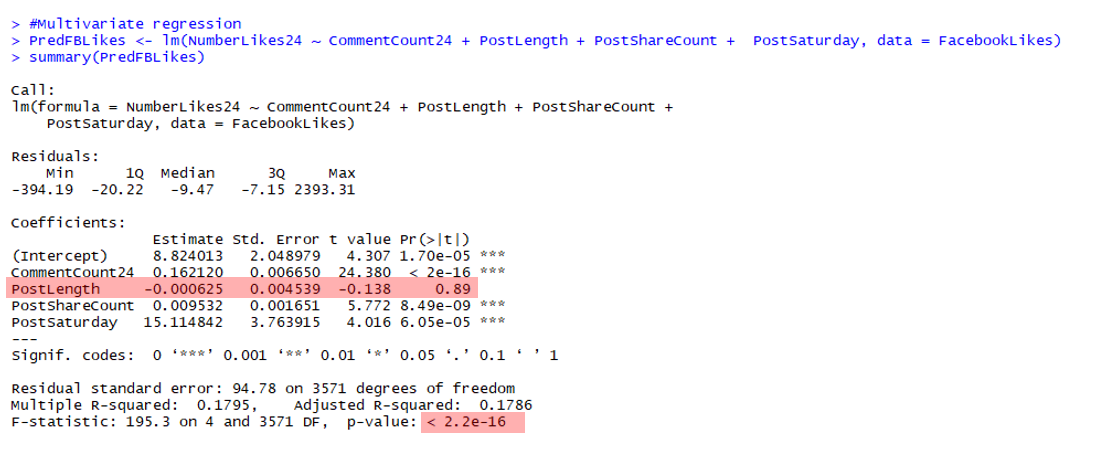
**Week 4 Regression Homework Answer Key**

1. Regression analysis can only be used for prediction. *FALSE*
2. Residuals are the difference between the observed values of y and the fitted values of y. *TRUE*

Using the Facebook\_Post\_Prediction\_Truncated.xlsx dataset, create a multivariate linear regression model using the post comment counts, post length, how many times the post has been shared, and if the post was made on Saturday as compared to the rest of the week to predict how many likes the Facebook post will receive in the first 24 hours.

\*\*\*Please note – there is NO cleaning done with this example (it’s really not correct to do this, but to save on the time spent on HW we skip this. If there is cleaning and processing answers obviously will differ)\*\*\*\*



1. What is the overall model p-value? *p<2.2e-16*
2. Is the model statistically significant at the p<0.05 level?  *YES*
3. Are there any variables in the model that are not statistically significant with the outcome after adjusting for the other variables? If so, which ones? *Yes. Post length is not statistically significant (p=0.89)*
4. Write a sentence that interprets the beta coefficient for the number of times a post has been shared.

*We estimate that there is an increase of 0.01 likes for every additional share of the post after adjusting for all other variables in the model.*

*OR (if you scale it)*

*We estimate that there is an increase of 1likes for every additional 100 shares of the post after adjusting for all other variables in the model.*

1. Write a sentence that interprets the beta coefficient for the post being made on a Saturday.

*We estimate that a post made on Saturday will result in 15 additional “likes” in the first 24 hours as compared to other days of the week after adjusting for other variables in the model.*

1. Given that a post has 10 comments in the first 24 hours, was 148 characters long, was shared by 3 people and was made on a Friday, how many likes can we expect that post to have in the first 24 hours?

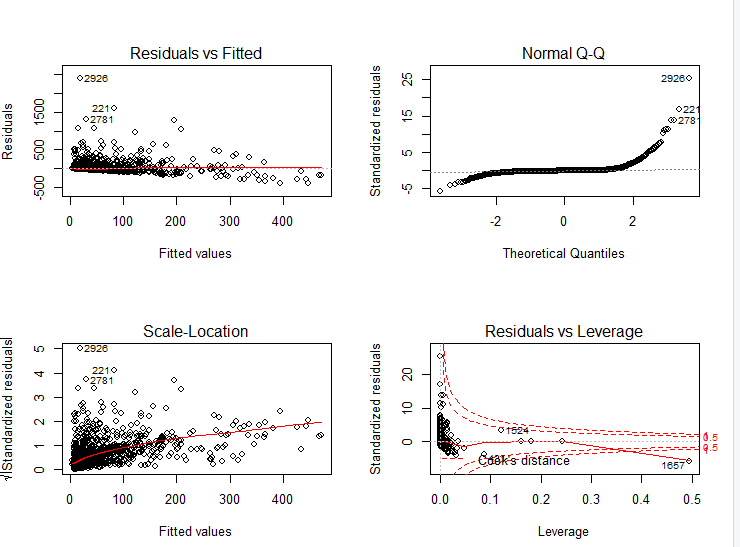
*MODEL: Predicted Likes in First 24 Hours = 8.824013 + (0.16212 \* Comment Count) + (-0.000625 \* Post Length) + (0.009532 \* Post Share Count) + (15.114841 \* Post made on Saturday)*



*We expected approximately 10 “likes” in the first 24 hours.*

1. Run the diagnostic tests we discussed in class on the above multivariate model. Explain your findings.





***Residuals vs Fitted*** *– There appears to be some indication of lack of fit near the lower values of “Facebook Likes in the first 24 hours”. There seems to be a pattern. Also, there is not equal number of residuals above and below the fitted line (most seem to be above). I would question the fit of this model.*

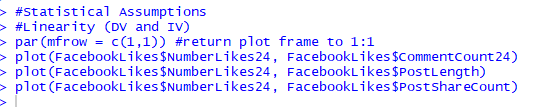
***Normal Q-Q*** *– The residuals do not appear to be normally distributed. There are issues on the tails.*

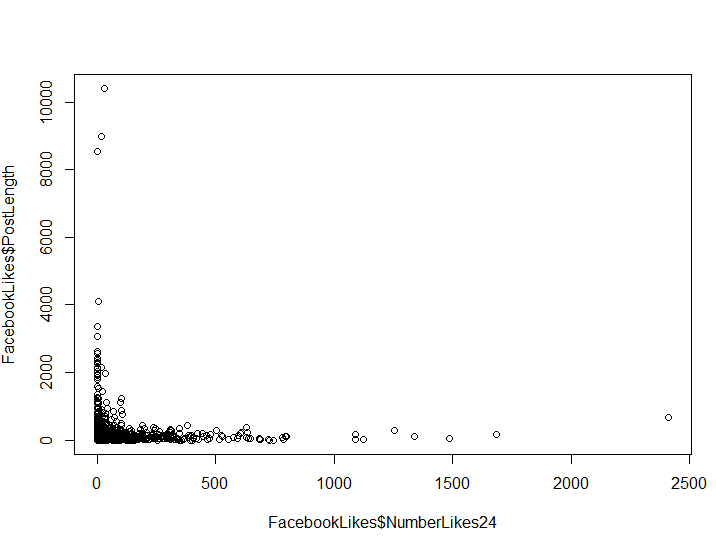
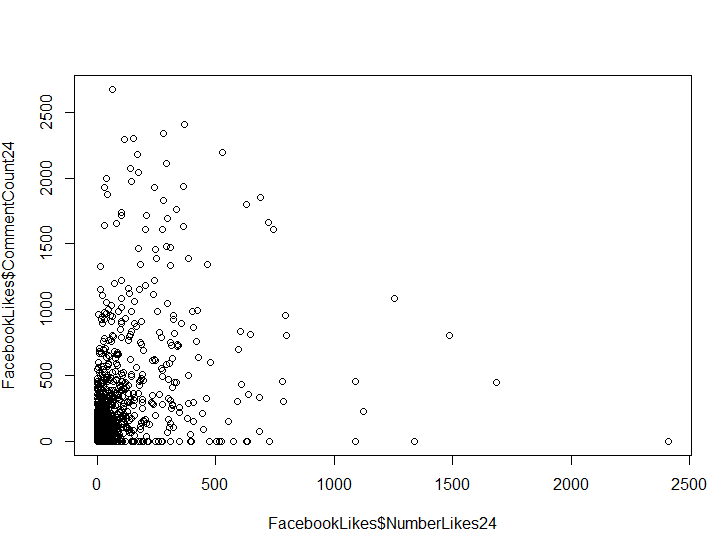
***Standardized Residuals vs Fitted (Scale-Location)*** *– same interpretation as Residuals vs Fitted*

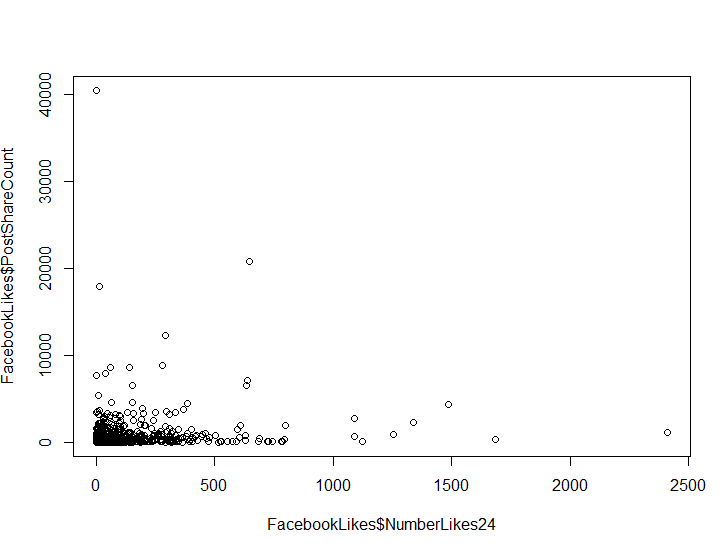
***Residuals vs Leverage*** *– There is one data value (1657) that is outside of the Cook’s Distance line. This indicates this data point has high leverage and may be influencing the model more than the other data values. In fact, upon further investigation, this data observation has 0 comments, 9 characters long, was share 10,428 times but has 0 comments. This seems odd and may be a data error. But it is adversely affecting our model. I would consider removing and running the model without this value.*

1. Test to see if the statistical assumptions for a linear model have been met. Explain your findings.

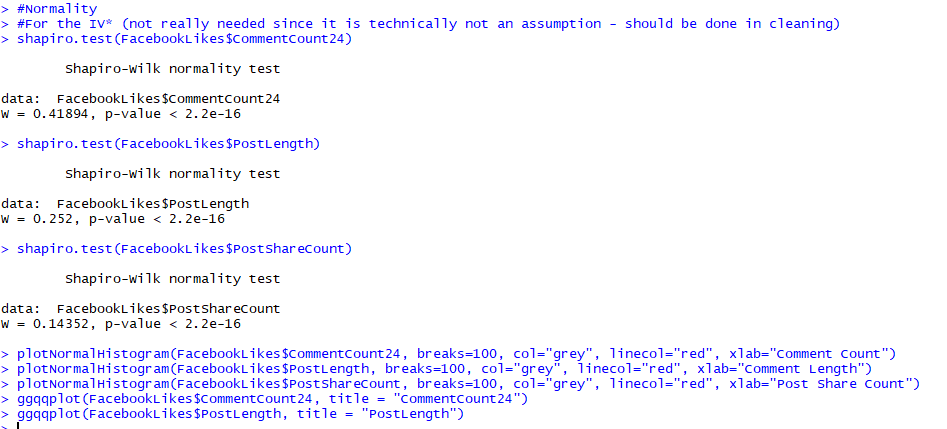
***Linearity*** *– There does not appear to be much of a linear relationship between the IVs and DV. This may be primarily due to the large number of zero values masking the relationship.*

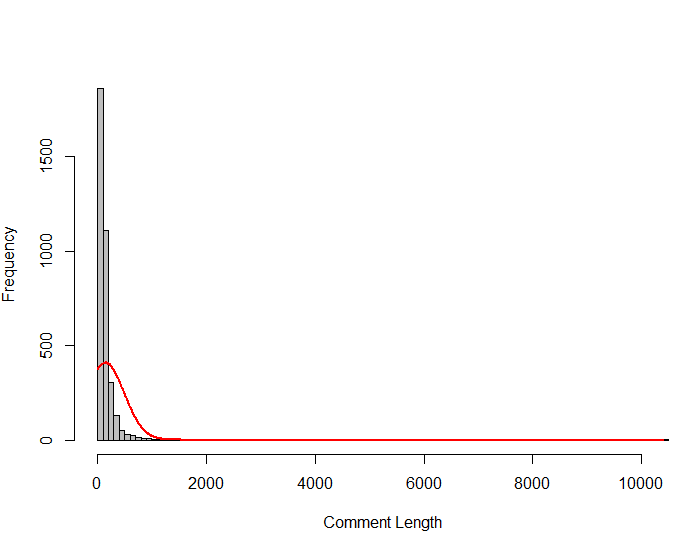
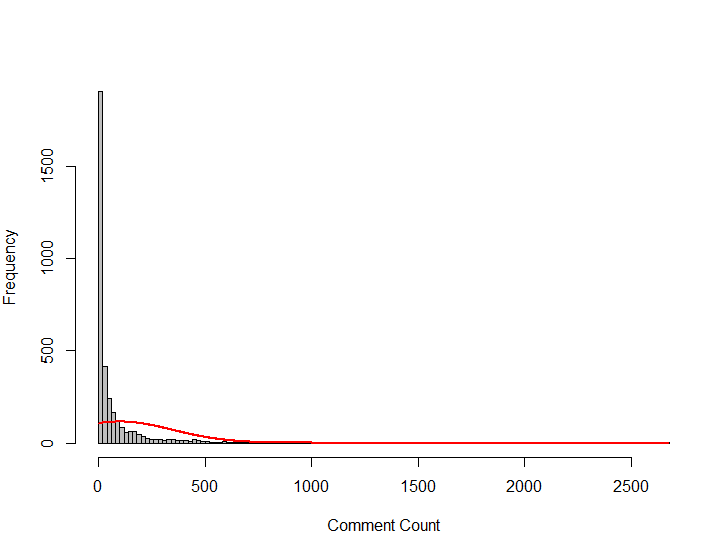


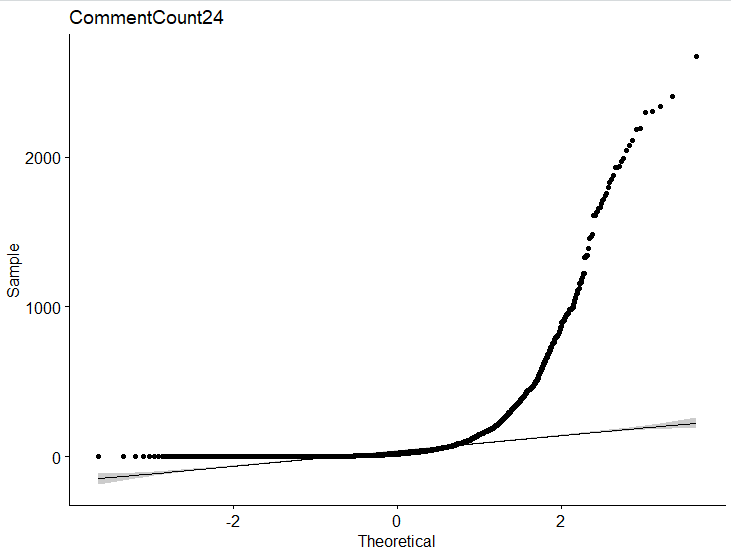
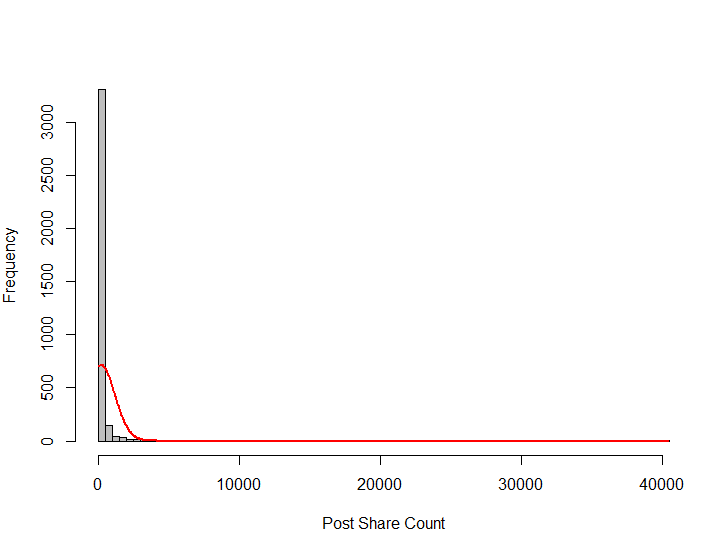


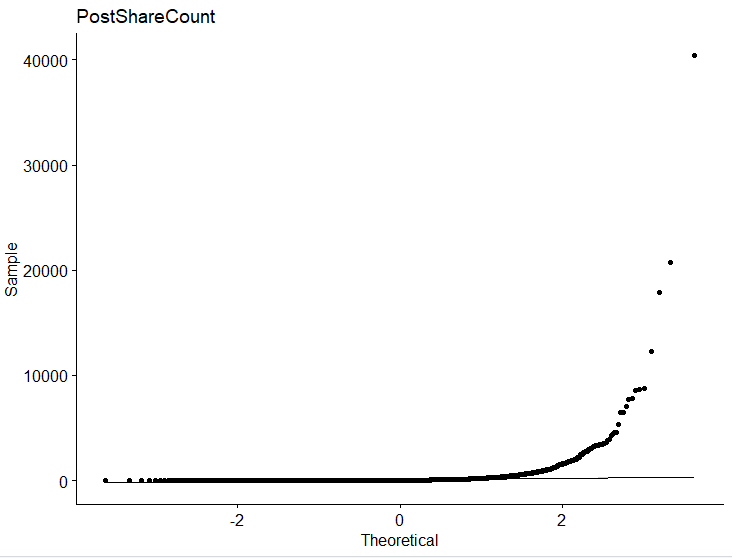
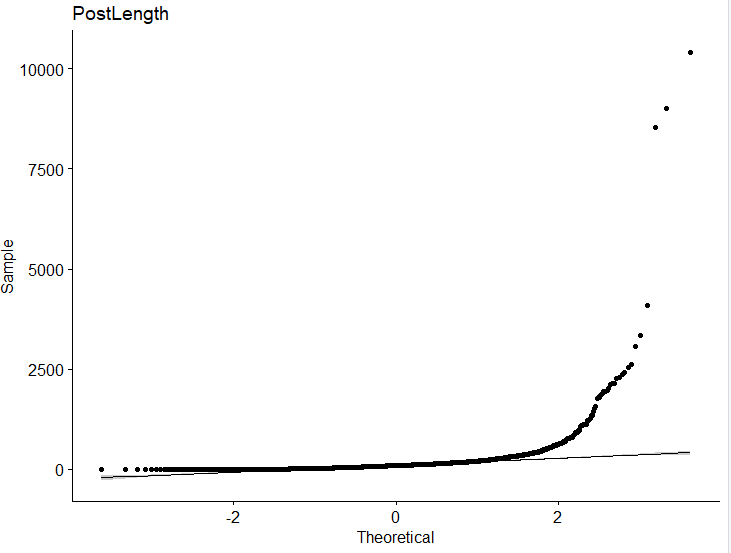


***Normality (for IV) –*** *While normality of the IV is not an assumption of linear regression, it does help to ensure the model works appropriate in the face of huge violations of this assumption. In the case of the 3 IV from our model, we can see they are highly skewed. This is due to the large amount of zero values. These variables should have been transformed prior to being put in the model.*

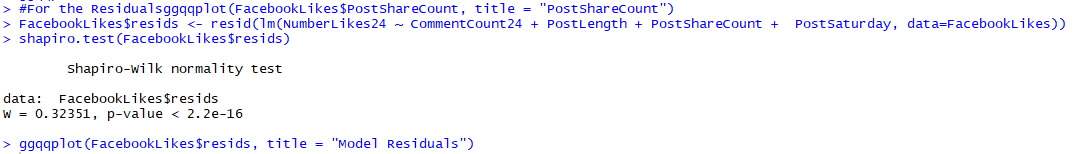


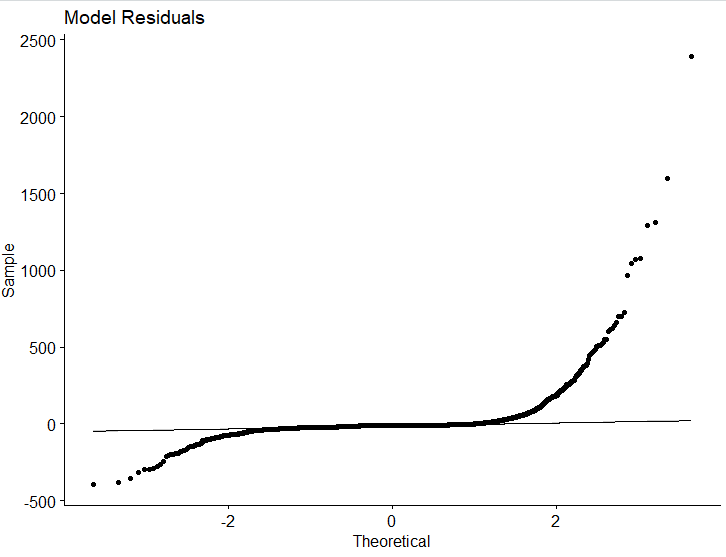




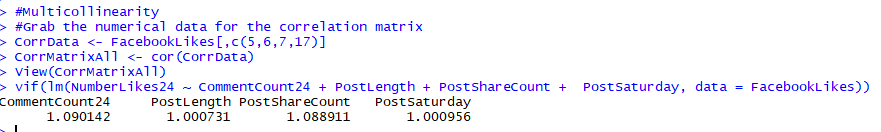


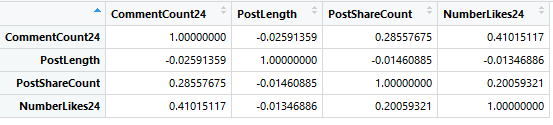
***Normality (Residuals)*** *– The residuals are also not normally distributed (but we knew this already from our diagnostic test. This may be compounded by our highly skewed IVs. I would proceed with caution as this model appears to have lots of problems.*



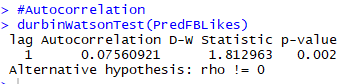


***Multicollinearity*** *– There appear to be no issues with multicollinearity. Our VIFs are low (~1) and the correlation matrix does not show any highly correlated variables.*

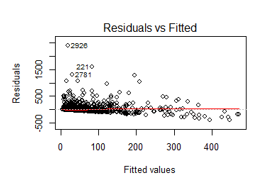




***Autocorrelation*** *– Although the p-value indicates there is a significant autocorrelation in our sample (deviation from test statistic of 2.0), we are ok with accepting this since the statistic (1.81) is within our acceptable range of 1.5 – 2.5. Therefore, we have met this assumption.*

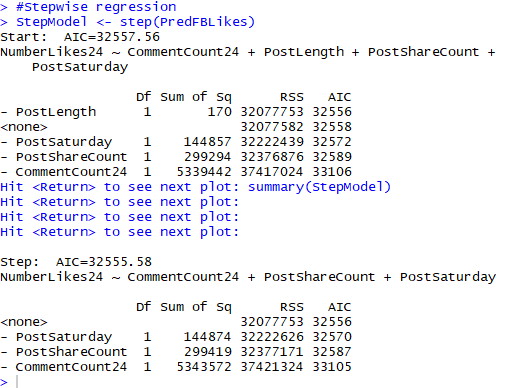


***Homoscedasticity*** *– Based on our Residuals vs Fitted plot we ran during our diagnostics, we do not appear to have any issues with this assumption as the residual values do NOT flare out as the fitted values increase.*



1. Perform stepwise regression with the above model to determine the optimal combination of variables to predict the number of Facebook post likes in the first 24 hours. What model is the best fit?

*Based on the stepwise regression, the best model looks to be the model that includes Post Share Count, Comment Count in the first 24 hours, and Posting on Saturday. It has removed the Post Length variable as it did not find this variable important to the model (based on AIC).*



1. Provide the code you used to run the analysis above.

#Load library

#REMINDER - packages must be installed first

library(xlsx)

library(ggplot2)

library(rcompanion)

library(ggpubr)

library(car)

#Create vector with File Path information

RegHWPath <-"C:/Users/nyzw/OneDrive - Chevron/DSDP/DSDP 2019/Lectures/Week 4 Regression/Facebook\_Post\_Prediction\_Truncated.xlsx"

#Import in file

FacebookLikes <- read.xlsx(RegHWPath,1)

#Look at data - visually look!

View(FacebookLikes)

#Look at the dataset dimensions

dim(FacebookLikes)

str(FacebookLikes)

#Multivariate regression

PredFBLikes <- lm(NumberLikes24 ~ CommentCount24 + PostLength + PostShareCount + PostSaturday, data = FacebookLikes)

summary(PredFBLikes)

#Model Diagnostics

par(mfrow = c(2,2))

PredFBLikes <- lm(NumberLikes24 ~ CommentCount24 + PostLength + PostShareCount + PostSaturday, data = FacebookLikes, plot(PredFBLikes))

#Statistical Assumptions

#Linearity (DV and IV)

par(mfrow = c(1,1)) #return plot frame to 1:1

plot(FacebookLikes$NumberLikes24, FacebookLikes$CommentCount24)

plot(FacebookLikes$NumberLikes24, FacebookLikes$PostLength)

plot(FacebookLikes$NumberLikes24, FacebookLikes$PostShareCount)

#Normality

#For the IV\* (not really needed since it is technically not an assumption - should be done in cleaning)

shapiro.test(FacebookLikes$CommentCount24)

shapiro.test(FacebookLikes$PostLength)

shapiro.test(FacebookLikes$PostShareCount)

plotNormalHistogram(FacebookLikes$CommentCount24, breaks=100, col="grey", linecol="red", xlab="Comment Count")

plotNormalHistogram(FacebookLikes$PostLength, breaks=100, col="grey", linecol="red", xlab="Comment Length")

plotNormalHistogram(FacebookLikes$PostShareCount, breaks=100, col="grey", linecol="red", xlab="Post Share Count")

ggqqplot(FacebookLikes$CommentCount24, title = "CommentCount24")

ggqqplot(FacebookLikes$PostLength, title = "PostLength")

ggqqplot(FacebookLikes$PostShareCount, title = "PostShareCount")

#For the Residuals

FacebookLikes$resids <- resid(lm(NumberLikes24 ~ CommentCount24 + PostLength + PostShareCount + PostSaturday, data=FacebookLikes))

shapiro.test(FacebookLikes$resids)

ggqqplot(FacebookLikes$resids, title = "Model Residuals")

#Multicollinearity

#Grab the numerical data for the correlation matrix

CorrData <- FacebookLikes[,c(5,6,7,17)]

CorrMatrixAll <- cor(CorrData)

View(CorrMatrixAll)

vif(lm(NumberLikes24 ~ CommentCount24 + PostLength + PostShareCount + PostSaturday, data = FacebookLikes))

#Autocorrelation

durbinWatsonTest(PredFBLikes)

#Stepwise regression

StepModel <- step(PredFBLikes)

summary(StepModel)