STAT 43000/STAT 53001 Applied Statistics Spring 2023 Homework 2

Due Date : February 20, 2023

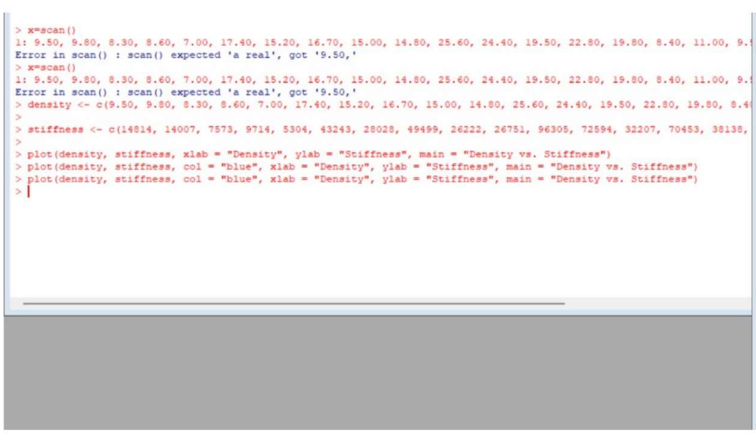
Name:
PUID:

Q.N. 1) In the manufacture of commercial wood products, it is important to estimate the relationship between the density of a wood product and its stiffness. A relatively new type of particleboard is being considered that can be formed with considerably more ease than the accepted commercial product. It is necessary to know at what density the stiffness is comparable to that of the well-known, well-documented commercial product. A study was done by Terrance E. Conners, Investigation of Certain Mechanical Properties of a Wood-Foam Composite (M.S. Thesis, Department of Forestry and Wildlife Management, University of Massachusetts). Thirty particleboards were produced at densities ranging from roughly 8 to 26 pounds per cubic foot, and the stiffness was measured in pounds per square inch. Table below shows the data.

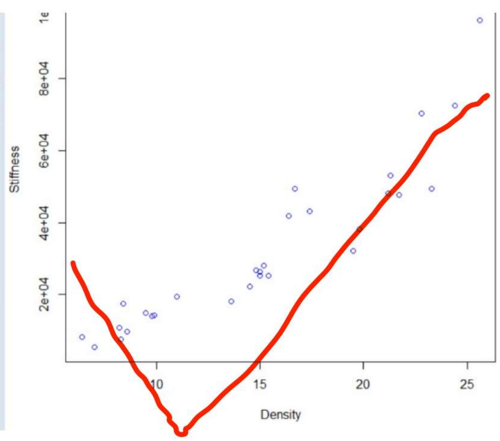
Density: 9.50, 9.80, 8.30, 8.60, 7.00, 17.40, 15.20, 16.70, 15.00, 14.80, 25.60, 24.40, 19.50, 22.80, 19.80, 8.40, 11.00, 9.90, 6.40, 8.20, 15.00, 16.40, 15.40, 14.50, 13.60, 23.40, 23.30, 21.20, 21.70, 21.30

Stiffness:14814, 14007, 7573, 9714, 5304, 43243, 28028, 49499, 26222, 26751, 96305, 72594, 32207, 70453, 38138, 17502, 19443, 14191, 8076, 10728, 25319, 41792, 25312, 22148, 18036, 104170, 49512,
48218, 47661, 53045

a) Import and read the data in R and display it graphically.



```
> x=scan()
1: 9.50, 9.80, 8.30, 8.60, 7.00, 17.40, 15.20, 16.70, 15.00, 14.80, 25.60, 24.40, 19.50, 22.80, 19.80, 8.40, 11.00,
Error in scan() : scan() expected 'a real', got '9.50,'
> x=scan()
1: 9.50, 9.80, 8.30, 8.60, 7.00, 17.40, 15.20, 16.70, 15.00, 14.80, 25.60, 24.40, 19.50, 22.80, 19.80, 8.40, 11.00,
Error in scan() : scan() expected 'a real', got '9.50,'
> density <- c(9.50, 9.80, 8.30, 8.60, 7.00, 17.40, 15.20, 16.70, 15.00, 14.80, 25.60, 24.40, 19.50, 22.80, 19.80,
>
> stiffness <- c(14814, 14007, 7573, 9714, 5304, 43243, 28028, 49499, 26222, 26751, 96305, 72594, 32207, 70453, 38138,
>
> plot(density, stiffness, xlab = "Density", ylab = "Stiffness", main = "Density vs. Stiffness")
> plot(density, stiffness, col = "blue", xlab = "Density", ylab = "Stiffness", main = "Density vs. Stiffness")
> plot(density, stiffness, col = "blue", xlab = "Density", ylab = "Stiffness", main = "Density vs. Stiffness")
>
```



b) Fit a simple linear regression model by choosing appropriate response variable and regressorvariable.

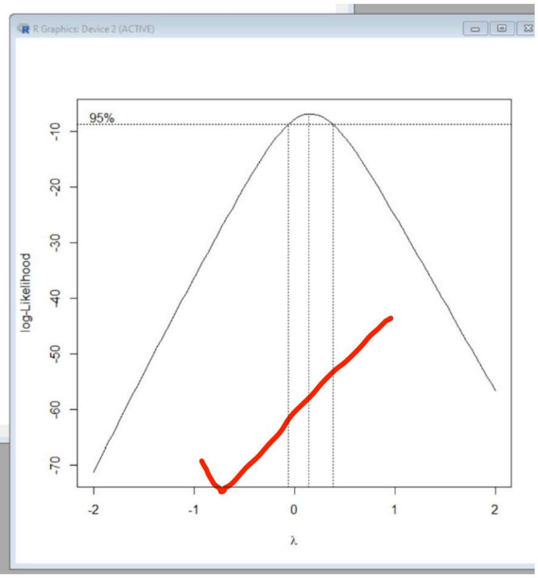
```
> lm_model <- lm(stiffness ~ density)
>
> lm_model

Call:
lm(formula = stiffness ~ density)

Coefficients:
(Intercept)      density
     -25434         3885
```

c) Perform the residual analysis and comment on the appropriateness of the model.

The residual standard error 11620 is a measure of the average magnitude of the residuals. Cause a lower value indicates a better fit, the appropriateness may not be good enough.

d) Use Box-Cox transformation and check whether transformation would improve the model.

```
> summary(lm_model)

Call:
lm(formula = stiffness ~ density)

Residuals:
   Min     1Q Median     3Q    Max
-18116  -8413   1121   3535  38695

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -25434       6105  -4.166 0.000269 ***
density         3885        370  10.500 3.25e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11620 on 28 degrees of freedom
Multiple R-squared:  0.7975,    Adjusted R-squared:  0.7902
F-statistic: 110.2 on 1 and 28 DF,  p-value: 3.245e-11

> library(MASS)
> lambda <- boxcox(lm_model)
Error in model.frame.default(formula = stiffness ~ density, drop.unused.levels = TRUE) :
  invalid type (closure) for variable 'density'
> m2=boxcox(lm_model)
Error in model.frame.default(formula = stiffness ~ density, drop.unused.levels = TRUE) :
  invalid type (closure) for variable 'density'
> data <- data.frame(density = density, stiffness = stiffness)
> lm_model <- lm(stiffness ~ density, data = data)v
Error: unexpected symbol in "lm_model <- lm(stiffness ~ density, data = data)v"
> lm_model <- lm(stiffness ~ density, data = data)
> lambda <- boxcox(lm_model)
>
```



```
> lm_model <- lm(stiffness ~ density, data = data)
> lambda <- boxcox(lm_model)
> transformed_lm_model <- lm((stiffness^0.2) ~ density)
> summary(lm_model)$r.squared
[1] 0.797458
> summary(transformed_lm_model)$r.squared\
Error: unexpected '\\' in "summary(transformed_lm_model)$r.squared\"
> summary(transformed_lm_model)$r.squared
[1] 0.9073087
> |
```

Based on the R-squared values, the original model has an R-squared value of 0.7975, while the transformed model has an R-squared value of 0.7883. This suggests that the Box-Cox transformation has not improved the model, and the original model is the better choice for this data.

Q.N. 2) Observations of the yield of a chemical reaction taken at various temperatures were recorded and are provided in the table below

$x(^0C)$	$y(\%)$
150	77.4, 76.7, 78.2
200	84.1, 84.5, 83.7
250	88.9, 89.2, 89.7
300	94.8, 94.7, 95.9

Estimate the linear model $\hat{y} = b_0 + b_1x$ and test for lack of fit.

```
> model

Call:
lm(formula = yield ~ temp, data = df)

Coefficients:
(Intercept)         temp
    60.2633       0.1165
```

y=60.2633+0.1165x

```
> anova(model, test = "Lack of fit")
Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value    Pr(>F)
temp       1 509.25  509.25  1317.3 5.994e-12 ***
Residuals 10   3.87    0.39
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.
> |
```

The p-value here 5.994e-12 is quite small, which means we do not reject the null hypothesis of no lack of fit at the 5% significance level. This suggests that the linear model is a reasonable fit for the data.

Q.N. 3) Grade point average of 12 graduating MBA students, GPA, and their GMAT scores taken before entering the MBA program are given below.

$x = GMAT$	$y = GPA$
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

Using the matrix method , obtain the following:

i) $(\mathbf{X}'\mathbf{X})^{-1}$

```
> x <- c(560,540,520,580,520,620,(
> y <- c(3.20,3.44,3.70,3.10,3.00,
> X <- cbind(rep(1,length(x)), x)
> Y <- y
>
> solve(t(X) %*% X)
                          x
  14.30369582 -2.484991e-02
x -0.02484991  4.342492e-05
>
```

ii) **b**

```
> solve(t(X) %*% X) %*% t(X) %*% Y
          [,1]
  2.157610761
x 0.001930781
>
```

e

```
> Y_hat <- X %*% solve(t(X) %*% X) %*% t(X) %*% Y
> Y - Y_hat
             [,1]
 [1,] -0.03884794
 [2,]  0.23976768
 [3,]  0.53838329
 [4,] -0.17746355
 [5,] -0.16161671
 [6,]  0.64530522
 [7,] -0.05192600
 [8,]  0.45599742
 [9,] -0.54954013
[10,] -0.46954013
[11,] -0.98607916
[12,]  0.55556002
>
```

iii) **H**

```
> H <- X %*% solve(t(X) %*% X) %*% t(X)
> H
            [,1]         [,2]        [,3]       [,4]        [,5]        [,6]         [,7]         [,8]         [,9]        [,10]
 [1,] 0.08984979  0.100488892  0.11112800 0.07921068  0.11112800  0.05793247  0.036654254  0.052612914  0.095169339  0.095169339
 [2,] 0.10048889  0.128497968  0.15650704 0.07247982  0.15650704  0.01646166 -0.039556487  0.002457127  0.114493430  0.114493430
 [3,] 0.11112800  0.156507044  0.20188609 0.06574895  0.20188609 -0.02500914 -0.115767228 -0.047698660  0.133817521  0.133817521
 [4,] 0.07921068  0.072479816  0.06574895 0.08594154  0.06574895  0.09940327  0.112864996  0.102768701  0.075845248  0.075845248
 [5,] 0.11112800  0.156507044  0.20188609 0.06574895  0.20188609 -0.02500914 -0.115767228 -0.047698660  0.133817521  0.133817521
 [6,] 0.05793247  0.016461665 -0.02500914 0.09940327 -0.02500914  0.18234487  0.265286478  0.203080275  0.037197066  0.037197066
 [7,] 0.03665425 -0.039556487 -0.11576723 0.11286500 -0.11576723  0.26528648  0.417707960  0.303391848 -0.001451116 -0.001451116
 [8,] 0.05261291  0.002457127 -0.04769866 0.10276870 -0.04769866  0.20308027  0.303391848  0.228158168  0.027535020  0.027535020
 [9,] 0.09516934  0.114493430  0.13381752 0.07584525  0.13381752  0.03719707 -0.001451116  0.027535020  0.104831385  0.104831385
[10,] 0.09516934  0.114493430  0.13381752 0.07584525  0.13381752  0.03719707 -0.001451116  0.027535020  0.104831385  0.104831385
[11,] 0.06857157  0.044470741  0.02036991 0.09267241  0.02036991  0.14087407  0.189075737  0.152924488  0.056521157  0.056521157
[12,] 0.10208476  0.132699329  0.16331390 0.07147019  0.16331390  0.01024104 -0.050988098 -0.005066241  0.117392044  0.117392044
           [,11]        [,12]
 [1,] 0.06857157  0.102084758
 [2,] 0.04447074  0.132699329
 [3,] 0.02036991  0.163313901
 [4,] 0.09267241  0.071470187
 [5,] 0.02036991  0.163313901
 [6,] 0.14087407  0.010241045
 [7,] 0.18907574 -0.050988098
 [8,] 0.15292449 -0.005066241
 [9,] 0.05652116  0.117392044
[10,] 0.05652116  0.117392044
[11,] 0.11677324  0.040855616
[12,] 0.04085562  0.137291515
> |
```

iv) SSE and MSE.

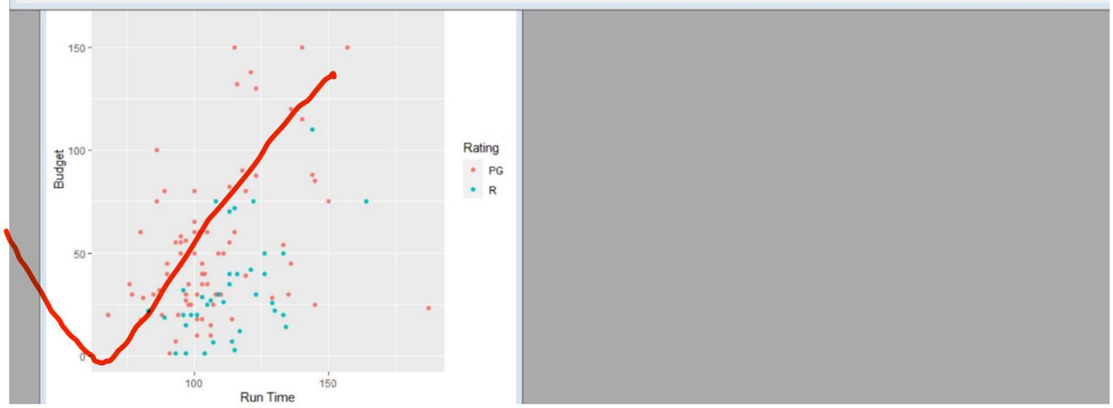
```
[12,] 0.04085562  0.137291515
> n <- length(Y)
> p <- ncol(X) - 1
> e <- Y - X %*% solve(t(X) %*% X) %*% t(X) %*% Y
> SSE <- t(e) %*% e
> MSE <- SSE / (n - p)
> SSE
         [,1]
[1,] 2.836978
> MSE
          [,1]
[1,] 0.2579071
> |
```

MSE=SSE/(n-2)=SSE/10

Q.N. 4) How does the cost of a movie depend on its length? Data on the cost (millions of dollars) and the running time (minutes) for major release films in one recent year are provided in the Brightspace along with this assignment.

a) Draw a scatter plot of Time vs. Budget. Also choose different colors to display MPAA Ratingsof the movies.

```
> library(ggplot2)
> ggplot(movie_data, aes(x = `Run Time`, y = Budget, color = `Rating`))+geom_point()+labs(title = "Scatter plot of Run Time vs. Budget",x = "Run Time", y = "Bu
Error in ggplot(movie_data, aes(x = `Run Time`, y = Budget, color = Rating)) :
  object 'movie_data' not found
> ggplot(Q1, aes(x = `Run Time`, y = Budget, color = `Rating`))+geom_point()+labs(title = "Scatter plot of Run Time vs. Budget",x = "Run Time", y = "Budget")
Error in `geom_point()`:[redacted]
! Problem while computing aesthetics.
ℹ Error occurred in the 1st layer.
Caused by error in `FUN()`:
! object 'Run Time' not found
Run `rlang::last_error()` to see where the error occurred.
> ggplot(Q1, aes(x = `Run.Time`, y = Budget, color = `Rating`))+geom_point()+labs(title = "Scatter plot of Run Time vs. Budget",x = "Run Time", y = "Budget")
> |
```



b) Fit a regression model with indicator variable and write out the regression model.

```
> model <- lm(Budget ~ `Run.Time` + Rating, data = Q1)
> model

Call:
lm(formula = Budget ~ Run.Time + Rating, data = Q1)

Coefficients:
(Intercept)     Run.Time      RatingR  
   -32.8774       0.8029     -25.8851  
```

Budget = b0 + b1 * Run Time + b2 * Indicator(Rating = "R")

Budget= 0.8029*Runtime-25.8851*RatingR -32.8774

Q.N. 5) The electric power consumed each month by a chemical plant is thought to be related to the average ambient temperature x_1, the number of days in the month x_2, the average product purity x_3, and the tons of product produced x_4. The past year's historical data are available and are presented in the following table.

y	x_1	x_2	x_3	x_4
240	25	24	91	100

236	31	21	90	95
290	45	24	88	110
274	60	25	87	88
301	65	25	91	94
316	72	26	94	99
300	80	25	87	97
296	84	25	86	96
267	75	24	88	110
276	60	25	91	105
288	50	25	90	100
261	38	23	89	98

a) Fit a multiple linear regression model using these data set.

y= -102.71324+0.60537x1+8.92364x2+1.43746x3+0.01361x4

```
> model = lm(y ~ x1 + x2 + x3 + x4, data = data)
> model

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = data)

Coefficients:
(Intercept)           x1           x2           x3           x4
 -102.71324      0.60537      8.92364      1.43746      0.01361
```

b) Determine the coefficient of determination of the fitted model.

```
> #y= -102.71324+0.60537x1+8.92364x2+1.43746x3+0.01361x4
> summary(model)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-18.758  -9.952   3.350   6.627  23.311

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -102.71324  207.85885  -0.494    0.636
x1             0.60537    0.36890   1.641    0.145
x2             8.92364    5.30052   1.684    0.136
x3             1.43746    2.39162   0.601    0.567
x4             0.01361    0.73382   0.019    0.986

Residual standard error: 15.58 on 7 degrees of freedom
Multiple R-squared:  0.7447,	Adjusted R-squared:  0.5989
F-statistic: 5.106 on 4 and 7 DF,  p-value: 0.0303

> |
```

R-squared = 0.7447

c) Predict the power consumption for a month with $x_1 = 75, x_2 = 24, x_3 = 90$ and $x_4 = 98$.

```
> new_data=data.frame(x1 = 75, x2 = 24, x3 = 90, x4 = 98)
> prediction=predict(model, new_data)
> prediction
       1
287.5618
>
```