APPLIED LINEAR MODELS 2023-24

Final assignment

1. Download or collect some data on a topic of interest of to you. You will use this dataset to work through the concepts and methods of the course, so the data should be worth your time and should have some complexity.
2. Discuss your applied goals in studying this example and how the data can address these goals.
3. Use suitable representation to graph the data and comment the relation among all variables (e.g. scatterplot matrix, boxplot, etc).
4. Consider all the available covariates (and potential interactions) to explain your response and assume a linear or a generalized linear regression model. Perform a subset selection to explore all the possible models and comment the results.
5. What is the best overall model obtained according to different criteria (e.g. *AIC*, *BIC*, *adjusted* R^2, *Mallow's* C_p) and Cross-validation error? Show some plots to provide evidence for your answer.
6. Identify potential collinearity issues in your selected model and, eventually, apply possible solutions to improve your model.
7. Perform regression diagnostics to validate all the hypotheses behind your model. Identify unusual observations (e.g. high leverage points, outliers, influential points).
8. Using the findings of your diagnostics analysis, improve your model.

9. Report the coefficients of the best model obtained, interpret the parameters and their uncertainties, also with the support of a graphical representation.
10. Test each of the β_j to be 0 against two-sided alternatives and discuss the conclusions.
11. Test a group of regressors (motivate your choices) and all regressors. Explicitly state and discuss the hypotheses tested and the conclusions.
12. How good does the model fit the data? Find and comment suitable measures of goodness of fit.
13. Suppose you have information about a new observation of your regressors. Provide the prediction of your response variable with associated uncertainty.
14. Simulate n data points from the fitted regression model, assuming the estimated parameters as the true parameters.

Final assignment

Sara Pascali

2024-02-26

1 Introduction

The data set was created by Jayanth S. and it was taken from the kaggle website. The dataset is called "IMDb ratings of the Top 50 movies" of a specific time, it studies the top 50 movies based on the classification provided by IMDb, also known as the Internet Movie Database. The data were collected from the official website of IMDb. This is an online database of information related to movies, television, series etc. It combines movie plot description, critic and user ratings and reviews and many more aspects.

1.1 Description of the data set

My dataset has 50 observations labeled with the rank from 1 to 50. The response variable that will be used through all the analysis is the "IMDbRating", which is the rating IMDb gave to each movie. This is a continuous variable restricted between 1 and 10, but since the dataset includes the 50 best movies the score is between 8.5 and 9.3.

While the predictors are the following:

- **Release Date**: date when the film has been released in theaters. The data set includes movies which have been released from 1942 to 2023. For analytical purposes, this variable is substituted with **Movie Age**, indicating the film's age (calculated as 2024 minus the year of release).
- **Duration**: duration of each movie in minutes.
- **Age Rating**: is an american method to rate a movie's suitability for certain audiences based on its content. This list of rate distinguished 4 categories: G as General audiences (all ages admitted), PG as Parental Guidance Suggested, in particular PG-13 suggests an accompanying adult under 13 and R as Restricted (Under 17 or Under 15 requires accompanying adult).
- **Number of people who voted**: the number of people who rated the movie. The least voted movie is "Harakiri" with 66.000 voters, while the most voted are "The Dark Knight" and "The Shawshank Redemption" with 2.800.000 with voters. Then I divided the variable for 1.000 to facilitate the analysis.

1.2 Applied goals

This dataset can be analyzed from various perspectives. Being able to predict the IMDb rating provides a valuable tool for film producers to understand which factors have the greatest influence on the final rating and thus the potential success of a film. Naturally, apart from the predictors present in my dataset, there are numerous other factors that could influence the rating, such as the genre, the director or the cast of the film. However, personally, I believe that the duration of the film and its year of release are the most interesting factors to analyze.

It also might be interesting to analyze the film's rating in relation to the box office for each individual movie or even the number of awards won by the movie. To do this, it is necessary to collect more data about the box office of each movie.

1.3 Graphical representation of my variables

Movie Age and Duration

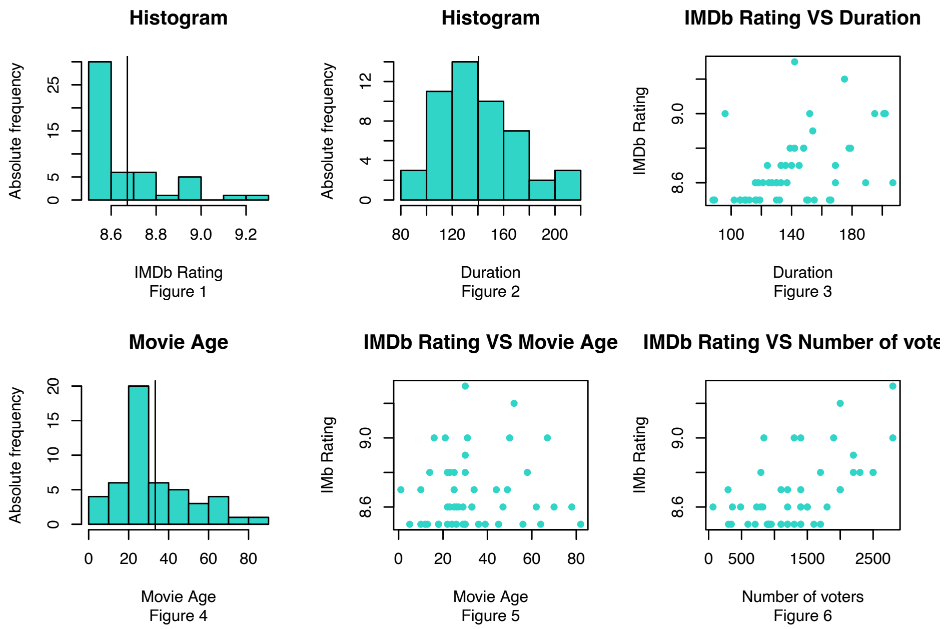
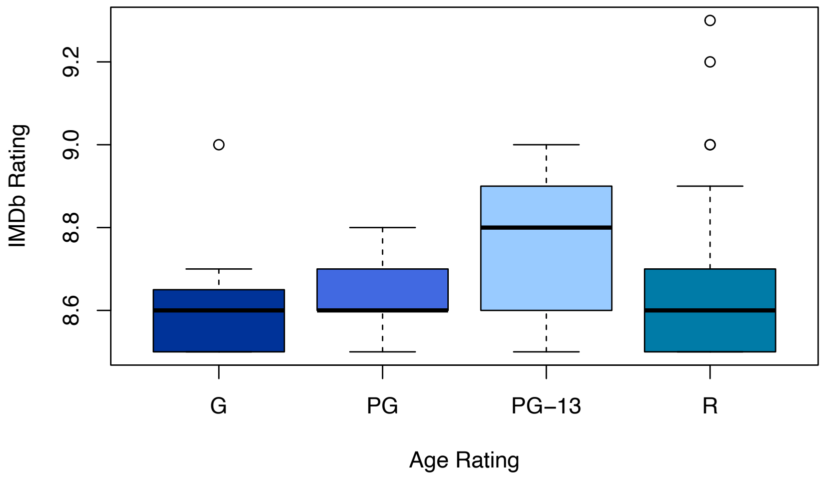


Figure 1 illustrates the response variable, constrained within the range of 8.5 to 9.3. Observing the graph, it becomes evident that the majority of films are within the range of 8.5 to 8.6, such as "Back to the Future" (8.5, 1985) and "Parasite" (8.6, 2019). Figures 2 and 4 show the two most interesting predictors: "Duration" and "Movie Age," with mean values respectively 140 minutes and 33 years (1991), Figures 3, 5 and 6 show the relationship between the predictors and the response variable. The first and the last plot (Figure 3 and 6), are the predictors where it seems to be present a slight positive correlation where it is stronger with "NumberOfVoters" variable. Furthermore, the correlation between IMDb rating and movie age does not reveal any definite pattern, suggesting that the age of the film—whether it is old or recent—does not significantly influence the dependent variable.

```
boxplot(Dataset$IMDbRating ~ Dataset$AgeRating, data = Dataset,
        col=c("#003399", "#4169E1", "#99CBFF", "#007BA7"), xlab= "Age Rating", ylab = "IMDb Rating")
```



2 Best subset selection

For a more precise interpretation of the coefficients, I centered the data of the independent variables by subtracting the mean from the variable values.

The best subset selection method enables us to assess which variables should be included in the final model. It is crucial to keep in mind the OCCAM principle, which advocates for favoring simpler hypotheses when two explanations account for the same phenomenon. This preference is due to the fact that a large number of variables means a great number of potential models, which can render decision-making challenging. In my study, I estimated the parameters through ordinary least squares regression.

```
ols_res = regsubsets(IMDbRating~MovieAge + Duration + AgeRating + NumberOfVoters, data = Dataset)
summ.ols_res= summary(ols_res)
#criterion approach
par(mfrow=c(1,4))
# AIC
AIC = matrix(NA, 6, 1)
for(j in 1:6){
  AIC[j] = summ.ols_res$bic[j] - (j+2)*log(n) + 2*(j+1)}
plot(AIC, type="b", pch=19, xlab="Number of predictors", ylab="", main="Drop in AIC")
abline(v=which.min(AIC),col = 2, lty=2)



#BIC
plot(summ.ols_res$bic, type="b", pch=19,
     xlab="Number of predictors", ylab="", main="Drop in BIC")
abline (v=which.min(summ.ols_res$bic),col = 2, lty=2)
mtext("Figure 8", side=1, line=4, cex=0.7)

#Cp Mallow
plot(summ.ols_res$cp, type="b", pch=19,
     xlab="Number of predictors", ylab="", main="Mallow' Cp")
abline (v=which.min(summ.ols_res$cp),col = 2, lty=2)
```

```
plot(summ.ols_res$adjr2, type="b", pch=19,
     xlab="Number of predictors", ylab="", main="Adjusted R^2")
abline (v=which.max(summ.ols_res$adjr2),col = 2, lty=2)
```

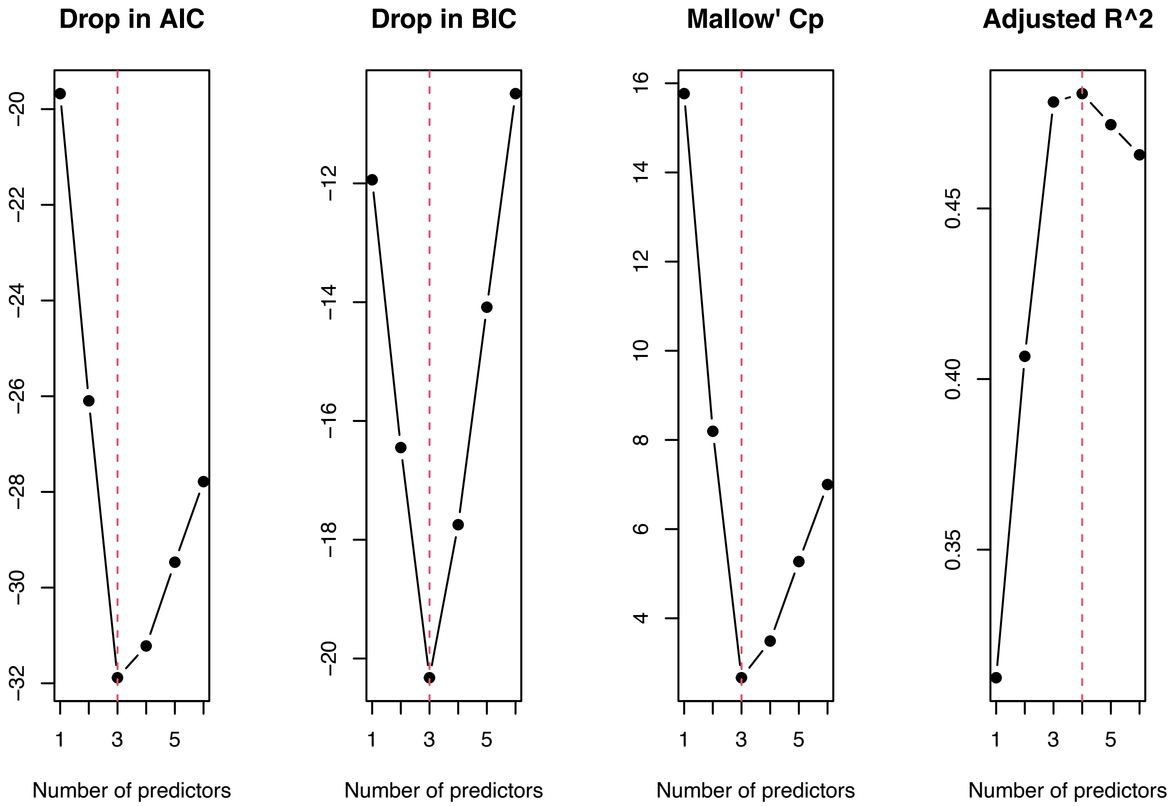
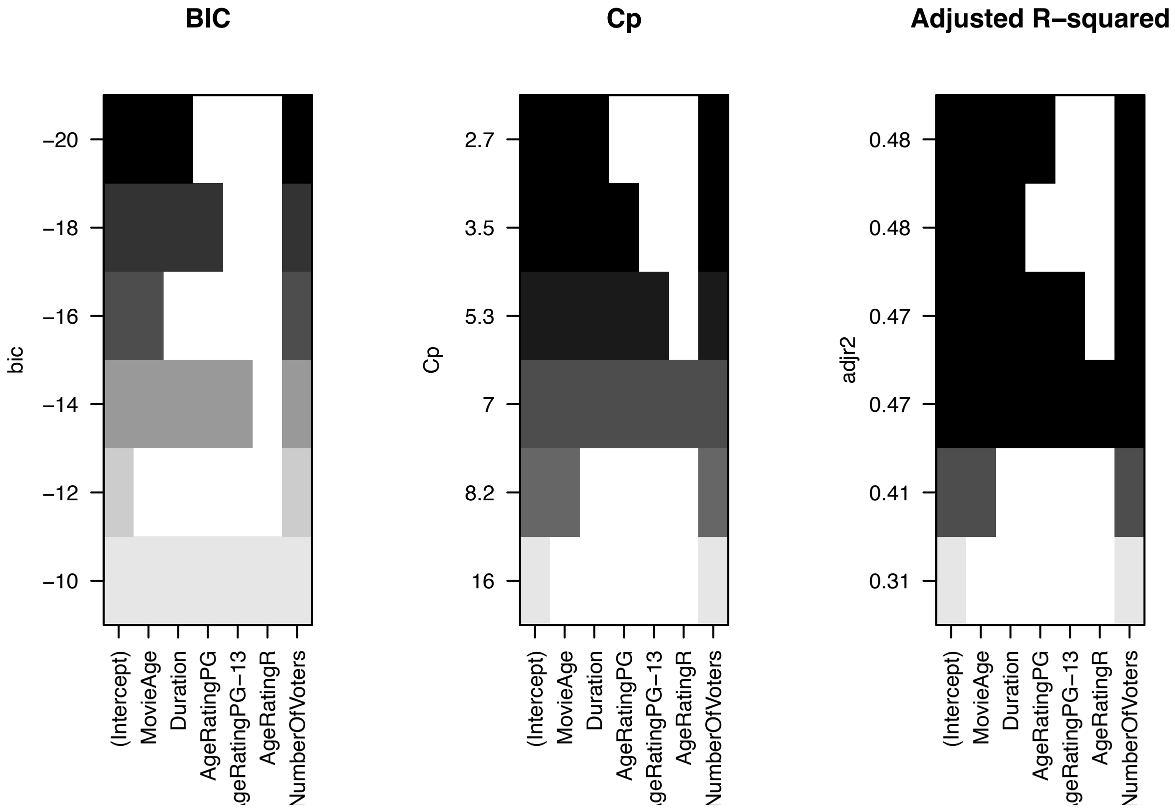


Figure 8

The Figure 8 illustrates five criteria used to assess the quality of each model relative to others, offering a methodology for model selection. Specifically, the AIC and BIC determine the variable to remove that has the most significant impact on the model ("drop"). Generally, the preferred model is characterized by the lowest AIC, BIC, or Cp value while maximizing the Adjusted R-square, which measures the goodness of fit while considering the number of predictor variables in the model. This metric diminishes the R-squared as additional variables are incorporated. Each plot has a dashed red line indicating the number of predictors that minimize/maximize the criterion. To determine the most suitable number of predictors for my model, I also considered the OCCAM principle and the elbow method. From this output I should choose a model with three predictors.



To facilitate clearer analysis, I integrated the output of the *regsubset*() function, which identifies the best predictor to include in my model by sequentially adding one covariate at a time, with the information presented in Figure 8. The plot exhibits all predictors on the x-axis, while the upper section illustrates where the criterion values are minimized (or maximized for the Adjusted R-squared). This plot facilitates the identification of the optimal number of predictors and which among them minimize/maximize the criterion values.

2.1 Cross-Validation: LOOCV

As an additional method for model selection, Cross-Validation serves as a statistical technique for evaluating and comparing learning algorithms. It involves partitioning data into two sets: one for training the model (the Training set) and the other for validating the model (the Test set).

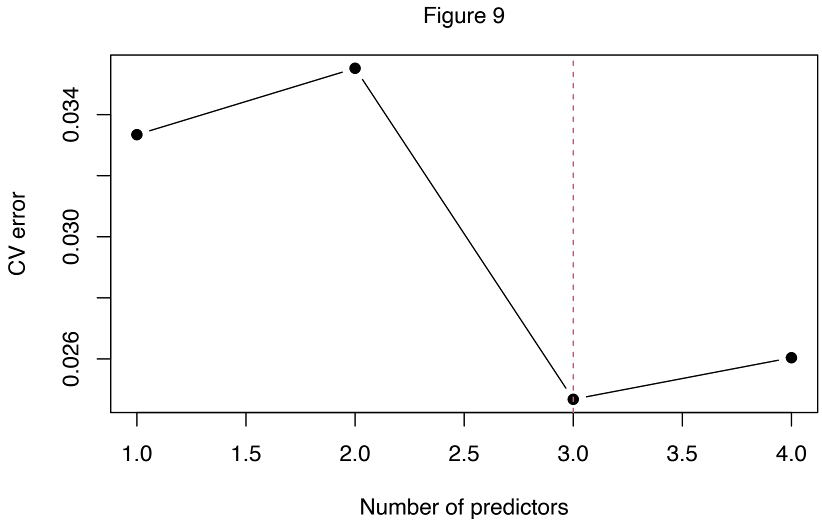
I processed the Leave-One-Out Cross Validation (LOOCV), it results in a reliable and unbiased estimate of model performance, despite its computational intense. Given the size of my dataset, which is not excessively large, I believe this method is the most appropriate choice.

```
k=n #number of folds equal to the number of observations
p = 4 #number of predictors
folds = sample(1:k,nrow(Dataset),replace =FALSE)
cv.errors = matrix (NA ,k, p, dimnames =list(NULL , paste (1:p)))
```

```
for(j in 1:k){
  ols_res = regsubsets(IMDbRating~MovieAge + Duration + AgeRating + NumberOfVoters + AgeRating, data = 
  for(i in 1:p) {
    mtrx = model.matrix(as.formula(ols_res$call[[2]]), Dataset[folds==j,])
    coeff = coef(ols_res , id = i)
    xvars = names(coeff)
    prediction = mtrx[,xvars ]%*% coeff
    cv.errors[j,i] = mean((Dataset$IMDbRating[folds==j] - prediction)^2)
}
}

cv.mean = colMeans(cv.errors)

plot(cv.mean ,type="b",pch=19,
xlab="Number of predictors",
ylab="CV error")
abline(v=which.min(cv.mean), col=2, lty=2)
mtext("Figure 9", side=3, line=1, cex=1)
```



The plot above (Figure 9) illustrates the Cross-Validation error corresponding to each model, in terms of number of predictors. The plot shows that the model that attains the minimum CV error is the one with three predictors, aligning with the outcomes of AIC, BIC, and Cp.

In conclusion, considering the insights garnered from the previous plots, I choose to include three predictors in my model: Movie Age, Duration, and Number of Voters, while excluding the categorical variable Age Rating.

Best overall Model

$$Y = \beta_0 + \beta_1 x_{MovieAge} + \beta_2 x_{Duration} + \beta_3 x_{NumberOfVoters}$$

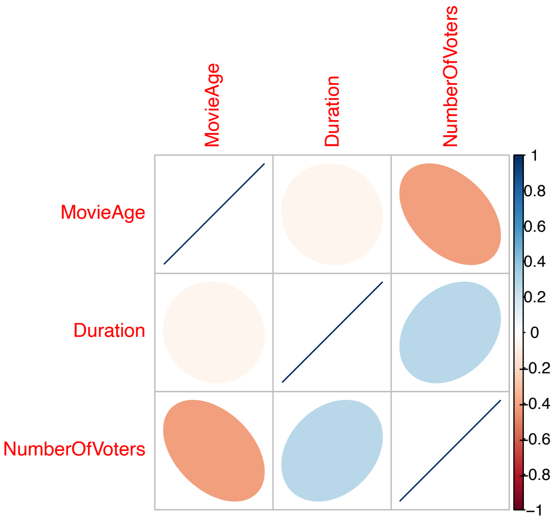
We shall now proceed to construct a linear model using the three predictors previously selected via the best subset selection method.

```
#Ordinary Least Squares
ols_best= lm(IMDbRating ~ MovieAge + Duration + NumberOfVoters, data = Dataset)
```

3 Multicollinearity

In any kind of analysis is important to check he presence of multicollinearity, which occurs when there is a high correlation among multiple independent variables within a model, which can adversely affect the model's outcomes. The correlation is checked examining the correlation matrix and conducting additional check by computing the Variance Inflation Factor (VIF), which is an indicator used in statistics to evaluate the presence of multicollinearity between the independent variables in a regression model.

```
#create the correlation matrix for all predictors
correlation_matrix = cor(Dataset[,-c(3,5)])
corrplot(correlation_matrix, method = "ellipse")
```



```
#VIF
vif(ols_best, type="predictor")
```

```
## VIFs computed for predictors
```

```
## [1] 1.217533 1.089099 1.315584
```

The correlation matrix above shows the correlations among each variable. The matrix is symmetric with the main diagonal with unity values, indicative of perfect correlation between each variable and itself. Initial inspection of the correlation matrix suggests the absence of collinearity issues. As I mentioned before, it is always better to compute the VIF and check if the value is below 10. In my study all the VIF are really small which confirms the absence of multicollinearity issues.

4 Diagnostic

Statistical diagnostics is a field that uses statistical methods and tools to analyze data and draw conclusions about specific phenomena or problems. Specifically, statistical diagnostics checks for:

4.1 The *linearity* assumption: if the residuals exhibit a non-random pattern, it suggests a potential violation of the linearity assumption.

4.1 The *homoschedaticity* assumption: this means that the estimation errors (difference between observed and model-predicted values) have a constant variance that is not dependent on the values of the independent variables.

4.2 The *normality* assumption: the residuals should be normally distributed.

4.3 *High leverage* points: these points may have a disproportionate influence on the linear approximation of the data, making the residuals particularly sensitive to changes in these points.

4.3 *Outlier*: points for which the model does not fit well.

4.3 *Influential point*: these are data points whose exclusion from the dataset would lead to a substantial alteration in the model fit. They may manifest as outliers or high leverage points.

4.1 Check for the homoschedasticity and linearity

```
#residuals
res = residuals(ols_best)
#fitted value
fitted = fitted.values(ols_best)
```

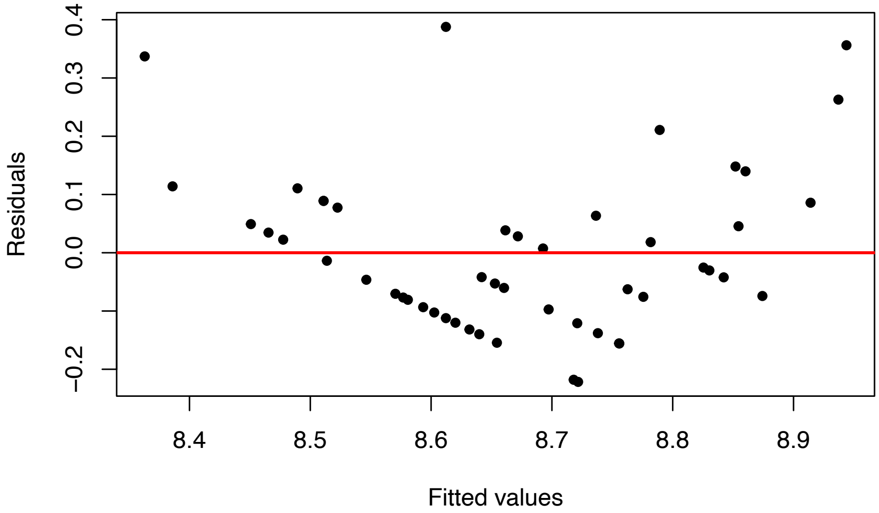


Figure 10

```
##                  Test stat Pr(>|Test stat|)
## MovieAge            0.1639          0.87055
## Duration            1.2253          0.22683
## NumberOfVoters      2.3793          0.02164 *
## Tukey test          6.6061        3.946e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

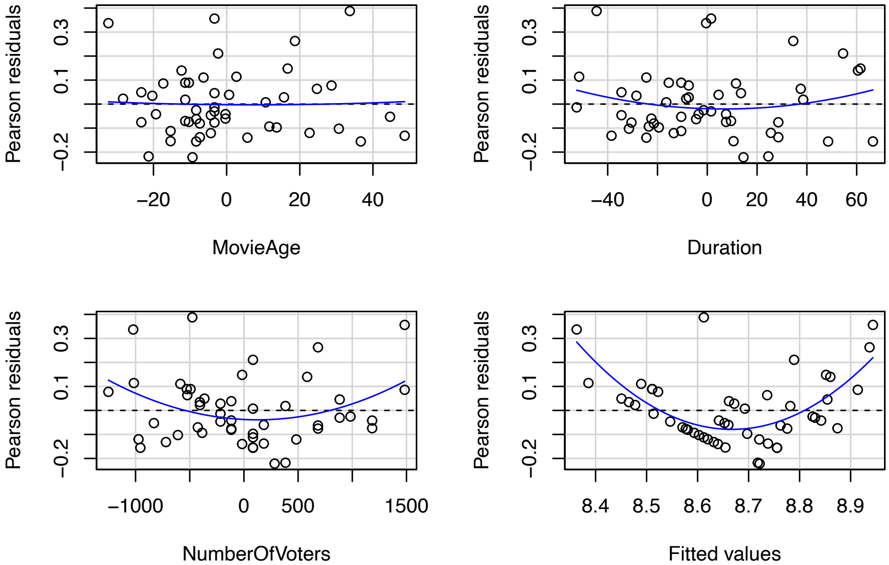


Figure 11

In the presence of heteroscedasticity, the dispersion pattern of residuals will exhibit variability that changes with the value of the independent variable. Analyzing Figure 10, it becomes evident that constant variance is observed.

Another crutial assumption to verify is the linearity assumption, which manifests when data points don't follow a define pattern. However, in my study, it is apparent that the data points form a convex curve, which confirms a non-linearity assumption verified, it can be seen also in the last graph on the right of Figure 11. The additional plots in Figure 11 show the distribution trend of residuals for each predictor. Notably, just the Movie Age predictor has linearity, whereas the remaining two predictors (Duration and Number Of Voters) do not have linearity, as evidenced by the blue curve. In order to fix the linearity I transformed all independent variables. Given the presence of negative values (due to data centering), conventional transformations such as logarithmic or square root are unsuitable. Instead, I processed a quadratic transformation and it actually solved the non-linearity.

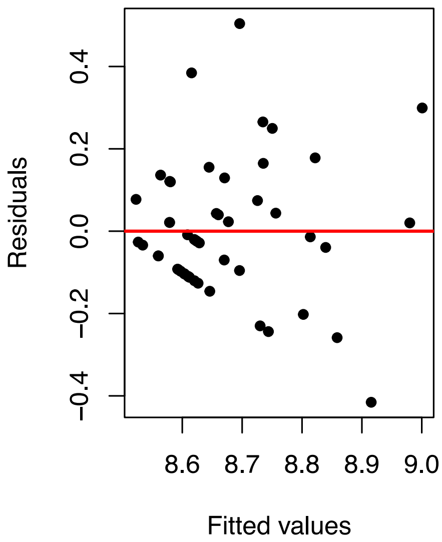


Figure 12

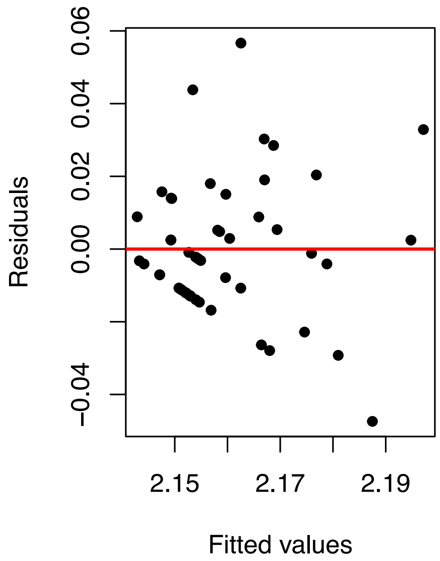


Figure 13: log(iMDb Rating)

Unfortunately, as shown in Figure 12, the quadratic transformation led to heteroscedasticity, classic funnel shape. Despite numerous attempts involving transformations of the response variable, including logarithmic transformation (Figure 13), square root transformation, as well as squaring IMDb Rating, neither the 1 and Y ratio resolved the issue. This condition could potentially damage the efficacy of model fitting.

4.2 Check the normality assumption

In order to check if there is a normal distribution I plotted the Normal QQ Plot and the Histogram.

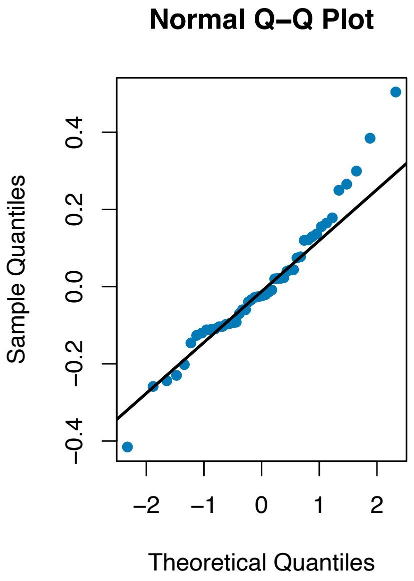
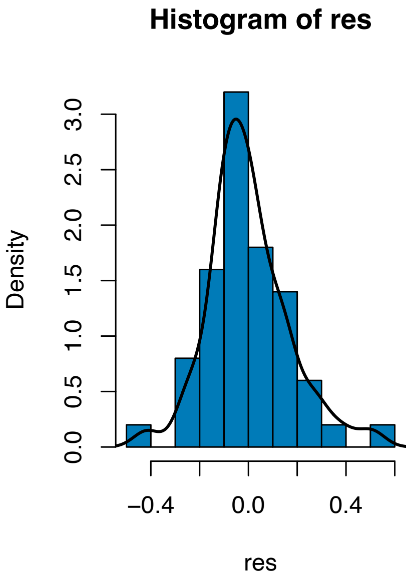


Figure 14



```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  res
## W = 0.9618, p-value = 0.1058
```

Figure 14 illustrates the presence of normality, because in the QQ-Plot the residuals follow the black line approximately. Additionally, the second plot demonstrates a distribution of residuals that is close to a normal distribution. This assumption is further supported by the Shapiro-Wilk test, which confirms the normality assumption, as indicated by a p-value greater than 0.05 (specifically, 0.1058).

4.3 Check for unusual observations

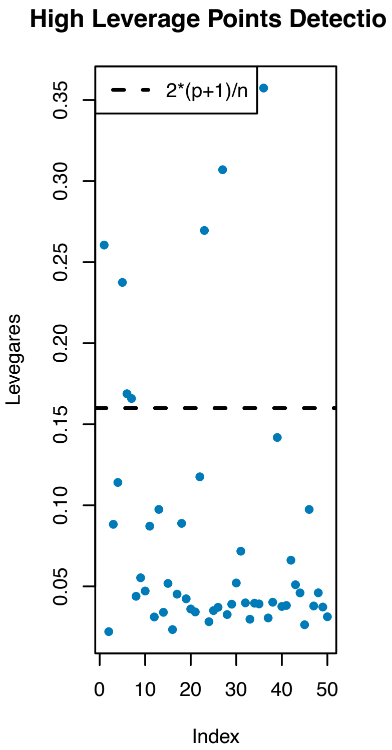


Figure 15

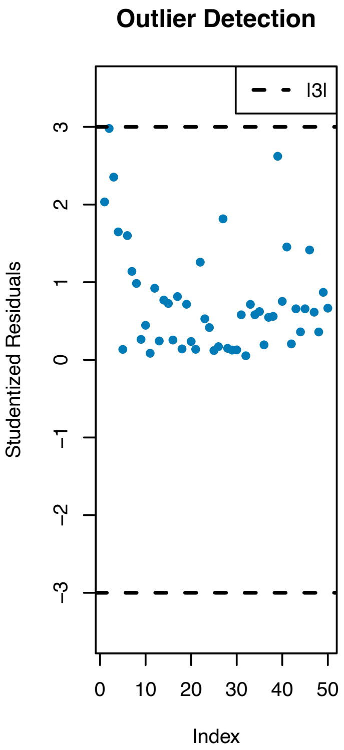


Figure 16

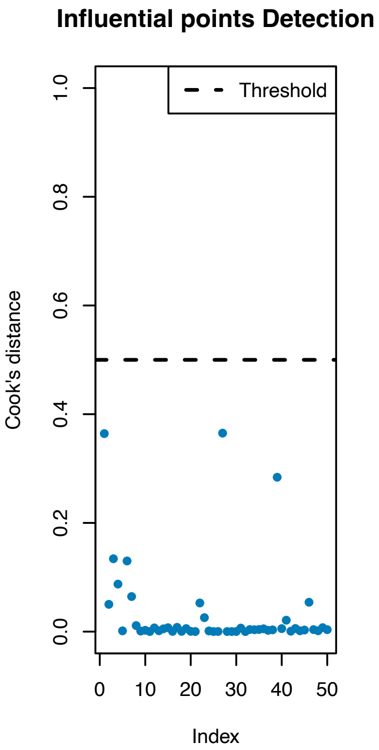


Figure 17

Figure 15 illustrates the detection of leverage for each observation. The black dashed line denotes the threshold beyond which points are considered high leverage points. The height of this threshold is denoted in the legend, where "n" represents the number of observations and "p" denotes the number of predictors in my model. The plot indicates the presence of several observations with high leverage. However, this is not a concern as these are not influential points because there are no points above the threshold in Figure 17.

Another crucial aspect to analyse is the existence of outliers (Figure 16). The plot reveals no outliers, even though there is one observation which is equal to 3 but, again, it is not a problem because it is not an influential point.

In conclusion, the diagnostic process involves examining influential points via the Cook's distance, as shown in Figure 17. The black dashed line represents the threshold beyond which points are classified as influential. In my case, no influential points surpass this threshold.

Before proceeding with the analysis, I conducted the same analysis done in Chapter 3. Despite transforming all predictors, the analysis yielded the same result as before. So we can continue the analys with the original model.

5 Interpretation of coefficients

```
##                       Estimate   Std. Error
## (Intercept)     8.672000e+00 2.420134e-02
## MovieAge       -9.622140e-05 5.018529e-05
## Duration        5.326315e-05 2.278438e-05
## NumberOfVoters  1.937204e-07 4.580281e-08
```

Since I transformed my predictors to make linear the residuals, the interpretation of the parameter slightly deviates from the usual. The coefficients now represent the effect of the squared independent variable on the dependent variable.

- The intercept (β_0) assumes a value equal to 8.575395. Given the data centering, interpretation varies from the norm. The value 8.575395 is the average IMDb Rating when the duration, age of the movie and the number of voters are at their mean value (after centering).
- β_1: the IMDb rating will decrease of 9.622140e-05 for a unit increase of the squared MovieAge variable, keeping constant the other variables.
- β_2: the IMDb rating will increase of 5.326315e-05 for a unit increase of the squared Duration variable, ceteris paribus
- β_3: the IMDb rating will increase of 1.937204e-07 for a unit increase of the squared Duration variable, ceteris paribus.

The Standard Error of each coefficient measures the variability in the estimated effect of the variable on the IMDb rating. A smaller standard error implies greater precision in estimating the predictor's effect. In my model, all variables exhibit relatively small standard errors, with the most precise predictor being the number of voters (5.213247e-09).

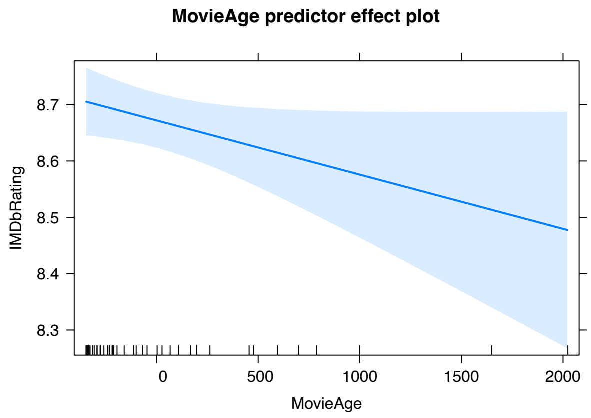
We can compute the confidence interval using *confint()* function. This allows for a deeper analysis of the parameter estimates and their associated uncertainty:

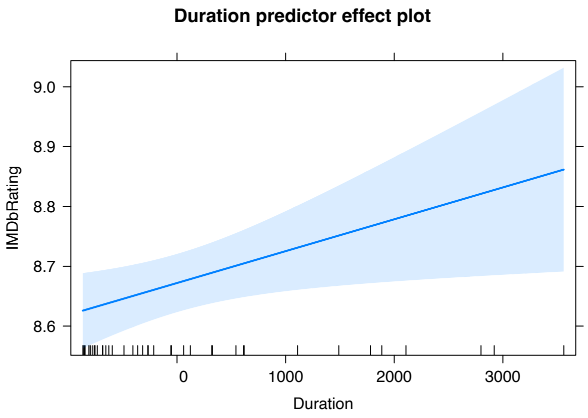
```
confint(ols_best)
```

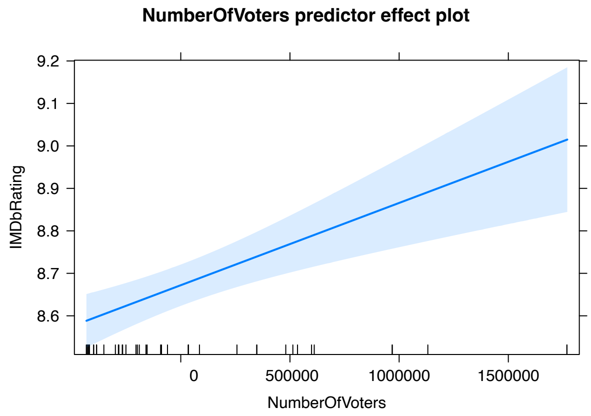
```
##                        2.5 %       97.5 %
## (Intercept)     8.623285e+00 8.720715e+00
## MovieAge       -1.972391e-04 4.796352e-06
## Duration        7.400577e-06 9.912573e-05
## NumberOfVoters  1.015241e-07 2.859167e-07
```

From the output, we observed that Duration and NumberOfVoters are statistically significant because the confidence interval does not include zero, instead the MovieAge variable seems to be not significance. This observation is further supported by the results of the t-test (Chapter 6.1).

To visualize the effect of each predictors we draw the *effects plot*. The shaded area on the graph provides a 95% pointwise confidence interval for the fitted values. The first plot shows the effect of the Movie Age on the IMDb Rating, as mentioned in the interpretation it has a negative effect which means that more recent is the movie, less will be the IMDb Rating and also the uncertainty increases (blue area is wider). Instead the effect plot of Duration and NumberOfVoters variables show the positive effect on IMDb Rating. Additionally, they reveal that uncertainty increase with longer durations and more voters.







10. Test each of the betaj to be 0 against two-sided alternatives and discuss the conclusions.

5.1 Test T

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

```
##                   t value     Pr(>|t|)
## (Intercept)    358.327225 6.475252e-81
## MovieAge        -1.917323 6.141936e-02
## Duration         2.337705 2.380795e-02
## NumberOfVoters   4.229443 1.102910e-04
```

The T test serves to determine if a parameter is statistically significant, indicating whether the predictor has an effect on the response variable. The table presented above displays the t-test values and corresponding p-values for each coefficient. Given that all p-values, except for MovieAge, are less than the significance level (α), we can conclude that they are statistically significant, thereby confirming their influence on IMDb Rating. Conversely, the MovieAge variable, as previously noted, it is not significance because the p-value is greater than 0.05 (0.06141936).

11. Test a group of regressors (motivate your choices) and all regressors. Explicitly state and discuss the hypotheses tested and the conclusions.

5.2 ANOVA Test

The ANOVA test is a particularly valuable method when investigating whether an independent variable significantly influences the dependent variable. If the p-value resulting from the ANOVA test is less than the significance level, it means that there are differences among the represented models, suggesting to keep the model with the original predictors.

The models tested are as follows:

$$\begin{cases} H_0 : Y = \beta_0 + \beta_1 x_{MovieAge} + \beta_2 x_{Duration} + \beta_3 x_{NumberOfVoters} \\ H_1 : Y = \beta_0 + \beta_2 x_{Duration} \end{cases}$$

```r
ols_reduced= lm(IMDbRating~Duration, data=Dataset)
anova(ols_reduced, ols_best)
```

```
## Analysis of Variance Table
##
## Model 1: IMDbRating ~ Duration
## Model 2: IMDbRating ~ MovieAge^2 + Duration^2 + NumberOfVoters^2
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     48 1.9041
## 2     46 1.3471  2   0.55694 9.5089 0.0003497 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The p-value is less than significance level, this suggests us to reject the null hypothesis and keep the bigger model because the two models differ.

12. How good does the model fit the data? Find and comment suitable measures of goodness of fit.

5.3 Index of goodness of fit

```
## [1] 0.2681675
```

R-Squared is the coefficient of determination, which is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. However, when using multiple regression models, simple R-squared can lead to misleading conclusions. This is because adding more predictor variables into the model can lead to an increase in R-squared, even if these added variables do not significantly improve the model's ability to predict the dependent variable. For this it is crucial to rely on the value of the *Adjusted R-squared* which takes into account the number of predictors in the model, in order to provide a more realistic measure of the model's ability to explain variation in the dependent variable. In my study the 27% of the variability of the IMDb Rating is explained by my model. It is not a great value but still two of the predictors are significant.

13. Suppose you have information about a new observation of your regressors. Provide the prediction of your response variable with associated uncertainty.

6 Prediction

I am conducting, using my model, a prediction of the IMDb Rating of the recent (2023) Oscar-nominated film "Poor Things". This movie has a duration of 142 minutes and has received votes from 136.000 people The output will provide the prediction interval.

```
# create a dataframe with the new obs of predictors
data_new = data.frame(MovieAge= 1, Duration= 142, NumberOfVoters=136)
# provide the response prediction
predict(ols_best, newdata= data_new, interval="prediction")
```

```
##         fit     lwr      upr
## 1 8.679493 8.33154 9.027447
```

Therefore, the prediction of IMDb rating for "Poor Things" falls within the range of 8.6 to 8.9. However, this estimation lacks precision as the actual IMDb rating is 8.2. This not precise prediction but is not surprising because my dataset represents the top 50 movies and not the entire "population". In general, 50 observations are not enough in order to make this kind of prediction.

(*reference: https://www.imdb.com/title/tt14230458/*)

7 Simulation

As the final step of my analysis, I estimated IMDb Ratings using the parameters of my model and then plot observed IMDb Ratings against the fitted values (Figure 18). if the points lies on the black bisector, it indicates that the parameters of my model accurately estimate the actual IMDb Ratings. However, the plot reveals that my model's estimation is not entirely accurate, as the cloud of points deviates from the bisector. This outcome was expected, as I did not solve the issue of non-constant variance highlighted in Chapter 5.3 and for the reasons mentioned in Chapter 8.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

```
n = nrow(Dataset) #number of observations
set.seed(12)
beta = coefficients(ols_best) #associate the betas to the values estimated through ols
X = model.matrix(ols_best)
y = X %*% beta + rnorm(n, 0, sigma(ols_best))
```

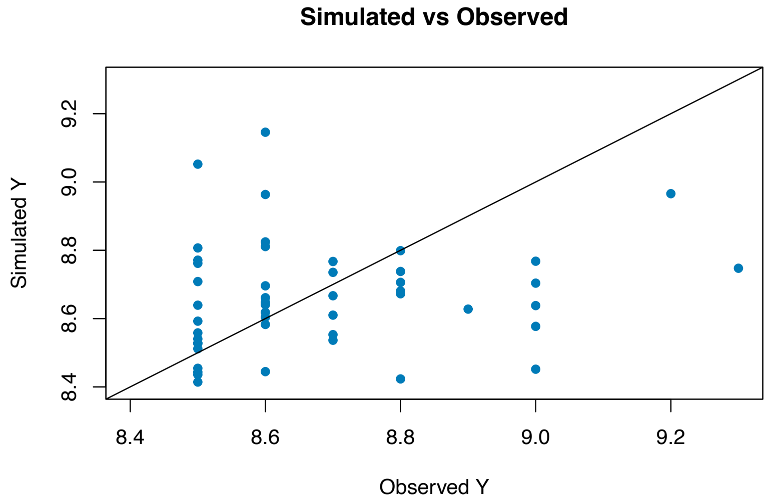


Figure 18