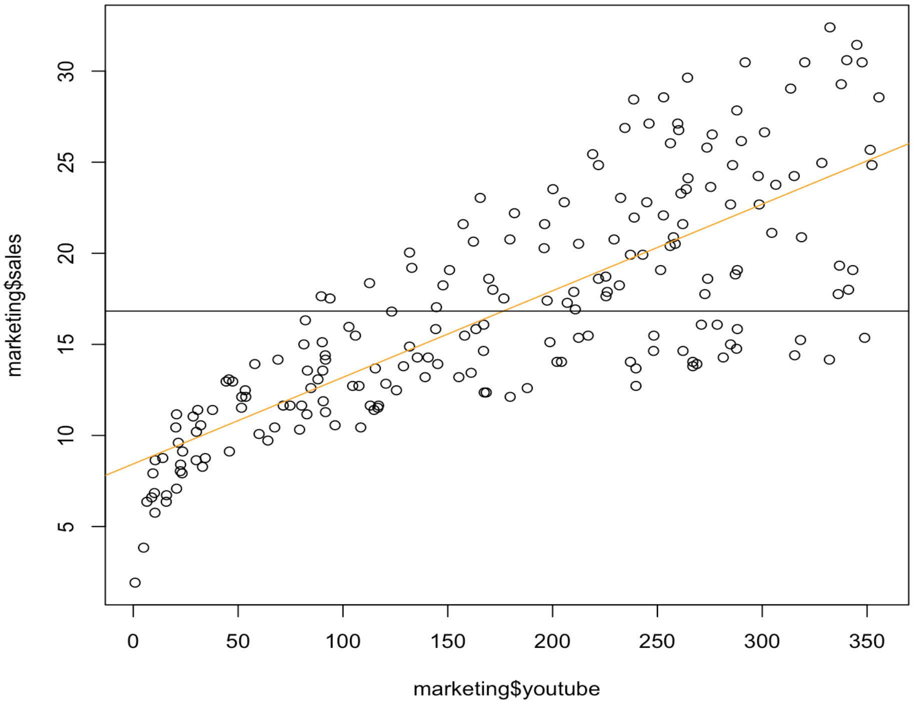
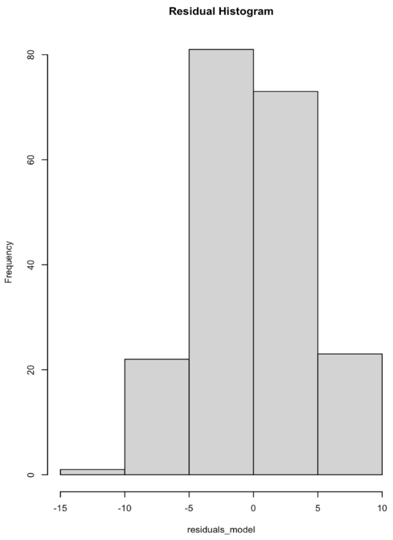
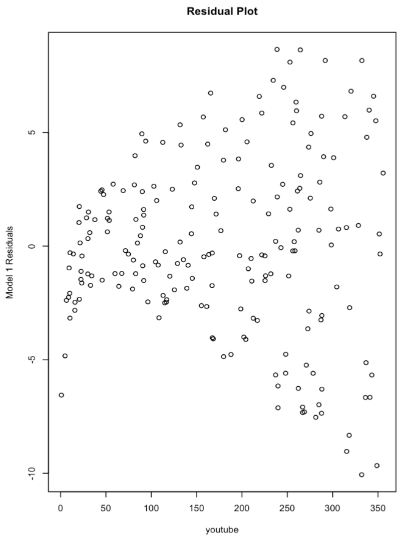
During the analysis of the data set "marketing" many conclusions were found, specifically on the predictors included in the set and sales prediction. The results of the findings include a simple linear regression model as well as a multiple linear regression model that can be used to accurately predict sales.

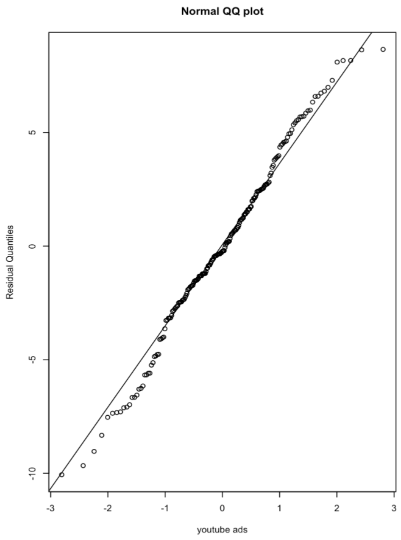
The most significant predictor was chosen to be used for the simple linear regression model. All three were tested, and the YouTube predictor was found to be the most significant. A model was made using each of the predictors so the accuracy of each model could be tested. Each model's adjusted R squared value and p-value was used to determine which model best fit the data. The model using Youtube as the predictor had the highest adjusted R squared value, which was 0.6099, and a p value of almost 0. Below is the scatter plot of Sales vs. YouTube Advertising:



According to the chosen model, the linear function for the expected value of sales is $\hat{y} = 8.439112 + 0.047537x$. That is the equation of the orange line present in the scatter plot. The intercept, 8.439112, means that without the predictor variable there is expected to be $8,439 in sales. As YouTube advertisements are increased by $1, there can be an expected $47 increase in sales. The regression standard error of the model is 0.3429913, meaning the difference between actual data points and ones predicted by the model is fairly small. The coefficient of determination value is 0.6118751, so about 61% of the variation in sales is able to be explained through the model, which indicates a relatively good fit. Because the slope value is positive and relatively significant, it can be concluded there is a positive relationship with moderate strength between the two variables. Below is the residual plot, histogram of residuals, and QQ plot:

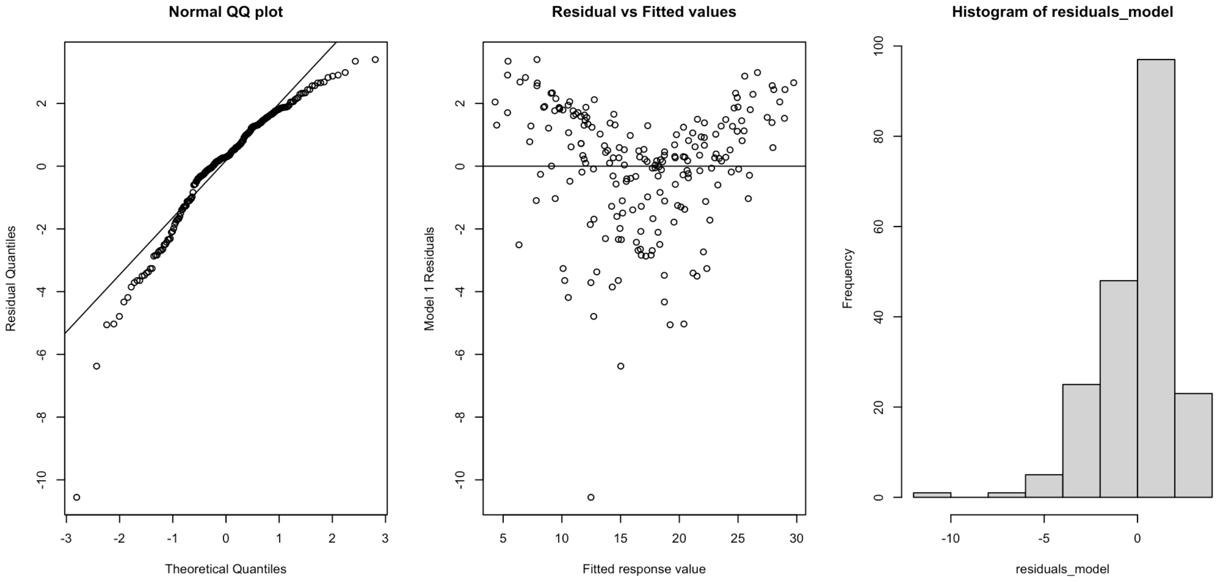






The residual plot and residual histogram do not suggest the model fits the data well at smaller values due to the clustering in the beginning of the plot. The normal QQ plot does suggest the residuals are roughly distributed aside from the beginning of the plot. These three models suggest that the model is more accurate for x values that are 200 and fits the model much less accurately in the lower bounds of the x values. Ultimately, they do still support that the model fits the majority of the data.

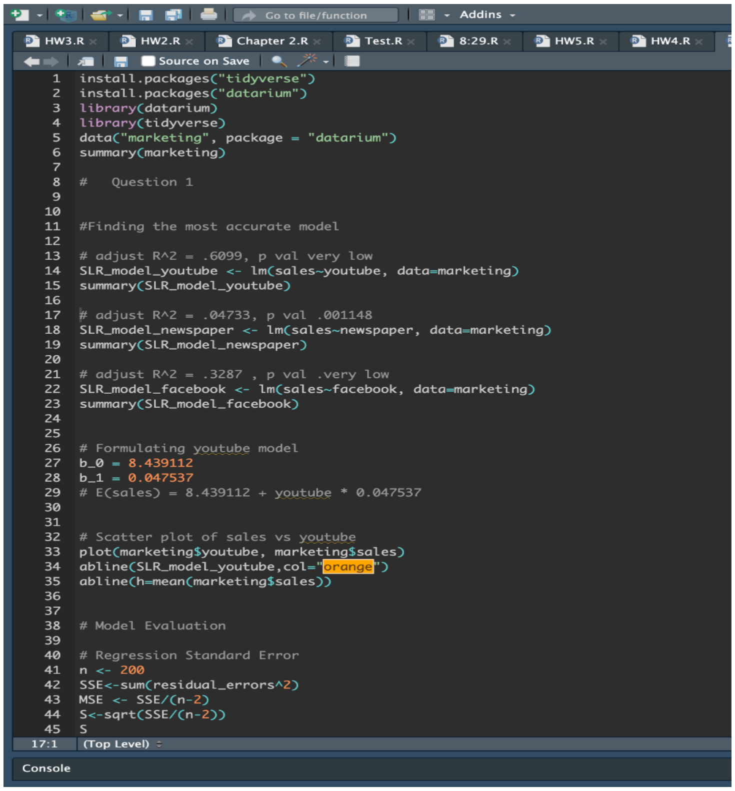
As for the multiple linear regression model, it was decided to drop the newspaper predictor. Several steps went into making that decision. First, a model was made with all three predictors. The model had an adjusted R squared value of 0.8956, so a large portion of the variability in sales is accounted for in the model. However, the newspaper predictor has a noticeably high p-value of 0.86 and its coefficient is almost 0 (-0.0010). This indicates that the predictor is not very significant to the model. YouTube and Facebook both have very low p-values, more significant coefficients, and three stars in the summary of the model. A model was created without the newspaper predictor and it had a higher adjusted R squared value, which was 0.8956. This value being higher suggests the model is a better fit of the data. The linear equation for the model is $E(sales) = 3.50532 + 0.04575(youtube) + 0.18799(facebook)$. So, without advertising on YouTube or Facebook $3,505 in sales can be expected. For every dollar spent on YouTube advertisements, sales are expected to increase $45.75. For every dollar spent on Facebook advertisement, sales are expected to increase about $188. The regression standard error of the model is 2.018, meaning that the observed sale values are roughly $2,018 away from the predicted values. Both of these values lend support to the model. Below are the plots used for model evaluation and residual analysis:



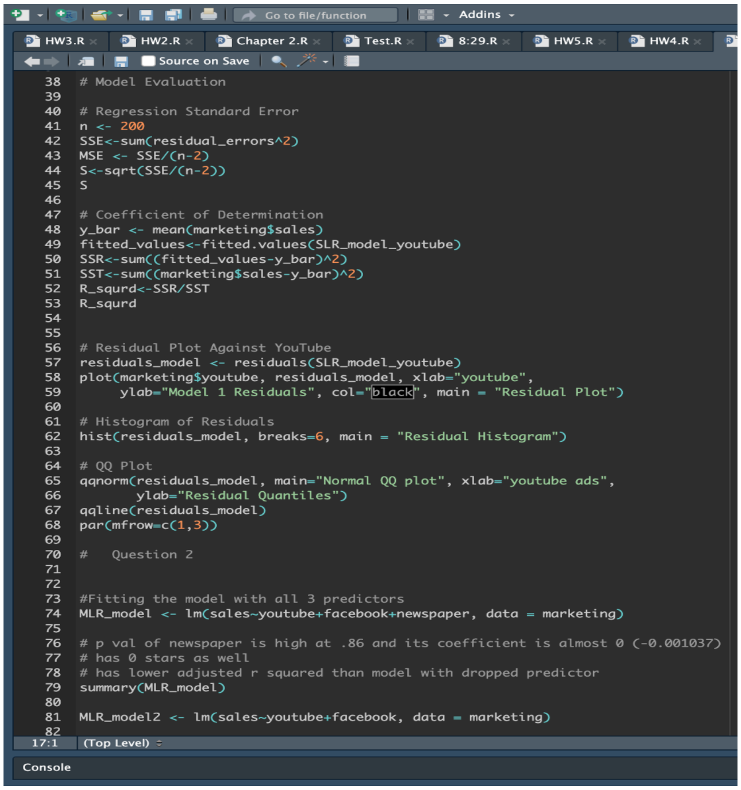
All of these plots ultimately support the model, but there are some flaws. The QQ plot indicates that the residuals are roughly normally distributed. The residual plot has some outliers and could have less clustering, but they are distributed random enough. This histogram is also a little skewed but is roughly normal.

Both the models fit the data well, but because the multiple linear regression model has a higher adjusted R squared and lower residual standard error it is reasonable to say it fits the data better. Using that model, we can be 95% confident that the expected sales are between $13,242 and $14,255 given $50.90 went to YouTube advertisements, $42.10 were used for advertisements on Facebook, and $87.60 were used to advertise in the newspaper. Likewise, we can be 95% confident that actual sales will be between $9,737 and $17,759 when there are $50.90 going toward YouTube advertisements, $42.10 are used for Facebook advertisements, and $87.60 are used to advertise in the newspaper.

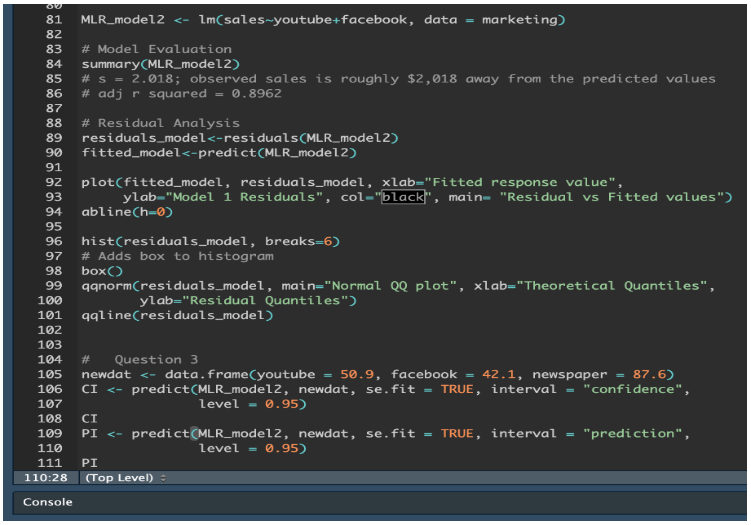
Appendix



```r
install.packages("tidyverse")
install.packages("datarium")
library(datarium)
library(tidyverse)
data("marketing", package = "datarium")
summary(marketing)

#   Question 1


#Finding the most accurate model

# adjust R^2 = .6099, p val very low
SLR_model_youtube <- lm(sales~youtube, data=marketing)
summary(SLR_model_youtube)

# adjust R^2 = .04733, p val .001148
SLR_model_newspaper <- lm(sales~newspaper, data=marketing)
summary(SLR_model_newspaper)

# adjust R^2 = .3287 , p val .very low
SLR_model_facebook <- lm(sales~facebook, data=marketing)
summary(SLR_model_facebook)


# Formulating youtube model
b_0 = 8.439112
b_1 = 0.047537
# E(sales) = 8.439112 + youtube * 0.047537


# Scatter plot of sales vs youtube
plot(marketing$youtube, marketing$sales)
abline(SLR_model_youtube,col="orange")
abline(h=mean(marketing$sales))


# Model Evaluation

# Regression Standard Error
n <- 200
SSE<-sum(residual_errors^2)
MSE <- SSE/(n-2)
S<-sqrt(SSE/(n-2))
S
```



```r
# Model Evaluation

# Regression Standard Error
n <- 200
SSE<-sum(residual_errors^2)
MSE <- SSE/(n-2)
S<-sqrt(SSE/(n-2))
S

# Coefficient of Determination
y_bar <- mean(marketing$sales)
fitted_values<-fitted.values(SLR_model_youtube)
SSR<-sum((fitted_values-y_bar)^2)
SST<-sum((marketing$sales-y_bar)^2)
R_squrd<-SSR/SST
R_squrd


# Residual Plot Against YouTube
residuals_model <- residuals(SLR_model_youtube)
plot(marketing$youtube, residuals_model, xlab="youtube",
     ylab="Model 1 Residuals", col="black", main = "Residual Plot")

# Histogram of Residuals
hist(residuals_model, breaks=6, main = "Residual Histogram")

# QQ Plot
qqnorm(residuals_model, main="Normal QQ plot", xlab="youtube ads",
       ylab="Residual Quantiles")
qqline(residuals_model)
par(mfrow=c(1,3))

#   Question 2


#Fitting the model with all 3 predictors
MLR_model <- lm(sales~youtube+facebook+newspaper, data = marketing)

# p val of newspaper is high at .86 and its coefficient is almost 0 (-0.001037)
# has 0 stars as well
# has lower adjusted r squared than model with dropped predictor
summary(MLR_model)

MLR_model2 <- lm(sales~youtube+facebook, data = marketing)
```



```r
MLR_model2 <- lm(sales~youtube+facebook, data = marketing)

# Model Evaluation
summary(MLR_model2)
# s = 2.018; observed sales is roughly $2,018 away from the predicted values
# adj r squared = 0.8962

# Residual Analysis
residuals_model<-residuals(MLR_model2)
fitted_model<-predict(MLR_model2)

plot(fitted_model, residuals_model, xlab="Fitted response value",
     ylab="Model 1 Residuals", col="black", main= "Residual vs Fitted values")
abline(h=0)

hist(residuals_model, breaks=6)
# Adds box to histogram
box()
qqnorm(residuals_model, main="Normal QQ plot", xlab="Theoretical Quantiles",
       ylab="Residual Quantiles")
qqline(residuals_model)


#   Question 3
newdat <- data.frame(youtube = 50.9, facebook = 42.1, newspaper = 87.6)
CI <- predict(MLR_model2, newdat, se.fit = TRUE, interval = "confidence",
              level = 0.95)
CI
PI <- predict(MLR_model2, newdat, se.fit = TRUE, interval = "prediction",
              level = 0.95)
PI
```

This is my own unaided work

Daria Casey