(classify, fold != f)
    model <- glm(training$dpisofirme ~., data=training, family = "binomial")
    # evaluate on the validation data
    testing <- filter(classify, fold == f)
    df <- data.frame(actual = testing$dpisofirme, log_odds= predict(model,testing)) %>% mutate(pred = i

# accuracy: correct/total
    acc<-df %>%summarize(acc = mean(pred == actual,na.rm=T))
    accuracy[counter]<-acc[1]
    #precision: true positives/all predicted positives
    prec<-df %>% filter(pred == '1') %>% summarize(prec = mean(actual == '1',na.rm=T))
    precision[counter]<-prec[1]
    rec<-df %>% filter(actual == '1') %>% summarize(recall = mean(pred == '1',na.rm=T))
    #recall: true positives/all actual positives
    recall[counter]<-rec[1]

    modeldf<-tidy(model)
    top<-modeldf%>%arrange(desc(estimate))%>%select(term)
```

```
    bottom<-modeldf%>%arrange(estimate)%>%select(term)
    topterm[counter]<-top[1,1]
    bottomterm[counter]<-bottom[1,1]
    counter<-counter+1
}
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
#calulcates mean of all of the validations
mean(as.numeric(accuracy))
```

```
## [1] 0.5943108
```

```
mean(as.numeric(recall))
```

```
## [1] 0.7201465
```

```
mean(as.numeric(precision))
```

```
## [1] 0.6616667
```

```
topterm
```

```
## [[1]]
## [1] "C_waterland"
##
## [[2]]
## [1] "C_waterland"
##
## [[3]]
## [1] "C_waterland"
##
## [[4]]
## [1] "C_waterland"
##
## [[5]]
## [1] "C_illiterate"
```

```
bottomterm
```

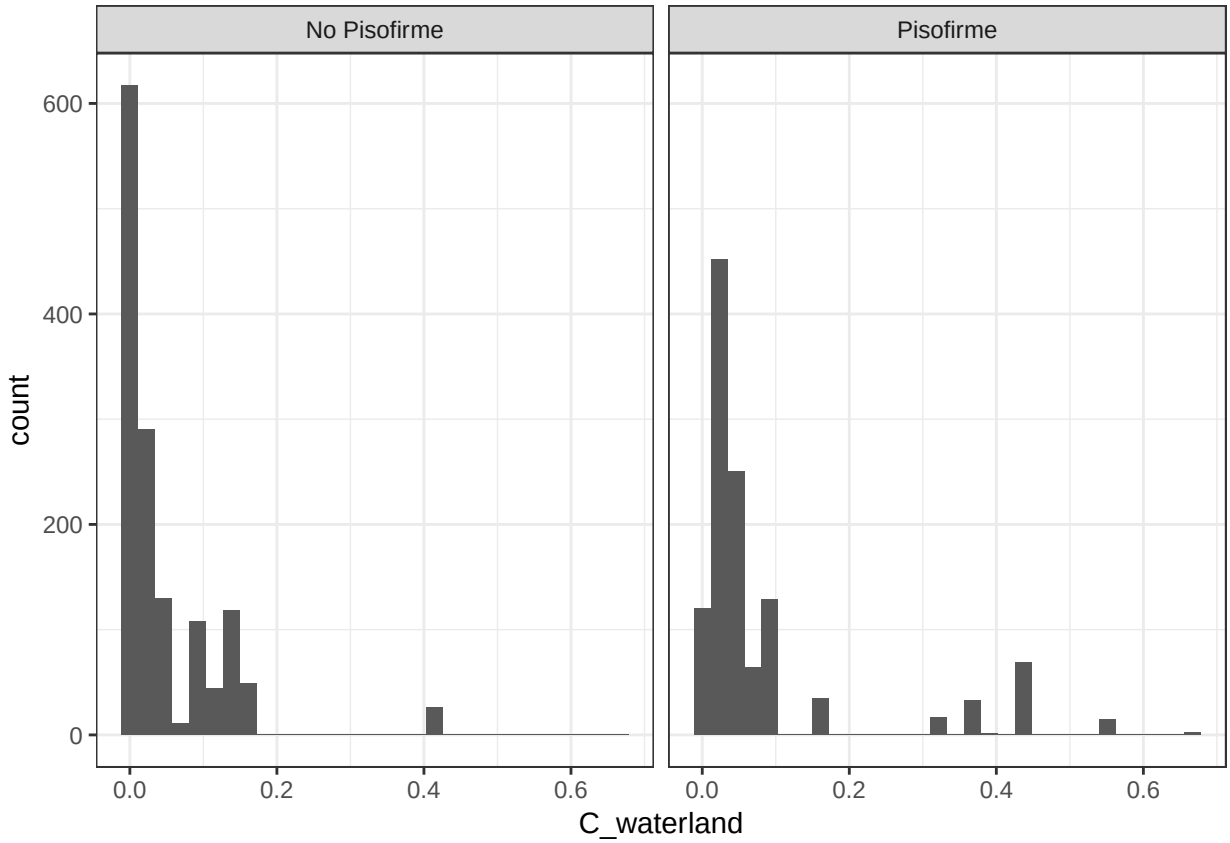
```
## [[1]]
## [1] "C_HHdirtfloor"
##
## [[2]]
## [1] "C_refrigerator"
##
## [[3]]
## [1] "C_HHdirtfloor"
##
## [[4]]
## [1] "C_refrigerator"
##
## [[5]]
## [1] "C_gasheater"
```

Visuals comparing some dependent variables with and without Dpisofirme

```
supp.labs <- c("No Pisofirme", "Pisofirme")
names(supp.labs) <- c(0, 1)
ggplot(household_dat)+geom_histogram(aes(x=C_waterland))+facet_grid(~dpisofirme,labeller=labeller(dpiso
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

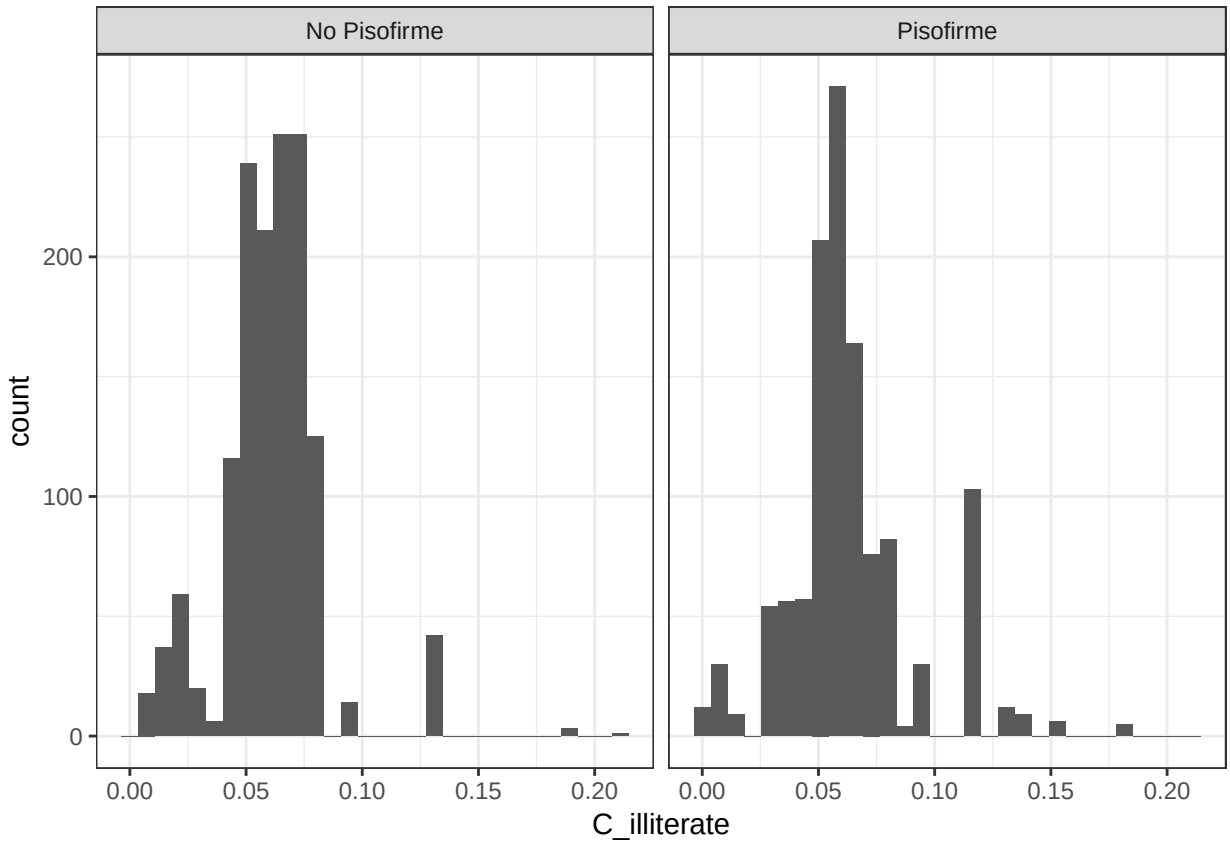
```
## Warning: Removed 203 rows containing non-finite values (stat_bin).
```




```
ggplot(household_dat)+geom_histogram(aes(x=C_illiterate))+facet_grid(~dpisofirme,labeller=labeller(dpis
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

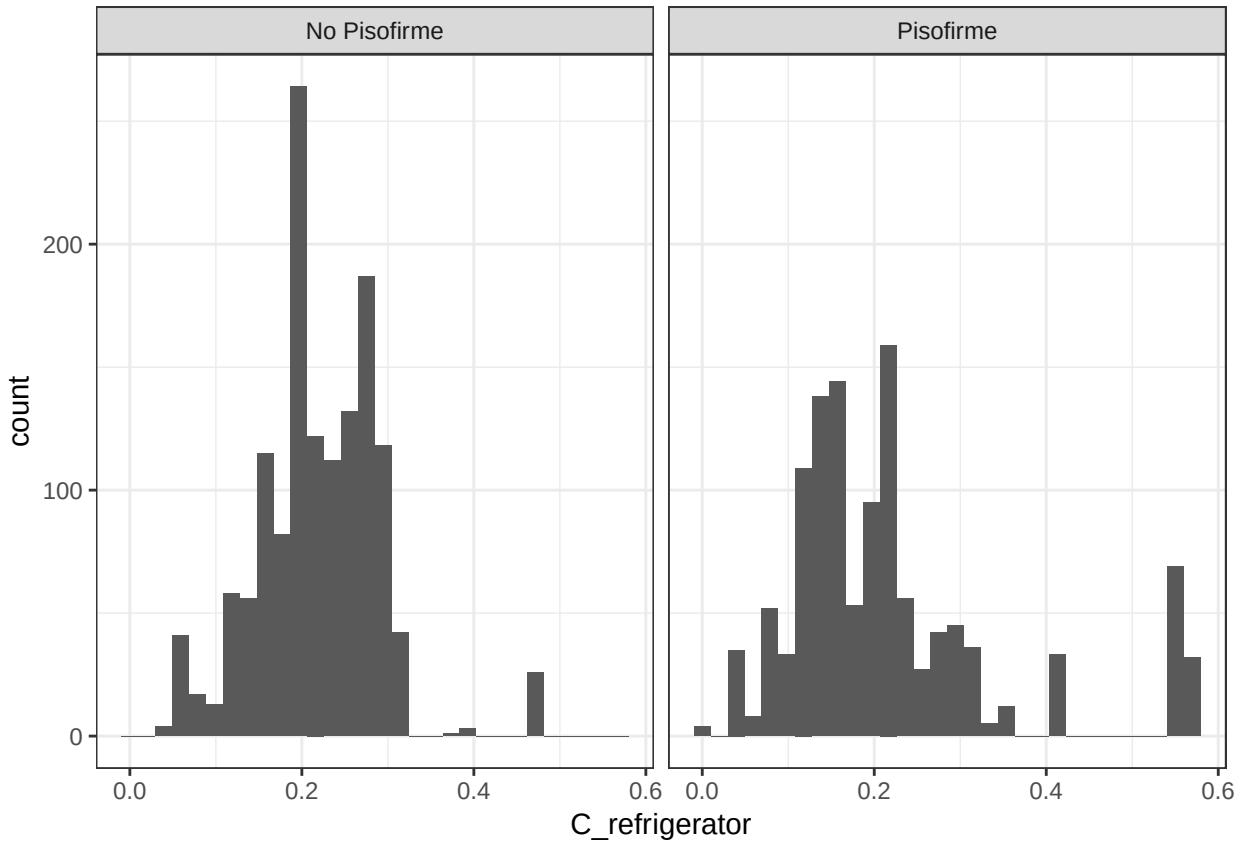
```
## Warning: Removed 203 rows containing non-finite values (stat_bin).
```




```
ggplot(household_dat)+geom_histogram(aes(x=C_refrigerator))+facet_grid(~dpisofirme,labeller=labeller(dp
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

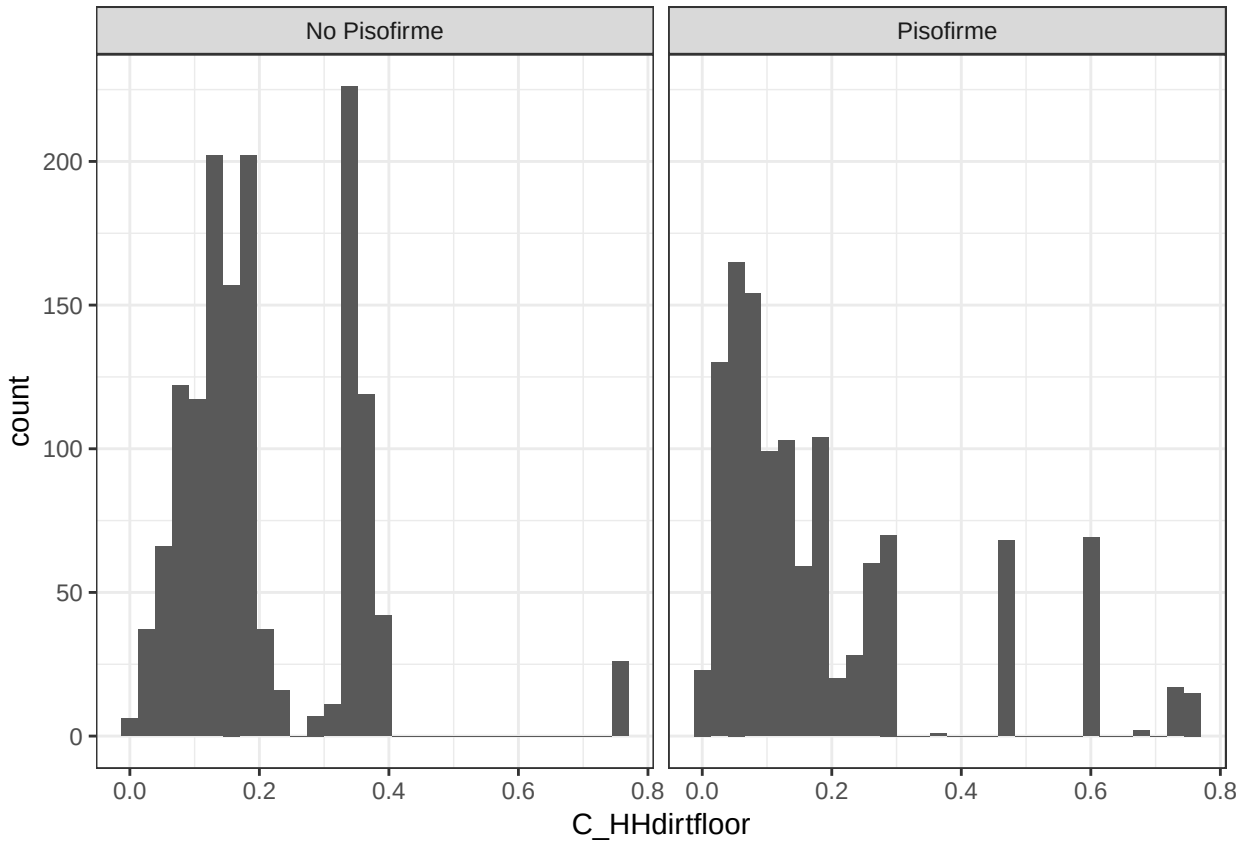
```
## Warning: Removed 203 rows containing non-finite values (stat_bin).
```




```
ggplot(household_dat)+geom_histogram(aes(x=C_HHdirtfloor))+facet_grid(~dpisofirme,labeller=labeller(dpis
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 203 rows containing non-finite values (stat_bin).
```



Extension: R Squared

As a further extension, we calculated R2 for the dependent variables for each of the three tables. This represents the proportion of the variance in the dependent variable that can be explained by the independent variable. Interestingly, we found that for each of the dependent variables, the R2 was relatively low. The highest R2 we found was for share of cement floors for model 3, and we found it to be 0.191. We compared the R2 across the multiple models, and generally saw an increase as the number of control variables in the model increased. Furthermore, we plotted the model coefficients against the R2 values to see if there was a correlation. When we first plotted this, the outliers made it difficult to notice any trends, so we then plotted it without the outliers. Generally, we found that when we disregard the outliers, we can see that coefficients closer to 1 or -1 tend to correspond to higher R2 values, while coefficients close to 0 have smaller R2 values. This is expected since higher absolute values of coefficients indicate that the given dependent variable has a stronger impact on the independent variable, and higher R2 values indicate that a higher proportion of the variance in the dependent variable can be explained by the independent variable.

R Squared

Here, we compute r squared for each dependent variable. This is the amount of change in the dependent variable that can be explained by the independent variable. We calculate this for each model.

```
# function for model 1, individual data set
model_1_i_rsq <- function(dependent) {
  dummy_i <- individual_dat$dpisofirme[!is.na(dependent)]
  dependent_updated <- dependent[!is.na(dependent)]
```

```
  return(summary(lm(dependent_updated ~ dummy_i))$r.squared)
}
# function for model 1, household data set
model_1_hh_rsq <- function(dependent) {
  dummy_hh <- household_dat$dpisofirme[!is.na(dependent)]
  dependent_updated <- dependent[!is.na(dependent)]
  return(summary(lm(dependent_updated ~ dummy_hh ))$r.squared)
}
# function for model 2, individual data set
model_2_i_rsq <- function(dependent) {
  # control variables
  x1<- individual_dat$S_HHpeople[!is.na(dependent)]
  x2<-individual_dat$S_rooms[!is.na(dependent)]
  x3<-individual_dat$S_age[!is.na(dependent)]
  x4<-individual_dat$S_gender[!is.na(dependent)]
  x5<-individual_dat$S_childma[!is.na(dependent)]
  x6<-individual_dat$S_childmaage[!is.na(dependent)]
  x7<-individual_dat$S_childmaeduc[!is.na(dependent)]
  x8<-individual_dat$S_childpa[!is.na(dependent)]
  x9<-individual_dat$S_childpaage[!is.na(dependent)]
  x10<-individual_dat$S_childpaeduc[!is.na(dependent)]
  x11<-individual_dat$S_waterland[!is.na(dependent)]
  x12<-individual_dat$S_waterhouse[!is.na(dependent)]
  x13<-individual_dat$S_electricity[!is.na(dependent)]
  x14<-individual_dat$S_hasanimals[!is.na(dependent)]
  x15<-individual_dat$S_animalsinside[!is.na(dependent)]
  x16<-individual_dat$S_garbage[!is.na(dependent)]
  x17<-individual_dat$S_washhands[!is.na(dependent)]
  x18<- individual_dat$dpisofirme[!is.na(dependent)]
  updated_dependent<- dependent[!is.na(dependent)]
  return(summary(lm( updated_dependent ~ x18 + x1 + x2 + x3  + x4+ x5+ x6+ x7+ x8+ x9+ x10+ x11+ x12+ x
}
# function for model 2, household data set
model_2_hh_sq <- function(dependent) {
  # control variables
  x1<- household_dat$S_HHpeople[!is.na(dependent)]
  x2<-household_dat$S_headage[!is.na(dependent)]
  x3<-household_dat$S_spouseage[!is.na(dependent)]
  x4<-household_dat$S_headeduc[!is.na(dependent)]
  x5<-household_dat$S_spouseeduc[!is.na(dependent)]
  x6<-household_dat$S_dem1[!is.na(dependent)]
  x7<-household_dat$S_dem2[!is.na(dependent)]
  x8<-household_dat$S_dem3[!is.na(dependent)]
  x9<-household_dat$S_dem4[!is.na(dependent)]
  x10<-household_dat$S_dem5[!is.na(dependent)]
  x11<-household_dat$S_dem6[!is.na(dependent)]
  x12<-household_dat$S_dem7[!is.na(dependent)]
  x13<-household_dat$S_dem8[!is.na(dependent)]
  x14<-household_dat$S_waterland[!is.na(dependent)]
  x15<-household_dat$S_waterhouse[!is.na(dependent)]
  x16<-household_dat$S_electricity[!is.na(dependent)]
  x17<-household_dat$S_hasanimals[!is.na(dependent)]
  x18<-household_dat$S_animalsinside[!is.na(dependent)]
```

```
  x19<-household_dat$S_garbage[!is.na(dependent)]
  x20<-household_dat$S_washhands[!is.na(dependent)]
  x21<- household_dat$dpisofirme[!is.na(dependent)]
  updated_dependent<- dependent[!is.na(dependent)]
  return(summary(lm(updated_dependent ~ x21 + x1 + x2 + x3  + x4+ x5+ x6+ x7+ x8+ x9+ x10+ x11+ x12+ x13
}
# function for model 3, individual data set
model_3_i_rsq <- function(dependent) {
  # control varibales
  x1<- individual_dat$S_HHpeople[!is.na(dependent)]
  x2<-individual_dat$S_rooms[!is.na(dependent)]
  x3<-individual_dat$S_age[!is.na(dependent)]
  x4<-individual_dat$S_gender[!is.na(dependent)]
  x5<-individual_dat$S_childma[!is.na(dependent)]
  x6<-individual_dat$S_childmaage[!is.na(dependent)]
  x7<-individual_dat$S_childmaeduc[!is.na(dependent)]
  x8<-individual_dat$S_childpa[!is.na(dependent)]
  x9<-individual_dat$S_childpaage[!is.na(dependent)]
  x10<-individual_dat$S_childpaeduc[!is.na(dependent)]
  x11<-individual_dat$S_waterland[!is.na(dependent)]
  x12<-individual_dat$S_waterhouse[!is.na(dependent)]
  x13<-individual_dat$S_electricity[!is.na(dependent)]
  x14<-individual_dat$S_hasanimals[!is.na(dependent)]
  x15<-individual_dat$S_animalsinside[!is.na(dependent)]
  x16<-individual_dat$S_garbage[!is.na(dependent)]
  x17<-individual_dat$S_washhands[!is.na(dependent)]
  x18<-individual_dat$S_cashtransfers[!is.na(dependent)]
  x19<-individual_dat$S_milkprogram[!is.na(dependent)]
  x20<-individual_dat$S_foodprogram[!is.na(dependent)]
  x21<-individual_dat$S_seguropopular[!is.na(dependent)]
  x22<- individual_dat$dpisofirme[!is.na(dependent)]
  updated_dependent<- dependent[!is.na(dependent)]
  return(summary(lm( updated_dependent ~ x22 + x1 + x2 + x3  + x4+ x5+ x6+ x7+ x8+ x9+ x10+ x11+ x12+ x
}
# function for model 3, household data set
model_3_hh_rsq <- function(dependent) {
  # control variables
  x1<- household_dat$S_HHpeople[!is.na(dependent)]
  x2<-household_dat$S_headage[!is.na(dependent)]
  x3<-household_dat$S_spouseage[!is.na(dependent)]
  x4<-household_dat$S_headeduc[!is.na(dependent)]
  x5<-household_dat$S_spouseeduc[!is.na(dependent)]
  x6<-household_dat$S_dem1[!is.na(dependent)]
  x7<-household_dat$S_dem2[!is.na(dependent)]
  x8<-household_dat$S_dem3[!is.na(dependent)]
  x9<-household_dat$S_dem4[!is.na(dependent)]
  x10<-household_dat$S_dem5[!is.na(dependent)]
  x11<-household_dat$S_dem6[!is.na(dependent)]
  x12<-household_dat$S_dem7[!is.na(dependent)]
  x13<-household_dat$S_dem8[!is.na(dependent)]
  x14<-household_dat$S_waterland[!is.na(dependent)]
  x15<-household_dat$S_waterhouse[!is.na(dependent)]
  x16<-household_dat$S_electricity[!is.na(dependent)]
```

```
  x17<-household_dat$S_hasanimals[!is.na(dependent)]
  x18<-household_dat$S_animalsinside[!is.na(dependent)]
  x19<-household_dat$S_garbage[!is.na(dependent)]
  x20<-household_dat$S_washhands[!is.na(dependent)]
  x21<- household_dat$dpisofirme[!is.na(dependent)]
  x22<-household_dat$S_cashtransfers[!is.na(dependent)]
  x23<-household_dat$S_milkprogram[!is.na(dependent)]
  x24<-household_dat$S_foodprogram[!is.na(dependent)]
  x25<-household_dat$S_seguropopular[!is.na(dependent)]
  updated_dependent<- dependent[!is.na(dependent)]
  return(summary(lm(updated_dependent ~ x21 + x1 + x2 + x3  + x4+ x5+ x6+ x7+ x8+ x9+ x10+ x11+ x12+ x13
}
```

```
# r square coefficients for table 4
T4_rsq <- data.frame(Dependent = c("share_cement_floors", "kitchen", "dining_room", "bedroom", "bathroom
T4_rsq$data <- list(c(household_dat$S_shcementfloor), c(household_dat$S_cementfloorkit), c(household_dat
T4_rsq <- T4_rsq %>% mutate(r_sq_m1 = unlist(map(data, model_1_hh_rsq)), r_sq_m2 = unlist(map(data, model
# r squared coefficients for table 5
T5_rsq <- data.frame(Dependent = c("parasite", "diarrhea", "anemia", "MacArthur", "Peabody", "height", "
T5_rsq$data <- list(c(individual_dat$S_parcount), c(individual_dat$S_diarrhea), c(individual_dat$S_anemi
T5_rsq <- T5_rsq %>% mutate(r_sq_m1 = unlist(map(data, model_1_i_rsq)), r_sq_m2 = unlist(map(data, model
# r squared coefficients for table 6
T6_rsq <- data.frame(Dependent = c("sat_floor", "sat_house", "sat_life", "depression", "stress"))
T6_rsq$data <- list(c(household_dat$S_satisfloor), c(household_dat$S_satishouse), c(household_dat$S_sati
T6_rsq <- T6_rsq %>% mutate(r_sq_m1 = unlist(map(data, model_1_hh_rsq)), r_sq_m2 = unlist(map(data, model
```

R Squared Plots

Here, we try to see if there is any trend between the coefficients for the dependents for each model and the r squared values.

```
T4_2 <- Table_4 %>% full_join(T4_rsq, by = "Dependent")
```

```
## Warning: Column `Dependent` joining factors with different levels, coercing
## to character vector
```

```
T5_2 <- Table_5 %>% full_join(T5_rsq, by = "Dependent")
```

```
## Warning: Column `Dependent` joining factors with different levels, coercing
## to character vector
```

```
T6_2 <- Table_6 %>% full_join(T6_rsq, by = "Dependent")
```

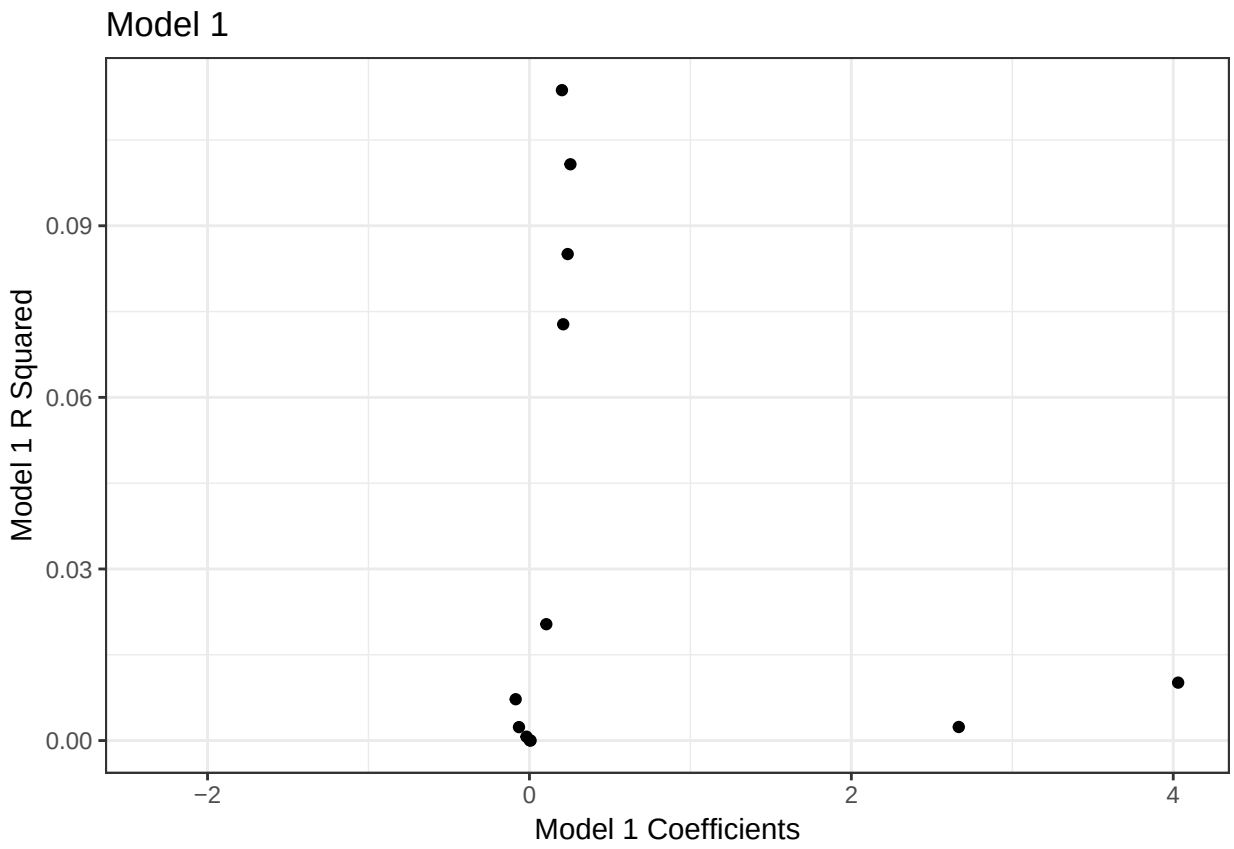
```
## Warning: Column `Dependent` joining factors with different levels, coercing
## to character vector
```

```
T_Tot <- T4_2 %>% full_join(T5_2) %>% full_join(T6_2)
```

```
## Joining, by = c("Dependent", "control_group_mean", "control_group_sd", "coeff_1", "sce_1", "coef_mean
## Joining, by = c("Dependent", "control_group_mean", "control_group_sd", "coeff_1", "sce_1", "coef_mean
```

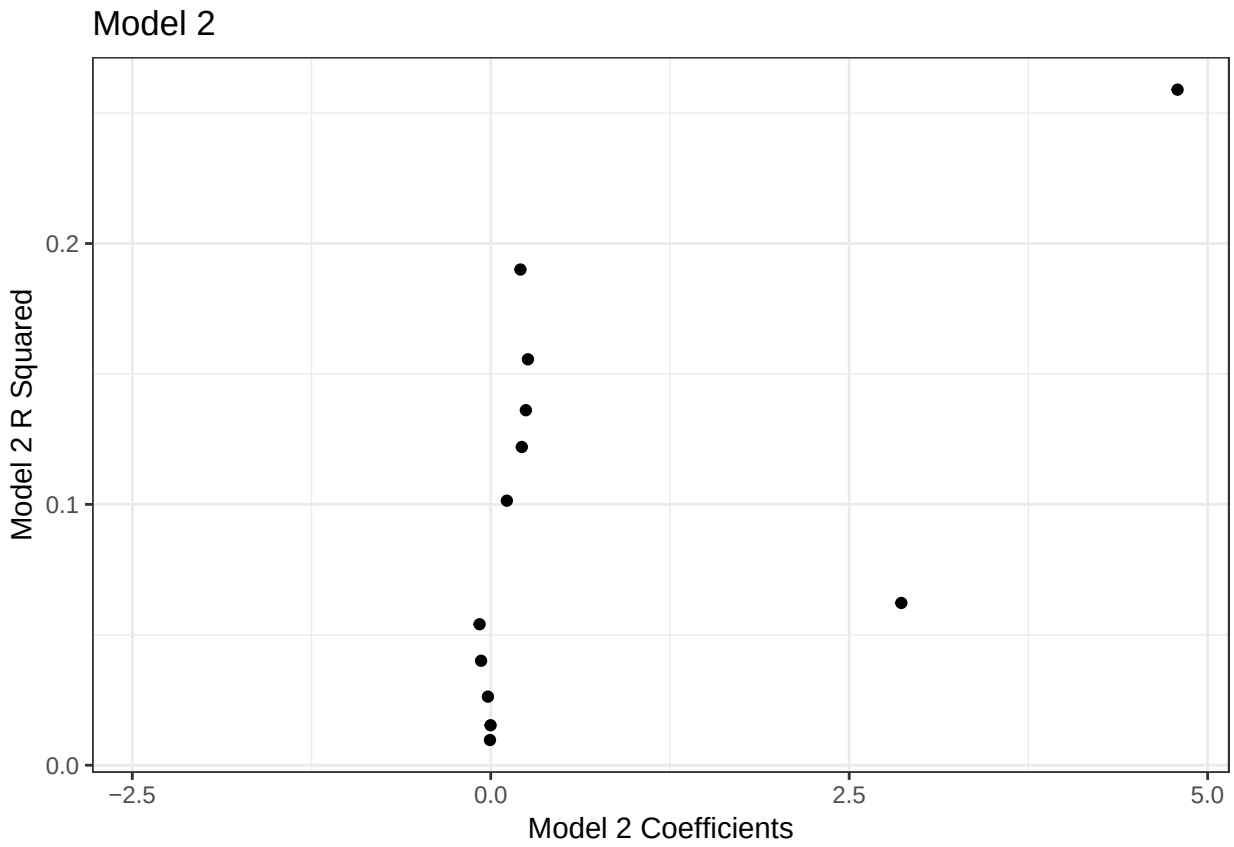
```
ggplot(data = T_Tot) + geom_point(aes(x = coeff_1, y = r_sq_m1)) + xlab("Model 1 Coefficients") + ylab(
```

```
## Warning: Removed 9 rows containing missing values (geom_point).
```



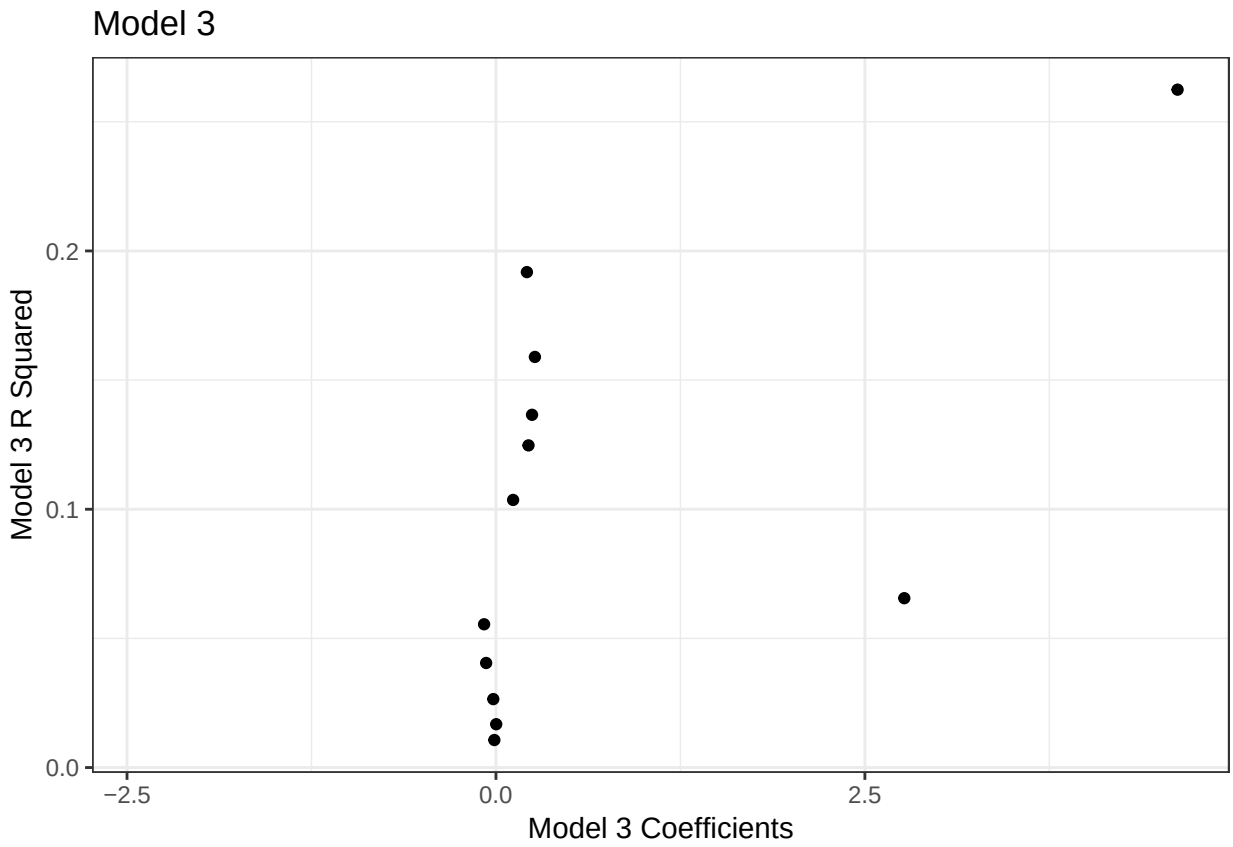

```
ggplot(data = T_Tot) + geom_point(mapping = aes(x = coeff_2, y = r_sq_m2)) + xlab("Model 2 Coefficients
```

```
## Warning: Removed 9 rows containing missing values (geom_point).
```



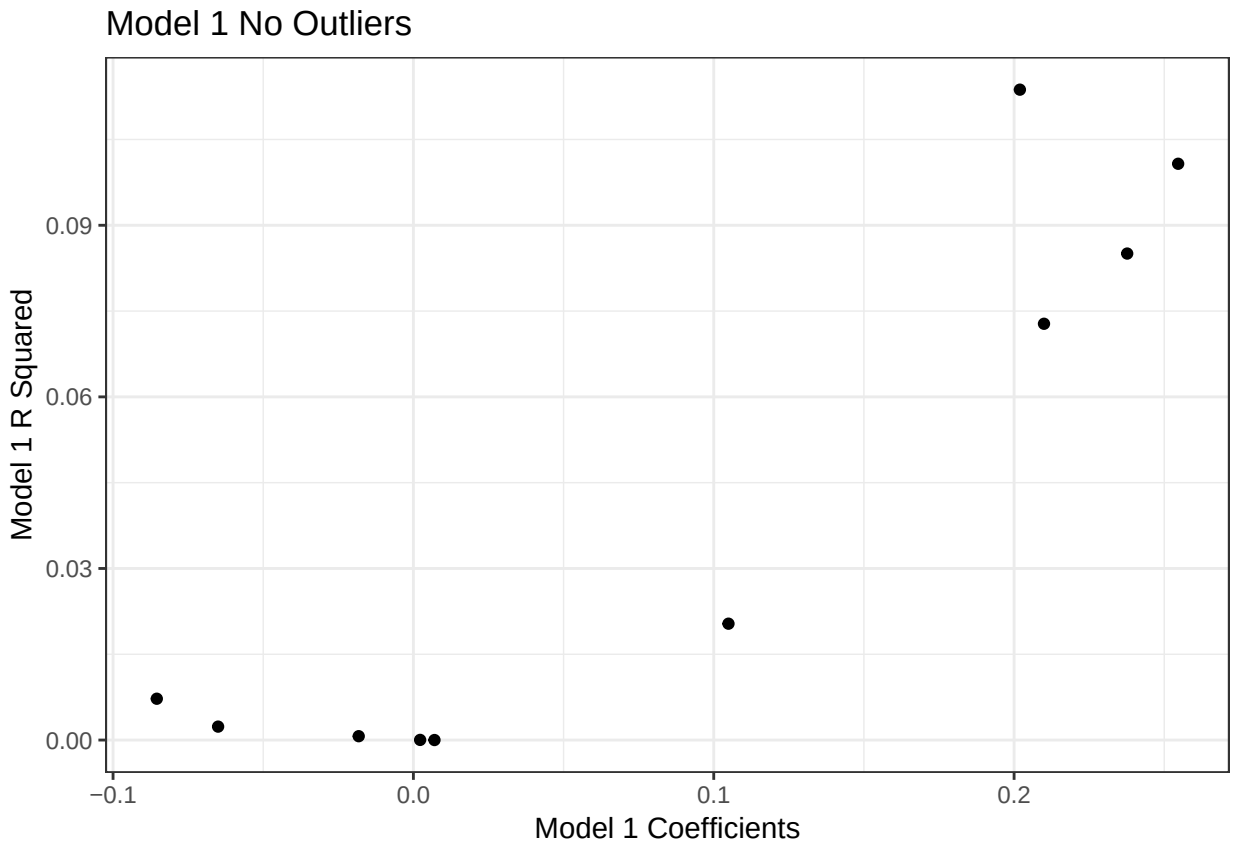

```
ggplot(data = T_Tot) + geom_point(mapping = aes(x = coeff_3, y = r_sq_m3)) + xlab("Model 3 Coefficients
```

```
## Warning: Removed 9 rows containing missing values (geom_point).
```



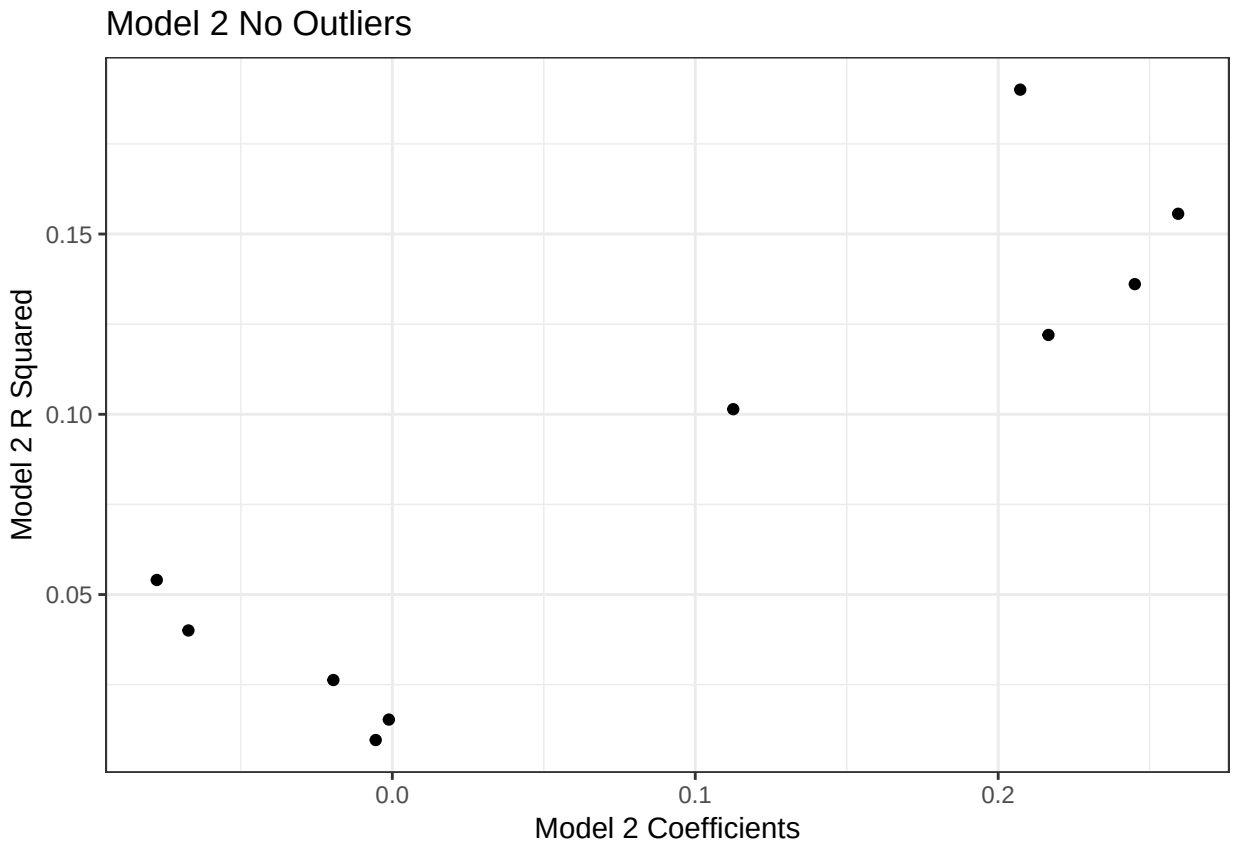

```
# here we remove outliers
T_Tot2 <- T_Tot %>% filter(coeff_1 > -1) %>% filter(coeff_1 < 1) %>% filter(coeff_2 > -1) %>% filter(co

ggplot(data = T_Tot2) + geom_point(mapping = aes(x = coeff_1, y = r_sq_m1)) + xlab("Model 1 Coefficient
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



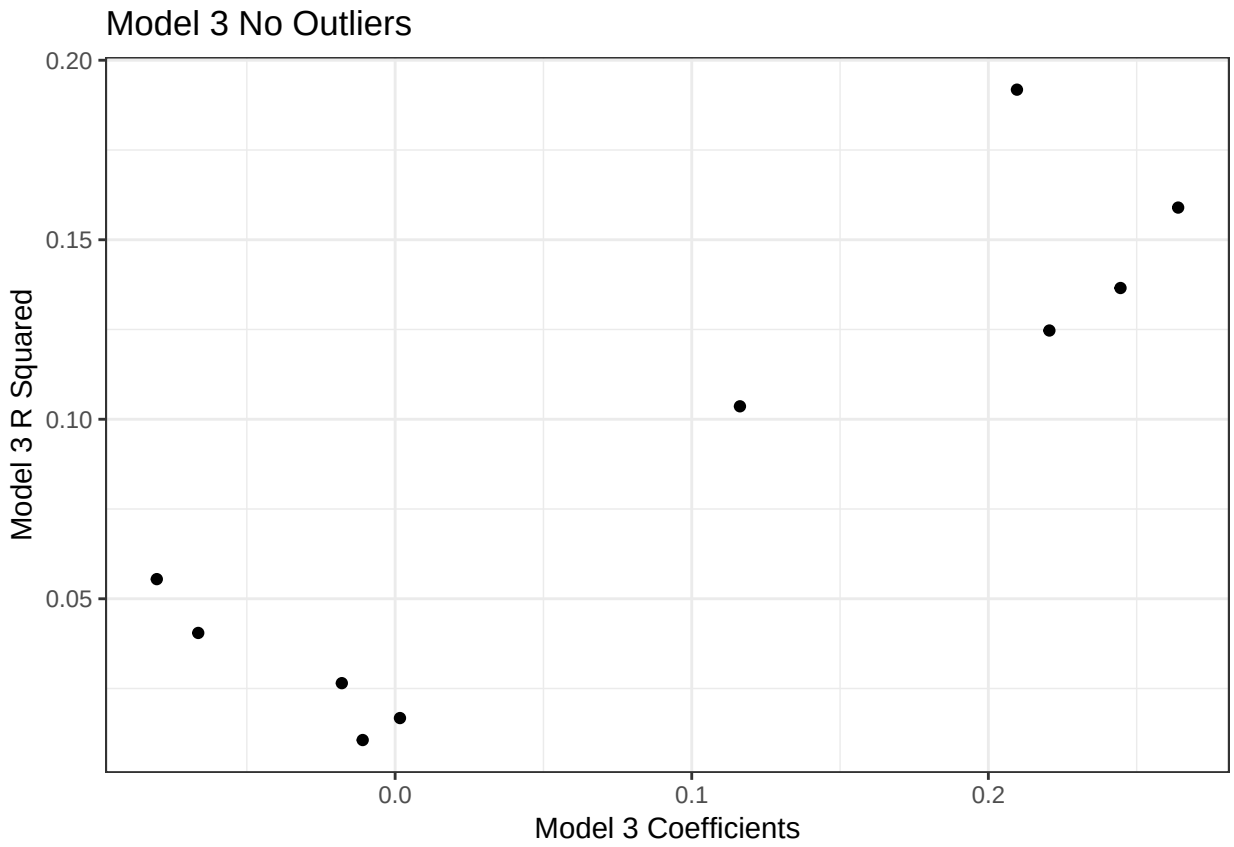

```
ggplot(data = T_Tot2) + geom_point(mapping = aes(x = coeff_2, y = r_sq_m2)) + xlab("Model 2 Coefficients
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```




```
ggplot(data = T_Tot2) + geom_point(mapping = aes(x = coeff_3, y = r_sq_m3)) + xlab("Model 3 Coefficients
```

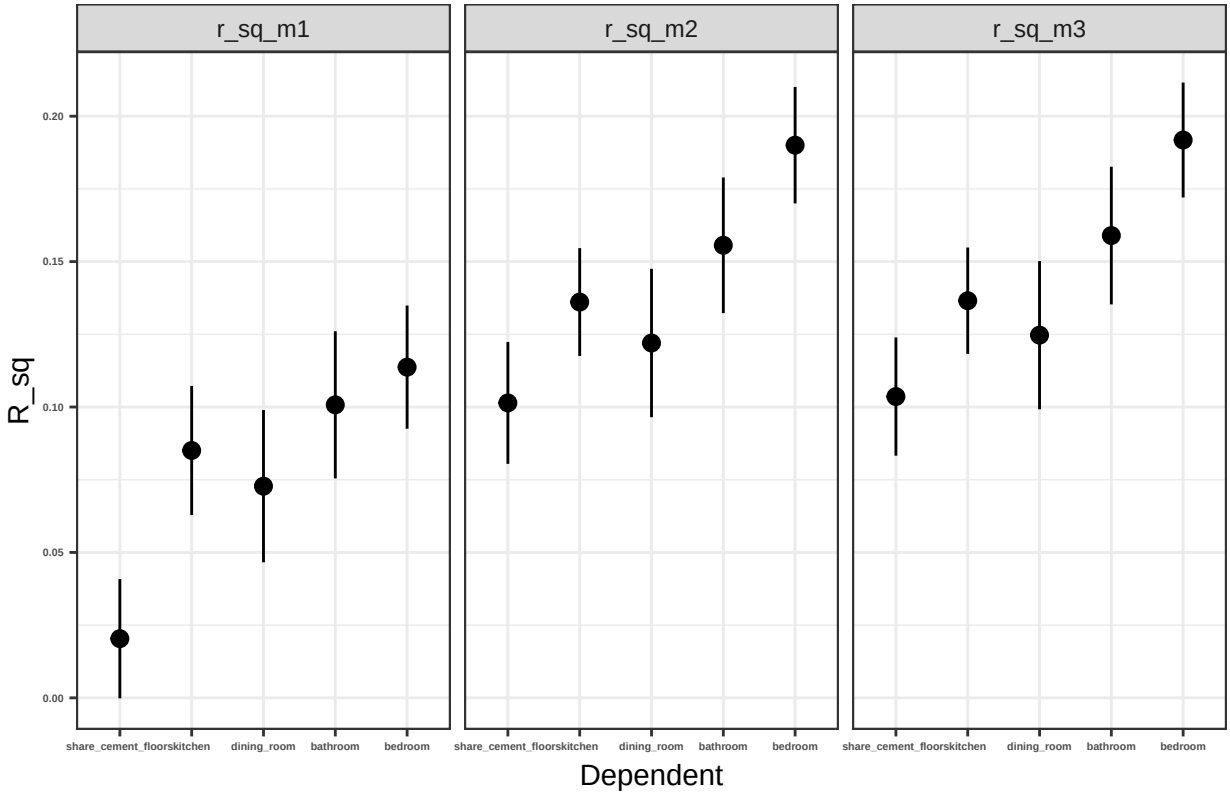
```
## Warning: Removed 2 rows containing missing values (geom_point).
```



Rsquared values for each table, separated by model using standard error as error bars (not clustered)

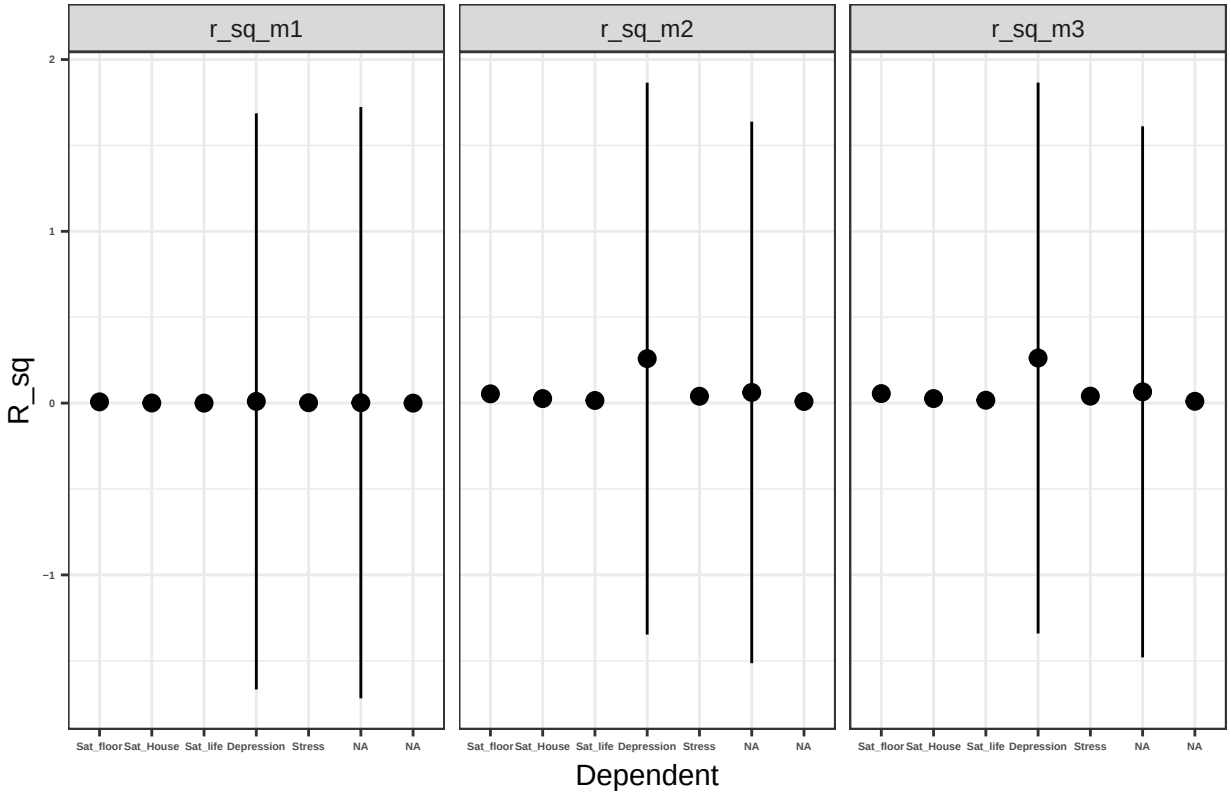
```
T6_rsq$r_sq_m1<-as.numeric(T6_rsq$r_sq_m1)
plot_r_6<-T6_rsq%>%gather("model_type","coeff",2:4)%>%mutate(sce=graphing[c(13:17,30:34,47:51),]$sce)%>%
T5_rsq$r_sq_m1<-as.numeric(T5_rsq$r_sq_m1)
plot_r_5<-T5_rsq%>%gather("model_type","coeff",2:4)%>%mutate(sce=graphing[c(6:12,23:29,40:46),]$sce)%>%m



T4_rsq$r_sq_m1<-as.numeric(T4_rsq$r_sq_m1)
plot_r_4<-T4_rsq%>%gather("model_type","coeff",2:4)%>%mutate(sce=graphing[c(1:5,18:22,35:39),]$sce)%>%mu

ggplot(plot_r_4,aes(x=Dependent,y=coeff))+geom_pointrange(aes(ymin=coeff-sce,ymax=coeff+sce))+scale_x_di
```

Table 4 R_sq by Model



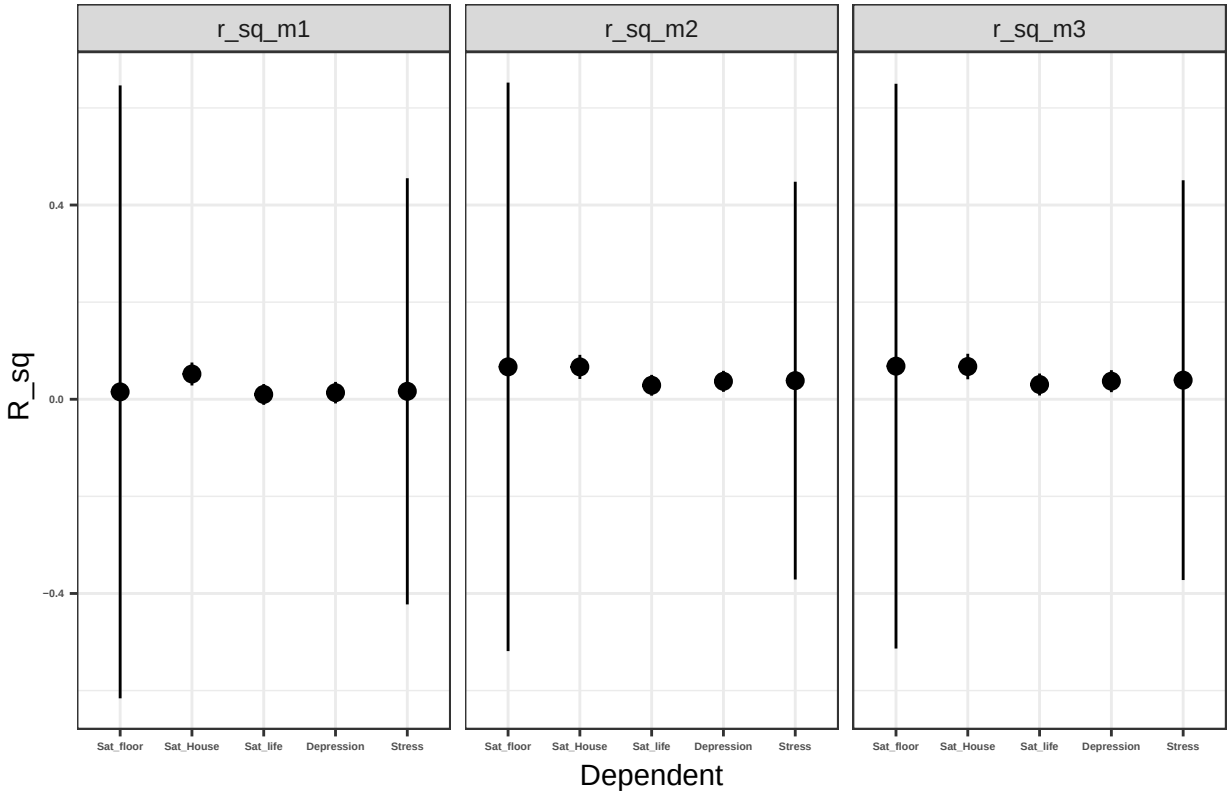

```
ggplot(plot_r_5,aes(x=Dependent,y=coeff))+geom_pointrange(aes(ymin=coeff-sce,ymax=coeff+sce))+scale_x_di
```

Table 5 R_sq by Model




```
ggplot(plot_r_6,aes(x=Dependent,y=coeff))+geom_pointrange(aes(ymin=coeff-sce,ymax=coeff+sce))+scale_x_di
```

Table 6 R_sq by Model



Issues

Our biggest struggle while recreating this paper was their use of clustering. The provided code was in Stata, which is not extremely transparent in its analysis. We were able to see that the data analysis was clustered by census block, however when we attempted to calculate it, we were puzzled by how to do it (especially with regards to regression). We learned that the regression was done through a Moran's I test which measures spatial correlation. Since Stata does many of these things through built in packages, it was difficult for us to reproduce it in R. Instead we calculated the overall regression coefficients, which for the most part gave us very similar values, save for Table 5. Our table 5 values were quite off from that of the paper, which we assume is due to our differing regression technique. We also had trouble calculating clustered standard error. At first, our coefficients for Table 5 differed noticeably from those presented in the paper, while our coefficients for Tables 4 and 6 closely matched those in the paper. This table relied on a different data set, the individuals data set, which the other two relied on the households data set, so we assumed that the error related to how we read in or analyzed this data. When we first tried the regression, we replaced all na values with zero. However, after closer examination of the work of the researchers, we noticed that na values should be replaced with zero only for control variables, rather than everywhere. If there is an na value for some dependent variable for a given individual, we disregarded that individual when doing the regression. After making this amendment, our coefficients for Tables 4, 5, and 6 closely matched those presented in the paper.

Works Cited: Original Paper

Cattaneo, Matias D., Sebastian Galiani, Paul J. Gertler, Sebastian Martinez, and Rocio Titiunik. 2009. "Housing, Health, and Happiness." American Economic Journal: Economic Policy, 1 (1): 75-105.

Works Cited: Cited Paper

A. Colin Cameron & Douglas L. Miller, 2015. "A Practitioner's Guide to Cluster-Robust Inference," Journal of Human Resources, University of Wisconsin Press, vol. 50(2), pages 317-372.

The following is a list of all packages used to generate these results. (Leave at very end of file.)

```
sessionInfo()
```

```
## R version 3.5.2 (2018-12-20)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS High Sierra 10.13.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] bindrcpp_0.2.2  estimatr_0.16   broom_0.5.1     haven_2.0.0
##  [5] forcats_0.3.0   stringr_1.3.1   dplyr_0.7.8     purrr_0.2.5
##  [9] readr_1.3.1     tidyr_0.8.2     tibble_2.0.1    ggplot2_3.1.0
## [13] tidyverse_1.2.1 modelr_0.1.2    scales_1.0.0    here_0.1
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_0.2.5 xfun_0.4         reshape2_1.4.3   lattice_0.20-38
##  [5] colorspace_1.4-0 generics_0.0.2   htmltools_0.3.6  yaml_2.2.0
##  [9] rlang_0.3.1      pillar_1.3.1     glue_1.3.0       withr_2.1.2
## [13] readxl_1.2.0     bindr_0.1.1      plyr_1.8.4       munsell_0.5.0
## [17] gtable_0.2.0     cellranger_1.1.0 rvest_0.3.2      evaluate_0.12
## [21] labeling_0.3     knitr_1.21       Rcpp_1.0.0       backports_1.1.3
## [25] jsonlite_1.6     hms_0.4.2        digest_0.6.18    stringi_1.2.4
## [29] grid_3.5.2       rprojroot_1.3-2  cli_1.0.1        tools_3.5.2
## [33] magrittr_1.5     lazyeval_0.2.1   Formula_1.2-3    crayon_1.3.4
## [37] pkgconfig_2.0.2  xml2_1.2.0       lubridate_1.7.4  assertthat_0.2.0
## [41] rmarkdown_1.11   httr_1.4.0       rstudioapi_0.9.0 R6_2.3.0
## [45] nlme_3.1-137     compiler_3.5.2
```