w241 Final Project

Cendy Lin

April 22, 2017

Load Packages

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
library(reshape2)
library(car)
library(stargazer)
```

```
## Warning: package 'stargazer' was built under R version 3.3.2
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
library(foreign)
library(multiwayvcov)
```

```
## Warning: package 'multiwayvcov' was built under R version 3.3.2
```

Prepare Data

```r
data = read.csv("./data.csv")

# Remove "full attritors" where we have no treatment
df = data[data$full_attrit==0,]

# view columns
names(df)
```

```
##  [1] "username"     "id"           "pilot"        "treat"
##  [5] "full_attrit"  "block"        "num_phones"   "age"
##  [9] "gender"       "os"           "phone_use"    "contact"
## [13] "relationship" "b1"           "b2"           "b3"
## [17] "b4"           "b5"           "b6"           "b7"
## [21] "t1"           "t2"           "t3"           "t4"
## [25] "t5"           "t6"           "t7"           "c1"
## [29] "c2"           "c3"           "c4"           "c5"
## [33] "c6"           "c7"
```

```r
# drop username row for anonimity
df = df[,-1]

# encode gender (male = 1, female = 0)
df$male = 1
df$male[df$gender == "Female"] = 0

# encode age (<44 = 1, >44 = 0)
df$age_code = 1
df$age_code[df$age == "45-54" | df$age == "55-64"] = 0
table(df$age_code)
```

```
##
##  0  1
## 10 46
```

```r
# encode OS (Apple = 1, Other = 0)
df$apple = 1
df$apple[grep("Android", df$os)] = 0 # regex, if contains 'Android', encode as 1
df$apple[df$os == "Blackberry"] = 0

# encode num_phones (personal use phone = 0, work & personal phone = 1)
df$personal_phone = 0
df$personal_phone[grep("only", df$num_phones)] = 1

# encode phone_use (at least once an hour = 1, less than once an hour = 0)
df$phone_usage = 1
df$phone_usage[grep("4", df$phone_use)] = 0 # regex, if contains '4', encode as 1

# encode relationship (friend/fam = 1, other = 0)
df$relation = 1
df$relation[df$relationship == ""] = 0

# label treat as treatment/control
```

```
df$treat_group = "treatment"
df$treat_group[df$treat == 0] = "control"

# drop columns
df = df[,c(-2, -4, -5, -6, -7, -8 ,-9, -10, -11, -12)]

# reorder columns so stress observations on the right
df = df[c("id", "treat", "treat_group", "male", "age_code", "apple", "personal_phone",
     "phone_usage", "relation", "c1", "c2", "c3", "c4", "c5", "c6", "c7",
     "b1", "b2", "b3", "b4", "b5", "b6", "b7",
     "t1", "t2", "t3", "t4", "t5", "t6", "t7")]

# change column names
colnames(df) <- c("id", "treat_code", "treat_group", "male", "age", "apple",
                  "personal_phone","phone_use", "relationship", "c1", "c2", "c3",
                  "c4", "c5", "c6", "c7", "b1", "b2", "b3", "b4", "b5", "b6",
                  "b7", "t1", "t2", "t3", "t4", "t5", "t6", "t7")
```

Codes: * comply: comply=1, non-comply=0 for treatment group only * male: male=1, female=0 * age: <44=0, >44=1 * apple: apple=1, other=0 * phone_usage: at least once an hour=0, less than once an hour=1 * peronal_phone: personal use phone = 0, work & personal phone = 1 * relation: friend/fam=0, other=1

Reshape into indivdual observations by day

```
df2 = melt(df, id.var = c("id", "treat_code", "treat_group", "male", "age", "apple",
                  "personal_phone","phone_use", "relationship", "c1", "c2", "c3",
                  "c4", "c5", "c6", "c7"), variable.name = "day")
```

```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```

```
# recode 'day' column as number day
index = levels(df2$day)
values = c(1:14)
df2$day = values[match(df2$day, index)]

# reformat compliance columns
df2$comply = NA
a = which(df2$day == 8)
df2$comply[a] = df2$c1[a] # turns into factor
b = which(df2$day == 9)
df2$comply[b] = df2$c2[b]
c = which(df2$day == 10)
df2$comply[c] = df2$c3[c]
d = which(df2$day == 11)
df2$comply[d] = df2$c4[d]
e = which(df2$day == 12)
df2$comply[e] = df2$c5[e]
f = which(df2$day == 13)
df2$comply[f] = df2$c6[f]
g = which(df2$day == 14)
```

```
df2$comply[g] = df2$c7[g]

# Replace original values instead of factors
values = c(NA, 0, 1, "missing")
index = c(1:4)
df2$comply = values[match(df2$comply, index)]

# drop c1-c7, rename 'value' to 'stress'
df2 = df2[,c(-10, -11, -12, -13, -14, -15, -16)]
colnames(df2)[11] <- "stress"

head(df2)
```

```
##   id treat_code treat_group male age apple personal_phone phone_use
## 1  1          1   treatment    0   1     1              0         0
## 2  2          0     control    0   1     1              1         1
## 3  4          1   treatment    0   1     1              0         1
## 4  5          0     control    1   1     1              1         1
## 5  6          0     control    0   1     1              0         1
## 6  7          0     control    1   1     1              1         1
##   relationship day stress comply
## 1            0   1      2   <NA>
## 2            0   1      3   <NA>
## 3            1   1      2   <NA>
## 4            0   1      2   <NA>
## 5            1   1      2   <NA>
## 6            1   1      3   <NA>
```

```
nrow(df2)
```

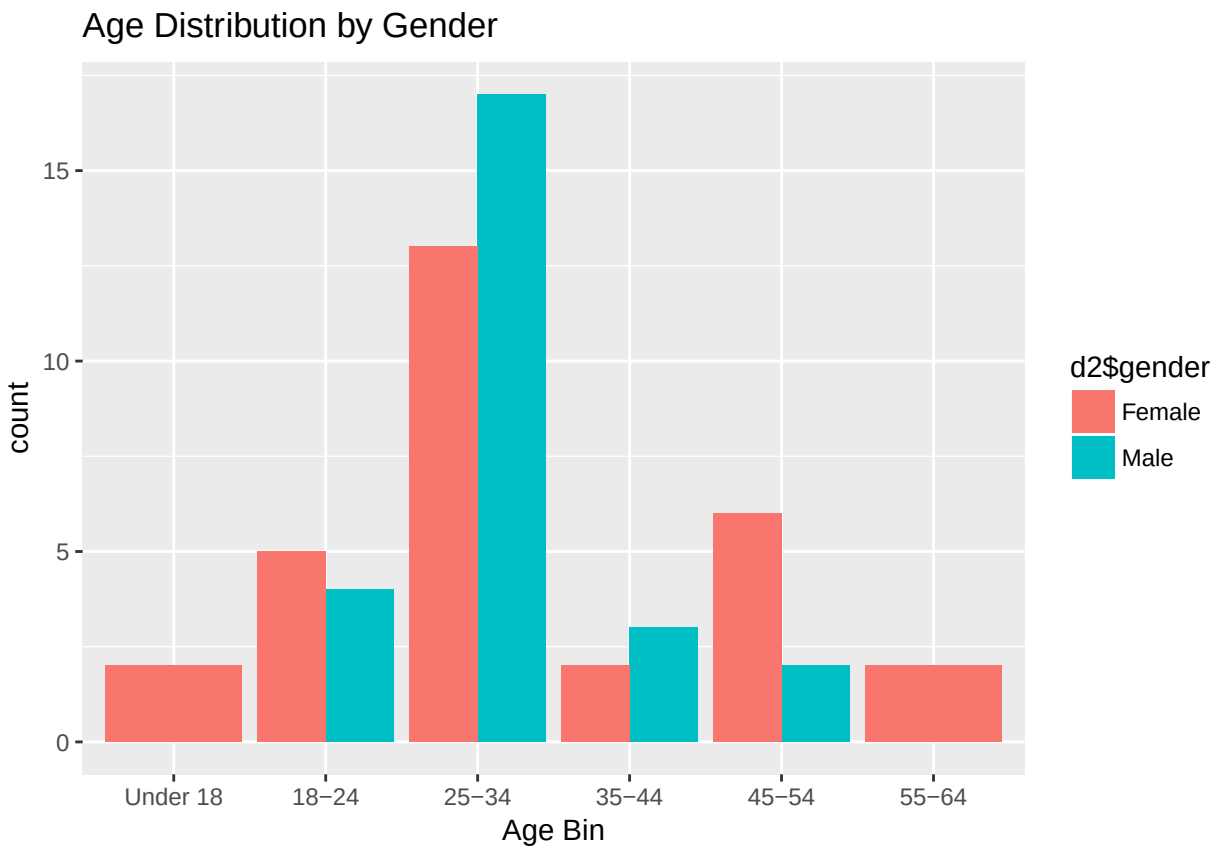
```
## [1] 784
```

Exploratory Analysis of Participants

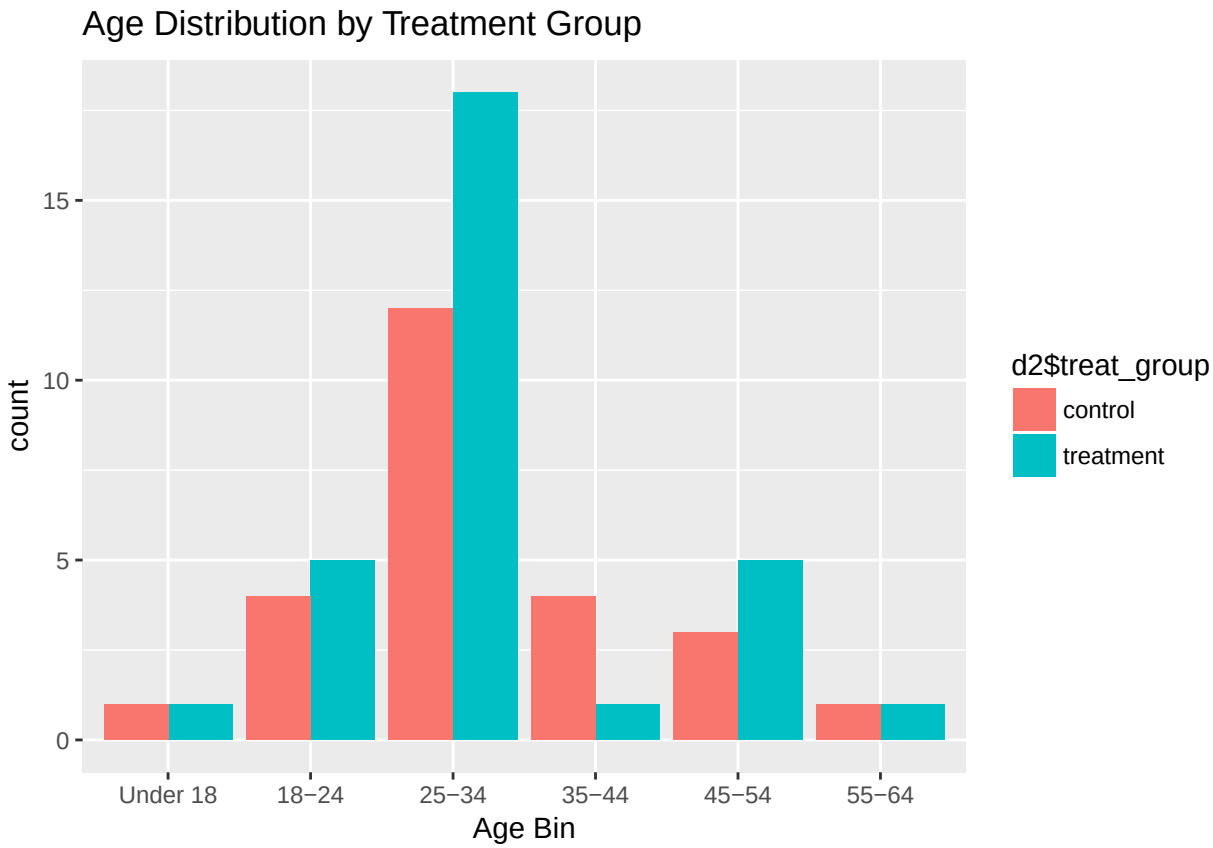
```
# Age and Gender
d2 = data[data$full_attrit == 0,]

# label treat as treatment/control
d2$treat_group = "treatment"
d2$treat_group[d2$treat == 0] = "control"

position = c("Under 18", "18-24", "25-34", "35-44", "45-54", "55-64")
ggplot(data.frame(d2$age), aes(x=d2$age, fill = d2$gender)) +
  geom_bar(position="dodge") +
  scale_x_discrete(limits = position) +
  ggtitle("Age Distribution by Gender") +
  labs(x="Age Bin")
```




```
# Age and Treatment Group
ggplot(data.frame(d2$age), aes(x=d2$age, fill = d2$treat_group)) +
  geom_bar(position="dodge") +
  scale_x_discrete(limits = position) +
  ggtitle("Age Distribution by Treatment Group") +
  labs(x="Age Bin")
```



Covariate balance check

```
# Not using individual observations

# gender
t.test(df$male ~ df$treat_code, var.equal=F) # not significant
```

```
##
##  Welch Two Sample t-test
##
## data:  df$male by df$treat_code
## t = -0.85787, df = 51.847, p-value = 0.3949
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3877852  0.1555271
## sample estimates:
## mean in group 0 mean in group 1
##        0.400000        0.516129
```

```
# age
t.test(df$age ~ df$treat_code, var.equal=F) # not significant
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  df$age by df$treat_code
## t = 0.32278, df = 52.83, p-value = 0.7481
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1749386  0.2420354
## sample estimates:
## mean in group 0 mean in group 1
##       0.8400000       0.8064516
```

```
# apple
t.test(df$apple ~ df$treat_code, var.equal=F) # not significant
```

```
##
##  Welch Two Sample t-test
##
## data:  df$apple by df$treat_code
## t = -1.0546, df = 42.993, p-value = 0.2975
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.30061858  0.09416697
## sample estimates:
## mean in group 0 mean in group 1
##       0.8000000       0.9032258
```

```
# personal phone
t.test(df$personal_phone ~ df$treat_code, var.equal=F) # not significant
```

```
##
##  Welch Two Sample t-test
##
## data:  df$personal_phone by df$treat_code
## t = -0.41837, df = 52.607, p-value = 0.6774
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2916205  0.1909754
## sample estimates:
## mean in group 0 mean in group 1
##       0.2400000       0.2903226
```

```
# phone use
t.test(df$phone_use ~ df$treat_code, var.equal=F) # not significant
```

```
##
##  Welch Two Sample t-test
##
## data:  df$phone_use by df$treat_code
## t = 0.21691, df = 52.997, p-value = 0.8291
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1383351  0.1718835
```

```
## sample estimates:
## mean in group 0 mean in group 1
##       0.9200000       0.9032258
```

```
# relationship
t.test(df$relationship ~ df$treat_code, var.equal=F) # not significant
```

```
##
##  Welch Two Sample t-test
##
## data:  df$relationship by df$treat_code
## t = 0.083541, df = 51.661, p-value = 0.9337
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2376637  0.2583089
## sample estimates:
## mean in group 0 mean in group 1
##       0.7200000       0.7096774
```

Initial DiD Analysis - ignoring attrition

```
# create rows for difference-in-difference
df2$stress = as.numeric(df2$stress)
```

```
## Warning: NAs introduced by coercion
```

```
df2$stress
```

```
##   [1]  2  3  2  2  2  3  4  2  2  3  3 NA  3  4 NA  2  2 NA  3  4  1 NA  2
##  [24]  3  3  2  2  3  2 NA  4  2 NA  2  3  3  4  3  4  4  4  4  3 NA  3  1
##  [47]  2  2  2  3 NA  5  4  4  4  2  4  3  2  3  1  3  3  4  3  2 NA  4  2
##  [70]  3 NA  4  2  2 NA  3  1 NA  3  3  3  2  1  2  3 NA  4  2  4  4  2  3
##  [93]  4  2  5  4  4  4  3 NA  3  2  3  1  2  3 NA  4  3  2 NA  3  4  2  3
## [116]  2  4  3  4 NA  4  3  2  4  3 NA NA  2  2  2  3  4  1 NA  2  1  2  3
## [139]  1  3  4  4  5  1  2  2  2  5  3  2  3  2  4  2  4  3  5  3  2  5 NA
## [162]  3  3  4  3  2  3  3  3  4  2  2  1  3  3  4  3  2  2  2 NA NA NA  2
## [185]  3  2  2  4  1  2  3  3  3  2  2  2  2  4  2  1  2  2  3  2  2  2 NA
## [208] NA  3  2  4  4  4  2  2  2 NA  3  2  3  4  2  3 NA  4  4  1  2  3  4
## [231]  3 NA  4  2 NA  1  2 NA NA  3  2  3  2  4  1  3  2  1  3  2  2  1  3
## [254]  3  2 NA  2  2  2  3  4  2 NA NA  3  2  4  3  2  2  1  1  2  4  2  5
## [277]  3  3  3  2  4  2 NA  2  3  3  2  3  3  2  2  1  2  2 NA  4  3  1  2
## [300]  2  1  1  3  1  2  2  2  3 NA  2  2 NA  1  1  2  2  3  2  5 NA  2  1
## [323]  3  3  2  2  1  2  4  2  2  5  4  2  2  1 NA  3 NA NA  4  3  3  3  2
## [346]  2  2 NA  2  3 NA  4  4  1 NA  1  1  2  4  3  2  2  2  2  4 NA  3  2
## [369]  1  3  3  5  4  2  5  4  2  2  3 NA  2  2  2  3  4  2  2  5  4  1  2
## [392]  2  3  4 NA  3  2  4  3  5  3  2 NA  3  2 NA  2  2 NA  1  4  5  1  3
## [415]  4  3  6  3  2  1  2  3 NA  2  4  2  3  5  5  3  3  3  2  4  5  3  4
## [438]  4  3  4  3  3  5  4  5  4  4  3  3  4  3  2  3  4  3  4  2  2  3  4
## [461]  2 NA NA  2  2  2  1  4  1  2  5 NA  4  3  3  3  3  3  4  4  4  2  4
## [484]  3  3  3  3  2  3  4  5  3  2  3  3  4 NA  4  3  5  5  4 NA  3  4  2
```

```
## [507] NA  2  2  4  3  4  2 NA NA  2  2  3  2  4  2  3 NA  5  2  3  4 NA  2
## [530] NA  4  3  2  3  3  2  2  2 NA  3  3  2  3  3  3  3  3  4  4  2  1  1
## [553] NA  2 NA  2  3  4  3 NA  4  2 NA  2 NA  2  4  3  4 NA NA  4  3  4 NA
## [576]  3  2  3 NA  5  2  3  2 NA  2 NA  2  2  2 NA  3  2  1  1  3  5  2  2
## [599]  3  3  2  3  3  3  4  2  2  1 NA  4  1  2  5  3  2  2  4  4 NA NA NA
## [622]  2  4 NA NA  2 NA  2  2 NA NA  4  2  3 NA  1  1  3 NA NA  2  1  2  1
## [645]  2 NA  3 NA  2  1  2  4  4  2  3 NA  3  2 NA  4  2  4  2  1 NA  3  2
## [668]  1  6  2 NA  1  1  2  3 NA NA  2  2  2  1  1 NA  2  3  3  2  3  2  2
## [691] NA  1  1  1  3 NA  2  2  2  1  1  2  2 NA  1  1  2  3 NA  1  2  2  2
## [714]  2  3  3  2  4  1  1 NA  2  2  2  5  2  2  1  3  3 NA NA NA  3  2 NA
## [737]  2  2 NA NA  2  4  2  4  2  2 NA  1  1 NA  2 NA  2  2  3  2  1  2  3
## [760]  4  2  1  2  3 NA NA  4  3  2  2  3  3  1  2  2  2 NA  2  1  3  3  1
## [783]  2  1
```

```
baseline = df2[df2$day < 8,]
treatment = df2[df2$day > 7,]
did = treatment$stress - baseline$stress

# Build dataframe with DiD as outcome only
treat = df2[c(1:length(did)),]$treat_group
treat_code = df2[c(1:length(did)),]$treat_code
id = df2[c(1:length(did)),]$id
day = df2[c(1:length(did)),]$day
diff = data.frame(id, treat, treat_code, day, did)

# count of NA differences in treatment groups
sum(is.na(diff$did[diff$treat == "control"])) #38
```

```
## [1] 38
```

```
sum(is.na(diff$did[diff$treat == "treatment"])) #61
```

```
## [1] 61
```

```
38/sum(diff$treat == "control")
```

```
## [1] 0.2171429
```

```
61/sum(diff$treat == "treatment")
```

```
## [1] 0.281106
```

```
# distribution by treatment group
ggplot(data.frame(diff$did), aes(x=diff$did, fill = diff$treat)) +
  geom_bar(position="dodge") +
  ggtitle("Difference in Difference of Stress Levels by Day of Week") +
  labs(x="Stress DiD")
```

```
## Warning: Removed 99 rows containing non-finite values (stat_count).
```


```
# boxplot by treatment group
ggplot(data.frame(diff$did), aes(x=diff$treat, y=diff$did, fill=diff$treat)) +
  geom_boxplot() +
  ggtitle("Difference in Difference of Stress Levels by Day of Week") +
  labs(x="Stress DiD")
```

```
## Warning: Removed 99 rows containing non-finite values (stat_boxplot).
```


```r
# model ignoring attrition and no clusters on ID
m1_wrong = lm(did ~ treat, data=diff)
summary(m1_wrong)
```

```
##
## Call:
## lm(formula = did ~ treat, data = diff)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8013 -0.8013  0.1987  0.7883  3.1987
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.2117     0.1021   2.073  0.03905 *
## treattreatment  -0.4104     0.1399  -2.933  0.00363 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.195 on 291 degrees of freedom
##   (99 observations deleted due to missingness)
## Multiple R-squared:  0.0287, Adjusted R-squared:  0.02537
## F-statistic:   8.6 on 1 and 291 DF,  p-value: 0.003629
```

```r
# model WITH clusters on ID but ignoring attrition
m1_wrong$cluster.vcov = cluster.vcov(m1_wrong, ~ id)
m1 = coeftest(m1_wrong, m1_wrong$cluster.vcov)

# model ignoring attrition and using fixed effects by individual ID
m2 = lm(did ~ treat + factor(id), data = diff)

m2$cluster.vcov = cluster.vcov(m2, ~ id)
m2a = coeftest(m2, m2$cluster.vcov)

# Compare models
stargazer(m1, m2, m2a, type="latex", omit = "id",
          dep.var.labels.include = FALSE,
          add.lines = list(c("Fixed effects?", "No", "Yes", "Yes"),
                           c("Clustered SE?", "Yes", "No", "Yes")),
          column.labels   = c("Clustered SE", "Fixed Effects", "Both"))
```

```
##
## % Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvar
## % Date and time: Mon, Apr 24, 2017 - 10:12:55 PM
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{3}{c}{\textit{Dependent variable:}} \\
## \cline{2-4}
## \\[-1.8ex] & \textit{coefficient} & \textit{OLS} & \textit{coefficient} \\
##  & \textit{test} & \textit{} & \textit{test} \\
##  & Clustered SE & Fixed Effects & Both \\
## \\[-1.8ex] & (1) & (2) & (3)\\
## \hline \\[-1.8ex]
##  treattreatment & $-$0.410$^{***}$ & $-$0.905 & $-$0.905$^{***}$ \\
##   & (0.157) & (0.650) & (0.000) \\
##   & & & \\
##  Constant & 0.212$^{**}$ & 0.571 & 0.571$^{***}$ \\
##   & (0.088) & (0.441) & (0.000) \\
##   & & & \\
## \hline \\[-1.8ex]
## Fixed effects? & No & Yes & Yes \\
## Clustered SE? & Yes & No & Yes \\
## Observations &  & 293 &  \\
## R$^{2}$ &  & 0.239 &  \\
## Adjusted R$^{2}$ &  & 0.070 &  \\
## Residual Std. Error &  & 1.168 (df = 239) &  \\
## F Statistic &  & 1.415$^{**}$ (df = 53; 239) &  \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{3}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

Dealing with Attrition