STAT 3032 - HW1

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:

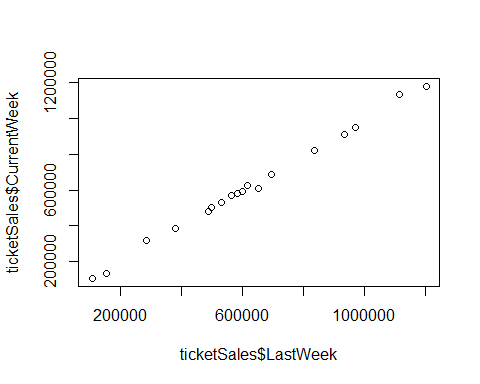
ticketSales = read.csv("playbill.csv", header=T)  
ticketSales[,]

## 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:

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:

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 we find that:

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:

### Finding a 95% Confidence Interval for

A confidence interval for is given by the following formula: $ - t\_{ ,n-2} *1 + t*{ ,n-2} $

The following snippet of code is used to find the 95% confidence interval for :

bet1\_hat = 0.9821  
x\_col = ticketSales$LastWeek  
y\_col = ticketSales$CurrentWeek  
x\_bar = mean(x\_col)  
#x\_bar  
y\_bar = mean(y\_col)  
#y\_bar  
sxx = sum((x\_col-x\_bar)^2)  
#sxx  
sxy = sum((x\_col-x\_bar)\*(y\_col-y\_bar))  
#sxy  
#sxy/sxx  
s = 18010  
t\_mult = 2.120  
beta1CIlower = bet1\_hat - (t\_mult\*(s/sqrt(sxx)))  
beta1CIupper = bet1\_hat + (t\_mult\*(s/sqrt(sxx)))  
#beta1CIlower  
#beta1CIupper

The values we found are as follows: x\_bar = 622186.6 y\_bar = 617842.8 SXX = 1.557916 \* (10 ^12) SXY = 1.53 \* (10 ^12) bet1\_hat = 0.9820815

The 95% confidence interval can thus be given as follows : (0.9515101, 1.01269). Yes we can say that 1 is a reasonable value for because 1 lies in the 95% Confidence Interval of

### Hypothesis Testing for

Doing the Hypothesis testing for the we get:

bet0\_hat = 6804.8860  
x\_col = ticketSales$LastWeek  
x\_bar = mean(x\_col)  
sxx = sum((x\_col-x\_bar)^2)  
n = length(x\_col)  
s = 18010  
t\_mult = 2.120  
beta0test = ((bet0\_hat - 10000)/(s \* sqrt(1 + (1/n) + ((x\_bar^2)/sxx))))  
beta0test

## [1] -0.1553558

2\*pt(-abs(beta0test), df = n-2)

## [1] 0.8784839

The values we found are as follows: x\_bar = 622186.6 SXX = 1.557916 \* (10 ^12) beta0\_hat = 6804.8860 beta0test = -0.1553558 p\_val = 0.8784839

The Hypothesis test for gives us a p-value = 0.8784839 which is greater than 0.05 which means that we cannot reject the null hypothesis which says = 10,000

### Prediction Interval for new Y

Finding the Prediction interval for the new Y we get:

#fitted model =   
# Y = 6804.8860 + 0.9821 \* X  
x\_curr = 400000  
y\_hat = 6804.8860 + (0.9821 \* x\_curr)  
mse = mean(mod$residuals^2) # mean((ticketSales$CurrentWeek - predict(mod))^2)  
  
x\_col = ticketSales$LastWeek  
x\_bar = mean(x\_col)  
sxx = sum((x\_col-x\_bar)^2)  
n = length(x\_col)  
s = 18010  
t\_mult = 2.120  
  
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

## [1] 362115

yPIUpper

## [1] 437174.8

The values we found are as follows: yPILower = 362115 yPIUpper = 437174.8

Because the prediction interval does not contain the value 450,000 then we cannot use that value as a prediction.

### Comment:

I think that is a statement that could be used frequently to describe the sales record for the broadway show, because the linear model also seems to follow a trend as seen. Other than that all the results from the previous computations seems to provide a good enough evidence to support this claim.