Week 4 - Regression Analysis - R

October 3, 2018

1 Data Warehousing and Data Mining

1.1 Labs

1.1.1 Prepared by Gilroy Gordon

Contact Information   SCIT ext. 3643
ggordonutech@gmail.com
gilroy.gordon@utech.edu.jm

1.1.2 Week 4 - Regression Analysis in R

Additional Reference Resources:

1.2 Objectives

```
> Data Transformation
> Data Mining
    > Linear Regression
    > Logistic Regression
    > Non-Linear Regression
> Model Evaluation and Prediction
    > Train/Test Split - 70/30
> Presentation
    > Scatter Plot
    > Q-Q Plots
```

```
In [2]: my.data =  read.csv('./data/Creditcardprom.csv',header=TRUE,check.names=FALSE) # read da
        my.data
```

Income Range	Magazine Promo	Watch Promo	Life Ins Promo	Credit Card Ins.	Sex	Age
40-50,000	Yes	No	No	No	Male	45
30-40,000	Yes	Yes	Yes	No	Female	40
40-50,000	No	No	No	No	Male	42
30-40,000	Yes	Yes	Yes	Yes	Male	43
50-60,000	Yes	No	Yes	No	Female	38
20-30,000	No	No	No	No	Female	55
30-40,000	Yes	No	Yes	Yes	Male	35
20-30,000	No	Yes	No	No	Male	27
30-40,000	Yes	No	No	No	Male	43
30-40,000	Yes	Yes	Yes	No	Female	41
40-50,000	No	Yes	Yes	No	Female	43
20-30,000	No	Yes	Yes	No	Male	29
50-60,000	Yes	Yes	Yes	No	Female	39
40-50,000	No	Yes	No	No	Male	55
20-30,000	No	No	Yes	Yes	Female	19

```
In [5]: reg = lm(`Age` ~ `Income Range` + `Magazine Promo` + `Watch Promo` + `Life Ins Promo` +
            data=my.data) # build the model
        reg


Call:
lm(formula = Age ~ `Income Range` + `Magazine Promo` + `Watch Promo` +
    `Life Ins Promo` + `Credit Card Ins.` + Sex, data = my.data)

Coefficients:
          (Intercept)   `Income Range`30-40,000   `Income Range`40-50,000
              41.1358                   14.1124                   12.5386
`Income Range`50-60,000       `Magazine Promo`Yes           `Watch Promo`Yes
              11.2295                   -1.7588                    3.2763
    `Life Ins Promo`Yes     `Credit Card Ins.`Yes                   SexMale
             -13.7447                    0.2248                   -6.9157



In [6]: summary(reg)


Call:
lm(formula = Age ~ `Income Range` + `Magazine Promo` + `Watch Promo` +
    `Life Ins Promo` + `Credit Card Ins.` + Sex, data = my.data)

Residuals:
     Min       1Q   Median       3Q      Max
 -10.4965  -3.2974  -0.2061   3.4555  13.8642

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)                 41.1358      7.1774    5.731   0.00122 **
`Income Range`30-40,000     14.1124     13.4316    1.051   0.33387
`Income Range`40-50,000     12.5386      7.4620    1.680   0.14389
`Income Range`50-60,000     11.2295     15.0835    0.744   0.48471
`Magazine Promo`Yes         -1.7588     11.5672   -0.152   0.88413
`Watch Promo`Yes             3.2763      6.4671    0.507   0.63050
`Life Ins Promo`Yes        -13.7447      9.2087   -1.493   0.18615
`Credit Card Ins.`Yes        0.2248      9.3598    0.024   0.98162
SexMale                     -6.9157      6.8149   -1.015   0.34936
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9.384 on 6 degrees of freedom
Multiple R-squared:  0.5825,Adjusted R-squared:  0.02589
F-statistic: 1.047 on 8 and 6 DF,  p-value: 0.4919
```

In [7]: coefficients(reg) # model coefficients

(Intercept) 41.135831381733 **'Income Range'30-40,000** 14.112412177986 **'Income Range'40-50,000** 12.5386416861827 **'Income Range'50-60,000** 11.2295081967213 **'Magazine Promo'Yes** -1.75878220140518 **'Watch Promo'Yes** 3.27634660421546 **'Life Ins Promo'Yes** -13.7447306791569 **'Credit Card Ins.'Yes** 0.224824355971896 **SexMale** -6.91569086651054

In [8]: confint(reg, level=0.95) # CIonfidence intervals for model parameters

	2.5 %	97.5 %
(Intercept)	23.573252	58.698410
'Income Range'30-40,000	-18.753558	46.978383
'Income Range'40-50,000	-5.720231	30.797514
'Income Range'50-60,000	-25.678380	48.137397
'Magazine Promo'Yes	-30.062761	26.545197
'Watch Promo'Yes	-12.548009	19.100703
'Life Ins Promo'Yes	-36.277710	8.788249
'Credit Card Ins.'Yes	-22.677751	23.127400
SexMale	-23.591091	9.759709

In [9]: fitted(reg) # predicted values

1 45 **2** 43.0210772833724 **3** 46.7587822014052 **4** 36.3302107728337 **5** 36.8618266978923 **6** 41.135831381733 **7** 33.0538641686183 **8** 37.4964871194379 **9** 46.5737704918033 **10** 43.0210772833724 **11** 43.2060889929742 **12** 23.751756440281 **13** 40.1381733021077 **14** 50.0351288056206 **15** 27.615925058548

In [10]: help(fitted)

In [12]: layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
        plot(reg)

3

```
Warning message:
not plotting observations with leverage one:
  1Warning message:
not plotting observations with leverage one:
  1
```
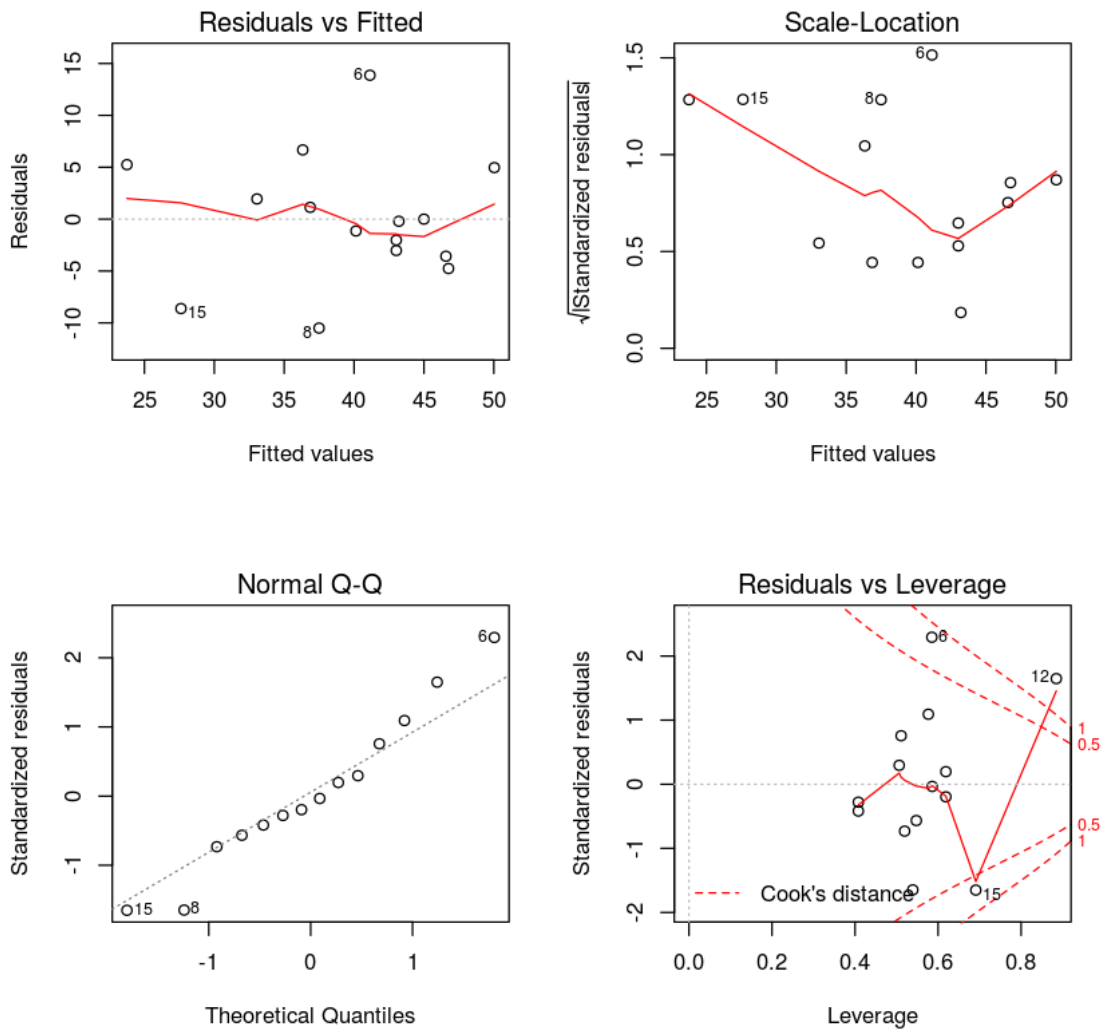

In [13]: help(prcomp)

In [14]: colnames(my.data)

1. 'Income Range' 2. 'Magazine Promo' 3. 'Watch Promo' 4. 'Life Ins Promo' 5. 'Credit Card Ins.' 6. 'Sex' 7. 'Age'

In [21]: X_columns = colnames(my.data)[1:6]
         X_columns

1. 'Income Range' 2. 'Magazine Promo' 3. 'Watch Promo' 4. 'Life Ins Promo' 5. 'Credit Card Ins.' 6. 'Sex'