Homework 10: Multiple Linear Regression

NAME: Michael Darfler

NETID: mbd25

DUE DATE: December 11, 2019 by 11:59pm

Instructions

For this homework:

1. All calculations must be done within your document in code chunks. Provide all intermediate steps.
2. DO NOT JUST INCLUDE A CALCULATION: Incude any formulas you are using for a calculation. You can put these immediately before the code chunk where you actually do the calculation.

Hollywood Movies 2011 Dataset

This dataset includes information for 118 movies released in 2011. Here is a brief description of each of the variables included in this dataset.

Variable	Description
`WorldGross`	Gross income for all viewers (in millions)
`AudienceScore`	Audience Rating
`BOAveOpenWeek`	Average box office income per theater in the opening week
`Budget`	Production Budget (in millions)
`Fantasy`	TRUE if the movie genre is Fantasy; FALSE if the movie genre is not Fantasy

Problem 1

Here we will explore a MLR with response equal to `WorldGross` and the following predictors: `AudienceScore`, `BOAveOpenWeek`, `Fantasy`, and `Budget`.

a. Read the data into this homework document and list the variable names.

```
hollywood = read.csv("https://raw.githubusercontent.com/mdarfler/BTRY_6010/master
/Homework/HW%2010/Hollywood(3).csv")

names(hollywood)
```

```
## [1] "AudienceScore" "BOAveOpenWeek" "WorldGross"    "Budget"
## [5] "Fantasy"
```

b. Fit a linear model with `WorldGross` as the response and `AudienceScore`, `BOAveOpenWeek`, `Fantasy`, and `Budget` as predictors. Also, include a summary of this model.

```
hollywood.lm <- lm(WorldGross ~ AudienceScore + BOAveOpenWeek + Fantasy + Budget,
data = hollywood)

summary(hollywood.lm)
```

```
##
## Call:
## lm(formula = WorldGross ~ AudienceScore + BOAveOpenWeek + Fantasy +
##     Budget, data = hollywood)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -337.46  -53.57   -6.59   48.22  533.89
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.263e+02  4.525e+01  -2.791  0.00616 **
## AudienceScore  1.679e+00  7.129e-01   2.356  0.02020 *
## BOAveOpenWeek  3.546e-03  1.220e-03   2.907  0.00439 **
## FantasyTRUE    8.306e+02  1.319e+02   6.296 6.01e-09 ***
## Budget         2.658e+00  2.432e-01  10.929  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 125.4 on 113 degrees of freedom
## Multiple R-squared:  0.6924, Adjusted R-squared:  0.6815
## F-statistic: 63.59 on 4 and 113 DF,  p-value: < 2.2e-16
```

c. State the expression for the estimated expected value of `WorldGross` using the model fit in (b).

$\hat{E}(Y|\text{ predictors}) =$ -126.322 + 1.679 AudienceScore + 0.004 BOAveOpenWeek + 830.596 FantasyTRUE + 2.658 Budget

d. What values can the covariate `FantasyTRUE` take on, and what is the meaning of each possible

value?

> `FantasyTRUE` can take on the values of `0` and `1`. If the value is `0` then the movie is **not** a fantsay film. If the value is `1` then it is a fantsay film.

e. Estimate the expected gross income for a non-fantasy movie that has an audience score equal to 90, a budget of 50 million dollars, and that has an opening week box office average of $10,000.

```
prediction <- predict(hollywood.lm, list(Fantasy = FALSE, AudienceScore =90, Budg
et = 50, BOAveOpenWeek = 10000))
```

> Given the predictors from above, the espected gross income for the film would be $193,165,667

f. Use the `confint()` function to create a 95% confidence interval for the partial slope of `Budget`. Interpret it in the context of this study.

```
ci <- confint(hollywood.lm, level = 0.95)
```

> We are 95% confident that the partial slope for budget between 2.176 and 3.139.

Problem 2

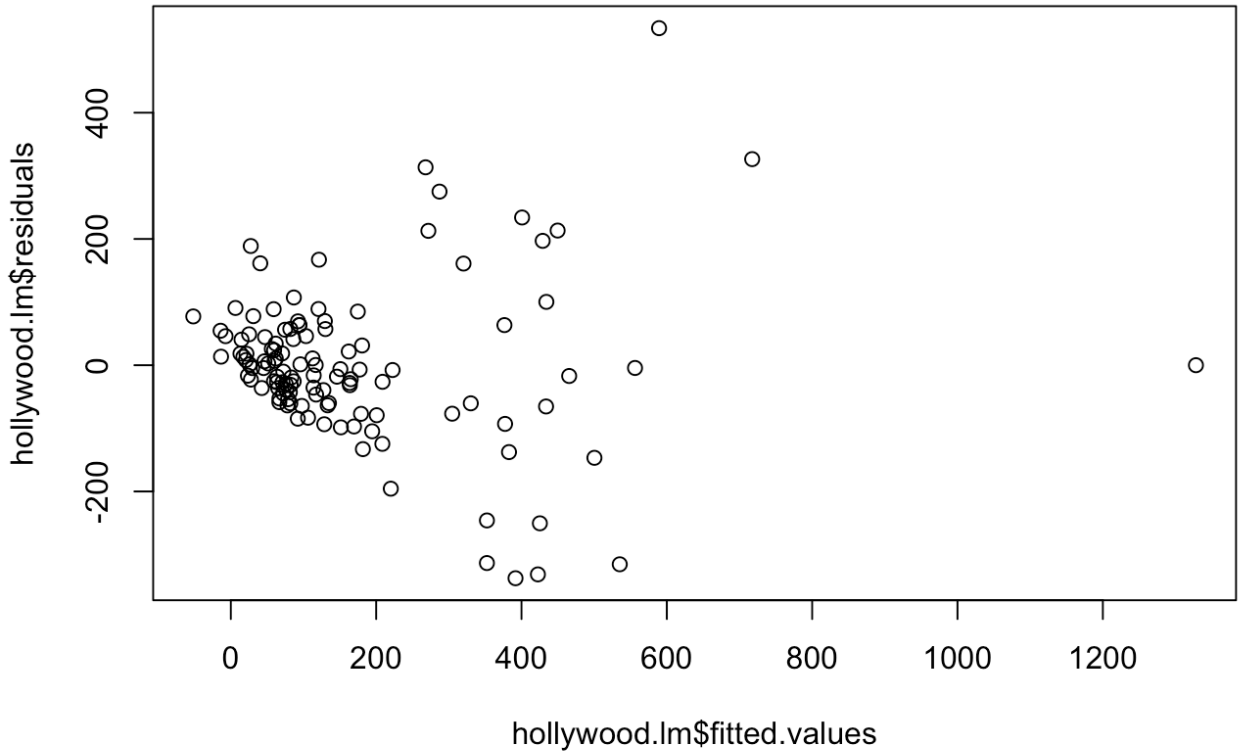
Here we will check the assumptions of the MLR fit in Problem 1.

a. Does it seem reasonable to assume these observations are independent?

> It does seem reasonable that these data are independent because each film is released on its

b. Create a scatterplot of the residuals (on the y-axis) vesus the fitted values (on the x-axis). Does the equal variance assumption seem reasonable?

```
plot(hollywood.lm$fitted.values, hollywood.lm$residuals)
```

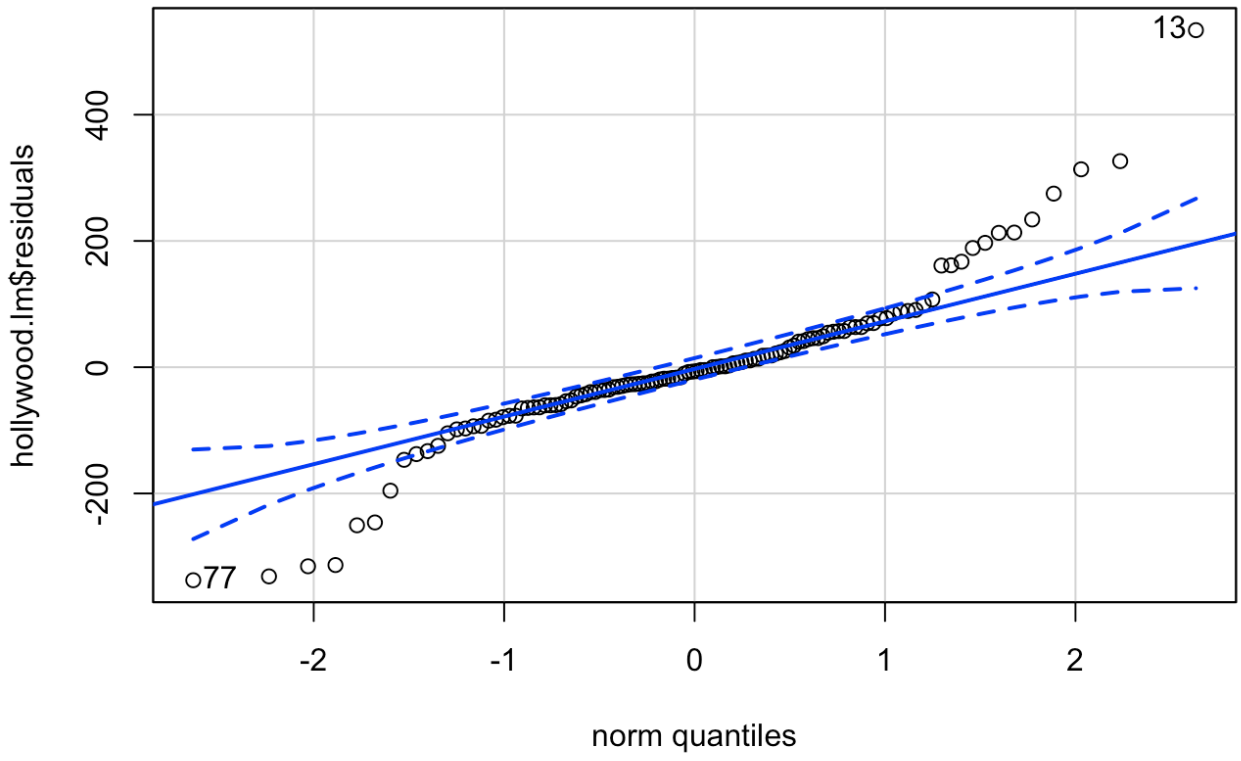


c. Create a Q-Q plot of the residuals. Does the normality assumption seem reasonable?

```
library(car)
```

```
## Loading required package: carData
```

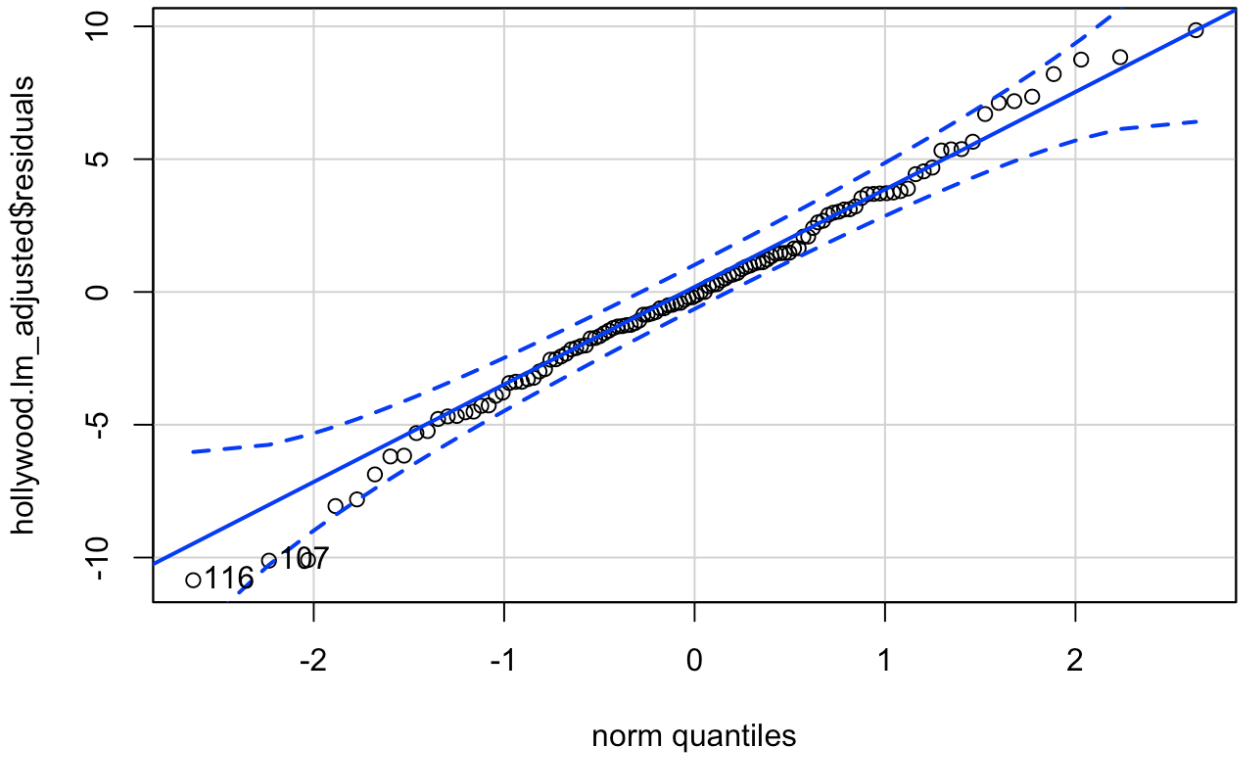
```
qqPlot(hollywood.lm$residuals)
```




```
## [1] 13 77
```

d. Try replacing `WorldGross` by `sqrt(WorldGross)` in the `lm` formula. Repeat part (c) and comment.

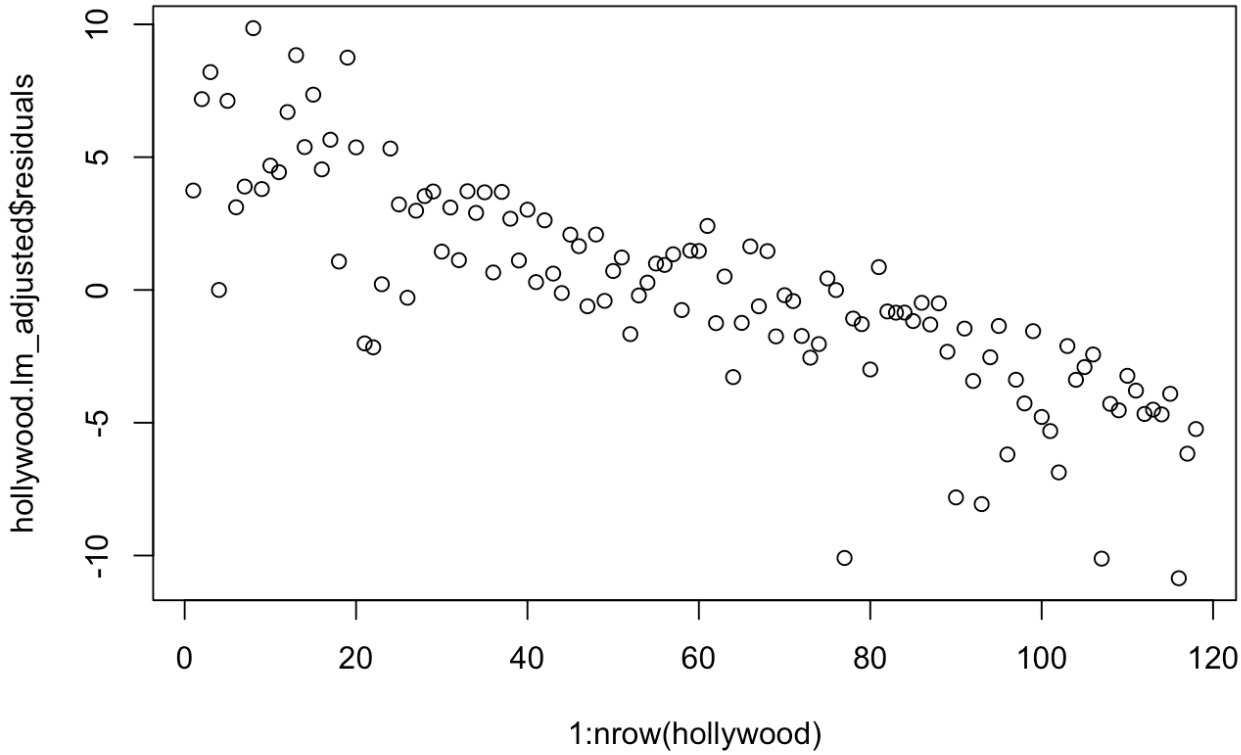
```
hollywood.lm_adjusted <- lm(sqrt(WorldGross) ~ AudienceScore + BOAveOpenWeek + Fa
ntasy + Budget, data = hollywood)

qqPlot(hollywood.lm_adjusted$residuals)
```




```
## [1] 116 107
```

e. Try plotting the residuals of the model in 2d versus the row number of the movie (that is, use `plot` with first argument `1:nrow(Hollywood)` and second argument the residuals). What do you observe? Explain what this indicates (and you might want to change your answer to 2a accordingly).

```
plot(1:nrow(hollywood), hollywood.lm_adjusted$residuals)
```


hollywood.lm_adjusted$residuals
10
5
0
-5
-10
0
20
40
60
80
100
120
1:nrow(hollywood)