Matt Williams

4/20/2022

Concepts of Statistical Data Analysis

Final Project Responses

1. A study examining a possible relationship of football playing and concussions on hippocampus volume, in µL, in the brain. The study included three groups: controls who had never played football (Control), football players with no history of concussions (FBNoConcuss), and football players with a history of concussions (FBConcuss). The data is available in FootballBrain, and the side-by-side boxplots shown below.

a. Use technology to find the summary statistics for each of the three groups. Which group has the largest mean hippocampus volume? Which group has the smallest?

```
> summary(Hipp[Group=="Control"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6175    6780    7385    7603    8510    9710
> summary(Hipp[Group=="FBNoConcuss"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4810    5965    6515    6459    7020    7790
> summary(Hipp[Group== "FBConcuss"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4490    5505    5710    5735    6035    7160
> |
```

-
- Using R, I get the following results when getting the summary information for each of the 3 groups individually. The “Control” group has the largest hippocampus volume (mean = 7603, median = 7385). The “FBConcuss” group has the smallest hippocampus volume (mean = 5735, median = 5710)

b. Use technology to construct an ANOVA table. What is the F-statistic? What is the p-value?

```
> anova(lm(Hipp ~ Group))
Analysis of Variance Table

Response: Hipp
           Df   Sum Sq  Mean Sq F value    Pr(>F)
Group       2 44348606 22174303  31.473 1.507e-10 ***
Residuals 72 50727336   704546
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> |
```

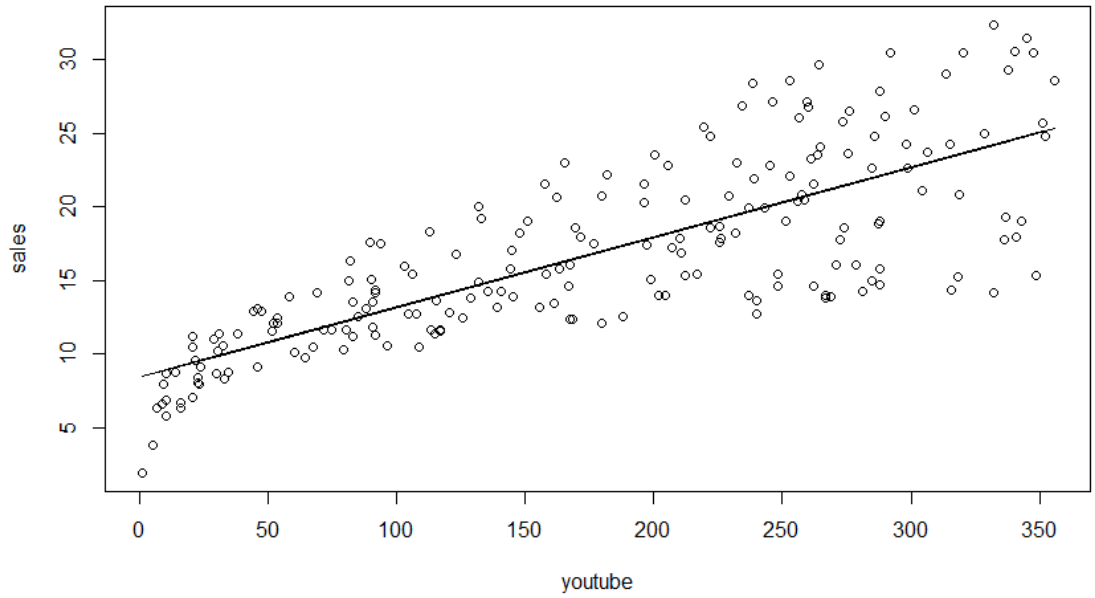
-

- Above we have the output of the ANOVA test for a linear regression model that predicts the volume of a person's hippocampus based on the group that person belongs to ("Control", "FBNoConcuss", "FBConcuss"). Here we can see that our F-statistic is equal to 31.473. We can also see that the p-value associated with this F-statistic is approximately equal to 0.

c. What is the conclusion of the test?
   - Since the p-value from part b was approximately equal to 0, if we were to make a conclusion for hypothesis testing, the conclusion would be that we reject the null hypothesis because our p-value is significant. The null hypothesis for this situation would state that "the mean hippocampus volume for all 3 groups is the same".

2. Marketing Data Set contains the impact of three advertising medias (YouTube, Facebook and newspaper) on sales. Data are the advertising budget in thousands of dollars along with the sales. The advertising experiment has been repeated 200 times.

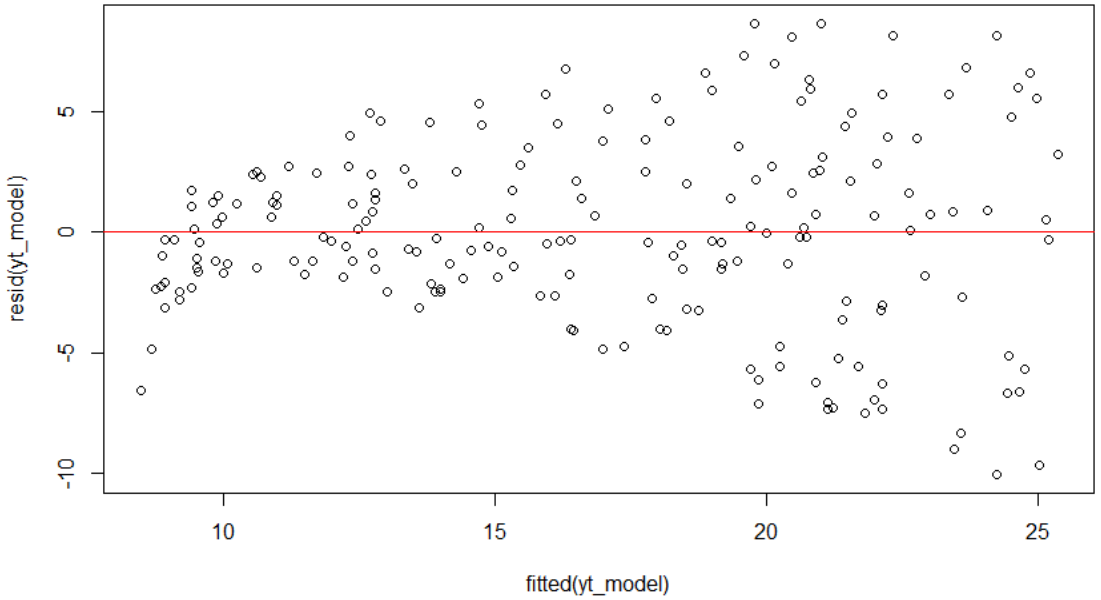
- (a) Use the marketing data to predict sales based on each of the advertising budget invested in YouTube, Facebook and newspaper by fitting a simple linear regression model. Verify whether the condition for fitting a regression model is met.
  - YouTube Model:



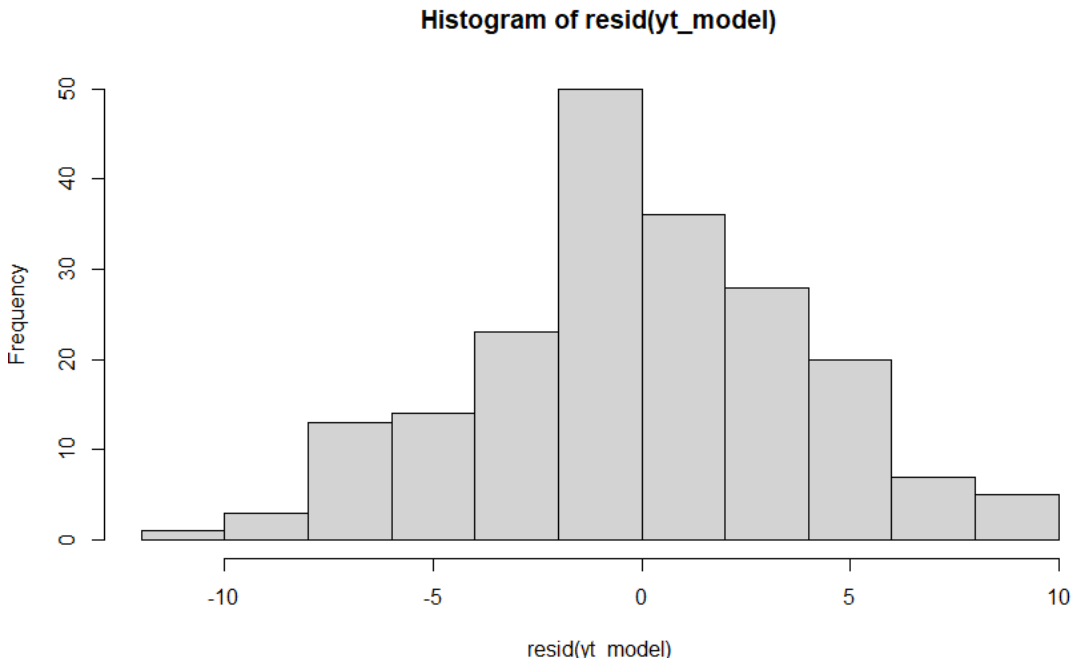
- The above plot shows the advertisement budget for YouTube vs. the sales numbers. We also see the regression line from our YouTube based linear regression model. The graph shows a linear relationship between the YouTube advertising budget and the sales numbers. This means the YouTube advertising budget data passes the first criteria to be valid for a linear regression model.

```
>
> mean(resid(yt_model))
[1] 7.007415e-17
>
```

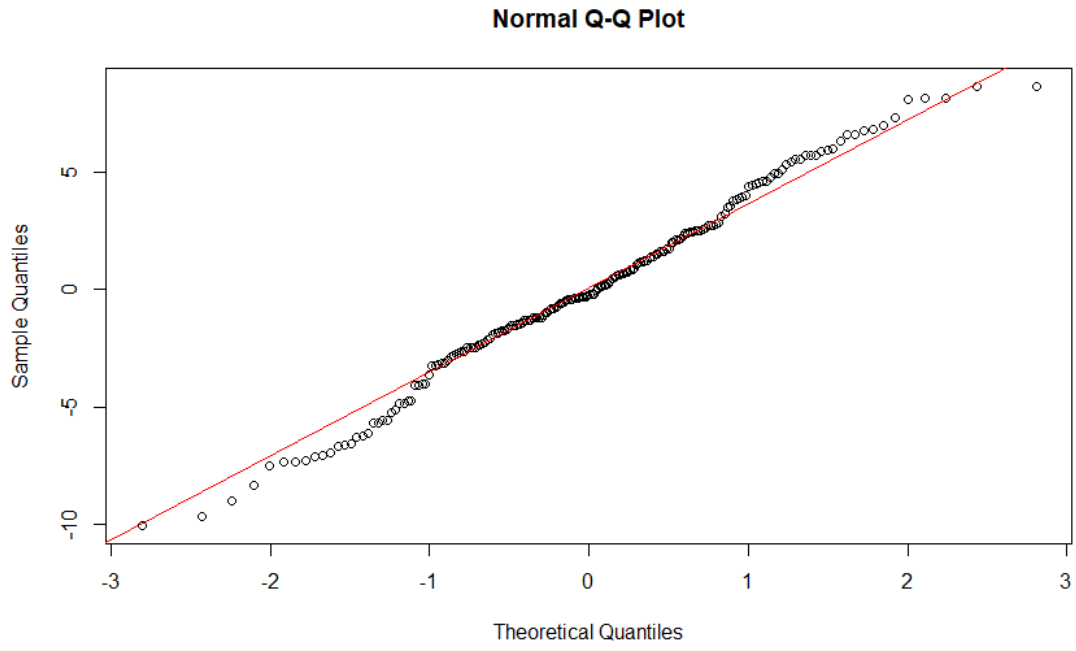
- The above output was provided after taking the average value of the residuals from our YouTube based model. We can see that the average value is approximately equal to 0. So, the YouTube advertising budget data passes the second criteria to be valid for a linear regression model.



- The above graph shows the Residual vs Fit plot for the YouTube based model. The plot shows that while the mean residual value is approximately 0, the variability of our residuals is not constant. As the prediction values calculated by our model increases, so does the variability of the residuals associated with those predictions. This means that the YouTube advertising budget data will not meet all the conditions needed to be considered valid to fit a linear regression model.

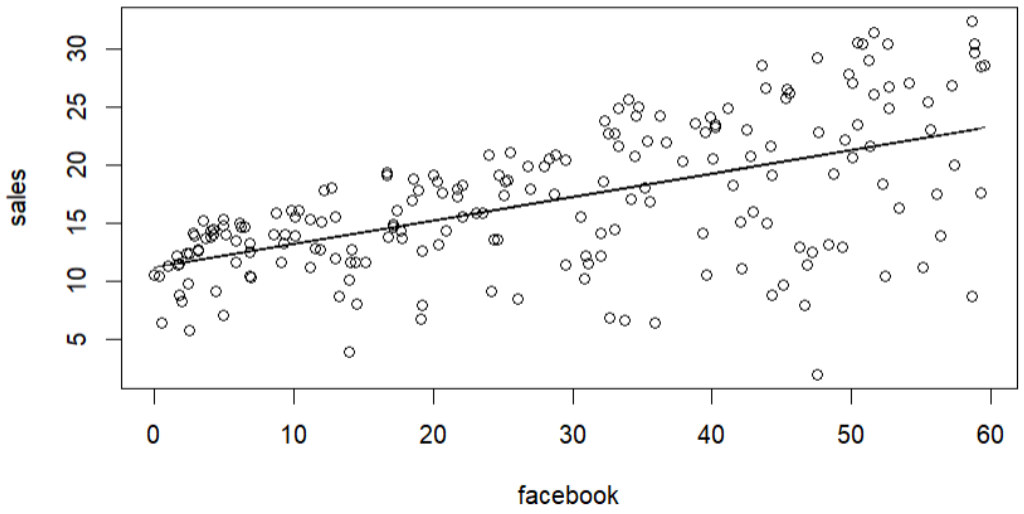


- The above graph shows a Histogram of our YouTube model's residuals. We can see from the graph that the mean residual value is approximately equal to 0 (as expected based on earlier findings). We can also see that the histogram looks like a normal distribution with a slight skew to the left. Since the skew in the normal distribution is minimal, I'd say the YouTube advertisement budget data passes the normality of the residuals criteria.



        - Above, we have our Quantile – Quantile plot based on the residuals of our YouTube model. We can use this QQ-plot in order to show abnormalities in the previously shown histogram of residuals. For this graph, we are ideally looking for the plot points to make a straight line at a 45-degree angle (represented as the red line). If so, then the distribution of our residuals should approximately be based on a normal distribution. As we can see from the graph, we can see that the data points do not make a perfect straight line. This is expected based on the histogram shown earlier and expected because of the previous Residual vs Fit plot results.
    - Based on the visualizations previously shown, using the YouTube advertising budget as a predictor seems promising, but unfortunately it doesn't pass all conditions in order for the data to be considered valid to fit a linear regression model. The one condition the model doesn't pass is that the variability in the residuals is not constant. However, I expect the data to be part of our final multi-linear regression model.

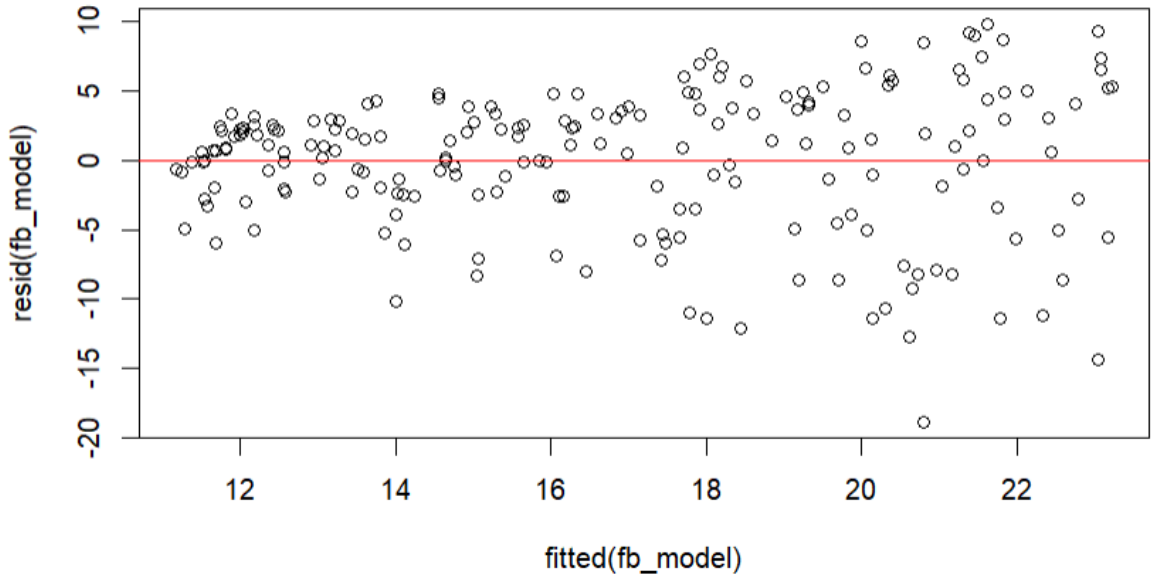
- Facebook model



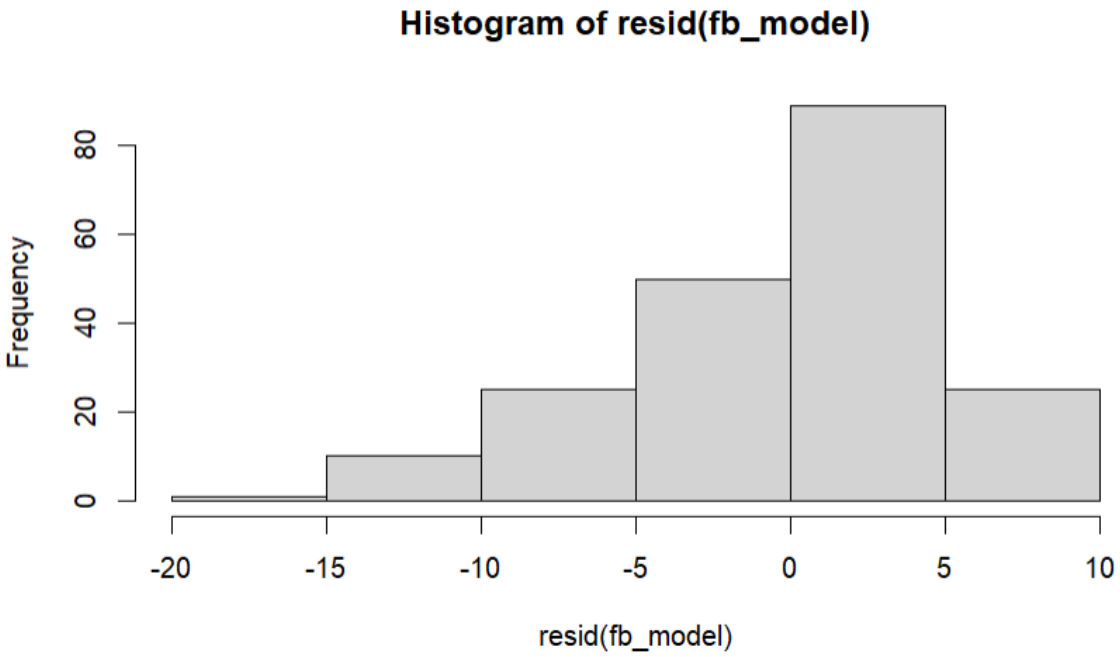
    - The above plot shows a scatter plot of the Facebook advertising budget vs. the sales numbers. Also shown is the regression line from our Facebook based linear regression model. The graph shows a linear relationship between the Facebook advertising budget and the sales numbers. However, the linear relationship isn't as strong as the YouTube based model. The Facebook advertising budget data passes the first criteria to be valid for a linear regression model.

```
> mean(resid(fb_model))
[1] 3.26128e-16
>
```

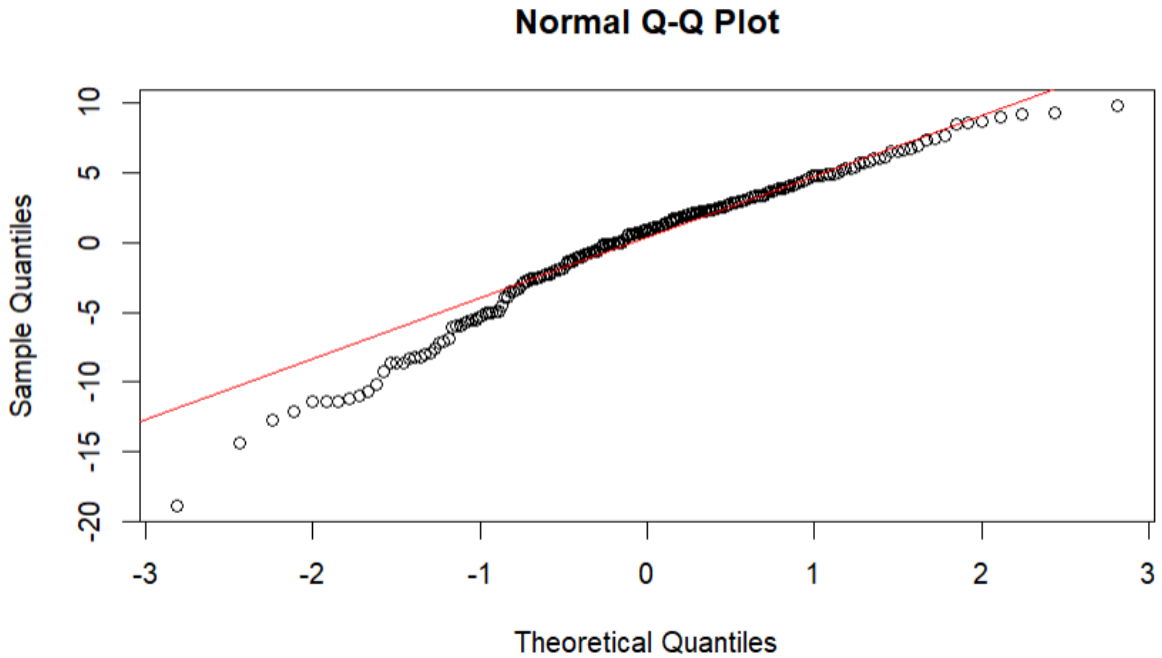
- I received the above output after calculating the average of the residuals from our Facebook based model. The output shows that the average residual value is approximately equal to 0. This means the Facebook advertisement budget data passes the second criteria to be considered valid for a linear regression model.



- The above graph shows the Residual vs fit plot for our Facebook based model. As with the same plot for the YouTube based model, the variability of the residuals is not constant. As the prediction values from our model increases, so does the variability of residuals associated with those predictions. This means that the Facebook advertising budget data will not pass all criteria to be considered valid to fit a linear regression model. I think its important to note that the change in variance for the residuals is a little better here than compared to the YouTube based model.



- The above graph shows a histogram of the residuals for our Facebook based model. We can see that the residuals are indeed centered around 0, but the graph is heavily skewed to the left. This further indicates that the Facebook related data is not valid for fitting a linear regression model.

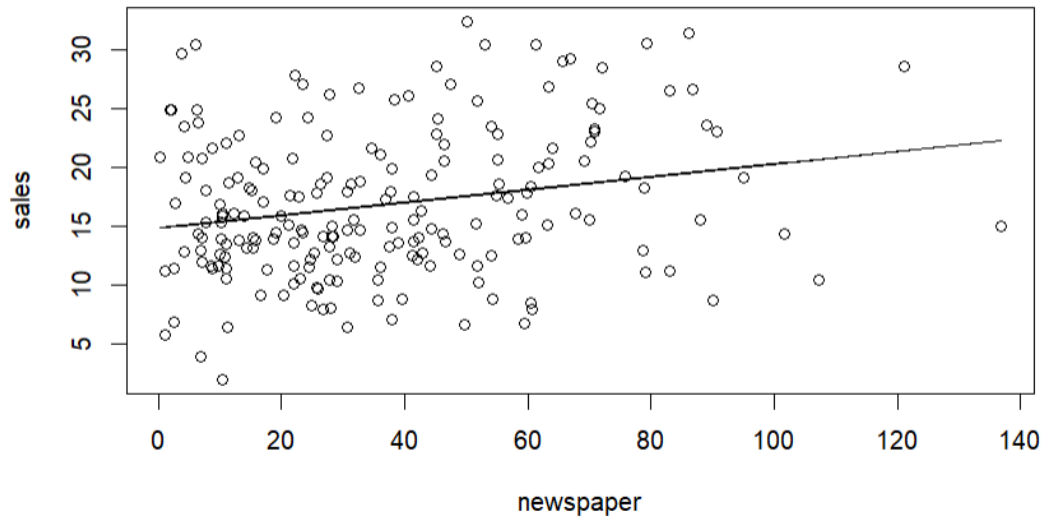


- The above graph shows the Quantile-Quantile plot for our Facebook based Linear Regression model. We can see that the plot doesn't line up with our 45-

degree line (represented in red) as would like to see. This is expected given the previous histogram of the residuals results.

- Based on the previous visualizations, I don't think that the Facebook advertisement budget data meets all the criteria to be valid for a linear regression model because the residuals of the model aren't constant, and the histogram of the residuals doesn't approximately make a normal distribution.

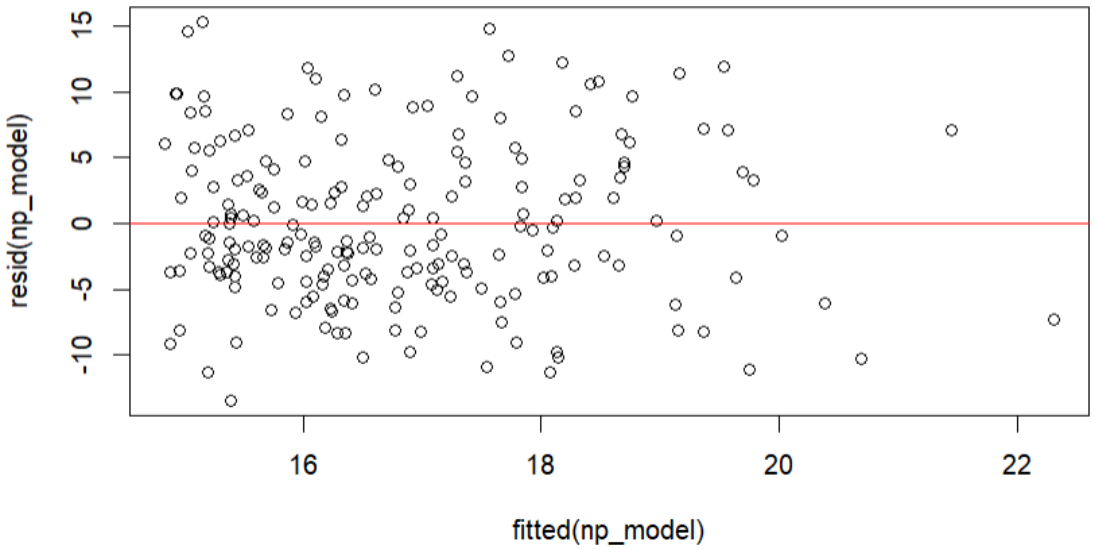
- Newspaper model



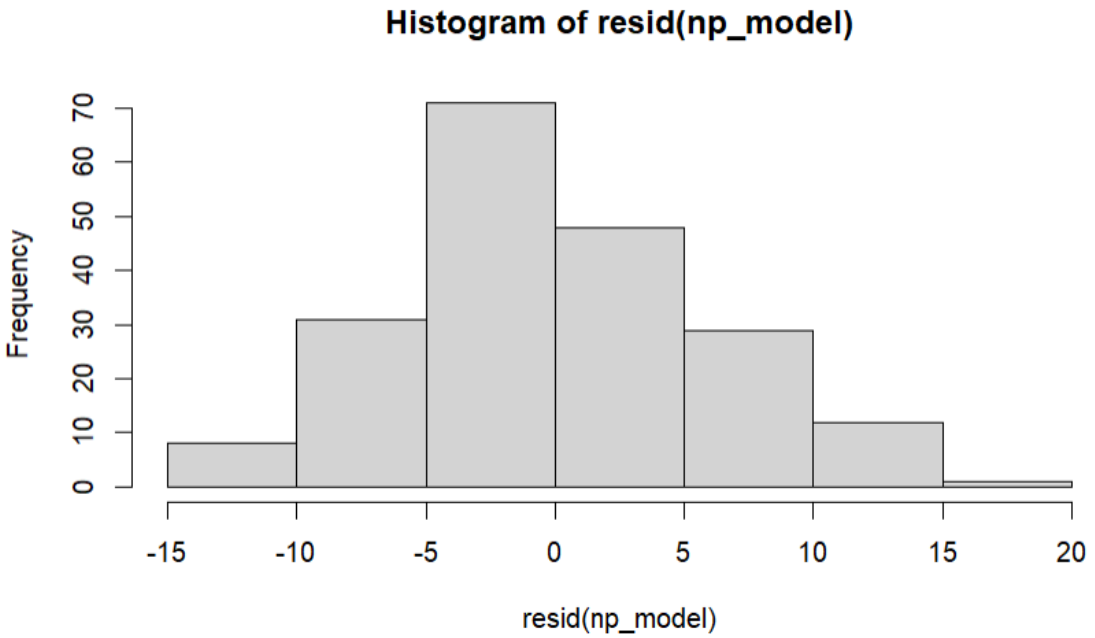
  - The above graph is a scatter plot of the Newspaper advertisement budget vs. the actual sale numbers. Also represented is the linear regression line from our linear regression model based on the Newspaper advertisement budget data. We can see from the graph that there really isn't a linear relationship between the newspaper advertising budget and the actual sales numbers. For this reason, the newspaper related data by itself will not be valid to fit a linear regression model.

```
> mean(resid(np_model))
[1] 9.23789e-17
>
```

  - I received the following output after calculating the average of the residual values for our Newspaper based model. As we can see, the average residual value for our Newspaper model is approximately equal to 0. This means the Newspaper data passes the second criteria to be considered valid for a linear regression model.



- Above we have the Residual vs Fit plot for our Newspaper based model. Not only does the graph show that the average residual value is approximately equal to 0, but it also shows that the residuals have a constant variance. This means that the Newspaper data passes the third criteria to be considered valid for a linear regression model.



- Above we have a histogram of the residuals associated with our Newspaper based model. Here we can see that the mean residual value is centered around 0 (as expected). We can also see that the distribution of the residuals looks like the normal distribution with a small skew to the right. This means that the

Newspaper data passes the final criteria to be considered valid for a linear regression model.

- Given the previous visualizations, the newspaper advertisement budget data doesn't pass all the conditions needed in order to be considered valid to fit a linear regression model because there is no linear relationship between the newspaper data and the actual sales numbers.

- (b) Use multiple regression model to predict sales based on the advertising budget invested in youtube, facebook and newspaper. What is your best model? Justify your answer.

```
>
> model_1 = lm(sales ~ youtube + facebook + newspaper)
> summary(model_1)

Call:
lm(formula = sales ~ youtube + facebook + newspaper)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5932  -1.0690   0.2902   1.4272   3.3951

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.526667   0.374290   9.422   <2e-16 ***
youtube      0.045765   0.001395  32.809   <2e-16 ***
facebook     0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.023 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

-

```
>
> model_2 = lm(sales ~ youtube + facebook)
> summary(model_2)

Call:
lm(formula = sales ~ youtube + facebook)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5572  -1.0502   0.2906   1.4049   3.3994

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.50532    0.35339   9.919   <2e-16 ***
youtube      0.04575    0.00139  32.909   <2e-16 ***
facebook     0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.018 on 197 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

-

- Above we can see 2 different outputs for 2 different linear regression models. The first is the summary of a multi-linear regression model using all 3 predictors. Here we can see we have a good, adjusted R-squared value of 0.8956 and we have a nice p-value that is

approximately equal to 0 for our entire model. However, not all of our predictors are significant. The p-value for the newspaper data is 0.86, which is not close to significant. For the next model, we are going to try removing the newspaper predictor.

- The second output shows a summary of the multi-linear regression model using just the YouTube and Facebook data for predictors. Here we can see our adjusted R-squared value increase a tiny bit to 0.8962, which is good. We can also see that our entire model has a p-value approximately equal to 0 and all of our predictors are significant with p-values also approximately equal to 0.
- Since we have a good model p-value, a good, adjusted R-squared, and all of the predictors are significant, I think the best multi-linear regression model for the sales numbers includes the YouTube and Facebook predictors.

Code Appendix:

- Problem 1:

```
### Problem 1 ####
### Part a ###
data <- read.csv("FootballBrain.csv", header = TRUE)
attach(data)

summary(Hipp[Group=="Control"])
summary(Hipp[Group=="FBNoConcuss"])
summary(Hipp[Group== "FBConcuss"])

### Part b ###
anova(lm(Hipp ~ Group))
```

- Problem 2b:

```
#Problem 2 Part b.
data <- read.csv("marketing.csv", header = TRUE)

attach(data)

model_1 = lm(sales ~ youtube + facebook + newspaper)
summary(model_1)

model_2 = lm(sales ~ youtube + facebook)
summary(model_2)
```

- Problem 2a:

```r
# Problem 2 part a.
data <- read.csv("marketing.csv", header = TRUE)
attach(data)
############ Youtube Model ############
yt_model <- lm(sales~youtube)
summary(yt_model)
mean(resid(yt_model))
# Scatter plot
plot(youtube, sales)
lines(youtube, fitted(yt_model))
#Residual vs fit plot
plot(fitted(yt_model), resid(yt_model))
abline( a = 0, b = 0, col = "red")
#histogram
hist(resid(yt_model))
#QQ-plot
qqnorm(resid(yt_model))
qqline(resid(yt_model), col = "red")

############ Facebook Model ############
fb_model <- lm(sales~facebook)
summary(fb_model)
mean(resid(fb_model))
#Scatter plot=
plot(facebook, sales)
lines(facebook, fitted(fb_model))=
#Residual vs fit plot
plot(fitted(fb_model), resid(fb_model))
abline(a=0, b=0, col="red")
#histogram
hist(resid(fb_model))
#QQ-plot
qqnorm(resid(fb_model))
qqline(resid(fb_model), col = "red")

############ Newspaper Model ############
np_model <- lm(sales~newspaper)
summary(np_model)
mean(resid(np_model))
#Scatter plot
plot(newspaper, sales)
lines(newspaper, fitted(np_model))
#Residual vs Fit plot
plot(fitted(np_model), resid(np_model))
abline(a=0, b=0, col = "red")
#histogram
hist(resid(np_model))
#QQ-plot
qqnorm(resid(np_model))
qqline(resid(np_model), col = "red")
```

-