STA 2312 Regression Modelling 1: Group Work Presentation.

Registration Numbers.

1). Allan Lusui: SCM211-0588/2017.

2). Karori Meshack : SCM211-0627/2017.

3). Christine Wangui: SCM211-0189/2017.

4). Jimmy Otieno: SCM211-0685/2017.

5). Raynauld Ronoh: SCM211-0248/2017.

Question.

Find an interesting set of data for which the methods of the course (Multiple Linear Regression) would be appropriate. Make up your own questions, answer them, and prepare a Results section (at most two pages)(Research on how to present results of a journal article) suitable for a research report. Prepare a 10-minute viva-voce presentation.

Note: Ensure that you have followed all the necessary steps in fitting a regression model

Problem.

Let's assume you are a small business owner at a regional delivery service(RDS) who offer same-day delivery of letters, packages and other small cargo. You are able to use google maps to group individual deliveries into one trip to reduce time and fuel cost. Therefore some trips will have more than one delivery.

As the owner you would like to estimate how a delivery will take based on three factors

1). The total distance of the trips in miles

2). The number of deliveries made during the trip

3). Daily price of gas/petrol in U.S dollars

```
#data
#total miles traveled
x1<-c(89,66,78,111,44,77,80,66,109,76)
#number of deliveries
```

```
x2<-c(4,1,3,6,1,3,3,2,5,3)
#daily gas price
x3<-c(3.84,3.19,3.78,3.89,3.57,3.57,3.03,3.51,3.54,3.25)
#total travel time(dependent variable)
y<-c(7,5.4,6.6,7.4,4.8,6.4,7,5.6,7.3,6.4)
install.packages("car")
```

#Model 1

```
#miles traveled vs total travel time
#visualizing data
plot(x1,y)
#fit the model
model1<-lm(y~x1)
summary(model1)
#checking for correlation between miles traveled and travel time
cor(x1,y, method ="pearson")
#getting the ANOVA table
anova(model1)
```

#**Results**

```
Call:
lm(formula = y ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.4425 -0.2512 -0.0211   0.2124   0.5939

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.185560   0.466951   6.822 0.000135 ***
x1          0.040257   0.005706   7.055 0.000107 ***
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3423 on 8 degrees of freedom

Multiple R-squared: 0.8615, Adjusted R-squared: 0.8442

F-statistic: 49.77 on 1 and 8 DF, p-value: 0.0001067

cor(x1,y, method ="pearson")

[1] 0.9281785

#**Model 2**

#number of deliveries vs total travel time

#visu data

plot(x2,y)

#fit the model

model2<-lm(y~x2)

summary(model2)

#checking for correlation between miles traveled and travel time

cor(x2,y, method ="pearson")

#getting the ANOVA table

anova(model2)

#**Results**

Call:

lm(formula = y ~ x2)

Residuals:

Min	1Q	Median	3Q	Max
-0.54367	-0.19061	0.05808	0.13614	0.65983

Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.84541    0.26535  18.261 8.32e-08 ***
x2           0.49825    0.07692   6.478 0.000193 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3681 on 8 degrees of freedom
Multiple R-squared:  0.8399,	Adjusted R-squared:  0.8199
F-statistic: 41.96 on 1 and 8 DF,  p-value: 0.0001926

cor(x2,y, method ="pearson")
[1] 0.9164434
```

#**Model 3**

```
#gas price vs total travel time
#visualizing data
plot(x3,y)
#fit the model
model3<-lm(y~x3)
summary(model3)
#checking for correlation between miles traveled and travel time
cor(x3,y, method ="pearson")
#getting the ANOVA table
anova(model3)
```

#**Results**

```
Call:
lm(formula = y ~ x3)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-1.6330 -0.5518   0.1116   0.6175   1.0051

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.5365     3.6490   0.969    0.361
x3            0.8113     1.0345   0.784    0.455

Residual standard error: 0.8864 on 8 degrees of freedom
Multiple R-squared:  0.0714,    Adjusted R-squared:  -0.04467
F-statistic: 0.6151 on 1 and 8 DF,  p-value: 0.4555

cor(x3,y, method ="pearson")
[1] 0.2672115
```

#**Model 4**

```
#visualizing data between the independent variables(x1,x2).Checking multicollinearity.
#miles traveled, number of deliveries vs total time traveled
#visualizing data
plot(x1+x2,y)
#fit the model
model4<-lm(y~x1+x2)
summary(model4)
cor(x1+x2,y)
#getting the ANOVA table
anova(model4)
library(car)
vif(model4)
```

#**Results**

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34711 -0.24290 -0.05702  0.17910  0.61792

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.73216    0.88697   4.208    0.004 **
x1           0.02622    0.02002   1.310    0.232
x2           0.18404    0.25091   0.733    0.487
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3526 on 7 degrees of freedom
Multiple R-squared:  0.8714,	Adjusted R-squared:  0.8347
F-statistic: 23.72 on 2 and 7 DF,  p-value: 0.0007627

cor(x1+x2,y)
[1] 0.9301224

vif(model4)
      x1       x2
11.59304 11.59304
```

#**Model 5**

```
#visualizing data between the independent variables(x1,x3).Checking multicollinearity.
#miles traveled, gas price vs total time traveled
#visualizing data
plot(x1+x3,y)
```

```
#fit the model
model5<-lm(y~x1+x3)
summary(model5)
cor(x1+x3,y)
#getting the ANOVA table
anova(model5)
library(car)
vif(model5)
```

#**Results**

```
Call:
lm(formula = y ~ x1 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49902 -0.22350 -0.00259  0.25122  0.48674

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.867570   1.482416   2.609 0.034966 *
x1           0.041370   0.006419   6.445 0.000352 ***
x3          -0.219123   0.449410  -0.488 0.640747
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3599 on 7 degrees of freedom
Multiple R-squared:  0.8661,	Adjusted R-squared:  0.8278
F-statistic: 22.63 on 2 and 7 DF,  p-value: 0.0008793

cor(x1+x3,y)
```

[1] 0.9272009

vif(model5)

```
      x1       x3
1.144939 1.144939
```

#**Model 6**

#visualizing data between the independent variables(x3,x2).Checking multicollinearity.

#gas price, number of deliveries vs total time traveled

#visualizing data

plot(x3+x2,y)

#fit the model

model6<-lm(y~x3+x2)

summary(model6)

cor(x3+x2,y)

#getting the ANOVA table

anova(model6)

library(car)

vif(model6)

#**Results**

Call:

lm(formula = y ~ x3 + x2)

Residuals:

```
     Min       1Q   Median       3Q      Max
-0.35980 -0.16634 -0.09405  0.24737  0.46784
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.32431    1.45757   5.025 0.001522 **
x3          -0.76499    0.44379  -1.724 0.128410
x2           0.56650    0.07946   7.129 0.000189 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3297 on 7 degrees of freedom
Multiple R-squared:  0.8876,	Adjusted R-squared:  0.8555
F-statistic: 27.63 on 2 and 7 DF,  p-value: 0.0004763

cor(x3+x2,y)
[1] 0.8764496

vif(model6)
      x3       x2
1.330221 1.330221
```

#**Model 7**

```
#visualizing data between the independent variables(x1,x2,x3).Checking multicollinearity.
#miles traveled, number of deliveries and gas price vs total time traveled
#visualizing data
plot(x1+x2+x3,y)
#fit the model
model7<-lm(y~x1+x2+x3)
summary(model7)
cor(x1+x2+x3,y)
#getting the ANOVA table
anova(model7)
install.packages("car")
```

```
library(car)
vif(model7)
```

#**Results**

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3183 -0.2123 -0.1218  0.2756  0.4304

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.21138    2.32057   2.677   0.0367 *
x1           0.01412    0.02221   0.636   0.5483
x2           0.38315    0.30006   1.277   0.2488
x3          -0.60655    0.52663  -1.152   0.2932
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3447 on 6 degrees of freedom
Multiple R-squared:  0.8947,	Adjusted R-squared:  0.842
F-statistic: 16.99 on 3 and 6 DF,  p-value: 0.002452

cor(x1+x2+x3,y)
[1] 0.9290681

vif(model7)
       x1        x2        x3
14.936013 17.353065  1.71380
```

Interpretation.

Model 1

Y = 3.186 + 0.0403(miles travelled)

Y = 3.186 + 0.0403(x1)

An increase in one mile will increase delivery time by 0.0403 hours

Y = 3.186 + 0.0403(84)=6.5708

Y = 6.5708 ± 2.31(0.3423)

Point estimator + T(α/2)[residual error]

5.7764 to 7.3615 hours (~95% Prediction interval)

There is a positive linear relationship between miles traveled and travel time since the p value(0.0001067) is less than the significance level (0.05)

MODEL 2

Y = 4.845 + 0.4983(deliveries)

Y = 4.845 + 0.4983(x2)

An increase in one delivery will increase delivery time by 0.4983hours

Y = 4.845 + 0.4983(4)=6.838 hours

There is a positive linear relationship between number of deliveries(x2) and travel time since the p value(0.0001926) is less than the significance level (0.05)

MODEL 3

WONT BOTHER TO CALCULATE SINCE GAS PRICE DOES NOT CONTRIBUTE TO TRAVEL TIME(NON-LINEAR)

The p value(0.4555) is greater than the significance level (0.05) therefore there is no definite relationship between gas price and travel time.

MODEL 4

Y = 3.732+ 0.0262(x1) + 0.184(x2)

The overall model(PV regression ANOVA) is significant since $r=\sqrt{R^2} = \sqrt{0.8714} = 0.9335$

indicating a strong linear relationship between x1+x2 and y.
The VIF between miles traveled and number of deliveries however is 11.59 which is greater than 10. This means there is a problematic amount of collinearity between x1 and x2.

MODEL 5

Y = 3.87 + 0.04137(x1) - 0.219(x3)

If gas price is held constant, then travel time is expected to increase by 0.04137 hours per extra mile travelled

If miles travelled are held constant travel time is expected to decrease by 0.219 hours per extra increase in gas price.

Formula suggests gas price will go up while travel time goes down.

The overall model(PV regression ANOVA) is significant since $r=\sqrt{R^2} = \sqrt{0.8661} = 0.9306$ indicating a strong linear relationship between x1+x3 and y.

The VIF between miles traveled and gas price is 1.18. This means there is absence of collinearity between x1 and x3.

MODEL 6

Y = 7.32 + 0.5565(x2) - 0.765(x3)

If gas price is held constant, then travel time is expected to increase by 0.5665 hours per delivery

If deliveries are held constant travel time is expected to decrease by 0.765 hours per extra gas price increase.

Formula suggests gas price will go up while travel time goes down.

The overall model(PV regression ANOVA) is significant since $r=\sqrt{R^2} = \sqrt{0.8876} = 0.9421$ indicating a strong linear relationship between x3+x2 and y.

The VIF between gas price and number of deliveries is 1.40. This means there is absence of collinearity between x3 and x2.

MODEL 7

Y = 6.2114+0.01412(x1)+0.38315(x2)-0.6067(x3)

The overall model(PV regression ANOVA) is significant since $r=\sqrt{R^2}=\sqrt{0.8947}=0.9459$ indicating a strong linear relationship between x1+x2+x3 and y.

The VIF=14.93,17.35,1.714 This means there is absence of collinearity between x3 and x1,x2 and a problematic amount of colinnearity between x1 and x2.

TABLE SUMMARY

Model	F-value	P-value	S	R-squared	Adj R-squared	x1	x2	x3	VIF
1	49.77	0.0001067	0.3423	0.8615	0.8442	X			1.00
2	41.959	0.0001926	0.3681	0.8399	0.8199		X		1.00
3	0.6151	0.4555	0.8864	0.0714	-0.04467			X	1.00
4	23.72	0.0007627	0.3526	0.8714	0.8347	X	X		11.59
5	22.63	0.0008793	0.3599	0.8661	0.8278	X		X	1.18
6	27.63	0.0004763	0.3297	0.8876	0.8555		X	X	1.40
7	16.99	0.002452	0.3447	0.8947	0.842	X	X	X	14.93,17.35

CONCLUSION.

When selecting best model for predicting we choose the one with:

Smallest error of regression **S**

Typically, you want to select models that have larger adjusted R-squared values. These statistics can help you avoid the fundamental problem with regular R-squared—it always

increases when you add an independent variable. This property tempts you into specifying a model that is too complex, which can produce misleading results.
High drop in R squared adjusted to R squared predicted shows ovefitting

Our best model therefore is Model1 , **Y** = 3.186 + 0.0403(x1).