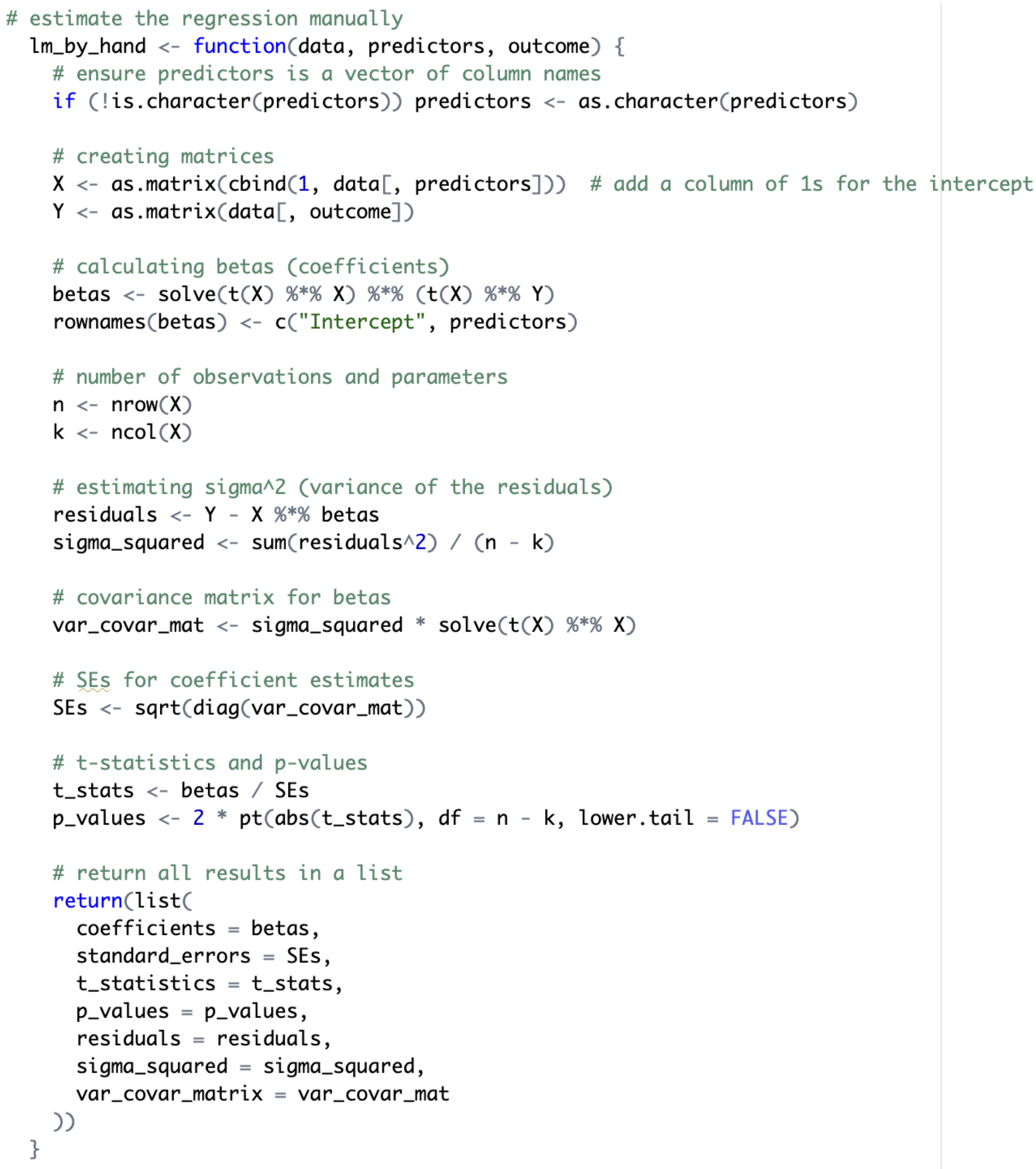
Claire Ott
20302999
Applied Statistics I

Problem Set 3

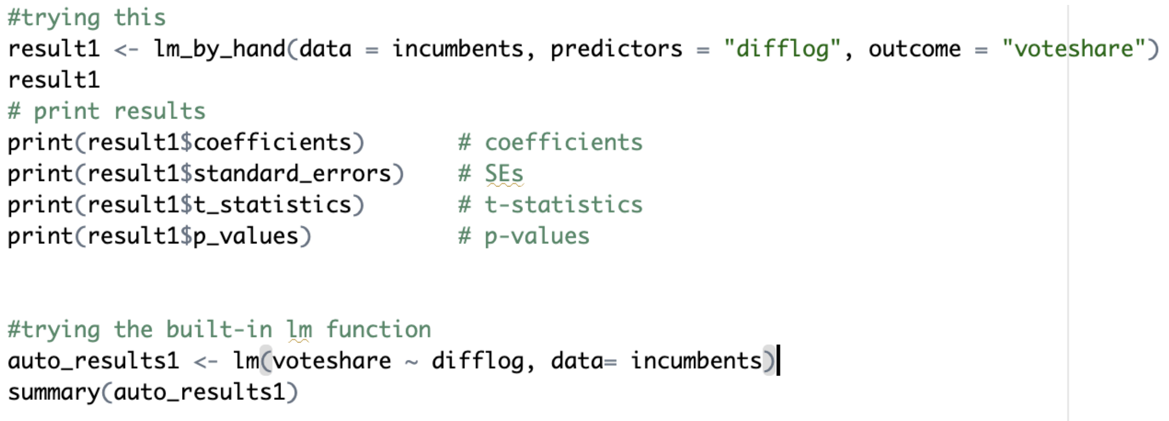
I did try to do my write-up in Latex but it kept crashing on my laptop and I don't think I did it right so I have made a PDF of my answers just in case.

Question 1

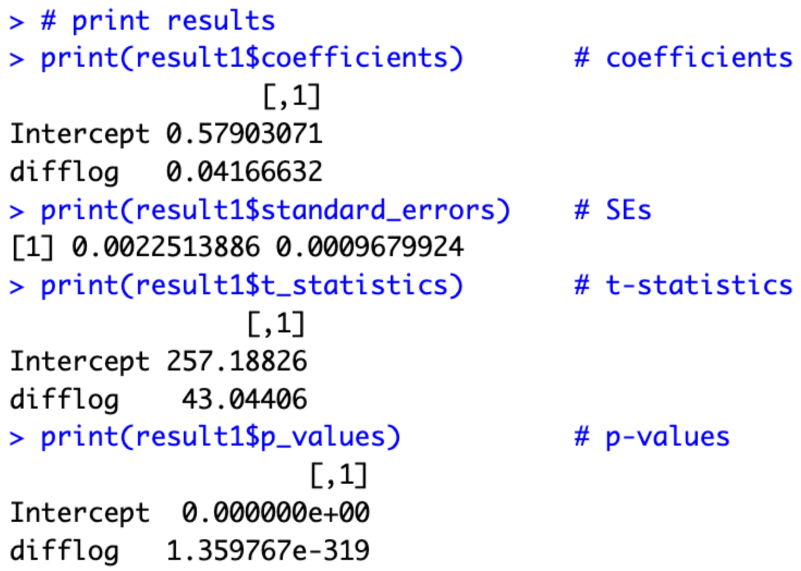
1.



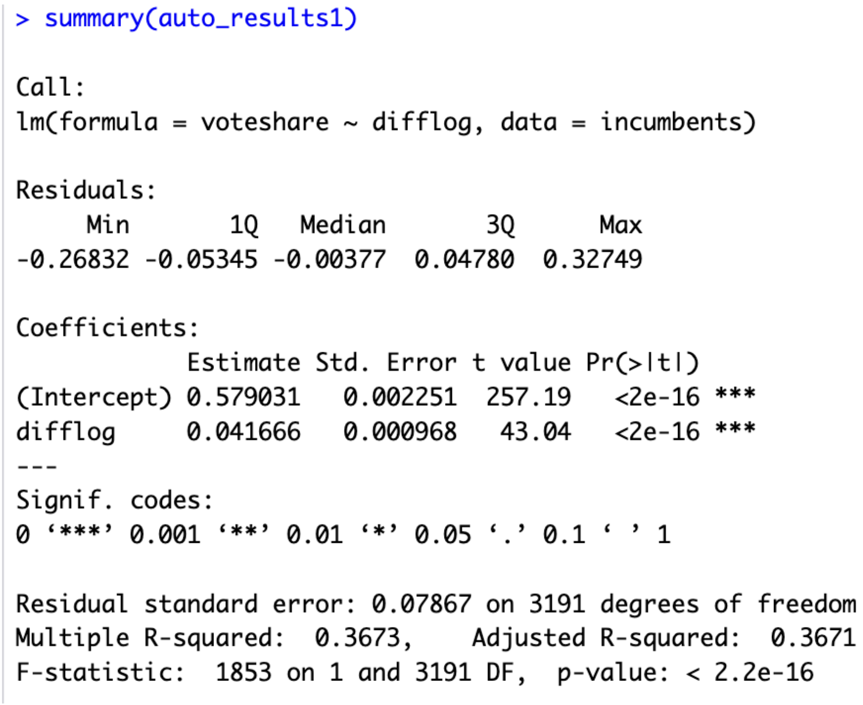
```
# estimate the regression manually
  lm_by_hand <- function(data, predictors, outcome) {
    # ensure predictors is a vector of column names
    if (!is.character(predictors)) predictors <- as.character(predictors)

    # creating matrices
    X <- as.matrix(cbind(1, data[, predictors]))  # add a column of 1s for the intercept
    Y <- as.matrix(data[, outcome])

    # calculating betas (coefficients)
    betas <- solve(t(X) %*% X) %*% (t(X) %*% Y)
    rownames(betas) <- c("Intercept", predictors)

    # number of observations and parameters
    n <- nrow(X)
    k <- ncol(X)

    # estimating sigma^2 (variance of the residuals)
    residuals <- Y - X %*% betas
    sigma_squared <- sum(residuals^2) / (n - k)

    # covariance matrix for betas
    var_covar_mat <- sigma_squared * solve(t(X) %*% X)

    # SEs for coefficient estimates
    SEs <- sqrt(diag(var_covar_mat))

    # t-statistics and p-values
    t_stats <- betas / SEs
    p_values <- 2 * pt(abs(t_stats), df = n - k, lower.tail = FALSE)

    # return all results in a list
    return(list(
      coefficients = betas,
      standard_errors = SEs,
      t_statistics = t_stats,
      p_values = p_values,
      residuals = residuals,
      sigma_squared = sigma_squared,
      var_covar_matrix = var_covar_mat
    ))
  }
```



```
#trying this
result1 <- lm_by_hand(data = incumbents, predictors = "difflog", outcome = "voteshare")
result1
# print results
print(result1$coefficients)       # coefficients
print(result1$standard_errors)    # SEs
print(result1$t_statistics)       # t-statistics
print(result1$p_values)           # p-values


#trying the built-in lm function
auto_results1 <- lm(voteshare ~ difflog, data= incumbents)
summary(auto_results1)
```

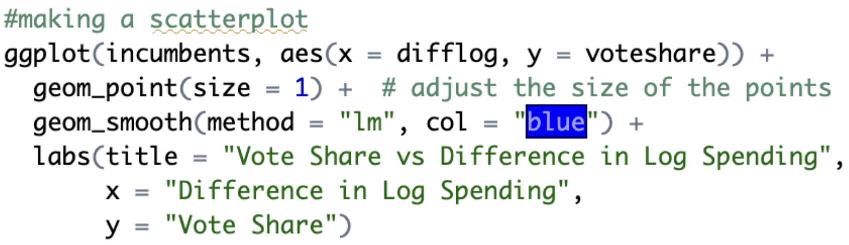


```
> # print results
> print(result1$coefficients)       # coefficients
                [,1]
Intercept 0.57903071
difflog   0.04166632
> print(result1$standard_errors)    # SEs
[1] 0.0022513886 0.0009679924
> print(result1$t_statistics)       # t-statistics
               [,1]
Intercept 257.18826
difflog    43.04406
> print(result1$p_values)           # p-values
                   [,1]
Intercept  0.000000e+00
difflog   1.359767e-319
```

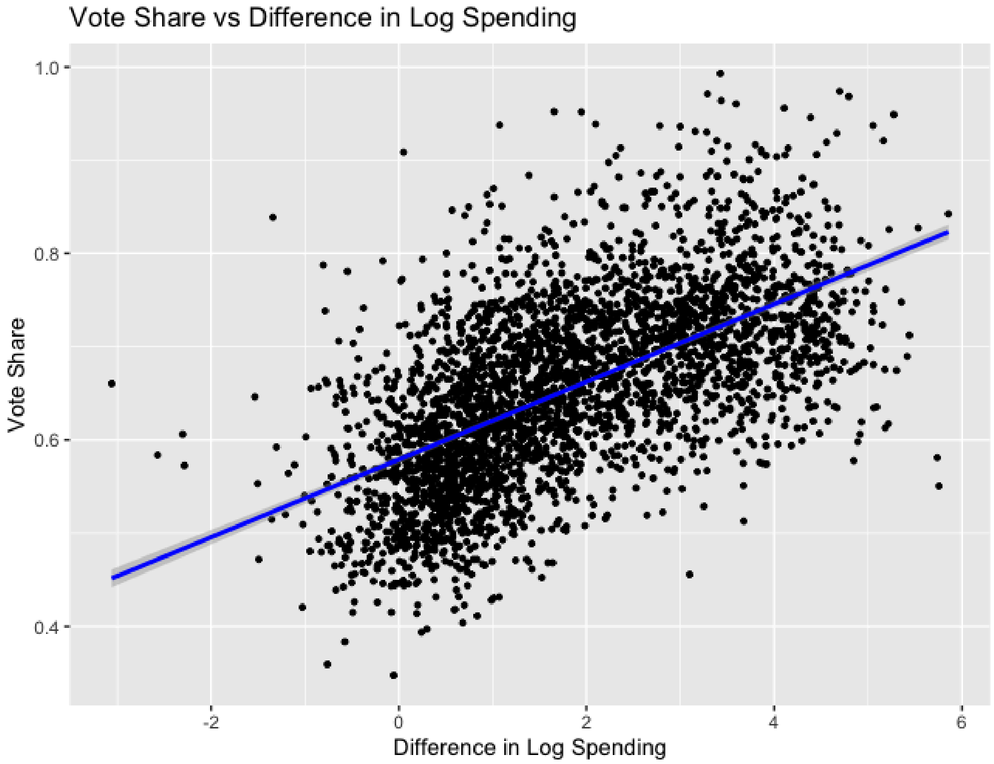


```
> summary(auto_results1)

Call:
lm(formula = voteshare ~ difflog, data = incumbents)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.26832 -0.05345 -0.00377  0.04780  0.32749 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.579031   0.002251  257.19   <2e-16 ***
difflog     0.041666   0.000968   43.04   <2e-16 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07867 on 3191 degrees of freedom
Multiple R-squared:  0.3673,	Adjusted R-squared:  0.3671 
F-statistic:  1853 on 1 and 3191 DF,  p-value: < 2.2e-16
```

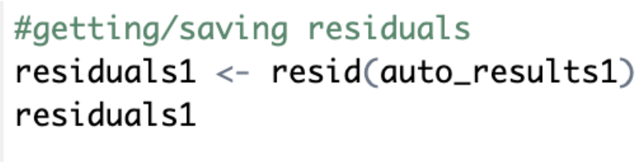
2. Getting the scatterplot

a. 

```
#making a scatterplot
ggplot(incumbents, aes(x = difflog, y = voteshare)) +
  geom_point(size = 1) +  # adjust the size of the points
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "Vote Share vs Difference in Log Spending",
       x = "Difference in Log Spending",
       y = "Vote Share")
```

b. 

3. Getting the residuals

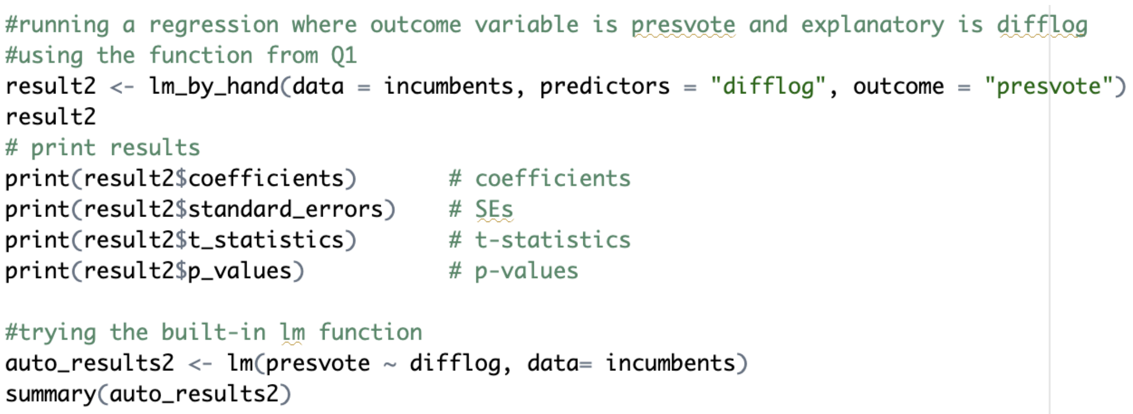
   a. 

```
#getting/saving residuals
residuals1 <- resid(auto_results1)
residuals1
```

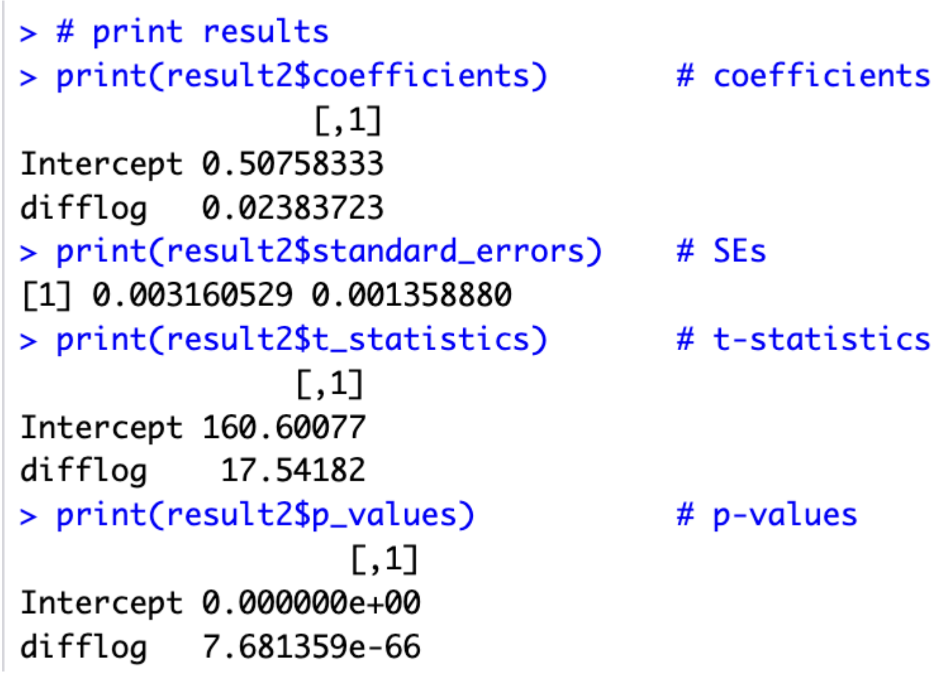
4. Writing the prediction equation
   a. Prediction = intercept + (slope x input value for difflog)
   b. y^ = 0.579031 + 0.0461666 x difflog

Question 2

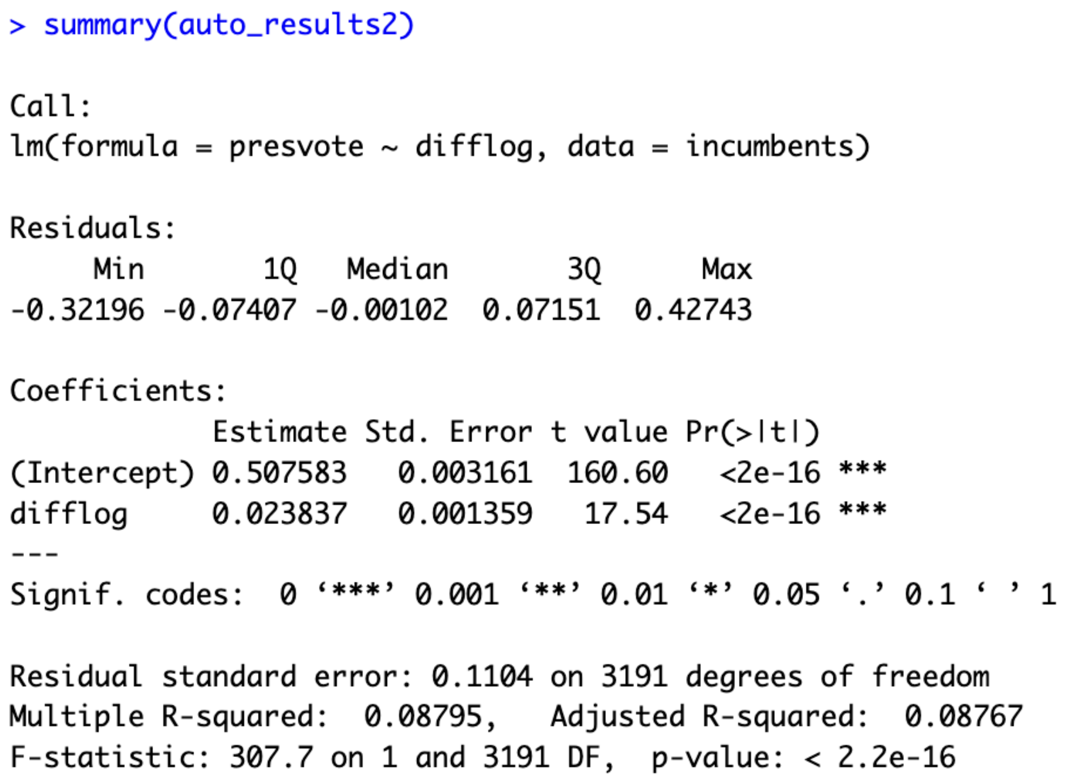
1.



```r
#running a regression where outcome variable is presvote and explanatory is difflog
#using the function from Q1
result2 <- lm_by_hand(data = incumbents, predictors = "difflog", outcome = "presvote")
result2
# print results
print(result2$coefficients)       # coefficients
print(result2$standard_errors)    # SEs
print(result2$t_statistics)       # t-statistics
print(result2$p_values)           # p-values

#trying the built-in lm function
auto_results2 <- lm(presvote ~ difflog, data= incumbents)
summary(auto_results2)
```

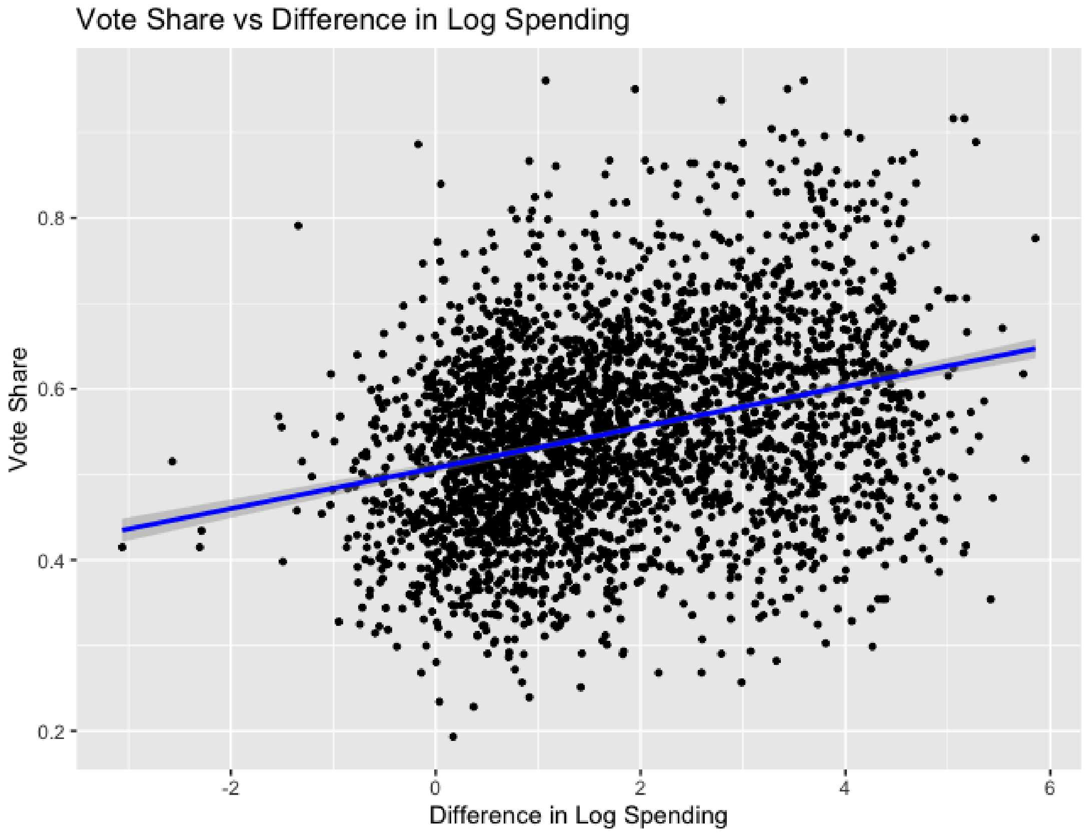


```
> # print results
> print(result2$coefficients)       # coefficients
                [,1]
Intercept 0.50758333
difflog   0.02383723
> print(result2$standard_errors)    # SEs
[1] 0.003160529 0.001358880
> print(result2$t_statistics)       # t-statistics
               [,1]
Intercept 160.60077
difflog    17.54182
> print(result2$p_values)           # p-values
                 [,1]
Intercept 0.000000e+00
difflog   7.681359e-66
```



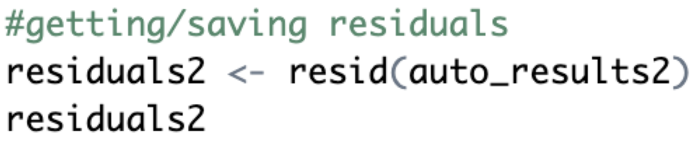
```
> summary(auto_results2)

Call:
lm(formula = presvote ~ difflog, data = incumbents)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.32196 -0.07407 -0.00102  0.07151  0.42743 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.507583   0.003161  160.60   <2e-16 ***
difflog     0.023837   0.001359   17.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1104 on 3191 degrees of freedom
Multiple R-squared:  0.08795,	Adjusted R-squared:  0.08767 
F-statistic: 307.7 on 1 and 3191 DF,  p-value: < 2.2e-16
```

2. Making a scatter plot



   a.

3. Getting and saving residuals



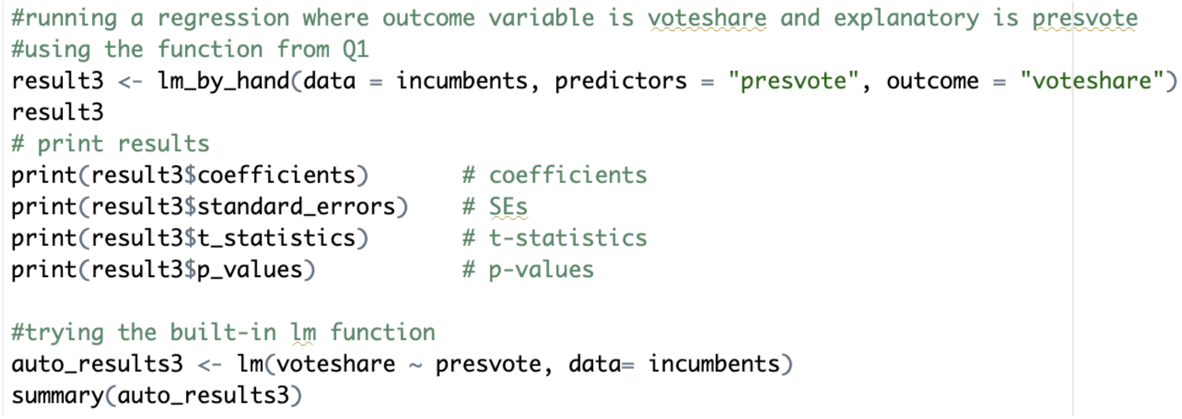
```
#getting/saving residuals
residuals2 <- resid(auto_results2)
residuals2
```

   a.

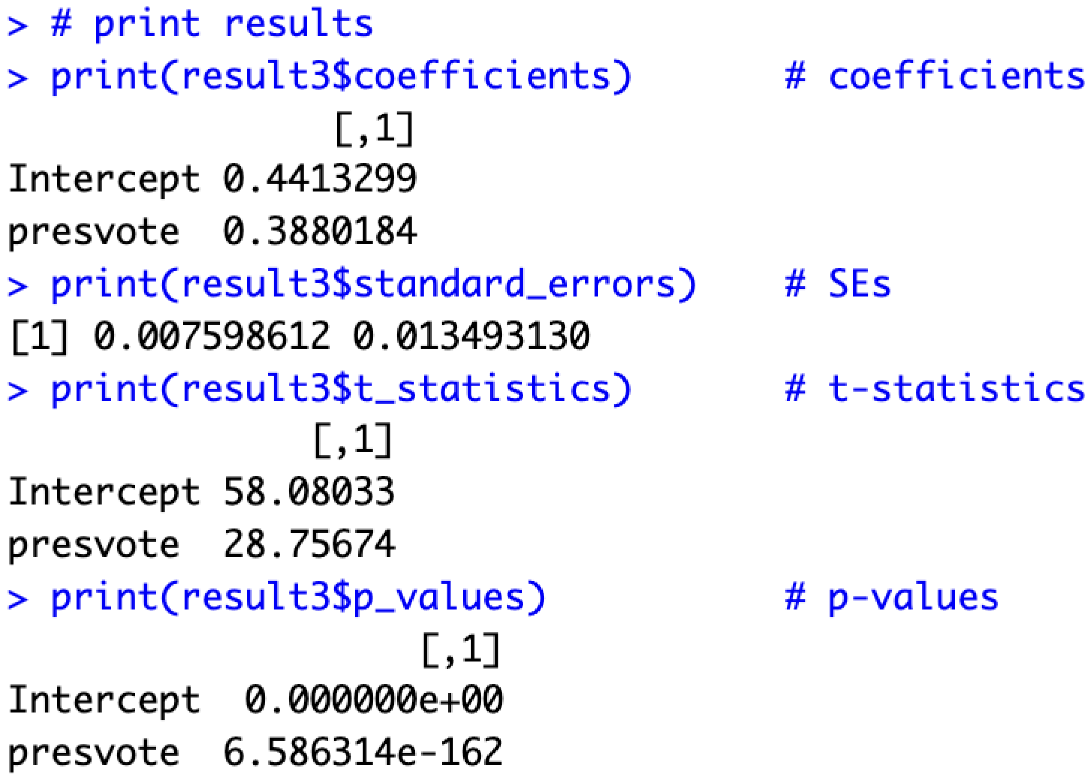
4. Writing the prediction equation
   a. Prediction = intercept + (slope x input value for difflog)
   b. y^ = 0.507583 + 0.023837 x difflog

Question 3

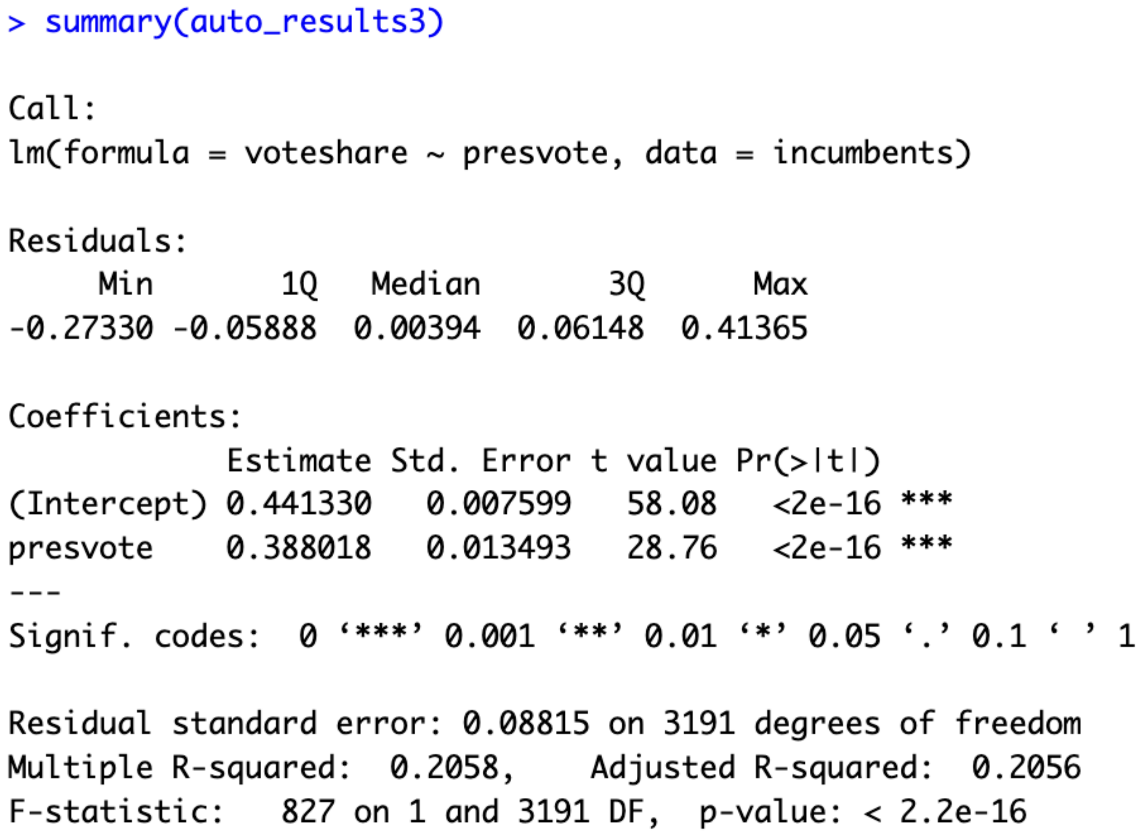
1.



```
#running a regression where outcome variable is voteshare and explanatory is presvote
#using the function from Q1
result3 <- lm_by_hand(data = incumbents, predictors = "presvote", outcome = "voteshare")
result3
# print results
print(result3$coefficients)       # coefficients
print(result3$standard_errors)    # SEs
print(result3$t_statistics)       # t-statistics
print(result3$p_values)           # p-values

#trying the built-in lm function
auto_results3 <- lm(voteshare ~ presvote, data= incumbents)
summary(auto_results3)
```

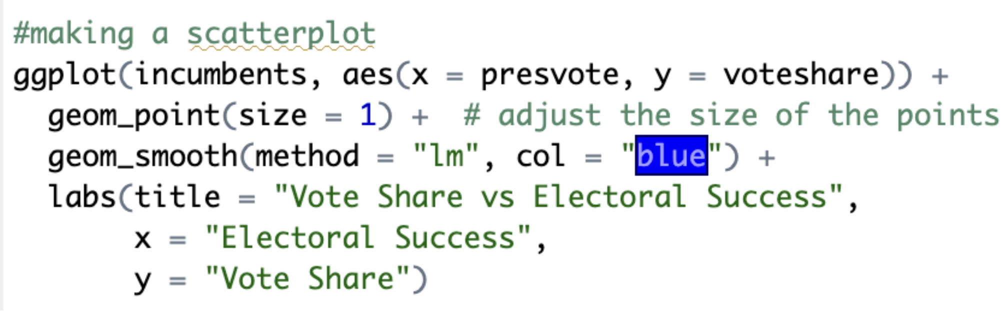


```
> # print results
> print(result3$coefficients)        # coefficients
                [,1]
Intercept 0.4413299
presvote  0.3880184
> print(result3$standard_errors)     # SEs
[1] 0.007598612 0.013493130
> print(result3$t_statistics)        # t-statistics
               [,1]
Intercept 58.08033
presvote  28.75674
> print(result3$p_values)            # p-values
                    [,1]
Intercept  0.000000e+00
presvote  6.586314e-162
```



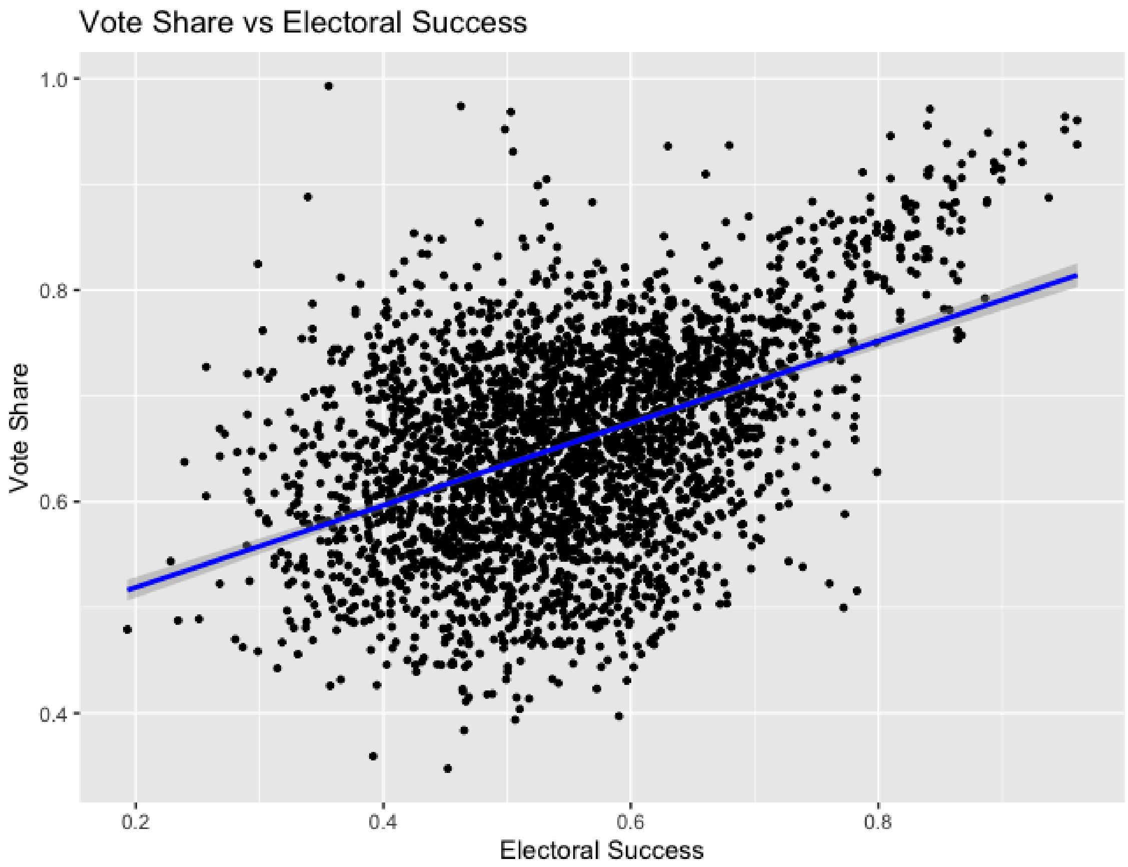
```
> summary(auto_results3)

Call:
lm(formula = voteshare ~ presvote, data = incumbents)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.27330 -0.05888  0.00394  0.06148  0.41365 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.441330   0.007599   58.08   <2e-16 ***
presvote    0.388018   0.013493   28.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08815 on 3191 degrees of freedom
Multiple R-squared:  0.2058,	Adjusted R-squared:  0.2056 
F-statistic:   827 on 1 and 3191 DF,  p-value: < 2.2e-16
```

2. Making a scatter plot

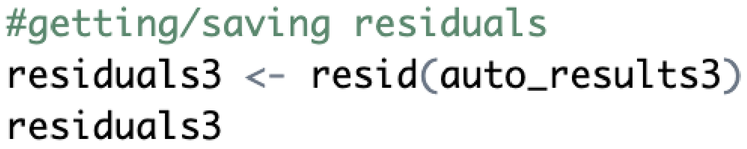


```
#making a scatterplot
ggplot(incumbents, aes(x = presvote, y = voteshare)) +
  geom_point(size = 1) +  # adjust the size of the points
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "Vote Share vs Electoral Success",
       x = "Electoral Success",
       y = "Vote Share")
```

a.



b.



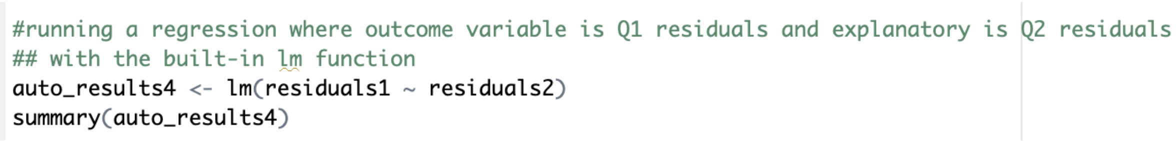
```
#getting/saving residuals
residuals3 <- resid(auto_results3)
residuals3
```

3.

4. Writing the prediction equation
    a. Prediction = intercept + (slope x input value for presvote)
    b. y^ = 0.441330 + 0.388018 x presvote

Question 4

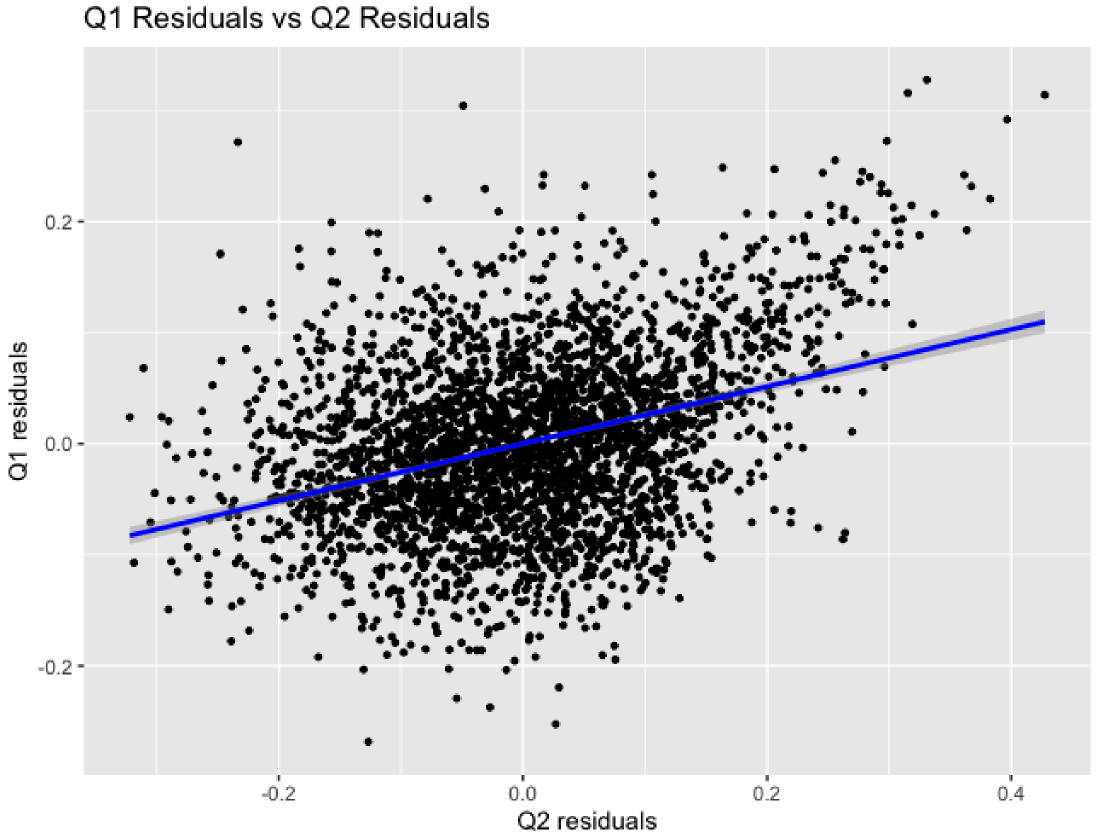
1.



```
#running a regression where outcome variable is Q1 residuals and explanatory is Q2 residuals
## with the built-in lm function
auto_results4 <- lm(residuals1 ~ residuals2)
summary(auto_results4)
```

```
> summary(auto_results4)

Call:
lm(formula = residuals1 ~ residuals2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.25928 -0.04737 -0.00121  0.04618  0.33126 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.876e-17  1.299e-03    0.00        1    
residuals2  2.569e-01  1.176e-02   21.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07338 on 3191 degrees of freedom
Multiple R-squared:   0.13,	Adjusted R-squared:  0.1298 
F-statistic:   477 on 1 and 3191 DF,  p-value: < 2.2e-16
```

2. Making a scatterplot



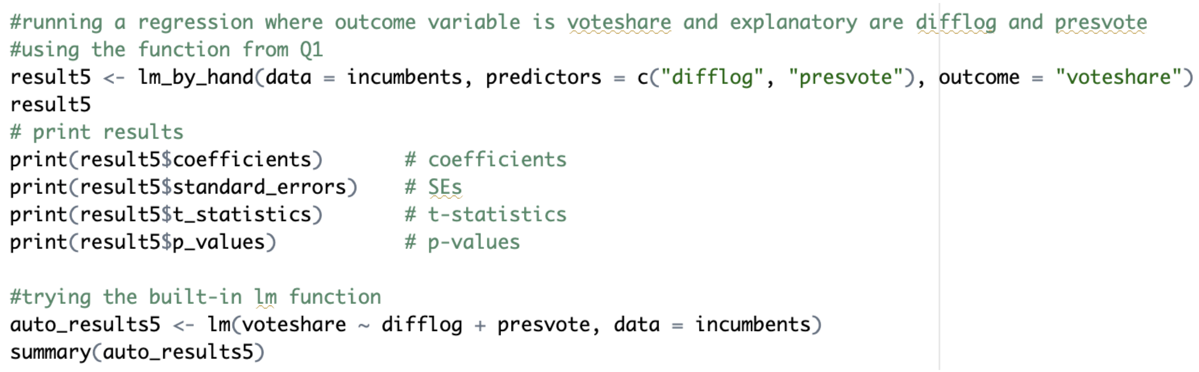
    a.

3. Writing the prediction equation
    a. Writing the prediction equation
        i. Prediction = intercept + (slope x input value for residuals2)

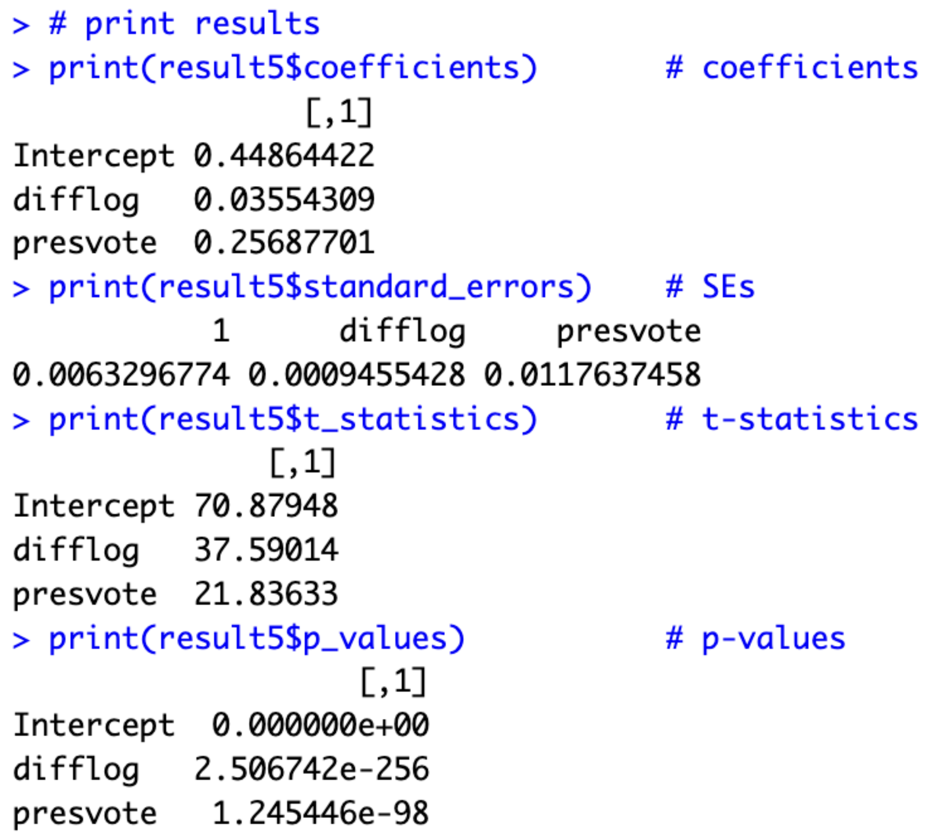
ii. y^ = 3.876e-17 + 2.569e-01 x residuals2

Question 5

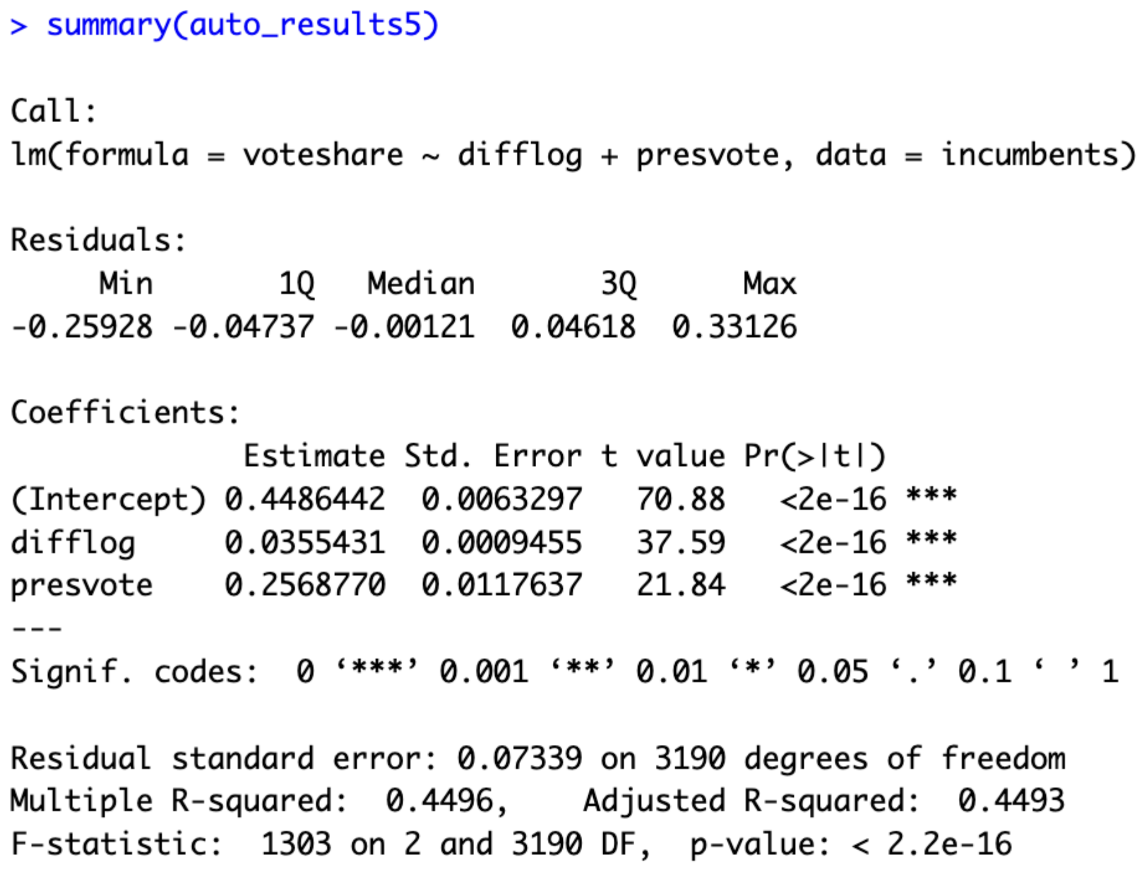
1.



```
#running a regression where outcome variable is voteshare and explanatory are difflog and presvote
#using the function from Q1
result5 <- lm_by_hand(data = incumbents, predictors = c("difflog", "presvote"), outcome = "voteshare")
result5
# print results
print(result5$coefficients)       # coefficients
print(result5$standard_errors)    # SEs
print(result5$t_statistics)       # t-statistics
print(result5$p_values)           # p-values

#trying the built-in lm function
auto_results5 <- lm(voteshare ~ difflog + presvote, data = incumbents)
summary(auto_results5)
```

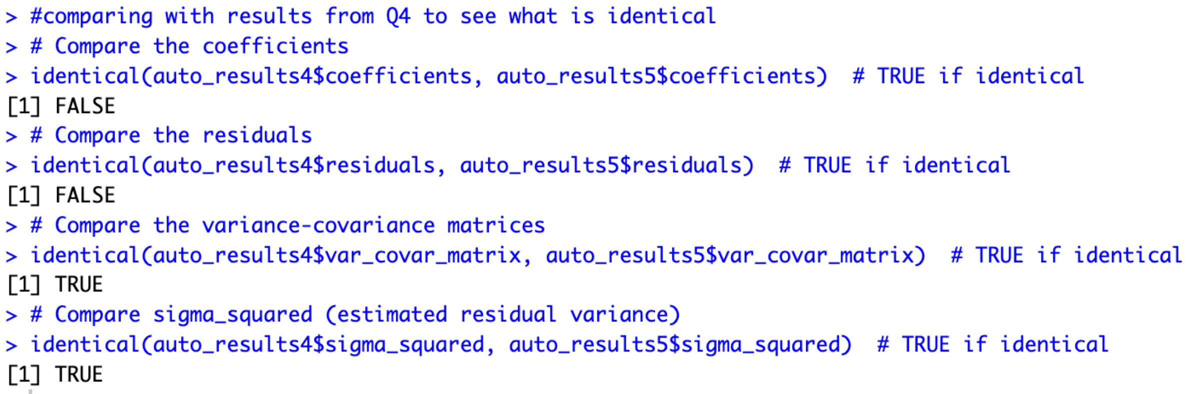


```
> # print results
> print(result5$coefficients)       # coefficients
                [,1]
Intercept 0.44864422
difflog   0.03554309
presvote  0.25687701
> print(result5$standard_errors)    # SEs
           1      difflog     presvote 
0.0063296774 0.0009455428 0.0117637458 
> print(result5$t_statistics)       # t-statistics
              [,1]
Intercept 70.87948
difflog   37.59014
presvote  21.83633
> print(result5$p_values)           # p-values
                   [,1]
Intercept  0.000000e+00
difflog   2.506742e-256
presvote   1.245446e-98
```



```
> summary(auto_results5)

Call:
lm(formula = voteshare ~ difflog + presvote, data = incumbents)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.25928 -0.04737 -0.00121  0.04618  0.33126 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.4486442  0.0063297   70.88   <2e-16 ***
difflog     0.0355431  0.0009455   37.59   <2e-16 ***
presvote    0.2568770  0.0117637   21.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07339 on 3190 degrees of freedom
Multiple R-squared:  0.4496,	Adjusted R-squared:  0.4493 
F-statistic:  1303 on 2 and 3190 DF,  p-value: < 2.2e-16
```

2. Writing the prediction equation
    a. voteshare = 0.4486442 + 0.03554309 * difflog + 0.256877 * presvote
3. Comparing outputs from Q4 and Q5 to see what is identical
    a.



```
> #comparing with results from Q4 to see what is identical
> # Compare the coefficients
> identical(auto_results4$coefficients, auto_results5$coefficients)  # TRUE if identical
[1] FALSE
> # Compare the residuals
> identical(auto_results4$residuals, auto_results5$residuals)  # TRUE if identical
[1] FALSE
> # Compare the variance-covariance matrices
> identical(auto_results4$var_covar_matrix, auto_results5$var_covar_matrix)  # TRUE if identical
[1] TRUE
> # Compare sigma_squared (estimated residual variance)
> identical(auto_results4$sigma_squared, auto_results5$sigma_squared)  # TRUE if identical
[1] TRUE
```

    b. The variance-covariance matrices and the sigma-squared values are identical in Q4 and Q5. This indicates that both models have the same level of unexplained variability in the outcome variable, voteshare. And because sigma-squared is identical, neither model is better than the other at explaining overall variance in voteshare. Additionally, because the variance-covariance matrices are the same, it suggests that the standard errors for the two models are unchanged/the same. This outcome indicates a possible collinearity between difflog and presvote, they might explain overlapping portions of voteshare. So given this information, I don’t believe there is sufficient evidence that difference in spending or presidential popularity has more of an effect on the incumbent’s vote share than the other.