Practice

Saksham Goel

February 3, 2018

Question 1 - Box Office Ticket Sales

Introducing the Dataset

This section is dedicated to see and understand the data that was provided from the weekly reports about the box office ticket sales for plays in Broadway in New York. The data being observed is of the week of October 11- 17, 2004. The dataset contains data about the gross box office results for the current week October 11-17, 2004 and that of the previous week October 3-10, 2004.

The data table is as follows:

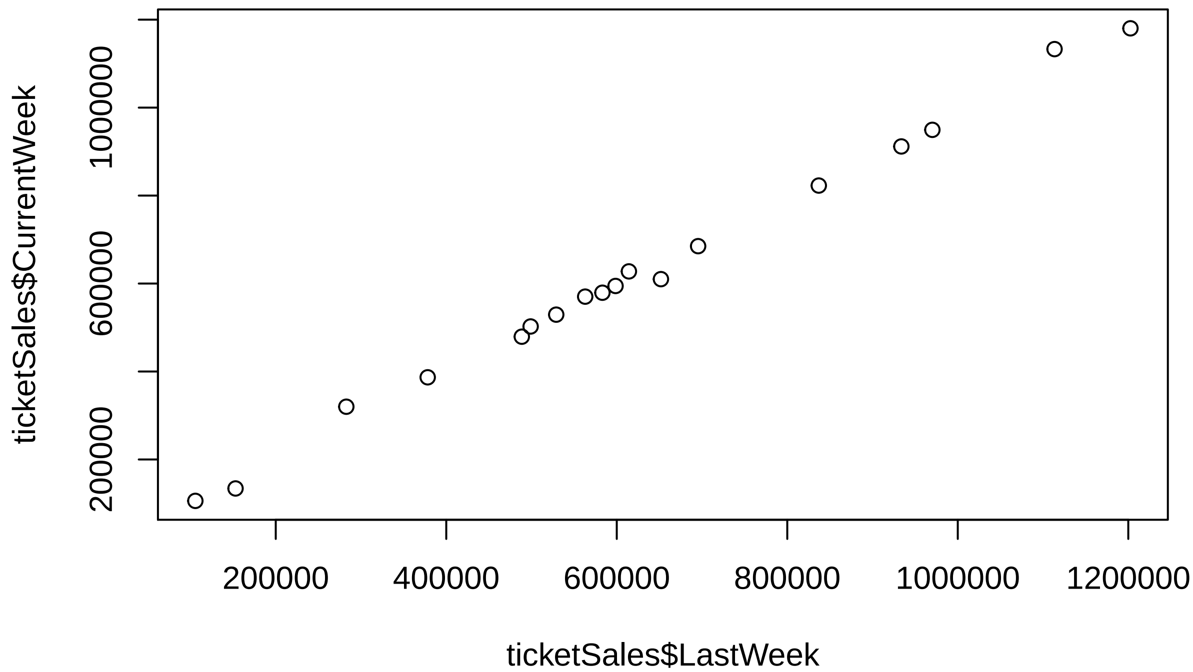
```r
ticketSales = read.csv("playbill.csv", header=T)
ticketSales[,]
```

```
##                    Production CurrentWeek LastWeek
## 1                 42nd Street      684966   695437
## 2                    Avenue Q      502367   498969
## 3            Beauty and Beast      594474   598576
## 4               Bombay Dreams      529298   528994
## 5                     Chicago      570254   562964
## 6                     Dracula      319959   282778
## 7          Fiddler on the Roof      579126   583177
## 8               Forever Tango      134042   152833
## 9             Golda's Balcony      105853   105698
## 10                  Hairspray      822775   836959
## 11                 Mamma Mia!      949462   970190
## 12                Movin' Out      610007   651808
## 13                       Rent      386797   378238
## 14              The Lion King     1133034  1113510
## 15 The Phantom of the Opera       627609   614246
## 16              The Producers      911727   933822
## 17                     Wicked     1180266  1202536
## 18             Wonderful Town      479155   488624
```

Visualizing the Dataset

The dataset can be visalized through a scatterplot in the figure below:

```r
plot(x = ticketSales$LastWeek, y = ticketSales$CurrentWeek)
```



Fitting a linear model

The linear trend in the scatterplot seems strong, which means that a linear regression model is appropriate. The linear model that we are trying to fit is of the form:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

here, X = Gross Box Office results for Previous Week Y = Gross Box Office results for Current Week.

Through this linear model we are trying to predict the value of actual gross box office reults of current week (Y) using the gross box office results of the previous week (X). The following snippet of the R-Code fits a linear model on the data, prints out the values of the intercept and slope and also provides a summary of the fitted model.

```
mod = lm( ticketSales$CurrentWeek ~ ticketSales$LastWeek )
mod
```

```
##
## Call:
## lm(formula = ticketSales$CurrentWeek ~ ticketSales$LastWeek)
##
## Coefficients:
##          (Intercept)  ticketSales$LastWeek
##            6804.8860                0.9821
```

After fitting the model and observing the values of the parameters $\hat{\beta}_0, \hat{\beta}_1$ we find that:

$$\hat{\beta}_0 = 6804.8860$$

$$\hat{\beta}_1 = 0.9821$$

The summary of the model is as follows:

```
summary(mod)
```

```
##
## Call:
## lm(formula = ticketSales$CurrentWeek ~ ticketSales$LastWeek)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36926  -7525  -2581   7782  35443
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.805e+03  9.929e+03   0.685    0.503
## ticketSales$LastWeek 9.821e-01  1.443e-02  68.071   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18010 on 16 degrees of freedom
## Multiple R-squared:  0.9966, Adjusted R-squared:  0.9963
## F-statistic:  4634 on 1 and 16 DF,  p-value: < 2.2e-16
```

The summary provides us with a lot of useful values that we will use in the upcoming sections to calculate the Confidence Intervals, Prediction Intervals and perform Hypothesis Testing. Some of the values are as follows:

Sum of Residuals Square (s) = 18010\ Degrees of freedom (n-2) = 16\

Combining all the information from the previous sections the final fitted model looks as follows:

$$Y = 6804.8860 + (0.9821 * X)$$

Finding a 95% Confidence Interval for β_1

A $100(1-\alpha)\%$ confidence interval for $\hat{\beta}_1$ is given by the following formula:

$$\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\frac{s}{\sqrt{S_{XX}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\frac{s}{\sqrt{S_{XX}}}$$

The following snippet of code is used to find the 95% confidence interval for $\hat{\beta}_1$:

```
bet1_hat = 0.9821 #found using the linear model
x_col = ticketSales$LastWeek
y_col = ticketSales$CurrentWeek
x_bar = mean(x_col)
#x_bar = 622186.6
y_bar = mean(y_col)
#y_bar = 617842.8
sxx = sum((x_col-x_bar)^2)
#sxx = 1.557916 * 10^12
sxy = sum((x_col-x_bar)*(y_col-y_bar))
#sxy = 1.53 * 10^12
#sxy/sxx = 0.9821
s = 18010 #found through the summary of the linear model
t_mult = 2.120 # found using the T Table, equal to t0.25 for 16 degrees of freedom
beta1CIlower = bet1_hat - (t_mult*(s/sqrt(sxx)))
beta1CIupper = bet1_hat + (t_mult*(s/sqrt(sxx)))
```

The values we found are as follows:

$\bar{x} = 622186.6$

$\bar{y} = 617842.8$

$S_{XX} = 1.557916 * (10^12)$

$S_{XY} = 1.53 * (10^12)$

$\hat{\beta_1} = 0.9820815$

The 95% confidence interval can thus be given as follows:

$$0.9515101 \leq \beta_1 \leq 1.01269$$

Yes we can say that 1 is a reasonable value for β_1 because 1 lies in the 95% Confidence Interval of β_1 ### Hypothesis Testing for β_0 We need to run a Hypothesis Test for β_0 given by:

$$Null(H_0)\beta_0 = 10000 Alternate(H_1)\beta_0 \neq 10000$$

The hypothesis test for β_0 is given by the following formula:

$$\hat{T} = \frac{\hat{\beta_0} - \beta_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}}$$

To find the p-value we then use the t-statistic we got from the previous formula and find the probability as follows:

$$p - val = 2 \cdot P\left(t \geq \mid \hat{T} \mid\right)$$

The following snippet of code is used to perform the Hypothesis Test for β_0:

```
bet0_hat = 6804.8860 #this value was found using the linear model through r in the previous section
x_col = ticketSales$LastWeek
x_bar = mean(x_col)
#x_bar = 622186.6
sxx = sum((x_col-x_bar)^2)
#sxx = 1.557916 * 10^12
s = 18010 #found through the summary of the linear model
t_mult = 2.120 # found using the T Table, equal to t0.25 for 16 degrees of freedom
n = length(x_col)
beta0test = ((bet0_hat - 10000)/(s * sqrt((1/n) + ((x_bar^2)/sxx))))
#beta0test =  -0.3217422
p_val = 2*pt(-abs(beta0test), df = n-2)
#p_val = 0.7518132
```

After performing the Hypothesis test for β_0 we find that the $p - value = 0.7518132$ which is greater than 0.05, hence we cannot reject the null hypothesis.

Point Estimate for new Y

Through the previous sections we know that:

$$Y = 6804.8860 + (0.9821 * X)$$

So to find the point estimate of Y using the fitted model we get:

```
X = 400000 #X value for which we need to find the estimated value
Y = 6804.8860 + (0.9821 * X)
#Y = 399644.9
```

Through this above snippet of code we get the point estimate of Y at $X = 400,000$ as $Y = 399644.9$.

Prediction Interval for new Y

There is no use of a Linear Model if we cannot predict new values of the parameter Y through X. In this section we will construct a prediction interval of the Y value using the formulas provided in the book. The prediction interval of Y is given by the formula as follows:

$$\hat{y}_{n+1} \pm t_{\alpha/2,n-2}\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

The following snippet of code helps us to find the upper and lower limits of the Prediction Interval for the new Y as follows:

```
#fitted model =
# Y = 6804.8860 + 0.9821 * X
x_curr = 400000 #the value at which we need to find the prediction interval
y_hat = 6804.8860 + (0.9821 * x_curr)
#y_hat = 399644.9.
mse = mean(mod$residuals^2)  # mean((ticketSales$CurrentWeek - predict(mod))^2)
#mse = 288241878
x_col = ticketSales$LastWeek
x_bar = mean(x_col)
#x_bar = 622186.6
sxx = sum((x_col-x_bar)^2)
#sxx = 1.557916 * 10^12
s = 18010 #found through the summary of the linear model
t_mult = 2.120 # found using the T Table, equal to t0.25 for 16 degrees of freedom
n = length(x_col)
yPILower = y_hat - (t_mult * sqrt(mse) * sqrt(1 + (1/n) + (((x_curr - x_bar)^2)/sxx) ))
yPIUpper = y_hat + (t_mult * sqrt(mse) * sqrt(1 + (1/n) + (((x_curr - x_bar)^2)/sxx) ))
#yPILower = 362115
#yPIUpper = 437174.8
```

Through the above snippet of code the resulting 95% prediction interval is given as follows: