Homework 10: Multiple Linear Regression

NAME: Andres Castano

NETID: ac986

DUE DATE: December 2, 2016 by 5:00pm

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.

```
Movies <- read.csv("Hollywood.csv")
names(Movies)
```

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

```
dim(Movies)
```

```
## [1] 118   5
```

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.

```r
Wgross.lm = lm(WorldGross~AudienceScore+BOAveOpenWeek+Budget+Fantasy, data=Movies)
summary(Wgross.lm)
```

```
##
## Call:
## lm(formula = WorldGross ~ AudienceScore + BOAveOpenWeek + Budget +
##     Fantasy, data = Movies)
##
## 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 **
## Budget         2.658e+00  2.432e-01  10.929  < 2e-16 ***
## FantasyTRUE    8.306e+02  1.319e+02   6.296 6.01e-09 ***
## ---
## 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).

The expression is as follows:

$\hat{E}(WorldGross|predictors)$ = *-126.3 + 1.679* `AudienceScore` *+ 0.003546* `BOAveOpenWeek` *+ 2.658* `Budget` *+ 830.6* `FantasyTRUE`

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

The covariate `FantasyTRUE` is a dummy variable that takes the value of 1 when the movie genre is Fantasy a 0 in other case. In the context of our regression the coefficient 830.6 means that on average movies which genre is Fantasy have a Gross income 830.6 millions greater compared with not fantasy movies. Despite that this effect is statistically significant, we need to be careful with the interpretation because in the sample there is only one (out of 118) movies of the fantasy genre, so the mean for each group is calculated on sample sizes radically different.

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.

$\hat{E}(WorldGross|predictors)$ = *-126.3 + (1.679)(90) + (0.003546)(10000) + (2.658)(50) + (830.6)(0)* =

```r
-126.3 + (1.679)*(90) + (0.003546)*(10000) +  (2.658)*(50) + (830.6)*(0)
```

```
## [1] 193.17
```

The estimated expected gross income for a movie with the characteristics mentioned is 193.17 millions.

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.

Here what we are calculating is 100(1-α)% confidence intervals for the budget coefficient:

100(1-0.05)% confidence interval for $Budget = \hat{Budget} \pm t_{0.025,113}\hat{se}(\hat{Budget})$

```
confint(Wgross.lm)
```

```
##                       2.5 %        97.5 %
## (Intercept)   -2.159766e+02  -36.66660497
## AudienceScore  2.670957e-01    3.09175139
## BOAveOpenWeek  1.129396e-03    0.00596279
## Budget         2.175829e+00    3.13930121
## FantasyTRUE    5.692309e+02 1091.96019004
```

As we can observe the 95% confidence interval for the partial slope `Budget` is (2.175829,3.13930121), which mean that we are 95% confident that the population `Budget` parameter is in this interval.

Problem 2

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

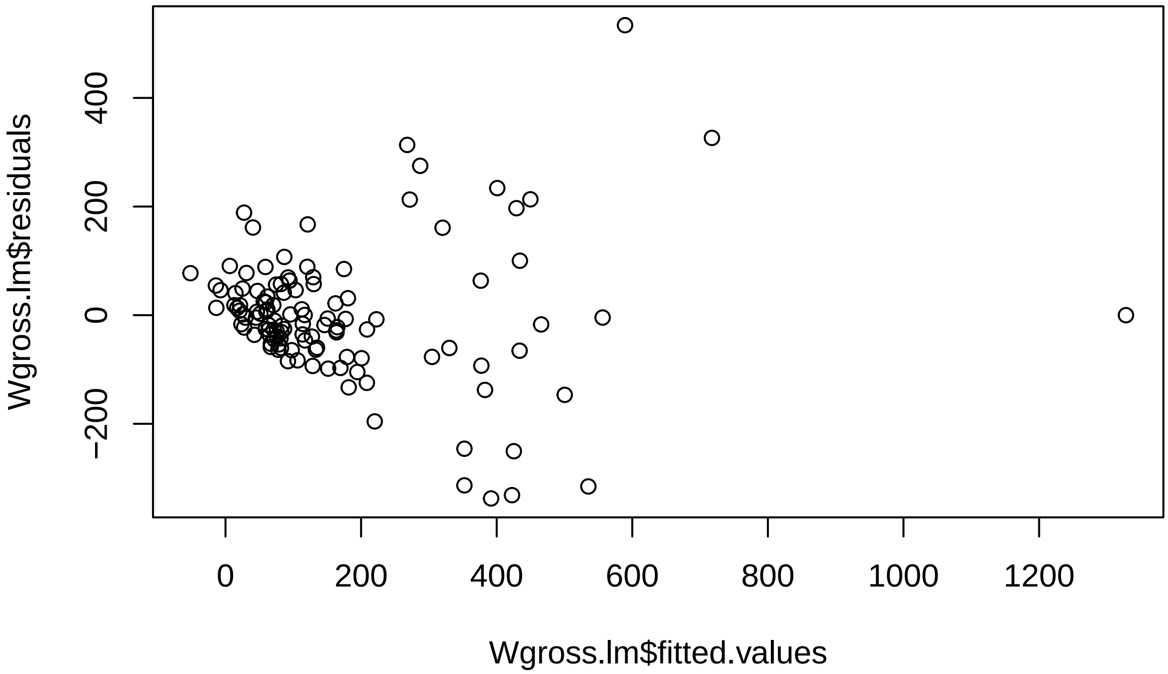
a) Does it seem reasonable to assume these observations are independent?

It would be reasonable to assume independence between observations if the movies were selected randomly among all the movies released in 2011. However, this information is not provided. I think tha one way to get an insight about this is examinig historical data of the released movies. For example, consulting in http://www.the-numbers.com/market/creative-types, I observed that the historical share of fantasy movies in the period 1995-2016 was 6.62% (see below), this value is particulary important because in our sample the fantasy movies only represent the 0.008% of the 118 movies. It is normal that due to variation from sample to sample, you would not expect to see exact 6.6% of fantasy movies in our dataset, however the difference might create suspicions about the way that the data was collected.

Creative Type	Movies	Share (%)
Contemporary Fiction	5714	50.41
Kids Fiction	434	3.83
Fantasy	750	6.62
Science Fiction	524	4.62
Historical Fiction	1156	10.20
Super Hero	84	0.74
Dramatizacion	788	6.95
Factual	1859	6.40
Multiple Creative Types	27	0.24
——————————-	———	————
Total	11336	100

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(Wgross.lm$fitted.values, Wgross.lm$residuals)
```



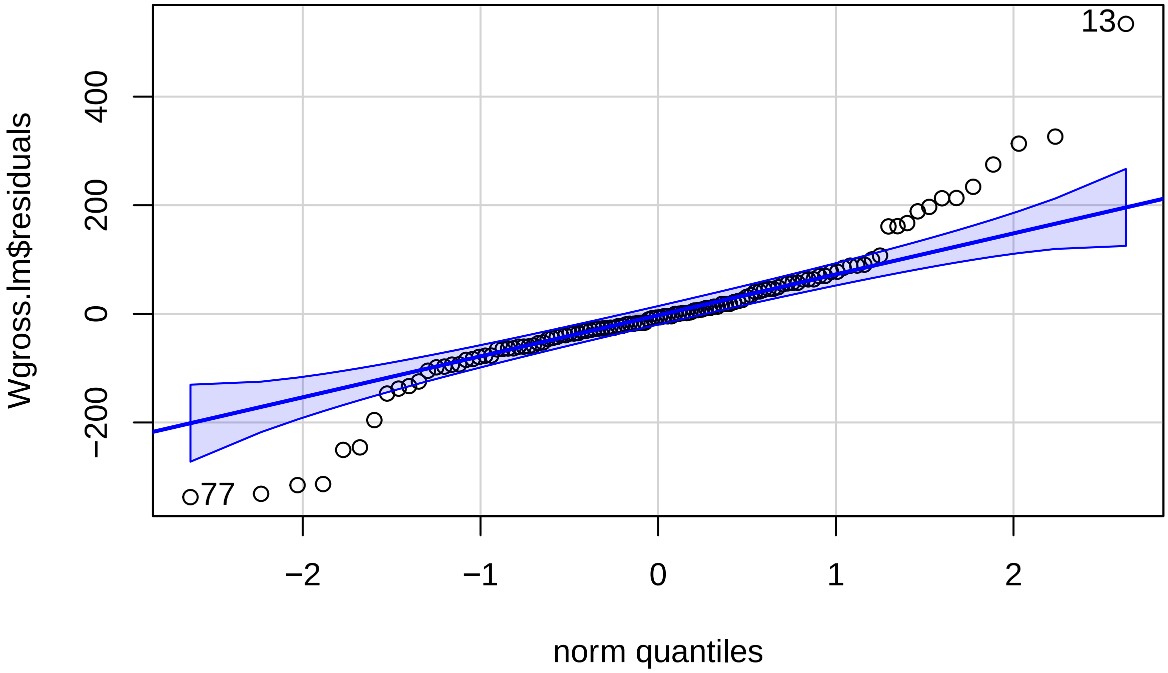
The equal variance could be questioned. There are some big outliers.

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

```
library(car)
```

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

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



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

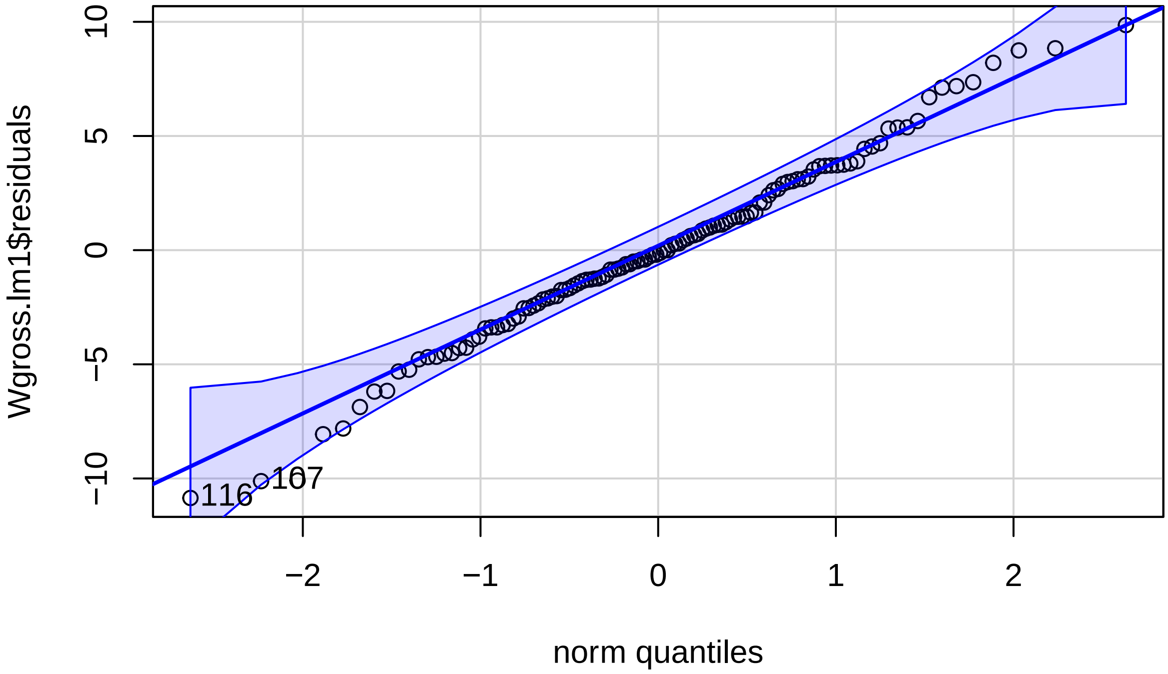
There are many points outside of the confidence intervals of the Q-Q plot. This suggest that the normality assumption does not seem reasonable.

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

```
sqrtWorldGross=sqrt(Movies$WorldGross)
Wgross.lm1 = lm(sqrtWorldGross~AudienceScore+BOAveOpenWeek+Budget+Fantasy, data=Movies)
summary(Wgross.lm1)
```

```
##
## Call:
## lm(formula = sqrtWorldGross ~ AudienceScore + BOAveOpenWeek +
##     Budget + Fantasy, data = Movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8545  -2.2857  -0.1571   2.6668   9.8568
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.356e+00  1.466e+00   0.925 0.356879
## AudienceScore 6.238e-02  2.309e-02   2.701 0.007974 **
## BOAveOpenWeek 1.025e-04  3.952e-05   2.595 0.010707 *
## Budget        8.555e-02  7.877e-03  10.861  < 2e-16 ***
## FantasyTRUE   1.469e+01  4.274e+00   3.437 0.000825 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.064 on 113 degrees of freedom
## Multiple R-squared:  0.644,  Adjusted R-squared:  0.6314
## F-statistic: 51.11 on 4 and 113 DF,  p-value: < 2.2e-16
```

```
library(car)
qqPlot(Wgross.lm1$residuals)
```

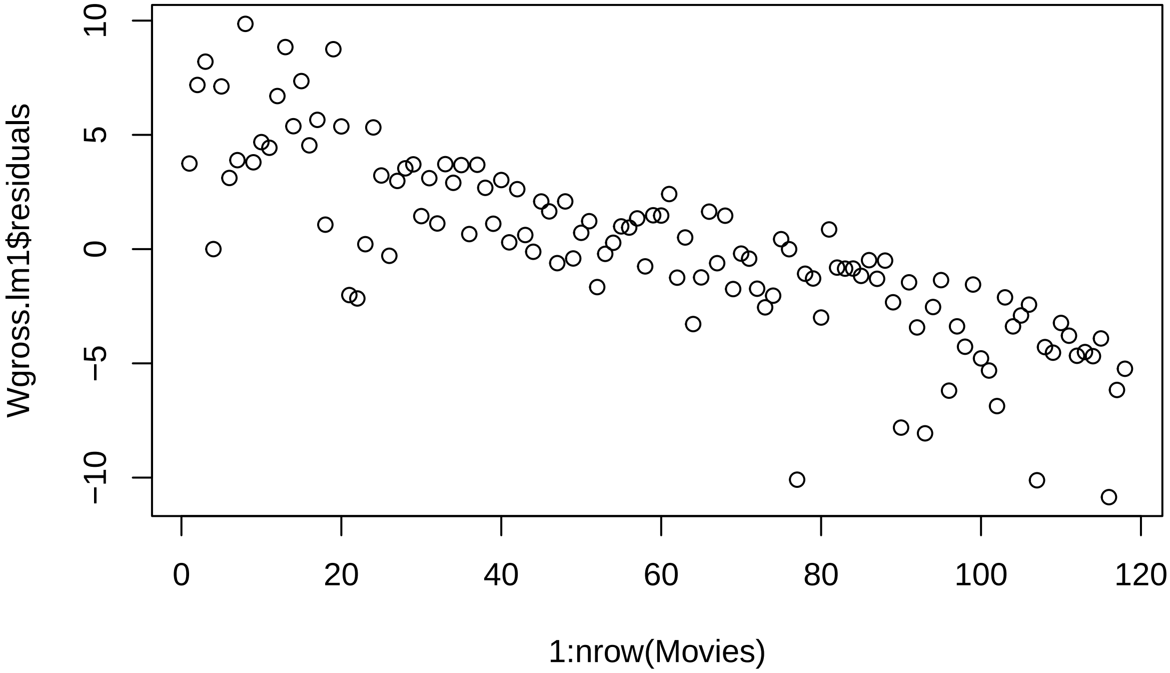


```
## [1] 116 107
```

With this transformation of the dependent (response) variable we could see that the normality assumption is reasonable. The transformation has helped us to get a better linear fit.

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(Movies),Wgross.lm1$residuals )
```



This result could suggest that the independence condition is in jeopardy. If the observartions were independent we should not observe a pattern in the residuals. This result supports my suspicions about the independence assumption: the data might not be a SRS of the movies.