Final Project Written Analysis

Contextualization

For this analysis, I will analyze the 3 hypothesis proposed by Eqvotey individually, and construct a single integrated model with all relevant variables visualized. While doing this, I run regression analysis for all 3, as well as t-tests for both experiments.

Hypothesis 1: *Some migrants feel more socially, emotionally, and/or economically attached to the communities they grew up in than the ones where they currently reside.*

To test this hypothesis we need to create a theoretical framework to aid in variable selection. This framework includes all variables that could possibly be included under the explanatory umbrella of social, emotional, and economic attachment. Because this model will be the main model created, there are a large number of variables from the dataset to consider. These include: *pol_active_village, more_at_home_village, num_calls_to_prior_residence, owns_home,* and *kms_to_home*. The variable *political_efficacy* can also be included in the framework as having an impact on emotional connectedness. While a bit of a stretch, I want to justify this considering the high correlation/R squared value between *political_efficacy* and *participation*, as seen in the exploratory corrplot (Figure 2).

Using this framework, variable selection for creating a model becomes much easier. While it will be important to check linear regression assumptions for all included data, while building the model we focus on excluding variables without significant p-values. Excluding variables without significant p-values yields the first 5 lines of the mtable (Figure 1). Note that these selected variables have a theoretical framework behind them, and were not added simply due to correlation with participation.

Graphing the distribution of number_of_calls_to_home and kms_to_home (Figure 3), it is clear that both variables are severely right skewed, and also not very normal. To address this to an extent, before including either of these in the model I choose to log transform them both. I will point out later that these variables are unacceptably collinear and exclude one of them.

Figure 1 (the main model) builds from a base case single-variable regression to a more complicated and fully specified model. Currently, we are only examining the first 5 columns of the model. Model 1 specifies the intercept at 5.699, indicating a base predicted participation value of ~5.7. As the model progresses from the base to model 5, all signs stay the same, so providing an interpretation of model 5 alone is sufficient. Participants rating that they feel 'more at home in their village' 1 point higher is associated with a 0.108 *decrease* in predicted participation. Owning a home predicts the predicted participation intercept to increase by 0.244, and an increase of 1% in log(kms_to_home + 1) predicts a 0.172 decrease in participation (changes as a percentage because this variable is log transformed). The shift from model 4 to model 5 needs to be explicitly mentioned, because the variable political_efficacy is comparatively strong in the model (large magnitude coefficient, large impact on R squared). Every 1 unit increase in political_efficacy predicts a 0.494 increase in participation.

Additionally, there is a **collinearity** issue that occurred in the model between log(number_of_calls_to_home + 1) and log(kms_to_home +1). These variables were initially flagged for me in the corrplot (fig.2), indicating the potential for issues. One way to identify that they are in fact collinear is by looking at the relative strength for each. Although p-values for both are technically significant, the strength/magnitude of the factor decreases heavily when both are included and strengthens when one is excluded. P-value exhibits the same behavior. I chose to exclude log(number_of_calls_to_home+1) instead of log(kms_to_home+1) because of the complete lack of normal distribution of the former's data. This is evident when comparing histograms for the two charts (Figure 3).

After removing these values, I generated a vif for the specified model (Figure 11). These values are satisfyingly low (close to 1, significantly less than 5), so I am not concerned about further collinearity issues.

To validate the model, we should examine the data to make sure it meets all **regression assumptions**. In Figure 4, it can be seen from the Q-Q plot that the data meets the "straight enough" condition, and the RVF plot shows it is random and displays homoscedasticity. For all regressions we run, it should also be assumed that we meet the independence assumption, as we are drawing data from a random sample.

Hypothesis 2: *Some migrants are unfamiliar with the local voter registration process and may face additional barriers to registration (because of bureaucrats who are skeptical of migrants, procedural idiosyncrasies, having to unregister in their origin community, etc.).*

The reghelp variable should be used to test hypothesis 2. This variable is a Boolean test variable, and because there was an experiment conducted, separates the population into both a treatment and control group.

For this reason, to test hypothesis 2, I believe that a t-test is warranted. However, also fitting a regression gives us the benefit of being able to control for various demographic effects, so I conducted both.

Figure 5 is a t-test comparing the two groups who either received or did not receive reghelp from the independent NGO. For this t-test, the null hypothesis was rejected (as seen by the extremely small p-value), indicating that it is very unlikely the difference in groups is due to random chance. The 95% confidence interval is from 0.44 to 0.80, meaning that we are 95% confident that the true difference in population mean is between 0.44 and 0.80.

Supplementing the t-test, I also created a regression model (Figure 6), controlling for some demographic factors (including gender and minority). I also controlled for education, as those who are less educated likely vote less (for a variety of social and economic reasons, many of which can be captured via collinearity by controlling for education). Figure 6 shows that the p-value significance holds up after controlling for these variables. Figure 7 shows that regression assumptions are met, and the RVF/QQ plots look acceptable. It was also important to demonstrate normal distribution as this is a requirement for an unpaired t-test.

Finally, I added the variable reghelp to Figure 1, looking at how it interacted with the rest of the base model. Controlling for these other variables, reghelp retained a high degree of significance, comparatively significant predictive power, and predicted a large positive (0.691) increase to the intercept.

Hypothesis 3: *Politicians are less likely to target migrant votes when campaigning, either because they dislike them or because they know less about their views and thus feel more unsure about how to get their votes.*

The variable *encouraged* should be used to test hypothesis 3. This variable is once again a Boolean test variable value, yielding both a treatment and control group. I am again going to use a t-test approach, and then fit a regression line to validate the t-test results. I point out the normality of the QQ plot in figure 10, indicating the normality assumption is met before conducting a t-test.

In figure 8, the first indication that something is awry with this t-test is the confidence interval for the difference between groups is between the values -0.25 and 0.11. This is not a very helpful result, as it is impossible to see if the treatment is even having a negative or positive effect. More concretely, though, the p-value is 0.42, unacceptably high. The null hypothesis for this t-test is not rejected, so the treatment does not appear to have any statistically significant impact on the treatment group. The fitted regression model (figure 9) also returns a p-value that is far too high, indicating a lack of predictive significance by the *encouraged* variable.

Looking at the QQ and RVF plots for this model (Figure 10), we also see that there is tapering to the sides, and the normality assumption is not met to a satisfactory extent. Finally, we still add this variable to the Figure 1 mtable and see that the p-values are not given as significant, and the variable should not be included in the final predictive model.

Limitations

Hypothesis 3 appears to be entirely refuted by the given data. However, I don’t think the treatment has a strong enough impact to lead to results. Simply encouraging a politician to behave a certain way does very little to change actual behavior, does nothing to dispel confusion about how to appeal to internal migrants, and has very little ‘substantive’ impact. So, the treatment is small and potentially ineffective enough to not necessarily create an effect, even if this is a real source driving down voter participation.

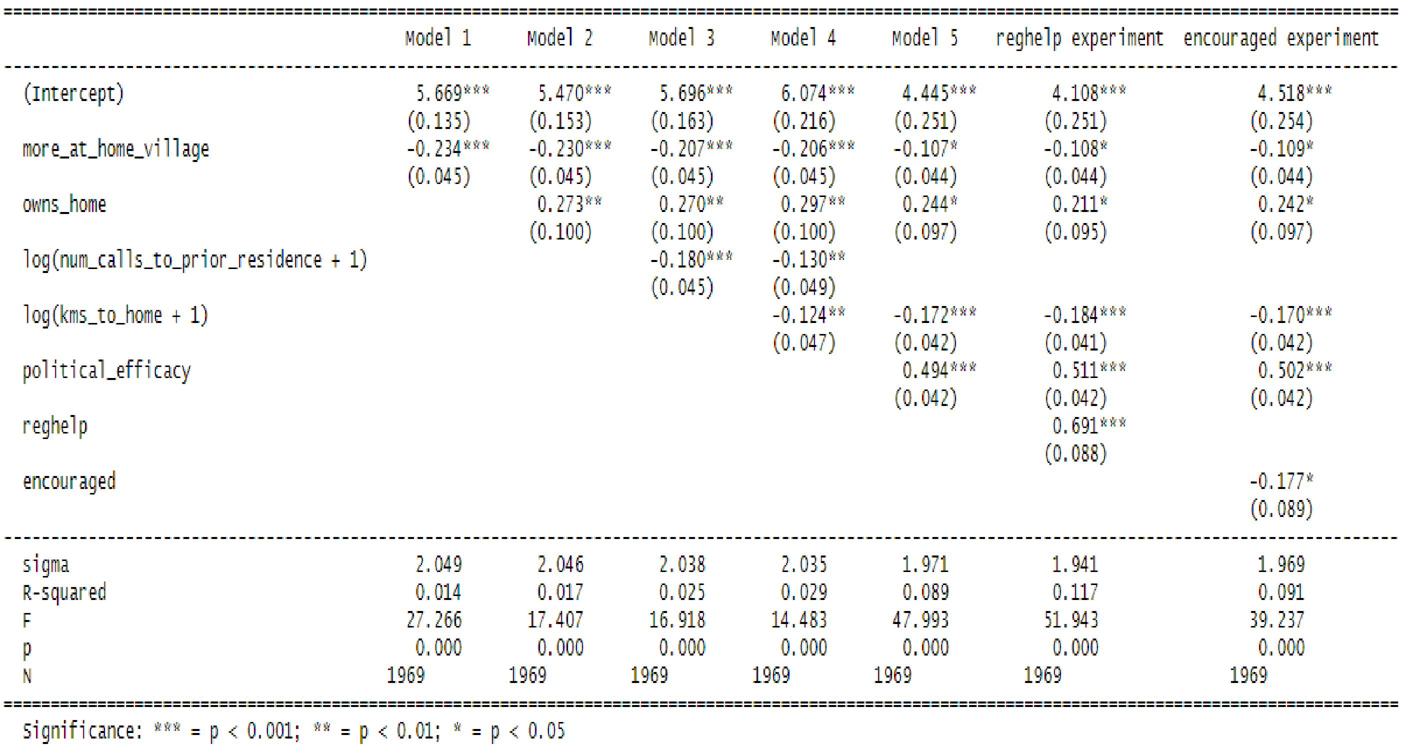
The in general low R squared values of the models can be viewed as a significant issue. Because even the best models I made are not significantly over 0.10 R squared, a large percentage of the variance still remains unexplained.

Some of the data is not entirely normally distributed (specifically, even taking the log of kms_to_home did not fully fix the expression). It is important to be aware of this when judging if regression assumptions were fully met.

Summary Remarks and Future Research

In figure 1, I would propose Model 5 as the final predictive model for hypothesis 1. It has enough statistical significance and a high enough R squared that the hypothesis does appear plausible. Hypothesis 2 is supported by the experiment, and the test for hypothesis 3 does not yield a statistically significant difference between the populations.

To inform future research, if I had to guess 1 variable that would have extremely high predictive power it would be the Boolean value “voted in last election”. Past behavior is an extremely high predictor of future behavior, as habit forming is a powerful force in individuals’ psyches. To determine if lack of voting history is a major contributing factor that could be addressed, we could look at the comparative voting history vs. age of the migrant population compared to the native population. We could then compare relative voting trends across similar age demographics in the population. If the lack of past voting history is a driving factor, we would expect to see similar voting behavior among new voters/those who just turned 18, and for the gap to expand among older migrants when compared to older native voters.



```
==========================================================================================================================
                                     Model 1    Model 2    Model 3    Model 4    Model 5   reghelp experiment  encouraged experiment
--------------------------------------------------------------------------------------------------------------------------
 (Intercept)                          5.669***   5.470***   5.696***   6.074***   4.445***     4.108***            4.518***
                                     (0.135)    (0.153)    (0.163)    (0.216)    (0.251)      (0.251)             (0.254)
 more_at_home_village                -0.234***  -0.230***  -0.207***  -0.206***  -0.107*      -0.108*             -0.109*
                                     (0.045)    (0.045)    (0.045)    (0.045)    (0.044)      (0.044)             (0.044)
 owns_home                                       0.273**    0.270**    0.297**    0.244*       0.211*              0.242*
                                                (0.100)    (0.100)    (0.100)    (0.097)      (0.095)             (0.097)
 log(num_calls_to_prior_residence + 1)                     -0.180***  -0.130**
                                                           (0.045)    (0.049)
 log(kms_to_home + 1)                                                 -0.124**   -0.172***    -0.184***           -0.170***
                                                                      (0.047)    (0.042)      (0.041)             (0.042)
 political_efficacy                                                               0.494***     0.511***            0.502***
                                                                                 (0.042)      (0.042)             (0.042)
 reghelp                                                                                       0.691***
                                                                                              (0.088)
 encouraged                                                                                                       -0.177*
                                                                                                                  (0.089)
--------------------------------------------------------------------------------------------------------------------------
 sigma                                2.049      2.046      2.038      2.035      1.971        1.941               1.969
 R-squared                            0.014      0.017      0.025      0.029      0.089        0.117               0.091
 F                                   27.266     17.407     16.918     14.483     47.993       51.943              39.237
 p                                    0.000      0.000      0.000      0.000      0.000        0.000               0.000
 N                                 1969       1969       1969       1969       1969         1969                1969
==========================================================================================================================
 Significance: *** = p < 0.001; ** = p < 0.01; * = p < 0.05
```

Figure 1

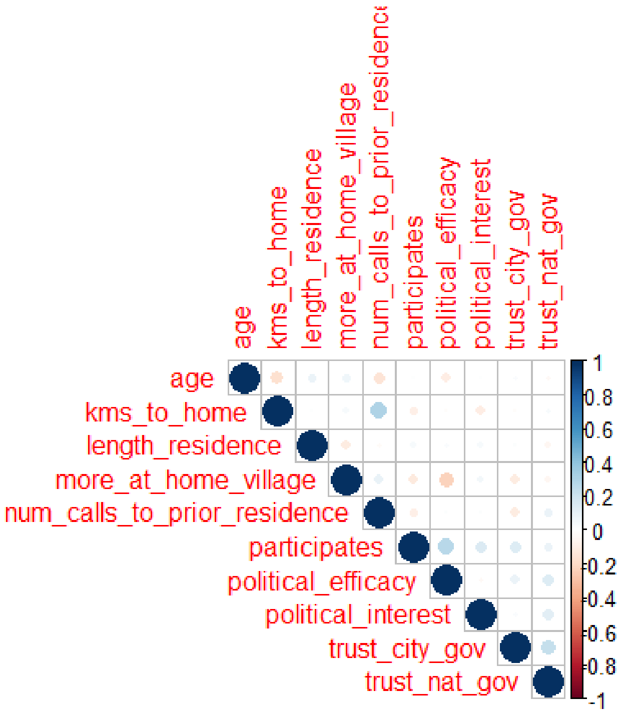


Figure 2

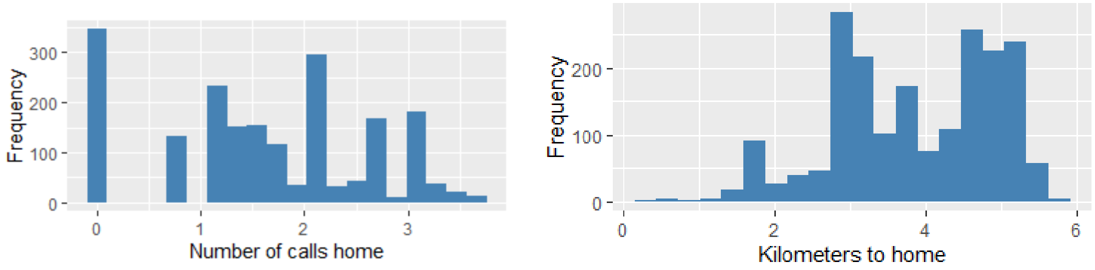


Figure 3

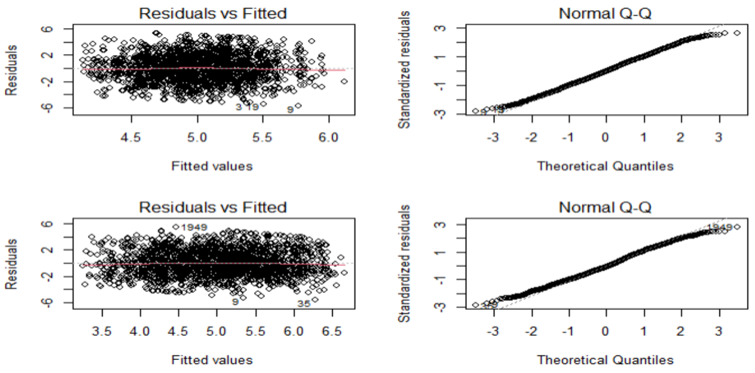
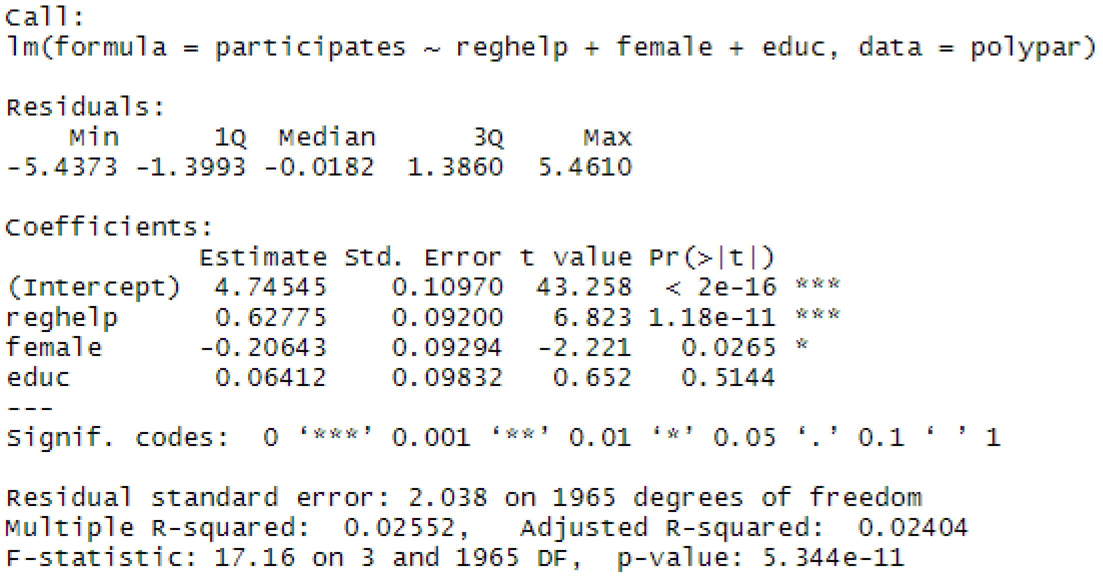


Figure 4



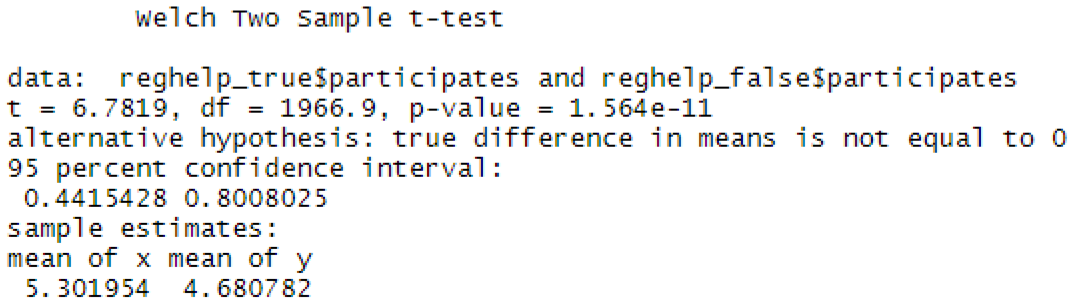
```
Call:
lm(formula = participates ~ reghelp + female + educ, data = polypar)

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
```

Figure 5



```
        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 6

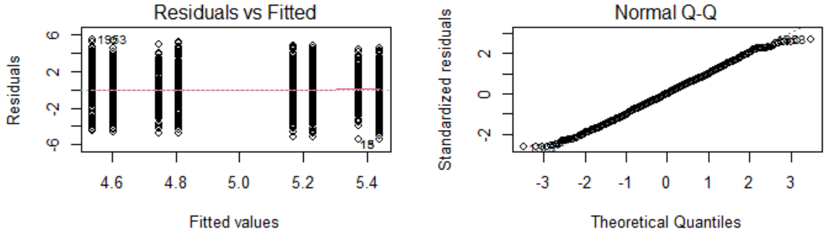
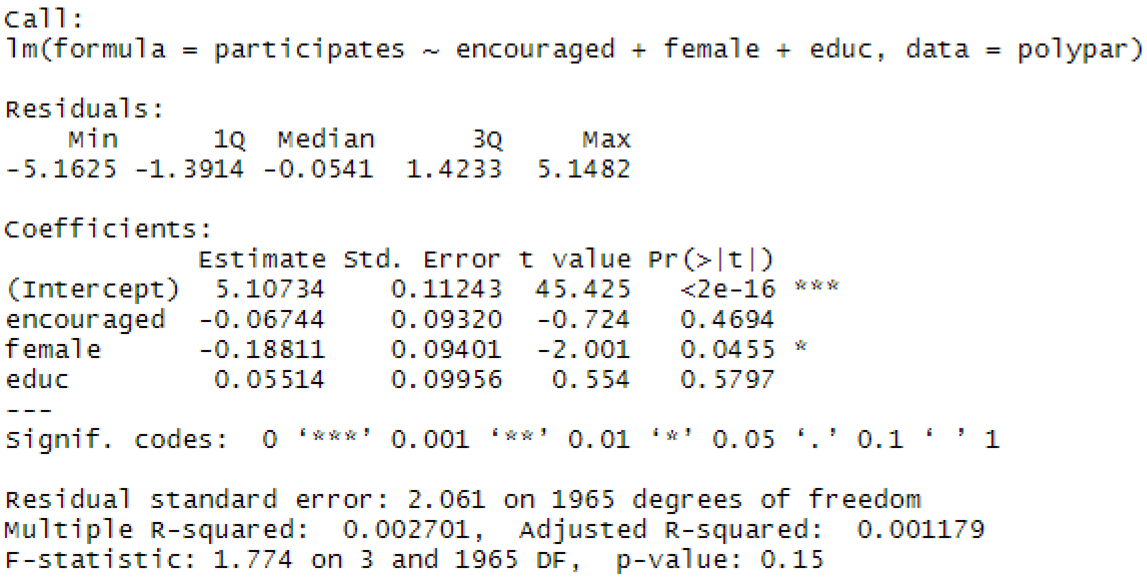


Figure 7



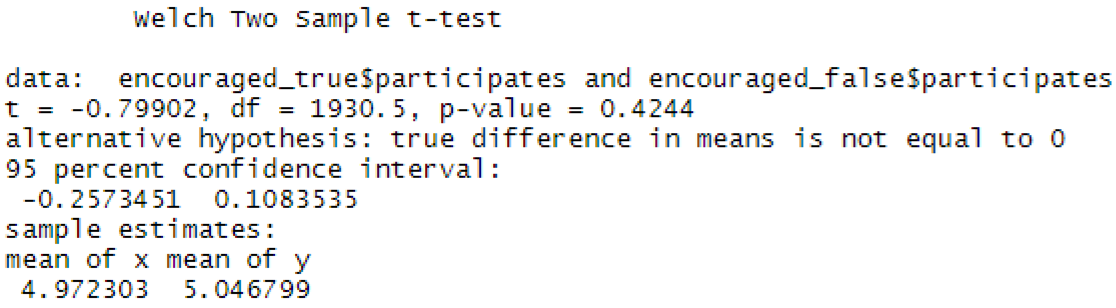
```
Call:
lm(formula = participates ~ encouraged + female + educ, data = polypar)

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
```

Figure 8



```
	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 9

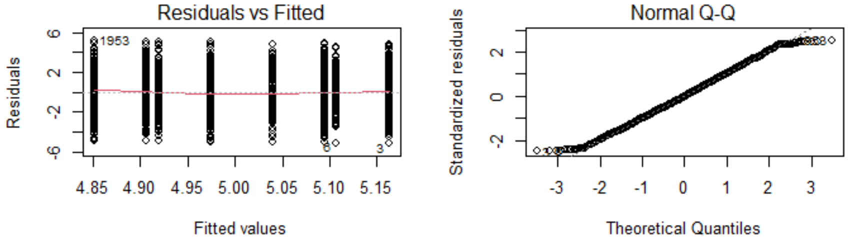
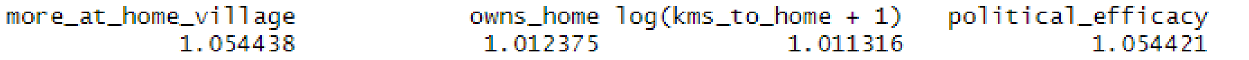


Figure 10



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

Figure 11