Final Project

Colden Johnson

2022-12-11

Part A - Short Answer Questions

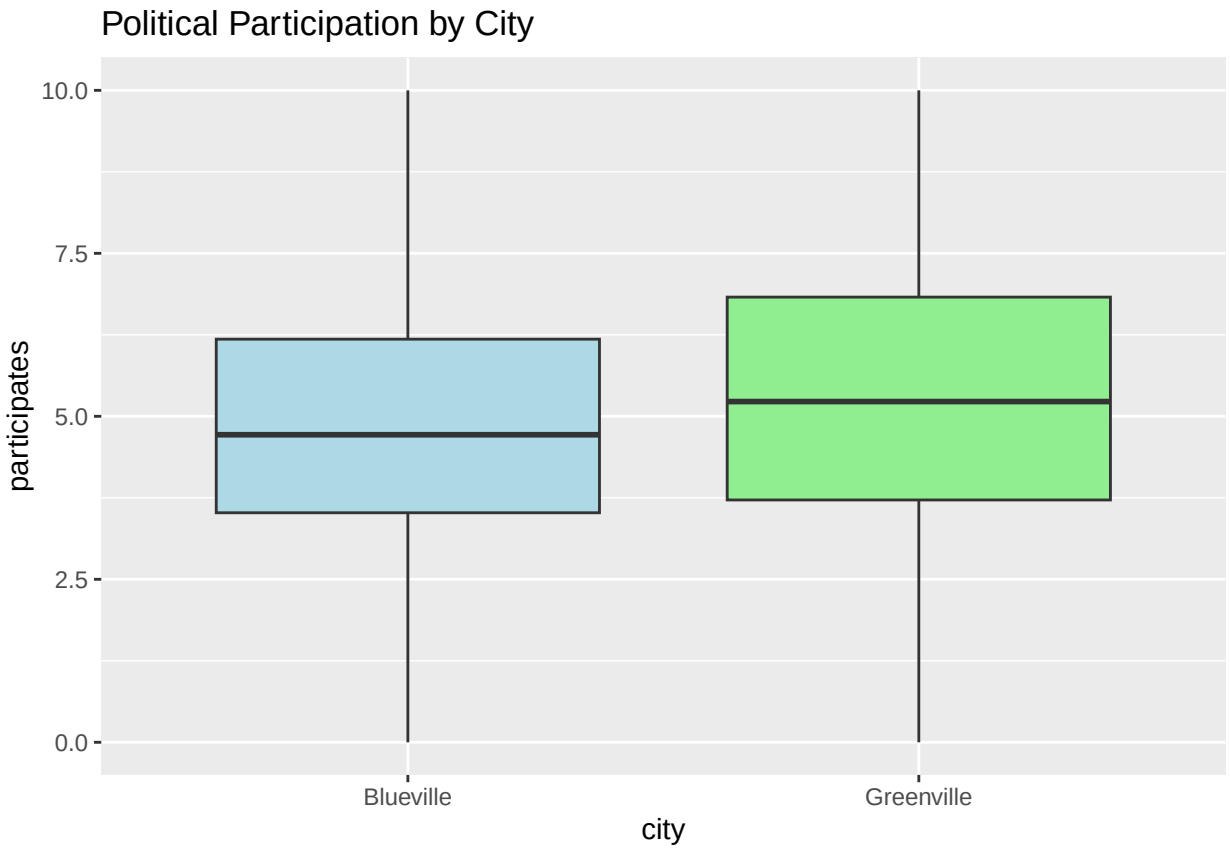
Question 1

```
grouped_polypar <- polypar %>%
  group_by(district) %>%
  summarize(mean_participation = mean(participates))
grouped_polypar <- grouped_polypar[order(grouped_polypar$mean_participation, decreasing = TRUE),]
grouped_polypar[1:10,]
```

```
## # A tibble: 10 x 2
##    district mean_participation
##       <dbl>              <dbl>
##  1       47               6.70
##  2       12               6.59
##  3       52               6.22
##  4       53               6.15
##  5       58               6.10
##  6       38               6.09
##  7       63               6.03
##  8       74               5.96
##  9       54               5.95
## 10       70               5.89
```

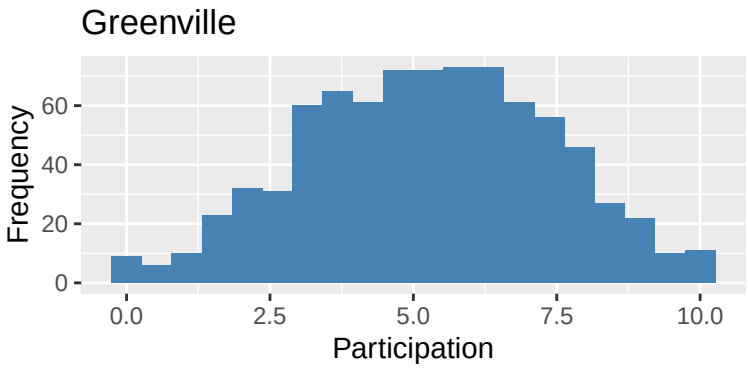
Question 2

```
ggplot(polypar, aes(city, participates)) +
  geom_boxplot(fill = c('light blue', 'light green')) +
  ggtitle('Political Participation by City')
```

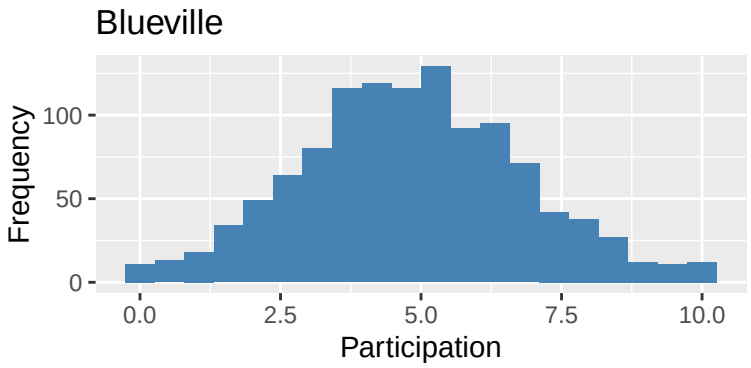



```
greenville <- polypar %>%
  filter(city == 'Greenville')
blueville <- polypar %>%
  filter(city == 'Blueville')

# before conducting a t-test, we need to a) check for normality and b) check for normal variance (we ca
ggplot(greenville, aes(x=participates)) +
  geom_histogram(bins = 20, fill = 'steelblue') +
  xlab('Participation') +
  ylab('Frequency') +
  ggtitle('Greenville')
```



```
ggplot(blueville, aes(x=participates)) +
  geom_histogram(bins = 20, fill = 'steelblue') +
  xlab('Participation') +
  ylab('Frequency') +
  ggtitle('Blueville')
```




```
var.test(blueville$participates, greenville$participates)
```

```
##
##  F test to compare two variances
##
## data:  blueville$participates and greenville$participates
## F = 0.87348, num df = 1148, denom df = 819, p-value = 0.03561
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7688901 0.9909493
## sample estimates:
## ratio of variances
##          0.8734822
```

```
t.test(greenville$participates, blueville$participates, method = 'two-side')
```

```
##
##  Welch Two Sample t-test
##
## data:  greenville$participates and blueville$participates
## t = 4.4861, df = 1689.9, p-value = 7.743e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2395637 0.6117814
## sample estimates:
## mean of x mean of y
##  5.255699  4.830026
```

Because the values of participates are bounded by 0 and 10, the tails of both boxplots end at these values. It appears that Greenville's mean and middle 2 quartiles (middle 50%) are both shifted somewhat higher than blueville's, and that in general greenville has a higher participates score. We can run a t-test to determine if this is a statistically significant difference.

Given the low p-value of the t-test, the null hypothesis is rejected. The 95% confidence interval falls between 0.24 and 0.61. This means we are 95% confident the true difference in means between blueville and greenville participation falls between 0.24 and 0.61.

Question 3

Variable Types

Age - quantitative discrete — discrete values of whole age numbers. However, number values that vary between 18-99. City - nominal categorical — There is no intrinsic order, but two or more distinct categories. City_home - ordinal categorical — because this is a ranked list, there is an intrinsic order (ordinal), and this order is separated into categories. District - nominal categorical (note: this variable could be treated as an identifier, if we grouped by district into categories) – There is no intrinsic order, but data is separated into a number of distinct categories. Educ - nominal categorical – There are two distinct categories, but no order for the categories to be ranked in. This is a boolean value.

Question 4

```
m1 <- lm(political_efficacy ~ political_interest, data = polypar)
m2 <- lm(political_efficacy ~ trust_city_gov * political_interest, data = polypar)
summary(m1)
```

```
##
## Call:
## lm(formula = political_efficacy ~ political_interest, data = polypar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7788 -0.7788  0.2212  1.2212  1.3407
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.83857    0.06703  42.345   <2e-16 ***
## political_interest -0.05975    0.04055  -1.474    0.141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.083 on 1967 degrees of freedom
## Multiple R-squared:  0.001103,   Adjusted R-squared:  0.0005948
## F-statistic: 2.171 on 1 and 1967 DF,  p-value: 0.1408
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = political_efficacy ~ trust_city_gov * political_interest,
##     data = polypar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.0634 -0.7649  0.2351  1.1035  1.4372
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       2.22653    0.18554  12.000  < 2e-16 ***
## trust_city_gov                    0.26724    0.07514   3.557 0.000384 ***
## political_interest                0.16945    0.11087   1.528 0.126564
## trust_city_gov:political_interest -0.10038    0.04453  -2.254 0.024301 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.078 on 1965 degrees of freedom
## Multiple R-squared:  0.01169,    Adjusted R-squared:  0.01018
## F-statistic: 7.747 on 3 and 1965 DF,  p-value: 3.83e-05
```

In the first model, the intercept of political_efficacy is 2.84. Every unit increase in political_interest is associated with a decrease of 0.06 in political_efficacy. This is semi-counterintuitive, but this doesn't mean much because a) the affect size is quite small, and b) the p-value is not significant.

In the second model, the intercept or predicted value when all 3 other values are 0 is predicted to be 2.23. As trust_city_gov increases by 1 unit, political efficacy is predicted to increase by 0.27. Every unit increase in political_interest is associated with a 0.17 increase in political_efficacy. The interaction between trust_city_gov:political_interest can be interpreted by as trust_city_gov increases by 1 unit the slope between trust_city_gov and political_interest decreases by 0.10. Political interest and trust_city_gov both appear to positively predict political efficacy, but when taken together have a lesser impact (according to this model). We can't say with confidence whether Eqvotey is right, given that the R squared value for both regressions is abysmally low (~1% variance explained).

Question 5

```r
set.seed(54321)
pconfint = function(phat,n,conf=0.95) {
  se = sqrt(phat*(1-phat)/n)
  al2 = 1-(1-conf)/2
  zstar = qnorm(al2)
  upperlimit = phat+zstar*se
  lowerlimit = phat - zstar*se
  return(c(round(lowerlimit,3), round(upperlimit, 3)))
}
phatPrpnReturn = mean(polypar$pol_active_village, na.rm = T)
pconfint(phat = phatPrpnReturn, n = 1969, conf = 0.95)
```

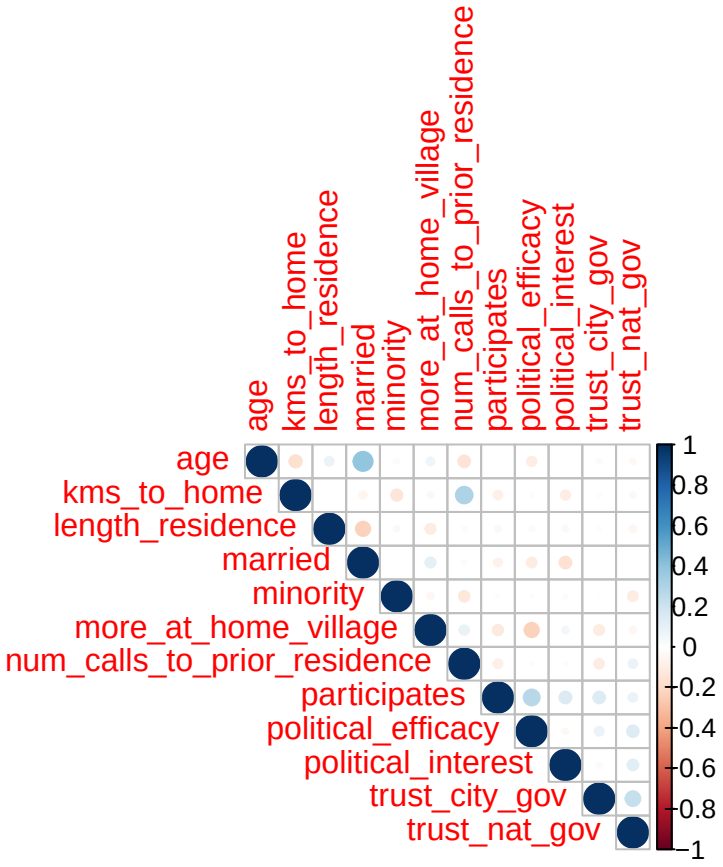
```
## [1] 0.147 0.179
```

The 95% confidence interval is between 0.147 and 0.179. The true value (0.20) given to us by divine intervention does NOT fall inside this 95% confidence interval.

Part B - Long Answer Question

Figure 2

```
# I'm going to start out by making a small corrplot. I recognize this is not necessary for analysis, bu
corrplotdata <- polypar[c(1,10:14,15,18,21:22, 25:26)]
corrplotdata <- drop_na(corrplotdata)
corrplotdata.cor = cor(corrplotdata)
corrplot(corrplotdata.cor, type = "upper")
```




```
# secondary corrplot?
```

Looking at this corrplot, we are trying to generate answers from some relatively weak correlations. This means, making sure our analysis is precise is very important, as small things (like not meeting normality assumptions, etc.) could throw off our recommendations.

Hypothesis 1 and Mtable Creation

Figure 1 / Main Model

```
# Base model
model1.1 <- lm(participates ~ more_at_home_village, data = polypar)
model1.2 <-  lm(participates ~ more_at_home_village + owns_home, data = polypar)
model1.3 <-  lm(participates ~ more_at_home_village + owns_home + log(num_calls_to_prior_residence+1), d
```

```r
model1.4 <-  lm(participates ~ more_at_home_village + owns_home + log(num_calls_to_prior_residence+1) +
model1.5 <- lm(participates ~ more_at_home_village + owns_home + log(kms_to_home + 1) + political_effica
model1.6 <- lm(participates ~ more_at_home_village + owns_home + log(kms_to_home + 1) + political_effica
model1.7 <- lm(participates ~ more_at_home_village + owns_home + log(kms_to_home + 1) + political_effica

# I am removing log(number_of_calls_to_home) due to collinearity. I chose this one due to the lack of n

# create mtable, name...
socialandEconMtable <- mtable("Model 1"=model1.1,"Model 2"=model1.2,"Model 3"=model1.3,"Model 4"=model1
                    summary.stats=c("sigma","R-squared","F","p","N"))
socialandEconMtable
```

```
##
## Calls:
## Model 1: lm(formula = participates ~ more_at_home_village, data = polypar)
## Model 2: lm(formula = participates ~ more_at_home_village + owns_home,
##     data = polypar)
## Model 3: lm(formula = participates ~ more_at_home_village + owns_home +
##     log(num_calls_to_prior_residence + 1), data = polypar)
## Model 4: lm(formula = participates ~ more_at_home_village + owns_home +
##     log(num_calls_to_prior_residence + 1) + log(kms_to_home +
##     1), data = polypar)
## Model 5: lm(formula = participates ~ more_at_home_village + owns_home +
##     log(kms_to_home + 1) + political_efficacy, data = polypar)
## reghelp experiment: lm(formula = participates ~ more_at_home_village + owns_home +
##     log(kms_to_home + 1) + political_efficacy + reghelp, data = polypar)
## encouraged experiment: lm(formula = participates ~ more_at_home_village + owns_home +
##     log(kms_to_home + 1) + political_efficacy + encouraged, data = polypar)
##
## ============================================================================================================
##                                          Model 1      Model 2      Model 3      Model 4      Model
## ------------------------------------------------------------------------------------------------------------
##   (Intercept)                             5.669***     5.470***     5.696***     6.074***     4.445
##                                          (0.135)      (0.153)      (0.163)      (0.216)      (0.25
##   more_at_home_village                   -0.234***    -0.230***    -0.207***    -0.206***    -0.10
##                                          (0.045)      (0.045)      (0.045)      (0.045)      (0.04
##   owns_home                                            0.273**      0.270**      0.297**      0.24
##                                                       (0.100)      (0.100)      (0.100)      (0.09
##   log(num_calls_to_prior_residence + 1)                             -0.180***    -0.130**
##                                                                     (0.045)      (0.049)
##   log(kms_to_home + 1)                                                           -0.124**     -0.17
##                                                                                  (0.047)      (0.04
##   political_efficacy                                                                           0.494
##                                                                                                (0.04
##   reghelp
##
##   encouraged
##
## ------------------------------------------------------------------------------------------------------------
##   sigma                                   2.049        2.046        2.038        2.035        1.97
##   R-squared                               0.014        0.017        0.025        0.029        0.089
##   F                                      27.266       17.407       16.918       14.483       47.993
##   p                                       0.000        0.000        0.000        0.000        0.000
```

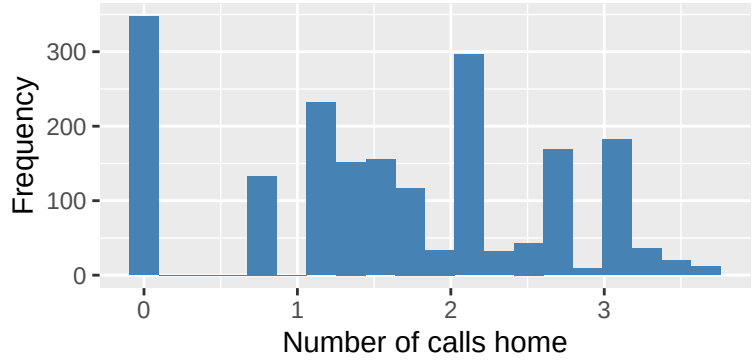
```
##   N                                      1969         1969         1969         1969         1969
## ================================================================================================
##   Significance: *** = p < 0.001; ** = p < 0.01; * = p < 0.05
```

```
# This is actually good -- reghelp has quite strong predictive power, even when included in the model (
# --- arguably should not be included, given that it is an experimental variable. However, it is ok to
# differently, encouraged does not have any predictive power, and is not shown as significant in the mo
```

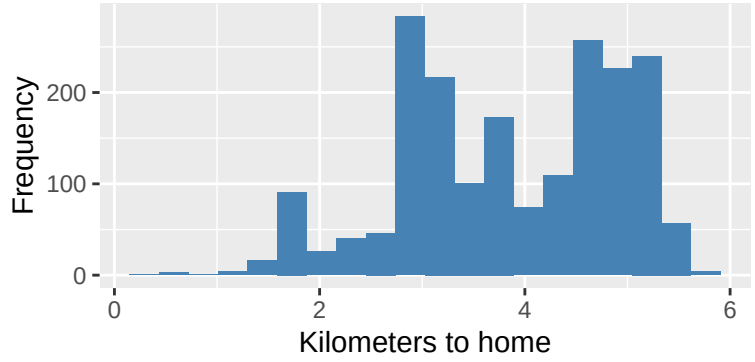
Figure 3.1, 3.2

```
# histograms to check for normal distribution and log transform

ggplot(polypar, aes(x=log(num_calls_to_prior_residence + 1))) +
  geom_histogram(bins = 20, fill = 'steelblue') +
  xlab('Number of calls home') +
  ylab('Frequency')
```




```
ggplot(polypar, aes(x=log(kms_to_home + 1))) +
  geom_histogram(bins = 20, fill = 'steelblue') +
  xlab('Kilometers to home') +
  ylab('Frequency')
```

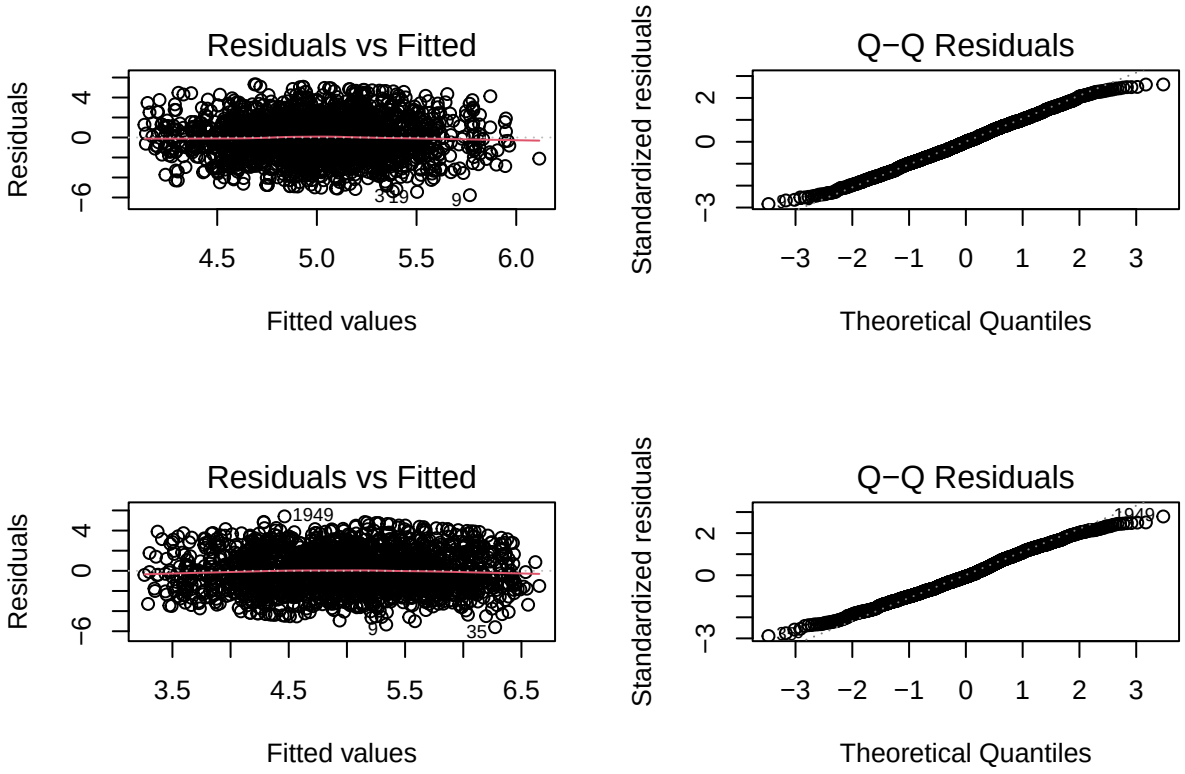



```
# Let me just note that, while log transforming both of these graphs did make the data slightly better,
```

Our regression assumptions are largely met, with the exception of normality for the two variables graphed above. For these variables, taking the log transform helped to a degree, but the values are still quite off what we would like to see in a normal bell curve. This does provide some limitations to the model.

Figure 4

```
# plots for mtable
par(mfrow = c(2, 2))
plot(model1.4, c(1,2))
plot(model1.6, c(1,2))
```



Hypothesis 2

Figure 5, Figure 6

```
# fit model 2
model2.1 <- lm(participates ~ reghelp + female + educ, data = polypar)
summary(model2.1)
```

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

```
## lm(formula = participates ~ reghelp + female + educ, data = polypar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4373 -1.3993 -0.0182  1.3860  5.4610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.74545    0.10970  43.258  < 2e-16 ***
## reghelp      0.62775    0.09200   6.823 1.18e-11 ***
## female      -0.20643    0.09294  -2.221   0.0265 *
## educ         0.06412    0.09832   0.652   0.5144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.038 on 1965 degrees of freedom
## Multiple R-squared:  0.02552,    Adjusted R-squared:  0.02404
## F-statistic: 17.16 on 3 and 1965 DF,  p-value: 5.344e-11
```

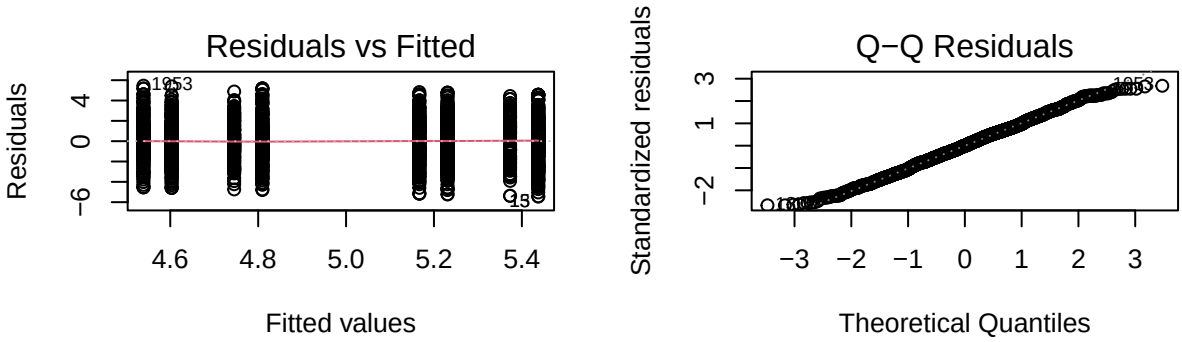
```
# filter
reghelp_true <- polypar %>%
  filter(reghelp == 1)
reghelp_false <- polypar %>%
  filter(reghelp == 0)

# t test for hypothesis 2
t.test(reghelp_true$participates, reghelp_false$participates, method = 'greater')
```

```
##
##  Welch Two Sample t-test
##
## data:  reghelp_true$participates and reghelp_false$participates
## t = 6.7819, df = 1966.9, p-value = 1.564e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.4415428 0.8008025
## sample estimates:
## mean of x mean of y
##  5.301954  4.680782
```

Figure 7

```
# plots for model 2
par(mfrow = c(2, 2))
plot(model2.1, c(1,2))
```



Hypothesis 3

Figure 8, Figure 9

```
# fit model 3
model3.1 <- lm(participates ~ encouraged + female + educ, data = polypar)
summary(model3.1)
```

```
##
## Call:
## lm(formula = participates ~ encouraged + female + educ, data = polypar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1625 -1.3914 -0.0541  1.4233  5.1482
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.10734    0.11243  45.425   <2e-16 ***
## encouraged  -0.06744    0.09320  -0.724   0.4694
## female      -0.18811    0.09401  -2.001   0.0455 *
## educ         0.05514    0.09956   0.554   0.5797
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.061 on 1965 degrees of freedom
## Multiple R-squared:  0.002701,   Adjusted R-squared:  0.001179
## F-statistic: 1.774 on 3 and 1965 DF,  p-value: 0.15
```

```
encouraged_true <- polypar %>%
  filter(encouraged == 1)
encouraged_false <- polypar %>%
  filter(encouraged == 0)

# t test for hypothesis 3
t.test(encouraged_true$participates, encouraged_false$participates, method = 'greater')
```

```
##
##  Welch Two Sample t-test
##
## data:  encouraged_true$participates and encouraged_false$participates
## t = -0.79902, df = 1930.5, p-value = 0.4244
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2573451  0.1083535
## sample estimates:
## mean of x mean of y
##  4.972303  5.046799
```

Figure 10

```
# plots for model 3
par(mfrow = c(2, 2))
plot(model3.1, c(1,2))
```

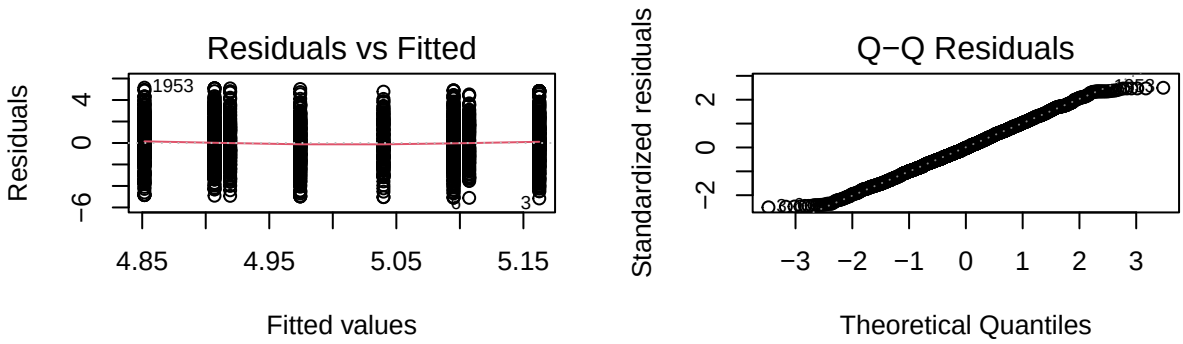


Figure 11

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:memisc':
##
##     recode
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```r
vif(model1.5)
```

```
## more_at_home_village            owns_home log(kms_to_home + 1)
##             1.054438             1.012375             1.011316
##   political_efficacy
##             1.054421
```

```r
# These are very satisfyingly low vif values
```