Lecture 3 – Program

1. Data structure and basic questions

2. Simple linear regression

3. Multiple linear regression

4. Least squares estimation

5. Hypothesis testing and confidence intervals

6. Proportion of explained variation (R^2)

7. Confounding

Data structure and basic questions

Data have the form:

unit	response	covariates
1	y_1	$x_{11} \cdots x_{1p}$
2	y_2	$x_{21} \cdots x_{2p}$
.	.	$\cdots$
.	.	$\cdots$
.	.	$\cdots$
n	y_n	$x_{n1} \cdots x_{np}$

- Objective: Explain how the response y is related to the covariates $x_1, \cdots, x_p$
  Sometimes predict new units where only the covariates are available.

- The covariates are treated as fixed

- Model: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$,
  where $\varepsilon_i \sim N(0, \sigma^2)$ is an error term (noise)

- Decomposition:
  Response = Systematic part + random part

Simple linear regression

Data $(x_i, y_i), \quad i = 1, ..., n$

$y_i =$ response
(or dependent variable)

$x_i =$ covariate
(or explanatory variable)
(or independent variable)

Model:

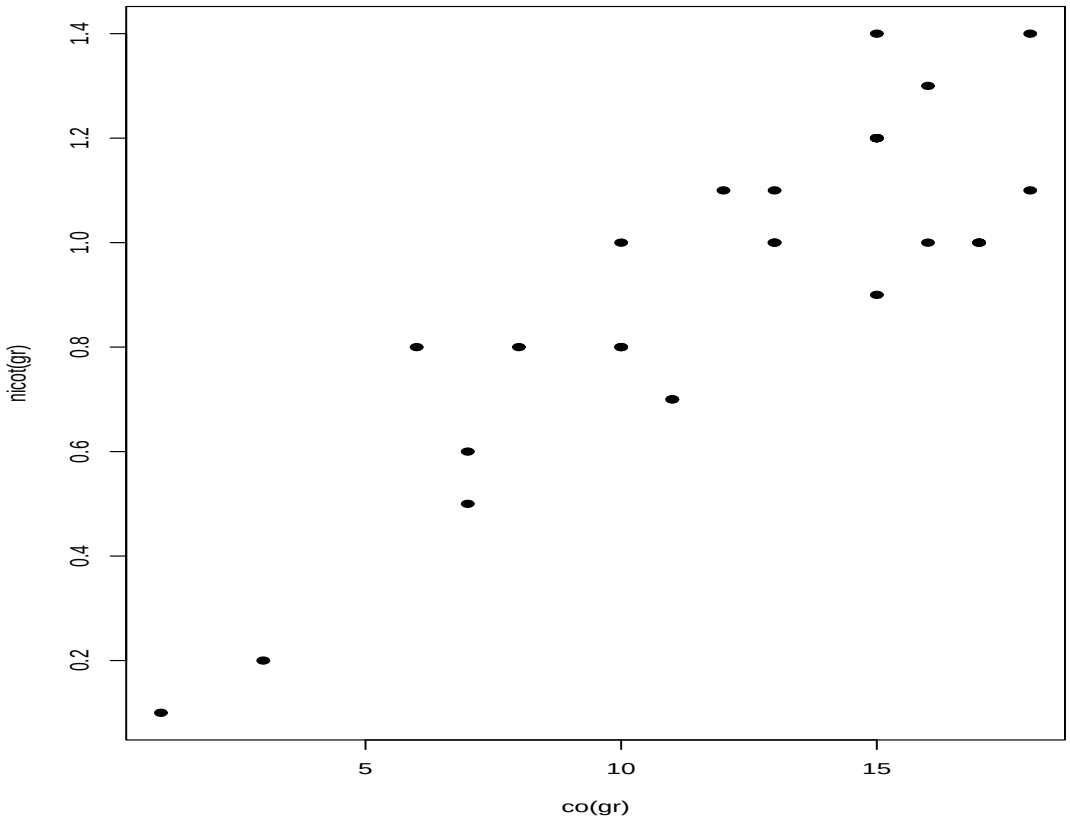
$$y_i = a + bx_i + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$; $i = 1, ..., n$; are independent error terms (noise)

Example: Nicotine in filter cigarettes

Response = amount of nicotine
Covariate = amount of CO

Amount of nicotine versus amount of CO :



Simple linear regression in R

```
mod1<-lm(nicot~co, data=sigarett)
summary(mod1)
```

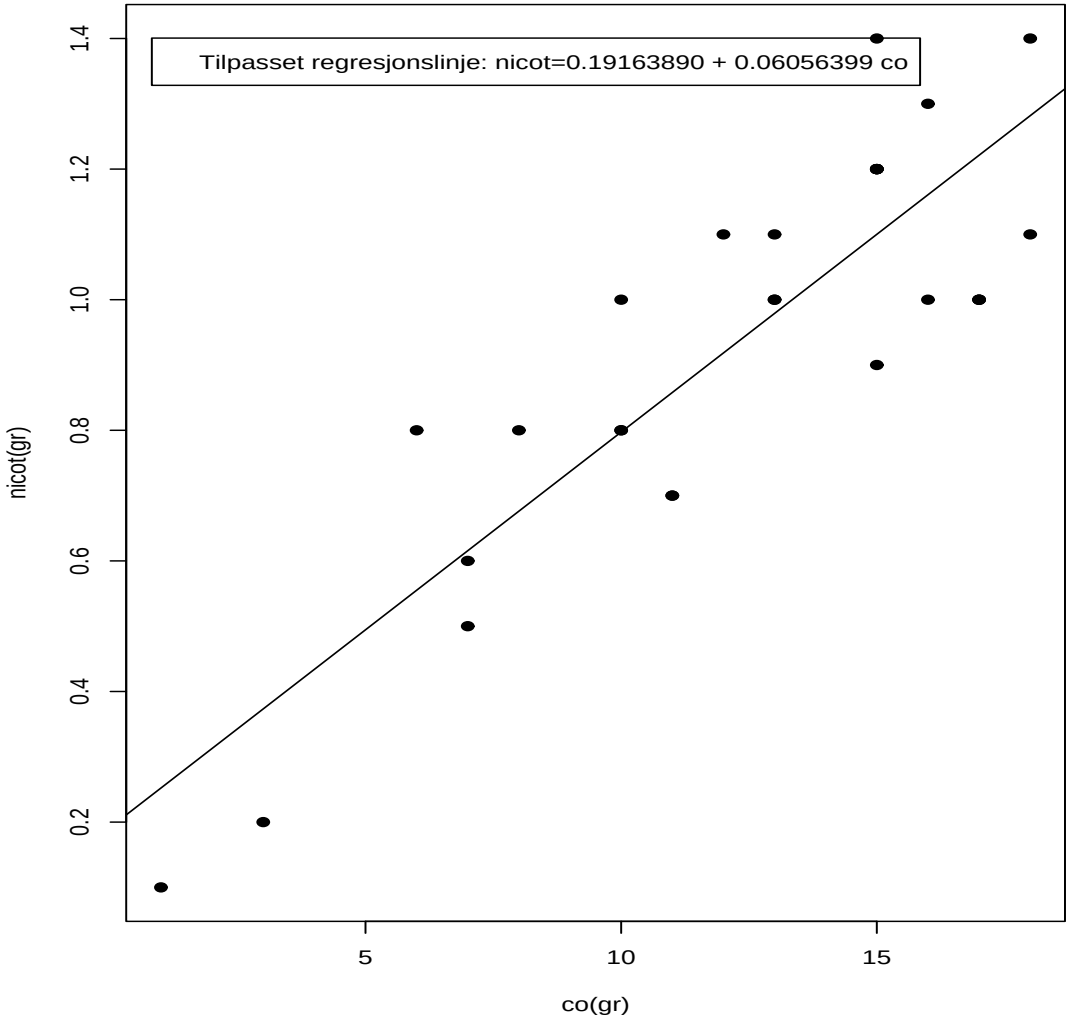
```
Residuals:
     Min       1Q   Median       3Q      Max
-0.22123 -0.15784  0.02103  0.11821  0.29990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.191639   0.089499   2.141   0.0414
co          0.060564   0.006814   8.888 1.67e-09

Residual standard error: 0.1611 on 27 degrees of freedom
Multiple R-Squared: 0.7453,    Adjusted R-squared: 0.7358
F-statistic:    79 on 1 and 27 DF,  p-value: 1.671e-09
```

(Slightly edited output)

Amount nicotine versus CO with fitted line



The computer program does not distinguish between planned experiments and observation studies.

But the difference is essential when interpreting the results.

Multiple linear regression

Data $(y_i, x_{i1}, \ldots, x_{ip})$ $\quad i = 1, ..., n$

y_i = response
x_{ip} = covariate no. j $\quad j = 1, \ldots, p$

Model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

where

$$\beta_j = \text{ regression coefficient for } x_{ij}$$

The ε_i's are independent and normally distributed error terms (noise) that have expectation zero and the *same* variance σ^2

Types of covariates and models

We assume that the response is a quantitative (numeric) variable.

Three possibilities for the covariates:

- Quantitative covariates
- Qualitative (categorical) covariates
- Mixture of quantitative and qualitative covariates

Examples:

Nicotine in filter cigarettes

$$\begin{aligned} y_i &= \text{ amount of nicotine} \\ x_{i1} &= \text{ amount of CO} \\ x_{i2} &= \text{ amount of tar} \end{aligned}$$

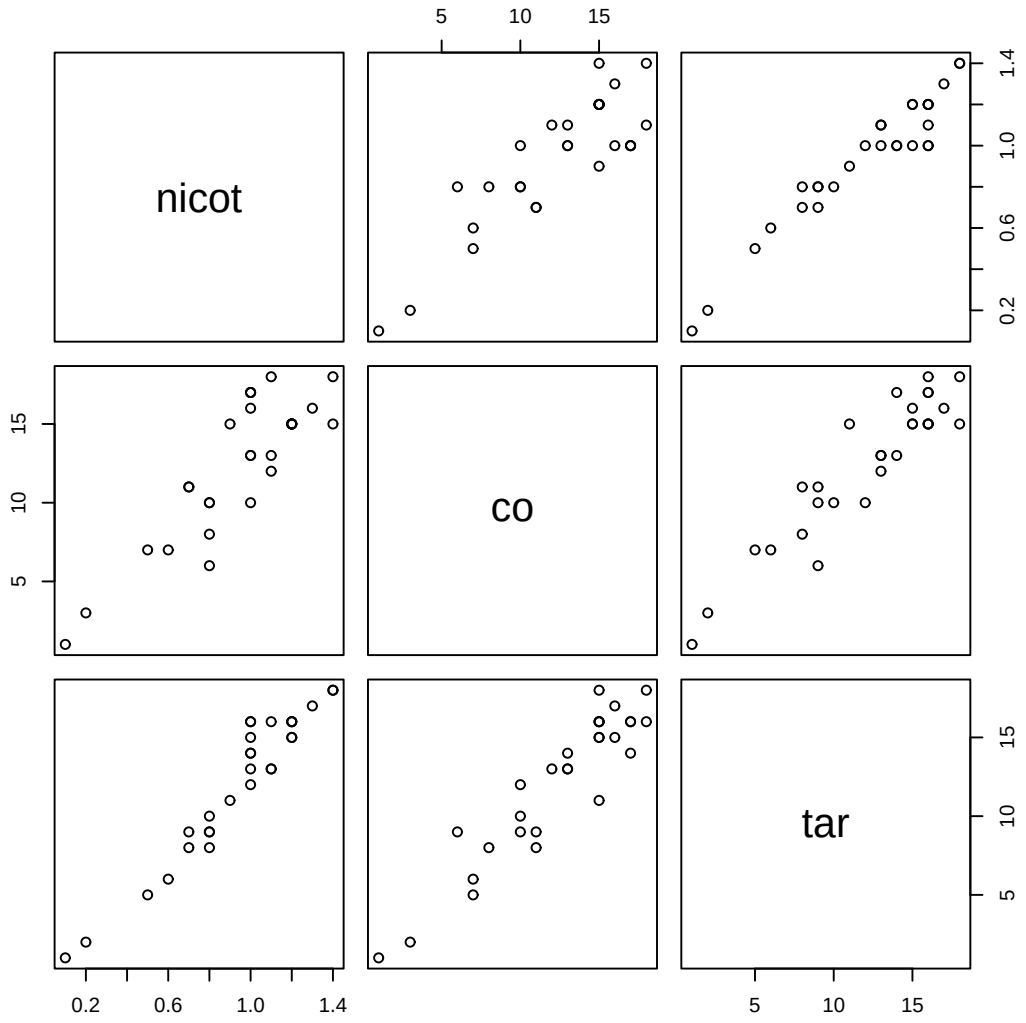
Coagulation of blood for rats

$$\begin{aligned} y_i &= \text{ time to coagulation} \\ x_{i1} &= \text{ 1 if diet 1; 0 otherwise} \\ x_{i2} &= \text{ 1 if diet 2; 0 otherwise} \\ x_{i3} &= \text{ 1 if diet 3; 0 otherwise} \\ x_{i4} &= \text{ 1 if diet 4; 0 otherwise} \end{aligned}$$

Lung function for children

$$\begin{aligned} y_i &= \text{ FEV1} \\ x_{i1} &= \text{ height in cm} \\ x_{i2} &= \text{ weight in kg} \\ x_{i3} &= \text{ age in years} \\ x_{i4} &= \text{ 1 if boy; 0 if girl} \end{aligned}$$

Scatter plot matrix of amount of nicotine, CO and tar



Multiple linear regression in R

```
mod2<-lm(nicot~co+tar, data=sigarett)
summary(mod2)
```

```
Residuals:
      Min         1Q    Median         3Q        Max
-0.174122 -0.013595 -0.003180  0.068722  0.112795

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.181645   0.046683   3.891 0.000621
co          -0.018642   0.009910  -1.881 0.071201
tar          0.081837   0.009559   8.562 4.84e-09

Residual standard error: 0.08399 on 26 degrees of freedom
Multiple R-Squared: 0.9333,     Adjusted R-squared: 0.9282
F-statistic: 181.9 on 2 and 26 DF,  p-value: 5.165e-16
```

(Slightly edited output)

Why multiple regression?

Typically several covariates need to be taken into account when trying to explain the variation of the response

Two objectives that may be conflicting:

- Explain the influence of a covariate on response

- Obtain a prediction rule for the response as a function of the covariates

Analysis of variance:
Only qualitative covariates
(more in Lecture 6)

Analysis of covariance:
Both qualitative and quantitative variables

Coding of qualitative covariates

Suppose a qualitative (categorical) covariate has r values.

The technical term for this type of covariate is **factor** and its possible values are called **factor levels.**

To include such a covariate in a multiple regression model it levels have to be coded (i.e. given numeric values), and that can be done in several ways.

First method

A) Two levels, e.g. female and male.
We use one dummy-variable:

$$x_{i1} = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$$

B) Three levels, e.g. placebo, treatment 1, and treatment 2. We use two dummy-variables:

$$x_{i1} = \begin{cases} 1 & \text{treatment 1} \\ 0 & \text{else} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{treatment 2} \\ 0 & \text{else} \end{cases}$$

We use $r-1$ dummy-variables to code a factor with r levels.

All the dummy variables will be equal to 0 for the *reference* level (female, placebo)

For the first example above we may use the model:

$$E(y_i) = \beta_0 + \beta_1 x_{i1}$$

Then

$$E(y_i) = \begin{cases} \beta_0 + \beta_1 & \text{for males} \\ \beta_0 & \text{for females} \end{cases}$$

Thus β_0 is the expected value of the response for females (reference), and β_1 as the difference of the expected value of the response for males and females.

Second method

In some applications it is more natural to treat all levels symmetrically, and not single out one as a reference level.

Consider as an example a situation where a company wants to compare the sales, y, of a product in four regions.

Let us introduce a dummy variable for each region so that ($i = 1, \ldots, 4$)

$$x_i = \begin{cases} 1 & \text{if region } i \\ 0 & \text{else} \end{cases}$$

Then for $i = 1, \ldots, 4$:

$$E(y_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_4 x_4$$

There is something suspect!
Considered as a system of linear equations, there are 4 equations, and 5 unknowns.

This can be solved by introducing a restriction on the β_j's. A common restriction is:

$$\beta_1 + \beta_2 + \beta_3 + \beta_4 = 0$$

Now $\beta_1, \cdots, \beta_4$ can be interpreted as measuring the expected sales above the country average, which is measured by β_0

Coding in R

R has options for choosing which method to use (in addition the covariate must be defined as a factor).

The command

```
options(contrasts=c("contr.treatment","contr.poly"))
```

defines the lowest level as a reference category (this is the default in R, but not in Splus)

The command

```
options(contrasts=c("contr.sum","contr.poly"))
```

specifies the sum constraint.

Estimation: least squares

Consider the sum of squared differences between the responses y_i and their expected values $\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$:

$$SS(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip} \right)^2$$

Given the data, the sum of squares $SS(\beta_0, \beta_1, \ldots, \beta_p)$ is a second order polynomial in the parameters $\beta_0, \beta_1, \ldots, \beta_p$

The values of values $\beta_0, \beta_1, \ldots, \beta_p$ that minimize the quadratic function $SS(\beta_0, \beta_1, \ldots, \beta_p)$ are the **least squares estimators** $\widehat{\beta}_0, \ldots, \widehat{\beta}_p$

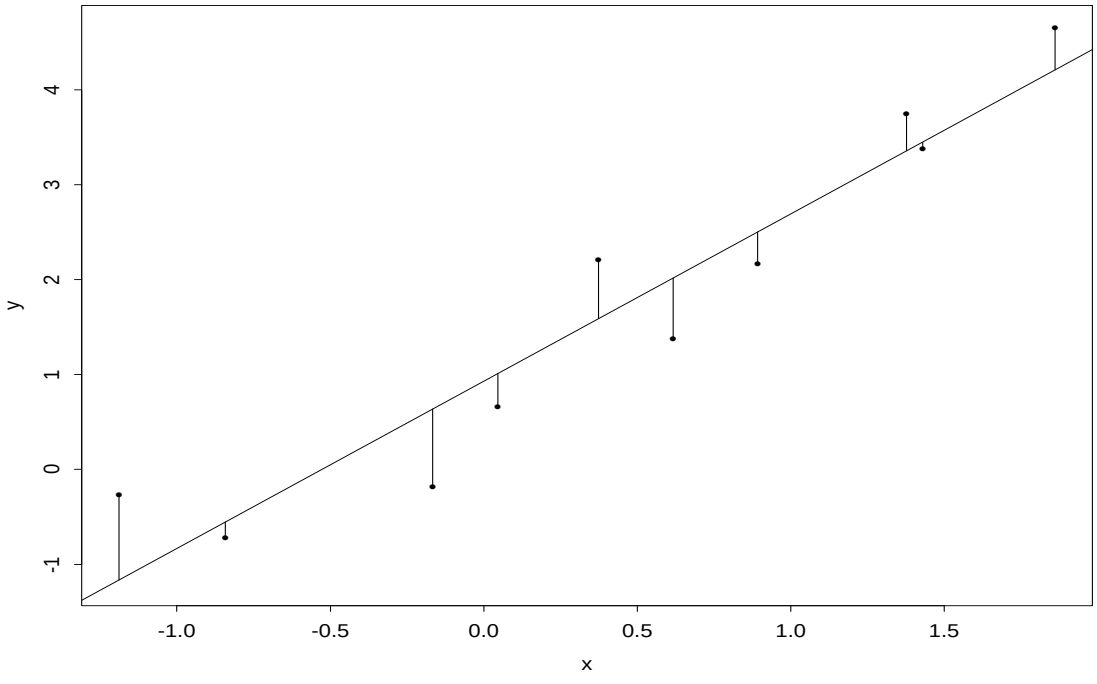
The minimum is found by setting the derivatives of $SS(\beta_0, \beta_1, \ldots, \beta_p)$ with respect to $\beta_0, \beta_1, \ldots, \beta_p$ equal to zero. This gives a linear system of equations that may be solved explicitly using matrix algebra.

Simulated example

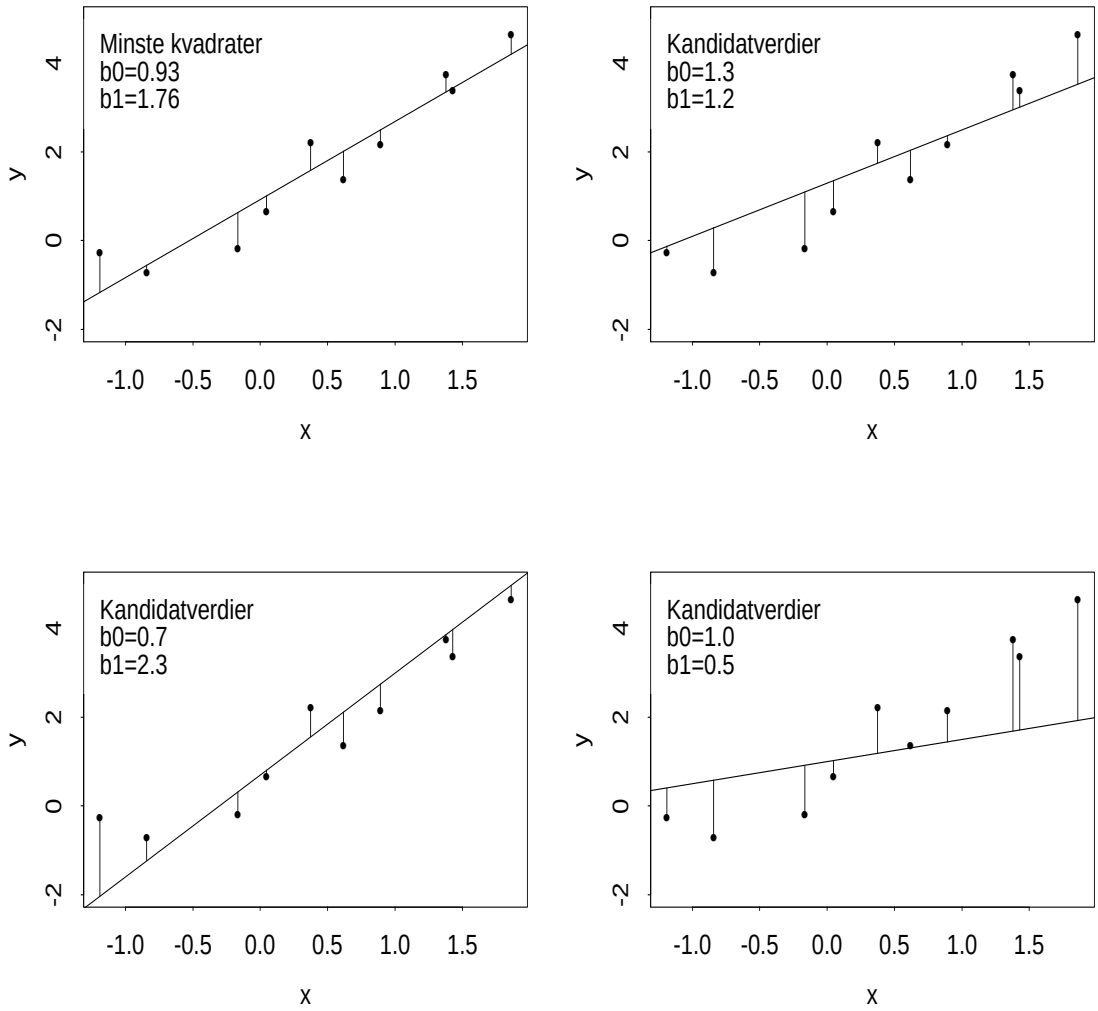
Generate 10 covariate values $x_i \sim \mathrm{N}(0, 1)$ and errors $\varepsilon_i \sim \mathrm{N}(0, 1)$

Obtain the responses by

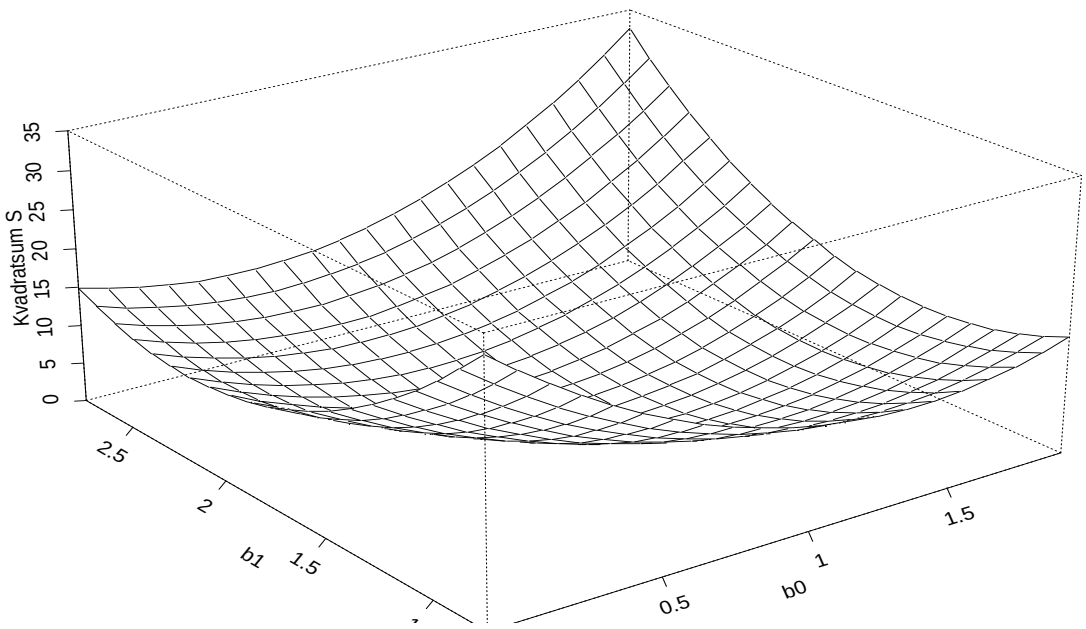
$$y_i = 1 + 2x_i + \varepsilon_i$$



Comparison with other values of β_0 and β_1



Plot of $SS(\beta_0, \beta_1)$ versus β_0 and β_1 :



Important quantities:

Fitted (or predicted) values:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip}$$

Residuals:

$$\begin{aligned} \widehat{e}_i &= y_i - \widehat{y}_i \\ &= y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \cdots - \widehat{\beta}_p x_{ip} \end{aligned}$$

Estimation of residual variance: $\sigma^2 = \text{Var}(\varepsilon_i)$

The minimum sum of squares, or the squared sum of residuals, is denoted by

$$SS_{unexpl} = SS_{res} = SS(\widehat{\beta}_0, \ldots, \widehat{\beta}_p) = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \cdots - \widehat{\beta}_p x_{ip})^2$$

The common estimate of σ^2 is

$$\widehat{\sigma}^2 = \frac{SS_{unexpl}}{n - p - 1}$$

The denominator is

$$\begin{aligned} n - p - 1 &= n - (p + 1) \\ &= \text{Number of observations} \\ &\quad - \text{Number of coefficients} \end{aligned}$$

This is the residual degrees of freedom (df).

Simple linear regression

Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Let $\bar{x}$ and $\bar{y}$ be the means of the x_i's and the y_i's. Then

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

The estimate of σ^2 is $\hat{\sigma}^2 = SS_{unexpl}/(n-2)$ wIth

$$SS_{unexpl} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Also

$$se(\hat{\beta}_1) = \sqrt{Var(\hat{\beta}_1)} = \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Two covariates (B& S p. 30)

Model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$

Let $\bar{x}_1$ and $\bar{x}_2$ be the means of the covariates, and let

$$v_j = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

be the empirical variances ($j = 1, 2$).

Introduce the empirical covariance

$$v_{12} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

and the correlation

$$\tau = \tau_{12} = \frac{v_{12}}{\sqrt{v_1 \, v_2}}$$

Then

$$se(\hat{\beta}_j) = \frac{\sigma}{\sqrt{(1-\tau^2)\,(n-1)\,v_j}}$$

$$\text{corr}(\hat{\beta}_2, \hat{\beta}_2) = -\tau$$

Inference in multiple regression

Test for effect of at least one covariate,
i.e. test of the null hypothesis

$$\mathsf{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

Define the sum of squares

Total variation:	$SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
Explained variation:	$SS_{expl} = SS_{tot} - SS_{unexpl}$

A test for H_0 is based on

$$F = \frac{SS_{expl}\,/p}{\widehat{\sigma}^2} = \frac{SS_{expl}\,/p}{SS_{unexpl}\,/(n-p-1)}$$

which is F-distributed with p and $n-p-1$
degrees of freedom under H_0.

ANOVA table for multiple regression

Source	SS	df	MS	F
Regression	SS_{expl}	p	$MS_{expl} = \frac{S_{expl}}{p}$	$F = \frac{MS_{expl}}{MS_{unexpl}}$
Residual	SS_{unexpl}	$n-p-1$	$MS_{unexpl} = \frac{S_{unexpl}}{n-p-1}$	
Total	SS_{tot}	$n-1$		

A large values of F indicates that H_0 should be rejected, i.e. we conclude that at least one of the β_j, $j = 1, \ldots, p$ is different from 0.

How large F should be for rejection of H_0 is determined from the F-distribution.
The P-value is given by

$$P = P(F_{p,n-k} > \text{observed value of } F)$$

Note that the F-statistic can be considered as the ratio of two estimators of the variance σ^2.

The denominator is the unbiased estimator, $\widehat{\sigma}^2$.
The numerator $SS_{expl}\,/p$ has expectation equal to σ^2 under H_0 and larger than σ^2 under the alternative.

Test for $H_0 : \beta_j = 0$

$\beta_j \neq 0$ means that covariate j has an influence on the response.

The test statistic is

$$t = \frac{\widehat{\beta}_j}{\widehat{se}_j}$$

where $\widehat{se}_j$ is the estimated standard error of $\widehat{\beta}_j$.

We may show that $\widehat{se}_j = t_j \, \widehat{\sigma}$, where t_j depends only on the covariates.

Recall that $\widehat{\sigma}^2$ is the sum of squares of the residuals divided by $n - p - 1$.

Under H_0 : $t \sim t_{n-p-1}$

H_0 is rejected for for large absolute values of t

Confidence interval for β_j

The 95% confidence interval has the form

$$\widehat{\beta}_j \pm t_{0.975}\, \widehat{se}_j$$

where $t_{0.975}$ is the 97.5%-percentile of the $t-$distribution with $n - p - 1$ df

$t_{0.975}$ can be found in R by `qt(0.975,n-p-1)`.

The t-distribution is close to N(0, 1) when the degrees of freedom is not too small.
Therefore, the 95% confidence interval is approximately

$$\widehat{\beta}_1 \pm 2\widehat{se}_j.$$

Confidence intervals with other confidence coefficients $1 - \alpha$, are obtained by using $t_{1-\alpha/2}$ instead of $t_{0.975}$.

Test for $H_0 : \beta_j = \beta_{j0}$

β_{j0} is any specified value (not necessarily 0).

Two possible procedures:

1. Use the confidence interval for β_j:
   If β_{j0} does not belong to the interval,
   reject H_0 at the 5% level

2. Use the test statistic

$$t = \frac{\widehat{\beta}_j - \beta_{j0}}{\widehat{se}_j}$$

   which is $t-$distributed $n-p-1$ df under H_0.
   P-value may be computed.

Multiple correlation and R_p^2

$$R_p^2 = \frac{SS_{expl}}{SS_{tot}} = 1 - \frac{SS_{unexpl}}{SS_{tot}}$$

If $p = 1$ (one covariate) we have $R_1^2 = r^2$
i.e. the squared Pearson correlation coefficient between covariate and response.

R_p^2 is the square of the **multiple correlation coefficient**

In multiple regression models R_p^2 is a measure of the proportional reduction of the total sum of squares by using the model.

SS_{unexpl} will decrease as we include more covariates in the regression.
Therefore R_p^2 will increase with p.

When all the covariates are uncorrelated

$$R_p^2 = r_1^2 + r_2^2 + \cdots + r_p^2$$

where r_j is the Pearson correlation coefficient between covariate j and the response.

Uncorrelated covariates

For planned experiments one can choose the values of the covariates so that they are uncorrelated. This is also called orthogonality.

Orthogonality is a useful property:

- R^2 can be decomposed.
- The estimates $\widehat{\beta}_j$ are the same as obtained by fitting a simple linear regression for each covariate.
- The part of the estimate for $\widehat{se}_j = t_j \, \widehat{\sigma}$ depending on the covariates is the same as in a simple linear regression.
- Estimated residual variance $\widehat{\sigma}^2$ is typically smaller. Hence, $\widehat{se}_j$ is also smaller.
- Therefore, shorter confidence intervals are often obtained.
- More precise predictions.

Correlated covariates

This is the typical case in *observational studies*.

Confounding exists if different interpretations of the relationship between the response and a covariate of primary interest change according to whether other covariates are are included or not.

Example

- Two covariates: covariate 2 has an effect, covariate 1 has little or no effect.
- The covariates are positively correlated.
- A simple linear regression on covariate 1 shows significant effect.
- When covariate 2 is included, there is little or no significant effect of covariate 1.

Nicotin, CO and tar revisited

Simple linear regression:

$$\text{nicot} = 0.192 + 0.060\,\text{CO}$$

P-value of coefficient of CO: < 0.0001

Multiple regression:

$$\text{nicot} = 0.182 - 0.019\,\text{CO} + 0.082\,\text{tar}$$

P-value and coefficient of CO: 0.071
P-value and coefficient of tar: < 0.0001

Formal argument:Two covariates

True model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{2i} + \varepsilon_i$

Use only covariate 1: $y_i = a + b x_{i1} + \varepsilon_i$

Then

$$\widehat{b} = \widehat{\beta}_1 + \widehat{\beta}_2 \, \tau \sqrt{\frac{v_2}{v_1}}$$

where τ is the empirical correlation between the covariates, and v_j is the empirical variance of covariate j (cf. R-exercise 3).

The estimated effect in the model with only covariate 1 is therefore a combination of

- the real effect of covariate 1
- the effect of covariate 2 via τ

Lesson:

In an observational study there is always a possibility that there are confounding variables that are not included in the model, so called lurking variables.

In a multiple regression one must always take account of covariates that may have an effect. Ignoring them means that the explanation of the relationship between the response and covariates of primary interest may be due to confounding.

It is also a useful information that a potential covariate turns out to have no significant effect.

Problems in multiple regression

Assume that there are two covariates that satisfy

$$x_{i2} = \gamma_0 + \gamma_1 x_{i1}$$

Thus there is an **exact** linear relationship between x_{i1} and x_{i2}, and the two covariates have correlation 1.

Including both in a regression model is *not* meaningful, and the estimates cannot be computed.

Usually a computer program, by default, will ignore one of the variables, but from the output it is not always obvious which one.

Problems in multiple regression, cont.

Usually in a multiple regression, we do not have correlation *exactly* equal to 1 between two or more covariates (or linear combinations of covariates).

But we may have correlations that are close to 1. This is called **almost** or **near co-linearity.**

Then one problem is that the estimates are correlated, so the interpretation is more difficult.

Example, two covariates

Remember

$$se_j = \frac{\sigma}{\sqrt{(1-\tau^2)v_j(n-1)}}$$

and

$$\text{corr}(\hat{\beta}_1, \hat{\beta}_2) = -\tau.$$

where τ is the empirical correlation between the covariates.

We see that when $\tau \neq 0$ the standard errors are larger, and also that the estimates are correlated. When the covariates are positively correlated, the estimates are negatively correlated. Hence, there is a trade-off between the coefficients how to express the variation in the covariates on the variation of the response.

Simulation as in B&S, p. 33

$\beta_1 = \beta_2 = 1$, $\tau = 0$ and0.9