SF2930 - Regression analysis

KTH Royal Institute of Technology, Stockholm

Lecture 4 – Multiple linear regression (MPV 3)

January 28, 2022

Todays lecture

- The multivariate normal distribution
- Maximum likelihood estimates
- Test for regression coefficients
- Coefficients of determination
- Confidence regions and sets
- Project 1 handout

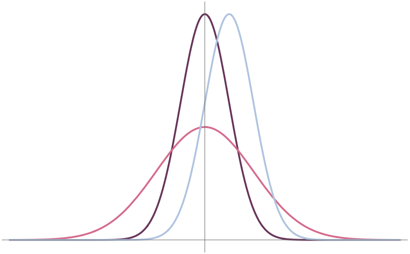
General assumption

General assumption

To evaluate the model, we need further assumptions on the errors, and will assume that they are independent with distribution $N(0, \sigma^2)$.

The normal distribution

X has a normal distribution if it has pdf $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma}$. We write $X \sim N(\mu, \sigma)$. Recall that if $X \sim N(\mu, \sigma)$, then $X + \mu' \sim N(\mu + \mu', \sigma)$ and $\sigma' X \sim N(\mu\sigma', \sigma^2\sigma'^2)$.

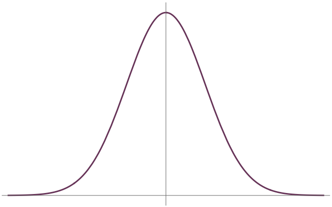


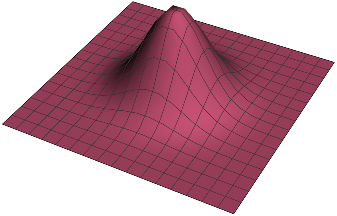
The multivariate normal distribution

The standard multivariate normal distribution

Let $X_1', X_2', \ldots, X_n' \sim N(0,1)$ be independent. Then

$$X = (X_1', X_2', \ldots, X_n')^T \sim N(0, I).$$





The multivariate normal distribution

The multivariate normal distribution with independent marginals

Let $X_1' \sim N(\mu_1, \sigma_1^2)$, $X_2' \sim N(\mu_2, \sigma_2^2), \ldots$ be independent, and let

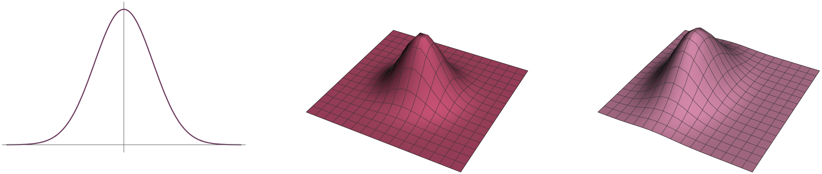
$$X = (X_1', X_2', \ldots, X_n')^T.$$

With

- $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^T$,
- $A = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$, and
- $X'' \sim N(0, I)$

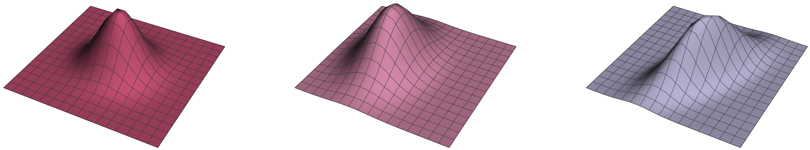
we equivalently have $X \coloneqq \boldsymbol{\mu} + AX''$.

We write $X \sim N(\boldsymbol{\mu}, A^2)$.



The multivariate normal distribution

The multivariate normal distribution
Let $X' \sim N(0, I)$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^T$, let A be a general invertible matrix, and let $X = \boldsymbol{\mu} + AX'$. We write $X \sim N(\boldsymbol{\mu}, A^T A)$.



The multivariate normal distribution

Properties

$$\mathbb{E}[X_i] = \mathbb{E}[\mathbf{e}_i^T X] = \mathbf{E}\big[\mathbf{e}_i^T(\boldsymbol{\mu} + AX')\big] = \mathbf{E}\big[\boldsymbol{\mu}^T \mathbf{e}_i + (AX')^T \mathbf{e}_i\big] = \mu_i$$

$$\begin{aligned}
\mathrm{Cov}[X_i, X_i] &= \mathbf{E}\big[(\mathbf{e}_i^T X - \mu_i)(\mathbf{e}_j^T X - \mu_j)\big] \\
&= \mathbf{E}\Big[\big(\mathbf{e}_i^T(\boldsymbol{\mu} + AX') - \mu_i\big)\big(\mathbf{e}_j^T(\boldsymbol{\mu} + AX') - \mu_j\big)\Big] = \mathbf{E}\big[(\mathbf{e}_i^T AX')(\mathbf{e}_j^T AX')\big] \\
&= \mathbf{E}\Big[\Big(\sum_k A_{ik} X'_k)(\sum_\ell A_{i\ell} X'_\ell\Big)\Big] = \sum_k A_{ik} A_{jk} = AA^T(i,j) = A^T A(i,j),
\end{aligned}$$

where the last equation follows from the fact that $\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1)$. We say that $X \sim N(\boldsymbol{\mu}, A^T A)$ has with *mean vector* $\boldsymbol{\mu}$ and *covariance matrix* AA^T.

If B is *positive definite*, then there is an invertible matrix A such that $B = A^T A$, and we may write $X \sim N(\boldsymbol{\mu}, B)$ instead of $X \sim N(\boldsymbol{\mu}, A^T A)$.

The multivariate normal distribution

Probability density function

One can verify that $X \sim N(\boldsymbol{\mu}, B) = N(\boldsymbol{\mu}, A^T A)$ has pdf

$$f(\mathbf{x}) := \frac{1}{\sqrt{(2\pi)^n \det B}} e^{(\mathbf{x}-\boldsymbol{\mu})^T B^{-1} (\mathbf{x}-\boldsymbol{\mu})/2}.$$

ML-esimates vs. LS estimates

A maximum likelihood estimate of β.
Assume that the model is $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$. Then the pdf of $\boldsymbol{\varepsilon}$ is given by

$$f(\boldsymbol{\varepsilon}) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}/2\sigma^2} =: L(\boldsymbol{\varepsilon}, \boldsymbol{\beta}, \sigma^2).$$

Hence

$$\begin{aligned}\log L(\boldsymbol{\varepsilon}, \boldsymbol{\beta}, \sigma^2) &= -\log(2\pi)^{n/2}\sigma^n - \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}/2\sigma^2 \\ &= -\log(2\pi)^{n/2}\sigma^n - (\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta)/2\sigma^2\end{aligned}$$

$\rightarrow \log L(\boldsymbol{\varepsilon}, \boldsymbol{\beta}, \sigma^2)$ is maximal when $(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta)$ is as small as possible, i.e., when $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. Consequently, the ML estimates are equal to the least squares estimates.

Significance of regression using ANOVA

Hypothesis

$$H_0\colon \beta_1 = \beta_2 = \ldots = \beta_k = 0 \qquad H_1\colon \beta_j \neq 0 \text{ for at least one } j$$

General idea

Recall the ANOVA identity

$$\underbrace{\sum (y_i - \bar{y})^2}_{SS_T} = \underbrace{\sum (\hat{y}_i - y_i)^2}_{SS_{Res}} + \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{SS_R}.$$

If H_0 is correct, then $y_i = \beta_0 + \varepsilon_y$. SS_R gives a measure on how much the residuals vary, while SS_{Res} measures how much the residuals vary in an "optimal" linear model. If H_0 is false, then SS_{Res} should be much smaller than SS_R.

Distribution of SS_R and SS_{Res}

Appendix C.3, they show that if H_0 is true, then SS_{Res} and SS_R are independent, $SS_R \sim \chi^2_k$, and $SS_{Res} \sim \chi_{n-k-1}$.

Test statistic

$$F_0 \coloneqq \frac{SS_R/k}{SS_{Res}/(n-k-1)} \sim F_{k,n-k-1}$$

Reject H_0 if $F_0 > F_{\alpha,k.n-k-1}$.

The coefficients of determination

Recall $SS_T = \sum(y_i - \bar{y})^2$, $SS_{Res} = \sum(\hat{y}_i - y_i)^2$, $SS_R = \sum(\hat{y}_i - \bar{y}_i)$.

The coefficient of determination
"The proportion of the variation explained by the regressors"

$$R^2 = SS_R/SS_T = 1 - SS_{Res}/SS_T$$

R^2 close to one means most of the variability is explained by the model.

Example

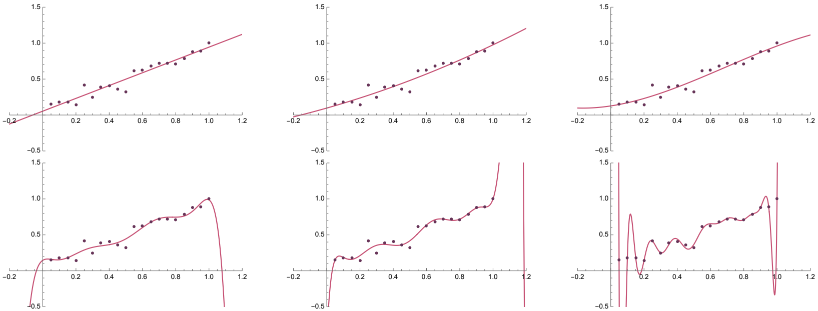
```r
df00.model <- lm(people_fully_vaccinated_per_hundred~gdp_per
    _capita, data = df00)
summary(df00.model)
```

```

Call:
lm(formula = people_fully_vaccinated_per_hundred ~ gdp_per_
    capita, data = df00)

Residuals:
    Min      1Q  Median      3Q     Max
-16.428  -6.176  -0.675   7.997  14.445

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.929e+01  6.075e+00   3.175  0.00588 **
gdp_per_capita 1.194e-03  1.957e-04   6.100 1.53e-05 ***
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.665 on 16 degrees of freedom
Multiple R-squared:  0.6993,  Adjusted R-squared:  0.6805
F-statistic: 37.21 on 1 and 16 DF,  p-value: 1.534e-05
```

The coefficients of determination

A problem with R^2

$$R^2 = 1 - SS_{Res}/SS_T = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y}_i)^2} = 1 - \frac{\sum e_i^2}{\sum(y_i - \bar{y}_i)^2}$$



- As we increase the degree number of regressors (in this case, the degree of the polynomial (regressors $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$, etc.), $SS_{Res} = \sum e_i^2$ decreases while SS_T is constant.
- R^2 is also affected by the positions of the points, as this affects SS_T.
- Encourages overfitting

The coefficients of determination

The adjusted coefficient of determination

$$R^2_{adj} = 1 - \frac{SS_{Res}/(n-k-1)}{SS_T/(n-1)}$$

Here the denominator does not depend on the number of variables in the model, and $SS_{Res}/(n-k-1)$ is the residual mean square, which do not necessarily decrease when we add a new variable.

Extra sum of squares

Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$.

Hypothesis

$$H_0\colon \boldsymbol{\beta}_2 = 0 \qquad H_1\colon \boldsymbol{\beta}_2 \neq 0.$$

Extra sum of squares

$$\begin{cases} \mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} & \text{the full model} \\ \mathbf{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} & \text{the reduced model} \end{cases}$$

The regression sum of squares that is due to adding β_2 to the reduced model is given by

$$SS_R(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1) = SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta_1}) \qquad \text{extra sum of squares due to } \boldsymbol{\beta}_2.$$

General idea

If the null hypothesis is true, then $SS_R(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1)$ should be small.

Extra sum of squares

$$H_0\colon \boldsymbol{\beta}_2 = 0 \qquad H_1\colon \boldsymbol{\beta}_2 \neq 0. \qquad \text{the hypothesis}$$

$$\begin{cases} \mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} & \text{the full model} \\ \mathbf{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} & \text{the reduced model} \end{cases}$$

$$SS_R(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1) = SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta_1}) \qquad \text{extra sum of squares due to } \boldsymbol{\beta}_2.$$

Properties

- $SS_{Res} \sim \chi^2_{n-(k+1)}$.
- $SS_R(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1) \sim \chi^2_r$, where r is the number of parameters in $\boldsymbol{\beta}_2$.
- If H_0 is true, then $SS_R(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1)$ is independent of $SS_{Res}(\boldsymbol{\beta})$.

Statistic

$$F_0 \coloneqq \frac{SS_R(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1)/r}{SS_{Res}/(n-k-1)} \sim F_{r,n-k-1}$$

$\rightarrow$ Reject if $F_0 > F_{\alpha,r,n-k-1}$.

Example

The function anova in R can be used to access the extra sums of squares sequentially, and to perform the corresponding tests.

```r
anova(df00.model2)
```

```
Analysis of Variance Table

Response: people_fully_vaccinated_per_hundred
                    Df Sum Sq Mean Sq F value    Pr(>F)    
gdp_per_capita       1 3475.8  3475.8 41.4921 1.111e-05 ***
hospital_beds_per_th 1  238.1   238.1  2.8417    0.1125    
Residuals           15 1256.6    83.8                      
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the column `Sum Sq` above contains $SS_R(\beta_j|\beta_0, \beta_1, \ldots, \beta_{j-1})$. Hence from the above table, we see that there is no real support for adding the regression variable `hospital_beds_per_thousand` to our model.

Tests for single coefficients

We want to test whether $\beta_j = 0$, since this would motivate removing β_j from our model.

Hypothesis

$$H_0\colon \beta_j = 0 \qquad H_1\colon \beta_j \neq 0$$

Test statistic

$$t_0 := \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 (X^T X)^{-1}(j+1, j+1)}} \sim t_{n-k-1}$$

Reject if $t_0 > t_{\alpha, n-k-1}$.

Comments

- This test is conditional on all other regressors being present, hence we cannot use this for all variables, and then remove all the variables that failed, but rather have to perform such tests in a sequence.
- We can use this to find confidence intervals for $\hat{\beta}_j$. However, this will only give confidence intervals for one variable. For intervals for several of the coefficients simultaneously, see MPV 3.4.3.

Example

```r
df00.model2 <- lm(people_fully_vaccinated_per_hundred~gdp_
    per_capita+hospital_beds_per_thousand, data = df00)
summary(df00.model2)
```

```
Call:
lm(formula = people_fully_vaccinated_per_hundred ~ gdp_per_
    capita + hospital_beds_per_thousand, data = df00)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.7639  -4.4811   0.0485   5.8690  12.7837 

Coefficients:
                             Estimate Std. Error t value Pr
    (>|t|)
(Intercept)       33.7305585 10.3210623   3.268  0.00519 **
gdp_per_capita     0.0011229  0.0001901   5.908 2.88e-05 ***
hosp_beds_per_th  -2.1185866  1.2567801  -1.686  0.11253
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.153 on 15 degrees of freedom
Multiple R-squared:  0.7472,  Adjusted R-squared:  0.7135
F-statistic: 22.17 on 2 and 15 DF,  p-value: 3.318e-05
```

Confidence sets for a single regression coefficient

Since

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$$

we have

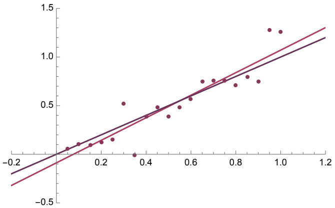
$$t_0 := \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 (X^T X)^{-1}(j+1, j+1)}} \sim t_{n-k-1}.$$

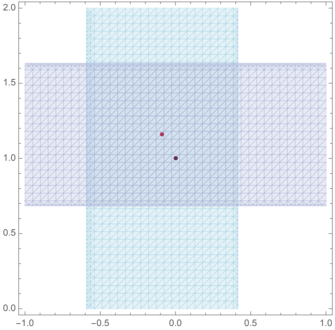
We can use to construct a confidence interval on level α exactly as before.

Confidence sets for a single regression coefficient

```
confint(df00.model2, level=0.95)
```

```
                                    2.5 %      97.5 %
(Intercept)                  11.7317350225 55.729382016
gdp_per_capita                0.0007177958  0.001528066
hospital_beds_per_thousand   -4.7973501033  0.560176819
```





Joint confidence sets

Important!!!
We do not have

$$P\big(\beta_1 \in CI_\alpha(\hat{\beta}_1) \text{ and } \beta_2 \in CI_\alpha(\hat{\beta}_2)\big) = P\big(\beta_1 \in CI_\alpha(\hat{\beta}_1)\big)P\big(\beta_2 \in CI_\alpha(\hat{\beta}_2)\big).$$

In fact, we only have

$$\begin{aligned} P\big(\beta_1 \in CI_\alpha(\hat{\beta}_1) \text{ and } \beta_2 \in CI_\alpha(\hat{\beta}_2)\big) &= 1 - P\big(\beta_1 \notin CI_\alpha(\hat{\beta}_1) \text{ or } \beta_2 \notin CI_\alpha(\hat{\beta}_2)\big) \\ \geq 1 - P\big(\beta_2 \notin CI_\alpha(\hat{\beta}_2)\big) - P\big(\beta_2 \notin CI_\alpha(\hat{\beta}_2)\big) &= 1 - (1-\alpha) - (1-\alpha). \end{aligned}$$

Bonferroni confidence intervals

A *joint confidence set* on level $1 - j(1-\alpha)$ is given by

$$CI_\alpha(\hat{\beta}_1) \times \ldots \times CI_\alpha(\hat{\beta}_j).$$

This marginals of this set is often referred to as *Bonferroni confidence intervals*.

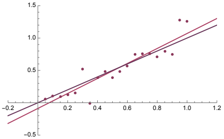
Joint confidence sets

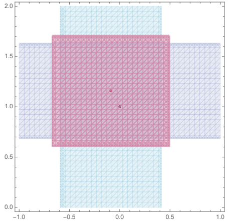
Bonferroni confidence intervals

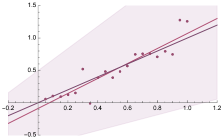
A *joint confidence set* on level $1 - j(1 - \alpha)$ is given by

$$CI_\alpha(\hat{\beta}_1) \times \ldots \times CI_\alpha(\hat{\beta}_j).$$

This marginals of this set is often referred to as *Bonferroni confidence intervals*.







Joint confidence sets

Since

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X^T X)^{-1})$$

implies that

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T X^T X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/(k+1)}{SS_{Res}/(n-k-1)} \sim F_{k+1,n-k-1},$$

we can use the F-distribution to calculate joint confidence regions directly.

Joint confidence sets

An elliptical confidence region

Since

$$P\left(\boldsymbol{\beta} \in \left\{\tilde{\boldsymbol{\beta}} \colon \frac{(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^T X^T X (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})/(k+1)}{SS_{Res}/(n-k-1)} \leq F_{\alpha,k+1,n-k-1}\right\}\right) = \alpha,$$

a confidence set (on confidence level α is given by

$$\left\{\tilde{\boldsymbol{\beta}} \colon \frac{(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^T X^T X (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})/(k+1)}{SS_{Res}/(n-k-1)} \leq F_{\alpha,k+1,n-k-1}\right\}$$

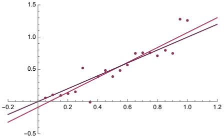
This often a smaller region than the corresponding Bonferroni confidence set, but is hard to understand and visualize if k is large.

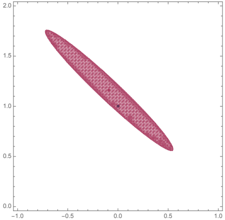
Joint confidence sets

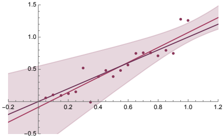
An elliptical confidence region

A confidence set on confidence level α is given by

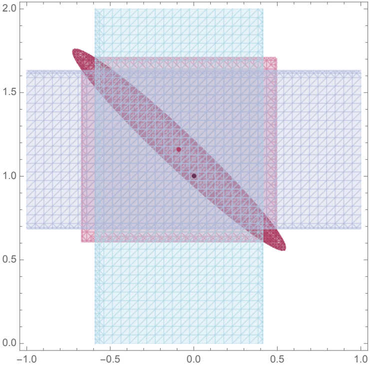
$$\left\{ \tilde{\boldsymbol{\beta}} \colon \frac{(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^T X^T X (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})/(k+1)}{SS_{Res}/(n-k-1)} \leq F_{\alpha,k+1,n-k-1} \right\}$$

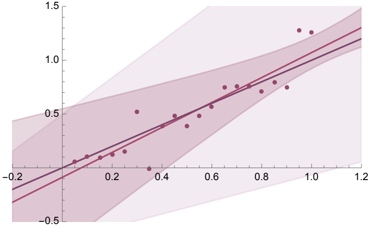






Joint confidence sets





Project 1

- Project 1 is now available on course web page.
- The purpose of this project is to developing a regression model for one out of two given datasets, by applying the ideas we have discussed in class to this dataset.
- Work in groups of 2, and joint the same "group" on project page before handing in a report
- Deadline is same day as exam
- Reported as 1.5 hec, pass/fail, must be passed to pass the course