PE-PSet1

Drazzel Feliu - 12174100

For this assignment, provide a write-up where you answer the questions below, selectively cutting and pasting output where needed. Be concise in your write-up; excess wordiness will be penalized. Also submit a log file that includes commands and results for your entire analysis.

Study motivation

Developing and middle-income countries increasingly provide health coverage to their residents. Mexico has become a leading example. Prior to 2004, roughly half the Mexican population had no health insurance. Health coverage was available through the social security system for salaried employees and their families, but the roughly 50 million Mexicans outside the salaried sector had no insurance.

In 2004, Mexico introduced Seguro Popular. Seguro Popular is a voluntary, non- contributory health insurance program for residents without coverage from the social security system. One of its explicit goals is to provide covered households with financial protection by reducing their incidence of catastrophic out-of-pocket health spending.

The Seguro Popular field experiment

The experiment involved a paired-cluster randomization design, by which 100 largely rural "health clusters" were formed into 50 pairs on the basis of similarities in demographic characteristics and healthcare infrastructure. Health clusters are defined as a healthcare facility and the population living within its catchment area. Within matched pairs of health clusters, one was randomly assigned to treatment, with the other serving as control. Treatment involved publicity campaigns designed to encourage local residents to enroll. Extensive information about the initial design of the experiment has been published by King et al (2009).

Pre-intervention baseline data were collected from participant households in 2005. Post-intervention follow-up data were collected in 2006, roughly 10 months after the intervention began. The 2005 survey included 32,515 households.

Questions

(1)

Did Seguro Popular reduce extreme expenditures on health?

(2)

Does it matter how we define "extreme expenditures?"

Data

Data for this assignment are in the file SP_dataset.dta on Canvas.

Analysis

(1)

Inspect the data. Are there truly 100 clusters in 50 pairs? Does each pair have a treatment and a control cluster? What is the modal level of schooling in the sample? The modal marital status? What share of sample households were participating in Oportunidades, Mexico's conditional cash transfer program? What is mean non- health expenditure (data are in Mexican pesos, where 12 pesos = approx. 1 dollar)? Does this make sense?

```
# data cleaning
# check data class & view summary
sapply(data, class)
summary(data)

# assuming that the data is missing completely at random, all NA values across all observations must be
data <- na.omit(data)

# view data summary
summary(data)
```

Answer

There are exactly 100 clusters and 50 cluster pairs. Each pair does indeed have a treatment and a control cluster. The modal level of education is elementary education (11484). The modal marital status is married (13398). The share of sample households participating in Opurtunidades is (11192). The mean non-health expenditure in 2005 is 9829.203 in pesos and 819.1002 in US dollars. The mean non-health expenditure in 2006 is 9737.579 in pesos and 811.4649 in US dollars.

```
# number of clusters
length(unique(data$cluster))
```

```
## [1] 100
```

```
# number of cluster pairs
length(unique(data$clust_pair))
```

```
## [1] 50
```

```
# number of treated and untreated in each cluster pair
cast(data, clust_pair ~ treatment)
```

```
## Using marstat6 as value column.  Use the value argument to cast to override this choice
```

```
## Aggregation requires fun.aggregate: length used as default
```

```
##    clust_pair   0   1
## 1           1 297 302
## 2           2 232 180
## 3           3 188 242
## 4           4 250 291
## 5           5 301 208
## 6           6 260 287
## 7           7 230 318
## 8           8 288 214
## 9           9 287 308
```

```
## 10         10 278 219
## 11         11 250 339
## 12         12 235 247
## 13         13 339 143
## 14         14 236 255
## 15         15 232 268
## 16         16 266  98
## 17         17 292 269
## 18         18 179 286
## 19         19 218 277
## 20         20 259 282
## 21         21 214 142
## 22         22 280 261
## 23         23 198 222
## 24         24 252 281
## 25         25 330 196
## 26         26 160 189
## 27         27 317 206
## 28         28 148 262
## 29         29 308 297
## 30         30 236 274
## 31         31 296 338
## 32         32 177 214
## 33         33 197 218
## 34         34 144 249
## 35         35 250 293
## 36         36 253 251
## 37         37 214 200
## 38         38 160 259
## 39         39 106 196
## 40         40 296 282
## 41         41 320 276
## 42         42 278 250
## 43         43 146 170
## 44         44  64 118
## 45         45 204 214
## 46         46 259 236
## 47         47 265 308
## 48         48 159 227
## 49         49  35 218
## 50         50 240 125
```

```r
# modal level of schooling
colSums(data[17:22])
```

```
## edu_info_051 edu_info_052 edu_info_053 edu_info_054 edu_info_055
##         4360        11484         4886         1305         1487
## edu_info_056
##          106
```

```r
# modal marital status
colSums(data[27:32])
```

```
## marstat1 marstat2 marstat3 marstat4 marstat5 marstat6
##     3139    13398     1289      252     2177     3373
```

```r
# of households participating in Oportunidades
sum(data$opor_05)
```

```
## [1] 11192
```

```r
# mean non-health expenditure in 2005 in pesos
mean(data$allbut_05)
```

```
## [1] 9829.203
```

```r
# mean non-health expenditure in 2005 in dollars
(mean(data$allbut_05))/12
```

```
## [1] 819.1002
```

```r
# mean non-health expenditure in 2006 in pesos
mean(data$allbut_06)
```

```
## [1] 9737.579
```

```r
# mean non-health expenditure in 2006 in dollars
(mean(data$allbut_06))/12
```

```
## [1] 811.4649
```

(2)

Which treatment parameters could be identified by these data?

Answer

The treatment parameters that can be identified by the data are the varying types of expenditures that differ between 2005 and 2006 (food expenditure, all non-medical expenditure, medical expenditure). Using an instrumental variable, assuming that monotonicity holds, we can identify the impact of treatment on the compliers thereby producing the local ATE. In this case, identifying those who would respond to treatment is a meaningful group worthy of being analyzed as this program is meant to identify those at the margins who would benefit from being insured against catastrophic out-of-pocket health spending.

(3)

Assess the adequacy of random assignment. You will need to think about how to use Stata's reg command to test for equal means among pre-intervention variables within cluster pairs. How important is it to account for dependence within clusters? What is the bottom line? Was random assignment properly executed?

Answer

```r
balance <- lm(data = data,treatment ~ sex + age + marstat + hhsize + food_yr_05 + allbut_05 + oop_yr3_0
summary(balance)
```

```
##
## Call:
## lm(formula = treatment ~ sex + age + marstat + hhsize + food_yr_05 +
##     allbut_05 + oop_yr3_05 + nkid_05 + nadult_05 + headwomen_05 +
```

```
##     edu_info_051 + edu_info_052 + edu_info_053 + edu_info_054 +
##     edu_info_055 + edu_info_056 + insp2005 + opor_05, data = data,
##     subset = cluster(data$clust_pair))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80696 -0.26041  0.03611  0.29816  0.72808
##
## Coefficients: (2 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.633e+00  1.876e-02  87.048  < 2e-16 ***
## sex          -3.569e-01  6.979e-03 -51.138  < 2e-16 ***
## age          -1.035e-03  2.991e-04  -3.461 0.000538 ***
## marstat      -4.673e-02  1.925e-03 -24.278  < 2e-16 ***
## hhsize       -7.197e-03  2.232e-03  -3.224 0.001267 **
## food_yr_05    1.596e-05  1.164e-06  13.717  < 2e-16 ***
## allbut_05    -2.289e-05  8.495e-07 -26.940  < 2e-16 ***
## oop_yr3_05    5.749e-06  4.176e-06   1.377 0.168604
## nkid_05      -1.400e-01  4.742e-03 -29.513  < 2e-16 ***
## nadult_05    -1.922e-01  5.102e-03 -37.660  < 2e-16 ***
## headwomen_05  3.735e-01  9.356e-03  39.924  < 2e-16 ***
## edu_info_051  1.944e-01  1.790e-02  10.859  < 2e-16 ***
## edu_info_052 -2.049e-01  1.067e-02 -19.207  < 2e-16 ***
## edu_info_053  1.824e-01  1.159e-02  15.744  < 2e-16 ***
## edu_info_054 -1.146e-01  1.132e-02 -10.123  < 2e-16 ***
## edu_info_055         NA         NA      NA       NA
## edu_info_056         NA         NA      NA       NA
## insp2005     -5.000e-01  1.734e-02 -28.837  < 2e-16 ***
## opor_05      -2.904e-02  1.009e-02  -2.878 0.004007 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3864 on 23611 degrees of freedom
## Multiple R-squared:  0.3953, Adjusted R-squared:  0.3949
## F-statistic: 964.6 on 16 and 23611 DF,  p-value: < 2.2e-16
```

Random assignment seems to have been properly executed across the board with the exception of one variable (oop_yr3_05). It's important to account for dependence within clusters to ensure that the treatment and control groups are reflective of each other within the clusters so that we can draw conclusions broadly about whether or not treatment was effective.

(4)

Construct two variables relating health spending to the household's total income. The first is the health budget share, equal to health spending divided by total spending. The second is the adjusted health budget share, equal to health spending divided by disposable income, where disposable income is defined as total spending less food expenditures.

Answer

```r
# create health budget share in 2005
data$hbshare05 <- data$oop_yr3_05/(data$allbut_05+data$oop_yr3_05)
```

```r
# create health budget share in 2006
data$hbshare06 <- data$oop_yr3_06/(data$allbut_06+data$oop_yr3_06)

# create adjusted health budget share in 2006
data$adjhbshare05 <- data$oop_yr3_05/(data$allbut_05+data$oop_yr3_05-data$food_yr_05)
# create adjusted health budget share in 2006
data$adjhbshare06 <- data$oop_yr3_06/(data$allbut_06+data$oop_yr3_06-data$food_yr_06)
```

(5)

On average, what share of total expenditures does health spending account for? What are the 75th and 90th percentiles of the budget share distribution? What are the corresponding figures as a share of disposable income rather than total expenditures?

Answer

Mean health budget shares (2005 & 2006) and mean adjusted health budget shares (2005 & 2006).

```r
# average health budget share in 2005
mean(data$hbshare05, na.rm = TRUE)
```

```
## [1] 0.02882707
```

```r
# average health budget share in 2006
mean(data$hbshare06, na.rm = TRUE)
```

```
## [1] 0.03294801
```

```r
# average adjusted health budget share in 2005
mean(data$adjhbshare05, na.rm = TRUE)
```

```
## [1] 0.04361944
```

```r
# average adjusted health budget share in 2006
mean(data$adjhbshare06, na.rm = TRUE)
```

```
## [1] 0.05067997
```

75th percentile for health budget share (2005 & 2006) and adjusted health budget shares (2005 & 2006)

```r
# 75th percentile for health budget share in 2005
quantile(data$hbshare05, .75, na.rm = TRUE)
```

```
## 75%
##   0
```

```r
# 75th percentile for health budget share in 2006
quantile(data$hbshare06, .75, na.rm = TRUE)
```

```
##         75%
## 0.009090049
```

```r
# 90th percentile for health budget share in 2005
quantile(data$hbshare05, .90, na.rm = TRUE)
```

```
##        90%
## 0.09379221
```

```r
# 90th percentile for health budget share in 2005
quantile(data$hbshare06, .90, na.rm = TRUE)
```

```
##       90%
## 0.1093665
```

90th percentile for health budget share (2005 & 2006) and adjusted health budget shares (2005 & 2006)

```r
# 75th percentile for adjusted health budget share in 2005
quantile(data$adjhbshare05, .75, na.rm = TRUE)
```

```
## 75%
##   0
```

```r
# 75th percentile for adjusted health budget share in 2006
quantile(data$adjhbshare06, .75, na.rm = TRUE)
```

```
##        75%
## 0.01693243
```

```r
# 90th percentile for adjusted health budget share in 2005
quantile(data$adjhbshare05, .90, na.rm = TRUE)
```

```
##       90%
## 0.1538067
```

```r
# 90th percentile for adjusted health budget share in 2006
quantile(data$adjhbshare06, .90, na.rm = TRUE)
```

```
##       90%
## 0.1861052
```

(6)

Construct two measures of extreme health spending: (i) a dummy equal to one if the household's health budget share exceeds 20 percent; and (ii) a dummy equal to one if the household spends more than 30 percent of its disposable income on health. How are these measures related?

Answer

```r
data$extreme20dummy05 <- ifelse(data$hbshare05 > .20,1,0)
data$extreme20dummy06 <- ifelse(data$hbshare06 > .20,1,0)

data$extreme30dummy05 <- ifelse(data$adjhbshare05 > .30,1,0)
data$extreme30dummy06 <- ifelse(data$adjhbshare06 > .30,1,0)
```

These measures are both reflections of how much of the household budget is consumed by medical spending. Given that the second dummy variable has both a higher threshhold and is made up of discretionary spending, it is likely more sensitive to the impact of health spending than that first dummy variable, as this one includes the fixed costs of food as part of its calculation.

(7)

Estimate the ITT effects of the intervention on extreme health spending. First, just regress the extreme health spending measures on the treatment dummy. Next, add cluster-pair dummies. In both cases, calculate standard errors that account for dependence among households within clusters. What happens to the coefficients when you add the cluster-pair dummies? To their standard errors? Why?

Answer

```r
# regress extreme health budget share spending in 2006 on treatment
itt <- lm(extreme20dummy06 ~ treatment, data = data)
sumitt <- summary(itt)
sumitt
```

```
##
## Call:
## lm(formula = extreme20dummy06 ~ treatment, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05932 -0.05932 -0.04438 -0.04438  0.95562
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.059325   0.002054  28.881  < 2e-16 ***
## treatment   -0.014947   0.002882  -5.186 2.17e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2214 on 23600 degrees of freedom
##   (26 observations deleted due to missingness)
## Multiple R-squared:  0.001138,   Adjusted R-squared:  0.001096
## F-statistic: 26.89 on 1 and 23600 DF,  p-value: 2.167e-07
```

```r
# regress extreme health budget share spending on treatment with cluster dummies
itt2 <- lm(extreme20dummy06 ~ treatment + factor(clust_pair), data = data)
sumitt2 <- summary(itt2)
sumitt2
```

```
##
## Call:
## lm(formula = extreme20dummy06 ~ treatment + factor(clust_pair),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14942 -0.06525 -0.04346 -0.02562  0.99417
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9.594e-02  9.093e-03  10.550  < 2e-16 ***
## treatment            -1.479e-02  2.902e-03  -5.095 3.52e-07 ***
## factor(clust_pair)2  -1.423e-02  1.406e-02  -1.012 0.311426
## factor(clust_pair)3   5.349e-02  1.393e-02   3.839 0.000124 ***
## factor(clust_pair)4  -1.106e-03  1.303e-02  -0.085 0.932343
## factor(clust_pair)5   3.781e-02  1.324e-02   2.855 0.004313 **
## factor(clust_pair)6  -4.973e-02  1.300e-02  -3.826 0.000130 ***
## factor(clust_pair)7  -4.529e-02  1.299e-02  -3.486 0.000491 ***
```

```
## factor(clust_pair)8   9.194e-06  1.329e-02   0.001 0.999448
## factor(clust_pair)9  -1.590e-02  1.272e-02  -1.250 0.211197
## factor(clust_pair)10 -2.503e-02  1.333e-02  -1.878 0.060376 .
## factor(clust_pair)11 -2.291e-02  1.275e-02  -1.797 0.072332 .
## factor(clust_pair)12 -2.185e-02  1.345e-02  -1.624 0.104308
## factor(clust_pair)13 -3.968e-02  1.345e-02  -2.949 0.003187 **
## factor(clust_pair)14 -6.170e-02  1.339e-02  -4.609 4.07e-06 ***
## factor(clust_pair)15 -3.769e-02  1.333e-02  -2.828 0.004690 **
## factor(clust_pair)16 -6.448e-02  1.461e-02  -4.412 1.03e-05 ***
## factor(clust_pair)17  1.831e-02  1.291e-02   1.418 0.156180
## factor(clust_pair)18 -3.082e-02  1.359e-02  -2.268 0.023332 *
## factor(clust_pair)19 -5.534e-02  1.334e-02  -4.147 3.38e-05 ***
## factor(clust_pair)20 -4.007e-02  1.303e-02  -3.074 0.002116 **
## factor(clust_pair)21 -3.908e-02  1.474e-02  -2.651 0.008033 **
## factor(clust_pair)22 -5.553e-02  1.303e-02  -4.262 2.03e-05 ***
## factor(clust_pair)23 -2.366e-02  1.399e-02  -1.691 0.090770 .
## factor(clust_pair)24 -2.435e-02  1.308e-02  -1.862 0.062658 .
## factor(clust_pair)25 -3.719e-02  1.313e-02  -2.833 0.004622 **
## factor(clust_pair)26  3.762e-03  1.479e-02   0.254 0.799233
## factor(clust_pair)27 -6.334e-02  1.315e-02  -4.817 1.47e-06 ***
## factor(clust_pair)28  8.635e-03  1.408e-02   0.613 0.539841
## factor(clust_pair)29 -6.554e-02  1.266e-02  -5.176 2.28e-07 ***
## factor(clust_pair)30 -6.250e-02  1.324e-02  -4.722 2.34e-06 ***
## factor(clust_pair)31 -2.812e-02  1.252e-02  -2.246 0.024692 *
## factor(clust_pair)32 -5.204e-02  1.428e-02  -3.644 0.000269 ***
## factor(clust_pair)33 -8.335e-02  1.403e-02  -5.941 2.87e-09 ***
## factor(clust_pair)34 -7.125e-02  1.429e-02  -4.987 6.16e-07 ***
## factor(clust_pair)35 -3.455e-02  1.302e-02  -2.654 0.007950 **
## factor(clust_pair)36 -5.683e-02  1.328e-02  -4.280 1.88e-05 ***
## factor(clust_pair)37 -7.183e-02  1.405e-02  -5.112 3.20e-07 ***
## factor(clust_pair)38 -7.009e-02  1.399e-02  -5.009 5.52e-07 ***
## factor(clust_pair)39 -5.313e-02  1.552e-02  -3.423 0.000621 ***
## factor(clust_pair)40 -2.285e-02  1.281e-02  -1.783 0.074528 .
## factor(clust_pair)41 -3.875e-02  1.271e-02  -3.049 0.002297 **
## factor(clust_pair)42 -6.242e-02  1.311e-02  -4.760 1.94e-06 ***
## factor(clust_pair)43 -7.532e-02  1.527e-02  -4.932 8.20e-07 ***
## factor(clust_pair)44 -5.888e-02  1.860e-02  -3.166 0.001547 **
## factor(clust_pair)45 -5.727e-02  1.400e-02  -4.091 4.32e-05 ***
## factor(clust_pair)46 -4.444e-02  1.334e-02  -3.331 0.000867 ***
## factor(clust_pair)47 -7.228e-02  1.284e-02  -5.631 1.81e-08 ***
## factor(clust_pair)48 -4.838e-02  1.434e-02  -3.374 0.000742 ***
## factor(clust_pair)49 -4.762e-02  1.650e-02  -2.886 0.003908 **
## factor(clust_pair)50 -7.717e-02  1.459e-02  -5.288 1.25e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2197 on 23551 degrees of freedom
##   (26 observations deleted due to missingness)
## Multiple R-squared:  0.01858,    Adjusted R-squared:  0.0165
## F-statistic: 8.919 on 50 and 23551 DF,  p-value: < 2.2e-16
```

The coefficients in both sets of regressions lose precision. Regressing extreme health budget share on only the treatment dummy provides a coefficient of -0.01495 with a standard error of 0.00282. When including the cluster pair dummy, the treatment coefficient becomes -0.01479. with a standard error of 0.00290. This follows logically, as any instrumental variable, no matter how effective it is, will diminish the precision of OLS estimators. Our instrument variable's F-statistic is less than 10, indicating the instrument is weak.

```r
# regress adjusted extreme health budget share spending in 2006 on treatment
itt3 <- lm(extreme30dummy06 ~ treatment, data = data)
sumitt3 <- summary(itt3)
sumitt3
```

```
##
## Call:
## lm(formula = extreme30dummy06 ~ treatment, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.06402 -0.06402 -0.04991 -0.04991  0.95009
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.064019   0.002149  29.796  < 2e-16 ***
## treatment   -0.014106   0.003015  -4.679 2.89e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2315 on 23585 degrees of freedom
##   (41 observations deleted due to missingness)
## Multiple R-squared:  0.0009275,  Adjusted R-squared:  0.0008851
## F-statistic:  21.9 on 1 and 23585 DF,  p-value: 2.895e-06
```

```r
# regress adjusted extreme health budget share spending on treatment with cluster dummies
itt4 <- lm(extreme30dummy06 ~ treatment + factor(clust_pair), data = data)
sumitt4 <- summary(itt4)
sumitt4
```

```
##
## Call:
## lm(formula = extreme30dummy06 ~ treatment + factor(clust_pair),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15371 -0.06949 -0.04714 -0.03178  0.99059
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.095557   0.009516  10.042  < 2e-16 ***
## treatment            -0.014034   0.003038  -4.620 3.86e-06 ***
## factor(clust_pair)2  -0.014182   0.014713  -0.964 0.335099
## factor(clust_pair)3   0.058152   0.014579   3.989 6.66e-05 ***
## factor(clust_pair)4   0.017353   0.013633   1.273 0.203096
## factor(clust_pair)5   0.045738   0.013859   3.300 0.000968 ***
## factor(clust_pair)6  -0.049744   0.013600  -3.658 0.000255 ***
## factor(clust_pair)7  -0.034381   0.013596  -2.529 0.011451 *
## factor(clust_pair)8  -0.001924   0.013911  -0.138 0.889973
## factor(clust_pair)9  -0.026014   0.013310  -1.954 0.050657 .
## factor(clust_pair)10 -0.012914   0.013948  -0.926 0.354534
## factor(clust_pair)11 -0.017870   0.013340  -1.340 0.180398
```

```
## factor(clust_pair)12 -0.011456   0.014073  -0.814 0.415630    
## factor(clust_pair)13 -0.029014   0.014087  -2.060 0.039441 *  
## factor(clust_pair)14 -0.049441   0.014009  -3.529 0.000418 ***
## factor(clust_pair)15 -0.037715   0.013947  -2.704 0.006852 ** 
## factor(clust_pair)16 -0.053317   0.015292  -3.486 0.000490 ***
## factor(clust_pair)17  0.015127   0.013524   1.119 0.263354    
## factor(clust_pair)18 -0.017971   0.014219  -1.264 0.206292    
## factor(clust_pair)19 -0.051250   0.013971  -3.668 0.000245 ***
## factor(clust_pair)20 -0.049339   0.013640  -3.617 0.000298 ***
## factor(clust_pair)21 -0.033333   0.015427  -2.161 0.030723 *  
## factor(clust_pair)22 -0.057363   0.013633  -4.208 2.59e-05 ***
## factor(clust_pair)23 -0.026068   0.014639  -1.781 0.074970 .  
## factor(clust_pair)24 -0.022492   0.013687  -1.643 0.100337    
## factor(clust_pair)25 -0.027589   0.013741  -2.008 0.044669 *  
## factor(clust_pair)26  0.004885   0.015536   0.314 0.753170    
## factor(clust_pair)27 -0.059436   0.013760  -4.320 1.57e-05 ***
## factor(clust_pair)28  0.023168   0.014739   1.572 0.115988    
## factor(clust_pair)29 -0.060568   0.013249  -4.572 4.87e-06 ***
## factor(clust_pair)30 -0.060566   0.013849  -4.373 1.23e-05 ***
## factor(clust_pair)31 -0.017097   0.013097  -1.305 0.191791    
## factor(clust_pair)32 -0.039282   0.014945  -2.628 0.008582 ** 
## factor(clust_pair)33 -0.068846   0.014701  -4.683 2.84e-06 ***
## factor(clust_pair)34 -0.068788   0.014949  -4.601 4.22e-06 ***
## factor(clust_pair)35 -0.029052   0.013620  -2.133 0.032937 *  
## factor(clust_pair)36 -0.048885   0.013894  -3.519 0.000435 ***
## factor(clust_pair)37 -0.059705   0.014701  -4.061 4.90e-05 ***
## factor(clust_pair)38 -0.067755   0.014653  -4.624 3.79e-06 ***
## factor(clust_pair)39 -0.043275   0.016246  -2.664 0.007731 ** 
## factor(clust_pair)40 -0.019373   0.013408  -1.445 0.148491    
## factor(clust_pair)41 -0.035366   0.013299  -2.659 0.007836 ** 
## factor(clust_pair)42 -0.060502   0.013721  -4.409 1.04e-05 ***
## factor(clust_pair)43 -0.072109   0.015998  -4.507 6.59e-06 ***
## factor(clust_pair)44 -0.058985   0.019460  -3.031 0.002439 ** 
## factor(clust_pair)45 -0.052417   0.014660  -3.576 0.000350 ***
## factor(clust_pair)46 -0.038360   0.013962  -2.747 0.006011 ** 
## factor(clust_pair)47 -0.056556   0.013438  -4.209 2.58e-05 ***
## factor(clust_pair)48 -0.043262   0.015005  -2.883 0.003940 ** 
## factor(clust_pair)49 -0.043938   0.017269  -2.544 0.010954 *  
## factor(clust_pair)50 -0.079787   0.015284  -5.220 1.80e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2299 on 23536 degrees of freedom
##   (41 observations deleted due to missingness)
## Multiple R-squared:  0.01682,    Adjusted R-squared:  0.01473 
## F-statistic: 8.055 on 50 and 23536 DF,  p-value: < 2.2e-16
```

Our second set of regressions shows a similar relationship to our initial set of regressions. Coefficients lose precision. And our standard errors grow. Our F-statistic is less than 10, again indicating a weak instrument.

(8)

Now adjust for baseline characteristics. Again, do this with and without the cluster-pair dummies in the regression. What happens to the estimated treatment effects? To their standard errors?

Answer

```r
itt5 <- lm(data = data, extreme20dummy06 ~ treatment + sex + age + hhsize + headwomen_05)
sumitt5 <- summary(itt5)
sumitt5
```

```
##
## Call:
## lm(formula = extreme20dummy06 ~ treatment + sex + age + hhsize +
##     headwomen_05, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.11386 -0.06144 -0.04851 -0.03659  0.99818
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.631e-02  8.006e-03   4.536 5.76e-06 ***
## treatment    -1.471e-02  2.877e-03  -5.115 3.16e-07 ***
## sex           4.921e-03  3.148e-03   1.563    0.118
## age           7.057e-04  9.196e-05   7.674 1.73e-14 ***
## hhsize       -3.645e-03  7.603e-04  -4.794 1.64e-06 ***
## headwomen_05  1.488e-03  3.710e-03   0.401    0.688
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2208 on 23596 degrees of freedom
##   (26 observations deleted due to missingness)
## Multiple R-squared:  0.006373,   Adjusted R-squared:  0.006163
## F-statistic: 30.27 on 5 and 23596 DF,  p-value: < 2.2e-16
```

```r
itt6 <- lm(data = data, extreme20dummy06 ~ treatment + sex + age + hhsize + headwomen_05 + factor(clust_
sumitt6 <- summary(itt6)
sumitt6
```

```
##
## Call:
## lm(formula = extreme20dummy06 ~ treatment + sex + age + hhsize +
##     headwomen_05 + factor(clust_pair), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19147 -0.06830 -0.04468 -0.02258  1.01637
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.0831810  0.0121404   6.852 7.48e-12 ***
## treatment           -0.0142749  0.0028962  -4.929 8.33e-07 ***
## sex                  0.0035315  0.0031404   1.125 0.260798
## age                  0.0005911  0.0000922   6.412 1.47e-10 ***
## hhsize              -0.0048860  0.0007655  -6.383 1.77e-10 ***
## headwomen_05         0.0024611  0.0037235   0.661 0.508641
## factor(clust_pair)2 -0.0071249  0.0140455  -0.507 0.611970
## factor(clust_pair)3  0.0575382  0.0139028   4.139 3.51e-05 ***
## factor(clust_pair)4 -0.0030653  0.0129964  -0.236 0.813544
```

```
## factor(clust_pair)5   0.0394207  0.0132124   2.984 0.002852 **
## factor(clust_pair)6  -0.0481278  0.0129759  -3.709 0.000209 ***
## factor(clust_pair)7  -0.0424301  0.0129681  -3.272 0.001070 **
## factor(clust_pair)8   0.0046063  0.0132655   0.347 0.728418
## factor(clust_pair)9  -0.0116769  0.0127021  -0.919 0.357955
## factor(clust_pair)10 -0.0189558  0.0133208  -1.423 0.154743
## factor(clust_pair)11 -0.0191398  0.0127248  -1.504 0.132561
## factor(clust_pair)12 -0.0211922  0.0134153  -1.580 0.114189
## factor(clust_pair)13 -0.0380273  0.0134342  -2.831 0.004649 **
## factor(clust_pair)14 -0.0598992  0.0133672  -4.481 7.46e-06 ***
## factor(clust_pair)15 -0.0330780  0.0133043  -2.486 0.012916 *
## factor(clust_pair)16 -0.0626574  0.0145875  -4.295 1.75e-05 ***
## factor(clust_pair)17  0.0191432  0.0128903   1.485 0.137533
## factor(clust_pair)18 -0.0290534  0.0135628  -2.142 0.032192 *
## factor(clust_pair)19 -0.0492855  0.0133265  -3.698 0.000218 ***
## factor(clust_pair)20 -0.0367350  0.0130147  -2.823 0.004768 **
## factor(clust_pair)21 -0.0351113  0.0147287  -2.384 0.017140 *
## factor(clust_pair)22 -0.0566907  0.0130025  -4.360 1.31e-05 ***
## factor(clust_pair)23 -0.0223241  0.0139704  -1.598 0.110067
## factor(clust_pair)24 -0.0194722  0.0130596  -1.491 0.135968
## factor(clust_pair)25 -0.0345788  0.0131095  -2.638 0.008353 **
## factor(clust_pair)26  0.0025783  0.0147558   0.175 0.861293
## factor(clust_pair)27 -0.0548883  0.0131393  -4.177 2.96e-05 ***
## factor(clust_pair)28  0.0107095  0.0140658   0.761 0.446432
## factor(clust_pair)29 -0.0672404  0.0126497  -5.316 1.07e-07 ***
## factor(clust_pair)30 -0.0617259  0.0132149  -4.671 3.01e-06 ***
## factor(clust_pair)31 -0.0299599  0.0125162  -2.394 0.016688 *
## factor(clust_pair)32 -0.0539430  0.0142575  -3.783 0.000155 ***
## factor(clust_pair)33 -0.0791680  0.0140087  -5.651 1.61e-08 ***
## factor(clust_pair)34 -0.0733023  0.0142657  -5.138 2.79e-07 ***
## factor(clust_pair)35 -0.0337655  0.0130099  -2.595 0.009455 **
## factor(clust_pair)36 -0.0551829  0.0132563  -4.163 3.16e-05 ***
## factor(clust_pair)37 -0.0722649  0.0140262  -5.152 2.60e-07 ***
## factor(clust_pair)38 -0.0679679  0.0139838  -4.860 1.18e-06 ***
## factor(clust_pair)39 -0.0520258  0.0154970  -3.357 0.000789 ***
## factor(clust_pair)40 -0.0251565  0.0127857  -1.968 0.049132 *
## factor(clust_pair)41 -0.0395403  0.0126830  -3.118 0.001826 **
## factor(clust_pair)42 -0.0623536  0.0130987  -4.760 1.94e-06 ***
## factor(clust_pair)43 -0.0702442  0.0152456  -4.608 4.10e-06 ***
## factor(clust_pair)44 -0.0549780  0.0185563  -2.963 0.003052 **
## factor(clust_pair)45 -0.0595994  0.0139881  -4.261 2.05e-05 ***
## factor(clust_pair)46 -0.0451092  0.0133335  -3.383 0.000718 ***
## factor(clust_pair)47 -0.0694909  0.0128226  -5.419 6.04e-08 ***
## factor(clust_pair)48 -0.0510828  0.0143231  -3.566 0.000363 ***
## factor(clust_pair)49 -0.0441453  0.0164713  -2.680 0.007364 **
## factor(clust_pair)50 -0.0756575  0.0145711  -5.192 2.09e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2191 on 23547 degrees of freedom
##   (26 observations deleted due to missingness)
## Multiple R-squared:  0.02394,    Adjusted R-squared:  0.0217
## F-statistic: 10.69 on 54 and 23547 DF,  p-value: < 2.2e-16
```

We've chosen the following variables to be our baseline: sex, age, hhsize, & headwomen_05. As with previous examples, we note similar differences between the coefficients of our baseline values in the regression without dummy variables for cluster pair and with. The coefficients become less precise. The standard errors increase.

(9)

Are any of your conclusions sensitive to the way that you measure extreme health expenditures?

Answer

I do not believe so. The coefficients for the health share are larger than the coefficients for the adjusted health share in either regression (dummy variables or not). But they do not exhibit what might be a statistically significant difference from each other. As such they both seem to be similarly effective parameters.

(10)

Now calculate the Wald estimator. How do you interpret it? What additional assumptions do you need to invoke for these estimates to be valid?

Answer

```r
firststage <- lm(data = data, insp2006 ~ treatment)
fs <- summary(firststage)
fs
```

```
##
## Call:
## lm(formula = insp2006 ~ treatment, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45131 -0.45131 -0.06642  0.54869  0.93358
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.066420   0.003667   18.11   <2e-16 ***
## treatment   0.384892   0.005145   74.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3954 on 23626 degrees of freedom
## Multiple R-squared:  0.1915, Adjusted R-squared:  0.1915
## F-statistic:  5596 on 1 and 23626 DF,  p-value: < 2.2e-16
```

```r
insp2006hat <- firststage$fitted.values
secondstage <- lm(data = data, extreme20dummy06 ~ insp2006hat)
ss <- summary(secondstage)
ss
```

```
##
## Call:
```

```
## lm(formula = extreme20dummy06 ~ insp2006hat, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05932 -0.05932 -0.04438 -0.04438  0.95562
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.061904   0.002434  25.436  < 2e-16 ***
## insp2006hat -0.038835   0.007488  -5.186 2.17e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2214 on 23600 degrees of freedom
##   (26 observations deleted due to missingness)
## Multiple R-squared:  0.001138,   Adjusted R-squared:  0.001096
## F-statistic: 26.89 on 1 and 23600 DF,  p-value: 2.167e-07
```

```
ss$coefficients[2]
```

```
## [1] -0.03883491
```

In order to calculate the Wald estimator, we need several assumptions to hold. We need there to be no interference or that there exist no spillover from one sample to another. We require that Z has no independent effect on Y, except through D. We need there to be a significant difference in D, given either value of Z. And we also need the monotonicity assumption, highlight prior, to hold.