DATA SCIENCE REPORT SERIES

A REVIEW OF STATISTICAL ANALYSIS CONCEPTS

Shintaro Hagiwara[1], Patrick Boily[2,3,4]

Abstract

Loosely speaking, a **statistic** is any function of a sample from the distribution of a random variable; statistics aim to extract information from an observed sample to summarise the essential features of a dataset. In this (far too) brief tour of a far-reaching and ubiquitous subject, we review ten areas of particular interest for analysts and consultants.

Keywords

Statistical analysis, hypothesis testing, ANOVA, MANOVA, ANCOVA, data reduction, model selection, multivariate statistics, goodness-of-fit tests, multiple linear regression, non-linear regression, Bayesian statistics, bootstrap, jackknife.

Funding Acknowledgement

Parts of this chapter were funded by Carleton University's Centre for Quantitative Analysis and Decision Support.

[1]School of Mathematics and Statistics, Carleton University, Ottawa
[2]Department of Mathematics and Statistics, University of Ottawa, Ottawa
[3]Data Action Lab, Ottawa
[4]Idlewyld Analytics and Consulting Services, Wakefield, Canada
Email: pboily@uottawa.ca

Contents

1. Introduction

In general, statistics can be divided into two categories based on their purposes: **descriptive statistics** and **inferential statistics**.

As its name implies, **descriptive statistics** aim to describe the collected data. Examples include:

- sample size (overall and/or subgroups);
- demographic breakdowns of participants;
- measures of central tendencies (e.g., mean, median, mode, etc.);
- measures of variability (e.g., sample variance, minimum, maximum, interquartile range, etc.);
- higher distribution moments (skew, kurtosis, etc.);
- non-parametric measures (various quantiles);
- derived measures (correlation coefficients), etc.

They can be presented as a single number, in a summary table, or even in graphical representations (e.g., histogram, pie chart, etc.). Descriptive statistics can be extended to summarise **multivariate** behaviours, *via* sample correlations, contingency tables, scatter plots, etc.

Descriptive statistics not only provide an easily undertandable **overview** of the dataset; they also give analysts a chance to study the collected sample and investigate two important questions:

- is the sample compatible with our understanding of the situation?
- is the sample representative of the underlying population?

Inferential statistics, on the other hand, aim to facilitate the process of inference (**induction**) to the general population from which the sample is drawn.

Our review of statistical methods is by necessity quite brief; further details can be found in [**?**, 2, 4, 6–9].

2. Hypothesis Testing

In a very broad sense, most of statistical inference is done through **hypothesis testing**:

- are the client's conjectures about their business situation compatible with the evidence provided by the data?
- is there a way to get a quantitative ruling in favour of one of several competing conjectures that relies on something other than gut feeling?
- can some of these conjectures be definitively eliminated?

Suppose that a researcher wants to determine if, as she believes, a new teaching method enables students to understand elementary statistical concepts better than the traditional lectures given in a university setting.

She recruits $N = 80$ second-year students to test her claim. The students are randomly assigned to one of two groups: students in group A are given the traditional lectures, whereas students in group B are taught using the new teaching method.

After three weeks, a short quiz is administered to the students in order to assess their understanding of statistical concepts – Table 1 summarises the results.

If we assume that both groups have similar background knowledge prior to being taught (which we attempt to do by randomising the group assignment), then the effectiveness of the teaching methods may be compared using two hypotheses: the **null hypothesis** H_0 and the **alternative** H_a.

Let μ_i represent the true performance of method i.

Group	Sample Size	Sample Mean	Sample Variance
A	$N_A = 40$	$\bar{y}_A = 75.1$	$S_A^2 = 6.7$
B	$N_B = 40$	$\bar{y}_B = 79.0$	$S_B^2 = 5.5$

Table 1. Summary of teaching method study example

One-sided testing pits

$$H_0 : \mu_A \geq \mu_B \quad \text{against} \quad H_\text{a} : \mu_A < \mu_B$$

(or the reverse); in **two-sided testing**, we have

$$H_0 : \mu_A = \mu_B \quad \text{against} \quad H_\text{a} : \mu_A \neq \mu_B.$$

Intuitively, it would seem that testing for inequality of method seems a looser approach (i.e. more general) than testing for the superiority of a specific method over the other.

Hypothesis testing can generate two types of error:

- we can mistakenly reject H_0 when it is, in fact, correct (**type I error**), or
- we can mistakenly accept H_0 when it is actually false (**type II error**).

In order to control the probability of making a type I error (called **significance level**, and denoted by α), we usually let the hypothesis of interest be the alternative hypothesis.

Since the researcher wants to claim that the new method is more effective than the traditional ones, then it is most appropriate for her to use one-sided hypothesis testing with

$$H_0 : \mu_A \geq \mu_B \quad \text{against} \quad H_\text{a} : \mu_A < \mu_B.$$

The testing procedure is simple:

1. calculate a **test statistic** under H_0;
2. reject H_0 in favour of H_a if the test statistic falls in the **critical region** (also called **rejection region**) of an associated distribution (see Figure 1), and
3. fail to reject H_0 otherwise, which is not quite the same thing as accepting it.

Using the summary table above, we can test the researcher's claim by using the **two-sample** t**−test**. Assuming that variability in two groups are roughly the same, the test statistic is given by:

$$t_0 = \frac{\bar{y}_B - \bar{y}_A}{S_p \sqrt{\frac{1}{N_A} + \frac{1}{N_B}}},$$

where the **pooled variance** S_p^2 is

$$S_p^2 = \frac{(N_A - 1)S_A^2 + (N_B - 1)S_B^2}{N_A + N_B - 2}.$$

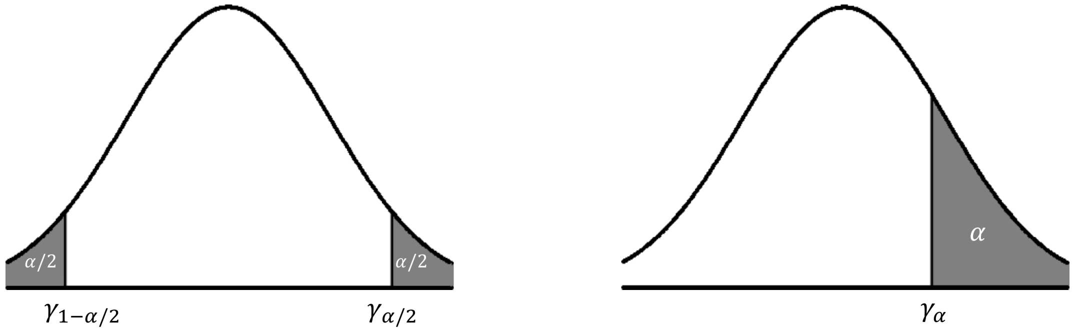


Figure 1. Critical regions for hypothesis testing at α (in grey); two-sided on the left, one-sided on the right; γ_k represent the critical value for the given test and underlying distribution.

In our example, the test statistic value is $t_0 = 7.02$. To reject (or not) the null hypothesis, we need to compare it against the **critical value** of the Student T distribution with $N-2 = 78$ degrees of freedom at $\alpha = 0.05$, which is

$$t^* = t_{1-\alpha,N-2} = t_{0.95,78} = 1.665.$$

Since $t_0 > t^*$ at $\alpha = 0.05$, we have enough evidence to believe that the new teaching method is indeed more effective than the traditional methods, at $\alpha = 0.05$.

In general, the challenge is to recognise which test statistic to use and how it is distributed under H_0. Various scenarios have been explored in the literature (see [**?**], for instance); it could be useful for analysts to be able to derive their own tests when the client's data does not meet the various assumptions.

Ad-hoc solutions come at a price, however – a fair number of clients (and reviewers), if they are familiar with statistical tests at all, do not understand how they are derived and thus only trust 'tried, tested, and true' methods (this also applies to other fields of quantitative analysis). Custom approaches are likely to be treated with **suspicion**.

2.1 Questions to Ponder

1. Distribution assumptions:
   - what distribution assumptions are we making by using a t—test?
   - how can we verify them?
   - if such assumptions are violated, what is our recourse?
2. Assumption of equal variance:
   - how can we verify the appropriateness of using pooled variance?
   - if it is not appropriate, can we modify the test to overcome the problem?
3. One-sided vs. two-sided tests:
   - when is it appropriate to use a one-sided test, and when is it better to employ a two-sided test?
   - are there drawbacks in using a two-sided test when a one-sided test would be indicated?

3. Analysis of Variance (ANOVA)

Analysis of variance (ANOVA) is a statistical method that partitions a dataset's variability into **explainable variability** (model-based) and **unexplained variability** (error) using various statistical models, to determine whether (multiple) treatment groups have significantly different group means.

The **total sample variability** of a feature y in a dataset is defined as

$$\text{SS}_{\text{tot}} = \sum_{k=1}^{N} (y_k - \bar{y})^2,$$

where $\bar{y}$ is the overall mean of the data.

Let us return to the teaching method example given in Section 2.

Figure 2 shows all the students' scores, ordered by participant ID. Since the assignment of ID is **arbitrary** (at least, in theory), we do not observe any patterns – if we were to guess someone's score with no knowledge except for their participant ID, then picking the sample mean is as good a guess as any other reasonable guesses.

Statistically speaking, this means that the **null model**

$$y_{i,j} = \mu + \varepsilon_{i,j},$$

where μ is the **overall mean**, $i = A, B$, and $j = 1, \ldots, 40$, does not explain any of the variability in the student scores (as usual, $\varepsilon_{i,j}$ represents the departure or noise from the model prediction).

But the students DID NOT all receive the same treatment: 40 randomly selected students were assigned to group A, and the other 40 to group B, and both group were taught using a different method.

When we add this information onto Figure 2 (on the right), we clearly see that the two study groups show different characteristics in term of their average scores.

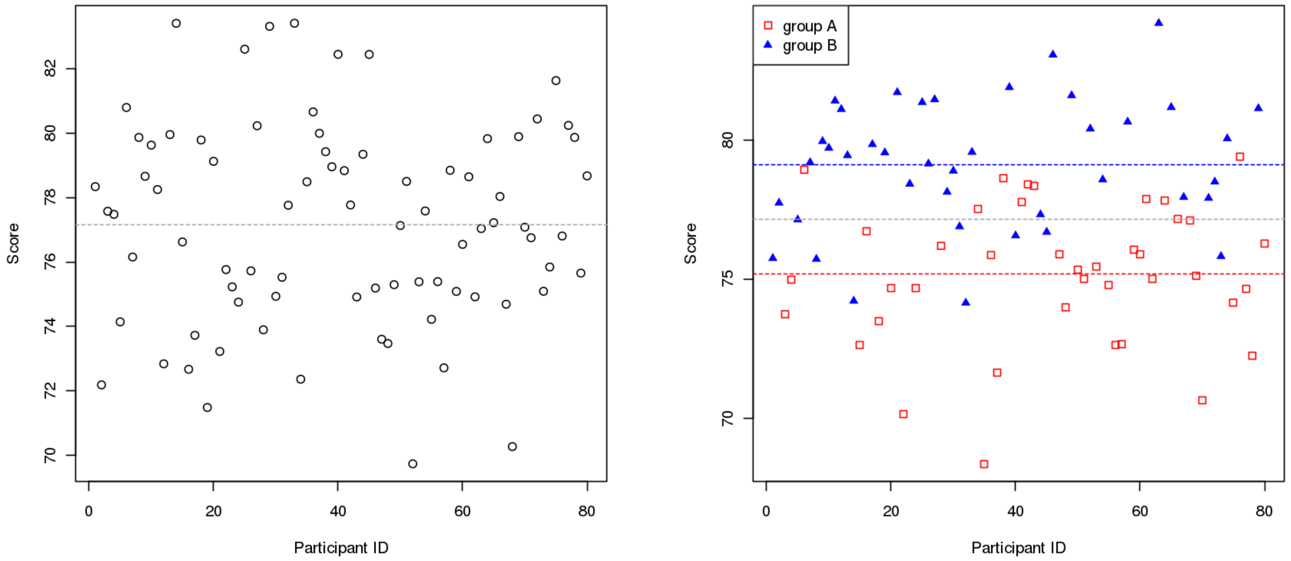


Figure 2. Effectiveness of new teaching method for two groups. The grey line is the overall sample mean (left), while the red and blue lines represent the average score for groups A and B, respectively (right).

With the group assignment information, we can refine our null model into the **treatment-based model**

$$y_{i,j} = \mu_i + \varepsilon_{i,j},$$

where μ_i, $i = A, B$ represent the group means.

Using this model, we can decompose SS_tot into **between-treatment sum of squares** and **error (within-treatment) sum of squares** as

$$\begin{aligned} \text{SS}_\text{tot} &= \sum_{i,j}(y_{i,j} - \bar{y})^2 = \sum_{i,j}(y_{i,j} - \bar{y}_i + \bar{y}_i - \bar{y})^2 \\ &= \sum_i N_i(\bar{y}_i - \bar{y})^2 + \sum_{i,j}(y_{i,j} - \bar{y}_i)^2 = \text{SS}_\text{treat} + \text{SS}_\text{e} \end{aligned}$$

The SS_treat component looks at the difference between each of the treatment means and the overall mean, which we consider to be **explainable**[1]; the SS_e component, on the other hand, looks at the difference between each observation and its own group mean, and is considered to be **random**.[2]

Thus, $\text{SS}_\text{treat}/\text{SS}_\text{tot} \times 100\%$ of the total variability can be explained using a treatment-based model. This ratio is called the **coefficient of determination**, denoted by R^2.

Formally, the ANOVA table incorporates a few more items – Table 2 summarises all the information that it contains; the specific table for the teaching methodology example is shown in 3.

The test statistic F_0 follows an F-distribution with

$$(\text{df}_\text{treat}, \text{df}_\text{e}) = (1, 78)$$

degrees of freedom. At a significance level of $\alpha = 0.05$, the critical value $F^* = F_{0.95,1,78} = 3.96$ is substantially smaller than the test statistic $F_0 = 49.28$, implying that the two-treatment model is statistically significant.

This, in turn, means that the model recognises a statistically significant difference between the students' scores, based on the teaching methods.

The coefficient of determination R^2 provides a way to measure the model's **significance**. From Table 3, we compute

$$R^2 = \frac{\text{SS}_\text{treat}}{\text{SS}_\text{tot}} = \frac{300.31}{775.69} \approx 0.39,$$

which means that 39% of the total variation in the data can be explained by our two-treatment model. Is this good enough? That depends on the specifics of the situation (in particular, on the client's needs).

3.1 Diagnostic Checks

As with most statistical procedures, ANOVA relies on certain assumptions for its result to be valid. Recall that our model is given by

$$y_{i,j} = \mu_i + \varepsilon_{i,j}$$

What assumptions are made? The main assumption is that the error terms follow independently and identically distributed (i.i.d.) normal distributions (i.e., $\varepsilon_{i,j} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$).

[1] That is to say, the treatment explains part of the difference in the observed group means.

[2] As the spread about the group means is fairly large (relatively-speaking), we suspect that the treatment-based model on its own does not capture all the variability in the data.

Source	Sum of Squares	df	Mean Square	F_0	p–value
Treatment (Model)	SS_{treat}	$p-1$	$\text{MS}_{\text{treat}} = \text{SS}_{\text{treat}}/(p-1)$	$\text{MS}_{\text{treat}}/\text{MS}_{\text{e}}$	$P(F_0 > F^*)$
Error	SS_{e}	$N-p$	$\text{MS}_{\text{e}} = \text{SS}_{\text{e}}/(N-p)$		
Total	SS_{tot}	$N-1$			

Table 2. A simple ANOVA table, with p treatments and N observations.

Source	Sum of Squares	df	Mean Square	F_0	p–value
Treatment (Model)	300.31	1	300.31	49.28	7.2×10^{-10} ***
Error	475.38	78	6.095		
Total	775.69	79			

Table 3. ANOVA table for the teaching methodology example, with $p = 2$ and $N = 80$, at $\alpha = 0.001$.

Assuming independence, we are required to verify three additional assumptions:

- normality of the error terms;
- constant variance (within treatment groups), and
- equal variances (across treatment groups).

Normality of the errors can be tested visually with the help of a **normal-QQ plot**, which compares the **standardized residuals quantiles** against the **theoretical quantiles** of the standard normal distribution $N(0,1)$ (a straight line indicates normality).

In other words, if the errors are normally distributed with mean 0 and variance σ^2, we would expect that the 80 standardized residuals $r_{i,j} = \frac{\varepsilon_{i,j}-0}{\sigma}$ should behave as though they had been drawn from $N(0,1)$.

Figure 3 (left) shows some departure in the lower tail, however, moderate departure from normality is usually acceptable as long as it is mostly a tail phenomenon.

To test the assumption of constant variance, we can run visual inspection using

- residuals vs. fitted values, and/or
- residuals vs. order/time.

The standardized residuals in both groups should be approximately distributed according to $N(0,1)$. Figure 3 (right) shows that variability from the mean in each treatment group is reasonably similar.[3]

More formally, equality of variance is often tested for using **Bartlett's test** (when normality of the residuals is met) or the **modified Levene's test** (when it is not).

Assuming that we felt the evidence of normal residuals was warranted in the two-treatment model of the teaching dataset, we get a p–value of 0.57 for Bartlett's test; otherwise, we get a p–value of 0.76 for Levene's test. In either case, the p–value falls above reasonable significance levels (0.05, say), which means that we cannot reject the null hypothesis of equal variance.

When there are $p > 2$ treatment groups, ANOVA provides a test for

$$H_0 : \mu_1 = \cdots = \mu_p \quad \text{vs.} \quad H_1 : \mu_i \neq \mu_j \text{ for at least one } i \neq j.$$

A significant F_0 value indicates that **there is at least one group which differs from the others**, but it does not specify which one(s) that may be.

Specialised methods such as **Scheffe's method** and **Tukey's test** can be used to identify the statistically different treatments.

Finally, while ANOVA can accommodate unequal treatment group sizes, it is recommended to keep those sizes equal across all groups – this makes the test statistic less sensitive to violations of the assumption of equal variances across treatment groups, providing yet another reason to involve the analysts/consultants in the **data collection process**.

4. Multiple Linear Regression

In the previous sections, we considered a simple scenario where a single, categorical, explanatory variable (Treatment A vs. Treatment B) was used to model a desired response variable (score Y).

Real-world data is, of course, much more intricate and complex, typically consisting of multiple response variables, with multiple quantitative and categorical/qualitative explanatory features.

In this section, we will review how to handle such cases.

4.1 Multiple Linear Regression in Matrix Form

Throughout, we suppose that the dataset consists of N observations with a single response output Y and p explanatory variables $X_1, \ldots, X_p$. The **first-order linear model** describing this scenario can be represented in matrix from by

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where the vectors $\boldsymbol{Y} = [y_1, \cdots, y_N]^\top$, $\boldsymbol{\beta} = [\beta_0, \cdots, \beta_p]^\top$, and $\boldsymbol{\varepsilon} = [\varepsilon_1, \cdots, \varepsilon_N]^\top$ are the **response vector**, the **coeffi-**

[3] If a difference is apparent and we cannot conclude that the variances are constant across groups, we need to apply a **variance stabilising transformation**, such as a **logarithmic transformation** or **square-root transformation** before proceeding.

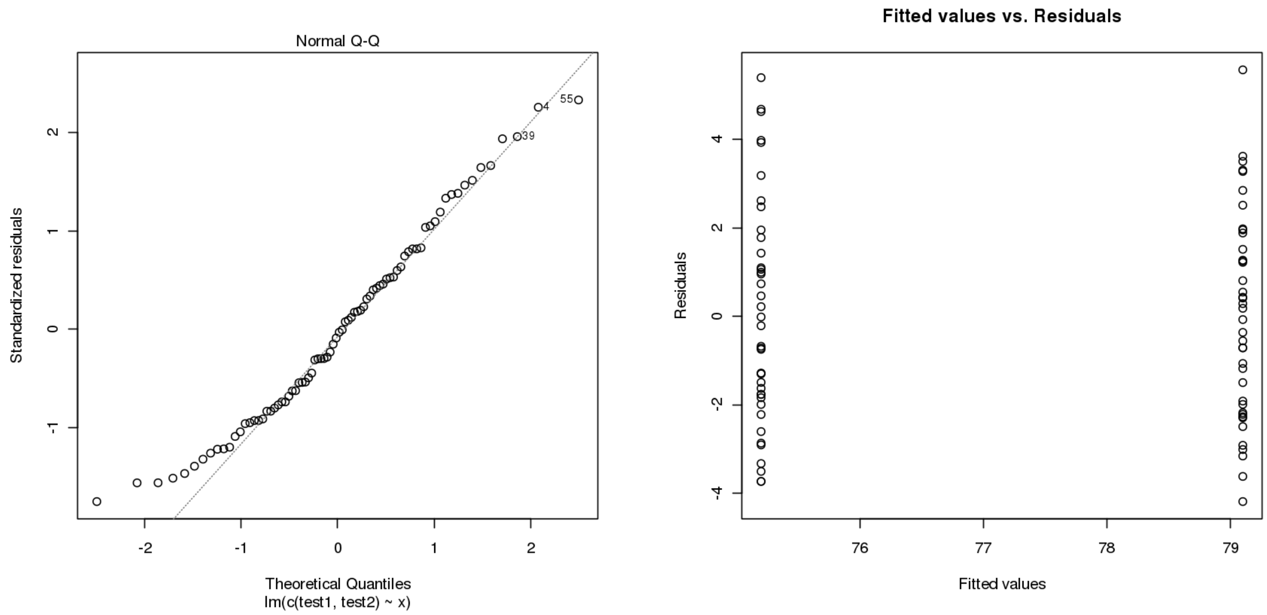


Figure 3. On the left: normal QQ-plot for the two-treatment teaching model (standardised residuals); note the moderate (but acceptable) departure in the lower tail. On the right: diagnostic check for constant variance in the two-treatment teaching model. The spread is fairly similar; we can safely assume constant variance (as well as equal variance across treatment groups).

Source	Sum of Squares	d.f.	Mean Square	F_0
Regression	SS_{reg}	$p-1$	$MS_{reg} = SS_{reg}/(p-1)$	MS_{reg}/MS_{e}
Error	SS_{e}	$N-p$	$MS_{e} = SS_{e}/(N-p)$	
Total	SS_{tot}	$N-1$		

Table 4. ANOVA table for first-order multiple regression model (1); with p explanatory variables and N observations.

cient vector, and the **error vector**, respectively, and

$$X = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,p} \end{bmatrix}$$

is the **design matrix**, with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, where $\boldsymbol{I}_n$ is the $N \times N$ **identity matrix**.

4.2 Qualitative Explanatory Variables

Some say that the colour of a vehicle is part of the assessment for car insurance premiums. Such a variable is **qualitative** (nominal, in fact) in nature, as there is no reasonable way to order colours for insurance purposes.

If we want to incorporate this feature in an insurance premium model taking into account k possible colour choices (red , black , . . . , green , yellow), then we need $k-1$ dummy variables $X_1, \ldots, X_{k-1}$ defined according to

$$X_1 = \begin{cases} 1 & \text{if red} \\ 0 & \text{otherwise} \end{cases} \cdots \quad X_{k-1} = \begin{cases} 1 & \text{if green} \\ 0 & \text{otherwise} \end{cases}$$

With **ordinal variables** (e.g., *on scale of 1 to 5, how likely are you to buy a new phone this year?*), we may choose to have 4 dummy variables as above, or a single continuous variable.

While the latter approach saves 4 degrees of freedom, we are imposing an assumption that equal spacings on the ordinal axis have an equal impact on the outcome, which may not be the case – if it isn't so, it might be preferable to use dummy variables.

4.3 Overall Significance of the Model

For the model presented in (1), **ordinary least square** (OLS) estimation yields **fitted values**

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$$

and residuals

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}^\top)\boldsymbol{Y}.$$

The ANOVA table has the same form as Table 2, although the sums of squares will be different:

$$\begin{aligned} SS_{tot} &= \|\boldsymbol{Y} - \bar{y}\mathbf{1}\|^2 = \sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 + \|\hat{\boldsymbol{Y}} - \bar{y}\mathbf{1}\|^2 \\ &= SS_{reg} + SS_{e} \end{aligned}$$

(see Table 4).

It is used in testing

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \quad \text{against} \quad H_1 : \beta_i \neq 0 \text{ for some } i.$$

If the test statistic F_0 is **significant**, it does not necessarily imply that all the independent variables $X_1, \ldots, X_p$ are useful in predicting y, only that at least one of them is.

We can examine significance of the β coefficients individually (using a t-test), or multiple coefficients simultaneously (e.g., by building **Bonferroni simultaneous confidence interval**). Choosing the best subset of the model will be discussed in the next section.

4.4 Model Adequacy Checks

There are some rare examples for which OLS does not yield a unique solution; but in the vast majority of instances, the data can be fitted to the model. How can we tell if the model is **adequate** to the situation at hand?

- **Assumptions on Residuals –** we cannot emphasise enough that **statistical significance** (i.e. when F_0 is in the critical region) **does not mean that the model is necessarily valid**; the conclusion only follows once the model has been determined to be an **adequate** fit for the data. A normal-QQ plot can help verify the assumption of normality, for instance, while the assumptions of independence and constant variance can be tested using scatterplots of fitted values against residuals.

- **Outliers and Influential Points** – in addition, **outliers** and **influential points** could affect the fitted values. While it is typically easier to classify some observations as outliers, influential points can distort the regression line significantly. Figure 4 shows the clear impact of an influential point. Outliers and influential points should be studied carefully, as there are a number of possible mechanisms that can account for their presence; it may be that these anomalies are due to data entry error, in which case we may try to correct/impute with a reasonable alternative, if possible. It may be the case that these unusual observations are worth studying on their own merit.

- **Multicollinearity** and **Variance Inflation Factor** – last but not least, it is important to take a look at the scatterplot matrix and the correlation matrix of the explanatory variables to detect **multicollinearity**. While it is hoped that the explanatory variables have some relationship with the response variable (otherwise any model is bound to be fruitless), high correlations and/or dependencies among the explanatory variables is contra-indicated as it introduces instability in the estimates of the regression coefficients are unstable. We can formally test for presence of multicollinearity using **variance inflation factors** (VIF); in its presence, data reduction and data transformation strategies might need to be implemented.

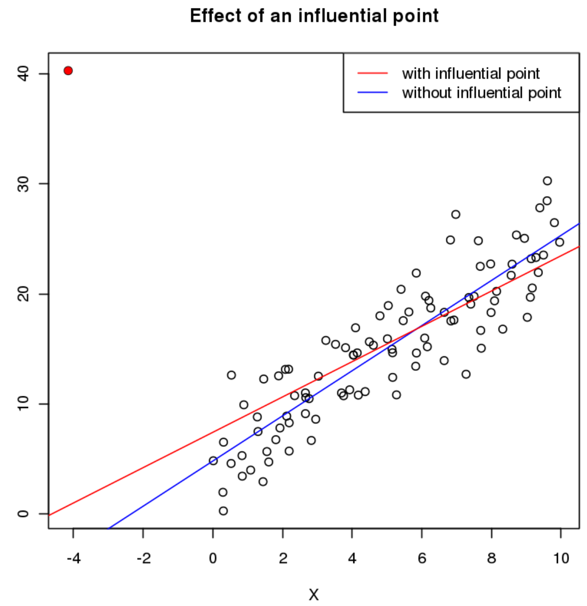


Figure 4. Illustrative example of the effect of an influential point. The red dot in the top left corner is an influential point – the slope of the regression line when it is included in the data (red) is quite different from the slope when it is not (blue).

5. Data Reduction/Model Selection

In a good model, a balance must be struck between **predictive ability** and **simplicity**. In practice, we look for the **simplest model** that explains the behaviour of the response variable Y in a **reasonably adequate manner** (a version of *Occam's Razor*).

If there are p predictor variables $X_1, \ldots, X_p$, then there are 2^p possible models from which to select the "best", ranging from the **simple average model**

$$y_i = \beta_0 + \varepsilon_i$$

to the **full model**

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \varepsilon_i.$$

5.1 Step-Wise Regression

As the number of predictors p grows, it is not feasible to fit all 2^p possible models to determine the optimal model.

Step-wise regression is an automated model selection procedure that builds a succession of models from which a choice can be made. There are numerous variants – the particular algorithm we present is called **forward selection**, for reasons that will shortly become clear (to fix the problem in conceptual space, assume that there are $p = 10$ predictor variables).

1. **Selecting the first variable:** fit p simple linear regressions

   $$y_i = \beta_0 + \beta_j x_{i,j} + \varepsilon_i, \quad j = 1, \ldots, p$$

   and choose the model with highest R^2 value. In other words, select the variable X_j that best describes the behaviour of Y **on its own**. If X_5 turns out to be that variable, fsay, then the tentative model is

   $$y_i = \beta_0 + \beta_5 X_{i,5} + \varepsilon_i.$$

   If this model is not statistically significant (tested at a predetermined significance level α), then the final model selection is

   $$y_i = \beta_0 + \varepsilon_i$$

   and the search is complete. Otherwise, proceed to step 2.

2. **Selecting the second variable:** fit all two-parameter regression models

   $$y_i = \beta_0 + \beta_5 x_{i,5} + \beta_j x_{i,j} + \varepsilon_i, \quad j = 1, \ldots, p, \quad j \neq 5.$$

   Select the model that has the highest value of the test statistic

   $$t'_k = \sqrt{\frac{\text{MS}_{\text{reg}}(X_5, X_k) - \text{MS}_{\text{reg}}(X_5)}{\text{MS}_{\text{e}}(X_5, X_k)}}.$$

   Say that $k = 3$ yields the largest such value. If the associated model's p—value is smaller than α, then our tentative model is updated to

   $$y_i = \beta_0 + \beta_3 X_{i,3} + \beta_5 X_{i,5} + \varepsilon_i$$

   and we proceed to step 3. Otherwise, the final model selection is

   $$y_i = \beta_0 + \beta_5 X_{i,5} + \varepsilon_i$$

   and the search is complete.

3. **All subsequent steps:** Repeat step 2 using

   $$t_k^{''} = \sqrt{\frac{\text{MS}_{\text{reg}}(X_5, X_3, X_k) - \text{MS}_{\text{reg}}(X_5, X_3)}{\text{MS}_{\text{e}}(X_5, X_3, X_k)}},$$

   and so forth, until no additional term improves the model significantly.

In contrast to forward selection which starts with the simple average model

$$y_i = \beta_0 + \varepsilon_i$$

and build a nested sequence of increasingly complex models, **backward elimination** begins with the full model

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \varepsilon_i$$

and keeps removing terms until removal of *any* variable causes a significant loss of its predictive power (calculated using $t_k^{(\ell)}$). In general, forward selection and backward elimination will not select the same final model.

In the **combined approach**, the process starts from the simple average model as in forward selection, but each time a new variable is added to the tentative model, a backward elimination search is performed to test whether any of the previously added variables are no longer significant, which can prevent **overfitting** (mistaking noise for a pattern).

The test statistic $t_k^{(\ell)}$ is the square root of the ratio of conditional MSR over MSE. In everyday terms, it is testing *whether the addition of X_k provides a significant improvement in predictive ability over the current tentative model's*. Other alternative include the **Akaike Information Criterion** (AIC), the **Bayesian Information Criteria** (BIC), **Mallow's C_p Criterion**, and the **R^2 criterion** – simply pick the model which optimises the desired criterion.

Note that step-wise regression is **flawed** in many ways which we will not explore at the moment; in practice, it has started being replaced by **regularisation methods** such as ridge regression and the LASSO.

6. Basics of Multivariate Statistics

To this point, we have only considering situations where the response has been **univariate**. In applications, the situation often calls for **multivariate** responses, where the response variables are thought to have some relationship to one another (e.g. a **correlation structure**).

It remains possible to analyse each response variable independently, but the dependence structure can be exploited to make **joint** (or simultaneous) inferences.

6.1 Properties of the Multivariate Normal Distribution

The probability density function of a multi-dimensional random vector $\mathbf{X} \in \mathbb{R}^p$ that follows a **multivariate normal distribution** with **mean vector $\boldsymbol{\mu}$** and **covariance matrix Σ**, denoted by $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$, is given by

$$f(\boldsymbol{X}) = \frac{1}{(2\pi)^{p/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})\right),$$

where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p,1} & \sigma_{p,2} & \cdots & \sigma_{p,p} \end{bmatrix}.$$

For such an $\boldsymbol{X}$, the following properties hold:

1. any linear combination of its components are normally distributed;
2. all subsets of components follow a (modified) multivariate normal distribution;
3. a diagonal covariance matrix implies the independence of its components;
4. conditional distributions of components follow a normal distribution, and
5. the quantity $(\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$ follows a χ^2_p.

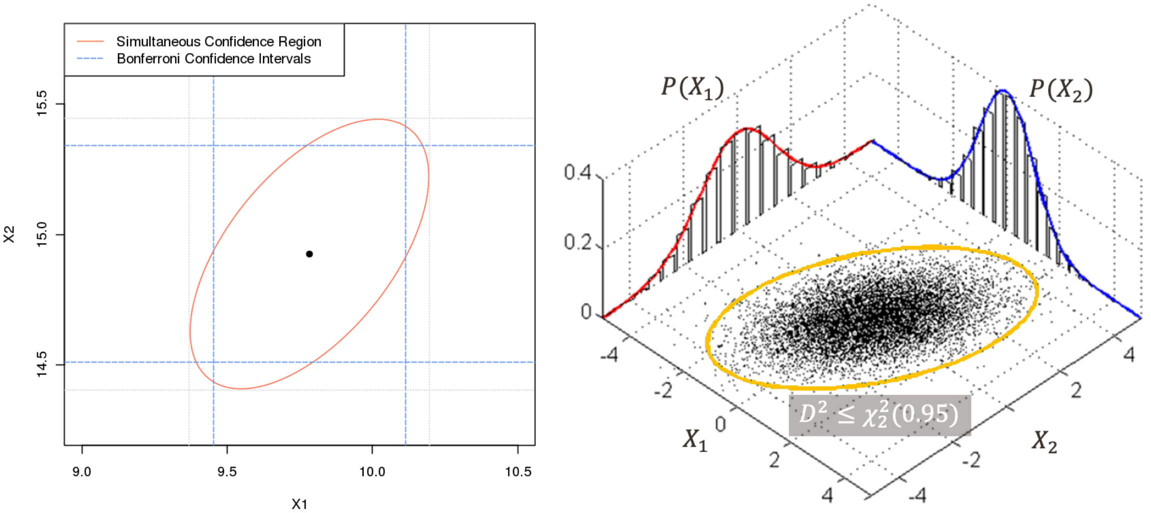


Figure 5. 95% confidence ellipse, Bonferroni and Hotelling's T^2 simulatenous confidence intervals for a bivariate normal random sample; showing the $p = 2$ framework for the Mahalanobis distance $D = \sqrt{(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})}$ [Wikipedia].

These properties make the multivariate normal distribution attractive, from a theoretical point of view (if not entirely realistic). For instance,

- using property 1, we can use **contrasts** to test which components are distinct from the others;
- property 5 is the multivariate analogue of the square of a standard normal random variable $Z \sim N(0,1)$ following a $Z^2 \sim \chi_1^2$ distribution;
- but two univariate normal random variables with zero covariance are not necessary independent (the joint p.d.f. of two such variables is not necessarily the p.d.f. of a multivariate normal distribution).

6.2 Hypothesis Testing for Mean Vectors

When the sample comes from a univariate normal distribution, we can test

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_1 : \mu \neq \mu_0$$

by using a $t-$statistic. Analogously, if the sample comes from a $p-$variate normal distribution, we can test

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu_0} \quad \text{against} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu_0}$$

by using **Hotelling's T^2 test statistic**

$$T^2 = N \cdot (\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$$

where $\bar{\boldsymbol{X}}$ denotes the **sample mean**, $\boldsymbol{S}$ the **sample covariance matrix**, and N the sample size.

Under H_0,

$$T^2 \sim \frac{(N-1)p}{(N-p)} F_{p,N-p}.$$

Thus, we do not reject H_0 at a significance level of α if

$$N \cdot (\bar{\boldsymbol{X}} - \boldsymbol{\mu_0})^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu_0}) \le \frac{(N-1)p}{(N-p)} F_{p,N-p}(\alpha)$$

and reject it otherwise.

6.3 Confidence Region and Simultaneous Confidence Intervals for Mean Vectors

In the $p-$variate normal distribution, any $\boldsymbol{\mu}$ that satisfies the condition

$$N \cdot (\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \le \frac{(N-1)p}{(N-p)} F_{p,N-p}(\alpha)$$

resides inside a $(1-\alpha)100\%$ **confidence region** (an ellipsoid in this case).

Simultaneous Bonferroni confidence intervals with overall error rate α can also be derived, using

$$(\bar{x}_j - \mu_j) \pm t_{N-1}(\alpha/p)\sqrt{\frac{s_{j,j}}{N}} \quad \text{for } j = 1,\ldots,p$$

Another approach is to use **Hotelling's T^2 simultaneous confidence intervals**, given by

$$(\bar{x}_j - \mu_j) \pm \sqrt{\frac{p(N-1)}{N-p} F_{p,N-p}(\alpha)}\sqrt{\frac{s_{j,j}}{N}} \quad \text{for } j = 1,\ldots,p$$

Figure 5 shows these regions for a bivariate normal random sample.

Note that the Hotelling's T^2 simultaneous confidence intervals form a rectangle (in grey) that confines the confidence region, while the Bonferroni confidence intervals (in blue) are slightly narrower.

Given that all the components of the mean vector are correlated (since the covariance matrix is generally non-diagonal), the confidence region should be used if the goal is to study the **plausibility of the mean vector as a whole**, while Bonferroni confidence intervals may be more suitable when **component-wise confidence intervals** are of needed.

Source	SSP	df	MSP	"$\mathbf{F}_0$"
Treatment	$\boldsymbol{B} = \sum_{i=1}^{I} N_i(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}})^\top$	$I-1$	$\boldsymbol{B}/(I-1)$	$\boldsymbol{W}^{-1}\boldsymbol{B}$
Error	$\boldsymbol{W} = \sum_{i=1}^{I} \sum_{j=1}^{n_i}(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)^\top$	$\sum_{i=1}^{I} N_i - I$	$\boldsymbol{W}/\sum_{i=1}^{I}(N_i-1)$	
Total	$\boldsymbol{B}+\boldsymbol{W} = \sum_{i=1}^{I} \sum_{j=1}^{n_i}(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}})(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}})^\top$	$\sum_{i=1}^{I} N_i - 1$	$\boldsymbol{B}+\boldsymbol{W}/(\sum_{i=1}^{I} N_i - 1)$	

Table 5. One-way MANOVA table; with I sub-populations.

7. Multivariate Analysis of Variance (MANOVA)

ANOVA is often used as a first attempt to determine whether the means from every sub-population are identical.

ANOVA can test means from more than two populations; the **multivariate ANOVA** (MANOVA) is quite simply a multivariate extension of ANOVA which tests whether the mean vectors from all sub-populations are identical.

Assume there are I sub-populations in the population, from each of which N_i p—dimensional responses are drawn, for $i = 1, \ldots, I$. Each observation can be expressed as:

$$\boldsymbol{X}_{i,j} = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \boldsymbol{\varepsilon}_{ij},$$

where $\boldsymbol{\mu}$ is the **overall mean vector**, $\boldsymbol{\tau}_i$ is the i^{th} **population-specific treatment effect**, and $\boldsymbol{\varepsilon}_{ij}$ is the **random error**, which follows a $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ distribution.

It is important to note that the covariance matrix $\boldsymbol{\Sigma}$ is assumed to be the same for each sub-population, and that

$$\sum_{i=1}^{I} N_i \boldsymbol{\tau}_i = \mathbf{0}$$

to ensure that the estimates are uniquely identifiable.

To test the hypothesis

$$H_0 : \boldsymbol{\tau}_1 = \cdots = \boldsymbol{\tau}_I = \mathbf{0} \quad \text{against} \quad H_1 : \text{some } \boldsymbol{\tau}_i \neq \mathbf{0},$$

we decompose the **total sum of squares and cross-products** SSP_{tot} into

$$\text{SSP}_{\text{tot}} = \text{SSP}_{\text{treat}} + \text{SSP}_{\text{e}}.$$

Based on this decomposition, we compute the test statistic known as **Wilks' lambda**

$$\Lambda^* = \frac{|\boldsymbol{W}|}{|\boldsymbol{B}+\boldsymbol{W}|},$$

where $\boldsymbol{B}, \boldsymbol{W}$ are as in Table 5, and reject H_0 if Λ^* is below some threshold, which depends on p, I, and N_i, $i = 1, \ldots, I$.

8. Goodness-of-Fit Tests

A (fictitious) 2017 survey asked a sample of $N = 200$ adults between the age of 25 to 35 about their highest educational achievement. The result is summarised in Table 6. In 1997, it was found that $p_1 = 13\%$ of adults had not complete high school, $p_2 = 32\%$ had obtained a high school degree but not a post-secondary degree, $p_3 = 37\%$ had either an undergraduate college or university diploma but no post-graduate degree, and $p_4 = 18\%$ had at least one post-graduate degree.

Year	<HS	HS	CU	CU+
2017	16%	55%	83%	46%
1997	13%	32%	37%	18%

Table 6. Respondents' educational achievements, from a (fictitious) survey, for 1997 and 2017.

Based on the result of this survey, is there sufficient evidence to believe that educational backgrounds of the population have changed since 1997?

Since each respondent's educational achievement can only be classified into one of these categories, they are **mutually exclusive**. Furthermore, these categories cover all possibilities on the educational front, so they are also **exhaustive**.

We can thus view the distribution of educational achievements as being **multinomial**. For such a distribution, with parameters $p_1, \cdots, p_k$, the expected frequency in each category is $m_j = Np_j$.

Let O_j denote the observed frequency for the j^{th} category. If there has been no real change since 1997, we would expect the sum of squared differences between the observed 2017 frequencies and the expected frequencies based on 1997 data to be small.

We can use this information to test the **goodness-of-fit** between the observations and the expected frequencies *via* Pearson's χ^2 test statistic

$$X^2 = \sum_{j=1}^{k} \frac{(O_j - m_j)^2}{m_j}$$

which follows a χ^2 distribution with $k-1$ df.

In the above example, the hypotheses of interest are

$$H_0 : \boldsymbol{p} = \boldsymbol{p}^* = (0.13, 0.32, 0.37, 0.18) \quad \text{vs} \quad H_1 : \boldsymbol{p} \neq \boldsymbol{p}^*.$$

Table 7 summarises the information under H_0.

Pearson's test statistic is $X^2 = 7.815$, with an associated p—value of 0.0295, which implies that there is enough statistical evidence (at the $\alpha = 0.05$ level) to accept that the population's educational achievements have changed over the last 20 years.

Category	O_j	$p_{j,0}$	$m_{j,0}$	$(O_j - m_{j,0})^2/m_{j,0}$
1	16	0.13	26	3.846
2	55	0.32	64	1.266
3	83	0.37	74	1.095
4	46	0.18	36	2.778
Total	200	1	200	7.815

Table 7. Summary table for goodness-of-fit data for educational achievements under H_0.

9. Analysis of Covariance (ANCOVA)

In Section 2, we looked at the effectiveness of new teaching method by assigning each group to a specific treatment and comparing the mean test scores. A crucial assumption for that model is that subjects in each group have **similar background knowledge** about statistics prior to the three week lectures.

If this assumption is wrong, however, we may be making incorrect decisions based on the model. Even if each group had similar background knowledge *on average*, there may be large variability from person-to-person, masking the true treatment effect.

9.1 Paired Comparison

One way to avoid such **subject-to-subject variability** is to administer both treatments to each individual, and then compare treatment effects by looking at the **difference in the outcomes**.

For instance, if a grocery chain is interested in measuring the effectiveness of two advertising campaigns, it could be reasonable to assume that there is a large variability in total sales, as well as popular items sold, at each store.

It may then be preferable to run both campaigns in each store and analyse the resulting data rather than to split the stores into two groups (in each of which a different advertising campaign is run) and then to compare the mean outcomes in the two groups.

Formally, let $X_{i,1}$ denote the total sales with campaign A and $X_{i,2}$ the total sales with campaign B. The quantity of interest is the **difference** $D_i = X_{i,1} - X_{i,2}$ for each store $i = 1, \dots, N$.

Assuming that the differences D_i follow an i.i.d. normal distribution with mean δ and variance σ_d^2, then we test for

$$H_0 : \delta = 0 \quad \text{against} \quad H_1 : \delta \neq 0$$

using the test statistic

$$t_0 = \sqrt{N}\frac{\bar{D}}{s_d},$$

which follows a Student's t distribution with $N-1$ degrees of freedom; thus we reject H_0 if the observed test statistic t_0 has p-value less than the significance level $\alpha/2$.

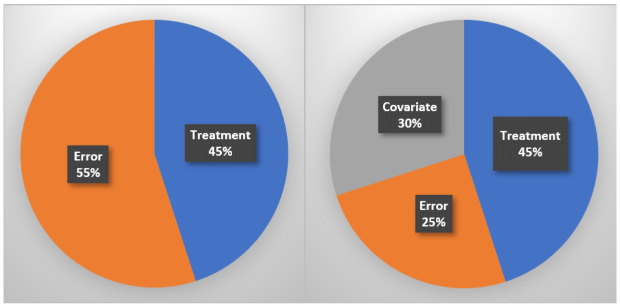


Figure 6. Breakdown of variability for ANOVA and ANCOVA.

9.2 Analysis of Covariance (ANCOVA)

ANOVA compares multiple group means and tests whether any of the group means differ from the rest, by breaking down the total variability into a treatment (explainable) variability component and an error (unexplained) variability component, and building a ratio F_0 to determine whether or not to reject H_0.

Analysis of covariance (ANCOVA) introduces **concomitant variables** (or **covariates**) to the ANOVA model, splitting the total variability into 3 components: SS_{treat}, SS_{con}, and SS_{e}, aiming to reduce error variability. The choice of covariates is thus crucial in running a successful ANCOVA.

In order to be useful, a concomitant variable must be related to response variable in some way, otherwise it not only fails to reduce error variability, but it also increases the model complexity:

- in the teaching method example, we could consider administering a pre-study test to measure the **prior knowledge level** of each participant and use this score as a concomitant variable;
- in the advertising campaign example, we could have used the **previous month's sales** as a covariate;
- in medical studies, we could use the **age** and **weight** of subjects, say.

Importantly, concomitant variables should not be affected by treatments. As an example, suppose that the patients in a medical study were asked:

> How strongly do you believe that you were given actual medication rather than a placebo?

If the treatment is indeed effective, then a participant's response to this question could be **markedly different** in the treatment group than in the placebo group.[4] This means that true treatment effect may be masked by concomitant variable due to unequal effects on treatment groups.

Note that **qualitative covariates** (such as gender, say) are not part of the ANCOVA framework – indeed, such covariates create new ANOVA treatment groups instead.

[4] The medication may have strong side-effects which cannot be ignored.

Figure 6 shows a potential breakdown of the total variability when moving from an ANOVA to an ANCOVA model – the error variability is further split into a **pure error** and a **covariate** component, while the **treatment** variability remains unchanged.

9.3 ANCOVA Model and Assumptions

Suppose that we are testing the effect of p treatments, with N_j subjects in each group. Then the ANCOVA model takes the form

$$y_{i,j} = \mu + \tau_j + \gamma(x_{i,j} - \bar{x}) + \varepsilon_{i,j} \tag{2}$$

where

- $y_{i,j}$ is the response of the i^{th} subject in the j^{th} treatment group;
- μ is the overall mean;
- τ_j is the j^{th} treatment effect, subject to a constraint
$$\sum_{j=1}^{p} \tau_j = 0;$$
- γ is the coefficient for the **covariate effect**;
- $(x_{i,j} - \bar{x})$ is the covariate value of the i^{th} subject in the j^{th} treatment group, adjusted by the mean, and
- $\varepsilon_{i,j}$ is the error of i^{th} subject in the j^{th} treatment group.

Additionally, four assumptions must be satisfied:

- **independence and normality of residuals** – the residuals follow an $i.i.d.$ normal distribution with mean of 0 and variance σ_ε^2;
- **homogeneity of residual variances** – the variance of the residuals is uniform across treatment groups;
- **homogeneity of regression slopes** – the regression effect (slope) is uniform across treatment groups, and
- **linearity of regression** – the regression relationship between the response and the covariate is linear.

The first of these assumptions can be tested with the help of a QQ-plot and a scatter-plot of residuals vs. fitted values, while the second may use the Bartlett or the Levene test. The final assumption is not as crucial as the other three assumptions. Various remedial methods can be applied should any of these assumptions fail.

The third assumption, however, is **crucial** to the ANCOVA model; it can be tested with the **equal slope test**, which requires an ANCOVA regression on equation (2) with an additional interaction term $x \times \tau$.

If the interaction is not significant, the third assumption is satisfied. In the event that the interaction term is statistically significant, a different approach (e.g. moderated regression analysis, mediation analysis) is required since using the original ANCOVA model is not prescribed.

An in-depth application of an ANCOVA model is highlighted in the next chapter.

10. Nonlinear Regression

From the use of tooth paste, cosmetics, cleaning solutions and so forth, we are exposed to numerous chemicals on a daily basis; thousands of new chemicals are introduced into commercial products each year, and government agencies (such as Health Canada and the Environmental Protection Agency in the U.S.) must determine whether these chemicals are safe for humans, animals, and the environment.

To test whether a chemical poses a risk of adverse effects, we must first determine whether it triggers adverse effects over a range of potential exposure levels, and if so, how much is considered safe (or how much would pose an unacceptable risk).[5]

Suppose that N laboratory rodents are divided into k groups, with group i consisting of N_i rodents. Over the course of the experiment, each group is given a certain amount of exposure to the chemical under investigation.

For each rodent, the experiment outcome records whether the rodent eventually develops a tumour or not; that is, the outcome is expressed as 0 (tumour absent) or 1 (tumour present).

Table 8 summarises the outcome of such an experiment.

Clearly, we cannot fit an ordinary linear regression to the data as the outcome is **dichotomous** (not a continuous variable). How could we then model the relationship between the adverse effect and the dose levels?

For each dose level d, the probability of adverse effect is $p_d = P(y = 1|d)$. The **conditional expectation** given the dose level is also $E(y = 1|d) = p_d$. Since the relationship resembles an S—shaped curve, we may use a logistic distribution to model the data:

$$E(y = 1|d) = p_d = \frac{\exp[\beta_0 + \beta_1 d]}{1 + \exp[\beta_0 + \beta_1 d]}$$

To obtain **maximum likelihood estimates** for β_0 and β_1, we need to rely on numerical methods such as the **Newton-Raphson method**; the dose-response model for the above example is shown in Figure 7 (on the left).

10.1 Relationship to Linear Regression

Since p_d is a probability, it has to lie in $[0,1]$. Let the odds of having an adverse effect be $\omega_d = p_d/(1-p_d)$; ω_d now lies in $[0,\infty)$, and the log odds $\ln \omega_d$ will span $\mathbb{R}$. The functional form of the **logistic regression model** is

$$\log(\omega_d) = \log\left(\frac{p_d}{1-p_d}\right) = \beta_0 + \beta_1 d,$$

which is a simple linear regression model.

[5] Traditionally (and not necessarily ethically), rodents were used to study whether a chemical is carcinogenic or not.

Dose Levels (d)	0	7000	15000	30000
Sample Size (n)	50	35	65	50
# of Observed Adverse Effect (y)	3	6	33	39
Rate of Observed Adverse Effect (p)	0.06	0.17	0.51	0.78

Table 8. Summary of experimental results involving C.I. Acid Red 114; $N = 200$.

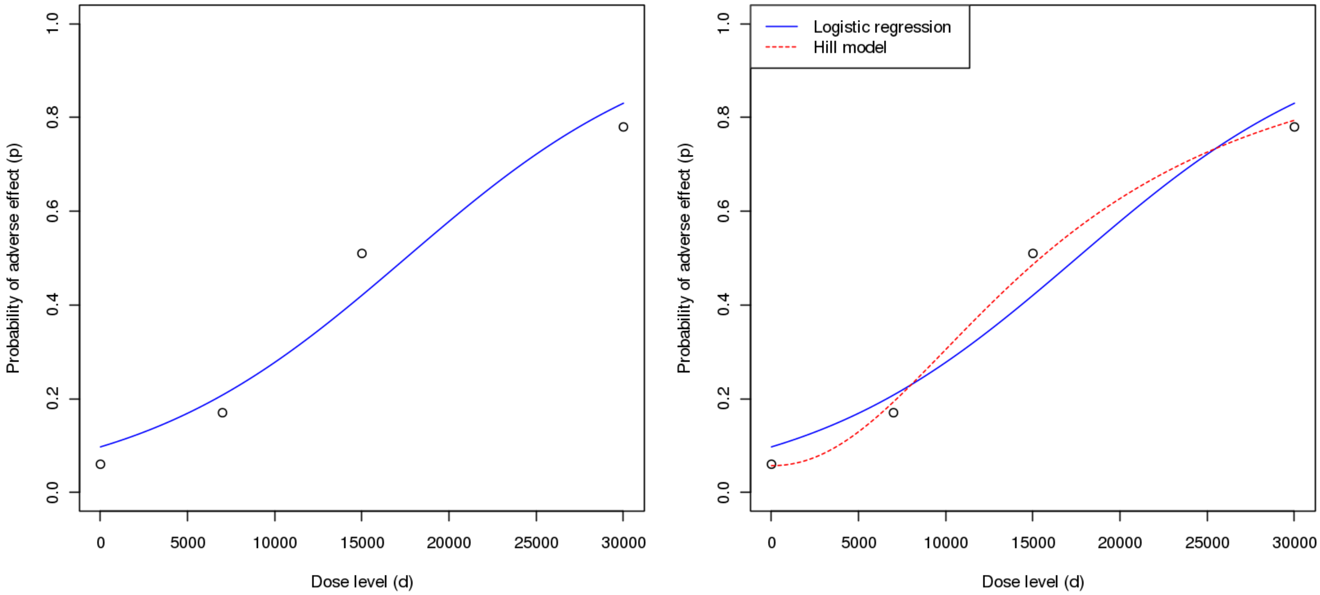


Figure 7. Dose-response model for C.I. Acid Red 114 using logistic regression (blue) and the Hill model (red).

10.2 Other Non-Linear Regression Models

Other **sigmoidal curves** can be used to model the relationship between predictors and a binary response variable.

Popular alternatives include:

- the **probit** link $P(y|x) = \Phi(\beta_0 + \beta_1 x)$, where Φ is the cumulative distribution function of the standard normal distribution, or
- the **complementary log-log** link

$$P(y|x) = 1 - \exp(-\exp(\beta_0 + \beta_1 x)).$$

In toxicology studies, one of the most widely used model is called the **Hill**model, and it is defined *via*

$$P(y|d, \alpha, \kappa, \eta) = \alpha + (1-\alpha)\frac{d^\eta}{d^\eta + \kappa^\eta};$$

part of its appeal to health scientists is the interpretation of its parameters – α represents the **background rate for adverse effect**, while κ denotes ED$_{50}$ (the **effective dose at which** 50% **of participants would exhibit the response of interest**) and η provides the **steepness of the dose-response curve**.

Figure 7 (on the right) compares the simple logistic model to the Hill model; we observe that the Hill model provides a closer fit to the observed proportions, and the curvature is more pronounced compared to the logistic model.

11. Bayesian Statistics

In classical statistics, model parameters such as μ and σ are treated as constants; **Bayesian statistics**, on the other hand assume that **model parameters are random variables**.

Bayes' Theorem lies at the foundation of such statistics:

$$P(H \mid D) = \frac{P(D \mid H) \times P(H)}{P(D)}, \tag{3}$$

where H represents the hypothesis and D denotes the observed data, which is sometimes written in shorthand as $P(H \mid D) \propto P(D \mid H) \times P(H)$; in other words, our **degree of belief in a hypothesis should be updated by the evidence provided by the data.**[6] More details are provided in [3].

Suppose we are interested in diagnosing whether a tumour is begin or malignant, based on several measurements obtained from video imaging. Bayes' Theorem (3) can be recast in a tumour data mould:

- **posterior:** $P(H \mid D) =$ based on collected data, how likely is a given tumour to be benign (or malignant)?

[6] Nobody disputes the validity of Bayes' Theorem, and it has proven to be a useful component in various models and algorithms, such as email spam filters, and the following example, but the **use** of Bayesian statistics is controversial in many quarters.

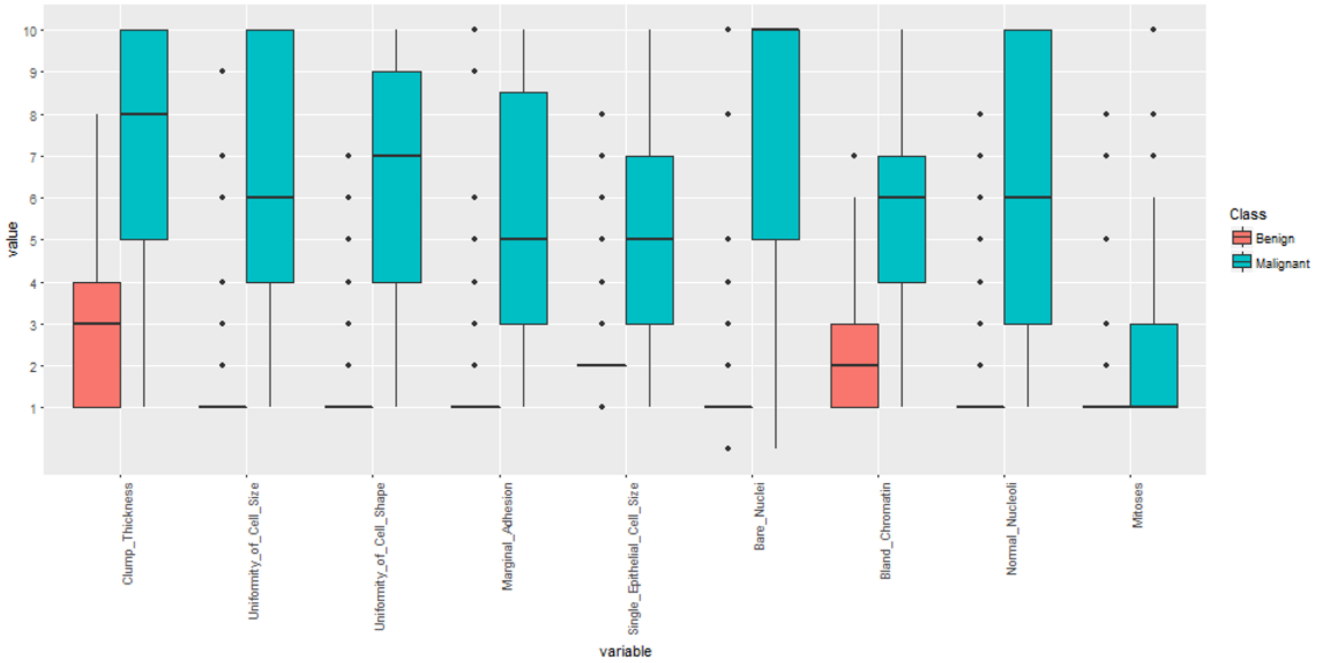


Figure 8. Boxplot visualisation of measurements for benign and malignant tumours.

Obs.	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
1	3	1	1	1	2	2	3	1	1

Table 9. Scores for an undiagnosed tumour.

- **prior:** $P(H)$ = in what proportion are tumours benign (or malignant) in general?
- **likelihood:** $P(D \mid H)$ = knowing a tumour is benign (or malignant), how likely is it that these particular measurements would have been observed?
- **evidence:** $P(D)$ = regardless of a tumour being benign or malignant, what is the chance that a tumour has the observed characteristics?

To answer the above question (that is, to compute the posterior), we will use a **naïve Bayes classifier** (NBC; see [1] for more information on classification methods).

11.1 Naïve Bayes Classification for Tumour Diagnoses

The procedure to apply NBC is straightforward.

1. **Objective function:** a simple way to determine whether a tumour is benign or malignant is to compare **posterior probabilities** and choose the one with highest probability. That is, we diagnose a tumour as **malignant** if

$$\frac{P(\text{malignant} \mid D)}{P(\text{benign} \mid D)} = \frac{P(D \mid \text{malignant}) \times P(\text{malignant})}{P(D \mid \text{benign}) \times P(\text{benign})} > 1,$$

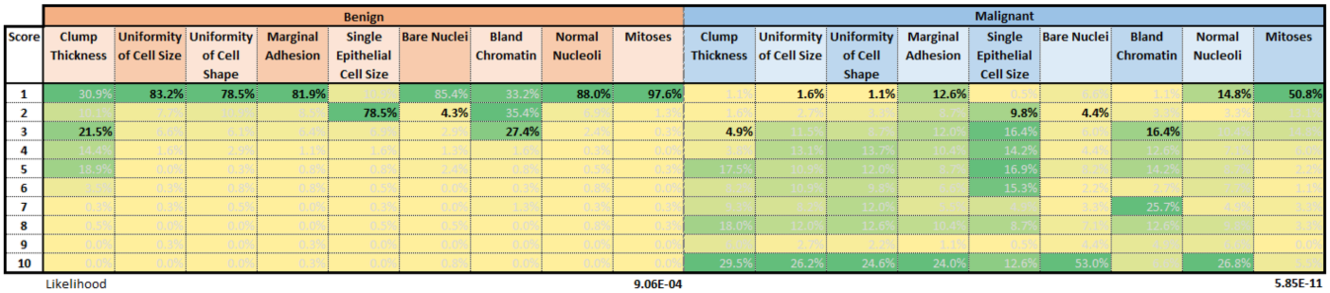
   and as **benign** otherwise.

2. **Dataset:** the classifier is built on a sample of $N = 458$ tumours with nine measurements, each scored on a scale of 1 to 10. The measurements include items such as *clump thickness* and *bare nuclei*; boxplots of these measurements are shown in Figure 8. We also have undiagnosed cases – an example of an **explanatory signature scores** is given in Table 9; this is an observation for which a prediction is required.

3. **Assumptions:** we assume that the scores of the measurements in each class are independent of one another (hence the **naive** qualifier); this assumption reduces the likelihood function to

$$\begin{aligned} P(D \mid H) &= P(x_1, x_2, \cdots, x_9 \mid H) \\ &= P(x_1 \mid H) \times \cdots P(x_9 \mid H). \end{aligned}$$

4. **Prior distribution:** we can ask subject matter experts to provide a rough estimate for the general ratio of benign to malignant tumours, or use the proportion of benign tumours in the sample as our prior. In situations where we have no knowledge about the distribution of priors, we may simply assume a **non-informative prior** (in this case, the prevalence rates would be the same for both responses).



	Benign									Malignant								
Score	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
1	30.9%	83.2%	78.5%	81.9%	10.9%	85.4%	33.2%	88.0%	97.6%	1.1%	1.6%	1.1%	12.6%	0.5%	6.6%	1.1%	14.8%	50.8%
2	10.1%	7.7%	10.9%	8.5%	78.5%	4.3%	35.4%	6.9%	1.3%	1.6%	2.7%	3.3%	8.7%	9.8%	4.4%	3.3%	3.3%	13.1%
3	21.5%	6.6%	6.1%	6.4%	6.9%	2.9%	27.4%	2.4%	0.3%	4.9%	11.5%	8.7%	12.0%	16.4%	6.0%	16.4%	10.4%	14.8%
4	14.4%	1.6%	2.9%	1.1%	1.6%	1.3%	1.6%	0.3%	0.0%	3.8%	13.1%	13.7%	10.4%	14.2%	4.4%	12.6%	7.1%	6.0%
5	18.9%	0.0%	0.3%	0.8%	0.8%	2.4%	0.8%	0.5%	0.3%	17.5%	10.9%	12.0%	8.7%	16.9%	8.2%	14.2%	8.7%	2.2%
6	3.5%	0.3%	0.8%	0.8%	0.5%	0.0%	0.3%	0.8%	0.0%	8.2%	10.9%	9.8%	6.6%	15.3%	2.2%	2.7%	7.7%	1.1%
7	0.3%	0.3%	0.5%	0.0%	0.3%	0.0%	1.3%	0.3%	0.3%	9.3%	8.2%	12.0%	5.5%	4.9%	3.3%	25.7%	4.9%	3.3%
8	0.5%	0.0%	0.0%	0.0%	0.5%	0.5%	0.0%	0.8%	0.3%	18.0%	12.0%	12.6%	10.4%	8.7%	7.1%	12.6%	9.8%	3.3%
9	0.0%	0.3%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	6.0%	2.7%	2.2%	1.1%	0.5%	4.4%	4.9%	6.6%	0.0%
10	0.0%	0.0%	0.0%	0.3%	0.0%	0.8%	0.0%	0.0%	0.0%	29.5%	26.2%	24.6%	24.0%	12.6%	53.0%	6.6%	26.8%	5.5%
Likelihood									9.06E-04									5.85E-11

Figure 9. Multinomial probabilities for benign and malignant tumours.

Class	Prior	Likelihood	Posterior	Ratio
Malignant	0.327	5.85×10^{-11}	1.92×10^{-11}	3.15×10^{-8}
Benign	0.673	9.06×10^{-4}	6.09×10^{-4}	

Table 10. Computation of posterior probabilities in the undiagnosed case of Table 9.

5. **Computation of likelihoods:** under independence, each measurement is assumed to follow a multinomial distribution (since scores are on $1-10$ scale). Multiplying probabilities from each multinomial distribution (one each for both classes) provides the overall likelihoods for benign and malignant tumours, respectively. The likelihood of the undiagnosed case being a benign tumour is seen to be 9.06×10^{-4}, while the likelihood of being a malignant tumour is 5.85×10^{-11}, based on the multinomial probabilities given in Table 9

6. **Computation of Posterior:** Multiplying the prior probability and likelihood, we get a quqntity that is proportional to the respective posterior probabilities. Looking at Table 10, we conclude that the tumour in the undiagnosed case is **likely benign** (note that we have no measurement on how much more likely it is to be benign than to be malignant – the classifier is **not calibrated**).

12. Resampling Methods

How do we determine the variability of a regression fit? It can be done by drawing different samples from the available data, fitting a regression model to each sample, and then examining the extent to which the various fits differ from one another.

Resampling methods provide additional information about a fitted model, by applying the same fitting approach to various sub-samples of the dataset. Following [5], we will consider two such methods:

- **cross-valiation**, which is used to estimate the test error associated with a modeling approach in order to evaluate model performance, and
- the **bootstrap**, which is used to provide a measure of accuracy, standard deviation, bias, etc. of various model parameter estimates.

12.1 Cross-Validation

For quantitative responses, the **test error** associated with a statistical model is the average error arising when predicting the response for observations that were not used to train the model.

The **training error**, on the other hand, is computed directly by comparing the model's predictions to the actual responses in the dataset.

In general, the training error underestimates the test error, dramatically so when the model complexity increases (i.e. the **variance-bias trade-off**, see Figure 10).

The **validation approach** is a simple strategy that is used to estimate the test error associated with a particular statistical model on a set of observations.

The latter is split into a **training set** and a **validation set** (also known as a hold-out set). The model is fit on the training set; the fitted model is used to make predictions on the validation set. The resulting validation set error provides an estimate for the test error.

This approach is easy to implement and interpret, but it has a number of drawbacks, most importantly:

- the validation error is highly dependent on the choice of the validation set, and is thus quite volatile;
- the model is fitted on a proper subset of the available observations, and we might expect that this would leadd to the validation error being larger than the test error in general;
- a number of classical statistical models can provide test error estimates without having to resort to the validation set approach.

*K***-Fold Cross Validation** is a widely-used approach to estimate the test error without losing some observations to a hold-out test.[7] The procedure is simple:

[7] It can also provide a basis for model selection.

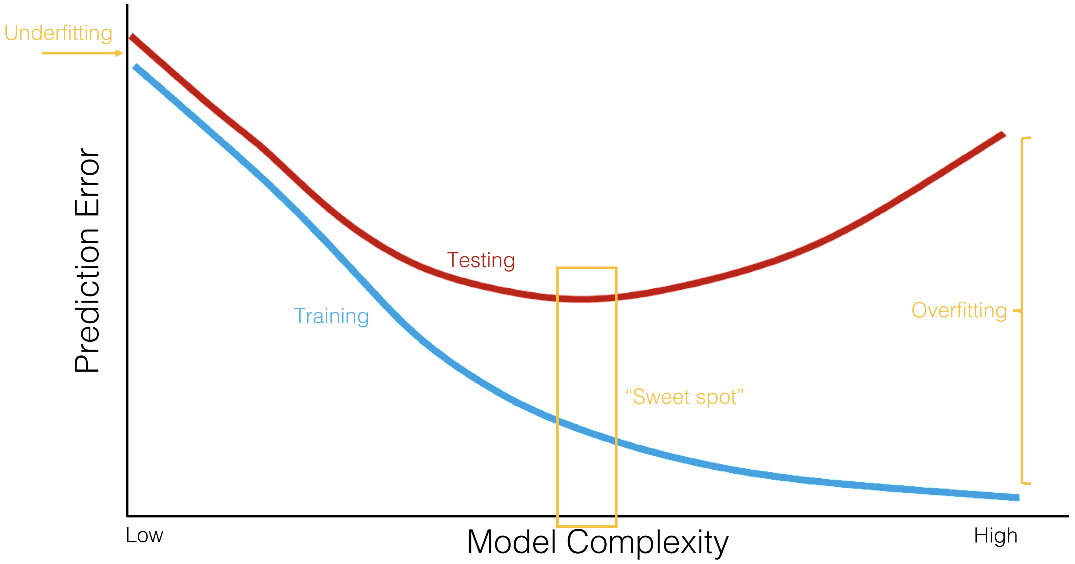


Figure 10. Underfitting and overfitting as a function of model complexity; error prediction on training sample (blue) and testing sample (red). High error prediction rates for simple models are a manifestation of underfitting; large difference between error prediction rates on training and testing samples for complex models are a manifestation of overfitting. Ideally, model complexity would be chosen to reach the situation's "sweet spot", but fishing for the ideal scenario might diminish explanatory power (based on [5]).

1. Divide the dataset **randomly** into K (roughly) equal-sized **folds** (typically, $K = 4, 5, 10$).
2. Each fold plays, in succession, the role of the **validation set**. If there are N observations in the dataset, partition
$$\{1,\ldots,N\} = \underbrace{\mathscr{C}_1}_{\text{fold 1}} \sqcup \cdots \sqcup \underbrace{\mathscr{C}_K}_{\text{fold } K}.$$
If $|\mathscr{C}_k| = n_k$, we expect $n_k \approx \frac{N}{K}$ for all $k = 1,\ldots,K$.
3. For all $k = 1,\ldots,K$, fit a model on $\{1,\ldots,N\} \setminus \mathscr{C}_k$ and denote the error on $\mathscr{C}_k$ by E_k.[8]
4. Write $\overline{E}$ for the average error on each of the folds.
5. The **cross-validation estimate** of the test error is
$$\text{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{N} E_k,$$
with standard error
$$\widehat{\text{se}}\left(\text{CV}_{(K)}\right) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(E_k - \overline{E})^2}.$$

These steps could also be **replicated** n times to generate a distribution of an evaluation metric (such as the standard error), see Figure 11 for an illustration.

12.2 The Bootstrap

The **bootstrap procedure** uses re-sampling of the available data to **mimic the process of obtaining new replicates**, which allows us to estimate the variability of a statistical model parameter of interest (a coefficient in a non-linear regression, a target optimization value, etc.) **without the need to generate new observations**.

Replicates are obtained by repeatedly sampling observations from the original dataset **with replacement**. A **bootstrap dataset** $\mathscr{X}^*$ for a training set $\mathscr{X}$ with N observations is a sample of N observations drawn from $\mathscr{X}$, with replacement.

The process is repeated M times to obtain bootstrap samples $\mathscr{X}_i^*$ and parameter estimates $\hat{\alpha}_i^*$, for $i = 1,\ldots,M$, from which we derive a **bootstrap estimate**
$$\hat{\alpha}^* = \frac{1}{M}\sum_{i=1}^{M}\hat{\alpha}_i^*,$$
with standard error
$$\widehat{\text{se}}\,(\hat{\alpha}^*) = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(\hat{\alpha}_i^* - \hat{\alpha}^*)^2}.$$

The bootstrap can also be used to build **approximate frequentist confidence intervals** for the parameter α, but there are complications, so caution is advised.

Finally, it should be noted that the appropriate bootstrap procedure might be more sophisticated than what has been described here in more complex scenarios. For instance, sampling with replacement at the observation level would not preserve the covariance structure of time series data.

[8] For a regression model, there are many options but we typically use
$$E_k = \sum_{i \in \mathscr{C}_k} \frac{(y_i - \hat{y}_i)^2}{n_k}.$$

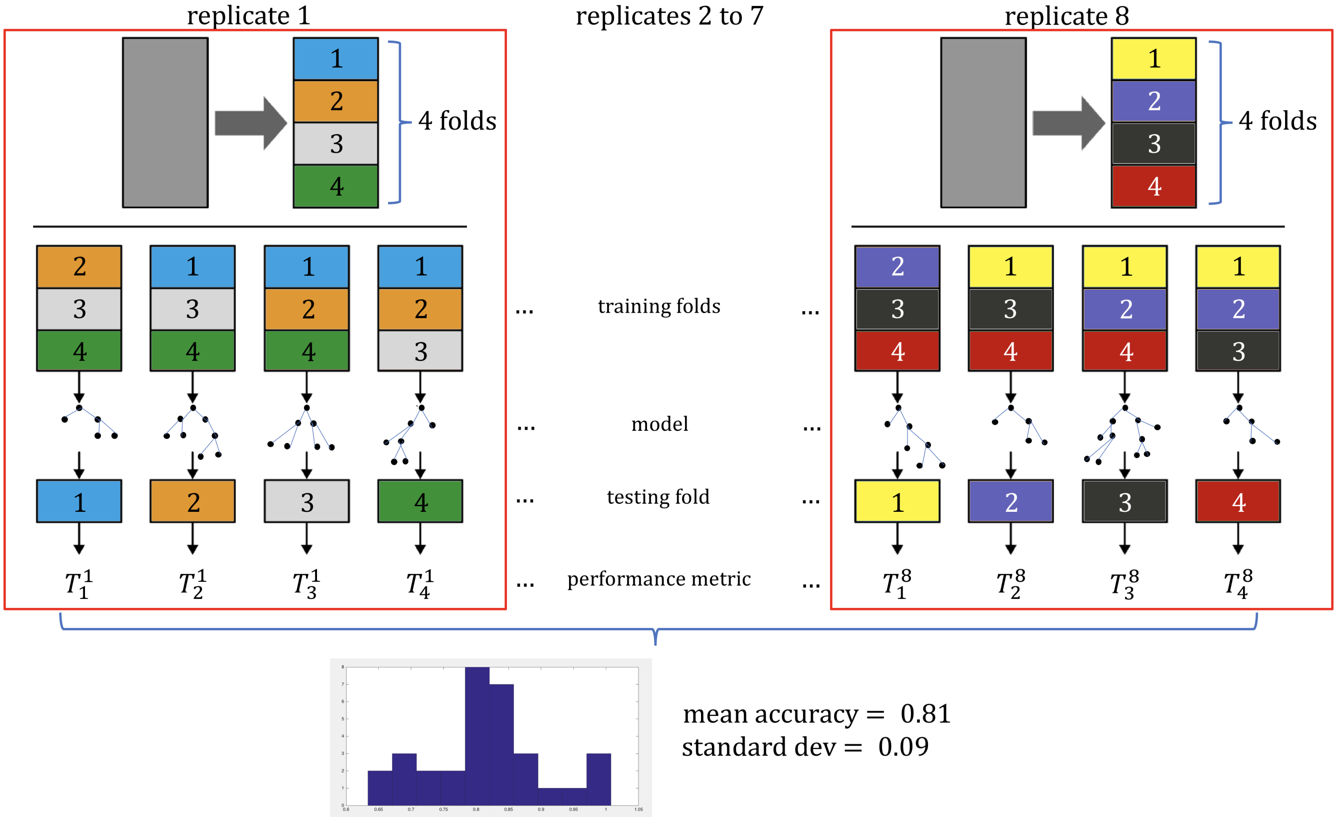


Figure 11. Schematic illustration of cross-fold validation, for 8 replicates and 4 folds; $8 \times 4 = 32$ models from a given family are built on various training sets (consisting of 3/4 of the available data – the training folds). Model family performance is evaluated on the respective holdout folds; the distribution of the performance metric values (in practice, some combination of the mean/median and standard deviation) can be used to compare various model families (based on [1,5]).

References

[1] P. Boily and J. Schellinck. Machine Learning 101. *Data Science Report Series*, 2021.

[2] P. Bruce and A. Bruce. *Practical Statistics for Data Scientists: 50 Essential Concepts*. O'Reilly, 2017.

[3] E. Ghashim and P. Boily. A Soft Introduction to Bayesian Data Analysis. *Data Science Report Series*, 2020.

[4] M. Hollander and D. Wolfe. *Nonparametric Statistical Methods*. Wiley, 2nd edition, 1999.

[5] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning, with Applications in R*. Springer, 2013.

[6] A. Reinhart. *Statistics Done Wrong: the Woefully Complete Guide*. No Starch Press, 2015.

[7] M. Rizzo. *Statistical Computing with R*. CRC Press, 2007.

[8] H. Sahai and M. Ageel. *The Analysis of Variance: Fixed, Random and Mixed Models*. Birkhäuser, 2000.

[9] D. Sivia and J. Skilling. *Data Analysis: A Bayesian Tutorial (2nd ed.)*. Oxford Science, 2006.