Module 3: Linear Regression

TMA4268 Statistical Learning V2020

Stefanie Muff, Department of Mathematical Sciences, NTNU

January xx, 2020

Introduction

Learning material for this module

- James et al (2013): An Introduction to Statistical Learning. Chapter 3.

We need more statistical theory than is presented in the textbook, which you find in this module page.

Module overview

Todo

- Extensions and challenges (self study)
  - qualitative predictors: dummy coding (needed)
  - non-additivity: including interactions (useful)
- Important results in MLR
- Summing up with team Kahoot! (if time permits!)

Linear regression

- Very simple approach for *supervised learning*.
- Parametric.
- Quantitative response vs. one or several explanatory variables.
- Aims:
  - **Prediction** - "black box"
  - **Explanation** - understanding the relationship between *explanatory variables* and the response
- Is linear regression too simple? Maybe, but very useful. Important to *understand* because many learning methods can be seen as generalization of linear regression.

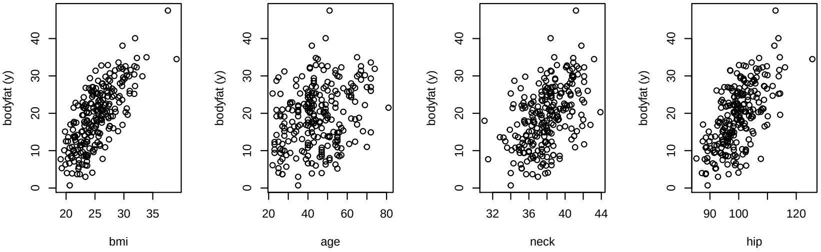
Motivating example: Prognostic factors for body fat

(From Theo Gasser & Burkhardt Seifert *Grundbegriffe der Biostatistik*)

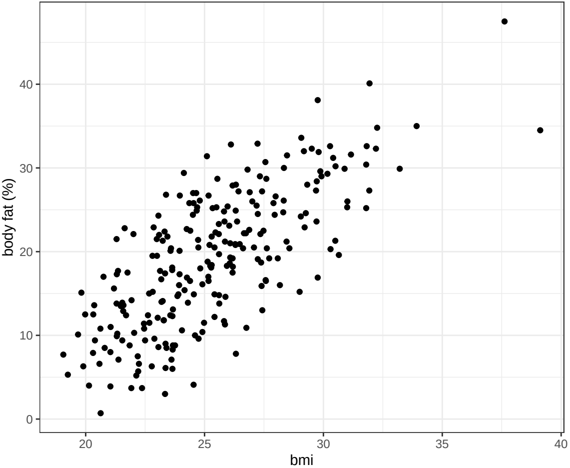
Body fat is an important indicator for overweight, but difficult to measure.

Question: Which factors allow for precise estimation (prediction) of body fat?

Study with 243 male participants, where body fat (%) and BMI and other predictors were measured. Some scatterplots:



For a good predictive model we need to dive into *multiple linear regression*. However, wer start with the simple case of *only one predictor variable*:



Interesting questions

1. How good is BMI as a predictor for body fat?
2. How strong is this relationship?
3. Is the relationship linear?
4. Are also other variables associated with `bodyfat`?
5. How well can we predict the bodyfat of a person?

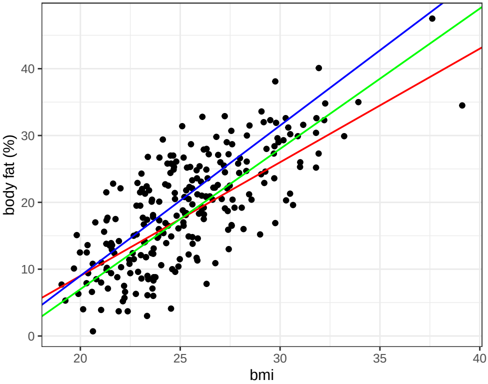
Simple Linear Regression

- One quantitative response Y is modelled
- from *one covariate* x (=simple),
- and the relationship between Y and x is assumed to be *linear*.

If the relation between Y and x is perfectly linear, all instances of (x, Y), given by (x_i, y_i), $i = 1, \ldots, n$, lie on a straight line and fulfill

$$y_i = \beta_0 + \beta_1 x_i \ .$$

But which is the “true” or “best” line, if the relationship is not exact?



Task: Estimate the intercept and slope parameters (by “eye”) and write it down (we will look at your answers later).

It is obvious that

- the linear relationship does not describe the data perfectly.
- another realization of the data (other 243 males) would lead to a slightly different picture.

$\Rightarrow$ We need a **model** that describes the relationship between BMI and bodyfat.

The simple linear regression model

In the linear regression model the dependent variable Y is related to the independent variable x as

$$Y = \beta_0 + \beta_1 x + \varepsilon \ , \qquad \varepsilon \sim N(0, \sigma^2) \ .$$

In this formulation Y is a random variable $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$ where

$$Y \quad = \quad \underbrace{\text{expected value}}_{\mathrm{E}(Y)=\beta_0+\beta_1 x} \quad + \quad \underbrace{\text{error}}_{\varepsilon} \ .$$

Note:

- The model for Y given x has three parameters: β_0 (intercept), β_1 (slope coefficient) and σ^2 .
- x is the independent/ explanatory / regressor variable.
- Y is the dependent / outcome / response variable.

Modeling assumptions

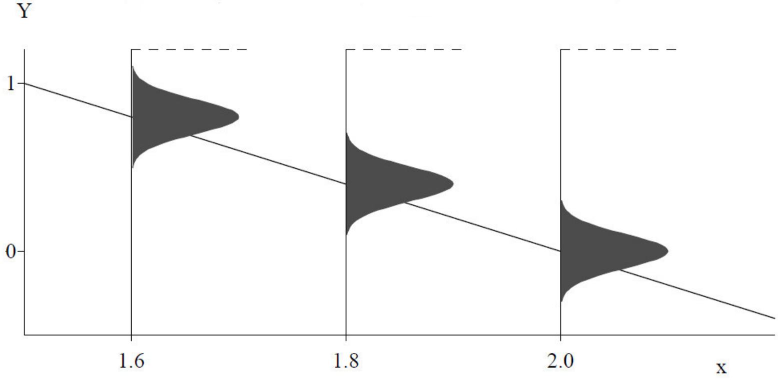
The central assumption in linear regression is that for any pairs (x_i, Y_i), the error $\varepsilon_i \sim N(0, \sigma^2)$. This implies

1. The expected value of ε_i is 0: $\mathrm{E}(\varepsilon_i) = 0$.
2. All ε_i have the same variance: $\mathrm{Var}(\varepsilon_i) = \sigma^2$.
3. All ε_i are normally distributed.
4. ε is independent of any variable, observation number etc.
5. $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent of each other.

Visualization of the regression assumptions

The assumptions about the linear regression model lie in the error term

$$\varepsilon \sim N(0, \sigma^2) \ .$$



Note: The true regression line goes through $\mathrm{E}(Y)$.

Parameter estimation ("model fitting")

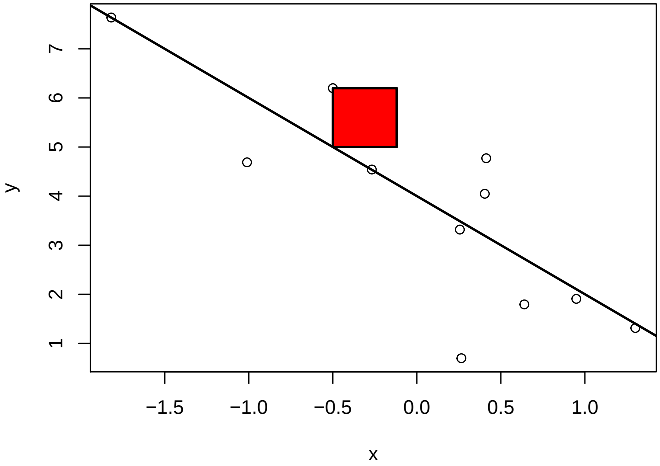
In a regression analysis, the task is to estimate the **regression coefficients** β_0, β_1 and the **residual variance** σ^2 for a given set of (x, y) data.

- **Problem:** For more than two points (x_i, y_i), $i = 1, \ldots, n$, there is generally no perfectly fitting line.
- **Aim**: We want to find the parameters (a, b) of the best fitting line $Y = a + bx$.
- **Idea:** Minimize the deviations between the data points (x_i, y_i) and the regression line.

But what are we actually going to minimize?

Least squares

Remember the **Least Squared Method**. Graphically, we are minimizing the sum of the squared distances over all points:



- Mathematically, β_0 and β_1 are estimated such that the sum of squared vertical distances (residual sum of squares)

$$RSS = \sum_{i=1}^{n} e_i^2 \ , \qquad \text{where} \quad e_i = y_i - (a + bx_i)$$

is being minimized.

- The respective "best" estimates are called $\hat{\beta}_0$ and $\hat{\beta}_1$.
- We can predict the value of the response for a (new) observation of the covariate at x.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

- The i-th *residual* of the model is the difference between the i-th *observed* response value and the i-th *predicted* value, and is written as:

$$e_i = Y_i - \hat{y}_i.$$

- We may regard the residuals as *predictions* (not estimates) of the error terms ε_i.

(The error terms are random variables and can not be estimated - they can be predicted. It is only for parameters that we speak about estimates.)

Least squares estimators:

Using n observed independent data points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \ ,$$

the least squares estiamtes for simple linear regression are given as

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{1}$$

and

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

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

Do-it-yourself “by hand”

Go to the Shiny gallery and try to “estimate” the correct parameters.

You can do this here:

https://gallery.shinyapps.io/simple_regression/

Example continued: Body fat

Assume a linear relationship between the % bodyfat (`bodyfat`) and the BMI (`bmi`), we can get the LS estimates using R as follows:

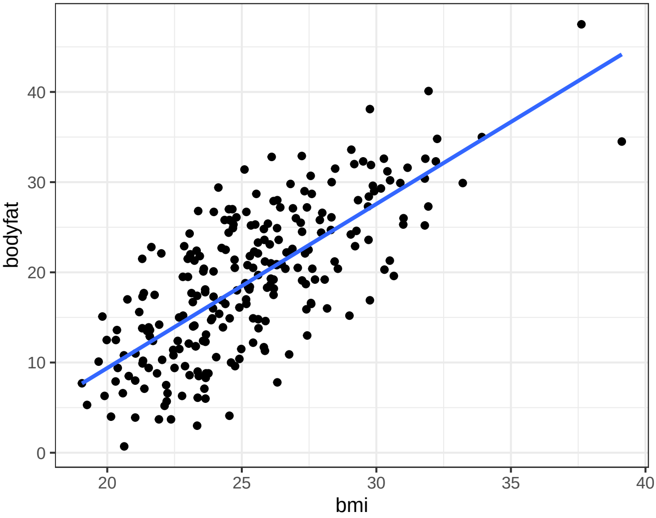
```
r.bodyfat = lm(bodyfat ~ bmi, data = d.bodyfat)
```

The estimates (and more information) can be obtained as follows:

```
summary(r.bodyfat)$coef
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -26.984368  2.7689004 -9.745518 3.921511e-19
## bmi           1.818778  0.1083411 16.787522 2.063854e-42
```

We see that the model fits the data quite well. It captures the essence. It looks that a linear relationship between `bodyfat` and `bmi` is a good approximation.



Questions:

- The blue line gives the estimated model. Explain what the line means in practice. Is this result plausible?
- Compare the estimates for β_0 and β_1 to the estimates you gave at the beginning - were you close?
- How does this relate to the *true* (population) model?
- By looking at the spread of the points around the line, can you detect any violations of the modelling assumptions?
- Finally: **What could the regression line look like if another set of 243 males were used for estimation?**

Uncertainty in the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

Note: $\hat{\beta}_0$ and $\hat{\beta}_1$ are themselves random variables and as such contain uncertainty!

Let us look again at the regression output, this time only for the coefficients. The second column shows the standard error of the estimate:

```
summary(r.bodyfat)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -26.984368  2.7689004 -9.745518 3.921511e-19
## bmi           1.818778  0.1083411 16.787522 2.063854e-42
```

$\rightarrow$ The logical next question is: what is the distribution of the estimates?

Distribution of the estimators for $\hat{\beta}_0$ and $\hat{\beta}_1$

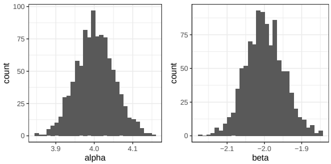
To obtain an intuition, we generate data points according to model

$$y_i = 4 - 2x_i + \varepsilon_i \ , \quad \varepsilon_i \sim N(0, 0.5^2).$$

In each round, we estimate the parameters and store them:

```
set.seed(1)
niter <- 1000
pars <- matrix(NA, nrow = niter, ncol = 2)
for (ii in 1:niter) {
    x <- rnorm(100)
    y <- 4 - 2 * x + rnorm(100, 0, sd = 0.5)
    pars[ii, ] <- lm(y ~ x)$coef
}
```

Doing it 1000 times, we obtain the following distributions for $\hat{\beta}_0$ and $\hat{\beta}_1$:



Accuracy of the parameter estimates

- The standard errors of the estimates are given by the following formulas:

$$\mathrm{Var}(\hat{\beta}_0) = \mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

  and

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

- $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)$ is in general different from zero.

Note: We will *derive a general version* of these formulas for multiple linear regression, because without matrix notation this is very cumbersome.

Under the assumption that $\varepsilon \sim N(0, \sigma^2)$, we have in addition that

$$\hat{\alpha} \sim N(\alpha, \sigma^2_{\beta_0}) \quad \text{and} \quad \hat{\beta} \sim N(\beta, \sigma^2_{\beta_1}) \ .$$

This implies that that $\hat{\beta}_0$ and $\hat{\beta}_1$ as defined in formulas (1) and (2).

Design issue with data collection

Recall that

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \; ,$$

thus for a given σ^2, the standard error is only dependent on the *design* of the x_i's!

- Would we like the $\mathrm{SE}(\hat{\beta}_1)^2$ large or small? Why?
- If it is possible for us to choose the x_i's, which strategy should we use to choose them?
- Assume x can take values from 1 to 10 and we choose $n = 10$ values. Which is the best design?
  - evenly in a grid: $[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$.
  - only lower and upper value: $[1, 1, 1, 1, 1, 10, 10, 10, 10, 10]$.
  - randomly drawn from a uniform distribution on $[1, 10]$.

```r
x1 = seq(1:10)
x2 = c(rep(1, 5), rep(10, 5))
x3 = runif(10, 1, 10)

sd1 = sqrt(1/sum((x1 - mean(x1))^2))
sd2 = sqrt(1/sum((x2 - mean(x2))^2))
sd3 = sqrt(1/sum((x3 - mean(x3))^2))

print(c(sd1, sd2, sd3))
```

```
## [1] 0.11009638 0.07027284 0.11505715
```

→ The second design - all observations at extremes - is best!

Residual standard error (RSE)

- **Problem**: σ is usually no known, but needs to be estimated.
- Remember: The residual sum of squares is $\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$.
- An estimate of σ, the residual standard error, RSE, is given by

$$\hat{\sigma} = \text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \ .$$

- It is related to the amount the response variables deviate from the estimated regression line.
- So actually we have

$$\hat{\text{SE}}(\hat{\beta}_1)^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \ ,$$

but we usually just write $\text{SE}(\hat{\beta}_1)^2$ (without the extra hat).

If the simple linear regression assumptions are fulfilled, that is, $\varepsilon_i \sim N(0, \sigma^2)$ and all ε_i independent, then it can be shown that

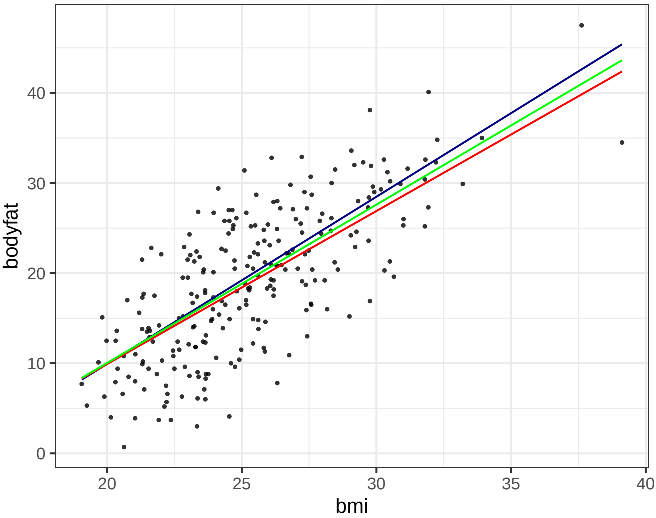
$$\frac{\mathrm{RSE}^2(n-2)}{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sigma^2} \sim \chi^2_{n-2}$$

The estimated standard errors can be seen using the `summary()` function:

```
summary(r.bodyfat)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -26.984368  2.7689004 -9.745518 3.921511e-19
## bmi           1.818778  0.1083411 16.787522 2.063854e-42
```

To illustrate this point further, again fit the bodyfat example, but each time with only half of the data (randomly selected points each time). See how the model fit varies:



Testing and Confidence Intervals

After the regression parameters and their uncertainties have been estimated, there are typically two fundamental questions:

1. **"Are the parameters compatible with some specific value?"** Typically, the question is whether the slope β_1 might be 0 or not, that is: "Is x an informative predictor or not?"

$\Rightarrow$ This leads to a **statistical test**.

2. "Which values of the parameters are compatible with the data?"

$\Rightarrow$ This leads us to determine **confidence intervals**.

Let's first go back to the output from the bodyfat example:

```r
summary(r.bodyfat)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -26.984368  2.7689004 -9.745518 3.921511e-19
## bmi           1.818778  0.1083411 16.787522 2.063854e-42
```

Besides the estimate and the standard error (which we discussed before), there is a `t value` and a probability `Pr(>|t|` that we need to understand.

How do these things help us to answer the two questions above?

Testing the effect of a covariate

Remember: in a statistical test you first need to specify the *null hypothesis*. Here, typically, the null hypothesis is

$$H_0 : \quad \beta_1 = 0 \; .$$

In words: $H_0 =$ "There is no relationship between X and Y."

- Note 1: However, you might want to test against another null hypothesis, like $\beta_1 = c$.
- Note 2: Included in H_0 is the assumption that the data follow the simple linear regression model!

Here, the *alternative hypothesis* is given by

$$H_A : \quad \beta_1 \neq 0$$

Remember: To carry out a statistical test, we need a *test statistic*. This is some type of **summary statistic** that follows a known distribution under H_0. For our purpose, we use the so-called **T-statistic**

$$T = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \ .$$

Note: If you want to test against another value than $\beta_1 = 0$, the formula is

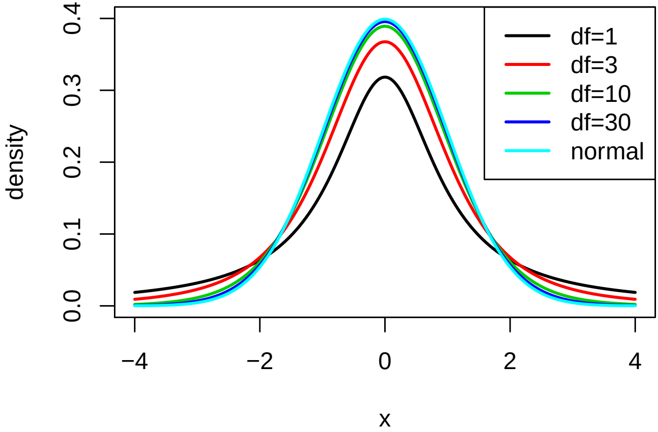
$$T = \frac{\hat{\beta}_1 - c}{SE(\hat{\beta}_1)}$$

Distribution of parameter estimators

We will *derive a general version* for multiple linear regression!
Brief recap:

- Under H_0, T has a t-distribution with $n - 2$ degrees of freedom ($n =$ number of data points; compare to Chapter 8.6 in (Walepole et al. 2012)).

Recap: The t-distribution



- The t-distribution has heavier tails than the normal distribution.
- For df ≥ 30 the t and Normal distribution are pretty similar.

Hypothesis tests for bodyfat example

So let's again go back to the bodyfat regression output:

```
summary(r.bodyfat)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -26.984368  2.7689004 -9.745518 3.921511e-19
## bmi           1.818778  0.1083411 16.787522 2.063854e-42
```

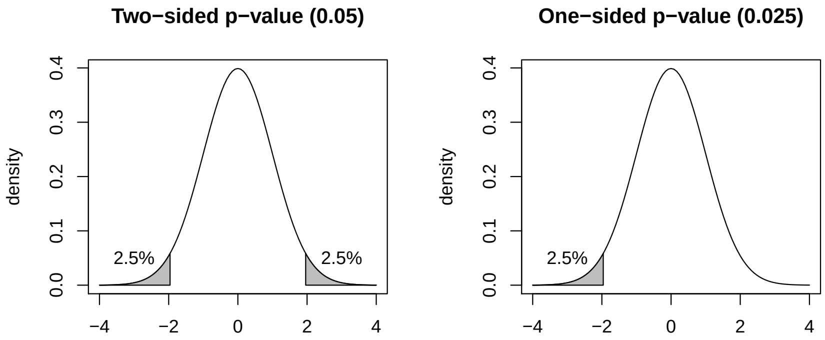
Task: Use the above formulas to derive the T-statistics.

- The last column contains the *p-values* of the tests $\beta_0 = 0$ and $\beta_1 = 0$.
- The p-value for `bmi` is very small ($p < 0.0001$). **What does this mean?**

Recap: Formal definition of the p-value

Formal definition of p-value: the probability to observe a data summary (e.g., an average) that is at least as extreme as the one observed, given that the Null Hypothesis is correct.

Example (normal distribution): Assume the observed test-statistic leads to a z-value = -1.96 $\Rightarrow$ $\mathrm{P}(|z| \geq 1.96) = 0.05$ and $\mathrm{P}(z \leq -1.96) = 0.025$.



Recap: Two types of errors

In the testing setup, we typically *reject the null hypothesis* if the p-value is small enough. Typical cutoffs for the *significance level* (α) are 5% or 1%.

However, this means we can make two types of errors:

- Type I error:
- Type II error:

Cautionary notes regarding p-values:

- The (mis)use of p-values is heavily under critique in the scientific world!!!
- Simple yes/no decisions do often stand on very wiggly scientific ground!!

(See reading tasks for this week.)

Confidence intervals

- Confidence intervals (CIs) are a much more informative way to report results than p-values!
- The t-distribution[1] can be used to create confidence intervals for the regression parameters. The lower and upper limits of a 95% confidence interval for β_j are

$$\hat{\beta}_j \pm t_{(1-\alpha/2),n-2} \cdot \mathrm{SE}(\hat{\beta}_j) \quad j = 0, 1.$$

- Interpretation of this confidence interval:
- There is a 95% probability that the interval will contain the *true* value of β_j.
- **It is the range of parameter estimates that are *compatible with the data* .**

[1] If n is large, the normal approximation to the t-distribution can be used (and is used in the textbook).

Doing this for the bodfat example "by hand" is not hard. We have $241 (= 243 - 2)$ degrees of freedom:

```
coefs <- summary(r.bodyfat)$coef
beta <- coefs[2, 1]
sdbeta <- coefs[2, 2]
beta + c(-1, 1) * qt(0.975, 241) * sdbeta
```

```
## [1] 1.605362 2.032195
```

Even easier: directly ask R to give you the CIs.

```
confint(r.bodyfat, level = c(0.95))
```

```
##                   2.5 %     97.5 %
## (Intercept) -32.438703 -21.530032
## bmi           1.605362   2.032195
```

Interpretation: for an increase in the bmi by one index point, roughly 1.82 percentage points more bodyfat are expected, and all true values for β_1 between 1.61 and 2.03 are compatible with the observed data.

Confidence and prediction ranges

- Based on the joint distribution of the intercept and slope it is possible to find the distribution for the linear predictor $\hat{\beta}_0 + \hat{\beta}_1 x$, and then confidence intervals for $\beta_0 + \beta_1 x$.

$\rightarrow$ **Confidence range**

- Accounting for the fact that we also have an error in the equation ε, we can also find the distribution of future observations.

$\rightarrow$ **Prediction range**

Todo ev say more about confidence and prediction ranges, or put this into exercise class.

Model accuracy

Measured by

1. The **residual standard error (RSE)**, which provides an **absolute measure** of *lack of fit* (see above).

2. The **coefficient of determination** R^2, which measures the proportion of y's variance explained by the model (between 0 and 1), is a **relative measure** of *lack of fit*:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2},$$

where

$$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

is the *total sum of squares*, a measure for the toal variability in Y.

R^2 in the bodyfat example

```r
summary(r.bodyfat)$r.squared
```

```
## [1] 0.5390391
```

Compare this to the squared correlation between the two variables:

```r
cor(d.bodyfat$bodyfat, d.bodyfat$bmi)^2
```

```
## [1] 0.5390391
```

$\rightarrow$ In simple linear regression, R^2 is the squared correlation between the independent and the dependent variable.

Multiple Linear Regression

Remember that the bodyfat dataset contained much more information than only bmi and bodyfat:

- `bodyfat`: % of body fat.
- `age`: age of the person.
- `weight`: body weighth.
- `height`: body height.
- `bmi`: bmi.
- `abdomen`: circumference of abdomen.
- `hip`: circumference of hip.

Model

We assume

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1 + ... + \beta_p X_p + \varepsilon \ , \tag{3}$$

where X_j is the jth predictor and β_j the respective regression coeffficient.

Assume we have n sampling units $(x_{1i}, \dots, x_{pi}, y_i)$, $1 \leq i \leq n$, such that each represent an instance of equation (3), we can use the data matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & ... & x_{1p} \\ 1 & x_{21} & ... & x_{2p} \\ \vdots & ... & ... & \vdots \\ 1 & x_{n1} & ... & x_{np} \end{bmatrix}$$

to write the model in matrix form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Notation

- $\mathbf{Y}$: $(n \times 1)$ vector of responses [e.g. one of the following: rent, weight of baby, ph of lake, volume of tree]
- $\mathbf{X}$: $(n \times (p+1))$ design matrix, and $\mathbf{x}_i^T$ is a $(p+1)$-dimensional row row vector for observation i.
- $\boldsymbol{\beta}$: $((p+1) \times 1)$ vector of regression parameters $(\beta_0, \beta_1, \ldots, \beta_p)^\top$.
- $\boldsymbol{\varepsilon}$: $(n \times 1)$ vector of random errors.
- We assume that pairs $(\mathbf{x}_i^T, y_i)$ $(i = 1, ..., n)$ are measured from *independent* sampling units.

Remark: other books, including the book in TMA4267 and TMA4315 define p to include the intercept. This may lead to some confusion about p or $p+1$ in formulas...

Classical linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Assumptions:

1. $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$.
2. $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2\mathbf{I}$.
3. The design matrix has full rank, $\mathrm{rank}(\mathbf{X}) = p + 1$. (We assume $n >> (p + 1)$.)

The classical *normal* linear regression model is obtained if additionally

4. $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$ holds. Here N_n denotes the n-dimensional multivarate normal distribution.

Design matrix in R

```r
r.bodyfat = lm(bodyfat ~ bmi + age, data = d.bodyfat)
head(model.matrix(r.bodyfat))
```

```
##   (Intercept)   bmi age
## 1           1 23.65  23
## 2           1 23.36  22
## 3           1 24.69  22
## 4           1 24.91  26
## 5           1 25.54  24
## 6           1 26.48  24
```

```r
head(d.bodyfat$bmi)
```

```
## [1] 23.65 23.36 24.69 24.91 25.54 26.48
```

```r
head(d.bodyfat$age)
```

```
## [1] 23 22 22 26 24 24
```

Distribution of the response vector

Assume that

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ , \quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}) \ .$$

Q:

- What is the mean $\mathrm{E}(\mathbf{Y})$?
- The covariance matrix $\mathrm{Cov}(\mathbf{Y})$ given $\mathbf{X}$?
- Thus what is the distribution of $\mathbf{Y}$?

A:

Parameter estimation

In multiple linear regression parameters in β are estimated with maximum likelihood and least squares. These two methods give the same estimator when we assume the normal linear regression model.

Least squares and maximum likelihood estimator for $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

The estimator is found by minimizing the RSS for a multiple linear regression model:

$$\begin{aligned}\mathrm{RSS} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_p x_{ip})^2 \\ &= \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})\end{aligned}$$

The estimator is found by solving the system of $(p+1)$ equations

$$\frac{\partial \mathrm{RSS}}{\partial \boldsymbol{\beta}} = \mathbf{0} \ .$$

$\rightarrow$ Derivation on the board.

Example continued

```r
r.bodyfat3 <- lm(bodyfat ~ bmi + age + neck + hip + abdomen, data = d.bodyfat)
summary(r.bodyfat3)
```

```
##
## Call:
## lm(formula = bodyfat ~ bmi + age + neck + hip + abdomen, data = d.bodyfat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3727 -3.1884 -0.1559  3.1003 12.7613
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.74965    7.29830  -1.062  0.28939
## bmi          0.42647    0.23133   1.844  0.06649 .
## age          0.01457    0.02783   0.524  0.60100
## neck        -0.80206    0.19097  -4.200 3.78e-05 ***
## hip         -0.31764    0.10751  -2.954  0.00345 **
## abdomen      0.83909    0.08418   9.968  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.392 on 237 degrees of freedom
## Multiple R-squared:  0.7185, Adjusted R-squared:  0.7126
## F-statistic:   121 on 5 and 237 DF,  p-value: < 2.2e-16
```

Reproduce the values under `Estimate` by calculating without the use of `lm`.

```
X = model.matrix(r.bodyfat3)
Y = d.bodyfat$bodyfat
betahat = solve(t(X) %*% X) %*% t(X) %*% Y
print(betahat)
```

```
##                    [,1]
## (Intercept) -7.74964673
## bmi          0.42647368
## age          0.01457356
## neck        -0.80206081
## hip         -0.31764315
## abdomen      0.83909391
```

Distribution of the regression parameter estimator

1. We assumed alassical *normal* linear regression model with $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$, with full-rank matrix $\mathbf{X}$, leading to
$$\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\sigma^2}\mathbf{I}) \ .$$
2. Then we "found" that an estimator for β is
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

What are

- The mean $\mathrm{E}(\hat{\boldsymbol{\beta}})$?
- The covariance matrix $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$?
- The distribution of $\hat{\boldsymbol{\beta}}$?

Distribution of the regression parameter estimator (summary)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

This can be written as $\hat{\boldsymbol{\beta}} = \mathbf{C}\mathbf{Y}$ where

- $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
- $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

Therefore:

- $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \mathbf{C}\mathrm{E}(\mathbf{Y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$.
- $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \mathbf{C}\mathrm{Cov}(\mathbf{Y})\mathbf{C}^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$.
- $\hat{\boldsymbol{\beta}}$ is multivariate normal $(p+1)$ dimensions.

So: $\hat{\beta} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$.
(Todo: Use PropertiesBetahatMLR.pdf for a derivation)

How does this compare to simple linear regression? Not so easy to see a connection!

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \text{ and } \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Often we use centered data (and also scaled) to ease interpretation.

Another data set: Ozone

New York, 1973: 111 observations of

- `ozone` : ozone concentration (ppm); **response variable**
- `radiation` : solar radiation (langleys)
- `temperature` : daily maximum temperature (F)
- `wind` : wind speed (mph)

```r
library(ElemStatLearn)
data(ozone)
head(ozone)
```

```
##   ozone radiation temperature wind
## 1    41       190          67  7.4
## 2    36       118          72  8.0
## 3    12       149          74 12.6
## 4    18       313          62 11.5
## 5    23       299          65  8.6
## 6    19        99          59 13.8
```

```r
ozone.lm = lm(ozone ~ temperature + wind + radiation, data = ozone)
```

```r
head(model.matrix(ozone.lm))
```

```
##   (Intercept) temperature wind radiation
## 1           1          67  7.4       190
## 2           1          72  8.0       118
## 3           1          74 12.6       149
## 4           1          62 11.5       313
## 5           1          65  8.6       299
## 6           1          59 13.8        99
```

```r
head(ozone$ozone)
```

```
## [1] 41 36 12 18 23 19
```

```r
summary(ozone.lm)
```

```
##
## Call:
## lm(formula = ozone ~ temperature + wind + radiation, data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.485 -14.210  -3.556  10.124  95.600
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -64.23208   23.04204  -2.788  0.00628 **
## temperature   1.65121    0.25341   6.516 2.43e-09 ***
## wind         -3.33760    0.65384  -5.105 1.45e-06 ***
## radiation     0.05980    0.02318   2.580  0.01124 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.17 on 107 degrees of freedom
## Multiple R-squared:  0.6062, Adjusted R-squared:  0.5952
## F-statistic: 54.91 on 3 and 107 DF,  p-value: < 2.2e-16
```

Remember: $\hat{\beta} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$. The covariance matrix can be obtained as follows:

```
vcov(ozone.lm)
```

```
##              (Intercept)  temperature          wind     radiation
## (Intercept) 530.93558002 -5.503192281 -1.043562e+01  0.0266688733
## temperature  -5.50319228  0.064218138  8.034556e-02 -0.0015749279
## wind        -10.43562350  0.080345561  4.275126e-01 -0.0003442514
## radiation     0.02666887 -0.001574928 -3.442514e-04  0.0005371733
```

Four important questions

1. Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor variables, what reaposne value should we predict, and how accurate is our prediction?

1. Relationship between predictors and response?

Question is whether we could as well omit all predictor variables at the same time, that is

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

vs.

$$H_1 : \text{at least one } \beta_j \text{ is non-zero.}$$

To answer this, we need the F-statistic

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} \sim F_{p,(n-p-1)} \ ,$$

where total sum of squares $\mathrm{TSS} = \sum_i (y_i - \bar{y})^2$, and residual sum of squares $\mathrm{RSS} = \sum_i (y_i - \hat{y}_i)^2$. Under the Normal regression assumptions, F follows an $F_{p,(n-p-1)}$ distribution (see (Walepole et al. 2012), Chapter 8.7).

- If H_0 is true, F is expected to be 1.
- Otherwise, we expect that the numerator is larger than the denominator (because the regression then explains a lot of variation) and thus F is greater than 1. For an observed value f_0, the p-value is given as

$$p = P(F_{p,n-p-1} > f_0) \ .$$

Checking the F-value in the R output:

```
summary(r.bodyfat)
```

```
##
## Call:
## lm(formula = bodyfat ~ bmi + age, data = d.bodyfat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0415  -3.8725  -0.1237   3.9193  12.6599
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -31.25451    2.78973 -11.203  < 2e-16 ***
## bmi           1.75257    0.10449  16.773  < 2e-16 ***
## age           0.13268    0.02732   4.857 2.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.329 on 240 degrees of freedom
## Multiple R-squared:  0.5803, Adjusted R-squared:  0.5768
## F-statistic: 165.9 on 2 and 240 DF,  p-value: < 2.2e-16
```

Conclusion?

More complex hypotheses

Sometimes we don't want to test if all β's are zero at the same time, but only a subset $1, \dots, q$:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0$ vs. H_1 : at least one different from zero.

Again, the F-test can be used, but now F is calculated like

$$F = \frac{(\mathrm{RSS}_0\text{-}\mathrm{RSS})/(q)}{\mathrm{RSS}/(n-p-1)} \sim F_{q,n-p-1} \ ,$$

where

- Large model: RSS with $p+1$ regression parameters
- Small model: RSS_0 with $q+1$ regression parameters

Example in R

- **Question:** Do `weight` and `height` explain something of `bodyfat`, on top of the variables `bmi` and `age`?
- Fit both models and use the `anova()` function to carry out the F-test:

```
r.bodyfat.large = lm(bodyfat ~ bmi + age, data = d.bodyfat)
r.bodyfat.small = lm(bodyfat ~ bmi + age + weight + height, data = d.bodyfat)
anova(r.bodyfat.large, r.bodyfat.small)
```

```
## Analysis of Variance Table
##
## Model 1: bodyfat ~ bmi + age
## Model 2: bodyfat ~ bmi + age + weight + height
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    240 6816.2
## 2    238 6702.9  2    113.28 2.0112 0.1361
```

Inference about a single predictor β_j

A special case is

$$H_0 : \beta_j = 0 \;\text{ vs. }\; H_1 : \beta_j \neq 0$$

- Nothing new: We did it for simple linear regression!
- However, now the F-statistic becomes

$$F = \frac{(\text{RSS}_0\text{-RSS})/(p-1)}{\text{RSS}/(n-p-1)} \sim F_{1,n-p-1} \ ,$$

and it is known that

$$F_{1,n-p-1} = t^2_{n-p-1} \ ,$$

thus we can use a T-statistics with $(n-p-1)$ degrees of freedom to get the p-value.

Going back again:

```r
summary(r.bodyfat)$coef
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -31.2545057 2.78973238 -11.203406 1.039096e-23
## bmi           1.7525705 0.10448723  16.773060 2.600646e-42
## age           0.1326767 0.02731582   4.857137 2.149482e-06
```

However:

- Only checking the individual p-values is dangerous. **Why?**
- Not possible if $n > p$ $\rightarrow$ need other approaches (see e.g., Module 6).

Inference about β_j: confidence interval

- Using that

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}(\hat{\beta}_j)} \sim t_{n-p-1} \ ,$$

  we can create confidence intervals for β_j in the same manner as we did for simple linear regression.

- For example, when using the typical confidence level $\alpha = 0.05$ we have

$$\hat{\beta}_j \pm t_{0.975,n-p-2} \cdot \mathrm{SE}(\hat{\beta}_j) \ .$$

```r
confint(r.bodyfat)
```

```
##                   2.5 %      97.5 %
## (Intercept) -36.7499929 -25.7590185
## bmi           1.5467413   1.9583996
## age           0.0788673   0.1864861
```

2. Deciding on important variables

Overarching question:

Which model is the best?

But:

- Not clear what *best* means → we need an objective criterion, like AIC, BIC, Mallows C_p, adjusted R^2.
- There are usually **many** possible models. For p predictors, we can build 2^p different models.
- **Cautionary note**: Model selection can also lead to biased parameters estimates.

→ This topic is the focus of Module 6.

3. Model Fit

We can again look at the two measures from simple linear regression:

- An absolute measure of lack of fit is again given by the estimate of σ, the residual standard error (RSE)

$$\hat{\sigma} = \mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{n-p-1}} \, .$$

- R^2 is again the fraction of variance explained (no change from simple linear regression)

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \, .$$

Simply speaking: "The higher R^2, the better."

However: Caveat with R^2

Let us look at the R^2s from the three bodyfat models (model 1: $y \sim bmi$; model 2: $y \sim bmi + age$; model 3: $y \sim bmi + age + neck + hip + abdomen$):

```
summary(r.bodyfatM1)$r.squared
```

```
## [1] 0.5390391
```

```
summary(r.bodyfatM2)$r.squared
```

```
## [1] 0.5802956
```

```
summary(r.bodyfatM3)$r.squared
```

```
## [1] 0.6004791
```

The models explain 54%, 58% and 72% of the total variability of y. It thus *seems* that larger models are "better". However, R^2 does always increase when new variables are included, but this does not mean that the model is more reasonable.

Adjusted R^2

When the sample size n is small with respect to the number of variables m included in the model, an *adjusted* R^2 gives a better ("fairer") estimation of the actual variability that is explained by the covariates:

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-m-1}$$

R_a^2 **penalizes for adding more variables** if they do not really improve the model!

$\rightarrow$ R_a may decrease when a new variable is added.

3.1 Model fit – broader sense

We will look at model validation / model checking later.

4. Predictions: Two questions

1. **Which other regression lines are compatible with the observed data?**

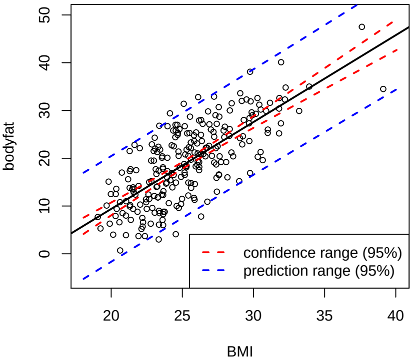
We can use $\hat{\beta}_0, \ldots, \hat{\beta}_p$ to estimate the *least squares plane*

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p$$

as an approximation of $f(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$. This leads to the confidence interval.

2. **Where do future observations with a given x coordinate lie?**

Even if we could predict $\hat{Y} = f(X)$, the *true* value Y varies around $\hat{Y}$. We can compute a prediction interval for new observations Y.



Note: The prediction range is much broader than the confidence range. Why?

Calculation of the confidence range

Given a realization of $X_1, \dots, X_p$, say $x_1^{(0)}, \dots x_p^{(0)}$. The question is:

Where does $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(0)} + \dots \hat{\beta}_p x_p^{(0)}$ lie with a certain confidence (i.e., 95%)?

This question is not trivial, because $\hat{\beta}_0, \dots \hat{\beta}_p$ are estimates from the data and contain uncertainty.

Plotting the confidence interval around all $\hat{Y}_0$ values one obtains the **confidence range** or **confidence band for the expected values** of Y.

$\rightarrow$ For the confidence range, only the uncertainty in the estimates $\hat{\beta}_0, \dots \hat{\beta}_p$ matters.

Calculation of the prediction interval

Given a new value of $X_1, \dots, X_p$, say $x_1^{(0)}, \dots x_p^{(0)}$. The question is:

Where does a **future observation** lie with a certain confidence (i.e., 95%)?

To answer this question, we have to consider not only the uncertainty in the predicted value $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(0)} + \dots \hat{\beta}_p x_p^{(0)}$, but also the irreducible error $\varepsilon_i \sim N(0, \sigma^2)$.

$\rightarrow$ The *prediction interval is always wider than the confidence range.*

Confidence and prediction intervals can be found in R using `predict` on an `lm` object (make sure that `newdata` is a `data.frame` with the same names as the original data).

```
fit = lm(bodyfat ~ bmi + age + abdomen, data = d.bodyfat)
newobs = d.bodyfat[1, ]
predict(fit, newdata = newobs, interval = "confidence", type = "response")
```

```
##        fit      lwr      upr
## 1 13.17595 11.99122 14.36069
```

```
predict(fit, newdata = newobs, interval = "prediction", type = "response")
```

```
##        fit      lwr     upr
## 1 13.17595 3.951613 22.4003
```

Difference between `interval="confidence"` and `interval="prediction"`?

Todo: Make an exercise about this!

Predictions: Model bias

Finally, we need to keep in mind that the model we work with is only an *approximation of the reality*. In fact,

In 2014, David Hand wrote:

> *In general, when building statistical models, we must not forget that the aim is to understand something about the real world. Or predict, choose an action, make a decision, summarize evidence, and so on, but always about the real world, not an abstract mathematical world: our models are not the reality – a point well made by George Box in his often-cited remark that "all models are wrong, but some are useful".*

(Box 1979)

Challenges - for model fit

1. Non-linearity of data
2. Correlation of error terms
3. Non-constant variance of error terms
4. Non-Normality of error terms
5. Outliers
6. High leverage points
7. Collinearity

Recap of modelling assumptions in linear regression

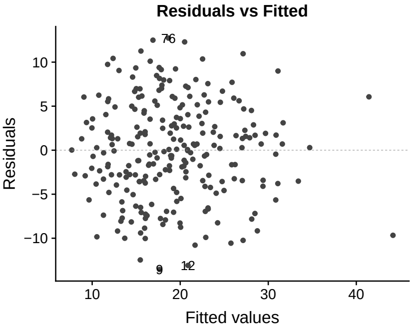
To make valid inference from our model, we must check if our model assumptions are fulfilled!

The assumption in linear regression is that the residuals follow a $N(0, \sigma^2)$ distribution, implying that :

1. The expected value of ε_i is 0: $\mathrm{E}(\varepsilon_i) = 0$.
2. All ε_i have the same variance: $\mathrm{Var}(\varepsilon_i) = \sigma^2$.
3. The ε_i are normally distributed.
4. The ε_i are independent of each other.

Model checking tool I: Tukey-Anscombe diagram

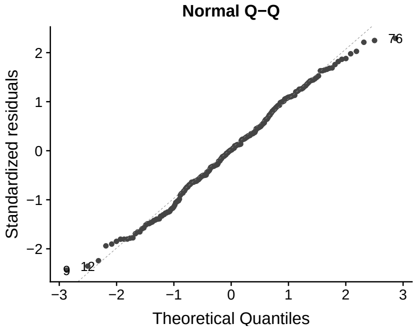
The Tukey-Anscombe diagram plots the residuals against the fitted values. For the bodyfat data it looks like this:



This plot is ideal to check if assumptions 1. and 2. (and partially 4.) are met. Here, this seems fine.

Model checking tool II: The QQ-diagram

To check assumption 3., the quantiles of the observed distribution are plotted against the quantiles of the respective theoretical (normal) distribution:

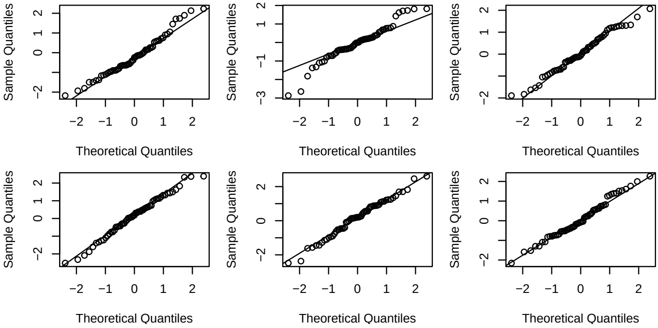


If the points lie approximately on a straight line, the data is fairly normally distributed. This is often “tested” by eye, and needs some experience.

Todo: Move this to exercises!

How do I know if a QQ-plot looks “good”?

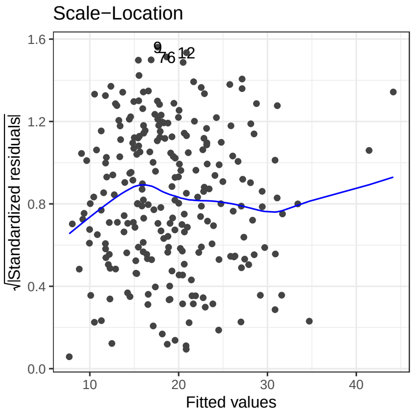
There is **no quantitative rule** to answer this question, experience is needed. However, you can gain this experience from simulations. To this end, generate the same number of data points of a normally distributed variable and compare to your plot.
Example: Generate 59 points $\varepsilon_i \sim N(0,1)$ each time:



Model checking tool III: The scale-location plot

The scale-location plot is particularly suited to check the assumption of equal variances (homoscedasticity; assumption 2.).

The idea is to plot the square root of the (standardized) residuals $\sqrt{|R_i|}$ against the fitted values $\hat{y}_i$. There should be *no trend*:



Model checking tool IV: The leverage plot

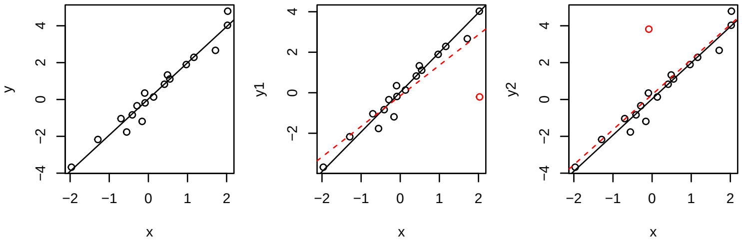
- Mainly useful to determine outliers.

- To understand the leverage plot, we need to introduce the idea of the **leverage**.

- In simple regression, the leverage of individual i is defined as $H_{ii} = (1/n) + (x_i - \overline{x})^2 \sum (x_i - \overline{x})^2$.

Q: When are leverages expected to be large/small?

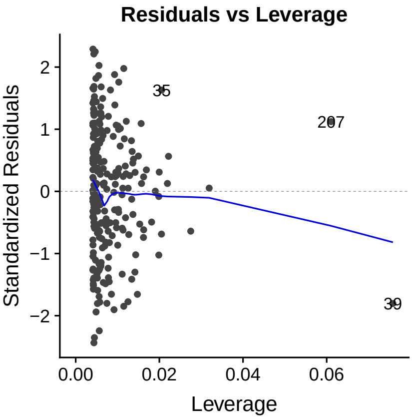
Illustration: Data points with x_i values far from the mean have a stronger leverage effect than when $x_i \approx \overline{x}$:



The outlier in the middle plot “pulls” the regression line in its direction and biases the slope.

Click here to do it manually!

In the leverage plot, (standardized) residuals $\tilde{R}_i$ are plotted against the leverage H_{ii} (still for the bodyfat):



Critical ranges are the top and bottom right corners!!

Leverages in multiple regression

- Leverage is defined as the diagonal elements of the hat matrix, i.e., the leverage of the i-th data point is h_{ii} on the diagonal of $\mathbf{H} = \mathbf{X}(\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T}$.
- A large leverage indicated that the observation (i) has a large influence on the estimation results, and that the covariate values ($\mathbf{x_i}$) are unusual.

Different types of residuals?

If can be shown that the vector of residuals, $\mathbf{e} = (e_1, e_2, \ldots, e_n)$ have a normal (singular) distribution with mean $\mathrm{E}(\mathbf{e}) = \mathbf{0}$ and covariance matrix $\mathrm{Cov}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

This means that the residuals (possibly) have different variance, and may also be correlated.

Q: Why is that a problem?

A:

We would like to check the model assumptions - we see that they are all connected to the error terms. But, but we have not observed the error terms ε so they can not be used for this. However, we have made "predictions" of the errors - our residuals. And, we want to use our residuals to check the model assumptions.

That is, we want to check that our errors are independent, homoscedastic (same variance for each observation), and not dependent on our covariates - and we want to use the residuals (observed) in place of the errors (unobserved). Then it would have been great if the residuals have these properties when the underlying errors have. To amend our problem we need to try to fix the residual so that they at least have equal variances. We do that by working with *standardized* or *studentized residuals*.

Standardized residuals:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

where h_{ii} is the ith diagonal element of the hat matrix $\mathbf{H}$.

In R you can get the standardized residuals from an `lm`-object (named `fit`) by `rstandard(fit)`.

Studentized residuals:

$$r_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

where $\hat{\sigma}_{(i)}$ is the estimated error variance in a model with observation number i omitted. This seems like a lot of work, but it can be shown that it is possible to calculated the studentized residuals directly from the standardized residuals.

In R you can get the studentized residuals from an `lm`-object (named `fit`) by `rstudent(fit)`.

Diagnostic plots in `R`

Todo: This must be part of the exercises.

Idea: Use `autoplot()` from the `ggfortify` package in R to plot the diagnostic plots.

Collinearity

In brief, collinearity refers to the situation when two or more predictors are correlated, thus encode (partially) for the same information.

Problems:

- Reduces the accuracy of the estimated coefficients $\hat{\beta}_j$ (large SE!).
- Consequently, reduces power in finding effects (p-values become larger).

Solutions:

- Detect it by calculating the *variance inflation factor* (VIF).
- Remove the problematic variable.
- Or combine the collinear variables into a single new one.

Todo: Read in the course book p.99-102 (self-study).

Other considerations in the regression model

1. Qualitative predictors (X_j):
   - Binary covariate (e.g., male/female, smoker/non-smoker)
   - Categorical covariate (e.g., black/white/green)?

2. Extensions of the linear model
   - Interactions
   - Non-linear terms

Binary predictors

So far, the covariates x were always continuous.

In reality, there are no restrictions assumed with respect to the X variables

One very frequent data type are **binary** variables, that is, variables that can only attain values 0 or 1.

If the binary variable x is the only variable in the model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the model has only two predicted outcomes (plus error):

$$Y_i = \begin{cases} \beta_0 + \varepsilon_i & \text{if } x_i = 0 \ , \\ \beta_0 + \beta_1 + \varepsilon_i & \text{if } x_i = 1 \ . \end{cases}$$

Qualitative predictors with more than 2 levels

Further reading

- Need details on the simple linear regression: From TMA4240/TMA4245 Statistics we have the thematic page for Simple linear regression (in Norwegian).
- Need more advanced thory: Theoretical version (no simple linear regression) from TMA4315 Generalized linear models H2018: TMA4315M2: Multiple linear regression
- Slightly different presentation (more focus on multivariate normal theory): Slides and written material from TMA4267
- And, same source, but now Slides and written material from TMA4267 Linear Statistical Models in 2017, Part 3: Hypothesis testing and ANOVA
- Videoes on YouTube by the authors of ISL, Chapter 2

R packages

If you want to look at the .Rmd file and `knit` it, you need to first install the following packages (only once).

```
# packages to install before knitting this R Markdown file to kn
# the Rmd
install.packages("knitr")
install.packages("rmarkdown")

# nice tables in Rmd
install.packages("kableExtra")

# cool layout for the Rmd
install.packages("prettydoc")  # alternative to github

# plotting
install.packages("ggplot2")  # cool plotting
install.packages("ggpubr")  # for many ggplots
install.packages("GGally")  # for ggpairs
# datasets
install.packages("ElemStatLearn")  # for ozone data set
```

Acknowledgements

Thanks to Julia Debik for contributing to this module page.

References

Box, G. E. P. 1979. "Robustness in the Strategy of Scientific Model Building." In *In Robustness in Statistics*, edited by R. L. Launer and G. N. Wilkinson, 201–36. New York: Academic Press.

Walepole, R. E., R. H. Myers, S. L. Myers, and K. Ye. 2012. *Probability & Statistics for Engineers and Scientists*. 9th ed. Boston: Pearson Education Inc.