20BRS1176_George Mathew EDA LAB G1

linear regression, R^2, RSE, t-stats, p-vals, f-stats, residual coefficients

```
d1 <- data.frame(X=c(0.01,0.48,0.71,0.95,1.19,0.01,0.48,1.44,0.71,1.96),
                 Y=c(127.6,124,110.8,103.9,101.5,130,122,92.3,113,83.7))
d2 <- data.frame(X=c(95.2,85.1,80.6,70.5,60.2,70.2,75.1),
                 Y=c(85.9,95.2,70.3,65.4,70.5,66,71.1))
d3 <- data.frame(X=c(88.1,76.5,79.2,85.4,90.2,74.3,67.7),
                 Y=c(85.9,95.2,70.3,65.4,70.5,66,71.1))
model1 <- lm(Y~X,data=d1)
model2 <- lm(Y~X,data=d2)
model3 <- lm(Y~X, data=d3)
print(summary(model1))
```

```
##
## Call:
## lm(formula = Y ~ X, data = d1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1069 -2.4475 -0.1551  1.4394  5.3242
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  130.593      1.594   81.94 5.49e-13 ***
## X            -24.827      1.618  -15.34 3.23e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.983 on 8 degrees of freedom
## Multiple R-squared:  0.9671, Adjusted R-squared:  0.963
## F-statistic: 235.4 on 1 and 8 DF,  p-value: 3.232e-07
```

```
print(summary(model2))
```

```
##
## Call:
## lm(formula = Y ~ X, data = d2)
##
## Residuals:
##      1      2      3      4      5      6      7
## -2.036 14.373 -7.359 -5.150  7.200 -4.339 -2.688
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  20.9261    23.7724   0.880   0.4190
## X             0.7039     0.3070   2.293   0.0704 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.594 on 5 degrees of freedom
## Multiple R-squared:  0.5125, Adjusted R-squared:  0.415
## F-statistic: 5.256 on 1 and 5 DF,  p-value: 0.07042
```

```
print(summary(model3))
```

```
##
## Call:
## lm(formula = Y ~ X, data = d3)
##
## Residuals:
##      1      2      3      4      5      6      7
## 10.526 20.501 -4.556 -9.817 -4.996 -8.571 -3.087
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.2468    49.8297   1.410    0.218
## X             0.0582     0.6186   0.094    0.929
##
## Residual standard error: 12.3 on 5 degrees of freedom
## Multiple R-squared:  0.001767,   Adjusted R-squared:  -0.1979
## F-statistic: 0.008851 on 1 and 5 DF,  p-value: 0.9287
```

Question 1: dataset1

```
d1
```

```
##       X     Y
## 1  0.01 127.6
## 2  0.48 124.0
## 3  0.71 110.8
## 4  0.95 103.9
## 5  1.19 101.5
## 6  0.01 130.0
## 7  0.48 122.0
## 8  1.44  92.3
## 9  0.71 113.0
## 10 1.96  83.7
```

1. Linear Regression:

```
print(model1)
```

```
##
## Call:
## lm(formula = Y ~ X, data = d1)
```

```
##
## Coefficients:
## (Intercept)            X
##      130.59       -24.83
```

2. Residual coefficients:

```
print(model1$residuals)
```

```
##           1           2           3           4           5           6
## -2.74463668  5.32421439 -2.16549679 -3.10693454  0.45162771 -0.34463668
##           7           8           9          10
##  3.32421439 -2.54153661  0.03450321  1.76868159
```

3. R^2

```
X=c(0.01,0.48,0.71,0.95,1.19,0.01,0.48,1.44,0.71,1.96)
Y=c(127.6,124,110.8,103.9,101.5,130,122,92.3,113,83.7)
Ycap=((-24.83*X)+130.59)
print(Ycap)
```

```
##  [1] 130.3417 118.6716 112.9607 107.0015 101.0423 130.3417 118.6716  94.8348
##  [9] 112.9607  81.9232
```

```
avgY=mean(Y)
SSR=sum((Ycap-avgY)^2)
SST=sum((Y-avgY)^2)
R2=SSR/SST
sprintf("The value of R^2 is: %f",R2)
```

```
## [1] "The value of R^2 is: 0.967344"
```

4. RSE

```
rmse=sqrt(mean((Y-Ycap)^2))
sprintf("The value of RMSE is %f",rmse)
```

```
## [1] "The value of RMSE is 2.668051"
```

Inference: The value of R2 is 0.96 which is close to 1 therefore there is a very high proportion of variability in the above dataset

Question 2: dataset 2

```
d2
```

```
##      X    Y
## 1 95.2 85.9
## 2 85.1 95.2
## 3 80.6 70.3
## 4 70.5 65.4
## 5 60.2 70.5
## 6 70.2 66.0
## 7 75.1 71.1
```

1. Linear Regression:

```
print(model2)
```

```
##
## Call:
## lm(formula = Y ~ X, data = d2)
##
## Coefficients:
## (Intercept)            X
##     20.9261       0.7039
```

2. Residual coefficients:

```
print(model2$residuals)
```

```
##         1         2         3         4         5         6         7
## -2.036206 14.373059 -7.359447 -5.150183  7.199859 -4.339016 -2.688066
```

3. R^2

```
X=c(95.2,85.1,80.6,70.5,60.2,70.2,75.1)
Y=c(85.9,95.2,70.3,65.4,70.5,66,71.1)
Ycap=((0.7039*X)+20.9261)
print(Ycap)
```

```
## [1] 87.93738 80.82799 77.66044 70.55105 63.30088 70.33988 73.78899
```

```
avgY=mean(Y)
SSR=sum((Ycap-avgY)^2)
SST=sum((Y-avgY)^2)
R2=SSR/SST
sprintf("The value of R^2 is: %f",R2)
```

```
## [1] "The value of R^2 is: 0.512489"
```

4. RMSE

```
rmse=sqrt(mean((Y-Ycap)^2))
sprintf("The value of RMSE is %f",rmse)
```

```
## [1] "The value of RMSE is 7.263485"
```

The value of R2 is 0.512 which is close to 1 therefore there is a medium range of proportion of variability in the above dataset

Question 3: dataset 3

```
d3
```

```
##      X    Y
## 1 88.1 85.9
## 2 76.5 95.2
## 3 79.2 70.3
## 4 85.4 65.4
## 5 90.2 70.5
## 6 74.3 66.0
## 7 67.7 71.1
```

1. Linear Regression:

```
print(model3)
```

```
##
## Call:
## lm(formula = Y ~ X, data = d3)
##
## Coefficients:
## (Intercept)            X
##     70.2468       0.0582
```

2. Residuals

```
print(model3$residuals)
```

```
##         1         2         3         4         5         6         7
## 10.525947 20.501048 -4.556087 -9.816917 -4.996270 -8.570915 -3.086806
```

3. R2

```
X=c(88.1,76.5,79.2,85.4,90.2,74.3,67.7)
Y=c(85.9,95.2,70.3,65.4,70.5,66,71.1)
Ycap=((0.0582*X)+70.2468)
print(Ycap)
```

```
## [1] 75.37422 74.69910 74.85624 75.21708 75.49644 74.57106 74.18694
```

```
avgY=mean(Y)
SSR=sum((Ycap-avgY)^2)
SST=sum((Y-avgY)^2)
R2=SSR/SST
sprintf("The value of R^2 is: %f",R2)
```

```
## [1] "The value of R^2 is: 0.001767"
```

4. RMSE

```
rmse=sqrt(mean((Y-Ycap)^2))
sprintf("The value of RMSE is %f",rmse)
```

```
## [1] "The value of RMSE is 10.393473"
```

INFERENCE: The value of R2 is 0.00176 which is close to 0 therefore the R2 value did not explain much of the variability in the outcome from the regression model

Question 4: dataset 4

```
library(MASS)
df=survey
df=df[complete.cases(df),]
X=df$Wr.Hnd
Y=df$Pulse
d <- data.frame(X,Y)
d
```

```
##        X   Y
## 1   18.5  92
## 2   19.5 104
## 3   20.0  35
## 4   18.0  64
## 5   17.7  83
## 6   17.0  74
## 7   20.0  72
## 8   18.5  90
## 9   17.0  80
## 10  19.5  66
## 11  18.0  89
## 12  19.4  74
## 13  21.0  78
## 14  21.5  72
## 15  20.1  72
## 16  18.5  64
## 17  21.5  62
## 18  21.0  90
## 19  20.8  62
## 20  19.5  79
## 21  18.8  78
## 22  17.1  72
## 23  20.1  70
## 24  22.2  66
## 25  19.4  72
## 26  22.0  80
## 27  17.8  72
## 28  20.1  80
## 29  23.2  84
## 30  22.5  96
## 31  18.0  60
## 32  18.0  50
## 33  22.0  55
## 34  20.5  68
```

```
## 35  17.0  78
## 36  20.5  56
## 37  22.5  65
## 38  15.5  70
## 39  19.5  62
## 40  22.8  66
## 41  18.5  72
## 42  19.6  70
## 43  17.3  64
## 44  18.0  64
## 45  17.0  68
## 46  16.5  40
## 47  15.6  88
## 48  17.5  68
## 49  17.0  76
## 50  18.3  68
## 51  19.2  76
## 52  23.0  90
## 53  17.7  76
## 54  18.2  70
## 55  18.3  75
## 56  18.0  60
## 57  20.5  75
## 58  18.2  70
## 59  21.3  65
## 60  20.0  68
## 61  17.5  60
## 62  19.4  68
## 63  18.9  60
## 64  17.5  72
## 65  17.5  80
## 66  19.5  80
## 67  17.5  64
## 68  19.7  67
## 69  18.5  76
## 70  19.2  80
## 71  17.2  75
## 72  20.5  60
## 73  16.0  60
## 74  16.9  70
## 75  17.0  70
## 76  23.0  83
## 77  18.5 100
## 78  21.0 100
## 79  22.5  76
## 80  18.5  92
## 81  19.8  59
## 82  18.5  66
## 83  16.0  68
## 84  18.8  66
## 85  17.5  74
## 86  16.4  90
## 87  22.0  86
## 88  19.0  60
```

```
## 89  15.4  80
## 90  17.9  85
## 91  23.1  90
## 92  22.0  72
## 93  19.5  68
## 94  18.0  84
## 95  19.0  65
## 96  21.4  96
## 97  20.0  68
## 98  18.5  75
## 99  22.5  64
## 100 19.5  60
## 101 18.0  92
## 102 18.0  64
## 103 21.8  76
## 104 13.0  80
## 105 16.3  92
## 106 21.5  69
## 107 18.9  68
## 108 20.5  76
## 109 18.9  74
## 110 18.5  84
## 111 17.5  80
## 112 20.2  72
## 113 16.5  60
## 114 17.6  81
## 115 19.5  70
## 116 16.5  65
## 117 19.0  72
## 118 20.5  80
## 119 18.0  48
## 120 17.5  68
## 121 19.0 104
## 122 20.5  76
## 123 16.7  84
## 124 17.0  70
## 125 19.0  68
## 126 14.0  87
## 127 17.5  79
## 128 18.5  70
## 129 18.0  90
## 130 20.5  72
## 131 17.0  79
## 132 18.5  65
## 133 18.0  62
## 134 18.5  63
## 135 20.0  92
## 136 22.0  60
## 137 17.9  68
## 138 17.6  72
## 139 17.0  76
## 140 15.0  80
## 141 16.0  71
## 142 19.1  80
```

```
## 143 17.5  80
## 144 16.2  61
## 145 21.0  48
## 146 18.5  86
## 147 17.0  80
## 148 17.5  83
## 149 17.5  76
## 150 17.5  84
## 151 17.5  97
## 152 18.6  74
## 153 17.5  83
## 154 17.0  65
## 155 18.0  68
## 156 18.2  88
## 157 23.2  75
## 158 15.9  70
## 159 17.5  88
## 160 18.8  80
## 161 20.0  68
## 162 18.6  70
## 163 18.6  71
## 164 18.8  80
## 165 18.0  85
## 166 18.5  88
## 167 21.0  90
## 168 17.6  85
```

1.Linear Regression:

```
model4 <- lm(Y~X,data=d)
print(model4)
```

```
##
## Call:
## lm(formula = Y ~ X, data = d)
##
## Coefficients:
## (Intercept)            X
##    75.58608     -0.08309
```

2.Residuals:

```
print(residuals(model4))
```

```
##           1           2           3           4           5           6
##  17.95106600  30.03415483 -38.92430075 -10.09047842   8.88459493  -0.17356725
##           7           8           9          10          11          12
##  -1.92430075  15.95106600   5.82643275  -7.96584517  14.90952158   0.02584595
##          13          14          15          16          17          18
##   4.15878808  -1.79966751  -1.91599187 -10.04893400 -11.79966751  16.15878808
##          19          20          21          22          23          24
## -11.85782969   5.03415483   3.97599265  -2.16525837  -3.91599187  -7.74150532
```

```
##          25          26          27          28          29          30
##  -1.97415405   6.24187691  -2.10709619   6.08400813  10.34158351  22.28342133
##          31          32          33          34          35          36
## -14.09047842 -24.09047842 -18.75812309  -5.88275634   3.82643275 -17.88275634
##          37          38          39          40          41          42
##  -8.71657867  -4.29820050 -11.96584517  -7.69165202  -2.04893400  -3.95753629
##          43          44          45          46          47          48
## -10.14864060 -10.09047842  -6.17356725 -34.21511167  13.71010838  -6.13202284
##          49          50          51          52          53          54
##   1.82643275  -6.06555177   2.00922818  16.32496574   1.88459493  -4.07386065
##          55          56          57          58          59          60
##   0.93444823 -14.09047842   1.11724366  -4.07386065  -8.81628527  -5.92430075
##          61          62          63          64          65          66
## -14.13202284  -5.97415405 -14.01569847  -2.13202284   5.86797716   6.03415483
##          67          68          69          70          71          72
## -10.13202284  -6.94922740   1.95106600   6.00922818   0.84305051 -13.88275634
##          73          74          75          76          77          78
## -14.25665609  -4.18187614  -4.17356725   9.32496574  25.95106600  26.15878808
##          79          80          81          82          83          84
##   2.28342133  17.95106600 -14.94091852  -8.04893400  -6.25665609  -8.02400735
##          85          86          87          88          89          90
##  -0.13202284  15.77657945  12.24187691 -14.00738959   5.69349061  10.90121270
##          91          92          93          94          95          96
##  16.33327463  -1.75812309  -5.96584517   9.90952158  -9.00738959  22.19202361
##          97          98          99         100         101         102
##  -5.92430075   0.95106600  -9.71657867 -13.96584517  17.90952158 -10.09047842
##         103         104         105         106         107         108
##   2.22525914   5.49407741  17.76827056  -4.79966751  -6.01569847   2.11724366
##         109         110         111         112         113         114
##  -0.01569847   9.95106600   5.86797716  -1.90768299 -14.21511167   6.87628605
##         115         116         117         118         119         120
##  -3.96584517  -9.21511167  -2.00738959   6.11724366 -26.09047842  -6.13202284
##         121         122         123         124         125         126
##  29.99261041   2.11724366   9.80150610  -4.17356725  -6.00738959  12.57716625
##         127         128         129         130         131         132
##   4.86797716  -4.04893400  15.90952158  -1.88275634   4.82643275  -9.04893400
##         133         134         135         136         137         138
## -12.09047842 -11.04893400  18.07569925 -13.75812309  -6.09878730  -2.12371395
##         139         140         141         142         143         144
##   1.82643275   5.66025508  -3.25665609   6.00091930   5.86797716 -13.24003832
##         145         146         147         148         149         150
## -25.84121192  11.95106600   5.82643275   8.86797716   1.86797716   9.86797716
##         151         152         153         154         155         156
##  22.86797716  -0.04062512   8.86797716  -9.17356725  -6.09047842  13.92613935
##         157         158         159         160         161         162
##   1.34158351  -4.26496497  13.86797716   5.97599265  -5.92430075  -4.04062512
##         163         164         165         166         167         168
##  -3.04062512   5.97599265  10.90952158  13.95106600  16.15878808  10.87628605
```

3. R2

```
Ycap=((-0.08309*X)+75.58608)
print(Ycap)
```

```
##   [1] 74.04891 73.96582 73.92428 74.09046 74.11539 74.17355 73.92428 74.04891
##   [9] 74.17355 73.96582 74.09046 73.97413 73.84119 73.79964 73.91597 74.04891
##  [17] 73.79964 73.84119 73.85781 73.96582 74.02399 74.16524 73.91597 73.74148
##  [25] 73.97413 73.75810 74.10708 73.91597 73.65839 73.71655 74.09046 74.09046
##  [33] 73.75810 73.88273 74.17355 73.88273 73.71655 74.29818 73.96582 73.69163
##  [41] 74.04891 73.95752 74.14862 74.09046 74.17355 74.21509 74.28988 74.13200
##  [49] 74.17355 74.06553 73.99075 73.67501 74.11539 74.07384 74.06553 74.09046
##  [57] 73.88273 74.07384 73.81626 73.92428 74.13200 73.97413 74.01568 74.13200
##  [65] 74.13200 73.96582 74.13200 73.94921 74.04891 73.99075 74.15693 73.88273
##  [73] 74.25664 74.18186 74.17355 73.67501 74.04891 73.84119 73.71655 74.04891
##  [81] 73.94090 74.04891 74.25664 74.02399 74.13200 74.22340 73.75810 74.00737
##  [89] 74.30649 74.09877 73.66670 73.75810 73.96582 74.09046 74.00737 73.80795
##  [97] 73.92428 74.04891 73.71655 73.96582 74.09046 74.09046 73.77472 74.50591
## [105] 74.23171 73.79964 74.01568 73.88273 74.01568 74.04891 74.13200 73.90766
## [113] 74.21509 74.12370 73.96582 74.21509 74.00737 73.88273 74.09046 74.13200
## [121] 74.00737 73.88273 74.19848 74.17355 74.00737 74.42282 74.13200 74.04891
## [129] 74.09046 73.88273 74.17355 74.04891 74.09046 74.04891 73.92428 73.75810
## [137] 74.09877 74.12370 74.17355 74.33973 74.25664 73.99906 74.13200 74.24002
## [145] 73.84119 74.04891 74.17355 74.13200 74.13200 74.13200 74.13200 74.04061
## [153] 74.13200 74.17355 74.09046 74.07384 73.65839 74.26495 74.13200 74.02399
## [161] 73.92428 74.04061 74.04061 74.02399 74.09046 74.04891 73.84119 74.12370
```

```
avgY=mean(Y)
SSR=sum((Ycap-avgY)^2)
SST=sum((Y-avgY)^2)
R2=SSR/SST
print(R2)
```

```
## [1] 0.0001910377
```

```
sprintf("The value of R^2 is: %f",R2)
```

```
## [1] "The value of R^2 is: 0.000191"
```

4. RMSE

```
rmse=sqrt(mean((Y-Ycap)^2))
sprintf("The value of RMSE is %f",rmse)
```

```
## [1] "The value of RMSE is 11.501982"
```

INFERENCE: The value of R2 is 0.00019 which is close to 0 therefore the R2 value did not explain much of the variability in the outcome from the regression model