Assignment2

Biplav Timalsina

February 13, 2018

STAT 757 Applied Regression Analysis

Exercise 2.8.1 [20 Points]

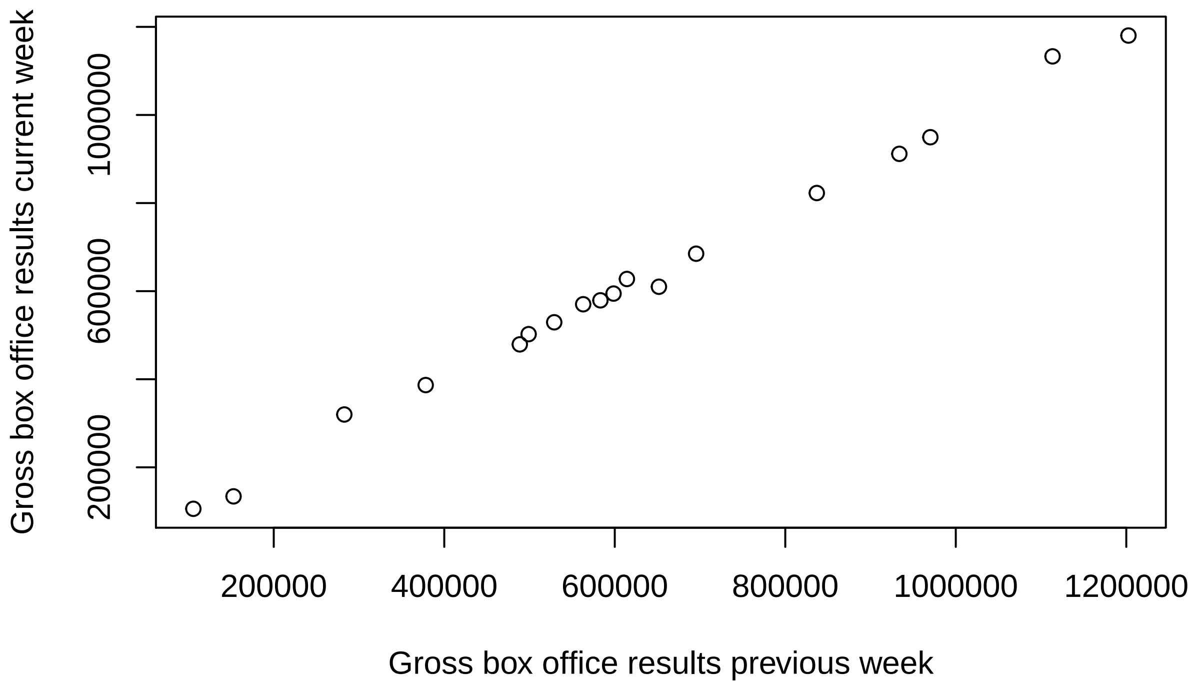
The web site www.playbill.com provides weekly reports on the box office ticket sales for plays on Broadway in New York. We shall consider the data for the week October 11-17, 2004 (referred to below as the current week). The data are in the form of the gross box office results for the current week and the gross box office results for the previous week (i.e., October 3-10, 2004). The data, plotted in Figure 2.6 , are available on the book web site in the file playbill.csv. Fit the following model to the data: $Y = b0 + b1x + e$ where Y is the gross box office results for the current week (in $) and x is the gross box office results for the previous week (in $). Complete the following tasks:

Ans:

First off let us begin by plotting our data to see any obvious references that we can make.

```
playbill=read.csv("F:/unr/4th sem/applied regression analysis/Assignments/HW2/playbill.csv",header=TRUE

plot(playbill$LastWeek,playbill$CurrentWeek,xlab="Gross box office results previous week",
     ylab="Gross box office results current week")
```



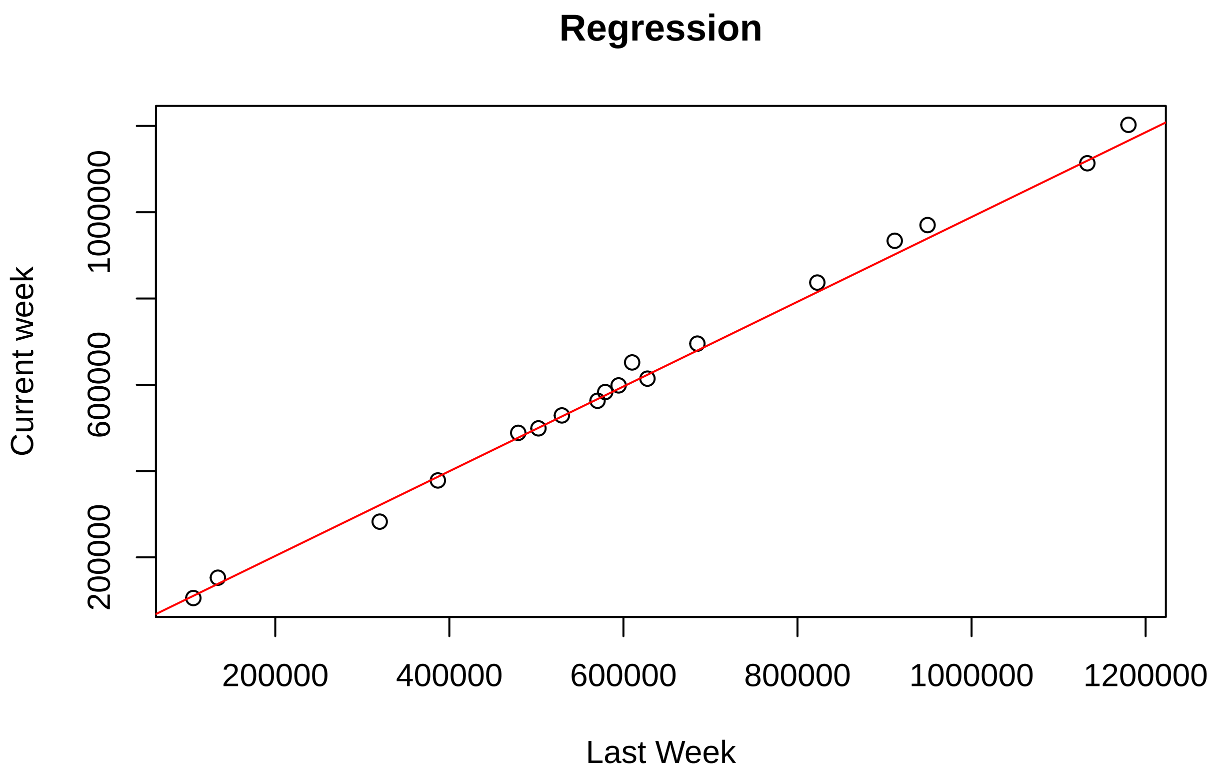
We notice that there is **some relationship (indicated by distribution of the data points along almost a straight line)** between the two variables under discussion, Revenue of current Week and revenue of previous Week, for plays in Broadway, NY. Before we answer any of the questions mentioned, let us fit our model in R.

```
myLinearModel<-lm(formula=playbill$CurrentWeek~playbill$LastWeek,data=playbill)
summary(myLinearModel)
```

```
##
## Call:
## lm(formula = playbill$CurrentWeek ~ playbill$LastWeek, data = playbill)
##
## 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
## playbill$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
```

Let us plot the regression line.

```
plot(playbill$CurrentWeek, playbill$LastWeek, xlab="Last Week", ylab="Current week", main="Regression")
abline(6.805e+03,9.821e-01,col="red")
```



It looks like the data is fitted properly by this model as indicated by the line that passes near all the data points very near.

Let us interpret the results obtained.

Residuals

Residuals are essentially the difference between observed values of variables and their predicted value, in our case, the difference between **actual observed response values (revenue current week)** and the **predicted revenues for current week.** Symmetrical distribution across the five points of Residuals signify a good fit. Here we can see that the data is not strongly symmetrical. This means our model predicts certain points that are far from the actual observed points.

Coefficients

This is where we get the slope and intercept of our linear model that R has generated.

Coefficient- Estimate It has two rows; the first one along (Intercept) is **intercept** (here 6.085e+03), and the second one along **playbill$LastWeek** (here,9.821e-01). The **intercept** in our case is essentially the average revenue of plays when we consider average revenue of all plays this week. [In this case, when x=0, y=6.805e+3 doesn't make much sense in that the plays we can only predict revenue of those plays which was being played last week too; any play that start this week cannot be predicted very well and predicted to be generate in average 6.805e+3]. The positive **slope** which is almost equal to 1 implies that there will be slight decrease in revenue for plays compared to last week. To be specific, we could say for a $1 difference in revenue generated by plays last week, it will only generate $0.98 this week.

Coefficient-Standard Error The coefficient-standard error (SE) estimates the average amount that the coefficient estimates vary from the actual average value of our response variables. Ideally, it should be less related to its coefficients. In our case, the SE for intercept is 9.929e+03 and for slope is 1.443e-02. In other words, we can say that the revenue for current week per $1 which was $0.98 could vary by $1.443e-02.

Coefficient- T Value The T value is a measure of how many standard deviations our coefficient estimates are from 0. We want it to be far from zero as this would indicate we could reject null hypothesis-that is, we could say there is a relationship between revenue this week and previous week. In our example, t value for Intercept is relatively close to 0 but slope is relatively far. So, we could say that we cannot reject null hypothesis that intercept is 0 (but it could be or couldn't be equal to 0). For slope, we can say that since it is 68 S.D.s away from 0, we can reject the null hypothesis that slope is 0, and that there is relationship between revenue generated last week and this week.

Coefficient- P value The Pr(>t) acronym found in the model relates to the probability of observing any value equal or larger than t. A small p-value indicates that it is unlikely we will observe a relationship between the predictor(revenue last week) and response(revenue this week) variables due to change. Typically, a p-value of 5% or less is a good cut-off point. Here, we cannot reject the null hypothesis of intercept being 0, but can tell that there is a relationship between the two variables as p value for slope is non-zero.

Residual Standard Error Residual standard error is the measure of the quality of linear regression fit. Theoretically, every linear model is assumed to contain an error term E. Due to the presence of this error term, we cannot perfectly predict our response variables from predictor variables. The Residual Standard Erros is the average amount that the response(revenue this week) could deviate from the true regression line. In our example, the actual revenue generated this week deviate from the true regression line by $18010. In other words, given the mean revenue generated is $6.085e+03, and the residual standard error is $18010, we can say the percentage error is 296.3%. We have to note that RSE was calculated using 16 degrees of freedom.

Multiple R-squared, Adjusted R-squared The R-squared statistic provides a measure of how well the model fits the actual data. It takes the form of proportion of variance. It always lies between 0 and 1 (i.e a number near 0 represents a regression that doesn't explain the variance in the response variable well and a number close to 1 does explain the observed variance in the response variable). For us, it is 0.996. or roughly 99.6% of the variance found in the response variable can be explained by the predictor variable. Lets think for a bit. If we were able to choose any metric to predict the revenue this week, would revenue in the last week be an important one that could explain it? In our case, it would definitely a yes. However, its hard to define what level of R-squared is appropriate to claim a model fits well. In multiple regression settings, the R-squared will always increase as more variables are included in the model. that is why, adjusted R-squared is preferred measure as it adjusts for the variables considered.

F-statistic F-statistic is a good indicator of whether there is a relationship between our predictor and the response variables. The further it is from 1 the better it is. For us, it is 4634, which is relatively larger than 1 given the size of our data.

(a) Find a 95% confidence interval for the slope of the regression model, b1 . Is 1 a plausible value for b1 ? Give a reason to support your answer.

Ans:

The formula for confidence interval for

$$\beta_1 = (\hat{\beta_1} - t(\alpha/2, n-2) * s.e * \beta_1, \hat{\beta_1} + t(\alpha/2, n-2) * s.e\beta_1)$$

We can calculate it manually by using the values from the coefficients, namely

$$\hat{\beta_1} = 9.821e-01, t(0.025, 16) = 2.12, se(\hat{\beta}) = 1.443e-02.$$

Thus, 95% confidence interval for b1 is given by (0.982-2.12 * 0.01443, 0.982+2.12*0.01443)=(0.9514,1.0125)

From R,

```
confint(myLinearModel,'playbill$LastWeek',level=0.95)
```

```
##                       2.5 %   97.5 %
## playbill$LastWeek 0.9514971 1.012666
```

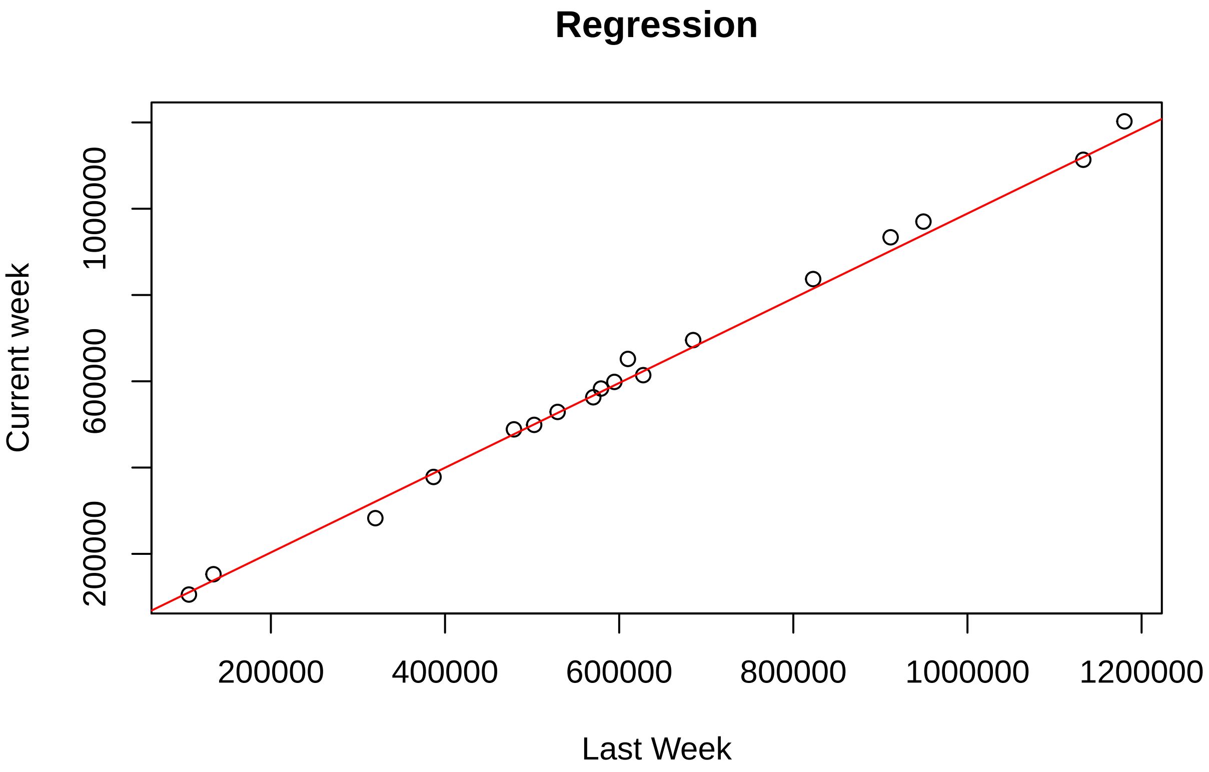
We can see the two results are almost same. Since 1 lies inside the confidence interval, it is a plausible value for b1. It also means that shows might have same sales in last week and this week.

We can also plot the confidence interval for line as below.

```
plot(playbill$CurrentWeek, playbill$LastWeek, xlab="Last Week", ylab="Current week", main="Regression")
abline(6.805e+03,9.821e-01,col="red")
conf_interval<-predict(myLinearModel, interval="confidence", level=0.95)
conf_interval
```

```
##          fit       lwr       upr
## 1   689780.7  680508.2  699053.2
## 2   496833.1  487078.0  506588.2
## 3   594655.3  585628.6  603682.0
## 4   526320.1  516881.7  535758.5
## 5   559681.4  550503.1  568859.7
## 6   284515.9  270778.5  298253.3
## 7   579532.2  570455.7  588608.7
## 8   156899.3  139957.6  173841.1
## 9   110608.9   92429.6  128788.3
## 10  828766.8  817626.5  839907.2
## 11  959610.5  945673.4  973547.6
## 12  646933.5  637890.2  655976.7
## 13  378265.4  366576.7  389954.2
## 14 1100362.4 1082847.7 1117877.1
## 15  610044.5  601043.5  619045.6
## 16  923894.2  910786.8  937001.5
## 17 1187793.2 1167893.3 1207693.2
## 18  486673.5  476791.8  496555.1
```

```
lines(conf_interval[,2],col="blue",lty=2)
lines(conf_interval[,3],col="blue",lty=2)
```



(b) Test the null hypothesis H0 : b0 = 10000 against a two-sided alternative. Interpret your result.

Ans:

```
modelSummary<-summary(myLinearModel)
modelCoeffs<-modelSummary$coefficients
beta.estimate<-modelCoeffs["(Intercept)","Estimate"]
std.error<-modelCoeffs["(Intercept)","Std. Error"]
t_value<-(beta.estimate-10000)/std.error
p_value<-2*pt(-abs(t_value),df=nrow(playbill)-ncol(playbill))
t_value
```

```
## [1] -0.3217858
```

```
p_value
```

```
## [1] 0.7520556
```

Here, since p_value >0.05, we cannot reject null hypothesis that b0=10000.

(c) Use the fitted regression model to estimate the gross box office results for the current week (in $) for a production with $400,000 in gross box office the previous week. Find a 95% prediction interval for the gross box office results for the current week (in $) for a production with $400,000 in gross box office the previous week. Is $450,000 a feasible value for the gross box office results in the current week, for a production with $400,000 in gross box office the previous week? Give a reason to support your answer.

Ans:

In R,

```r
a<-playbill$CurrentWeek
b<-playbill$LastWeek
myLinearModel<-lm(formula=a~b,data=playbill)

beta0.estimate<-modelCoeffs["(Intercept)","Estimate"]
beta1.estimate<-modelCoeffs["playbill$LastWeek","Estimate"]
predictedValue<-beta0.estimate+400000*beta1.estimate
predictedValue
```

```
## [1] 399637.5
```

```r
predict(myLinearModel,data.frame(b=400000),interval="prediction",level=0.95)
```

```
##        fit      lwr      upr
## 1 399637.5 359832.8 439442.2
```

Notice here the predicted value using formula and the fit value generated using predict formula is same (399637.5). Since $450,000 is not included in the 95% prediction interval, it is not feasible for a production with $400000 in gross box office to generate $450000 current week.

(d) Some promoters of Broadway plays use the prediction rule that next week's gross box office results will be equal to this week's gross box office results. Comment on the appropriateness of this rule.

We have seen that the confidence interval for slope of the linear regression model developed was (0.9514,1.0125), which implies that it includes a slope of one, which further implies that plays could generate equal revenue next week as comparedcurrent week. Lets also observe the confidence interval for intercept of linear regression model. intercept= (6805- 2.12 * 9929, 6805 + 2.12*9929) = (-14244.48,27854.48) Hence, 0 lies in the confidence interval and is plausible value for intercept. So for slope=1, and intercept =0, the rule can be appropriate.

2. A story by James R. Hagerty entitled With Buyers Sidelined, Home Prices Slide published in the Thursday October 25, 2007 edition of the Wall Street Journal contained data on so-called fundamental housing indicators in major real estate markets across the US. The author argues that. prices are generally falling and overdue loan payments are piling up . Thus, we shall consider data presented in the article on

Y = Percentage change in average price from July 2006 to July 2007 (based on the S&P/Case-Shiller national housing index); and

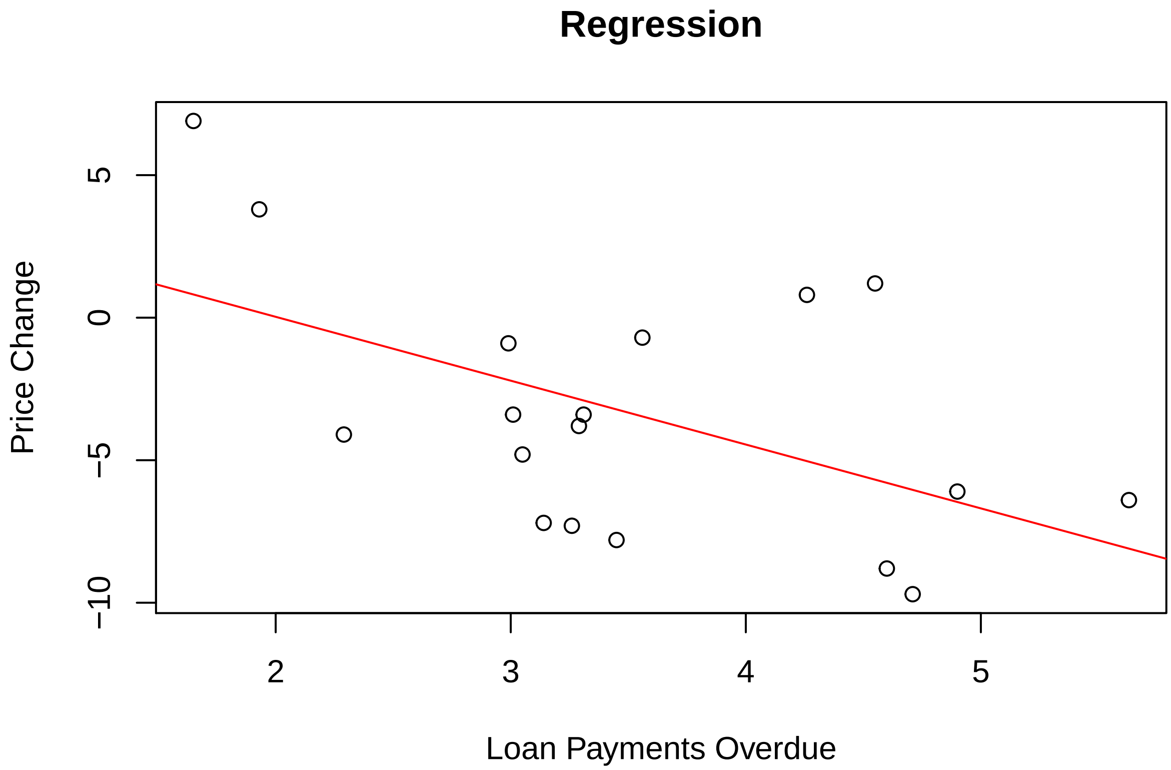
X = Percentage of mortgage loans 30 days or more overdue in latest quarter (based on data from Equifax and Moody's).

The data are available on the book web site in the file indicators.txt. Fit the following model to the data: Y = b0 + b1x + e . Complete the following tasks:

Ans:

Fitting the linear regression model, Y=b0+b1x+e

```
myData<-read.table("F:/unr/4th sem/applied regression analysis/Assignments/HW2/indicators.txt",header=T
plot(myData$LoanPaymentsOverdue,myData$PriceChange, xlab="Loan Payments Overdue",
     ylab="Price Change",main="Regression")
abline(4.51,-2.24,col="red")
```



From the plot above, we can observe that there might not be huge correlation between the data as they are very scattered. But let us now using linear regression to generate a line of fit.

```
myLinearModel<-lm(PriceChange~LoanPaymentsOverdue,data=myData)
summary(myLinearModel)
```

```
## 
## Call:
## lm(formula = PriceChange ~ LoanPaymentsOverdue, data = myData)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.6541 -3.3419 -0.6944  2.5288  6.9163 
## 
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)  
## (Intercept)           4.5145     3.3240   1.358   0.1933  
## LoanPaymentsOverdue  -2.2485     0.9033  -2.489   0.0242 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.954 on 16 degrees of freedom
## Multiple R-squared:  0.2792, Adjusted R-squared:  0.2341 
## F-statistic: 6.196 on 1 and 16 DF,  p-value: 0.02419
```

We can see that the slope is -2.2485 and intercept is 4.5145. We can gain various insights from these values as in question 1.

(a) Find a 95% confidence interval for the slope of the regression model, b1 . On the basis of this confidence interval decide whether there is evidence of a significant negative linear association.

Lets use the in-built function of R to predict the 95% confidence interval for the slope.

```
confint(myLinearModel,'LoanPaymentsOverdue',level=0.95)
```

```
##                         2.5 %     97.5 %
## LoanPaymentsOverdue -4.163454 -0.3335853
```

From the above result, we can see that the confidence interval totally lies in the negative domain, i.e there is negative correlation between the percentage change in price change and percentage of mortage loans. Or in other words, we could say, payment increase/decrease in loan payment over due might actually be cause/related to price change decrease/increase respectively.

b) Use the fitted regression model to estimate E (Y | X =4). Find a 95% confidence interval for E (Y | X =4). Is 0% a feasible value for E (Y | X =4)? Give a reason to support your answer.

We could manually calculate E(Y | X=4), as

```
modelSummary<-summary(myLinearModel)
modelCoeffs<-modelSummary$coefficients
intercept<-modelCoeffs["(Intercept)","Estimate"]
slope<-modelCoeffs["LoanPaymentsOverdue","Estimate"]
```

```
predictedValue<-intercept+4*slope
predictedValue
```

```
## [1] -4.479585
```

To find the 95% confidence interval for E(Y | X=4),

```
predict(myLinearModel,data.frame(LoanPaymentsOverdue=4),interval="confidence",level=0.95)
```

```
##         fit       lwr       upr
## 1 -4.479585 -6.648849 -2.310322
```

As 0% doesn't fall in 95% confidence interval, we can say that 0% is not feasible for E(Y | X=4).

3. The manager of the purchasing department of a large company would like to develop a regression model to predict the average amount of time it takes to process a given number of invoices. Over a 30-day period, data are collected on the number of invoices processed and the total time taken (in hours). The data are available on the book web site in the file invoices.txt. The following model was fit to the data: 0 1 Y = b + b x + e where Y is the processing time and x is the number of invoices. A plot of the data and the fitted model can be found in Figure 2.7 . Utilizing the output from the fit of this model provided below, complete the following tasks.

(a) Find a 95% confidence interval for the start-up time, i.e., b0 .

####Ans: Intercept=0.641 Std. Error = 0.122 t(0.025,28)=2.048 Confidence interval = (0.641 - 2.048*0.122, 0.641+2.048*0.122)=(0.391,0.891)

(b) Suppose that a best practice benchmark for the average processing time for an additional invoice is 0.01 hours (or 0.6 minutes). Test the null hypothesis H0 : b1 = 0.01 against a two-sided alternative. Interpret your result.

####Ans: H0: b1: 0.01

```
slope<-0.0112916
stdError<-0.0008184
numRow<-30
numCol<-3

t_value<-(slope-0.01)/stdError
p_value<-2*pt(-abs(t_value),df=numRow-numCol)
t_value
```

```
## [1] 1.578201
```

```
p_value
```

```
## [1] 0.1261651
```

Since, p_value is >0.05, we cannot reject null hypothesis, ie. it could be that b1 be equal to 0.01.

(c). Find a point estimate and a 95% prediction interval for the time taken to process 130 invoices.

####Ans: The point estimate for the time take to process 130 invoices can be calculated by using the model that has been provided in question,

time taken= 0.641 + 0.0113* number of invoices

Using this formula, for 130 invoices, the time taknen=0.641+0.113*130 = 2.11

```
myData<-read.table("F:/unr/4th sem/applied regression analysis/Assignments/HW2/invoices.txt",header=TRUI
myLinearModel<-lm(Time~Invoices,data=myData)
summary(myLinearModel)
```

```
##
## Call:
## lm(formula = Time ~ Invoices, data = myData)
##
## 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 ***
## 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
```

We can see the model summary above.

For 95% prediction interval,

```
predict(myLinearModel,data.frame(Invoices=130),interval="prediction",level=0.95)
```

```
##        fit      lwr    upr
## 1 2.109624 1.422947 2.7963
```

According to result, it takes between 1.423 hours and 2.793 hours to process 130 invoices with 95% confidence interval.

4. Straight-line regression through the origin: In this question we shall make the following assumptions:

(1) Y is related to x by the simple linear regression model Yi = bxi + ei (i = 1,2,. . . ,n) , i.e., E(Y|X=x_i)= Bx_i

(2) The errors e1, e2,. . . , en are independent of each other

(3) The errors e1, e2,. . . , en have a common variance s2

(4) The errors are normally distributed with a mean of 0 and variance s2 (especially when the sample size is small), i.e., e | X~ N(0,s2)

In addition, since the regression model is conditional on X we can assume that the values of the predictor variable, x_1 , x_2 , ., x_n are known fixed constants.

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

To show the above, we start with the least-square solution of RSS, which is given by,

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}i)^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{\beta} x_i)^2$$

To obtain the value of β, we can do partial derivative of this equation with respect to β, like follows and equating it to 0, we get,

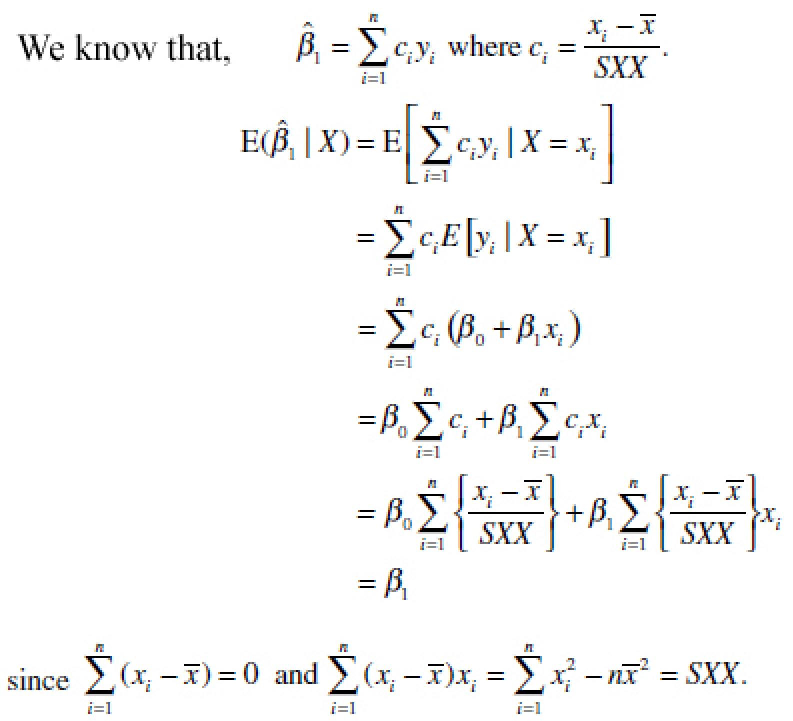
$$\frac{\partial}{\partial \beta} \text{RSS} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta} x_i)$$

$$or, \sum_{i=1}^{n} x_i y_i - \beta \sum_{i=1}^{n} x_i^2 = 0$$

Simplifying, we get,

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

b) For i)

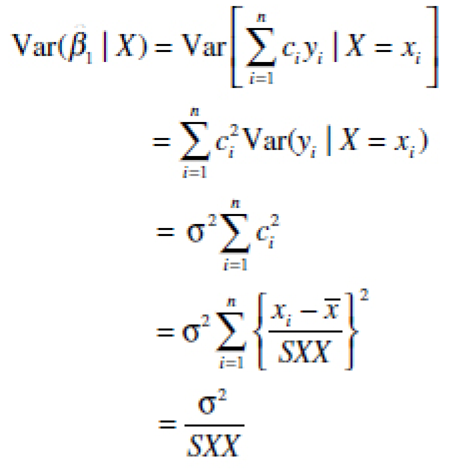


We know that, $\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$ where $c_i = \dfrac{x_i - \overline{x}}{SXX}$.

$$
\begin{aligned}
\mathrm{E}(\hat{\beta}_1 \mid X) &= \mathrm{E}\left[\sum_{i=1}^{n} c_i y_i \mid X = x_i\right] \\
&= \sum_{i=1}^{n} c_i E\left[y_i \mid X = x_i\right] \\
&= \sum_{i=1}^{n} c_i \left(\beta_0 + \beta_1 x_i\right) \\
&= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i \\
&= \beta_0 \sum_{i=1}^{n} \left\{\frac{x_i - \overline{x}}{SXX}\right\} + \beta_1 \sum_{i=1}^{n} \left\{\frac{x_i - \overline{x}}{SXX}\right\} x_i \\
&= \beta_1
\end{aligned}
$$

since $\sum_{i=1}^{n} (x_i - \overline{x}) = 0$ and $\sum_{i=1}^{n} (x_i - \overline{x}) x_i = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = SXX.$

For ii)



$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_1 \mid X) &= \mathrm{Var}\left[\sum_{i=1}^{n} c_i y_i \mid X = x_i\right] \\
&= \sum_{i=1}^{n} c_i^2 \mathrm{Var}(y_i \mid X = x_i) \\
&= \sigma^2 \sum_{i=1}^{n} c_i^2 \\
&= \sigma^2 \sum_{i=1}^{n} \left\{\frac{x_i - \overline{x}}{SXX}\right\}^2 \\
&= \frac{\sigma^2}{SXX}
\end{aligned}
$$

For iii) Once we have obtained the two above things, we can directly represent the information in the notation,

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

Note = sign represents follows that distribution here.I couldn't write the 'similar' sign.

5. Two alternative straight line regression models have been proposed for Y . In the first model, Y is a linear function of x1 , while in the second model Y is a linear function of x2 . The plot in the first column of Figure 2.8 is that of Y against x1 , while the plot in the second column below is that of Y against x2 . These plots also show the least squares regression lines. In the following statements RSS stands for residual sum of squares, while SSreg stands for regression sum of squares. Which one of the following statements is true?

(a) RSS for model 1 is greater than RSS for model 2, while SSreg for model 1 is greater than SSreg for model 2.

(b) RSS for model 1 is less than RSS for model 2, while SSreg for model 1 is less than SSreg for model 2.

(c) RSS for model 1 is greater than RSS for model 2, while SSreg for model 1 is less than SSreg for model 2.

(d) RSS for model 1 is less than RSS for model 2, while SSreg for model 1 is greater than SSreg for model 2.

Give a detailed reason to support your choice.

Ans:

To begin, we know SST=SS_reg + RSS

SS_reg=Variability explained by model, RSS=Unexplained variability RSS = SUM(i=1 to n) (y_i - y_cap)^2

Here, Model 1 seems better fit than Model 2, because the line of best fit is near to all the actual points. Thus, RSS is small for Model 1 compared to Model2.

Secondly, for regression sum of squares, which is sum of square of difference of estimated y_i and y_mean, we can see that at any point x_i, y_i's seem to be not symmetrical along the line of best fit for Model1, meaning the SS_reg is larger than Model2, where at any x_i, y_i's seem to be symmetrical along the line of best fit. In other words, the line of best is close to mean line for Model 2 than compared to Model1. Hence, SS_reg os Model1 is larger than Model2.

Thus, d is the answer.

*6. In this problem we will show that SST=SSreg + RSS. To do this we will show that

Ans:

a)

$$
\begin{aligned}
y_i - \hat{y}_i &= (y_i - \bar{y}) - (\hat{y}_i - \bar{y}) \\
&= (y_i - \bar{y}) - (\hat{\beta} x_i - \hat{\beta}\bar{x}) \\
&= (y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})
\end{aligned}
\qquad \square
$$

b)

$$
\begin{aligned}
\hat{y}_i - \bar{y} &= \hat{\beta}(x_i - \bar{x}) \\
&= \hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x}
\end{aligned}
$$

$$= \hat{y}_i - \beta_0 - \bar{y} + \beta_0$$
$$= \hat{y}_i - \bar{y}$$

c)

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}i - \bar{y})$$

$$= \sum_{i=1}^{n} \{(y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x})\}(\hat{\beta}(x_i - \bar{x}))$$
$$= \hat{\beta}_1 \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$
$$= \hat{\beta}_1(\mathrm{SXY} - \hat{\beta}_1{}^2\mathrm{SXX})$$
$$= \hat{\beta}_1(SXY - SXY)$$
$$= 0$$

7.A statistics professor has been involved in a collaborative research project with two entomologists. The statistics part of the project involves fitting regression models to large data sets. Together they have written and submitted a manuscript to an entomology journal. The manuscript contains a number of scatter plots with each showing an estimated regression line (based on a valid model) associated individual 95% confidence intervals for the regression function at each x value, as well as the observed data. A referee has asked the following question:

I don't understand how 95% of the observations fall outside the 95% CI as depicted in the figures.

Briefly explain how it is entirely possible that 95% of the observations fall outside the 95% CI as depicted in the figures.

Ans:

Regression methods are used to predict the mean of the y_i's at certain x_i's. If the data is highly scattered far away on both sides of the regression line, then in such case the observations may fall outside the 95% CI. 95% CI is 95% confident that the mean will fall within that range, not the observations. It is prediction interval that will contain 95% of the observations. So, it is possible that 95% observations lie outside of 95% CI of regression line.