Topics in Data Science

Lecture 8

Department of Mathematics and Statistics

PURDUE UNIVERSITY NORTHWEST

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 family doctor wishes to examine the variables that affect his female patients' total cholesterol. He randomly selects 14 of his female patients and asks them to determine their average daily consumption of saturated fat . He then measures their total cholesterol level (y in mg/dL), age (x_1 in years) and fat consumption (x_2 in g.). The data are provided below

Age (x_1)	Saturated Fat (x_2)	Total Cholesterol (y)
25	19	180
25	28	195
28	19	186
32	16	180
32	24	210
32	20	197
38	31	239
42	20	183
48	26	204
51	24	221
51	32	243
58	21	208
62	21	228
65	30	269

Source: *National Center for Health Statistics*

Example

A child's birth weight 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,]
>which(fitted(model)==max(fitted(model)))
```

Updating a Model

Updating the model using `update()` function is quick and easy
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 update() function.
`update(model1, formula=.~.+ new.terms)`
Example:

```
>library(UsingR)
>data(babies)
>attach(babies)
>newdata=subset(babies, gestation<350 &age<99 &ht<99&wt1<999&dage<99&dht<99&dwt<999)
>model1=lm(wt~gestation+age, data=newdata)
>model2=update(model1, formula=.~.+dwt, data=newdata)
```

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 $R_{Adj.}^2$ will only increase on adding a variable to the model if the addition of the variable reduces the residual mean squares.

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://CS 52540//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)
```

We can create the indicator variable and perform the analysis

Creating Indicator Variables

FEV (forced expiratory volume) is an index of pulmonary function that measures the volume of air expelled after one second of constant effort. The data contains determinations of FEV on 654 children ages 6-22 who were seen in the Childhood Respiratory Desease Study in 1980 in East Boston, Massachusetts. The data are part of a larger study to follow the change in pulmonary function over time in children.

ID–ID number, Age – years

FEV – liters,

Height - - inches

Sex - - Male or Female,

Smoker – Non = nonsmoker, Current = current smoker

The data is avaiable in the website

`http://www.statsci.org/data/general/fev.txt`.

Now we create indicator variables for Sex and Smoker as below

```
>data=read.table("http://www.statsci.org/data/general/fev.txt",header=T
> attach(data)
> D1=(Sex=="Female")*1
> D2=(Smoker=="Current")*1
```

Example

Government statisticians in England conducted a study of the relationship between smoking and lung cancer. The data concern 25 occupational groups and are condensed from data on thousands of individual men. The explanatory variables are work category and the number of cigarettes smoked per day by men in each occupation relative to the number smoked by all men of the same age. This smoking ratio is 100 if men in an occupation are exactly average in their smoking, it is below 100 if they smoke less than average, and above 100 if they smoke more than average. The response variable is the standardized mortality ratio for deaths from lung cancer. It is also measured relative to the entire population of men of the same ages as those studied, and is greater or less than 100 when there are more or fewer deaths from lung cancer than would be expected based on the experience of all English men.

Occupation: occupation category (Outdoor/Factory/Office)

Smoke.Index: relative ranking of amount of smoking, with 100 being average

Cancer.Index: relative ranking of deaths due to lung cancer, with 100 being average

Example- Smoking and Cancer

	Smoke.Index	Cancer.Index	Occupation
1	77	84	Outdoor
2	137	116	Outdoor
3	117	123	Factory
4	94	128	Factory
5	116	155	Factory
6	102	101	Factory
7	111	118	Office
8	93	113	Outdoor
9	88	104	Factory
10	102	88	Factory
11	91	104	Factory
12	104	129	Factory
13	107	86	Factory
14	112	96	Factory
15	113	144	Factory
16	110	139	Office
17	125	113	Outdoor
18	133	125	Outdoor
19	115	146	Outdoor
20	105	115	Office
21	87	79	Office
22	91	85	Office
23	100	120	Outdoor
24	76	60	Office
25	66	51	Office

Example

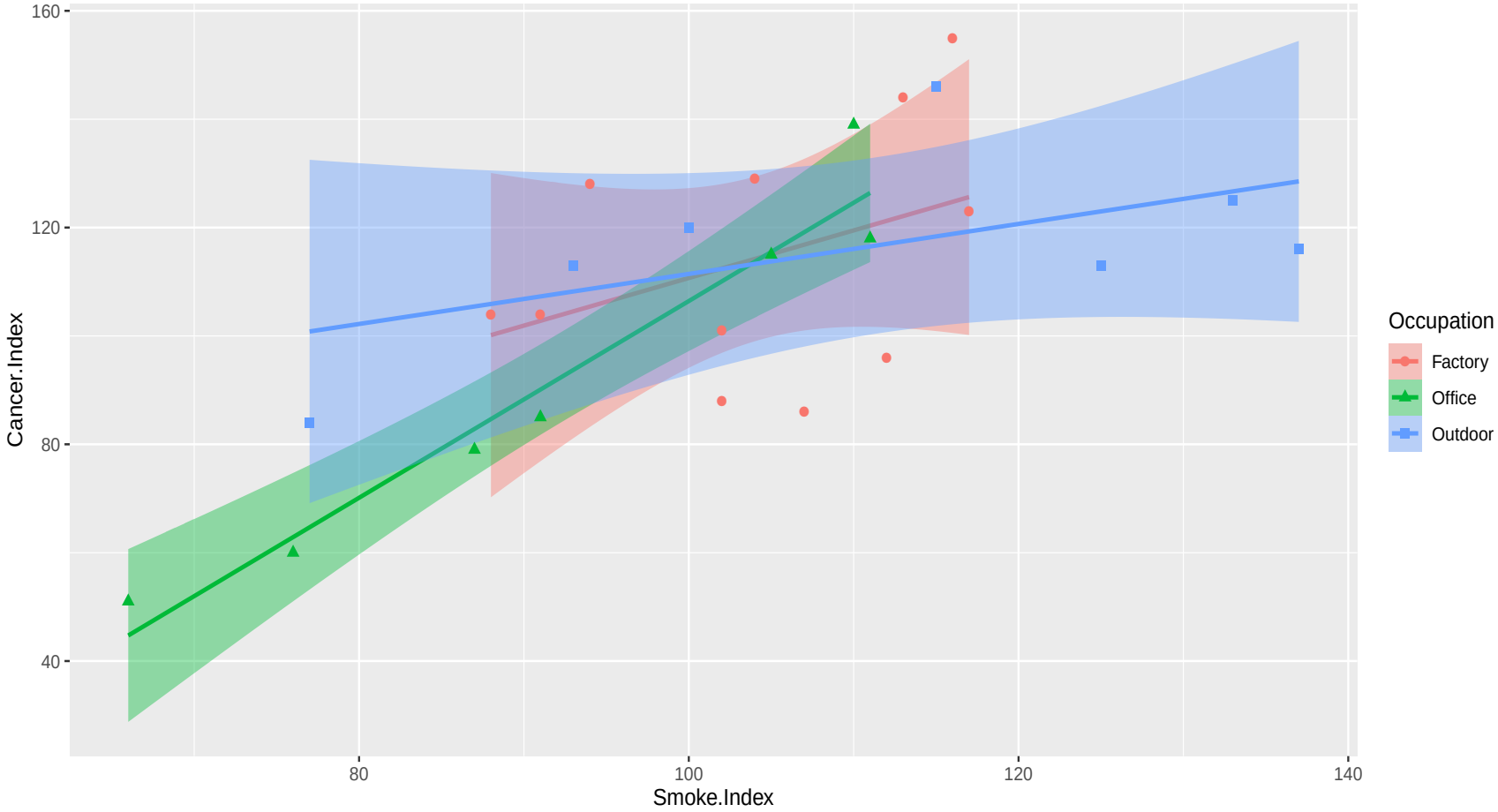
```
> Smoke.Index <- c( 77 , 137 , 117 , 94 , 116 , 102 , 111 , 93 , 88 , 102,
+ 91 , 104 , 107 , 112 , 113 , 110 , 125 , 133 , 115 , 105 , 87 , 91 ,
+ 100 , 76 , 66 )
> Cancer.Index <- c( 84 , 116 , 123 , 128 , 155 , 101 , 118 , 113 , 104 ,
+ 88 , 104 , 129 , 86 , 96 , 144 , 139 , 113 , 125 , 146 , 115 , 79 , 85 ,
+ 120 , 60 , 51 )
> Occupation <- c("Outdoor" , "Outdoor" , "Factory" , "Factory" ,
+ "Factory" , "Factory" , "Office" , "Outdoor" , "Factory" , "Factory" ,
+ "Factory" , "Factory" , "Factory" , "Factory" , "Factory" , "Office" ,
+ "Outdoor" , "Outdoor" , "Outdoor" , "Office" , "Office" , "Office" ,
+ "Outdoor" , "Office" , "Office" )
>plot(Smoke.Index, Cancer.Index)

> model1= lm(Cancer.Index ~ Smoke.Index)
> Occupation <- as.factor(Occupation)
> model2= lm(Cancer.Index ~ Occupation)
> model3= lm(Cancer.Index ~ Smoke.Index + Occupation)
```

Question: What does -ve coefficients indicate for occupation(see output of model 3)?

using ggplot package

```
> data=data.frame(Smoke.Index, Cancer.Index, Occupation)
> library(ggplot2)
> p=ggplot(data, aes(x=Smoke.Index,y=Cancer.Index,shape=Occupation, colour=Occupation, fill=Occupation ))
> p+ geom_smooth(method="lm")+geom_point(size=2)
```

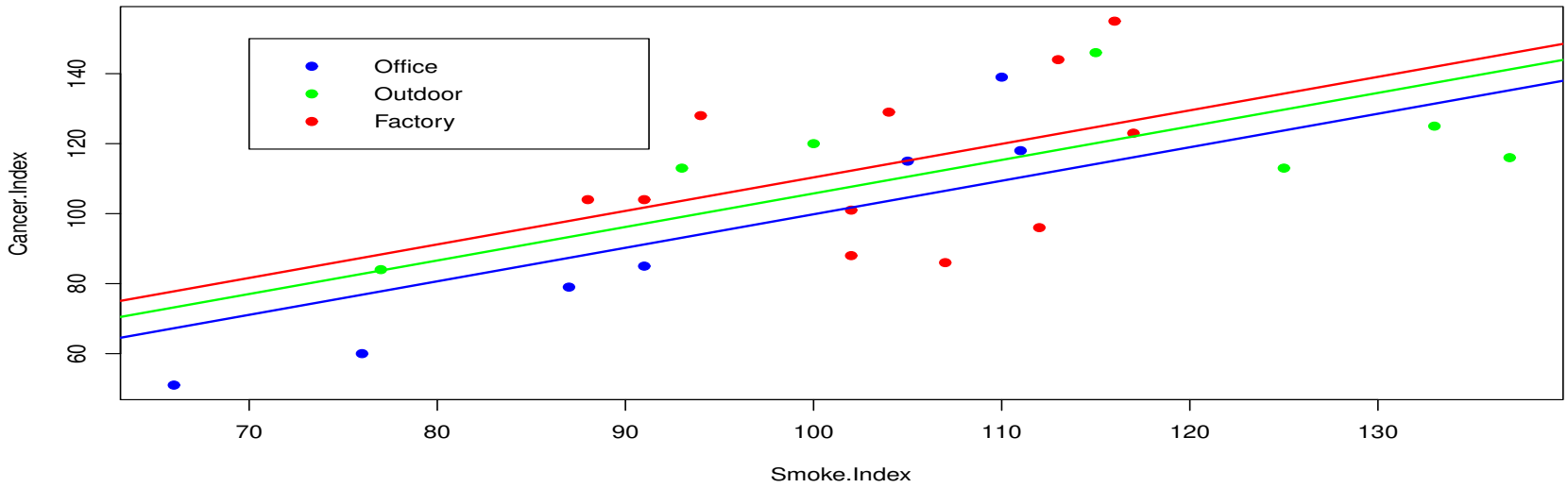


Parallel plots

```
> plot(Smoke.Index, Cancer.Index, col=ifelse(Occupation=="Office", "blue", ifelse(Occupation=="Outdoor", "green", "red")))
> plot(Smoke.Index, Cancer.Index,col=ifelse(Occupation=="Office", "blue",ifelse(Occupation=="Outdoor","green", "red")),pch=19)
> model=lm(Cancer.Index~Smoke.Index+Occupation)
> model

Coefficients:
      (Intercept)        Smoke.Index   OccupationOffice  OccupationOutdoor
          14.5868             0.9577           -10.5420            -4.5897

> abline(14.5868,0.9577, col="red", lwd=2)
> abline(14.5868-10.5420,0.9577, col="blue", lwd=2)
> abline(14.5868-4.5897,0.9577, col="green", lwd=2)
> legend(70,150, col=c("blue","green", "red"), c("Office","Outdoor","Factory"), pch=19)
```



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

Example

Belsey, Kuh and Wesch studied the economic data for 50 different countries 1960-1970. The data are named as *savings* in *faraway* library. The descriptions of the variables :
sr: savings rate - personal saving divided by disposable income
pop15: percent population under age of 15
pop75: percent population over age of 75
dpi: per-capita disposable income in dollars
ddpi: percent growth rate of dpi We will fit a polynomial regression model:

```
library(faraway)
data(savings)
attach(savings)
model1=lm(sr~ddpi)
```

The p-value of ddpi is significant so move to a quadratic term:

```
model2=lm(sr~ddpi+I(ddpi^2))
```

The p-value of $ddpi^2$ is significant so move to a cubic term:

```
model3=lm(sr~ddpi+I(ddpi^2)+I(ddpi^3))
```

Stick with quadratic model as cubic term is not significant.

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

Model Selection

The Principle of Parsimony:

A model should contain the smallest number of variables necessary to fit the data.

When there are many possible explanatory variables, often times several models are nearly equally good at explaining variation in the response variable.

- R^2 and adjusted R^2 measure closeness of fit, but are poor criteria for variable selection.
- AIC and BIC are sometimes used as objective criteria for model selection.
- Stepwise regression searches for best models, but does not always find them.
- Models selected by AIC or BIC are often over fit.
- Tests after model selection are not valid, typically.
- Parameter interpretation is complex.

R^2 and Adjusted R^2

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 $R^2_{Adj.}$ is given by

$$R^2_{Adj.} = 1 - \frac{\frac{SSE}{n-p}}{\frac{TSS}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{TSS}$$

which can be written as

$$R^2_{Adj.} = 1 - \frac{\frac{SSE}{n-p}}{\frac{TSS}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2).$$

Remark: In general the 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.
But $R^2_{Adj.}$ will only increase on adding a variable to the model if the addition of the variable reduces the residual mean squares.

Akaike's Information Criterion (AIC)

Akaike's Information Criterion (AIC) is based on maximum likelihood and a penalty for each parameter.
The expression for AIC for a general model is

$$AIC = -2 \log L + 2p$$

where L is the likelihood and p is the number of parameters.
In multiple linear regression model with with $p-1$ regressor variables and n observations it becomes

$$AIC = n \log \left(\frac{SSE}{n} \right) + 2p$$

where SSE is the residual sum of squares.
A model having minimum AIC is considered to be a better model.

Bayesian Information Criterion (BIC)

Schwartz's Bayesian Information Criterion (BIC) is similar to AIC but penalizes additional parameters more.
The general form is

$$BIC = -2 \log L + (\log n)p$$

where n is the number of observations. L is the likelihood and p is the number of parameters.
In multiple linear with $p-1$ regressor variables it becomes

$$BIC = n \log \left(\frac{SSE}{n} \right) + (\log n)p$$

where SSE is the residual sum of squares .
A model having minimum BIC is considered to be a better model.
Observe that for both the AIC and BIC the first term is the same which decreases as p increases and the second term increases with the number of parameters, p. If $n \geq 8$ the penalty term for BIC is larger than that for AIC , hence BIC criteria tends to favor more parsimonious models.

Mallow's C_p Statistics

The Mallow's C_p statistic is based on the normalized expected total error of estimation. After substituting the sample estimator s^2 for σ^2, we have

$$C_p = \frac{SSE}{s^2} + 2p - n$$

A value of C_p near p suggests that the model bias is small which means there is no significant over-fitting or under-fitting of the model. Values of c_p near or below p is generally desirable.
If the p-term model has negligible bias then SSE=0. Consequently, $E(SSE) = (n-p)\sigma^2$ and

$$E(C_p|Bias = 0) = \frac{(n-p)\sigma^2}{\sigma^2} - n + 2p = p$$

PRESS Statistic

For a specified model the PRESS (prediction sum of squares) statistic is formed by predicting each observation based on a model developed by using the other observations. Let us define the PRESS residual as

$$e_{(i)} = y_i - \hat{y}_{(i)} \quad i = 1, 2, 3, \cdots n$$

where $\hat{y}_{(i)}$ is the fitted value of the i^{th} response based on all $n-1$ observations except the i^{th} one. Note that large PRESS residuals are potentially useful in identifying observations where the model does not fit the data well or observations for which the model is likely to provide poor future predictions. We define the PRESS statistic as

$$PRESS = \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2$$

Note that models with smaller PRESS statistics are preferred.
PRESS can be computed in R using library MPV :

```
library(MPV)
PRESS(model)
```

Variable Selection

Some testing based procedures for variable selection are

- Forward Selection
- Backward Elimination
- Stepwise Regression

Forward selection assumes a model with an intercept only and adds the most significant (smallest p-value) variables one at a time. The function *add*1() in R is used to find out which variable will be added in the model
Backward elimination starts with all the variables in the model and eliminates variables with the largest (least significant) p-values:
The function *drop*1() in R is used to find out which variable will be added in the model
This is a combination of backward elimination and forward selection. This allows to reenter the variable that has been dropped in the backward elimination method.
In order to perform the stepwise selection we need library named *MASS*.

Example-Survival Time

A hospital studied the survival time(y) of patients who had undergone a liver operation. The regressor variables for the data includes

x_1: blood clotting score

x_2: prognostic index

x_3: enzyme function test score

x_4: liver function test score

x_5: age, in years

x_6: gender,(0=male, 1=female)

x_7 : moderate use of alcohol

x_8: severe use of alcohol

Sample of the data is as below (Complete data in the Brightspace)

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	y
6.7	62	81	2.59	50	0	1	0	695
5.1	59	66	1.70	39	0	0	0	403
7.4	57	83	2.16	55	0	0	0	710
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
6.4	59	85	2.33	63	0	1	0	550
8.8	78	72	3.20	56	0	0	0	651

Example-Survival Time- Forward Selection

```
> add1(lm(y~1), scope=(~.+x1+x2+x3+x4+x5+x6+x7+x8),test="F")
Single term additions

Model:
y ~ 1
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>              8369521 647.36
x1      1   1005152 7364369 642.45  7.0974  0.010255 *
x2      1   1479767 6889754 638.85 11.1685  0.001547 **
x3      1   2798310 5571211 627.38 26.1186 4.671e-06 ***
x4      1   3804272 4565248 616.63 43.3322 2.289e-08 ***
x5      1    118863 8250658 648.59  0.7491  0.390726
x6      1    251809 8117712 647.71  1.6130  0.209721
x7      1    271062 8098458 647.58  1.7405  0.192857
x8      1   1454057 6915463 639.06 10.9336  0.001717 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that x4 has the smallest p-value so it will be added in the model.

Example-Survival Time-Forward Selection

```
> add1(lm(y~1+x4), scope=(~.+x1+x2+x3+x5+x6+x7+x8),test="F")
Single term additions
Model:
y ~ 1 + x4
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>              4565248 616.63
x1      1       685 4564563 618.62  0.0077 0.9306162
x2      1    285592 4279656 615.14  3.4034 0.0708749 .
x3      1    896004 3669244 606.83 12.4539 0.0008935 ***
x5      1      3726 4561522 618.59  0.0417 0.8390782
x6      1      6162 4559086 618.56  0.0689 0.7939538
x7      1    225396 4339852 615.90  2.6488 0.1097942
x8      1    939077 3626171 606.19 13.2076 0.0006480 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that x8 has the smallest p-value so it will be added in the model.

The process continues until we obtain the p values for each of the regressor variables greater than desired level of α.

Example-Survival Time-Backward Elimination

```
> drop1(lm(y~x1+x2+x3+x4+x5+x6+x7+x8), test="F")
Single term deletions
Model:
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>              1825906 581.14
x1      1    263780 2089686 586.43  6.5009 0.0142584 *
x2      1    930187 2756093 601.38 22.9247 1.857e-05 ***
x3      1   1307757 3133663 608.31 32.2301 9.390e-07 ***
x4      1     51016 1876922 580.63  1.2573 0.2681093
x5      1      5231 1831137 579.30  0.1289 0.7212314
x6      1      2990 1828896 579.23  0.0737 0.7872685
x7      1       572 1826478 579.16  0.0141 0.9060073
x8      1    576636 2402542 593.96 14.2114 0.0004736 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that x7 has the largest p-value so it is eliminated from the model.

Example-Survival Time-Backward Elimination

```
> drop1(lm(y~x1+x2+x3+x4+x5+x6+x8), test="F")
Single term deletions

Model:
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x8
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>               1826478 579.16
x1      1    263219 2089697 584.43  6.6292   0.01331 *
x2      1    936544 2763022 599.51 23.5869 1.420e-05 ***
x3      1   1309433 3135911 606.35 32.9782 7.027e-07 ***
x4      1     51951 1878429 578.68  1.3084   0.25860
x5      1      4878 1831356 577.31  0.1229   0.72755
x6      1      2958 1829436 577.25  0.0745   0.78613
x8      1    726329 2552807 595.24 18.2926 9.472e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Note that x6 has the largest p-value so it is eliminated from the model.
The process continues until we obtain the p values for each of the regressor variables less than desired level of α.

Example-Survival Time-Stepwise Regression

```
> library(MASS)
> step <- stepAIC(model, direction="both")
Start:  AIC=581.14
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8
       Df Sum of Sq     RSS    AIC
- x7    1       572 1826478 579.16
- x6    1      2990 1828896 579.23
- x5    1      5231 1831137 579.30
- x4    1     51016 1876922 580.63
<none>              1825906 581.14
- x1    1    263780 2089686 586.43
- x8    1    576636 2402542 593.96
- x2    1    930187 2756093 601.38
- x3    1   1307757 3133663 608.31
Step:  AIC=579.16
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x8
       Df Sum of Sq     RSS    AIC
- x6    1      2958 1829436 577.25
- x5    1      4878 1831356 577.31
- x4    1     51951 1878429 578.68
<none>              1826478 579.16
+ x7    1       572 1825906 581.14
- x1    1    263219 2089697 584.43
- x8    1    726329 2552807 595.24
- x2    1    936544 2763022 599.51
- x3    1   1309433 3135911 606.35
Step:  AIC=577.25
y ~ x1 + x2 + x3 + x4 + x5 + x8
       Df Sum of Sq     RSS    AIC
- x5    1      4281 1833716 575.38
- x4    1     63596 1893032 577.09
<none>              1829436 577.25
+ x6    1      2958 1826478 579.16
+ x7    1       540 1828896 579.23
- x1    1    260399 2089835 582.44
- x8    1    723371 2552807 593.24
- x2    1    934511 2763947 597.53
- x3    1   1306483 3135918 604.35
Step:  AIC=575.38
```

Model Selection using regsubsets

The R package *leaps* is needed for the function *regsubsets*() for R^2_{Adj}.

```
> library(leaps)
> subsets=regsubsets(y~x1+x2+x3+x4+x5+x6+x7+x8, data=data)
> subsets
Subset selection object
Call: regsubsets.formula(y ~x1+x2+x3+x4+x5+x6+x7+x8,data=data)
8 Variables  (and intercept)
   Forced in Forced out
x1     FALSE      FALSE
x2     FALSE      FALSE
x3     FALSE      FALSE
x4     FALSE      FALSE
x5     FALSE      FALSE
x6     FALSE      FALSE
x7     FALSE      FALSE
x8     FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
```

We can add the option *nbest*= "." to choose the number of best models.

To get some additional information about the data we must use `summary` command

Model Selection using regsubsets

```
> summary(subsets)
         x1  x2  x3  x4  x5  x6  x7  x8
1  ( 1 ) " " " " " " "*" " " " " " " "
2  ( 1 ) " " " " " " "*" " " " " " "*"
3  ( 1 ) "*" "*" "*" " " " " " " " " "
4  ( 1 ) "*" "*" "*" " " " " " " " "*"
5  ( 1 ) "*" "*" "*" "*" " " " " " " "*"
6  ( 1 ) "*" "*" "*" "*" "*" " " " " "*"
7  ( 1 ) "*" "*" "*" "*" "*" "*" " " "*"
8  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*"
```

The useful part of this is the last, a "matrix" indicating which variables are included in each of the models. Models are listed in order of size (the first column), and within a size, in order of fit (best model first). The included variables are indicated by asterisks in quotes and variables not in a model have empty quotes.

Model Selection using regsubsets

An option is available which makes this matrix perhaps more readable:

```
> summary(subsets, matrix.logical=TRUE)
```

The subsets obtained above can be plotted using the code below in R

```
> plot(subsets)
```

Model Selection using AIC /BIC/Cp

AIC and BIC:

There is a function that gives AIC or Schwartz' BIC, from an lm object:

```
> model=lm(y~x1+x2+x3+x4+x5+x6+x7+x8)
> AIC(model)
[1] 736.3899
```

The Mallow's c_p and adjusted $R^2_{Adj.}$ can be obtained as

```
> library(leaps)
> subsets=regsubsets(y~x1+x2+x3+x4+x5+x6+x7+x8, data=data)
> summary(subsets)$cp
> summary(subsets)$adjr2
```

Model building steps

- Fit the largest model possible to the data
- Perform a through analysis of the model
- Determine if a transformation of the data is necessary
- Determine if all possible regressions is feasible
  - If all possible regressions is feasible perform all possible regressions using such criteria as Mallow's C_p, Adjusted R^2 and the PRESS statistic to rank the best subset models.
  - if all possible regression is not feasible , use stepwise selection techniques to generate the largest model
- Compare and contrast the best models recommended by each criteria
- Perform a through analysis of the "best" models
- Explore the need for further transformation
- Make the recommendation.

Model Validation

The final step in the model building process is the validation of the selected regression model. Model validation involves checking a candidate model against independent data. Three types of model validation procedures are

- Collection of new (or fresh) data to check the model and its predictive ability
- Comparison of the results with theoretical expectations, earlier empirical results and simulation results.
- Data Splitting, that is, set aside some of the original data and use these observations to investigate the model's predictive performance.

The final intended use of the model often indicates the appropriate validation methodology. Thus, validation of a model for prediction purpose should be as accurate as possible. If possible all the validation techniques must be used. The best means of model validation is through the collection of the new (fresh) data. Sometimes these new observations are called "conformation runs".

Diagnostics for Leverage and Influence

An outlying cases is defined as a particular observation $(y_i, x_{i1}, x_{i2}, \cdots, x_{ip})$ that differs from the majority of the cases in the data set. Since several variables are involved in each case, one must distinguish among outliers in the y(response) dimension and outliers in the x(covariate) dimension. Note that the detection of outlier will be tricky if more than two dimensions are involved.
Outliers in the y are linked to the regression model as one tries to explain the response as a function of the covariates. Outliers in the y dimension may be due to many different reasons

- The random component in the regression may be unusually large.
- The response y or the one of the response variable may have been recorded incorrectly.
- Both x and y variables are correct but there is another covariate that is missing from the model and that can explain the "strange" observation.

Several procedures for detecting outliers has been developed. A simplest step is to compute the studentized residuals d_i and it has approximately standard normal distribution as long as the model assumptions are satisfied. Large value of d_i are unexpected. For example $|d_i| > 2.5$ is indicative of an outlier.

Identifying Influential Cases

After identifying cases that are outlying with respect to their y values and /or their x values, we want to know whether they are influential. A case is said to be influential if its exclusion from the model causes major changes in the fitted regression model. We can identify the influential cases using three different measures:

- DFFITS Measure
- Cook's Distance
- DFBETAS Measure

We can use R to calculate DFFITS, Cook's D and DFBETAS as below:

```
dffits(model)
cooks.distance(model)
dfbetas(model)
```

In fact, we can use a sigla code `influence.measures(model)` to obtain all of the DFFITS, Cook's D and DFBETAS values.

- If the absolute value of DFFITS exceeds 1 for small data set and $2\sqrt{p/n}$ for large data set then it must be considered as an influential case.
- We usually consider points for which $D_i > 1$ to be influential.
- If $|DFBETAS| > 1$ for small to medium data sets and $|DFBETAS| > 2/\sqrt{n}$ for large data set.

Example

This data refer to the dataset `swiss` available in R. The following variables were collected (primarily from military records) for each of the 47 French- speaking provinces in Switzerland:
`Fertility:` (standardized)
`Agriculture` : Percent of males involved in agriculture as an Occupation
`Examination:` Percent of draftees who received the highest mark on their army examination
`Education:` Percent of draftees educated beyond primary school
`Catholic:` Percent Catholic (as opposed to Protestant)
`InfantMortality:` Percent of live births who live less than one year

```
> data(swiss)
> Fertility=swiss$Fertility
> Agriculture =swiss$Agriculture
> Exam=swiss$Examination
> Education=swiss$Education
> Catholic=swiss$Catholic
> Mortality=swiss$Infant.Mortality
> model=lm(Fertility~Agriculture+Exam+Education+Catholic+Mortality)
>influence.measures(model)
>extractAIC(model)
>extractAIC(model,k=log(n)) ## BIC
```

Example-Swiss Data

```
> library(leaps)
> subset=regsubsets(Fertility~ Agriculture+ Exam+ Education+ Catholic+
      + Mortality, data=swiss)
> summary(subset)
> plot(subset,scale="Cp")
> plot(subset,scale="r2")
> plot(subset,scale="adjr2")
> subset=regsubsets(Fertility~ Agriculture+ Exam+ Education+ Catholic+
      + Mortality, data=swiss, nbest=10)
> plot(subset)
```