TAMS24 — Computer Lab 2

Problem 1.
Simple linear regression

A geyser is a hot spring, which more or less regularly erupts. During an eruption, the water can spray high into the air. Old Faithful Geyser in Wyoming is one such source which has become a tourist attraction. Time between two consecutive eruptions is usually long, it is therefore interested in being able to predict the time until the next eruption. It is believed that this time depends on the length of the previous eruption. To construct such a model we put

$$\begin{aligned} x &= \text{ the length of the last eruption (enhet:min) and} \\ y &= \text{ time till the next eruption (enhet:min).} \end{aligned}$$

We will use the model

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$$

where

$$\varepsilon_1, ..., \varepsilon_n \text{ are independent and } N(0, \sigma).$$

Go to our course webpage and download the file *uppg1.m*, where values of x and y can be found.

a) Run *uppg1.m* in the command window. We first plot y against x in order to see if there is a linear relation between them.

```
plot(x,y,'b*')
```

We can even find the correlation coefficient $\rho_{X,Y}$ using the command `corr`.

Now we do a full analysis of the linear regression using `regstats`, that is (try to understand the meaning of each command),

```
stats = regstats(y,x,'linear','all');
betahat = stats.tstat.beta
se = stats.tstat.se
t = stats.tstat.t
s2 = stats.mse
fstat = stats.fstat
```

b) What is the regression line ? To see how the regression fits the points, we do:

```
figure
scatter(x,y,'*')
xlabel('x'), ylabel('y')
hold on
lsline % ls = least square, this is how we obtain the regression line
```

c) In the output of Matlab, find the standard errors of the coefficients, namely, $d(\hat{\beta}_0)$ and $d(\hat{\beta}_1)$, (se = standard error), in Lecture we used $s\sqrt{h_{00}} = d(\hat{\beta}_0)$ and $s\sqrt{h_{11}} = d(\hat{\beta}_1)$.

We know that σ^2 can be estimated by s^2. Write down the formula of the estimate s^2.

d) With a significance level $\alpha = 0.01$, test the hypotheses

$$H_0 : \beta_1 = 0 \text{ mot } H_1 : \beta_1 \neq 0$$

Reject H_0 ?

e) It is assumed that the error terms $\varepsilon_j \sim N(0, \sigma)$, but is this really true ? We will study this by looking at the residuals:

```
residualer = stats.r;

figure
scatter(x,residualer,'filled')
title('Residualer')
```

The idea is: if there is no obvious pattern in the plot of residuals, then it is reasonable to say $\varepsilon_j \sim N(0, \sigma)$. But if there is an obvious pattern, then the error terms are not normal. Do you think $\varepsilon_j \sim N(0, \sigma)$?

f) Suppose that we just had an eruption for 4 minutes (namely, $x = 4$), find a 95% **prediction interval** for the time till next eruption (namely y). You need to use the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ which can be found as

```
XtXinv = stats.covb/stats.mse
```

Write your solutions on the last solution page. Save your codes/commands and graphs.

Problem 2.
Evaluation of the linearity of the data

It is believed that 'current' (x) and 'weld seam strength' (y) are related, during the use of welding machines. In a study, we have the following measurements:

x	y		
4000	600	750	550
5000	1850	2000	1750
6000	2700	2650	2650
7000	3650	3800	3600
8000	4400	4550	4350
9000	4900	4800	4850

Go to our course webpage to download the file *uppg2.m*, which contains the above data.

We don't know yet if a linear regression (with a normal error) between y and x is appropriate, so we first try this.

$$\textbf{Model 1:} \quad Y = \beta_0 + \beta_1 x + \varepsilon \text{ where } \varepsilon \sim N(0, \sigma_1)$$

Then we do the linear regression as follows

```
regr = regstats(y,x,'linear','all');
```

Now we produce the plot of residuals

```
residualer = regr.r;

figure
scatter(x,residualer,'filled')
```

a) Find an estimate $\hat{\sigma}_1 = s_1$ of σ_1.

b) In the plot of residuals, can you see any pattern ? Do you think $\varepsilon \sim N(0, \sigma_1)$?

c) The plot of residuals in b) suggests to consider the following model (why ?)

$$\textbf{Model 2:} \quad Y = \beta_0' + \beta_1' x + \beta_2' x^2 + \varepsilon' \text{ where } \varepsilon' \sim N(0, \sigma_2)$$

Use a new variable x2 to denote x^2: `x2 = x.^2`, then we have a multiple linear regression. To show that Model 2 works better, we will plot the original points and the regression line on one graph.

```
x1=x;
x2=x.^2;
regre = regstats(y,[x1, x2],'linear','all');
betahat=regre.tstat.beta;

xx = [4000:10:9000]';
xx2 = xx.^2;
yy = [ones(length(xx),1) xx xx2]*betahat;

figure
scatter(x,y,'*') % the original points
hold on
plot(xx,yy) % the points on the regression line
```

Does Model 2 seem to work better ?

d) Make the plot of residuals yourself. Do you think $\varepsilon' \sim N(0, \sigma_2)$?

End of Lab 2

There is an additional problem on the next page on how to perform step-by-step model selection. If you have time and are willing to, you can do by yourself.

Problem 3 (optional).
Stegvis regression enligt framåtvalsprincipen

En amerikansk fastighetsmäklare vill utveckla en regressionsmodell med vars hjälp man kan prissätta stora enfamiljshus i utkanterna av en stad. I datamaterialet finns uppgifter om 30 hus som nyligen sålts i det aktuella området.

Property i	Property Taxes (dollars) x_{i1}	House Size (sq. feet) x_{i2}	Lot Size (acres) x_{i3}	Lot Size Sq. x_{i4}	Attractiveness Index x_{i5}	Selling Price ($000) y_i
1	7337	3000	3.6	12.96	64	550
2	4204	2300	1.2	1.44	69	461
⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	4612	2900	1.1	1.21	74	477

Öppna filen *uppg3.m*. Ni ska nu välja en regressionsmodell enligt framåtvalsprincipen, dvs. först ska ni välja den förklaringsvariabel som har störst korrelation med y för att sedan lägga till en variabel i taget enligt vissa kriterier.

Step 1. Sök först upp den x-variabel som är **starkast korrelerad** med Y.

```
correlation=corr([x1, x2, x3, x4, x5], y)
```

Gör sedan en regressionsanalys med denna bästa ensamma förklaringsvariabel.

Anteckna modellen, residualkvadratsumman (`sse`) och dess frihetsgrad (`dfe`). Undersök om ”sambandet är signifikant” (nivå 5%), dvs. avgör med ett test om β-koefficienten framför x-variabeln är skild från noll (vilket görs enklast med hjälp av t-kvoten som i uppgift 1).

Om β-koefficienten är signifikant skild från noll är det klart att denna första förklaringsvariabel ska ingå i modellen.

Step 2. Nästa steg är då att kombinera **var och en av de övriga** förklaringsvariablerna **med den först valda** och genomföra regressionsanalyser för att hitta det par, som bäst förklarar variationen hos Y, dvs. ger minst residualkvadratsumma. Du behöver alltså göra fyra olika regressionsanalyser.

Anteckna modellen, residualkvadratsumman och dess frihetsgrad för varje analys. Undersök för det bästa paret om den nya förklaringsvariabeln gör signifikant nytta (t-test på nivå 5%).

Step 3. Om även den andra förklaringsvariabeln du hittat skall ingå i modellen, så är nästa steg att man **studerar alla modeller med tre förklarande variabler**, där de båda först valda ingår.

Fortsätt enligt framåtvalsprincipen och plocka in flera förklaringsvariabler i modellen (nivå 5% för samtliga test). Proceduren slutar då β-koefficienten framför den senaste förklaringsvariabeln inte är signifikant skild från noll.

Vilken modell ger framåtvalsprincipen?

SOLUTION PAGE. Write down your name and personal number.

1)

.....

2)

.....

Problem 1 **OK** .....

a) Correlation coefficient $\rho_{X,Y}$ = ...

b) Regression line is: .....

c) $d(\hat{\beta}_0)$ =..... $d(\hat{\beta}_1)$ =.....

Formula s^2 = ——

d) Test statistic TS =..... ; critical region C =: ..... ; Reject H_0? .....

e) Are the error terms normal ?

f) The prediction interval is: .....

$(\mathbf{X}'\mathbf{X})^{-1}$ =

Problem 2 **OK** .....

a) Do you think s_1 = ...

b) $\varepsilon \sim N(0, \sigma_1)$?

c) Does Model 2 seem to work better ?

d) Do you think $\varepsilon' \sim N(0, \sigma_2)$?

Problem 3 (optional) **OK** .....

Bästa ensamma förklaringsvariabel: .....

Modell enligt framåtvalsprincipen (skriv in den teoretiska modellen **inte** den med skattade parametrar):

.....

Skattade parametrar: .....