Homework 6

Arjun Ganesh

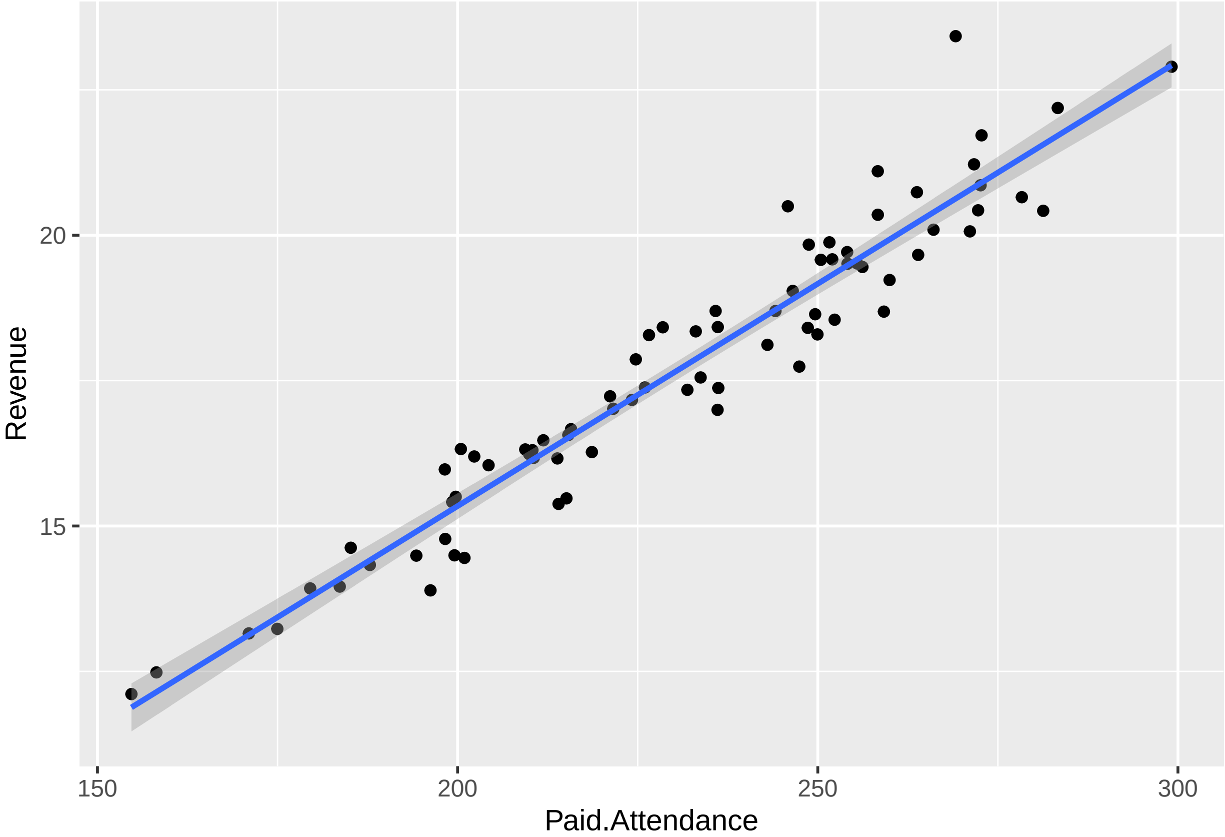
4/1/2020

1. [15] Viewing this a business, we want to build a simple linear regression model using Revenue as the response and Paid attendance as a predictor.

(A) [3] Create a scatter plot for the two variables with the regression line superimposed. Comment on what the scatter plot reveals (be sure to comment of the form of association, direction and strength - by calculating and explaining the correlation coefficient - of the relationship).

```
Bway<- read.csv( "/cloud/project/BroadwayShows.csv")
Bway <-na.omit(Bway)

library(ggplot2)
ggplot(Bway, aes(x=Paid.Attendance, y= Revenue))+geom_point()+
  geom_smooth(method ="lm")
```



```
cor(Bway$Paid.Attendance,Bway$Revenue)
```

```
## [1] 0.9610977
```

The correlation value is 0.96 which demonstrates a clear strong positive association between the variables.

(B) [3] Find and write down the simple linear regression model [Note: be sure to include the output of your code.]

```
linear_model<- lm(Revenue~Paid.Attendance, Bway)
summary(linear_model)
```

```
##
## Call:
## lm(formula = Revenue ~ Paid.Attendance, data = Bway)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.22826 -0.48244  0.02748  0.37756  2.79352
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.05261    0.58696    0.09    0.929
## Paid.Attendance  0.07645    0.00252   30.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6981 on 76 degrees of freedom
## Multiple R-squared:  0.9237, Adjusted R-squared:  0.9227
## F-statistic: 920.2 on 1 and 76 DF,  p-value: < 2.2e-16
```

(C) [2] Interpret, in context, the meaning of the slope coefficient in your model.

The meaning of the slope coefficient, 0.07645, is for every one thousand paid attendees, revenue increases by 76,450 dollars. So for every one attendee the revenue increases by 76.45 dollars.

(D) [3] Interpret, in context, the meaning of the R-Squared value in your model.

The r-squared value of 0.9237 which shows that model does fit our data and that 92.37% of the variabilty is accounted for. The higher the r squared value the better for us.

(E) [4] Use you model to predict (using the predict() function) the revenue for Broadway shows if 200 thousand people paid to watch a show at a given night.

```
what_if<-data.frame(Paid.Attendance=200)
predict(linear_model,what_if)
```

```
##        1
## 15.34279
```

2. [15] Viewing this a business, we want to build a multiple regression model using Revenue as the response variable and Paid attendance, Number of shows and average ticket price as predictor variables.

(A) [4] Use the AIC criteria to find and write down the multiple regression model [Note: be sure to include the output of your code.]

```
library(MASS)
Mult_reg <-lm(Revenue~Paid.Attendance+Number.of.Shows+Avg.Ticket.Price ,Bway)
summary(Mult_reg)
```

```
##
## Call:
## lm(formula = Revenue ~ Paid.Attendance + Number.of.Shows + Avg.Ticket.Price,
##     data = Bway)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26452 -0.03214 -0.00207  0.03789  0.31920
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.832e+01  3.127e-01 -58.582   <2e-16 ***
## Paid.Attendance   7.596e-02  6.291e-04 120.751   <2e-16 ***
## Number.of.Shows   7.028e-03  4.418e-03   1.591    0.116
## Avg.Ticket.Price  2.384e-01  3.907e-03  61.014   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09313 on 74 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9986
## F-statistic: 1.863e+04 on 3 and 74 DF,  p-value: < 2.2e-16
```

```r
step.model<-stepAIC(Mult_reg, direction = "both", trace = FALSE)
summary(step.model)
```

```
##
## Call:
## lm(formula = Revenue ~ Paid.Attendance + Number.of.Shows + Avg.Ticket.Price,
##     data = Bway)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26452 -0.03214 -0.00207  0.03789  0.31920
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.832e+01  3.127e-01 -58.582   <2e-16 ***
## Paid.Attendance   7.596e-02  6.291e-04 120.751   <2e-16 ***
## Number.of.Shows   7.028e-03  4.418e-03   1.591    0.116
## Avg.Ticket.Price  2.384e-01  3.907e-03  61.014   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09313 on 74 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9986
## F-statistic: 1.863e+04 on 3 and 74 DF,  p-value: < 2.2e-16
```

(B) [4] Interpret, in context, the meaning of the slope coefficients corresponding to all the predictor variables in the AIC model.

For paid attendance revenue will increase by 75,960 dollars for every 1000 people. For every additional show, revenue will increase by 7,028. For every $1 increase in ticket price, revenue will increase by 238,400 dollars.

(C) [3] Is the AIC model useful in predicting the revenue for broadway shows? Explain, briefly.

Yes AIC model is useful in specific prediction of a variable because it uses more relevant predictor variables necessary to give an output.

(D) [4] Pick some reasonable values for the predictor variables in your AIC model and use it to make a prediction (using the predict() function) for revenue.

```r
rand <- data.frame(Paid.Attendance = 300, Number.of.Shows = 20, Avg.Ticket.Price =75)
predict(Mult_reg,rand)
```

```
##        1
## 22.48796
```