TAMS65 - Lecture 10: Linear Regression analysis

Zhenxia Liu

Matematisk statistik
Matematiska institutionen

LINKÖPINGS UNIVERSITET

Contents

Simple linear regression - Introduction
What is the simple linear regression about?

For example, in one study, people want to know the relation between **damage costs** and **distance** to the nearest fire station in case of fires.

Then we can define two populations X and Y in the following way.

$X = \{$the distance from fire station (in miles)$\}$

$Y = \{$the damage cost (in thousands of dollars)$\}$

Note: X and Y are Not independent. Note: check by their covariance.

We assume that X and Y are linear. How to check the linear relation? I: $|\rho(X, Y)| \approx 1$. II: plot n observations.

Simple linear regression - Introduction

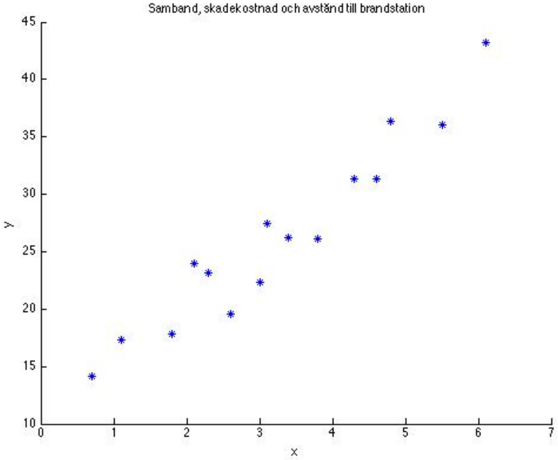
Take a sample,choose n fire stations:

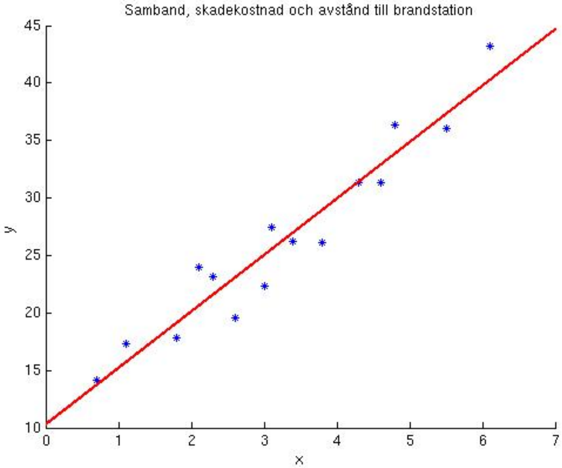
$\{x_1, x_2, \ldots, x_n\}$ $\{y_1, y_2, \ldots, y_n\}$

Paired data: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.

For example, we have 15 observations (fire stations) as following:

$$\begin{aligned}(x_i, y_i) :&(3.4, 26.2), (1.8, 17.8), (4.6, 31.3), (2.3, 23.1), (3.1, 27.5),\\ &(5.5, 36.0), (0.7, 14.1), (3.0, 22.3), (2.6, 19.6), (4.3, 31.3),\\ &(2.1, 24.0), (1.1, 17.3), (6.1, 43.2), (4.8, 36.4), (3.8, 26.1).\end{aligned}$$


Samband, skadekostnad och avstånd till brandstation
45
40
35
30
25
20
15
10
y
0
1
2
3
4
5
6
7
x



We want to study the linear relation between X and Y.

Simple linear regression - Introduction

We want to study the linear relation between X and Y.

Goal: Find $y = \beta_0 + \beta_1 x$ which fits the points best.

Linear regression will give us an answer.

Simple linear regression - Model

Simple linear regression

Model:

$$Y = \beta_0 + \beta_1 x + \varepsilon, \text{ where } \varepsilon \sim N(0, \sigma)$$

This model is called **simple linear regression** (**Enkel linjär regression**).

- β_0, β_1 are unknown parameters
- x: variable for observations (Not a r.v.)
- ε is an error (a r.v.)
- Y is a r.v.
- There are several equivalent forms of the **Model**.

Simple linear regression - Model

Equivalent forms of the **Model**

- **Model:** $Y = \beta_0 + \beta_1 x + \varepsilon,$ where $\varepsilon \sim N(0, \sigma)$
  - $\mu = E(Y) = \beta_0 + \beta_1 x$, $\hat{\beta}_0, \hat{\beta}_1 = ?$
  - $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x$ or $y = \hat{\beta}_0 + \hat{\beta}_1 x$ is called **estimated regression line** (**skattad regressionslinje**).
  - x is called **explanatory variable** (**förklaringsvariable**).
- **Model:** $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma), i = 1, \ldots, n.$
  - $\varepsilon_1, \ldots, \varepsilon_n$ are independent.
  - $\mu_i = E(Y_i) = \beta_0 + \beta_1 x_i$, $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$,
  - $y_i - \hat{\mu}_i = e_i =$ residual, error

Simple linear regression - Model

Equivalent forms of **Model**

- **Model:**

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{=\boldsymbol{Y}} = \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}}_{=\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{=\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{=\boldsymbol{\varepsilon}}$$

- **Model:** $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N(\vec{0}, \sigma^2 \mathbf{I}_{n\times n})$.
  - $\mathbf{X}$ is a constant matrix (observations).
  - $\boldsymbol{\beta}$ is unknown parameter vector.
  - $\boldsymbol{Y}$ is a normal vector, and $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_{n\times n})$.

Multiple linear regression - Introduction

Multiple linear regression

Populations: $X_1, X_2, \ldots, X_k$ e.g. distance, water, ...

Population: Y e.g. damage cost.
Not independent

Sample: choose n fire stations, after measure/observe, we have
$((x_{11}, x_{12}, \ldots, x_{1k}), y_1)$
$((x_{21}, x_{22}, \ldots, x_{2k}), y_2)$
$\vdots$
$((x_{n1}, x_{n2}, \ldots, x_{nk}), y_n)$

Goal: Find $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ which fits the points best.

Multiple linear regression - Model

Multiple linear regression

Model:

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon, \text{ where } \varepsilon \sim N(0, \sigma)$$

This model is called **Multiple linear regression (multipel linjär regression.)**

- $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters.
- $x_1, x_2, \ldots, x_k$ are variables for observations (Not r.v.).
- ε is an error (a r.v.)
- Y is a r.v.
- There are several equivalent forms of the **Model**.

Multiple linear regression - Model

Equivalent forms of the **Model**

- **Model:** $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$, where $\varepsilon \sim N(0, \sigma)$.
  - $\mu = E(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k = ?$
  - $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$ or $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$ is called **estimated regression line**.
  - $x_1, x_2, \ldots, x_k$ are called **explanatory variables** (**förklaringsvariabler**).

- **Model:** $Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma)$ for $i = 1, \ldots, n$.
  - $\varepsilon_1, \ldots, \varepsilon_n$ are independent.
  - $\mu_i = E(Y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$,
  - $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}$,
  - $y_i - \hat{\mu}_i = e_i =$ residual, error

Multiple linear regression - Model

- **Model:**

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{=\boldsymbol{Y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}}_{=\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{=\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{=\boldsymbol{\varepsilon}}$$

- **Model:**

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ where } \boldsymbol{\varepsilon} \sim N(\vec{0}, \sigma^2 \mathbf{I}_{n \times n})$$

  - $\mathbf{X}$ is a constant matrix (observations).
  - $\boldsymbol{\beta}$ is unknown parameter vector.
  - $\boldsymbol{Y}$ is a normal vector, and $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_{n \times n})$. **Exercise.**

5 Questions

We will analyze the model by discussing the following 5 questions:

- Q_1 : Find the estimated regression line,

$$\hat{\mu} = y = \hat{\beta}_0 + \hat{\beta}_1 x + \ldots + \hat{\beta}_k x_k? \text{ That is, } \quad \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} =?$$

- Q_2 : How well does the model fit observations / data (the estimated regression line)?
- Q_3 : $\hat{\sigma^2} =?$

5 Questions

- Q_4 : Does y depend on $\{x_1, x_2, \ldots, x_k\}$? i.e. at least one variable (say x_j) is useful? i.e. at least one $\beta_j \neq 0$?

- Q_5 : Does y depend on a specific variable, say x_j? i.e. Is x_j useful? i.e. $\beta_j \neq 0$? where $j = 1, 2, \ldots, k$.

5 Questions - Q_1

$Q_1: \quad Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \Rightarrow$

$\mu = E(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

For the $i-$th observation, we have

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}, i = 1, 2, \ldots, n.$$

Replace μ_i by y_i, then $y_i \approx \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$

That is,

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{=\boldsymbol{y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}}_{=\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{=\boldsymbol{\beta}}$$

Therefore we get

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}$$

5 Questions - Q_1

Q_1 : If $det\left(\mathbf{X}'\mathbf{X}\right) \neq 0$, then we get the point estimate

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\boldsymbol{y}.$$

Note:

- This method is called least square method (LSM.)
- The point estimator

$$\hat{\boldsymbol{B}} = \begin{pmatrix} \hat{B}_0 \\ \hat{B}_1 \\ \vdots \\ \hat{B}_k \end{pmatrix} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\boldsymbol{Y} \sim N\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right).$$

5 Questions - Q_1

- The point estimator $\hat{B}_j \sim N(\beta_j, \sigma\sqrt{h_{jj}})$, where we assume that

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} h_{00} & h_{01} & \dots & h_{0k} \\ h_{10} & h_{11} & \dots & h_{1k} \\ \vdots & & & \\ h_{k0} & h_{k1} & \dots & h_{kk} \end{pmatrix}$$

  Exercise: Prove that $\hat{B}_j \sim N(\beta_j, \sigma\sqrt{h_{jj}})$.
  Hint: $A = \begin{pmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{pmatrix}$

- The point estimator

$$\hat{B}_j - \hat{B}_i \sim N\left(\beta_j - \beta_i, \sigma\sqrt{A(\mathbf{X}'\mathbf{X})^{-1}A'}\right),$$

  Where $A = \begin{pmatrix} 0 & \dots & 1 & 0 & \dots & -1 & 0 & \dots & 0 \end{pmatrix}$

5 Questions - Q_2

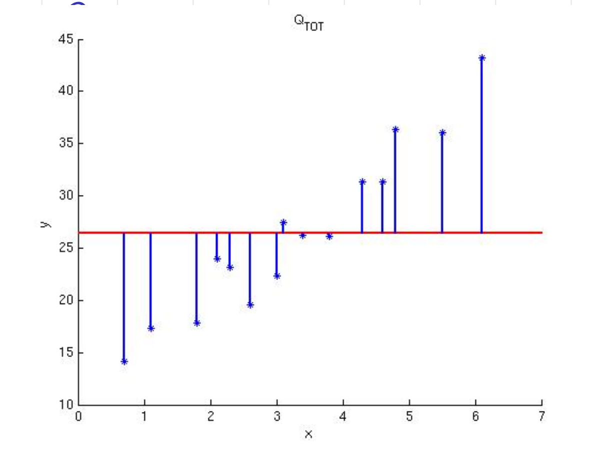
Q_2 : Note: We have $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik}$. Then we can divide the sum of squares of total in the following way:

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{=SS_{TOT}} = \underbrace{\sum_{i=1}^{n}(\hat{\mu}_i - \overline{y})^2}_{=SS_R} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}_{=SS_E}.$$

- SS_{TOT} is the sum of squares of Total.
- SS_R is the sum of squares of regression.
- SS_E is the sum of squares of residuals / errors.

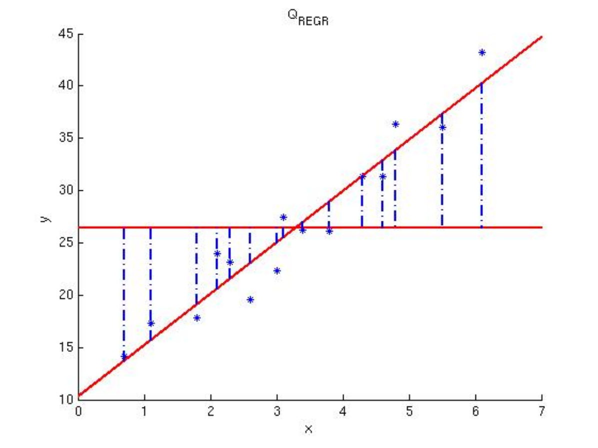
Note: $SS_{TOT} \Leftrightarrow Q_{TOT}$, $SS_R \Leftrightarrow Q_{REGR}$ and $SS_E \Leftrightarrow Q_{RES}$.

5 Questior



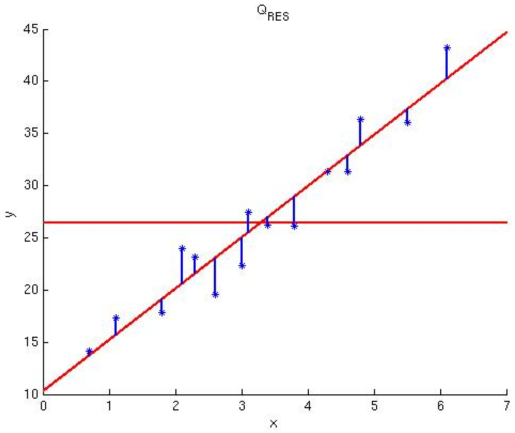
$$SS_{TOT} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

5 Questior



$$SS_R = \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2$$

5 questions - Q_2



$$SS_E = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$$

5 Questions - Q_2

Q_2 :

Coefficient of Determination (förklaringsgraden)

$$R^2 = \frac{SS_R}{SS_{TOT}} = \frac{SS_R}{SS_R + SS_E}$$

- $0 \leq R^2 \leq 1$
- $R^2 \approx 1$ implies that the Model (the estimated regression line) fits the observations / data well.
- Generally, the higher R^2, the better Model fits the data.

5 Questions - Theorem

Theorem

Under the conditions of the above model it holds that

- $$\frac{SS_E}{\sigma^2} \sim \chi^2(n-k-1).$$
- SS_E , SS_R *and* $\hat{\boldsymbol{B}}$ *are independent.*
- *If* $\beta_1 = \beta_2 = ... = \beta_k = 0$ *then*

$$\frac{SS_R}{\sigma^2} \sim \chi^2(k).$$

- $\hat{\sigma}^2 = s^2 = \frac{SS_E}{n-k-1}$.

5 Questions - Q_3

Q_3 :

$$\sigma^2 \approx \hat{\sigma}^2 = s^2 = \frac{SS_E}{n-k-1},$$

Where n is the number of observations, and k is the number of explanatory variables.

- $n-k-1$ is the degrees of freedom of SS_E

Q_4 : Hypothesis

$$H_0:\ \beta_1 = \ldots = \beta_k = 0 \quad \text{against} \quad H_1:\ \text{at least one } \beta_i \neq 0.$$

The Sampling distribution

$$\frac{SS_R/k}{SS_E/(n-k-1)} \sim F(k, n-k-1)$$

Test statistics is

$$TS = \frac{SS_R/k}{SS_E/(n-k-1)} \text{ and } C = (F_\alpha(k, n-k-1), \infty)$$

If $TS \in C$, then reject H_0. That is, at least one variable is useful.

- k is degrees of freedom of SS_R.

5 Questions - Q_5

Q_5 :

Note that: The point estimator $\hat{B}_j \sim N(\beta_j, \sigma\sqrt{h_{jj}}), j = 1, 2, \ldots, k$ where we assume that

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} h_{00} & h_{01} & \ldots & h_{0k} \\ h_{10} & h_{11} & \ldots & h_{1k} \\ \vdots & & & \\ h_{k0} & h_{k1} & \ldots & h_{kk} \end{pmatrix}$$

Hence, we get the sampling distribution

$$\frac{\hat{B}_j - \beta_j}{\sigma\sqrt{h_{jj}}} \sim N(0,1) \iff \frac{\hat{B}_j - \beta_j}{S\sqrt{h_{jj}}} \sim t(n-k-1)$$

5 Questions - Q_5

Method 1: Make a test on β_j with level α : $\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$

$$TS = \frac{\hat{\beta}_j - 0}{s\sqrt{h_{jj}}} \text{ or } \frac{\hat{\beta}_j - 0}{d(\hat{\beta}_j)}$$

$$C = (-\infty, -t_{\alpha/2}(n-k-1)) \cap (t_{\alpha/2}(n-k-1), \infty)$$

If $TS \in C$, reject H_0, i.e. y depends on x_j.

Method 2: Get a confidence interval for β_j with confidence $1-\alpha$

$$I_{\beta_j} = \hat{\beta}_j \mp t_{\alpha/2}(n-k-1)s\sqrt{h_{jj}}$$

If $0 \notin I_{\beta_j}$, then $\beta_j \neq 0$, i.e. y depends on x_j.

Method III: In MATLAB, $p-$value $< \alpha \iff$ reject H_0.

Example

A company has measured three performance variables x_1, x_2 and x_3 for its sellers. The values of these have been standardized such that 100 represents an average performance for a person in the industry. Furthermore, they have undergone a test where they measured creativity (x_4), ability to reason mechanically"(x_5) and abstract (x_6) as well as mathematical ability (x_7).

Nr	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	88.8	91.8	87.6	1	10	10	16
2	99.0	101.3	103.0	5	12	9	23
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
49	114.3	109.5	117.1	18	12	12	45
50	116.0	118.5	112.5	18	16	11	50

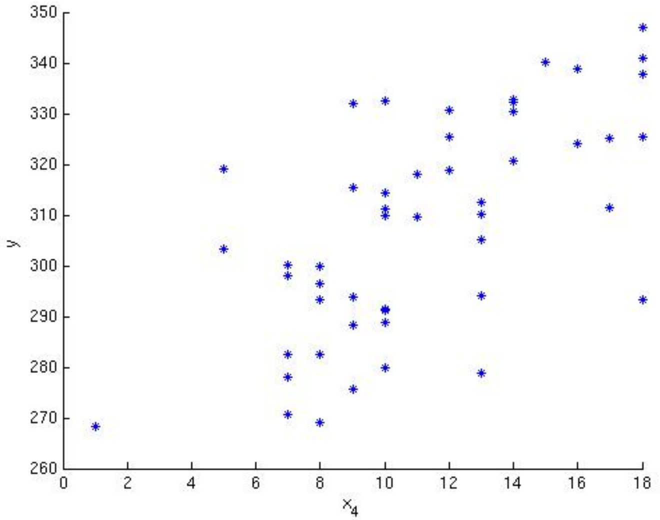
Let's use $y = x_1 + x_2 + x_3$ as the overall performance metric. When recruiting staff, it is interesting to predict the Y value using x_4 and x_7.

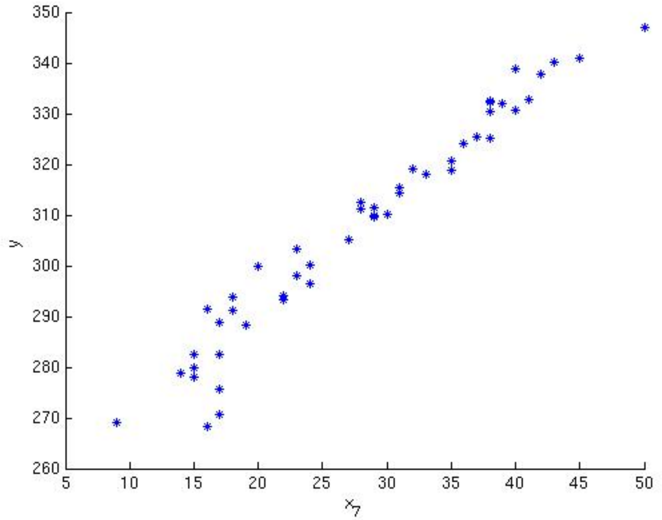
We plot y against x_4 and x_7.

```
y = [268.2 303.3 ... 340.9 347.0]';
x4 = [1 5 ... 18 18]';
x7 = [16 23 ... 45 50]';

figure;
scatter(x4,y,'*');
xlabel('x_4'); ylabel('y')

figure;
scatter(x7,y,'*');
xlabel('x_7'); ylabel('y')
```


350
340
330
320
310
300
290
280
270
260
y
0
2
4
6
8
10
12
14
16
18
x4


350
340
330
320
310
300
290
280
270
260
y
5
10
15
20
25
30
35
40
45
50
x7

We also can check the linear relation by correlation.

```
>> korrelation = corr([x4 x7],y)

korrelation =

    0.6393
    0.9771
```

Both plots and correlations show tendencies to linear relation, so we can analyze the data according to the Model below

$$Y = \beta_0 + \beta_4 x_4 + \beta_7 x_7 + \varepsilon$$

where $\varepsilon \sim N(0, \sigma)$.

Example

Answer the followings:

(a) Write out the estimated regression line according to the model.

(b) How does the model fit the data

(c) Estimate σ.

(d) Make a test on $\begin{cases} H_0 : \beta_4 = \beta_7 = 0 \\ H_1 : \text{ at least one } \beta_j \neq 0, j = 4, 7. \end{cases}$
with a significance level $\alpha = 5\%$.

(e) Is the variable x_7 useful/important for the model with a significance level $\alpha = 5\%$?

Direct output from MATLAB

.

```
>> regr = regstats(y,[x4 x7],'linear','all')
```

```
regr =
            ...
        source: 'regstats'
          beta: [3x1 double]
          covb: [3x3 double]
          yhat: [50x1 double]
             r: [50x1 double]
           mse: 21.5633
       rsquare: 0.9551
            ...
         tstat: [1x1 struct]
         fstat: [1x1 struct]
            ...
```

```
>> s2 = regr.mse

s2 =

   21.5633
```

```
>> betahat = regr.beta

betahat =

  248.6924
    0.1169
    2.0603
```

```
>> regr.tstat

ans =

    beta: [3x1 double]
      se: [3x1 double]
       t: [3x1 double]
    pval: [3x1 double]
     dfe: 47

>> regr.tstat.se

ans =

    2.1654
    0.2189
    0.0862
```

```
>> regr.tstat.beta

ans =

  248.6924
    0.1169
    2.0603
```

```
>> regr.tstat.t

ans =

  114.8471
    0.5342
   23.9132

>> regr.tstat.pval

ans =

    0.0000
    0.5957
    0.0000
```

```
>> format long
>> Cbetahat = regr.covb

Cbetahat =

   4.689050677582475  -0.198384946199334  -0.072464543761939
  -0.198384946199334   0.047903470928113  -0.012093152553998
  -0.072464543761939  -0.012093152553998   0.007423313673958

>> XtXinv = Cbetahat/s2

XtXinv =

   0.217455265812369  -0.009200124753437  -0.003360551571997
  -0.009200124753437   0.002221528987479  -0.000560821344011
  -0.003360551571997  -0.000560821344011   0.000344257027525
```

```
>> fstat = regr.fstat

fstat =

     sse: 1.0135e+03
     dfe: 47
     dfr: 2
     ssr: 2.1555e+04
       f: 499.8054
    pval: 0
```

Summarized information of above output from MATLAB is also given in the following way.

Estimated regression line: $y = 248.69 + 0.12x_4 + 2.06x_7$

i	$\widehat{\beta}_i$	$d(\widehat{\beta}_i)$
0	248.6924	2.1654
4	0.1169	0.2189
7	2.0603	0.0862

Variance analysis:

	Degrees of freedom	sum of squares
REGR	2	21555
RES	47	1013.5
TOT	49	22568.5

Note: Degrees of freedom (frihetsgrader), sum of squares (kvadratsumma).

Example

(a) The estimated regression line is
$y = \hat{\beta}_0 + \hat{\beta}_4 x_4 + \hat{\beta}_7 x_7 = 248.6924 + 0.1169 x_4 + 2.0603 x_7$.

(b) $R^2 = \frac{SS_R}{SS_{TOT}} = 0.9551 \approx 1$ which means the model fits the data very well.

(c) $\hat{\sigma} = \sqrt{s^2} = \sqrt{\frac{SS_E}{n-k-1}} = \sqrt{1013.5/47} \approx 4.6437$.

(d)

$$TS = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{21555/2}{1013.5/47} = 499.8$$

$C = (F_{0.05}(2,47), \infty) = (3.18, \infty)$, we reject H_0 since $TS \in C$. i.e. at least one of x_4 and x_7 is useful/important to the model.

Example

(e) The sampling distribution for β_7 is

$$\frac{\hat{B}_7 - \beta_7}{s\sqrt{h_{77}}} \sim t(n-k-1).$$

Method 1: $H_0 : \beta_7 = 0$ against $H_1 : \beta_7 \neq 0$

$$TS = \frac{\hat{\beta}_7 - 0}{d(\hat{\beta}_7)} = \frac{\hat{\beta}_7}{s\sqrt{h_{77}}} = \frac{2.0603}{0.0862} = 23.91$$

$C = (-\infty, -t_{0.025}(47)) \cup (t_{0.025}(47), \infty) =$
$(-\infty, -2.013) \cup (2.013, \infty)$. $TS \in C$, reject H_0. i.e. $\beta_7 \neq 0$, which means the variable x_7 is useful/important for the model.

Example

(e) Method 2: Confidence interval

$$
\begin{aligned}
I_{\beta_7} &= \hat{\beta}_7 \mp t_{0.025}(47) d(\hat{\beta}_7) \\
&= \hat{\beta}_7 \mp t_{0.025}(47) s\sqrt{h_{77}} = (1.887\ ,\ 2.234)
\end{aligned}
$$

We can see that $0 \notin I_{\beta_7}$, i.e. $\beta_7 \neq 0$, which means the variable x_7 is useful/important for the model.

Method 3: $p - Value = 0 < \alpha, \Rightarrow$ reject H_0. i.e. the variable x_7 is useful/important for the model.

Practice after the lecture:

Exercises:

(I) 14.4a-d, 14.7, PS-26, PS-27, PS-25.

(II) 14.2, PS-28, PS-24.

Thank you!

http://courses.mai.liu.se/GU/TAMS65/

LINKÖPINGS UNIVERSITET