# STA 210: HW 3

*Jackson Hubbard*

*Septemnber 27*

## Question 1: Ex. 7.24

### a

```
#use ex0724 data set
male_births <- ex0724

model_denmark <- lm(Denmark~ Year, data= male_births)
tidy(model_denmark)
```

```
## # A tibble: 2 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  0.599     0.0408        14.7  2.40e-18
## 2 Year        -0.0000429 0.0000207     -2.07 4.42e- 2
```

```
model_Netherlands <- lm(Netherlands~ Year, data= male_births)
tidy(model_Netherlands)
```

```
## # A tibble: 2 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  0.672     0.0279        24.1  1.37e-26
## 2 Year        -0.0000808 0.0000142     -5.71 9.64e- 7
```

```
model_Canada <- lm(Canada~ Year, data= male_births)
tidy(model_Canada)
```

```
## # A tibble: 2 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  0.734    0.0548        13.4  3.98e-11
## 2 Year        -0.000111 0.0000277     -4.02 7.38e- 4
```

```
model_USA <- lm(USA~ Year, data= male_births)
tidy(model_USA)
```

```
## # A tibble: 2 x 5
##   term          estimate  std.error statistic  p.value
```

```
##   <chr>            <dbl>      <dbl>     <dbl>    <dbl>
## 1 (Intercept)  0.620     0.0186         33.3  2.52e-18
## 2 Year        -0.0000543 0.00000939     -5.78 1.44e- 5
```

Looking at the output, we see that the estimates and standard error's from the textbook are replicated correctly.

**b**

The tables also show the test statistics for the the test that the slope of the regression model on the proportion of male births by year is zero. Looking at each of the tables, we see that each of the 4 countries have test statistics that are sufficiently high. This means that all of the resulting p values are smaller than 0.05 which means that there is sufficient evidence to reject the null hypothesis of the slopes being equal to 0. Further, since eachof the estimates for the slope are negative, we see that the proportion of male births is declining.

**c**

The United States has the largest of the 4 test statistics even though its slope is only the third largest. This can be understood by looking at the equation for the test statistic, which is $(\hat{\beta}_1/\beta_1)/SE(\hat{\beta}_1)$. In this formula, the numerator represents the slope coefficient. Thus, the United States has the third largest numerator. However, the United States also has the smallest standard error its slope. This results in the largest test statistic.

**d**

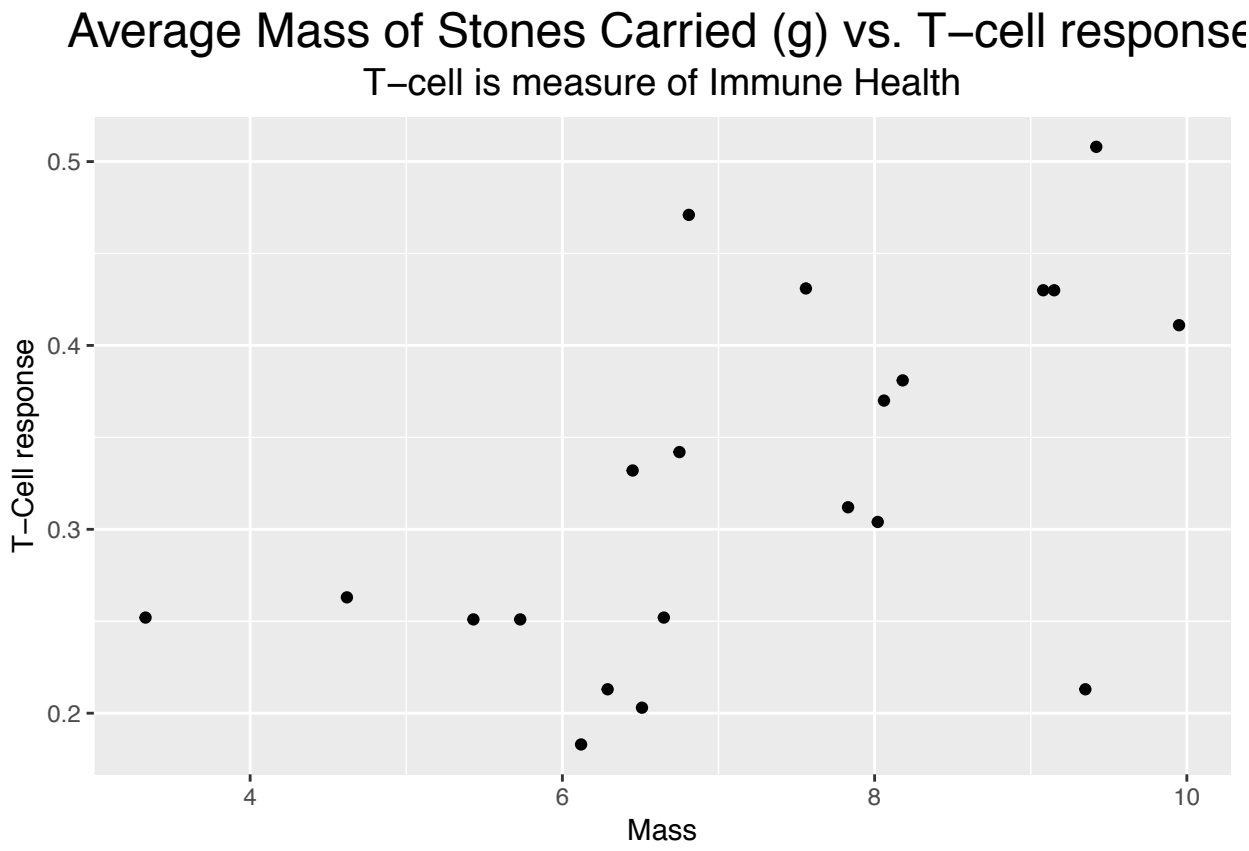
The United States has a lower standard error of the estimated slope than Canada even though their sample sizes are the same because the formula for standard error is $\sigma/\sqrt{n}$. Thus even though the n value is the same, the United States has a smaller standard deviation, which results in a smaller stanard error.

**e**

Since the proportion of male births is reported as an average, the sample size comes into play. As the sample size increases, the standard deviation decreases because the formula for standard deviations has n in the denominator.

## Question 2: Ex. 7.27

```r
#use ex0727 data set
q2 <- ex0727

q2 %>% ggplot(aes(x = Mass, y = Tcell)) + geom_point() + labs(title = "Average Mass of S
y = "T-Cell response", subtitle = "T-cell is measure of Immune Health") +
  theme(plot.title = element_text(hjust = 0.5,size=18),
      plot.subtitle=element_text(hjust=0.5,size=14))
```




Looking at the scatterplot, we see that there is a small positive correlation between mean mass size and T cell response so I will create a linear model.

```r
model1 <- lm(Tcell ~ Mass, data = q2)
tidy(model1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   0.0875    0.0787      1.11 0.280
## 2 Mass          0.0328    0.0106      3.08 0.00611
```

```r
glance(model1)
```

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared  sigma statistic p.value    df logLik   AIC   BIC
```

```
## *     <dbl>         <dbl>  <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
## 1     0.334         0.299 0.0810      9.51 0.00611     2   24.0 -42.1 -38.9
## # ... with 2 more variables: deviance <dbl>, df.residual <int>
```

$H_0 : \beta_1 = 0$
$H_A : \beta_1 \neq 0$ $\beta_1$ is the slope of the model predicting the T-cell response based on the average mass of stones carried by the bird.

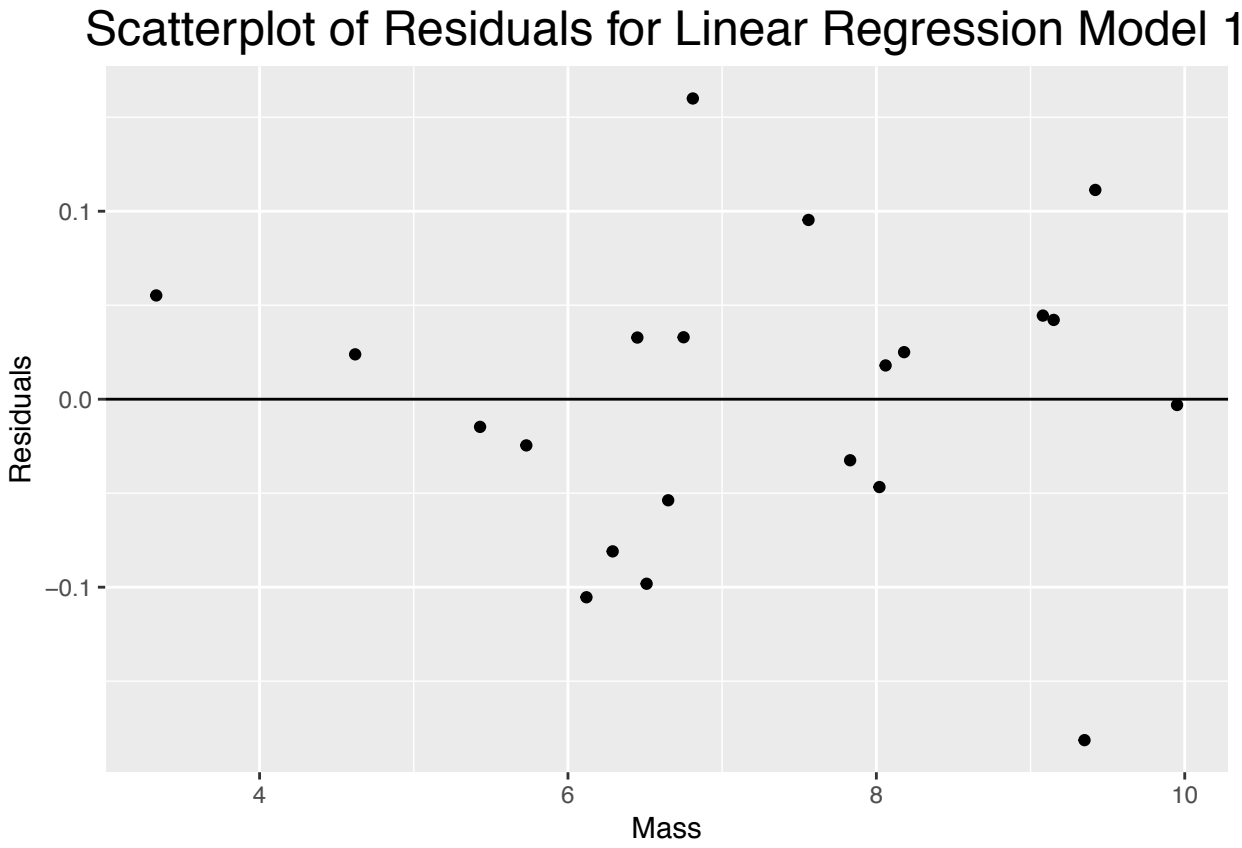
Looking at the output we see that the estimate for $\beta_1$ is 0.03282149. This slope has a p value of 0.006105022, which is lower than 0.05. This means that it is significant. There is sufficient evidence to reject the null hypothesis, which means there is sufficient evidence that the slope is not equal to 0.

The interpretation of this slope is that for every for every 1 gram increase in the mean stone mass that a bird carries, their T-cell response is increases by 0.03282149 mm on average. The interpretation of the intercept is that if a wheatears bird carried 0 grams of stones to his nest, his average T-cell repsonse is 0.08749698 mm.

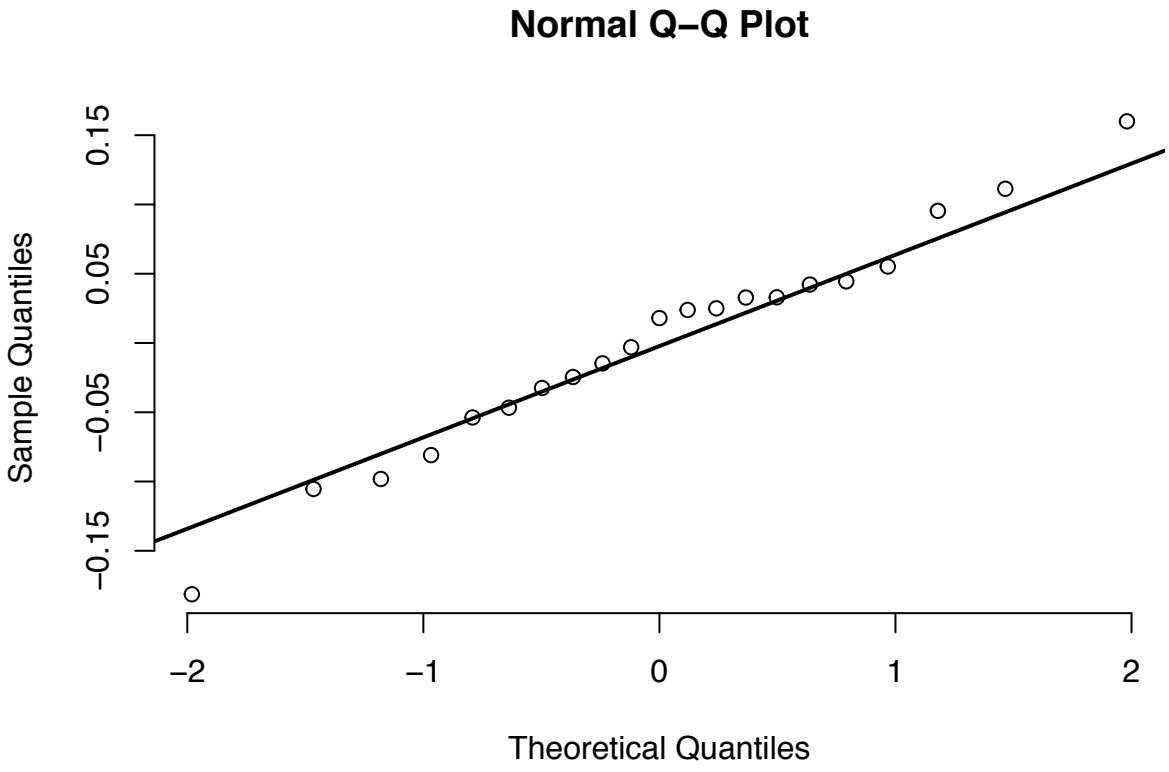
Now, looking at the other table outputted, we see that the R squared value is .334. This means that 33.4% of the variation in T-cell response is explained by the model.

Checking assumptions of linear model:

```
q2 <- q2 %>% mutate(resid1 = resid(model1))


q2 %>% ggplot(aes(x = Mass, y = resid1)) + geom_point() + geom_hline(yintercept = 0) +
plot.subtitle=element_text(hjust=0.5,size=14))
```




```r
qqnorm(q2$resid1, pch = 1, frame = FALSE)

qqline(q2$resid1, col = "black", lwd = 2)
```




To check the assumptions, first we look at the residuals scatterplot. We see that there is no systemized pattern and that the residuals appear to be random. This means that we were correct in building a linear model, as there is indeed a linear relationship between Mass of stones carried and the t-cell reposnse. This graph also shows that the assumption of constant variance is also met. Next, looking at the QQ plot, we see that the Normality assumption is also met. We also can conclude that the observations are indpendent, as there are no cluster or serial effects.

Based on the results of the model and the confirmation of the validity of the assumptions we can conclude that there is sufficient evidence to reject the null hypothsis of there being no relationship between average mass of stones carried by wheatears birds and its health.

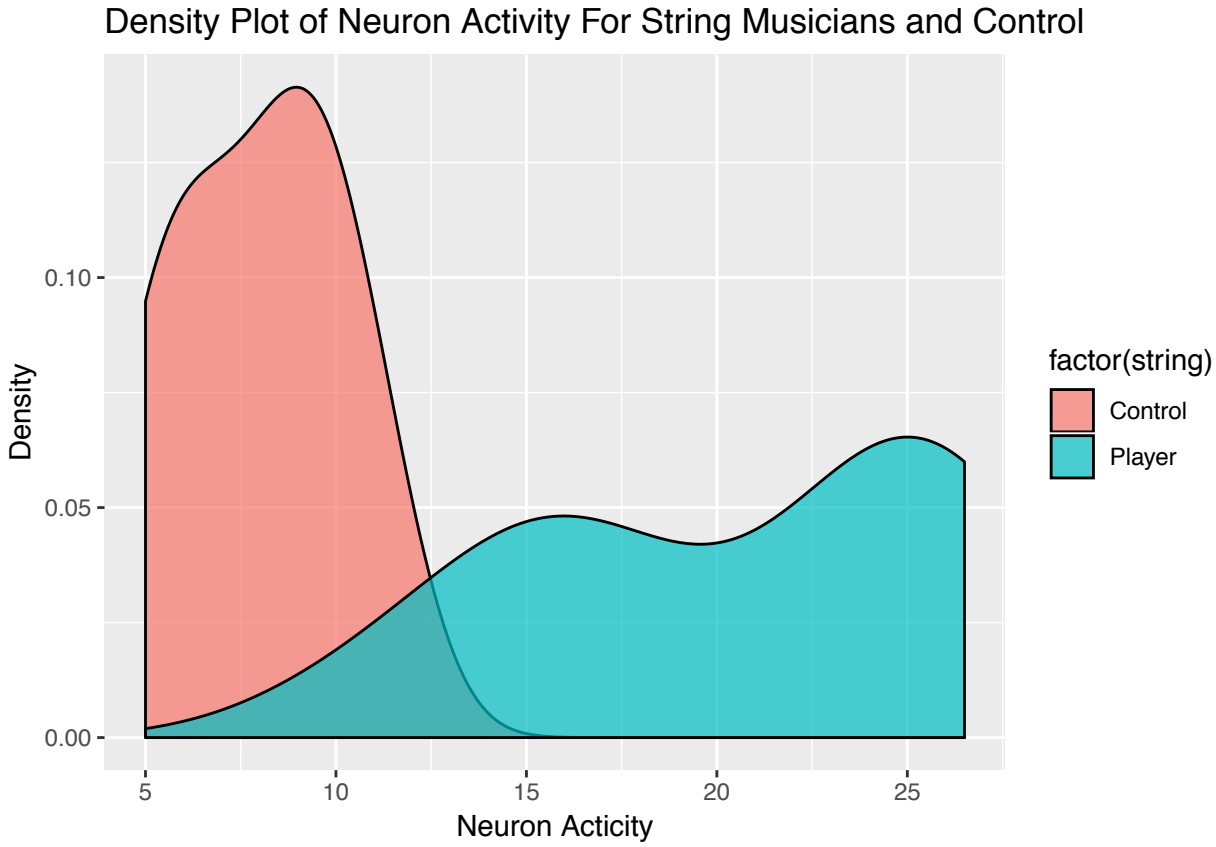
With the assumptions verified, we can say that the linear regression equation is T-cell = 0.0874 + 0.0328 * Mean Mass of Rocks carried

## Question 3: Ex. 7.28

Question #1- Is neuron activity different for string musicians (defined as playing years > 1) and the controls (playing years =0)?

```
#use ex0728 data set

q3 <- ex0728
q3 <- q3 %>% mutate(string = case_when(
Years == 0 ~ "Control",
```

```
Years > 0 ~ "Player" )
)


ggplot(q3, aes(Activity)) + geom_density(aes(fill = factor(string)),alpha=0.7) + labs(t
```




```
q3 %>%
group_by(string) %>%
summarise(n = n(), mean = mean(Activity), median = median(Activity), sd = sd(Activity))
```

```
## # A tibble: 2 x 5
##   string      n  mean median    sd
##   <chr>   <int> <dbl>  <dbl> <dbl>
## 1 Control     6   8     8.25  2.26
## 2 Player      9  20.6  23     5.59
```

Looking at the density plot and the summary statistics we see that there is a difference in mean Neuron activity between the Sting musicians and the control group. But to confirm this I will perform a 2 sample t test.

$H_0$ : $\mu_1$ - $\mu_2 = 0$
$H_A$ : $\mu_1$ - $\mu_2 \neq 0$ $\mu_1$ = mean activity of musicians $\mu_2$ = mean activity of contorl

When performing the t test, we want a two sided test (based off of our hypothesises) and we

need to set the equal variance parameter to FALSE (from the summary statistics we see that they are not equal)

```
musicians <- q3 %>% filter(Years > 0)
control <- q3 %>% filter(Years == 0)
t.test(musicians$Activity,control$Activity, alternative="two.sided", var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  musicians$Activity and control$Activity
## t = 6.067, df = 11.313, p-value = 7.205e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   8.051473 17.170750
## sample estimates:
## mean of x mean of y
##  20.61111   8.00000
```

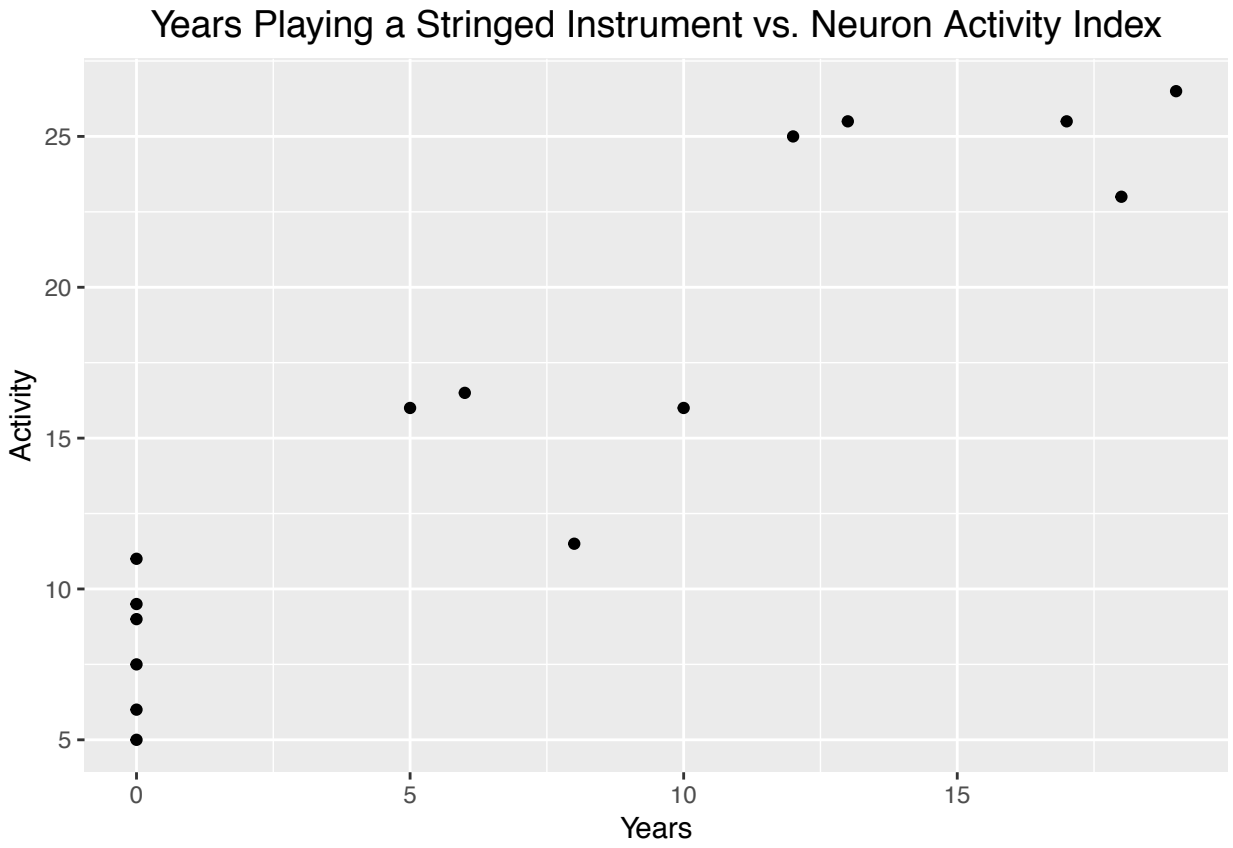
Looking at the results of the t-test, we see that the p-value is very small. This means there is sufficient evidence to reject the null hypothesis. There is sufficient evidence that the neuron activity is different between string players and the control (people who have never played a string instrument). We also see a 95% confidence interval (8.051473 17.170750), which does not include 0, which is further reason to believe there is a difference.

Question #2 Whether amount of neuron activity is associated with the number of years the person has played the instrument

Perform a linear regression model:

$H_0 : \beta_1 = 0$
$H_A : \beta_1 \neq 0$ $\beta_1$ is the slope of the model predicting the neuron response based on the number of years playing a Stringed instrument.

```
q3 %>% ggplot(aes(x = Years, y = Activity)) + geom_point() +
labs(title = "Years Playing a Stringed Instrument vs. Neuron Activity Index" ) + theme(p
```




```
model_neuron <- lm(Activity ~ Years, data = q3)
tidy(model_neuron)
```

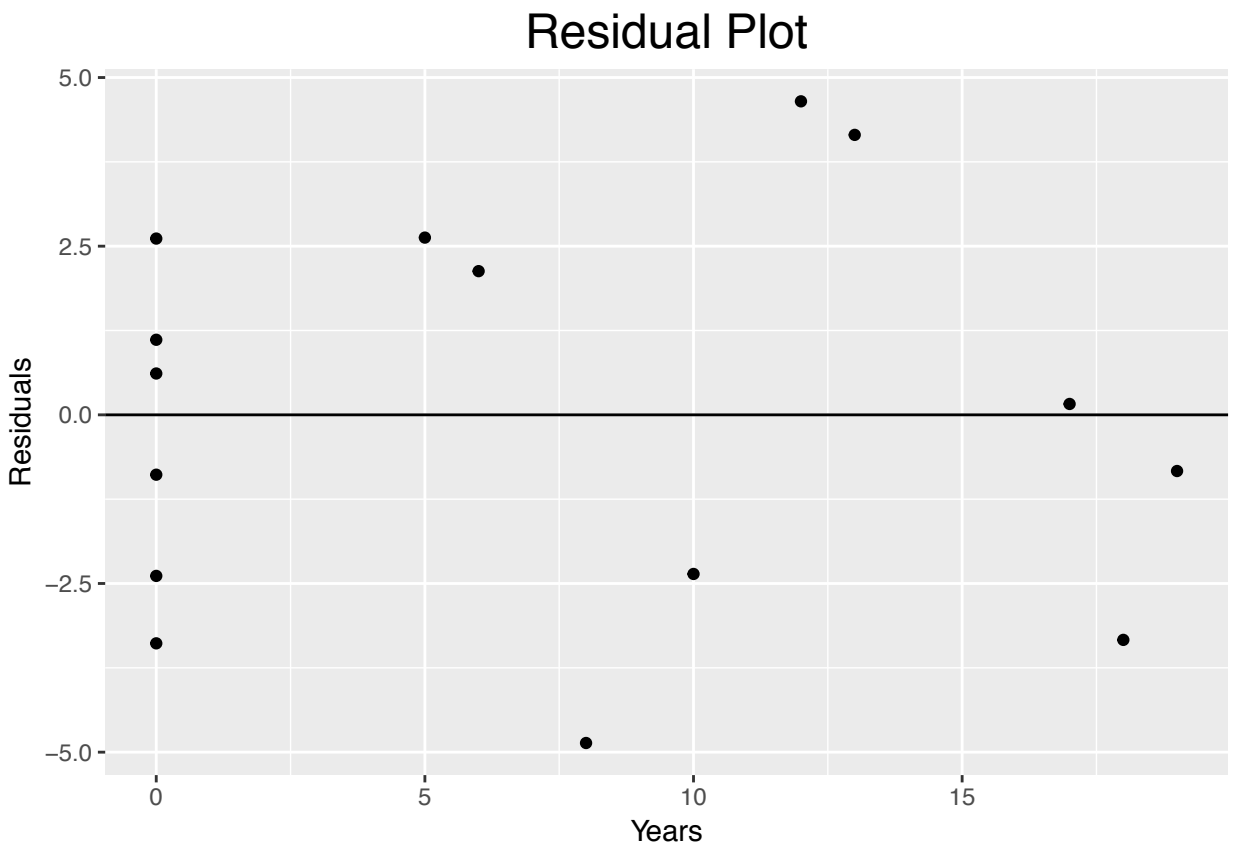
```
## # A tibble: 2 x 5
##   term        estimate std.error statistic     p.value
##   <chr>          <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept)    8.39      1.11       7.52 0.00000435
## 2 Years          0.997     0.111      8.98 0.000000618
```

```
glance(model_neuron)
```

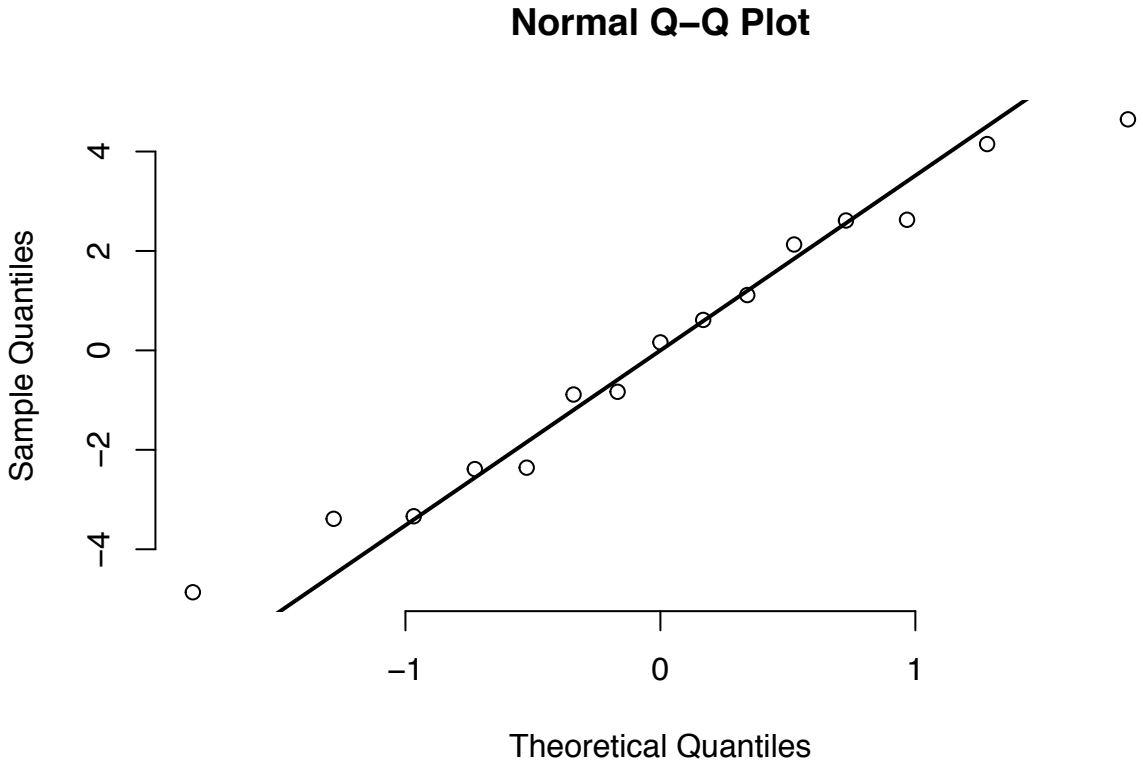
```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
## *     <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
## 1     0.861         0.850  3.01      80.6 6.18e-7     2  -36.7  79.5  81.6
## # ... with 2 more variables: deviance <dbl>, df.residual <int>
```

Looking at the outputed tables, we see that the p value is low. This means that there is sufficient evidence to reject the null hypothesis of there being no relationship between the number of years a person has played a string instrument and their neuron activity. We also see that the R squared is 0.861 which means that 86% of the variation in nueron activity is explained by the model.

```
q3 <- q3 %>% mutate(resid = resid(model_neuron))
q3 %>% ggplot(aes(x = Years, y = resid)) + geom_point() + geom_hline(yintercept=0) + la
theme(plot.title = element_text(hjust = 0.5,size=18))
```




```
qqnorm(q3$resid, pch = 1, frame = FALSE)
qqline(q3$resid, col = "black", lwd = 2)
```

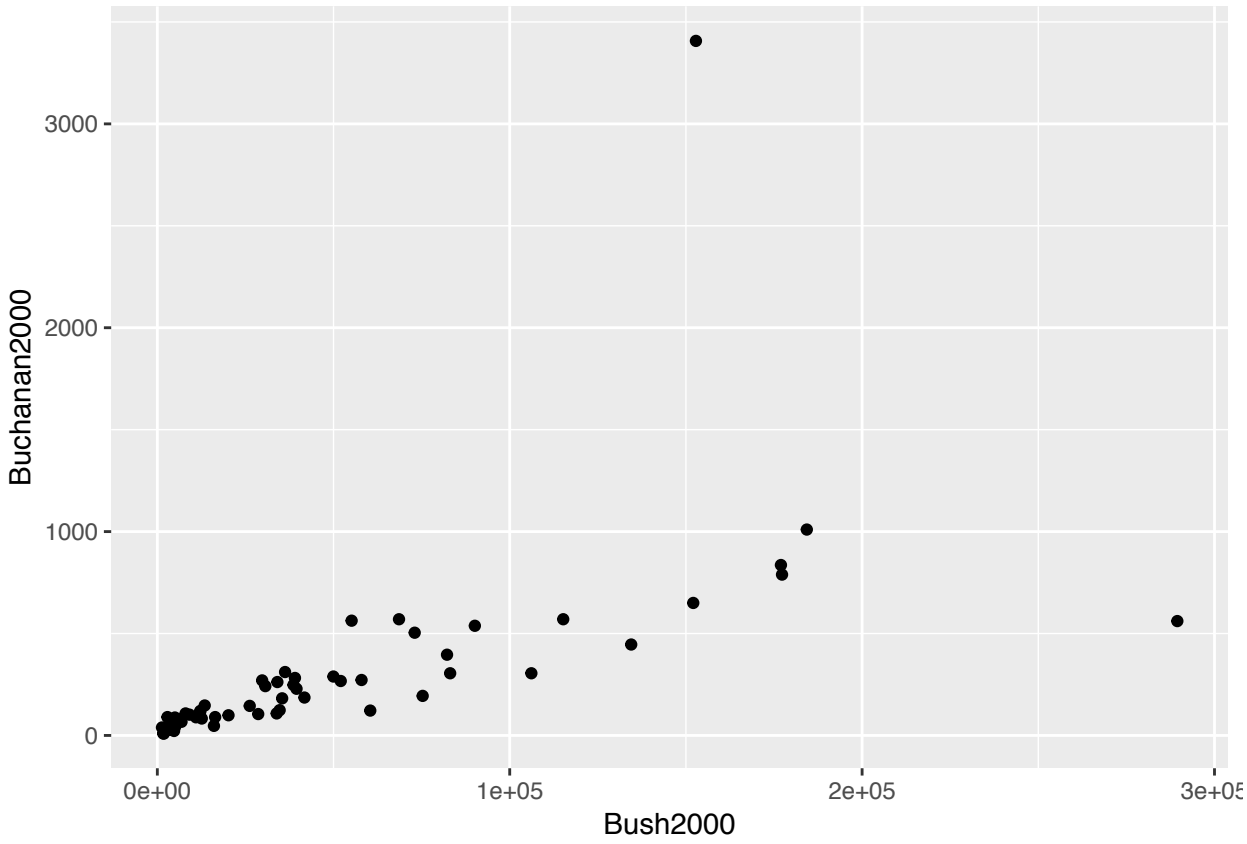



Looking at the residual plot, we see that there is no apparent pattern in the residuals. This means that the constant variance assumption is met and that the linear relationship assumption is also met. Also, from the QQ plot we see that the Normality assumpiton is met. The independence assumption is also met since the tests are done on different individual subjects and not taken over time.

With the assumptions verified, we can say that the linear regression equation is Activity = 8.39 + 0.997 * Years.

## Question 4: Ex. 8.25

```
#use ex0825 data set
q4 <- ex0825
ggplot(q4, aes(Bush2000, Buchanan2000)) + geom_point()
```




```
q4
```

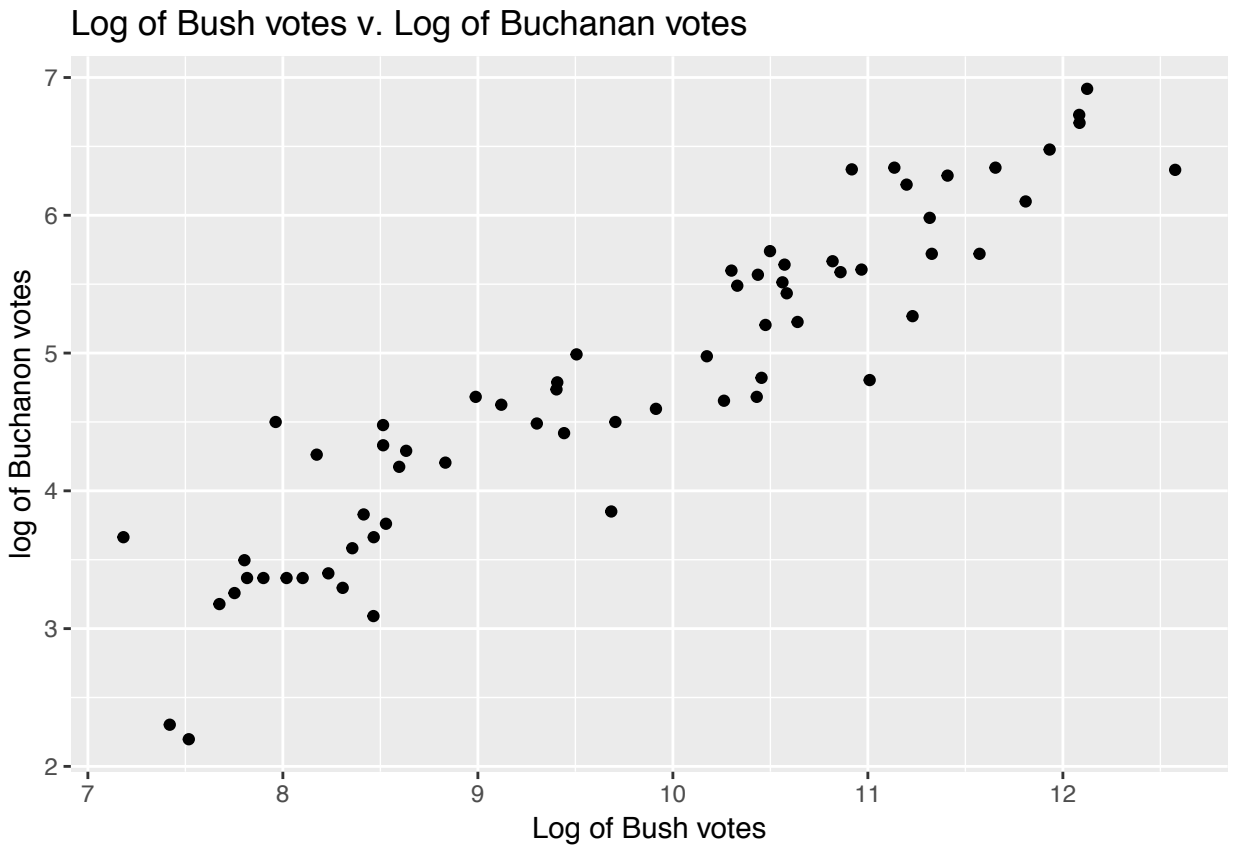
```
##          County Buchanan2000 Bush2000
## 1       Alachua           262    34062
## 2         Baker            73     5610
## 3           Bay           248    38637
## 4      Bradford            65     5413
## 5       Brevard           570   115185
## 6       Broward           789   177279
## 7       Calhoun            90     2873
## 8     Charlotte           182    35419
## 9        Citrus           270    29744
## 10         Clay           186    41745
## 11      Collier           122    60426
## 12     Columbia            89    10964
## 13         Dade           561   289456
## 14      De Soto            36     4256
## 15        Dixie            29     2698
## 16        Duval           650   152082
## 17     Escambia           504    73029
## 18      Flagler            83    12608
## 19     Franklin            33     2448
## 20      Gadsden            39     4750
```

```
## 21    Gilchrist          29     3300
## 22       Glades           9     1840
## 23         Gulf          71     3546
## 24     Hamilton          24     2153
## 25       Hardee          30     3764
## 26       Hendry          22     4743
## 27     Hernando         242    30646
## 28    Highlands          99    20196
## 29 Hillsborough         836   176967
## 30       Holmes          76     4985
## 31 Indian River         105    28627
## 32      Jackson         102     9138
## 33    Jefferson          29     2481
## 34    Lafayette          10     1669
## 35         Lake         289    49963
## 36          Lee         305   106141
## 37         Leon         282    39053
## 38         Levy          67     6860
## 39      Liberty          39     1316
## 40      Madison          29     3038
## 41      Manatee         272    57948
## 42       Marion         563    55135
## 43       Martin         108    33864
## 44       Monroe          47    16059
## 45       Nassau          90    16404
## 46     Okaloosa         267    52043
## 47   Okeechobee          43     5058
## 48       Orange         446   134476
## 49      Osceola         145    26216
## 50        Pasco         570    68581
## 51     Pinellas        1010   184312
## 52         Polk         538    90101
## 53       Putnam         147    13439
## 54    St. Johns         229    39497
## 55    St. Lucie         124    34705
## 56   Santa Rosa         311    36248
## 57     Sarasota         305    83100
## 58     Seminole         194    75293
## 59       Sumter         114    12126
## 60     Suwannee         108     8014
## 61       Taylor          27     4051
## 62        Union          26     2326
## 63      Volusia         396    82214
## 64      Wakulla          46     4511
## 65       Walton         120    12176
```

```
## 66    Washington            88     4983
## 67    Palm Beach          3407   152846
```

Looking at the scatterplot, we see that there is a large outlier. So comparing it to the data, we see that it is from the Palm Beach county.

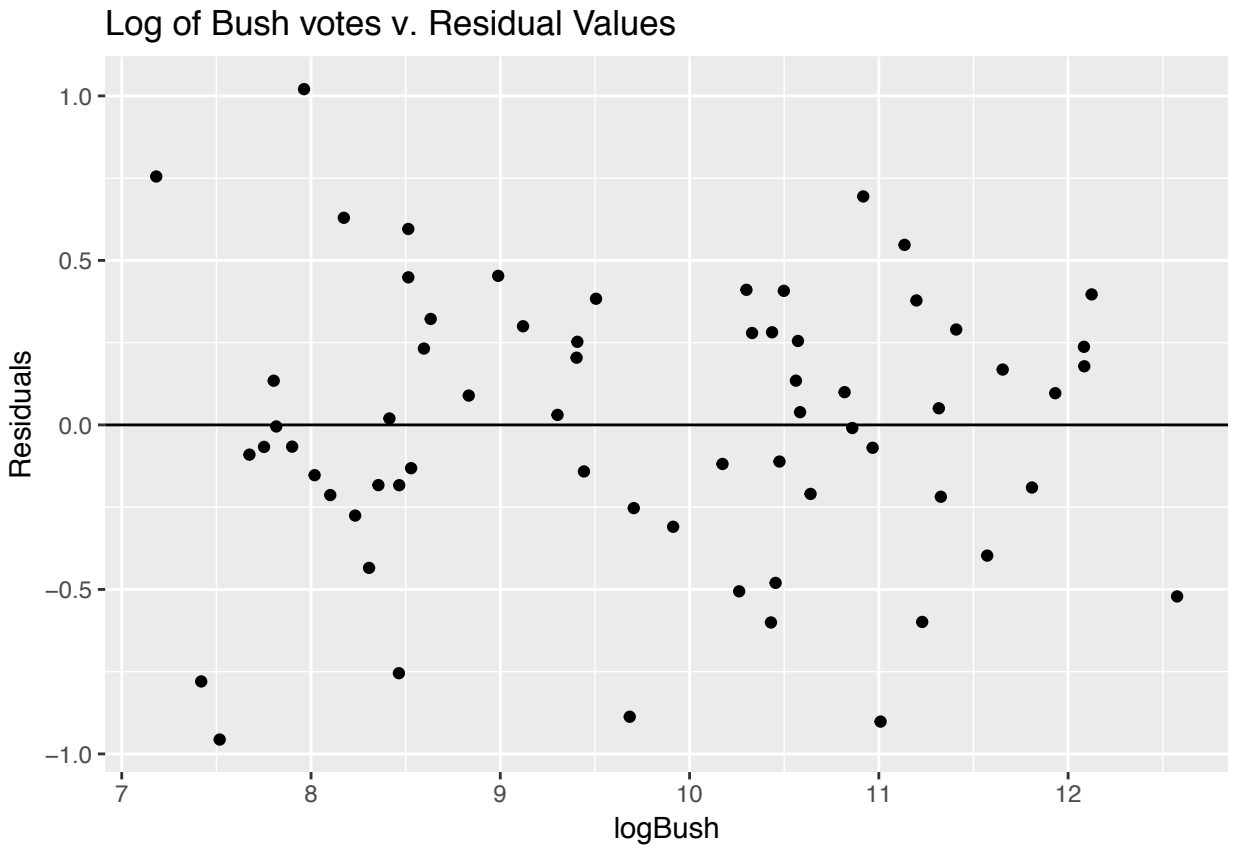
```
# Create a new data set with the Palm Beach observation removed
# for the regression analysis

q4 <- q4[-c(67),]
q4 <- q4 %>% mutate(logBuchanan = log(Buchanan2000), logBush = log(Bush2000))
# ggplot(q4, aes(Bush2000, Buchanan2000)) + geom_point()
# ggplot(q4, aes(logBush, Buchanan2000)) + geom_point()
# ggplot(q4, aes(Bush2000, logBuchanan)) + geom_point()
ggplot(q4, aes(logBush, logBuchanan)) + geom_point() + labs(x= "Log of Bush votes", y= "
```




To generate a proper linear regression model, I needed to do some transformations of the data. This involved some trial and error, but after a couple of tries logging the x and the y, I found that by logging both the x and the y variables resulted in the most linear relationship. Thus this is the model I am going to use.

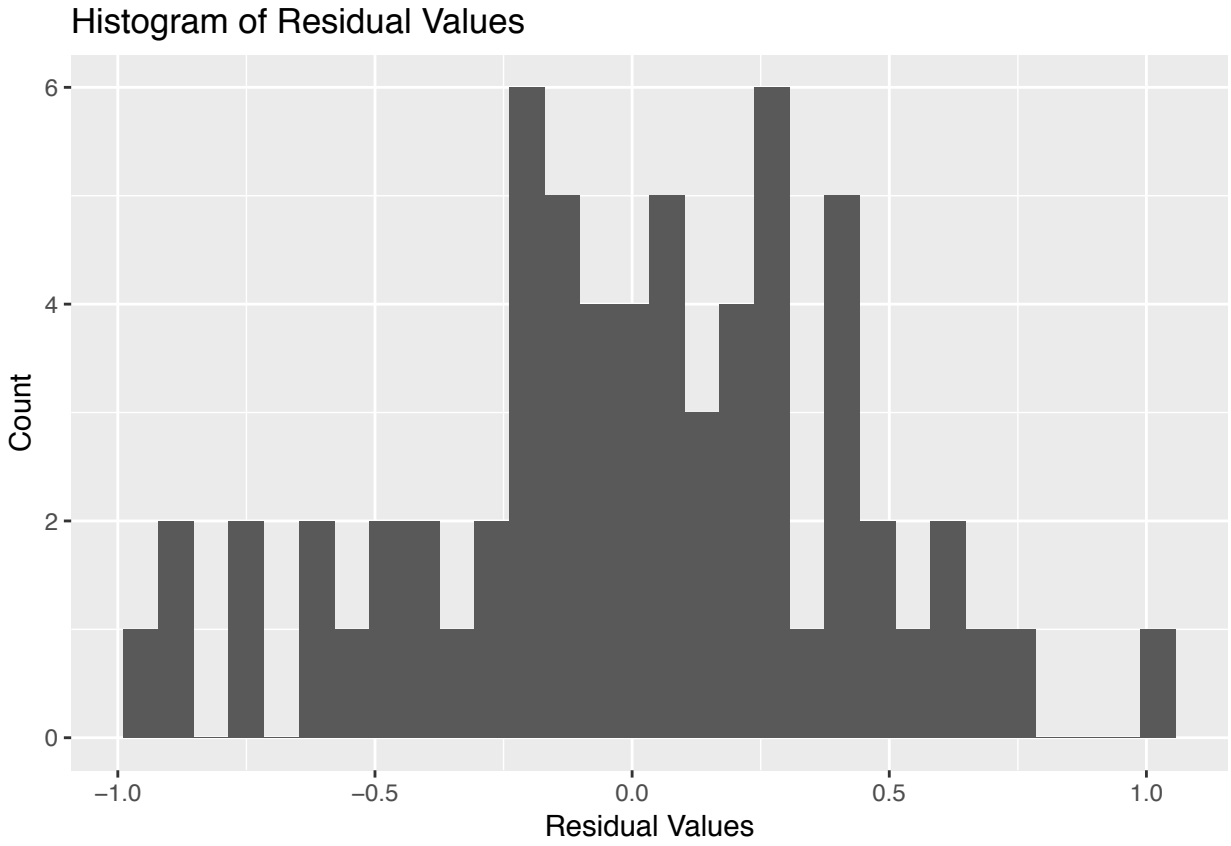
```
modelvotes <- lm(logBuchanan ~ logBush, data = q4)
q4 <- q4 %>% mutate(resid2000 = resid(modelvotes))
```

```
ggplot(q4, aes(logBush, resid2000)) + geom_point() + geom_hline(yintercept = 0) + labs(
```




```
ggplot(q4, aes(resid2000)) + geom_histogram() + labs(x = "Residual Values", y = "Count"
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




After doing a linear regression model, I must check the assumptions. Looking at the residual graph, we see that the resiuals do not have a pattern to them. This means that the constant variance and the linearity assumptions are met. Further, the histogram of the values of the residual is also normal, so the normality assumption is also met. The independence assumption is also met since the votes are idnependent.

```
PalmBeach <- 152846
newdata2000 <- data.frame(logBush = log(PalmBeach))
predict.lm(modelvotes, newdata2000, interval = "prediction", conf.level = 0.95)
```

```
##        fit      lwr     upr
## 1 6.384143 5.524656 7.24363
```

Taking a 95% confidence interval from the model for the true amount of Buchanan votes expected for Palm Beach county we see that: The predicted amount of Buchanan votes expected for Palm beach is about 592.37 ( e raised to 6.38).

The 95% confidence interval for the mean amount of votes is (250.80, 1399.16) (e raised to 5.52 and e raised to 7.24).

Using this confidence interval and combining it with the actual results, I am 95% confident that the amount of votes that went to Buchannan that were meant for Gore is within the interval (2007.84, 3156.20).