Assignment_5

Md Shamiul Islam_ID 0743469

2022-08-01

Question 1

Required Data: movies.csv

Looking for a way to predict box office receipts, an MGM producer collects the production costs, promotional costs, and book adaptation sales for 10 randomly sampled blockbuster movies, as well as their box office ticket sales (in millions of dollars). This year, he's pulling out all the stops on his newest feature film. He is planning to spend 15 million on production costs, 20 million on promotional costs, and hopes to make 5 million on sales of book adaptations.

Q1(a):

Fit a model to help predict the expected box office ticket sales. You do not need to perform any variable selection. Fit the response using the provided predictors and state the model equation

Answer Q1(a):

Data

Let's have a look at our data set and try to guess appropriate model fit according to the given question.

```
movies <- read.csv("movies.csv")
head(movies, 5)
```

```
##     Box Production Promotional Books
## 1  85.1        8.5    5.100000   4.7
## 2 106.3       12.9    5.800000   8.8
## 3  50.2        5.2    2.100000  15.1
## 4 130.6       10.7    8.399999  12.2
## 5  54.8        3.1    2.900000  10.6
```

```
str(movies)
```

```
## 'data.frame':    10 obs. of  4 variables:
##  $ Box        : num  85.1 106.3 50.2 130.6 54.8 ...
##  $ Production : num  8.5 12.9 5.2 10.7 3.1 ...
##  $ Promotional: num  5.1 5.8 2.1 8.4 2.9 ...
##  $ Books      : num  4.7 8.8 15.1 12.2 10.6 ...
```

According to the question and data set, we are interested in variable `Box` and need to establish an equation with other predictor variables and proper co-efficient. We will expand on the equation below by fitting a **multiple regression** model.

$$Box_i = \beta_0 + \beta_1 Production_i + \beta_2 Promotion_i + \beta_3 Book_i + \epsilon_i$$

Variables defined as below at i-th state:

- Box_i is box office ticket sale
- β_0 is intercept
- β_1 is slope of Production
- β_2 is slop of Promotion
- β_3 is slop of Book
- $Production_i$ is the cost of production
- $Promotion_i$ is the cost of promotion
- $Book_i$ is the book adaption sale
- ϵ_i represents noise.

We shall assume significance level $\alpha = 0.05$ as the question paper does not mention any significance level.

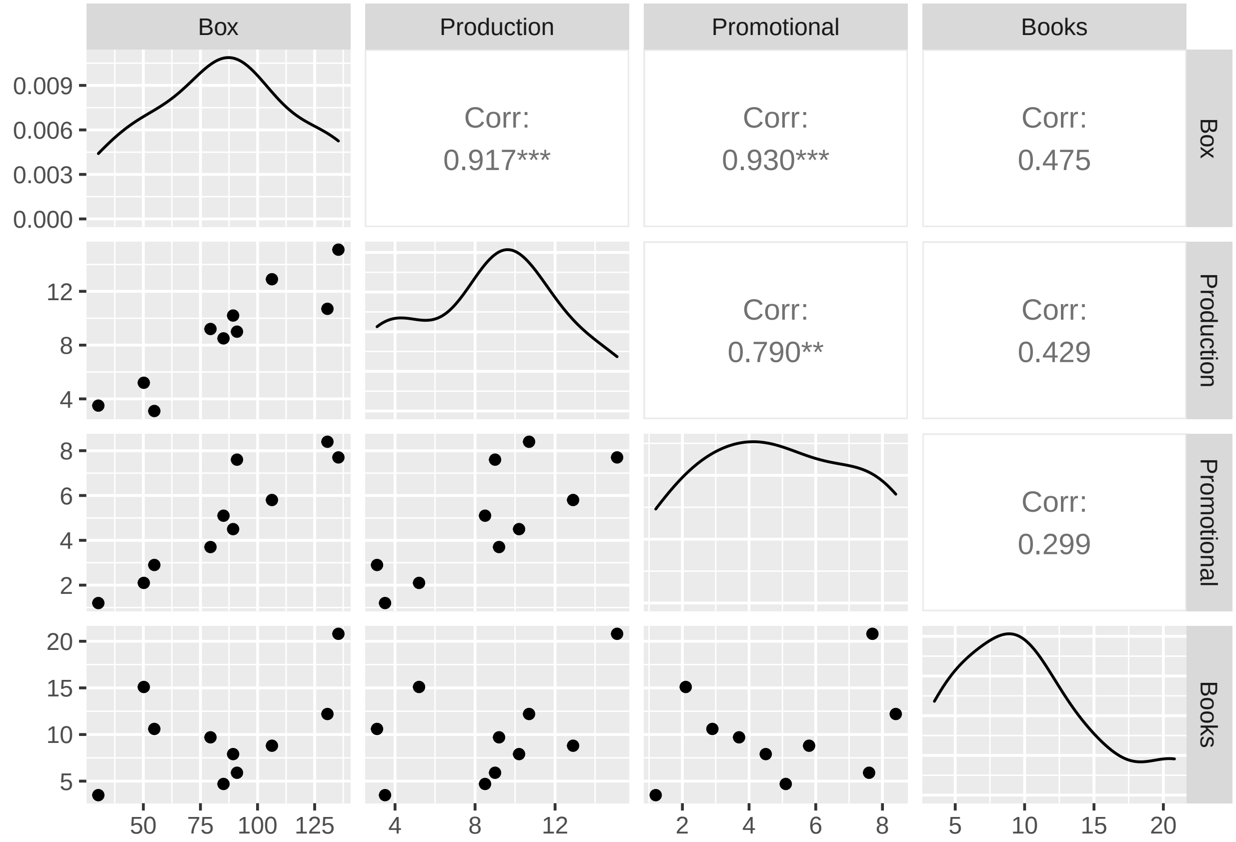
EDA:

Exploratory Data Analysis (EDA) is visual or tabular presentation which gives us overall view of our data. We can visualize pair relationship of response variable and predictor variable and make assumptions of our planed model. I have used `ggpairs` function for this analysis.

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(movies)
```



Multiple Regrassion Conditions:

Liniarity:

A linear relationship should exist between each predictor and the response. From the plot above, we see that data points are reasonably linearly distributed. Hence, multiple regression should be a good model fit.

Uncorelaetd predictors:

Predictor **Promotional** & **Production** cost has fairy strong correlation among them. Correlation between Other predictors reasonably week, so the predictors are multicollinearity free.

Normality:

Assume that observations and independent. All the variables seemingly normally distributed except the variable **Books** which seems skewed. The logic applies to the distribution of residuals.

Model Fitting:

Assuming the above conditions are met, we are fitting multiple regression model as below:

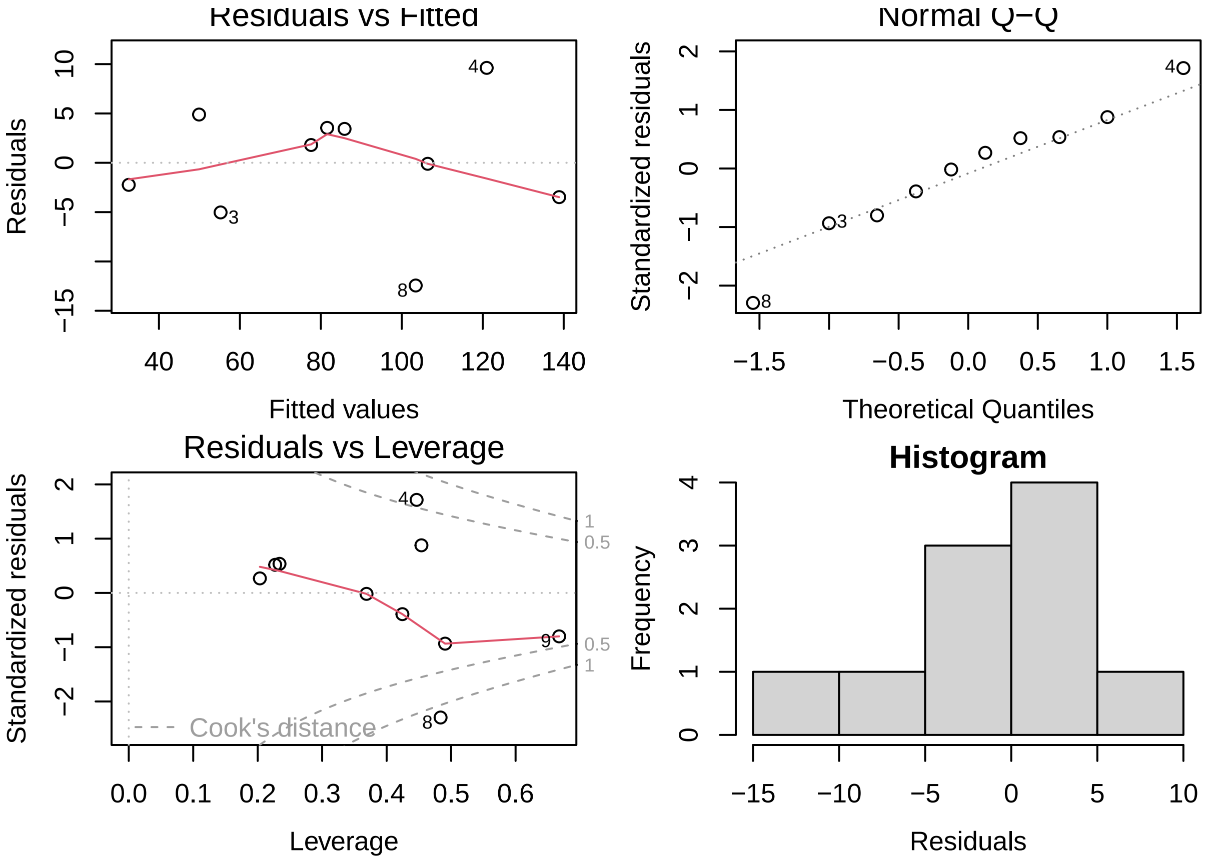
```r
fit1 <- lm(Box ~ Production + Promotional + Books, data = movies)
summary(fit1)
```

```
##
## Call:
## lm(formula = Box ~ Production + Promotional + Books, data = movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4384  -3.1695   0.8499   3.5134   9.6207
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.6760     6.7602   1.135   0.2995
## Production    3.6616     1.1178   3.276   0.0169 *
## Promotional   7.6211     1.6573   4.598   0.0037 **
## Books         0.8285     0.5394   1.536   0.1754
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.541 on 6 degrees of freedom
## Multiple R-squared:  0.9668, Adjusted R-squared:  0.9502
## F-statistic: 58.22 on 3 and 6 DF,  p-value: 7.913e-05
```

The summary of the model shows an overall model $p-value = 7.913e-05$ which indicates a good model fit. However, `Books` seems to be statistically insignificant as it carries a $p-value$ higher than the significance level of $\alpha = 0.05$. $Adjusted\ R^2 = 0.9502$ which means 95.02% of the variability of predictor variables has been explained by the model. It seems like a good fit. Diagnostic plots will help us visualise the model better. Let’s draw diagnostic plots.

Diagnostic plot:

```r
par(mar = c(4,4,1,1), mgp = c(2.5, 1, 0), mfrow = c(2, 2))
plot(fit1, which = c(1,2,5))
hist(fit1$residuals, breaks = 5, main = "Histogram", xlab = "Residuals", ylab = "Frequency")
```



The **constant variability** appears to be fine based on the Fitted Values and Residuals plot because the residuals are distributed evenly; however, determining linearity is difficult due to the small number of observations.

We can have a sense of **Linearity** from the **variability**, residuals seem to have moderately equal variability. The 'Residual Vs Leverage' plot shows a reasonable linear distribution of the residuals.

The QQ plot and histogram show that the data are not that skewed, which indicates that the residuals are approximately **normally distributed**.

Model Interpretation & Equation:

The model we fitted seems to be a good model fit as all the assumptions are met, model statistics are fine, and diagnostic plots also align with conditions. We shall construct a model equation and use the same for prediction.

```
summary(fit1)
```

```
##
## Call:
## lm(formula = Box ~ Production + Promotional + Books, data = movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4384  -3.1695   0.8499   3.5134   9.6207
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.6760     6.7602   1.135   0.2995
## Production    3.6616     1.1178   3.276   0.0169 *
## Promotional   7.6211     1.6573   4.598   0.0037 **
## Books         0.8285     0.5394   1.536   0.1754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.541 on 6 degrees of freedom
## Multiple R-squared:  0.9668, Adjusted R-squared:  0.9502
## F-statistic: 58.22 on 3 and 6 DF,  p-value: 7.913e-05
```

$$\hat{Box} = 7.6760 + 3.6616 * Production + 7.6211 * Promotional + 0.8285 * Books$$

Coefficient Interpretations

- **Intercept** ($\beta_0 = 7.6760$): Given all other factors constant, if there is zero production cost, promotional cost, and book sales, the average ticket sale is 7.6760 million.
- **Slope** ($\beta_1 = 3.6616$): Given all other factors constant, for every 1 million increase in Production cost, on average, we would expect Box office sales to increase by 3.6616 million.
- **Slope** ($\beta_2 = 7.6211$): Given all other factors constant, for every 1 million increase in promotional cost, on average, we would expect Box office sale to increase by 7.6211 million.
- **Slope** ($\beta_3 = 0.8285$): Given all other factors constant, on average, for every 1 million increase in book sale, we would expect Box office sale to increase by 0.8285 million.

Q1 (b)

Comment on the model fit. Does this seem to be a good model? Why or why not? What would you change?

Answer Q1(b):

Model Output:

```
summary(fit1)
```

```
##
## Call:
## lm(formula = Box ~ Production + Promotional + Books, data = movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4384  -3.1695   0.8499   3.5134   9.6207
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.6760     6.7602   1.135   0.2995
## Production    3.6616     1.1178   3.276   0.0169 *
## Promotional   7.6211     1.6573   4.598   0.0037 **
## Books         0.8285     0.5394   1.536   0.1754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.541 on 6 degrees of freedom
## Multiple R-squared:  0.9668, Adjusted R-squared:  0.9502
## F-statistic: 58.22 on 3 and 6 DF,  p-value: 7.913e-05
```

The model seems to be a good one. If we look at the overall p value of model, a small $p = 7.913e - 05$ value indicates that the model is statically significant. At the same time *adjusted* $R^2 = 0.9502$ indicates that 95.02% of the variability has been addressed by the model, which is very good. Large residual indicates high average distance of observed values from the regression line. The predictor **Books** has an insignificant p-value, which means it has the least effect on the response variable. Overall, the model seems to be a good fit. However, we may apply **Principle of Parsimony** and remove insignificant predictor **Books** to see the effect on the response variable to take a firm decision.

Model without **Books** variable:

```
rev_model <- lm(Box ~ Production + Promotional, data = movies)
summary(rev_model)
```

```
##
## Call:
## lm(formula = Box ~ Production + Promotional, data = movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4168  -2.5696   0.8052   2.1200  11.0463
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.848      6.765   1.751  0.12334
## Production     4.228      1.153   3.667  0.00800 **
## Promotional    7.436      1.806   4.117  0.00448 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.241 on 7 degrees of freedom
## Multiple R-squared:  0.9537, Adjusted R-squared:  0.9405
## F-statistic: 72.14 on 2 and 7 DF,  p-value: 2.131e-05
```

We found minor changes in outcome of the revised model, lets analyse in details.

1. Overall p value decreased form 7.913e-05 to 2.131e-05 which is a good indication.
2. *Adjusted* R^2 decreased form 0.9502 to 0.9405 which is not a good indication.
3. Residual increases form 7.541 to 8.241 which is also not a good indication.

We found minor changes in the outcome of the revised model. Let's analyse in detail.

In this revised model, we find a good indication with the $p - value$ but a bad indication with the other two parameters of model fit. Hence, we can say that the model we fitted earlier is a better fit than the revised version of the model.

Q1(c)

Calculate a 93% parametric confidence interval for the true model parameter associated with the Production variable

Answer Q1(c)

For β_1 = Coefficient of Production (true model parameter):

$$b_1 \pm t_6^{\star} SE_{b_1}$$

From the model summary, we will find standard error (SE), degree of freedom, and coefficient.

```
summary(fit1)
```

```
##
## Call:
## lm(formula = Box ~ Production + Promotional + Books, data = movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4384  -3.1695   0.8499   3.5134   9.6207
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.6760     6.7602   1.135   0.2995
## Production    3.6616     1.1178   3.276   0.0169 *
## Promotional   7.6211     1.6573   4.598   0.0037 **
## Books         0.8285     0.5394   1.536   0.1754
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.541 on 6 degrees of freedom
## Multiple R-squared:  0.9668, Adjusted R-squared:  0.9502
## F-statistic: 58.22 on 3 and 6 DF,  p-value: 7.913e-05
```

93% CI Calculation:

```
library(s20x)
alpha <- 0.07
SE <-  1.1178
df1<- 6
beta_1 <- fit1$coefficients[2]

t_star <- qt((alpha/2), df = df1, lower.tail = FALSE)

CI <- beta_1 + c(-1, 1) * t_star * SE
round(CI, 3)
```

```
## [1] 1.201 6.122
```

```
ciReg(fit1, conf.level = 0.93, print.out = TRUE)
```

```
##             93 % C.I.lower    93 % C.I.upper
## (Intercept)       -7.20363          22.55569
## Production         1.20137           6.12184
## Promotional        3.97320          11.26890
## Books             -0.35869           2.01563
```

Interpretation:

We are 93% confident that, for the true model (slop) parameter associated with the Production variable lies in the interval of `2.971 to 4.352`

Q(d)

Provide the predicted box office ticket sales for his movie.

Answer Q(d)

Producer is planning to spend `15` million on production costs, `20` million on promotional costs, and hopes to make USD5 million on sales of book adaptations.

```
Production <- 15   # Production cost
Promotional <- 20 # Promotional cost
Books <- 5 # Book sale

new_data <- data.frame(Production, Promotional, Books)
```

We have created a new data frame with the new values of predictor variables to predict the data based on our observed data. The observed data is already fitted into the model defined as `fit1`.

```
predict(fit1, newdata = new_data )
```

```
##        1
## 219.1635
```

As per the prediction, the box office ticket sales for the movie would be **219.1635** million given the conditions of the said variables Production, Promotion and Books.

Q1(e)

Calculate the 90% prediction interval for d).

Answer Q(e)

Here is the prediction interval for the new data frame created in the Q1(d)

```
pred_interval <- predict(fit1, newdata =  new_data,
                interval = "prediction", level = 0.90)
pred_interval
```

```
##        fit      lwr      upr
## 1 219.1635 176.6536 261.6733
```

Interpretation:

We are 90% confident that the coefficient (slope) associated with the new `Prediction` value lies between 176.6536 to 261.6733.