Problem Set 2

Johnny Chapman Hersh MGSC 310

## Load Packages

library(readr)  
library('tidyverse')

## ── Attaching packages ─────────────────────────────────────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.2.1 ✔ purrr 0.3.2  
## ✔ tibble 2.1.3 ✔ dplyr 0.8.3  
## ✔ tidyr 0.8.3 ✔ stringr 1.4.0  
## ✔ ggplot2 3.2.1 ✔ forcats 0.4.0

## ── Conflicts ────────────────────────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

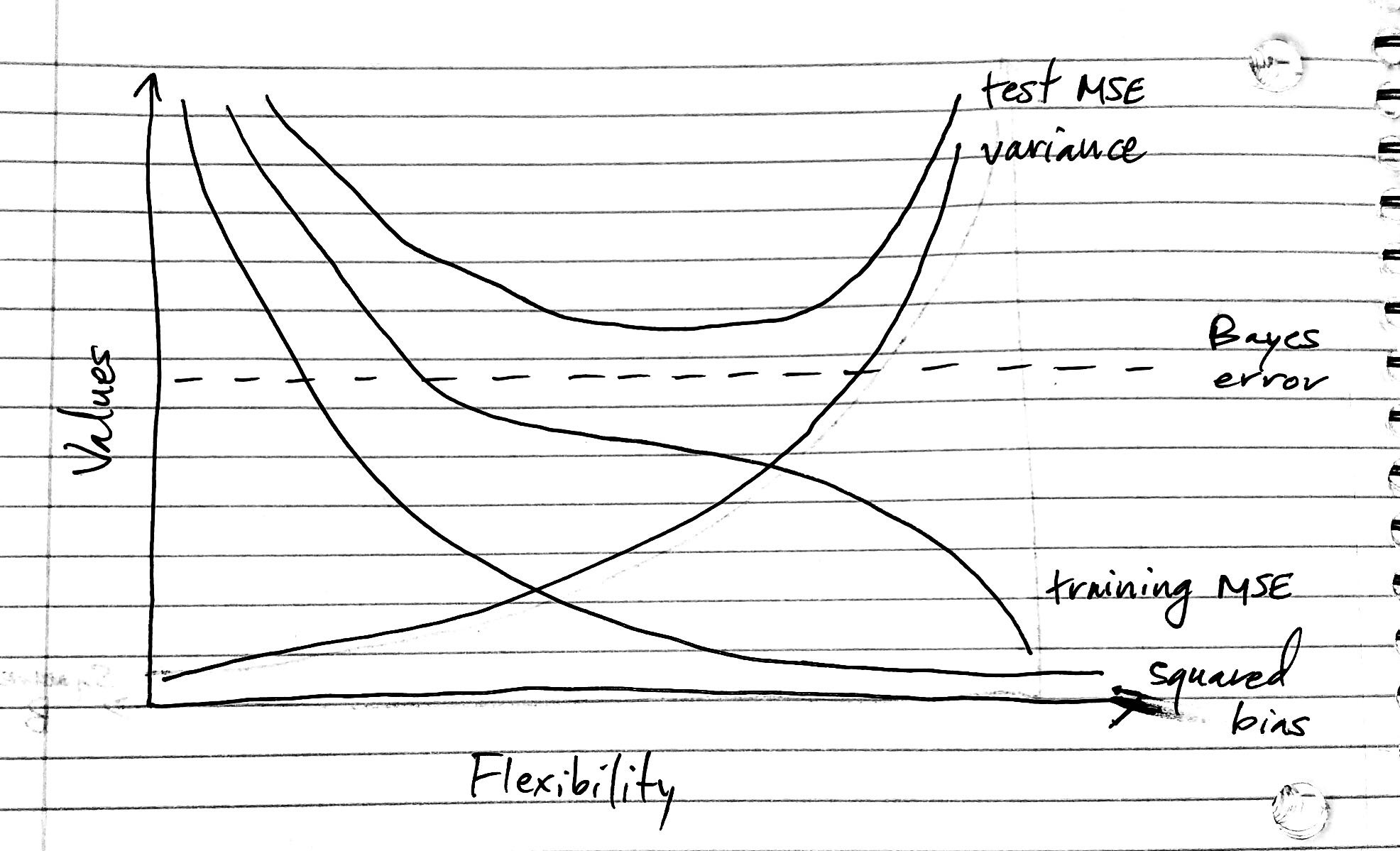
library('ggthemes')  
library('knitr')  
library('magrittr')

##   
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':  
##   
## set\_names

## The following object is masked from 'package:tidyr':  
##   
## extract

### Question 1, ISLR Chapter 2, question 3:



Graph sketch

1. The training MSE data decreases as flexibility increases because as the flexibility increases the function f fits the data more closely. The test MSE declines initially as flexibility increases, but then levels off and starts to increase. This happens when there is a small training MSE and large test MSE, that overfits the data. The function tries to find the patterns in the data even if data is caused by chance and not follow the properties of the function. The squared bias decreases as the variance increases. The bias is the tendency of an in-sample statistic to over or under estimate the statistic in the population. The Bayes error is constant so it is a parallel line.

### Question A:

Download the Top 5000 movies on IMDB from the following link:

movies <- read.csv("Datasets/movie\_metadata.csv")

### Question B:

options(scipen = 10)  
  
movies <- movies %>% filter(budget < 400000000) %>% filter(content\_rating != "",  
content\_rating != "Not Rated")  
movies <- movies %>%  
mutate(genre\_main = unlist(map(strsplit(as.character(movies$genres),"\\|"),1)),  
grossM = gross / 1000000,  
budgetM = budget / 1000000,  
profitM = grossM - budgetM,  
cast\_total\_facebook\_likes000s = cast\_total\_facebook\_likes / 1000)  
movies <- movies %>% mutate(genre\_main = factor(genre\_main) %>% fct\_drop())

### Question C:

Split the movies dataset into a testing and training set, with the training set 80% of the size of the original dataset. Be sure to use set.seed(1861) to ensure your results are comparable to mine and your classmates.

set.seed(1861)  
train\_idx <- sample(1:nrow(movies), 0.8 \* nrow(movies))  
movies\_train <- movies %>% slice(train\_idx)  
movies\_test <- movies %>% slice(-train\_idx)

### Question D:

How many rows are in the test and trainig datasets?

dim(movies\_train)

## [1] 3396 33

dim(movies\_test)

## [1] 849 33

3,396 Rows in training data 849 rows in test data

### Question E:

cormat <- cor(movies\_train %>% select\_if(is.numeric) %>% drop\_na())  
print(cormat[, "profitM"])

## num\_critic\_for\_reviews duration   
## 0.2416979638 0.1063892917   
## director\_facebook\_likes actor\_3\_facebook\_likes   
## 0.1070303835 0.1665145723   
## actor\_1\_facebook\_likes gross   
## 0.0472643456 0.7859808359   
## num\_voted\_users cast\_total\_facebook\_likes   
## 0.4861649378 0.0988924207   
## facenumber\_in\_poster num\_user\_for\_reviews   
## -0.0196624052 0.3693420547   
## budget title\_year   
## 0.0005501269 -0.1205250671   
## actor\_2\_facebook\_likes imdb\_score   
## 0.1196320100 0.2609408853   
## aspect\_ratio movie\_facebook\_likes   
## -0.0602042210 0.2386962438   
## grossM budgetM   
## 0.7859808359 0.0005501269   
## profitM cast\_total\_facebook\_likes000s   
## 1.0000000000 0.0988924207

There is a strong correlation between grossM and profitM. With a grossM correlation coefficient of 0.786.

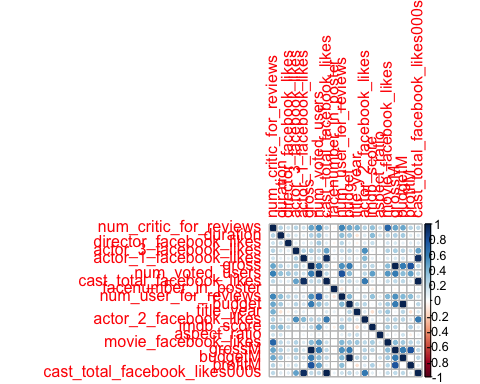
### Question F:

Extra Credit. Use the corrplot package to produce a plot of the correlation matrix

library('corrplot')

## corrplot 0.84 loaded

corrplot(cormat)



### Question G:

Let’s regress profitM against imdb\_score and store this as mod1. Use the summary() function over mod1 to print the regerssion summary. Be sure to estimate our model against the training dataset.

mod1 <- lm(profitM ~ imdb\_score, data = movies\_train)  
summary(mod1)

##   
## Call:  
## lm(formula = profitM ~ imdb\_score, data = movies\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -389.02 -26.38 -9.65 16.38 490.35   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -72.1702 5.8946 -12.24 <2e-16 \*\*\*  
## imdb\_score 13.3319 0.9019 14.78 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 52.53 on 3027 degrees of freedom  
## (367 observations deleted due to missingness)  
## Multiple R-squared: 0.06732, Adjusted R-squared: 0.06701   
## F-statistic: 218.5 on 1 and 3027 DF, p-value: < 2.2e-16

### Question H:

Interpret the coefficient for imdb\_score, being specific about the impact regarding magnitude and sign.

This means that for every 1 point increase in imdb\_score there is a $13 million increase in profit.

### Question I:

What is the p-value associated with the estimate of imdb\_score? In your own words, what does a p-value mean? What does this estimate p-value imply about the relationship between imdb\_score and profit?

The p-value is a statistic to find the significance of the results. This p-value is < 2.2e-16, which is less than 0.05 indicating you should reject the null hypothesis.

### Question J:

Include cast\_total\_facebook\_likes000s as a predictor in addition to imdb\_score. Store this model as mod2 and use summary() to output the results.

mod2 <- lm(profitM ~ imdb\_score + cast\_total\_facebook\_likes000s, data = movies\_train)  
summary(mod2)

##   
## Call:  
## lm(formula = profitM ~ imdb\_score + cast\_total\_facebook\_likes000s,   
## data = movies\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -386.71 -25.96 -9.25 16.22 492.26   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -72.09743 5.87758 -12.27 < 2e-16 \*\*\*  
## imdb\_score 12.95458 0.90359 14.34 < 2e-16 \*\*\*  
## cast\_total\_facebook\_likes000s 0.20710 0.04805 4.31 0.0000168 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 52.37 on 3026 degrees of freedom  
## (367 observations deleted due to missingness)  
## Multiple R-squared: 0.07301, Adjusted R-squared: 0.0724   
## F-statistic: 119.2 on 2 and 3026 DF, p-value: < 2.2e-16

### Question K:

What is the estimated impact of cast facebook likes on movie profits?

This means that for every 1 like there is a $207.1 thousand increase in profit.

### Question L:

What is the impact of content rating on a movie’s expected profit? To answer this question we will have to clean content\_rating a little bit. Use the fct\_lump() function to create factor levels for the four most common factor levels, leaving the rest as a category “other”. Call this variable rating\_simple and store it in the movies\_train data frame.

movies\_train %<>% mutate(rating\_simple =   
 fct\_lump(content\_rating, n = 4))  
  
table(movies\_train$rating\_simple)

##   
## G PG PG-13 R Other   
## 92 509 1107 1559 129

### Question M:

Estimate a model with profitM on the left-hand-side and imdb\_score, cast\_total\_facebook\_likes000s and rating\_simple on the right-hand side. Interpret the coefficient for rating\_simpleR.

mod2 <- lm(profitM ~ imdb\_score + cast\_total\_facebook\_likes000s + rating\_simple, data = movies\_train)  
summary(mod2)

##   
## Call:  
## lm(formula = profitM ~ imdb\_score + cast\_total\_facebook\_likes000s +   
## rating\_simple, data = movies\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -378.42 -24.53 -7.55 17.12 486.11   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -64.55692 8.27213 -7.804 8.19e-15 \*\*\*  
## imdb\_score 14.08094 0.90685 15.527 < 2e-16 \*\*\*  
## cast\_total\_facebook\_likes000s 0.19101 0.04778 3.998 6.55e-05 \*\*\*  
## rating\_simplePG -1.11971 6.29605 -0.178 0.858857   
## rating\_simplePG-13 -10.21472 6.01564 -1.698 0.089605 .   
## rating\_simpleR -23.03335 5.95321 -3.869 0.000112 \*\*\*  
## rating\_simpleOther -18.26624 9.17181 -1.992 0.046509 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 51.73 on 3022 degrees of freedom  
## (367 observations deleted due to missingness)  
## Multiple R-squared: 0.0968, Adjusted R-squared: 0.095   
## F-statistic: 53.98 on 6 and 3022 DF, p-value: < 2.2e-16

The coefficient of rating\_simpleR means that there is an expected $23.03 million decrease profit compared to the expected profit for a rating\_simpleG.

### Question N:

Why does the coefficient for G not appear in the regression table above?

The coefficient for G does not appear because R uses rating\_simpleG as a reference.