R Assignment:04

Keya Adhyaru

Question 1 - Movies

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.

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:

Fitting a Linear Regression Model on Box Office Collection using all other predictors as:

```
#Loading movies.csv file
library(readr)
movies <- read_csv("Path/to/movies.csv")
```

```
## Rows: 10 Columns: 4
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (4): Box, Production, Promotional, Books
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
box.lm <- (lm(Box ~ Production+ Promotional+ Books, data= movies))

summary(box.lm)
```

```
##
## 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
```

By conducting multiple linear regression model, we observe different intercept for different predictor variable and further form following model equation:

Box Office Tickets = 7.676 + 3.662* Production + 7.621* Promotional + 0.828*Books

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

Answer:

The R^2 of the model = 0.9667, which implies that 96.67% of variation in the box office collection can be explained by variation in all other predictor variables.

Adjusted, which implies that all the variables are significant in predicting dependent variable.

p value of the model = 0.0000 < 0.05 (at 5% significance level), hence we have enough evidence to reject the null hypothesis and conclude that model is significant.

This shows that model is a good fit on the data.

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

Answer:

Constructing 93% confidence interval for the true slope in the linear model:

```
#new.dat <- data.frame(movies$Production)
#predict(box.lm , newdata = new.dat, interval = 'confidence')
confint(box.lm, parm = "Production", level = 0.93)
```

```
##                3.5 %   96.5 %
## Production 1.201367 6.121841
```

The 93% confidence interval for the true model parameter associated with the Production variable is (1.201, 6.121)

d) Provide the predicted box office ticket sales for his movie. Answer:

According to the model equation, the predicted box office ticket sales for his movie can be calculated as follows:

Box Office Tickets = 7.676 + 3.662* 15 + 7.621* 20 + 0.828*5 7.676+ 54.93 + 152.42+ 4.14 Box Office Ticket= 219.66

e) Calculate the 90% prediction interval for d).

Answer:

Calculating 90% prediction interval for the model

```
Production = 15
Promotional = 20
Books = 5
test <- as.data.frame(cbind(Production,Promotional,Books))
predict(box.lm,  newdata = test,  interval = "prediction",level = 0.90)
```

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

The 90% prediction interval for the model lies between 176.65 and 261.67

Question 2 - Credit Cards

A financial institution that issues credit cards for subprime borrowers wants to identify its credit card applicants who do not exceed a default chance threshold of 30% to approve an application. It randomly selected 41 past credit card holders and investigated their monthly salary, monthly debt, and marital status at the time of issuance of its credit card and whether they defaulted after taking the credit card. (In the data, Default = 1 means the customer defaulted and 0 otherwise; Marital = 1 means married and 0 otherwise.)

a) Identify the response variable and the predictor variables. Answer:

Observing the given data, we want to find the status of default of credit card for subprime borrowers so from the problem statement we find that 'Default' is the response variable and 'Salary', 'Dept', 'Marital' are predictor variable.

b) Determine the logistic regression model for the purpose of the institution.

Answer:

```
#Loading DefaultRate.csv file
library(readr)
DefaultRate <- read_csv("Path/to/DefaultRate.csv")
```

```
## Rows: 41 Columns: 4
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (4): Default, Salary, Debt, Marital
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Conducting logistic regression model for binomial response variable:

```
default.glm <- (glm(Default ~ Salary+ Debt+ Marital, data = DefaultRate, family = binomial))
summary(default.glm)
```

```
##
## Call:
## glm(formula = Default ~ Salary + Debt + Marital, family = binomial,
##     data = DefaultRate)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1919  -0.6491  -0.3812   0.6160   3.0572
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.979576   1.209569   0.810  0.41802
## Salary      -0.002301   0.000792  -2.905  0.00367 **
## Debt         0.003624   0.001256   2.886  0.00390 **
## Marital     -2.631395   1.122488  -2.344  0.01907 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##   Null deviance: 49.572  on 40  degrees of freedom
## Residual deviance: 33.825  on 37  degrees of freedom
## AIC: 41.825
##
## Number of Fisher Scoring iterations: 5
```

c) Does the institution issue its credit card to a married customer with monthly salary of 2000 and monthly debt of 1400.

Answer

```
new.data <- data.frame(Salary=2000, Debt=1400, Marital=1)
predict.glm(default.glm,newdata= new.data, type='response')
```

```
##         1
## 0.2351833
```

The probability that institution issue its credit card to a married customer with monthly salary of 2000 and monthly debt of 1400 is 0.2351833 < 0.3 so institute will issue its credit card to this customer.

d) Does the institution issue its credit card to a married customer with monthly salary of 3208 and monthly debt of 2200.

Answer:

```
new.data <- data.frame(Salary=3208, Debt=2200, Marital=1)
predict.glm(default.glm,newdata= new.data, type='response')
```

```
##         1
## 0.2574774
```

The probability that institution issue its credit card to a married customer with monthly salary of 3208 and monthly debt of 2200 is 0.2574774 < 0.3 so institute will issue its credit card to this customer.

e) Does the institution issue its credit card to a single customer with monthly salary of 3408 and monthly debt of 1700.

Answer:

```
new.data <- data.frame(Salary=3408, Debt=1700, Marital=0)
predict.glm(default.glm,newdata= new.data, type='response')
```

```
##         1
## 0.3318034
```

The probability that institution issue its credit card to a married customer with monthly salary of 3408 and monthly debt of 1700 is 0.3318034 > 0.3 so institute will not issue its credit card to this customer.