Homework1

Venakata Kanishk Chaganti

1/28/2022

Exercise 1. The following is a regression summary from R for a linear regression model between an explanatory variable x and a response variable y. The data contain n = 50 points. Assume that all the conditions for SLR are satisfied.

(a) Write the equation for the least squares regression line.

Answer: Y= -1.1016 + 2.2606 X

(b) R performs a t-test to test whether the slope is significantly different than 0. State the null and alternative hypothesis for this test. Based on the p-value what is the conclusion of the test (i.e., reject or do not reject the null hypothesis)?

Answer : As represented in the Coefficients table we can conclude that slope(2.2606) is greater than zero and probability (p) value(2e-16 ***) is less than 0.5. Hence we use an alternate hypothesis.

(c) Calculate the missing p-value for the intercept.

Answer : p-value = 0.009573, which is smaller than 0.5.

(d) Calculate the missing t-statistic for the slope.

Answer : t-statistic can be calculated by using,

t= beta0/SE(beta0)

t= 2.2606/0.0981

t= 23.0438

(e) Calculate a 95% confidence interval for the slope of the regression line. Does this interval agree with the results of the hypothesis test?

Answer: upper limit= 2.4568

lower limit= 2.0644

Exercise 3

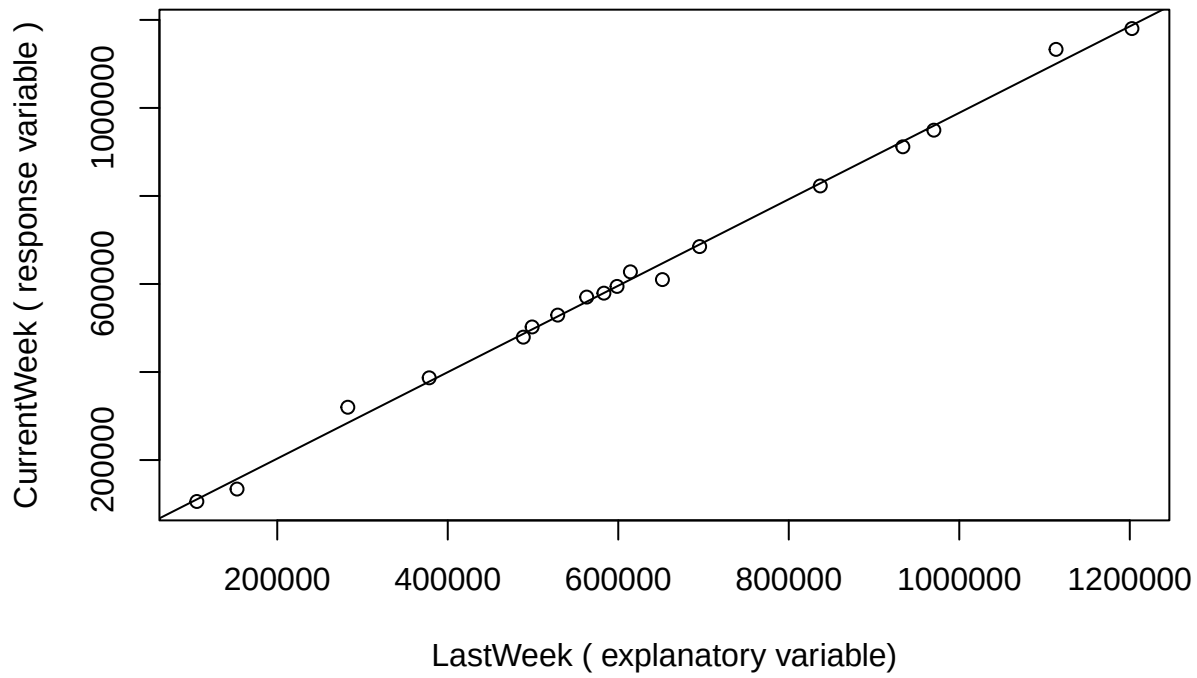
(A) Use read.csv() to load the playbill.csv data file into R. Make a scatter plot of the response versus the explanatory variable, and superimpose the least squares regression line.

```
data_set <- read.csv("playbill.csv")
x<- lm(data_set$CurrentWeek ~ data_set$LastWeek, data=data_set)
```

(b) Calculate a 95% confidence interval for the intercept and slope of the regression model, beta0 and beta1 [hint: use the confint() function]. Is 1 a plausible value for beta1?

Answer :

```
plot(data_set$CurrentWeek ~ data_set$LastWeek, xlab = "LastWeek ( explanatory variable)", ylab ="Curren
abline(x)
```




```
confint(x)
```

```
##                           2.5 %       97.5 %
## (Intercept)       -1.424433e+04 27854.099443
## data_set$LastWeek  9.514971e-01     1.012666
```

```
summary(x)
```

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

Yes, 1 a plausible value for beta1

(c) Use the fitted regression model to estimate the gross box office results for the current week (in dollars) 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 dollars) 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?

Answer :

```
newdata=data.frame(Duration =c(400000))
predict(x, newdata =newdata,interval="prediction", level = 0.95)
```

```
## Warning: 'newdata' had 1 row but variables found have 18 rows
```

```
##          fit        lwr       upr
## 1   689780.7  650496.37  729065.0
## 2   496833.1  457432.08  536234.1
## 3   594655.3  555428.26  633882.3
## 4   526320.1  486996.28  565643.9
## 5   559681.4  520419.22  598943.6
## 6   284515.9  243945.06  325086.8
## 7   579532.2  540293.69  618770.7
## 8   156899.3  115134.52  198664.2
## 9   110608.9   68326.94  152890.9
## 10  828766.8  789000.18  868533.5
## 11  959610.5  918971.60 1000249.4
## 12  646933.5  607702.61  686164.3
## 13  378265.4  338341.68  418189.2
## 14 1100362.4 1058361.92 1142362.9
```

```
## 15  610044.5  570823.37  649265.6
## 16  923894.2  883532.30  964256.1
## 17 1187793.2 1144743.40 1230843.0
## 18  486673.5  447240.93  526106.0
```

(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.

Answer :

As we predicted earlier, the predicted value for 400,000$ is

Exercise 4

For this question use the oldfaith data set form the alr4 package. To access this data set first install the package using install.packages("alr4") (this only needs to be done once). Then load the package into R with the command library(alr4). Documentation for the data set can be read in the help menu by entering the command help(oldfaith).

The oldfaith data set gives information about eruptions of Old Faithful Geyser during October 1980. Variables are Duration in seconds of the current eruption, and the Interval, the time in minutes to the next eruption. The data wer collected by volunteers and were provided by the late Roderick Hutchinson. Apart from missing data for the period from midnight to 6 a.m., this is a complete record of eruptions for that month.

Old Faithful Geyser is an important tourist attraction, with up to several thousand people watching it erupt on pleasant summer days. The park service uses data like these to obtain a prediction equation for the time to the next eruption.

```
library("alr4")
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
data("oldfaith")
head(oldfaith)
```

```
##   Duration Interval
## 1      216       79
## 2      108       54
## 3      200       74
## 4      137       62
## 5      272       85
## 6      173       55
```

(a) Use the lm() function to perform a simple linear regression with Interval as the response and Duration as the predictor. Use the summary() function to print the results.

Answer:

```
Duration <- oldfaith$Duration
Interval <- oldfaith$Interval
LMmodel <- lm(Interval ~ Duration, data=oldfaith)
summary(LMmodel)
```

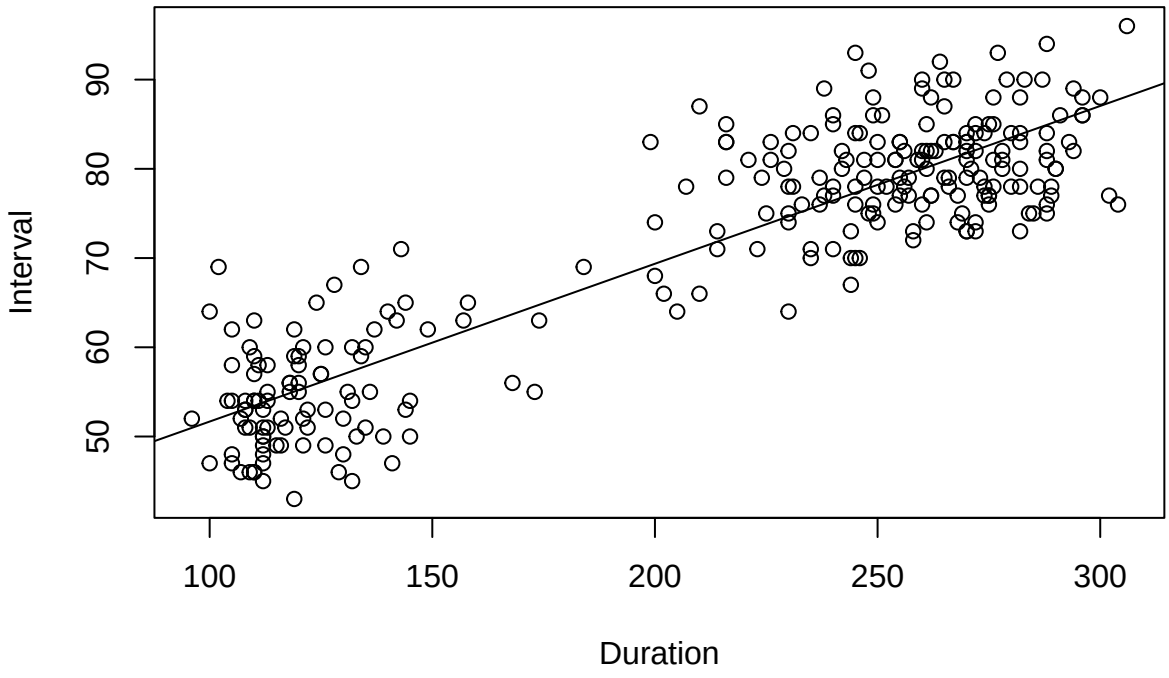
```
##
## Call:
## lm(formula = Interval ~ Duration, data = oldfaith)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3337  -4.5250   0.0612   3.7683  16.9722
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.987808   1.181217   28.77   <2e-16 ***
## Duration     0.176863   0.005352   33.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.004 on 268 degrees of freedom
## Multiple R-squared:  0.8029, Adjusted R-squared:  0.8022
## F-statistic:  1092 on 1 and 268 DF,  p-value: < 2.2e-16
```

(b) Make a scatter plot of Interval versus Duration. Superimpose the least squares regression line on the scatter plot.

Answer :

```
plot(Duration, Interval)
abline(LMmodel)
```



(c) An individual has just arrived at the end of an eruption that lasted 250 seconds. What is the predicted

amount of time the individual will have to wait until the next eruption? Calculate a 95% prediction interval for the time the individual will have to wait for the next eruption.

Answer :

```
newdata=data.frame(Duration =c(250))
predict(LMmodel, newdata =newdata,interval="prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 78.20354 66.35401 90.05307
```

(d) Interpret the coefficient of determination (R2).

Answer : Coefficient of determination measures the percentage of variability with in the y-value. As represented in output of summary function we can observe that coefficient of determination is 80%. Hence we can say that there is 80$ variability in y-value.