**BU.510.650 Homework #1**

**Data Analytics Page 1 of 2**

**Spring 2021 The Johns Hopkins University**

**Arnab Bisi Carey Business School**

**Homework #2**

Due: 02/13/21, 11:59pm

1. (a) Run the R codes provided in the file “R-code-Session02.R” on the data sets Advertising.csv and FuelEfficiency.csv and submit your output. (1 point)

<R-code-Session02.docx>

(b) For the Advertising.csv data set, when you fit a multiple linear regression model to predict Sales based on TV, Newspaper, and Radio advertising budget, which variables have a significant relationship with Sales? Write down and submit the fitted multiple linear regression model. (2 points)

According to results generated by “R-code-Session02.R”, in the Advertising.csv data set, TV and Radio variable have significant relationship with Sales. Output shown below:

my.lm.5 = lm(Sales ~ TV + Radio)  
summary(my.lm.5)

##   
## Call:  
## lm(formula = Sales ~ TV + Radio)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.7977 -0.8752 0.2422 1.1708 2.8328   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.92110 0.29449 9.919 <2e-16 \*\*\*  
## TV 0.04575 0.00139 32.909 <2e-16 \*\*\*  
## Radio 0.18799 0.00804 23.382 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.681 on 197 degrees of freedom  
## Multiple R-squared: 0.8972, Adjusted R-squared: 0.8962   
## F-statistic: 859.6 on 2 and 197 DF, p-value: < 2.2e-16

Sales = 2.92110 + 0.04575\*TV + 0.18799\*Radio

1. (a) Download the “Boston” data set from R library “MASS” and run the R codes provided in the file “R-code-Session02.R”, and submit your output. (1 point)

[R-code-Session02.docx](file:///D:\duter\Documents\JHU%20MSIS\Spring%202021\Data%20Analytics\Data_Analytics\HW2\R-code-Session02.docx)

(b) For the Boston data set, provide the R code for fitting a multiple linear regression model to predict “medv” based on all the predictor variables. Which variable(s) do not have a significant linear relationship with “medv”? (1 point)

According to results generated by “R-code-Session02.R”, in the Boston data set, Indus and age variable do not have significant relationship with Sales. Output shown below:

lm.fit.2 = lm(medv ~ ., data = Boston)  
summary(lm.fit.2)

##   
## Call:  
## lm(formula = medv ~ ., data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.595 -2.730 -0.518 1.777 26.199   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.646e+01 5.103e+00 7.144 3.28e-12 \*\*\*  
## crim -1.080e-01 3.286e-02 -3.287 0.001087 \*\*   
## zn 4.642e-02 1.373e-02 3.382 0.000778 \*\*\*  
## indus 2.056e-02 6.150e-02 0.334 0.738288   
## chas 2.687e+00 8.616e-01 3.118 0.001925 \*\*   
## nox -1.777e+01 3.820e+00 -4.651 4.25e-06 \*\*\*  
## rm 3.810e+00 4.179e-01 9.116 < 2e-16 \*\*\*  
## age 6.922e-04 1.321e-02 0.052 0.958229   
## dis -1.476e+00 1.995e-01 -7.398 6.01e-13 \*\*\*  
## rad 3.060e-01 6.635e-02 4.613 5.07e-06 \*\*\*  
## tax -1.233e-02 3.760e-03 -3.280 0.001112 \*\*   
## ptratio -9.527e-01 1.308e-01 -7.283 1.31e-12 \*\*\*  
## black 9.312e-03 2.686e-03 3.467 0.000573 \*\*\*  
## lstat -5.248e-01 5.072e-02 -10.347 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.745 on 492 degrees of freedom  
## Multiple R-squared: 0.7406, Adjusted R-squared: 0.7338   
## F-statistic: 108.1 on 13 and 492 DF, p-value: < 2.2e-16

3. GPA’s (Grade Point Averages) for 16 graduating MBA students, and their GMAT scores taken

before entering the MBA program are given below. Use this data to respond to questions (a)-(c).

|  |  |
| --- | --- |
| x=GMAT | y=GPA |
| 560 | 3.20 |
| 540 | 3.44 |
| 520 | 3.70 |
| 580 | 3.10 |
| 520 | 3.00 |
| 620 | 4.00 |
| 660 | 3.38 |
| 630 | 3.83 |
| 550 | 2.67 |
| 550 | 2.75 |
| 600 | 2.33 |
| 537 | 3.75 |
| 610 | 3.85 |
| 570 | 3.30 |
| 590 | 3.50 |
| 650 | 3.65 |

1. Create a linear regression model that uses GMAT scores as a predictor of GPA. Obtain and interpret the coefficient of determination *R2.*

The output model only has an R-squared of 0.0818, which implies only 8.18% GPA can be explained using GMAT. Detail results shown below:

GPA.lm <- lm(GPA ~ GMAT)  
summary(GPA.lm)

##   
## Call:  
## lm(formula = GPA ~ GMAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.07025 -0.21217 0.04427 0.35855 0.54359   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.571451 1.588509 0.989 0.339  
## GMAT 0.003048 0.002729 1.117 0.283  
##   
## Residual standard error: 0.4707 on 14 degrees of freedom  
## Multiple R-squared: 0.0818, Adjusted R-squared: 0.01622   
## F-statistic: 1.247 on 1 and 14 DF, p-value: 0.2829

1. Calculate the fitted value for the fourth student on the list (GMAT = 580).

The predicted GPA value where a student has a GMAT score of 580 is 3.339291

GPA.pred <-  
 predict(GPA.lm, data.frame(GMAT = 580), interval = "confidence")  
GPA.pred

## fit lwr upr  
## 1 3.339291 3.086886 3.591697

1. Test whether GMAT is an important variable using a significance level of 0.05.

H\_0: Coefficient\_GMAT = 0

H\_a: Coefficient\_GMAT =/= 0

According to GMAT variable’s p value (0.283) provided below, which is greater than the significance level of 0.05; thus, we do not reject H–0.

**summary**(GPA.lm)

##   
## Call:  
## lm(formula = GPA ~ GMAT)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.07025 -0.21217 0.04427 0.35855 0.54359   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.571451 1.588509 0.989 0.339  
## GMAT 0.003048 0.002729 1.117 0.283  
##   
## Residual standard error: 0.4707 on 14 degrees of freedom  
## Multiple R-squared: 0.0818, Adjusted R-squared: 0.01622   
## F-statistic: 1.247 on 1 and 14 DF, p-value: 0.2829

(Total 3 points for parts (a)-(b))

4. Use the rnorm() function to create a vector of 150 observations drawn from a N(0,1)

distribution (call this vector x), and another vector of 150 observations drawn from a

N(0, 0.2) distribution (call this vector Error). Use these to create a vector *y* according to

the model *Y = -1.5 + 0.8 X + Error.* Consider this data to answer questions (a) and (b).

1. Fit a least squares linear model to predict y using x. How do  and  compare to the actual values of  and  ?

As the model shown below, and are approximate to and . Since and are both significant and the model yield a 0.94595 adjusted R-squared; thus, the fitted model is statistically identical to the original formular.

model <- lm(y ~ x)  
summary(model)

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.46686 -0.12056 -0.01488 0.13997 0.50503   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.52264 0.01572 -96.84 <2e-16 \*\*\*  
## x 0.82022 0.01550 52.93 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1925 on 148 degrees of freedom  
## Multiple R-squared: 0.9498, Adjusted R-squared: 0.9495   
## F-statistic: 2802 on 1 and 148 DF, p-value: < 2.2e-16

1. What are the 95% confidence intervals for  and  and what is the prediction interval for a case with x = 1?

The 95% confidence intervals for are: -1.553709 and -1.491568 and for are: 0.7895934 and 0.8508385.

As shown below, with x=1, the prediction intervals are -1.085306 and -0.3195389.

confint(model, '(Intercept)', level = .95)

## 2.5 % 97.5 %  
## (Intercept) -1.553709 -1.491568

confint(model, 'x', level = .95)

## 2.5 % 97.5 %  
## x 0.7895934 0.8508385

pred <- predict(model, data.frame(x = 1), interval = "prediction")  
pred

## fit lwr upr  
## 1 -0.7024223 -1.085306 -0.3195389

(Total 2 points for parts (a)-(b))