Problem Set 1

Se Hyun Kim

Problem 1A

```
a1 <-  ig %>%
  group_by(type, late.luck) %>%
  summarize(kept_mean = mean(kept), n = n())
```

```
## `summarise()` regrouping output by 'type' (override with `.groups` argument)
```

```
a1 <- as_tibble(a1)

tbl1 <- data.frame(
  "row" = c("Low", "High"),
  "not" = c(a1[[1,3]], a1[[3,3]]),
  "lucky" = c(a1[[2,3]], a1[[4,3]]))

kbl(tbl1,
  align=rep('c'),
  col.names = c("", "Not Lucky", "Lucky"),
  caption = "Demo Table",
  booktabs = T,
  digits = 2) %>%
  kable_styling(full_width = F)
```

The coefficient on `late.luck` suggests that if the last four payments before the retention decision were larger than average, the participant is 6% more likely to keep his/her allocator. It *does not* mean that if there were 100 such cases, 6 more cases will keep the allocator.

Problem 1B

```
areg <- lm(kept ~ late.luck + type, ig)
stargazer(areg,
          type = "text",
```

Table 1: Demo Table

	Not Lucky	Lucky
Low	0.56	0.60
High	0.75	0.81

```
          column.sep.width = "1pt",
          no.space = TRUE,
          header = FALSE,
          title = "Regression Results for Problem 1A",
          covariate.labels  = c("late.luck",
                                "type"),
          dep.var.labels = c("kept"))
```

```
##
## Regression Results for Problem 1A
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                  kept
## ------------------------------------------------
## late.luck                       0.061*
##                                 (0.032)
## type                            0.191***
##                                 (0.039)
## Constant                        0.558***
##                                 (0.037)
## ------------------------------------------------
## Observations                      990
## R2                                0.027
## Adjusted R2                       0.025
## Residual Std. Error         0.435 (df = 987)
## F Statistic              13.868*** (df = 2; 987)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Problem 1C

Estimate is a particular value from an estimator (a rule to produce a numerical value that represents the estimand) for a certain sample. Conditional means that the occurrence of an event given that another event(s) have already happened. Expectation is the expected value. Assuming Y and X are discrete random variables, the conditional expectation is the values of Y given certain value of X. Therefore, estimates of conditional expectation is the particular value of Y given a certain value of X.

Problem 1D

Yes, the estimated conditional expectations (ECE) can be recovered from the regression output. Simply, the equation would be `kept` $= \beta_0 + \beta_1$`late.luck` $+ \beta_2$`type`. For 'Not lucky'X'low', `late.luck` and `type` will be 0. `kept` will be β_0 which is 0.558. For 'Not Luck'X'high', `late.luck` will be 0 and `type` will be 1. `kept` will be $\beta_0 + \beta_2$ which is 0.749. In the like manner, other ECEs can be recovered. The regression is controlling for `type` and `late.luck`.

Problem 1E

For example, by the explanation of the experiment, the value of `late.luck` is associated with `type`. The higher payment giving allocator you have, you are more likely to get a result above average. Therefore, an interaction term could improve the fit.

Problem 1F

```
reg1f <- lm(kept ~ late.luck + type + late.luck*type, ig)
stargazer(reg1f,
          type = "text",
          column.sep.width = "1pt",
          no.space = TRUE,
          header = FALSE,
          title = "Regression Results for Problem 1F",
          covariate.labels  = c("late.luck",
                                "type",
                                "Interaction"),
          dep.var.labels = c("kept"))
```

```
##
## Regression Results for Problem 1F
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                  kept
## ------------------------------------------------
## late.luck                        0.038
##                                 (0.090)
## type                            0.185***
##                                 (0.045)
## Interaction                      0.027
##                                 (0.096)
## Constant                        0.563***
##                                 (0.041)
## ------------------------------------------------
## Observations                      990
## R2                               0.027
## Adjusted R2                      0.024
## Residual Std. Error       0.435 (df = 986)
## F Statistic           9.263*** (df = 3; 986)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

However, adding the interaction term makes it difficult to recover the ECEs. Also, the test fails to reject the null hypothesis. There is no interaction term.

Problem 1G

No, the variables are not independent. Being lucky, that is getting a higher than average payment depends on the type of allocator participants have. The `type` variable is a confounder. Therefore, causal effect cannot be argued.

Problem 2A

Y_{1i} and Y_{0i} represent potential outcomes. Y_{1i} represents the outcome given treatment for individual unit i and Y_{0i} represents the outcome without treatment for individual unit i. In reality, only one of the potential

Table 2: Table 2C

i	D_i	Y_1i	Y_0i	t_i	Y_i
1	1	10	4	6	10
2	1	1	2	-1	1
3	0	3	3	0	3
4	0	5	2	3	2

outcomes is observed. For example, imagine a researcher handing out pieces of paper to 20 students. Each of these pieces have two values written on it: Y_{1i} and Y_{0i}. If a student is assigned to a treatment group, they inform the researcher of the value of Y_{1i}. If a student is assigned to a control group (without treatment), they inform the researcher of the value of Y_{0i}. In short, every individual unit has two potential outcomes, when they are given treatment and when they are not. However, single individual unit cannot be both given treatment and not given treatment. Only one of the two potential outcomes can be observed.

Problem 2B

The difference between Y_{1i} and "Y_i for a unit that actually received the treatment" (hereafter Y_{ac}) is that Y_{1i} represent a *potential outcome* and Y_{ac} denotes the *observed value* for an individual unit that *was* given treatment. As explained above, a researcher cannot observe both of the potential outcomes. The student has both Y_{1i} and Y_{0i} in the handed out piece of paper. However, once the student is assigned to the treatment group and hands in the value of Y_{1i}, Y_{1i} becomes the actually observed treatment value Y_{ac}.

Problem 2C

```
## Warning: The ‘x‘ argument of ‘as_tibble()‘ can’t be missing as of lifecycle 3.0.0.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_warnings()‘ to see where this warning was generated.
```

The average treatment effect (ATE) is the average of the treatment effect given that the researcher knows all the potential outcomes. From Table 2C, we can observe the difference between the two potential outcomes for every individual unit. Therefore, the ATE ($\mathbb{E}[\tau_i]$) is the average value of τ_i which is $\mathbb{E}[Y_{1i} - Y_{0i}] = \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}]$ (because linearity of expectation) $= \frac{6+(-1)+0+3}{4}$: 2. The average treatment effect among the treated (ATT) is the average value of τ_i among units given treatment ($D_i = 1$). Therefore, ATT is expressed as follows: $\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1] = \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 1]$ or $\mathbb{E}[\tau_i|D_i = 1]$. From Table 2C, ATT will be $\frac{6+(-1)}{2} = 2.5$

Problem 2D

The ATT and the ATE will be the same if selection into treatment is not associated with potential outcomes. For example, a researcher wants to see if hospitals have a positive effect on health. To do so, the researcher simply looks at the average life expectancy of those who have been hospitalize for more than three days and those who haven't. The researcher finds that those who haven't been hospitalized have longer average life expectancy! However, in this case, it is likely that the sick are more likely to be hospitalized. Therefore, there is a selection bias: the division of the population into the treatment group and the control group is bias because the sick are more likely to be given treatment (being hospitalized). In short, the average potential outcome of not being treated for the those who are treated and the average potential outcome of not being treated for those who are not treated are different. This can be expressed much more clearly by the following equation.

From Problem 2C, $ATT = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$. This simply states that the ATT equals the average of potential outcomes given treatment minus the average of potential outcomes not given treatment. It can be expressed as $ATT = \mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] + \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{1i}|D_i = 1] = \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1] + \mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0]$. Note that we are basically adding 0 to the original equation ($\mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{1i}|D_i = 1]$). From Problem 2C, we can see that $\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1]$ is ATT. The rest, $\mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0]$ is the selection bias, the bias like the sick being more likely being hospitalized. Based on this selection bias, we can see that the indepedance of Y_{0i} from D_i can elimiate the selection bias because $\mathbb{E}[Y_{0i}|D_i = 1]$ will equal $\mathbb{E}[Y_{0i}|D_i = 0]$. Therefore, if the selection into treatment is not associated with potential outcomes, ATT will equal ATE.

Problem 2E

Two events Y and X are independent iff $Pr(Y) \cap Pr(X) = Pr(Y)Pr(X)$. Therefore, $Pr(Y|X) = Pr(Y)$ because $Pr(Y|X) = \frac{Pr(Y \text{ and } X)}{Pr(X)} = \frac{Pr(Y \cap X)}{Pr(X)} = \frac{Pr(Y)Pr(X)}{Pr(X)}$. Similarly, two random variables Y and X and independent iff $f_{X,Y}(x, y) = f_X(x)f_Y(y)$. This implies that $f_{Y|X}(y|x) = f_Y(y)$. Therefore, no matter the value of X $\mathbb{E}[Y|X = x]$ will equal $\mathbb{E}[Y]$. If the treatment is randomly assigned, the potential outcomes for the treatment and the potential outcomes for the non-treatment will be independent. That is, $\{Y_{1i}, Y_{0i}\} \perp\!\!\!\perp D_i$. We can say that the identification of ATE is as follows. $\mathbb{E}[Y_i|D_i = 1] = \mathbb{E}[D_i \cdot Y_{1i} + (1 - D_i) \cdot Y_{0i}|D_i = 1] = \mathbb{E}[Y_{1i}|D_i = 1] = \mathbb{E}[Y_{1i}]$ (assuming SUTVA). Similarly, $\mathbb{E}[Y_i|D_i = 0] = \mathbb{E}[Y_{0i}]$. Therefore, ATE, the difference between the two groups of potential outcomes ($\mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}]$) will equal the difference of the observed outcomes of the two groups ($\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$).This argument demonstrates that $\alpha_{ATE(x)} = E[Y_1 - Y_0|X = x]$ is identifiable given random assignment.

Problem 3A

```
# comparing the mean
x %>%
  group_by(treat.invite) %>%
  summarize(across(c(head.edu, mosques, pct.poor, total.budget), mean))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 5
##   treat.invite head.edu mosques pct.poor total.budget
##          <dbl>    <dbl>   <dbl>    <dbl>        <dbl>
## 1            0     11.6    1.47    0.400         83.4
## 2            1     11.5    1.42    0.411         83.2
```

```
# comparing the SD
x %>%
  group_by(treat.invite) %>%
  summarize(across(c(head.edu, mosques, pct.poor, total.budget), sd))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

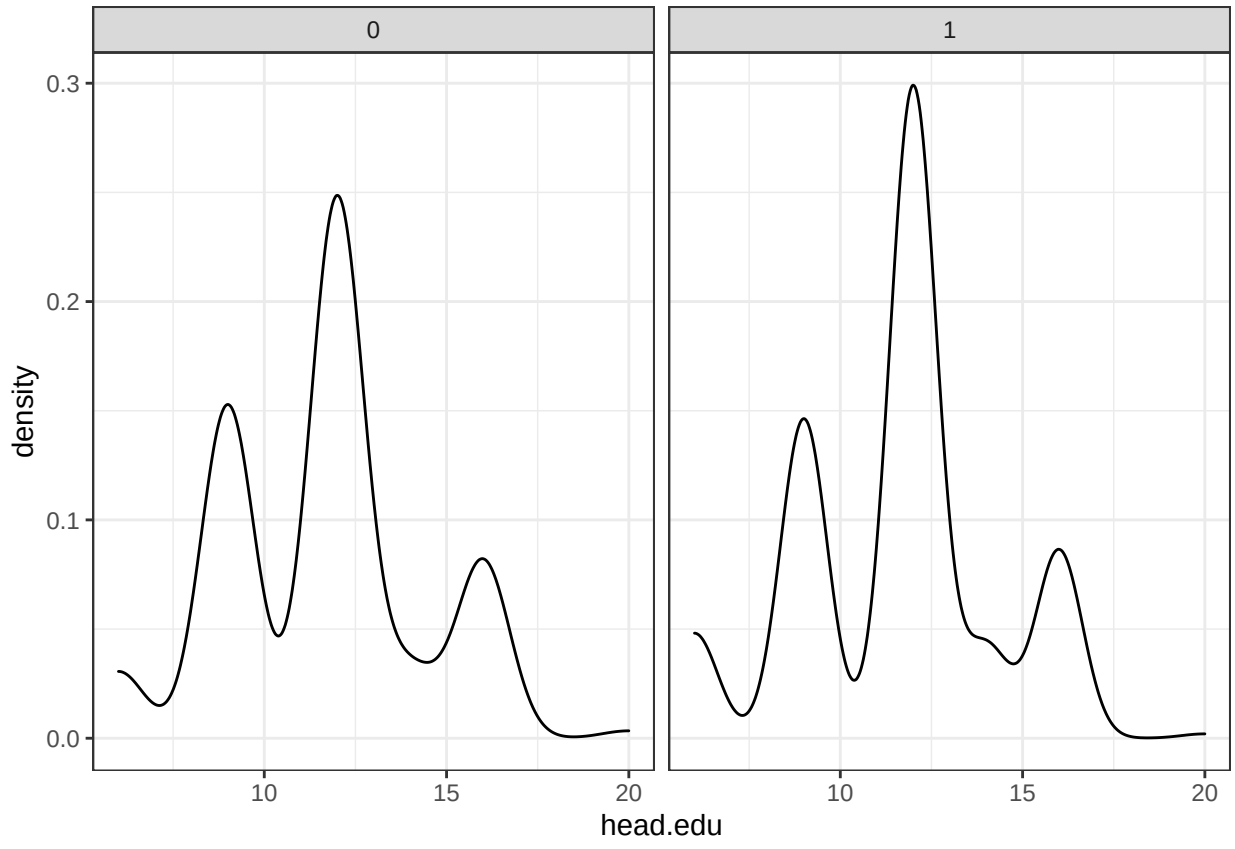
```
## # A tibble: 2 x 5
##   treat.invite head.edu mosques pct.poor total.budget
##          <dbl>    <dbl>   <dbl>    <dbl>        <dbl>
## 1            0     2.72   0.826    0.212         42.9
## 2            1     2.72   0.838    0.213         60.4
```

```r
# using regression to test the null hypothesis
fit1 <- lm(treat.invite ~ # check with others
             head.edu +
             mosques +
             pct.poor +
             total.budget,
           data = x)
summary(fit1)$coef
```

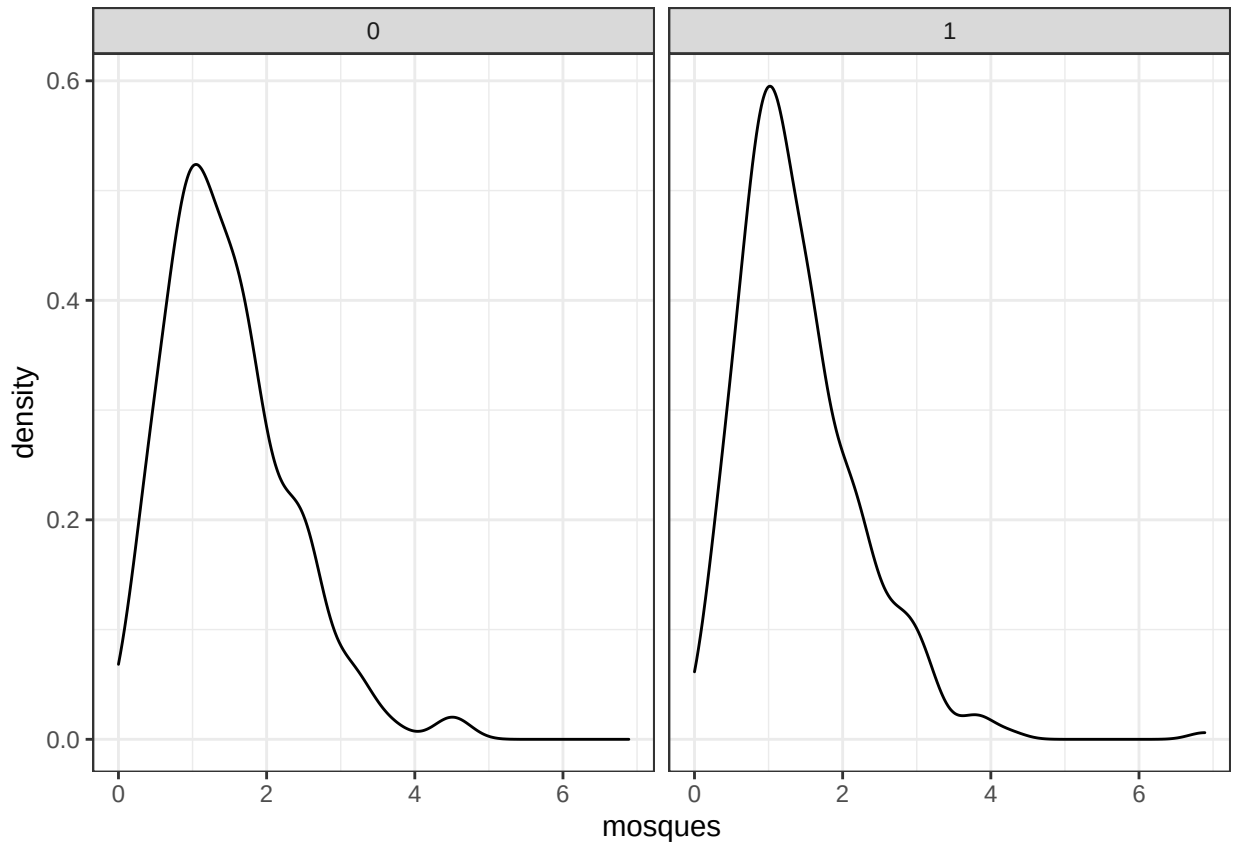
```
##                   Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)   6.766925e-01 0.1172396496  5.7718742 1.430358e-08
## head.edu     -8.971669e-04 0.0080993073 -0.1107708 9.118457e-01
## mosques      -1.870989e-02 0.0264620311 -0.7070467 4.798901e-01
## pct.poor      5.744237e-02 0.1043511204  0.5504720 5.822589e-01
## total.budget -4.698076e-05 0.0004012573 -0.1170839 9.068439e-01
```

Problem 3B

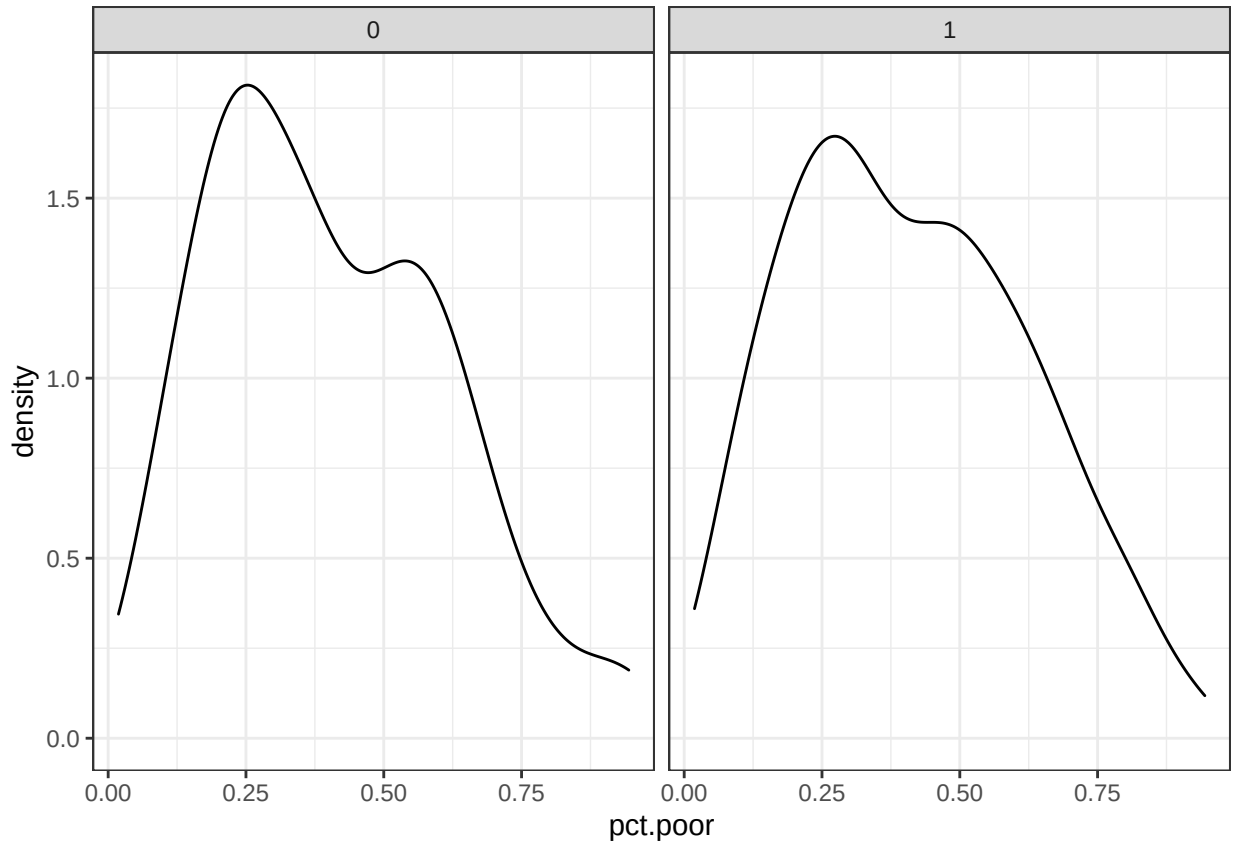
```r
# head.edu
ggplot(x, aes(x = head.edu)) +
  geom_density() +
  facet_grid(~treat.invite) +
  theme_bw()
```



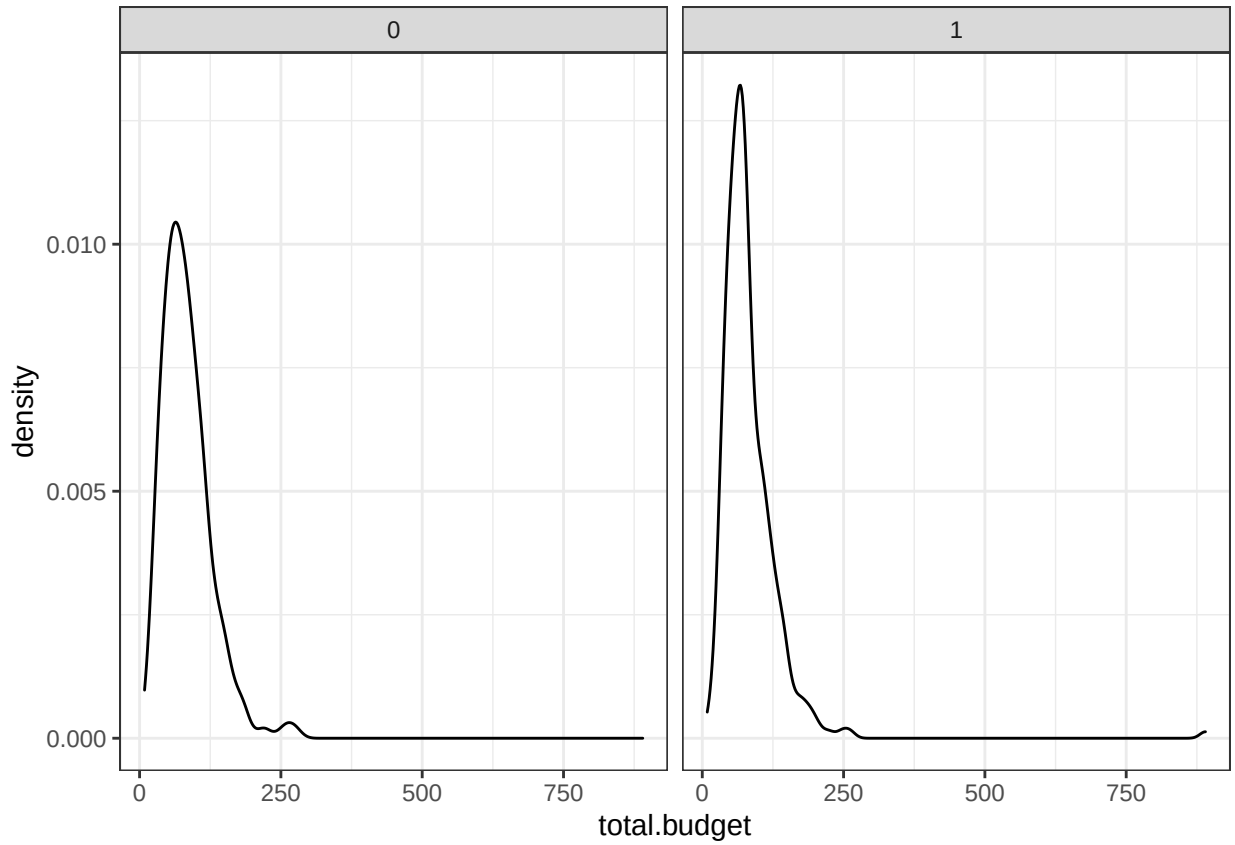
```
# mosques
ggplot(x, aes(x = mosques)) +
  geom_density() +
  facet_grid(~treat.invite) +
  theme_bw()
```




```
# pct.poor
ggplot(x, aes(x = pct.poor)) +
  geom_density() +
  facet_grid(~treat.invite) +
  theme_bw()
```




```
# total.budget
ggplot(x, aes(x = total.budget)) +
  geom_density() +
  facet_grid(~treat.invite) +
  theme_bw()
```



Problem 3C

Based on the table in 3A and plots in 3B, the pre-treatment covariates are balanced between the treated and untreated. First, the sample is sufficiently large with 472 observations. Second, the means and the std. error are similar (with the exception of std. error for age). Third, the test results for each of the pre-treatment covariate regressions fail to reject the null hypotheses.
If the pre-treatment covariates are not balanced, the observed change cannot be attributed to the treatment. The difference in pre-treatment covariates might be the cause of the observed difference.

Problem 3D

```
reg3d <-lm(treat.invite ~
             head.edu +
             mosques +
             pct.poor +
             total.budget,
           data = x)
summary(reg3d)
```

```
##
## Call:
## lm(formula = treat.invite ~ head.edu + mosques + pct.poor + total.budget,
##     data = x)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7075 -0.6474  0.3263  0.3449  0.4480
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.767e-01  1.172e-01   5.772 1.43e-08 ***
## head.edu     -8.972e-04  8.099e-03  -0.111    0.912
## mosques      -1.871e-02  2.646e-02  -0.707    0.480
## pct.poor      5.744e-02  1.044e-01   0.550    0.582
## total.budget -4.698e-05  4.013e-04  -0.117    0.907
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4762 on 467 degrees of freedom
## Multiple R-squared:  0.00163,    Adjusted R-squared:  -0.006921
## F-statistic: 0.1906 on 4 and 467 DF,  p-value: 0.9433
```

The p-value of the omnibus F test is 0.9433. As the p-value is larger than the conventionally accepted significance level (0.05) used to decide whether a test fails to reject the null hypothesis. Therefore, from the problem 3d, all of the pre-treatment covariates are not statistically significant. One can conclude that the randomization has been successful.

Problem 3E

```
xt <- as_tibble(filter(x, treat.invite == 1))
xnt <- as_tibble(filter(x, treat.invite == 0))
ATE <- mean(xt$pct.missing) - mean(xnt$pct.missing)
SE <- sqrt(var(xt$pct.missing)/311 + var(xnt$pct.missing)/161)
ATE
```

```
## [1] -0.02494953
```

```
SE
```

```
## [1] 0.03310019
```

Problem 3F

```
reg3f <- lm(pct.missing ~ treat.invite, x)
summary(reg3f)
```

```
##
## Call:
## lm(formula = pct.missing ~ treat.invite, data = x)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.33064 -0.21249 -0.01284  0.18281  1.42154
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.25277    0.02716   9.306   <2e-16 ***
## treat.invite -0.02495    0.03346  -0.746    0.456
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3447 on 470 degrees of freedom
## Multiple R-squared:  0.001181,   Adjusted R-squared:  -0.0009438
## F-statistic: 0.5559 on 1 and 470 DF,  p-value: 0.4563
```

Yes, while the ATE is similar, the standard error is different. By dividing the group into two, we have created clusters: treated and untreated. But the standard error calucated isn't based on clustering. If we use cluster robust standard error, we get the same standard error.

```
coeftest(reg3f, vcovHC, type = "HC2") # cluster robust standard error
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.252767   0.026549  9.5209   <2e-16 ***
## treat.invite -0.024950   0.033100 -0.7538   0.4514
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problem 3G

```
reg3g <- lm(pct.missing ~
              treat.invite +
              head.edu +
              mosques +
              pct.poor +
              total.budget,
            data = x)
summary(reg3g)
```

```
##
## Call:
## lm(formula = pct.missing ~ treat.invite + head.edu + mosques +
##     pct.poor + total.budget, data = x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28605 -0.21411 -0.01291  0.18932  1.42530
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   0.3904455  0.0869471   4.491 8.96e-06 ***
## treat.invite -0.0264183  0.0331558  -0.797   0.4260
## head.edu     -0.0055082  0.0058032  -0.949   0.3430
## mosques      -0.0481914  0.0189702  -2.540   0.0114 *
## pct.poor     -0.1177125  0.0747921  -1.574   0.1162
## total.budget  0.0005307  0.0002875   1.846   0.0655 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3412 on 466 degrees of freedom
## Multiple R-squared:  0.0294, Adjusted R-squared:  0.01898
## F-statistic: 2.823 on 5 and 466 DF,  p-value: 0.01594
```

Yes. The coefficient is different because we included the covariates. The OLS in this case also considers the pre-treatment covariates when determining the slope.

Problem 3H

```
xpl <- as_tibble(filter(x, pct.poor >= 0.5))
reg3hl <- lm(pct.missing ~
              treat.invite,
            data = xpl)
summary(reg3hl)
```

```
##
## Call:
## lm(formula = pct.missing ~ treat.invite, data = xpl)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28652 -0.18691  0.02655  0.16013  1.35274
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.25696    0.04600   5.586 9.69e-08 ***
## treat.invite -0.07325    0.05651  -1.296    0.197
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3411 on 161 degrees of freedom
## Multiple R-squared:  0.01033,    Adjusted R-squared:  0.004181
## F-statistic:  1.68 on 1 and 161 DF,  p-value: 0.1968
```

```
xph <- as_tibble(filter(x, pct.poor < 0.5))
reg3hh <- lm(pct.missing ~
               treat.invite,
             data = xph)
summary(reg3hh)
```

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

```
## lm(formula = pct.missing ~ treat.invite, data = xph)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23283 -0.22572 -0.02433  0.18494  1.42371
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.250594   0.033616   7.455 9.26e-13 ***
## treat.invite 0.000692   0.041474   0.017    0.987
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3461 on 307 degrees of freedom
## Multiple R-squared:  9.068e-07,  Adjusted R-squared:  -0.003256
## F-statistic: 0.0002784 on 1 and 307 DF,  p-value: 0.9867
```

```
# Making dummy variable
xh <- mutate(x, poor.h = ifelse(pct.poor >= 0.5, 1 , 0))
reg3h <- lm(pct.missing ~ treat.invite + poor.h + treat.invite*poor.h, xh)
summary(reg3h)
```

```
##
## Call:
## lm(formula = pct.missing ~ treat.invite + poor.h + treat.invite *
##     poor.h, data = xh)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28652 -0.21462 -0.02035  0.17527  1.42371
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.250594   0.033451   7.491 3.43e-13 ***
## treat.invite         0.000692   0.041270   0.017    0.987
## poor.h               0.006363   0.057232   0.111    0.912
## treat.invite:poor.h -0.073942   0.070413  -1.050    0.294
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3444 on 468 degrees of freedom
## Multiple R-squared:  0.006967,   Adjusted R-squared:  0.0006015
## F-statistic: 1.094 on 3 and 468 DF,  p-value: 0.3511
```

```
# for the SE
SE3H <- sqrt(diag(vcovHC(reg3h, type = "HC2")))
SE3H
```

```
##         (Intercept)        treat.invite              poor.h treat.invite:poor.h
##          0.03454477          0.04202629          0.05323696          0.06787224
```

```r
#Testing the Null
n=100
estimated_1 <-
  replicate(10000,
            mean(xt[xt$pct.poor>=0.5,]$pct.missing[sample(1:nrow(xt[xt$pct.poor>=0.5,]),n/2)])-mean(xnt

estimated_2 <-
  replicate(10000,
            mean(xt[xt$pct.poor<0.5,]$pct.missing[sample(1:nrow(xt[xt$pct.poor<0.5,]),n/2)])-mean(xnt[x

t.test(estimated_1,estimated_2)
```

```
##
##  Welch Two Sample t-test
##
## data:  estimated_1 and estimated_2
## t = -108.82, df = 18256, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.07540520 -0.07273671
## sample estimates:
##     mean of x     mean of y
## -0.0737154372  0.0003555176
```

There is a difference. For villages with more than half of households below the poverty line will see a 7% increase in corruption if community monitoring is increased. But for villages with less than half of households below the poverty line will see a 0.1% decrease in corruption if community monitoring is increased.

Problem 3I

SUTVA (Stable Unit Treatment Value Assumption) implicitly assumes that potential outcomes for a unit are not affected by treatment assignment for another unit. If community monitoring is seen as effective, people from one village could have told people in a different village of the effect. SUTVA could have been better maintained if the treatment was audits by professional engineers. However, since people can organize oversight by themselves to some degree (not sure about the political climate of Indonesia at the time of the experiment), it is possible but the SUTVA was violated. If so, the results proliferate, and causal inference becomes "exponentially" difficult. In this case, we cannot argue for a causal effect.

Problem 4

Each student has ${}_{10}P_4$ potential outcomes because only 4 treatments are randomly given. Since there are 10 students, the total potential outcome is ${}_{10}P_4 * 10$ which is 50400.