Multiple Regression

Example : Hospital Manpower Data

A monthly manpower problem was studied. Previous study showed that the **monthly man-hours (Y)** has a high correlation with the **monthly occupied bed days**. However, there are more variables that are related to the monthly man-hours, for instance:

Y:	Monthly Man-hours
X_1:	Average Daily Patient Load
X_2:	Monthly X-ray Exposures
X_3:	Monthly Occupied Bed Days
X_4:	Eligible Population in the Area /1000
X_5:	Average Length of Patients' Stay in Days

The related data is listed in Table 1:

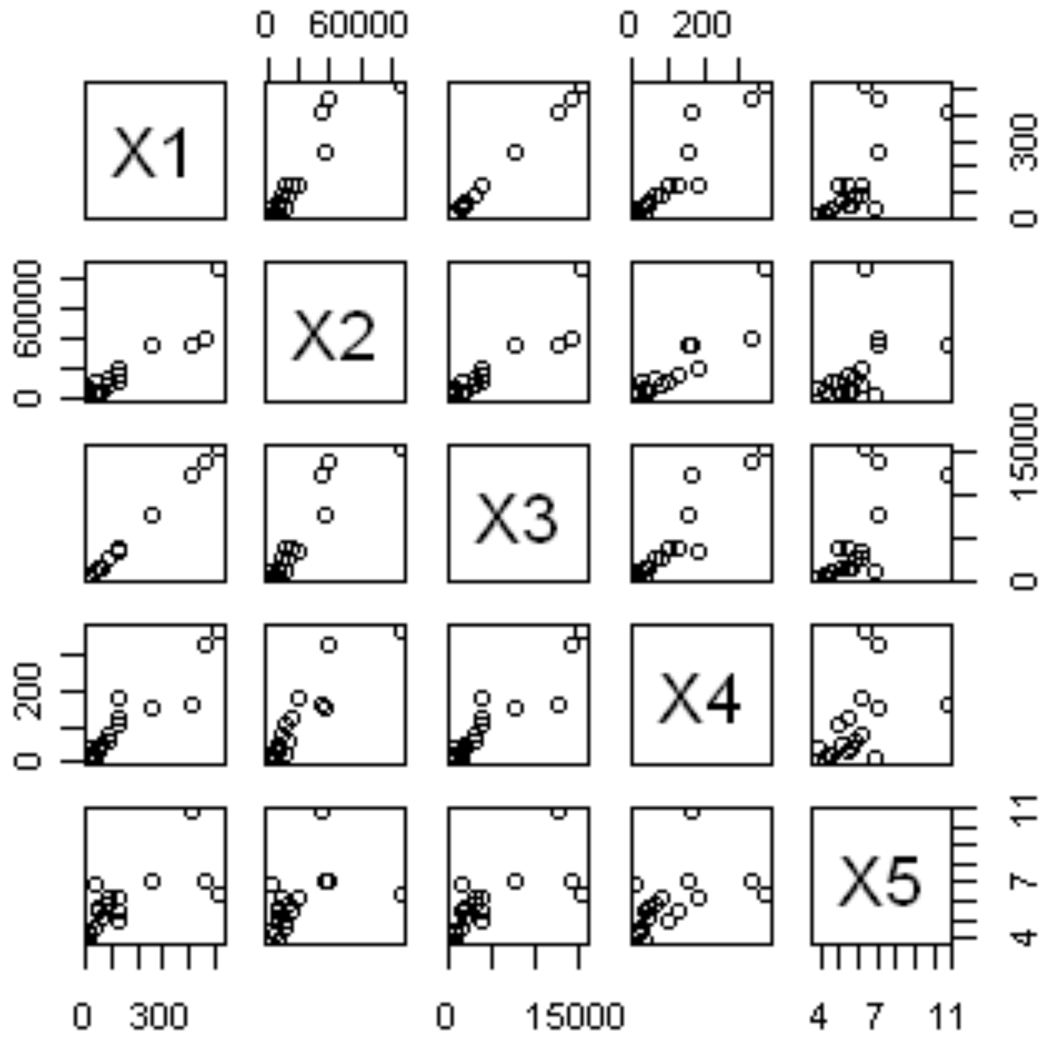
Example : Hospital Manpower Data

Obs	*X1*	*X2*	*X3*	*X4*	*X5*	*Y*
1	15.57	2463	472.92	18.0	4.45	566.52
2	44.02	2048	1339.75	9.5	6.92	696.82
3	20.42	3940	620.25	12.8	4.28	1033.15
4	18.74	6505	568.33	36.7	3.90	1603.62
5	49.20	5723	1497.60	35.7	5.50	1611.37
6	44.92	11520	1365.83	24.0	4.60	1613.27
7	55.48	5779	1687.00	43.3	5.62	1854.17
8	59.28	5969	1639.92	46.7	5.15	2160.55
9	94.39	8461	2872.33	78.7	6.18	2305.58
10	128.02	20106	3655.08	180.5	6.15	3503.93
11	96.00	13313	2912.00	60.9	5.88	3571.89
12	131.41	10771	3921.00	103.7	4.88	3741.40
13	127.21	15543	3865.67	126.8	5.50	4026.52
14	252.90	36194	7684.10	157.7	7.00	10343.81
15	409.20	34703	12446.33	167.4	10.78	11732.17
16	463.70	39204	14098.40	331.4	7.05	15414.94
17	510.22	86533	15524.00	371.6	6.35	18854.45

Download the data set manpower1.txt from ISpace➪

Scatter plot for multiple variable

plot(manpower[,-6])



Example : Hospital Manpower Data

If we choose all independent variables as the regressors, *the Model Equation is:*

$$Y = 1954.10 + -17.10x1 + 0.0559\,x2 + 1.627\,x3 - 4.07x4 - 392.58\,x5$$

Use R

manpower<-read.table(file.choose(),header=T)

lm(Y~.,data=manpower)

```
Call:
lm(formula = Y ~ ., data = manpower)

Coefficients:
(Intercept)          X1          X2          X3          X4          X5
 1954.09898   -17.09758     0.05589     1.62707    -4.06894  -392.58143
```

Example : Hospital Manpower Data

If we choose *x*2, *x*3 and *x*5 as the regressors, *the Model Equation is:*

$Y = 1523.39 + 0.0530\, x2 + 0.9785\, x3 - 320.951\, x5$

Use R

lm(Y~X2+X3+X5,data=manpower)

```
Call:
lm(formula = Y ~ X2 + X3 + X5, data = manpower)

Coefficients:
(Intercept)          X2           X3          X5
 1523.38924      0.05299      0.97848   -320.95083
```

Interpretation: ***β_j* is the net change in *Y* for each unit change in X_j holding all other values constant.**

```
lm.manpower<-lm(Y~X2+X3+X5,data=manpower)
reg1<-summary(lm.manpower)
```

Residuals:

Min	1Q	Median	3Q	Max
-687.40	-380.60	-25.03	281.91	1630.50

Test the Significance of each Independent Variable

$H0: b_j = 0 \qquad Ha: b_j \neq 0$

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1523.38924	786.89772	1.936	0.0749	.
X2	0.05299	0.02009	2.637	0.0205	*
X3	0.97848	0.10515	9.305	4. 12e-07	***
X5	-320.95083	153.19222	-2.095	0.0563	.

$$t = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}$$

Residual standard error: 614.8 on 13 degrees of freedom

Multiple R-Squared: 0.9901, Adjusted R-squared: 0.9878

F-statistic: 432 on 3 and 13 DF, p-value: 2.894e-13

reg1<-summary(lm.manpower)

```
Residuals:
    Min       1Q  Median      3Q     Max
-687.40 -380.60  -25.03  281.91 1630.50

Coefficients:
                Estimate    Std. Error   t value   Pr(>|t|)
(Intercept) 1523.38924  786.89772   1.936     0.0749 .
X2             0.05299      0.02009      2.637     0.0205 *
X3             0.97848      0.10515      9.305 4.  12e-07 ***
X5          -320.95083    153.19222    -2.095     0.0563 .
```

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-p}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-p}}$$

Residual standard error: 614.8 on 13 degrees of freedom

```
Multiple R-Squared: 0.9901,     Adjusted R-squared: 0.9878
F-statistic:   432 on 3 and 13 DF,  p-value: 2.894e-13
```

Unexplained or Random Variation

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{SS_{\text{Total}}} = \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{SS_{\text{Res}}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{SS_{\text{Reg}}}$$

Explained variation by regression model

reg1<-summary(lm.manpower)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1523.38924	786.89772	1.936	0.0749	.
X2	0.05299	0.02009	2.637	0.0205	*
X3	0.97848	0.10515	9.305	4. 12e-07	***
X5	-320.95083	153.19222	-2.095	0.0563	

Residual standard error: 614.8 on 13 degrees of freedom

$$R^2 = \frac{SS_{\text{Reg}}}{SS_{\text{Total}}}$$ the proportion of the total variation of the dependent variable that is explained by the multiple regression model

Multiple R-Squared: 0.9901,   Adjusted R-squared: 0.9878

F-statistic: 432 on 3 and 13 DF, p-value: 2.894e-13

reg1<-summary(lm.manpower)

```
Residuals:
    Min      1Q  Median      3Q     Max
-687.40 -380.60  -25.03  281.91 1630.50

Coefficients:
              Estimate   Std. Error  t value  Pr(>|t|)
(Intercept) 1523.38924  786.89772   1.936    0.0749 .
X2             0.05299    0.02009   2.637    0.0205 *
X3             0.97848    0.10515   9.305 4. 12e-07 ***
X5          -320.95083  153.19222  -2.095    0.0563 .

Residual standard error: 614.8 on 13 degrees of freedom
Multiple R-Squared: 0.9901,     Adjusted R-squared: 0.9878
```

The Overall F Test

$$F = \frac{SS_{Reg}/(p-1)}{SS_{Res}/(n-p)}$$

H_0: $\beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$ versus

H_a: At least one of $\beta_1, \beta_2, \ldots, \beta_{p-1} \neq 0$

F-statistic: 432 on 3 and 13 DF, p-value: 2.894e-13

Extracting information from lm results

Example : Hospital Manpower Data

```
reg1<-summary(lm.manpower)
names(reg1)
```

```
[1] "call"          "terms"         "residuals"     "coefficients"
[5] "aliased"       "sigma"         "df"            "r.squared"
[9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

```
reg1$sigma
```

```
[1] 614.7794
```

```
reg1$r.squared
```

```
[1] 0.9900682
```

```
reg1$coefficients
```

```
                Estimate   Std. Error   t value     Pr(>|t|)
(Intercept) 1523.38923568 786.89772473  1.935943 7.492387e-02
X2             0.05298733   0.02009194  2.637243 2.050318e-02
X3             0.97848162   0.10515362  9.305258 4.121293e-07
X5          -320.95082518 153.19222065 -2.095086 5.631250e-02
```

ANOVA table

Explained variation by regression model

Source	DF	SS	MS	F	p
Regression	$P\text{-}1$	SS_{Reg}	$MS_{\text{Reg}} = SS_{\text{Reg}}/p\text{-}1$	$F = \dfrac{MS_{\text{Reg}}}{MS_{\text{Res}}}$	
Error	$n-p$	SS_{Res}	$MS_{\text{Res}} = SS_{\text{Res}}/n\text{-}p$		
Total	$n-1$	SS_{Total}			

Unexplained or Random Variation

```
lm.manpower<-lm(Y~X2+X3+X5,data=manpower)
anova(lm.manpower)
```

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X2	1	441952483	441952483	1169.3296	4.041e-14 ***
X3	1	46187675	46187675	122.2046	5.556e-08 ***
X5	1	1658984	1658984	4.3894	0.05631 .
Residuals	13	4913399	377954		

aov1<-anova(lm.manpower)

```
Analysis of Variance Table

Response: Y
           Df    Sum Sq    Mean Sq   F value    Pr(>F)
X2          1 441952483  441952483 1169.3296  4.041e-14 ***
X3          1  46187675   46187675  122.2046  5.556e-08 ***
X5          1   1658984    1658984    4.3894    0.05631 .
Residuals 13   4913399     377954
```

SS.reg=sum(aov1[1:3,2])

F.ratio<-MS.reg/aov1[4,3]

MS.reg=SS.reg/3

ANOVA:

Source	DF	*SS*	*MS*	*F*	*p*
Model	3	489799142	163266381	431.9745	0.0001
Error	13	4913398.50	377953.731		
Total	16	494712540			

Comparison of two models

Example : Hospital Manpower Data

```
lm.manpower<-lm(Y~X2+X3+X5,data=manpower)
lm.manpower2<-lm(Y~X2+X3+X5+X2*X3,data=manpower)
anova(lm.manpower,lm.manpower2)
```

```
Analysis of Variance Table

Model 1: Y ~ X2 + X3 + X5
Model 2: Y ~ X2 + X3 + X5 + X2 * X3

  Res.Df    RSS      Df    Sum of Sq    F       Pr(>F)
1    13    4913399
2    12    4528502    1     384897    1.0199    0.3325
---
```

SSres

p-value

$$F_{df1,df2} = \frac{(RSS_1 - RSS_2)/(df_1 - df_2)}{RSS_2/df_2}$$

Textbook: Page 96-97

Comparison of two models

Example : Hospital Manpower Data

```
lm.manpower1<-lm(Y~X2+X5,data=manpower)
lm.manpower2<-lm(Y~.,data=manpower)
 anova(lm.manpower1,lm.manpower2)
```

```
Analysis of Variance Table

Model 1: Y ~ X2 + X5
Model 2: Y ~ X1 + X2 + X3 + X4 + X5

  Res.Df      RSS        Df     Sum of Sq        F              Pr(>F)
1     14    37639593
2     11     4543605     3       33095988    26.708      2.381e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Prediction

predict(lm.manpower)

predict(lm.manpower,newdata)

Example : Hospital Manpower Data *(cont…)*

Suppose that X2(Monthly X-ray Exposures) is 5000, X3(monthly Occupied Bed Days) is 1000 and X5 (Average Length of Patients' Stay in Days) is 5. The estimated monthly man-hours is given by

Use R:

```
>lm.manpower<-lm(Y~X,data=manpower)
>newdata<-data.frame(X2=5000,X3=1000,X5=5)
>predict(lm.manpower,newdata)
[1] 1162.053
```

Multiple Regression including Categorical or Indicator Variables

Example The Electronics World Case

Store	Number of Households x	Location	Sales Volume y
1	161	Street	157.27
2	99	Street	93.28
3	135	Street	136.81
4	120	Street	123.79
5	164	Street	153.51
6	221	Mall	241.74
7	179	Mall	201.54
8	204	Mall	206.71
9	214	Mall	229.78
10	101	Mall	135.22

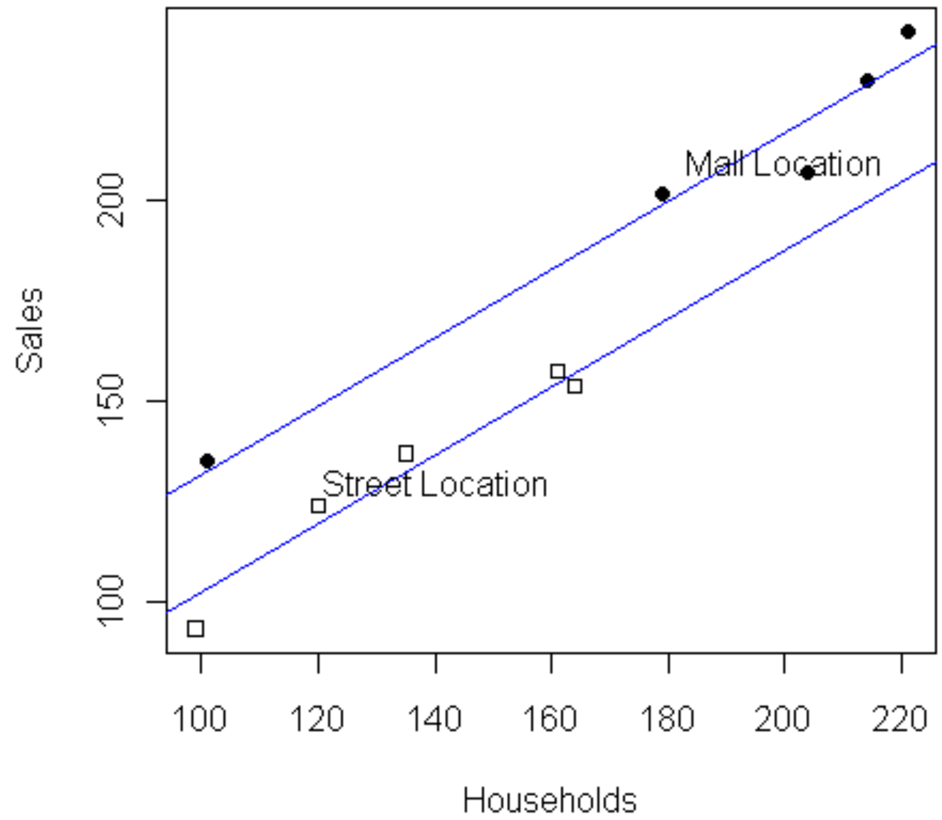
```
ele<-read.csv('d:\\Electronics1.csv',header=T)
lm1<-lm(Sales~Households+Location,data=ele)
```

```
Coefficients:
  (Intercept)      Households  LocationStreet
       46.576           0.851         -29.216
```

Example The Electronics World Case

```
plot(Sales~Households,pch=c(16,22)[Location],data=ele)
abline(lm1$coef[1],lm1$coef[2],col='blue')
abline(lm1$coef[1]+lm1$coef[3],lm1$coef[2],col='blue')
text(200,210,'Mall Location')
text(140,130,'Street Location')
```

For any given number of households, we estimate that the mean monthly sales volume in a mall location is $29,216 greater than the mean monthly sales volume in a street location.



Example Electronic World

Store	Number of Households, x	Location	Sales Volume, y
1	161	Street	157.27
2	99	Street	93.28
3	135	Street	136.81
4	120	Street	123.79
5	164	Street	153.51
6	221	Mall	241.74
7	179	Mall	201.54
8	204	Mall	206.71
9	214	Mall	229.78
10	101	Mall	135.22
11	231	Downtown	224.71
12	206	Downtown	195.29
13	248	Downtown	242.16
14	107	Downtown	115.21
15	205	Downtown	197.82

Example Electronic World

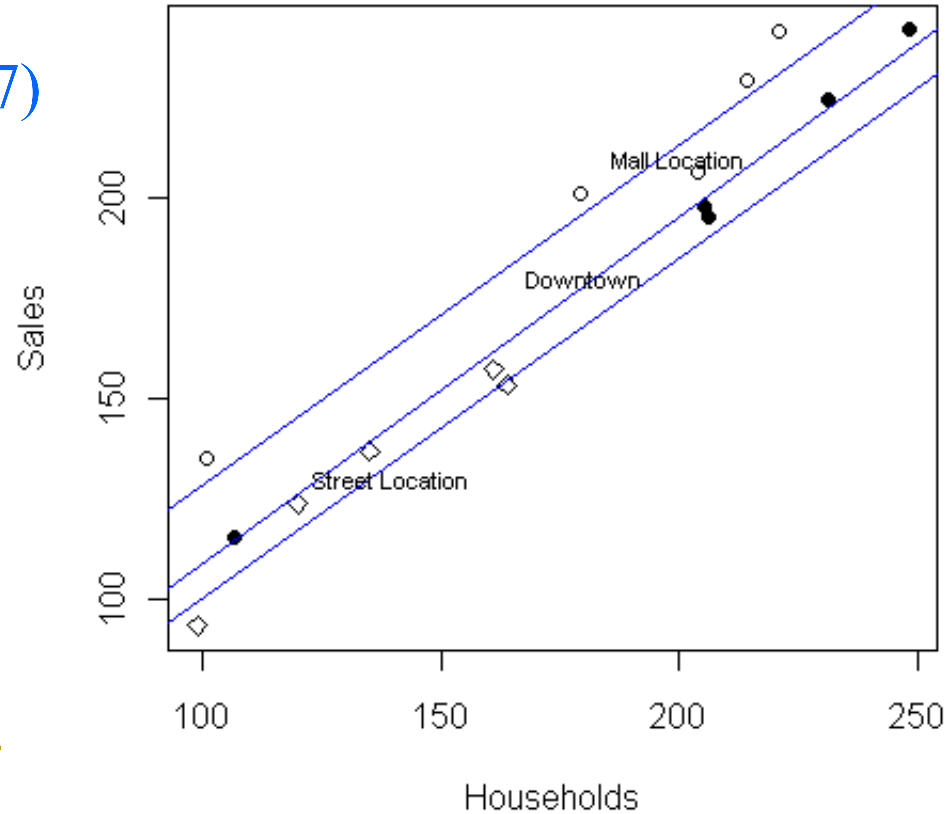
```
ele2<-read.csv('d:\\Electronics2.csv',header=T)
lm2<-lm(Sales~Households+Location,data=ele2)
```

```
Call:
lm(formula = Sales ~ Households + Location, data = ele2)

Coefficients:
  (Intercept)     Households    LocationMall  LocationStreet
      21.8415         0.8686         21.5100         -6.8638
```

Example Electronic World

```
plot(Sales~Households,pch=c(16,21,23)[Location],data=ele2)
abline(lm2$coef[1],lm2$coef[2],col='blue')
abline(lm2$coef[1]+lm2$coef[3],lm1$coef[2],col='blue')
abline(lm2$coef[1]+lm2$coef[4],lm1$coef[2],col='blue')
text(200,210,'Mall Location',cex=0.7)
text(180,180,'Downtown',cex=0.7)
text(140,130,'Street Location',cex=0.7)
```



Example : Hospital Manpower Data *(cont…)*

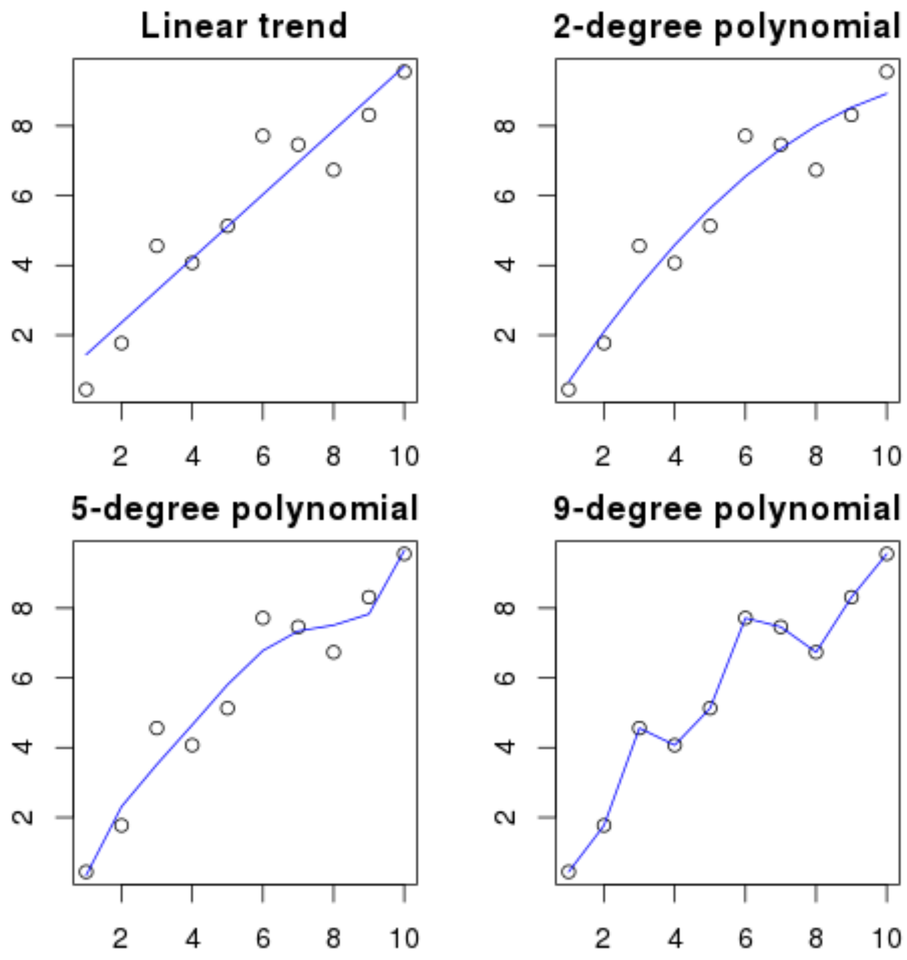
```
a=c(1,3,8)
manpower1=manpower[-a,]
lm.manpower<-lm(Y~X2+X3+X5,data=manpower1)
predict(lm.manpower,newdata=manpower[a,])
```

```
        1         3         8
 645.4575  922.1461 1754.6666
```

Cross Validation for model selection and Determination of Model Performance

overfitting

- The higher the polynomial order, the better it will fit the existing dots.
- However, the high order polynomials, despite looking like to be better models for the dots, are actually overfitting them. It models the noise rather than the true data distribution.



Cross Validation for model selection and Determination of Model Performance

1. Training sample & Validation sample (Randomly split the data into two groups:

$$\left.\begin{matrix} y_1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ y_m & x_{m1} & \cdots & x_{mk} \end{matrix}\right\} \text{Training sample (fitting sample)}$$

$$\left.\begin{matrix} y_{m+1} & x_{m+1,1} & \cdots & x_{m+1,k} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & \cdots & x_{nk} \end{matrix}\right\} \text{Validation Sample}$$

2. Based on the training sample, we get this regression model:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

3. Using the regression model obtained in step 2, make prediction for validation data

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{j1} + \cdots + \hat{\beta}_k x_{jk}, \text{ for } j = m+1, \ldots, n$$

Calculate MSE for validation data.

$$MSE = \frac{1}{(n-m)} \sum_{j=m+1}^{n} (y_j - \hat{y}_j)^2$$

4. Repeat step1-step 3 at least n times (n≥100) and then calculate average MSE of n times trails.

The mean of MSE is a good criteria to determine the best candidate model.

Example:

use cross-validation to evaluate the predictability of the following model:

```
lm.manpower<-lm(Y~X2+X3+X5,data=manpower1)
```

- Select a random sample of 12 observations.
- Based on the selected 12 observations, develop the above regression model.
- Use the regression model to predict the monthly man-hours of the remaining unselected 5 observations and calculate the corresponding sum of squared residuals

$$SSE = \sum_{i=1}^{n} (Y - \hat{Y})^2$$

- Repeat step1-step 3 at least n times (n≥100) and then calculate average MSE of n times trails.

```
manpower=read.table(file.choose(),header=T)
```

PRESS Statistics

We establish the model without the i^{th} observation $\{y_i, x_{i1}, \ldots, x_{ik}\}$, and predict y_i by this model, denote it as:

$$\boxed{\hat{y}_{i,-i}}$$

The criteria:

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

can be used for model / variable selection.

Computation of *PRESS*

$$PRESS = \sum_{i=1}^{n} \left[\frac{e_i}{1 - h_{ii}} \right]^2 .$$

In R:

```
lm1<-lm(Y~X2+X3+X5,data=manpower1)
reg1<-summary(lm1)
reg1$residuals
x<-model.matrix(lm.manpower)
hat(x)
```

$\# \textit{ Residuals} : e_i = y_i - \hat{y}_i$

$\# \textit{ Hat matrix} : \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

$h_{ii} = [diag(H)]_i$

```
Press=sum((reg1$residuals/(1-hat(x)))^2)
```

or

```
Press=sum((residuals(lm1)/(1 - lm.influence(lm1)$hat))^2)
```

Conceptual Predictive Criteria (The C_p *Statistics*)

$$C_p = p + \frac{(S_p^2 - s^2)(n-p)}{s^2}$$

σ^2 ***unknown;***

s^2 : $\hat{\sigma}^2$ ***based on the full model.***

In R:

```
lm1<-lm(Y~X2+X3+X5,data=manpower1)
lm2<-lm(Y~., data=manpower1)
reg1<-summary(lm1)
reg2<- summary(lm2)
Cp=4+(reg1$sigma^2-reg2$sigma^2)*(17-
4)/(reg2$sigma^2)
```

Stepwise regression procedure

- Goal is to develop a model with the best set of independent variables
  - Easier to interpret if unimportant variables are removed
  - Lower probability of collinearity

- Stepwise regression procedure

(Sequential Variable Selection Procedures)

- **Forward Selection**
- **Backward Elimination**
- **Stepwise Regression**

Stepwise regression:

Suppose that there are k independent variables $x_1, \ldots, x_k$ in the problem. We want to find the "best" model.

Forward Selection:

1. Start with no variables in the model.

2. For each independent variable, fit simple regression model including only one independent variable. Choose the simple regression model with the largest correlation in absolute value with the response y.

3. Next, add a variable that improves the model the most.

4. Repeat step 3 until none improves the model.

Forward Selection

Add to the model using the following criterion:

Adding the variable will increase adjust R^2 more than any other single variable.

Or:

Adding the variable will decrease AIC at most. (Akaike information criterion)

$$\mathrm{AIC} = 2k \; - \; 2\ln(L)$$

The number of estimated parameters in the model

likelihood function of the parameters in model

Remark: Other criterion, such as F-tests, t-tests, SS_{Res} are also possibly used in selection process.

The probability density is, under the model

$$\prod_{i=1}^{n} p(y_i|x_i;\beta_0,\beta_1,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-(\beta_0+\beta_1 x_i))^2}{2\sigma^2}}$$

of observing that data set (x_1,y_1), (x_2,y_2),…(x_n,y_n).

$$\prod_{i=1}^{n} p(y_i|x_i;b_0,b_1,s^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(y_i-(b_0+b_1 x_i))^2}{2s^2}}$$

Likelihood function of the parameters in model

$$\begin{aligned}
L(b_0, b_1, s^2) &= \log \prod_{i=1}^{n} p(y_i | x_i; b_0, b_1, s^2) \\
&= \sum_{i=1}^{n} \log p(y_i | x_i; b_0, b_1, s^2) \\
&= -\frac{n}{2} \log 2\pi - n \log s - \frac{1}{2s^2} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2
\end{aligned}$$

Forward Selection

- Stop entering variables into the model when there are no more variables that result in a significant increase in adjust R^2

- Stop entering variables into the model when there are no more variables that result in a decrease of AIC.

Backward Elimination

1. Fit the full model with all possible variables

2. Remove one variable that has the smallest contribution in the current model.

3. Repeat step 2 until none improves the model.

The stopping rule for the backward procedure is similar to that for the forward procedure

Stepwise Regression (Bidirectional)

The stepwise technique starts as in the forward procedure.

At each stage in the process, after a new variable is added, a test is made to check if some variables can be deleted without appreciably increasing the residual sum of squares (RSS).

Sequential Variable Selection Procedures

1. **Forward Selection**
2. **Backward Elimination**
3. **Stepwise Regression**

R:

step(object, scope, scale = 0, direction = c("both", "backward", "forward"), trace = 1, keep = NULL, steps = 1000, k = 2, ...)

Example

```
manpower=read.table(file.choose(),header=T)
reg1=lm(Y~.,data=manpower)
step(reg1,direction = c("backward"))
```

OR

```
library(MASS)
 step <- stepAIC(reg1, direction="both")
```