TAMS65 - Lecture 11:
Linear Regression - continued

Zhenxia Liu
Matematisk statistik
Matematiska institutionen

LINKÖPINGS UNIVERSITET

Content

- Repetition
- Confidence interval for $\mu_0 = E(Y_0)$
- Prediction interval for Y_0
- Example 1 - continued
- Example 2

Repetition

We want to study **linear relation** among variables

$$y \text{ and } \{x_1, x_2, \ldots, x_k\}$$

Take a sample (n observations):
$((x_{i1}, x_{i2}, \ldots, x_{ik}), y_i)$, $i = 1, 2, \ldots, n$.

Pre-judgment on **linear relation**:

I: correlation (y, x_j) II plot (y, x_j) $j = 1, 2, \ldots, k$.

Multiple/simple linear regression:

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

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

Repetition

5 Questions:

- Q_1 : The estimated regression line

$$y = \hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k,$$

- $$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 & \hat{\beta}_1 & \ldots & \hat{\beta}_k \end{pmatrix}' = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{y},$$

  where

  - $\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}$
  - $\boldsymbol{y} = \begin{pmatrix} y_1 & y_2 & \ldots & y_n \end{pmatrix}'$

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

Repetition

- Q_2 : How well does the model (estimated regression line) fit observations / data ?
  - $$R^2 = \frac{SS_R}{SS_{TOT}} = \frac{SS_R}{SS_R + SS_E}$$
  - It fits well if $R^2 \approx 1$.
- $Q_3 : \sigma^2 \approx \hat{\sigma}^2 = s^2 = \frac{SS_E}{n-k-1}$.

Repetition

- Q_4 : Does y depend on $\{x_1, x_2, \ldots, x_k\}$? i.e. at least one variable is useful?
  - Test on $H_0: \ \beta_1 = ... = \beta_k = 0$.
  - The sampling distribution $\frac{SS_R/k}{SS_E/(n-k-1)} \sim F(k, n-k-1)$
- Q_5 : Does y depend on a specific variable, say x_j? i.e. Is x_j useful?
  - Check whether $\beta_j = 0$ or not.
  - The sampling distribution

$$\frac{\hat{B}_j - \beta_j}{S\sqrt{h_{jj}}} = \frac{\hat{B}_j - \beta_j}{d(\hat{\beta}_j)} \sim t(n-k-1)$$

Example 1 - Continued Example (Lecture 10)

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.

In lecture 10, we have analyzed the data according to the model

$Y = \beta_0 + \beta_4 x_4 + \beta_7 x_7 + \varepsilon$, where $\varepsilon \sim N(0, \sigma)$.

- $$\hat{\boldsymbol{\beta}} = \begin{pmatrix}\hat{\beta}_0 & \hat{\beta}_4 & \hat{\beta}_7\end{pmatrix}' = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{y} = \begin{pmatrix}248.6924 & 0.1169 & 2.0603\end{pmatrix}$$
- $(\mathbf{X}'\mathbf{X})^{-1}$

A new problem: Suppose we want to estimate or predict the overall performance metric (i.e. y) on new employees who have creativity $x_4 = 11$ and mathematical ability $x_7 = 30$.

What can we say about the y value for such employees? Note: We do **NOT** have observed value for such y.

Substitute x_4 and x_7 values to the model

$$Y_0 = \beta_0 + 11\beta_4 + 30\beta_7 + \varepsilon_0 = \begin{pmatrix} 1 & 11 & 30 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_4 \\ \beta_7 \end{pmatrix} + \varepsilon_0, \varepsilon_0 \sim N(0, \sigma).$$

• Y_0 is the overall performance metric of **a new employee (individual)** who has creativity $x_4 = 11$ and mathematical ability $x_7 = 30$.

We also can get

$$\mu_0 = E(Y_0) = \beta_0 + 11\beta_4 + 30\beta_7$$

• $\mu_0 = E(Y_0)$ is the average of the overall performance metric of all new employees who have creativity $x_4 = 11$ and mathematical ability $x_7 = 30$.

To study the overall performance metric of these **new employees** who have creativity $x_4 = 11$ and mathematical ability $x_7 = 30$, we consider the followings:

- $(1-\alpha)$ confidence interval (C.I.) for $\mu_0 = E(Y_0)$: I_{μ_0}.
- $(1-\alpha)$ **prediction interval (prediktionsintervall)** (P.I.) for Y_0: I_{Y_0}.

Note: $I_{\mu_0} \subseteq I_{Y_0}$

Confidence interval for $\mu_0 = E(Y_0)$

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

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

At the beginning, we have applied n observations, and got the followings: $\Rightarrow$ The point estimate $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$

$\Rightarrow$ The point estimator $\hat{\boldsymbol{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y} \sim N\left(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right)$

Now we have a **new observation**: $x_1 = u_1, x_2 = u_2, \ldots, x_k = u_k$

We substitute the **new observation** to the model.

$Y_0 = \beta_0 + \beta_1 u_1 + \ldots + \beta_k u_k + \varepsilon_0$, where $\varepsilon_0 \sim N(0, \sigma)$

We also get $\mu_0 = E(Y_0) = \beta_0 + \beta_1 u_1 + \ldots + \beta_k u_k$

Confidence interval for $\mu_0 = E(Y_0)$

We rewrite the model and the mean

$$Y_0 = (1 \ u_1 \ \ldots \ u_k) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \varepsilon_0 = \boldsymbol{u}'\boldsymbol{\beta} + \varepsilon_0$$

and $\mu_0 = E(Y_0) = (1 \ u_1 \ \ldots \ u_k) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = \boldsymbol{u}'\boldsymbol{\beta}$.

We **denote** the new observation by $\boldsymbol{u} = \begin{pmatrix} 1 \\ u_1 \\ \vdots \\ u_k \end{pmatrix}$

Confidence interval for $\mu_0 = E(Y_0)$

For new observation $\boldsymbol{u} = (1 \; u_1 \; \ldots \; u_k)'$, we have $\mu_0 = E(Y_0) = \boldsymbol{u}'\boldsymbol{\beta}$.

$\Downarrow$

The point estimate of μ_0 is $\hat{\mu}_0 = \boldsymbol{u}'\widehat{\boldsymbol{\beta}}$

$\Downarrow$

The point estimator of μ_0 is $\hat{M}_0 = \boldsymbol{u}'\widehat{\boldsymbol{B}}$.

$\Downarrow$

$\mathrm{E}(\hat{M}_0) = \mathrm{E}\left(\boldsymbol{u}'\widehat{\boldsymbol{B}}\right) = \boldsymbol{u}'\,\mathrm{E}\left(\widehat{\boldsymbol{B}}\right) = \boldsymbol{u}'\boldsymbol{\beta} = \mu_0$

$\Downarrow$

$V(\hat{M}_0) = V\left(\boldsymbol{u}'\widehat{\boldsymbol{B}}\right) = \boldsymbol{u}'\mathbf{C}_{\widehat{\boldsymbol{B}}}\boldsymbol{u} = \boldsymbol{u}'\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{u}.$

$\Downarrow$

Moreover, $\hat{M}_0$ is normally distributed. Why?

$\Downarrow$

$$\hat{M}_0 = \boldsymbol{u}'\widehat{\boldsymbol{B}} \sim N\left(\boldsymbol{u}'\boldsymbol{\beta}, \sigma\sqrt{\boldsymbol{u}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{u}}\right).$$

Confidence interval for $\mu_0 = E(Y_0)$

Construct confidence intervals for $\mu_0 = E(Y_0) = \boldsymbol{u}'\boldsymbol{\beta}$ according to the following sampling distribution

$$\frac{\hat{M}_0 - \mu_0}{S\sqrt{\boldsymbol{u}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{u}}} = \frac{\boldsymbol{u}'\widehat{\boldsymbol{B}} - \boldsymbol{u}'\boldsymbol{\beta}}{S\sqrt{\boldsymbol{u}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{u}}} \sim t(n-k-1).$$

Therefore, $(1-\alpha)$ confidence interval (C.I.) for μ_0

$$I_{\mu_0} = \boldsymbol{u}'\hat{\boldsymbol{\beta}} \mp t_{\alpha/2}(n-k-1)s\sqrt{\boldsymbol{u}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{u}}$$

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

Note: $\hat{\mu}_0 = \boldsymbol{u}'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 u_1 + \ldots + \hat{\beta}_k u_k$

Prediction interval for Y_0

For new observation $\boldsymbol{u} = (1 \; u_1 \; \ldots \; u_k)'$, we have $Y_0 = \boldsymbol{u}'\boldsymbol{\beta} + \varepsilon_0$.

$\Downarrow$

The point estimator of is $Y_0 = \boldsymbol{u}'\widehat{\boldsymbol{B}} + \varepsilon_0$.

$\Downarrow$

$\mathrm{E}(Y_0) = \mathrm{E}\left(\boldsymbol{u}'\widehat{\boldsymbol{B}} + \varepsilon_0\right) = \boldsymbol{u}'\,\mathrm{E}\left(\widehat{\boldsymbol{B}}\right) = \boldsymbol{u}'\boldsymbol{\beta} = \mu_0$

$\Downarrow$

$$\begin{aligned} V(Y_0) &= V(\boldsymbol{u}'\widehat{\boldsymbol{B}}) + V(\varepsilon_0) \\ &= \boldsymbol{u}'\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{u} + \sigma^2 = \sigma^2(\boldsymbol{u}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{u} + 1). \end{aligned}$$

$\Downarrow$

Moreover, Y_0 is normally distributed. Why?

$\Downarrow$

$$Y_0 = \boldsymbol{u}'\widehat{\boldsymbol{B}} + \varepsilon_0 \sim N\left(\boldsymbol{u}'\boldsymbol{\beta}, \sigma\sqrt{\boldsymbol{u}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{u} + 1}\right).$$

Prediction interval for Y_0

The sampling distribution is

$$\frac{Y_0 - \boldsymbol{u}'\boldsymbol{\beta}}{S\sqrt{\boldsymbol{u}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{u} + 1}} \sim t(n-k-1).$$

Therefore, $(1-\alpha)$ prediction interval (P.I.) for Y_0

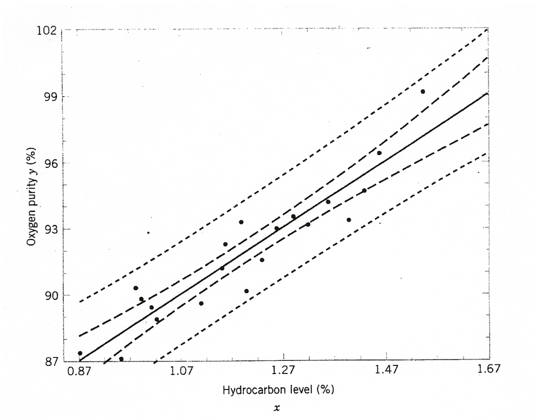
$$I_{Y_0} = \boldsymbol{u}'\hat{\boldsymbol{\beta}} \mp t_{\alpha/2}(n-k-1)s\sqrt{\boldsymbol{u}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{u} + 1}$$

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

Note

- $\hat{\mu}_0 = \boldsymbol{u}'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 u_1 + \ldots + \hat{\beta}_k u_k$
- P.I. I_{Y_0} is wider than C.I. I_{μ_0}.

Confidence interval and prediction interval relation



The chart shows: 1 the observation points, 2 the estimated regression, 3 confidence interval for $\mu_0 = E(Y_0)$ and 4 prediction interval for Y_0.

Example 1 - continued

Example 1 - continued, At the beginning, we have $n = 50$ observations. Now we have new observation: new employees who have creativity $x_4 = 11$ and mathematical ability $x_7 = 30$.

Then we get
$Y_0 = \beta_0 + 11\beta_4 + 30\beta_7 + \varepsilon_0$, where $\varepsilon_0 \sim N(0, \sigma)$.

$\mu_0 = E(Y_0) = \beta_0 + 11\beta_4 + 30\beta_7$. We denote **the new observation** by

$$\boldsymbol{u}' = \begin{pmatrix} 1 & 11 & 30 \end{pmatrix}.$$

(a) Construct 95% confidence interval for μ_0.
(b) Construct 95% prediction interval for Y_0.

Example 1 - continued

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
```


Example 1 - continued

```
>> betahat = regr.beta

betahat =

  248.6924
    0.1169
    2.0603
```

Example 1 - continued

```
>> 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
```


Example 1 - continued

```
>> u= [1 11 30]';
>> u'*XtXinv*u

ans =

   0.021913672127010
```


Example 1 - continued

$$I_{\mu_0} = \boldsymbol{u}'\hat{\boldsymbol{\beta}} \mp t_{\alpha/2}(n-k-1)s\sqrt{\boldsymbol{u}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{u}} = (310.4, 313.2)$$

Where

$$\widehat{\mu}_0 = \widehat{\beta}_0 + 11\widehat{\beta}_4 + 30\widehat{\beta}_7 = \begin{pmatrix} 1 & 11 & 30 \end{pmatrix} \begin{pmatrix} 248.6924 \\ 0.1169 \\ 2.0603 \end{pmatrix} = 311.789;$$

$$t_{0.025}(50-2-1) = t_{0.025}(47) \approx 2.01;$$

$$s = \sqrt{s^2} = \sqrt{21.5633};$$

Example 1 - continued

$$I_{Y_0} = \boldsymbol{u}'\hat{\boldsymbol{\beta}} \mp t_{\alpha/2}(n-k-1)s\sqrt{\boldsymbol{u}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{u}+1} = (302.3, 321.2)$$

Example 1 - continued

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

betahat = regr.beta;
u= [1 11 30]';

s2 = regr.mse;
s = sqrt(s2);
dfe = regr.fstat.dfe;

t = tinv(0.975,dfe);

Cbetahat = regr.covb;
XtXinv = Cbetahat/s2;

% Confidence interval for mu0=E(Y0) = beta0 + 11beta4 + 30beta7
I_EY0 = [u'*betahat-t*s*sqrt(u'*XtXinv*u), u'*betahat+t*s*sqrt(u'*XtXinv*u)]

% Prediktion interval for Y0 = beta0 + 11beta4 + 30beta7 + epsilon0
I_Y0 = [u'*betahat-t*s*sqrt(1+u'*XtXinv*u), ...
                                  u'*betahat+t*s*sqrt(1+u'*XtXinv*u)]
```


Example 2

Example 2, The plot on the next page contains a company's sales Y (unit: thousands of dollars) of televisions for the various quarters for four consecutive years from (year 1 to year 4).

The quarters have been numbered from 1 to 16 (x_1).

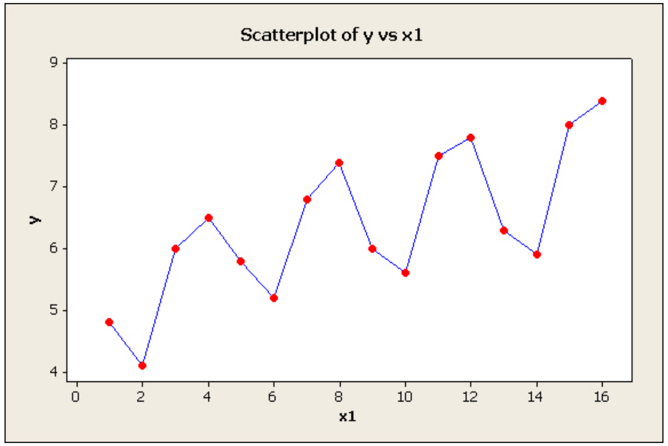
The data has been plotted against quarterly numbers and you see a clear seasonal pattern for each year and possibly also an increase in sales.

Example 2

```
MTB > print c1-c6
```

Data Display

Row	x1	Kvart	y	x2	x3	x4
1	1	1	4,8	0	0	0
2	2	2	4,1	1	0	0
3	3	3	6,0	0	1	0
4	4	4	6,5	0	0	1
5	5	1	5,8	0	0	0
6	6	2	5,2	1	0	0
7	7	3	6,8	0	1	0
8	8	4	7,4	0	0	1
9	9	1	6,0	0	0	0
10	10	2	5,6	1	0	0
11	11	3	7,5	0	1	0
12	12	4	7,8	0	0	1
13	13	1	6,3	0	0	0
14	14	2	5,9	1	0	0
15	15	3	8,0	0	1	0
16	16	4	8,4	0	0	1



By applying linear regression, we can both take into account differences between quarters and find the long-term trend.

Example 2

Data have been analyzed according to the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon, \varepsilon \sim N(0, \sigma)$$

where

$$x_1 = \text{ quarter number}$$

and for $i = 2, 3, 4$

$$x_i = \begin{cases} 1 & \text{for quarter number } i \\ 0 & \text{others.} \end{cases}$$

Example 2

```
y = [4.8 4.1 6.0 6.5 5.8 5.2 6.8 7.4 6.0 ...
                        5.6 7.5 7.8 6.3 5.9 8.0 8.4]';
x1 = [1:16]';
x2 = [0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0]';
x3 = [0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0]';
x4 = [0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1]';

regr = regstats(y,[x1 x2 x3 x4], 'linear','all');
```

Question:

- Can we demonstrate with the model that sales increase over time? Justify your answer using a suitable two-sided 95% confidence interval.

Example 2

```
>> betahat = regr.tstat.beta

betahat =

    4.7056
    0.1456
   -0.6706
    1.0587
    1.3631
```

```
>> se = regr.tstat.se

se =

    0.1376
    0.0121
    0.1537
    0.1551
    0.1575
```


Example 2

$$\frac{\widehat{B}_1 - \beta_1}{d(\widehat{\beta}_1)} \sim t(n-k-1) = t(11), n = 16, k = 4.$$

$$I_{\beta_1} = \widehat{\beta}_1 \mp t_{0.025}(11) d(\widehat{\beta}_1) = (0.12, 0.17) > 0,$$

where $\widehat{\beta}_1 = 0.1456$, $t_{0.025}(11) = 2.20$ and $d(\widehat{\beta}_1) = 0.0121$. So the sales is increasing over time.

Practice after the lecture:

Exercises:

(I) PS-30, PS-31, PS-38, PS-36.

(II) PS-35, 14.4e, PS-33, PS-34.

Thank you!

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

LINKÖPINGS UNIVERSITET