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:

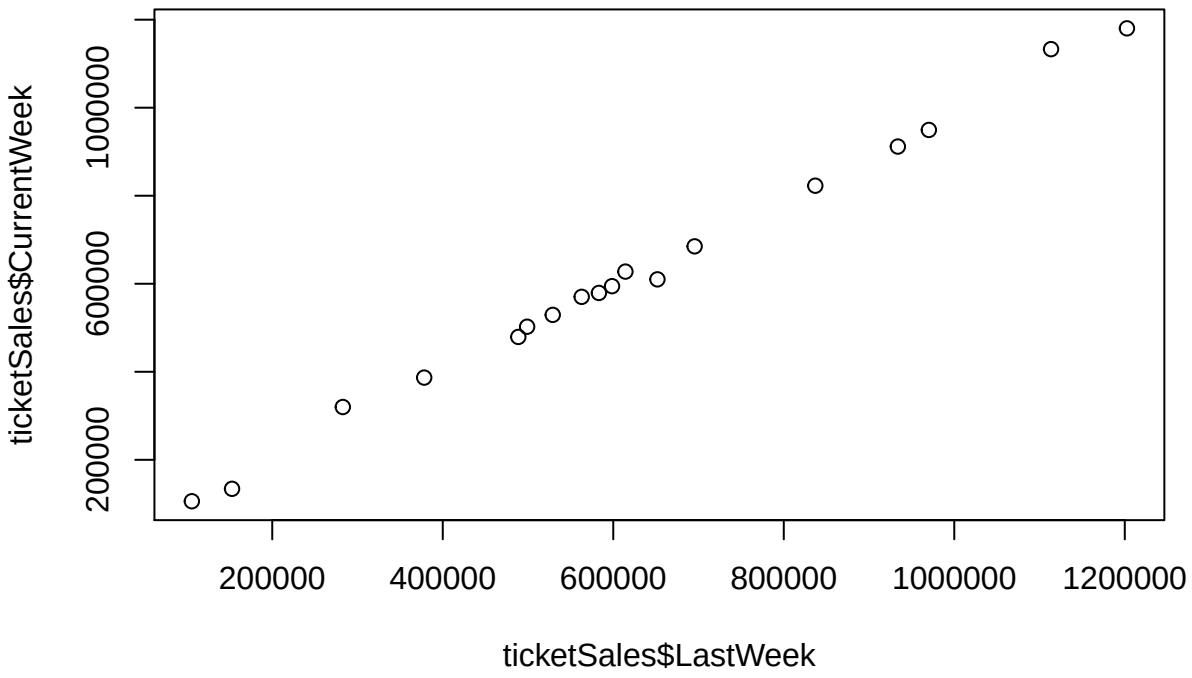
```
ticketSales = read.csv("playbill.csv", header=T)
ticketSales[,]
```

```
##                   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:

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

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:

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

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. This result means that it is very likely (probability of 75.18%) that we will see a value of β_0 as much as here. Because the probability is high we really cannot reject the Null Hypothesis H_0 and our hypothesis testing suggests that the value of β_0 can be 10,000.

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.
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 * s * sqrt(1 + (1/n) + (((x_curr - x_bar)^2)/sxx) ))
yPIUpper = y_hat + (t_mult * s * sqrt(1 + (1/n) + (((x_curr - x_bar)^2)/sxx) ))
#yPILower = 359833
#yPIUpper = 439456.8
```

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

$$359833 \leq Y \leq 439456.8$$

Because the prediction interval does not contain the value $450,000$ in it, we cannot use that value as a prediction. We are basing all of our decision based on the abover prediction interval. If the data would have been different there could have been enough evidence to say that the value lies in the prediction interval, however because for the current dataset it does not exist in the prediction interval, we can say with 95% confidence that this value would not be a good 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. I am arguing in favor for this statement because of the linear model that I was able to fit which contains a slope term of 0.9821

which when formed a 95% confidence interval for the average value, contains 1 in it which shows that the data on average is expected to have a equality coorelation to a very high degree.

Question 2 - Processing Time of Invoices

Introducing the Dataset

This section is dedicated to see and understand the data that was provided from a large company about the number of invoices processed and the total processing time for those invoices. The data being observed is of a whole month. The data table is as follows:

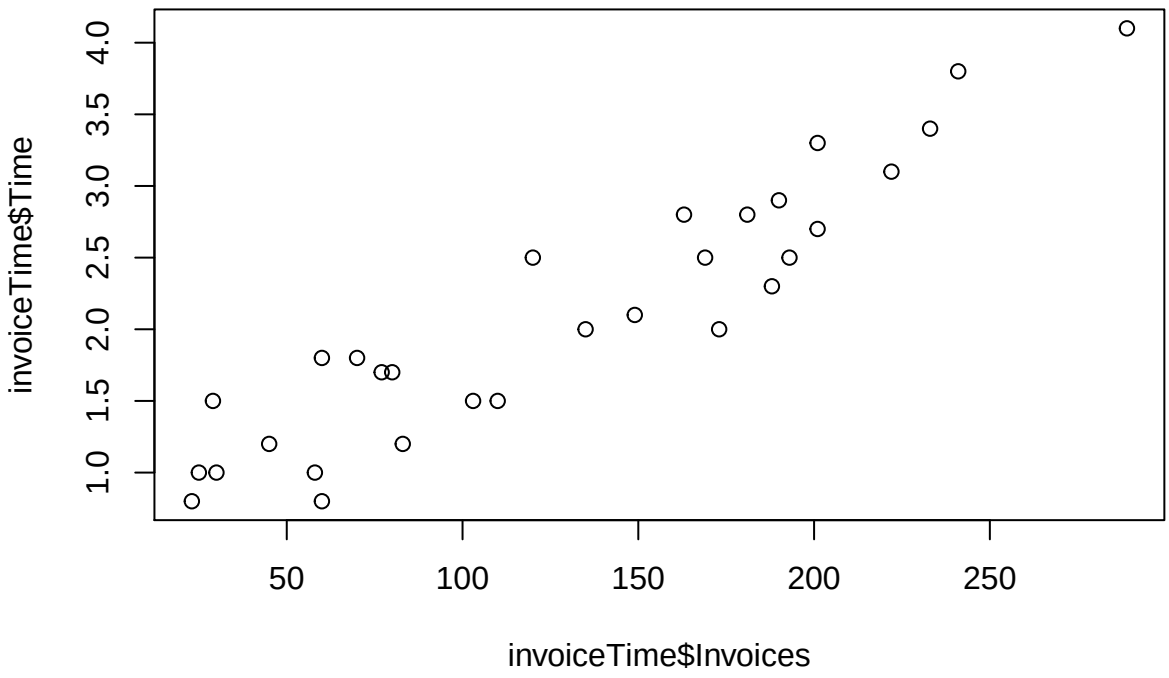
```
invoiceTime = read.table("invoices.txt", header = T)
invoiceTime[,]
```

```
##    Day Invoices Time
## 1    1      149  2.1
## 2    2       60  1.8
## 3    3      188  2.3
## 4    4       23  0.8
## 5    5      201  2.7
## 6    6       58  1.0
## 7    7       77  1.7
## 8    8      222  3.1
## 9    9      181  2.8
## 10  10       30  1.0
## 11  11      110  1.5
## 12  12       83  1.2
## 13  13       60  0.8
## 14  14       25  1.0
## 15  15      173  2.0
## 16  16      169  2.5
## 17  17      190  2.9
## 18  18      233  3.4
## 19  19      289  4.1
## 20  20       45  1.2
## 21  21      193  2.5
## 22  22       70  1.8
## 23  23      241  3.8
## 24  24      103  1.5
## 25  25      163  2.8
## 26  26      120  2.5
## 27  27      201  3.3
## 28  28      135  2.0
## 29  29       80  1.7
## 30  30       29  1.5
```

Visualizing the Dataset

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

```
plot(x = invoiceTime$Invoices, y = invoiceTime$Time)
```



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 = Number of Invoices Y = Total Time required to process those Invoices

Through this linear model we are trying to predict the value of total time required to process invoices (Y) using the number of invoices (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(invoiceTime$Time ~ invoiceTime$Invoices )
mod
```

```
##
## Call:
## lm(formula = invoiceTime$Time ~ invoiceTime$Invoices)
##
## Coefficients:
##          (Intercept)  invoiceTime$Invoices
##              0.64171               0.01129
```

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

$$\hat{\beta}_0 = 0.64171$$

$$\hat{\beta}_1 = 0.01129$$

The summary of the model is as follows:

```
summary(mod)
```

```
##
## Call:
## lm(formula = invoiceTime$Time ~ invoiceTime$Invoices)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59516 -0.27851  0.03485  0.19346  0.53083
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.6417099  0.1222707   5.248 1.41e-05 ***
## invoiceTime$Invoices 0.0112916  0.0008184  13.797 5.17e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3298 on 28 degrees of freedom
## Multiple R-squared:  0.8718, Adjusted R-squared:  0.8672
## F-statistic: 190.4 on 1 and 28 DF,  p-value: 5.175e-14
```

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) = 0.3298\ Degrees of freedom (n-2) = 28\

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

$$Y = 0.6417099 + (0.0112916 * X)$$

Finding a 95% Confidence Inrerval β_0

A $100(1-\alpha)\%$ confidence interval for $\hat{\beta_0}$ is given by the following formula:

$$\hat{\beta_0} - t_{\frac{\alpha}{2},n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}} \leq \beta_0 \leq \hat{\beta_0} + t_{\frac{\alpha}{2},n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}$$

First finding the confidence interval by just using the summary table given on Page 40. We see: $\hat{\beta_0} = 0.6417099$ StdError$(\beta_0) = 0.1222707$ $n = 30$ $s = 0.3298$ finding the value of the t multiplier $t_{\alpha/2,n-2}$ using qt() we get t_mult = 2.048. So the below R code helps in computing the values:

```
bet0_hat = 0.6417099 #using the linear model fit from prevous section
stderror = 0.1222707
n = 30
t_mult = 2.048 #found using the T Table, equal to t0.25 for 28 degrees of freedom
beta0CIlower = bet0_hat - ( t_mult * stderror )
beta0CIupper = bet0_hat + ( t_mult * stderror )
#beta0CIlower = 0.3912
#beta0CIupper = 0.8921
```

The 95% confidence interval found using the above values is as follows:

$$0.3912 \leq \beta_0 \leq 0.8921$$

The following snippet of code is used to find the 95% confidence interval for $\hat{\beta_0}$ by using the R code from scratch and not using the value from the Summary:

```r
bet0_hat = 0.6417099 #using the linear model fit from prevous section
x_col = invoiceTime$Invoices
x_bar = mean(x_col)
#x_bar = 130.0333
sxx = sum((x_col-x_bar)^2)
#sxx = 162367
n = length(x_col)
#n = 30
s = 0.3298 #found through the summary of the linear model
t_mult = 2.048 #found using the T Table, equal to t0.25 for 28 degrees of freedom
beta0CIlower = bet0_hat - ( t_mult * s * sqrt( (1/n) + ((x_bar^2)/(sxx)) ) )
beta0CIupper = bet0_hat + ( t_mult * s * sqrt( (1/n) + ((x_bar^2)/(sxx)) ) )
#beta0CIlower = 0.3912792
#beta0CIupper = 0.8921406
```

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

$$0.3912792 \leq \beta_0 \leq 0.8921406$$

We can easily observe that the two methods give the same value for the confidence interval.

Hypothesis Testing for β_1

We need to run a Hypothesis Test for β_0 given by:

$$Null(H_0)\beta_1 = 0.01 Alternate(H_1)\beta_1 \neq 0.01$$

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

$$\hat{T} = \frac{\hat{\beta_1} - \beta_1}{\frac{s}{\sqrt{S_X X}}}$$

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

$$p - val = 2 \cdot P\big(t \geq | \hat{T} | \big)$$

First we will do the Hypothesis testing using the values from the summary table on Page 40. We see: $\hat{\beta_1} = 0.0112916$ StdError$(\beta_1) = 0.0008184$ $n = 30$ $s = 0.3298$ So the below R code helps in computing the values:

```r
bet1_hat = 0.0112916
stderror = 0.0008184
n = 30
beta1test = ( ( bet1_hat - 0.01 )/( stderror ) )
#beta1test = 1.57807
p_val = 2*pt(-abs(beta1test), df = n-2)
#p_val = 0.1257
```

After performing the Hypothesis test for β_1 we find that the $p - value = 0.1257$. Now we will compare the p-value for this section with p-value we found using just R code in the code sample below.

The following snippet of code is used to perform the Hypothesis Test for β_1:

```r
bet1_hat = 0.0112916 #found using the linear model in the previous sections
x_col = invoiceTime$Invoices
y_col = invoiceTime$Time
x_bar = mean(x_col)
#x_bar = 130.0333
y_bar = mean(y_col)
```

```
#y_bar = 2.11
sxx = sum((x_col-x_bar)^2)
#sxx = 162367
sxy = sum((x_col-x_bar)*(y_col-y_bar))
#sxy = 1833.39
#sxy/sxx = 0.0112916
n = length(x_col)
s = 0.3298 #found through the summary of the linear model
t_mult = 2.048 #found using the T Table, equal to t0.25 for 28 degrees of freedom
beta1test = ( ( bet1_hat - 0.01 )/( s/sqrt( sxx ) ) )
#beta1test = 1.57807
p_val = 2*pt(-abs(beta1test), df = n-2)
#p_val = 0.1257819
```

After performing the Hypothesis test for β_1 we find that the $p - value = 0.1257819$. We see that both ways give us the same p-value 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:

$$Y = 0.6417099 + (0.0112916 * X)$$

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

```
X = 130 #X value for which we need to find the estimated value
Y = 0.6417099 + 0.0112916 * X
#Y = 2.109618
```

Through this above snippet of code we get the point estimate of Y at $X = 130$ as $Y = 2.109618$ which is roughly equal to $127 mins$.

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 = 0.6417099 + 0.0112916 * X
x_curr = 130 #the value at which we need to find the prediction interval
y_hat = 0.6417099 + (0.0112916 * x_curr)
#y_hat = 2.109618
x_col = invoiceTime$Invoices
x_bar = mean(x_col)
#x_bar = 130.0333
sxx = sum((x_col-x_bar)^2)
#sxx = 162367
n = length(x_col)
```

```
s = 0.3298 #found through the summary of the linear model
t_mult = 2.048 #found using the T Table, equal to t0.25 for 28 degrees of freedom
yPILower = y_hat - (t_mult * s * sqrt(1 + (1/n) + (((x_curr - x_bar)^2)/sxx) ))
yPIUpper = y_hat + (t_mult * s * sqrt(1 + (1/n) + (((x_curr - x_bar)^2)/sxx) ))
#yPILower = 1.423023
#yPIUpper = 2.796213
```

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

$$1.423023 \leq Y \leq 2.796213$$

Question 3 - Simple Linear Model

Suppose the linear model we fit is as follows:

$$Y_i = \beta \cdot x_i + e_i$$

and assume that the least squares estimate $\hat{\beta}$ is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Then

Proof 1:

$$E(\hat{\beta}|X) = \beta$$

The proof is as follows:

$$E(\hat{\beta}|X) = E(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}|X) = \frac{\sum_{i=1}^{n} x_i E(y_i|X)}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i \beta x_i}{\sum_{i=1}^{n} x_i^2} = \beta \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} = \beta$$

Proof 2:

$$Var(\hat{\beta}|X) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

The proof is as follows:

$$Var(\hat{\beta}|X) = Var(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}|X) = Var(\frac{\sum_{i=1}^{n} x_i(\beta x_i + e_i)}{\sum_{i=1}^{n} x_i^2}|X) = Var(\frac{\sum_{i=1}^{n} \beta x_i^2}{\sum_{i=1}^{n} x_i^2} + \frac{\sum_{i=1}^{n} x_i e_i}{\sum_{i=1}^{n} x_i^2}|X) =$$

$$Var(\frac{\sum_{i=1}^{n} \beta x_i^2}{\sum_{i=1}^{n} x_i^2}|X) + Var(\frac{\sum_{i=1}^{n} x_i e_i}{\sum_{i=1}^{n} x_i^2}|X) = 0 + \frac{\sum_{i=1}^{n} x_i^2 Var(e_i|X)}{(\sum_{i=1}^{n} x_i^2)^2} = \frac{\sum_{i=1}^{n} x_i^2 \sigma^2}{(\sum_{i=1}^{n} x_i^2)^2} = \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{(\sum_{i=1}^{n} x_i^2)^2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

Proof 3:

$$\hat{\beta}|X \sim N(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2})$$

The proof is as follows:

Combining the above two proofs we can say that:

$$\hat{\beta}|X \sim N(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2})$$

Question 4 - Comparision between RSS and SSreg

My Answer: Option d

Explanation: First consider the formulas for RSS and SSreg

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_i)^2$$

RSS calculates the variability between the actual (observed) values of Y and the predicted values of Y, while SSreg calculates the variability between the predicted value of Y and mean of Y. In Plot 1 we see that the fitted model is a very close fit to the data, which reduces the value for the difference between actual Y and predicted Y, while in Plot 2 the fitted model is very bad due to which we see large differences between actual Y and fitted Y. This shows that the RSS value for Plot 1 should be less than Plot 2. Now considering SSreg, we see that in Plot 1 the slope of the fitted model is much more steep, hence leading to more difference between predicted Y and mean Y, while the slope of the fitted model in Plot 2 is much more small, due to which there would be less difference between predited Y and mean Y which will reduce the SSreg for Plot 2 as compared to Plot 1 which is why I chose the option D.