TAMS65 - Lecture 12:
Linear Regression - continued

Zhenxia Liu
Matematisk statistik
Matematiska institutionen

LINKÖPINGS UNIVERSITET

Content

- Repetition
- Residual analysis
- Compare two linear regression models
- Forward selection
- Quiz

Repetition

To study **linear relation** among variables y and $\{x_1, x_2, \ldots, x_k\}$

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

Repetition

- $\hat{\boldsymbol{\beta}} = \begin{pmatrix}\hat\beta_0 & \hat\beta_1 & \ldots & \hat\beta_k\end{pmatrix}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$
- $\sigma^2 \approx \hat{\sigma^2} = s^2 = \frac{SS_E}{n-k-1}$.
- Test on H_0 : $\beta_1 = ... = \beta_k = 0$. The sampling distribution
$$\frac{SS_R/k}{SS_E/(n-k-1)} \sim F(k, n-k-1)$$
- Test on $H_0 : \beta_j = 0$. The sampling distribution
$$\frac{\hat{B}_j - \beta_j}{S\sqrt{h_{jj}}} = \frac{\hat{B}_j - \beta_j}{d(\hat{\beta}_j)} \sim t(n-k-1)$$

Residual analysis

A question

$$\text{Is } \varepsilon \sim N(0, \sigma) \text{ reasonable?}$$

Residual = error = $e_i = y_i - \hat{\mu}_i$, where

$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}, i = 1, \ldots, n.$

- $$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y}$$

- Residual plot.

Note: $\hat{\boldsymbol{y}} = ?$

Residual analysis

According to the model assumption $\varepsilon_i \sim N(0, \sigma), i = 1, \ldots, n$, the residuals should

1. have mean 0.
2. have constant variance.
3. be independent of each other.
4. be normally distributed.

Remark:

• 1. 2. and 3. are important for the regression model.

• 4. is important for the continued analysis when making inference such as: All t- and F- tests are based on the normal distribution assumption.

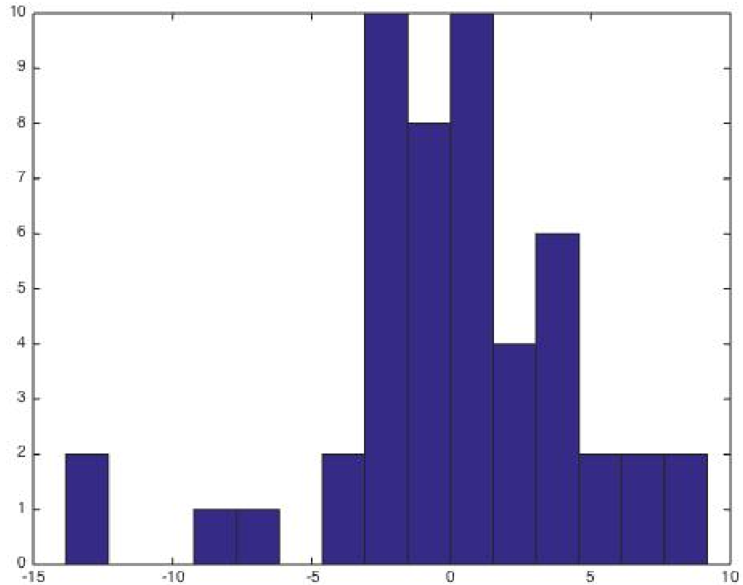
Residual analysis

The easiest way is to do the residual analysis visually, that is, study different residual plots i.e. plots of $e_1, \ldots, e_n$.

We do the following residual plots: Histogram for residuals, Residuals versus observation order, Residual versus **fitted value** and Normal Probability Plot.

- Histogram for residuals
  - The classical bell-shaped, symmetric histogram, i.e. most of the frequency counts bunched in the middle and the counts dying off out in the tails, which indicates the normal distribution.

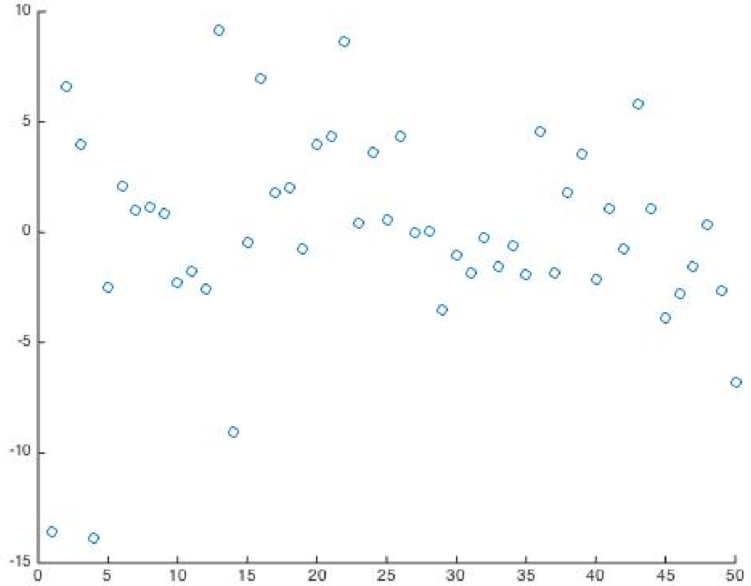
Histogram for residuals



Residual analysis

- Residuals versus observation order
  - The points bounce randomly around the residual 0 line which indicates that the variations of observations are not due to time.

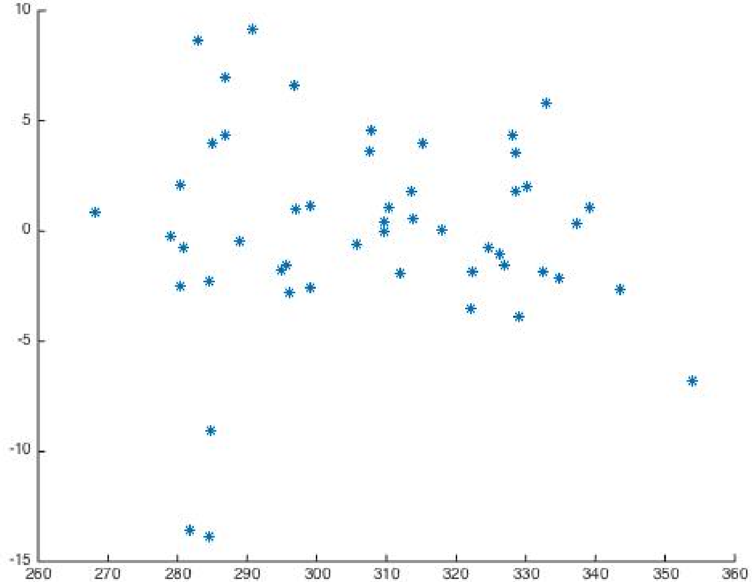
Residuals versus observation order



Residual analysis

- Residual versus **fitted value**
  - **Fitted value** $\hat{\mu}_i$.
  - The points appear to be randomly scattered around 0, so the assumption of mean 0 is reasonable. The vertical width of the scatter doesn't appear to increase or decrease across the fitted values, so the assumption of variance is constant.

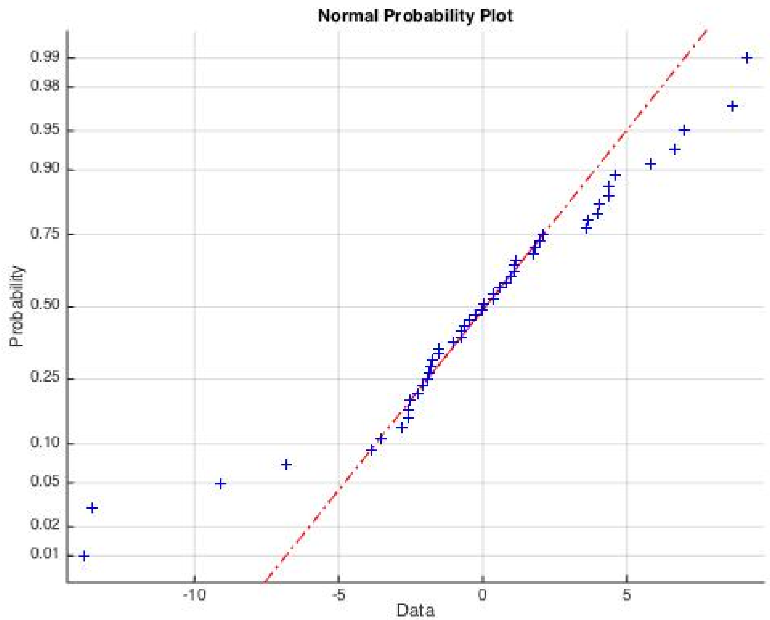
Residual versus fitted value



Residual analysis

- Normal Probability Plot
  - The points form an approximate straight line, which indicates the normal distribution.

Normal Probability Plot



Residual plots i MATLAB.

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

yhat = regr.yhat;
r = regr.r;

figure; hist(r,15)
figure; scatter(1:length(r),r)
figure; scatter(yhat,r,'*')
figure; normplot(r)
```


Compare two linear regression models 15

Compare two linear regression models

Model 1:

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

Model 2:

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \beta_{k+1} x_{k+1} + \ldots + \beta_{k+p} x_{k+p} + \varepsilon,$$

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

Question:
Do the new variables $x_{k+1}, x_{k+2}, \ldots, x_{k+p}$ give new /useful information to Y? or Is the model 2 better than the model 1?

Compare two linear regression models

Compare two linear regression models

Method 1: Compare two coefficient of determinations R_1^2, R_2^2. But this method is not always right, it only gives a general estimation.

Method 2: Make a formal test.

$$\begin{cases} H_0 : \beta_{k+1} = \beta_{k+2} = \ldots = \beta_{k+p} = 0 \\ H_1 : \text{at least one } \beta_{k+i} \neq 0, i = 1, 2, \ldots, p. \end{cases}$$

The sampling distribution is

$$\frac{(SS_E^{(1)} - SS_E^{(2)})/p}{SS_E^{(2)}/(n-k-p-1)} \sim F(p, n-k-p-1),$$

where $SS_E^{(1)} = SS_E$ from Model 1, $SS_E^{(2)} = SS_E$ from Model 2.

Compare two linear regression models

Then we get

$$TS = \frac{(SS_E^{(1)} - SS_E^{(2)})/p}{SS_E^{(2)}/(n-k-p-1)} \text{ and } C = (F_\alpha(p, n-k-p-1), \infty)$$

If $TS \in C$, then reject H_0, i.e. at least one $\beta_{k+i} \neq 0$, which means at least one x_{k+i} is useful or The model 2 is better than the model 1.

Why do we use one sided?

Is $(SS_E^{(1)} - SS_E^{(2)})$ positive?

Example 2

Example 2,

$Y =$ the time used by a bus at a bus stop, e.g., passengers get on or get off the bus,

$x_1 =$ the number of passengers getting on the bus,

$x_2 =$ the number of passengers getting off the bus,

Take a sample

$$\begin{aligned} y &= (4, 24, \ldots, 25), n = 20 \\ x_1 &= (0, 2, \ldots, 1) \\ x_2 &= (1, 3, \ldots, 8) \end{aligned}$$

Time is in seconds.

Example 2

Analyses from Matlab

Model 1: $Y = \beta_0 + \beta_1 x_1 + \varepsilon, \quad \varepsilon \sim N(0, \sigma)$

j	$\hat{\beta}_j$	$d(\hat{\beta}_j)$
0	8.7359	1.7630
1	9.3967	0.6437

	Degrees of freedom	Sum of squares
REGR	1	7523
RES	18	635.5
TOT	19	8158.5

Example 2

Analyses from Matlab

Model 2: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma)$

j	$\hat{\beta}_j$	$d(\hat{\beta}_j)$
0	5.4055	1.6786
1	9.3761	0.5050
2	1.4642	0.4183

	Degrees of freedom	Sum of squares
REGR	2	7789.2
RES	17	369.3
TOT	19	8158.5

Question: Does x_2 affect Y with $\alpha = 1\%$? I am sorry I made a mistake on α in the Lecture Video.

Example 2

Test on $H_0 : \beta_2 = 0$.

Method I: Compare two models. The sampling distribution is

$$\frac{(SS_E^{(1)} - SS_E^{(2)})/p}{SS_E^{(2)}/(n-k-p-1)} \sim F(p, n-k-p-1)$$

$$TS = \frac{(SS_E^{(1)} - SS_E^{(2)})/p}{SS_E^{(2)}/(n-k-p-1)} = \frac{(635.5 - 369.3)/1}{369.3/17} \approx 12.25$$

$C = (F_{0.01}(1,17), \infty) = (8.41, \infty)$ then $TS \in C$, we reject H_0. i.e. x_2 affects Y.

Example 2

Method II: Only consider Model 2. The sampling distribution

$$\frac{\hat{B}_2 - \beta_2}{S\sqrt{h_{22}}} = \frac{\hat{B}_2 - \beta_2}{d(\hat{\beta}_2)} \sim t(n-k-1)$$

$$TS = \frac{\hat{\beta}_2 - 0}{d(\hat{\beta}_2)} = \frac{1.4642}{0.4183} = \approx 3.5$$

$C = (-\infty, -t_{0.005}(17)) \cup (t_{0.005}(17), \infty) = (-\infty, -2.9) \cup (2.9, \infty)$ then $TS \in C$, we reject H_0. i.e. x_2 affects Y.

Note: Since here Model 2 only contains one extra explonatory varialbe, so we have two methods. Otherwise, if we have more than one extra explonatory variables, we can only use method I, i.e. compare two models.

Forward selection

Sometimes we have a response variable y and a whole set of conceivable explanatory variables $x_1, x_2, \ldots, x_k$. But we do not know which explanatory variables that are relevant. Then we can select explanatory variables using so-called **forward selection**.

Forward selection(framåtvalsprincipen)

Forward selection(framåtvals- principen) is a type of stepwise regression which usually begins with an empty model and adds in variables one by one. In each forward step, you add the one variable that gives the single best improvement to your model.

- We can also begin with a given model.
- There are many other methods for incremental regression. For example, one type of of stepwise regression is called **backward elimination (bakåtelimination)** Next, we will use an example to explain how **the Forward selection** works.

Forward selection

In Los Angeles, people want to construct statistical models with meteorological morning data, which can predict the maximum air pollution level during the day. The purpose is to be able to warn in the morning and possibly via traffic restrictions can prevent excessive levels of pollution. People have collected data on a certain oxidant y (a photochemical pollutant) as well morning values of four meteorological variables, wind speed w, temperature t, humidity h and sun insolation i:

Note:

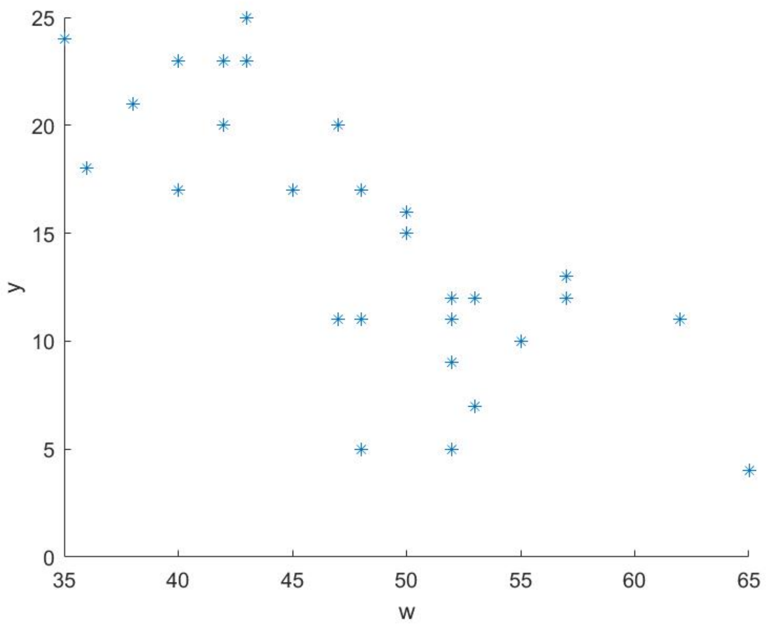
- A response variable y.

- Explanatory variables w, t, h, i.

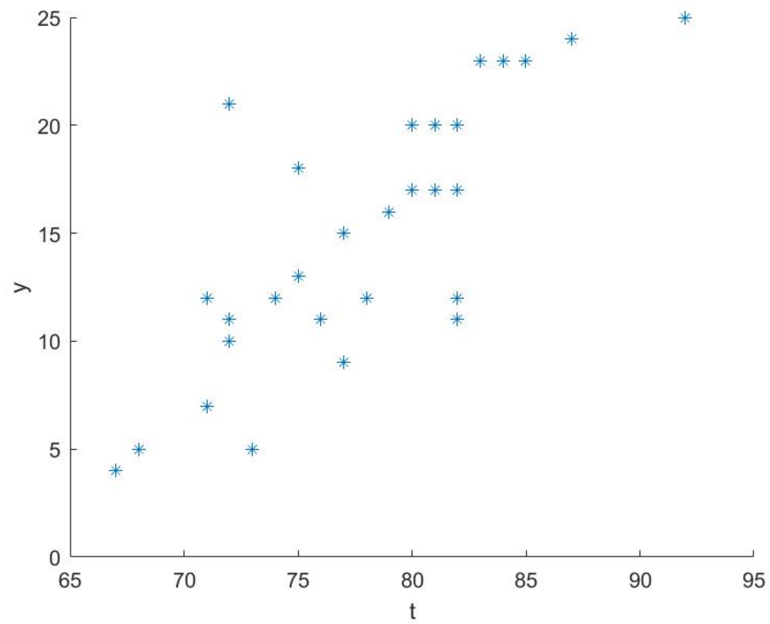
Forward selection

Day	Wind Speed	Temperature	Humidity	Insolation	Oxidant
1	50	77	67	78	15
2	47	80	66	77	20
3	57	75	77	73	13
4	38	72	73	69	21
5	52	71	75	78	12
6	57	74	75	80	12
7	53	78	64	75	12
8	62	82	59	78	11
9	52	82	60	75	12
10	42	82	62	58	20
11	47	82	59	76	11
12	40	80	66	76	17
13	42	81	68	71	20
14	40	85	62	74	23
15	48	82	70	73	17
16	50	79	66	72	16
17	55	72	63	69	10
18	52	72	61	57	11
19	48	76	60	74	11
20	52	77	59	72	9
21	52	73	58	67	5
22	48	68	63	30	5
23	65	67	65	23	4
24	53	71	53	72	7
25	36	75	54	78	18
26	45	81	44	81	17
27	43	84	46	78	23
28	42	83	43	78	23
29	35	87	44	77	24
30	43	92	35	79	25

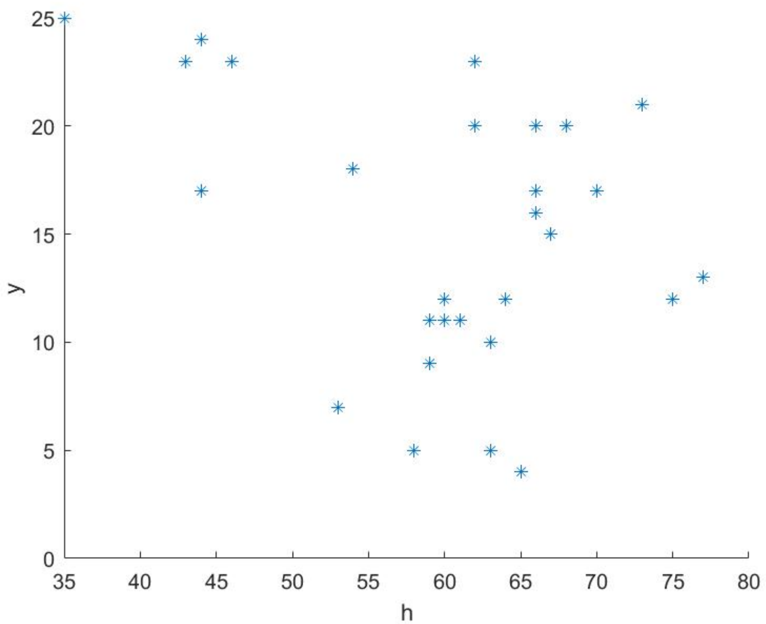
Forward selection



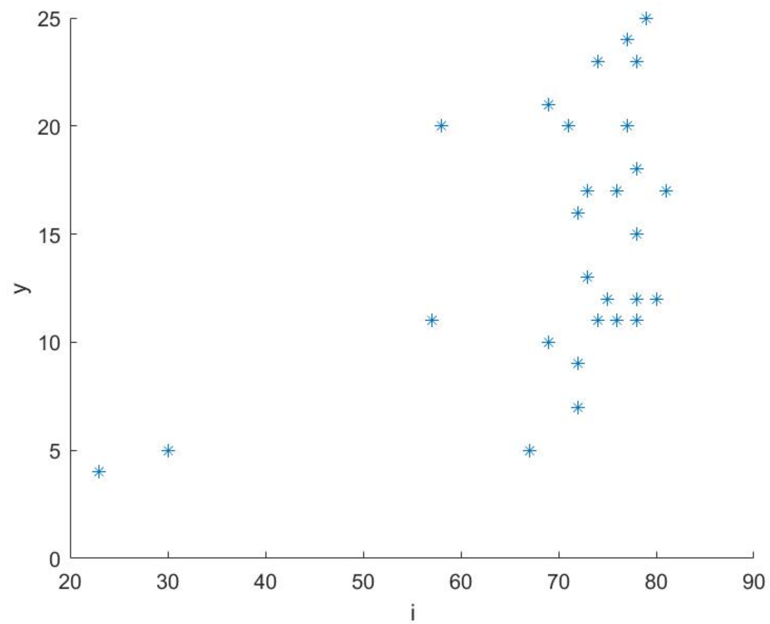
Forward selection



Forward selection



Forward selection



Forward selection - Step I

Step I - 1: We are looking for the best single explanatory variable by correlations.

```
>> correlation = corr([w t h i],y)

correlation =

    -0.7657
    0.7570
    -0.3522
    0.5051
```

w is the best single explanatory variable.

Forward selection - Step I

Step I - 2: Do a regression analysis with *w* and check if *w* is useful.

```
>> regr = regstats(y,w,'linear','all');
>> t = regr.tstat.t

t =
  9.2528
 -6.2996

>> P = regr.tstat.pval

P =
   1.0e-06 *

    0.0005
    0.6128
```

$TS = -6.2996$ and $p-$value $< \alpha = 5\%$, i.e. reject $H_0 : \beta_1 = 0$ and w is useful.

Forward selection - Step II

Step II - 1: Do all analyzes with w and **one** another explanatory variable. Compare SS_E for these three analyzes and select the one that has the least.

```
disp('-- w och h --')
regr = regstats(y,[w h], ...
              'linear','all');
sse = regr.fstat.sse

disp('-- w och t --')
regr = regstats(y,[w t], ...
              'linear','all');
sse = regr.fstat.sse

disp('-- w och i --')
regr = regstats(y,[w i], ...
              'linear','all');
sse = regr.fstat.sse
```

```
-- w och h --

sse =
   431.0908

-- w och t --

sse =
  234.8971

-- w och i --

sse =
  357.2503
```


Forward selection - Step II

$$SS_E^{w,t} = 235 \quad < \quad SS_E^{w,i} = 357 \quad < \quad SS_E^{w,h} = 431$$

Select t.

Forward selection - Step II

Step II - 2: Choose $Y = \beta_0 + \beta_1 w + \beta_2 t + \varepsilon$ and then test if t is useful.

```
disp('-- w och t --')
regr = regstats(y,[w t], ...
              'linear','all');
t = regr.tstat.t
P = regr.tstat.pval
```

```
-- w och t --

t =
   -0.4680
   -4.9401
    4.8121

P =
    0.6435
    0.0000
    0.0001
```

$TS = 4.8121$ and $p-\text{value} = 0.0001 < \alpha$, i.e. reject H_0 and t is useful. Take t into the model.

Forward selection - Step III (**repeat Step II**)

Step III - 1: Do all analyzes with w, t and **one** another explanatory variable. Compare SS_E for these two analyzes and select the one that has the least.

```
disp('-- w, t och i --')
regr = regstats(y,[w t i], ...
              'linear','all');
sse = regr.fstat.sse

disp('-- w, t och h --')
regr = regstats(y,[w t h], ...
              'linear','all');
sse = regr.fstat.sse
```

```
-- w, t och i --

sse =

  230.3194

-- w, t och h --

sse =

  214.8067
```

Forward selection - Step III(**repeat Step II**)

$$SS_E^{w,t,h} = 215 \quad < \quad SS_E^{w,t,i} = 230$$

Select h.

Forward selection - Step III (**repeat Step II**)

Step III - 2: Choose $Y = \beta_0 + \beta_1 w + \beta_2 t + \beta_3 h + \varepsilon$ and then test if h is useful.

```
regr = regstats(y,[w t h], ...
                'linear','all');
t = regr.tstat.t
P = regr.tstat.pval
```

```
t =
   -1.2705
   -5.2413
    5.1165
    1.5594

P =
    0.2152
    0.0000
    0.0000
    0.1310
```

$TS = 1.5594$ and $p-\text{value} = 0.1310 > \alpha = 5\%$, i.e. we don't reject H_0. **Stop!** Actually, If we reject H_0 in Step II, then we will repeat Step II with extra variables until we don't reject H_0!

Forward selection

Forward selection gives the model

$$Y = \beta_0 + \beta_1 w + \beta_2 t + \varepsilon.$$

Now we want to test whether the model with all four explanatory variables is significantly better.

```
>> regr = regstats(y,[w t], ...
                   'linear','all');
>> fstat = regr.fstat

fstat =
     sse: 234.8971
     dfe: 27
     dfr: 2
     ssr: 819.9029
       f: 47.1214
    pval: 1.5633e-09
```

```
>> regr = regstats(y,[w t h i], ...
                   'linear','all');
>> fstat = regr.fstat

fstat =
     sse: 213.0881
     dfe: 25
     dfr: 4
     ssr: 841.7119
       f: 24.6879
    pval: 2.2791e-08
```


Forward selection

Model 1: $Y = \beta_0 + \beta_1 w + \beta_2 t + \varepsilon$, where $\varepsilon \sim N(0, \sigma)$

Model 2:
$Y = \beta_0 + \beta_1 w + +\beta_2 t + \beta_3 h + \beta_4 i + \varepsilon$, where $\varepsilon \sim N(0, \sigma)$

Test on $H_0 : \beta_3 = \beta_4 = 0$.

The sampling distribution is

$$\frac{(SS_E^{(1)} - SS_E^{(2)})/p}{SS_E^{(2)}/(n-k-p-1)} \sim F(p, n-k-p-1)$$

$$TS = \frac{(SS_E^{(1)} - SS_E^{(2)})/p}{SS_E^{(2)}/(n-k-p-1)} = \frac{(234.8971 - 213.0881)/2}{213.0881/25} = 1.28$$

$C = (F_{0.05}(2, 25), \infty) = (3.41, \infty)$ then $TS \notin C$, we don't reject H_0.

Thus, we choose the small model.

Quiz

Distance Quiz is given on Lisam.

- When? May, 6, 2020. **13:15-13:30**
  - Where? **Lisam - Quiz - Distance Quiz (13:15-13:30)**
- Students who need extra time.
  - May, 6, 2020. **13:15-13:36**
  - Where?
    Lisam - Quiz - Distance Quiz (13:15-13:36) - Extra time
  - **Email** me and attach your **certificate** on the extra time from Liu(Linköping University).

Quiz

- The quiz is based on the project and lectures.
  - It contains 5 questions with multiple choice options.
  - You will randomly get 5 questions from the Quiz question bank.
- Pass the Quiz = at least 3 questions out of 5 are right.

A sample Quiz question: Circle the right answer:

How many explanatory variables (förklaringsvariabler) does the following model have

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon, \varepsilon \sim N(0, \sigma)$$

a) 0 b) 1 c) 2

Project

Coming timetable for the Distance Project

- The Project will be released to the Lisam on **Apr. 10, 2020**.
- **Submit first version** of your report to your **teaching assistants** not later than May 1, 2020.
- **Submit final version** of your report to the lisam: Lisam - Submissions
  - Deadline for submission is at 23:00 May 15, 2020.
  - The submit entrance will open at 0:00 May 7, 2020.
  - The submit entrance will close at 8:00 May, 16, 2020.

Project

- To pass the Project you should do the followings.
  - Title your attached project in pdf.file,and name it as Project.pdf
  - Choose **2 assignments** out of 7 assignments, and make a detailed report on these two assignments.
  - Write down only solutions to the rest of 5 assignments. Here you don't need to show extra information.

For details, please read **Instructions to the Distance Project** on the Lisam: Lisam - Course documents - 7 Project - Instructions to the Distance Project.

TAMS65 VT2

Pass the course =

=Pass the written Exam

+ Pass the project(VT2) + Pass the quiz(VT2)

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

LINKÖPINGS UNIVERSITET

Practice after the lecture:

Exercises:

(I) PS-39, PS-41.

(II) PS-40, PS-42.

Thank you!