Multiple Linear Regression Model Building and Selection

Summary of simple linear regression

- Steps in building simple linear regression models
- Model summary
- Residual analysis
- Checking need for refinement
- Refined model building

In this lecture

- Multiple linear regression
  - Build linear model with one dependent and multiple independent variables
  - Look at summary of the model to discard insignificant variable
  - Model selection

Loading data

- Dataset 'nyc' is given in ".csv" format
- To load data from the file the function used is `read.csv( )`

read.csv()

Reads a file in table format and creates a data frame from it

SYNTAX

```
read.csv(file,row.names=1)
```

file	the name of the file which the data are to be read from. Each row of the table appears as one line of the file.
row.names	a vector of row names. This can be a vector giving the actual row names, or a single number giving the column of the table which contains the row names, or character string giving the name of the table column containing the row names.

Loading data

- Assuming that 'nyc.csv' is in your current working directory

```
nyc <- read.csv("nyc.csv")
```

- The data is saved into a data frame 'nyc'

Viewing data

- `View(nyc)` will display the dataframe in a tabular format

	Price	Food	Decor	Service	East
1	43	22	18	20	0
2	32	20	19	19	0
3	34	21	13	18	0
4	41	20	20	17	0

- `head(nyc)` and `tail(nyc)` will display the first and last six rows from the dataframe

Description of dataset

Menu pricing in restaurants of NYC

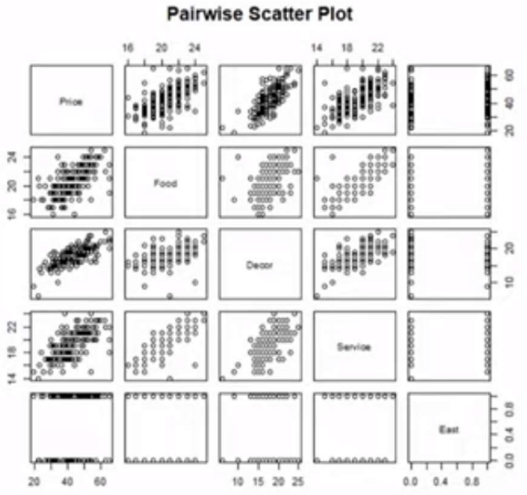
y : Price of dinner
x_1: Customer rating of the food (Food)
x_2: Customer rating of the décor (Décor)
x_3: Customer rating of the service (Service)
x_4: If the restaurant is east or west (East)

Objective: Build a linear model

Pairwise scatter plot

```
plot(nyc, main="Pairwise Scatter Plot")
```

```
> round(cor(nyc),3)
        Price  Food Decor Service  East
Price   1.000 0.627 0.724   0.641 0.187
Food    0.627 1.000 0.504   0.795 0.180
Decor   0.724 0.504 1.000   0.645 0.036
Service 0.641 0.795 0.645   1.000 0.209
East    0.187 0.180 0.036   0.209 1.000
```



Building multiple linear regression model

- Dependent variable (y) depends on p independent variables x_i, i=1,2..p

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p + \epsilon$$

- For i^{th} observation,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \cdots + \hat{\beta}_p x_{p,i} + \epsilon_i$$

Building multiple linear regression model

- Building multiple linear model using the function `lm()`
- Syntax: `lm(formula,data)`

```
lm(dependent var~indep.var1+ indep.var2)
```

```
nycmod_1 <- lm(Price~Food+Decor+Service+East,data = nyc)
```

or

```
nycmod_1<-lm(Price~.,data=nyc)
```

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p + \epsilon$$

Model summary

```
nycmod_1 <- lm(Price~Food+Decor+Service+East,data = nyc)
summary(nycmod_1)
```

```
Call:
lm(formula = Price ~ Food + Decor + Service + East, data = nyc)

Residuals:
     Min       1Q   Median       3Q      Max
-14.0465  -3.8837   0.0373   3.3942  17.7491

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.023800   4.708359  -5.102 9.24e-07 ***
Food          1.538120   0.368951   4.169 4.96e-05 ***
Decor         1.910087   0.217005   8.802 1.87e-15 ***
Service      -0.002727   0.396232  -0.007   0.9945
East          2.068050   0.946739   2.184   0.0304 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.738 on 163 degrees of freedom
Multiple R-squared:  0.6279,    Adjusted R-squared:  0.6187
F-statistic: 68.76 on 4 and 163 DF,  p-value: < 2.2e-16
```

$\hat{\beta}_i$ $i = 0, 1 \ldots 4$

New model dropping Service

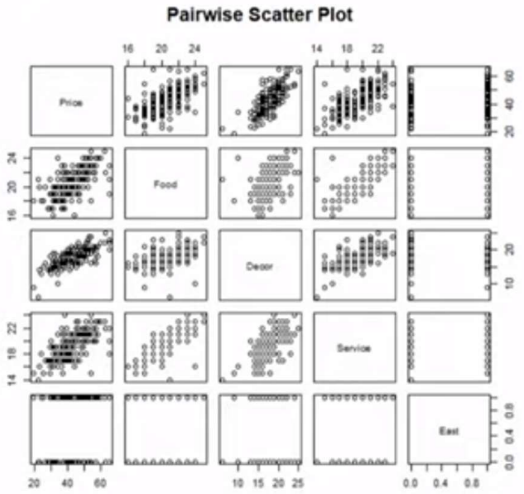
```
nycmod_2 <- lm(Price~Food+Decor+East,data = nyc)

summary(nycmod_2)
```

```
Call:
lm(formula = Price ~ Food + Decor + East, data = nyc)

Residuals:
     Min       1Q   Median       3Q      Max
-14.0451  -3.8809   0.0389   3.3918  17.7557

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.0269     4.6727  -5.142 7.67e-07 ***
Food          1.5363     0.2632   5.838 2.76e-08 ***
Decor         1.9094     0.1900  10.049  < 2e-16 ***
East          2.0670     0.9318   2.218   0.0279 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.72 on 164 degrees of freedom
Multiple R-squared:  0.6279,    Adjusted R-squared:  0.6211
F-statistic: 92.24 on 3 and 164 DF,  p-value: < 2.2e-16
```

Pairwise scatter plot

```
plot(nyc, main="Pairwise Scatter Plot")
```

```
> round(cor(nyc),3)
        Price  Food Decor Service  East
Price   1.000 0.627 0.724   0.641 0.187
Food    0.627 1.000 0.504   0.795 0.180
Decor   0.724 0.504 1.000   0.645 0.036
Service 0.641 0.795 0.645   1.000 0.209
East    0.187 0.180 0.036   0.209 1.000
```



New model dropping Food

```
nycmod_3 <- lm(Price~Service+Decor+East,data = nyc)

summary(nycmod_3)

Call:
lm(formula = Price ~ Service + Decor + East, data = nyc)

Residuals:
    Min      1Q  Median      3Q     Max
-14.514  -3.668  -0.769   3.704  18.778

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.5308     4.3220  -3.362 0.000963 ***
Service       1.1513     0.2973   3.872 0.000156 ***
Decor         1.8956     0.2276   8.331 3.05e-14 ***
East          2.1535     0.9927   2.169 0.031489 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.018 on 164 degrees of freedom
Multiple R-squared:  0.5882,    Adjusted R-squared:  0.5807
F-statistic: 78.09 on 3 and 164 DF,  p-value: < 2.2e-16
```