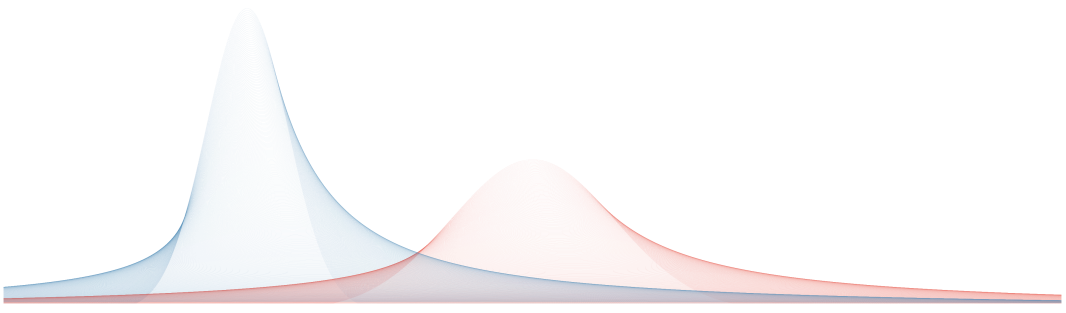
Frans Rodenburg

Elements of Biostatistics



Contents

```
## Warning: package 'knitr' was built under R version 4.0.3
```

Preface

- **The Biostatistics books are a work in progress. Suggestions can be sent to biostatistics.bookseries@gmail.com (F.J. Rodenburg).**

The goal of this series is to provide a free, intuitive and well-structured introduction to statistics. The target audience primarily includes students or researchers in the life sciences. As such, emphasis lies on methods used in life sciences and all examples will concern research questions in biology. Naturally, the use of these methods extends beyond the life sciences.

This book is aimed at beginners* and novices in statistics. A short recapitulation of ***Introduction to Biostatistics*** is included, but basic knowledge of statistical concepts and programming in R is assumed.[1]

[1] Refer to the syllabus *Statistics in R & RStudio* for a brief introduction to statistical programming.

Biostatistics Series

This book belongs to a three part series:

1. **Introduction to Biostatistics**: basic concepts, statistical tests, ANOVA & simple linear regression;
2. **Elements of Biostatistics**: linear regression, basic multivariate techniques, ANCOVA & GLM;
3. **Advanced Biostatistics**: mixed models, cross-validation, bootstrapping & advanced multivariate techniques.

Outline

This book focuses on expanding your knowledge of statistical models, in particular using regression analysis. While there is a short recapitulation of the first book in chapter ***1***, consider reading the first book if the learning curve of this book is too steep. I have made references to it in this book where relevant.

1. Recap of Introduction to Biostatistics
2. Techniques 1: Linear Regression
3. Techniques 2: GLM
4. Techniques 3: PCA
5. Background 1: Common mistakes in analyses
6. Techniques 4: Multiple testing correction
7. Techniques 5: Equivalence testing
8. Background 2: Handling missing data

Some chapters, paragraphs and sections are denoted by an asterisk (*). These serve for completeness, correctness, or clarification and are non-essential.

1 Recap of the First Book

This chapter outlines the background knowledge that is assumed for the rest of this book. Please refer to the previous book for details.

Biology is an **empirical science**: Evidence is primarily collected through observations. Whether these are in line with hypotheses is at least to some extent subject to chance. Statistics is concerned with expressing/controlling that chance.

In God we trust, all others must bring data. — W. Edwards Deming

1.1 What is Statistics?

A word with several meanings, statistics refers both to abstractions of the data (e.g. a mean, a percentage) and to the field of research involved with empirical analysis.

Emperical science makes little use of proof based on axioms, like $a + b = b + a$. Instead, theories are supported by observations or experiments. Theories with strong experimental evidence and a scientific explanation may be extremely appealing, but are not *proven*.

Uncertain results are by no means unique to biology. In fact, one could argue that *all* knowledge is but empirical evidence, insomuch that the human brain tries to make sense of the world based only on observations through senses. Since observations are all we have, understanding the strength of their evidence is crucial.

Before collecting any evidence, consider the purpose of the study: What are you trying to demonstrate? Which category[1] best describes the research question? Carefully consider how the proposed study design is supposed to answer the research question.

[1] The categories being: **exploration**, **estimation**, **confirmation** & **prediction**.

- **Paragraph 5.1 discusses the dangers of vague research goals.**

1.2 Probability Theory

Probability theory is the basis of all statistical inference. While you don't need extensive knowledge of mathematics to perform, or understand statistical analysis, you should have a decent understanding of how chances propagate, how the chance of one event can depend on another, and what a probability distribution is (paragraph ***1.6***).

Name	Formula	Interpretation	Example
Complement	$P(A^c) = 1 - P(A)$	Probability of *not* A.	The chance of not rolling ⚅.
At least one success	$1 - P(\text{no event})^{\text{trials}}$	Probability of at least one event after multiple trials.	The chance of at least one false positive among multiple tests.
Multiplication principle	$k_1 \times k_2 \times \cdots \times k_p$	Combinations between categorical variables.	Total combinations of different diets and exercise routines.
Binomial coefficient	$\binom{n}{k} = \frac{n!}{k!(n-k)!}$	Number of possible combinations of size k from a total of n categories.	Number of combinations of two antibiotics that could be prescribed to a patient.
Permutations	n^t	Possible rearrangments.	Codons (amino acid triplets).
Conditional probability	$P(A) = P(A\|B)P(B)$	Chance of A given B.	Chance of developing cancer given a familial history.
Bayes' theorem	$P(A\|B) = \frac{P(B\|A)P(A)}{P(B)}$	Chance given prior knowledge.	Bayesian statistics.

Table 1.1: Overview of probability covered in the first book.

1.3 Study Design

Experiments and observations serve to draw conclusions about some *population*. This could be a population in the biological sense of the word, like that of the studied organism, but it could also be implicit:

1. Assessing the effect of some treatment on white blood cells;
2. Elucidating the structure of a protein;
3. Flipping a coin to find out whether it is fair.

The implied population of (1) consists of all white blood cells; (2) consists of all proteins of that kind and (3) involves all possible coin flips with that particular coin. Evidently, we cannot measure all white blood cells or proteins in existence, let alone perform an infinite number of coin flips. Hence we rely on *samples*: a random subset of the population, meant to represent it.

If the sample is an accurate representation, conclusions can be made about the population, bearing in mind the uncertainty from using just a sample. This process is called *statistical inference*. There are no methods to guarantee a sample represents its population well. However, a random, possibly stratified[2] sample from the population of interest has a minimal chance of bias.

Even if a sample accurately portrays its own population, is does not necessarily represent the *population of interest*: Clinical trials on patients from The Netherlands generalize well to the Dutch population, but not necessarily to the entire human race. Hence why scientific reports must state how samples were drawn.

Moreover, claims of causal relationships can only be supported by evidence from interventional studies. A study using only historical data, or otherwise purely observational data, can at best support claims of association, but not the *direction* of association (e.g. does A cause B or vice versa).

To avoid systematic effects confounding the results, apply random treatment assignment wherever it is practically feasible. This could mean the order in which you load the lanes on a western blot, the agar plates to which you assign either of two growth media, etc.

Statistics is not just about drawing conclusions, but also about designing a study such that the research question is actually answerable. The only thing more disappointing than not being able to conclude something interesting, is not being able to make any meaningful conclusions at all, due to limitations in experimental design.

[2] Sampling is *stratified* when it is performed separately per group (e.g. men and women). This is important if the effect of the group is expected to affect the outcome by a relevant amount. Continuous variables (e.g. age) can also be stratified for, by ensuring the range of the variable is well represented along other variables in the sample.

1.4 Exploring Data

Numerical and visual data inspection are useful in every analysis. Exploratory data analysis can reveal relationships between variables, outliers, and whether data is read correctly into the software at all. Especially in larger data sets, summaries and figures are faster, easier and more reliable than inspecting (a portion of) the whole data set. Both summaries and figures are used, because they reveal different aspects of the data:

> *Numerical quantities focus on expected values, graphical summaries on unexpected values.* — John Tukey

Gaining insight in an unfamiliar process can help form new ideas for research. Multivariate techniques like PCA[3] are often used to this end. Note, however, that hypotheses cannot be tested on the same data used to form them.[4]

[3] See chapter **4**.

[4] See paragraph **5.1**.

1.5 Estimation

Estimation tries to find a population parameter using a sample. Since a sample does not contain all information about its population, a key point in parameter estimation is expressing the uncertainty of the estimate, through standard errors, or confidence intervals.

Parameter estimation from the first book is summarized below:

Name	Population parameter	Sample estimate	Interpretation
Mean	$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$	$\bar{y} = \frac{1}{n-1}\sum_{i=1}^{n} y_i$	Average of all observations; expected value.
Variance	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$	Spread of the observations around the mean.
Standard deviation	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$	Square-root of the variance.

Continued on next page

Name	Population parameter	Sample estimate	Interpretation
Standard error		$\text{SE}_{\text{mean}} = \frac{s}{\sqrt{n}}$	Measure of inaccuracy (here: inaccuracy of the sample mean).
Confidence interval		$\text{CI}_{95\%}(\bar{y}) \approx \bar{y} \pm 1.96 \cdot \text{SE}$	Inaccuracy of an estimate (here: the sample mean) as a range.
Covariance	$\text{Cov}[X, Y] = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$	$q_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$	Measure of linear association (positive of negative). The actual number has no meaning.
Correlation	$\rho_{X,Y} = \frac{\text{Cov}[X, Y]}{\sigma_X \cdot \sigma_Y}$	$r_{X,Y} = \frac{q_{X,Y}}{s_X \cdot s_Y}$	Standardized measure of linear association.

Table 1.2: Overview of estimation covered in the first book.

1.6 Distributions

Probability distributions describe the chances of different values from random variables. The total area under a probability density function covers all possible values and is thus equal to 1. The chance to observe a value within a certain range is equal to the fraction of the total area covered by that range.

The simplest distribution is the uniform distribution (fig. 1.1): All values are equally likely. Think for example of a dice roll (discrete), or the chance of a particular position on a perfectly round ball to end up on top after kicking it in the air.

Biological processes are affected by the countless effects of genetic variation and environmental factors. The sum of all these effects cause a *central tendency*: Variables like height, blood pressure and IQ are often close to some average value, with further deviations in either direction exceedingly rare. This is often well approximated by a *normal distribution* (fig. 1.2). The mean (μ) is the location of this average value and the variance (σ^2) is the extent to which observations tend to deviate from it.

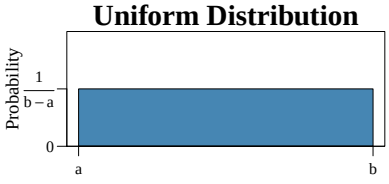


Figure 1.1: The continuous uniform distribution with minimum a and maximum b.

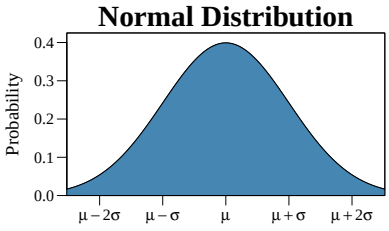


Figure 1.2: The normal distribution with mean μ and variance σ^2.

Most of the tests and models discussed in the previous book assume approximate[5] or exact[6] normality of the underlying process. However, many processes are inherently non-normal (e.g. counts and ratios). Other common probability distributions and linear models supporting them are covered in chapter 3.

[5] *T*-test, ANOVA, simple linear model

[6] *F*-test

1.7 Hypothesis Testing

Confirmational statistics try to answer yes or no questions:

1. Are men taller than women on average?
2. Does smoking affect cancer incidence rates?
3. Is there a difference in the efficacy of two treatments?

Since samples do not span their entire population, empirical evidence cannot give a hard yes or no. Hence, confirmational statistics rely on null-hypotheses and measures of uncertainty as evidence for a statement. One such measure is the ***p*-value**:

A sample always has some difference from its population. If the difference from the null-hypothesis is sufficiently large, the null-hypothesis is rejected.

> What is the chance of observing at least this large a difference from the null-hypothesis in the sample, if the actual difference in the population were zero?

Another measure of uncertainty is the **confidence interval**, although contrary to the *p*-value, its use is not restricted to hypothesis testing:

> If samples were drawn from this population repeatedly, and x% confidence intervals were constructed each time, then x% of these intervals should contain the true value.

While confirmational statistics can have its use in some research, hypothesis testing (and the *p*-value in particular) has been heavily criticized for being misused and overused. Not every research question can be meaningfully reduced to a yes or no question, and doing so does not further science. Only use tests when an empirical yes or no provides useful information.

Lastly, keep in mind that there is no way to collect evidence *for* the null-hypothesis. A high *p*-value means the results *might as well* have arised under the null-hypothesis. If you do want to demonstrate there is no difference, use an equivalence test (chapter 7).

1.8 Statistical Tests

Below is a summary of the significance tests covered in the first book:

Name	Formula	Interpretation	Example
T-test	$t = \frac{\bar{y}_1 - \bar{y}_2}{\text{SE}_{\text{pooled}}}$	Comparison of means.	Is there a difference in average height between men and women?
χ^2-test	$\chi^2 = \sum_{i=1}^{k} \frac{(\mathbf{O}_i - \mathbf{E}_i)^2}{\mathbf{E}_i}$	Comparison of ratios.	Is there a difference in cancer prevalence between smokers and non-smokers?
F-test	$F = \frac{s_1^2}{s_2^2}$	Comparison of variances.	Do male and female height vary to a different extent?

Table 1.3: Overview of tests covered in the first book.

Of the three tests you learned about in the previous book, one is particularly common: The t-test. It answers the question: *Is there a significant difference between two means?* Even comparisons in linear regression models often boil down to a set of t-tests.

While the t-test has many variants, the principle remains the same:

- A one sample t-test compares a mean directly to the value under the null-hypothesis ($t = \frac{\bar{y} - \mu_0}{\text{SE}}$);
- A t-test for equal variance replaces $\text{SE}_{\text{pooled}}$ with SE;
- A paired t-test is simply a one-sample t-test for the differences between two paired samples (with $\mu_0 = 0$).

The odd one out is the the Wilcoxon rank sum test. This test does not compare means,[7] but determines the probability that a randomly selected member of one group is larger than a randomly selected member of the other. While this test does not make any assumption of normality, it is less powerful if normality *can* be assumed.

Once you are familiar with generalized linear models (chapter 3), the χ^2-test becomes all but deprecated, in favor of a binomial GLM.

The F-test is can occasionally be useful for comparing variances, but care must be taken not to confuse non-normality for a difference in variance.

[7] Note that this implies you cannot say group A has a larger mean than group B, since this is not what is tested for. The correct interpretation is: *There is significant evidence for a location shift*. If this sounds a little convoluted, it is. In this book you will learn better alternatives for non-normal processes.

1.9 Statistical Models

A statistical model is an approximation for the process that generates the outcome: the *data-generating process*. This process generally consists of a systemic and stochastic part:

$$\underset{\text{e.g. cancer development}}{\textbf{biological process}} = \underset{\substack{\text{e.g. smoking behaviour,}\\ \text{genetic predisposition}}}{\textbf{systematic part}} + \underset{\substack{\text{e.g. random mutations,}\\ \text{environmental effects}}}{\textbf{stochastic part}} \tag{1.1}$$

The systematic part is what a model should capture: be it an average, a linear relationship, or some other consistently observed trend or feature. It is the information about the process we wish to learn.

The stochastic part is the random deviation from what is expected: A random Dutch man's height is likely *around* 1.8 m, but unlikely exactly 1.8 m. At a given point in time, the expectation is constant, but the random deviation will be different on each random draw from the population (fig. 1.3).

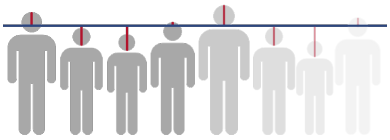


Figure 1.3: While each individual has a different random deviation, and while each sample drawn from the population will have a slightly different sample mean, the population average is constant.

The exact data-generating process is rarely known or even determinable. Even something as basic as human height is the result of countless genetic and environmental factors. Yet scientists and laymen alike often mistake empirical results for proof. Hence the famous quote:

> *All models are wrong, some are useful.* — George Box

This quote is a reminder that you cannot find the "perfect model" that exactly describes the data-generating process—at least not by collecting data. The second part reassures that we should not judge whether a model is wrong, but whether it accomplishes its goal to a satisfactory extent.

Suppose we model height as a function of sex and normally distributed error: If height is the result of innumerable genetic effects and a lifetime of environmental influences, then of course this model is wrong: Height is not determined by sex alone. However, we consistently observe about a 0.1 m difference between the *average* height of males and females,[8] with the deviations from these averages approximately following a normal distribution. While the model is plain wrong, it works fine as a tool for estimating differences in means between males and females.

For inferential models, a reasonable approximation using only the largest contributors can be very useful in understanding the underlying process.[9] Is lung cancer caused by smoking? Only if we disregard the effects of unrelated DNA damage, limitations of

[8] ourworldindata.org/human-height

[9] A model that explains a process well with as little variables as possible is called *parsimonious*.

DNA repair mechanisms, random acquirement of certain mutations, aggregation of tissue damage and chronic inflammation. So of course smoking alone does not cause lung cancer, but since it is such a major contributor, it is still a useful variable in assessing the increased risk, even if we cannot measure all other contributors and include them in the model.

Predictive models take this one step further: It does not even matter how wrong the model is, as long as it makes good predictions: Suppose you are a journalist and want to assess the severity of a fire, but you are not allowed to get close enough to it. You could call the fire department and ask how many firemen were sent to the area. The number of firemen quite reasonably predict the severity of the fire, but the direction of causality is completely disregarded (firemen are presumably not the cause of the fire).

Models and tests make assumptions about the underlying data-generating process. Some of these are explicit: If you include a variable for age in your model, then you are assuming that age has an effect on the outcome. Other assumptions are implicit: If you use a t-test, then you assume that the distribution of the outcome within groups can be reasonably approximated with a normal distribution.

Statistical assumptions are no different then other assumptions in life: You *assume* something to be true, without necessarily having evidence for it. You cannot prove that an assumption is correct. To maximally ensure that a model is giving results that you can trust, you should be able to defend your choice of assumptions from theory and check for violations from the assumptions in practice.

1.9.1 *Regression Models*

Many research questions in biology fit the framework of regression analysis. A regression model attempts to describe the process by estimating *parameters*: Unknown constants connecting the explanatory variable(s) to the response (y). Consider the simple linear model:

$$\underset{\text{response variable}}{y_i} = \overbrace{\beta_0 + \beta_1}^{\text{parameters}} \cdot \underset{\text{explanatory variable}}{x} + \underset{\text{residual}}{\epsilon_i} \tag{1.2}$$

If the model accurately represents the real process, then the fixed part ($\beta_0 + \beta_1 x$) describes the systematic part of the process. The remainder (ϵ_i) then consists purely of random, unpredictable deviations. If the model is too simple, its residuals will still contain useful information. Conversely, a model that is too complex will overfit: Its parameters will depend (at least in part) on the random deviations rather than the systematic part, causing bias towards the sample (see **2.6.4**).

1.9.2 *Is there a 'Best' Kind of Model? (*)*

The most general form of a model is a non-parametric model. Non-parametric implies that no assumptions are made *a priori* about the data-generating process. Since these models have no predefined structure, they are generally much harder to interpret, making them primarily suitable for the purpose of prediction. Examples include kernel density estimation or neural networks.

Neural networks are actually densely parameterized models with weights that change throughout the fitting, such that there is essentially no prior assumed structure of the data generating process. In theory, an arbirarily complex neural network can learn any such process, which is called the *general approximation theorem*.

This often causes the misconception that neural networks[10] will eventually obsolesce all other methods, at least in terms of predictive performance. However, even when the goal is purely prediction, there is a catch: all these parameters need to be learned by the model. More parameters require more data.

[10] Often synonymously used with: deep learning, or artificial intelligence (AI).

Furthermore, just because an advanced algorithm *should* be able to learn any kind of function, does not guarentee that it *will* within a reasonable amount of time. Nor does the general approximation theorem imply that any particular amount of data will suffice in learning the underlying process.

Lastly, there is the *no free lunch theorem*, which states that if an algorithm performs well on a certain class of problems, then it must perform worse on the set of remaining problems. In other words, there will never be a singular algorithm which outperforms all others regardless of the problem fed to it.

With that in mind, there are problems dominated by AI:

- Image classification;
- Speech recognition;
- Natural language processing.

With large, labeled data sets (e.g. $n > 10.000$) and sufficient computing power, deep neural networks tend to achieve predictive accuracy on these problems beyond what is possible with other methods. However, there is no general-purpose best model.

2 Linear Regression

Regression analysis is any analysis involving a response variable and one or more explanatory variables. If the effect of each explanatory variable can be expressed in terms of an intercept and a slope, then the regression is linear.

Linear regression is a powerful and versatile tool, that expresses a relationship in terms of intercepts and slopes (fig. 2.1). One might argue this restricts us to a narrow class of scientific problems, but linear models can capture many kinds of relationships by transformation (paragraph **2.4**) and generalization (chapter **3**). Moreover, if the assumptions hold, linear models are more efficient and interpretable than non-linear models (e.g. tree-based models, neural networks).

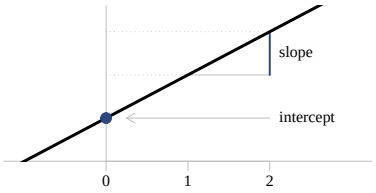


Figure 2.1: The intercept is where the line crosses the y-axis. The slope is the increase in y, for a unit increase in x.

2.1 Simple Linear Regression

A linear model with one response variable and one explanatory variable is called a **simple linear model**:

$$\underset{\text{response variable}}{y} = \underset{\text{intercept}}{\beta_0} + \underset{\text{slope}}{\beta_1} \cdot \underset{\text{explanatory variable}}{x} + \underset{\text{residual}}{\epsilon} \tag{2.1}$$

For example, a tree's circumference grows approximately linearly with its age (fig. 2.2). This model has **two parameters**:

- The **intercept** (β_0) is where the line crosses the y-axis ($x = 0$);
- The **slope** (β_1) is the steepness of the line.

Given some observations, the parameters can be estimated (fig. 2.2). According to this model, after one year the average circumference is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x = 21.7 + 7.2 \cdot 12 = 108.1 \text{ mm}$$

Predictions could also be made for a week, a month, or 100 years, but their accuracy heavily depends on whether the estimated relationship still holds beyond the range of the observations.

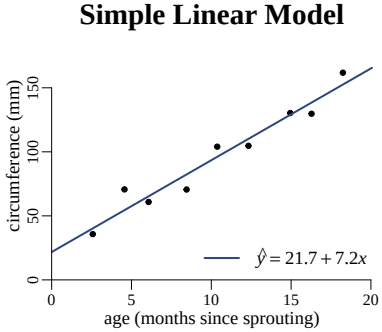


Figure 2.2: Linear relationship between tree age and circumference.

If there is a true linear relationship, estimates are *on average* correct. However, individual trees differ, and so there is also a stochastic part: the residuals (fig. 2.3). These are assumed to be normally distributed with a mean of zero and constant variance: $\epsilon \sim \mathcal{N}(0, \sigma^2)$. This has several implications:

- **Normality** implies that larger deviations from the line in either direction should be rarer than smaller deviations;
- **Zero mean** means the residual term is *on average* 0. Because of this, the residual term is ignored for point estimates ($\hat{y}$);
- **Constant variance** means the spread of the residuals should be more or less the same along the regression line.

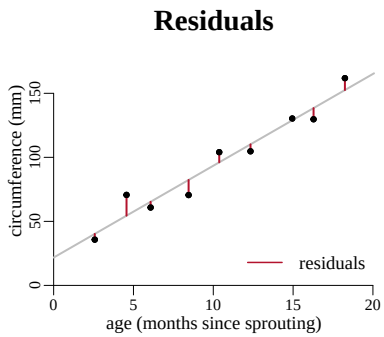


Figure 2.3: The residuals of a linear model are the vertical distances to the regression line (ϵ in eq. 2.1).

In the next paragraphs, we will go over the assumptions (**2.2**), how to assess problems (**2.3**), how to overcome certain problems (**2.4**) and how to estimate the parameters (**2.5**).

Simple linear regression is no statement about the difficulty of the technique, but rather about the presence of a *single* explanatory variable. Linear regression with multiple explanatory variables is also possible and will be covered in paragraph **2.6**.

2.1.1 *Simple Linear Regression in R*

To demonstrate the concepts explained so far, a practical example of a simple linear model in R is given below:

1. Suppose there is a presumed linear relationship between the sepals and petals of *Iris* flowers (fig. 2.4): If a flower's sepals are longer, its petals are expected to be proportionately longer as well. A simple linear model can capture this relationship.

   In this example, the response variable is sepal length and the explanatory variable petal length. However, since neither variable is "explained" by the other in the causal sense of the word, the other way around would be valid too. The only difference would be the interpretation of the intercept and slope.

   The following code loads the `iris` data set into the object space and estimates the coefficients of a simple linear model:

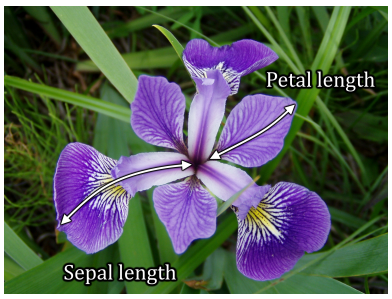


Figure 2.4: The length of sepals and petals of *Iris* flowers.

```
data(iris)
LM <- lm(Sepal.Length ~ Petal.Length, data = iris)
```

Interpret this code as follows: Regress sepal length (y) on petal length (x), from the data set `iris`. The model is then: $y_i = \beta_0 + \beta_1 x + \epsilon_i$.

That's all there is to it! Estimates are now stored into the object `LM`.

2. Printing `LM` returns the estimated intercept ($\hat{\beta}_0$) and slope ($\hat{\beta}_1$):

```
print(LM)
```

```
  (Intercept)  Petal.Length
       4.3066        0.4089
```

3. According to this model, *Iris* flower sepal length is on average: $4.3 + 0.4 \times (\text{petal length})$. A milimeter increase in petal length thus corresponds to about 0.4 mm longer sepals (fig. 2.5).

   You can plot the regression line as follows:

```
plot(Sepal.Length ~ Petal.Length, data = iris)
abline(coef = coef(LM), col = 4, lwd = 2)
```

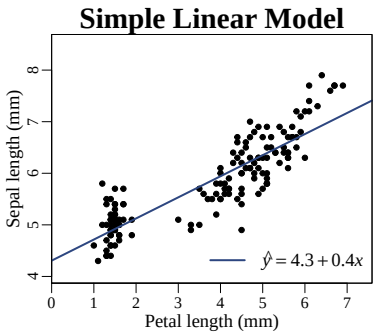


Figure 2.5: Visualization of the simple linear model in this example.

4. A summary can be printed, which contains additional information about the model. First, consider the coefficients tab:

```
summary(LM)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.30660    0.07839   54.94   <2e-16 ***
Petal.Length  0.40892    0.01889   21.65   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

   The summary table above can be interpreted as follows:

- `Estimate` is the estimated intercept and slope;
- `Std. Error` is the standard error, a measure of precision of the estimates (the smaller, the more precise);
- `t value` belongs to a one-sample t-test, to assess whether the estimate differs significantly from zero: $t = \frac{\hat{y} - \mu_0}{\text{SE}} = \frac{\texttt{Estimate} - 0}{\texttt{Std. Error}}$;
- `Pr(>|t|)` is a two-sided p-value corresponding to the t-value. Both are lower than the machine precision.[1] This is lower than any reasonable choice for the level of significance (α), so the estimates are significantly non-zero. This does not mean that they are important, just that they are very unlikely to be this large and consistent in the sample, if they were actually 0 in the population of *Iris* flowers. Of course, this is assuming there is a linear relationship at all;
- The significance codes (e.g. `***`) can be ignored. While popular in scientific literature, they have no added value or meaning.[2]

[1] The machine precision is the lowest number that R considers larger than 0. By default, it is $2 \cdot 10^{-16}$.

[2] `*` means $p < 0.05$, `**` means $p < 0.01$, `***` means $p < 0.001$.

Under the coefficients tab, there is some additional information:

```
Residual standard error: 0.4071 on 148 degrees of freedom
Multiple R-squared:   0.76, Adjusted R-squared:  0.7583
F-statistic: 468.6 on 1 and 148 DF,  p-value: < 2.2e-16
```

Interpret the information above as follows:

- The standard error of the residuals is `0.4071`;
- There were $n = 150$ observations and $p = 2$ parameters (intercept & slope), so there are $n - p = 148$ degrees of freedom left;
- R^2 `= 0.76`, which means 76% of the variance in sepal length is explained by the petal length. In other words, given the petal length of *Iris* flowers, over three-quarters of the variance in sepal length can be accounted for;
- Adjusted R^2 is used in multiple regression (ignore for now);
- The F-statistic belongs to an F-test. It compares the variance explained by the explanatory variable to the variance of the residuals. If this test is significant, then the model explains a significant amount of the variance in the response variable. You can do the same test manually by running `anova(LM)`. Though beyond the scope of this book, it can be shown that in the case of simple linear regression, this test is equivalent to the t-test for the slope;
- The p-value is extremely small (`< 2.2e-16`), so this model explains a significant amount of the variance in sepal length.

5. 95% confidence intervals (CI) can also be easily generated for the estimates. Recall that for normally distributed variables, these are approximately equal to `Estimate` $\pm 1.96 \times$ `Std. Error`:

```
confint(LM)
```

```
                2.5 %    97.5 %
(Intercept)  4.1516972 4.4615096
Petal.Length 0.3715907 0.4462539
```

6. A different confidence level can also be chosen:

```
confint(LM, level = 0.99)
```

```
                0.5 %    99.5 %
(Intercept)  4.1020508 4.5111560
Petal.Length 0.3596262 0.4582184
```

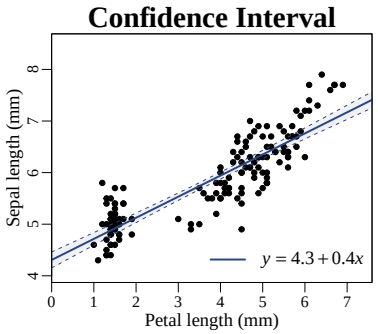


Figure 2.6: Visualization of the 95% CI for the intercept & slope.

7. **Confidence intervals do not show where the data are expected**. Rather, they show the uncertainty of the intercept and slope. An interval for where the data are expected is a *prediction interval* (PI). While these are a little bit more complicated to calculate, it can be done in a few lines of code in R:

```
plot(Sepal.Length ~ Petal.Length, data = iris)
x_values  <- seq(0, 8, 0.1) # 0.0, 0.1, 0.2, ..., 8.0
newdata   <- data.frame(Petal.Length = x_values)
y_predict <- predict(LM, newdata, interval = "predict")
lines(y_predict[, 1] ~ x_values)          # Regression
lines(y_predict[, 2] ~ x_values, lty = 2) # Lower bound
lines(y_predict[, 3] ~ x_values, lty = 2) # Upper bound
```

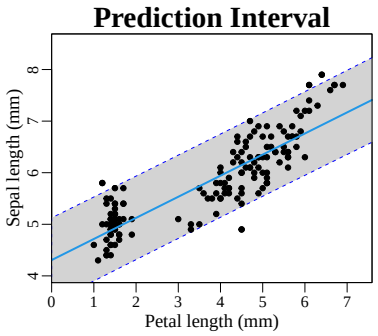


Figure 2.7: Visualization of the 95% prediction interval of the model.

When performing your own analyses, feel free to reuse the code here, or make use of readily available implementations,[3] as long as you understand what is being shown.

[3] For example, the function `plotFit` in the package `investr`.

2.2 Assumptions

Estimates, standard errors, p-values & confidence intervals obtained from linear models are only meaningful if there is indeed a linear relationship, and the assumptions of the linear model hold. Most assumptions concern the residuals (fig. 2.3). Below is a summary:

1. **Linearity of the relationship** between x and y;
2. **Normality of the residuals** with mean equal to zero;[4]
3. **Constant variance** of the residuals along the regression line;
4. **Independence of observations**;
5. **Fixed explanatory variable(s)**, unchanged if measured again.

Although not really an assumption, often also mentioned is:

6. **No outliers**, since linear regression is not robust to outliers.

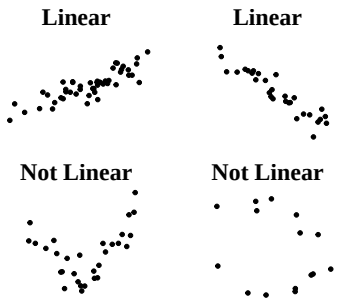
[4] If the mean were unequal to zero, the observations would systematically deviate from the line (i.e. there exists a better line).

2.2.1 Linearity of the Relationship

Linearity implies that the relationship can be expressed in terms of intercept(s) and slope(s). Naturally, if there is no linear relationship, a linear model is not appropriate.

However, transforming the data (e.g. squaring the explanatory variable) can sometimes yield a linear relationship, even though the original variables were not linearly related (paragraph **2.4**).

- **Linearity can be assessed with a residuals vs fitted plot (2.3.1).**



2.2.2 *Normality of the Residuals*

Residuals are assumed to be normally distributed around the regression line (fig. 2.8). This means most observations are near the regression line, some farther away and a rare few far away. Moreover, since the normal distribution is symmetric, the chance of deviations in either direction from the line is equal.

If this assumption is not reasonable because the response variable is inherently not normal, a GLM (chapter 3) can offer an alternative.

- **Non-normality can be diagnosed with a QQ-plot (2.3.2).**

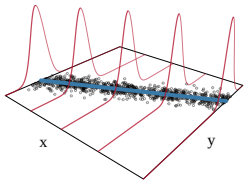


Figure 2.8: Imagine a normal distribution along the regression line, describing the chance that **observations** are at a certain distance from the line.

2.2.3 *Constant Variance*

The variance of the residuals should be approximately constant along the regression line. In figure 2.8, this means the width of the normal distribution should remain approximately the same along the line. In a scatter plot, non-constant variance looks like a cone-shape (fig. 2.9).

Non-constant variance can be caused by correlation between the explanatory variable and the error (ϵ). Consider for example a device that measures less accurate, the higher the value of interest is.

However, non-linearity, non-normality, non-constant variance and outliers can be difficult to discern from visual inspection alone. Consider the most probable cause from a biological point of view.

- **Non-constant variance can be seen in a scale-location plot (2.3.3).**

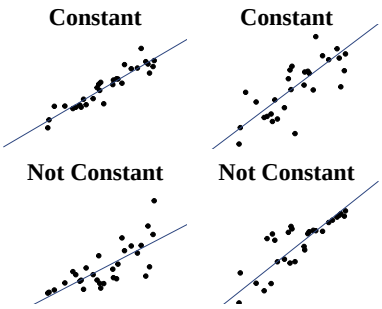


Figure 2.9: Variance should not increase or decrease along the regression line.

2.2.4 *Independence of Observations*

Independent observations come from distinct experimental units: If 5 leaves of one plant were to be measured (fig. 2.10), conclusions can only be drawn about that particular plant. After all, it is completely unknown how far that plant deviates from the 'average plant'.

On the other hand, if 5 different plants were measured (fig. 2.11), conclusions can be drawn about a larger group of plants, because we now have an idea of how much individual plants differ from one another and what the average plant is like. Even if experimental units are genetically the same and live in comparable environments, they are still not identical. Measuring experimental units more than once is *pseudoreplication* and only decreases measurement error.

In repeated measures studies, observations are not independent by design, such as patients observed before and after a treatment. This can be dealt with by a mixed model (***Advanced Statistics***).

- **Independence of observations is inherent to the experimental design and cannot be (easily) diagnosed.**

Figure 2.10: Pseudoreplication occurs when the same experimental units are measured more than once.

Figure 2.11: True replication involves distinct experimental units.

2.2.5 *Fixed Explanatory Variable(s)*

To understand this assumption, the difference between a fixed and random effect in the context of mixed models must be explained first. Mixed models are properly introduced in ***Advanced Biostatistics***.

Fixed vs Random

Continuing the leaf/plant example (fig. 2.11), one might wonder: What if we measure multiple leaves from multiple plants and just add a variable "`plant`" to correct for the difference between plants? This is an example of a random effect and is the motivation behind mixed models. An ordinary linear model cannot do this; it assumes all effects to be fixed.[5] A rule of thumb for fixed effects is that they should remain unchanged if the experiment were to be carried out in duplo. Examples of fixed effects include:

[5] If you were to include an effect for plants, the standard errors (and thus the p-values and confidence intervals) would be incorrect.

- Age: if measured again immediately, its value is the same;
- Gender: idem dito;
- Dose of a drug: set by the experimenter;
- Temperature: should usually not change in a second experiment.[6]

[6] Just because it cannot *technically* be measured without error, does not mean the real temperature is different when measured twice at the same place and time.

Examples of random effects include:

- Patient: if we repeat an experiment, we *will* want to measure new patients, to learn about a larger group of patients in total;
- Plant: idem dito;
- Hospital: unless we want to say something about one hospital specifically, we could measure patients from any hospital;
- Location: idem dito.

This has an important implication for experimental design: Dependent measures should be avoided unless they are required to answer the research question. There are no simple tricks to avoid having to use more complex models in the presence of random effects.

An exception is when there is interest in a single variable across two time points, in which case a paired t-test will suffice.

- **Whether variables are fixed depends on the experimental design and the population of interest. It cannot be diagnosed.**

2.3 *Visual Diagnostics*

Even if the assumptions of a linear model seem reasonable at first, they may not always hold in practice. Visual diagnostics can help decide whether the model is appropriate.

In R you can simply run `plot(model)`, which will return 4 plots that can diagnose linearity, normality, constant variance, and the presence of outliers, respectively.

For example, try running:

```
par(mfrow = c(2, 2))
LM <- lm(Sepal.Length ~ Sepal.Width, data = iris)
plot(LM)
par(mfrow = c(1, 1))
```

The next sections explain how to interpret each of these plots, and what consitutes a violation. If a violation persists, an alternative model should be considered.

Diagnostic plots serve as guidelines. Small deviations from what is expected can generally be ignored, large deviations are a cause for concern. Whether a violation is large enough to be considered serious can be reasonably learned from experience. Statistical tests for assumptions should be avoided.

- **Diagnostic plots are easier to interpret with larger sample size.**

2.3.1 *Residuals vs Fitted*

This diagnostic plot is primarily used to diagnose deviations from linearity. It shows the distances of the observations from the regression line (residuals) along the regression line (fitted values).

If the true relationship is linear, the smoothing line will be more or less flat (fig. 2.12). In case of quadratic, exponential, or logarithmic relationships, a parabolic shape will be visible (fig. 2.13).

Exponential and logarithmic relationships also tend to cause non-constant variance and should therefore be visible in the scale-location plot as well (fig. 2.17).

The residuals vs fitted plot itself can also reveal non-constant residual variance. In the case of non-constant variance, there is not necessarily a strong trend in the smoothing line. However, the residuals will form a cone shape along the smoothing line (fig 2.14).

Occasionally, there may be a single influential observation yiedling the impression of non-linearity or non-constant variance. For this reason, residuals on the outskirts of the plot are labeled (see the numbered points in figures 2.12 - 2.14). To assess whether such points are statistically outliers, use the Cook's distance (fig. 2.19).

- **Non-linearity and/or non-constant variance can in certain cases be remedied through transformation (paragraph 2.4).**

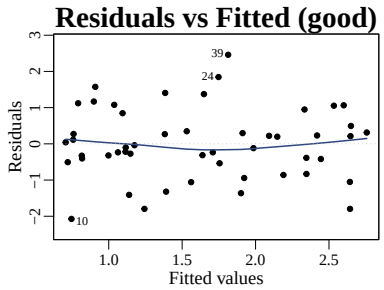


Figure 2.12: Diagnostic plot to check for deviations from a linear relationship.

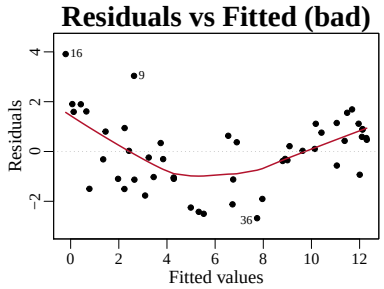


Figure 2.13: Quadratic and exponential relationships results in a parabolic shape.

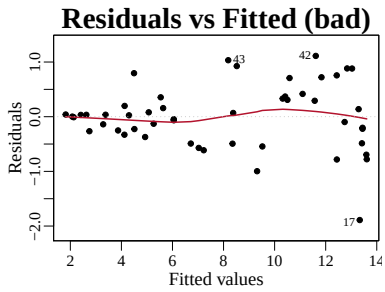


Figure 2.14: The residual variance increases along the smoothing line.

2.3.2 QQ-Plot

A quantile-quantile plot can be used to assess whether the assumption of normality is reasonable. It is difficult to judge whether the residuals follows a bell-shaped curve, so the quantiles of a theoretical normal distribution are used instead, plotted against the quantiles of the residuals, which form a straight line. Deviations from this line represent deviations from normality and 95% confidence intervals can be included to judge the severity of the deviations. Deviations beyond these confidence intervals are considered significant.

An example of approximately normally distributed residuals is shown in figure 2.15. Although there are minor deviations from the blue line representing a theoretical normal distribution, none are beyond the confidence intervals. A problematic example is shown in figure 2.16 (these are Poisson distributed; see chapter 3).

In the presence of non-linear effects or outliers, the residuals may also deviate from normality. Always take all four diagnostic plots into consideration.

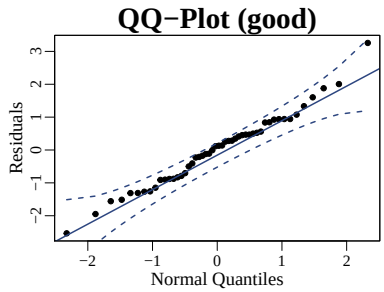


Figure 2.15: QQ-plot for normality.

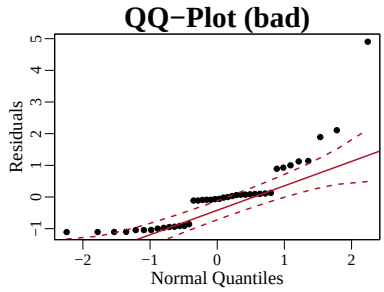


Figure 2.16: Non-normal residuals.

- **GLMs can model inherently non-normal processes (chapter 3).**

2.3.3 Scale-Location

Constant residual variance along the regression line can be checked using the scale-location plot (fig. 2.17). A modified version of the residuals vs fitted plot, the scale-location plot shows the square root of the absolute values of the residuals along the regression line.

This plot makes it easier to determine whether variance is constant: Whereas non-constant variance yields a cone-like shape in the residuals vs fitted plot (e.g. fig. 2.14), it simply results in an increasing or decreasing trend in the scale-location plot (2.18).

If caused by logarithmic or exponential effects, non-constant variance can be resolved through transformation. Other sources of non-constant variance include non-normal processes (e.g. counts, ratios). These can be modeled with a GLM.

If there is an unknown source of non-constant variance (e.g. an unknown important variable not included in the model), robust linear regression can offer a solution by reweighting observations based on their variance.

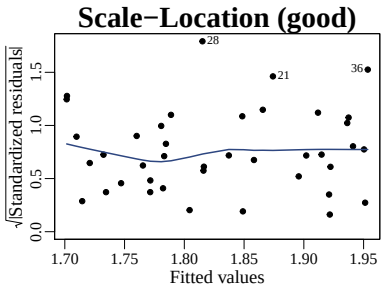


Figure 2.17: Constant residual variance.

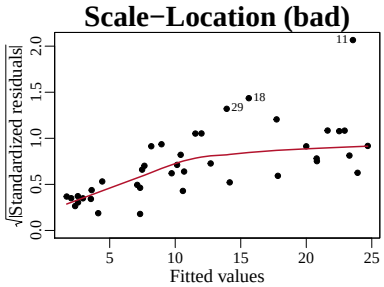


Figure 2.18: The residual variance increases along the fitted values.

- **Depending on the source of non-constant variance, transformation, GLM or robust linear regression can resolve the problem.**

2.3.4 *Cook's Distance*

Linear regression is not robust to outliers, meaning that a single extreme observation can yield a large difference in the intercept and slope. These outliers can have different causes: Some processes simply produce an extreme result every now and then, while at other times, the outlier is the result of an error in data collection.

The presence of potential outliers or strongly influential observations can be assessed through the Cook's distance plot (fig. 2.19). Any observations beyond 0.5 Cook's distance (the first boundary) are noteworthy and any observations beyond 1.0 Cook's distance (the second boundary) are considered statistical outliers.

The x-axis represents the *leverage*, which is the extent to which an observation can influence the regression line. High leverage implies the explanatory variable is unusually small or large, but need not be cause for concern if it is in agreement with the rest of the observations (fig. 2.21). If an observation is outlying, but has low leverage, then its effect on the regression line is minimal. Outliers with high leverage are greater cause for concern (fig. 2.22).

Outliers can be caused by an incorrect entry, such as a misplaced decimal seperator (e.g. 54 instead of 5.4), or by an error or anomaly that occurred during experimentation. Always consider the cause of outlyingness and consult the labjournal, if available. Only outliers known to have arised from some accidental error in experimentation or data collection should be removed from the data set. Outliers of unknown origin should never be removed from the data set, since these extreme observations could be legitimate, in which case they might be your most important information about the underlying process.

In the presence of one or more high leverage outliers, if there are no solid grounds for removing the observation (e.g. no labjournal entry explaining the strange value), consider robust linear regression as an alternative. This method estimates a regression line that is less affected by differences in the size of the residuals.

- **Only remove an outlier if it can be traced back to an error in experimentation (e.g. from a labjournal entry).**

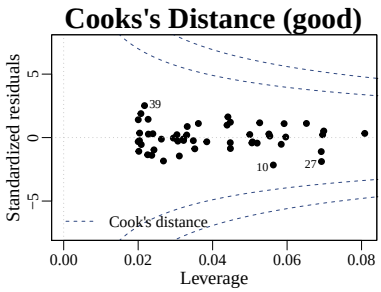


Figure 2.19: Diagnostic plot to check for potential outliers. The first boundary is 0.5 Cook's distance and the second 1.0.

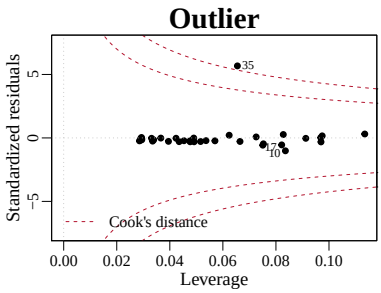


Figure 2.20: An outlier beyond 1.0 Cook's distance.

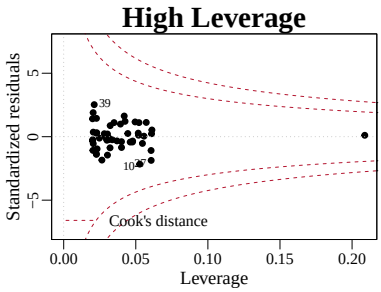


Figure 2.21: An observation with high leverage, though not statistically outlying. This occurs when there is an extreme observation, but it falls nicely on the regression line.

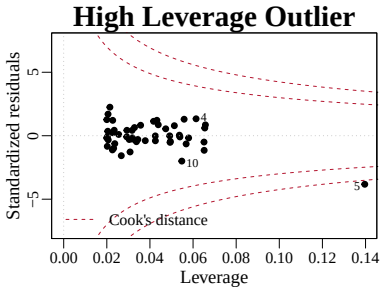


Figure 2.22: A potential outlier with high leverage. These are the most influential cases and thus the greatest cause for concern.

2.3.5 *Still Not Sure? (*)*

Simulating data can help to get a better idea of typical diagnostic plots. The code below generates data that are by definition linear. However, as you will see, the uncertainty of the sample will almost always cause small deviations. Run the code below as many times as you like for different values of n, so that you can get a feel for what a "typical" set of diagnostic plots look like:

```
n       <- 30                          # The sample size
x       <- runif(n)                    # Draw n random x values
beta_0  <- rnorm(1)                    # Draw a random intercept
beta_1  <- rnorm(1)                    # Draw a random slope
sigma2  <- rexp(1)                     # Draw a random variance
epsilon <- rnorm(n, 0, sigma2)         # Draw n random residuals
y  <- beta_0 + beta_1 * x + epsilon # Simulate y values
LM <- lm(y ~ x)                        # Fit a linear model

par(mfrow = c(2, 2))
plot(LM, which = 1)
car::qqPlot(residuals(LM), main = expression("Normal Q-Q"))
plot(LM, which = c(3, 5))
par(mfrow = c(1, 1))
```

The least informative of the 4 base plots is the QQ-plot, as it has no indication whatsoever of severity. The code above therefore replaces it with the QQ-plot from the package `car`, which draws confidence intervals as well. Install the package using `install.packages("car")` if it is missing.

Especially for small sample sizes (say, $n \leq 10$), the validity of your model will largely depend on theoretical substantiation. Try lowering the sample size `n` in the example above to something small, and observe how common it is to find "serious violations", even though the true relationship was simulated to be perfectly linear. With only a few data points, there is not much to be learned from visual inspection. If this is unsatisfactory, because you cannot postulate a linear relationship *a priori*, consider collecting more data first.

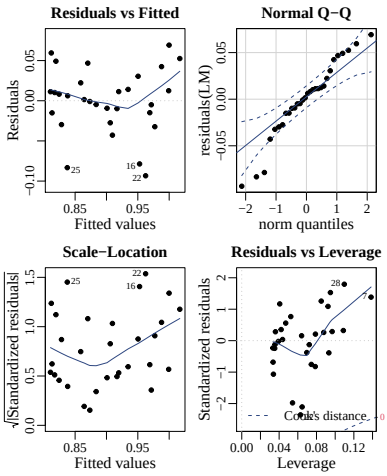


Figure 2.23: The outcome of a particular run of the example code. Even though the relationship was simulated to be perfectly linear, samples always show small deviations from the assumptions.

2.4 Transformation

A linear model of the form $y = ax + b$ can capture quadratic, cubic, logarithmic, exponential, or even sinusoidal relationships. That might seem a little strange at first, but it is fairly easy to demonstrate why:

1. Suppose there is a true relationship: $y = ax^2 + b$;
2. Create a new variable: $z = x^2$;
3. Include this transformed variable in the model: $y = az + b$.

The resulting relationship between y and the transformed variable z is linear, even though the relationship between the original variables is not. It is also possible to visualize why this works, by replacing the x-axis of the original variable (fig. 2.24) with that of the transformed variable (fig. 2.25).

Rather than transforming the explanatory variable, it is also possible to transform the response variable in the opposite 'direction'. That is to say, a squared transformation on x has a similar result to a square-root transformation on y (fig. 2.26). Note that transforming y also affects the residuals, while transforming x does not. Which is better to transform depends on whether one is easier to justify than the other from a biological perspective (section ***2.4.1***).

- **Transformations solve non-linearity. If the underlying problem is non-normality (e.g. count data or ratios), see GLM (chapter 3).**

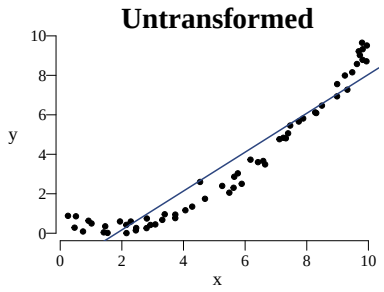


Figure 2.24: The relationship between y and x is quadratic in this example.

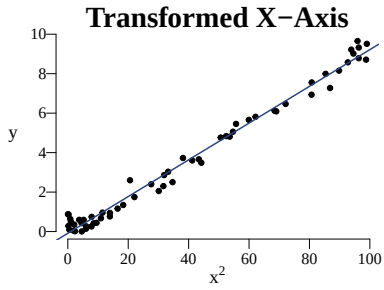


Figure 2.25: The relationship between y and x^2 is linear.

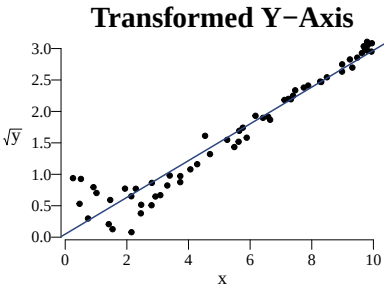


Figure 2.26: The relationship between $\sqrt{y}$ and x is also approximately linear, but introduces non-constant variance in this example.

2.4.1 Justifying a Transformation

Although there often exists some transformation that yields a linear relationship, the proposed transformation should also make sense from a biological point of view, for example:

- Surface area of leaves increase quadratically with length (x^2);
- Bacteria can grow exponentially over time (e.g. 2^x for doubling);
- Effectiveness of a drug tends to plateau with concentration ($\log(x)$).

Examples that are harder to justify include:

- Power transformations of 3 or higher (e.g. x^3, x^4, x^5);[7]
- Non-integer power transformations (e.g. $x^{1.47\cdots}$);
- Multiple transformations (e.g. $\log(\frac{1}{x^2})$).

In some cases it is justifiable to transform x but not y and in other cases the other way around. If no transformation can be justified within the context of your research, then transformation should generaly be avoided.

[7] Exceptions exist, such as volume (m^3), which increases cubically with length, width or height (m).

2.4.2 *Box-Cox*

Box-Cox transformations are a set of possible power transformations of y, such that the relationship with x is linear. A power transformation is any transformation y^λ (except $\lambda = 0$). The likelihood of λ can be calculated and a confidence interval can help choose the nearest useful transformation. If the confidence interval is very wide (i.e. includes multiple possible transformations), a Box-Cox transformation is unlikely to resolve non-linearity and it is better not to transform.

In R, the function `boxcox()` calculates the likelihood of λ and plots the result (fig. 2.27). A 95% confidence interval for power transformations on y yielding an approximately linear relationship with x is also included (dashed line).

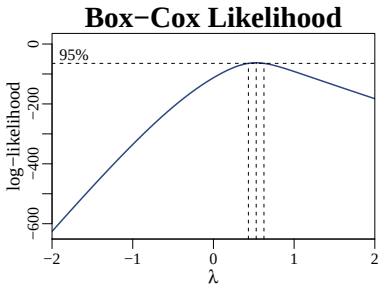


Figure 2.27: Likelihood and 95% confidence interval for λ. The confidence interval contains $\frac{1}{2}$, so a suitable transformation would be $y^{\frac{1}{2}} = \sqrt{y}$.

In principle, any λ will do, but a transformation of e.g. $y^{-0.354...}$ is difficult to interpret and hard to justify. Hence, a list of suggested transformations is included in the table below:

95% CI of λ includes:	Suggested	Alternative
-2	$\frac{1}{y^2}$	$\frac{1}{\sqrt{x}}$
-1	$\frac{1}{y}$	$\frac{1}{x}$
0	$\log(y)$	$\exp(x)$
$\frac{1}{2}$	$\sqrt{y}$	x^2
1	y	x
2	y^2	$\sqrt{x}$

Table 2.1: Suggested transformations of y and corresponding alternative transformations of x. The logarithm is used at $\lambda = 0$, since $y^0 = 1$ for all values of y. Any base can be used for the logarithm of y, or the exponent of x.

As you can see from the table above, if the confidence interval includes 1, then the original variables produce the best fit. For example:

```
LM <- lm(Petal.Length ~ Petal.Width, data = iris)
boxcox(LM)
```

Box-Cox transformations still require theoretical justification. Just because you can find a better fit in the *sample*, does not mean that the population relationship is e.g. quadratic. Regardless, Box-Cox can be quite useful to choose among a set of candidate transformations. For example, from a theoretical point of view, it is sometimes difficult to choose between a logarithmic or square-root transformation, as both will incur a plateau (fig. 2.28).

Note that in case of logarithmic transformations, $\log(y+1)$ is often used instead of $\log(y)$, since $\log(0)$ is undefined. This trick is often used when a logarithmic transformation is required and zeroes occur in the data. If the response variable can be negative, then neither a logarithm nor a square-root transformation are appropriate. In that case, it may be better to transform the explanatory variable.

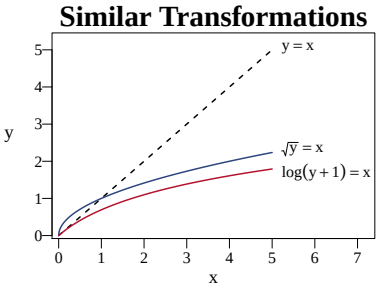


Figure 2.28: $\log(y+1)$ and $\sqrt{y}$ both result in diminishing returns, albeit at a different rate.

2.5 Ordinary Least Squares

In ordinary linear regression, the parameters are estimated such that the model has the smallest possible squared distances to the observations. This objective is called *ordinary least squares* (OLS):

$$\hat{y}_i = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \epsilon_i$$
$$\text{OLS} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

As long as there are more observations than parameters ($n > p$), the OLS solution is *unique*. This means there is only one such solution, and any other coefficients would increase the sum of squared residuals (figure 2.29).

In simple linear regression, OLS estimates of the intercept and slope can be obtained as follows:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (2.2)$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$$

Where $\bar{y}$ is the mean of the response variable and $\bar{x}$ the mean of the explanatory variable.

2.5.1 What Statistical Software Actually Does (*)

Equation 2.2 is nice for explanation, because it can be easily performed by hand in case you wanted to check what was actually happening. However, statistical software does not find the estimates one by one, but simultaneously. This can be done by performing the following matrix multiplication and inversion:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\mathbf{y} \qquad (2.3)$$

Where $\mathbf{X}$ is the design matrix,[8] and $\mathbf{X}^\mathrm{T}$ the matrix transpose of $\mathbf{X}$. This can be done for any number of explanatory variables, as long as there are more observations than explanatory variables. However, matrix inversion can be slow and numerically unstable, so R's `lm` function uses QR-decomposition, which is beyond the scope of this book.

Linear regression can be solved in many ways. Any of these will result in (practically) the same estimates as long as they minimize the OLS objective, because there is only one solution to it. Minor differences arise only due to precision used in the calculation.

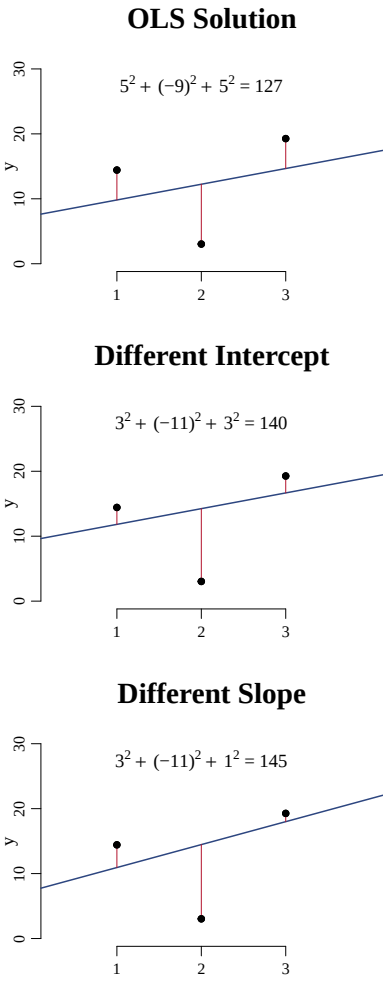


Figure 2.29: OLS finds the model with the smallest sum of squared residuals. Any change in the estimated coefficients increases this sum.

[8] In simple linear regression:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

The column of 1s represents the intercept (it is added everywhere). The slope is multiplied by x_i, the value of x for observation i.

2.6 Multiple Linear Regression

Multiple linear regression is the extension of the simple linear model to more than one explanatory variable. It is used to asses multiple effects on the outcome simultaneously. Consider for example the effect age (x_1) and bodyweight (x_2) on systolic blood pressure (y). On average, blood pressure is higher in older, but also in heavier individuals. Within a certain range of age and bodyweight, their relationship with blood pressure is approximately linear. Since there are two variables affecting the outcome, there are also two slopes:

$$y_i = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \epsilon_i \tag{2.4}$$

According to this model, someone's systolic blood pressure (y_i) is equal to the intercept (β_0) plus a slope (β_1) times that person's age (x_1) plus another slope (β_2) times that person's weight (x_2) plus some random deviation from the model (ϵ_i). Try plugging in $\hat{\beta}_0 = 82.2$, $\hat{\beta}_1 = 0.40$ and $\hat{\beta}_2 = 0.23$ and predict your own blood pressure.[9]

Multiple linear regression can have any number of variables. For a linear model with p explanatory variables, this can be written as:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j \cdot x_j + \epsilon_i \tag{2.5}$$

Once there are multiple explanatory variables, phenomena like interaction (**2.6.2**), overfitting (**2.6.4**) and model selection (**2.6.5**) become relevant. Because of this, a seventh assumption is made:

7. **Explanatory variables are not (highly) colinear (section 2.6.3).**

Hyperplane (*)

A slope in two directions is still possible to imagine (fig. 2.30), the regression *line* just becomes a *plane* which can be traversed in two dimensions. The residuals are then the vertical distances to the plane (i.e. how much the actual outcome differs from the height of the plane; fig. 2.31). For more than 2 slopes, there is no intuitive visualisation. The generalization of a line or plane to any number of dimensions is called a *hyperplane*.

While we can still talk about an intercept and slopes, there is no real way of visualizing hyperplanes. This is one of the reasons residual diagnostics are important: They reveal what cannot be seen from the model output. If the assumptions are valid, the residual plots should show little or no deviations and the model describes the systematic, linear part of the process as it should.

[9] Recall that the expected value of the error term (ϵ_i) is 0.

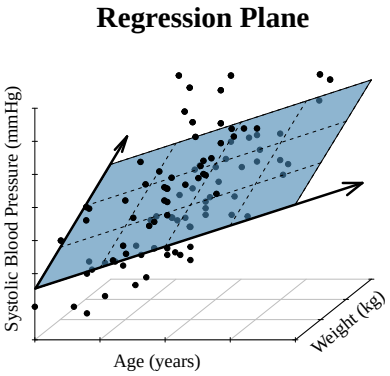


Figure 2.30: Systolic blood pressure linearly increases with age, but also with bodyweight. A line in two directions forms a plane.

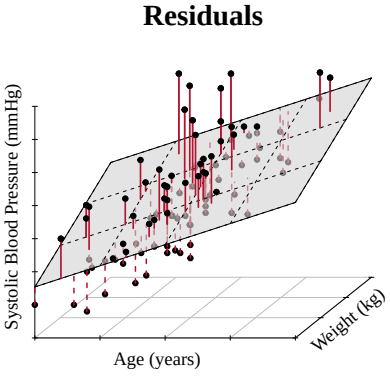


Figure 2.31: The residuals of figure 2.30 are the vertical distances to the plane. Negative residuals are indicated by dashed linepieces.

2.6.1 *Dummy Variables*

In the example at the beginning (eq. 2.4), age and weight are both continuous. However, continuity of explanatory variables is not an assumption of linear models. Categorical explanatory variables (e.g. gender) can also be included, by using *dummy variables*, which are equal to 1 if the observation has a certain category and 0 otherwise. For example, gender can be modeled by including a variable for being male:

$$y_i = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_{\text{male}} + \epsilon_i$$

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

One might wonder how this can be linear, but recall that linearity only means that all coefficients must be describable as intercepts and slopes. The slope of 'being male' (β_2) is only added if the observation is male ($x_{\text{male}} = 1$). For females ($x_{\text{male}} = 0$), the expected value is: $\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot 0 = \beta_0 + \beta_1 \cdot x_1$. The visual interpretation is that the model now has a second line for males, shifted up by β_2 (fig. 2.32).

If there are more than two categories (e.g. three species of *Iris* flowers), multiple dummy variables can be included, e.g.:

$$y_i = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_{versicolor} + x_{virginica} + \epsilon_i$$

$$x_{versicolor} = \begin{cases} 1 & \text{if } versicolor \\ 0 & \text{otherwise} \end{cases}$$

$$x_{virginica} = \begin{cases} 1 & \text{if } virginica \\ 0 & \text{otherwise} \end{cases}$$

A third dummy variable is not required, because an observation that is neither *versicolor* nor *virginica* must be *setosa* (fig. 2.33).[10]

Adding categorical variables is very simple in R. When a `factor` is added to a linear model, R automatically converts it to dummy variables. The first group[11] is taken to be the reference group and is the group for which all dummy variables are equal to zero.

- **Categorical *response* variables can also be modelled by linear models, but require a different method (see section 3.2.1).**

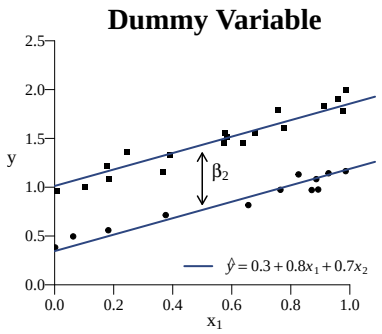


Figure 2.32: A dummy variable adds a fixed amount (β_2) to the intercept of the regression line if $x_{\text{male}} = 1$.

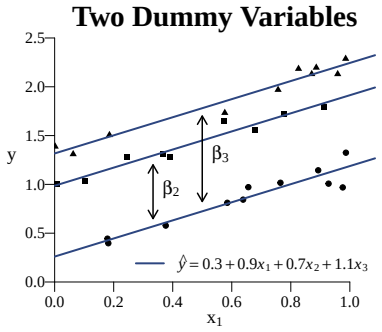


Figure 2.33: Two dummy variables, describe three classes (● = *setosa*, ■ = *versicolor* , ▲ = *virginica*).

[10] In fact, a third dummy would be colinear with the intercept (**2.6.3**).

[11] To see which group is first, use `levels(x)`, where `x` is the categorical variable to be converted to dummy variables. To change which group is first, use `relevel(x, ref = "...")`, where `...` is the group to be first. To convert a variable to categorical, use `as.factor(x)`.

ANOVA & Linear Regression (*)
Linear models with continuous *and* categorical explanatory variables are called analysis of **co**variance (AN**C**OVA). Sound familiar? Consider the following: A linear model with only dummy variables is identical to an ANOVA (***1.6.1***). The only difference is the tests typically associated with ANOVA (omnibus, post-hoc) and linear regression (significance of intercept and slopes). Using `aov()` or `lm()` makes no difference for the estimates and standard errors.

2.6.2 *Interaction*

When the effect of one explanatory variable depends on the value of another explanatory variable (fig. 2.34), those variables *interact*.

Positive interaction is similar to the concept of synergy (i.e. the sum is greater than its parts), and negative interaction is similar to cancellation (i.e. an increase in one explanatory variable reduces the effect of the another). Both are common occurrences in biological processes. Some examples are given at the end of this section.

Interaction is easiest to visualize using a model with one continuous and one categorical variable (i.e. ANCOVA). In figure 2.34, the effect (or slope) of x is different for the two categories. Equivalently, the difference between the categories depends on the value of x.

Interactions can also occur between categorical variables (fig. 2.35) or continuous variables (fig. 2.36). While the latter may be harder to represent visually, it is conceptually the same: A change in one variable corresponds to a change in the effect on the outcome of the other variable.

Principle of Marginality (*)
In the presence of an interaction term, there are no main effects of variables in that interaction. The individual variables cannot be interpreted without considering the interaction effect as well. Consider for example the effect of x in figure 2.34. It is not possible to state what the slope of x is. The slope can be reported for either category, but there is no 'overall' slope.

This may seems evident when inspecting the figure (after all, where in the figure would this overall slope be?), but this is important to take into consideration when interpreting the output of a model. In most statistical software, a slope for x *will* be reported, but keep in mind that this is a marginal effect. The actual slope of x depends on the value of the interacting variable. In the case of figure 2.36, the reported slope of x_1 will be that of x_1 when $x_2 = 0$.

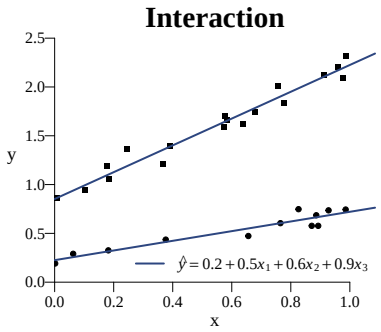


Figure 2.34: The slope of x depends on the category (the lines are not parallel).

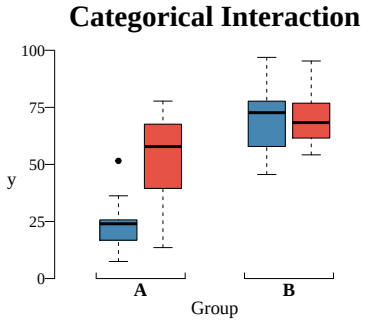


Figure 2.35: Example of an interaction between categorical variables. The difference in y between A and B differs for the red and blue group.

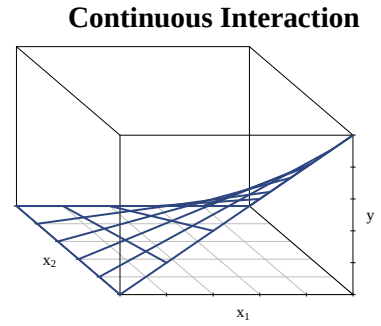


Figure 2.36: Example of an interaction between two continuous variables. The blue grid represents the regression plane. The slope of x_1 depends on the value of x_2 and vice versa. Without interaction, the plane would be flat.

Higher-Order Interactions

Interactions between more than two variables are technically possible, but also more difficult to interpret. In case all variables in the interaction are categorical, it is essentially a separate effect for every possible combination of categories. An example is the classical N, P, K experiment, where three treatments are applied to soil,[12] to find the optimal combination that maximizes the yield of a particular crop. The justification for a three-way interaction here is that the effect of one addition might depend on the effect of other additions to the soil. If potassium is the bottleneck for yield, adding just nitrogen might not have any effect. In this case, nitrogen's effect depends on the addition of potassium. Similarly, if one of the additives is already the most abundant, its effect might only be visible when all three treatments are applied on the same soil.

[12] Extra nitrogen (N), phosphate (P), and/or potassium (K).

Generally, higher-order interactions are harder to justify from a biological point of view. Moreover, including a higher-order interaction term can easily drain power because of the large number of degrees of freedom it costs to estimate an effect for every possible combination.

- **Avoid including higher-order interactions unless there is a theoretical justification, such as in the N, P, K experiment.**

Justifying an Interaction

Below are some examples where an interaction is justifiable from a theoretical point of view:

1. A drug only works in females;
2. A plant grows faster if and only if *all* nutrients are more abundant;
3. Baking soda is a cleaning agent, vinegar is a cleaning agent, but combined, they react (explosively) and do not clean effectively.

In the case of (1), the effect of dose interacts with sex,[13] because the slope will be zero for males. For (2), the effect of individual nutrients postively interact with each other. Lastly, (3) implies the amount of baking soda negatively interacts with the amount of vinegar.

[13] Whether this interaction is positive or negative depends on which class is coded 1 and which is coded 0 in the dummy variable.

2.6.3 *Colinearity*

Colinearity occurs when there is (too much) redundancy among explanatory variables. Specifically, if one explanatory variable is (approximately) a *linear combination* of other variables,[14] including all variables in the model will result in unstable estimates, inflated standard errors, or even failure to estimate the coefficients entirely.

[14] For example: $x_1 = \frac{1}{2}x_2$, or $x_1 = x_2 - x_3$, or $x_1 = (x_2 + x_3 + x_4)/3$ are all examples where x_1 is a linear combination of other explanatory variables.

An example of perfect colinearity is if you were to include a variable x_1 for the temperature in Celsius and another variable x_2 for the

temperature in Fahrenheit. Since $x_1 = \frac{5}{9}(x_2 - 32)$, all information in x_2 is already contained in x_1. Another is example is if you were to include a difference between two explanatory variables as a third explanatory variable.

Imperfect colinearity can also be problematic, such as including measures of 'almost the same thing'. For example, Weight, BMI, fat percentage, visceral fat, total fat, LDL cholesterol concentration, and $\frac{\text{HDL cholesterol}}{\text{Total cholesterol}}$ ratio are all used in obesity studies, but contain very similar information. The more of these are included, the more likely colinearity problems are. Considerations for variable selection in cases like these include: (1) Which measures add the most unique information; (2) Which measures are most likely related to the biological process; (3) Which measures are well-recorded (e.g. LDL concentration may not have been measured in all patients).

Since interactions (section **2.5.2**) are muliplicative combinations, including too many interactions can also result in colinearity issues.

Another example, genomics[15] experiments are inherently prone to colinearity due to regulatory effects between genes. Since one gene could affect the expression of another, including the expression of a large number of genes into a single model can result in many non-unique effects.

[15] Large-scale gene expression studies.

The best solution to colinearity is usually to exclude the colinear term(s) from the model. If you are not sure what term causes the colinearity, use the condition number as a guideline (next heading). Alternatively, if colinearity is unavoidable (e.g. with -omics research due to regulatory effects), then regularized models such as ridge regression can still estimate the coefficients.[16]

[16] Covered in Advanced Statistics.

Condition Number

There exist many measures to detect colinearity. A simple yet effective approach to diagnose colinearity is by calculating the condition number of the design matrix. In short: A large condition number is problematic. Whichever term[17] in the model leads to the largest decrease in condition number when removed from the model, should be removed. A simple implementation of the condition number for the purpose of diagnosing colinearity is included in the package `mctest`. As a rule of thumb, a condition number larger than 30 is problematic.

[17] Model terms are any of the explanatory variables or included interactions between them.

Although we do not go into the details of linear algebra in this book, **2.5** showed how the OLS objective can be solved through inversion. When there is colinearity, the design matrix contains one or more very small eigenvalues (0 in case of perfect colinearity). Small eigenvalues lead to numerical instability when performing matrix inversion (and 0 results in complete uninvertability).

2.6.4 *Overfitting*

Including too many explanatory variables, or explanatory variables which are too weakly correlated with the outcome can cause a model to *overfit*: The model will follow the trend of this particular sample too closely. Such a model generalizes poorly to the population.

Figure 2.37 illustrates this problem, by adding six polynomial terms for age (i.e. x, x^2, x^3, x^4, x^5, x^6). As a result, this model fits the sample perfectly, the residual distances to the model are zero! However, a perfect fit is not a good thing: Imagine another tree were to be measured, do you think it will be closer to a simple linear model (gray) or to the polynomial regression model (blue)?

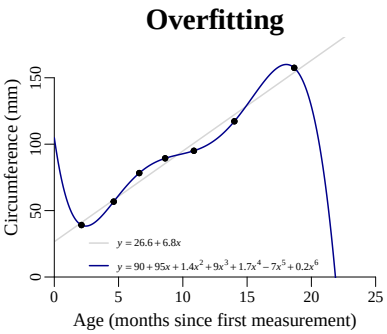


Figure 2.37: Example of overfitting. This model (blue) explains 100% of the variance in the data, but is strongly biased towards the sample. The simple model from figure 2.2 is shown in gray.

Consider the bizarre interpretation of the polynomial model: Trees shrink and widen for 18 months, after which circumference magically shrinks to zero and beyond. The simple model has a more plausible interpretation: Tree circumference gradually increases over time.

This example is a bit contrived, as a 6^{th} degree polynomial can hardly be justified from a theoretical point of view. However, even when serious thought has been put into which variables to include, overfitting can still occur when the sample size is small compared to the number of parameters in the model.

Some introductory statistics books provide a heuristic[18] minimum sample size to avoid overfitting (e.g. ≥ 5 observations per explanatory variable). However, this largely depends on the problem you are trying to solve. The only way to verify that a model neither underfits nor overfits is through validation.[19] This is part of prediction and will be covered in ***Advanced***, along with methods to include more variables than the sample size would ordinarily allow.

[18] A heuristic is any method that 'seems to work', rather than having a sound logical or empirical basis.

[19] Assessing the performance of the model on an independent data set.

- **Overfitting can be avoided by limiting the number of variables.**

Overfitting in Very Large Models (*)

In deep learning,[20] overfitting is sometimes described as the model memorizing your data. Imagine memorizing the answers to a multiple choice exam (e.g. A, C, C, B, ...) versus learning the subject matter—any new questions cannot be answered by memorizing old answers and will contain many errors.

When the number of parameters grows very large, it becomes exceedingly difficult to prevent overfitting. For example, an artificial neural network can have as many as thousands or millions of parameters. Even with an equally large number of observations, such models almost always have some form of regularization to limit the total size of the parameter estimates.

[20] Deep learning uses neural networks with multiple hidden layers.

2.6.5 *Underfitting*

Figure 2.37 illustrates how adding more terms will *always* result in a better fit, but not always in a better model with regard to generalization to the population. The opposite however, can also occur — a model that uses too few parameters oversimplifies the process. This is called the bias-variance tradeoff (fig. 2.38), where the variance is what we wish to explain and the bias the peculiarities of the specific sample we drew from the population. How to stay close to the optimum model complexity is addressed in the following paragraph.

- **An ideal model neither underfits nor overfits.**

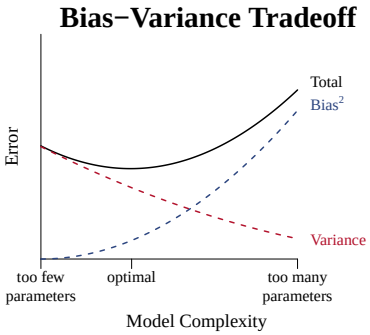


Figure 2.38: An oversimplified model (left) results in underfitting: The remaining variance in the response variable is still high. On the other extreme, an overparameterized mode (right) overfits the data, resulting in high bias. An optimal model has the lowest possible total error.

2.7 *Model Selection*

Sections ***2.6.4*** discussed reasons to generally limit the number of explanatory variables. However, choosing the right model may also imply including more variables than are of primary interest. This paragraph discusses the most important considerations when deciding on the right model.

Model selection is a difficult part of analysis, because it lies at the crossroads of the subject area (e.g. biology) and statistics. Simultaneously taking into account considerations of both fields means there is no single method to determine the best model. If the sample size is sufficiently large, evaluating the predictive performance of the model is usually the best method. However, comparing too many models can still lead to spurious findings.

- **This paragraph concerns model selection for causal inference. Model selection in prediction is covered in *Advanced*.**

2.7.1 *Omitted-Variable Bias & Confounding*

When multiple explanatory variables each have a substantial effect on the outcome, all should be included in the model. Omitting any explanatory variables which substantial affect the outcome will result in an overestimation of the remaining effects (fig. 2.39).

Confounding is a special case of omitted-variable bias, which occurs when an apparent relationship between the explanatory and response variable is actually caused by a third, unobserved variable that affects both. When a confounder is omitted, one might conclude a causal relationship where there is none (fig. 2.40). Confounding is closely related to the principle that correlation does not mean causation.

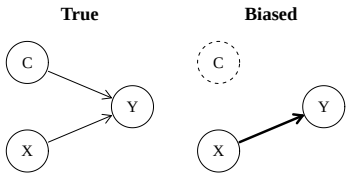


Figure 2.39: Omitted-variable bias can lead to overestimation of remaining effects.

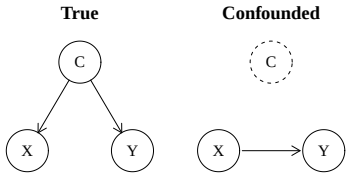


Figure 2.40: If an unobserved variable C affects two observed variables X and Y, this yields the impression that there exists a relation between X and Y, even when there is none.

For example, the more firemen are sent to a fire, the more damage is done. While this is a true, strong correlation, it does not imply that firemen cause damage. The confounding variable is the severity of the fire: The greater the fire, the more damage is done *and* the more firemen are sent to it.

If one or more confounders need to be included in the model that would result in overfitting, consider collecting more observations. If this is not possible, regularized models can help to some extent.

- **Confounders should be included in the model, to remove their bias on the effect of other explanatory variables.**

2.7.2 *Mediation*

Suppose you are interested in the effect of fossil fuel usage on the global temperature. This is not a direct relationship, because fossil fuel usage increases CO_2 levels, which in turn affects the global temperature. In this example, CO_2 is a mediator. If your interest lies in the (indirect) contribution of fossil fuel usage on the global temperature, then CO_2 should *not* be included in the model. If you did, is would appear as though the effect of fossil fuel is not there (fig. 2.41), because the *direct* effect is caused by CO_2. A regression model with global temperature as the response variable would show a relatively large coefficient for CO_2, and an almost zero coefficient for fossil fuel usage.

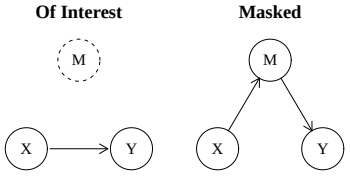


Figure 2.41: Including a mediator (M) in the model can conceal the effect of primary interest ($X \rightarrow Y$).

- **Mediators should not be included in the model, since they can mask the effect of interest.**

If the goal of the research is to *demonstrate* the mediating effect, then a mediation analysis can be performed. Although there are packages and R implementations aplenty, these are advanced techniques, which require delicate interpretation. Moreover, the validity of mediation analysis hinges on the study design. If you suspect your research question involves mediation, consult a statistician, preferably before any data collection takes place.

2.7.3 *Occam's Razor*

A model that captures the systematic part of a process as best as possible using as few explanatory variables as possible is said to be *parsimonious*. Occam's Razor, or the law of parsimony, is a guide for deciding on the best model, by stating that the best model is the model with the greatest parsimony. Many popular model selection criteria (e.g. AIC) aim to maximize parsimony.

Parsimony can be a useful heuristic: Including more explanatory variables will *always* increase the variance explained,[21] but only if the inclusion leads to a large enough increase in explanatory power, will the model be more parsimonious.

However, parsimony is not always justifiable, or even desirable. A model chosen by maximizing parsimony will be biased towards significance and may miss out on predictive power. Renowned contemporary statistician Prof. F.E. Harrell has even said, *"Parsimony is your enemy."* referring to the overuse of variable selection for the sake of parsimony alone.[22] If the research objective is confirmatory or predictive, a parsimonious model is rarely justifiable. If the goal is to learn which variables have the largest contribution to a process, then Occam's razor can be a useful guide.

[21] Recall from section ***2.6.4*** that this is not necessarily a good thing.

[22] `https://stats.stackexchange.com/a/17572/176202`

2.7.4 *Model Selection Criteria*

After considerations of confounding, mediation, and potential collinearity issues have been made, there are various methods to help decide between (a small set of) candidate models, including:

- R^2_{adj}: Variance explained penalized for relative model complexity;
- AIC: Log-likelihood penalized for the model complexity;
- BIC: Log-likelihood penalized for the relative model complexity;
- PRESS (*): An estimate of the out-of-sample error.

The first three methods should only be used if the interest lies in the model itself. Each to a different extent, they maximize parsimony. The last method is a predictive measure. In ***Advanced Statistics***, regularization is introduced, which outperforms variable selection for most purposes. Variable selection should not be used at all if you want to perform hypothesis tests, as the resulting p-values will be optimistically biased.

Another important caveat is that you cannot compare models with different reponse variables, since the extent to which a model can explain one response has no relation to the extent it can explain another. This also means that you cannot compare a model with a transformed response variable to a model with the original response variable.

Adjusted R-Squared (R^2_{adj})

R^2_{adj} is a correction to the variance explained (R^2) for the number of obervations (n) and the number of parameters in the model (p):

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{2.6}$$

The output of `summary()` includes R^2_{adj} if the input is a linear model.

This allows for a comparison of variance explained between models, while taking into account the model complexity relative to the sample size. Without this correction, R^2 is always be highest in the most complex model, which is not parsimonious and may be overfitting.

R^2_{adj} has no meaningful interpretation: Use it only to choose between models based on their explained variance. If you wish to express the variance explained by the final model, report R^2 instead.

A disadvantage of both R^2 and R^2_{adj} is that they are more often misleading than useful:

- Both can be very high with severely violated assumptions;[23]
- Both will be low if the data-generating process simply has large variance, even if the model perfectly captures the systematic part.

[23] Try running the following code:

```
x <- runif(30, 0, 10)
y <- x^2 + rnorm(30)
LM <- lm(y ~ x)
summary(LM)
```

This model clearly violates linearity, yet $R^2 > 90\%$.

Despite their shortcomings, R^2 and R^2_{adj} are still frequently reported in scientific literature and are also part of the standard output of a linear model in R.

Akaike's Information Criterion (AIC)

AIC uses not the variance explained, but the likelihood as a measure to optimize. Specifically, it is a bias correction for likelihood, penalizing for the number of variables:

$$\text{AIC} = 2p - 2\log(\mathcal{L}) \tag{2.7}$$

The likelihood ($\mathcal{L}$) is beyond the scope of this book. R has the function `AIC()`. The input should be a (linear) model.

This measure penalizes more strongly for model complexity than R^2_{adj}, especially when the sample size is large, since it does not take n into account. This tends to favor more parsimonious models than other measures. The smaller AIC, the better the model.

An advantage of AIC over R^2_{adj} is that it is still a valid measure for non-normal linear regression (next chapter).

Bayesian Information Criterion (BIC)

An alternative correction to the log-likelihood, BIC is more lenient towards complex model as the sample size grows larger.

$$\text{BIC} = p\log(n) - 2\log(\mathcal{L}) \tag{2.8}$$

R has the function `BIC()`. The input should be a (linear) model.

This correction makes more sense when the sample size is sufficiently large, as the model can have more parameters without overfitting.

PRESS (*)

Predicted residual error sum of squares (PRESS) is an estimate of the the out-of-sample error by repeatedly fitting the model on $n-1$ observations and calculating the residual error of the remaining observation. After cycling through all n observations, the PRESS statistic can be calculated as:

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{-i})^2 \tag{2.9}$$

See the next paragraph for an Rimplementation.

Where y_i is the observations left out and $\hat{y}_{-i}$ is the predicted value of y_i using a model fitted on all observations except y_i.

PRESS is a form of leave-one-out cross-validation, formally introduced in ***Advanced Statistics***. Cross-validation is a way to estimate how the model would perform on new, unseen observations.

Brief Explanation of Cross-Validation

(...) to be included.

2.8 Examples in R

These lengthy examples illustrate complete analyses. You can run code yourself by copying it to the console in RStudio.

2.8.1 Simple Linear Regression

A simple example was already covered in ***2.1.1***. This example is more elaborate, including explanation you might add to a scientific report.

Running the Model

In this example, we regress tree diameter (y) on height (x).

```
data(trees) # A standard data set in R, see ?trees
simple_linear_model <- lm(Girth ~ Height, trees)
```

Example theoretical justification:

> The diameter of a tree is roughly proportional to its height. Deviations are caused by a myriad of small contributing factors, including various soil nutrients, rainfall, exposure to sunlight and temperature. The sum of all these effects may well be approximately normally distributed.

Visual Diagnostics

You could simply run `plot(simple_linear_model)`, but the script below uses the more elaborate QQ-plot from the package `car`.

```
library("car") # If missing: install.packages("car")
par(mfrow = c(2, 2)) # Make a 2 by 2 raster of plots
plot(simple_linear_model, which = 1)
qqPlot(residuals(simple_linear_model),
       main = expression("Normal Q-Q"))
plot(simple_linear_model, which = c(3, 5))
par(mfrow = c(1, 1)) # Reset to default
```

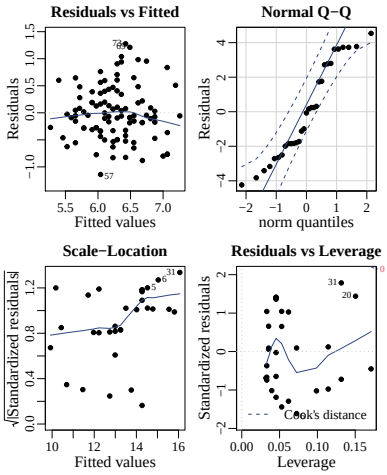


Figure 2.42: Visual diagnostics from the model. No serious deviations are observed.

Summary

With the model choice justified and no problems revealed by the visual diagnostics, we can summarize the findings of this model.

```
summary(simple_linear_model)
```

```
Call:
lm(formula = Girth ~ Height, data = trees)
```

The first part of the summary just shows the model that was fit. This can occasionally be useful to see what the output belongs to.

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.2386 -1.9205 -0.0714  2.7450  4.5384
```

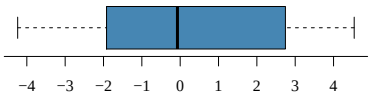


Figure 2.43: Boxplot of the residuals.

The next part is a numerical summary of the residuals. You could construct a boxplot with these values (fig. 2.43), to give an indication of approximate normality. The median should be close to the mean (0) and the distribution should be symmetric (`1Q` $\approx$ `3Q`). However, the QQ-plot (fig. 2.42) is a much better diagnostic for normality.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.18839    5.96020  -1.038  0.30772
Height       0.25575    0.07816   3.272  0.00276 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This part shows the estimated coefficients ($\hat{\beta}_0$, $\hat{\beta}_1$), their standard errors, t-values and p-values. The cryptic legend below it shows whether the p-values exceed some arbitrary thresholds.

If you dislike clutter, you can run: `options(show.signif.stars = FALSE)` before the summary. This will disable the silly stars until you restart R.

Before drawing any conclusions, some final considerations have to be made using the last part of the summary:

```
Residual standard error: 2.728 on 29 degrees of freedom
Multiple R-squared:  0.2697,    Adjusted R-squared:  0.2445
F-statistic: 10.71 on 1 and 29 DF,  p-value: 0.002758
```

1. The first part is the estimated standard deviation (σ) of the error term (ϵ).[24] It gives an indication of the remaining spread in the response variable, after accounting for the effect of explanatory variable. If it is small compared to the original standard deviation of the response variable, then the model explains a lot of the variance in the response.
2. The degrees of freedom are $n - p = 31 - 2 = 29$, because this model has two parameters (β_0, β_1). We do not need this information for simple linear regression.
3. The model explains $R^2 = 1 - \frac{\text{var(residuals)}}{\text{var(response variable)}} \approx 27.0\%$ of the variance in the reponse variable. This is not very much; around three-quarters remains unexplained. R^{adj} is meaningless in simple linear regression.
4. An F-test for the difference in variance from (3) has a p-value of $p \approx 0.0028$. If you set $\alpha = 0.05$, that means the model explains significantly more variance than the null-model (no explanatory variables). Remember that this does not necessarily mean that *a lot* of variance is explained, just that it is significantly non-zero. Also note how for simple linear regression, the p-value is the same as that for the slope... do you understand why?

[24] Unfortunately, the language used is imprecise: It says standard *error*, but this is an estimate of the residual standard *deviation*. The residuals are assumed to follow a normal distribution. If they do, this value is an estimate of the standard deviation of that normal distribution (σ in $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Confidence Intervals

It is good practice to add a measure of uncertainty to your estimates. A simple way to obtain such a measure is as follows:

```
confint(simple_linear_model)
```

```
                  2.5 %    97.5 %
(Intercept) -18.37837106 6.0015820
Height        0.09589546 0.4155988
```

Visualizing the Model

For simple linear regression, you can summarize the entire model in a single plot. 95% confidence intervals and 95% prediction intervals around the regression line have been added.

```
plot(Girth ~ Height, data = trees)
x_values <- seq(min(trees$Height), max(trees$Height), 0.1)
newdata  <- data.frame(Height = x_values)

# Add confidence intervals
y_conf <- predict(simple_linear_model, newdata)
lines(y_predict[, 1] ~ x_values)          # Regression
lines(y_predict[, 2] ~ x_values, lty = 2) # Lower bound
lines(y_predict[, 3] ~ x_values, lty = 2) # Upper bound

# Add prediction intervals
y_predict <- predict(simple_linear_model, newdata,
                     interval = "predict")
lines(y_predict[, 1] ~ x_values)          # Regression
lines(y_predict[, 2] ~ x_values, lty = 2) # Lower bound
lines(y_predict[, 3] ~ x_values, lty = 2) # Upper bound
```

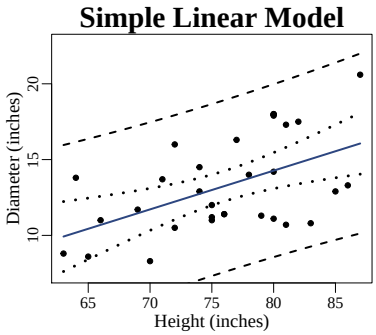


Figure 2.44: Simple linear model for the relationship between tree height and diameter (—), with 95% confidence bands (· · ·) and a 95% prediction interval (- - -).

Example conclusion:

> On average, an inch increase in tree height corresponds to an increase of around 0.26 inch in diameter (95% CI: 0.096–0.42, $p \approx 0.0028$). A small but significant amount of the variance in tree diameter is explained by tree height (27%).

A figure like 2.44 would make a nice addition to the report. You could also include the visual diagnostics as supplementary material.

2.8.2 *Multiple Linear Regression*

(...) to be included. See the GLM examples for now, which also deal with multiple linear regression.

3 Generalized Linear Regression

*Ordinary linear models are quite limited by the assumptions under which inference is valid (paragraph **2.2**). A generalized linear model is less restrictive, by relaxing one of the assumptions (normality of the residuals).*

In the linear models discussed so far, given the explanatory effects, the remainder of the response should be normally distributed noise. The normal distribution is symmetric, continuous, and goes all the way from $-\infty$ to $+\infty$. In real life, almost nothing adheres to this.

While certain response variables have a stochastic part that can be reasonably approximated by a normal distribution (e.g. BMI, blood pressure, IQ), others usually do not (fig. 3.1). For example, consider a study where the primary outcome is whether a patient has a disease (1) or not (0). Since there are only two possible values, the conditional distribution will never resemble a normal distribution.

Rather than assuming the error term to be normally distributed, a generalized linear model (GLM) can have errors with any probability distribution in the exponential dispersion family.

The following paragraphs introduce the exponential dispersion family (**3.1**), the generalized linear model (**3.2**), visual diagnostics for GLMs (**3.3**), model selection for GLMs (**3.4**) and an algorithm for fitting GLMs (**3.5**).

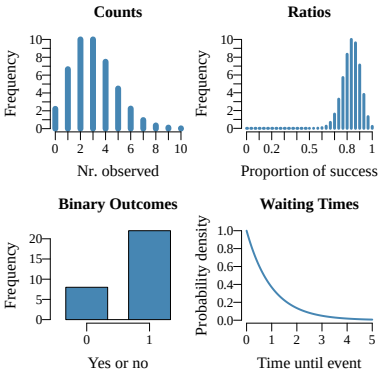


Figure 3.1: Examples of non-normal processes that a GLM can model.

3.1 The Exponential Dispersion Family

The normal distribution is a member of the exponential dispersion family. What does this mean? There are many commonly occurring probability distributions that can all be rewritten as a member of a larger class of distributions. By extending the linear model to encompass any distribution within this class, a GLM allows us to model many more different kinds of response variables.

3.1.1 Poisson distribution

Named after the mathematician who first described it, the Poisson distribution is **the distribution of independent counts**. This means that if you were to count the number of independent random events occurring within a fixed amount of time, then the chance of counting any particular number of events is Poisson distributed. An example is the number of earthquakes in a year, given the yearly average. For a mean number of events λ, the **probability density function** is:

$$P(k \text{ events} \mid \text{average of } \lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{3.1}$$

Figure 3.2 shows three examples of Poisson distributions: When the rate is high enough, the Poisson distribution somewhat resembles a normal distribution (except that it is still discrete). When the rate is low, the distribution is right skewed (assymetric). The reason is simple: Something cannot be counted less than zero times, so as λ tends to 0, all probability density is squeezed into the left corner.

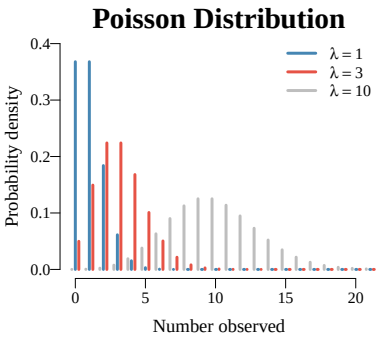


Figure 3.2: The Poisson distribution at different rates. The Poisson distribution has only one parameter (λ), which is both the mean and variance.

3.1.2 Binomial Distribution

The chance of k **successes out of** n **trials**, with chance of success p, follows a binomial distribution (fig. 3.3). More generally, any kind of ratio $A : B$ is binomially distributed. Think for example of the number of times you roll ⚅ out of 12 rolls with a 6-sided die. If the die is fair $p = \frac{1}{6}$, and you would expect $k = n \cdot p = 2$. But what is the chance of a different number of successes? This can be calculated with the **probability density function**:

$$P(k \text{ successes} \mid n \text{ trials, chance of } p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.2}$$

Ratios and Hidden Ratios

If 1 in 10 people are ill, the ratio diseased : healthy $= 1 : 9$. However, sometimes data is presented as counts (e.g. *number* of deaths), while in reality, they are ratios (e.g. number of deaths *per 1000 patients*). This is called a hidden ratio and should be considered binomial. Likewise, percentages and proportions can also represent ratios.

Categorical Response

A categorical response variable with two categories (e.g. yes/no, case/control) follows a binomial distribution with $n = 1$. If there are more than two categories, the multinomial distribution can be used, which is beyond the scope of this book.

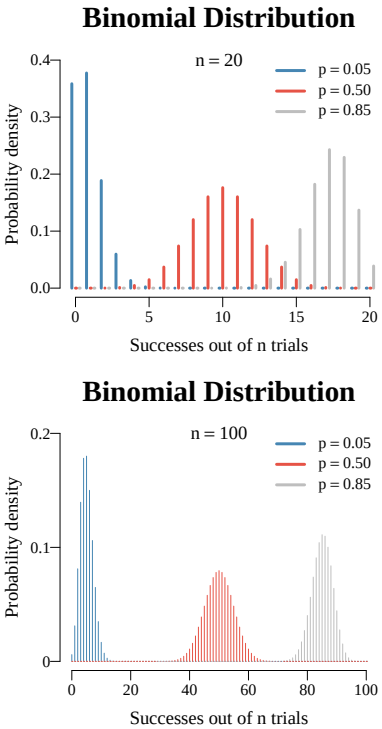


Figure 3.3: The binomial distribution at different numbers of trials (n) and probabilities of success (p). The mean is $n \cdot p$ and the variance $n \cdot p(1-p)$.

3.1.3 Exponential Distribution (*)

The time (t) between random independent events occurring at a rate of (λ) is exponentially distributed. If a Poisson distributed process has on average λ events in a given unit of time, then the average time between two events is $\frac{1}{\lambda}$. The **probability density function** is:

$$p(\text{time of } t \mid \text{rate of } \lambda) \;=\; \lambda e^{-\lambda t} \tag{3.3}$$

Time is continuous—it can be divided endlessly into smaller parts (minutes, seconds, microseconds, ...). Therefore, the chance of any exact amount of time t is zero. For continuous distributions, probability *density* (p) is simply the height of the curve at a particular value. To calculate the *probability* (P), a range of t must be chosen. For example: What is the chance of waiting at least t? This area under the curve can be calculated with the **cumulative density function**:

$$P(T \leq t \mid \lambda) \;=\; 1 - e^{-\lambda t} \tag{3.4}$$

Imagine trains arrive randomly at an average interval of 15 minutes. By plugging $\lambda = \frac{1}{15}$ into the formula above, the chance of a train arriving *within* 10 minutes would be $1 - e^{-\frac{10}{15}} \approx 48.7\%$. For the probability of *more* than 10 minutes, recall the complement from paragraph **1.2**, and calculate: $P(T > t) = 1 - P(T \leq t)$. In general, the chance of the train arriving within an interval $a \leq t \leq b$ can be calculated by subtracting $P(T \leq a)$ from $P(T \leq b)$.

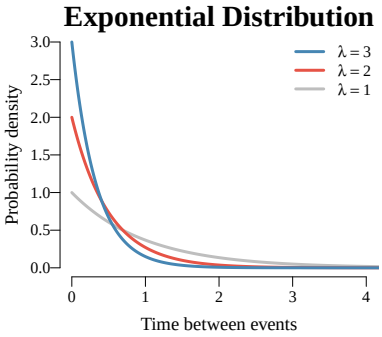


Figure 3.4: The exponential distribution at different rates. The mean is $\frac{1}{\lambda}$ and the variance $\frac{1}{\lambda^2}$.

3.1.4 Negative Binomial Distribution

Where the binomial distribution describes the chance of k successes out of n trials (a ratio), the *negative* binomial distribution describes the chance of r failures before k successes have occurred (a count). Its **probability density function** is:

$$P(r \mid k,\, p) \;=\; \binom{k+r-1}{k} p^k (1-p)^r \tag{3.5}$$

Like the Poisson distribution, this is a **distribution of counts**. Unlike the Poisson distribution, the negative binomial has a separate parameter for the variance. This is useful when the counts are not entirely independent, e.g. due to some unobserved variable. The Poisson distribution must have a variance equal to the mean. If there is an extra source of uncertainty, the negative binomial can account for the resulting excess variance (paragraph **3.2.5**).

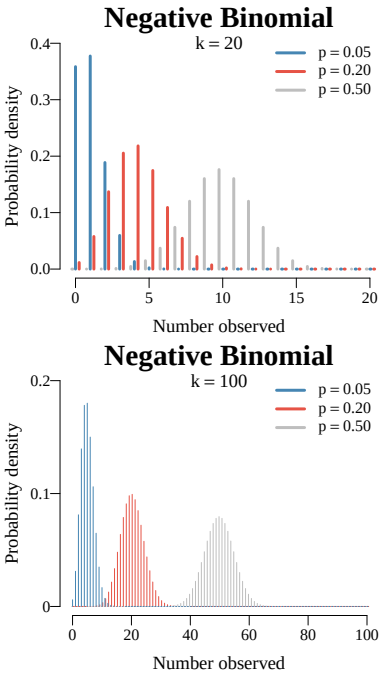


Figure 3.5: Negative binomial distributions at different rates and variances. The mean is $\frac{rp}{1-p}$ and variance $\frac{rp}{(1-p)^2}$

3.1.5 *Normal Distribution (*)*

To contrast against the other distributions in the exponential dispersion family, here is a summary of the normal distribution (fig. 3.6).

From a practical point of view, the normal distribution is used for **processes that are continuous, with a central tendency**. Using a QQ-plot, the residuals can then be compared to a theoretical normal distribution, to see if approximate normality holds. For a mean (μ) and variance (σ^2), the **probability density function** is:

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \tag{3.6}$$

Since the normal distribution is continuous, the lowercase p means probability density—the height of the function at a given value of x. To calculate probabilities, the **cumulative density function** should be used (fig. 3.7). This is the integral of eq. 3.6. It has no simple form, but can be well-approximated in R with `pnorm()`. For example, the chance that a value is between -1 and 1 with $\mu = 0$ and $\sigma^2 = 1$ is:

```
pnorm(1, mean = 0, sd = 1) - pnorm(-1, mean = 0, sd = 1)
```

```
[1] 0.6826895
```

The normal distribution is the distribution of sample means: If you were to draw multiple samples from the same population, each will have a (slightly) different mean. As the number of samples goes to infinity, the sample means will follow a normal distribution.

That might sound a little convoluted, but observe that the mean is just the sum divided by the number of observations. Because of this, a large number of independent effects (e.g. genetic, environmental) will, when summed together, also tend to a normal distribution. This phenomenon is formalized in the central limit theorem (CLT), and it occurs regardless of what the distributions of the individual effects are. Hence why the normal distribution is so commonly used.

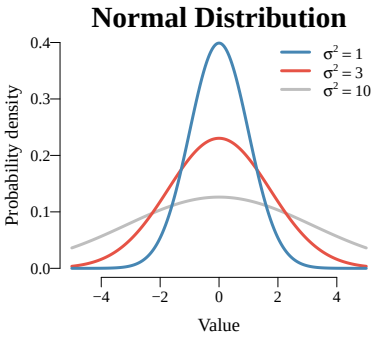


Figure 3.6: The normal distribution with different variances. For the residual term in linear regression, $\mu = 0$.

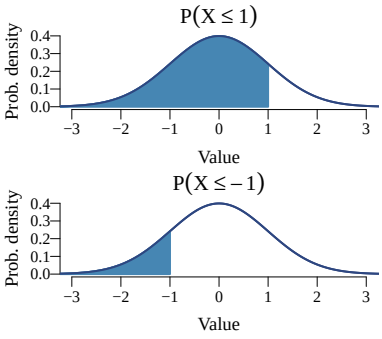


Figure 3.7: What is calculated by `pnorm(1)` and `pnorm(-1)`.

When Does Something Tend to Normal?

Since most things are affected by a sum of many independent effects, that begs the question: When is something normally distributed? The CLT provides no guarantees of convergence to a normal distribution unless $n \to \infty$. The Berry–Esseen theorem gives an indication of what the rate of convergence might be, under stronger assumptions.

However, this is both beyond the scope of this book, and rarely something you should be concerned about during research. Instead, try to identify the most theoretically appropriate distribution for the process you are modelling, and then check for violations from the proposed error structure using visual diagnostics.

3.1.6 *Gamma Distribution (*)*

(. . .) to be included.

3.1.7 *Tweedie Distribution (*)*

(. . .) to be included. (package `statmod`)

3.2 Generalized Linear Model

A GLM estimates a linear relationship between the parameter(s) of the chosen error structure (e.g. normal, Poisson, binomial), and the explanatory variable(s). In order to translate estimates on the scale of the linear equation to that of the response variable, a link function is required. In summary, a GLM has three components:

1. An **error structure** (the conditional distribution the response);
2. A **linear equation** (for the parameters of the error structure);
3. A **link function** (translating the linear equation to the response).

3.2.1 Error structure

The error structure is the probability distribution of the response, conditional on the explanatory variable(s). In an ordinary linear model, the error structure is assumed to be a normal distribution. In a GLM, it can be any distribution in the exponential dispersion family (paragraph ***3.1***). This is the reason to use a GLM and the quintessential difference with a normal linear model.

Common types of data modelled with GLMs include: counts, ratios, binary outcomes and waiting times.

3.2.2 Linear equation

The linear equation[1] is the relationship to be estimated. It is identical to the systematic part of a linear model and most concepts from chapter **2** apply to it as well (e.g. multiple regression, interaction).

[1] Also referred to as: linear predictor, or linear part.

Rather than estimating a linear relationship with the response (y) directly, a GLM estimates the parameter(s) of the error structure, given the link function. Hence, fitted values are not called $\hat{y}$, but are given a new name: η. Simple linear regression with a GLM is then:

$$\eta = \beta_0 + \beta_1 x \tag{3.7}$$

And multiple linear regression in a GLM is written as:

$$\eta = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \tag{3.8}$$

Note that there are no residuals in the linear equation, since these are already accounted for by the error structure.

Estimation of the linear equation of a GLM does not work by OLS (paragraph **2.5**), because that would require the error structure to be a normal distribution. Instead, GLMs are fitted by iteratively reweighted least squares (paragraph **3.3**).

3.2.3 *Link function*

The link function maps the linear equation to the error structure. If this mapping translates the linear equation to the mean of the response, it is called a **canonical link function**. Below is a summary of canonical link functions for the exponential dispersion family.

Response	Error structure	Canonical link	R syntax
Continuous	Normal	μ	`"identity"`
Binary	Bernoulli†	$\log\left(\frac{p}{1-p}\right)$	`"logit"`
Ratio	Binomial		
Count	Poisson	$\log(\mu)$	`"log"`
	Negative binomial		
Time to event	Exponential‡	$\frac{1}{\mu}$	`"inverse"`
	Gamma		

Table 3.1: Common error structures and their canonical link functions. †: The Bernoulli distribution is a binomial distribution with one trial. ‡: The exponential distribution is a Gamma with shape equal to 1.

Another way to think of the purpose of the link function is to 'stretch out' the error structure to better reflect the importance on the scale of the linear equation. Consider the following examples:

- The logarithm for count data (fig. 3.8). The link function stretches out the lower counts and squeezes the higher counts. In effect, this means that a difference between counting 0 or 1 has a greater effect on the scale of the linear equation, than say, a difference between 20 and 21.
- The logit for a binary response (fig. 3.9). The link function stretches probabilities close to 0 and close to 1 and squeezes probabilities close to 0.5. In effect, this means that a difference between 1% and 2% (a doubling), corresponds to a larger difference on the scale of the linear equation than, say, a difference between 50% and 51%. The same goes for probabilities close to 100%.

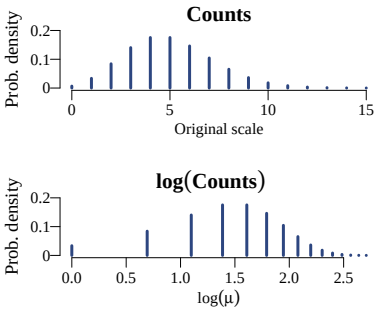


Figure 3.8: The logarithm stretches the left part of the distribution and squeezes the right.

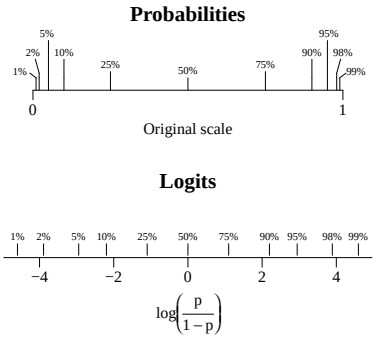


Figure 3.9: Logit stretches the extremes of the distribution. On the original scale, labels near the extremes barely fit in the figure. When translated by the link function, values near the extremes are further away from one another.

Alternative Link Functions

Of course, many functions can appropriately 'stretch' the support of the probability distribution. Recall from paragraph ***2.4.2***, figure 2.28, that the effect of the logarithm is similar to that of the square-root. For logit, alternatives include probit and the complementary-log-log.

In general, you will want to start with the canonical link function. Occasionally, choosing a different link function can improve the fit. You can view the standard implemented link functions in the `glm()` function under `link` section of the help page (`?glm`).

3.2.4 GLM in R

This example demonstrates the concepts explained so far in R.

1. A data set you can load without installing any new packages in R is `eagles`. In this study, 160 attempts by Bald Eagles to steal salmon from another Bald Eagle were recorded. The goal of the experiment was to determine how size and age affected the successfulness of stealing attempts. The following code loads the data into the environment prints it to the console.

Also see the help file:
`?MASS::eagles`

Below is a summary of abbreviations.
`y`: succesful piracy attempts
`n`: total piracy attempts
`P`: pirate size (`S`: small, `L`: large)
`A`: pirate age (`I`: immature, `A`: adult)
`V`: victim size (`S`: small, `L`: large)

```
data(eagles, package = "MASS")
print(eagles)
```

```
   y  n P A V
1 17 24 L A L
2 29 29 L A S
3 17 27 L I L
4 20 20 L I S
5  1 12 S A L
6 15 16 S A S
7  0 28 S I L
8  1  4 S I S
```

2. For the response, we have the number of successful attempts (`y`) and the total number of attempts (`n`). This is a ratio, which we can model with a binomial GLM. To do so in R, we will turn the response variable into a matrix of two columns, successes and failures, using column bind (`cbind`).

```
successes    <- eagles$y
failures     <- eagles$n - eagles$y
piracy_ratio <- cbind(successes, failures)
```

3. For the explanatory variables, we have the size of the pirating bird (`P`), the age of the pirating bird (`A`) and the size of the victim (`V`). First, let's set the reference group to be a small, young pirating bird stealing from a small victim bird. This is not necessary, but gives the model the easiest interpretation in my opinion.

To view the order of categories, use:
`levels()`

```
eagles$P <- relevel(eagles$P, "S") # Set reference: small
eagles$A <- relevel(eagles$A, "I") # Set reference: immature
eagles$V <- relevel(eagles$V, "S") # Set reference: small
```

4. From a theoretical point of view, it would make sense that larger eagles are more succesful than smaller eagles at stealing food. It would also make sense that an immature eagle is less experienced and therefore worse at stealing food.

```
GLM <- glm(formula = cbind(y, n - y) ~ P + A + V,
           family  = binomial(link = logit),
           data    = eagles)
```

```
# library(mctest)
# X <- model.matrix(~ P + A + V, data = eagles)
# omcdiag(X, car::logit(eagles$y/(eagles$n - eagles$y)))
#eagles$Laplace_smoothed_logit <- car::logit((eagles$y + 1) / (eagles$n - eagles$y + 1))
#boxplot(eagles$Laplace_smoothed_logit ~ eagles$V + eagles$P)
#boxplot(eagles$Laplace_smoothed_logit ~ eagles$A + eagles$P)
#boxplot(eagles$Laplace_smoothed_logit ~ eagles$A + eagles$V)
```

5. Coefficients, summary... scale of response.

```
coefficients(GLM)
```

```
(Intercept)          PL          AA          VL
  0.6204617   4.5569563   1.0972964  -4.9331366
```

```
summary(GLM)
```

```

Call:
glm(formula = cbind(y, n - y) ~ P + A + V, family = binomial(link = logit),
    data = eagles)

Deviance Residuals:
      1       2       3       4       5       6       7       8
-0.9775  0.3303  0.7263  0.4744  0.7009  1.1119 -0.8633 -1.6329

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6205     0.6797   0.913   0.3613
PL            4.5570     1.0545   4.322 1.55e-05 ***
AA            1.0973     0.5465   2.008   0.0446 *
VL           -4.9331     1.1193  -4.407 1.05e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 128.2673  on 7  degrees of freedom
Residual deviance:   6.9563  on 4  degrees of freedom
AIC: 27.696

Number of Fisher Scoring iterations: 5
```

6. The original paper about these data also included an interaction between the age of the eagle and the size of the pirating eagle.[2] Let's fit their model too for later comparison.

[2] Knight, R. L. and Skagen, S. K. (1988) Agonistic asymmetries and the foraging ecology of Bald Eagles. *Ecology* **69**, 1188–1194. DOI: 10.2307/1941273

```
GLM_interaction <- glm(formula = cbind(y, n - y) ~ P * A + V,
                       family  = binomial(link = logit),
                       data    = eagles)
```

7. The age and size of a bird are already quite similar measures, since adult eagles are generally larger than immature eagles. Including an interaction means we are looking at combination of age and size. Though theoretically justifiable, the number of observations is quite limited.

3.2.5 *Over- and Underdispersion*

When the model overfits the data, the residuals can get very small (even zero), as we saw in figure 2.37. In the context of a GLM, this is called **underdispersion** of the residuals, because there is a smaller spread in the data than *can* be modelled with the given distribution (e.g. Poisson or binomial). The opposite can also occur: Too much noise in the data, an extreme observation, or a model that is too simple (due to a variable left out, or a variable unknown entirely). This is known as **overdispersion**. In figures 3.10 and 3.11, examples of over and underdispersion in the Poisson and binomial distribution are shown, respectively.

Note: An outlier can create the illusion of overdispersion, so always perform visual diagnostics.

3.2.6 *Model Selection*

Contrary to most features of the linear predictor, model selection is slightly different. Although we can still select a model based on AIC, there is no R^2_{adj} for a GLM. The reason for this is that if we use an asymmetric distribution, like Poisson or binomial, the 'variance' explained is a poor measure of goodness-of-fit. After all, the Poisson

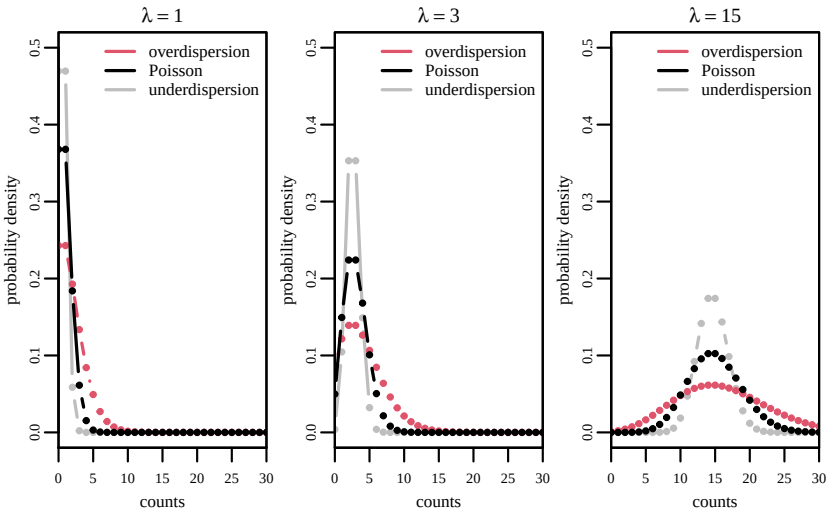


Figure 3.10: Poisson distribution (black) and two distributions with over and underdispersion (red and grey, respectively). Whatever value of λ we choose determines both location and dispersion, so a model that has larger (or smaller) dispersion requires an additional parameter.

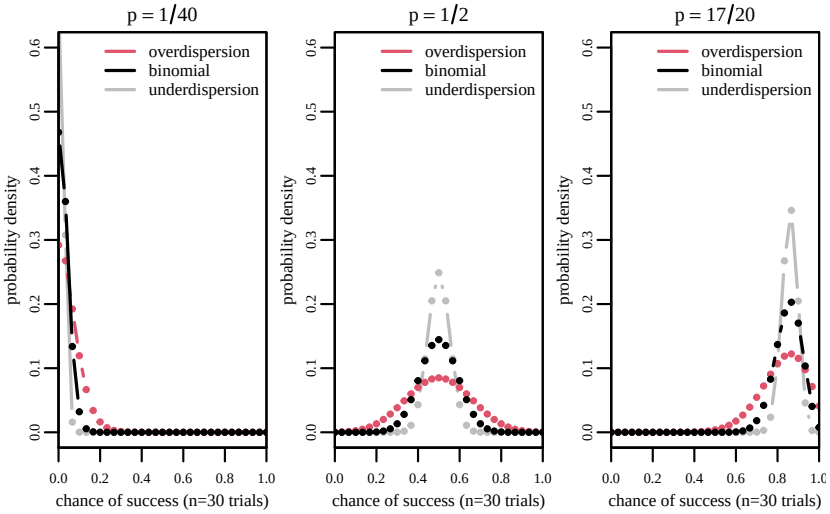


Figure 3.11: Binomial distribution (black) and two distributions with over and underdispersion (red and grey, respectively). For given values of p and n (here: $n = 30$), the dispersion is fixed in the binomial distribution. Hence, a model that has larger (or smaller) dispersion requires an additional parameter (quasi-binomial).

and binomial distribution have different measures of dispersion than the normal distribution. (. . .)

3·3 *Iteratively Reweighted Least Squares (*)*

(. . .) to be included.

3·4 *Beta Regression (*)*

(. . .) to be included.

3·5 *GLM in R*

(. . .) to be included.

4 Principal Component Analysis

Suppose we measure expression of 40 genes and want to 'have a look' at the data. What should we do? Scroll through rows and columns in Excel and look at the individual numbers? Should we produce 40 histograms and attach them to the report? What if instead we measure 1500 metabolites; where do we even begin?

This is a challenging problem and there is no single correct answer. However, arguably the best known and most widely used multivariate technique is principal component analysis (PCA). There are scenarios where it is more informative and scenarios where it is less informative, but the great thing about PCA is that *it is never wrong to do*, the only condition is that the input is numeric data (i.e. not nominal[1] or ordinal[2]) and has no missing values. Just like it is never wrong to visually inspect your data, you should consider PCA to be one of the basic tools for summarizing and interpreting data — especially when the data frame itself is too big to look at. Moreover, advanced multivariate techniques often rely on PCA or some method that works similarly. That being said, PCA is most useful when the input is multivariate normally distributed.[3]

[1] Nominal data is categorical and without order (e.g. male & female or red, blue & green).

[2] Ordinal data is categorical, but with order (e.g. good, better, best).

[3] Multivariate normality implies the distribution of individual variables is normal.

What is PCA?

- Linear, independent combinations of a (possibly correlated) set of variables that contain the highest variance in descending order (PC1, PC2, . . .).

Why Would I Need That?

- Visualizing high-dimensional data
  - Summarizing large data sets can be difficult
  - Reporting only the first few PCs allows you to summarize data of any number of variables

- Dimension reduction
  - When variables are correlated, the choice of which variable to remove in your analysis can be somewhat arbitrary; *combining* variables may then be preferable over *selecting* variables
  - Instead of selecting (i.e. removing) variables, PCA seeks the combination(s) with highest variance
- Dealing with highly correlated data
  - High correlation of explanatory variables causes many techniques to fail or give poor estimates
  - While your original variables might be correlated, PCs are by definition uncorrelated (independent)

Nomenclature

- **Principal component** (PC): A linear combination of the original explanatory variables;
- **Scores**: Location of the original observations on these principal components;
- **Loadings**: Contribution of the original variables to the principal components.

4.1 PCA: Visualization

Imagine you are at the zoo and want to take a picture of the animals in a cage. Some animals are taller, others wider, some more colourful, etc. PCA takes the picture from an angle such that the greatest variability in animals is shown from one side to the other.

Reducing expression levels of 40 genes to a two-dimensional plot may seem a little strange at first. What happened to my data; how do I interpret this with respect to my original 40 genes? In this example we will see what happens in two and three dimensions. From there, you might be able to imagine what it is like going from 4, 5, or any number of variables to a smaller set of principal components.

4.1.1 The 2-Dimensional Case

First we will consider the simplest scenario: PCA on two variables, for example the diameter of a bacterial colony and the expression of a GFP construct. Of course, we don't need PCA when we only have 2 variables, but this example is meant to show how PCA works. In figure 4.1 (left), GFP expression is represented by colour and point size is proportional to the colony diameter. As you can see,

there is a strong negative correlation between these two variables ($\rho = -67.3\%$). It appears that smaller colonies have stronger GFP expression for some reason. This implies that a linear combination of these variables (i.e. a diagonal line through this figure) can have a larger variance than any of the two variables seperately.[4] PCA seeks the *largest* possible variance using any such combination, which turns out to be the blue arrow. This is the first principal component (PC1) and is the blue x-axis in the new space (figure 4.1). Since there are only two variables, the second principal component (PC2) covers the remaining variance. The total variance explained is $83.7\% + 16.7\% = 100\%$ (everything). Keep this in mind for the next paragraph.

[4] This is true for any set of variables with pairwise correlations $\rho \neq 0$.

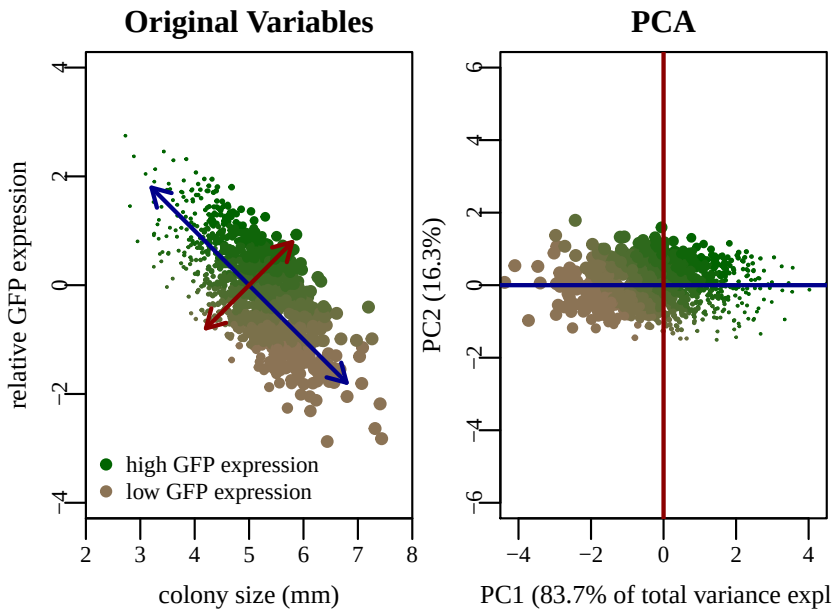


Figure 4.1: In the two-dimensional case (left), PCA finds the linear combination of the two variables with the greatest variance (PC1). Then, a second linear combination is made (PC2), independent of the first with the remaining variance. These PCs then become the axes of the new coordinate system (right). In this case, PCA is identical to rotating the plot.

4.1.2 *The 3-Dimensional Case*

Suppose this GFP construct is related to antibiotic production and we also measured the sizes of halos around the colonies as a reflection of their antibiotic production. Naturally, this should be correlated to GFP expression. However, it also turns out to be (weakly) correlated to the colony size (figure 4.2).

We can also try and visualize all three variables together in a 3D scatterplot (figure 4.3, left), adding a circle around each datapoint to represent its halo. PC1 is again the largest variance we can describe in one line, that can go in any direction through the three-dimensional space. PC2 is the second largest variance, perpendicular

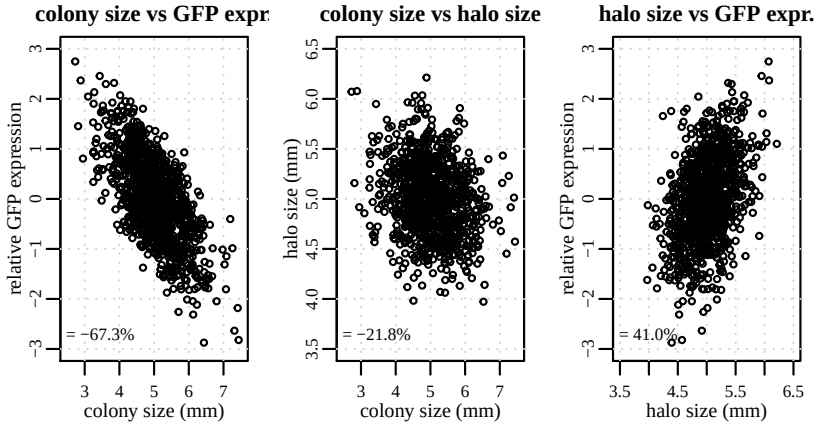


Figure 4.2: All three variables plotted against each other.

to PC1 and since there were originally 3 dimensions, a final line can be construed that is perpendicular to both PC1 and PC2 (PC3).

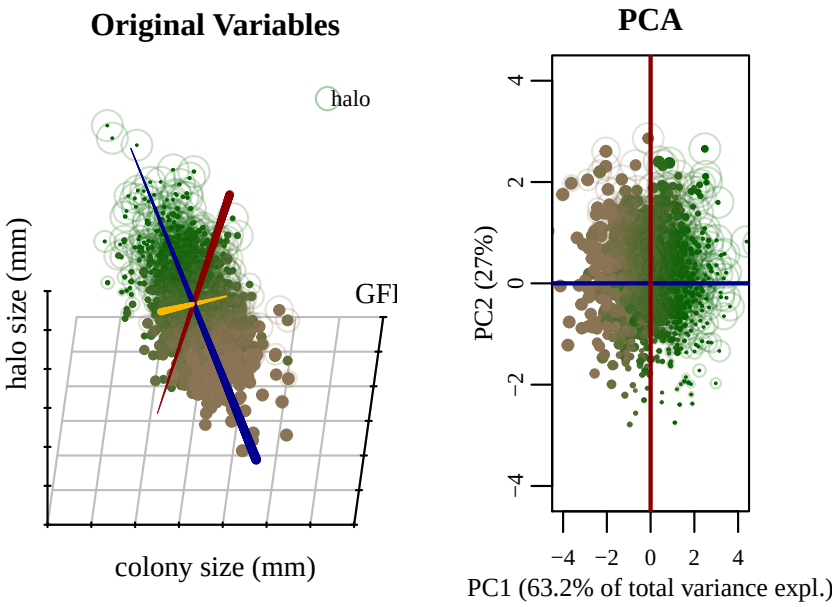


Figure 4.3: Example of PCA on three variables (i.e. three dimensions). The greatest variance, using any combination of the three variables becomes the first principal component (PC1). The greatest remaining variance, independent of (i.e. perpendicular to) PC1 becomes the second principal component (PC2). Since there are three variables, a third principal component (PC3) exists, which is is perpendicular to both PC1 and PC2. The sizes of the linepieces representing the principal components on the have been made proportional to the amount of variance they explain. This example is still identical to rotation.

Now we can demonstrate one use of PCA: Using only two principal components, we have explained $63.2\% + 27.0\% = 90.2\%$ of the total variance (figure 4.3, right). We have reduced three dimensions to two PCs. The majority of the information in the entire data set, can now be visualized in a single 2D figure! Adding a third PC would only increase the variance explained by 9.8%.

4.2 Biplot

In figures 4.1 and 4.3 we plotted the datapoints in the new dimensions: the principal components. The locations of our datapoints in these new dimensions are called the **scores**. We can also indicate the contribution of the original variables to the principal components. These contributions are called the **loadings**. A plot that contains both the scores and loadings of a PCA is called a biplot.

4.2.1 Beyond Three Dimensions

Up till now, we have discussed colony size, GFP expression and halo size in our example. What do we do when another variable is introduced? Continuing from our previous example: suppose we did this experiment on plates with varying amounts of carbon source. We seem to be running out of options for indicating another variable (figure 4.6). However, a biplot offers a fairly simple solution to this problem, as is shown in figure 4.7. If the number of variables becomes too large to display the names, just name your variables $1, 2, 3, ..., p$.

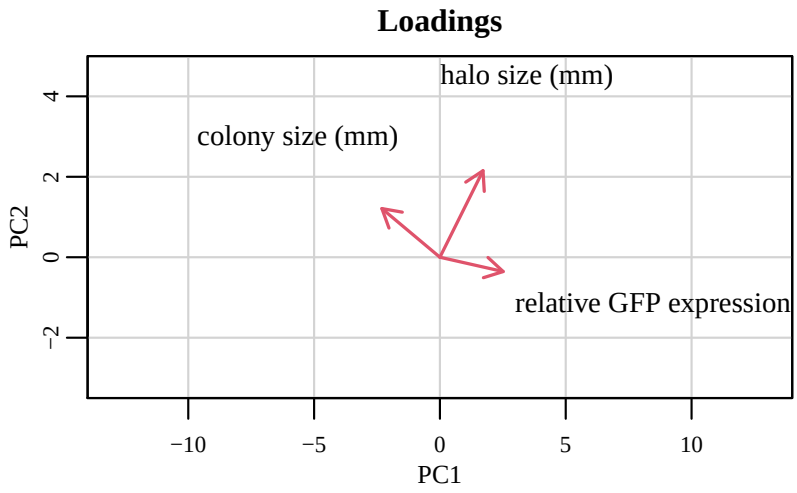
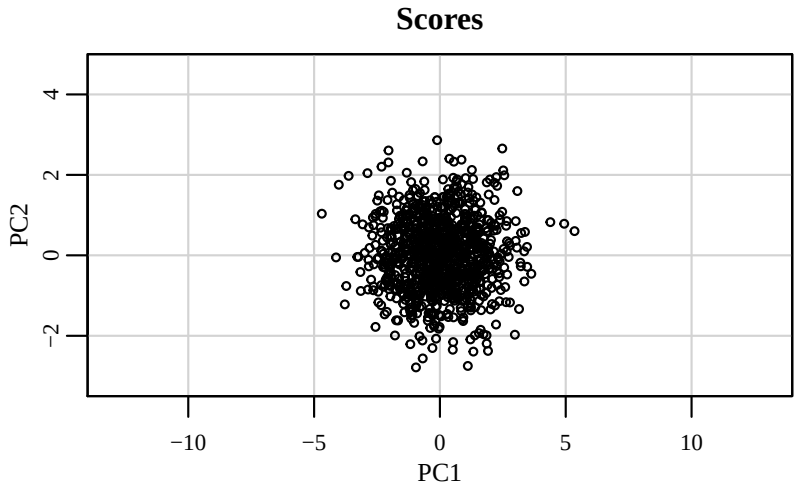


Figure 4.4: On the left, the loadings are shown, which are the contributions of the original variables to the principal components. On the right, the scores (same as what was depicted in figure 4.3) are shown, which are the locations of the observations in the principal components.



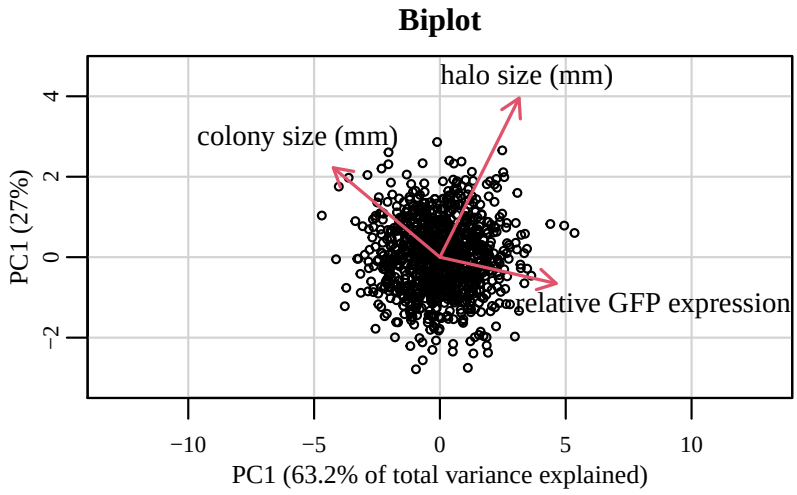


Figure 4.5: A combination of the scores and loadings.

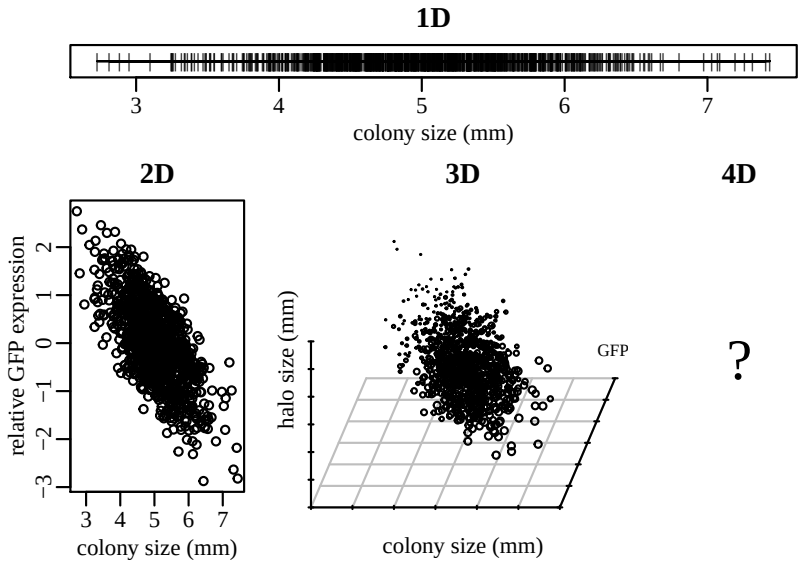


Figure 4.6: Plotting 1 or 2 dimensions on paper is a trivial task. With some effort, a reasonably recognizable 3D plot can be made. Anything beyond 3D has no easily understandable form on paper.

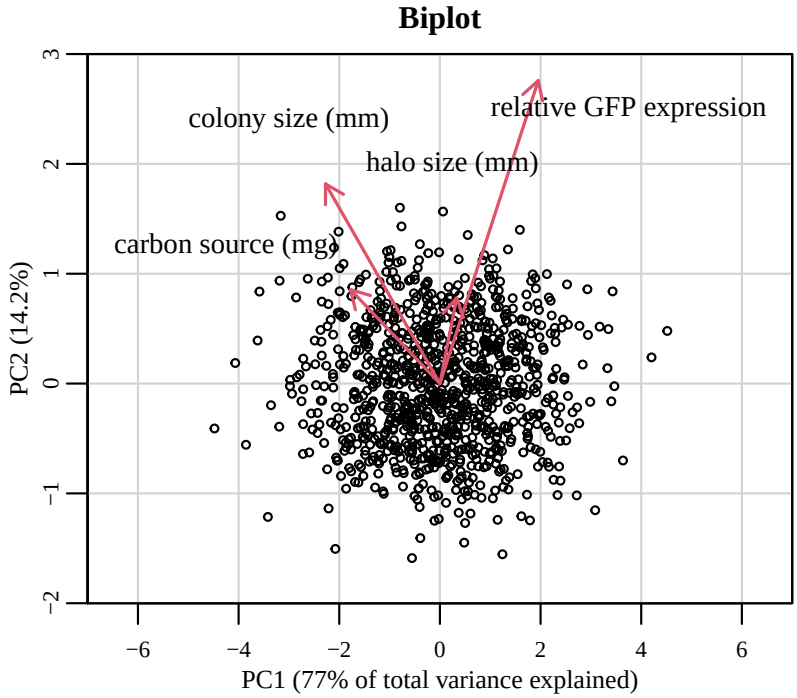


Figure 4.7: Contribution of 4 original dimensions to PC1 and PC2. This plot shows 77.0% + 14.2% = 91.2% of the total variance among four original dimensions in 2D space.

4.3 Choices to be made in PCA

The are several choices to be made when using PCA. Moreover, these choices are important regardless of the type of multivariate method used, be it PCA or some other technique.

4.3.1 Centering

Centering data means to concentrate it around some point (usually the origin). In PCA, centering means to remove the mean of every variable from its observations. You will find lengthy discussions if you Google: centering in PCA, but the bottom line is that mean-centering is almost always desired and hence the default in the R functions `prcomp()` and `princomp()`.[5]

[5] The difference is the type of method used for finding the principal components: SVD on centered $\mathbf{X}$ or EVD on $\mathbf{X}^T\mathbf{X}$.

4.3.2 Scaling

Scaling means to force variables to have similar variances (dispersion around the mean), which can be achieved by dividing variables by their variances. In strong contrast to centering, scaling is a very important choice that can severely affect the results.

When To Scale Variables

- When variables are measured on a different scale (e.g. distances in mm, cm, km)
  - Otherwise, variables with larger values (e.g. mm) will have larger variance
  - This will cause PCA to attribute greater importance to such variables
- When you believe all variables are equally important

When NOT To Scale Variables

- When variables are measured on the same scale (e.g. measuring 40 genes with the same method)
  - Otherwise, variables with smaller variances are artificially increased in importance and visa versa
  - This will cause PCA to attribute equal importance to all variables
- When variables cannot be assumed to be equally important

Examples

- The `iris` data set:
  - Measurements of length and width of *Iris* flowers' sepals and petals
  - All variables are in mm and have essentially the same meaning
  - **No scaling** should be applied
- The `airquality` data set:
  - Measurements of ozone (ppb), solar radiation (lang), wind (mph) and temperature (? F) over time
  - Variables differ by several orders of magnitude and are thus on vastly different scale
  - **Scaling** should be applied
- (Log)-expression levels of 40 genes with qPCR
  - All genes are measured with the same method
  - Genes may vary in (log)-expression level, but these numbers have the same interpretation
  - **No scaling** should be applied

4.3.3 *Number of Principal Components*

So far we have restricted ourselves to the first two principal components (containing the most variance by definition). However, these may not explain a satisfactory amount of total variance. Say we want to explain at least 95% of our total variance in the data. We then simply look at the cumulative proportion of variance explained and choose a number of components such that this exceeds 95%.

If we have no reason to choose any % of variance specifically, we can use a screeplot (figure 4.8) to identify where the increase in variance becomes sufficiently small (i.e. beyond the 'elbow').

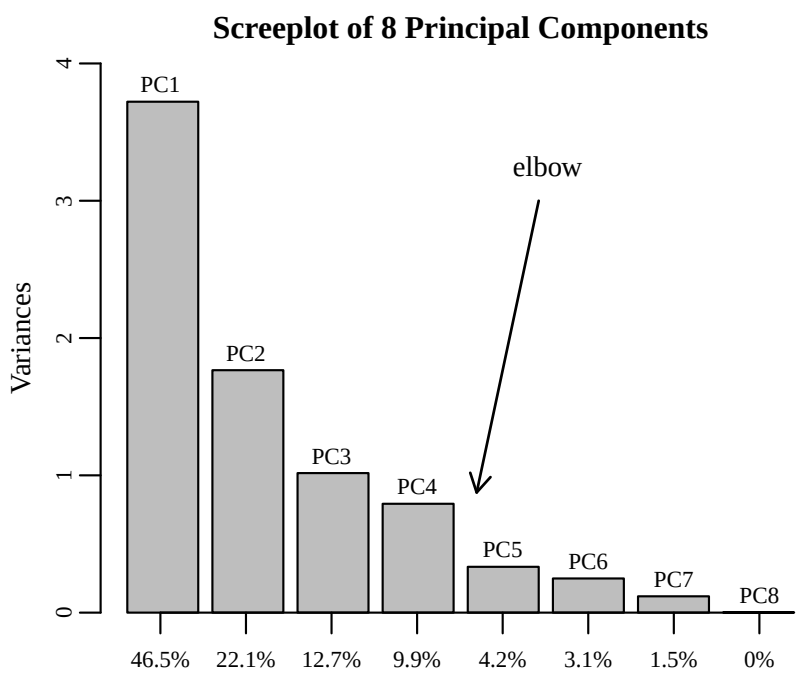


Figure 4.8: Screeplot of a PCA on data with 8 variables to 8 PCs. The first PC has by far the greatest amount of variance explained. Beyond PC_4, the cumulative proportion of explained variance increases slowly and could arguably be called the elbow. PC8 has contributes less then 0.1% to the cumulative variance explained.

4.4 Examples of PCA in R

Basic script

```
data(iris)
X <- iris[, -5]  # the (numeric) explanatory variables
Y <- iris[, 5]  # the response variable (here: Species)
PCA <- prcomp(X, center = TRUE, scale = FALSE)  # default
PCA <- prcomp(X)  # identical, since we are using the default
scores <- PCA$x  # the scores
loadings <- PCA$rotation  # the loadings
summary(PCA)  # cumulative proportion of variance explained per principal component
loadings  # inspect the contributions of individual variables to each component
plot(PCA)  # produces a screeplot
biplot(PCA)  # produces a biplot (PC1 vs PC2)
biplot(PCA, choices = c(1, 3))  # produces a biplot (PC1 vs PC3)
biplot(PCA, choices = c(2, 3))  # produces a biplot (PC2 vs PC3)
```

Manually produce a biplot (offers more flexibility)

```
plot(scores[, 1:2])  # PC1 vs PC2
arrows(0, 0, loadings[, 1], loadings[, 2])
```

Custom biplot example

```
Pch <- 19 # change point character
Col <- c("darkred", "darkblue", "darkgreen") # change colorscheme
var.expl <- round(summary(PCA)$importance * 100, 1)[2 ,] # variance explained (%)
plot(scores[, 1:2], pch = Pch, col = Col[iris$Species], main = "PCA",   # plot the scores
     xlab = paste0("PC1 (", var.expl[1], "% of variance explained)"),
     ylab = paste0("PC2 (", var.expl[2], "%)"))
arrows(0, 0, loadings[, 1], loadings[, 2], length = 0.1, lwd = 2)       # add the loadings
text(loadings[, 1] * 1.1, loadings[, 2] * 1.1, labels = 1:4)            # add labels (1:4)
legend("topright", legend = levels(iris$Species), col = Col, pch = Pch) # legend
```

4.4.1 Other Uses of PCA

Principal component analysis (PCA) is dimension reduction. It is often used to gain a first impression of your data, but can also be used as input for more advanced methods. In –omics research, systems biology and more generally in all high-dimensional data, PCA can be used to summarize and visualize your data. One other use of PCA is to find out which variables are responsible for the majority of the variance in the data.

Principal Component Regression

When (a large number of) explanatory variables p are (strongly) correlated, PCA can extract $k \leq p$ uncorrelated principal components. Using these as explanatory variables in regression models is called principal component regression (PCR). Often the first k components are chosen such that their total explained variance is greater than 95%. Make sure you **do not** include the response variable in the PCA part of PCR, lest the results are biased. A correct way to find principal components that correlate with the response variable is known as supervised PCA. Alternatively, one might opt for regularization (e.g. ridge regression).

- The advantage of PCR is a smaller number of parameters to estimate from uncorrelated input.
- The disadvantage of PCR is poorer interpretability than e.g. ridge regression, since the coefficients in PCR are based on combinations of all variables.

4.5 Summary

PCA seeks the largest variance among a set of (possibly correlated) variables and creates new dimensions (i.e. the principal components) which are independent. These dimensions contain the largest variance in the original data in descending order, such that we can use the first k principal components instead of all data.

(. . .)

5 Pitfalls in Statistics

5.1 The Objective is not Decided Yet

It is sometimes difficult to form a clear hypothesis or research objective without having engaged in any experimentation yet. What could go wrong, right? In principle, this pragmatic approach is fine, so long as the study is only preliminary, a pilot, *exploratory*.[1] One can search for new potentially interesting hypotheses by searching for patterns and associations in the data after experimentation. This is the quintessential example of exploration as research objective, which is a valid approach with one major caveat:

[1] Contains no mentions of significance.

- **The same data cannot be used to form *and* test hypotheses.**

Re-use of data is guarenteed to lead to many false positives and misleading conclusions. Examples of misuse of exploration include:

1. Deciding to test one-sided *after* inspecting the data;
2. Testing many hypotheses and reporting only significant ones;
3. Fitting a large number of models, followed by inference on the 'best' model.

In case (1) or (2) occurs, the actual false positive rate is much higher than the chosen level of significance (α).[2] In case of (3), one is likely to find a model that describes the sample extremely well, but does not generalize to the population (also see **5.3**).

[2] Chapter ***X*** elaborates on this problem.

All of these problems are forms of bias towards the sample. There always exists some spurious correlation or coincidental significance. Deciding on the research objective *before* data collection, can largely avoid this type of bias.

> *To be sure of hitting the target, shoot first and call whatever you hit the target.*
> — Ashleigh Brilliant

5.2 Continuous or Discrete?

A common misconception is (...) to be included. Never discretize continuous variables.

5.3 Stepwise Regression

(...) to be included. Don't seek the best model through stepwise inclusion/exclusion of variables. If you do, its p-values are optimistically biased.

5.4 Ritualized Statistics

There is no (correct) step-by-step guide to statistical analysis. You have to make considerations and decisions based on your knowledge of both biology and statistics.

For example, suppose the objective is confirmation. There is no harm in using a rough guideline, such as:

1. Make assumptions and form a null-hypothesis;
2. Collect data to build evidence against this null-hypothesis;
3. Check the validity of assumptions through diagnostics;
4. Perform an appropriate test;
5. Interpret the results.

However, compare this to the (bad) guideline below:

1. Assume normality and form null-hypothesis;
2. Collect data to build evidence against null-hypothesis;
3. Check for deviations from normality, then compare the groups' variances to assess whether a Welch t-test is required;
4. Perform a t-test, or if normality is deviated from, a Wilcoxon test;
5. Interpret the results.

The problem with this guideline is that it tries to fit the problem to a limited set of statistical techniques, rather than trying to find an appropriate technique for this specific problem. There is a reason you are reading a book, rather than a flowchart. Like all science, statistics has its nuances, its rules and exceptions. What you learned to do in this book is useful, what you learned *not* to do more so.

If the problem does not fit into a category you are familiar with, it is better to ask for help than to change the problem to something

you can do in a familiar way. A good first place to go is the forum CrossValidated. Perhaps someone else has had a similar question, or you can try asking your own.

5.5 What Should I Correct For?

(...) to be included. Refer to ***2.6.3*** - ***2.7.3*** for now.

5.6 Personalized Medicine

(...) to be included. In short, don't fall for the trap of dividing everything into smaller and smaller groups. Instead, consider an efficient way to model individual effect, such as a random intercept (see Advanced Statistics—Mixed Models).

6 Multiple Testing

Multiple testing occurs when interest lies in more than one hypothesis, or similarly, (the significance of) the effect of more than one explanatory variable on the response variable. For example, we might be interested in which of many genes are significantly differentially expressed in cancer patients.

*This chapter assumes you are familiar with the error rates α and β. Refer to chapter **6** of Introduction to Biostatistics.*

6.1 FWER

First we will discuss the Family-Wise Error Rate (FWER) and its archetypal correction: Bonferroni. Suppose we are interested in the effects of *age* and *gender* on a drug's efficacy, we can define two type I errors:

1. We find a significant effect of *age* on drug efficacy, even though there is none.
2. We find a significant effect of *gender* on drug efficacy, even though there is none.

This could occur for example because our sample is not representative of the population. What would now be the chance of a type I error? It can't be 5%, because there are two variables and thus two possibilities of a type I error. It also probably isn't $2 \cdot 5\% = 10\%$, because through this logic, measuring 20 variables would *guarantee* a false positive ($20 \cdot 5\% = 100\%$), which obviously doesn't have to be the case.

The FWER is the overall chance of a false positive, calculated by multiplying each chance of **not** making a type I error and subtracting that from 1:

$$\text{FWER} = 1 - (1 - \alpha)^k \tag{6.1}$$

Where k is the number of p-values. For example, one variable has a 95% of **not** making a type I error, so its FWER is $1 - 0.95 = 0.05$, the same as α. Our two-variable example of *age* and *gender* has two instances of 95% chance to **not** make a type I error, so its FWER is $1 - 0.95 \cdot 0.95 = 1 - 0.95^2 \approx 9.8\%$. Although the two variable case is approximately $2 \cdot 5\%$, when testing 20 variables the FWER is $1 - 0.95^{20} \approx 0.64 \neq 1$.

The simplest correction for the FWER is the Bonferroni correction, which multiplies all p-values by the number of p-values, again denoted by k.

$$p_{\text{Bonferroni},i} = p_i \cdot k \tag{6.2}$$

Which is equivalent to dividing α by k. The number of false positives in the 10 variable case would then be $1 - 0.995^{10} \approx 0.05$, which is the accepted number of 5%.

6.2 FDR

An alternative to FWER is the False Discovery Rate (FDR). Instead of penalizing the p-value by a flat amount, dependent on the amount of tests performed, the FDR becomes more lenient as more tests remain significant. The justification for this is as follows: If the most significant p-value gets the harshest penalty, and it is still significant (< 0.05), it is technically the same as the FWER. However, as the number of significant p-values increases, the real chance of making a type I error decreases.

In FDR, the i-th adjusted p-value is defined as:

$$p_{\text{adj},i} = p_i \cdot \frac{k}{i} \tag{6.3}$$

Where k is the number of tests and p the original p-values, sorted from low to high.

Because we want to respect the ordering of the p-values, we add the restriction:

$$p_{\text{FDR},i} = \min(p_{\text{adj},i},\ p_{\text{adj},i+1}) \tag{6.4}$$

Where each $p_{\text{adj},i}$ is the i-th value from eq. (6.3). Respecting the ordering means that the most significant gene is still the lowest, the second most still the second lowest, etcetera.

6.2.1 FDR correction in R

In `R`, you can simply correct with the function `p.adjust(x, method = 'fdr')`, where `x` is a vector of p-values. The output will be a vector

with corrected p-values.

```
# 10 T-tests, comparing cases to controls for 10 different genes:
p.values <- numeric(10)  # a vector for storing the p-values
p.values[1] <- t.test(control.gene01, case.gene01)$p.value  # p-value of the first gene
p.values[2] <- t.test(control.gene02, case.gene02)$p.value  # p-value of the second gene
p.values[3] <- t.test(control.gene03, case.gene03)$p.value  # p-value of the third gene
# ... you get the point
p.values[10] <- t.test(control.gene10, case.gene10)$p.value  # p-value of the last gene

p.adjust(p.values, method = "fdr")  # FDR corrected p-values


# Linear regression with 6 explanatory variables (works the same with
# interactions):
model <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6)
p.values <- anova(model)$"Pr(>F)"  # anova() is just a function for comparing models,
# not an ANOVA in the traditional sense.
p.values <- p.values[!is.na(p.values)]  # removes 'NA', or missing p-values

p.adjust(p.values, method = "fdr")  # FDR corrected p-values
```

6.3 FWER vs FDR (*)

FWER assumes a worst case scenario: all null-hypotheses are true / there are no true effects. FDR assumes that if the lowest (i.e. most significant) p-value remains significant after the FWER penalty, there apparently *are* effects. Hence, it makes little sense to correct the remaining p-values with the same penalty. The more of these effects remain significant, the lower the overall type I error rate and thus the penalty decreases.

FDR is generally a better choice than the Bonferroni adjustment (FWER), because Bonferroni costs a lot of power. However, there is one scenario where FDR is not appropriate: Suppose we do a lot of tests that are known to have true effects. Since every significant test makes the next more lenient and eq. (6.4) preserves order, these significant p-values may actually result in other p-values remaining significant as well. This inflates the type I error rate.

(...) example to be included.

6.4 lFDR (*)

(...) to be included.

7 Equivalence Testing (*)

(. . .) to be included.

7.1 Non-Inferiority

(. . .) to be included.

7.2 Two One-Sided Tests of Equivalence (TOST)

(. . .) to be included.

8 Missing Values (*)

Missing values are a difficult problem. Generally, a distinction is made between missing values that are:

- **Missing completely at random** (MCAR): There is no reason to suspect that these values are missing *because* they are extreme or because of (an extreme value of) another variable;
- **Missing at random** (MAR): There is no reason to suspect that these values are missing because they are extreme, but they might be missing due to (an extreme value of) another variable;
- **Missing not at random** (MNAR): Data which is missing is missing because it is extreme.

Naturally, MCAR is considered the least problematic and MNAR the most. When missing data is MCAR, imputation can be used to fill in the missing values. They can be replaced by the mean of non-missing values for example.

A Example Analyses

(. . .) to be included.

B Further Reading

Ideally, every scientist would read every good book on statistics ever written, but obviously there is no time for that. However, beyond the statistics course in your education, you may find yourself limited by the delibarately shallow explanations in this book series. For a much more comprehensive introduction to linear models, I highly recommend:

- Fox (2008): Applied Regression Analysis and Generalized Linear Models.

To understand the expressions in Fox (2008), you may want to learn the basics of linear algebra first. There is a free video series on **YouTube** by Gilbert Strang from MIT OpenCourseWare.

(. . .)

Acknowledgements

- **Dr. H.G.J. van Mil** for invaluable support and vision to improve the quality of teaching statistics
- **L.H. Tran, BSc** for proof-reading of earlier versions
- **V. de Bakker, MSc** for proof-reading of earlier versions