Multiple Linear Regression Model

October 31, 2013

The Multiple Linear Regression Model

A regression model that involves more than one regressor variable is called a multiple regression model. Most real life applications of regression analysis use models that are more complex than the simple regression model. As a result the linear regression model must be extended to the multiple linear regression situation. Consider an experiment in which data are generated of the type:

$\mathbf{y}$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\cdots$	$\mathbf{x}_k$
y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

Each row of the above array represent a data point. If the experiment is willing to assume that in the region of the $x's$ defined by the data, y_i is related approximately linearly to the regressor variables, then the model formulation will be as below:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_k x_{ik} + \epsilon_i \qquad (i = 1, 2 \cdots, n);$$

Equivalently the model can be written as,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \epsilon$$

Note that this is a multiple linear regression model with k regressor variables.

Multiple Linear Regression Model

Consider a multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \epsilon$$

Remark: As in the case of simple regression, ϵ_i is a model error, assumed uncorrelated from observation to observation, with mean 0 and constant variance σ^2. In addition x_{ji} are not random and are measured with negligible error.

Remember that *Linear model is defined as a model that is linear in the parameters* β_i*,i.e; linear in the coefficients*

Example

A soft drink bottler is analyzing the vending machine service routes in his distribution system. He is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. Two key variables that determines the delivery time (y) in minutes are the number of cases(x_1) and distance walked by the route driver (x_2) in ft.

y	x_1	x_2
16.68	7	560
12.03	3	340
13.75	6	150
8.00	2	110
79.24	30	1460
40.33	16	688
13.50	4	255
24.00	9	448
15.35	6	200
9.50	3	36
17.90	10	140
18.75	9	450
10.75	4	150
11.50	3	220
14.88	4	80
18.11	7	330
17.83	7	210
21.5	5	605
21.00	10	215
19.75	6	462
29.00	10	776
19.00	7	132
35.10	17	770
52.32	26	810
19.83	8	635

We will discuss this example in the class. Data can be accessed on Blackboard.

Example

A child's birth wight depends on many things. The babies data set in UsingR package investigate the relationship between several potential variables.

Fitting the birth-weight model is straight forward. The basic model formula is

```
wt~gestation+age+ht+wt1+dage+dht+dwt
```

Note that there are few values coded as 9, 99 or 999 which should deleted from the data before developing the model

```
>library(UsingR)
>data(babies)
>attach(babies)
>newdata=subset(babies, gestation<350 &age<99 &ht<99&wt1<999&dage<99&dht<99&dwt<999)
>model=lm(wt~gestation+age+ht+wt1+dage+dht+dwt, data=newdata)
>plot(fitted(model),resid(model))
```

Identify the observation which appears to be an outlier?

```
>which(fitted(model)==min(fitted(model)))
>babies[261,]
```

Updating the model using update() function

When comparing models, we may be interested in adding or subtraction a term and refitting. rather than typing the entire model formula again we can use update the model using updat() function . The useage is

```
 update(model1, formula=.~.+ new.terms)
```

Coefficient of Multiple Determination

The coefficient of multiple determination, is denoted by R^2 , and is defined by

$$R^2 = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS}$$

and the adjusted coefficient of multiple determination, denoted by R_a^2 is given by

$$R_a^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{TSS}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{TSS}$$

and the coefficient of multiple correlation r is the positive square root of R^2
Note that in general value of R^2 never decreases when a regressor is added to the model, regardless of the value of the contribution of that variable. Therefore, it is difficult to judge whether an increase in R^2 is really telling us anything important.
Bur $R^2_{Adj.}$ will only increase on adding a variable to the model if the addition of the variable reduces the residual mean squares.

Extra Sum of Squares

An extra sum of squares measures the marginal reduction in the error sum of squares when one or several predictor variables are added to the regression model, given that other predictor variables are already in the model. Let there are two predictor variables x_1 and x_2 in the model. Let $SSE(x_1)$ be the error sum of squares if only x_1 predictor variable is taken into account and let $SSE(x_1, x_2)$ be the error sum of square if both of them are considered. Then the difference in the error sum of square on adding a new predictor is called an extra sum of squares and is given by

$$SSR(x_2|x_1) = SSE(x_1) - SSE(x_1, x_2)$$

It can be expressed in terms of the regression sum of squares (SSR) as

$$SSR(x_2|x_1) = SSR(x_1, x_2) - SSR(x_1).$$

This notion can be extended to three or more variables and we have

$$SSR(x_3|x_1, x_2) = SSE(x_1, x_2) - SSE(x_1, x_2, x_3)$$

or,

$$SSR(x_3|x_1, x_2) = SSR(x_1, x_2, x_3) - SSR(x_1, x_2)$$

and

$$\begin{aligned} SSR(x_2, x_3|x_1) &= SSE(x_1) - SSE(x_1, x_2, x_3) \\ Or \qquad SSR(x_2, x_3|x_1) &= SSR(x_1, x_2, x_3) - SSR(x_1) \end{aligned}$$

Type I and Type III Sum of Squares

The regression sum of squares can have several decomposition in terms of the ESS. For example

$$SSR(x_1, x_2, x_3, x_4) = SSR(x_1) + SSR(x_2|x_1) + SSR(x_3|x_1, x_2) + SSR(x_4|x_1, x_2, x_3)$$

This particular decomposition is called TYPE I sums of squares (variables added in order).
Note that Type I sum of squares examines the sequential incremental improvement in the fit of the model as each independent variable is added to the model. Type I sum of squares is also called the hierarchical decomposition sum of squares method.
Type III sum of squares refers to variable added last. So Type III sum of squares for a particular predictor is the reduction in error sum of squares achieved when the predictor variable is included to the model. These don't add to SSR.

$$\begin{gathered} SSR(x_1|x_2, x_3) \\ SSR(x_2|x_1, x_3) \\ SSR(x_3|x_1, x_2) \end{gathered}$$

Also note that sometimes Type I sum of squares is called the sequential sum of squares and Type III is called partial sum of squares.

Example-Soft drink data

First we regressed y on x_1 and we call it model 1. Observed that in this model SSR= 5382.4 and SSE= 402.1 and TSS= SSR+SSE= 5784.5. In this model there are 1 degrees of freedom for regression and 23 degrees of freedom for the error.

```
> data=read.table("C://STAT 4001//softdrink.txt", header=T)
> y=data$y
> x1=data$x1
> x2=data$x2
> model1=lm(y~x1) #  y is regressied on x1 only
> anova(model1)
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x1         1 5382.4  5382.4  307.85 8.22e-15 ***
Residuals 23  402.1    17.5
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

Example-Soft drink data

Conside we regressed y on both x_1 and x_2 . In this model there are 2 degrees of freedom for regression and 22 degrees of freedoms for the error.

```
> data=read.table("C://STAT 40001//softdrink.txt", header=T)
> y=data$y
> x1=data$x1
> x2=data$x2
> model2=lm(y˜x1+x2) # we conside both x1 and x2
> anova(model2)
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 5382.4  5382.4 506.619 < 2.2e-16 ***
x2         1  168.4   168.4  15.851 0.0006312 ***
Residuals 22  233.7    10.6
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

Example- Softdrink data

We have $SSR(x_1) = 5382.4$ and $SSR(x_2|x_1) = 168.4$.
Note that

$$SSR(x_1, x_2) = SSR(x_1) + SSR(x_2|x_1)).$$

Hence, $SSR = 5382.4 + 168.4 = 5550.8$
Note that

$$SSR(x_2|x_1) = SSR(x_1, x_2) - SSR(x_1) = 5550.8 - 5382.4 = 168.4$$

is the extra increase in the regression sum of squares achieved by adding x_2 to a model that has x_1 as the only predictor variable.
Since we have

$$SSR(x_2|x_1) = SSE(x_1) - SSE(x_1, x_2) = 402.1 - 233.7 = 168.4$$

Here 168.4 is referred to as the extra reduction in error sum of squares achieved by adding x_2 to a model that has x_1 as the only predictor variable. Note that, in R $SSR(x_2|x_1)$ is directly produce the extra sum of squares.
Please make a comparison of this model with the previous model. It is important to note that SSR and SSE remain the same.

Example

A hospital administrator wished to study the relationship between patients satisfaction (y) and patient's age (x_1, in years), severity of illness (x_2, an index) and anxiety level (x_3, an index). 46 patients are randomly selected and data is collected. The data is as below

y	x_1	x_2	x_3	y	x_1	x_2	x_3	y	x_1	x_2	x_3
48	50	51	2.3	57	36	46	2.3	66	40	48	2.2
70	41	44	1.8	89	28	43	1.8	36	49	54	2.9
46	42	50	2.2	54	45	48	2.4	26	52	62	2.9
77	29	50	2.1	89	29	48	2.4	67	43	53	2.4
47	38	55	2.2	51	34	51	2.3	57	53	54	2.2
66	36	49	2.0	79	33	56	2.5	88	29	46	1.9
60	33	49	2.1	49	55	51	2.4	77	29	52	2.3
52	44	58	2.9	60	43	50	2.3	86	23	41	1.8
43	47	53	2.5	34	55	54	2.5	63	25	49	2.0
72	32	46	2.6	57	32	52	2.4	55	42	51	2.7
59	33	42	2.0	83	36	49	1.8	76	31	47	2.0
47	40	48	2.2	36	53	57	2.8	80	34	49	2.2
82	29	48	2.5	64	30	51	2.4	37	47	60	2.4
42	47	50	2.6	66	43	53	2.3	83	22	51	2.0
37	44	51	2.6	68	45	51	2.2	59	37	53	2.1
92	28	46	1.8								

Example- Hospital

We would like to test whether x_3 could be dropped from the regression model retaining x_1 and x_2. In order to test

$$H_0 : \beta_3 = 0 \quad Vs. \quad H_a : \beta_3 \neq 0,$$

```
> data=read.table("C://STAT 40001//hospital.txt", header=T)
> y=data$y
> x1=data$x1
> x2=data$x2
> x3=data$x3
> model1=lm(y~x1+x2+x3)
> anova(model1)
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 8275.4  8275.4 81.8026 2.059e-11 ***
x2         1  480.9   480.9  4.7539   0.03489 *
x3         1  364.2   364.2  3.5997   0.06468 .
Residuals 42 4248.8   101.2
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

Decision: There is no enough evidence that the anxiety level is a significance variable. So x_3 could be dropped from the regression model retaining x_1 and x_2.

Example- Hospital

We would like to test whether x_2 could be dropped from the regression model retaining x_1 and x_3 in the model. In order to test

$$H_0 : \beta_2 = 0 \quad Vs. \quad H_a : \beta_2 \neq 0,$$

We can not simply look at the value of p- from the previous output and make a decision rather we have to rerun the model as

```
> y=data$y
> x1=data$x1
> x2=data$x2
> x3=data$x3
> model2=lm(y~x1+x3+x2)
> anova(model2)
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 8275.4  8275.4 81.8026 2.059e-11 ***
x3         1  763.4   763.4  7.5464  0.008819 **
x2         1   81.7    81.7  0.8072  0.374070
Residuals 42 4248.8   101.2
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

Regression Model with Quantitative and Qualitative Predictors

While estimating the regression model it is possible to include the qualitative regressor variables like gender, pain, party association etc. In order to consider gender in consideration we define the regressor variable gender x_1 as

$$x_1 = \begin{cases} 1 & \text{if gender is female} \\ 0 & \text{if gender is male} \end{cases}$$

This is also known an indicator variable.

Suppose there is one more regressor variable x_2. So a simple linear regression model will be

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

The response function for this regression model is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Note that for male the response function is

$$E(y) = \beta_0 + \beta_2 x_2$$

whereas the response function for female is

$$E(y) = \beta_0 + \beta_1 + \beta_2 x_2$$

Note that these two response functions are parallel but have different intercept.

Indicator variable with More than two levels

If a qualitative predictor variable has more than 2 classes we require additional indicator variables in the regression model. For example, suppose an electric utility is investigating the effect of the size of a single family house and the type of the air conditioning used in the house on the total electricity consumption during warm weather months. Let y be the total electricity consumption (in KWH) during June-Sept. and x_1 be the size of the house. There are four types of air conditioning systems

a) No air conditioning

b) Window units

c) heat pump

d) central air conditioning

These four levels of this factors can be modeled by three indicator variables, x_2, x_3 and x_4 defined as below

Type of Air Conditioning	x_2	x_3	x_4
No air conditioning	0	0	0
Window Units	1	0	0
Heat Pump	0	1	0
Central air conditioning	0	0	1

The regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

Example

The data set hsb2 (High school and beyond) is available at http://www.ats.ucla.edu/stat/r/modules/hsb2.csv. We will study this data

```
hsb2 <- read.table("http://www.ats.ucla.edu/stat/r/modules/hsb2.csv", header=T, sep=",")
attach(hsb2)
# creating the factor variable
race.f <- factor(race)
is.factor(race.f)
lm(write ~ (race.f))
```

We can create the indicator variable and perform the analysis

```
hsb2 <- read.table("http://www.ats.ucla.edu/stat/r/modules/hsb2.csv", header=T, sep=",")
attach(hsb2)
contr.treatment(4)  # Creating a contrast matrix
contrasts(hsb2$race.f) = contr.treatment(4)
summary(lm(write ~ race.f, hsb2))
```

Polynomial Regression Model and Multicollinearity

The linear regression model

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

is a general model for fitting any relationship that is linear in the parameters $\boldsymbol{\beta}$. This includes the important class of polynomial regression models. For example,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

which is a second order polynomial in one variable
Similarly,

$$y = \beta_0 + \beta_1 x +_1 \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

which is a second order polynomial in two variables
But both of them are linear regression model of the form $\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
Remark: Polynomials are widely used in situations where the response is curvilinear.

Polynomial Model

A k^{th} order polynomial regression model in one variable is of the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon$$

If we set $x_j = x^j, j = 1, 2, \cdots, k$ then this model reduces to

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

which is a multiple linear regression model in the k regressors.
While fitting a polynomial model in one variable one should consider the following important notes

- Order of the model
- Extrapolation
- Ill Conditioning
- Hierarchy

Example

Table below is a data concerning the strength of kraft paper(y) and the percentage of hardwood in the batch of pulp from which the paper was produced.

```
> x=c(1,1.5,2,3,4,4.5,5,5.5,6,6.5,7,8,9,10,11,12,13,14,15)
> y=c(6.3,11.1,20,24,26.1,30,33.8,34,38.1,39.9,42,46.1,53.1,
    52,52.5,48,42.8,27.8,21.9)
> x2=x^2
> model=lm(y~x+x2)
> model
Call:
lm(formula = y ~ x + x2)
Coefficients:
(Intercept)            x           x2
    -6.6742      11.7640      -0.6345
> anova(model)
Analysis of Variance Table
Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 1043.43 1043.43   53.40 1.758e-06 ***
x2         1 2060.82 2060.82  105.47 1.894e-08 ***
Residuals 16  312.64   19.54
```

Multicollinearity

In a multiple linear regression model if there is no linear relationship between the regressor variables they are said to be orthogonal. when the regressors are orthogonal, inference such as prediction, selection of a set of variables, can be made easily. Most of the time the regressor variables fail to be orthogonal. When there are near linear dependencies among the regressors, the model is said to have multicollinearity.Primary source of multicollinearity

- The data collection method employed
- Constrains on the model or in the population
- Model specification
- An overdefined model

The presence of multicollinearity has a potential effects on the least square estimators of the regression coefficients.The sample correlation r_{ij} measures the linear association between x_i and x_j. A matrix of the correlations

$$\mathbf{C} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{pmatrix}$$

provides an indication of the pairwise association s among the explanatory variables. If the off-diagonal elements of $\mathbf{C}$ are large in absolute value then there is strong pairwise linear association among the corresponding variables.

variance inflation factors

The variance inflation factors (VIF) measure how the correlation among the regressor variables inflates the variance of the estimates. If these factors are much larger than one then there is multicollinearity.
In general case of p regressors, it can be shown that the VIF of the coefficient estimate corresponding to the j^{th} regressor x_j is

$$VIF_j = \frac{1}{(1 - R_j^2)}$$

where R_j^2 is the coefficient of determination from the regression of x_j on all other regressors. If x_j is linearly dependent on the other regressors then R_j^2 will be large and VIF_j also will be large. In practice values of VIF larger than 10 are taken as solid evidence of multicollinearity

Example

A manager of pizza outlet has collected monthly sales data over 16 months period. During this time span, the outlet has been running a series of different advertisements. The manager has kept track of the cost of these advertisement (x_2) (in hundreds of dollars) as well as the number of advertisements(x_1) that have appeared and the sales in thousand in dollars (y). The data is as in the table below

```
M     x1 x2  y
Jan. 11 14.0 49.4
Feb. 8 11.8 47.5
Mar. 11 15.7 52.6
Apr. 14 15.5 49.3
May 17 19.5 61.1
Jun. 15 16.8 53.2
Jul. 12 12.8 47.4
Aug. 10 13.6 49.4
Sep. 17 18.2 62.0
Oct. 11 16.0 47.9
Nov. 8 13.0 47.3
Dec. 18 20.0 61.5
Jan. 12 15.1 54.2
Feb. 10 14.2 44.7
Mar. 13 17.3 53.6
Apr. 12 15.9 55.4
```

The estimated regression line is

$$\hat{y} = 24.8231 + 0.6626x_1 + 1.2329x_2$$

Example

```
> y<-data$y
> x1<-data$x1
> x2<-data$x2
> model=lm(y~x1+x2)
> summary(model)
lm(formula = y ~ x1 + x2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.8231     5.6611   4.385 0.000738 ***
x1            0.6626     0.5386   1.230 0.240393
x2            1.2329     0.6962   1.771 0.100011
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> library(car) # Note that car package should be installed
> vif(model)
      x1       x2
5.339243 5.339243
```

Thus, $VIF_1 = VIF_2 = 5.339$. This means the variance is inflated 5.34-fold. Since VIF is greater than 1 there is a evidence of multicollinearity.