STK4900/9900 - Lecture 3

Program

1. Data structure and basic questions
2. The multiple linear regression model
3. Categorical predictors
4. Planned experiments and observational studies

- Section 2.5
- Sections 4.1, 4.2 (except 4.2.4), 4.3 (except 4.3.4-5)
- Supplementary material on planned experiments and uncorrelated predictors

Data structure and basic questions

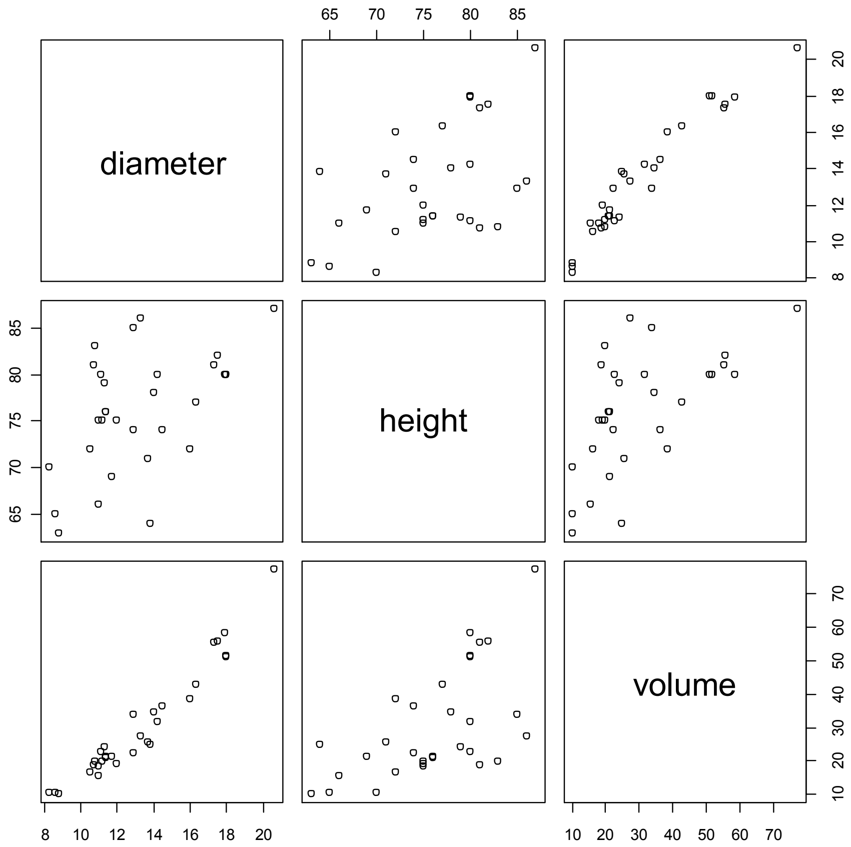
Data have the form:

unit	outcome	predictors (covariates)
1	y_1	$x_{11}, x_{21}, \ldots., x_{p1}$
2	y_2	$x_{12}, x_{22}, \ldots., x_{p2}$
M	M	M
n	y_n	$x_{1n}, x_{2n}, \ldots., x_{pn}$

Objectives:

- Study the effect of one predictor while adjusting for the effects of the other predictors
- Identify important predictors for an outcome
- Predict the outcome for a new unit where only the values of the predictors are available

Example: We have data on the diameter (in inches 4.5 feet above ground level), height (in feet) and volume (in cubic feet) of a sample of 31 trees from a forest in the US. We want to study how the volume of a tree is related to its diameter and height



A scatterplot matrix gives an overview of the data

R-commands:

```
trees=read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v11/trees.txt",header=T)
plot(trees)
```

For the tree data we may fit a simple linear regression model with volume as the outcome using either diameter or height as the covariate

Height as predictor:

R-commands:

```
fit.height=lm(volume~height, data=trees)
summary(fit.height)
```

R-output (edited)

```
              Estimate    Std. Error
(Intercept)   -87.12      29.27
height          1.54       0.38

Multiple R-squared: 0.358
```

Diameter as predictor:

R-commands:

```
fit.diameter=lm(volume~diameter, data=trees)
summary(fit.diameter)
```

R-output (edited)

```
              Estimate    Std. Error
(Intercept)   -36.94       3.37
diameter        5.07       0.25

Multiple R-squared: 0.935
```

Diameter accounts for more of the variability in the volumes than height

But we would like to use both covariates

Multiple linear regression

Data: $(y_i, x_{1i}, x_{2i}, \ldots, x_{pi}) \qquad i = 1, \ldots, n$

y_i = outcome for unit no. i

x_{ji} = predictor (covariate) no. j for unit no. i

Model:

$$\begin{aligned} y_i &= E(y_i \mid \mathbf{x}_i) + \varepsilon_i \\ &= \underbrace{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots. + \beta_p x_{pi}}_{\text{systematic part (linear predictor)}} + \underbrace{\varepsilon_i}_{\text{random part (noise)}} \end{aligned}$$

Here $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})$

The x_{ji}'s are considered to be fixed quantities, and the $\mathbf{e}_i$'s are independent error terms that are assumed to be $N(0, \sigma_\varepsilon^2)$-distributed

Interpretation of regression coefficients

β_j is the change in $E(y \mid \mathbf{x})$ for an increase of one unit in the covariate x_j *holding all other covariates constant*

The effect of each covariate in a multiple linear regression model is adjusted for the effects of all the other covariates in the model

Least squares

Also for multiple linear regression do we use the method of least squares, i.e. the estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots., \hat{\beta}_p$ are obtained as the values of $b_0, b_1, \ldots., b_p$ that minimize the sum of squares

$$\sum_{i=1}^{n} \left(y_i - E(y_i \mid \mathbf{x}_i) \right)^2 = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_{1i} - \ldots. - b_p x_{pi} \right)^2$$

For the tree data we may fit a multiple linear regression model with volume as the outcome and both diameter and height as predictors

R-commands:

```
fit.both=lm(volume~diameter+height, data=trees)
summary(fit.both)
```

R-output (edited):

```
              Estimate    Std. Error    t value      Pr(>|t|)
(Intercept)   -57.99      8.64          -6.71        2.75e-07
diameter        4.71       0.264        17.82        < 2e-16
height          0.339     0.130           2.61         0.015

Residual standard error: 3.88 on 28 degrees of freedom
Multiple R-squared: 0.948
```

Note that the effects of diameter and height are modified when adjusted for the effect of the other

The regression model is *linear in the parameters* β_j

But the model allows for *non-linear effects of the covariates*

For example we may for the tree data include a quadratic term for diameter, i.e. we may consider the model:

$$\text{volume} = \beta_0 + \beta_1 \,\text{diameter} + \beta_2 \,(\text{diameter})^2 + \beta_3 \,\text{height} + \varepsilon$$

R-commands:

```
fit.both=lm(volume~diameter+I(diameter^2)+height, data=trees)
summary(fit.both)
```

R-output (edited):

```
                 Estimate    Std. Error     t value     Pr(>|t|)
(Intercept)      -9.92        10.08          -0.98       0.334
diameter         -2.89          1.310        -2.20       0.036
I(diameter^2)     0.269         0.046         5.85       3.13e-06
height            0.376         0.088         4.27       0.00022

Residual standard error: 2.63 on 27 degrees of freedom
Multiple R-squared: 0.977
```

Transformations

Sometimes it may be useful to perform the regression analysis based of transformations of the outcome and/or the covariates

The formula for the volume of a cone indicates that the volume of a tree is (approximately) proportional to $\text{height} \times (\text{diameter})^2$

This suggest the linear regression model:

$$\log(\text{volume}) = \beta_0 + \beta_1 \log(\text{height}) + \beta_2 \log(\text{diameter}) + \varepsilon$$

R-commands:

```
fit.log=lm(log(volume)~log(height)+log(diameter), data=trees)
summary(fit.log)
```

R-output (edited):

```
                   Estimate     Std. Error     t value     Pr(>|t|)
(Intercept)        -6.63        0.800          -8.29       5.06e-09
log(height)         1.12         0.204          5.46       7.81e-06
log(diameter)       1.98         0.075          26.4       < 2e-16

Residual standard error: 0.117 on 28 degrees of freedom
Multiple R-squared: 0.978
```

Fitted values and residuals

In a similar manner as for simple linear regression, we have:

Fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_p x_{pi}$

Residuals: $r_i = y_i - \hat{y}_i$

Sums of squares

$$TSS = \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2$$ (total sum of squares)

$$MSS = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2$$ (model sum of squares)

$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$ (residual sum of squares)

Decomposition: $TSS = MSS + RSS$

Coefficient of determination

The *coefficient of determination* is given as for simple linear regression:

$$R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS}$$

This may be interpreted as the proportion of the total variability in the outcomes that is accounted for by the predictors

The *multiple correlation coefficient* is given by

$$r = \sqrt{R^2}$$

One may show that this is the Pearson correlation coefficient between the outcomes (y_i) and the fitted values ($\hat{y}_i$)

Residual standard error

Unbiased estimator of σ_{ε}^{2} :

$$\hat{\text{Var}}(\varepsilon) = s_{y|\mathbf{x}}^{2} = \frac{RSS}{n-p-1}$$

$s_{y|\mathbf{x}}$ is the residual standard error

The denominator is

$$n - p - 1 = n - (p+1)$$
$$= \text{number of observation} - \text{number of } \beta_j\text{'s}$$

This is the residual degrees of freedom (df)

Standard error of the estimates

The variance of $\hat{\beta}_j$ is estimated by :

$$\hat{\mathrm{Var}}(\hat{\beta}_j) = \frac{s^2_{y|\mathbf{x}}}{(n-1)\, s^2_{x_j} (1 - r^2_j)} \qquad (*)$$

Here $s^2_{x_j} = \sum_{i=1}^{n} (x_{ji} - \overline{x}_j)^2 / (n-1)$ is the sample variance of the x_{ji}'s and r^2_j is the multiple correlation coefficient for a multiple linear regression where x_j is regressed on the other predictors in the model

Standard error: $se(\hat{\beta}_j) = \sqrt{\hat{\mathrm{Var}}(\hat{\beta}_j)}$

Formula (*) is similar to the one for simple linear regression

The formula shows that $se(\hat{\beta}_j)$ becomes larger if x_j is correlated with the other predictors in the model

Hypothesis tests

Overall test

Consider the null hypothesis that *none* of the predictors have an effect , i.e. the null hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$

Test statistic

$$F = \frac{MSS / p}{RSS / (n - p - 1)}$$

We reject H_0 for large values of F

The test statistic is F-distributed with p and $n - p - 1$ df under H_0

This is a generalization of the F-test for one-way ANOVA

Test for the effect of a single predictor

Quite often it is not very interesting to test the null hypothesis that none of the covariates have an effect

It may be of more interest to test the null hypothesis $H_0 : \beta_j = 0$ versus the alternative $H_A : \beta_j \neq 0$

To this end we may use the test statistic

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

We reject H_0 for large values of $|t|$

Under H_0 the test statistic is t-distributed with $n - p - 1$ df

Note that the t-test is the same as for simple linear regression (i.e. with only on covariate), except for the degrees of freedom

Confidence intervals

95% confidence interval for β_j :

$$\hat{\beta}_j \pm c \cdot se(\hat{\beta}_j)$$

where *c* is the upper 97.5% percentile in the t-distribution
with $n - p - 1$ df

Note that the confidence interval is the same as for simple linear regression, except for the degrees of freedom

Binary categorical predictors

For the tree example both predictors are numerical

In general the predictors in a linear regression model may be numerical and/or categorical

However, special care needs to be exercised when using categorical predictors

For ease of presentation, we start out by considering a single binary predictor, i.e. a categorical predictor with only two levels (female/male, treatment/control, etc)

This corresponds to the situation where we compare two groups

We assume that the data for the two groups are random samples from $N(\mu_1, \sigma_\varepsilon^2)$ and $N(\mu_2, \sigma_\varepsilon^2)$, respectively

We will reformulate the situation as a regression problem

Example:

In Lectures 1 and 2 we considered a study of bone mineral density (in g/cm^2) for rats given isoflavone and for rats in a control group

We then used a t-test and the corresponding confidence interval to study the effect of isoflavone

R-output:

```
Two Sample t-test
data:  treat and cont
t = 2.844,  df = 28,  p-value = 0.0082
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0045    0.0279
sample estimates:
mean of x              mean of y
0.2351                0.2189
```

First reformulation as a regression problem

The observations may be denoted (with $n = n_1 + n_2$):

$$\text{Group 1:} \quad y_1, y_2, \ldots., y_{n_1}$$

$$\text{Group 2:} \quad y_{n_1+1}, y_{n_1+2}, \ldots., y_n$$

We introduce the binary covariate

$$x_i = \begin{cases} 0 & \text{for } i = 1, 2, \ldots, n_1 \qquad \text{(group 1, reference)} \\ 1 & \text{for } i = n_1 + 1, \ldots, n \qquad \text{(group 2)} \end{cases}$$

Then we may write

$$y_i = \mu_1 + (\mu_2 - \mu_1) \cdot x_i + \varepsilon_i \qquad i = 1, 2, \ldots., n$$

where the $\boldsymbol{e}_i$'s are independent error terms that are $N(0, \sigma_\varepsilon^2)$-distributed

This has the form of a simple linear regression model with

$$\beta_0 = \mu_1 = \text{expected outcome in the reference group}$$

$$\beta_1 = \mu_2 - \mu_1 = \text{difference in expected outcome}$$

R-commands for bone density example:

```
bonedensity=
read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v11/bonedensity.txt",header=T)
bonedensity$group=factor(bonedensity$group)
lm.density=lm(density~group,data=bonedensity)
summary(lm.density)
```

R-output:

```
               Estimate      Std. Error      t value      Pr(>|t|)
(Intercept)    0.2189        0.00402         54.34        < 2e-16
group2         0.0162        0.00569         2.844        0.0082

Residual standard error: 0.0156 on 28 degrees of freedom
```

Note that we define "group" to be a categorical covariate (or "factor")

The intercept is the mean in group 1 (the reference group)

The estimate for group2 is the difference between the means in the two groups

The t-value (2.844) and p-value (0.0082) equals the t-test

An alternative reformulation as a regression problem

We may write the model as

$$y_i = \begin{cases} \mu_1 + \varepsilon_i & \text{for } i \text{ in group 1} \\ \mu_2 + \varepsilon_i & \text{for } i \text{ in group 2} \end{cases}$$

We introduce the "grand mean" $\bar{\mu} = \dfrac{\mu_1 + \mu_2}{2}$

Then the model may be reformulated as

$$y_i = \begin{cases} \bar{\mu} + (\mu_1 - \bar{\mu}) + \varepsilon_i & \text{for } i \text{ in group 1} \\ \bar{\mu} + (\mu_2 - \bar{\mu}) + \varepsilon_i & \text{for } i \text{ in group 2} \end{cases}$$

We now introduce the covariate

$$x_i = \begin{cases} 1 \ \text{ for } i = 1,2,...,n_1 & \text{(group 1)} \\ -1 \ \text{ for } i = n_1 + 1,...,n & \text{(group 2)} \end{cases}$$

Then the model may be written

$$y_i = \bar{\mu} + (\mu_1 - \bar{\mu}) \cdot x_i + \varepsilon_i \qquad i = 1,2,....,n$$

This has the form of a simple linear regression model with

$$\beta_0 = \bar{\mu} = \text{grand mean}$$

$$\beta_1 = \mu_1 - \bar{\mu} = \text{deviation from grand mean in group 1}$$

R-commands for bone density example:

```
options(contrasts=c("contr.sum","contr.poly"))
 lm.density.sum=lm(density~group,data=bonedensity)
summary(lm.density.sum)
```

R-output:

```
                Estimate      Std. Error     t value    Pr(>|t|)
(Intercept)     0.2270        0.00285        79.70      < 2e-16
group1          -0.0081       0.00285        -2.844     0.0082

Residual standard error: 0.0156 on 28 degrees of freedom
```

We get the formulation on the previous slide by using "sum-contrast"

The formulation on slide 19 is denoted "treatment-contrast" and is specified by the command: options(contrasts=c("contr.treatment","contr.poly"))

The intercept estimate is the "grand mean"

The group1 estimate is the difference between the mean in group1 and the "grand mean"

Treatment-contrast is default in R and we will stick to it in the following. But note that other software may use sum-contrast as default

Multilevel categorical predictors

We then consider a categorical predictor with K levels

This corresponds to the situation where we compare K groups

We denote the observations for all groups combined by

$$y_1, y_2, \ldots, y_n$$

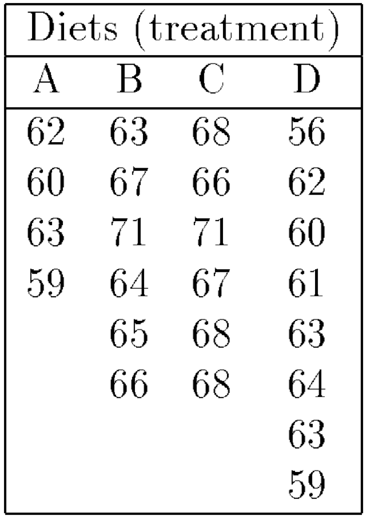
Here the first n_1 observations are from group 1, the next n_2 observations are from group 2, etc.

We assume that all observations are independent and that the observations from group k are $N(\mu_k, \sigma_\varepsilon^2)$-distributed

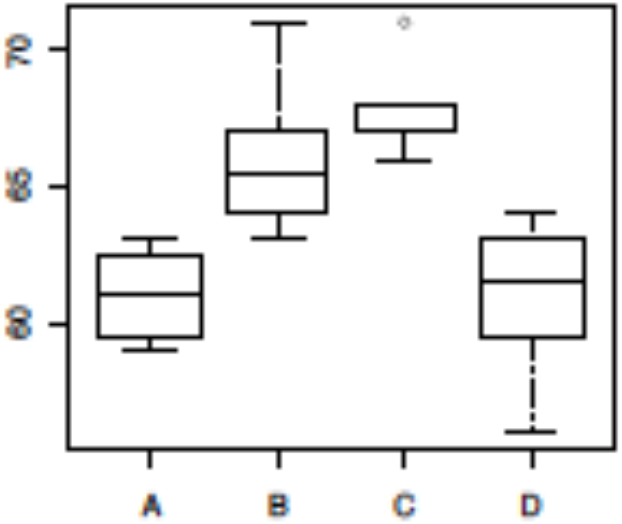
We will reformulate the situation as a regression problem

Example:

In Lecture 2 we considered an experiment were 24 rats were randomly allocated to four different diets, and the blood coagulation time (in seconds) were measured for each animal



Diets (treatment)			
A	B	C	D
62	63	68	56
60	67	66	62
63	71	71	60
59	64	67	61
	65	68	63
	66	68	64
			63
			59



We will study the effect of diet on the blood coagulation time

Reformulation as a regression problem

With K groups we need to introduce $K - 1$ predictor variables

$$x_{1i} = \begin{cases} 1 & \text{for } i \text{ in group 2} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{2i} = \begin{cases} 1 & \text{for } i \text{ in group 3} \\ 0 & \text{otherwise} \end{cases}$$

…

$$x_{K-1,i} = \begin{cases} 1 & \text{for } i \text{ in group } K \\ 0 & \text{otherwise} \end{cases}$$

Note that all $x_{ji} = 0$ for i in group for 1, which is the reference group

Then we may write

$$y_i = \mu_1 + (\mu_2 - \mu_1) \cdot x_{1i} + (\mu_3 - \mu_1) \cdot x_{2i} + \ldots + (\mu_K - \mu_1) \cdot x_{K-1,i} + \varepsilon_i$$

where the $\mathbf{e}_i$'s are independent error terms that are $N(0, \sigma_\varepsilon^2)$-distributed

This has the form of a multiple linear regression model with

$\beta_0 = \mu_1$ (expected outcome in the reference group)

$\beta_j = \mu_{j+1} - \mu_1$ (difference in expected outcome between group $j+1$ and the reference)

R-commands for blood coagulation example:

```
rats=read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v11/rats.txt",header=T)
rats$diet=factor(rats$diet)
fit.rats=lm(time~diet,data=rats)
summary(fit.rats)
anova(fit.rats)
```

R-output (edited):

```
              Estimate     Std. Error     t value      Pr(>|t|)
(Intercept)   61.00        1.18           51.55        < 2e-16
diet2          5.00        1.53            3.27        0.0038
diet3          7.00        1.53            4.58        0.0002
diet4          0.00        1.45            0.00        1.0000

Residual standard error: 2.366 on 20 degrees of freedom
Multiple R-squared: 0.671,    Adjusted R-squared: 0.621
F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.66e-05

Analysis of Variance Table
             Df     Sum Sq    Mean Sq     F value     Pr(>F)
diet          3     228       76.0        13.57       4.66e-05
Residuals    20     112        5.6
```

We get a more detailed picture than in Lecture 2

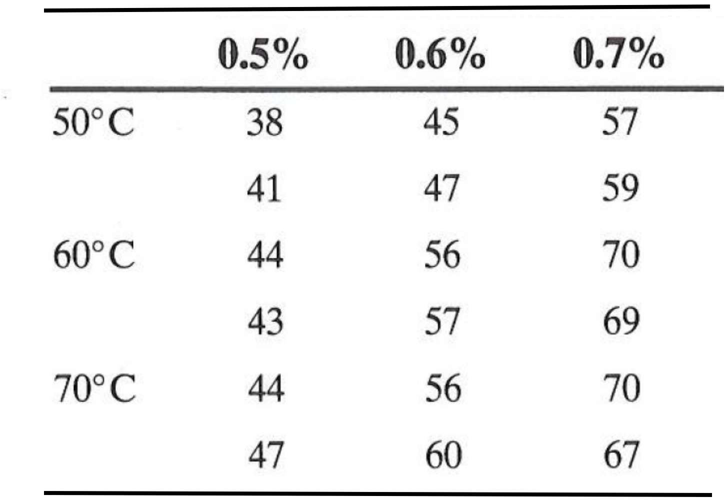
Planned experiments and observational studies

The methods for multiple linear regression are valid both for *planned experiments* (where the values of the predictors are under the control of the experimenter) and *observational studies* (where we condition on the observed values of the predictors)

But the interpretation of the results is more complicated for observational studies, as we will now discuss

We start out by considering the situation with two covariates

Planned experiment

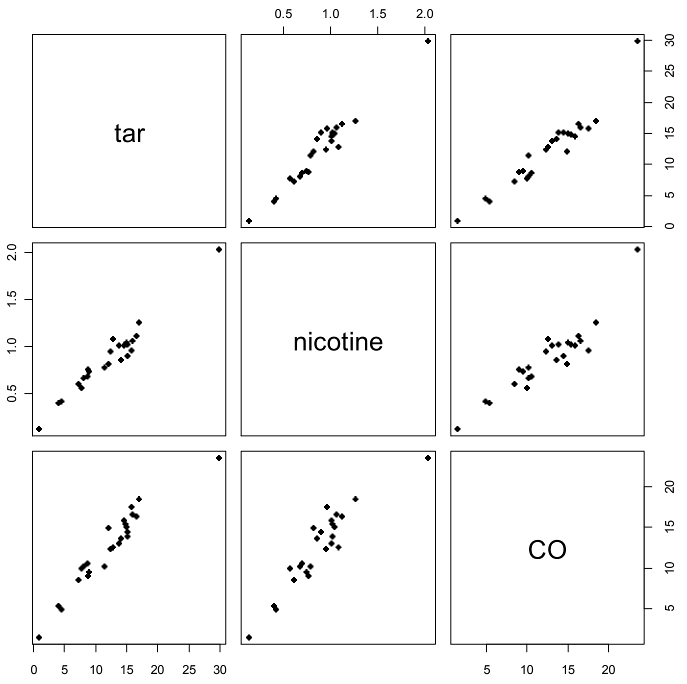


	0.5%	0.6%	0.7%
50°C	38	45	57
	41	47	59
60°C	44	56	70
	43	57	69
70°C	44	56	70
	47	60	67

An experiment has been conducted to study how the extraction rate of a certain polymer depend on temperature and the amount of catalyst used. The extraction rate was recorded twice for each of three levels of temperatures and three levels of the catalyst

Observational study

For 25 brands of cigarettes the content of tar, nicotine, and carbon monoxide have been measured (details)



We want to study how the amount of CO emitted from the cigarette smoke depends on the content of tar and nicotine

Polymer example

Note that:

- The estimates are the same in the model with two predictors as they are in the simple linear regression models with only one predictor at a time
- R^2 for the model with two predictors is the sum of R^2 -values for the two one-predictor models
- The reason is that the two predictors are uncorrelated

```
              Estimate    Std. Error
(Intercept)   25.39         17.74
temp           0.475         0.293

Residual standard error: 10.15
Multiple R-squared: 0.141
```

```
              Estimate    Std. Error
(Intercept)   -13.61         8.73
cat           112.50        14.42

Residual standard error: 4.99
Multiple R-squared: 0.792
```

```
              Estimate    Std. Error
(Intercept)   -42.11      7.20
temp            0.475     0.084
cat           112.50      8.44

Residual standard error: 2.92
Multiple R-squared: 0.933
```

```
polymer=read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v11/
polymer.txt",header=T)
```

Cigarette example

Note that:

- When only nicotine is used as predictor, it has large effect on CO
- The effect of nicotine disappears when adjusted for the effect of tar
- The reason is that the two predictors are strongly correlated

	Estimate	Std. Error
(Intercept)	1.66	0.99
nicotine	12.40	1.05

Residual standard error: 1.83
Multiple R-squared: 0.857

	Estimate	Std. Error
(Intercept)	2.74	0.68
tar	0.80	0.05

Residual standard error: 1.40
Multiple R-squared: 0.917

	Estimate	Std. Error
(Intercept)	3.09	0.84
nicotine	-2.65	3.79
tar	0.96	0.24

Residual standard error: 1.41
Multiple R-squared: 0.919

```
cigarettes=read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v11/cigarettes.txt",
header=T)
```

Planned experiments and uncorrelated predictors

For *planned experiments* one may choose the values of the predictors so that they are uncorrelated. This is also called orthogonality

Orthogonality is a useful property:

- R^2 is given as $R^2 = r_1^2 + r_2^2 + ... + r_p^2$ where r_j is the Pearson correlation between predictor j and the outcome
- The estimates $\hat{\beta}_j$ are the same as obtained by fitting a simple linear regression for each covariate.
- The standard errors $se(\hat{\beta}_j)$ are typically smaller (cf. slide 13)
- Therefore, shorter confidence intervals and more precise predictions may be obtained

Observational studies and correlated predictors

For *observational studies* the predictors will be correlated

Then, as illustrated above for two covariates, the effect of one covariate may change when other covariates are included in the model

Therefore special care has to be exercised when analysing data from observational studies

We will have a closer look at this in Lecture 4

Randomization

Another difference between planned and observational studies is that for planned studies we are able to randomize which study subjects receive different treatments.

For instance, comparing a proposed treatment with a placebo treatment we at random select n/2 individuals that get the proposed treatment and the remaining n/2 get the placebo.

This way there will be no systematic initial difference between the two groups and a difference in outcomes between treatment groups can, due to the randomization, be attributed to a causal effect.

Spurious effects

In an observational study randomization will not be possible and observed differences between groups can be spurious, i.e. due to initial differences between the groups.

These initial differences will be then be correlated with the groups that we want to compare. Thus this is related to the discussion of correlated covariates and confounding in Lecture 4.