Spring 2021

Section 5 (DSSA)

Regression – Part 3

Revision on ANOVA "Analysis of variance" Table :

	DF	SS	MS	F	SIGNIFICANCE
Regression	k	SSR	$MSR = SSR/k$	MSR/MSE	$P(F > MSR/MSE)$
Residual	$n - k - 1$	SSE	$MSE = SSE/(n - k - 1)$		
Total	$n - 1$	SST			

- Don't forget the relationship SST=SSR+SSE
- Df(SST)=df(SSE)+df(SSR)

Example 1:

Complete the missing cells (Highlighted in yellow) in the following Analysis of variance table developed for 80 observation data set with simple linear model.

ANOVA Summary			
	Sum	Degrees of Freedom	Mean Square
SSR (Regression)	4000		
SSE (SQ Error)			
SST (Total)	10000		
F Statistic (Calculated)			
Probability (significance)	0.023		

Answer:

ANOVA SUmmary			
	Sum	Degrees of Freedom	Mean Square
SSR (Regression)	4000	1	4000
SSE (SQ Error)	6000	78	76.92307692
SST (Total)	10000	79	
F Statistic (Calculated)	52		
Probability (significance)	0.023		

Multiple Regression Models

They are extensions to the simple linear model and allow the creation of models with several independent variables

- **Some manual calculations for multiple regressions (slope, intercept, f-test) are outside of our scope**
- **QM tool/similar tool are utilized**

The following model represent the notations multiple regression for population

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta kXk + \varepsilon$$

Where:

- Y = dependent variable (response variable)
- Xi = ith independent variable (predictor or explanatory variable)
- β0 = intercept (value of Y when all Xi = 0)
- βi = coefficient of the ith independent variable
- k = number of independent variables
- ε = random error

And the following model represents estimate, for a sample taken from the population

$$Y^\wedge = \beta0 + \beta1X1 + \beta2X2 + \ldots + \beta kXk$$

Where:

- Y^= predicted value of Y
- β0 = sample intercept (and is an estimate of β0)
- βi = sample coefficient of the ith variable (and is an estimate of βi)

Adjusted r^2 in multiple linear regressions:

- The adjusted r^2 takes into account the number of independent variables in the model.
- The adjusted r^2 value is often used to determine the usefulness of an additional variable.

$$r^2 = \frac{SSE/(n-k-1)}{SST/(n-1)}$$

Example 2:

For certain data set, a linear regression model was developed with 2 independent variables (with adjusted r^2=0.6115).

And after adding one more independent variable that we are sure, has no effect on the dependent variable. What do you think the adjusted new r^2value will be?

a) Larger than old r^2
b) The same
c) Smaller than old r^2

Answer:

(C) (As the new r^2will only increase if the newly added independent variable is significant to y)

Testing model significance

- For the whole model / or for each parameter (independent variable)
- To reject null hypothesis (the model/ parameter is significant)

$$P - Value < \propto$$

Example 3

The dataset "Cereal" contains, among other variables, the consumer reports ratings of 77 cereals available in many grocery stores. Considering "Rating" as the dependent variable, while 'grams of sugars', 'grams of fat' and 'manufacturer' as the independent variables.

Generate the regression model and the ANOVA table (using excel (data analysis->regression) and excel QM). Comment on the results. (Let alpha = 0.05)

Answer

As 'manufacturer' is qualitative data, dummy variables are needed to represent the qualitative parameter in terms of binary variable.

- A dummy variable is assigned a value of 1 if a particular condition is met and a value of 0 otherwise.
- The number of dummy variables must equal one less than the number of categories of the qualitative variable.
- Three dummy variables are used to describe the Four manufactures (General Mills, Kellogg's, Post, Quaker Oats)
- X3=1 if the manufacture is General mills and 0 other wise
- X4= 1 if the manufacture is Kellogg's and 0 other wise.
- X5=1 if the manufacture is Post and 0 other wise
- While x1 is sugar and x2 is fat
- No variable is needed for "Quaker Oats" condition since if X3,X4 and X5 = 0, then the manufacturer must b Quaker Oats,

Now the data is ready !!

- First we will use excel QM to generate the model and ANOVA Table
- Second we will use excel -data analysis -regression to generate the model ,statistical table , ANOVA Table and a final table that represents information for each parameter (independent variable)

Utilize excel QM

1. Open the Excel QM.
2. Click on the "By chapter tab" and choose (chapter 4: regression models), then choose multiple regression for both simple or multiple regression examples.
3. Make sure to choose ANOVA from Options

4. Enter the number of past observations =77 and the number of independent (X) variables=4. You can also enter a name or title for the problem. This will initialize the size of the spreadsheet.
5. Enter the data in the shaded part under Y and X1 and the calculations will be automatically added.

The regression model

y=57.5170609--2.2114651X1-1.5662556X2-3.3042068X3+6.36783522X4+4.30979775X5

ANOVA table

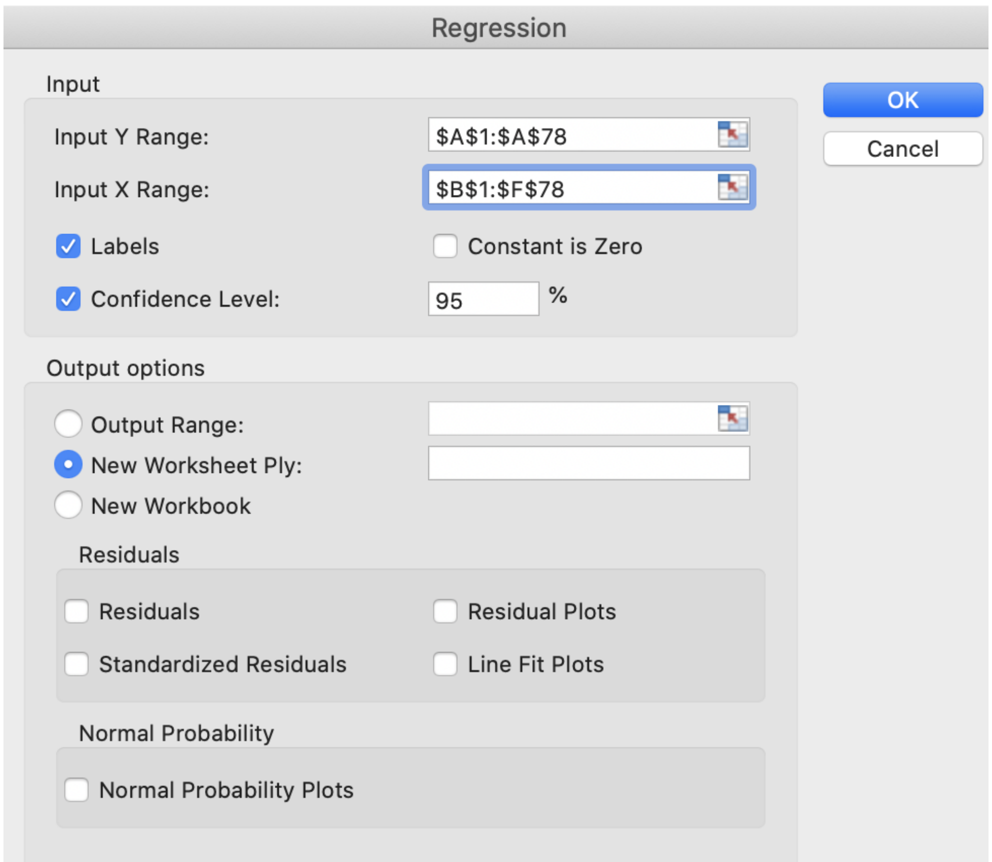
ANOVA Summary

	Sum	Degrees of Freedom	Mean Square
SSR (Regression)	10428.80368	5	2085.76074
SSE (SQ Error)	4567.99672	71	64.337982
SST (Total)	14996.8004	76	
F Statistic	32.4188088		
Probability	0.00000000		

- **In the ANOVA table, the $P-value=0$**
    - **Since $P-value<\propto$, therefore our model is significant**

Utilize excel Data analysis

1. Open the Excel.
2. Click on the data bar.
3. Click on data analysis
4. Choose regression.
5. Enter the ratings in input y range
6. Enter all independent variables in input x range .



Output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.83390762
R Square	0.69540191
Adjusted R Square	0.67395134
Standard Error	8.02109606
Observations	77

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	10428.8037	2085.76074	32.4188088	4.6525E-17
Residual	71	4567.99672	64.337982		
Total	76	14996.8004			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	57.5170609	2.69600445	21.3341862	1.3356E-32	52.1413807	62.8927411
grams of sugars (X1)	-2.2114651	0.21800289	-10.1442	1.8764E-15	-2.6461505	-1.7767796
grams of fat (X2)	-1.5662556	1.03753522	-1.5095927	0.13558502	-3.6350421	0.50253083
Manufacturer (X3)	-3.3042068	2.72142638	-1.2141452	0.22871538	-8.7305768	2.12216333
Manufacturer (x4)	6.36783522	2.75422071	2.31202793	0.02368076	0.8760751	11.8595953
Manufacturer (x5)	4.30979775	3.16808948	1.36037753	0.17801392	-2.0071933	10.6267888

From the previous tables

- We can also get as before the regression model and the significance of the model:

y=57.5170609--2.2114651X1-1.5662556X2 3.3042068X3+6.36783522X4+4.30979775X5

Significance F $= P - value = 0$

Since, $P - value < \propto$, therefore our model is significant

- Form the statistical table

  - Multiple R = correlation coefficient = r = 0.83390762
    - Strong linear relationship between the variables.
  - R Square = Coefficient of Determination (r^2) = 0.69540191
    - 69 % of the variability in Y explained by regression equation
  - Adjusted R Square = 0.67395134
    - value is often used to determine the usefulness of an additional variable
  - Standard Error = standard deviation for errors = 8.02109606
  - Observations = 77

- From the final table we can conclude which parameter is significant whose:

  $\mathrm{P} - value < \propto$

  - Only X1 and X4 are significant, (rejecting null hypothesis) meaning that those variables are responsible for the change in y.

  - As for X2, X3, and X5: since $\mathrm{p} - value > \propto$, (accepting null hypothesis)
    - Therefore, those variables are not responsible for the change in y.

  - Also, we get the interval of confidence for each parameter (lower 95% and upper 95%)
    - The interval for each parameter coefficient could take regarding the population regression model (the range for βI)
    - For example, the interval of confidence for the sugar parameter is $52.14 < \beta 1 < 62.8$

Comparing adjusted r^2

In the previous section, the model was developed for only x1 with the following statistical table

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.75967466
R Square	0.57710559
Adjusted R Square	0.57146699
Standard Error	9.1956969
Observations	77

SUMMARY OUTPUT

The adjusted r^2 is larger in the new model 0.67395134 > 0.57146699 meaning that some of the added parameters in the new model are significant to y.