Statistical Modelling

Contents

Preliminary remarks

While we show how to implement statistical tests and models in SAS in class, these note will illustrate the concepts using R: visit the R-project website to download the program. The most popular graphical cross-platform front-end is RStudio Desktop.

The most famous quote about statistical models is probably due to George Box, who claimed that "all models are wrong, but some are useful". This standpoint is reductive: Peter McCullagh and John Nelder wrote in the preamble of their book (emphasis mine)

> Modelling in science remains, partly at least, an art. Some principles do exist, however, to guide the modeller. The first is that all models are wrong; some, though, are better than others and we can search for the better ones. At the same time we must recognize that eternal truth is not within our grasp.

And this quote by David R. Cox adds to the point:

> …it does not seem helpful just to say that all models are wrong. The very word model implies simplification and idealization. The idea that com- plex physical, biological or sociological systems can be exactly described by a few formulae is patently absurd. The construction of idealized representations that capture important stable aspects of such systems is, however, a vital part of general scientific analysis and statistical models, especially substantive ones, do not seem essentially different from other kinds of model.

Chapter 1

Introduction to statistical inference

Statistical modelling requires a good graps of statistical inference: as such, we begin with a review of hypothesis testing and graphical exploratory data analysis.

The purpose of statistical inference is to draw conclusions based on data. Scientific research relies on hypothesis testing: once an hypothesis is formulated, the researcher collects data, performs a test and concludes as to whether there is evidence for the proposed theory.

There are two main data type: experimental data are typically collected in a control environment following a research protocol with a particular experimental design: they serve to answer questions specified ahead of time. This approach is highly desirable to avoid the garden of forking paths (researchers unfortunately tend to refine or change their hypothesis in light of data, which invalidates their findings — preregistration alleviates this somewhat). While experimental data are highly desirable, it is not always possible to collect experimental data: for example, an economist cannot modify interest rates to see how it impacts consumer savings. When data have been collected beforehand without intervention (for other purposes), these are called observational. These will be the ones most frequently encountered.

A stochastic model will comprise two ingredients: a distribution for the random data and a formula linking the parameters or the conditional expectation of a response variable Y to a set of explanatories X. A model can serve to either predict new outcomes (predictive modelling) or else to test research hypothesis about the effect of the explanatory variables on the response (explanatory model). These two objectives are of course not mutually exclusive even if we distinguish in practice inference and prediction.

A predictive model gives predictions of Y for different combinations of explanatory variables or future data. For example, one could try to forecast the enery consumption of a house as a function of weather, the number of inhabitants and its size. Black boxes used in machine learning are often used solely for prediction: these models are not easily interpreted and they

often ignore the data structure.

By constrast, explicative models are often simple and interpretable: regression models are often used for inference purpose and we will focus on these.

- Are consumer ready to spend more when they pay by credit card rather than by cash?
- Is there wage discrimination towards women in a US college?
- University degree: "is the university experience worth the cost''?
- What are the criteria impacting health insurance premiums?
- Is the price of gasoline more expansive in the Gaspé peninsula than in the rest of Quebec? A report of the Régie de l'énergie examines the question
- Are driving tests in the UK easier if you live in a rural area? An analysis of The Guardian hints that it is the case.
- Does the risk of transmission of Covid19 increase with distancing? A (bad) meta-analysis says two meters is better than one (or how to draw erroneous conclusions from a bad model).

1.1 Hypothesis testing

An hypothesis test is a binary decision rule used to evaluate the statistical evidence provided by a sample to make a decision regarding the underlying population. The main steps involved are:

- define the model parameters
- formulate the alternative and null hypothesis
- choose and calculate the test statistic
- obtain the null distribution describing the behaviour of the test statistic under $\mathcal{H}_0$
- calculate the p-value
- conclude (reject or fail to reject $\mathcal{H}_0$) in the context of the problem.

A good analogy for hypothesis tests is a trial for murder on which you are appointed juror.

- The judge lets you choose between two mutually exclusive outcome, guilty or not guilty, based on the evidence presented in court.
- The presumption of innocence applies and evidences are judged under this optic: are evidence remotely plausible if the person was innocent? The burden of the proof lies with the prosecution to avoid as much as possible judicial errors. The null hypothesis $\mathcal{H}_0$ is not guilty, whereas the alternative $\mathcal{H}_a$ is guilty. If there is a reasonable doubt, the verdict of the trial will be not guilty.
- The test statistic (and the choice of test) represents the summary of the proof. The more overwhelming the evidence, the higher the chance the accused will be declared guilty. The prosecutor chooses the proof so as to best outline this: the choice of evidence (statistic) ultimately will maximise the evidence, which parallels the power of the test.

- The final step is the verdict. This is a binary decision, guilty or not guilty. For an hypothesis test performed at level α, one would reject (guilty) if the p-value is less than α.

The above description provides some heuristic, but lack crucial details developed in the next section written by Juliana Schulz.

1.1.1 Hypothesis

In statistical tests we have two hypotheses: the null hypothesis (H_0) and the alternative hypothesis (H_1). Usually, the null hypothesis is the 'status quo' and the alternative is what we're really interested in testing. A statistical hypothesis test allows us to decide whether or not our data provides enough evidence to reject H_0 in favour of H_1, subject to some pre-specified risk of error. Usually, hypothesis tests involve a parameter, say θ, which characterizes the underlying distribution at the population level ans whose value is unknown. A two-sided hypothesis test regarding a parameter θ has the form

$$\mathcal{H}_0 : \theta = \theta_0 \qquad \text{versus} \qquad \mathcal{H}_a : \theta \neq \theta_0.$$

We are testing whether or not θ is precisely equal to the value θ_0. The hypotheses are a statistical representation of our research question.

For example, for a two-sided test for the regression coefficient β_j associated to an explanatory variable X_j, the null and alternative hypothesis are explicative d'intérêt X_j, les hypothèses sont

$$\mathcal{H}_0 : \beta_j = \beta_j^0 \qquad \text{versus} \qquad \mathcal{H}_a : \beta_j \neq \beta_j^0,$$

where β_j^0 is some value that reflects the research question of interest. For example, if $\beta_j^0 = 0$, the underlying question is: is covariate X_j impacting the response Y once other variables have been taken into account?

Note that we can impose direction in the hypotheses and consider alternatives of the form $\mathcal{H}_a : \theta > \theta_0$ or $\mathcal{H}_a : \theta < \theta_0$.

1.1.2 Test statistic

A test statistic T is a functional of the data that summarise the information contained in the sample for θ. The form of the test statistic is chosen such that we know its underlying distribution under H_0, that is, the potential values taken by T and their relative probability if H_0 is true. Indeed, Y is a random variable and its value change from one sample to the next. This allows us to determine what values of T are likely if H_0 is true. Many statistics

we will consider are Wald statistic, of the form

$$T = \frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})}$$

where $\hat{\theta}$ is an estimator of θ, θ_0 is the postulated value of the parameter and $\text{se}(\hat{\theta})$ is an estimator of the standard deviation of the test statistic $\hat{\theta}$.

For example, to test whether the mean of a population is zero, we set

$$\mathcal{H}_0 : \mu = 0, \qquad \mathcal{H}_a : \mu \neq 0,$$

and the Wald statistic is

$$T = \frac{\overline{X} - 0}{S_n/\sqrt{n}}$$

where $\overline{X}$ is the sample mean of $X_1, \ldots, X_n$,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + \cdots + X_n}{n}$$

and the standard error (of the mean) $\overline{X}$ is $S_n/\sqrt{n}$; the sample variance S_n is an estimator of the standard deviation σ,

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Its important to distinguish between procedures/formulas and their numerical values. An estimator is a rule or formula used to calculate an estimate of some parameter or quantity of interest based on observed data. For example, the sample mean $\bar{X}$ is an estimator of the population mean μ. Once we have observed data we can actually compute the sample mean, that is, we have an estimate — an actual value. In other words,

- an estimator is the procedure or formula telling us how to use sample data to compute an estimate. Its a random variable since it depends on the sample.
- an estimate is the numerical value obtained once we apply the formula to observed data

1.1.3 Null distribution and p-value

The p-value allows us to decide whether the observed value of the test statistic T is plausible under H_0. Specifically, the p-value is the probability that the test statistic is equal or more extreme to the estimate computed from the data, assuming H_0 is true. Suppose that based

on a random sample $X_1, \dots, X_n$ we obtain a statistic whose value $T = t$. For a two-sided test $\mathcal{H}_0 : \theta = \theta_0$ vs. $\mathcal{H}_a : \theta \neq \theta_0$, the p-value is $\mathsf{Pr}_0(|T| \geq |t|)$. If the distribution of T is symmetric around zero, the p-value is

$$p = 2 \times \mathsf{Pr}_0(T \geq |t|).$$

Consider the example of a two-sided test involving the population mean $H_0 : \mu = 0$ against the alternative $H_1 : \mu \neq 0$. Assuming the random sample comes from a normal (population) $\mathsf{No}(\mu, \sigma^2)$, it can be shown that if H_0 is true (that is, if $\mu = 0$), the test statistic

$$T = \frac{\overline{X}}{S/\sqrt{n}}$$

follows a Student-t distribution with $n - 1$ degrees of freedom, denoted St_{n-1}. This allows us to calculate the p-value (either from a table, or using some statistical software). The Student-t distribution is symmetric about zero, so the p-value is $P = 2 \times \mathsf{Pr}(T_{n-1} > |t|)$, where $T \sim \mathsf{St}_{n-1}$.

1.1.4 Conclusion

The p-value allows us to make a decision about the null hypothesis. If $\mathcal{H}_0$ is true, the p-value follows a uniform distribution. Thus, if the p-value is small, this means observing an outcome more extreme than $T = t$ is unlikely, and so we're inclined to think that H_0 is not true. There's always some underlying risk that we're making a mistake when we make a decision. In statistic, there are two type of errors:

- type I error: we reject H_0 when H_0 is true,
- type II error: we fail to reject H_0 when H_0 is false.

These hypothesis are not judged equally: we seek to avoid error of type I (judicial errors, corresponding to condamning an innocent). To prevent this, we fix a the level of the test, α, which captures our tolerance to the risk of commiting a type I error: the higher the level of the test α, the more often we will reject the null hypothesis when the latter is true. The value of $\alpha \in (0, 1)$ is the probability of rejecting $\mathcal{H}_0$ when $\mathcal{H}_0$ is in fact true,

$$\alpha = \mathsf{Pr}_0 \left(\text{ reject } \mathcal{H}_0 \right).$$

The level α is fixed beforehand, typically 1%, 5% or 10%. Keep in mind that the probability of type I error is α only if the null model for $\mathcal{H}_0$ is correct (sic) and correspond to the data generating mechanism.

The focus on type I error is best understood by thinking about medical trial: you need to prove a new cure is better than existing alternatives drugs or placebo, to avoid extra costs or harming patients (think of Didier Raoult and his unsubstantiated claims that hydrochloroquine, an antipaludean drug, should be recommended treatment against Covid19).

Decision \ true model	$\mathcal{H}_0$	$\mathcal{H}_a$
fail to reject $\mathcal{H}_0$	✓	type II error
reject $\mathcal{H}_0$	type I error	✓

To make a decision, we compare our p-value P with the level of the test α:

- if $P < \alpha$, we reject $\mathcal{H}_0$;
- if $P \geq \alpha$, we fail to reject $\mathcal{H}_0$.

Do not mix up level of the test (probability fixed beforehand by the researcher) and the p-value. If you do a test at level 5%, the probability of type I error is by definition α and does not depend on the p-value. The latter is conditional probability of observing a more extreme likelihood given the null distribution $\mathcal{H}_0$ is true.

1.1.5 Power

There are two sides to an hypothesis test: either we want to show it is not unreasonable to assume the null hypothesis, or else we want to show beyond reasonable doubt that a difference or effect is significative: for example, one could wish to demonstrate that a new website design (alternative hypothesis) leads to a significant increase in sales relative to the status quo. Our ability to detect these improvements and make discoveries depends on the power of the test: the larger the power, the greater our ability to reject $\mathcal{H}_0$ when the latter is false.

Failing to reject $\mathcal{H}_0$ when $\mathcal{H}_a$ is true corresponds to the definition of type II error, the probability of which is $1-\gamma$, say. The power of a test is the probability of rejecting $\mathcal{H}_0$ when $\mathcal{H}_0$ is false, i.e.,

$$\gamma = \mathsf{Pr}_a(\text{reject}\mathcal{H}_0)$$

Depending on the alternative models, it is more or less easy to detect that the null hypothesis is false and reject in favor of an alternative.

We want a test to have high power, i.e., that γ be as close to 1 as possible. Minimally, the power of the test should be α because we reject the null hypothesis α fraction of the time even when $\mathcal{H}_0$ is true. Power depends on many criteria, notably

- the effect size: the bigger the difference between the postulated value for θ_0 under $\mathcal{H}_0$ and the observed behavior, the easier it is to detect it. (Figure 1.3);
- variability: the less noisy your data, the easier it is to detect differences between the curves (big differences are easier to spot, as Figure 1.2 shows);
- the sample size: the more observation, the higher our ability to detect significative differences because the standard error decreases with sample size n at a rate (typically)

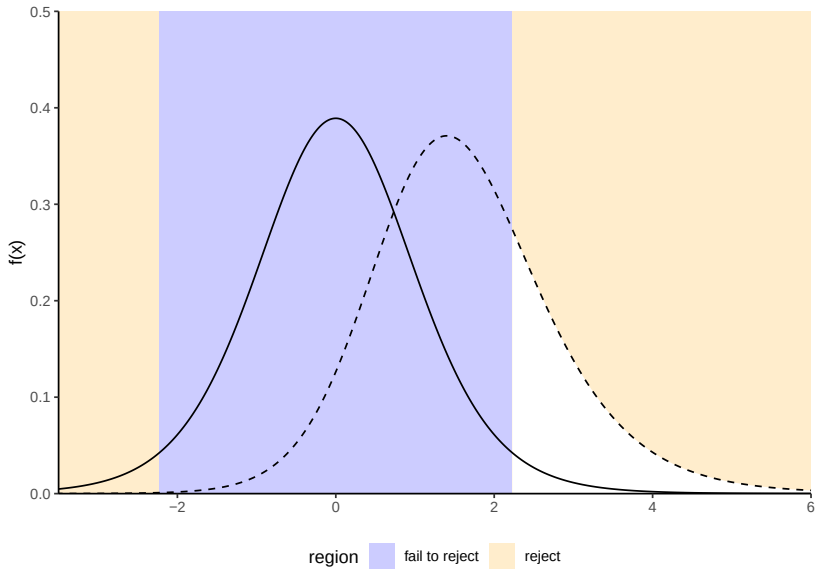

Figure 1.1: Comparison between null distribution (full curve) and a specific alternative for a *t*-test (dashed line). The power corresponds to the area under the curve of the density of the alternative distribution which is in the rejection area (in white).

of $n^{-1/2}$. The null distribution also becomes more concentrated as the sample size increase.

- the choice of test statistic: for example, rank-based statistics discard information about the actual values and care only about relative ranking. Resulting tests are less powerful, but are typically more robust to model misspecification and outliers. The statistics we will choose are standard and amongst the most powerful: as such, we won't dwell on this factor.

To calculate the power of a test, we need to single out a specific alternative hypothesis. In very special case, analytic derivations are possible: for example, the one-sample t-test statistic $T = \sqrt{n}(\overline{X}_n - \mu_0)/S_n \sim \mathcal{T}_{n-1}$ for a normal sample follows a noncentral Student-t distribution with noncentrality parameter Δ if the expectation of the population is $\Delta + \mu_0$. In general, such closed-form expressions are not easily obtained and we compute instead the power of a test through Monte Carlo methods. For a given alternative, we simulate repeatedly samples from the model, compute the test statistic on these new samples and the associated p-values based on the postulated null hypothesis. We can then calculate the proportion of tests that lead to a rejection of the null hypothesis at level α, namely the percentage of p-values smaller than α.

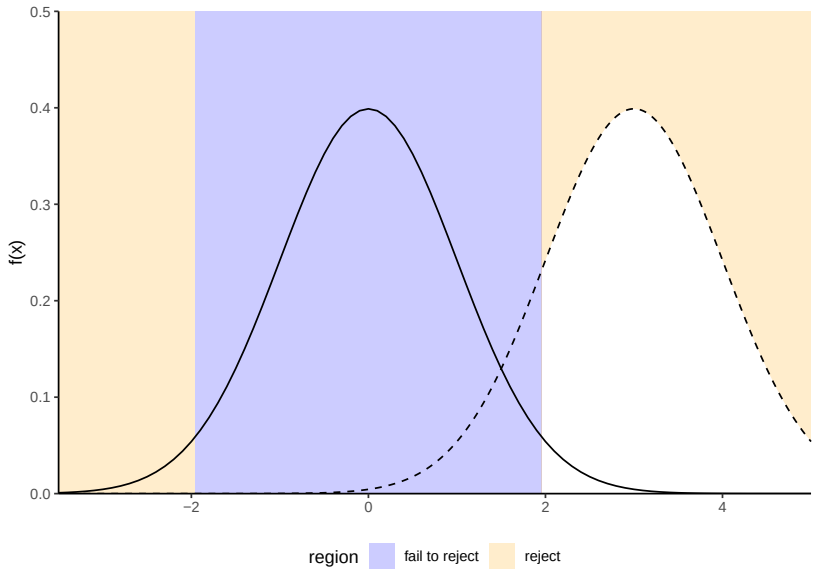

Figure 1.2: Increase in power due to an increase in the mean difference between the null and alternative hypothesis. Power is the area in the rejection region (in white) under the alternative distribution (dashed): the latter is more shifted to the right relative to the null distribution (full line).

1.1.6 Confidence interval

A confidence interval is an alternative way to present the conclusions of an hypothesis test performed at significance level α. It is often combined with a point estimator $\hat{\theta}$ to give an indication of the variability of the estimation procedure. Wald-based $(1-\alpha)$ confidence intervals for a parameter θ are of the form

$$\hat{\theta} \pm \mathfrak{q}_{\alpha/2}\, \mathrm{se}(\hat{\theta})$$

where $\mathfrak{q}_{\alpha/2}$ is the $1-\alpha/2$ quantile of the null distribution of the Wald statistic

$$T = \frac{\hat{\theta} - \theta}{\mathrm{se}(\hat{\theta})},$$

and where θ represents the postulated value for the fixed, but unknown value of the parameter. The bounds of the confidence intervals are random variables, since both $\hat{\theta}$ and $\mathrm{se}(\hat{\theta})$ are random variables: their values depend on the sample, and will vary from one sample to another.

For example, for a random sample $X_1, \ldots, X_n$ from a normal distribution $\mathsf{No}(\mu, \sigma)$, the $(1-\alpha)$

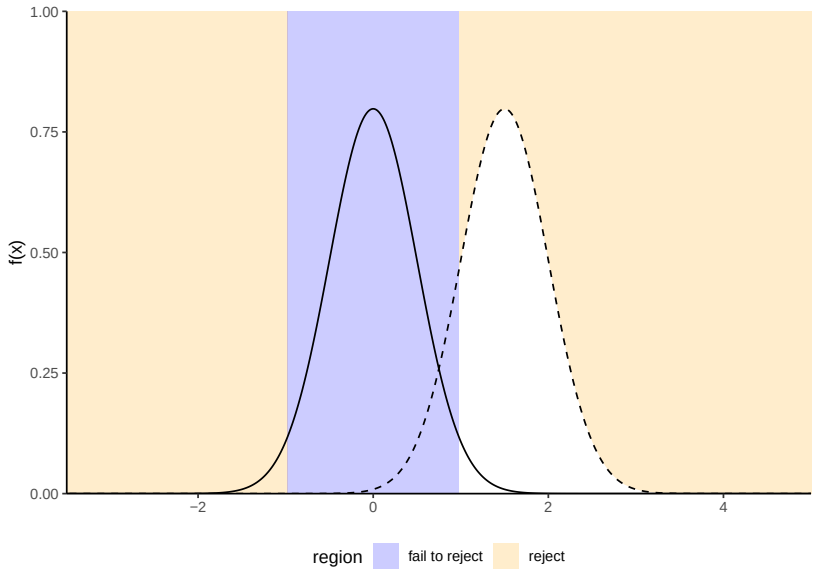

Figure 1.3: Increase of power due to an increase in the sample size or a decrease of standard deviation of the population: the null distribution (full line) is more concentrated. Power is given by the area (white) under the curve of the alternative distribution (dashed). In general, the null distribution changes with the sample size.

confidence interval for the population mean μ is

$$\overline{X} \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}$$

where $t_{n-1,\alpha/2}$ is the $1-\alpha/2$ quantile of a Student-t distribution with $n-1$ degrees of freedom.

Before the interval is calculated, there is a $1-\alpha$ probability that θ is contained in the random interval $(\hat{\theta}-\mathfrak{q}_{\alpha/2}\ \mathrm{se}(\hat{\theta}), \hat{\theta}+\mathfrak{q}_{\alpha/2}\ \mathrm{se}(\hat{\theta}))$, where $\hat{\theta}$ denotes the estimator. Once we obtain a sample and calculate the confidence interval, there is no more notion of probability: the true value of the parameter θ is either in the confidence interval or not. We can interpret confidence interval's as follows: if we were to repeat the experiment multiple times, and calculate a $1-\alpha$ confidence interval each time, then roughly $1-\alpha$ of the calculated confidence intervals would contain the true value of θ in repeated samples (in the same way, if you flip a coin, there is roughly a 50-50 chance of getting heads or tails, but any outcome will be either). Our confidence is in the procedure we use to calculate confidence intervals and not in the actual values we obtain from a sample.

If we are only interested in the binary decision rule reject/fail to reject $\mathcal{H}_0$, the confidence interval is equivalent to a p-value since it leads to the same conclusion. Whereas the $1-\alpha$

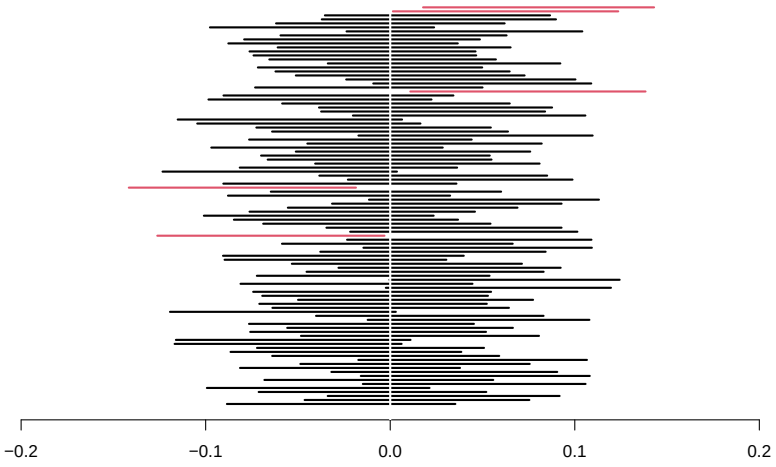

Figure 1.4: 95% confidence intervals for the mean of a standard normal population $\mathsf{No}(0, 1)$, with 100 random samples. On average, 5% of these intervals fail to include the true mean value of zero (in red).

confidence interval gives the set of all values for which the test statistic doesn't provide enough evidence to reject $\mathcal{H}_0$ at level α, the p-value gives the probability under the null of obtaning a result more extreme than the postulated value and so is more precise for this particular value. If the p-value is smaller than α, our null value θ will be outside of the confidence interval and vice-versa.

Example 1.1 (Online purchases of millenials). Suppose a researcher studies the evolution of online sales in Canada. She postulates that generation Y members make more online purchase than older generations. A survey is sent to a simple random sample of $n = 500$ individuals from the population with 160 members of generation Y and 340 older people. The response ariable is the total amount of online goods purchased in the previous month (in dollars).

In this example, we consider the difference between the average amount spent by Y members and those of previous generations: the mean difference in the samples is -16.49 dollars and thus millenials spend more. However, this in itself is not enough to conclude that the different is significative, nor can we say it is meaningful. The amount spent online varies from one individual to the next (and plausibly from month to month), and so different random samples would yield different mean differences.

The first step of our analysis is defining the parameters corresponding to quantities of interest and formulating the null and alternative hypothesis as a function of these parameters. We will consider a test for the difference in mean of the two populations, say μ_1 for the expected

amount spent by generation Y and μ_2 for older generations, with respective standard errors σ_1 and σ_2. We next write down our hypothesis: the researcher is interested in whether millenials spend more, so this is the alternative hypothesis, $\mathcal{H}_a : \mu_1 > \mu_2$. The null consists of all other values $\mathcal{H}_0 : \mu_1 \leq \mu_2$, but only $\mu_1 = \mu_2$ matters for the purpose of testing (why?)

The second step is the choice of test statistic. We consider the Welch (1947) statistic for a difference in mean between two samples,

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^{1/2}},$$

where $\overline{X}_i$ is the sample mean, S_i^2 is the unbiased variance estimator and n_i is the sample size for group i ($i = 1, 2$). If the mean difference between the two samples is zero, then $\overline{X}_1 - \overline{X}_2$ has mean zero and the difference has variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$. For our sample, the value of statistic is $T = -2.76$ Since the value changes from one sample to the next, we need to determine if this value is compatible with the null hypothesis by comparing it to the null distribution of T (when $\mathcal{H}_0$ is true and $\mu_1 - \mu_2 = 0$). We perform the test at level $\alpha = 0.05$.

The third step consists in obtaining a benchmark to determine if our result is extreme or unusual. To make comparisons easier, we standardize the statistic so its has mean zero and variance one under the null hypothesis $\mu_1 = \mu_2$, so as to obtain a dimensionless measure whose behaviour we know for large sample. The (mathematical) derivation of the null distribution is beyond the scope of this course, and will be given in all cases. Asymptotically, T follows a standard normal distribtion $\mathsf{No}(0, 1)$, but there exists a better finite-sample approximation when n_1 or n_2 is small; we use Satterthwaite (1946) and a Student-t distribution as null distribution.

It only remains to compute the p-value. If the null distribution is well-specified and $\mathcal{H}_0$ is true, then the random variable P is uniform on $[0, 1]$; we thus expect to obtain under the null something larger than 0.95 only 5% of the time for our one-sided alternative since we consider under $\mathcal{H}_0$ the event $\mathsf{Pr}(T > t)$. The p-value is 1 and, at level 5%, we reject the null hypothesis to conclude that millenials spend significantly than previous generation for monthly online purchases, with an estimated average difference of -16.49.

Example 1.2 (Price of Spanish high speed train tickets). The Spanish national railway company, Renfe, manages regional and high speed train tickets all over Spain and The Gurus harvested the price of tickets sold by Renfe. We are interested in trips between Madrid and Barcelona and, for now, ask the question: are tickets more expensive one way or another? To answer this, we consider a sample of 10000 tickets, but restrict attention to AVE tickets sold at Promo rate. Our test statistic will again be the mean difference between the price (in euros) for a train ticket for Madrid–Barcelona (μ_1) and the price for Barcelona–Madrid (μ_2), i.e., $\mu_1 - \mu_2$. The null hypothesis is that there are no difference in price, so $\mathcal{H}_0 : \mu_1 - \mu_2 = 0$. We again use Welch test statistic for two samples.

```
# Library for manipulating data, including the pipe operator (%>%)
library(poorman)
# Load data
data(renfe, package = "hecstatmod")
head(renfe, n = 5)
```

```
## # A tibble: 5 x 7
##   price type    class      fare     dest              duration wday
##   <dbl> <fct>   <fct>      <fct>    <fct>                <dbl> <fct>
## 1 143.  AVE     Preferente Promo    Barcelona-Madrid       190 6
## 2 182.  AVE     Preferente Flexible Barcelona-Madrid       190 2
## 3  86.8 AVE     Preferente Promo    Barcelona-Madrid       165 7
## 4  86.8 AVE     Preferente Promo    Barcelona-Madrid       190 7
## 5  69.0 AVE-TGV Preferente Promo    Barcelona-Madrid       175 4
```

```
# Sub-sample with only Promo tickets
renfe_promo <- renfe %>% subset(fare == "Promo")
# two-sample t-test and mean difference
ttest <- t.test(price~dest, data = renfe_promo)
ttest #print result
```

```
##
##  Welch Two Sample t-test
##
## data:  price by dest
## t = -1, df = 8040, p-value = 0.2
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.100  0.209
## sample estimates:
## mean in group Barcelona-Madrid mean in group Madrid-Barcelona
##                           82.1                           82.6
```

Rather than use the asymptotic distribution, whose validity stems from the central limit theorem, we could consider another approximation under the less restrictive assumption that the data are exchangeable: under the null hypothesis, there is no difference between the two destinations and so the label for destination (a binary indicator) is arbitrary. The reasoning underlying permutation tests is as follows: to create a benchmark, we will consider observations with the same number in each group, but permuting the labels. We then compute the test statistic on each of these datasets. If there are only a handful in each group (fewer

than 10), we could list all possible permutations of the data, but otherwise we can repeat this procedure many times, say 9999, to get a good approximation. This gives an approximate distribution from which we can extract the p-value by computing the rank of our statistic relative to the others.

```
# p-value (permutation test)
n <- nrow(renfe_promo)
B <- 1e4
ttest_stats <- numeric(B)
ttest_stats[1] <- ttest$statistic
set.seed(20200608) # set seed of pseudo-random number generator
for(i in 2:B){
  # Recalculate the test statistic, permuting the labels
  ttest_stats[i] <- t.test(price ~ dest[sample.int(n = n)],
                           data = renfe_promo)$statistic
}
# Graphics library
library(ggplot2)
# Plot the empirical permutation distribution
ggplot(data = data.frame(statistic = ttest_stats),
       aes(x=statistic)) +
  geom_histogram(bins = 30, aes(y=..density..), alpha = 0.2) +
  geom_density() +
  geom_vline(xintercept = ttest_stats[1]) +
  ylab("density") +
  stat_function(fun = dnorm, col = "blue")
```

The so-called bootstrap approximation to the p-value of the permutation test, 0.186, is the proportion of statistics that are more extreme than the one based on the original sample. It is nearly identical to that obtained from the Satterthwaite approximation, 0.182 (the Student-t distribution is numerically equivalent to a standard normal with that many degrees of freedom), as shown in Figure 1.5. Even if our sample is very large ($n = 8059$ observations), the difference is not statistically significative. With a bigger sample (the database has more than 2 million tickets), we could estimate more precisely the average difference, up to 1/100 of an euro: the price difference would eventually become statistically significative, but this says nothing about practical difference: 0.28 euros relative to an Promo ticket priced on average 82.56 euros is a negligible amount.

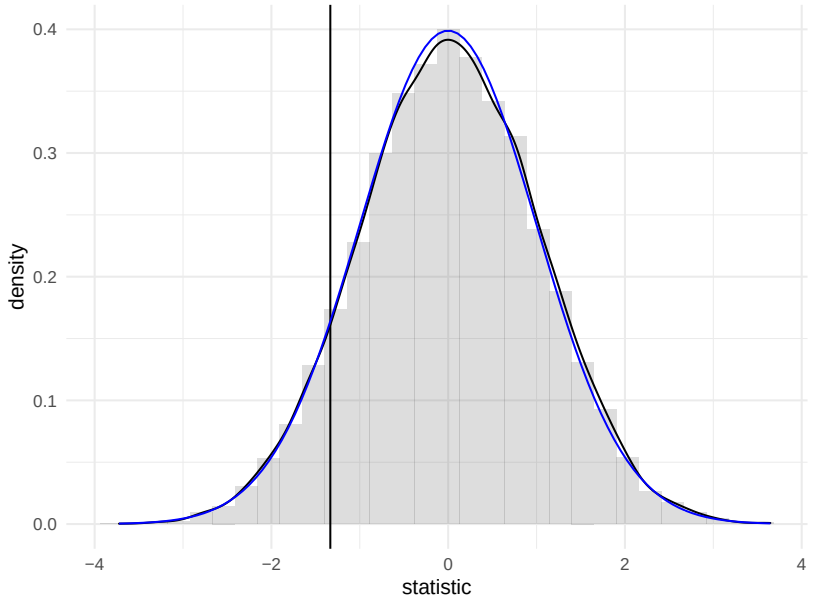

Figure 1.5: Permutation-based approximation to the null distribution of Welch two-sample t-test statistic (histogram and black curve) with standard normal approximation (blue curve) for the price of AVE tickets at promotional rate between Madrid and Barcelona. The value of the test statistic calculated using the original sample is represented by a vertical line.

1.2 Exploratory Data Analysis

Before fitting a model, it is advisable to understand the structure of the data to avoid interpretation errors. Basic knowledge of graphs is required and we will spend some time addressing this. Further references include

- Chapter 3, R for Data Science by Garrett Grolemund and Hadley Wickham
- Section 1.6 of OpenIntro Introductory Statistics with Randomization and Simulation
- Fundamentals of Data Visualization by Claus O. Wilke
- Chapter 1 of Data Visualization: A practical introduction by Kieran Healy

If exploratory data analysis is often neglected in statistics (perhaps because it has little to no mathematical foundations), it is crucial. More than a rigorous approach, it is an art: Grolemund and Wickham talk of "state of mind". The purpose of graphical exploratory data analysis is the extraction of useful information, often through a series of preliminary questions that are refined as the analysis progresses. Of particular interest and the relations and interactions between difference variables and the distribution of the variables themselves. The major steps for undertaking an exploratory analysis are:

1. Formulate questions about the data
2. Look for answers using frequency table, descriptive statistics and graphics.
3. Refine the questions in light of the finding

In a report, you should highlight the most import features in a summary so that the reader can grasp your understanding and so that you guide him or her in the interpretation of the data.

1.2.1 Polish your work

Pay as much attention to figures and tables than to the main text. These should always include a legend that describes and summarize the findings in the graph (so that the latter is standalone), name of variables (including units) on the axes, but also proper formatting so that the labels and numbers are readable (good printing quality, not too small). One picture is worth 1000 words, but make sure the graph tells a coherent story and that it is mentioned in the main text. Also ensure that only the necessary information is displayed: superfluous information (spurious digits, useless summary statistics) should not be presented.

1.2.2 Variable type

The data we will handled are stored in tables or frames. If the data frame is stocked in wide format, each line corresponds to an observation and each column to a variable: the entries of the data base contain the (numeric) values.

- a variable represents a characteristic of the population, for example the sex of an individual, the price of an item, etc.
- an observation is a set of measures (variables) collected under identical conditions for an individual or at a given time.

The choice of statistical model and test depends on the underlying type of the data collected. There are many choices: quantitative (discrete or continuous) if the variables are numeric, or qualitative (binary, nominal, ordinal) if they can be described using an adjective; I prefer the term categorical, which is more evocative.

Most of the models we will deal with are so-called regression models, in which the mean of a quantitative variable is a function of other variables, termed explanatories. There are two types of numerical variables

- a discrete variable takes a countable number of values, prime examples being binary variables or count variables.
- a continuous variable can take (in theory) an infinite possible number of values, even when measurements are rounded or measured with a limited precision (time, width, mass). In many case, we could also consider discrete variables as continuous if they take enough values (e.g., money).

Categorical variables take only a finite of values. They are regrouped in two groups, nominal if there is no ordering between levels (sex, color, country of origin) or ordinal if they are ordered (Likert scale, salary scale) and this ordering should be reflected in graphs or tables. We will bundle every categorical variable using arbitrary encoding for the levels: for modelling, these variables taking K possible values (or levels) must be transformed into a set of $K-1$ binary 0/1 variables, the omitted level corresponding to a baseline. Failing to declare categorical variables in your favorite software is a common mistake, especially when these are saved in the database using integers rather than strings.

1.2.3 Graphs

The main type of graph for representing categorical variables is bar plot (and modifications thereof). In a bar plot, the frequency of each category is represented in the y-axis as a function of the (ordered) levels on the x-axis. This representation is superior to the ignominious pie chart, a nuisance that ought to be banned (humans are very bad at comparing areas and a simple rotation changes the perception of the graph)!

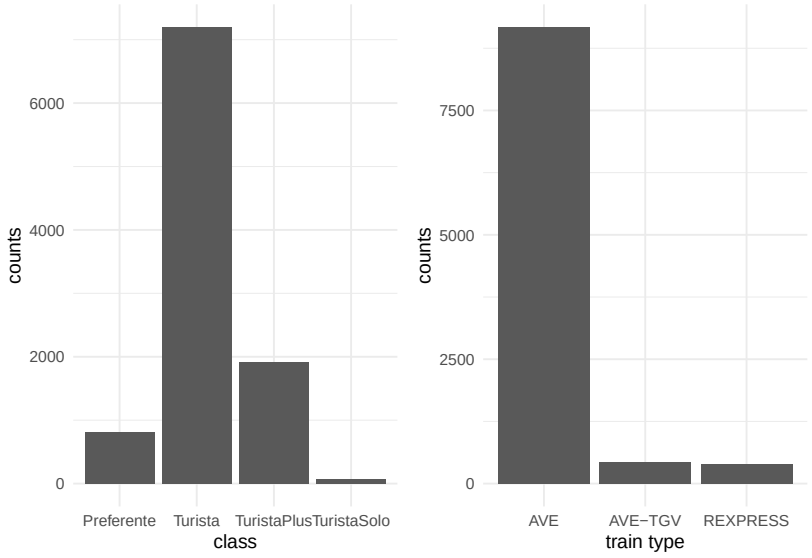

Figure 1.6: Bar plot of ticket class for Renfe tickets data

Continuous variables can take as many distinct values as there are observations, so we cannot simply count the number of occurences by unique values. Instead, we bin them into distinct intervals so as to obtain an histogram. The number of class depends on the number of observations: as a rule of thumb, the number of bins should not exceed $\sqrt{n}$, where n is the sample size. We can then obtain the frequency in each class, or else normalize the histogram so that the area under the bands equals one: this yields a discrete approximation

of the underlying density function. Varying the number of bins can help us detect patterns (rounding, asymmetry, multimodality).

Since we bin observations together, it is sometimes difficult to see where they fall. Adding rugs below or above the histogram will add observation about the range and values taken, where the heights of the bars in the histogram carry information about the (relative) frequency of the intervals.

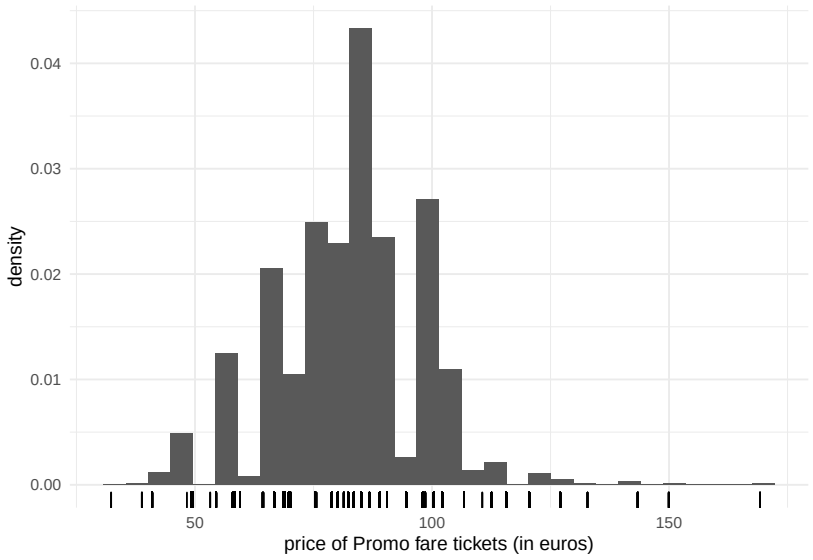

Figure 1.7: Histogram of Promo tickets for Renfe ticket data

If we have a lot of data, it sometimes help to focus only on selected summary statistics. A box-and-whiskers plot (or boxplot) represents five numbers

- The box gives the quartiles q_1, q_2, q_3 of the distribution. The middle bar q_2 is thus the median, so 50% of the observations are smaller or larger than this number.
- The length of the whiskers is up to 1.5 times the interquartiles range $q_3 - q_1$ (the whiskers extend until the latest point in the interval, so the largest observation that is smaller than $q_3 + 1.5(q_3 - q_1)$, etc.)
- Observations beyond the whiskers are represented by dots or circles, sometimes termed outliers. However, beware of this terminology: the larger the sample size, the more values will fall outside the whiskers. This is a drawback of boxplots, which was conceived at a time where the size of data sets was much smaller than what is current standards.

We can represent the distribution of a response variable as a function of a categorical variable by drawing a boxplot for each category and laying them side by side. A third variable, categorical, can be added via a color palette, as shown in Figure 1.9.

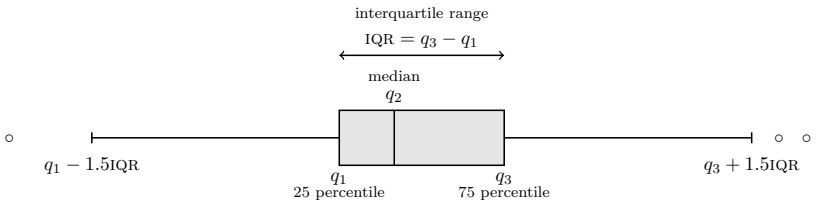

Figure 1.8: Box-and-whiskers plot

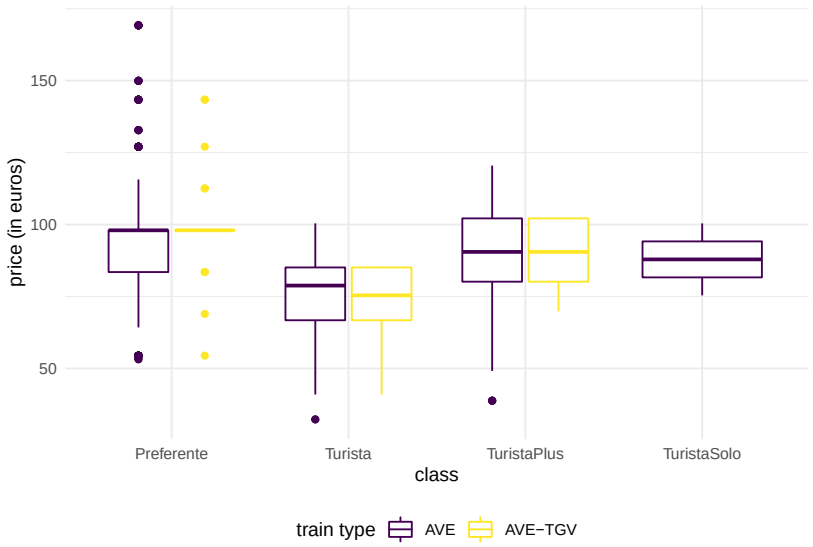

Figure 1.9: Box-and-whiskers plots for Promo fare tickets as a function of class and type for the Renfe tickets data.

Scatterplots are used to represent graphically the co-variation between two continuous variables: each tuple gives the coordinate of the point. If only a handful of large values are visible on the graph, a transformation may be useful: oftentimes, you will encounter graphs where the x- or y-axis is on the log-scale when the underlying variable is positive. If the number of data points is too large, it is hard to distinguish points because they are overlaid: adding transparency, or binning using a two-dimensional histogram with the frequency represented using color are potential solutions. The left panel of Figure 1.10 shows the 100 simulated observations, whereas the right-panel shows a larger sample of 10 000 points using hexagonal binning, an analog of the bivariate density.

Sometimes, continuous data have a particular structure, mostly when observations are collected over space or time. Time series are ordered and the response should be plotted on the y-axis as a function of time (on the x-axis). It is customary to draw segments between

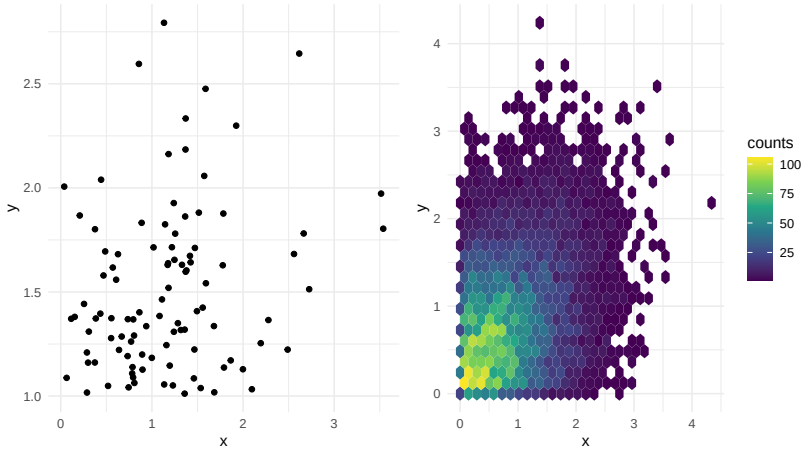

Figure 1.10: Scatterplot (left) and hexagonal heatmap of bidimensional bin counts (right) of simulated data.

observations, but this display is sometimes misleading.

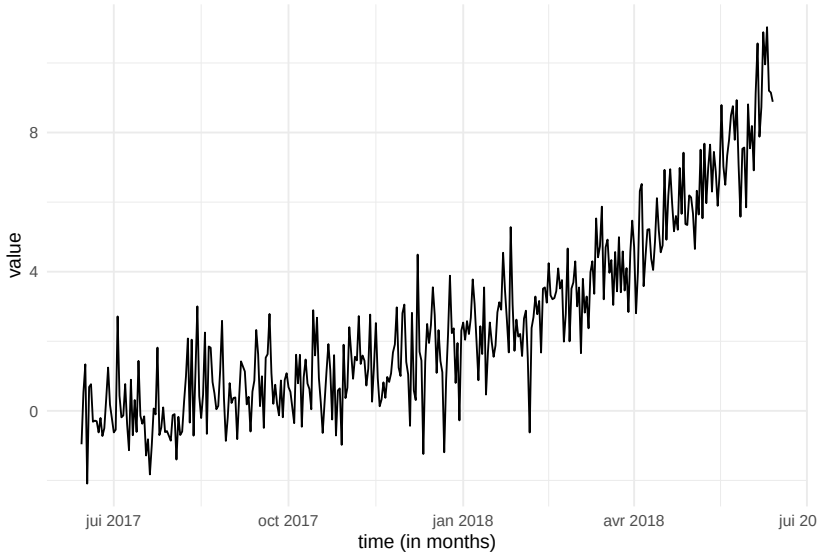

Figure 1.11: Graphical representation of a time series.

1.2.4 Exploratory data analysis

Rather than describe in details the exploratory analysis procedure, we proceed with an example that illustrates the process on the Renfe ticket dataset that was introduced previously.

Example 1.3 (Exploratory data analysis of Renfe tickets). First, read the documentation accompanying the dataset! The data base `renfe` contains the following variables:

- `price` price of the ticket (in euros);
- `dest` binary variable indicating the journey, either Barcelona to Madrid (`0`) or Madrid to Barcelona (`1`);
- `fare` categorical variable indicating the ticket fare, one of `AdultoIda`, `Promo` or `Flexible`;
- `class` categorical variable giving the ticket class, either `Preferente`, `Turista`, `TuristaPlus` or `TuristaSolo`;
- `type` categorical variable indicating the type of train, either Alta Velocidad Española (`AVE`), Alta Velocidad Española jointly with TGV (parternship between SNCF and Renfe for trains to/from Toulouse) `AVE-TGV` or regional train `REXPRESS`; only trains labelled `AVE` or `AVE-TGV` are high-speed trains.
- `duration` length of train journey (in minutes);
- `wday` categorical variable (integer) denoting the week day, ranging from Sunday (`1`) to Saturday (`7`).

There are no missing values and a quick view of the first row of the data frame (`head(renfe)`) shows that the data are stored in long format, meaning each line corresponds to a different ticket. We will begin our exploratory analysis with vague questions, for example

1. What are the factors determining the price and travel time?
2. Does travel time depend on the type of train?
3. What are the distinctive features of train types?
4. What are the main differences between fares?

Except for `price` and `duration`, all the other (explanatory) variables are categoricals. These need to be cast into factors (`factor`), especially integer-valued levels such as `wday`.

The database is clean and this preliminary preprocessing step has been done already. We can check the type of encoding using the command `str`, which also shows the data; the function `summary` is used to obtain descriptive statistics (min, max, mean, quartiles for continuous variables or else frequency for categorical variables); the function also returns the number of missing values (`NA`) of each column.

Data manipulation is often messy and R base syntax is particularly inelegant: data frames are list whose elements are accessed using `$`: for example `renfe$price`. A more legible and modular alternative is the pipe operator (`%>%`), with which one creates a logical chain of

command (this function is not part of R base packages, but the libraries `tidyverse` and the minimal alternative `poorman` include it).

```
renfe %>% count(class)
```

```
##         class    n
## 1  Preferente  809
## 2     Turista 7197
## 3 TuristaPlus 1916
## 4 TuristaSolo   78
```

```
# `count` is a shortcut for the following syntax
renfe %>% group_by(type) %>% tally()
```

```
##       type    n
## 1      AVE 9174
## 2  AVE-TGV  429
## 3 REXPRESS  397
```

```
renfe %>% group_by(fare) %>% tally()
```

```
##        fare    n
## 1 AdultoIda  397
## 2  Flexible 1544
## 3     Promo 8059
```

By counting the number of train tickets in each category, we notice there are as many `REXPRESS` tickets as there are tickets sold at `AdultoIda` fare. Using a contingency table to get the number in respective sub-categories of each of those variables confirms that all tickets in the database for RegioExpress trains are sold with the `AdultoIda` fare and that there is only a single class, `Turista`. There are few such tickets, only 397 out of 10 000. This raises a new question: why are such trains so unpopular?

```
##        fare     type    n
## 1 AdultoIda REXPRESS  397
## 2  Flexible      AVE 1446
## 3  Flexible  AVE-TGV   98
## 4     Promo      AVE 7728
## 5     Promo  AVE-TGV  331
```

We have only scratched the surface, but one could also notice that there are only 17 duration values on tickets (`renfe %>% distinct(duration)` or `unique(renfe$duration)`). This leads us to think the duration on the ticket (in minutes) is the expected travel time. The majority of those travel time (15 out of 17) are smaller than 3h15, but the other two exceed 9h! Looking at Google Maps, Madrid and Barcelona are 615km apart by car, 500km as the crow flies. this means some trains travel at about 200km/h, while others are closer to 70km/h. What are these slower trains? the variable `type` is the one most likely to encode this feature, and a quick look shows that the RegioExpress trains fall in the slow category (mystery solved!)

```
renfe %>%
  subset(duration > 200) %>%
  group_by(type, dest) %>%
  summarise("average duration" = mean(duration),
            "std. dev" = sd(duration),
            "average price" = mean(price),
            "std. dev" = sd(price))
```

```
##       type             dest average duration std. dev average price std. dev
## 1 REXPRESS Barcelona-Madrid               544        0          43.2        0
## 2 REXPRESS Madrid-Barcelona               562        0          43.2        0
```

The regular trains running between two cities take more than 9h, but one way (Madrid to Barcelona) is 18 minutes slower than in the other direction. More striking, we see that the price of the RegioExpress tickets is fixed: 43.25 euros regardless of direction. This is the most important finding so far, because these are not a sample for price: there is no variability! Graphics could have lead to the discovery (the boxplot of price as a function of train type would collapse to a single value).

We could have suspected that trains labeled `AVE` are faster: after all, it is the acronym of Alta Velocidad Española, literally Spanish high speed. What is the distinction between the two high speed train types. According to the SNCF website, AVE-TGV trains are partnership between Renfe and SNCF that operate between France and Spain.

```
renfe %>%
  subset(type %in% c("AVE","AVE-TGV")) %>%
  group_by(type, dest) %>%
  summarise("average duration" = mean(duration),
            "std. dev" = sd(duration),
            "average price" = mean(price),
```

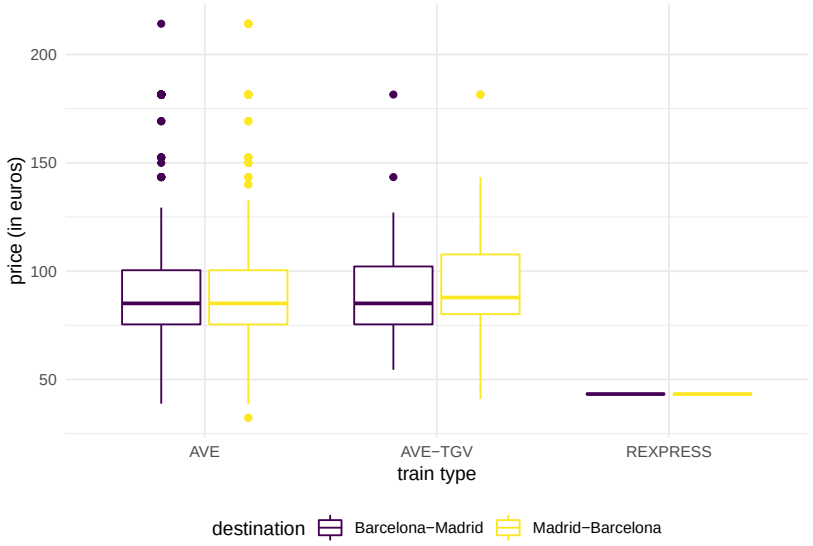

Figure 1.12: Boxplot of ticket price as a function of destination and train type.

```
            "std. dev" = sd(price))
```

```
##      type             dest average duration std. dev average price std. dev
## 1     AVE Barcelona-Madrid                171     15.9          87.4     19.8
## 2     AVE Madrid-Barcelona                170     16.6          88.2     20.8
## 3 AVE-TGV Barcelona-Madrid                175      0.0          87.0     16.8
## 4 AVE-TGV Madrid-Barcelona                179      0.0          90.6     20.2
```

The price of high speed trains are on average more than twice as expensive as regular trains. There is strong evidence of heterogeneity (standard deviation of 20 euros), which should raise scrutiny and suggests that high speed train tickets are dynamically priced. There is a single duration time for AVE-TGV tickets. We do not see meaningful differences in price depending on the type or the direction, but fares of ticket class availability may differ depending on whether the train is run in partnership with SNCF.

We have not looked at ticket fare and class, except for RegioExpress trains. Figure 1.14 shows large disparity in the variance of price according to fare: Promo fare takes many more distinct values than AdultoIda (duh) and Flexible fares. First class tickets (`Preferente`) are more expensive, but there are fewer observations falling in this group. Turista class is the least expensive for high-speed trains and the most popular. `TuristaPlus` offer an alternative to the latter with more comfort, whereas `TuristaSolo` gives access to individual seats.

Fare-wise Promo and PromoPlus give access to rebates that can go up to 70% and 65%, respectively. Promo tickets cannot be cancelled or exchanged, while both are possible with PromoPlus by paying a penalty amounting to 30-20% of the ticket price. Flexible fare ticket a sold at the same price as regular high-speed train tickets, but offer additional benefits (and no rebates!)

```
renfe %>% subset(fare  != "AdultoIda") %>%
ggplot(aes(y = price, x = class, col = fare)) +
  geom_boxplot() +
  labs(y = "price (in euros)",
       x = "class",
       color = "fare") +
  theme(legend.position = "bottom")
```

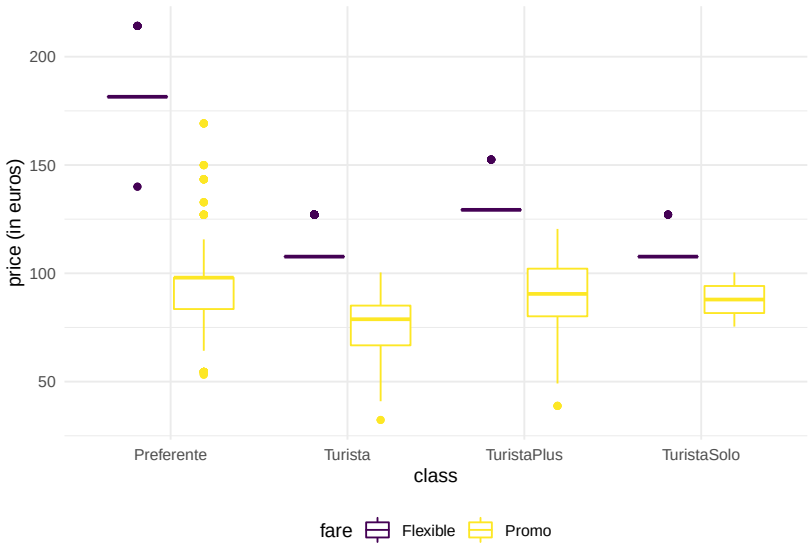

Figure 1.13: Boxplot of ticket price as a function of fare and class for high-speed Renfe trains.

```
ggplot(data = renfe, aes(x = price, y=..density.., fill = fare)) +
    geom_histogram(binwidth = 5) +
    labs(x = "price (in euros)", y = "density") +
    theme(legend.position = "bottom")
```

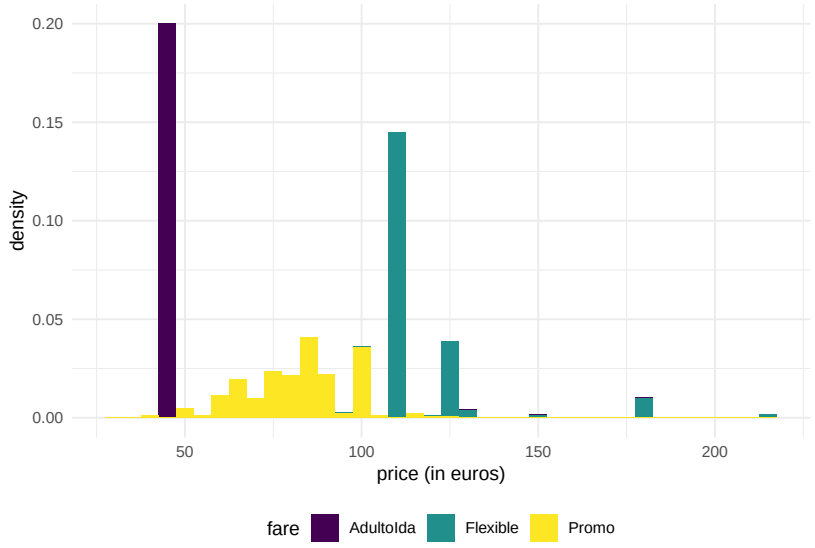

Figure 1.14: Histograms of ticket price as a function of fare for Renfe trains.

```
# Check the spread of Flexible tickets
renfe %>% subset(fare  == "Flexible") %>% count(price, class)
```

```
##   price       class    n
## 1   108     Turista 1050
## 2   108 TuristaSolo   67
## 3   127     Turista  285
## 4   127 TuristaSolo    9
## 5   129 TuristaPlus   31
## 6   140  Preferente    2
## 7   152 TuristaPlus   10
## 8   182  Preferente   78
## 9   214  Preferente   12
```

Note how Flexible tickets prices are spread: the boxplot is crushed and the interquartile range seems zero, even if some of the values are larger: this is indicative either constant price or of (too few) tickets in the category. We can find out which of these two possibities is most likely by counting the number of Flexible fare tickets for the different types.

Neither duration, nor type or destination explains why some Flexible tickets are more or less expensive than the average. Promo tickets, on the other hand, are cheaper on average and Preferente more expensive.

We can summarize our findings:

- more than 91% of trains are high-speed trains.
- travel time depends on the type of train: high-speed train take at most 3h20.
- duration records expected travel time: there are only 17 unique values, 13 of which are for AVE trains.
- the price of RegioExpress train ticket is fixed (43.25€); all such tickets are sold with AdultoIda fare and there only one class (Turista). 57% of these trains go from Barcelona to Madrid and travel time is 9h22 from Barcelona to Madrid, 9h04 in the other direction.
- `Turista` is the cheapest and most popular class. `Preferente` class tickets are more expensive and are less often sold. `TuristaPlus` offers more comfort and `TuristaSolo` let you get individual seats.
- according to the Renfe website, `Flexible` fare tickets "come with additional offers and passengers can exchange or cancel their tickets if they miss their train"; as a counterpart, these tickets are more expensive and most tickets have a fixed fare (a handful are cheaper or more expensive, but this price difference is unexplained).
- the distribution of `Promo` fare high-speed trains ticket prices are more or less symmetric, but `Flexible` tickets seem left-truncated (the minimum price for these tickets in the sample is 107.7€).
- it appears that tickets sold by Renfe (`Promo` fare) are dynamically priced: the latter can be up to 70% cheaper than regular high-speed train tickets when purchased through the official agency or Renfe's website. These tickets cannot be refunded or exchanged.
- there is no indication that prices depend on the direction of travel.

Chapter 2

Linear regression

A linear regression is a model for the conditional mean of a response variable Y as a function of p explanatory variables (also termed regressors or covariates),

$$\mathsf{E}(Y \mid \mathrm{X}) = \beta_0 + \beta_1 \mathrm{X}_1 + \cdots + \beta_p \mathrm{X}_p. \tag{2.1}$$

The mean of Y is conditional on the values of the observed covariate X; this amounts to treating them as non-random, known in advance.

In practice, any model is an approximation of reality. An error term is included to take into account the fact that no exact linear relationship links X and Y (otherwise this wouldn't be a statistical problem), or that measurements of Y are subject to error. The random error term ε will be the source of information for our inference, as it will quantify the goodness of fit of the model.

We can rewrite the linear model in terms of the error for a random sample of size n: denote by Y_i the value of the response for observation i, and X_{ij} the value of of the jth explanatory variable of observation i. The model is

$$Y_i = \beta_0 + \beta_1 \mathrm{X}_{i1} + ... + \beta_p \mathrm{X}_{ip} + \varepsilon_i, \qquad i = 1, ..., n, \tag{2.2}$$

where ε_i is the additive error term specific to observation i. While we may avoid making distributional assumption about ε_i, we nevertheless fix its expectation to zero to encode the fact we do not believe the model is systematically off,, so $\mathsf{E}(\varepsilon_i \mid X_i) = 0$ $(i = 1, ..., n)$.

One important remark is that the model is linear in the coefficients $\beta \in \mathbb{R}_{p+1}$, not in the explanatory variables! the latter are arbitrary and could be (nonlinear) functions of other explanatory variables, for example $\mathrm{X} = \log(\texttt{annees})$, $\mathrm{X} = \texttt{horsepower}^2$ or $\mathrm{X} = \mathsf{I}_{\texttt{man}} \cdot \mathsf{I}_{\texttt{full}}$. The mean of the response is specified as a linear combination of explanatory variables. This is at the core of the flexibility of the linear regression, which is used mainly for the following purposes:

1. Evaluate the effects of covariates X on the mean response of Y.
2. Quantify the influence of the explanatories X on the response and test for their significance.
3. Predict the response for new sets of explanatories X.

2.1 Introduction

Linear regression is the most famous and the most widely used statistical model around. The name may appear reductive, but many tests statistics (t-tests, ANOVA, Wilcoxon, Kruskal–Wallis) can be formulated using a linear regression, while models as diverse as trees, principal components and deep neural networks are just linear regression model in disguise. What changes under the hood between one fancy model to the next are the optimization method (e.g., ordinary least squares, constrained optimization or stochastic gradient descent) and the choice of variables entering the model (spline basis for nonparametric regression, indicator variable selected via a greedy search for trees, activation functions for neural networks).

This chapter explores the basics of linear regression, parameter interpretation and testing for coefficients and sub-models. Analysis of variance will be presented as special case of linear regression.

To make concepts and theoretical notions more concrete, we will use data from a study performed in a college in the United States. The goal of the administration who collected these information was to investigate potential gender inequality in the salary of faculty members. The data contains the following variables:

- `salary`: nine-month salary of professors during the 2008–2009 academic year (in thousands USD).
- `rank`: academic rank of the professor (`assistant`, `associate` or `full`).
- `field`: categorical variable for the field of expertise of the professor, one of `applied` or `theoretical`.
- `sex`: binary indicator for sex, either `man` or `woman`.
- `service`: number of years of service in the college.
- `annees`: number of years since PhD.

Before drafting a model, a quick look at the data is in due order. If salary increases with year, there is more heterogeneity in the salary of higher ranked professors: logically, assistant professors are either promoted or kicked out after at most 6 years according to the data. The limited number of years prevents large variability for their salaries.

Salary increases over years of service, but its variability also increases with rank. Note the much smaller number of women in the sample: this will impact our power to detect differences between sex. A contingency table of sex and academic rank can be useful to see if the proportion of women is the same in each rank: women represent 16% of assistant

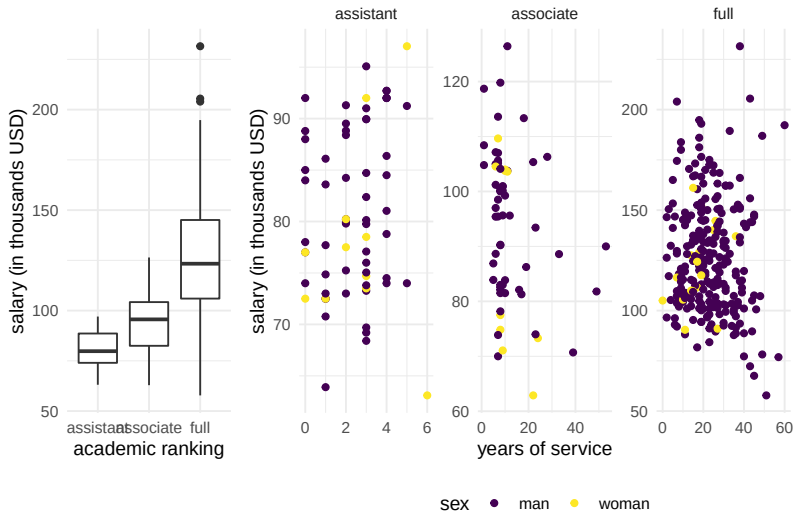

Figure 2.1: Exploratory data analysis of `college` data: salaries of professors as a function of the number of years of service and the academic ranking

Table 2.1: Contingency table of the number of prof in the college by sex and academic rank.

	assistant	associate	full
man	56	54	248
woman	11	10	18

professors and 16% of associate profs, but only 7% of full professors and these are better paid on average.

The simple linear regression model only includes a single explanatory variable and defines a straight line linking two variables X and Y by means of an equation of the form $y = a + bx$; Figure 2.2 shows the line passing through the scatterplot for years of service.

2.2 Ordinary least squares

The ordinary least square estimators $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ are the values that simultaneously minimize the Euclidean distance between the random observations Y_i and the fitted values

$$\widehat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{X}_{i1} + \cdots + \hat{\beta}_p \mathrm{X}_{ip}, \qquad i = 1, \ldots, n.$$

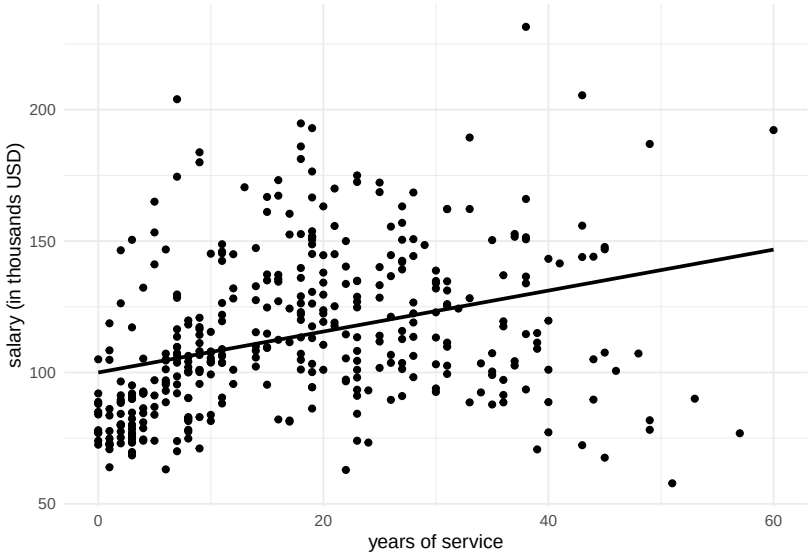

Figure 2.2: Simple linear regression model for the salary of professors as a function of the number of years of service; the line is the solution of the least squares problem.

In other words, the least square estimators are the solution of the convex optimization problem

$$\widehat{\beta} = \min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} (Y_i - \widehat{Y_i})^2 = \min_{\beta} \|Y - \mathrm{X}\beta\|^2$$

This system of equations has an explicit solution which is better expressed using matrix notation: this amounts to expressing equation (2.2) with one observation per line.

Consider the matrices

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \ \mathrm{X} = \begin{pmatrix} \mathrm{X}_{11} & \mathrm{X}_{12} & \cdots & \mathrm{X}_{1p} \\ \mathrm{X}_{21} & \mathrm{X}_{22} & \cdots & \mathrm{X}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{X}_{n1} & \mathrm{X}_{n2} & \cdots & \mathrm{X}_{np} \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

The model in compact form is

$$Y = \mathrm{X}\beta + \varepsilon.$$

The ordinary least squares estimator solves the unconstrained optimization problem

$$\widehat{\beta} = \min_{\beta \in \mathbb{R}^{p+1}} (y - \mathrm{X}\beta)^\top (y - \mathrm{X}\beta).$$

and a proof is provided in the Appendix. If the $n \times (p+1)$ matrix X is full-rank, we obtain a unique solution to the optimization problem,

$$\widehat{\beta} = (\mathrm{X}^\top \mathrm{X})^{-1} \mathrm{X}^\top Y. \tag{2.3}$$

What does the solution to the least squares problem represent in two dimensions? The estimator is the one minimizing the sum of squared residuals: the ith ordinary residual $e_i = y_i - \widehat{y}_i$ is the vertical distance between a point y_i and the fitted value $\widehat{y}_i$ on the line; the blue segments on Figure 2.3 represent the individual vectors of residuals.

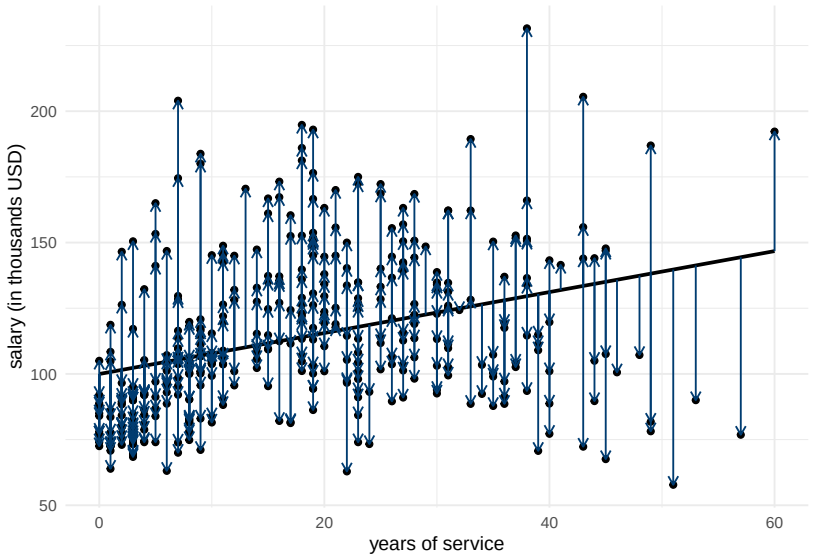

Figure 2.3: Illustration of ordinary residuals added to the regression line (blue vectors).

Remark (Geometry of least squares). If we consider the n observations as a (column) vector, the term $\mathrm{X}\widehat{\beta}$ is the projection of the response vector y on the linear span generated by the columns of X, $\mathcal{S}_{\mathrm{X}}$. The ordinary residuals are thus orthogonal to $\mathcal{S}_{\mathrm{X}}$ and to the fitted values, meaning $e^\top \widehat{y} = 0$. A direct consequence of this fact is that the linear correlation between e and $\widehat{y}$ is zero; we will use this property to build graphical diagnostics.

Remark (Complexity of ordinary least squares). This is an aside: in machine learning, you will often encounter linear models fitted using a (stochastic) gradient descent algorithm. Unless your sample size n or the number of covariates p is significant (think at the Google scale), an approximate should not be prefered to the exact solution! From a numerical perspective, obtaining the least square estimates requires inverting a $(p+1) \times (p+1)$ matrix $\mathrm{X}^\top \mathrm{X}$, which is the most costly operation. In general, direct inversion should be avoided because it is not the most numerically stable way of obtaining the solution. R uses the QR decomposition,

which has a complexity of $O(np^2)$. Another more stable alternative, which has the same complexity but is a bit more costly is use of a singular value decomposition.

Any good software will calculate ordinary least square estimates for you. Keep in mind that there is an explicit and unique solution provided your design matrix X doesn't contain collinear columns. If you have more than one explanatory variable, the fitted values lie on a hyperplan (which is hard to represent graphically). Mastering the language and technical term (fitted values, ordinary residuals, etc.) is necessary for the continuation.

2.3 Interpretation of the model parameters

What do the β parameters of the linear model represent? In the simple case presented in Figure 2.2, the equation of the line is $\widehat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{X}_1$, β_0 is the intercept (the mean value of Y when $\mathrm{X}_1 = 0$) and β_1 is the slope, i.e., the average increase of Y when X_1 increases by one unit.

In some cases, the intercept is meaningless because the value $\mathrm{X}_1 = 0$ is impossible (e.g., X_1 represents the height of a human). In the same vein, there may be no observation in a neighborhood of $\mathrm{X}_1 = 0$, even if this value is plausible, in which case the intercept is an extrapolation.

If the columns of X are arbitrary, it is customary to include an intercept: this amounts to including 1_n as column of the design matrix X. Because the residuals are orthogonal to the columns of X, their mean is zero, since $n^{-1} 1_n^\top e = \bar{e} = 0$. In general, we could also obtain mean zero residuals by including a set of vectors in X that span 1_n.

In our example, the equation of the fitted line of Figure 2.2 is

$$\widehat{\texttt{salary}} = 99.975 + 0.78\texttt{service}.$$

The average salary of a new professor would then be 99974.653 dollars, whereas the average annual increase for each additional year of service is 779.569 dollars.

If the response variable Y should be continuous (for the least square criterion to be meaningful), we place no such restriction on the explanatories. The case of dummies in particular is common: these variables are encoded using binary indicators (0/1). Consider for example the sex of the professors in the study:

$$\texttt{sex} = \begin{cases} 0, & \text{for men,} \\ 1, & \text{for women.} \end{cases}$$

The equation of the simple linear model that includes the binary variable `sex` is $\texttt{salary} = \beta_0 + \beta_1 \texttt{sex} + \varepsilon$. Let μ_0 denote the average salary of mean and μ_1 that of women. The intercept

β_0 can be interpreted as usual: it is the average salary when $\texttt{sex} = 0$, meaning that $\beta_0 = \mu_0$. We can write the equation for the conditional expectation for each sex,

$$\mathsf{E}(\texttt{salary} \mid \texttt{sex}) = \begin{cases} \beta_0, & \texttt{sex} = 0 \text{ (men)}, \\ \beta_0 + \beta_1 & \texttt{sex} = 1 \text{ (women)}. \end{cases}$$

A linear model that only contains a binary variable X as regressor amounts to specifying a different mean for each of two groups: the average of women is $\mathsf{E}(\texttt{salary} \mid \texttt{sex} = 1) = \beta_0 + \beta_1 = \mu_1$ and $\beta_1 = \mu_1 - \mu_0$ represents the difference between the average salary of men and women. The least square estimator $\widehat{\beta}_0$ is the sample mean of men and $\widehat{\beta}_1$ is the difference of the sample mean of women and men. The parametrization of the linear model with β_0 and β_1 is in terms of contrasts and is particularly useful if we want to test for mean difference between the groups, as this amounts to testing $\mathcal{H}_0 : \beta_1 = 0$. If we wanted our model to directly output the sample means, we would need to replace the design matrix $\mathrm{X} = [1_n, \texttt{sex}]$ by $[1_n - \texttt{sex}, \texttt{sex}]$. The fitted model would be the same because they span the same 2D subspace, but this is not recommended because software treat cases without intercept differently and it can lead to unexpected behavior (more on this latter).

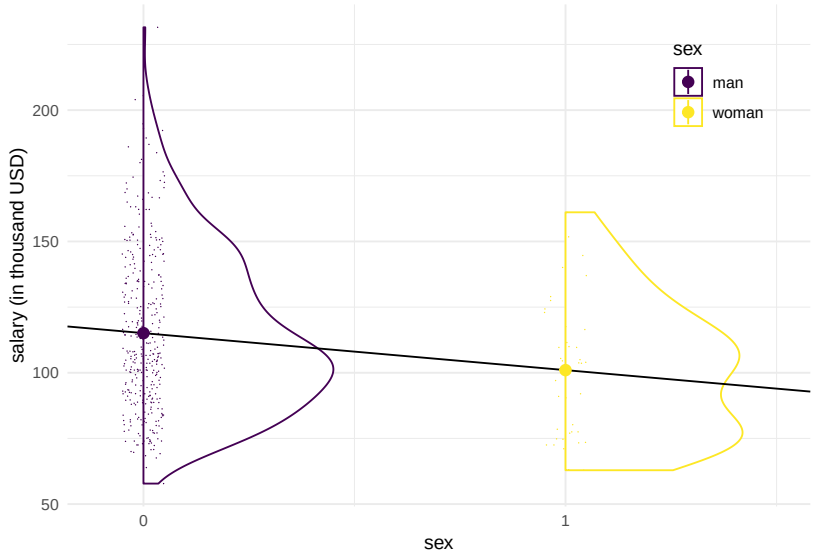

Figure 2.4: Simple linear model for the `college` data using the binary variable sex as regressor: even if the equation defines a line, only its values in 0/1 are realistic.

If we fit the model with sex only to the `college` data, we find that the average salary of men is $\widehat{\beta}_0 = 1.151 \times 10^5$ USD and the mean difference estimate of the salary between women and men is $\widehat{\beta}_1 = 14088.009$ dollars. Since the estimate is negative, this means women are paid less. Bear in mind that the model is not adequate for determining if there are gender

inequalities in the salary distribution: 2.2 shows that the number of years of service and the academic rank strongly impact wages, yet the distribution of men and women is not the same within each rank.

Even if the linear model defines a line, the latter is only meaningful when evaluated at 0 or 1; Figure 2.4 shows it in addition to sample observations (jittered) and a density estimate for each sex. The colored dot represents the mean, showing that the line does indeed pass through the mean of each group.

A binary indicator is a categorical variable with two levels, so we could extend our reasoning and fit a model with a categorical explanatory variable with k levels. To do this, we add $k-1$ indicator variables plus the intercept: if we want to model a different mean for each of the k groups, it is logical to only include k parameters in the mean model. We will choose, as we did with sex, a reference category or baseline whose average will be encoded by the intercept β_0. The other parameters $\beta_1, \ldots, \beta_{k-1}$ are contrasts relative to the baseline. The college data includes the ordinal variable `rank`, which has three levels (assistant, associate and full). We thus need two binary variables, $\mathrm{X}_1 = \mathsf{I}(\texttt{rank} = \texttt{associate})$ and $\mathrm{X}_2 = \mathsf{I}(\texttt{rank} = \texttt{full})$; the ith element of the vector X_1 is one for an associate professor and zero otherwise. The linear model is

$$\texttt{salary} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \varepsilon,$$

and the conditional expectation of salary

$$\mathsf{E}(\texttt{salary} \mid \texttt{rank}) = \begin{cases} \beta_0, & \texttt{rank} = \texttt{assistant}, \\ \beta_0 + \beta_1 & \texttt{rank} = \texttt{associate}, \\ \beta_0 + \beta_2 & \texttt{rank} = \texttt{full}, \end{cases}$$

Thus β_1 (respectively β_2) are the difference between the average salary of associate (respectively full) professors and assistant professors. The choice of the baseline category is arbitrary and all choices yield the same model: only the interpretation changes from one parametrization to the next. For an ordinal variable, it is recommended to choose the smallest or the largest category to ease comparisons.

The models we have fitted so far are not adequate because they ignore variables that are necessarily to correctly explain variations in salaries: Figure 2.1 show for example that rank is critical for explaining the salary variations in the college. We should thus fit a model that include those simultaneously to investigate the gender gap (which consists of differences that are unexplained by other factors). Before doing this, we come back to the interpretation of the parameters in the multiple linear regression setting.

Consider the model $Y = \beta_0 + \beta_1 \mathrm{X}_1 + \cdots + \beta_p \mathrm{X}_p + \varepsilon$. The intercept β_0 represents the mean value of Y when all of the covariates are set to zero,

$$\beta_0 = \mathsf{E}(Y \mid \mathrm{X}_1 = 0, \mathrm{X}_2 = 0, \ldots, \mathrm{X}_p = 0).$$

For categorical variables, this yields the baseline, whereas we fix the continous variables to zero: again, this may be nonsensical depending on the study. The coefficient β_j ($j \geq 1$) can be interpreted as the mean increase of the response Y when X_j increases by one unit, all other things being equal (ceteris paribus); e.g.,

$$\begin{aligned}\beta_1 &= \mathsf{E}(Y \mid \mathrm{X}_1 = x_1 + 1, \mathrm{X}_2 = x_2, \ldots, \mathrm{X}_p = x_p) \\ &\qquad - \mathsf{E}(Y \mid \mathrm{X}_1 = x_1, \mathrm{X}_2 = x_2, \ldots, \mathrm{X}_p = x_p) \\ &= \big\{\beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2 + \cdots + \beta_p \mathrm{X}_p\big\} \\ &\qquad - \big\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p \mathrm{X}_p\big\}\end{aligned}$$

It is not always possible to fix the value of an explanatory if multiple columns of X contains functions/transformations of it. For example, if we included a polynomial of order k for some variable X,

$$Y = \beta_0 + \beta_1 \mathrm{X} + \beta_2 \mathrm{X}^2 + \ldots + \beta_k \mathrm{X}^k + \varepsilon.$$

If we include a term of order k, X^k, we must always include the lower order terms $1, \mathrm{X}, \ldots, \mathrm{X}^{k-1}$ to make sure the resulting model is interpretable (otherwise, it amounts to a particular class of polynomials with some zero coefficients). Interpreting nonlinear effects (even polynomials, for which $k \leq 3$ in practice), is complicated because the effect of an increase of one unit of X depends of the value of the latter.

Example 2.1 (Auto data). We consider a linear regression model for the fuel autonomy of cars as a function of the power of their motor (measured in horsepower) from the `auto` dataset. The postulated model,

$$\texttt{mpg}_i = \beta_0 + \beta_1 \texttt{horsepower}_i + \beta_2 \texttt{horsepower}_i^2 + \varepsilon_i,$$

includes a quadratic term. Figure 2.5 shows the scatterplot with the fitted regression line, above which the line for the simple linear regression for horsepower is added.

It appears graphically that the quadratic model fits better than the simple linear alternative: we will assess this hypothesis formally later. For the degree two polynomial, Figure 2.5 show that fuel autonomy decreases rapidly when power increases between 50 to 100, then more slow until 189.35 hp. After that, the model postulates that autonomy increases again as evidenced by the scatterplot, but beware of extrapolating (weird things can happen beyond the range of the data, as exemplified by Hassett's cubic model for the number of daily cases of Covid19 in the USA).

The representation in Figure 2.5 may seem counter-intuitive given that we fit a linear model, but it is a 2D projection of 3D coordinates for the equation $\beta_0 + \beta_1 x - y + \beta_2 z = 0$, where $x = \texttt{horsepower}$, $z = \texttt{horsepower}^2$ and $y = \texttt{mpg}$. Physics and common sense force $z = x^2$, and so the fitted values lie on a curve in a 2D subspace of the fitted plan, as shown in grey in the 3D Figure 2.6.

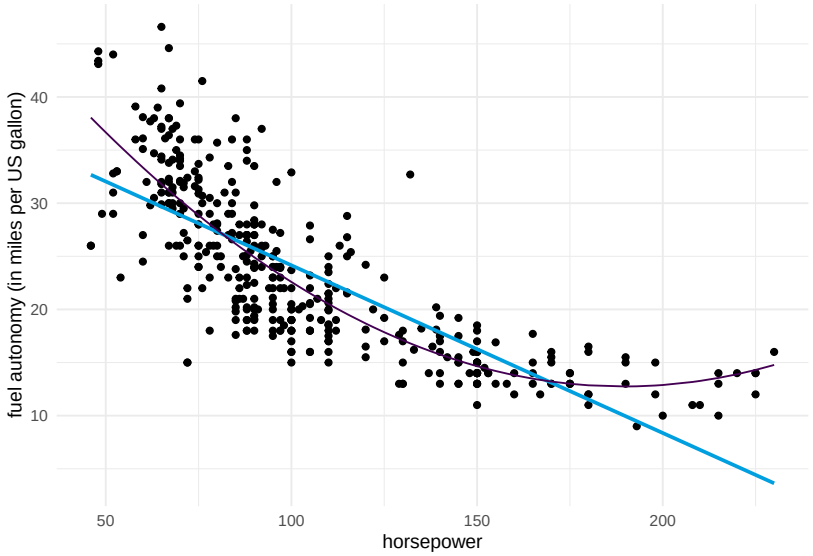

Figure 2.5: Linear regression models for the fuel autonomy of cars as a function of motor power

Remark (There are better alternatives to polynomials for modelling nonlinear effects). Generally speaking, one uses flexible basis vectors (splines) rather than polynomials for smoothing when the relation between the response Y and an explanatory variable X is nonlinear; these models involve many covariates and it is customary to add a penalty term to control for overfitting and wiggliness. A better (physical) understanding of the system, or a theoretical model can also guide the choice of functions to use.

The coefficient β_j in Eq. (2.1) represents the marginal contribution of X_j when all the other covariates are included in the model and which is not explained by them. This can be represented graphically by projecting Y and X_j in the orthogonal complement of X_{-j} (the matrix containing all but the jth column X_j). The added-variable plot is a graphical tool showing this projection: the residuals from the linear regression of Y onto $\mathcal{S}(\mathrm{X}_{-j})$ are mapped to the y-axis, whereas the residuals from the linear regression of X_j as a function of X_{-j} are shown on the x-axis. The regression line passes through $(0, 0)$ and its slope is $\widehat{\beta}_j$. This graphical diagnostic is useful for detecting collinearity and the impact of outliers.

Example 2.2 (Wage inequality in an American college). We consider a multiple regression model for the `college` data that includes sex, academic rank, field of study and the number of years of service as regressors.

If we multiply `salary` by a thousand to get the resulting estimates in US dollars, the postu-

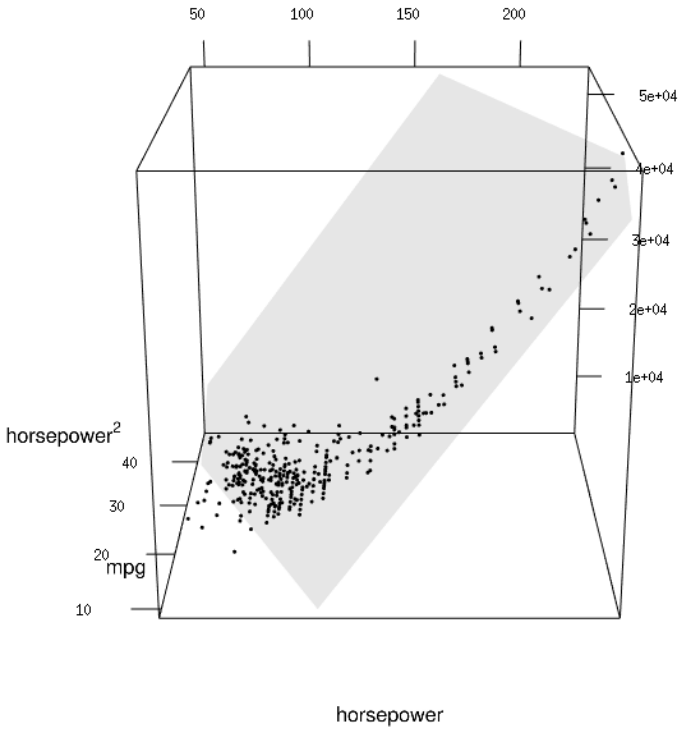

Figure 2.6: 3D graphical representation of the linear regression model for the *extttauto* data.

Table 2.2: Estimated coefficients of the linear model for the `college` (in USD, rounded to the nearest dollar).

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
86596	-4771	-13473	14560	49160	-89

lated model is

$$\begin{aligned}\texttt{salary} \times 1000 = \beta_0 &+ \beta_1 \texttt{sex}_{\texttt{woman}} + \beta_2 \texttt{field}_{\texttt{theoretical}} \\ &+ \beta_3 \texttt{rank}_{\texttt{associate}} + \beta_4 \texttt{rank}_{\texttt{full}} + \beta_5 \texttt{service} + \varepsilon.\end{aligned}$$

The interpretation of the coefficients is as follows:

- The estimated intercept is $\widehat{\beta}_0 = 86596$ dollars; it corresponds to the mean salary of men assistant professors who just started the job and works in an applied domain.

- everything else being equal (same field, academic rank, and number of years of service), the estimated salary difference between a woman and is estimated at $\widehat{\beta}_1 = -4771$ dollars.
- ceteris paribus, the salary difference between a professor working in a theoretical field and one working in an applied field is β_2 dollars: our estimate of this difference is -13473 dollars, meaning applied pays more than theoretical.
- ceteris paribus, the estimated mean salary difference between associate and assistant professors is $\widehat{\beta}_3 = 14560$ dollars.
- ceteris paribus, the estimated mean salary difference between full and assistant professors is $\widehat{\beta}_4 = 49160$ dollars.
- au sein d'un même échelon, chaque année supplémentaire de service mène à une augmentation de salary annuelle moyenne de $\widehat{\beta}_5 = -89$ dollars.

All other factors taken into account, women get paid less than men. It remains to see if this difference is statistically significant. Perhaps more surprising, the model estimates that salary decreases with every additional year of service: this seems counterintuitive when looking at Figure 2.2, which showed a steady increas of salary over the years. However, this graphical representation is misleading because Figure 2.1 showed that academic ranking mattered the most. Once all the other factors are accounted for, years of service serves to explain the salary of full professors who have been working unusual amounts of time and who gain less than the average full professor, as shown by the added-variable plot of Figure 2.7.

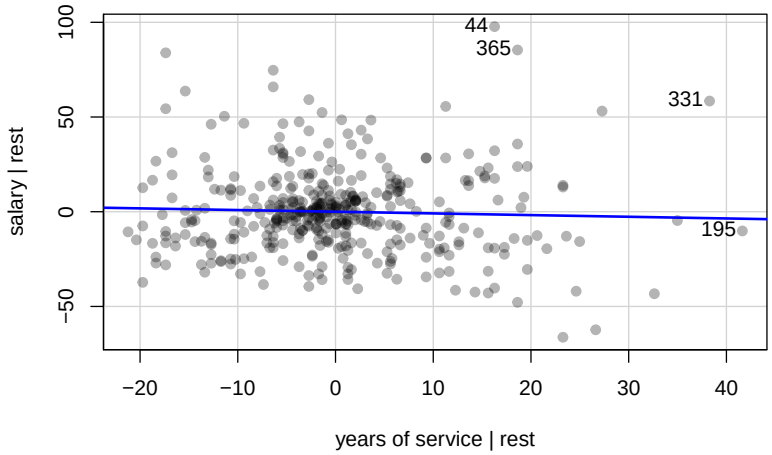

Figure 2.7: Added-variable plot for years of service in the linear regression model of the `college` data.

Details about implementation of linear models using R are provided in the Appendix.

2.4 Tests for parameters of the linear model

Chapter 3

Likelihood-based inference

The goal of this chapter is to familiarize you with likelihood-based inference.

The starting point of likelihood-based inference is a statistical model: we postulate that (a function of) the data has been generated from a probability distribution with p-dimensional parameter vector θ. The purpose of the analyst is to estimate these unknown parameters on the basis of a sample and make inference about them.

The likelihood $L(\theta)$ is a function of θ that gives the probability (or density) of observing a sample under a postulated distribution, treating the observations as fixed. In most settings we consider, observations are independent and so the joint probability of the sample values is the product of the probability of the individual observations[1]: for $y_1, \ldots, y_n$ assuming Y_i $(i = 1, \ldots, n)$ follows a distribution whose mass function or density is $f(y; \theta)$, this is just

$$L(\theta; y) = \prod_{i=1}^{n} f(y_i; \theta) = f(y_1; \theta) \times \cdots \times f(y_n; \theta).$$

From a pure optimization perspective, the likelihood is a particular choice of objective function that reflects the probability of the observed outcome. One shouldn't however maximize directly the likelihood, since computing the product of a lot of potentially small numbers is subject to numerical overflow and is unstable (for discrete distributions, the mass function gives probabilities that are by definition between zero and one). Instead, one should work with the log-likelihood function, $\ell(\theta) = \log\{L(\theta)\}$. Since logarithm is a strictly increasing function, maximizing the natural logarithm (denoted $\log \equiv \ln$ throughout) of the likelihood leads to the same solution. Another reason why working with the log-likelihood is preferable

[1]If this seems foreign, think about repeated coin tosses with Bernoulli distribution of unknown parameter p and convince yourself that the trials are independent, so the probability of obtaining two consecutive heads is 0.25 for a fair coin.

is because product over n likelihood contributions becomes a sum and this facilitates numerical and analytical derivations of the maximum likelihood estimators (the log of a product is equal to the sum of logs, i.e., $\log(ab) = \log(a) + \log(b)$ for $a, b > 0$.)

The maximum likelihood estimator $\widehat{\theta}$ is the value of θ that maximizes the likelihood, i.e., the value under which the random sample is the most likely to be generated. The scientific reasoning behind this is: "whatever we observe, we have expected to observe" so we choose between competing models the one that makes the most sense.

Several properties of maximum likelihood estimator makes it appealing for inference.

- The maximum likelihood estimator is consistent, i.e., it converges to the correct value as the sample size increase (asymptotically unbiased).
- The maximum likelihood estimator is invariant to reparametrizations
- Under regularity conditions, the maximum likelihood estimator is asymptotically normal, so we can obtain the null distribution of classes of hypothesis tests and derive confidence intervals based on $\widehat{\theta}$.
- The maximum likelihood estimator is efficient, meaning it has the smallest asymptotic mean squared error (or the smallest asymptotic variance).

The score function $U(\theta; y) = \partial\ell(\theta; y)/\partial\theta$ is the gradient of the log-likelihood function and, under regularity conditions, the maximum likelihood estimator solves $U(\theta; Y) = 0_p$. This property can be used to derive gradient-based algorithms for optimization and for verifying that the solution found is a global maximum.

Remark. While least squares admit a closed-form solution, the maximum of the log-likelihood is generally found numerically by solving the score equation. The algorithms used in most software are reliable and efficient for regression models we consider in this course. However, for more complex models, like generalized linear mixed models, the convergence of optimization algorithms is oftentimes problematic and scrutiny is warranted.

The observed information matrix is the hessian $j(\theta; y) = -\partial^2\ell(\theta; y)/\partial\theta\partial\theta^\top$ evaluated at the maximum likelihood estimate $\widehat{\theta}$. Under regularity conditions, the Fisher information matrix is

$$i(\theta) = \mathsf{E}\left\{U(\theta; Y)U(\theta; Y)^\top\right\} = -\mathsf{E}\left\{j(\theta; Y)\right\}$$

The Fisher (or expected) and observed information matrices encodes the curvature of the log-likelihood and provides information about the variability of $\widehat{\theta}$.

The properties of the log-likelihood are particularly convenient for inference because they provide omnibus testing procedures that have a known asymptotic distribution. The starting point for the distributional theory surrounding likelihood-based statistics is the asymptotic normality of the score $U(\theta) \dot{\sim} \mathsf{No}(0, i(\theta))$, which follows from a central limit theorem. The variance of $U(\theta_0)$ is exactly $i(\theta_0)$, while that of $\widehat{\theta}$ is approximately $i(\theta_0)^{-1}$ under the null

hypothesis $\mathcal{H}_0$. This result is particularly useful: we often use the inverse of the observed information as estimate of the covariance matrix of the maximum likelihood estimator. To obtain the standard errors of $\widehat{\theta}$, one simple computes the square root of the diagonal elements of the inverse of the observed information, i.e., $[\text{diag}\{j^{-1}(\widehat{\theta})\}]^{1/2}$.

Example 3.1 (Exponential model for waiting times of the Montreal metro). Consider the waiting time Y between consecutive subways arriving at Station Édouard-Montpetit on the blue line in Montreal during rush hour. We postulate that these waiting times follow an exponential distribution with scale θ, denoted $Y \sim \mathsf{E}(\theta)$. The purpose of statistical inference is to use the information from a random sample of size n to estimate the unknown parameter θ. The density of Y evaluated at y, $f(y;\theta) = \theta^{-1}\exp(-y/\theta)$, encodes the probability of the observed waiting time for a given parameter value and, if the records are independent, the probability of observing $y_1, \ldots, y_n$ is the product of probabilities of individual events. The likelihood is thus

$$L(\theta; y) = \prod_{i=1}^{n} f(y;\theta) = \prod_{i=1}^{n} \theta^{-1}\exp(-y_i/\theta),$$
$$\ell(\theta; y) = -n\log(\theta) - \theta^{-1}\sum_{i=1}^{n} y_i$$

To find the maximum of the function, we differentiate the log-likelihood $\ell(\theta; y)$ and set the gradient to zero,

$$\frac{\partial \ell(\theta; y)}{\partial \theta} = -\frac{n}{\theta} + \theta^{-2}\sum_{i=1}^{n} y_i = 0.$$

Solving for θ gives $\widehat{\theta} = \overline{y}$, so the maximum likelihood estimator is the sample mean $\overline{Y}$. The observed information is $j(\theta) = -n\theta^{-2} + 2\theta^{-3}n\overline{y}$ and $i(\theta) = \mathsf{E}\{j(\theta)\} = n\theta^{-2}$.

For the sample of waiting time in the subway, the maximum likelihood estimate is $\widehat{\theta} = 3.058$, the observed information is $j(\widehat{\theta}) = i(\widehat{\theta}) = 2.139$ and the standard error of $\widehat{\theta}$ is $j(\widehat{\theta})^{-1/2} = 0.684$.

Example 3.2 (Normal samples and ordinary least squares). Suppose we have an independent normal sample of size n with mean μ and variance σ^2, where $Y_i \sim \mathsf{No}(\mu, \sigma^2)$ are independent. Recall that the density of the normal distribution is

$$f(y; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}.$$

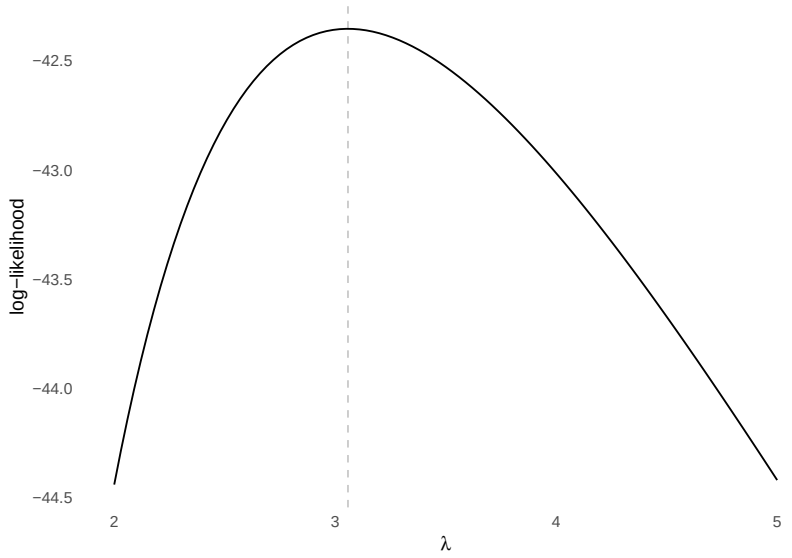

Figure 3.1: Log-likelihood for a sample of size 20 of waiting times (in minutes)

For an n-sample y, the likelihood is

$$\begin{aligned} L(\mu, \sigma^2; y) =& \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\} \\ =& (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}. \end{aligned}$$

and the log-likelihood is

$$\ell(\mu, \sigma^2; y) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

One can show that the maximum likelihood estimators for the two parameters are

$$\hat{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

The fact that the estimator of the theoretical mean μ is the sample mean is fairly intuitive and one can show the estimator is unbiased for μ. The (unbiased) sample variance estimator,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{Y}_i - \overline{Y})^2$$

Since $\widehat{\sigma}^2 = (n-1)/nS^2$, it follows that the maximum likelihood estimator of σ^2 is biased, but both estimators are consistent and will thus get arbitrarily close to the true value σ^2 for n sufficiently large.

The case of normally distributed data is intimately related to linear regression and ordinary least squares: assuming normality of the errors, the least square estimators of β coincide with the maximum likelihood estimator of β.

Recall the linear regression model,

$$Y_i = \beta_0 + \beta_1 \mathrm{X}_{i1} + \beta_2 \mathrm{X}_{i2} + ... + \beta_p \mathrm{X}_{ip} + \varepsilon_i, \qquad (i = 1, ..., n),$$

where the errors $\varepsilon_i \sim \mathsf{No}(0, \sigma^2)$. The linear model has $p+2$ parameters (β and σ^2) and the log-likelihood is

$$\ell(\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left\{y - \mathrm{X}\beta)^\top (y - \mathrm{X}\beta)\right\}^2.$$

Maximizing the log-likelihood with respect to β is equivalent to minimizing the sum of squared errors $\|Y - \widehat{Y}\|^2$. Since this objective function is the same as that of least squares, it follows that the least-square estimator $\widehat{\beta}$ for the mean parameters is the maximum likelihood estimator for normal errors with common variance σ^2, regardless of the value of the latter. The maximum likelihood estimator $\widehat{\sigma}^2$ is thus

$$\widehat{\sigma}^2 = \max_{\sigma^2} \ell(\widehat{\beta}, \sigma^2).$$

The log-likelihood, excluding constant terms that don't depend on σ^2, is

$$\ell(\widehat{\beta}, \sigma^2) \propto -\frac{1}{2}\left\{n\log\sigma^2 + \frac{1}{\sigma^2}(y - \mathrm{X}\widehat{\beta})^\top (y - \mathrm{X}\widehat{\beta})\right\}.$$

Differentiating each term with respect to σ^2 and setting the gradient equal to zero yields the maximum likelihood estimator

$$\widehat{\sigma}^2 = \frac{1}{n}(Y - \mathrm{X}\widehat{\beta})^\top (Y - \mathrm{X}\widehat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{\mathsf{SS}_e}{n};$$

where SS_e is the sum of squared residuals. The usual unbiased estimator of σ^2 calculated by software is $S^2 = \mathsf{SS}_e/(n-p-1)$, where the denominator is the sample size n minus the number of mean parameters β, $p+1$.

Oftentimes, we wish to compare two models: the model implied by the null hypothesis, which is a restriction or simpler version of the full model. Models are said to be nested if we can obtain one from the other by imposing restrictions on the parameters.

We consider a null hypothesis $\mathcal{H}_0$ that imposes restrictions on the possible values of θ can take, relative to an unconstrained alternative $\mathcal{H}_1$. We need two nested models: a full model, and a reduced model that is a subset of the full model where we impose q restrictions. For example, the full model could be a regression model with four predictor variables and the reduced model could include only the first two predictor variables, which is equivalent to setting $\mathcal{H}_0 : \beta_3 = \beta_4 = 0$. The testing procedure involves fitting the two models and obtaining the maximum likelihood estimators of each of $\mathcal{H}_1$ and $\mathcal{H}_0$, respectively $\widehat{\theta}$ and $\widehat{\theta}_0$ for the parameters under $\mathcal{H}_0$. The null hypothesis $\mathcal{H}_0$ tested is: 'the reduced model is an adequate simplification of the full model' and the likelihood provides three main classes of statistics for testing this hypothesis: these are

- likelihood ratio tests statistics, denoted R, which measure the drop in log-likelihood (vertical distance) from $\ell(\widehat{\theta})$ and $\ell(\widehat{\theta}_0)$.
- Wald tests statistics, denoted W, which consider the standardized horizontal distance between $\widehat{\theta}$ and $\widehat{\theta}_0$.
- score tests statistics, denoted S, which looks at the scaled gradient of ℓ, evaluated only at $\widehat{\theta}_0$ (derivative of ℓ).

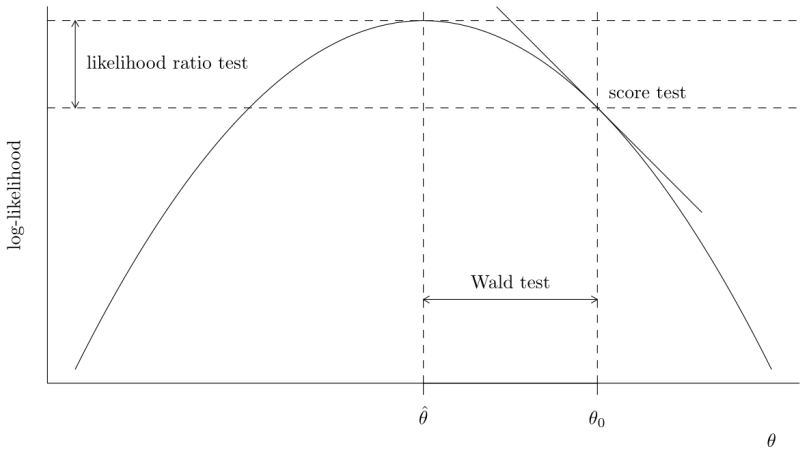

Figure 3.2: Log-likelihood curve; the three likelihood-based tests, namely Wald, likelihood ratio and score tests, use different information about the function.

The three main classes of statistics for testing a simple null hypothesis $\mathcal{H}_0 : \theta = \theta_0$ against the alternative $\mathcal{H}_a : \theta \neq \theta_0$ are the likelihood ratio, the score and the Wald statistics, defined

respectively as

$$\begin{aligned} R &= 2\left\{\ell(\widehat{\theta}) - \ell(\theta_0)\right\}, \\ S &= U^\top(\theta_0) i^{-1}(\theta_0) U(\theta_0), \\ W &= (\widehat{\theta} - \theta_0)^\top i(\theta_0)(\widehat{\theta} - \theta_0), \end{aligned}$$

where $\widehat{\theta}$ is the maximum likelihood estimate under the alternative and θ_0 is the null value of the parameter vector. Asymptotically, all the test statistics are equivalent (in the sense that they lead to the same conclusions about $\mathcal{H}_0$). If $\mathcal{H}_0$ is true, the three test statistics follow asymptotically a χ^2_q distribution under a null hypothesis $\mathcal{H}_0$, where the degrees of freedom q are the number of restrictions.

For scalar θ with $q = 1$, signed versions of these statistics exist, e.g.,

$$W(\theta_0) = (\widehat{\theta} - \theta_0)/\mathsf{se}(\widehat{\theta}) \dot{\sim} \mathsf{No}(0, 1)$$

for the Wald statistic or the directed likelihood root

$$R(\theta_0) = \text{sign}(\widehat{\theta} - \theta) \left[2\left\{\ell(\widehat{\theta}) - \ell(\theta)\right\}\right]^{1/2} \dot{\sim} \mathsf{No}(0, 1).$$

The likelihood ratio test statistic is normally the most powerful of the three likelihood tests. The score statistic S only requires calculation of the score and information under $\mathcal{H}_0$ (because by definition $U(\widehat{\theta}) = 0$), so it can be useful in problems where calculations of the maximum likelihood estimator under the alternative is costly or impossible.

The Wald statistic W is the most widely encountered statistic and two-sided 95% confidence intervals for a single parameter θ are of the form

$$\widehat{\theta} \pm q_{1-\alpha/2}\text{se}(\widehat{\theta}),$$

where $q_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution; for a 95% confidence interval, the 0.975 quantile of the normal distribution is 1.96. The Wald-based confidence intervals are by construction symmetric: they may include implausible values (e.g., negative values for variances). The Wald-based confidence intervals are not parametrization invariant: if we want intervals for a nonlinear continuous function $h(\theta)$, then in general $\mathsf{CI}_W\{h(\theta)\} \neq h\{\mathsf{CI}_W(\theta)\}$. So, if $[0.1, 0.9]$ is a 95% Wald-based confidence interval for $\widehat{\beta}$, the Wald-based confidence interval for $\exp(\widehat{\beta})$ is not $[\exp(0.1), \exp(0.9)]$.

These confidence intervals can be contrasted with the (better) ones derived using the likelihood ratio test: these are found through a numerical search to find the limits of

$$\theta : 2\{\ell(\widehat{\theta}) - \ell(\theta)\} \leq \chi^2_1(1-\alpha),$$

where $\chi^2_1(1-\alpha)$ is the $(1-\alpha)$ quantile of the χ^2_1 distribution. If θ is multidimensional, confidence intervals for θ_i are derived using the profile likelihood. Likelihood ratio-based confidence intervals are parametrization invariant, so $\mathsf{CI}_R\{h(\theta)\} = h\{\mathsf{CI}_R(\theta)\}$. Because the likelihood is zero if a parameter value falls outside the range of possible values for the parameter, the intervals only include plausible values of θ. In general, the intervals are asymmetric and have better coverage properties.

The F-tests covered in linear regression are equivalent to the likelihood ratio test for the linear model, but software reports the p-values calculated using an F distribution, which is the exact null under the assumption of normally distributed errors with constant variance. The chi-square distribution and the F-distribution are equivalent for n large.

The Wald statistic and the likelihood ratio test for testing individual coefficients of the mean model β_j in a linear regression are equivalent: they give the same p-value.

3.1 Profile likelihood

Consider a parametric model with log-likelihood function $\ell(\theta)$ whose p-dimensional parameter vector $\theta = (\psi, \lambda)$ that can be decomposed into a q-dimensional parameter of interest ψ and a $(p-q)$-dimensional nuisance vector λ.

For example, we may be interested in obtaining confidence intervals for a single β_j in a logistic regression, treating the other parameters β_{-j} as nuisance

In these cases, we can consider the profile likelihood ℓ_{p}, a function of ψ alone, which is obtained by maximizing the likelihood pointwise at each fixed value ψ_0 over the nuisance vector φ_{ψ_0},

$$\ell_{\mathsf{p}}(\psi) = \max_{\varphi} \ell(\psi, \varphi) = \ell(\psi, \widehat{\varphi}_{\psi}).$$

Figure 3.3 shows a fictional log-likelihood contour plot with the resulting profile curve (in black), where the log-likelihood value is mapped to colors. If one thinks of these contours lines as those of a topographic map, the profile likelihood corresponds in this case to walking along the ridge of both mountains along the ψ direction, with the right panel showing the elevation gain/loss.

The maximum profile likelihood estimator behaves like a regular likelihood for most quantities of interest and we can derive test statistics and confidence intervals in the usual way. One famous example of profile likelihood is the Cox proportional hazard covered in Chapter 7.

Example 3.3 (Box–Cox transformation). Sometimes, the assumption of normality of the error

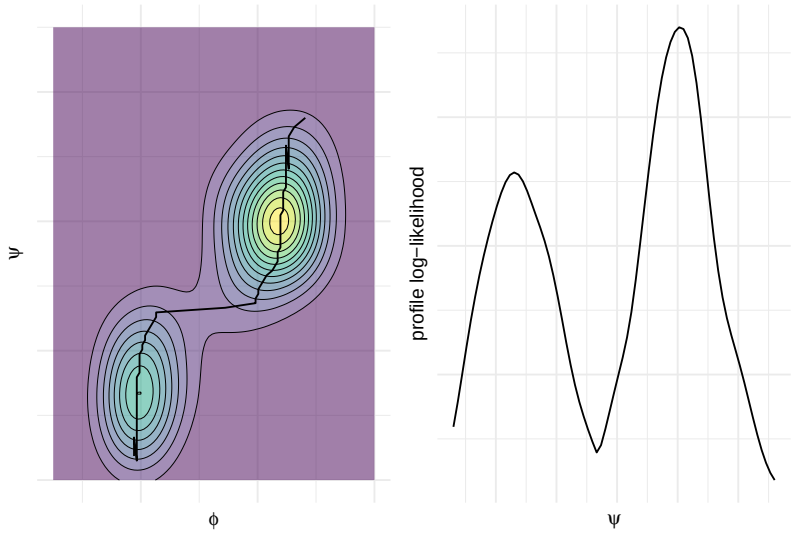

Figure 3.3: Two-dimensional log-likelihood surface with a parameter of interest ψ and a nuisance parameter φ; the contour plot shows area of higher likelihood, and the black line is the profile log-likelihood, also shown as a function of ψ on the right panel.

doesn't hold. If the data are strictly positive, one can consider a Box–Cox transformation,

$$y_i(\lambda) = \begin{cases} (y^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \log(y), & \lambda = 0. \end{cases}$$

If we assume that $y(\lambda) \sim \mathsf{No}(\mathrm{X}\beta, \sigma^2 \mathrm{I}_n)$, then the likelihood of y is

$$L(\lambda, \beta, \sigma; y, \mathrm{X}) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\{y(\lambda) - \mathrm{X}\beta\}^\top \{y(\lambda) - \mathrm{X}\beta\}\right] J(\lambda, y),$$

where J denotes the Jacobian of the Box–Cox transformation, $\prod_{i=1}^n y_i^{\lambda-1}$. For each given value of λ, the maximum likelihood estimator is that of the usual regression model, with y replaced by $y(\lambda)$, namely $\widehat{\beta}_\lambda = (\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}^\top y(\lambda)$ and $\hat{\sigma}^2_\lambda = n^{-1}\{y(\lambda) - \mathrm{X}\widehat{\beta}_\lambda\}^\top\{y(\lambda) - \mathrm{X}\widehat{\beta}_\lambda\}$.

The profile log-likelihood is

$$\ell_{\mathsf{p}}(\lambda) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2_\lambda) - \frac{n}{2} + (\lambda - 1)\sum_{i=1}^n \log(y_i)$$

The maximum profile likelihood estimator is the value λ minimizes the sum of squared residuals from the linear model with $y(\lambda)$ as response.

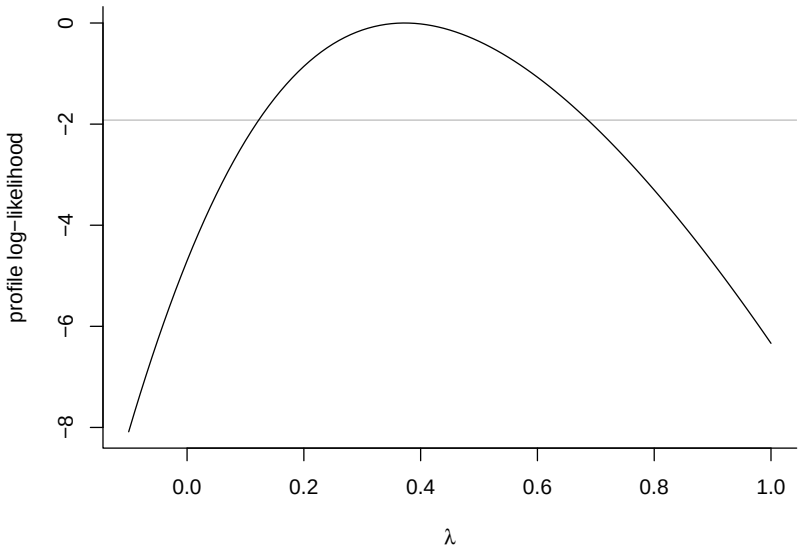

Figure 3.4: Profile log-likelihood for the Box–Cox transformation for the waiting time data

Figure 3.4 shows the profile log-likelihood for the linear model with an intercept-only, rescaled to be zero at the maximum. The function shows that a value of approximately 0.37 would provide residuals that are closer to normally distributed. The 95% profile-likelihood based confidence interval is given by the two values of λ, $(0.12, 0.13)$, at which the curve intersects the horizontal grey line drawn at $-\chi^2_1/2$. The Box–Cox is not a panacea and should be reserved to cases where the transformation reduces heteroscedasticity (unequal variance) or creates a linear relation between explanatories and response: theory provides a cogent explanation of the data (e.g., the Cobb–Douglas production function used in economics can be linearized by taking a log-transformation). Rather than an ad hoc choice of transformation, one could choose a log transformation if the value 0 is included within the 95% confidence interval since this improves interpretability.

3.2 Likelihood-based tools for model comparison

The likelihood can also serve as building block for model comparison: the larger $\ell(\widehat{\theta})$, the better the fit. However, the likelihood doesn't account for model complexity in the sense that more complex models with more parameters lead to higher likelihood. This is not a problem for comparison of nested models using the likelihood ratio test because we look only at relative improvement in fit. There is a danger of overfitting if we only consider the likelihood of a model.

AIC and BIC are information criteria measuring how well the model fits the data, while

penalizing models with more parameters,

$$\begin{aligned}\mathsf{AIC} &= -2\ell(\hat{\theta}) + 2k \\ \mathsf{BIC} &= -2\ell(\hat{\theta}) + k\log(n),\end{aligned}$$

where k is the number of parameters in the model. The smaller the value of AIC (or of BIC), the better the model fit.

Note that information criteria do not constitute formal hypothesis tests on the parameters, but they can be used to compare non nested-models, even these estimates are particularly noisy. If we want to compare likelihood from different probability models, we need to make sure they include normalizing constant. The BIC is more stringent than AIC, as its penalty increases with the sample size, so it selects models with fewer parameters. The BIC is consistent, meaning that it will pick the true correct model from an ensemble of models with probability one as $n \to \infty$. In practice, this is of little interest if one assumes that all models are approximation of reality (it is unlikely that the true model is included in the ones we consider). AIC often selects overly complicated models in large samples, whereas BIC is sometimes too conservative in that it chooses models that are overly simple.

Chapter 4

Generalized linear models

Chapter 5

Correlated and longitudinal data

Chapter 6

Linear mixed models

Chapter 7

Survival analysis

Chapter 8

Basic concepts

8.1 Population and samples

Statistics is the science of uncertainty quantification: of paramount importance is the notion of randomness. Generally, we will seek to estimate characteristics of a population using only a sample (a sub-group of the population of smaller size).

The population of interest is a collection of individuals which the study targets. For example, the Labour Force Survey (LFS) is a monthly study conducted by Statistics Canada, who define the target population as "all members of the selected household who are 15 years old and older, whether they work or not." Asking every Canadian meeting this definition would be costly and the process would be long: the characteristic of interest (employment) is also a snapshot in time and can vary when the person leaves a job, enters the job market or become unemployed.

In general, we therefore consider only samples to gather the information we seek to obtain. The purpose of statistical inference is to draw conclusions about the population, but using only a share of the latter and accounting for sources of variability. George Gallup made this great analogy between sample and population:

> One spoonful can reflect the taste of the whole pot, if the soup is well-stirred

A sample is a random sub-group of individuals drawn from the population. Creation of sampling plans is a complex subject and semester-long sampling courses would be required to evens scratch the surface of the topic. Even if we won't be collecting data, keep in mind the following information: for a sample to be good, it must be representative of the population under study. Selection bias must be avoided, notably samples of friends or of people sharing opinions.

Because the individuals are selected at random to be part of the sample, the measurement of the characteristic of interest will also be random and change from one sample to the next. However, larger samples of the same quality carry more information and our estimator will be more precise. Sample size is not guarantee of quality, as the following example demonstrates.

Example 8.1. The Literary Digest surveyed 10 millions people by mail to know voting preferences for the 1936 USA Presidential Election. A sizeable share, 2.4 millions answered, giving Alf Landon (57%) over incumbent President Franklin D. Roosevelt (43%). The latter nevertheless won in a landslide election with 62% of votes cast, a 19% forecast error. Biased sampling and differential non-response are mostly responsible for the error:) the sampling frame was built using "phone number directories, drivers' registrations, club memberships, etc.' ', all of which skewed the sample towards rich upper class white people more susceptible to vote for the GOP.

In contrast, Gallup correctly predicted the outcome by polling (only) 50K inhabitants. Read the full story here.

8.2 Random variable

Suppose we wish to describe the behaviour of a stochastic phenomenon. To this effect, one should enumerate the set of possible values taken by the variable of interest and their probability: this is what is encoded in the distribution. We will distinguish between two cases: discrete and continuous variables. Random variables are denoted using capital letters: for example $Y \sim \mathsf{No}(\mu, \sigma^2)$ indicates that Y follows a normal distribution with parameters μ and σ^2, which represent respectively the expectation and variance of Y.

The (cumulative) distribution function $F(y)$ gives the cumulative probability that an event doesn't exceed a given numerical value y, $F(y) = \mathsf{Pr}(Y \leq y)$.

If Y is discrete, then it has atoms of non-zero probability and the mass function $f(y) = \mathsf{Pr}(Y = y)$ gives the probability of each outcome y. In the continuous case, no numerical value has non-zero probability and so we consider intervals instead: the density function gives the probability of Y falling in a set B, via $\mathsf{Pr}(Y \in B) = \int_B f(y)\mathrm{d}y$. It follows that the distribution function of a continuous random variable is simply $F(y) = \int_{-\infty}^{y} f(x)\mathrm{d}x$.

8.3 Moments

One of the first topics covered in introductory statistics is descriptive statistics such as the mean and standard deviation. These are estimators of (centered) moments, which characterise a random varaible. In the case of the standard normal distribution, the expectation and variance fully characterize the distribution.

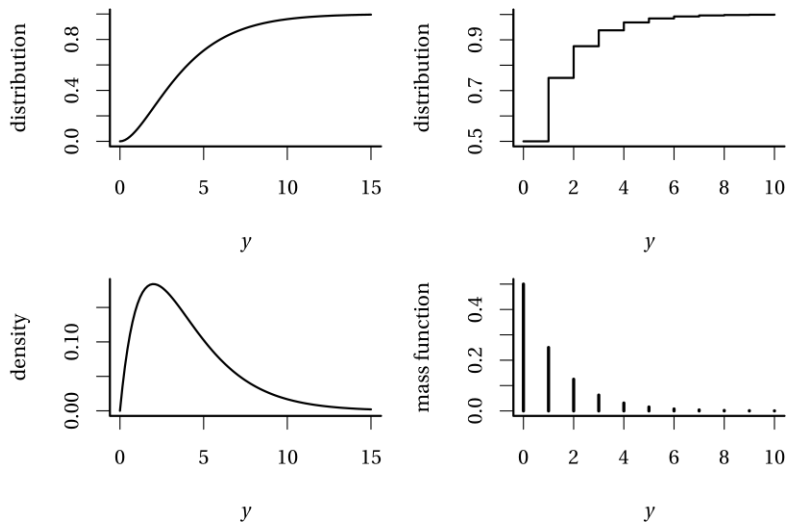

Figure 8.1: (Cumulative) distribution functions (top) and density/mass functions (bottom) of continuous (left) and discrete (right) random variables.

Let Y be a random variable with density (or mass) function $f(x)$. This function is non-negative and satisfies $\int_{\mathbb{R}} f(x)\mathrm{d}x = 1$: the integral over a set B gives the probability of Y falling inside $B \in \mathbb{R}$.

The expectation of a continuous random variable Y is

$$\mathsf{E}(Y) = \int_{\mathbb{R}} xf(x)\mathrm{d}x.$$

Expectation is the "theoretical mean" in the discrete case, we set rather $\mu = \mathsf{E}(Y) = \sum_{x\in\mathcal{X}} x\mathsf{Pr}(X = x)$, where $\mathcal{X}$ stands for the support of Y, which is the set of numerical values at which the probability of Y is non-zero. More generally, we can look at the expectation of a function $g(x)$ for Y, which is nothing but the integral (or sum in the discrete case) of $g(x)$ weighted by the density or mass function of $f(x)$. In the same fashion, provided the integral is finite, the variance is

$$\mathsf{Va}(Y) = \mathsf{E}\{Y - \mathsf{E}(Y)\}^2 \equiv \int_{\mathbb{R}} (x-\mu)^2 f(x)\mathrm{d}x.$$

An estimator $\hat{\theta}$ for a parameter θ is unbiased if its bias $\mathsf{bias}(\hat{\theta}) = \mathsf{E}(\hat{\theta}) - \theta$ is zero. The unbiased estimator of the mean of Y is $\overline{Y}_n = n^{-1}\sum_{i=1}^n Y_i$ and that of the variance is $S_n = (n-1)^{-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$. While unbiasedness of an estimator is a desirable property, but may not be optimal. There may even be cases where no unbiased estimator exists for a parameter!

Often, rather, we seek to balance bias and variance: recall that an estimator is a function of random variables and thus it is itself random: even if it is unbiased, the numerical value obtained will vary from one sample to the next. We often seek an estimator that minimises the mean squared error,

$$\mathsf{MSE}(\hat{\theta}) = \mathsf{E}\{(\hat{\theta} - \theta)^2\} = \mathsf{Va}(\hat{\theta}) + \{\mathsf{E}(\hat{\theta})\}^2.$$

The mean squared error is an objective function consisting of the sum of the squared bias and the variance.

An alternative to this criterion is to optimize a function such as the likelihood of the sample: the resulting estimator is termed maximum likelihood estimator. These estimator are asymptotically efficient, in the sense that they have the lowest mean squared error of all estimators for large samples. Other properties of maximum likelihood estimators make them attractive default choice for estimation.

The likelihood of a sample is the joint density of the n observations, which requires a distribution to be considered. Many such distributions describe simple physical phenomena and can be described using a few parameters: we only cover the most frequently encountered.

Example 8.2 (Bernoulli distribution). We consider a binary event such as coin toss (heads/tails). In general, the two events are associated with success/failure. By convention, failures are denoted by zeros and successes by ones, the probability of success being π so $\mathsf{Pr}(Y = 1) = \pi$ and $\mathsf{Pr}(Y = 0) = 1 - \pi$ (complementary event). The mass function of the Bernoulli distribution is thus

$$\mathsf{Pr}(Y = y) = \pi^y(1 - \pi)^{1-y}, \quad y = 0, 1.$$

A rapid calculation shows that $\mathsf{E}(Y) = \pi$ and $\mathsf{Va}(Y) = \pi(1 - \pi)$. Many research questions have binary responses, for example:

- did a potential client respond favourably to a promotional offer?
- is the client satisfied with service provided post-purchase?
- will a company go bankrupt in the next three years?
- did a study participant successfully complete a task?

Example 8.3 (Binomial distribution). If the data five the sum of independent Bernoulli events, the number of sucessess Y out of m trials is binomial, denoted $\mathsf{Bin}(m, \pi)$; the mass function of the binomial distribution is

$$\mathsf{Pr}(Y = y) = \binom{m}{y}\pi^y(1 - \pi)^{1-y}, \quad y = 0, 1.$$

The likelihood of a sample from a binomial distribution is (up to a normalizing constant that doesn't depend on π) the same as that of $m$independent Bernoulli trials. The expectation of the binomial random variable is $\mathsf{E}(Y) = m\pi$ and its variance $\mathsf{Va}(Y) = m\pi(1 - \pi)$.

As examples, we could consider the number of successful candidates out of m who passed their driving license test or the number of customers out of m total which spent more than 10$ in a store.

More generally, we can also consider count variables whose realizations are integer-valued, for examples the number of

- insurance claims made by a policyholder over a year,
- purchases made by a client over a month on a website,
- tasks completed by a study participant in a given time frame.

Example 8.4 (Geometric distribution). The geometric distribution is a model describing the number of Bernoulli trials with probability of success π required to obtain a first success. The mass function of $Y \sim \mathsf{Geo}(\pi)$ is

$$\mathsf{Pr}(Y = y) = \pi(1 - \pi)^{y-1}, \quad y = 1, 2, ...$$

For example, we could model the numbers of visits for a house on sale before the first offer is made using a geometric distribution.

Example 8.5 (Poisson distribution). If the probability of success π of a Bernoulli event is small in the sense that $m\pi \to \lambda$ wen the number of trials m increases, then the number of success followss approximately a Poisson distribution with mass function

$$\mathsf{Pr}(Y = y) = \frac{\exp(-\lambda)\lambda^y}{\Gamma(y + 1)}, \quad y = 0, 1, 2, ...$$

where $\Gamma(\cdot)$ denotes the gamma function. The parameter λ of the Poisson distribution is both the expectation and the variance of the distribution, meaning $\mathsf{E}(Y) = \mathsf{Va}(Y) = \lambda$.

Example 8.6 (Negative binomial distribution). The negative binomial distribution arises as a natural generalization of the geometric distribution if we consider the number of Bernoulli trials with probability of success π until we obtain m success. Let Y denote the number of failures: the order of success and failure doesn't matter, but for the latest trial which is a success. The mass function is thus

$$\mathsf{Pr}(Y = y) = \binom{m - 1 + y}{y} \pi^m (1 - \pi)^y.$$

The negative binomial distribution also appears as the unconditional distribution of a two-stage hierarchical gamma-Poisson model, in which the mean of the Poisson distribution is

random and follows a gamma distribution. In notation, this is $Y \mid \Lambda = \lambda \sim \mathsf{Po}(\lambda)$ and Λ follows a gamma distribution with shape r and scale θ, whose density is

$$f(x) = \theta^{-r} x^{r-1} \exp(-x/\theta)/\Gamma(r).$$

The unconditional number of success is then negative binomial.

In the context of generalized linear models, we will employ yet another parametrisation of the distribution, with the mass function

$$\mathsf{Pr}(Y = y) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} \left(\frac{r}{r+\mu}\right)^r \left(\frac{\mu}{r+\mu}\right)^y, y = 0, 1, \ldots, \mu, r > 0,$$

where Γ is the gamma function and the parameter $r > 0$ is not anymore integer valued. The expectation and variance of Y are $\mathsf{E}(Y) = \mu$ et $\mathsf{Va}(Y) = \mu + k\mu^2$, where $k = 1/r$. The variance of the negative binomial distribution is thus higher than its expectation, which justifies the use of the negative binomial distribution for modelling overdispersion.

Example 8.7 (Student-*t* distribution). If $X \sim \mathsf{No}(0, 1)$ independent of $Y \sim \chi^2_\nu$, then

$$T = \frac{X}{\sqrt{Y/\nu}}$$

follows a Student-t distribution with ν degrees of freedom, denoted St_ν. The density of T is

$$f(y; \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}} \left(1 + \frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

the distribution has polynomial tails, is symmetric around 0 and unimodal. Its bell-curve shape resembles that of the normal distribution, and, as $\nu \to \infty$, the Student distribution converges to a normal distribution. It has heavier tails than the normal distribution and only the first $\nu - 1$ moments of the distribution exist, so a Student distribution with $\nu = 2$ degrees of freedom has infinite variance.

8.3.1 Quantiles-quantiles plots

Models are (at best) an approximation of the true data generating mechanism and we will want to ensure that our assumptions are reasonable and the quality of the fit decent. Quantile-quantile plots are graphical goodness-of-fit diagnostics that are based on the following principle: if Y is a continuous random variable with distribution function F, then the mapping $F(Y) \sim \mathsf{U}(0, 1)$ yields uniform variables. Similarly, the quantile transform applied to a uniform variable provides a mean to simulating samples from F, viz. $F^{-1}(U)$. Consider then a random sample of size n from the uniform distribution ordered from smallest to largest,

with $U_{(1)} \leq \cdots \leq U_{(n)}$. One can show these ranks have marginally a Beta distribution, $U_{(k)} \sim \mathsf{Beta}(k, n+1-k)$ with expectation $k/(n+1)$.

In practice, we don't know F and, even if we did, one would need to estimate the parameters. We consider some estimator $\widehat{F}$ for the model and apply the inverse transform to an approximate uniform sample $\{i/(n+1)\}_{i=1}^n$. The quantile-quantile plot shos the data as a function of the (first moment) of the transformed order statistics:

- on the x-axis, the theoretical quantiles $\widehat{F}^{-1}\{\text{rank}(y_i)/(n+1)\}$
- on the y-axis, the empirical quantiles y_i

If the model is adequate, the ordered values should follow a straight line with unit slope passing through the origin. Whether points fall on a 45 degree line is difficult to judge by eye and so it is advisable to ease the interpretation to subtract the slope: the detrended plot is easier to interpret and was proposed by Tukey (but beware of the scale of the y-axis!). Figure 8.2 shows two representations of the same data using simulated samples from a standard normal distribution.

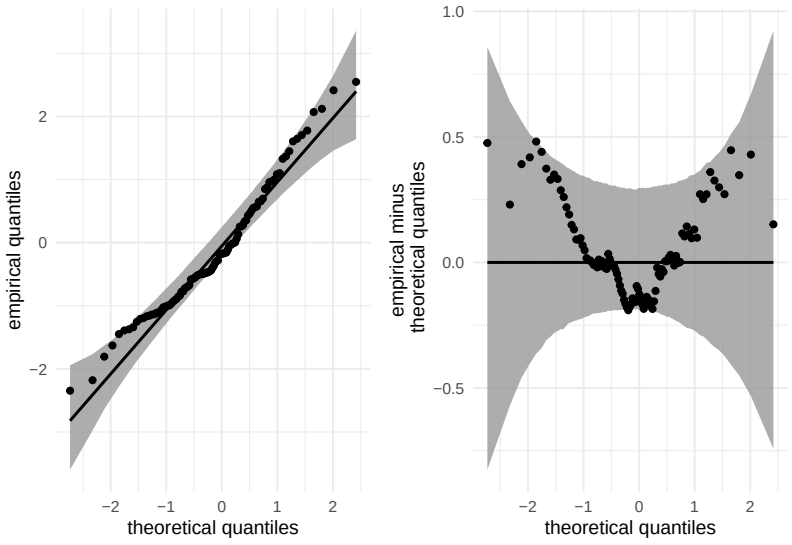

Figure 8.2: Normal quantile-quantile plot (left) and detrended version (Tukey's) of the same plot (right).

Even if we knew the true distribution of the data, the sample variability makes it very difficult to spot if deviations from the model are abnormal or compatible with the model. A simple point estimate with no uncertainty measure can lead to wrong conclusions. As such, we add approximate pointwise or simultaneous confidence intervals. The simplest way to do this is by simulation (using a parametric bootstrap), by repeating the following steps B times:

1. simulate a (bootstrap) sample $\{Y_i^{(b)}\}(i = 1, \dots, n)$ from $\widehat{F}$
2. re-estimate the parameters of F to obtain $\widehat{F}_{(b)}$
3. calculate and save the plotting positions $\widehat{F}_{(b)}^{-1}\{i/(n+1)\}$.

The result of this operation is an $n \times B$ matrix of simulated data. We obtain a symmetric $(1-\alpha)$ confidence interval by keeping the empirical quantile of order $\alpha/2$ and $1-\alpha/2$ from each row. The number B should be larger than 999, say, and be chosen so that $B/alpha$ is an integer.

For the pointwise interval, each order statistic from the sample is a statistic and so the probability of any single one falling outside the confidence interval is approximately α. However, order statistics are not independent (they are ordered), so its common to see neighboring points falling outside of their respective intervals. [It is also possible to use the bootstrap samples to derive an (approximate) simultaneous confidence intervals, in which we expected values to fall $100(1-\alpha)\%$ of the time inside the bands in repeated samples; see Section 4.4.3 of these course notes. The intervals shown in Figure 8.2 are pointwise and derived (magically) using a simple function. The uniform order statistics have larger variability as we move away from 0.5, but the uncertainty in the quantile-quantile plot largely depends on F.

Interpretation of quantile-quantile plots requires some experience: this post by Glen_b on StackOverflow nicely summarizes what can be detected (or not) from them.

8.4 Laws of large numbers

An estimator for a parameter θ is consistent if the value obtained as the sample size increases (to infinity) converges to the true value of θ. Mathematically speaking, this translates into convergence in probability, meaning $\hat{\theta} \xrightarrow{\mathsf{Pr}} \theta$. In common language, we say that the probability that $\hat{\theta}$ and θ differ becomes negligible as n gets large.

Consistency is the a minima requirement for an estimator: when we collect more information, we should approach the truth. The law of large number states that the sample mean of n (independent) observations with common mean μ, say $\overline{Y}_n$, converges to μ, denoted $\overline{Y}_n \to \mu$. Roughly speaking, our approximation becomes less variable and asymptotically unbiased as the sample size (and thus the quantity of information available for the parameter) increases. The law of large number is featured in Monte Carlo experiments: we can approximate the expectation of some (complicated) function $g(x)$ by simulating repeatedly independent draws from Y and calculating the sample mean $n^{-1}\sum_{i=1}^{n} g(Y_i)$.

If the law of large number tells us what happens in the limit (we get a single numerical value), the result doesn't contain information about the rate of convergence and the uncertainty at finite levels.

8.5 Central Limit Theorem

The central limit theorem is perhaps the flagship result of probability theory: for a random sample of size n with (independent) random variables whose expectation is μ and variance σ^2, then the sample mean converges to μ, but

- the estimator $\overline{Y}$ is centered around μ,
- the standard error is $\sigma/\sqrt{n}$; the rate of convergence is thus $\sqrt{n}$. For a sample of size 100, the standard error of the sample mean will be 10 times smaller than that of the underlying random variable.
- the sample mean, once properly scaled, follows approximately a normal distribution

Mathematically, the central limit theorem states $\sqrt{n}(\overline{Y} - \mu) \xrightarrow{\text{d}} \mathsf{No}(0, \sigma^2)$. If n is large (a rule of thumb is $n > 30$, but this depends on the underlying distribution of Y), then $\overline{Y} \dot{\sim} \mathsf{No}(\mu, \sigma^2/n)$.

How do we make sense of this result? Let us consider the mean travel time of high speed Spanish trains (AVE) between Madrid and Barcelona that are operated by Renfe.

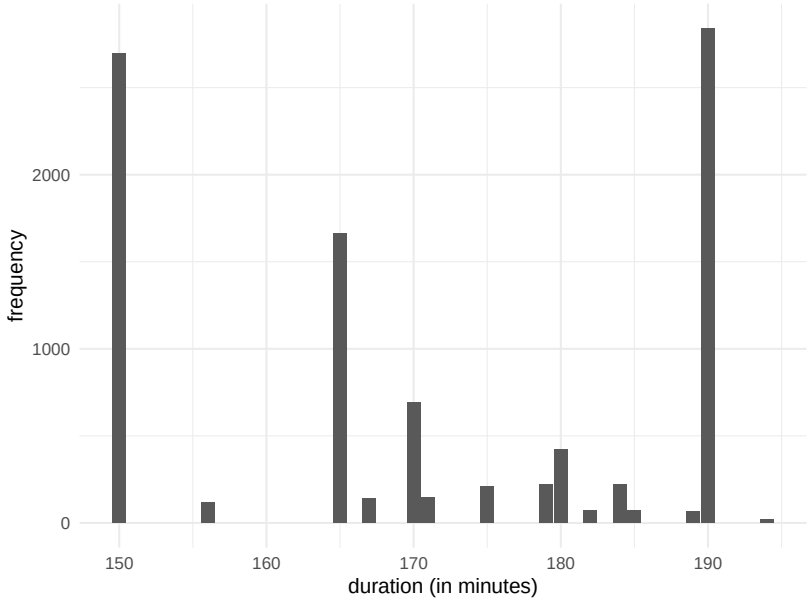

Figure 8.3: Empirical distribution of travel times of high speed trains.

Our exploratory data analysis showed previously that the duration is the one advertised on the ticket: there are only 15 unique travel time. Based on 9603 observations, we estimate the mean travel time to be 170 minutes and 41 seconds. Figure 8.3 shows the empirical distribution of the data.

Consider now samples of size $n = 10$, drawn repeatedly from the population: in the first sample, the sample mean is 170.9 minutes, whereas we get an estimate of 164.5 minutes in our second , 172.3 minutes in the third, etc.

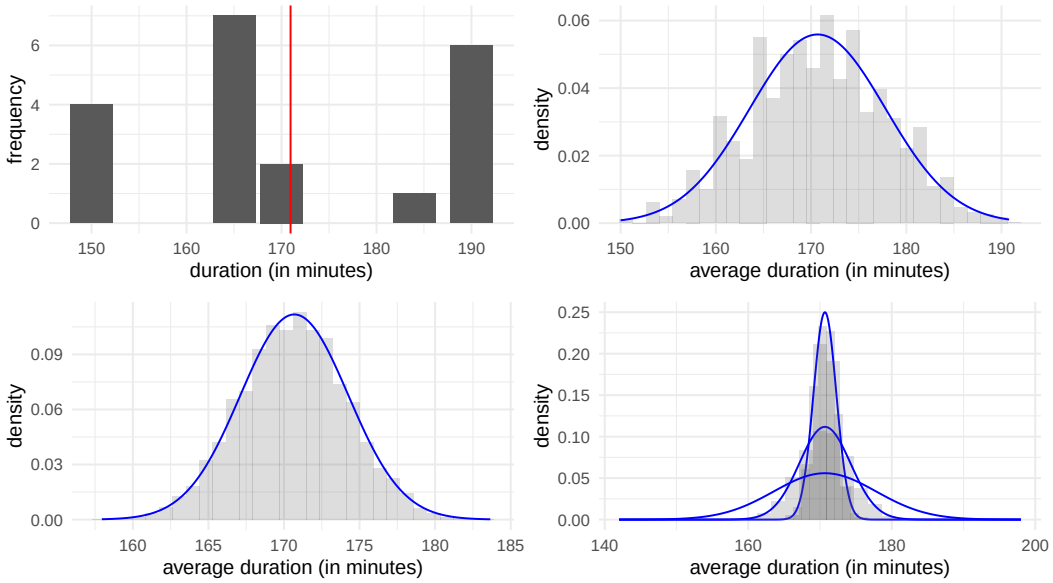

Figure 8.4: Graphical representation of the central limit theorem. The upper left panel shows a sample of 20 observations with its sample mean (vertical red). The three other panels show the histograms of the sample mean from repeated samples of size 5 (top right), 20 (bottom left) and 20, 50 and 100 overlaid, with the density approximation provided by the central limit theorem.

We draw $B = 1000$ different samples, each of size $n = 5$, from two millions records, and calculate the sample mean in each of them. The top right panel of 8.4 shows the result for $n = 5$, but also for $n = 20$ (bottom left). The last graph of Figure 8.4 shows the impact of the increase in sample size: whereas the normal approximation is okay-ish for $n = 5$, it is indistinguishable from the normal approximation for $n = 20$. As n increases and the sample size gets bigger, the quality of the approximation improves and the curve becomes more concentrated around the true mean. Even if the distribution of the travel time is discrete, the mean is approximately normal.

We considered a single distribution in the example, but you could play with other distributions and vary the sample size to see when the central limit theorem kicks in usng this applet.

The central limit theorem underlies why scaled test statistics which have sample mean zero and sample variance 1 have a standard null distribution in large sample: this is what guarantees the validity of our inference!

Chapter 9

Mathematical derivations

This section regroups optional derivations which are provided for the sake of completeness.

9.1 Derivation of the ordinary least squares estimator

Consider the optimization problem

$$\widehat{\beta} = \min_{\beta \in \mathbb{R}^{p+1}} (y - \mathrm{X}\beta)^\top (y - \mathrm{X}\beta).$$

We can compute the derivative of the right hand side with respect to β, set it to zero and solve for $\widehat{\beta}$,

$$\begin{aligned} 0_n &= \frac{\partial}{\partial \beta}(y - \mathrm{X}\beta)^\top (y - \mathrm{X}\beta) \\ &= \frac{\partial (y - \mathrm{X}\beta)}{\partial \beta} \frac{\partial (y - \mathrm{X}\beta)^\top (y - \mathrm{X}\beta)}{\partial (y - \mathrm{X}\beta)} \\ &= \mathrm{X}^\top (y - \mathrm{X}\beta) \end{aligned}$$

using the chain rule. Distributing the terms leads to the so-called normal equation

$$\mathrm{X}^\top \mathrm{X}\beta = \mathrm{X}^\top y.$$

If the $n \times p$ matrix X is full-rank, the quadratic form $\mathrm{X}^\top \mathrm{X}$ is invertible and we obtain the solution to the least square problems provided in Equation (2.3).

R

R is an object-oriented interpreted language. It differs from usual programming languages in that it is designed for interactive analyses.

You can find several introductions to R online. Have a look at the R manuals or better at contributed manuals. A nice official reference is An introduction to R. You may wish to look up the following chapters of the R language definition (Evaluation of expressions and part of the Objects chapter). Another good (small reference) is the cheatsheet Getting started in R.

.1 Basics of R

.1.1 Help

Help can be accessed via `help` or simply `?`, e.g., `help("Normal")`. See R page about help files.

.1.2 Basic commands

Basic R commands are fairly intuitive, especially if you want to use R as a calculator. Elementary functions such as `sum`, `min`, `max`, `sqrt`, `log`, `exp`, etc., are self-explanatory.

Some (unconventional) features of the language:

- R is case sensitive.
- Use `<-` for assignments to a variable, and `=` for matching arguments inside functions
- Indexing in R starts at 1, not 0.
- Most functions in R are vectorized and loops are typically inefficient.
- Integers are obtained by appending `L` to the number, so `2L` is an integer and `2` a double (`numerical`).

Besides integers and doubles, the common types are

- logical (`TRUE` and `FALSE`);
- null pointers (`NULL`), which can be assigned to arguments;

- missing values, namely `NA` or `NaN`. These can also be obtained a result of invalid mathematical operations such as `log(-2)`.

Beware! In R, invalid calls will often returns something rather than an error. It is therefore good practice to check that the output is sensical.

.1.3 Linear algebra in R

R is an object oriented language, and the basic elements in R are (column) vector. Below is a glossary with some useful commands for performing basic manipulation of vectors and matrix operations:

- `c` concatenates elements to form a vector
- `cbind` (`rbind`) binds column (row) vectors
- `matrix` and `vector` are constructors
- `diag` creates a diagonal matrix (by default with ones)
- `t` is the function for transpose
- `rep` creates a vector of duplicates, `seq` a sequence. For integers i, j with $i < j$, `i:j` generates the sequence $i, i+1, ..., j-1, j$.

Subsetting is fairly intuitive and general; you can use vectors, logical statements. For example, if `x` is a vector, then

- `x[2]` returns the second element
- `x[-2]` returns all but the second element
- `x[1:5]` returns the first five elements
- `x[(length(x) - 5):length(x)]` returns the last five elements
- `x[c(1, 2, 4)]` returns the first, second and fourth element
- `x[x > 3]` return any element greater than 3. Possibly an empty vector of length zero!
- `x[ x < -2 | x > 2]` multiple logical conditions.
- `which(x == max(x))` index of elements satisfying a logical condition.

For a matrix `x`, subsetting now involves dimensions: `[1,2]` returns the element in the first row, second column. `x[,2]` will return all of the rows, but only the second column. For lists, you can use `[[` for subsetting by index or the `$` sign by names.

.1.4 Packages

The great strength of R comes from its contributed libraries (called packages), which contain functions and datasets provided by third parties. Some of these (`base`, `stats`, `graphics`, etc.) are loaded by default whenever you open a session.

To install a package from CRAN, use `install.packages("package")`, replacing `package` by the package name. Once installed, packages can be loaded using `library(package)`; all the

functions in `package` will be available in the environment.

There are drawbacks to loading packages: if an object with the same name from another package is already present in your environment, it will be hidden. Use the double-colon operator `::` to access a single object from an installed package (`package::object`).

.1.5 Datasets

- datasets are typically stored inside a `data.frame`, a matrix-like object whose columns contain the variables and the rows the observation vectors.
- The columns can be of different types (`integer`, `double`, `logical`, `character`), but all the column vectors must be of the same length.
- Variable names can be displayed by using `names(faithful)`.
- Individual columns can be accessed using the column name using the `$` operator. For example, `faithful$eruptions` will return the first column of the `faithful` dataset.
- To load a dataset from an (installed) R package, use the command `data` with the name of the `package` as an argument (must be a string). The package `datasets` is loaded by default whenever you open R, so these are always in the search path.

The following functions can be useful to get a quick glimpse of the data:

- `summary` provides descriptive statistics for the variable.
- `str` provides the first few elements with each variable, along with the dimension
- `head` (`tail`) prints the first (last) n lines of the object to the console (default is $n = 6$).

We start by loading a dataset of the Old Faithful Geyser of Yellowstone National park and looking at its entries.

```
# Load Old faithful dataset
data(faithful, package = "datasets")
# Query the database for documentation
?faithful
# look at first entries
head(faithful)
```

```
##   eruptions waiting
## 1      3.60      79
## 2      1.80      54
## 3      3.33      74
## 4      2.28      62
## 5      4.53      85
```

```
## 6      2.88      55
```

```
str(faithful)
```

```
## 'data.frame': 272 obs. of  2 variables:
##  $ eruptions: num  3.6 1.8 3.33 2.28 4.53 ...
##  $ waiting  : num  79 54 74 62 85 55 88 85 51 85 ...
```

```
# What kind of object is faithful?
class(faithful)
```

```
## [1] "data.frame"
```

Other common classes of objects:

- `matrix`: an object with attributes `dim`, `ncol` and `nrow` in addition to `length`, which gives the total number of elements.
- `array`: a higher dimensional extension of `matrix` with arguments `dim` and `dimnames`.
- `list`: an unstructured class whose elements are accessed using double indexing `[[ ]]` and elements are typically accessed using `$` symbol with names. To delete an element from a list, assign `NULL` to it.
- `data.frame` is a special type of list where all the elements are vectors of potentially different type, but of the same length.

.1.6 Graphics

The `faithful` dataset consists of two variables: the regressand `waiting` and the regressor `eruptions`. One could postulate that the waiting time between eruptions will be smaller if the eruption time is small, since pressure needs to build up for the eruption to happen. We can look at the data to see if there is a linear relationship between the variables.

An image is worth a thousand words and in statistics, visualization is crucial. Scatterplots are produced using the function `plot`. You can control the graphic console options using `par` — see `?plot` and `?par` for a description of the basic and advanced options available.

Once `plot` has been called, you can add additional observations as points (lines) to the graph using `point` (`lines`) in place of `plot`. If you want to add a line (horizontal, vertical, or with known intercept and slope), use the function `abline`.

Other functions worth mentioning for simple graphics:

- `boxplot` creates a box-and-whiskers plot

- `hist` creates an histogram, either on frequency or probability scale (option `freq = FALSE`). `breaks` control the number of bins. `rug` adds lines below the graph indicating the value of the observations.
- `pairs` creates a matrix of scatterplots, akin to `plot` for data frame objects.

There are two options for basic graphics: the base graphics package and the package `ggplot2`. The latter is a more recent proposal that builds on a modular approach and is more easily customizable — I suggest you stick to either and `ggplot2` is a good option if you don't know R already, as the learning curve will be about the same. Even if the display from `ggplot2` is nicer, this is no excuse for not making proper graphics. Always label the axis and include measurement units!

.2 Linear models in R using the `lm` function

The function `lm` is the workshorse for fitting linear models in R. It takes as input a formula: suppose you have a data frame containing columns `x` (a regressor) and `y` (the regressand); you can then call `lm(y ~ x)` to fit the linear model $y = \beta_0 + \beta_1 x + \varepsilon$. The explanatory variable `y` is on the left hand side, while the right hand side should contain the predictors, separated by a `+` sign if there are more than one. If you provide the data frame name using `data`, then the shorthand `y ~ .` fits all the columns of the data frame (but `y`) as regressors.

If you include categorical variables, make sure they are transformed to factors. Normally, strings are cast to factor (unless you specify `stringsAsFactors = FALSE`) upon import of the data, but the danger here lies with variables that are encoded using integers (sex, revenue class, level of education, marital status, etc.) It is okay if we keep binary variables as is if they are encoded using 0/1, but it is often better to cast them to factor to get more meaningful labels given the lack of obvious ordering.

By default, the baseline level for a factor is based on the alphabetical order, while SAS uses the first value it encounters. Once the variables are cast to factor, `summary` will print the counts for each respective categories; these could be likewise be obtained using `table`.

To fit higher order polynomials or transformations, use the `I` function to tell R to interpret the input "as is". Thus, `lm(y~x+I(x^2))`, would fit a linear model with design matrix $(1_n, \mathrm{x}^\top, \mathrm{x}^2)^\top$. A constant is automatically included in the regression, but can be removed by writing `-1` or `+0` on the right hand side of the formula (but don't do that!). The `lm` function output will display ordinary least squares estimates along with standard errors, t values for the Wald test of the hypothesis $\mathrm{H}_0 : \beta_i = 0$ and the associated P-values. Other statistics and information about the sample size, the degrees of freedom, etc., are given at the bottom of the table.

Many methods allow you to extract specific objects from `lm` objects. For example, the

functions `coef`, `resid`, `fitted`, `model.matrix` will return the coefficients $\hat{\beta}$, the ordinary residuals e, the fitted values $\hat{y}$ and the design matrix X, respectively.

```
data(college, package = "hecstatmod") #load data
class(college$rank) #check that rank is a factor
```

```
## [1] "factor"
```

```
#if not, use the following "<-" to assign, "$" to access the column of a data.frame/list
# college$rank <- factor(college$rank, labels = c("assistant","associate","full"))
linmod <- lm(salary ~ sex + rank + service + field, data = college)
coef(linmod) #coefficients
```

```
##      (Intercept)         sexwoman    rankassociate         rankfull
##          86.5963          -4.7712          14.5604          49.1596
##          service fieldtheoretical
##          -0.0888         -13.4734
```

```
summary(linmod) #summary table
```

```
##
## Call:
## lm(formula = salary ~ sex + rank + service + field, data = college)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -64.20 -14.26  -1.53  10.57  99.16
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        86.5963     2.9603   29.25  < 2e-16 ***
## sexwoman           -4.7712     3.8780   -1.23  0.21931
## rankassociate      14.5604     4.0983    3.55  0.00043 ***
## rankfull           49.1596     3.8345   12.82  < 2e-16 ***
## service            -0.0888     0.1116   -0.80  0.42696
## fieldtheoretical -13.4734     2.3155   -5.82  1.2e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 22.7 on 391 degrees of freedom
## Multiple R-squared:  0.448,Adjusted R-squared:  0.441
## F-statistic: 63.4 on 5 and 391 DF,  p-value: <2e-16
```

```
confint(linmod) #confidence intervals for model parameters
```

```
##                    2.5 % 97.5 %
## (Intercept)       80.776 92.416
## sexwoman         -12.396  2.853
## rankassociate      6.503 22.618
## rankfull          41.621 56.698
## service           -0.308  0.131
## fieldtheoretical -18.026 -8.921
```

```
yhat <- fitted(linmod) #fitted values
e <- resid(linmod) #ordinary residuals
```

Bibliography

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. Biometrics Bulletin, 2(6), 110–114.

Welch, B. L. (1947). The generalization of "Student's" problem when several population variances are involved. Biometrika, 34(1–2), 28–35.