Faculty of Health Sciences

Day 3: Univariate linear regression, correlation and regression to the mean

Paul Blanche

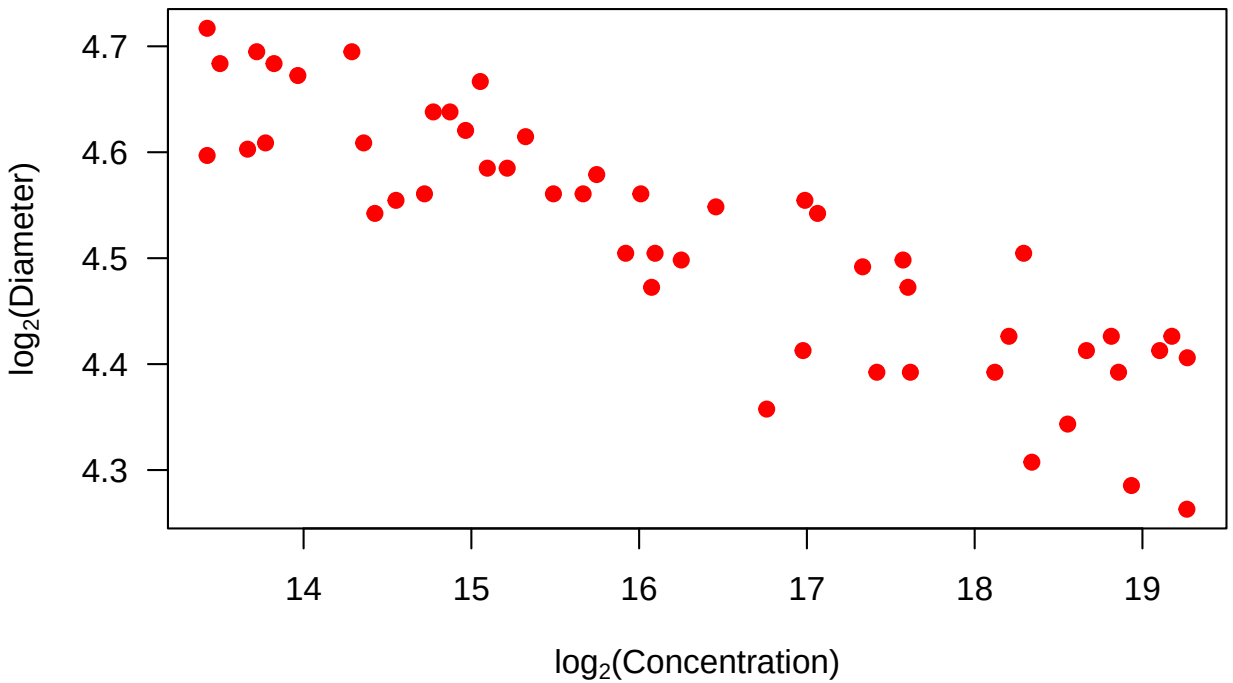
Section of Biostatistics, University of Copenhagen

April 24, 2023

Outline

Case study: Cell cultivation

In an experiment with the unicellar organism tetrahymena, we are interested in determining how cell concentration (n. of cells in 1 mL of the growth media) may affect the cell size (average cell diameter, in μm).



Original Reference: Hellung-Larsen, Leick, Tommerup, Kronborg (1990) Chemotaxis in tetrahymena. Europ J Prostitol 25:229–233.
Statistical Textbook: Andersen & Skovgaard. Regression with linear predictors. Springer, 2010.

Remarks on the case study and log-transformation

- It is common, and often sensible, to log-transform some data, to analyze them, especially **outcomes** (e.g. concentrations, CD4 counts ..)[1]. It is less common to transform predictors, but not unusual and sometimes useful or even necessary. [2]
- We will log-transform in our case study:

$$\text{outcome} = \log_2(\text{Diameter})$$
$$\text{predictor} = \log_2(\text{Concentration})$$

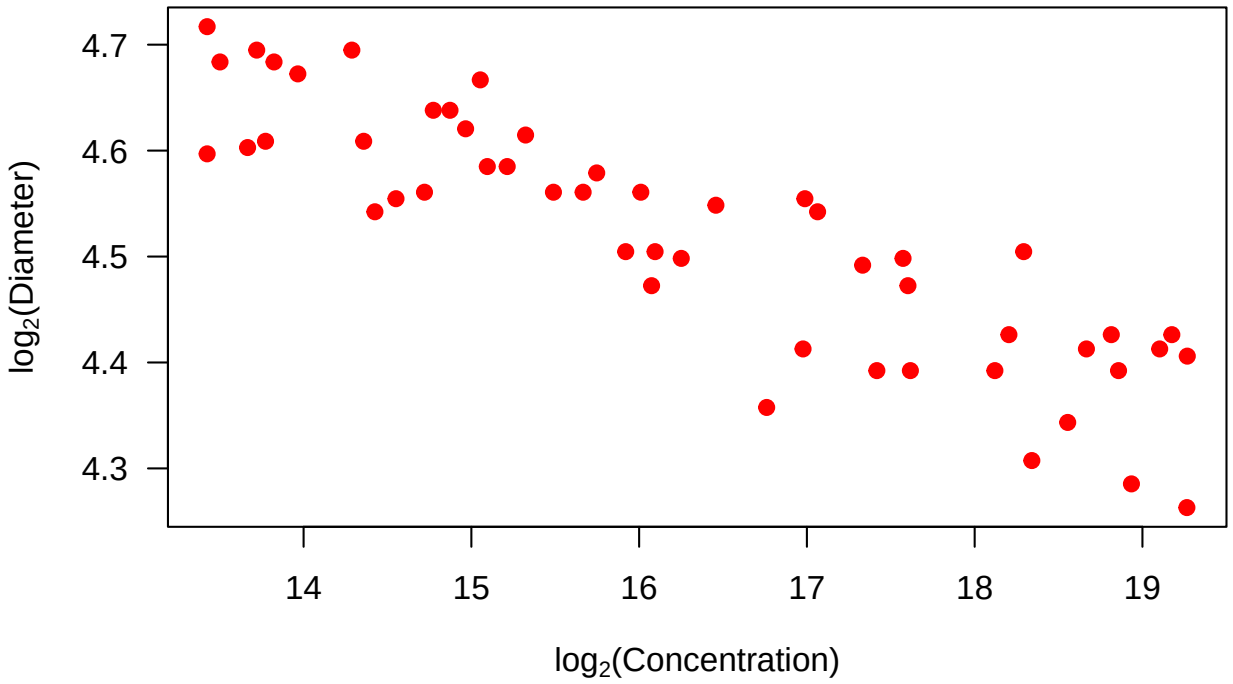
- But, it is not always needed and important to log-transform!

DO NOT SYSTEMATICALLY LOG-TRANSFORM WITHOUT A GOOD REASON!

- It is best to pre-specify the choice of transforming or not based on background knowledge (i.e. your experience of that of others reported in the literature).

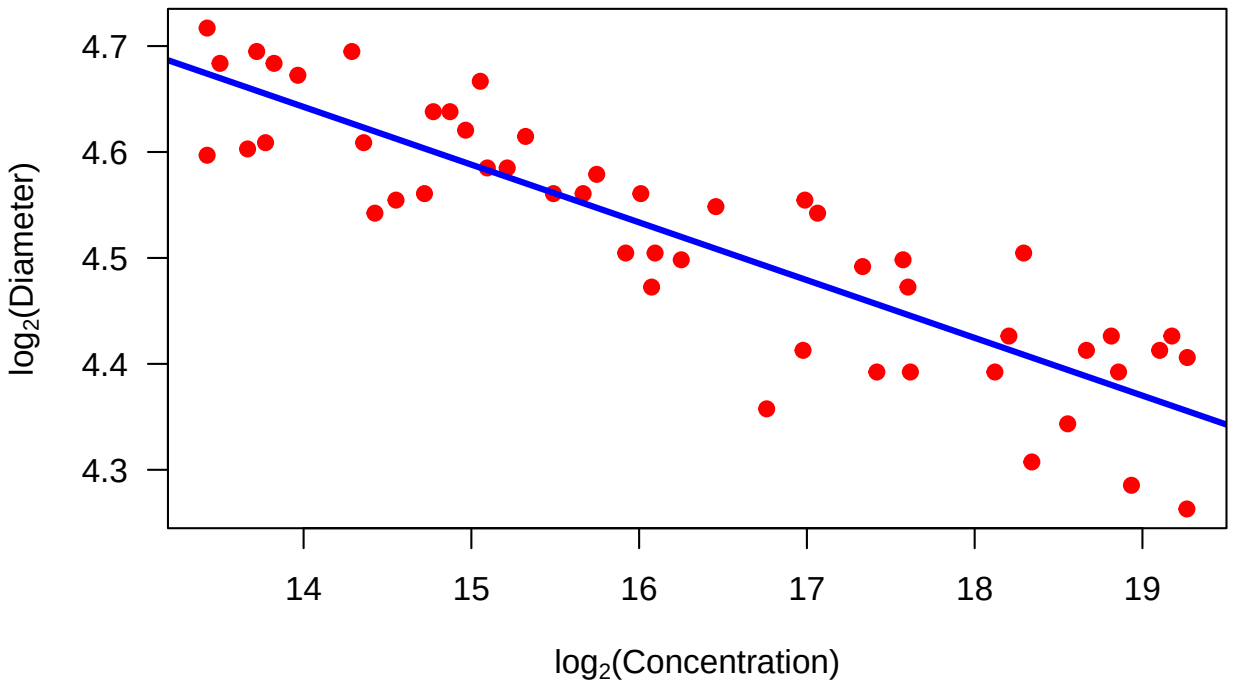
[1] See e.g. Bland & Altman. "Statistics notes: Transforming data." BMJ 312.7033 (1996): 770; and also Keene "The log transformation is special." Statistics in Medicine 14.8 (1995): 811-819.

[2] See e.g. Appendix B in Andersen & Skovgaard. Regression with linear predictors. Springer, 2010.



- Is there an **association**?
- How can we **describe** it?
- How well can we **predict** diameter when we know the concentration?

Same picture with fitted regression line



- Overall the association looks **linear**.
- Even though the line doesn't fit all measurments spot on, the residual variation looks **'random'**.

We will see how to carefully check these 2 **model assumptions**.

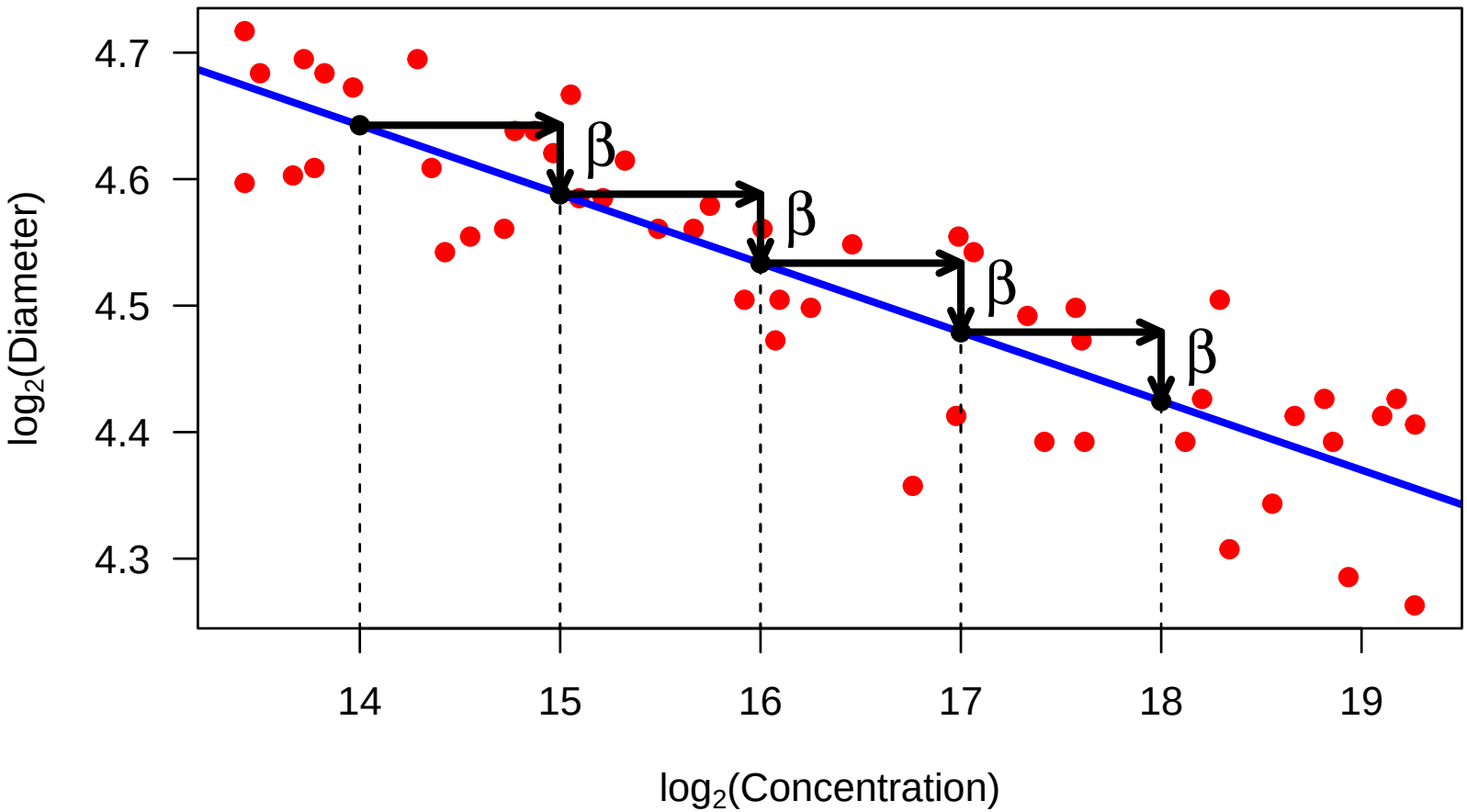
The linear model

$$y = \alpha + \beta x + \varepsilon$$

- y is the response/ouctcome, **in this case:** log2(diameter).
- x is the explanatory variable/predictor, log2(concentration).
- β is the regression coefficient (or **slope**): It tells us how much y increases when x increases by one unit.
- α is the intercept: the expected value of y when $x = 0$. It does not always have a meaningful interpretation.
- ε is an individual 'error' term, assumed **normally distributed** with **zero mean** and standard deviation σ_ε. The standard deviation σ_ε quantifies the **'unexplained'** variation of the outcome y (i.e. the differences in outcome y which is not explained by x).

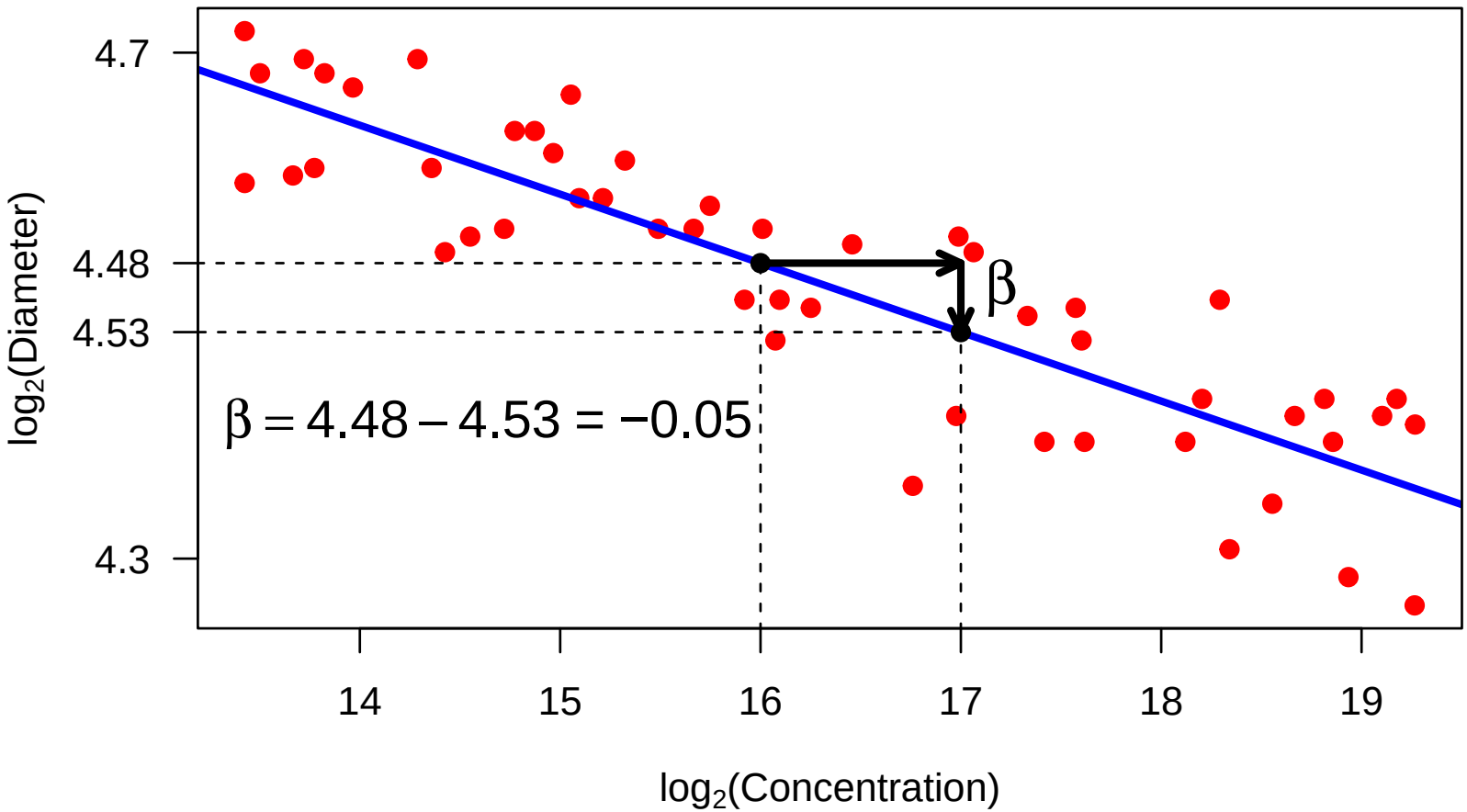
Interpretation of the slope

The slope is the difference in the mean outcome y between subgroups whose difference in their values of x is one unit.



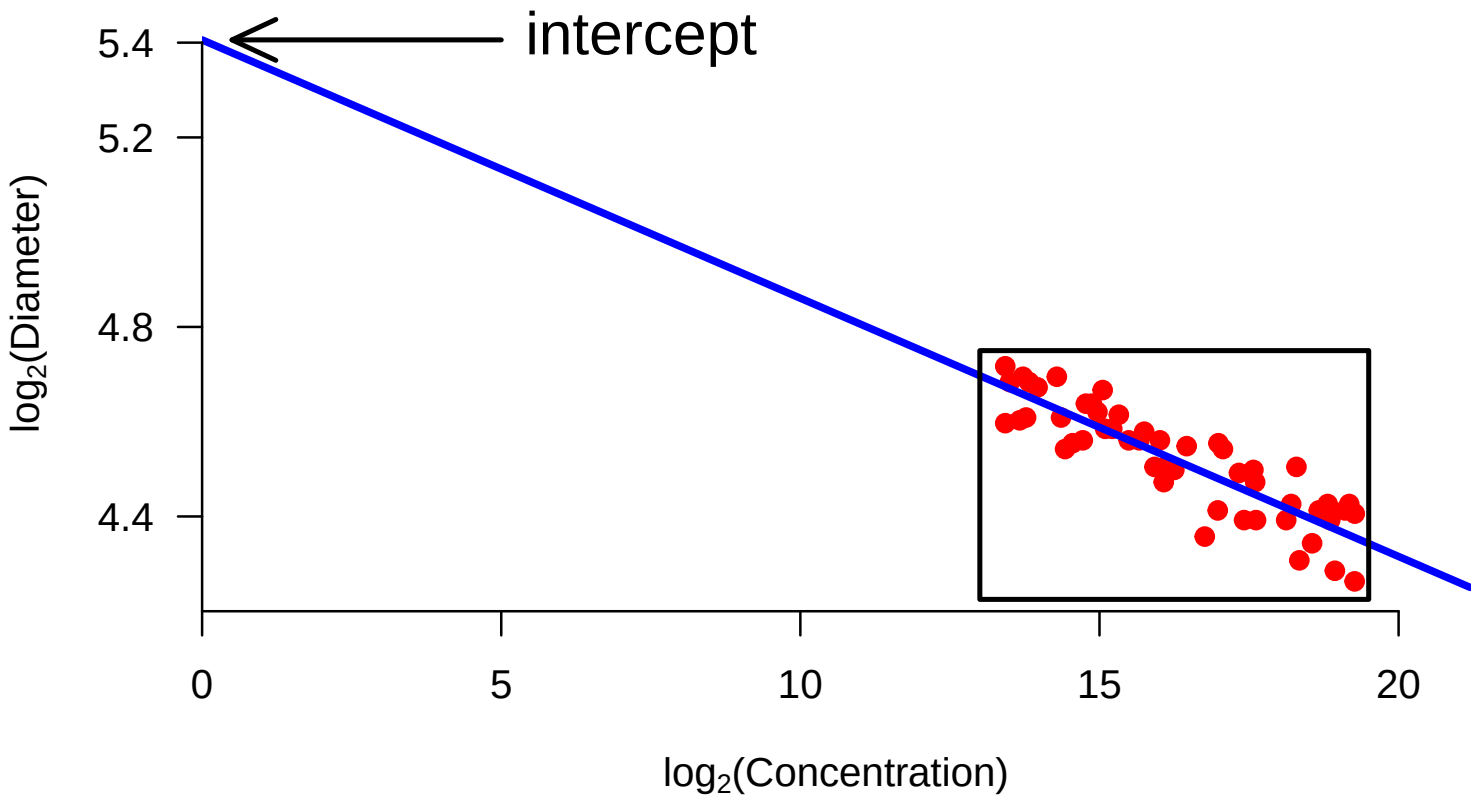
Interpretation of the slope

The slope is the difference in the mean outcome y between subgroups whose difference in their values of x is one unit.



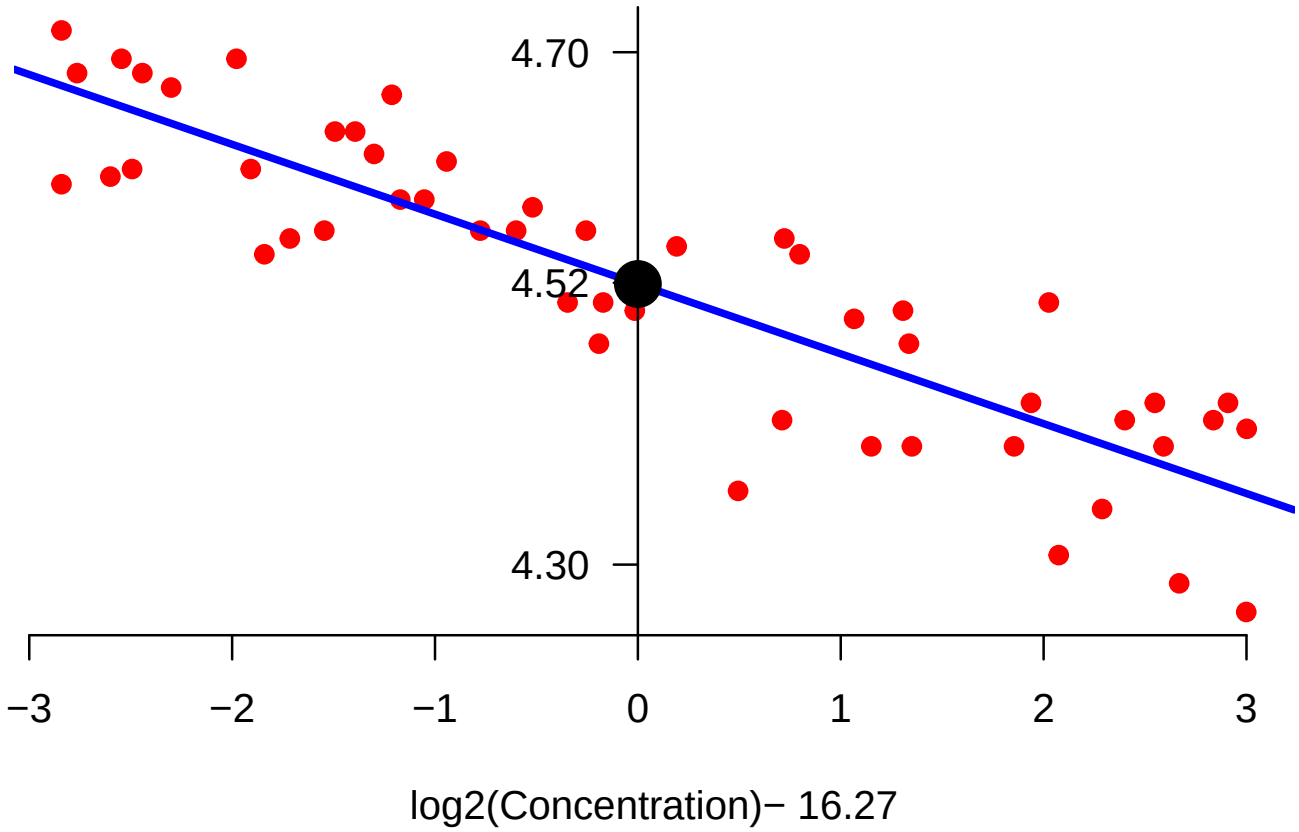
Interpretation of the intercept

The intercept is the expected (fitted) value of y when $x = 0$. Here it does not have a meaningful interpretation: the average diameter when there is only $2^0 = 1$ cell is not meaningful.



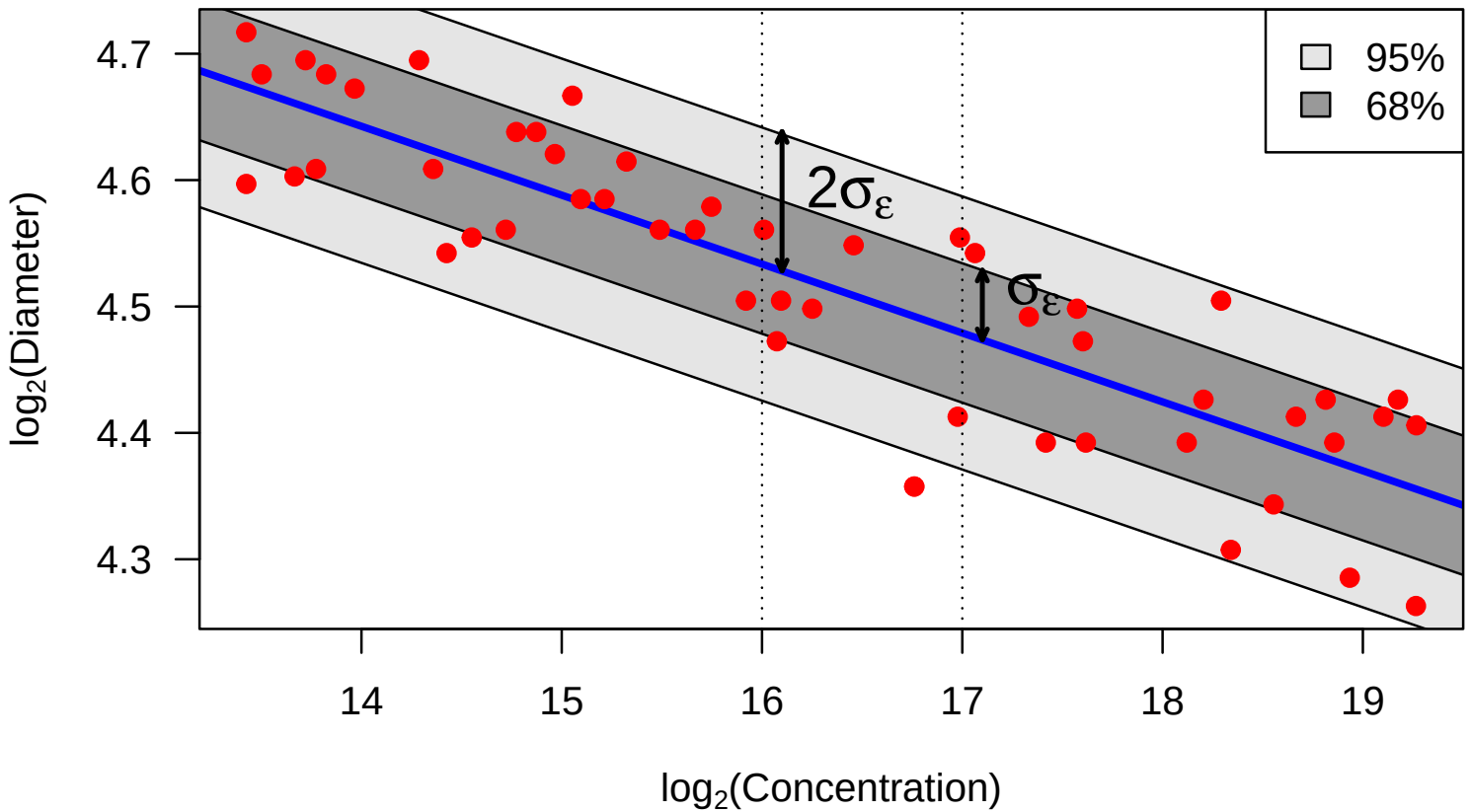
To improve the interpretation of the intercept

After **centering** the exploratory variable, the intercept becomes the expected (fitted) value of y for the average value of x. This is also the average of the outcome y.



Interpretation of the sd of the error term (σ_ε)

The standard deviation of the error term ε, that is σ_ε, tells us how much vertically spread are the points above and below the regression line.



Note: for this example 98% and 67% of the y-coordinates of the red points are within one and two times σ_ε vertical distance from the regression line, respectively.

Outline

How do we find the *best fitting* line? (1/2)

Answer: By the least squares method, which also corresponds to the maximum likelihood method (here).

That is, we find the parameter values $\hat{\alpha}$, $\hat{\beta}$ which minimize

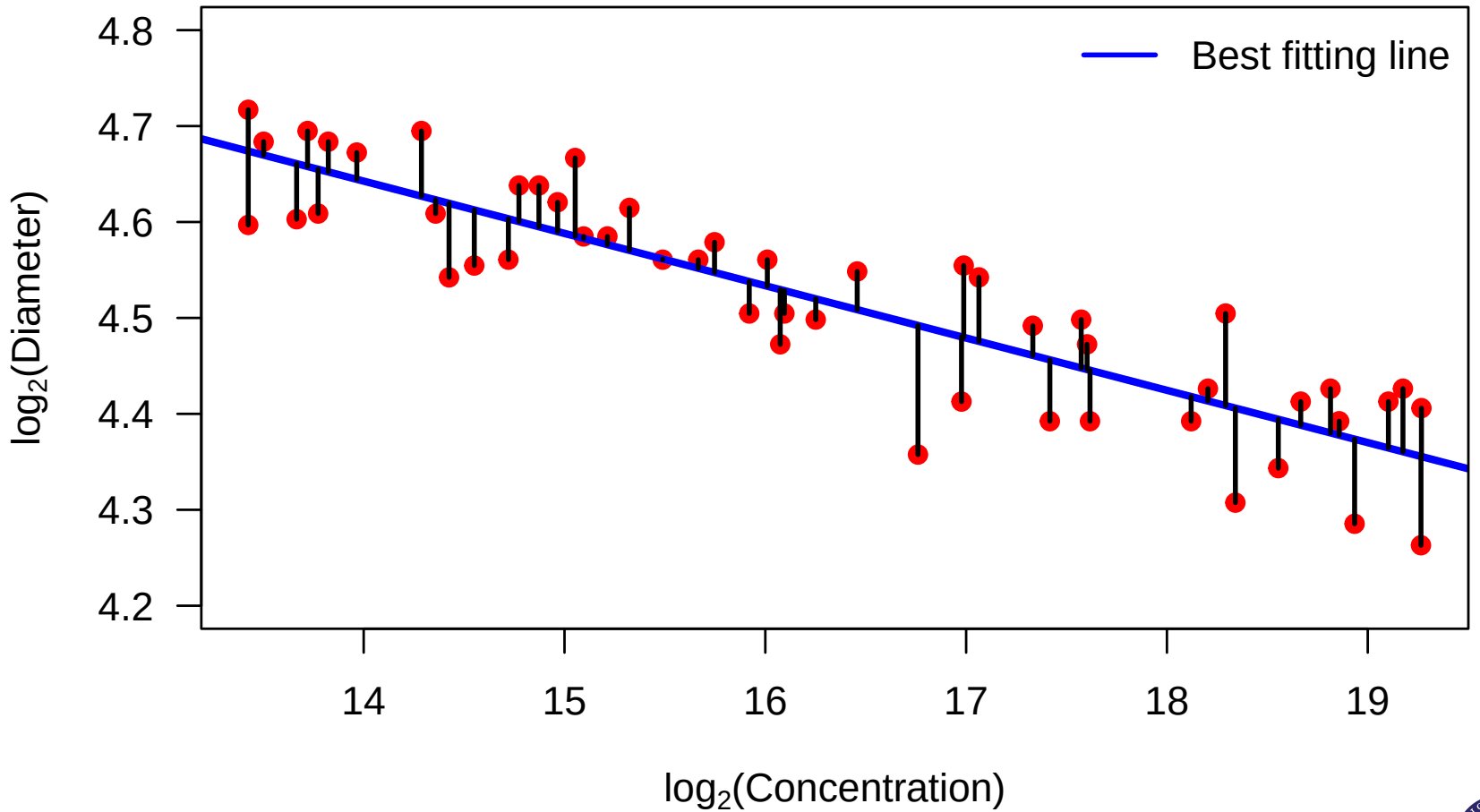
$$\sum_{i=1}^{n} \Big(y_i - (\alpha + \beta x_i) \Big)^2$$

$$= \sum \Big(\text{observation} - \text{expected from the linear model} \Big)^2$$

Simple formulas exist for computing $\hat{\alpha}$ and $\hat{\beta}$ and their standard error (see appendix), but in practice we use a software like R, of course.

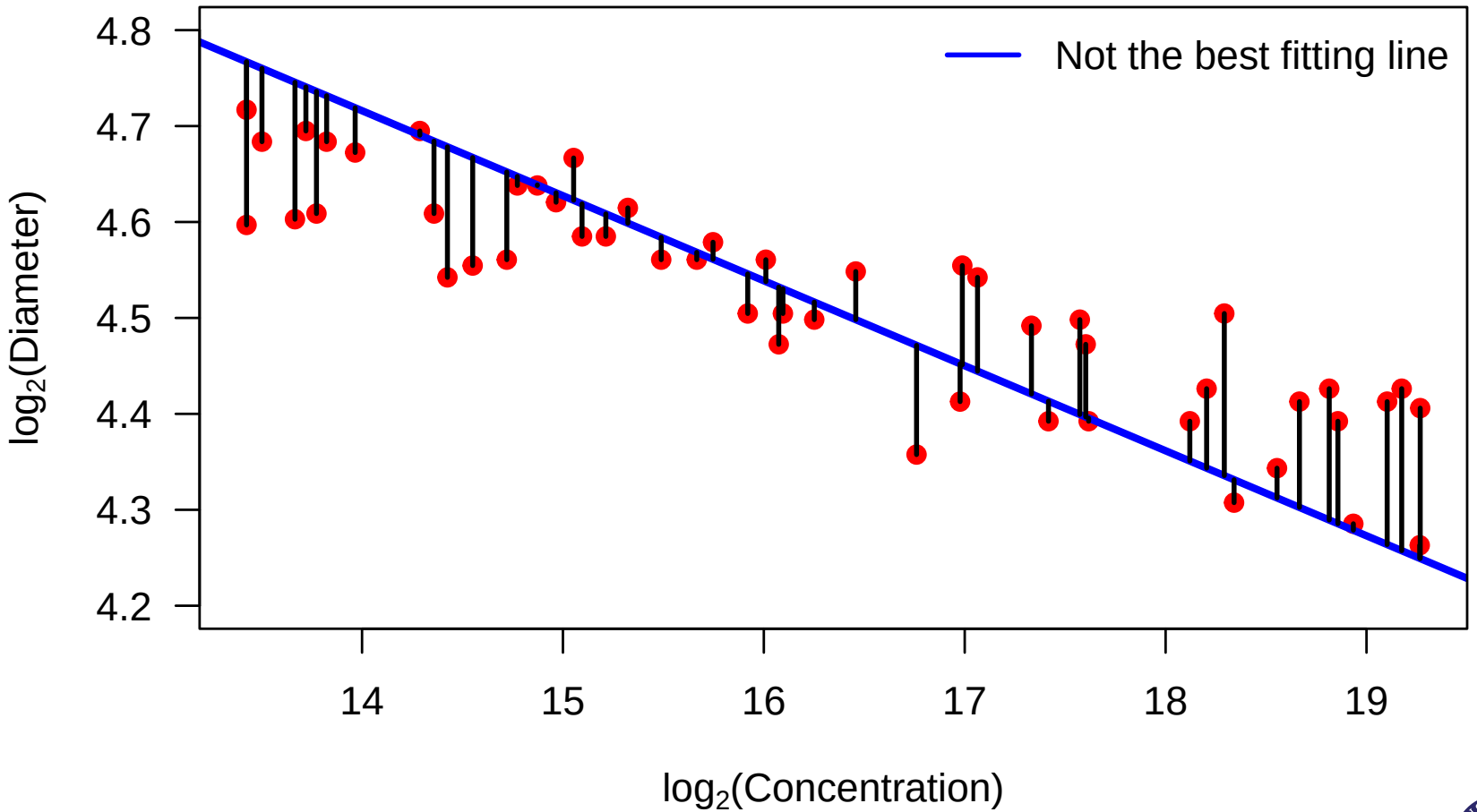
How do we find the *best fitting* line? (2/2)

We minimize sum of the squares of the size of the horizontal bars over all possible blue lines.



How do we find the *best fitting* line? (2/2)

We minimize sum of the squares of the size of the horizontal bars over all possible blue lines.



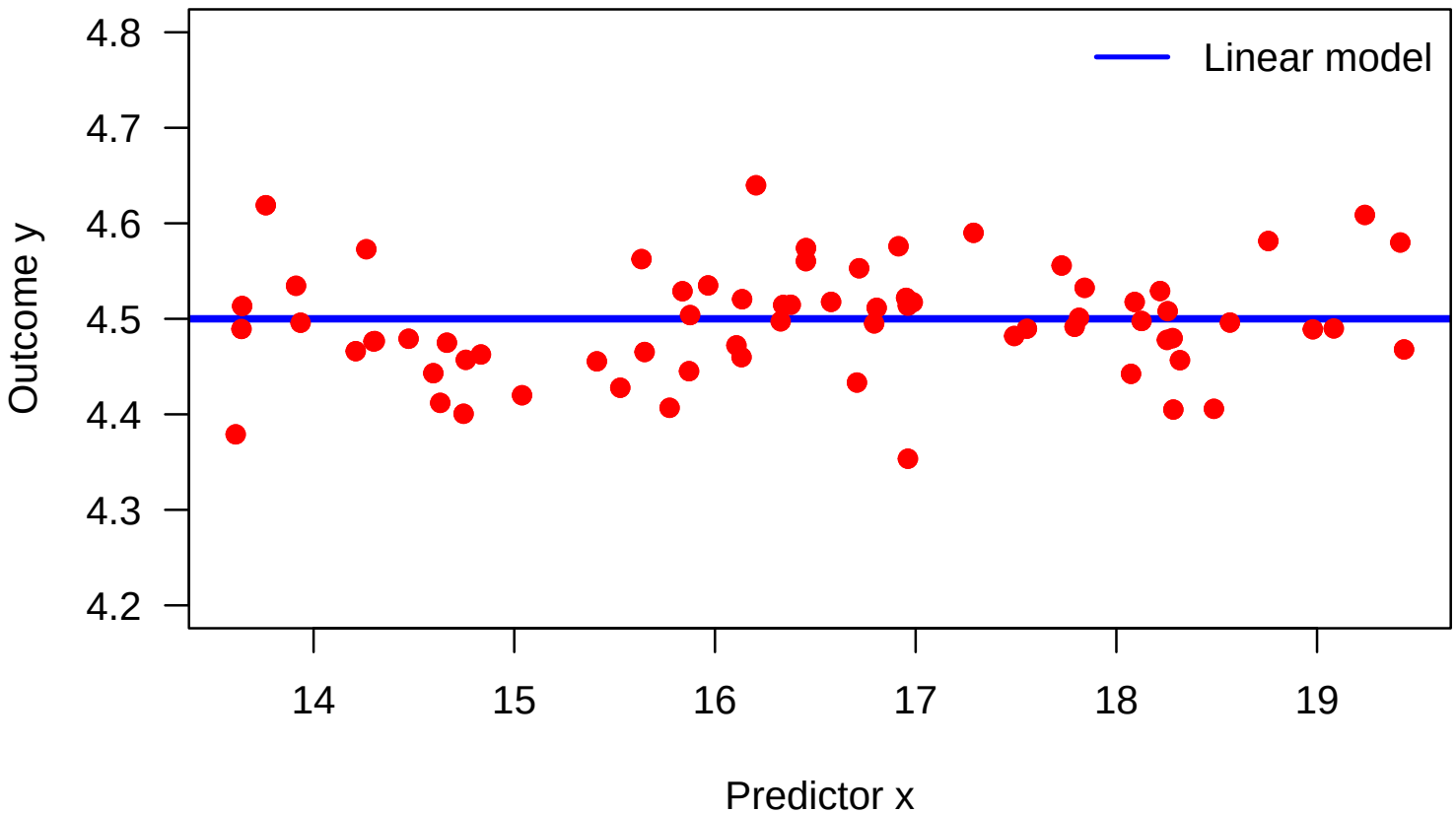
More on model fitting

- The deviations $r_i = y_i - (\hat{\alpha} + \hat{\beta} x_i)$ are called the residuals. They are estimates of the individual 'error' term ε_i.
- Finding the best fitting line is the same as minimizing the residual variance, $s^2 = \frac{1}{n-2} \sum_{i=1}^{n} r_i^2$.
- We estimate the standard deviation of the 'error' term, i.e., σ_ε, by $s = \sqrt{s^2}$.

Quantification and test of association (1/2)

- If **no association** exists between x and y, then the true regression line will be horizontal, that is $\beta = 0$.

Hypothetical example:



Quantification and test of association (2/2)

- We can test the nul hypothesis $\boxed{H_0 : \beta = 0}$ by using:

$$t = \frac{\hat{\beta}}{\text{s.e.}(\hat{\beta})}$$

  which has a t-distribution with $n - 2$ degrees of freedom in case the null hypothesis is true.

- We can get a confidence interval for β from:

$$\hat{\beta} \pm t'_{n-2} \times \text{s.e.}(\hat{\beta})$$

- Inference for α, and especially test for $H_0 : \alpha = 0$, is less often of interest, but otherwise similar.

Case study: inference with R (more in R-demo)

Just run the simple R code:

```
fit <- lm(log2diam~log2conc,data=th)
summary(fit)
```

which returns (among other things):

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.405816   0.068406   79.03   <2e-16 ***
log2conc    -0.054515   0.004178  -13.05   <2e-16 ***

Residual standard error: 0.05514 on 49 degrees of freedom
```

Interpretation : see next slides.

Interpretation (1/2)

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.405816   0.068406   79.03   <2e-16 ***
log2conc    -0.054515   0.004178  -13.05   <2e-16 ***

Residual standard error: 0.05514 on 49 degrees of freedom
```

- `(Intercept)` is $\hat{\alpha}$.
- `log2conc` is the slope $\hat{\beta}$ (i.e. the effect of the predictor).
- Warning: "`Residual standard error`" is maybe not the best chosen term in the R-output: this is acutally the residual standard deviation! (i.e. the estimated value for σ_ε)

Interpretation (2/2)

- The intercept 5.41 **should not** be interpreted here (as already explained).
- There is a significant association between concentration and cell diameter (p-value$<$0.0001).
- We estimate that, in average, $\log 2(\text{diameter})$ decreases by -0.0545 every time $\log 2(\text{concentration})$ increases by one unit.
- Since data was log2-transformed, a better way of saying this is that the median diameter decreases exponentially with an estimated factor $2^{-0.0545} \approx 0.9629$, that is, a decrease by 3.71%, every time the concentration is doubled, with 95% CI= (3.15; 4.27). [3]

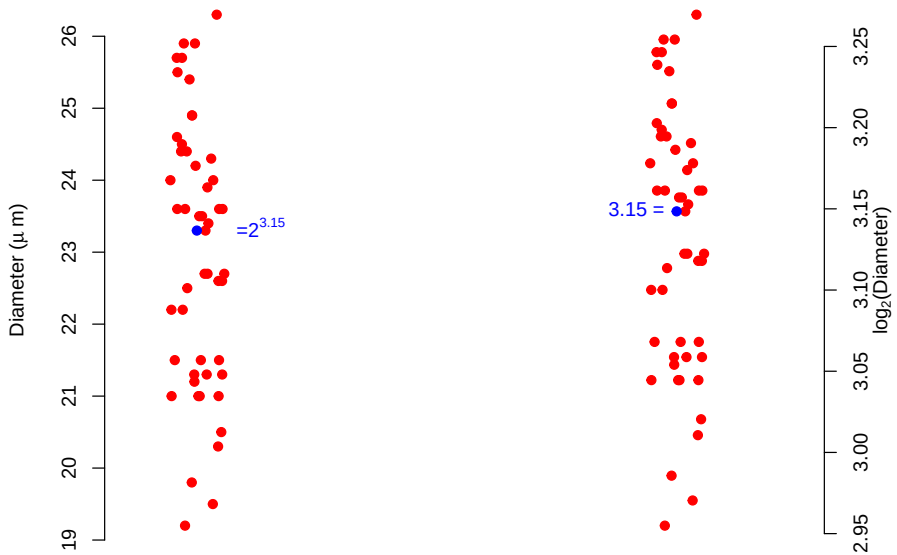
[3] See R-demo for computational details. Also, see more about that on Lecture 7, especially why we can also interpret the median decrease as a mean decrease, even though we model medians on the original scale...

Why a median change interpretation?

- When the data are normally distributed, the mean is the same as the median.

- Hence we have,
$$\text{median}\{\log_2(d)\} = \text{mean}\{\log_2(d)\} = \hat{\alpha} + \hat{\beta} \cdot \log_2(c)$$
$$\Leftrightarrow \quad 2^{\text{median}\{\log_2(d)\}} = 2^{\hat{\alpha}} \cdot c^{\hat{\beta}}$$
$$\Leftrightarrow \quad \text{median}(d) = 2^{\hat{\alpha}} \cdot c^{\hat{\beta}}$$

- Since, $2^{\text{median}\{\log_2(d)\}} = \text{median}(d)$, as only the ranking matters. [4]



[4] Note: this equation is NOT true with "mean" instead of "median".

Why an exponential decrease?

"The median diameter changes with an estimated factor of $2^{\hat{\beta}}$ every time the concentration is doubled"

. . . because according to the linear model we have the following relationship between a median diameter d and a concentration c,

$$\text{median}(d) = 2^{\hat{\alpha}} \cdot c^{\hat{\beta}}$$

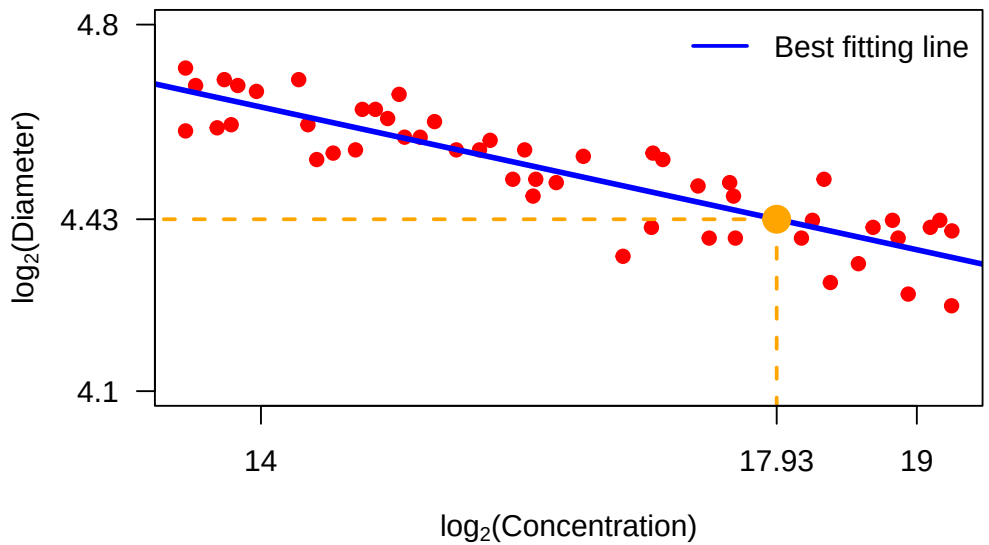
Hence, the diameter for a concentration which is doubled, say $c' = 2c$, is

$$\approx 2^{\hat{\alpha}} \cdot (c')^{\hat{\beta}} = 2^{\hat{\alpha}} \cdot (2c)^{\hat{\beta}} = 2^{\hat{\alpha}} \cdot 2^{\hat{\beta}} \cdot c^{\hat{\beta}} = 2^{\hat{\beta}} \cdot \underbrace{2^{\hat{\alpha}} \cdot c^{\hat{\beta}}}_{=d} = 2^{\hat{\beta}} d \, .$$

Outline

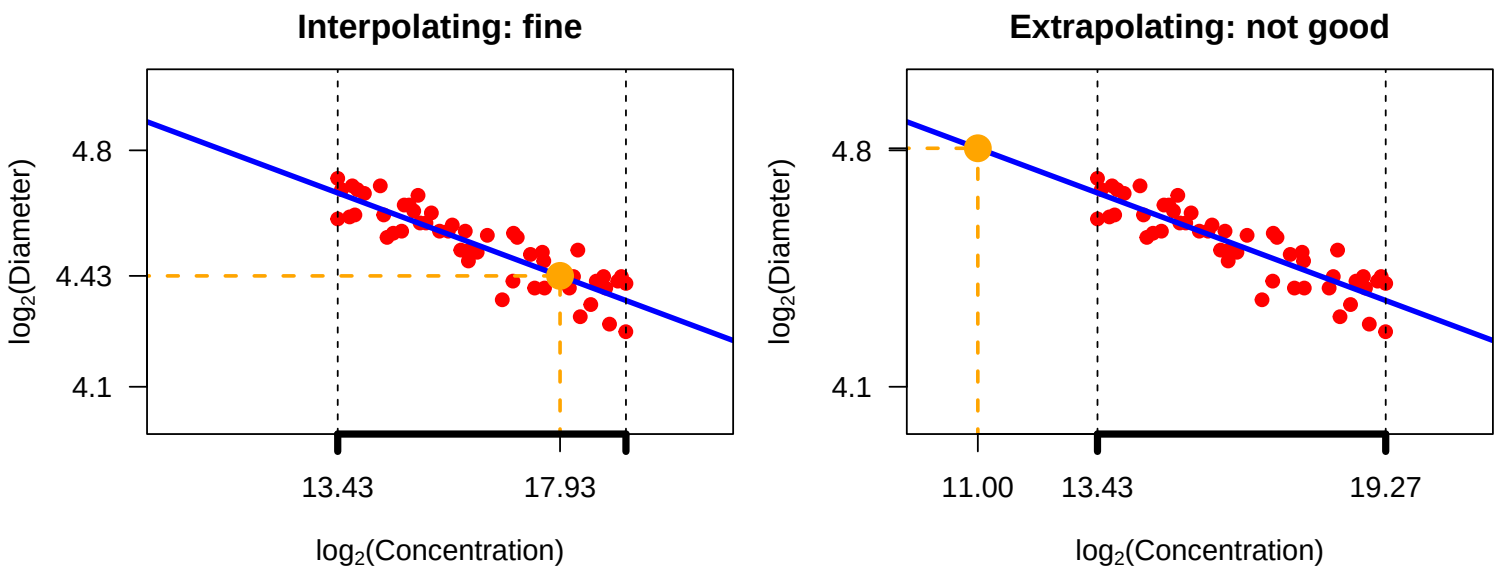
Predicted values (1/2)

- To predict what value of y we can expect for a specific value of x, we plug in to the estimated regression equation, i.e. $\hat{y}(x_0) = \hat{\alpha} + \hat{\beta}x_0$.
- **Example:** For a concentration of 250000 cells/ml, we have $x_0 = \log_2(250000) = 17.93$, hence we would expect a $\log 2$-diameter of $5.41 - 0.0545 \times 17.93 = 4.43$, i.e. a diameter around $2^{4.43} = 21.53\mu m$.



Predicted values (2/2)

- Note that interpolating between the concentrations observed in the experiment is usually 'safe/valid'.
- Extrapolating beyond the range of the observations is usually '**not safe/valid**'. You need convincing subject matter expertise to justify that it makes sense.



Prediction intervals

However, not all responses are on the average.

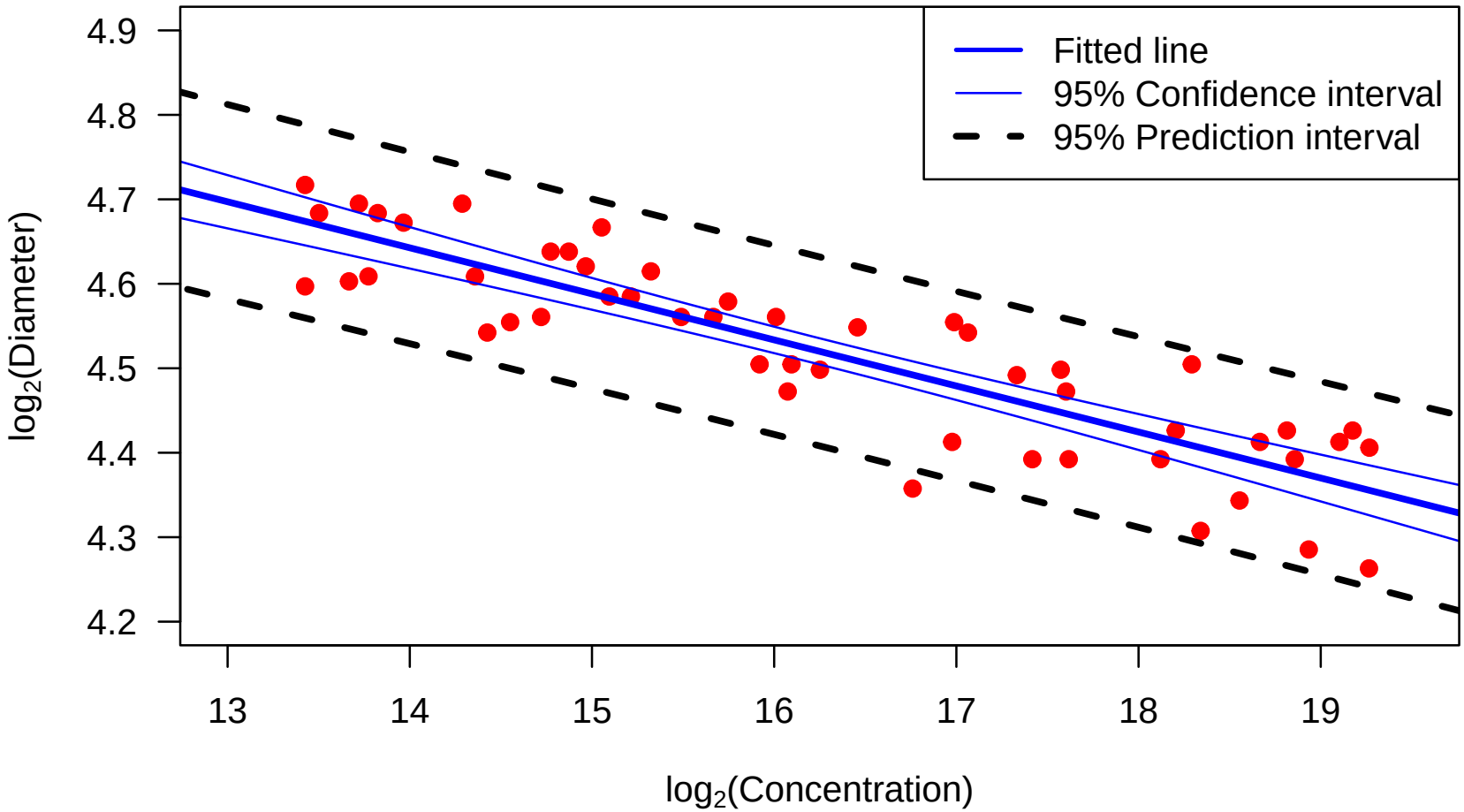
- **Remember:** 95% of individual responses vary within $\pm 1.96\sigma_\varepsilon$ (and 68% within $\pm\sigma_\varepsilon$), where σ_ε is the standard deviation of the error terms (because of the assumption of the normal distribution of ε).

There are **two sources of uncertainty** we need to consider to compute prediction intervals:

1. The statistical uncertainty in our predicted value, which we estimate by a standard error. This is **small for large sample sizes**.
2. The natural variation in the responses (i.e. unexplained variation), which we estimate by the residual standard deviation s. This has **nothing to do with the sample size** and this is usually large even with large sample sizes.[5]

[5] Do you remember "SE vs SD" in Lecture 1? This is similar.

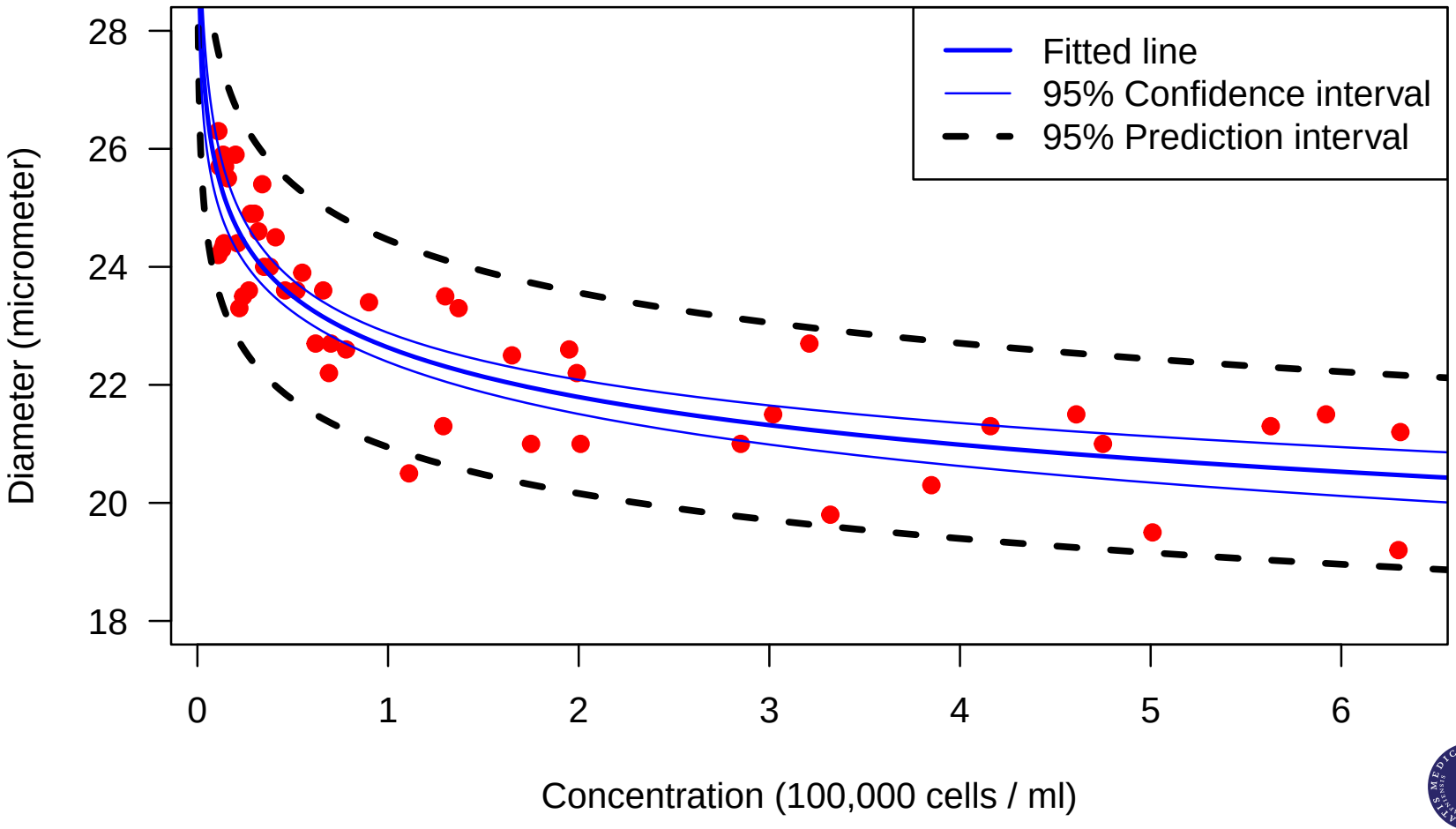
Confidence vs prediction interval



Note: confidence intervals are narrower for predicted values closer to the average of the predictor variable. See appendix for formulas, R-demo for code.

A maybe nicer picture

$$\text{Estimated median diameter} = 2^{5.41} \cdot \text{Concentration}^{-0.0545}$$



Outline

Model assumptions

The statistical model assumed by the linear regression analysis is:

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

where the error terms ε_i describes the individual deviations from the regression line, assumed to be random, normally distributed with mean 0 and standard deviation σ_ε.

The **model assumptions** (1,2 & 4 are important, 3 not always):

1. Observations are independent (no pairing and clustering).
2. The *true association* is linear.
3. The error terms, ε's, are normally distributed.
4. The error terms, ε's, have the same standard deviation, regardless of the value of x.

... should be checked.

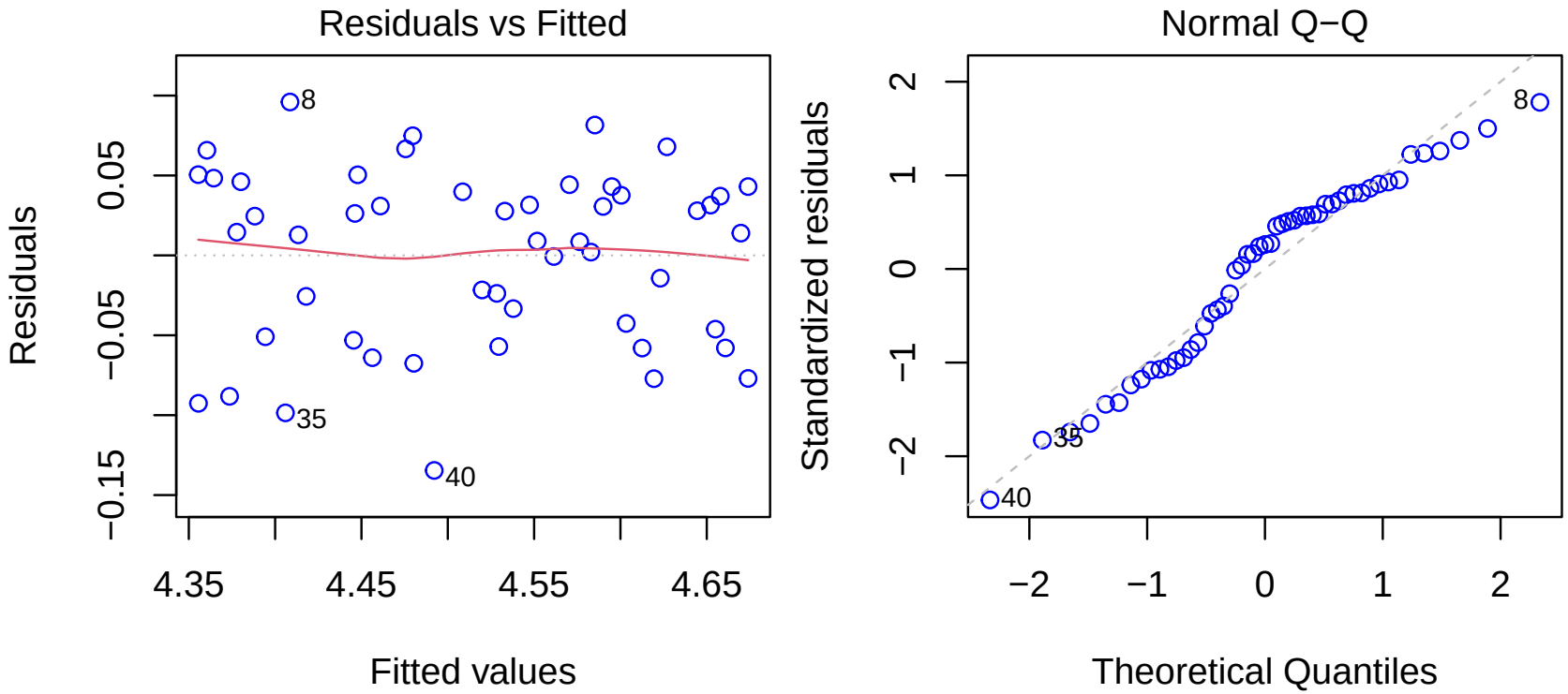
What, when and how should we check?

1. **Independence** should be ensured by the study design.
2. **Linearity** is checked in a residualplot, that is a scatterplot of the residuals against the fitted values (i.e. the predicted values).
3. **Normal distribution** is checked by making a QQplot of the standardized residuals.
4. **Homogenity of variance** is assessed from the residualplot.

IMPORTANT: it is not necessary that the error terms are normally distributed if the sample size is large. Confidence intervals and tests are valid as long as the other assumptions hold. However, prediction intervals are not valid if error terms are not normal.[6]

[6] Similar to remarks for CI and prediction interval for a mean seen in Lecture 1.

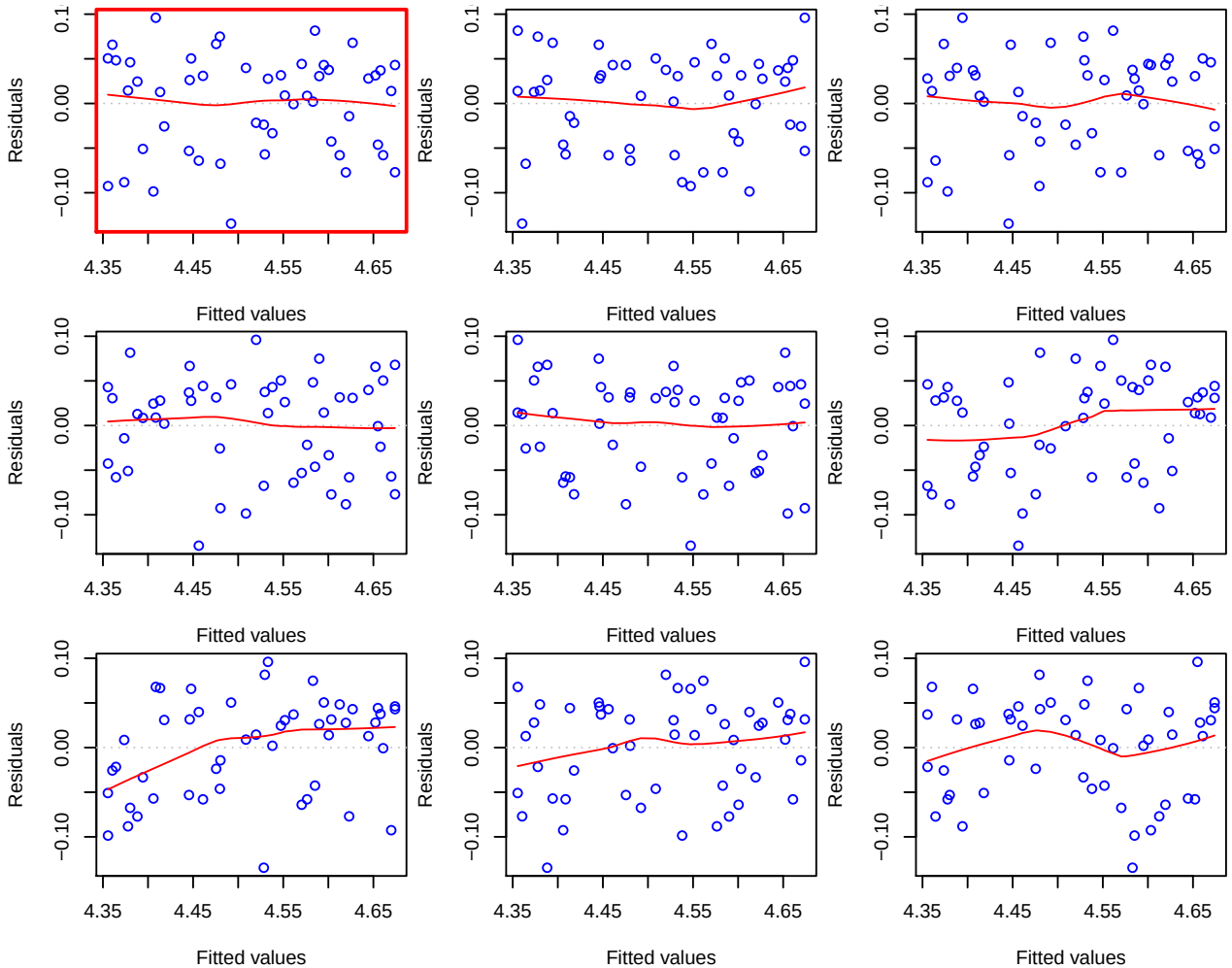
Case study: residual- and QQplot



Note: The points in the residual plot (left) should be randomly and symmetrically scattered around the zero-line with the same variability across the range of fitted values.

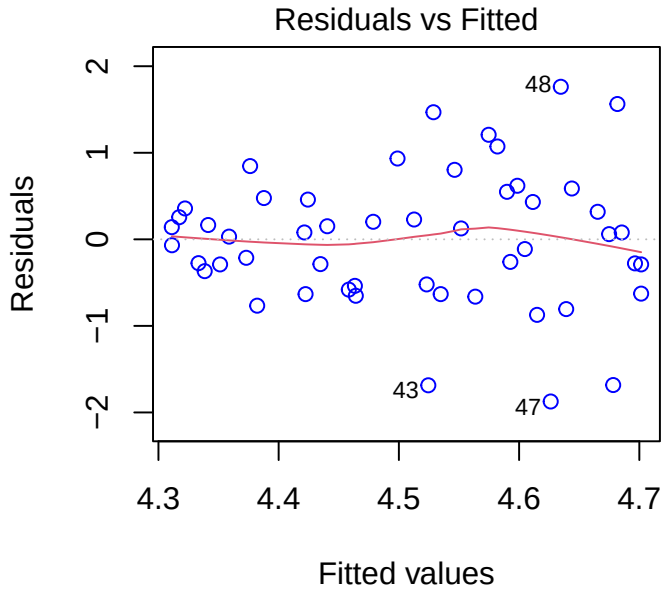
Here everything quite good: we cannot really see more deviations than those 'expected' due to random variation (small sample size).

Wally plot to "visualize" random variation



Note: nine comparison plots produced by randomly permutting the y-axis coordinates (if the model holds, there is no association between 'Residuals' and 'Fitted values').

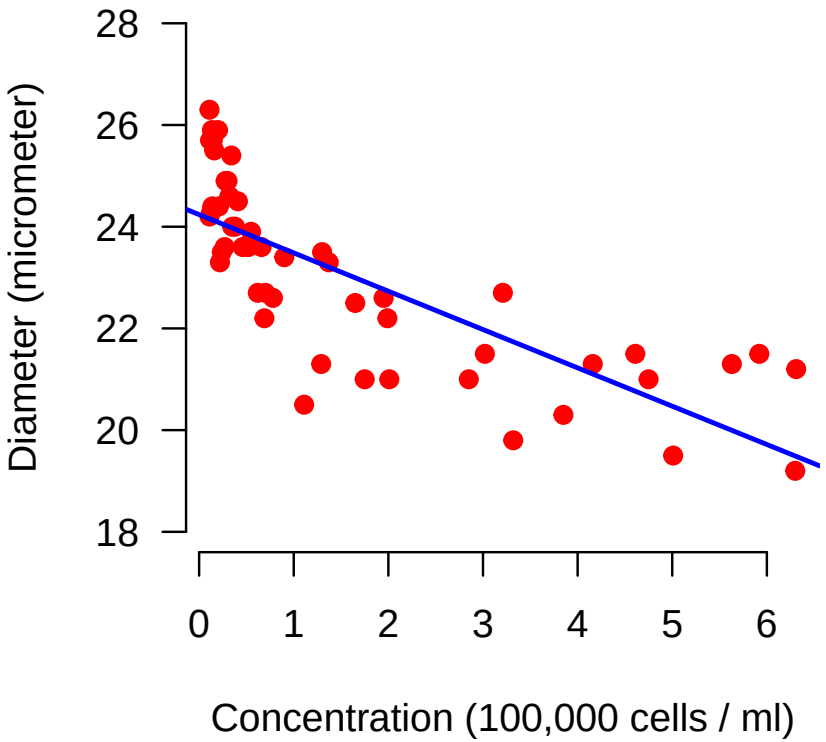
'Typical' example of a problematic residual-plot (hypothetical data)

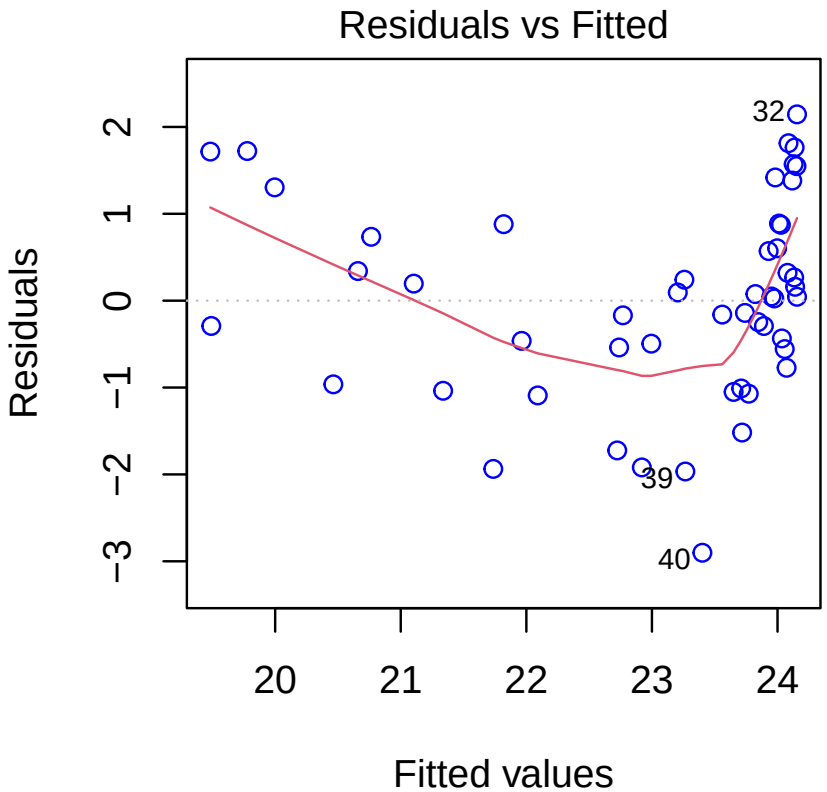


- **NOT the same variability across the range of fitted values.**
- Higher variability for larger fitted values.
- Appropriate transformation of the data can often prevent this. [7]
- Be careful to not overinterpret with smallish sample sizes. Random variation will result in plots having a random pattern in that case.
  → Wally plot can help preventing overinterpretation.

[7] See e.g. Bland & Altman. "Statistics notes: Transforming data." BMJ 312.7033 (1996): 770.

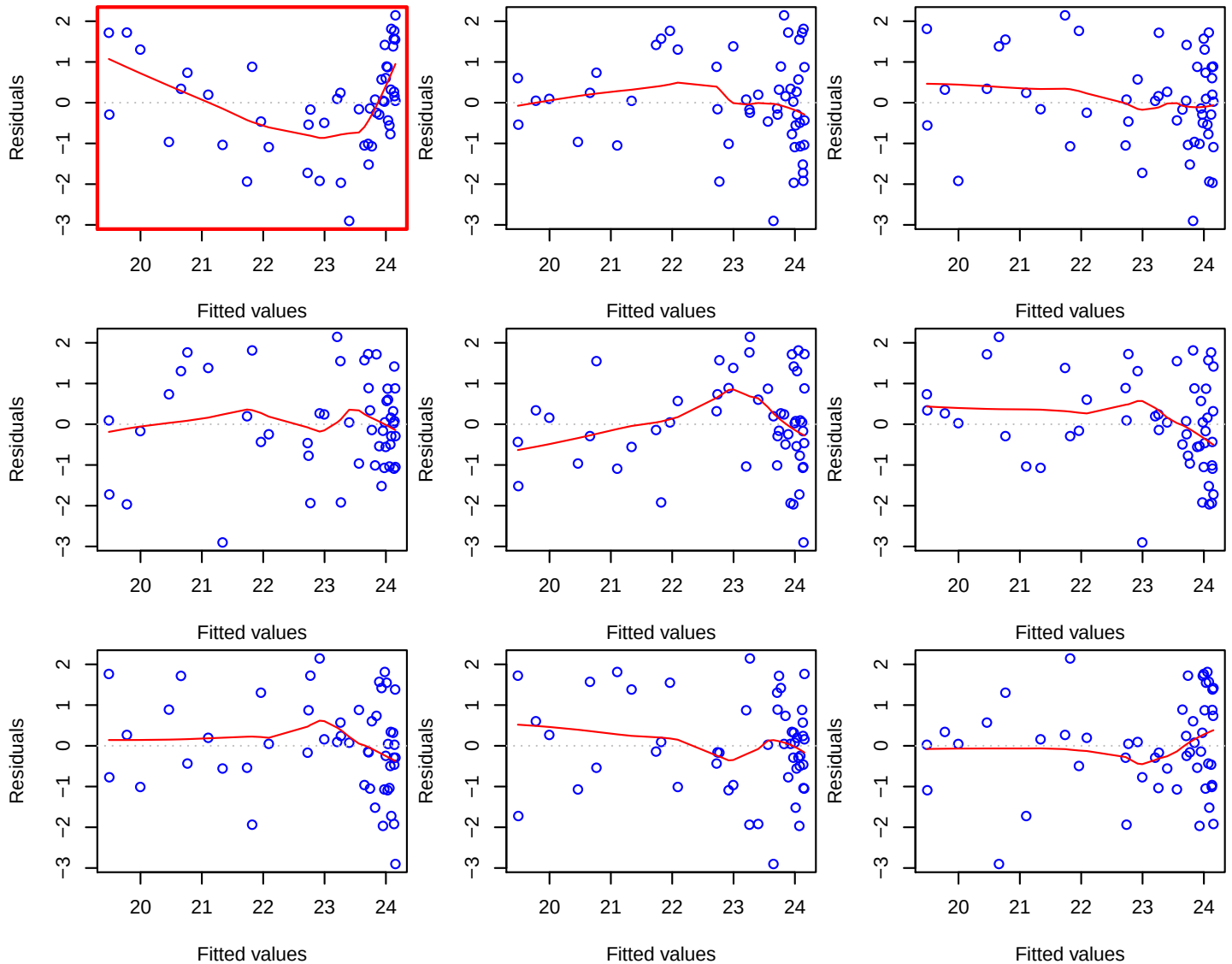
Why not model the data on the original scales?





. . . then interpretation would be easier.

BUT: the association does NOT look linear on the original scale.


Residuals
Fitted values
−3
−2
−1
0
1
2
20
21
22
23
24

Outline

Regression vs correlation

In linear regression, we model a **directed** relationship, either:

- **A causal relation:** We assume that x has an effect on y, *not* the other way around.
- **A prediction problem:** We know x and want to predict y.

But sometimes we just want to know:

- Are two different outcomes associated? (without any specific directed relationship)
- In this case we can use a correlation coefficient as a crude measure of the strength of the association.

Pearson's correlation

Measures the strength of linear association between two outcomes.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

My favorite **interpretation**: this is just the **regression coefficient** (i.e. the estimated slope) obtained **after standardizing** both the outcome and the predictor variable.

[8] See e.g. Rodgers and Nicewander. "Thirteen ways to look at the correlation coefficient." The American Statistician 42.1 (1988): 59-66.

Pearson's correlation

Measures the strength of linear association between two outcomes.

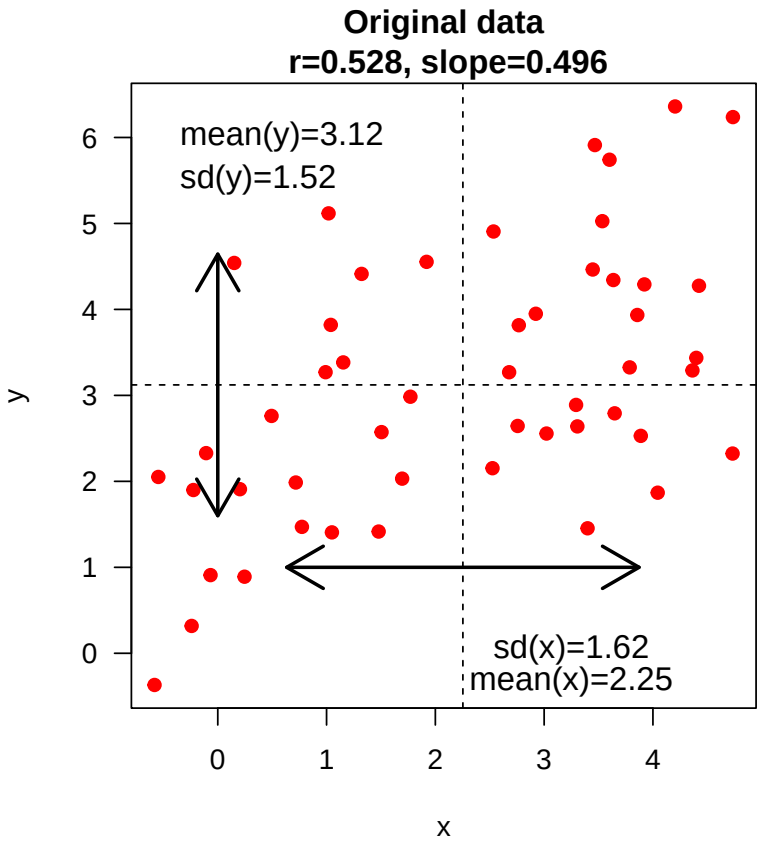
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

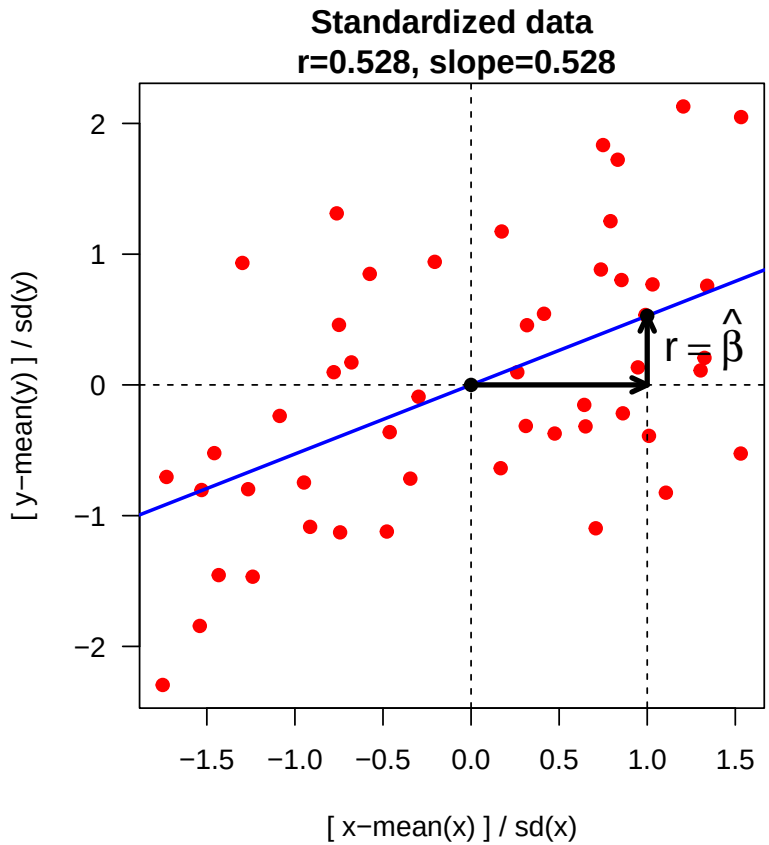
My favorite **interpretation**: this is just the **regression coefficient** (i.e. the estimated slope) obtained **after standardizing** both the outcome and the predictor variable.

There are many other interesting interpretations. Some require specific assumptions (bivariate normal distribution), others none.[8]

[8] See e.g. Rodgers and Nicewander. "Thirteen ways to look at the correlation coefficient." The American Statistician 42.1 (1988): 59-66.

Pearson's correlation as an estimated slope





Note: standardizing preserves the association (same scatter plots) but changes the unit of the variables (i.e. the values displays on the x and y-axes).

Pearson's correlation properties

Properties:

- r is symmetrical in x and y
- r is always between -1 and $+1$
- r has the same sign as the regression coefficient β (no matter whether you regress y on x or the other way around).

Interpretion of Pearsons correlation coefficient

$r = 0$, no correlation

- occurs when x and y are independent (but not only in this case).

$r > 0$, positive correlation

- Larger/smaller values of x and y tend to coincide.

$r < 0$, negative correlation

- Larger values of x tend to coincide with smaller values of y and vice versa.

$r = \pm 1$, perfect linear association

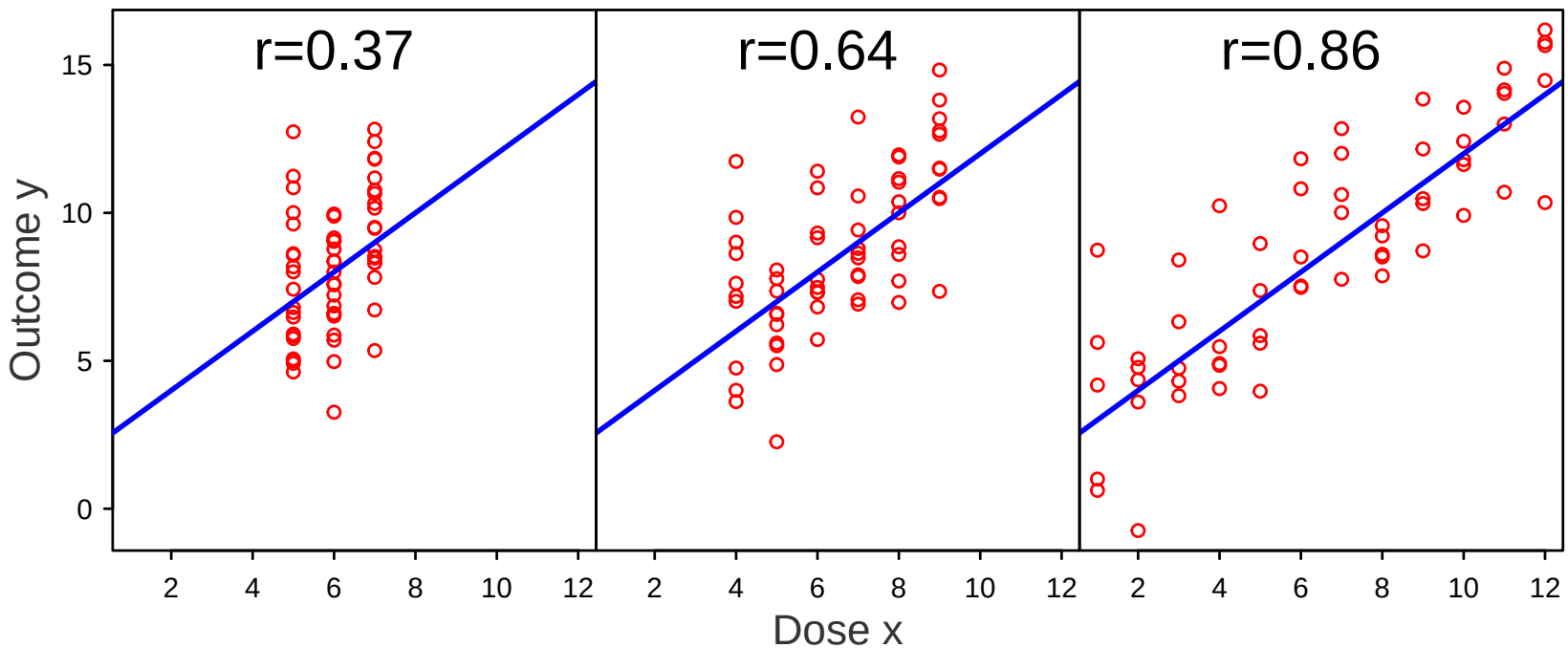
Arbitrary rule of thumb: $|r| < 0.3$: weak correlation, $0.3 < |r| < 0.5$: moderate correlation, $|r| > 0.5$: strong correlation.

Be careful! (1/2)

- The correlation makes sense mostly when both x and y are **random**. It doesn't really make sense to report a correlation coefficient if the values of x were dictated by the study protocol (e.g. doses).

Be careful! (1/2)

- The correlation makes sense mostly when both x and y are **random**. It doesn't really make sense to report a correlation coefficient if the values of x were dictated by the study protocol (e.g. doses).



Example: same linear association (i.e. same slope) but different study designs to split n=60 subjects in 3, 6 or 12 dose groups.

Be careful! (2/2)

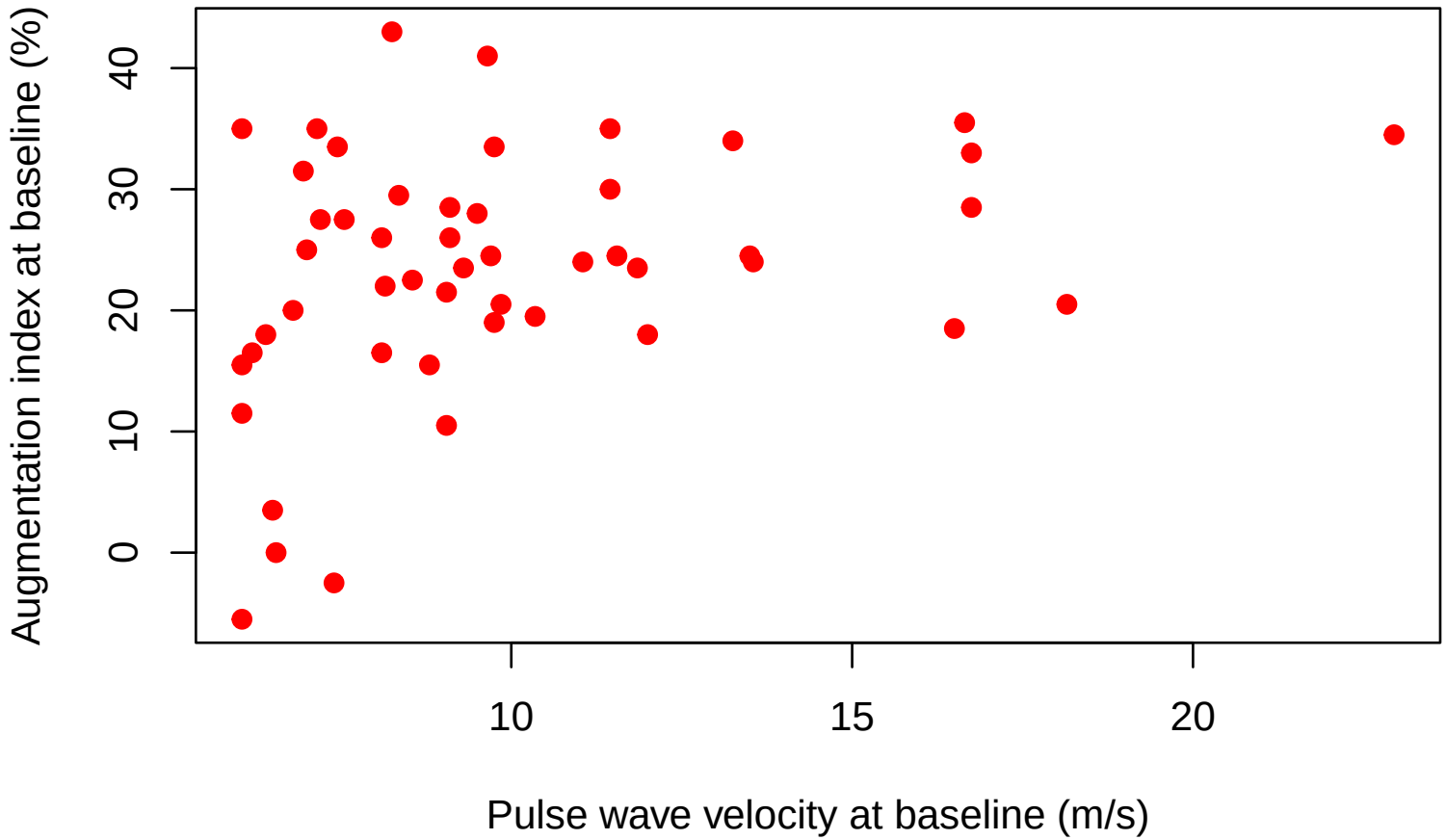
- The strength of the correlation **depends on the study population**. E.g. height and weight is stronger correlated in children than in adults (different SDs).
- Interpretation should **depend on the study aims**. An 0.9 correlation may be poor if we are comparing two methods of clinical measurement supposed to measure the same thing.
- Association is not the same as agreement. A device that hasn't been properly calibrated may correlate almost perfectly with one that has, but still measurements may show a large systematic deviation.

Note: show a 'Bland-Altman plot' instead of reporting a correlation in the example contexts of the last two items. [9]

[9] Bland & Altman. "Statistical methods for assessing agreement between two methods of clinical measurement." The lancet 327.8476 (1986): 307-310.

Case study: CKD

Are these two outcomes[10] related?



How strong is the association?

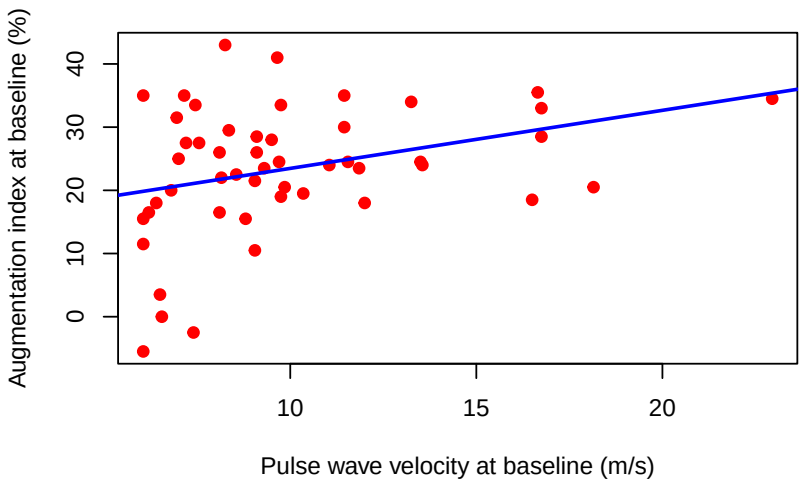
[10] Data from: Boesby et al: Eplerenone Attenuates Pulse Wave Reflection in Chronic Kidney Disease Stage 3–4 - A Randomized Controlled Study, PLOS ONE 2013.

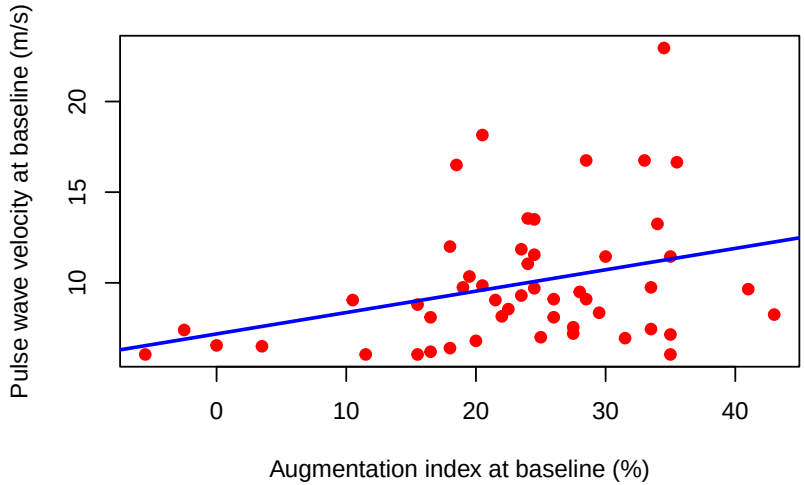
Analyzing correlation in R

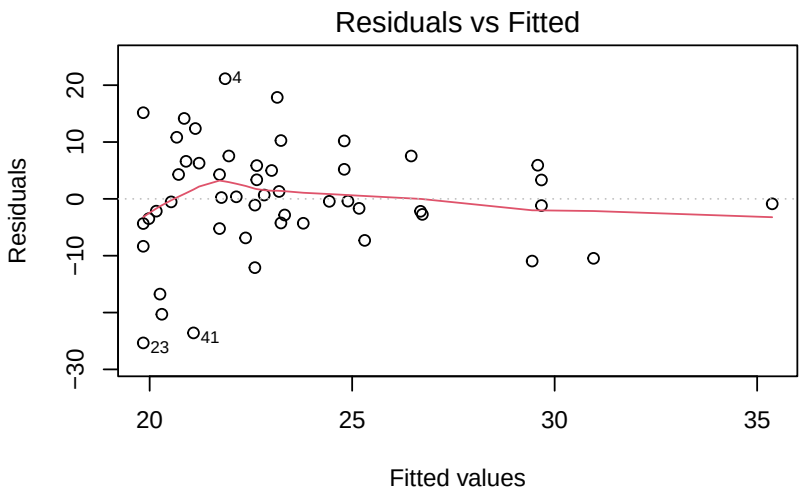
```
> cor.test(ckd$pwv0, ckd$aix0)


Pearson's product-moment correlation

data:  ckd$pwv0 and ckd$aix0
t = 2.4151, df = 48, p-value = 0.01959
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.05594193 0.55652213
sample estimates:
      cor
0.3291641
```

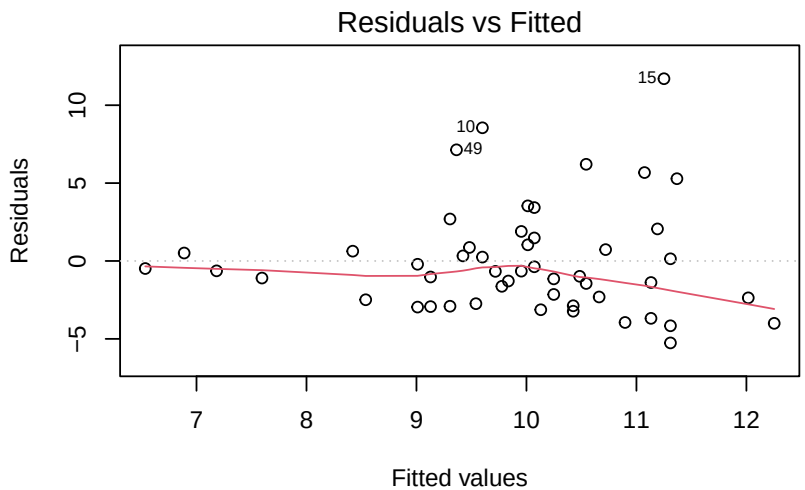
BUT: Are these outcomes really **linearly** related?

Linear associations in CKD data





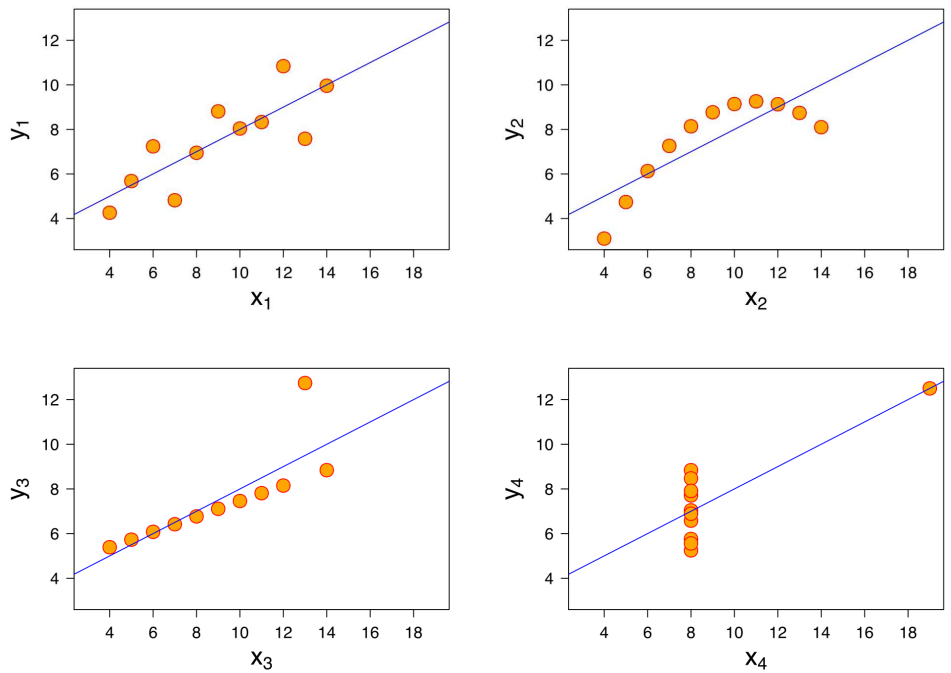




A linear association is maybe not the best way of summarizing the association

Summarizing implies loss of information

Anscombe's example of 4 datasets sharing the same 6 key statistics: Pearson correlation (r=0.816), slope and mean and sd of x and y (see e.g. wikipedia article "Anscombe's quartet" for more details).



Keep in mind: summarizing a scatter plot by a single (or few) number(s) cannot give the full picture of the association. It especially applies to the correlation coefficient, which is often computed and interpreted without much thinking. Summarizing implies loss of information, but hopefully ease of understanding

Spearman's rank correlation

Spearman's rank correlation is an interesting alternative to Pearson correlation for summarizing a monotonic association which is not necessarily linear.

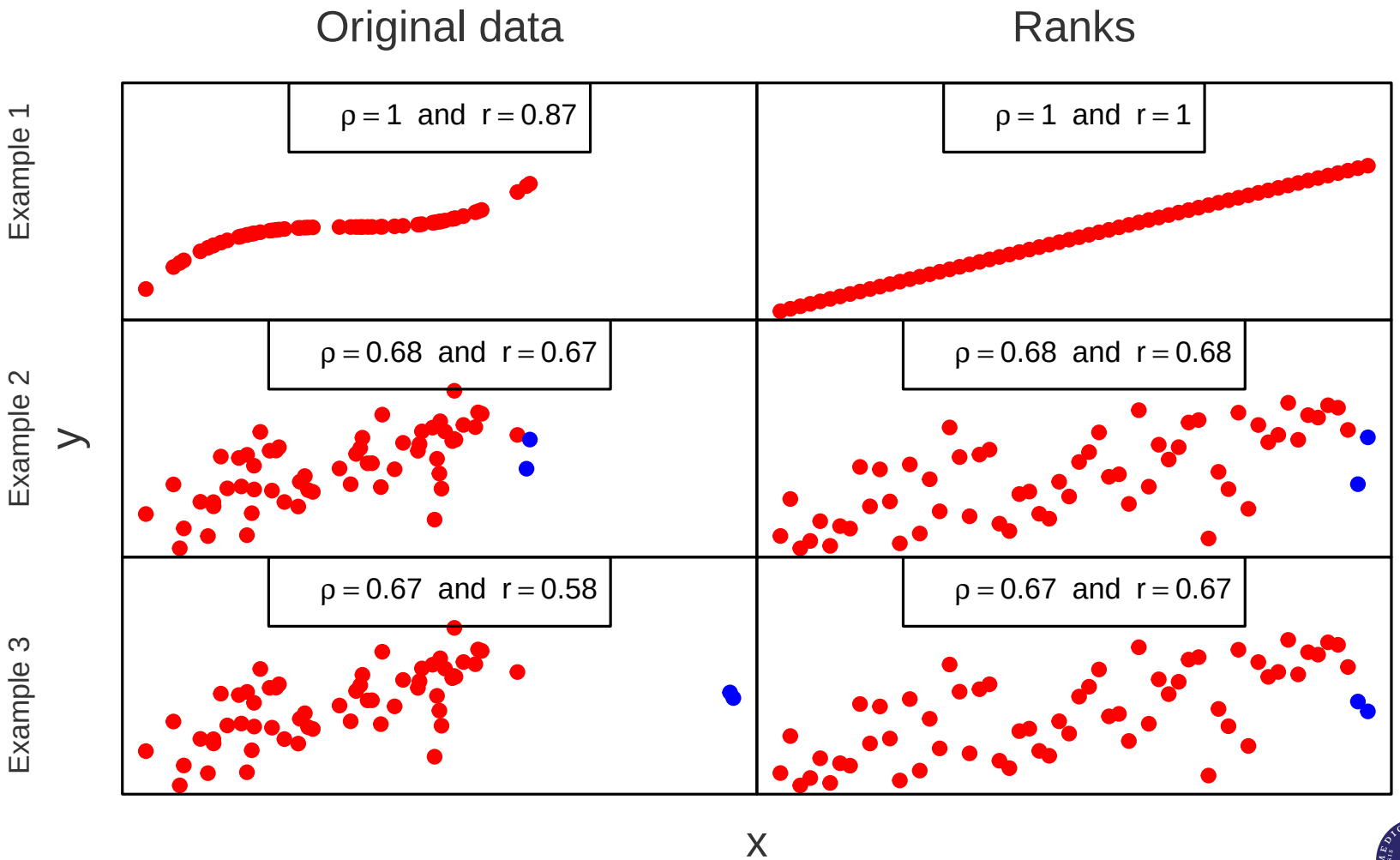
- The formula is the same as for Pearson's correlation, except that the original data has been replaced by ranks.

$$\rho = \frac{\sum_{i=1}^{n}(\text{rank}(x_i) - \overline{\text{rank}(x)})(\text{rank}(y_i) - \overline{\text{rank}(y)})}{\sqrt{\sum_{i=1}^{n}(\text{rank}(x_i) - \overline{\text{rank}(x)})^2 \sum_{i=1}^{n}(\text{rank}(y_i) - \overline{\text{rank}(y)})^2}}.$$

The rank[11] of an observation is it's number on the list when all data has been ordered from the largest value to the smallest.

[11] We have already met ranks in Lecture 2, when presenting the Wilcoxon test.

Spearman's rank correlation vs Pearson's Correlation



Spearman's correlation in R

Test the null hypothesis $H_0 : \rho = 0$.

```
> cor.test(ckd$pwv0, ckd$aix0, method='spearman')

Spearman's rank correlation rho

data:  ckd$pwv0 and ckd$aix0
S = 13982, p-value = 0.01981
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3285996
```

Limitation: we don't get a confidence interval (with this function).

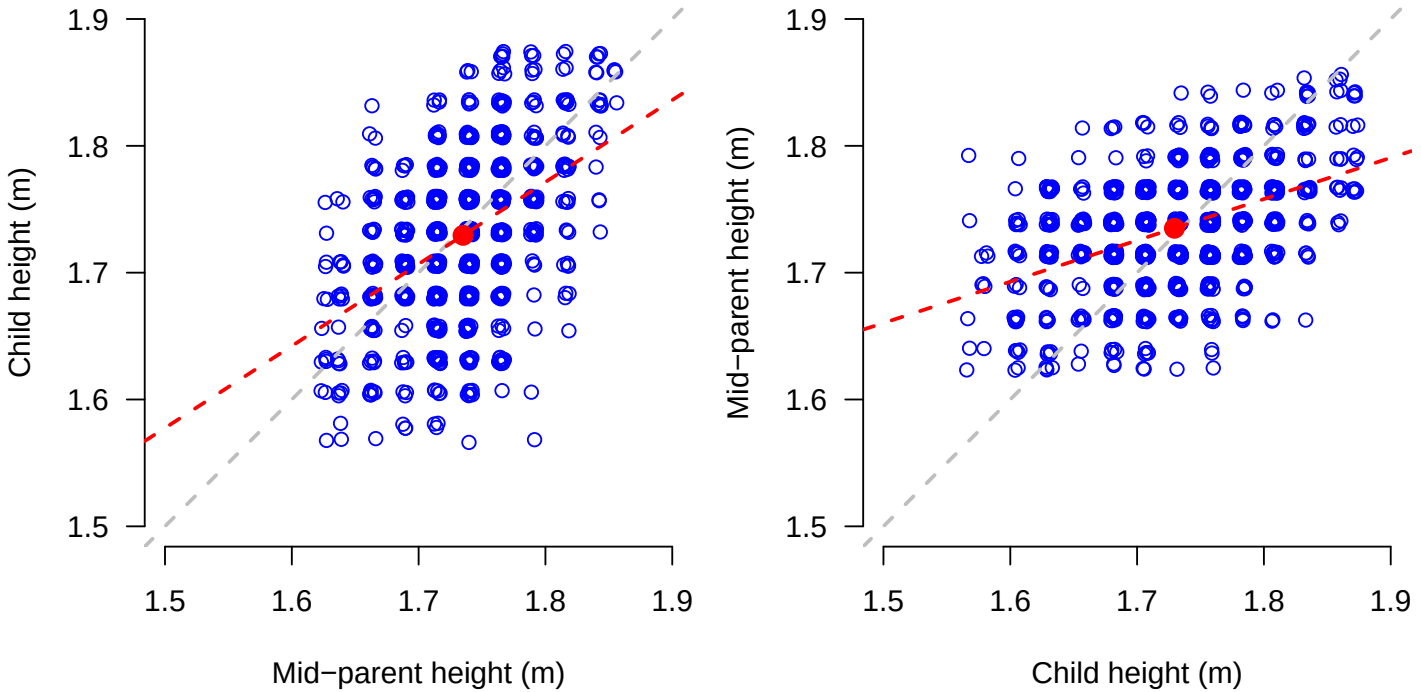
Outline

Why bother discussing regression to the mean?

Regression to the mean is *"so trivial that all should be capable of learning it and so deep that many scientists spend their whole career being fooled by it."* [12]

[12] Senn. "Francis Galton and regression to the mean." Significance 8.3 (2011): 124-126.

Historical (nice) example [14]



- *"we may expect that an adult child is closer to average height than its parents"* (left plot, slope < 1)[13]
- *"but also, paradoxically, that parents are closer to average height than is their child"* (right plot, slope < 1)

[13] Note: women height has been multiplied by 1.08 ("male equivalent")

[14] Galton's data (1886).

"Regression to the mean is a consequence of the observation that, on average, extremes do not survive."

"In our height example, extremely tall parents tend to have children who are taller than average and extremely small parents tend to have children who are smaller than average, but in both cases the children tend to be closer to the average than were their parents. If that were not the case the distribution of height would have to get wider over time!"

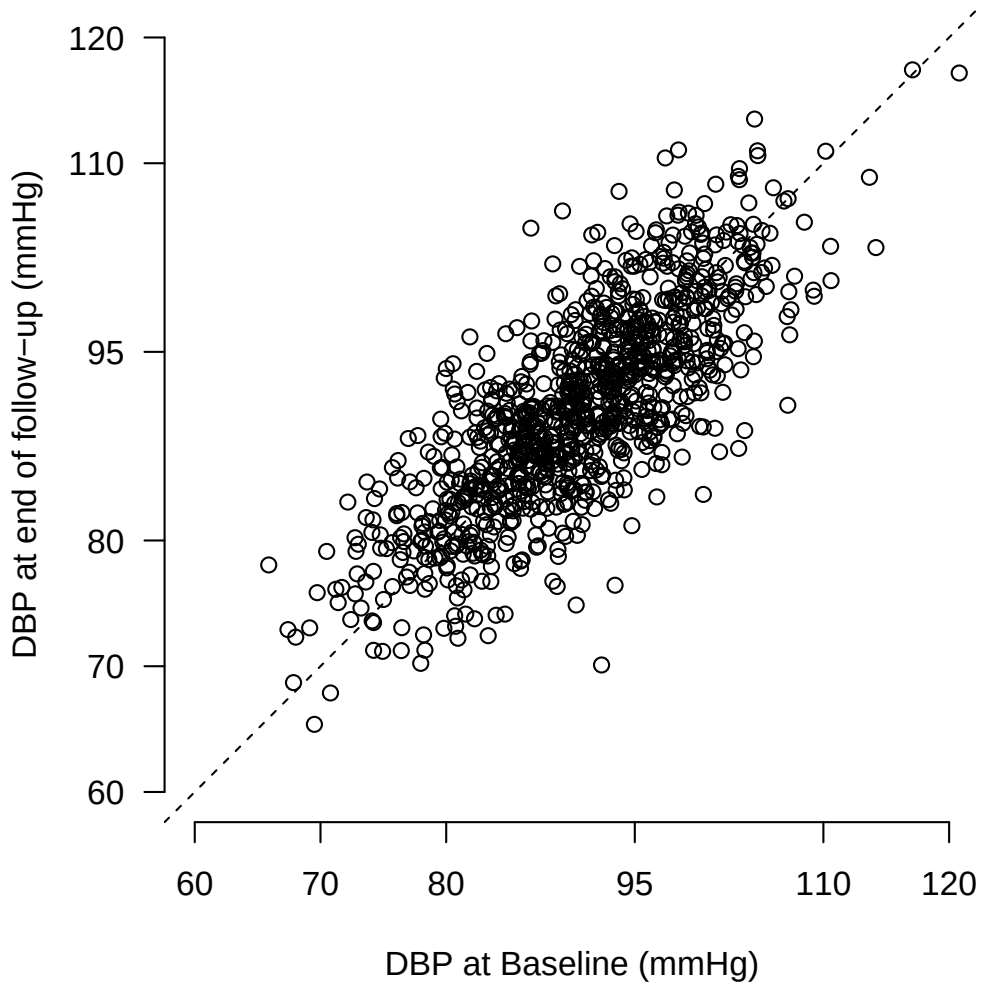
Got it? Then...

"Do you think that there is good evidence that the placebo effect is genuine?"

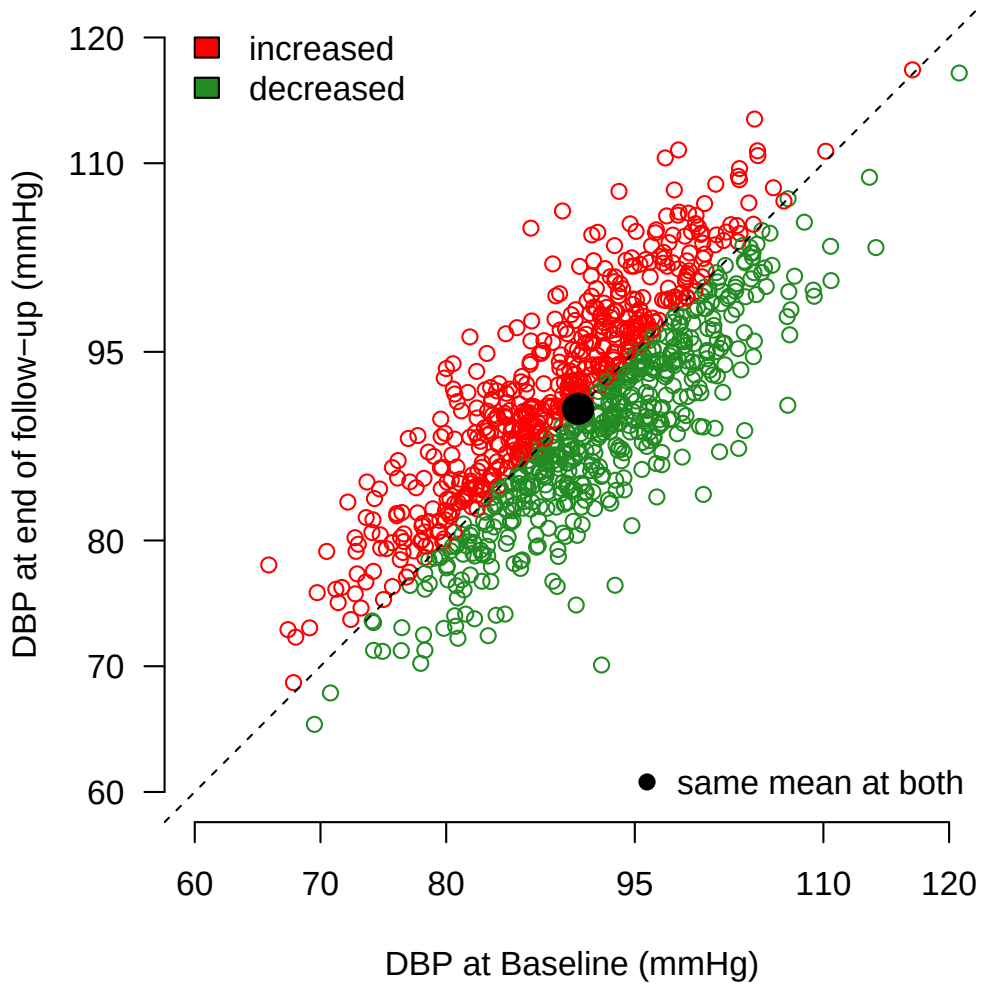
"If so, stick around for a while because I will try and show you that you (and ten thousand physicians with you) are wrong." [15]

[15] Senn. "Francis Galton and regression to the mean." Significance 8.3 (2011): 124-126.

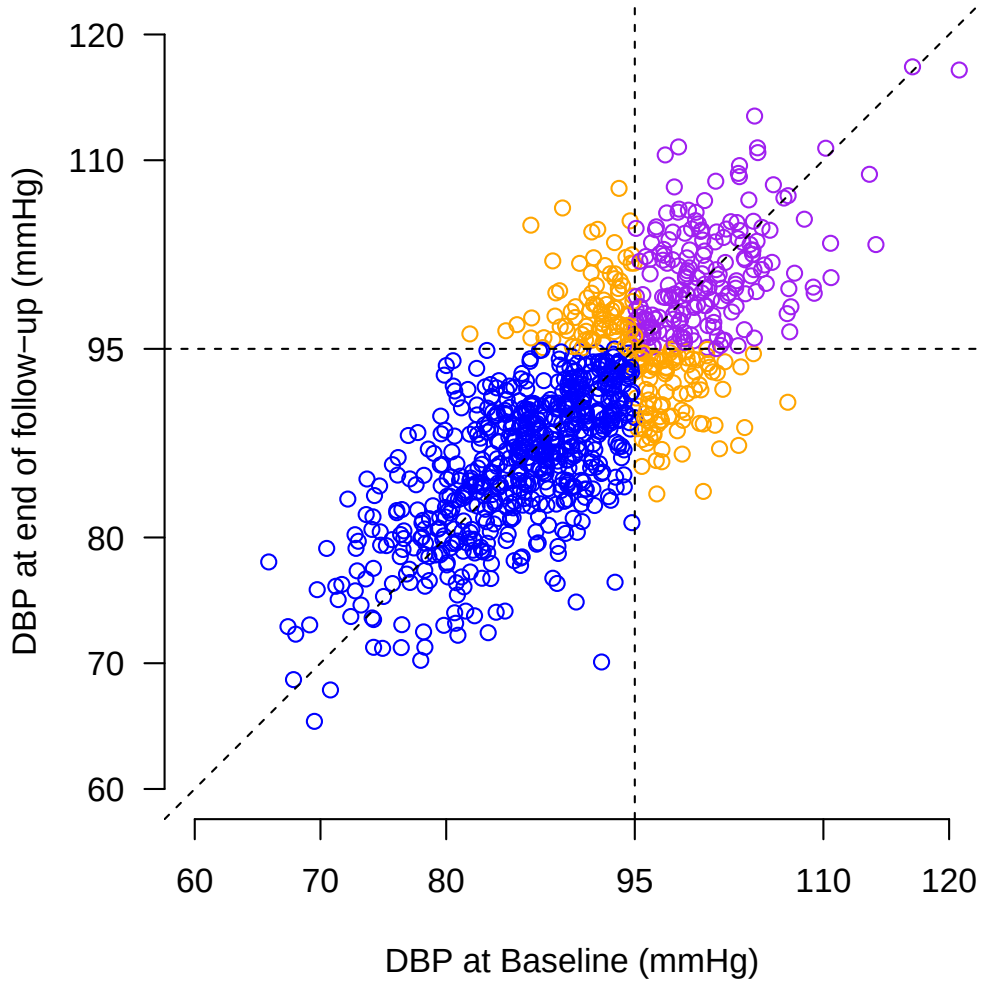
Diastolic blood pressure: "Random sample" (n=1000)



Diastolic blood pressure: "Random sample" (n=1000)

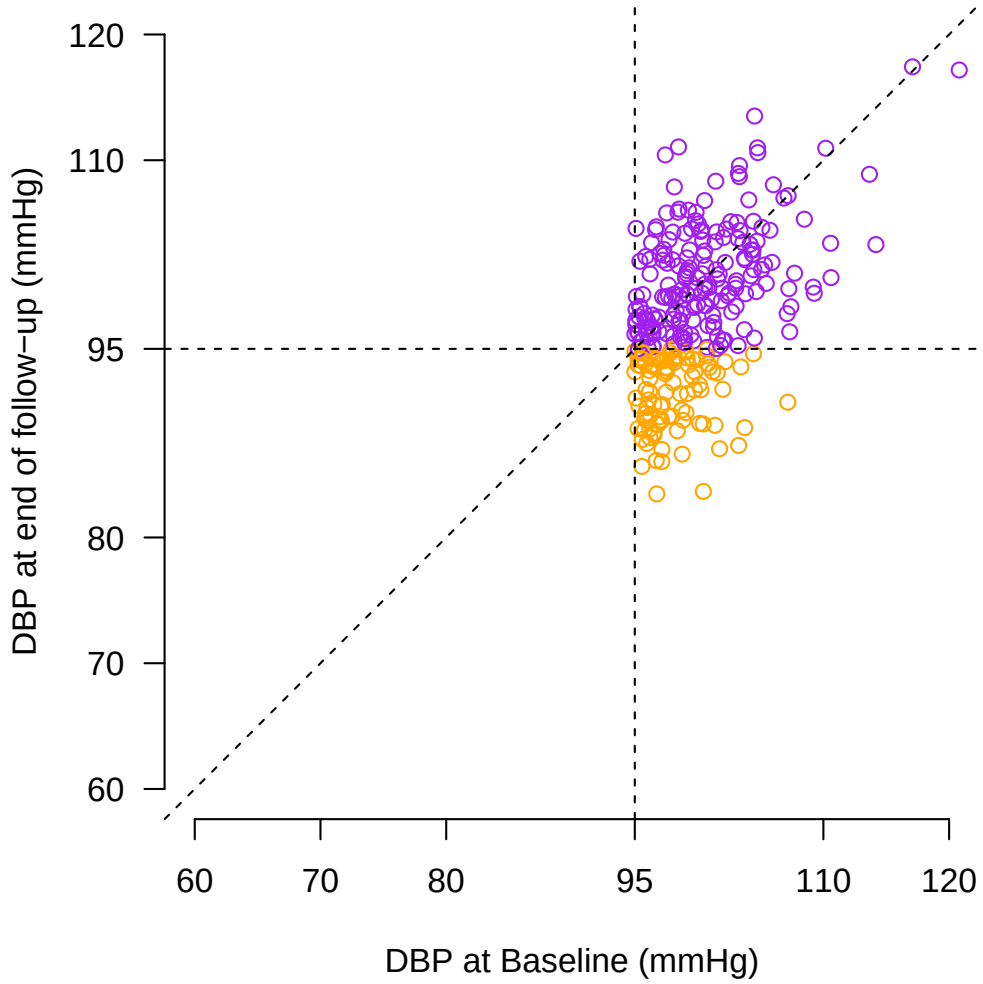


Diastolic blood pressure: "Random sample" (n=1000)



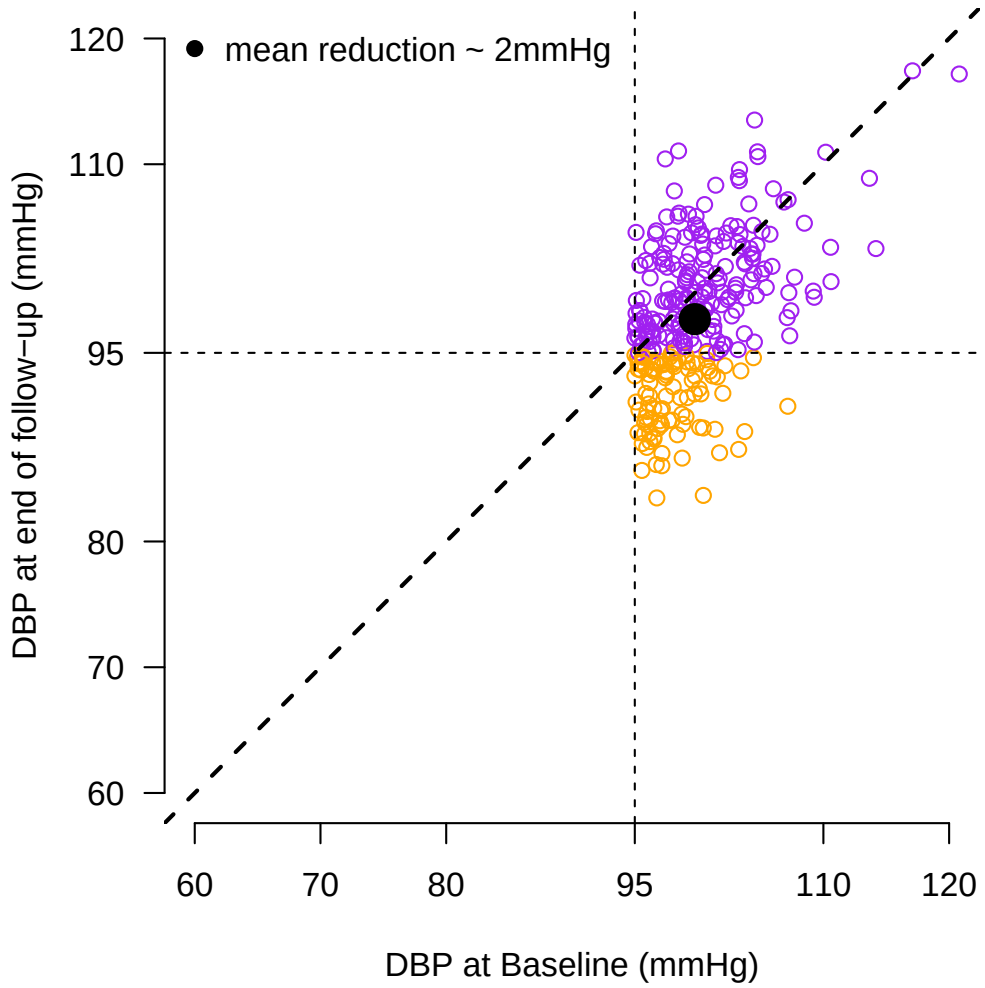
- "Hypertensive" if > 95mmg.

Diastolic blood pressure: "Clinical trial"



- Only "Hypertensive" patients are included in the trial.

Diastolic blood pressure: "Clinical trial"



- Only "Hypertensive" patients are included in the trial.

Consequences on baseline follow-up studies

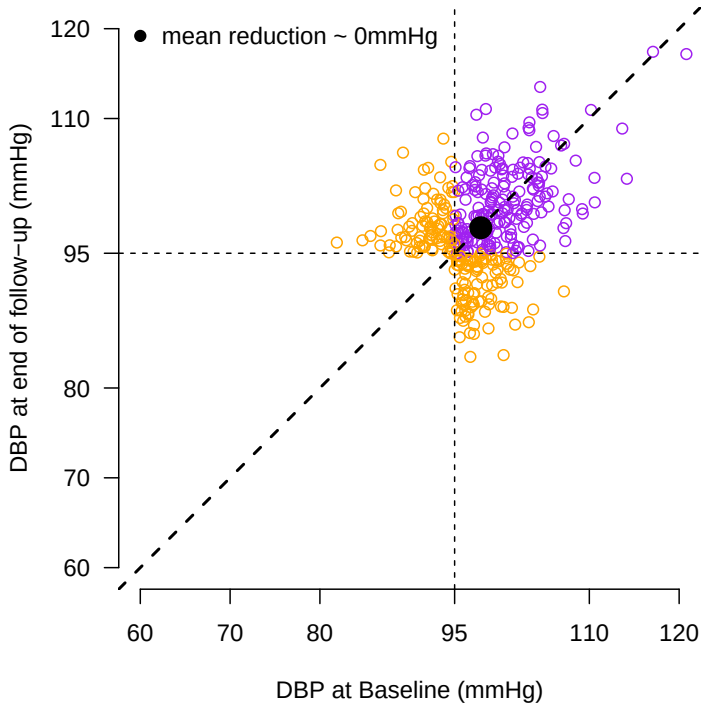
- We can (almost) always expect a spontaneous improvement from baseline when we include "symptomatic" patients.
- This (usually) has nothing to do with a genuine placebo effect [16] but it is only a statistical "oddity" or in SS own words *"a consequence of this stupid (but very common) way of looking at the data"*.
- Regression to the mean makes it clear that a control group is needed for stronger (causal) conclusions.[17]

[16] Only in area of pain control does there seems to be reliable evidence of a placebo effect.

[17] Regression to the mean will result in improvements in the two groups, and the comparison of the two improvements can be used to draw stronger conclusions.

More details

[In this hypothetical trial,] *"We can only see patients who remain hypertensive or who become normotensive. We left out the patients who were normotensive but became hypertensive. They are shown in* [the right Figure]. *If we had their data they would correct the misleading picture in* [the previous Figure], *but the way we have gone about our study means that we will not see their outcome values."*



Outline

The linear model
- ILO: to describe the model, its parameters and assumptions

Model fitting and inference
- ILO: to outline model fitting and interpret standard results

Prediction
- ILO: to describe what we can (or cannot) predict, why and how

Checking the model assumptions
- ILO: to list the model assumptions and know how to assess them
- ILO: to explain why they are not all equally important

Correlation
- ILO: to interpret a correlation and critically discuss its usefullness

Regression to the mean
- ILO: to recall the phenomenon and its potential to be misleading

Appendix: Formulas and linear models in R

Estimated regression coefficients

The best fitting line can be solved explicitly:

- The estimated slope is given by:

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

- And the intercept can be computed by

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

  where $\hat{\beta}$ from the previous formula is inserted.

Note that the fitted line always passes through the point $(\bar{x}, \bar{y})$, where $\bar{x}$ and $\bar{y}$ are the sample means.

Standard errors

The standard errors for $\hat{\alpha}$ and $\hat{\beta}$ are given by.

$$\text{s.e.}(\hat{\alpha}) = s \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$\text{s.e.}(\hat{\beta}) = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where s is the residual standard deviatoin.

A bigger sample size n will of course give rise to smaller standard errors, but the specific values of the x's also has an impact.

- s.e$(\hat{\beta})$ is larger if x doesn't vary much.
- s.e$(\hat{\alpha})$ is larger if x doesn't vary much, and/or if $\bar{x}$ is far away from 0.
- Both are larger if the residual variance is large.

Uncertainty in prediction

- The standard error of the expected value at x_0 is:

$$\text{s.e.}(\hat{y}(x_0)) = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

- This is the uncertainty related to **estimating the average response** at x_0.

- If we want to **predict individual responses** at $x = x_0$ **with 95% certainty**, then we need:

$$\text{s.d.}(y_{\text{new}}(x_0) - \hat{y}(x_0)) = s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x - \bar{x})^2}}.$$

- where the residual variance has been added to the estimation uncertainty.

Connection between regression and correlation

Recall that the estimated regression coefficient $\hat{\beta}$ is given by:

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

while Pearson's correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

From this it follows that

$$r = \hat{\beta} \cdot \frac{s_x}{s_y}$$

where $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ and $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$ are the sample standard deviations for x and y.

Linear models in R

- We use the `lm`-function to do linear regression
  (and a lot more: ANOVA, multiple regression, . . .)
- The model must be specified by a model formula, e.g.:

```
fit <- lm(log2diam ~ log2conc, data=th)
```

- where $\sim$ should be read as "potentially depending on" or "is potentially predicted by".
- The response goes on the left and the predictor on the right.
- `lm` returns a so-called *model object* of the class "`lm`".
  You don't have to understand all of its contents to use it!

Extractor functions

R-functions that extract information from model objects, e.g.:

- `summary(fit)` — table of estimates, tests, and more.
- `confint(fit)` — confidence intervals.
- `abline(fit)` — add the fitted line to an existing plot.
- `residuals(fit)` — vector containing the residuals
- `predict(fit, frame)` — predict y's for supplied x values.
- `plot(fit)` – diagnostic plots (e.g. model assumptions).

For an essential summary of your analysis use:

- `publish(fit)` – from the publish-package.