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 ??);
- variability: the less noisy your data, the easier it is to detect differences between the curves (big differences are easier to spot, as Figure ?? 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)

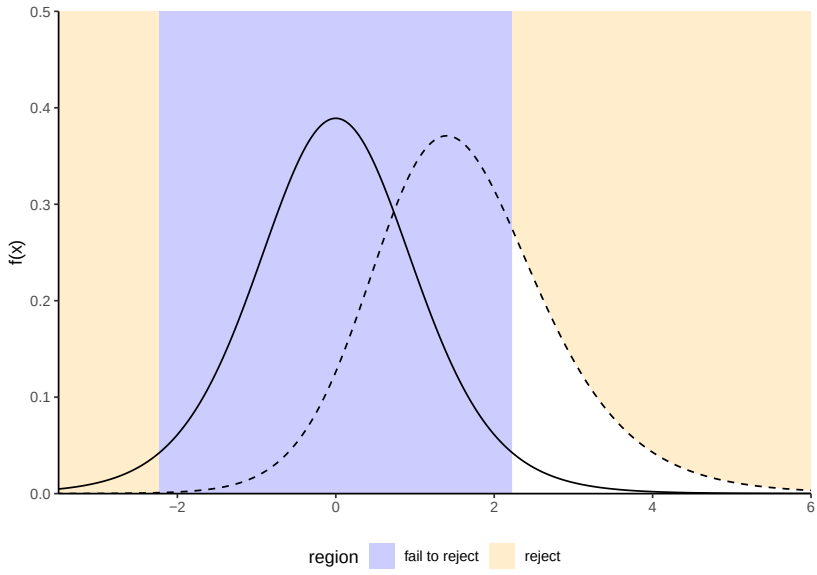

Figure 1.1: Comparison between null distribution (full curve) and a specific alternative for a *t*-test (dashed line). The power corresponds to the area under the curve of the density of the alternative distribution which is in the rejection area (in white).

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.

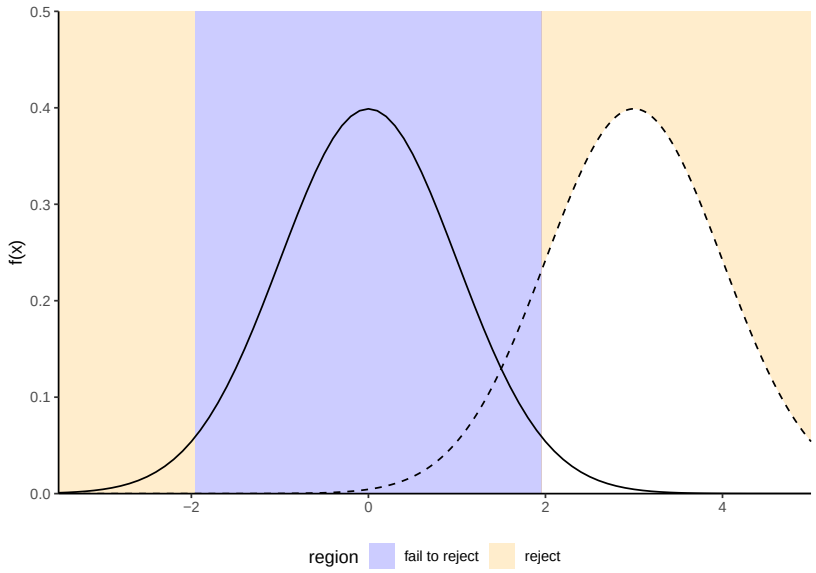

Figure 1.2: Increase in power due to an increase in the mean difference between the null and alternative hypothesis. Power is the area in the rejection region (in white) under the alternative distribution (dashed): the latter is more shifted to the right relative to the null distribution (full line).

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 $(1-\alpha)$

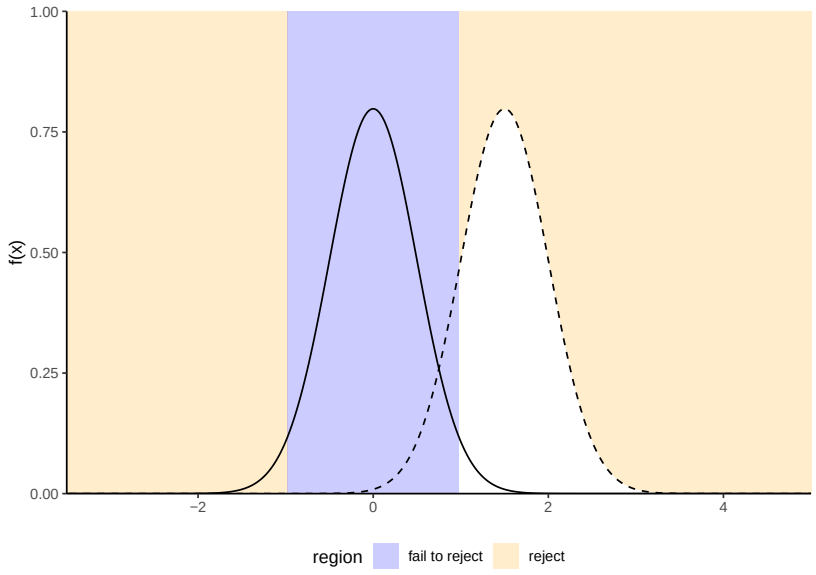

Figure 1.3: Increase of power due to an increase in the sample size or a decrease of standard deviation of the population: the null distribution (full line) is more concentrated. Power is given by the area (white) under the curve of the alternative distribution (dashed). In general, the null distribution 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 $1-\alpha$

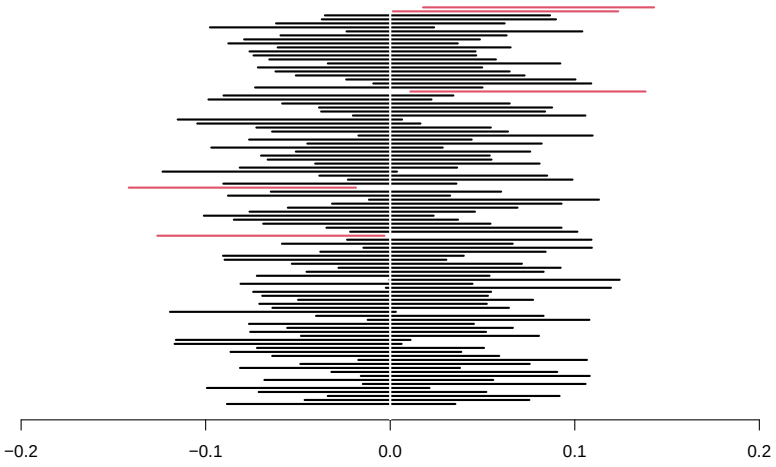

Figure 1.4: 95% confidence intervals for the mean of a standard normal population $\mathsf{No}(0, 1)$, with 100 random samples. On average, 5% of these intervals fail to include the true mean value of zero (in red).

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 ? 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 ? 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 ??. 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.

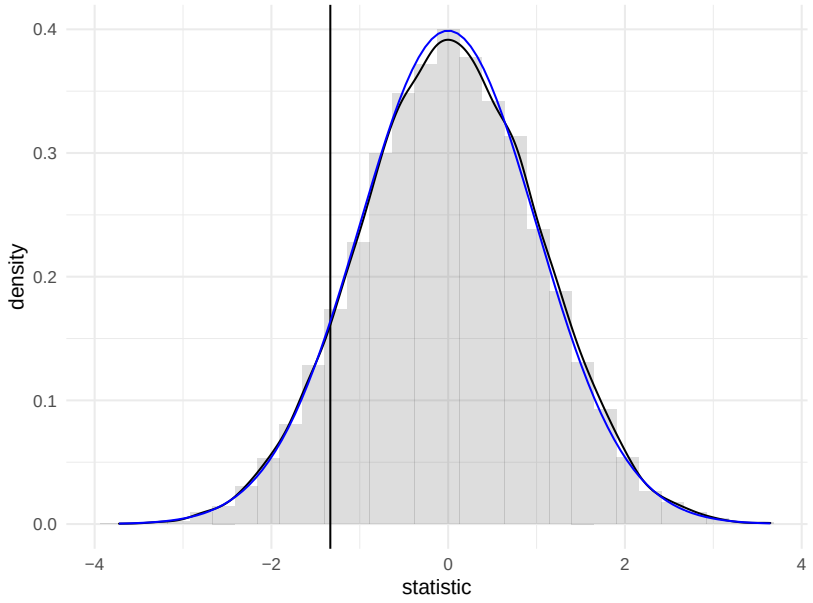

Figure 1.5: Permutation-based approximation to the null distribution of Welch two-sample t-test statistic (histogram and black curve) with standard normal approximation (blue curve) for the price of AVE tickets at promotional rate between Madrid and Barcelona. The value of the test statistic calculated 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 long format, each line corresponds to an observation and each column to a variable: the entries of the data base contain the (numeric) values.

The alternative is wide format, whereby the columns represent categorical variables and the entries are values of the response for a specific category (notably contingency tables). Figure ?? shows the difference between the two structures. Software typically require long formatted database for modelling purposes.

- 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

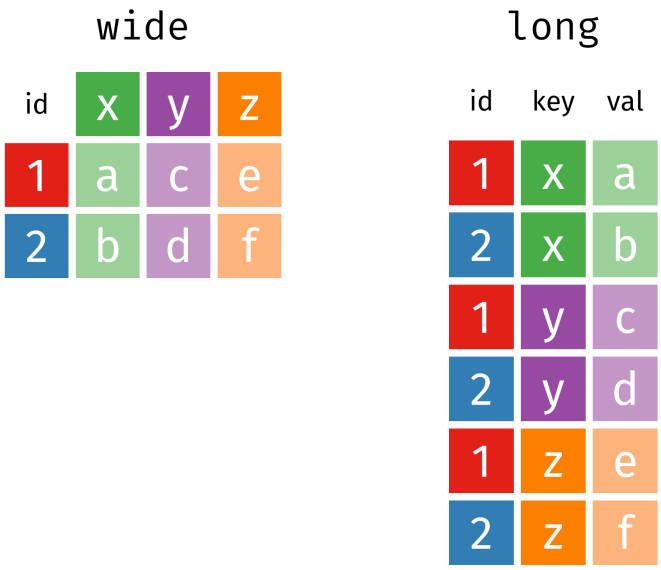

Figure 1.6: Long versus wide-format for data tables (illustration by Garrick Aden-Buie).

- 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)!

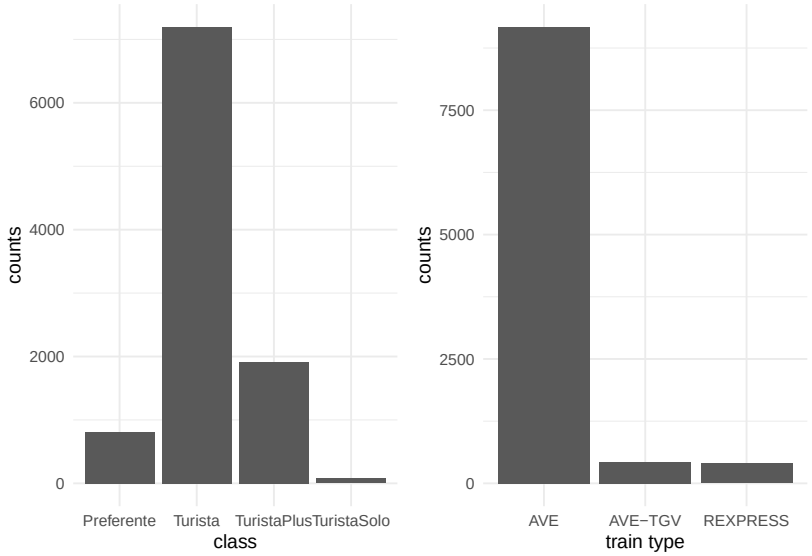

Figure 1.7: Bar plot of ticket class for Renfe tickets data

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.

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.)

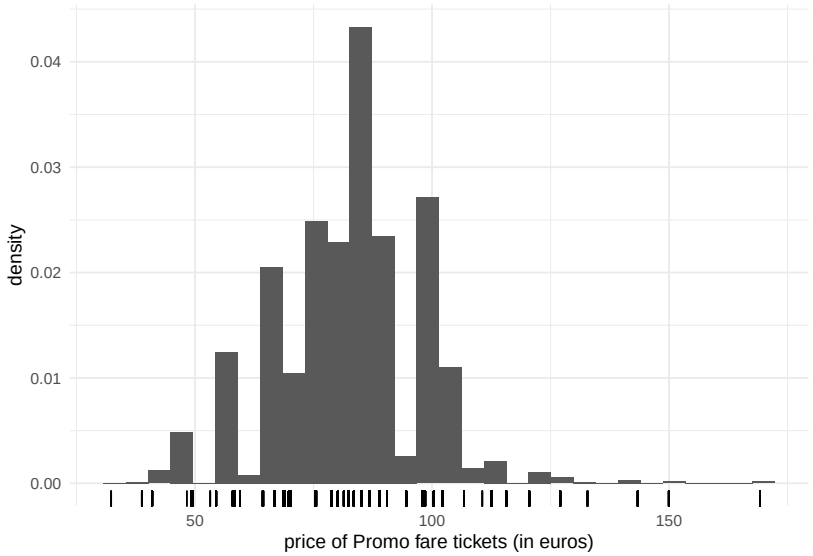

Figure 1.8: Histogram of Promo tickets for Renfe ticket data

- 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.

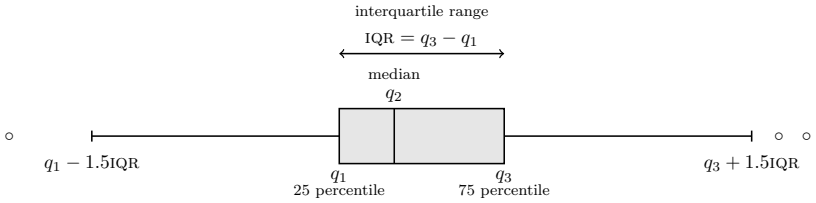

Figure 1.9: Box-and-whiskers plot

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 ??.

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

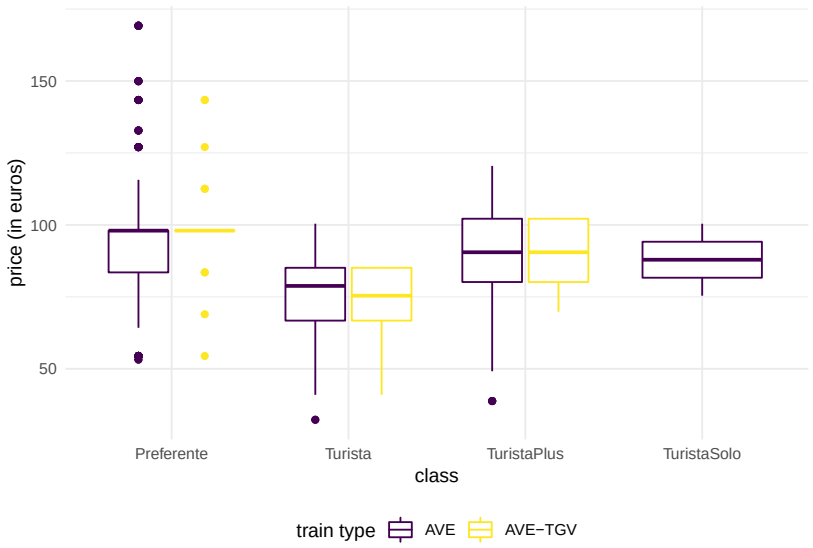

Figure 1.10: Box-and-whiskers plots for Promo fare tickets as a function of class and type for the Renfe tickets data.

transparency, or binning using a two-dimensional histogram with the frequency represented using color are potential solutions. The left panel of Figure ?? 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.

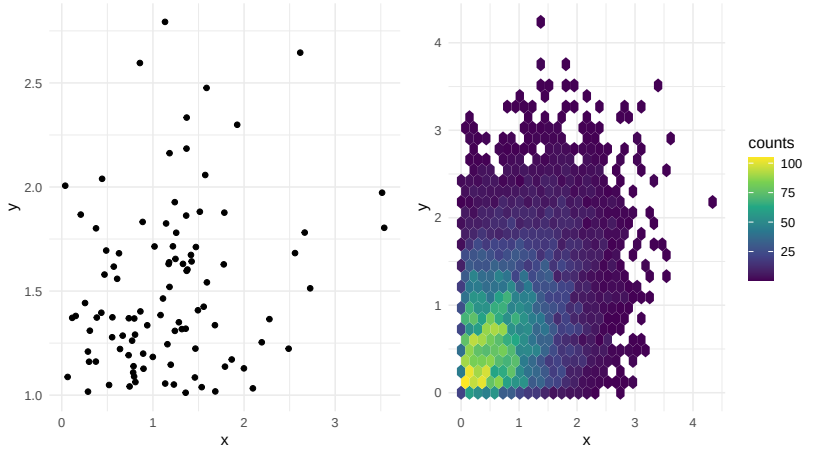

Figure 1.11: Scatterplot (left) and hexagonal heatmap of bidimensional bin counts (right) of simulated data.

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 observations, but this display is sometimes misleading.

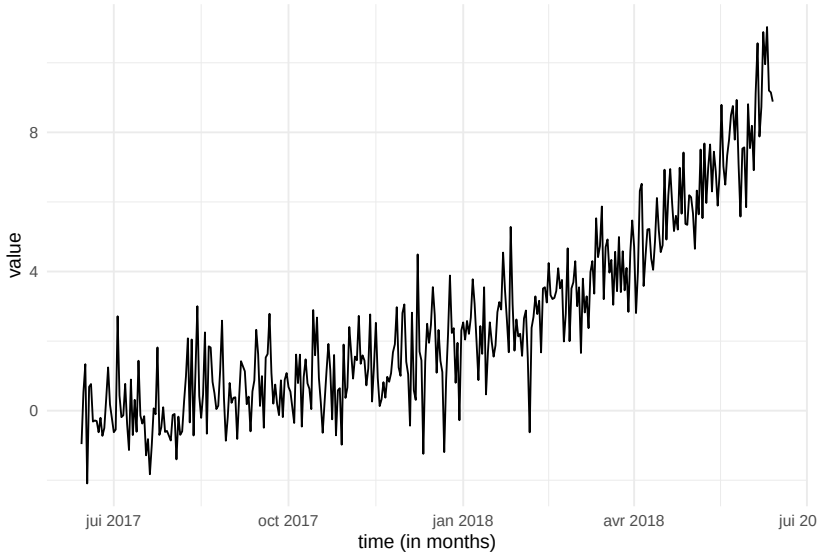

Figure 1.12: Graphical representation of a time series.

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).

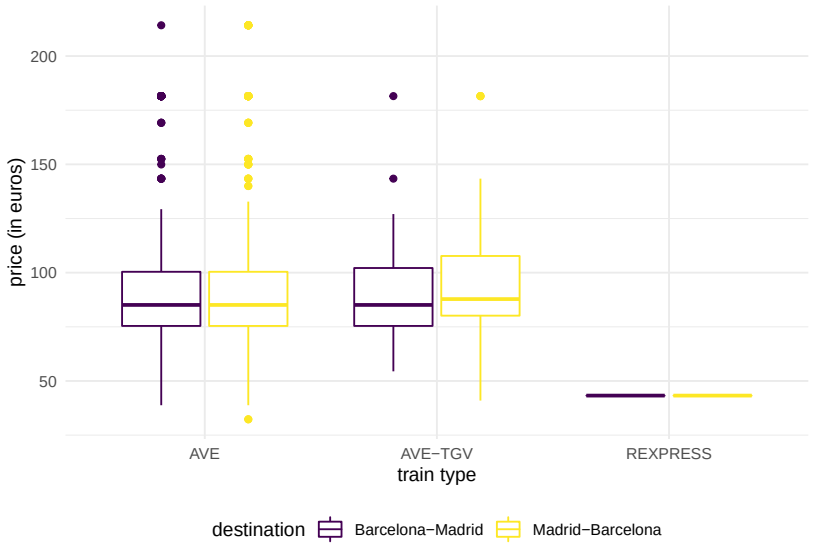

Figure 1.13: Boxplot of ticket price as a function of destination and train type.

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),
            "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 ?? 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")
```

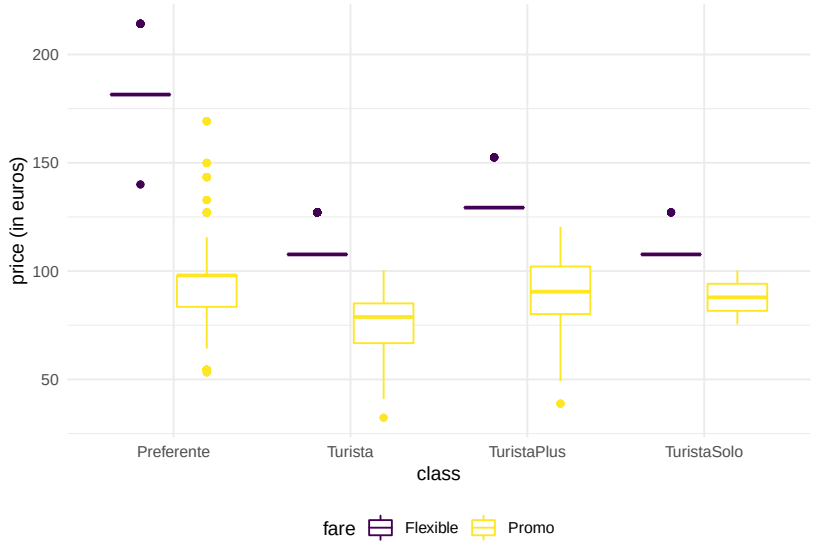

Figure 1.14: Boxplot of ticket price as a function of fare and class for high-speed Renfe trains.

```
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")
```

```
# 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

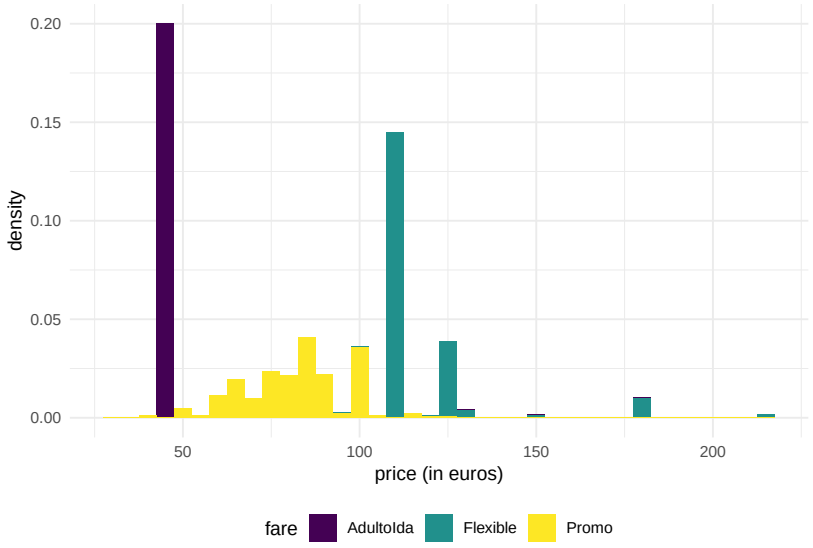

Figure 1.15: Histograms of ticket price as a function of fare for Renfe trains.

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 $\mathrm{X} = \ln(\texttt{years})$, $\mathrm{X} = \texttt{horsepower}^2$ 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.
- `years`: 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

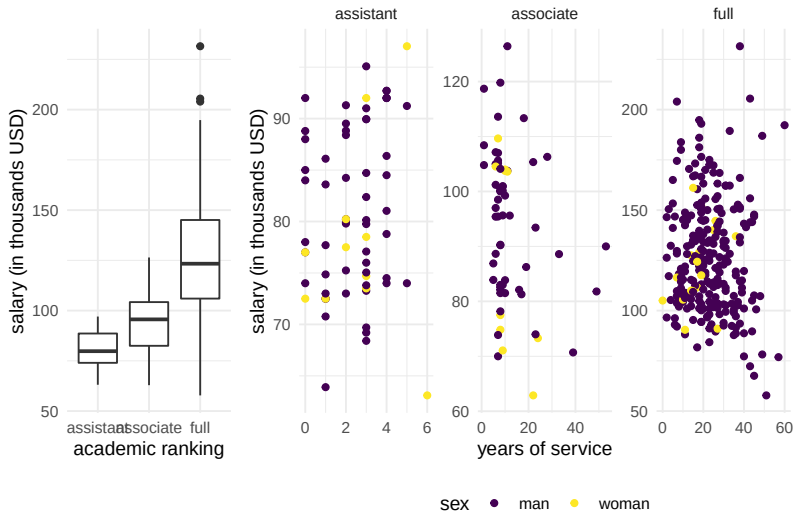

Figure 2.1: Exploratory data analysis of `college` data: salaries of professors as a function of the number of years of service and the academic ranking

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

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

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 ?? shows the line passing through the scatterplot for years of service.

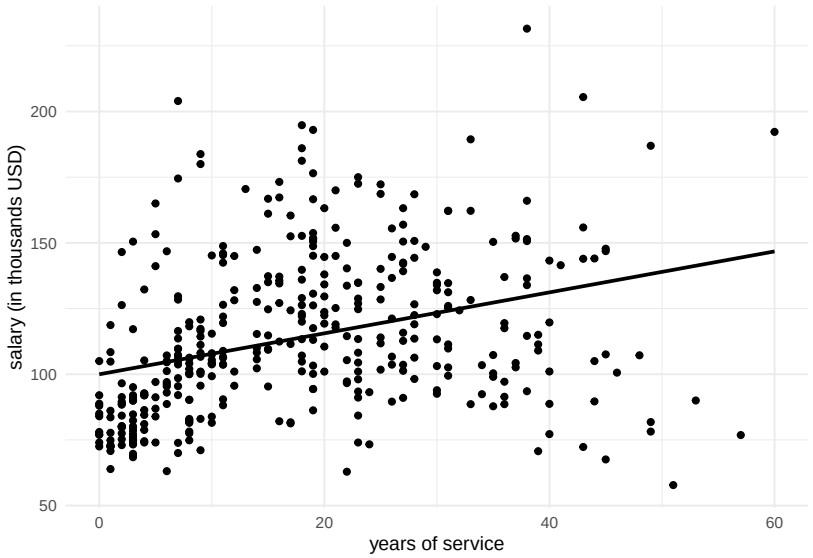

Figure 2.2: Simple linear regression model for the salary of professors as a function of the number of years of service; the line is the solution of the least squares problem.

2.2 Ordinary least squares

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

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

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

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

This system of equations has an explicit solution which is better expressed using matrix notation: this amounts to expressing equation (??) 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 ?? represent the individual vectors of residuals.

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 $\mathrm{O}(np^2)$. 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.

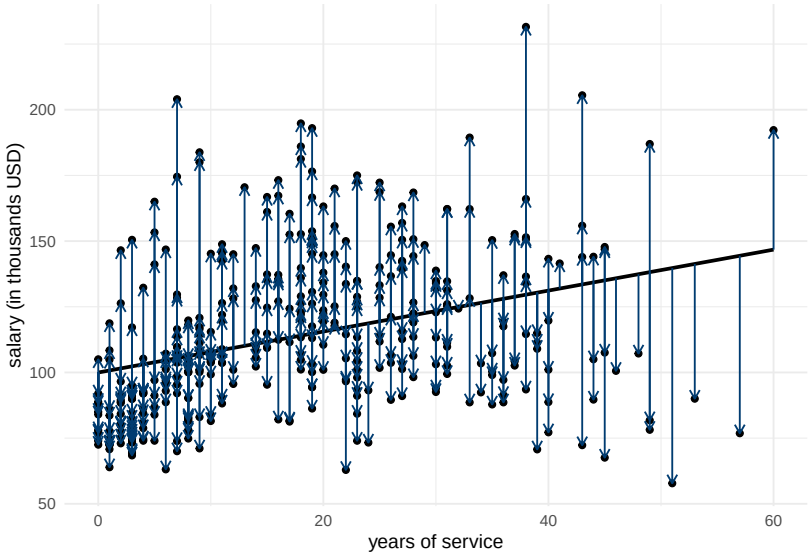

Figure 2.3: Illustration of ordinary residuals added to the regression line (blue vectors).

2.3 Interpretation of the model parameters

What do the β parameters of the linear model represent? In the simple case presented in Figure ??, 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 ?? 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 behaviour (more on this latter).

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: ?? 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 ?? 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 $k-1$

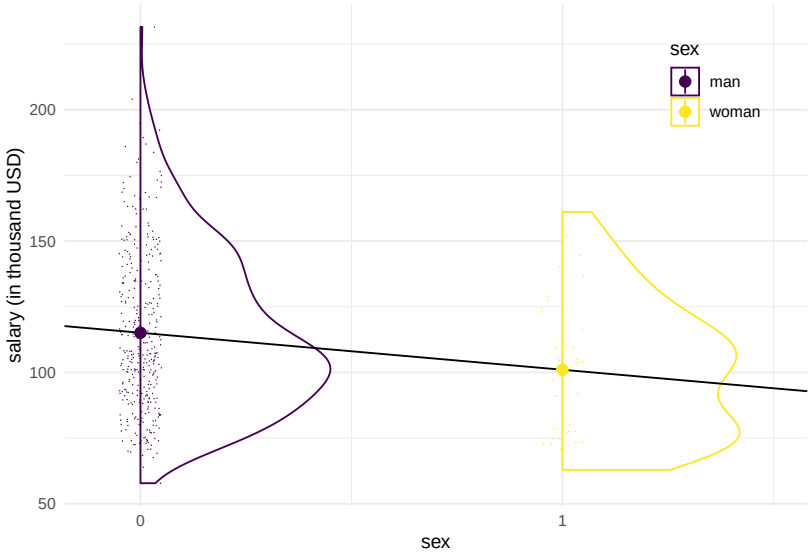

Figure 2.4: Simple linear model for the `college` data using the binary variable sex as regressor: even if the equation defines a line, only its values in 0/1 are realistic.

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 ?? 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 ?? shows the scatterplot with the fitted regression line, above which the line for the simple linear regression for horsepower is added.

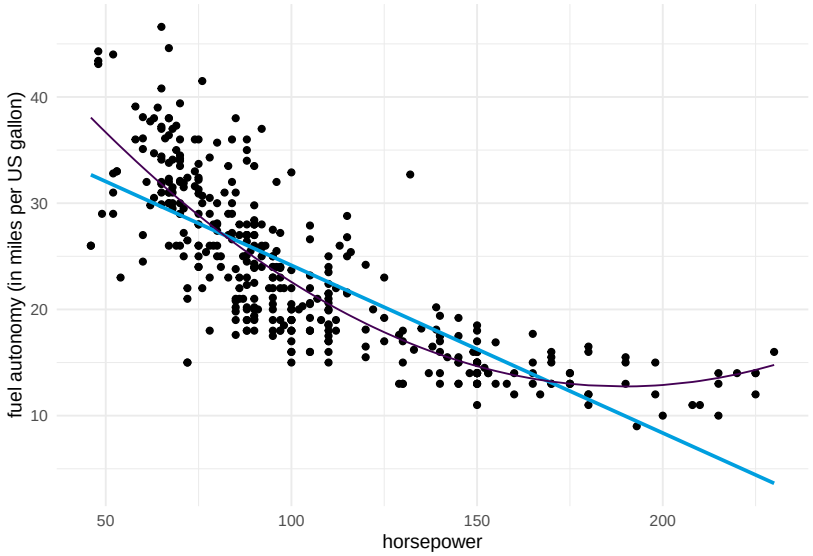

Figure 2.5: Linear regression models for the fuel autonomy of cars as a function of motor power

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 ?? 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 ?? 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 ??.

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. (??) represents the marginal contribution of X_j when all the other

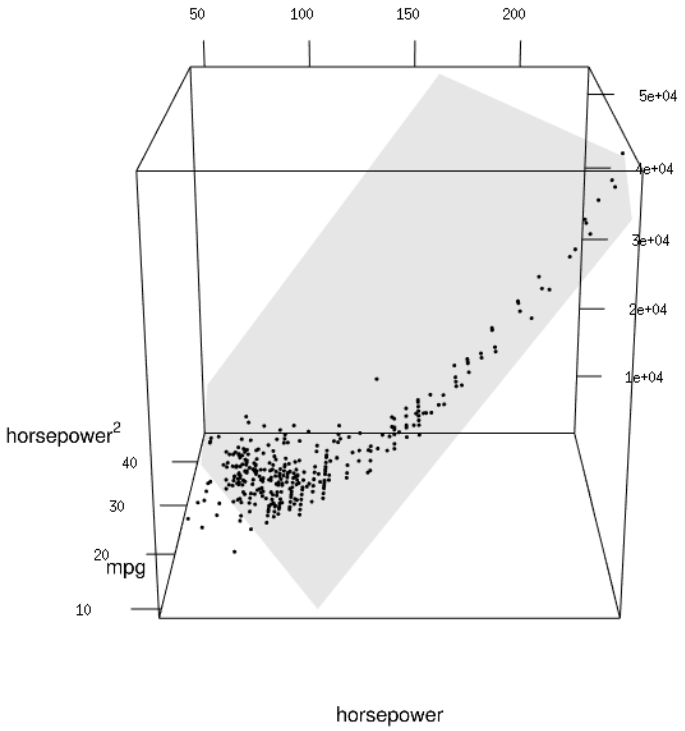

Figure 2.6: 3D graphical representation of the linear regression model for the *extttauto* data.

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-

lated model is

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

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

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 ??, which showed a steady increas of salary over the years. However, this graphical representation is misleading because Figure ?? 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 ??.

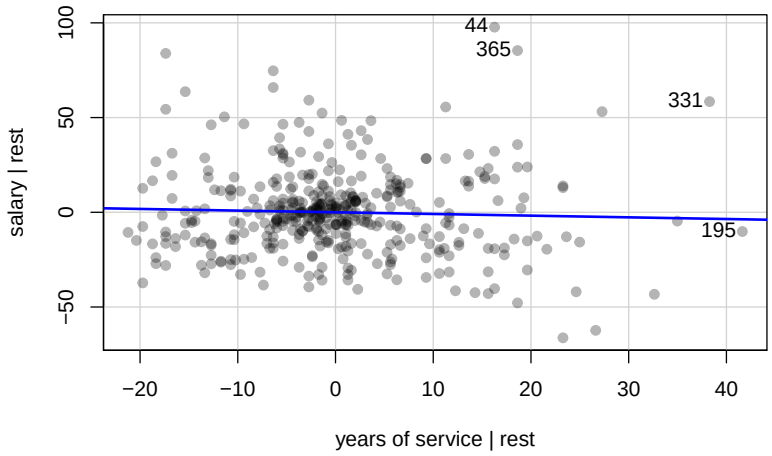

Figure 2.7: Added-variable plot for years of service in the linear regression model of the `college` data.

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

2.4 Tests for parameters of the linear model

We need to make further assumptions in order to carry out inference and build testing procedures for the mean model parameters of the linear model. In order to get a tractable distribution for test statistics, it is customary to assume that the disturbances ε are independent normal random variables with mean zero and common variance σ^2. It follows that $Y_1, \ldots, Y_n$ are conditionally independent random variables with

$$\mathsf{E}(Y_i \mid \mathrm{X}_i) = \beta_0 + \sum_{j=1}^{p} \mathrm{X}_{ij}\beta_j, \quad \mathsf{Va}(Y_i \mid \mathrm{X}_i) = \sigma^2, \quad i = 1, \ldots, n.$$

Under this hypothesis, the least square estimators for the mean parameters coincide with the maximum likelihood estimators. The advantage of imposing this (more stringent than necessary) assumption is that we can use our toolbox for testing. The theory underlying likelihood tests is presented in the chapter on likelihood-based inference. Assuming normal errors leads to exact distributions for the test statistics (which also coincide with the asymptotic ones in large samples).

Of particular interest are tests of restrictions for components of β. The large sample properties of the maximum likelihood estimator imply that

$$\widehat{\beta} \dot{\sim} \mathsf{No}_{p+1}\left\{\beta, \sigma^2(\mathrm{X}^\top\mathrm{X})^{-1}\right\}$$

for sufficiently large sample size. One can thus easily estimate the standard errors from the matrix upon replacing σ^2 by an estimator, typically the unbiased estimator of the variance.

In an inferential setting, it's often important to test whether the effect of an explanatory variable is significant: if X_j is binary or continuous, the test for $\mathcal{H}_0 : \beta_j = 0$ corresponds to a null marginal effect for X_j. The null model is a linear regression in which we remove the $(j+1)$st column of X, so both models are nested. The Wald test statistic is reported by most software $W = \widehat{\beta}_j/\mathsf{se}(\widehat{\beta}_j)$ and the null distribution is Student-t with $n-p-1$ degrees of freedom, which explains the terminology (t values). In addition to coefficient estimates, it is possible to obtain confidence intervals for β_j, which are the usual $\widehat{\beta}_j \pm t_{n-p-1,\alpha/2}\mathsf{se}(\widehat{\beta}_j)$, with $t_{n-p-1,\alpha/2}$ denoting the $1-\alpha/2$ quantile of the St_{n-p-1} distribution.

For categorical variables with more than two levels, testing if $\beta_j = 0$ is typically not of interest because the contrast represent the different between the category X_j and the baseline: these two categories could have a small difference, but the categorical variable as a whole may still be a useful predictor given the other explanatories. The hypothesis of zero contrast is awkward because it implies a null model in which selected categories are merged.

2.4.1 F-tests for comparison of nested linear models

Consider the full linear model which contains p predictors,

$$\mathbb{M}_1 : Y = \beta_0 + \beta_1\mathrm{X}_1 + \cdots + \beta_g\mathrm{X}_g + \beta_{k+1}\mathrm{X}_{k+1} + ... + \beta_p\mathrm{X}_p + \varepsilon.$$

Suppose without loss of generality that we want to test $\mathcal{H}_0 : \beta_{k+1} = \beta_{k+2} = ... = \beta_p = 0$ (one could permute columns of the design matrix to achieve this configuration). The global hypothesis specifies that $(p-k)$ of the β parameters are zero. The restricted model corresponding to the null hypothesis contains only the covariates for which $\beta_j \neq 0$,

$$\mathbb{M}_0 : Y = \beta_0 + \beta_1\mathrm{X}_1 + ... + \beta_k\mathrm{X}_k + \varepsilon.$$

Let $\mathsf{SS}_e(\mathbb{M}_1)$ be the residuals sum of squares for model $\mathbb{M}_1$,

$$\mathsf{SS}_e(\mathbb{M}_1) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i^{\mathbb{M}_1})^2,$$

where $\hat{Y}_i^{\mathbb{M}_1}$ is the ith fitted value from $\mathbb{M}_1$. Similarly define $\mathsf{SS}_e(\mathbb{M}_0)$ for the residuals sum of square of $\mathbb{M}_0$. Clearly, $\mathsf{SS}_e(\mathbb{M}_0) \geq \mathsf{SS}_e(\mathbb{M}_1)$ (why?)

The F-test statistic is

$$F = \frac{\{\mathsf{SS}_e(\mathbb{M}_0) - \mathsf{SS}_e(\mathbb{M}_1)\}/(p-k)}{\mathsf{SS}_e(\mathbb{M}_1)/(n-p-1)}$$

Under $\mathcal{H}_0$, the F statistic follows a Fisher distribution with $(p-k)$ and $(n-p-1)$ degrees of freedom, $\mathsf{F}(p-k, n-p-1)$ — $p-k$ is the number of restrictions, $n-p-1$ is sample size minus the number of β's in $\mathbb{M}_1$.

It turns out that both F and t-statistics are equivalent for testing a single coefficient β_j: the F-statistic is the square of the Wald statistic and they lead to the same inference — the p-value for the test are identical. While it may get reported in tables, the test for $\beta_0 = 0$ is not of interest; we keep the intercept merely to centre the residuals.

For normal linear regression, the likelihood ratio test for comparing models $\mathbb{M}_1$ and $\mathbb{M}_0$ is a function of the sum of squared residuals: the usual formula simplifies to $R = n \ln\{\mathsf{SS}_e(\mathbb{M}_0)/\mathsf{SS}_e(\mathbb{M}_1)\}$. This is reminiscent of the F-statistic formula and the two are in fact intimately related modulo null distribution and scaling. The t-tests and F-tests presented above could thus both be viewed as particular cases of likelihood-based tests.

Consider the `college` data example and the associated linear model with `rank`, `sex`, years of `service` and `field` as covariates.

Table of linear regression coefficients with associated standard errors, Wald tests and p-values based on Student-t distribution

term

estimate

std. error

Wald stat.

p-value

(Intercept)

86.596

2.96

29.25

< 0.001

sex [woman]

-4.771

3.878

-1.23

0.22

field [theoretical]

-13.473

2.315

-5.82

< 0.001

rank [associate]

14.56

4.098

3.55

< 0.001

rank [full]

49.16

3.834

12.82

< 0.001

service

-0.089

0.112

-0.8

0.43

Table ?? shows the estimated coefficients (in thousands of dollars). The coefficients are the least squares estimates $\widehat{\beta}$, the standard errors are the square root of the diagonal elements of $S^2(\mathrm{X}^\top\mathrm{X})^{-1}$. The Wald (or t-) statistic is simply $W = \hat{\beta}/\mathsf{se}(\hat{\beta})$ for $\mathcal{H}_0 : \beta_j = 0$: given two of the three estimates, we could easily recover the third using the formula for the test. The p-values are for the two-sided alternative test with $\mathcal{H}_a : \beta_j \neq 0$.

The interpretation is usual: p-values that are less than our prescribed level α do not contribute significantly given the other variables already in the model. Neither years of service nor sex are statistically different from zero given all the other variables. The test for $\beta_{\texttt{sex}}$ is comparing the model with all covariates (including service), and vice-versa. Note that the conclusion changes depending on the model: both coefficients would be statistically significant had we removed rank from the set of covariates, because they are correlated. The gender imbalance among ranks explains most of the gap between sex, whereas year of service is largely redundant once we account for the jumps due to change of academic rank.

Type 3 sum of square decomposition table: sum of square decomposition comparing nested models with and without covariates, F-statistic and associated p-value.

variable

sum of square

df

F stat.

p-value

(Intercept)

439059.2

1

855.71

< 0.001

sex

776.7

1

1.51

0.22

field

17372.5

1

33.86

< 0.001

rank

102883.1

2

100.26

< 0.001

service

324.5

1

0.63

0.43

Residuals

200620.4

391

Table ?? gives the F-statistics values and the associated p-values. The sum of squares represent the difference $\mathsf{SS}_e(\mathbb{M}_0) - \mathsf{SS}_e(\mathbb{M}_1)$ for various null models $\mathbb{M}_0$, except the last line for residuals which reports $\mathsf{SS}_e(\mathbb{M}_1)$. You can verify that (up to rounding) these p-pvalues are identical to those of the Wald test in the output when `df=1`. The only categorical variable here with more than one level is `rank`, and it is strongly significant: removing it from the model leads to a sharp decrease in fit.

We could have also easily computed the likelihood ratio test to compare the models: for example, the log-likelihood for the full model is -1799.027 and that of the model without rank is -1881.202, so the likelihood ratio statistic would be 164.349 and this is strongly significant when compared to the χ^2_2 distribution (both likelihood ratio test and F-test give a p-value of 2.2×10^{-16}).

2.5 Coefficient of determination

When we specify a model, the error term ε accounts for the fact no perfect linear relationship characterizes the data (if it did, we wouldn't need statistic to begin with). Once we have fitted a model, we estimate the variance σ^2; one may then wonder which share of the total variance in the sample is explained by the model.

The total sum of squares, defined as the sum of squared residuals from the intercept-only model, serves as comparison — the simplest model we could come up with would involving every observation by the sample mean of the response and so this gives (up to scale) the variance of the response, $\mathsf{SS}_c = \sum_{i=1}^n (y_i - \overline{y})^2$. We can then compare the variance of the original data with that of the residuals from the model with covariate matrix X, defined as $\mathsf{SS}_e = \sum_{i=1}^n e_i^2$ with $e_i = y_i - \widehat{\beta}_0 - \sum_{j=1}^p \widehat{\beta}_j \mathrm{X}_j$. We define the coefficient of determination, or squared multiple correlation coefficient of the model, R^2, as

$$R^2 = 1 - \frac{\mathsf{SS}_e}{\mathsf{SS}_c} = \frac{\sum_{i=1}^n (y_i - \overline{y})^2 - \sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \overline{y})^2}.$$

The coefficient of determination can be interpreted as the square of Pearson's linear correlation between the response y and the fitted values $\widehat{y}$; see the Appendix for a derivation of this fact.

Its important to note that R^2 is not a goodness-of-fit criterion: some phenomena are inherently noisy and even a good model will fail to account for much of the response's variability. Moreover, one can inflate the value of R^2 by including more explanatory variables: the coefficient is non-decreasing in the dimension of X, so a model with $p+1$ covariate will necessarily have a higher R^2 values than only p of the explanatories. For model comparisons, it is better to employ information criteria, or else rely on the predictive performance if this is the purpose of the regression. Lastly, a model with a high R^2 may imply high correlation, but the relation may be spurious: linear regression does not yield causal models!

2.6 Predictions

When we compute least square estimates, we obtain fitted values $\widehat{y}$ as $\mathrm{X}\widehat{\beta}$, where X denotes the $n \times (p+1)$ matrix of original observations. Recalling that $\mathsf{E}(Y_i \mid \mathrm{X}_i) = \beta_0 + \sum_{j=1}^p \beta_j \mathrm{X}_{ij}$, we can obtain an estimate of the mean surface for any value of $\mathrm{x}_{n+1} \in \mathbb{R}^p$ by replacing the unknown coefficients β by our estimates $\widehat{\beta}$ – this actually yields the best linear unbiased predictor of the mean.

If we want to predict the value of a new observation, say Y_{n+1}, with explanatories $\mathrm{x}_{n+1} = (1, x_1, \ldots, x_p)$, the prediction of the value will also be $\widehat{y}_{n+1} = \mathrm{x}_{n+1}\widehat{\beta}$ because

$$\mathsf{E}(Y_{n+1} \mid \mathrm{x}_{n+1}) = \mathrm{x}_{n+1}\beta + \mathsf{E}(\varepsilon_{n+1} \mid \mathrm{x}_{n+1}) = \mathrm{x}_{n+1}\beta.$$