# Kurs Bio144:

# Datenanalyse in der Biologie

Stefanie Muff (Lecture) & Owen L. Petchey (Exercises)

Week 2: Simple linear regression

2./3. March 2017

# Overview

- Introduction of the linear regression model
- Parameter estimation
- Simple model checking
- Goodness of the model: Correlation and $R^2$
- Tests and confidence intervals
- Confidence and prediction ranges

# Course material covered today

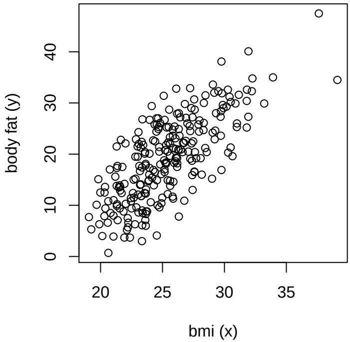
- Chapter 2 of *Lineare Regression*, p.7-20 (Stahel script), or chapters 13.1 - 13.4 in the Stahel book "Statistische Datenanalyse".
- Something from *The New Statistics with R*

# The body fat example

Remember: Aim is to find prognostic factors for body fat, without actually measuring it.

Even simpler question: How good is BMI as a predictor for body fat?

```
> plot(bodyfat ~ bmi,d.bodyfat,xlab="bmi (x)", ylab="body fat (y)")
```



# Linear relationship

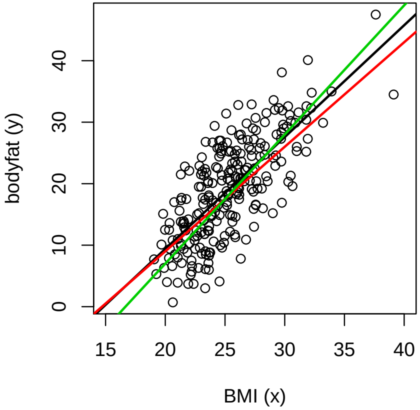
- The most simple relationship between an *explanatory variable* ($X$) and a *target/outcome variable* ($Y$) is a linear relationship. All points $(x_i, y_i)$, $i = 1, \ldots, n$, on a straight line follow the equation

$$y_i = \alpha + \beta x_i \ .$$

- Here, $\alpha$ is the axis intercept and $\beta$ the slope of the line. $\beta$ is also denoted as the regression coefficient of $X$.

- If $\alpha = 0$ the line goes through the origin.
- Interpretation of linear dependency: proportional increase in $y$ with increase (decrease) in $x$.

But which is the “true” or “best” line?



It is obvious that another realization of the data (other 241 males) would lead to a slightly different picture.

$\Rightarrow$ The model should take this into account!

**Solution:** Add an error term $e_i$ to the predictor

$$(bodyfat)_i = \alpha + \beta \cdot bmi_i + e_i \ ,$$

where $e_i$ treated as a random variable with a normally distributied

$$e_i \ \sim \mathrm{N}(0, \sigma_e^2) \quad \text{for } i = 1, \ldots, n \ .$$

This is a model for *bodyfat* given *bmi*. The assumption is that the target value $bodyfat_i$ is the sum of a predicted value $(\alpha + \beta \cdot bmi_i)$ plus an error term $e_i$.

# The simple linear regression model

Generally:

The linear regression model for the data $\mathbf{y} = (y_1, \ldots, y_n)$ given $\mathbf{x} = (x_1, \ldots, x_n)$ is

$$y_i = \alpha + \beta x_i + e_i \ , \qquad e_i \sim \mathrm{N}(0, \sigma_e^2) \text{ independent.}$$

The assumption is that

$$y_i \quad = \quad \underbrace{\text{prediction}}_{\alpha+\beta x_i} \quad + \quad \underbrace{\text{error}}_{e_i}$$

Note:

- The model for $\mathbf{y}$ given $\mathbf{x}$ has three parameters: $\alpha$, $\beta$ and $\sigma_e^2$ .
- $\mathbf{x}$ is the independent or explanatory variable.
- $\mathbf{y}$ is the dependent or outcome variable.

This is a general approach in statistics:

- Formulate a model and modelling assumptions that seem plausible for your data. A model emerges in our mind.
- Estimate the parameters.
- Only now it is a *specific* model.

# Visualization of regression assumptions

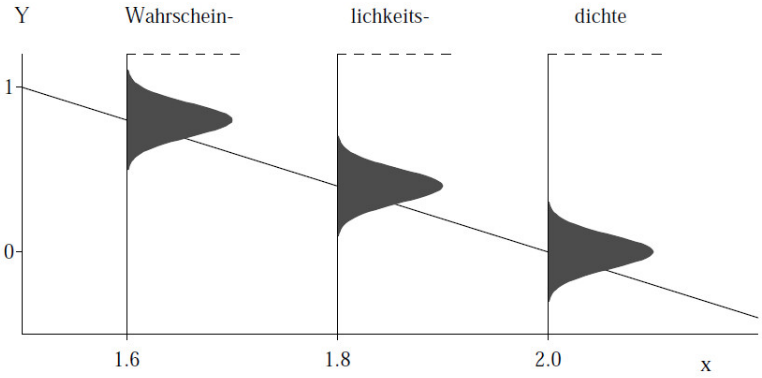


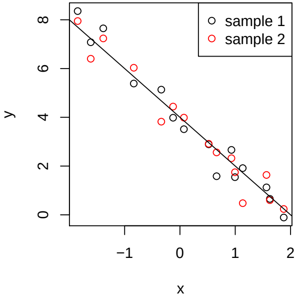

Abbildung 2.1.h: Veranschaulichung des Regressionsmodells $Y_i = 4 - 2x_i + E_i$ für drei Beobachtungen $Y_1$, $Y_2$ und $Y_3$ zu den $x$-Werten $x_1 = 1.6$, $x_2 = 1.8$ und $x_3 = 2$

# Insight from data simulation

(Simulation are *always* a great way to understand statistics!!)

Generate an independent (explanatory) variable **x** and **two** samples of a dependent variable **y** assuming that

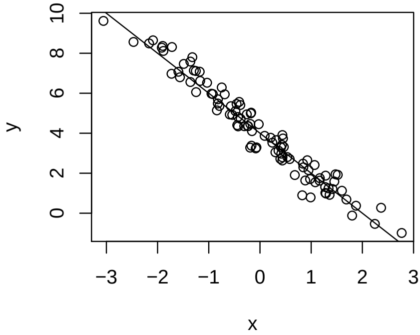
$$y_i = 4 - 2x_i + e_i \ , \quad e_i \sim \text{N}(0, 0.5^2) \ .$$




Note the different y-coordinates for the two samples.

Or one larger sample:

```
> par(mfrow=c(1,1))
> x <- rnorm(100)
> y <- 4 - 2*x + rnorm(100,0,sd=0.5)
> plot(x,y);abline(c(4,-2))
```




Random variation is always present. This leads us to the next question.

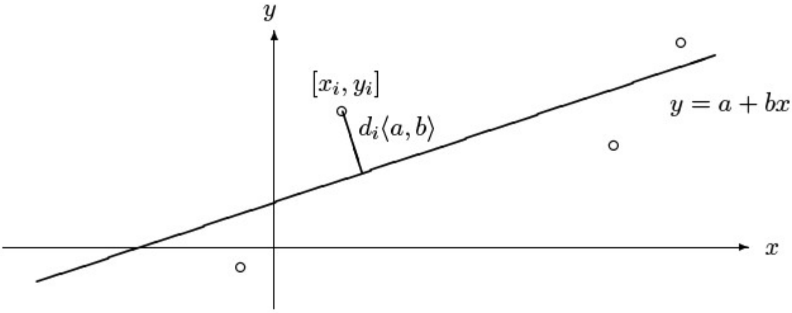
# Parameter estimation

Remember: There are three parameters $\alpha$, $\beta$ and $\sigma_e^2$ that want to be estimated.

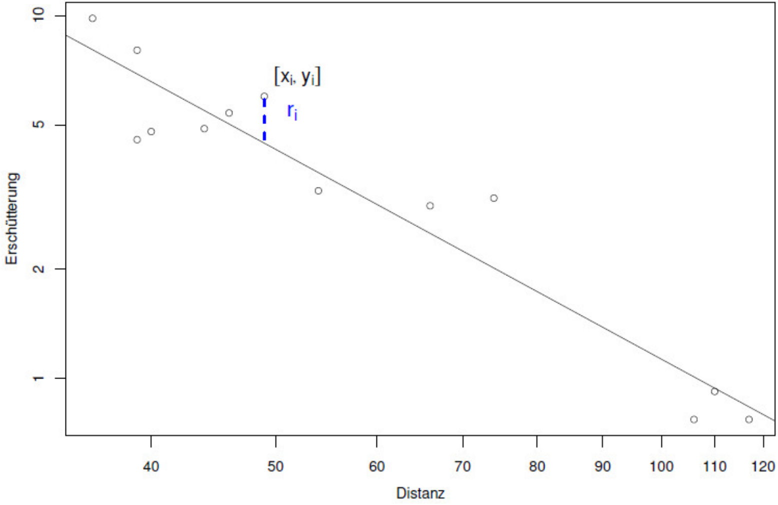
- **Problem:** For more than two points there is generally no perfectly fitting line.
- **Aim:** We want to find the best fitting line.
- **Idea:** Minimize the deviations between the points and the line.

But how?

# Should we minimize these distances...



# … or these?



10
5
2
1
Erschütterung
$[x_i, y_i]$
$r_i$
40
50
60
70
80
90
100
110
120
Distanz

# Least squares

For multiple reasons (theoretical aspects and mathematical convenience), the parameters are estimated using the least squares approach. In this, the second type of distances are minimized:

The parameters are estimated such that the sum of squared vertical distances

$$\sum_{i=1}^{n} r_i^2 \ , \qquad r_i = y_i - (\alpha + \beta x_i)$$

is being minimized.

Note: The vertical deviations $r_i$ are called the residuals.

# Formulas for regression line parameters

$$
\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{cov(\mathbf{x},\mathbf{y})}{var(\mathbf{x})} \\
\hat{\alpha} &= \overline{y} - \hat{\beta}\overline{x} \\
\hat{\sigma}_e^2 &= \frac{1}{n-2}\sum_{i=1}^{n} R_i^2 \quad \text{with residuals } R_i = y_i - (\hat{\alpha} - \hat{\beta}x_i)
\end{aligned}
$$

The hat on the parameters ($\hat{\alpha}$,$\hat{\beta}$,$\hat{\sigma}_e$) indicates that these are estimates.

(The derivation of the parameters can be looked up in the Stahel script 2.A b.
Idea: Minimization through derivating equations and setting them =0.)

# 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/

# Estimation using R

Let's estimate the regression parameters from the bodyfat example

```
> r.bodyfat <- lm(bodyfat ~ bmi,d.bodyfat)
> summary(r.bodyfat)
```

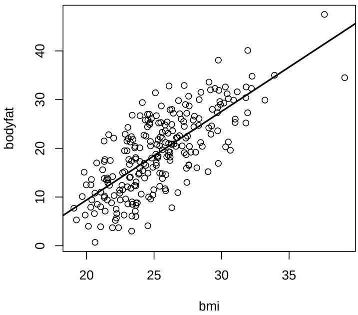
```
Call:
lm(formula = bodyfat ~ bmi, data = d.bodyfat)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5485  -3.5583   0.0785   4.0384  12.7330

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -26.9844     2.7689  -9.746   <2e-16 ***
bmi           1.8188     0.1083  16.788   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.573 on 241 degrees of freedom
Multiple R-squared:  0.539,       Adjusted R-squared:  0.5371
F-statistic: 281.8 on 1 and 241 DF,  p-value: < 2.2e-16
```

$\Rightarrow \hat{\alpha} = -26.98$ , $\hat{\beta} = 1.82$, $\hat{\sigma}_e = 5.57$.

Interpretation: for an increase in the bmi by one index point, roughly 1.82% percentage points more bodyfat are expected.

Plotting the resulting line into the scatterplot is simple:

```
> plot(bodyfat ~ bmi,d.bodyfat)
> abline(r.bodyfat,lwd=2)
```



# Are the modellign assumptions met?

Before we continue to look into the results, we need to check if the modelling assumptions are met!

Why? Because otherwise we draw invalid conclusions from the results.

The assumption we took here is that the errors $e_i \sim \mathrm{N}(0, \sigma_e^2)$. This implies four things:

a) The expected value of $e_i$ is 0: $\mathrm{E}(e_i) = 0$.

b) All $e_i$ have the same variance: $\mathrm{Var}(e_i) = \sigma_e^2$.

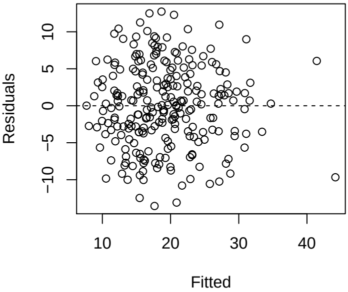
c) The $e_i$ are normally distributed.

d) The $e_i$ are independent of each other.

For the moment, we introduce two simple graphical model checking tools.

# Model checking tools I: Tukey-Anscombe diagram

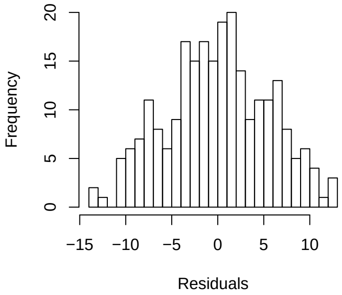
The Tukey-Anscombe diagram plots the residuals against the fitted values:




This plot is ideal to check if assumptions a) and b) (and partially d)) are met. Here, this seems fine.

# Model checking tools II: Histogram

Look at the histogram of the residuals:




The normal distribution assumptions seems ok as well.

# Uncertainty in the estimates $\hat{\alpha}$ and $\hat{\beta}$

Let us look again at the regression output, this time only for the coefficients:

```
> 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
```

The second column shows a standard error of the estimate. The estimates thus seem to contain uncertainty!

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

# Distribution of $\hat{\alpha}$ and $\hat{\beta}$

Again, a simulation can help to get an idea. We generate data points according to the model

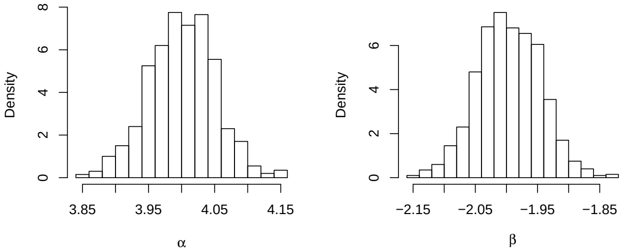
$$y_i = 4 - 2x_i + e_i \ , \quad e_i \sim \text{N}(0, 0.5^2).$$

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

```
> 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 this 1 000 times, we obtain the following distributions for $\hat{\alpha}$ and $\hat{\beta}$.

```
> par(mfrow=c(1,2))
> hist(pars[,1],xlab =expression(alpha),main="",freq=F,nclass=20)
> hist(pars[,2],xlab=expression(beta),main="",freq=F,nclass=20)
```




This looks suspiciously normal...

In fact, from theory:

$$\hat{\beta} \sim \mathsf{N}(\beta, \sigma^{(\beta)2}) \qquad \text{and} \qquad \hat{\alpha} \sim \mathsf{N}(\alpha, \sigma^{(\alpha)2})$$

The standard deviations $\sigma^{(\beta)2}$ and $\sigma^{(\alpha)2}$ are defined as

$$\sigma^{(\beta)2} = \sigma_e^2 / SSQ^{(X)} \qquad \sigma^{(\alpha)2} = \sigma_e^2 \left( \frac{1}{n} + \overline{x}^2 / SSQ^{(X)} \right)$$

with the sum-of-squares for $X$ given as

$$SSQ^{(X)} = \sum_{i=1}^{n} (x_i - \overline{x})^2 \ .$$

(See also Stahel 2.2.h)

Don't worry, you do not need to know these formulas by heart!

**You should know that**

- the parameters estimates $\hat{\alpha}$ and $\hat{\beta}$ are normally distributed.
- the formulas to calculate the variances depend on the residual variance $\sigma_e^2$, the sample size $n$ and $SSQ^{(X)}$.

# How good is the regression model?

This is, per se, a difficult question....

One often considered index is the **coefficient of determination (Bestimmtheitsmass)** $R^2$. Let us again look at the regression output form the bodyfat example:

```
> summary(r.bodyfat)$r.squared
[1] 0.5390391
```

This is the $R^2$ from the regression of bodyfat against bmi.
Compare this to the correlation between the two variables:

```
> cor(d.bodyfat$bodyfat,d.bodyfat$bmi)
[1] 0.7341928
```

... and square it:

```
> cor(d.bodyfat$bodyfat,d.bodyfat$bmi)^2
[1] 0.5390391
```

We conclude:

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

Generally, $R^2$ indicates the proportion of variability of the response variable $\mathbf{y}$ that is explained by the ensemble of all covariates. **The larger $R^2$, the more variability of $\mathbf{y}$ is captured by the covariates, thus the "better" is the model.** (However, we will qualify this statement later in the course...)

$R^2$ becomes more interesting in *multiple* linear regression.

# Testing and Confidence Intervals

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

1. **"Are the parameters compatible with some specific value?"**
   Typically, the question is whether the slope $\beta$ might be 0 or not, that is: "Is there an effect of the covariate **x** 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:

```
> 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

$$H_0 : \quad \beta = \beta_0 = 0 \ .$$

Included in $H_0$ is the assumption that the data follow the simple linear regression model.
(However, you might want to test against another null hypothesis, see Stahel 2.3 a,b).

Here, the *alternative hypothesis* is given by

$$H_A : \quad \beta \neq 0 \ ,$$

Remember: to carry out a statistical test, we need a *test statistic*.

What is a test statistic?? (Clicker exercise here?)

Remember: to carry out a statistical test, we need a *test statistic*.

What is a test statistic?? (Clicker exercise here?)

It is some type of summary statistic that follows a known distribution under $H_0$. For our purpose, we use the so-called $T$-statistic

$$T = \frac{\hat{\beta} - \beta_0}{se^{(\beta)}} \quad \text{with} \quad se^{(\beta)} = \sqrt{\hat{\sigma}_e^2 / SSQ^{(X)}} \ . \tag{1}$$

Again: typically, $\beta_0 = 0$, so the formula simplifies to.... (please think:-))

Under $H_0$, $T$ has a $t$-distribution with $n-2$ degrees of freedom ($n =$ number of data points).
(Question: do you remember why this is a t-distribution? Check Mat183, keyword: t-test.)

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
```

2 tasks:

1. Please use equation (1) to find out how the first three columns (Estimate, Std. Error and t value) are related! Check your ideas by doing some calculations...

2. Then think how we get the fourth column from the third.
   Hint: last column contains the *p*-**value** of the test $\beta = 0$.

For task 2 above we can use the built-in "distribution table" of the t-distribution in R:

```
> 2*pt(16.787522,241,lower.tail=F)
[1] 2.063861e-42
```

Conclusion: there is very strong evidence that the BMI is associated with bodyfat, because $p$ is extremely small.

This basically answers question 1 from slide 30.

(Remark: if you forgot the details of the $p$-value, I have a special task for you: go back to the book "Statistische Datenanalyse" from W. Stahel and read chapter 8.7.)

# A cautionary note on the use of $p$-values

Maybe you have seen that in statistical testing, often the criterion $p \leq 0.05$ is used to test whether $H_0$ should be rejected. This is often done in a black-or-white manner.

However, we will put a lot of attention to a more reasonable and cautionary interpretation of $p$-values in this course!

# Confidence intervals of regression parameters

Question 2 from slide 30: "Which values of the parameters are compatible with the data?"

To answer this question, we can determine the confidence intervals of the regression parameters.

Let us collect the facts we know about $\hat{\beta}$:

- $\hat{\beta}$ is estimated with a standard error of $\sigma^{(\beta)}$.
- The distribution of $\hat{\beta}$ is normal, namely $\hat{\beta} \sim \mathsf{N}(\beta, \sigma^{(\beta)2})$.
- However, since we need to estimate $\sigma_e^2$ from the data, we have a $t$-distribution.

Doing some calculations (blackboard) leads us to the 95% confidence interval

$$[\hat{\beta} - c \cdot \sigma^{(\beta)}; \hat{\beta} + c \cdot \sigma^{(\beta)}] \ ,$$

where $c$ is the 97.5% quantile of the $t$-distribution with $n-2$ degrees of freedom.

Doing this for the bodfat example "by hand" is not hard. We have 241 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: all true values for $\beta$ between 1.61 and 2.03 are compatible with the observed data.

# Confidence and Prediction Ranges

- Remember: When another sample from the same population was taken, the regression line would look slightly different.
- There are two questions to be asked:

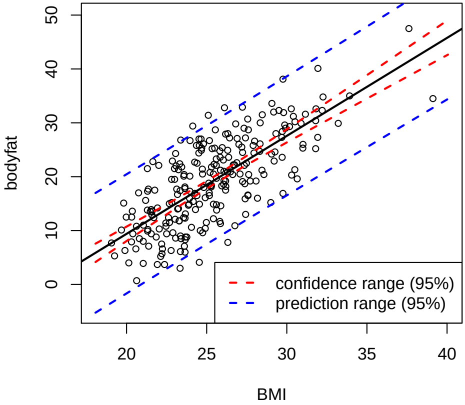
1. Which other regression lines are compatible with the observed data?

   $\Rightarrow$ This leads to the **confidence range**.

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

   $\Rightarrow$ This leads to the **prediction range**.

# Bodyfat example




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

# Calculation of the confidence range

Given a fixed value of $x$, say $x_0$. The question is:

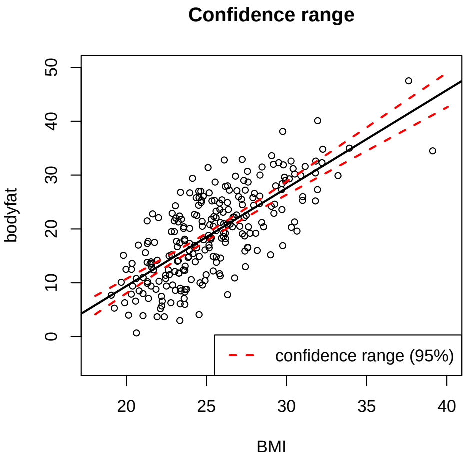
Where does $\hat{y}_0 = \hat{\alpha} + \hat{\beta} x_0$ lie with a certain confidence (i.e., 95%)?

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

The details of the calculation are given in Stahel 2.4b. (Idea: $\hat{y}_0 \pm q \cdot se^{(y_0)}$.)

For *each* $x_0$ one obtains a confidence interval for the expected value $\hat{y}_0 = \hat{\alpha} + \hat{\beta} x_0$. Plotting this interval for all values of $x_0$ one obtains the **confidence range** or **confidence band for the expected values** of $y$.

Note: For the confidence range, only the uncertainty in the estimates $\hat{\alpha}$ and $\hat{\beta}$ are decisive.


Confidence range
bodyfat
BMI
0
10
20
30
40
50
20
25
30
35
40
confidence range (95%)

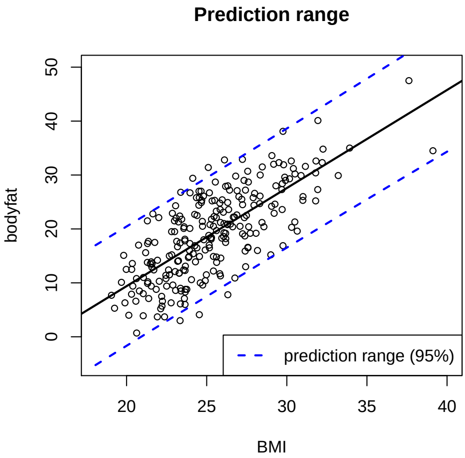
# Calculation of the prediction range

Given a fixed value of $x$, say $x_0$. 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 $\hat{y}_0 = \hat{\alpha} + \hat{\beta} x_0$, but also the error in the equation $e_i \sim \mathrm{N}(0, \sigma_e^2)$.

This is the reason why the prediction range is always wider than the confidence range.


Prediction range
bodyfat
BMI
0
10
20
30
40
50
20
25
30
35
40
prediction range (95%)

# Further reading

“Points of significance: Simple linear regression” is a nice, 2-page overview of simple linear regression, summarized in Nature Methods in 2015:

http://www.nature.com/nmeth/journal/v12/n11/full/nmeth.3627.html

# Summary

To do