Chapter5:Multiple Regression Analysis

Kevin Kam Fung Yuen
PhD, Senior Lecturer, School of Business, Singapore University of Social Sciences
kfyuen@suss.edu.sg, kevinkf.yuen@gmail.com

May 28, 2019

Contents

1 Definitions

For a data set, $Y = \{y_1, \ldots, y_i, \ldots, y_n\}$ is a vector of values for an output variable (also called dependent variable, regressand, response, explained variable, or label);$x_j = \{x_{1j}, \ldots, x_{1j}, \ldots, x_{nj}\}$, $j = 1, , m$, is a vector of values for an input variable j (also called independent variable , regressor, predictor, explanatory variable, factor, or feature) where n is the number of individuals and m0 is the number of input variables.

$x_0 = \{1\}_{nx1}$ is a vector of 1 with size of n; $X = \{x_{ij} : i \in \{1, \ldots n\}, j \in \{0, \ldots m\}\}$ is a matrix of values including and all input variables ($X_{j\neq 0}$). $\hat{Y} = \{\hat{y}_1, \ldots, \hat{y}_i, \ldots, \hat{y}_n\}$ is a set of fitted values (model, estimated or predicted values).

$W = \{w_0, \ldots, w_j, \ldots, w_m\}$ is a set of an intercept (w_0) and regression coefficients (or parameters) ($\{w_{j\neq 0}\}$). $\hat{W}$ is used to estimate W.

The estimated regression function is shown as below. $\hat{y}_i = \hat{w}_0 + \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{w}_j x_{ij}, i = 1, .., n$

In matrix form,

$$\hat{Y} = \hat{W} \cdot X$$

$\hat{Y}$ is used to estimate the true Y defined as below.

$$Y = W \cdot X$$

Since there are residual errors ε , that is

$$Y = W \cdot X = \hat{Y} + \varepsilon = \hat{W} \cdot X + \varepsilon$$

Explicitly,

$$y_i = \hat{y}_i + \varepsilon_i = \left(\widehat{w}_0 + \sum_{j=1}^{m} \hat{w}_j x_{ij} \right) + \varepsilon_i, i = 1, .., n$$

To illustrate the concepts above, let's demonstrate in R code.

```
data <-read.csv(".\\data\\airquality.csv");
#browse the data
head(data,5);

#column 1 is date, and should be removed. we update the data.
data=data[,-1];
head(data,5)
```

We call the *lm()* function for multiple linear regression. suppose we analyse the factors for PM2.5, which is in column 3 of the data matrix. So we choose *PM2.5* as dependent variable, Y. The other factors are independent variables,X.

```
model.full=lm(PM2.5 ~ .,data = data)
(summary.full=summary(model.full))
```

The *lm()* and *summary()* functions are used to calculate and display all statistics for regression. According to the results above, we have the regression model like this:

$$Y = -54.8 + 0.15X_1 + 0.06X_2 - 0.4X_3 + 0.07X_4 + 0.47X_5 + 0.16X_6 + 0.84X_7 + 0.28X_8 - 0.05X_9 + 0.01X_{10} + 0.004X_{11}$$

However, the above model is not really fit due to some statistics. Some notations are defined as below 1. t value: t_W 2. Estimate: $\hat{W}$
3. Std. Error of Estimate:$SE\,(W)$ 4. Pr(>|t|):$p\,(t_w)$
The following sections will explain how to understand the statistics.

2 Estimates of Regression

$\hat{W}$ is the estimate of W and has the form below

$$\hat{W} = \frac{Y}{X} = YX^{-1}$$

If matrix rank of X is equal to $m+1$, W has a unique solution. Otherwise, there will be the advance topics of regression problem, which is out of discussion in this unit.

The QR decomposition is one of the common approaches to find the inversion matrix, i.e.X^{-1} . The concept of QR is advanced level of mathematics, which we do not discuss here, but we can skip the calculation step and simply solve this problem by using qr.solve(X,Y) in R easily.

$\hat{W}$ is used to calculate the fitting (or model) values as below. $\hat{Y}_i = \hat{W} \cdot X$

Therefore $\hat{Y}$ is used to estimate Y with some errors for the estimation. To measure the fit of the multiple regression models, some statistics methods are used and presented in the next section.

```r
#assign data into dependent variable
Y=as.matrix(data[2])
head(Y,5)

#number of rows
(n=length(Y))
#assign data into independent variable
X=as.matrix(data[c(-2)]);
head(X,5)
```

Add vector of 1 to calculate

```r
X=cbind(rep(1,n),X);
head(X,5)
```

```r
#check matrix rank
matrixrank=qr(X)$rank
# True Value means the matrix is invetible
matrixrank == ncol(X)
```

```r
#estimate of W
W.E=qr.solve(X,Y)
W.E
```

The results above are the same as the results from *lm()* and *summary()*.

#Calculate estimate of Y

```r
Y.E =  X %*% W.E
head(Y.E,5)
head(Y,5)
```

3 Multiple Coefficient of Determination

The square root of R^2 is the multiple correlation coefficient, R, which represents the linear correlation between observed values (Y) and regression model values ($\hat{Y}$).

The Squared Multiple Correlation Coefficient (R^2) for a linear regression model is the ratio of the variances of the model variances and observed variances of the dependent variable. In other words, R^2 is used to measure the portion of variability of the data being explained by the regression model and has the form below.

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}$$

$\bar{Y}$ is the mean of Y . For the sample of small size, R^2 can be adjusted to offset bias and has the form below.

$$R^2_{adjust} = 1 - \left(1 - R^2\right) \frac{n-1}{n-m-1}$$

When the sample size n is larger, R^2 and R^2_{adjust} are closer. Therefore R^2_{adjust} is suitable for the analysis of small sample. However, regression does not mean causation; small R does not mean no relationship when R^2_{adjust} , R and R^2_{adjust} are used to draw conclusions.

```
# Same as SSR: error sum of squares
(SSE = sum((Y-Y.E)^2))

##Total sum of squares (SST)
(SST= sum((Y-mean(Y))^2))

## the regression sum of squares (SSR) or model sum of square (MSS)
(SSR = sum((Y.E-mean(Y))^2))

# R Square
(R2=  SSR/SST)

#adjusted R Square
(m=ncol(X)-1) #do not count the column with 1
(n=length(Y))  ##number of rows
(R2.adjust= 1-(1-R2)*((n-1)/(n-m-1)))
```

4 F-test

F statistics is used to test the null hypothesis (H_0) and alternative hypothesis (H_a) established as below.

$H_0 : w_0 = w_1 = \ldots = w_m = 0$

H_a: At least one w_i is not equal to 0.

To show H_a for the regression model, H_0 should be tested to be rejected at significant level α by the F statistics.

F test does not tell which variable should be rejected. So we introduce t test as below.

5 t-test for Regression Coefficients (Slopes) and Intercept

We test whether there is any statistical significant relationship between the dependent variable Y and intercept w_0 and each independent variable x_j (j = 1, 2, ..., m). The following hypotheses are established.

$$H_o : w_j = 0$$

$$H_a : w_j \neq 0$$

If the p value is less than the significant level, e.g. $\alpha \leq 0.05$, we reject the null hypothesis H_o that $w_j = 0$. In other words, there is a significant relationship between x_j and Y in the linear regression model. The input variable(s) without favoured statistical significance should be removed to form a better regression model.

6 Reduced and Full Models for Regression

"Reduced" and "Full" is a relative concept for the number of variable used. A reduced model is with less factors for the regression model. A full model includes all factors in reduced model and some new factors

not used in the reduced model. Therefore, we can use F test if either a reduced model or full model is in favour. We have the hypotheses as below. H_0 : Reduced model is in favour H_a : Full model is in favour

If $p_F \leq \alpha$, such as $\alpha = 0.05$, should be rejected and the full model is favoured. Otherwise, the reduced model is favoured and H_a is in favour. We may repeat the steps above until a good enough and simpler regression model without "unrelated" variables is found.

NO2, NOX, SO2, Cloud and Wind Speed should be removed. O3 in marginal value will go through the further process removing the 5 factors. Ultimately, only 6 factors are chosen.

As P(>F) =0.3967 > 0.05, we do not reject H_0 . That means the reduced model is in favour. We just use 6 factors instead of 11 factors.

7 References

Kevin Kam Fung Yuen, Towards multiple regression analyses for relationships of air quality and weather, Journal of Advances in Information Technology, Volume 8, No. 2, May 2017, pp.135-140.