5240 R Assignment 02

Digvijay Jondhale 0862899

Q1: Visualisation, Interpretion, Fitting, And Prediction of movie dataset

Introduction

The Movies dataset contains two variables, namely Box (Revenue made at box office (M dollars)) and Production (Amount invested in production (M dollars)).From this we have to analyze the relationship between the two variables, fit a model and interpret the results.

Loading the data

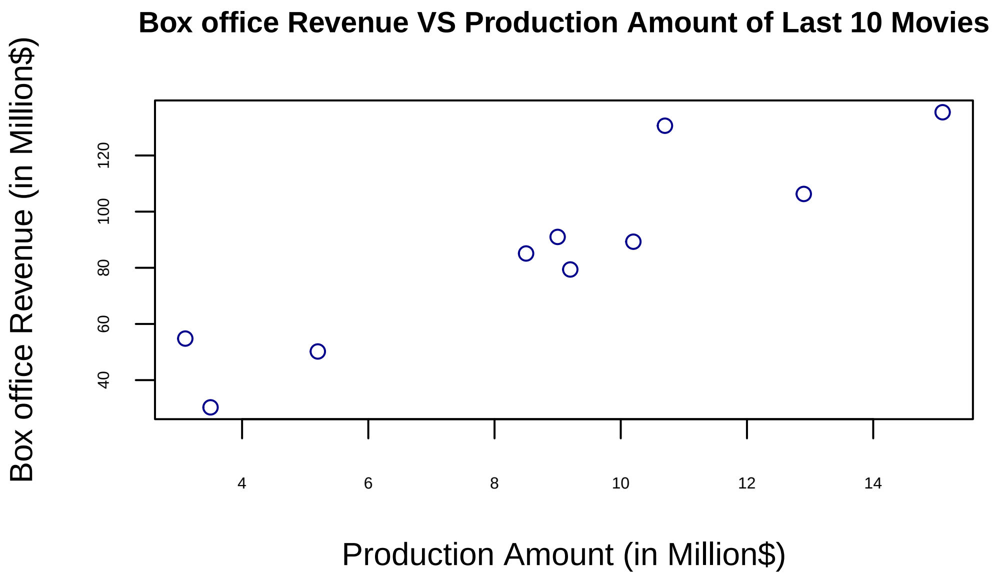
```
movies_db <- read.csv(file="movies.csv", header = TRUE)
```

a) Box office Revenue VS Production Budget

The plot below shows the relationship between Box office revenue on y axis and production amount on x axis for 10 movies of the dataset.

- Since the box office revenue is a future event and is of the interest to predict using the production amount, so box office will be the dependent variable, x will be independent variable, and the relationship will be y is a function of x.

```
plot(x= movies_db$Production, y=movies_db$Box,
     col="darkblue",
     xlab="Production Amount (in Million$)",
     ylab = "Box office Revenue (in Million$)",
     main = "Box office Revenue VS Production Amount of Last 10 Movies",
     cex.axis = 0.5,
     cex.main = 0.9)
```



b) Association between the variable from the above plot :

```
cor(y=movies_db$Box , x=movies_db$Production)
```

```
[1] 0.9174448
```

From the above plot we can clearly see the following : * A positive Relationship : As production Amount (x) increase , there seems to be increase in box-office revenue (y).

- And the covariance calculated above is 0.917 , which means it is a strong linear relationship. So there is strong possibility of a movies generating more revenue at box office if the investors invest large amount of money.

c) Fitting a Linear Model :

In this section we will try to fit a linear model (using lm() function) between the two variables, where y will be the output variable and x will be independent variable.

```r
model_fit <- lm(Box ~ Production,data = movies_db)
```

Summary of the fitted model :

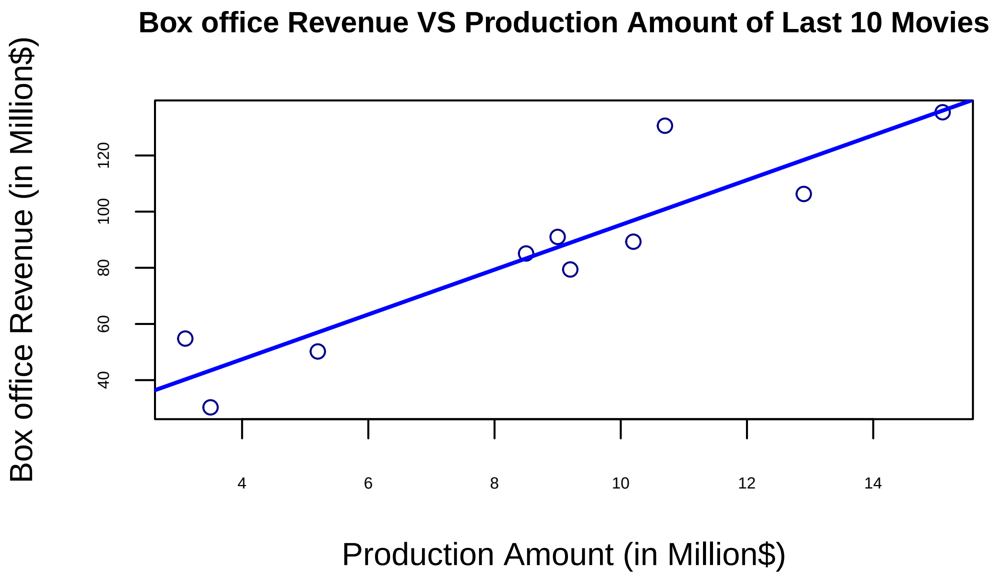
```r
summary(model_fit)
```

```
Call:
lm(formula = Box ~ Production, data = movies_db)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.136  -9.029  -3.689   3.208  29.723 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   15.513     11.603   1.337 0.217989    
Production     7.978      1.223   6.522 0.000184 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.26 on 8 degrees of freedom
Multiple R-squared:  0.8417,	Adjusted R-squared:  0.8219 
F-statistic: 42.54 on 1 and 8 DF,  p-value: 0.0001838
```

d) Equation of the Model :

$$box\,\hat{office} = 15.513 + 7.978 \cdot \hat{production}$$

```r
plot(x= movies_db$Production, y=movies_db$Box,
     col="darkblue",
     xlab="Production Amount (in Million$)",
     ylab = "Box office Revenue (in Million$)",
     main = "Box office Revenue VS Production Amount of Last 10 Movies",
     cex.axis = 0.5,
     cex.main = 0.9)
abline(a = 15.513, b = 7.978, col = "blue", lwd = 2)
```



e) About the Intercept :

The intercept from the summary is 15.513 , which means if a movies invests **zero** amount of money in production , then also it will be able to make **15.513 million$** revenue at box-office.

Note : *Now in reality this seems weird, but considering our sample data size and given statistics this is the result.*

f) About the slope :

The slope term in the model describes the rate by which the box office revenue changes with respect to the production amount.

- The slope is 7.978, which means for every 1 million dollars invested in production , the box office revenue will generate 7.978 million$.

g) Predicting Box office revenue for 20 million$ investment :

To predict the revenue , we can either use the equation above or use the predict() function of R inbuilt library.

By using predict() function :

```
# This (x) is the production amount in million$
production_amt <- data.frame(Production = 20)
prediction <- predict(model_fit, newdata = production_amt)
print(prediction)
```

```
       1
175.0715
```

The movie will be able to generate **175.07** million dollars in revenue at the box office if the investors invests 20 million$ in the movie.

Now this does not mean that if the investors keep on investing any amount of money into the movies , then they can receive same or always increasing amounts of revenue at box office, this will at some point get saturated and the the relationship will falten at some point.

Q2 Squirrels

a) Prbability of Catching 7 squirrels :

For all the solution below , size referes to the sample size which is 12 , and prob referes to the given probability which is 12.

```
pro_7sq <-dbinom(7, size = 12, prob = 0.70)
print(pro_7sq)
```

```
[1] 0.1584958
```

Probability of catching 7 squirrels os 0.158 In the above , we have used simple probability by using dbinom.

b) Prbability of Catching at most 4 squirrels :

```
pro_at_most_4 <- sum(dbinom(0:4, size = 12, prob = 0.70))
print(pro_at_most_4)
```

```
[1] 0.009489371
```

Probability of catching at most 4 squirrels is : 0.0094. Here we have calculated the Probability of each number of squirrels from 0 , 4 and added them. We have used Probability Mass Function.

c) Prbability of Catching between 4 and 10 squirrels :

```
pro_4_to_10 <- sum(dbinom(5:9, size = 12, prob = 0.70))
print(pro_4_to_10)
```

```
[1] 0.7376953
```

The Probability of David catching 4 to 10 squirrels is 0.73 and that is quite high!.

$4<P(X)<10$

d) Prbability of catching odd number of squrrels :

```
pro_odd <- sum(dbinom(seq(1, 11, by = 2), size = 12, prob = 0.70))
print(pro_odd)
```

```
[1] 0.4999916
```

Probability of David catching odd number of squirrels is 0.499

$P = 1,3,5,7,9,11$ i.e $P(1) + P(3) + ... P(11)$

e) Probability of catches less than 3 or more than 8 squirrels?

```
pro_less_3_more_8 <- (sum(dbinom(0:2, size = 12, prob = 0.70))
                      + sum(dbinom(9:12, size = 12, prob = 0.70)))

print(pro_less_3_more_8)
```

```
[1] 0.4927221
```

Probability of catches less than 3 or more than 8 squirrels is **0.4927221** $P(X<3 \text{ or } X>8)$