Visualizing Multivariate Data and Models in R

Michael Friendly

2023-09-10

Table of contents

Preface **6**
ONE, TWO, MANY 6
Flatland 7
EUREKA! 10
What I assume 12
Conventions used in this book 13
References 13

1 Introduction **14**
1.1 Why use a multivariate design 14
1.2 Linear models: Univariate to multivariate 15
1.3 Visualization is harder 16
1.4 Problems in understanding and communicating MLM results 18
References 19

2 Getting Started **20**
2.1 Why plot your data? 20
2.1.1 Anscombe's Quartet 21
2.1.2 A real example 25
2.2 Plots for data analysis 29
2.3 Data plots 30
2.4 Model plots 30
2.5 Diagnostic plots 30
References 30

3 Plots of Multivariate Data **31**
3.1 Bivariate summaries 31
3.1.1 The Data Ellipse 31
3.2 R functions for data ellipses 36
3.3 Quantitative data: 36
3.4 Categorical data: 36
3.5 Generalized pair plots 37

References 37

4 PCA and Biplots **38**
4.1 Principal components analysis 38
4.2 Biplots 38

5 Overview of Linear models **39**
5.1 Regression 39
5.2 ANOVA 39
5.3 ANCOVA 39
References 39

6 Plots for univariate response models **40**
6.1 The "regression quartet" 40
6.2 Other Diagnostic plots 47
6.2.1 Spread-level plot 47
6.3 Coefficient plots 47
6.4 Added-variable plots 47
6.5 Marginal plots 47
6.6 Outliers, leverage and influence 47
6.6.1 The leverage-influence quartet 48
6.6.2 Measuring leverage 52
6.6.3 Outliers: Measuring residuals 56
6.6.4 Measuring influence 57

7 Collinearity & Ridge Regression **59**
7.1 What is collinearity? 60
7.1.1 Visualizing collinearity 61
7.1.2 Data space and β space 63
7.2 Measuring collinearity 66
7.2.1 Variance inflation factors 66
7.2.2 Collinearity diagnostics 69
7.2.3 Tableplots 72
7.2.4 Collinearity biplots 74
7.3 Remedies for collinearity: What can I do? 77
7.4 Ridge regression 82
7.4.1 What is ridge regression? 82
7.4.2 Univariate ridge trace plots 82
7.4.3 Bivariate ridge trace plots 82
References 82

8 Hotelling's T^2 **83**
8.1 T^2 as a generalized t-test 84
8.2 T^2 properties 85
Example 87
8.3 HE plot and discriminant axis 90
8.3.1 `heplot()` 92
8.4 Discriminant analysis 94
8.5 Exercises 97
References 97

9 Visualizing Multivariate Models **98**
9.1 HE plot framework 98
9.1.1 HE plot details 99
9.2 Canonical discriminant analysis 99

10 Brief review of the multivariate linear model **100**
10.1 ANOVA -> MANOVA 101
10.2 MRA -> MMRA 101
10.3 ANCOVA -> MANCOVA 101
10.4 Repeated measures designs 101

11 Case studies **102**
11.1 Neuro- and Social-cognitive measures in psychiatric groups 102
11.1.1 Research questions 103
11.1.2 Data 103
11.1.3 A first look 105
11.1.4 Bivariate views 106
11.2 Fitting the MLM 110
11.2.1 HE plot 111
11.2.2 Canonical space 113
11.3 Social cognitive measures 115
11.3.1 Model checking 117
11.3.2 Canonical HE plot 119
References 122

12 Visualizing Tests for Equality of Covariance Matrices **123**
12.1 Homogeneity of Variance in Univariate ANOVA . 124
12.2 Homogeneity of variance in MANOVA 125
12.3 Assessing heterogeneity of covariance matrices: Box's M test 127

12.4 Visualizing heterogeneity 128
12.5 References 129

13 Summary **130**

References **131**

Preface

This is an early draft of material that may or may not appear in the preface.

ONE, TWO, MANY

There is an old and helpful idea I learned from John Hartigan in my graduate days at Princeton:

> In statistics and data visualization *all* methods can be classified by the number of dimensions contemplated, on a scale of **ONE**, **TWO**, **MANY**.

By this, he meant that, at a global level, all data, statistical summaries, and graphical displays could be classified as:

- **univariate**: a single variable, considered in isolation (age, COVID cases, pizzas ordered). Univariate numerical summaries are means, medians, measures of variablilty, and so forth. Univariate displays include dot plots, boxplots, histograms and density estimates.
- **bivariate**: two variables, considered jointly. Numerical summaries include correlations, covariances and two-way tables of frequencies or measures of association for categorical variables. Bivariate displays include scatterplots and mosaic plots.
- **multivariate**: three or more variables, considered jointly. Numerical summaries include correlation and covariance matrices, consisting of all pairwise values, but also derived measures from the analysis of these matrices (eigenvalues, eigenvectors). Graphical displays of multivariate data can sometimes be shown in 3D, but often involve multiple views of the data projected into 2D plots.

As a quasi-numerical scale, I refer to these as **1D**, **2D** and **nD**. This admits the possibility of half-integer cases, such as 1.5D, where the main focus is on a single variable, but it is classified by a simple factor (gender). His point in this classification was that once you reached three variables, all higher dimensions involved similar summaries and data displays.

Univariate and bivariate methods and displays are well-known. This book is about how these ideas can be extended to an n-dimensional world. Three-dimensional data displays are now fairly easy to produce, even if they are sometimes difficult to understand. But how can we even think about four or more dimensions? The difficulty can be appreciated by considering the tale of *Flatland*.

Flatland

> To comport oneself with perfect propriety in Polygonal society, one ought to be a Polygon oneself. — Edwin A. Abbott, *Flatland*

In 1884, an English schoolmaster, Edwin Abbott Abbott, shook the world of Victorian culture with a slim volume, *Flatland: A Romance of Many Dimensions* (Abbott 1884). He described a two-dimensional world, *Flatland*, inhabited entirely by geometric figures in the plane. His purpose was satirical, to poke fun at the social and gender class system at the time: Women were mere line segments, while men were represented as polygons with varying numbers of sides— a triangle was a working man; gentlemen and professionals had more sides. Abbot published this under the pseudonym, "A Square", suggesting his place in the hierarchy.

> True, said the Sphere; it appears to you a Plane, because you are not accustomed to light and shade and perspective; just as in Flatland a Hexagon would appear a Straight Line to one who has not the Art of Sight Recognition. But in reality it is a Solid, as you shall learn by the sense of Feeling. — Edwin A. Abbott, *Flatland*

But how did it feel to be a member of a flatland society? How could a point (a child?) understand a line (a woman)? How does a Triangle "see" a Hexagon or even a infinitely-sided Circle? Abbott introduces these ideas through dreams and visions:

- A Square dreams of visiting a one-dimensional *Lineland* where men appear as lines, and women are merely "illustrious points", but the inhabitants can only see the Square as lines.
- In another vision, the Square is visited by a Sphere, to illustrate what a 2D flatlander could understand from a 3D sphere.

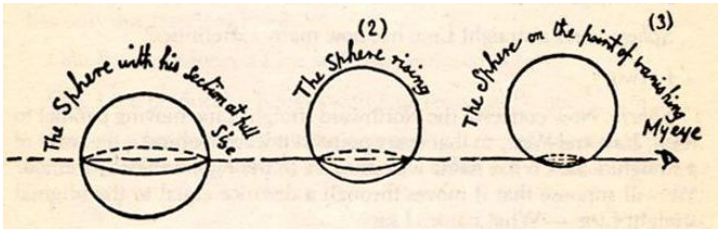

Figure 1: A 2D flatlander seeing a sphere pass through Flatland. Source: Abbott (1884)

Abbott goes on to state what could be considered as a demonstration (or proof) by induction of the difficulties of seeing in 1, 2, 3 dimensions, and how the idea motion over time (one more dimension) could allow citizens of any 1D, 2D, 3D world to contemplate one more dimension.

> In One Dimensions, did not a moving Point produce a Line with two terminal points? In two Dimensions, did not a moving Line produce a Square with four terminal points? In Three Dimensions, did not a moving Square produce - did not the eyes of mine behold it - that blessed being, a Cube, with eight terminal points? And in Four Dimensions, shall not a moving Cube - alas, for Analogy, and alas for the Progress of Truth if it be not so - shall not, I say the motion of a divine Cube result in a still more divine organization with sixteen terminal points? — Edwin A. Abbott

For Abbot, the way for a citizen of any world to image one more dimension was to consider how a higher-dimensional object would change over time.[1] [2]: In his famous TV series, *Cosmos*, Carl Sagan provides an intriguing video presentation Flatland and the 4th dimension. However, as far back as 1754 (Cajori 1926), the idea of adding a fourth dimension appears in Jean le Rond d'Alembert's "Dimensions", and one realization of a four-dimensional object is a *tesseract*, shown in the figure below.

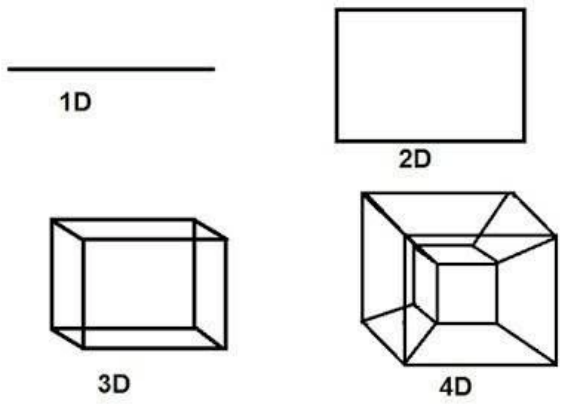

Figure 2: Geometrical object in 1 to 4 dimensions

But to really see a tesseract you have to view it in an animation over time: ::: {#fig-tesseract}

Animation of a tesseract. :::

Yet the deep mathematics of more than three dimensions only emerged in the 19th century. In Newtonian mechanics, space and time were always considered independent of each other. Our familiar three-dimensional space, of length, width, and height

[1] If group sizes are greatly unequal **and** homogeneity of variance is violated, then the F statistic is too liberal (p values too large) when large sample variances are associated with small group sizes. Conversely, the F statistic is too conservative if large variances are associated with large group sizes.

[2] If group sizes are greatly unequal **and** homogeneity of variance is violated, then the F statistic is too liberal (p values too large) when large sample variances are associated with small group sizes. Conversely, the F statistic is too conservative if large variances are associated with large group sizes.

had formed the backbone of Euclidean geometry for millenea. However, the idea that space and time are indeed interwoven was first proposed by German mathematician Hermann Minkowski (1864–1909) in 1908. This was a powerful idea. It bore fruit when Albert Einstein revolutionized the Newtonian conceptions of gravity in 1915 when he presented a theory of general relativity which was based primarily on the fact that mass and energy warp the fabric of four-dimensional spacetime.

The parable of *Flatland* can provide inspiration for statistical thinking and data visualization. Once we go beyond bivariate statistics and 2D plots, we are in a multivariate world of possibly MANY dimensions. It takes only some imagination and suitable methods to get there.

EUREKA!

Even modest sized multivariate data can have secrets that can be revealed in the right view. As an example, David Coleman at RCA Laboratories in Princeton, N.J. generated a data set of five (fictitious) measurements of grains of pollen for the 1986 Data Exposition at the Joint statistical Meetings. The first three variables are the lengths of geometric features 3848 observed sampled pollen grains – in the x, y, and z dimensions: a `ridge` along x, a `nub` in the y direction, and a `crack` in along the z dimension. The fourth variable is pollen grain `weight`, and the fifth is `density`. The challenge was to "find something interesting" in this data set.

Those who solved the puzzle were able to find an orientation of this 5-dimensional data set, such that zooming in revealed a magic word, "EUREKA" spelled in points, as in the following figure.

This can be seen better in a 3D animation. `rgl` (Adler and Murdoch 2023) is used to create a 3D scatterplot of the first three variables. Then the `animation` package (Xie 2021) is use to record a sequence of images, adjusting the `rgl::par3d(zoom)` value.

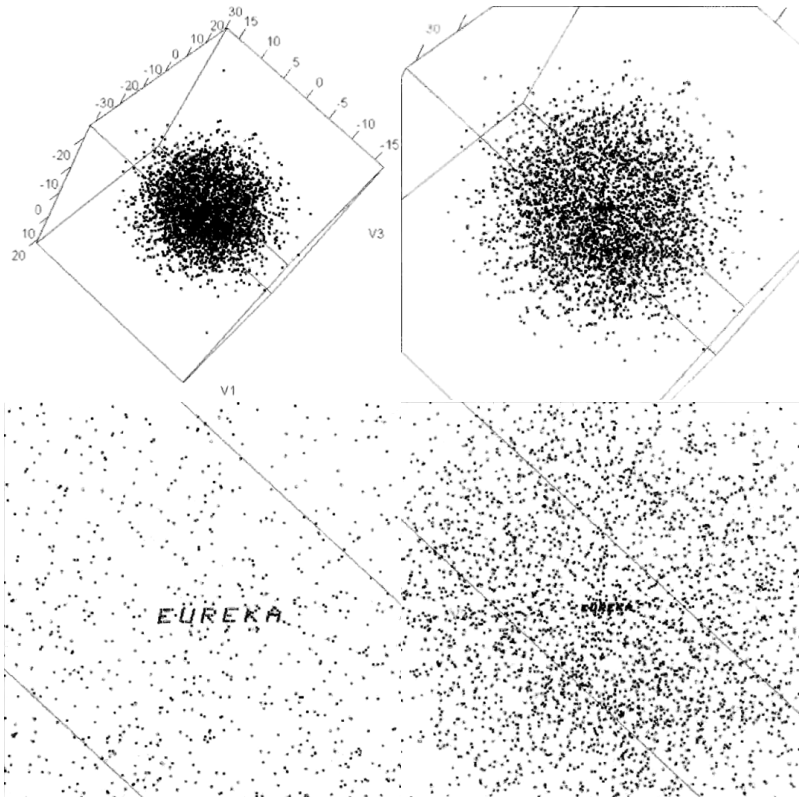

Figure 3: Four views of the `pollen` data, zooming in, clockwise from the upper left to discover the word "EUREKA".

```
library(animation)
library(rgl)
data(pollen, package = "animation")
oopt = ani.options(interval = 0.05)
## adjust the viewpoint
uM =
  matrix(c(-0.370919227600098, -0.513357102870941,
           -0.773877620697021, 0, -0.73050606250763, 0.675815105438232,
           -0.0981751680374146, 0, 0.573396027088165, 0.528906404972076,
           -0.625681936740875, 0, 0, 0, 0, 1), 4, 4)
open3d(userMatrix = uM,
       windowRect = c(10, 10, 510, 510))

plot3d(pollen[, 1:3])

# zoom in
zm = seq(1, 0.045, length = 200)
par3d(zoom = 1)
for (i in 1:length(zm)) {
  par3d(zoom = zm[i])
  ani.pause()
}
ani.options(oopt)
```

Figure 4: Animation of zooming in on the `pollen` data.

What I assume

I assume the reader to have at least a basic familiarity with R. While R fundamentals are outside the scope of the current paper, I believe that this language provides a rich set of resources, far beyond that offered by other statistical software packages, and is well worth learning.

For those not familiar with R, I recommend Matloff (2011), Wickham (2014), and Cotton (2013) for introductions to programming in the language, and Fox and Weisberg (2018) and

Teetor (2011) for learning about how to conduct basic statistical analyses.

Conventions used in this book

The following typographic conventions are used in this book:

- *italic* : indicates terms to be *emphasized* or defined in the text, ...
- **bold** : is used for names of R packages (well, not so far)
- `fixed-width` : is used in program listings as well as in text to refer to variable and function names, R statement elements and keywords.
- *`fixed-width italic`* : isn't used yet, but probably should be.

For R functions in packages, we use the notation `package::function()`, for example: `car::Anova()` to identify where those functions are defined

References

1 Introduction

This material may or may not survive; it was taken from an earlier article.

1.1 Why use a multivariate design

A particular research outcome (e.g., depression, neuro-cognitive functioning, academic achievement, self-concept, attention deficit hyperactivity disorders) might take on a multivariate form if it has several observed measurement scales or related aspects by which it is quantified, or if there are multiple theoretically distinct outcomes that should be assessed in conjunction with each other (e.g., using depression, generalized anxiety, and stress inventories to model overall happiness). In this situation, the primary concern of the researcher is to ascertain the impact of potential predictors on two or more response variables simultaneously.

For example, if academic achievement is measured for adolescents by their reading, mathematics, science, and history scores, the following questions are of interest:

- Do predictors such as parent encouragement, socioeconomic status and school environmental variables affect *all* of these outcomes?
- Do they affect them in the *same* or *different* ways?
- How many different aspects of academic achievement can be distinguished in the predictors? Equivalently, is academic achievement *unidimensional* or *multidimensional* in relation to the predictors?

Similarly, if psychiatric patients in various diagnostic categories are measured on a battery of tests related to social skills and cognitive functioning, we might want to know:

- Which measures best discriminate among the diagnostic groups?
- Which measures are most predictive of positive outcomes?
- Further, how are the *relationships* between the outcomes affected by the predictors?

Such questions obviously concern more than just the separate univariate relations of each response to the predictors. Equally, or perhaps more importantly, are questions of how the response variables are predicted *jointly*.

> **Note**
>
> Structural equation modeling (SEM) offers another route to explore and analyze the relationships among multiple predictors and multiple responses. They have the advantage of being able to test potentially complex systems of linear equations in very flexible ways; however, these methods are often far removed from data analysis *per se* and except for path diagrams offer little in the way of visualization methods to aid in understanding and communicating the results. The graphical methods we describe here can also be useful in a SEM context.

1.2 Linear models: Univariate to multivariate

For classical linear models for ANOVA and regression, the step from a univariate model for a single response, y, to a multivariate one for a collection of p responses, $\mathbf{y}$ is conceptually very easy. That's because the univariate model,

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + u_i,$$

or, in matrix terms,

$$\mathbf{y} = \mathbf{X}\ \beta + \mathbf{u}, \quad \text{with} \quad \mathbf{u} \sim \mathcal{N}(0, \sigma^2\mathbf{I}),$$

generalizes directly to an analogous multivariate linear model (MLM),

$$\mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_p}] = \mathbf{X}\ \mathbf{B} + \mathbf{U} \quad \text{with} \quad \mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$$

for multiple responses (as will be discussed in detail). The design matrix, $\mathbf{X}$ remains the same, and the vector β of coefficients becomes a matrix $\mathbf{B}$, with one column for each of the p outcome variables.

Happily as well, hypothesis tests for the MLM are also straightforward generalizations of the familiar F and t-tests for univariate response models. Moreover, there is a rich geometry underlying these generalizations which we can exploit for understanding and visualization.

1.3 Visualization is harder

However, with two or more response variables, visualizations for multivariate models are not as simple as they are for their univariate counterparts for understanding the effects of predictors, model parameters, or model diagnostics. Consequently, the results of such studies are often explored and discussed solely in terms of coefficients and significance, and visualizations of the relationships are only provided for one response variable at a time, if at all. This tradition can mask important nuances, and lead researchers to draw erroneous conclusions.

The aim of this book is to describe and illustrate some central methods that we have developed over the last ten years that aid in the understanding and communication of the results of multivariate linear models (Friendly 2007; Friendly and Meyer 2016). These methods rely on *data ellipsoids* as simple, minimally sufficient visualizations of variance that can be shown

in 2D and 3D plots. As will be demonstrated, the *Hypothesis-Error (HE) plot* framework applies this idea to the results of multivariate tests of linear hypotheses.

Further, in the case where there are more than just a few outcome variables, the important nectar of their relationships to predictors can often be distilled in a multivariate juicer— a **projection** of the multivariate relationships to the predictors in the low-D space that captures most of the flavor. This idea can be applied using *canonical correlation plots* and with *canonical discriminant HE plots*.

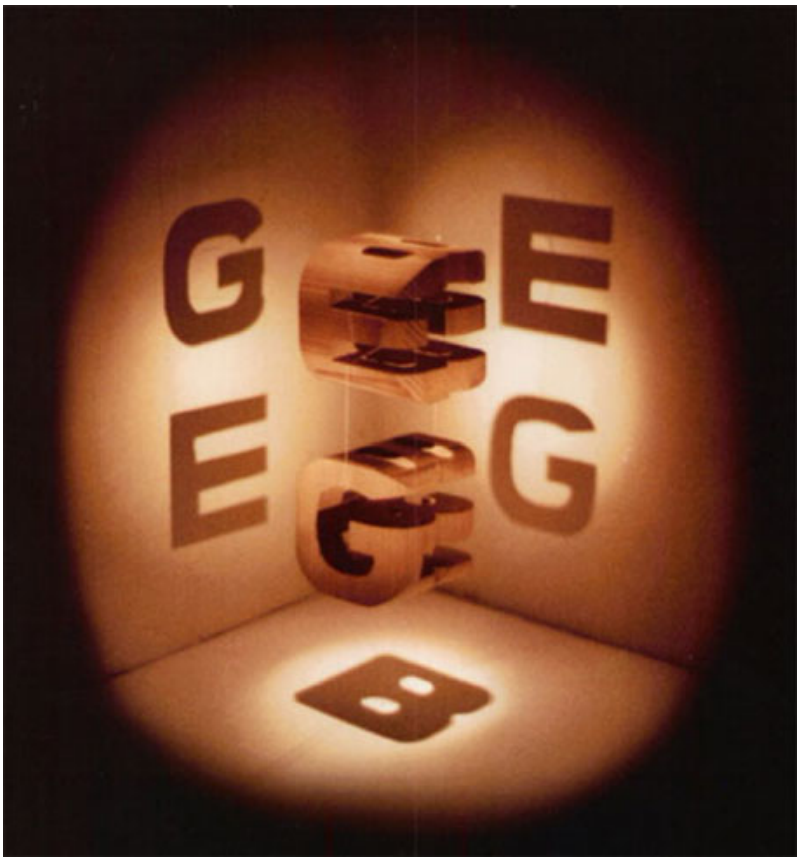

Figure 1.1: **Projection**: The cover image from Hofstadter's *Goedel, Bach and Escher* illustrates projection of 3D solids onto each 2D plane.

1.4 Problems in understanding and communicating MLM results

In my consulting practice within the Statistical Consulting Service at York University, I see hundreds of clients each year ranging from advanced undergraduate thesis students, to graduate students and faculty from a variety of fields. Over the last two decades, and across each of these groups, I have noticed an increasing desire to utilize multivariate methods. As researchers are exposed to the utility and power of multivariate tests, they see them as an appealing alternative to running many univariate ANOVAs or multiple regressions for each response variable separately.

However, multivariate analyses are more complicated than such approaches, especially when it comes to understanding and communicating results. Output is typically voluminous, and researchers will often get lost in the numbers. While software (SPSS, SAS and R) make tabular summary displays easy, these often obscure the findings that researchers are most interested in. The most common analytic oversights that we have observed are:

- **Atomistic data screening**: Researchers have mostly learned the assumptions (the Holy Trinity of normality, constant variance and independence) of univariate linear models, but then apply *univariate* tests (e.g., Shapiro-Wilk) and diagnostic plots (normal QQ plots) to every predictor and every response.
- **Bonferroni everywhere**: Faced with the task of reporting the results for multiple response measures and a collection of predictors for each, a common tendency is to run (and sometimes report) each of the separate univariate response models and then apply a correction for multiple testing. Not only is this confusing and awkward to report, but it is largely unnecessary because the multivariate tests provide protection for multiple testing.
- **Reverting to univariate visualizations**: To display results, SPSS and SAS make some visualization methods available through menu choices or syntax, but usually

these are the wrong (or at least unhelpful) choices, in that they generate separate univariate graphs for the individual responses.

This book to discusses a few essential procedures for multivariate linear models, how their interpretation can be aided through the use of well-crafted (though novel) visualizations, and provides replicable sample code in R to showcase their use in applied behaviorial research. A later section [ref?] provides some practical guidelines for analyzing, visualizing and reporting such models to help avoid these and other problems.

```
#cat("Packages used here:\n")
write_pkgs(file = .pkg_file)
#> 7  packages used here:
#>  base, datasets, graphics, grDevices, methods, stats, utils
```

References

2 Getting Started

2.1 Why plot your data?

> Getting information from a table is like extracting sunlight from a cucumber. Farquhar and Farquhar (1891)

At the time the Farhquhar brothers wrote this pithy aphorism, graphical methods for understanding data had advanced considerably, but were not universally practiced, prompting their complaint.

The main graphic forms we use today—the pie chart, line graphs and bar—were invented by William Playfair around 1800 (Playfair 1786, 1801). The scatterplot arrived shortly after (Herschel 1833) and thematic maps showing the spatial distributions of social variables (crime, suicides, literacy) were used for the first time to reason about important societal questions (Guerry 1833) such as "is increased education associated with lower rates of crime?"

In the last half of the 18th Century, the idea of correlation was developed (Galton 1886; Pearson 1896) and the period, roughly 1860–1890, dubbed the "Golden Age of Graphics (Funkhouser 1937) became the richest period of innovation and beauty in the entire history of data visualization. During this time there was an incredible development of visual thinking, represented by the work of Charles Joseph Minard, advances in the role of visualization within scientific discovery, as illustrated through Francis Galton, and graphical excellence, embodied in state statistical atlases produced in France and elsewhere. See Friendly (2008); Friendly and Wainer (2021) for this history.

2.1.1 Anscombe's Quartet

In 1973, Francis Anscombe (Anscombe 1973) famously constructed a set of four data sets illustrate the importance of plotting the graphs before analyzing and model building, and the effect of unusual observations on fitted models. Now known as *Anscombe's Quartet*, these data sets had identical statistical properties: the same means, standard devitions, correlations and regression lines.

His purpose was to debunk three notions that had been prevalent at the time:

- Numerical calculations are exact, but graphs are rough;
- For any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;
- Performing intricate calculations is virtuous, whereas actually looking at the data is cheating.

The data set `datasets::anscombe` has 11 observations, recorded in wide format, with variables `x1:x4` and `y1:y4`. ::: {.cell layout-align="center"}

```
data(anscombe)
head(anscombe)
#>   x1 x2 x3 x4   y1   y2    y3   y4
#> 1 10 10 10  8 8.04 9.14  7.46 6.58
#> 2  8  8  8  8 6.95 8.14  6.77 5.76
#> 3 13 13 13  8 7.58 8.74 12.74 7.71
#> 4  9  9  9  8 8.81 8.77  7.11 8.84
#> 5 11 11 11  8 8.33 9.26  7.81 8.47
#> 6 14 14 14  8 9.96 8.10  8.84 7.04
```

:::

The following code transforms this data to long format and calculates some summary statistics for each `dataset`.

```
anscombe_long <- anscombe |>
  pivot_longer(everything(),
               names_to = c(".value", "dataset"),
```

```
                names_pattern = "(.)(.)"
  ) |>
  arrange(dataset)

anscombe_long |>
  group_by(dataset) |>
  summarise(xbar      = mean(x),
            ybar      = mean(y),
            r         = cor(x, y),
            intercept = coef(lm(y ~ x))[1],
            slope     = coef(lm(y ~ x))[2]
          )
#> # A tibble: 4 x 6
#>   dataset  xbar  ybar     r intercept slope
#>   <chr>   <dbl> <dbl> <dbl>     <dbl> <dbl>
#> 1 1           9  7.50 0.816      3.00 0.500
#> 2 2           9  7.50 0.816      3.00 0.5
#> 3 3           9  7.5  0.816      3.00 0.500
#> 4 4           9  7.50 0.817      3.00 0.500
```

As we can see, all four data sets have nearly identical univariate and bivariate statistical measures. You can only see how they differ in graphs, which show their true natures to be vastly different.

Figure 2.1 is an enhanced version of Anscombe's plot of these data, adding helpful annotations to show visually the underlying statistical summaries.

This figure is produced as follows, using a single call to `ggplot()`, faceted by `dataset`. As we will see later (Section 3.1.1), the data ellipse (produced by `stat_ellipse()`) reflects the correlation between the variables.

```
desc <- tibble(
  dataset = 1:4,
  label = c("Pure error", "Lack of fit", "Outlier", "Influence")
)

ggplot(anscombe_long, aes(x = x, y = y)) +
  geom_point(color = "blue", size = 4) +
```

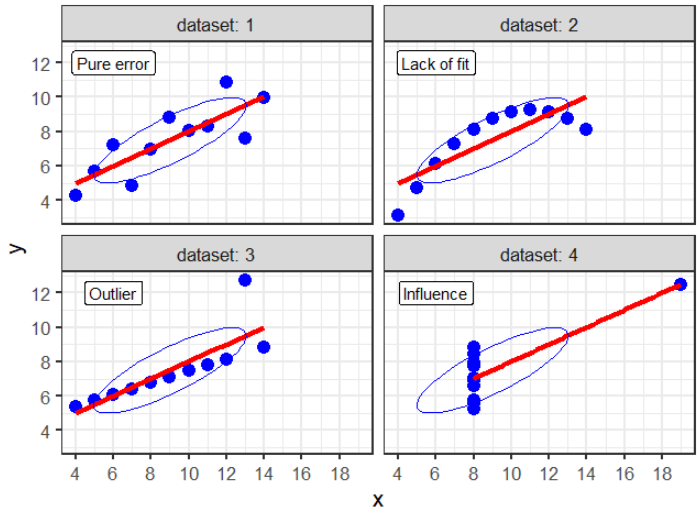

Figure 2.1: Scatterplots of Anscombe's Quartet. Each plot shows the fitted regression line and a 68% data ellipse representing the correlation between x and y.

```
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE,
              color = "red", linewidth = 1.5) +
  scale_x_continuous(breaks = seq(0,20,2)) +
  scale_y_continuous(breaks = seq(0,12,2)) +
  stat_ellipse(level = 0.5, color=col, type="norm") +
  geom_label(data=desc, aes(label = label), x=6, y=12) +
  facet_wrap(~dataset, labeller = label_both)
```

The subplots are labeled with the statistical idea they reflect:

- dataset 1: **Pure error**. This is the typical case with well-behaved data. Variation of the points around the line reflect only measurement error or unreliability in the response, y.
- dataset 2: **Lack of fit**. The data is clearly curvilinear, and would be very well described by a quadratic, `y ~ poly(x, 2)`. This violates the assumption of linear regression that the fitted model has the correct form.
- dataset 3: **Outlier**. One point, second from the right, has a very large residual. Because this point is near the

extreme of x, it pulls the regression line towards it, as you can see by imagining a line through the remaining points.

- dataset 4: **Influence**. All but one of the points have the same x value. The one unusual point has sufficient influence to force the regression line to fit it **exactly**.

One moral from this example:

> **Linear regression only "sees" a line. It does its' best when the data are really curvilinear. Because the line is fit by least squares, it pulls the line toward discrepant points to minimize the sum of squared residuals.**

Datasaurus Dozen

The method Anscombe used to compose his quartet is unknown, but it turns out that that there is a method to construct a wider collection of data sets with identical statistical properties. After all, in a bivariate data set with n observations, the correlation has $(n-2)$ degrees of freedom, so it is possible to choose this number of (x, y) pairs to yield any given value. As it happens, it is possible to create any number of data sets with the same means, standard deviations and correlations with nearly any shape you like — even a dinosaur!

The *Datasaurus Dozen* was first publicized by Alberto Cairo in a blog post and are available in the **datasauRus** package Davies, Locke, and D'Agostino McGowan (2022). As shown in **?@fig-datasaurus**, the sets include a star, cross, circle, bullseye, horizontal and vertical lines, and, of course the "dino". The method (Matejka and Fitzmaurice 2017) uses *simulated annealing*, an iterative process that perturbs the points in a scatterplot, moving them towards a given shape while keeping the statistical summaries close to the fixed target value.

The **datasauRus** package just contains the data sets, but a general method, called *statistical metamers*, for producing such data sets has been described by Elio Campitelli and implemented in the **metamer** package.

Quartets

The essential idea of a statistical "quartet" is to illustrate four quite different data sets or circumstances that seem superficially the same, but yet are paradoxically very different when you look behind the scenes. For example, in the context of causal analysis Gelman, Hullman, and Kennedy (2023), illustrated sets of four graphs, within each of which all four represent the same average (latent) causal effect but with much different patterns of individual effects. As an example of machine learning models, Biecek et al. (2023), introduced the "Rashamon Quartet", a synthetic dataset for which four models from different classes (linear model, regression tree, random forest, neural network) have practically identical predictive performance. In all cases, the paradox is solved when their visualization reveals the distinct ways of understanding structure in the data. The **quartets** package contains these and other variations on this theme.

2.1.2 A real example

In the mid 1980s, a consulting client had a strange problem. She was conducting a study of the relation between body image and weight preoccupation in exercising and non-exercising people (Davis 1990). As part of the design, the researcher wanted to know if self-reported weight could be taken as a reliable indicator of true weight measured on a scale. It was expected that the correlations between reported and measured weight should be close to 1.0, and the slope of the regression lines for men and women should also be close to 1.0. The data set is `car::Davis`.

She was therefore very surprise to see the following numerical results: For men, the correlation was nearly perfect, but not so for women.

```
data(Davis, package="carData")
Davis <- Davis |>
```

```
  drop_na()          # drop missing cases
Davis |>
  group_by(sex) |>
  select(sex, weight, repwt) |>
  summarise(r = cor(weight, repwt))
#> # A tibble: 2 x 2
#>   sex       r
#>   <fct> <dbl>
#> 1 F     0.501
#> 2 M     0.979
```

Similarly, the regression lines showed the expected slope for men, but that for women was only 0.26.

```
Davis |>
  nest(data = -sex) |>
  mutate(model = map(data, ~ lm(repwt ~ weight, data = .)),
         tidied = map(model, tidy)) |>
  unnest(tidied) |>
  filter(term == "weight") |>
  select(sex, term, estimate, std.error)
#> # A tibble: 2 x 4
#>   sex   term   estimate std.error
#>   <fct> <chr>     <dbl>     <dbl>
#> 1 M     weight    0.990    0.0229
#> 2 F     weight    0.262    0.0459
```

What could be wrong here?, the client asked. The consultant replied with the obvious question:

> *Did you plot your data?*

The answer turned out to be one discrepant point, a female, whose measured weight was 166 kg (366 lbs!). This single point exerted so much influence that it pulled the fitted regression line down to a slope of only 0.26.

```
Davis |>
  ggplot(aes(x = weight, y = repwt, color = sex, shape=sex)) +
  geom_point(size = ifelse(Davis$weight==166, 6, 2)) +
```

```
  labs(y = "Measured weight (kg)", x = "Reported weight (kg)") +
  geom_smooth(method = "lm", formula = y~x, se = FALSE) +
  theme(legend.position = c(.8, .8))
```

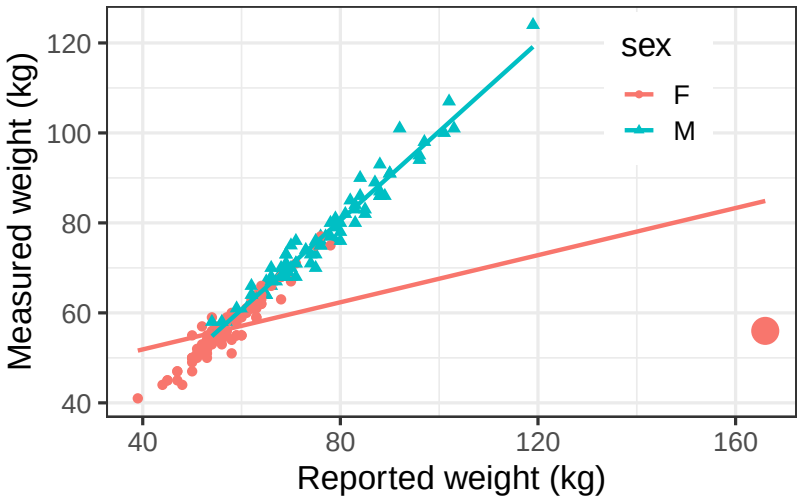

Figure 2.2: Regression for Davis' data on reported weight and measures weight for men and women. Separate regression lines, predicting reported weight from measured weight are shown for males and females. One highly unusual point is highlighted.

In this example, it was arguable that x and y axes should be reversed, to determine how well measured weight can be predicted from reported weight. In `ggplot` this can easily be done by reversing the `x` and `y` aesthetics.

```
Davis |>
  ggplot(aes(y = weight, x = repwt, color = sex, shape=sex)) +
  geom_point(size = ifelse(Davis$weight==166, 6, 2)) +
  labs(y = "Measured weight (kg)", x = "Reported weight (kg)") +
  geom_smooth(method = "lm", formula = y~x, se = FALSE) +
  theme(legend.position = c(.8, .8))
```

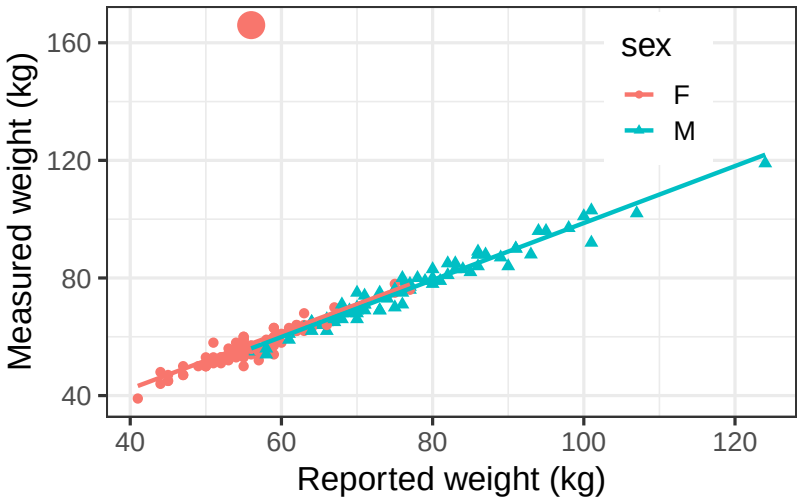

Figure 2.3: Regression for Davis' data on reported weight and measures weight for men and women. Separate regression lines, predicting measured weight from re[ported] weight are shown for males and females. The highly unusual point no longer has an effect on the fitted lines.

In Figure 2.3, this discrepant observation again stands out like a sore thumb, but it makes very little difference in the fitted line for females. The reason is that this point is well within the range of the x variable (`repwt`). To impact the slope of the regression line, an observation must be unusual in_both_ x and y. We take up the topic of how to detect influential observations and what to do about them in Chapter XX.

The value of such plots is not only that they can reveal possible problems with an analysis, but also help identify their reasons and suggest corrective action. What went wrong here? Examination of the original data showed that this person switched the values, recording her reported weight in the box for measured weight and vice versa.

2.2 Plots for data analysis

Visualization methods take an enormous variety of forms, but it is useful to distinguish several broad categories according to their use in data analysis:

- **data plots** : primarily plot the raw data, often with annotations to aid interpretation (regression lines and smooths, data ellipses, marginal distributions)
- **reconnaissance plots** : with more than a few variables, reconnaissance plots provide a high-level, bird's-eye overview of the data, allowing you to see patterns that might not be visible in a set of separate plots. Some examples are scatterplot matrices, showing all bivariate plots of variables in a data set; correlation diagrams, using visual glyphs to represent the correlations between all pairs of variables and "trellis" or faceted plots that show how a focal relation of one or more variables differs across values of other variables.
- **model plots** : plot the results of a fitted model, such as a regression line or curve to show uncertainty, or a regression surface in 3D, or a plot of coefficients in model together with confidence intervals. Other model plots try to take into account that a fitted model may involve more variables than can be shown in a static 2D plot. Some examples of these are added variable plots, and marginal effect plots, both of which attempt to show the net relation of two focal variables, controling or adjusting for other variables in a model.
- **diagnostic plots** : indicating potential problems with the fitted model. These include residual plots, influence plots, plots for testing homogeneity of variance and so forth.
- **dimension reduction plots** : plot representations of the data into a space of fewer dimensions than the number of variables in the data set. Simple examples include principal components analysis (PCA) and the related biplots, and multidimensional scaling (MDS) methods.

We give some more details and a few examples in the sections that follow.

2.3 Data plots

Data plots portray the data in a space where the coordinate axes are the observed variables.

- 1D plots include line plots, histograms and density estimates.
- 2D plots are most often scatterplots, but contour plots or hex-binned plots are also useful when the sample size is large.

2.4 Model plots

2.5 Diagnostic plots

```
cat("Writing packages to ", .pkg_file, "\n")
#> Writing packages to  C:/R/Projects/Vis-MLM-quarto/bib/pkgs.txt
write_pkgs(file = .pkg_file)
#> 18  packages used here:
#>  base, broom, datasets, dplyr, forcats, ggplot2, graphics, grDevices, lubridate, methods,
```

References

3 Plots of Multivariate Data

- Bivariate summaries
  - smoothers
  - data ellipses
- Quantitative data:
  - scatterplot matrices
  - parallel coordinate plots
- Categorical data:
  - mosaic plots
- Generalized pair plots

3.1 Bivariate summaries

```
+ smoothers
+ data ellipses
```

3.1.1 The Data Ellipse

The *data ellipse* (Monette 1990), or *concentration ellipse* (Dempster 1969) is a remarkably simple and effective display for viewing and understanding bivariate relationships in multivariate data. The data ellipse is typically used to add a visual summary to a scatterplot, that shows all together the means, standard deviations, correlation, and slope of the regression line for two variables. Under the classical assumption that the data are bivariate normally distributed, the data ellipse is also a **sufficient** visual summary, in the sense that it captures **all** relevant features of the data. See Friendly, Monette, and Fox (2013) for

a complete discussion of the role of ellipsoids in statistical data visualization.

It is based on the idea that in a bivariate normal distribution, the contours of equal probability form a series of concentric ellipses. If the variables were uncorrelated and had the same variances, these would be circles, and Euclidean distance would measure the distance of each observation from the mean. When the variables are correlated, a different measure, *Mahalanobis distance* is the proper measure of how far a point is from the mean.

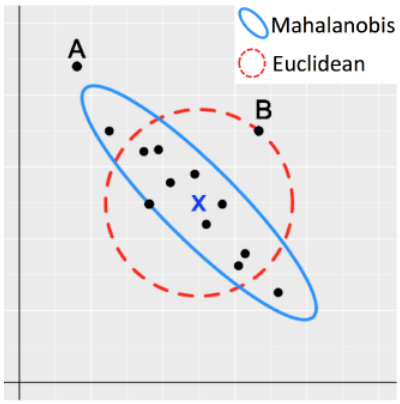

Figure 3.1: 2D data with curves of constant distance from the centroid. Which of the points A and B is further from the mean (X)? *Source*: Ou Zhang

To illustrate, Figure 3.1 shows a scatterplot with labels for two points, "A" and "B". Which is further from the mean, "X"? A contour of constant Euclidean distance, shown by the red dashed circle, ignores the apparent negative correlation, so point "A" is further. The blue ellipse for Mahalanobis distance takes the correlation into account, so point "B" has a greater distance from the mean.

Mathematically, Euclidean (squared) distance for p variables, $j = 1, 2, \ldots, p$, is just a generalization of the square of a univariate standardized (z) score, $z^2 = [(y - \bar{y})/s]^2$,

$$D^2_E(\mathbf{y}) = \sum_j^p z_j^2 = \mathbf{z}^T\mathbf{z} = (\mathbf{y} - \bar{\mathbf{y}})^T \, \text{diag}(\mathbf{S})^{-1} (\mathbf{y} - \bar{\mathbf{y}}) \; ,$$

where $\mathbf{S}$ is the sample variance-covariance matrix, $\mathbf{S} = (n-1)^{-1} \sum_{i=1}^{n} (\mathbf{y}_i - \bar{\mathbf{y}})^T (\mathbf{y}_i - \bar{\mathbf{y}})$.

Mahalanobis' distance takes the correlations into account simply by using the covariances as well as the variances,

$$D_M^2(\mathbf{y}) = (\mathbf{y} - \bar{\mathbf{y}})^T S^{-1} (\mathbf{y} - \bar{\mathbf{y}}) \; .$$

For p variables, the data *ellipsoid* $\mathcal{E}_c$ of size c is a p-dimensional ellipse, defined as the set of points $\mathbf{y} = (y_1, y_2, \ldots y_p)$ whose squared Mahalanobis distance, $D_M^2(\mathbf{y})$ is less than or equal to c^2.

When $\mathbf{y}$ is (at least approximately) bivariate normal, $D_M^2(\mathbf{y})$ has a large-sample χ_2^2 distribution (χ^2 with 2 df), so taking $c^2 = \chi_2^2(0.68) = 2.28$ gives a "1 standard deviation bivariate ellipse," an analog of the standard interval $\bar{y} \pm 1s$, while $c^2 = \chi_2^2(0.95) = 5.99 \approx 6$ gives a data ellipse of 95% coverage.

Properties

The essential ideas of correlation and regression and their relation to ellipses go back to Galton (1886). Galton's goal was to to predict (or explain) how a heritable trait, Y, (e.g., height) of children was related to that of their parents, X. He made a semi-graphic table of the frequencies of 928 observations of the average height of father and mother versus the height of their child, shown in Figure 3.2. He then drew smoothed contour lines of equal frequencies and had the wonderful visual insight that these formed concentric shapes that were tolerably close to ellipses. He then calculated summaries, Ave$(Y|X)$, and, for symmetry, Ave$(X|Y)$, and plotted these as lines of means on his diagram. Lo and behold, he had a second visual insight: the lines of means of $(Y|X)$ and $(X|Y)$ corresponded approximately to the loci of horizontal and vertical tangents to the concentric ellipses. To complete the picture, he added lines showing the major and minor axes of the family of ellipses, with the result shown in Figure 3.2.

For two variables, x and y, the remarkable properties of the data ellipse are illustrated in Figure 3.3, a modern reconstruction of Galton's diagram.

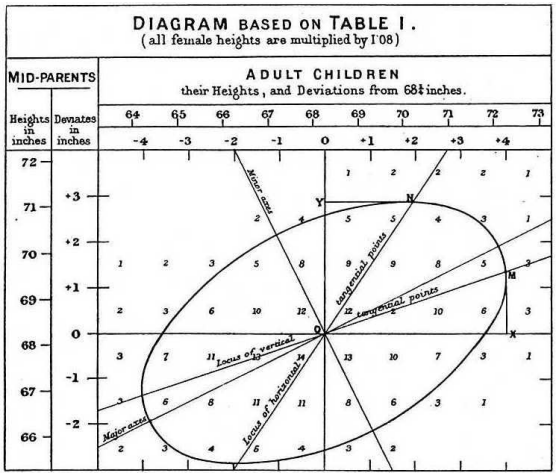

Figure 3.2: Galton's 1886 diagram, showing the relationship of height of children to the average of their parents' height. The diagram is essentially an overlay of a geometrical interpretation on a bivariate grouped frequency distribution, shown as numbers.

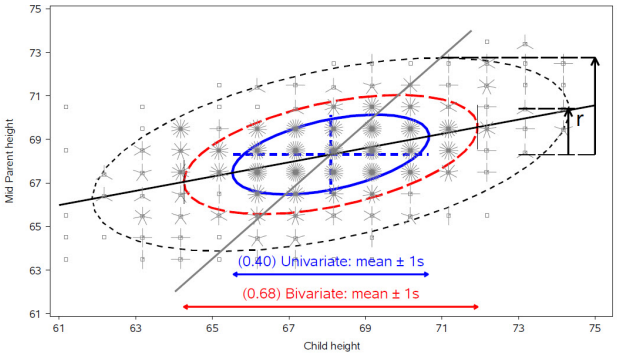

Figure 3.3: Sunflower plot of Galton's data on heights of parents and their children (in.), with 40%, 68% and 95% data ellipses and the regression lines of y on x (black) and x on y (grey).

- The ellipses have the mean vector $(\bar{x}, \bar{y})$ as their center.
- The lengths of arms of the central cross show the standard deviations of the variables, which correspond to the shadows of the ellipse covering 40% of the data. These are the bivariate analogs of the standard intervals $\bar{x} \pm 1 s_x$ and $\bar{y} \pm 1 s_y$.
- More generally, shadows (projections) on the coordinate axes, or any linear combination of them, give any standard interval, $\bar{x} \pm k s_x$ and $\bar{y} \pm k s_y$. Those with $k = 1, 1.5, 2.45$, have bivariate coverage 40%, 68% and 95%, corresponding to these quantiles of the χ^2 distribution with 2 degrees of freedom, i.e., $\chi^2_2(.40) \approx 1^2$, $\chi^2_2(.68) \approx 1.5^2$, and $\chi^2_2(.95) \approx 2.45$.
- The regression line predicting y from x goes through the points where the ellipses have vertical tangents. The *other* regression line, predicting x from y goes through the points of horizontal tangency.
- The correlation $r(x, y)$ is the ratio of the vertical segment from the mean of y to the regression line to the vertical segment going to the top of the ellipse as shown at the right of the figure. It is $r = 0.46$ in this example.
- The residual standard deviation, $s_e = \sqrt{MSE} = \sqrt{\Sigma(y - \bar{y})^2 / n - 2}$, is the half-length of the ellipse at the mean $\bar{x}$

Because Galton's values of `parent` and `child` height were recorded in class intervals of 1 in., they are shown as sunflower symbols in Figure 3.3, with multiple 'petals' reflecting the number of observations at each location. This plot is constructed using `sunflowerplot()` and `car::dataEllipse()` for the ellipses.

```
data(Galton, package = "HistData")

sunflowerplot(parent ~ child, data=Galton,
      xlim=c(61,75),
      ylim=c(61,75),
      seg.col="black",
```

```
        xlab="Child height",
      ylab="Mid Parent height")

y.x <- lm(parent ~ child, data=Galton)     # regression of y on x
abline(y.x, lwd=2)
x.y <- lm(child ~ parent, data=Galton)     # regression of x on y
cc <- coef(x.y)
abline(-cc[1]/cc[2], 1/cc[2], lwd=2, col="gray")

with(Galton,
     car::dataEllipse(child, parent,
         plot.points=FALSE,
         levels=c(0.40, 0.68, 0.95),
         lty=1:3)
    )
```

3.2 R functions for data ellipses

A number of packages provide functions for drawing data ellipses of data, with various features.

- `car::scatterplot()`
- `car::dataEllipse()`
- `heplots::covEllipses()`
- `ggplot2::stat_ellipse()`

3.3 Quantitative data:

```
+ scatterplot matrices
+ parallel coordinate plots
```

3.4 Categorical data:

```
+ mosaic plots
```

3.5 Generalized pair plots

```
#> 7  packages used here:
#>  base, datasets, graphics, grDevices, methods, stats, utils
```

References

4 PCA and Biplots

4.1 Principal components analysis

4.2 Biplots

5 Overview of Linear models

This chapter reviews the characteristics of the standard univariate response models for quantitative outcomes.

5.1 Regression

5.2 ANOVA

5.3 ANCOVA

References

6 Plots for univariate response models

For a univariate linear model fit using `lm()`, `glm()` and similar functions, the standard `plot()` method gives basic versions of *diagnostic* plots of residuals and other calculated quantities for assessing possible violations of the model assumptions. Some of these can be considerably enhanced using other packages.

Beyond this,

- tables of model coefficients, standard errors and test statistics can often be usefully supplemented or even replaced by suitable plots providing essentially the same information.
- when there are two or more predictors, you can more easily understand their separate impact on the response by plotting the *marginal* effects of one or more focal variables, averaging over other variables not shown in a given plot.
- when there are highly correlated predictors, some specialized plots are useful to understand the nature of *multicollinearity*.

The classic reference on regression diagnostics is Belsley, Kuh, and Welsch (1980). My favorite modern texts are the brief Fox (2020) and the more complete Fox and Weisberg (2018), both of which are supported by the **car** package (Fox, Weisberg, and Price 2023).

6.1 The "regression quartet"

For a fitted model, plotting the model object with `plot(model)` provides for any of six basic plots, of which four are produced

by default, giving rise to the term *regression quartet* for this collection. These are:

- **Residuals vs. Fitted**: For well-behaved data, the points should hover around a horizontal line at residual = 0, with no obvious pattern or trend.
- **Normal Q-Q plot**: A plot of sorted standardized residuals e_i (obtained from`rstudent(model)`) against the theoretical values those values would have in a standard normal $\mathcal{N}(0, 1)$ distribution.
- **Scale-Location**: Plots the square-root of the absolute values of the standardized residuals $\sqrt{|e_i|}$ as a measure of "scale" against the fitted values $\hat{y}_i$ as a measure of "location". This provides an assessment of homogeneity of variance, which appears as a tendency for scale to vary with location.
- **Residuals vs. Leverage**: Plots standardized residuals against leverage to help identify possibly influential observations. Leverage, or "hat" values (given by `hat(model)`) are proportional to the squared Mahalanobis distances of the predictor values $\mathbf{x}_i$ from the means, and measure the potential of an observation to change the fitted coefficients if that observation was deleted. Actual influence is measured by Cooks's distance (`cooks.distance(model)`) and is proportional to the product of residual times leverage. Contours of constant Cook's D are added to the plot.

One key feature of these plots is providing **reference** lines or smoothed curves for ease of judging the extent to which a plot conforms to the expected pattern; another is the **labeling** of observations which deviate from an assumption.

The base-R `plot(model)` plots are done much better in a variety of packages. I illustrate some versions from the **car** (Fox, Weisberg, and Price 2023) and **performance** (Lüdecke et al. 2021) packages, part of the **easystats** (Lüdecke et al. 2022) suite of packages.

Packages:

```
library(car)
library(easystats)
```

Example: Duncan's occupational prestige

In a classic study in sociology, Duncan (1961) used data from the U.S. Census in 1950 to study how one could predict the prestige of occupational categories — which is hard to measure — from available information in the census for those occupations. His data is available in `carData:Duncan`, and contains

- `type`: the category of occupation, one of `prof` (professional), `wc` (white collar) or `bc` (blue collar);
- `income`: the percentage of occupational incumbents with a reported income $>$ \$3500 (about \$40,000 in current dollars);
- `education`: the percentage of occupational incumbents who were high school graduates;
- `prestige`: the percentage of respondents in a social survey who rated the occupation as "good" or better in prestige.

These variables are a bit quirky in they are measured in percents, 0-100, rather dollars for `income` and years for `education`, but this common scale permitted Duncan to ask an interesting sociological question: Assuming that both income and education predict prestige, are they equally important, as might be assessed by testing the hypothesis $H_0 : \beta_{\text{income}} = \beta_{\text{education}}$.

A quick look at the data shows the variables and a selection of the occupational categories, which are the `row.names()` of the dataset.

```
data(Duncan, package = "carData")
set.seed(42)
car::some(Duncan)
#>                 type income education prestige
#> accountant      prof     62        86       82
#> professor       prof     64        93       93
#> engineer        prof     72        86       88
#> factory.owner   prof     60        56       81
```

```
#> store.clerk        wc     29        50       16
#> carpenter          bc     21        23       33
#> machine.operator   bc     21        20       24
#> barber             bc     16        26       20
#> soda.clerk         bc     12        30        6
#> janitor            bc      7        20        8
```

Let's start by fitting a simple model using just income and education as predictors. The results look very good! Both `income` and `education` are highly significant and the $R^2 = 0.828$ for the model indicates that `prestige` is very well predicted by just these variables.

```
duncan.mod <- lm(prestige ~ income + education, data=Duncan)
summary(duncan.mod)
#>
#> Call:
#> lm(formula = prestige ~ income + education, data = Duncan)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -29.54  -6.42   0.65   6.61  34.64
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  -6.0647     4.2719   -1.42     0.16
#> income        0.5987     0.1197    5.00  1.1e-05 ***
#> education     0.5458     0.0983    5.56  1.7e-06 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 13.4 on 42 degrees of freedom
#> Multiple R-squared:  0.828,  Adjusted R-squared:  0.82
#> F-statistic:  101 on 2 and 42 DF,  p-value: <2e-16
```

Beyond this, Duncan was interested in the coefficients and whether income and education could be said to have equal impacts on predicting occupational prestige. A nice display of model coefficients with confidence intervals is provided by `parameters::model_parameters()` and we can test Duncan's

hypothesis with `car::linearHypothesis()`. The latter is constructed as a test of a restricted model in which the two coefficients are forced to be equal against the unrestricted model. Duncan was very happy with this result.

```
parameters::model_parameters(duncan.mod)
#> Parameter   | Coefficient |   SE |         95% CI | t(42) |      p
#> ------------------------------------------------------------------
#> (Intercept) |       -6.06 | 4.27 | [-14.69, 2.56] | -1.42 | 0.163
#> income      |        0.60 | 0.12 | [  0.36, 0.84] |  5.00 | < .001
#> education   |        0.55 | 0.10 | [  0.35, 0.74] |  5.56 | < .001

car::linearHypothesis(duncan.mod, "income = education")
#> Linear hypothesis test
#>
#> Hypothesis:
#> income - education = 0
#>
#> Model 1: restricted model
#> Model 2: prestige ~ income + education
#>
#>   Res.Df  RSS Df Sum of Sq    F Pr(>F)
#> 1     43 7519
#> 2     42 7507  1      12.2 0.07    0.8
```

But, should Duncan be **so** happy? It is unlikely that he ran any model diagnostics or plotted his model; we do so now. Here is the regression quartet for this model. Each plot shows some trend lines, and importantly, labels some observations that stand out and might deserve attention.

```
op <- par(mfrow = c(2,2),
          mar = c(4,4,3,1)+.1)
plot(duncan.mod, lwd=2, pch=16)
par(op)
```

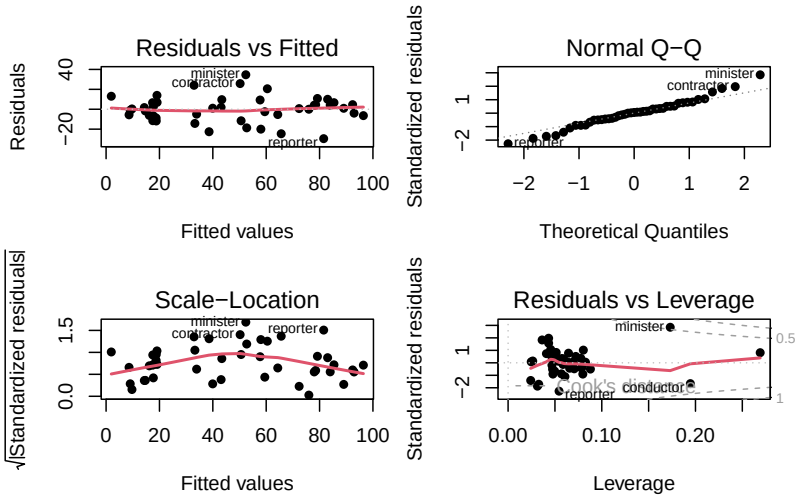

Figure 6.1: Regression quartet of diagnostic plots for the `Duncan` data. Several possibly unusual observations are labeled.

Example: Occupational prestige

CUT THIS EXAMPLE

These examples use the data on the prestige of 102 occupational categories and other measures from the 1971 Canadian Census, recorded in `carData::Prestige`. Our interest is in understanding how `prestige` (the Pineo-Ported prestige score, from a social survey) is related to census measures of the average education, income, percent women of incumbents in those occupations. Occupation type is a factor with levels `"bc"` (blue collar), `"wc"` (white collar) and `"prof"` (professional). **TODO**: These data should be introduced earlier with descriptive plots, scatterplots, ...

```
data(Prestige, package="carData")
# `type` is really an ordered factor. Make it so.
Prestige$type <- ordered(Prestige$type,
                         levels=c("bc", "wc", "prof"))
str(Prestige)
#> 'data.frame':    102 obs. of  6 variables:
```

```
#>  $ education: num  13.1 12.3 12.8 11.4 14.6 ...
#>  $ income   : int  12351 25879 9271 8865 8403 11030 8258 14163 11377 11023 ...
#>  $ women    : num  11.16 4.02 15.7 9.11 11.68 ...
#>  $ prestige : num  68.8 69.1 63.4 56.8 73.5 77.6 72.6 78.1 73.1 68.8 ...
#>  $ census   : int  1113 1130 1171 1175 2111 2113 2133 2141 2143 2153 ...
#>  $ type     : Ord.factor w/ 3 levels "bc"<"wc"<"prof": 3 3 3 3 3 3 3 3 3 3 ...
```

We fit a main-effects model using all predictors (ignoring `census`, the Canadian Census occupational code):

```
prestige.mod <- lm(prestige ~ education + income + women + type,
                   data=Prestige)
```

`plot(model)` produces four separate plots. For a quick look, I like to arrange them in a single 2x2 figure.

```
op <- par(mfrow = c(2,2),
          mar=c(4,4,3,1)+.1)
plot(prestige.mod, lwd=2, cex.lab=1.4)
par(op)
```

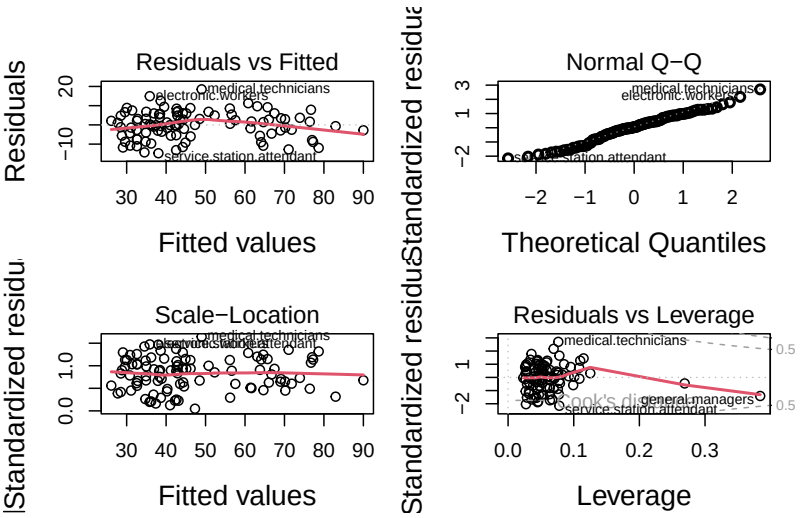

Figure 6.2: Regression quartet of diagnostic plots for the `Prestige` data. Several possibly unusual observations are labeled.

6.2 Other Diagnostic plots

6.2.1 Spread-level plot

6.3 Coefficient plots

6.4 Added-variable plots

6.5 Marginal plots

6.6 Outliers, leverage and influence

In small to moderate samples, "unusual" observations can have dramatic effects on a fitted regression model, as we saw in the analysis of Davis's data on reported and measured weight (Section 2.1.2) where one erroneous observations hugely altered the fitted line.

An observation can be unusual in three archetypal ways, with different consequences:

- Unusual in the response y, but typical in the predictor(s), $\mathbf{x}$ — a badly fitted case with a large absolute residual, but with x not far from the mean, as in Figure 2.3. This case does not do much harm to the fitted model.
- Unusual in the predictor(s) $\mathbf{x}$, but typical in y — an otherwise well-fitted point. This case also does litle harm, and in fact can be considered to improve precision, a "good leverage" point.
- Unusual in **both** $\mathbf{x}$ and y — This is the case, a "bad leverage" point, revealed in the analysis of Davis's data, Figure 2.2, where the one erroneous point for women was highly influential, pulling the regression line towards it and affecting the estimated coefficient as well as all the fitted values.

Influential cases are the ones that matter most. As suggested above, to be influential an observation must be unusual in **both** $\mathbf{x}$ and y, and affects the estimated coefficients, thereby also altering the predicted values for all observations. A heuristic formula capturing the relations among leverage, "outlyingness" on y and influence is

$$\text{Influence}_{\text{coefficients}} = X_{\text{leverage}} \times Y_{\text{residual}}$$

As described below, leverage is proportional to the squared distance $(x_i - \bar{x})^2$ of an observation x_i from its mean in simple regression and to the squared Mahalanobis distance in the general case. The Y_{residual} is best measured by a *studentized* residual, obtained by omitting each case i in turn and calculating its residual from the coefficients obtained from the remaining cases.

6.6.1 The leverage-influence quartet

These ideas can be illustrated in the "leverage-influence quartet" by considering a standard simple linear regression for a sample and then adding one additional point reflecting the three situations described above. Below, I generate a sample of $N = 15$ points with x uniformly distributed between (40, 60) and $y \sim 10 + 0.75x + \mathcal{N}(0, 1.25^2)$, duplicated four times.

```
library(tidyverse)
library(car)
set.seed(42)
N <- 15
case_labels <- paste(1:4, c("OK", "Outlier", "Leverage", "Influence"))
levdemo <- tibble(
    case = rep(case_labels,
               each = N),
    x = rep(round(40 + 20 * runif(N), 1), 4),
    y = rep(round(10 + .75 * x + rnorm(N, 0, 1.25), 4)),
    id = " "
)
```

```
mod <- lm(y ~ x, data=levdemo)
coef(mod)
#> (Intercept)           x
#>      13.332       0.697
```

The additional points, one for each situation are set to the values below.

- `Outlier`: (52, 60) a low leverage point, but an outlier (`O`) with a large residual
- `Leverage`: (75, 65) a "good" high leverage point (`L`) that fits well with the regression line
- `Influence`: (70, 40) a "bad" high leverage point (`OL`) with a large residual.

```
extra <- tibble(
  case = case_labels,
  x  = c(65, 52, 75, 70),
  y  = c(NA, 65, 65, 40),
  id = c("  ", "O", "L", "OL")
)

#' Join these to the data
both <- bind_rows(levdemo, extra) |>
  mutate(case = factor(case))
```

We can plot these four situations with `ggplot2` in panels faceted by `case` as shown below. The standard version of this plot shows the regression line for the original data and that for the ammended data with the additional point. Note that we use the `levdemo` dataset in `geom_smooth()` for the regression line with the original data, but specify `data = both` for that with the additional point.

```
ggplot(levdemo, aes(x = x, y = y)) +
  geom_point(color = "blue", size = 2) +
  geom_smooth(data = both,
              method = "lm", formula = y ~ x, se = FALSE,
              color = "red", linewidth = 1.3, linetype = 1) +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE,
```

```
            color = "blue", linewidth = 1, linetype = "longdash" ) +
  stat_ellipse(data = both, level = 0.5, color="blue", type="norm", linewidth = 1.4) +
  geom_point(data=extra, color = "red", size = 4) +
  geom_text(data=extra, aes(label = id), nudge_x = -2, size = 5) +
  facet_wrap(~case, labeller = label_both) +
  theme_bw(base_size = 14)
```

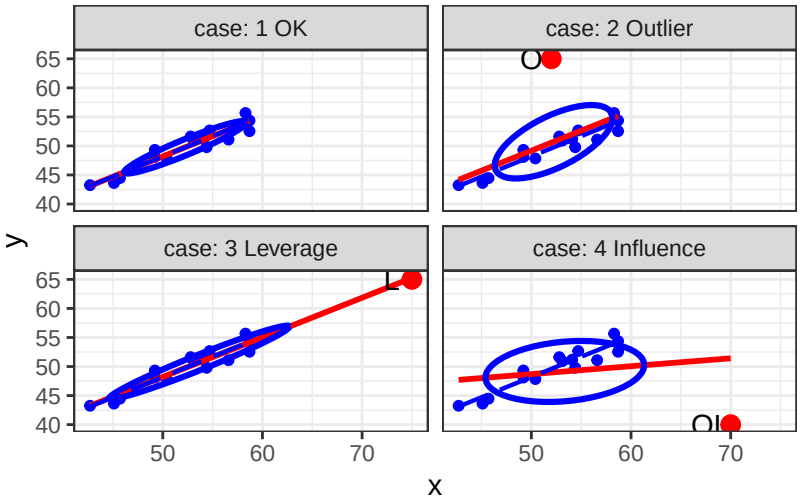

Figure 6.3: Leverage influence quartet with data 50% ellipses. Case (1) original data; (2) adding one low-leverage outlier, "O"; (3) adding one "good" leverage point, "L"; (4) adding one "bad" leverage point, "OL". The dashed line is the fitted line for the original data, while the solid line reflects the additional point. The data ellipses show the effect of the additional point on precision.

The standard version of this graph shows only the fitted regression lines in each panel. As can be seen, the fitted line doesn't change very much in panels (2) and (3); only the bad leverage point, "OL" in panel (4) is harmful. Adding data ellipses to each panel immediately makes it clear that there is another part to this story— the effect of the unusual point on *precision* (standard errors) of our estimates of the coefficients.

Now, we see *directly* that there is a big difference in impact between the low-leverage outlier [panel (2)] and the high-leverage,

small-residual case [panel (3)], even though their effect on coefficient estimates is negligible. In panel (2), the single outlier inflates the estimate of residual variance (the size of the vertical slice of the data ellipse at $\bar{x}$), while in panel (3) this is decreased.

To allow direct comparison and make the added value of the data ellipse more apparent, we overlay the data ellipses from Figure 6.3 in a single graph, shown in Figure 6.4.
Here, we can also see why the high-leverage point "L" added in panel (c) of 6.3 is called a "good leverage" point. By increasing the standard deviation of x, it makes the data ellipse somewhat more elongated, giving increased precision of our estimates of β.

```
colors <- c("black", "blue", "darkgreen", "red")
with(both,
     {dataEllipse(x, y, groups = case,
          levels = 0.68,
          plot.points = FALSE, add = FALSE,
          center.pch = "+",
          col = colors,
          fill = TRUE, fill.alpha = 0.1)
     })

case1 <- both |> filter(case == "1 OK")
points(case1[, c("x", "y")], cex=1)

points(extra[, c("x", "y")],
       col = colors,
       pch = 16, cex = 2)

text(extra[, c("x", "y")],
     labels = extra$id,
     col = colors, pos = 2, offset = 0.5)
```

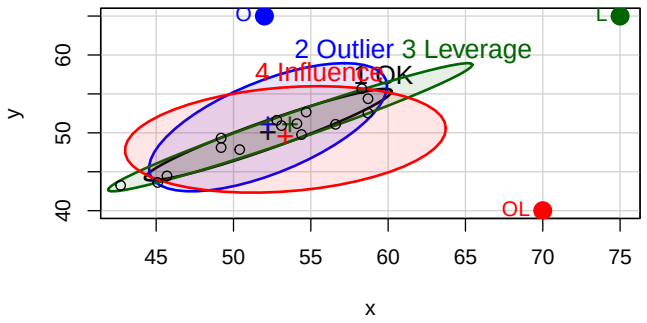

Figure 6.4: Data ellipses in the Leverage-influence quartet. This graph overlays the data ellipses and additional points from the four panels of Figure 6.4. It can be seen that only the OL point affects the slope, while the O and L points affect precision of the estimates in opposite directions.

6.6.2 Measuring leverage

Leverage is thus an index of the *potential* impact of an observation on the model due to its' atypical value in the X space of the predictor(s). It is commonly measured by the "hat" value, h_i, so called because it puts the hat ($\hat{\bullet}$) on $\mathbf{y}$, i.e., the vector of fitted values can be expressed as

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{H}\mathbf{y} \\ &= [\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\,\mathbf{y}\ , \end{aligned}$$

where $h_i \equiv h_{ii}$ are the diagonal elements of the Hat matrix $\mathbf{H}$. In simple regression, hat values are proportional to the squared distance of the observation x_i from the mean, $h_i \propto (x_i - \bar{x})^2$,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma_i (x_i - \bar{x})^2}\ ,$$

and range from $1/n$ to 1, with an average value $\bar{h} = 2/n$. Consequently, observations with h_i greater than $2\bar{h}$ or $3\bar{h}$ are commonly considered to be of high leverage.

With $p \geq 2$ predictors, it is demonstrated below that $h_i \propto D^2(\mathbf{x} - \bar{\mathbf{x}})$, the squared distance of $\mathbf{x}$ from the centroid $\bar{\mathbf{x}}$[1]. [2]: See this Stats StackExchange discussion for a proof. The analogous formula is

$$h_i = \frac{1}{n} + \frac{1}{n-1} D^2(\mathbf{x} - \bar{\mathbf{x}}) \ ,$$

where $D^2(\mathbf{x} - \bar{\mathbf{x}}) = (\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{S}_X^{-1} (\mathbf{x} - \bar{\mathbf{x}})$. From Section 3.1.1, it follows that contours of constant leverage correspond to data ellipses or ellipsoids of the predictors in $\mathbf{x}$, whose boundaries, assuming normality, correspond to quantiles of the χ^2_p distribution

Example:

To illustrate, I generate $N = 100$ points from a bivariate normal distribution with means $\mu = (30, 30)$, variances $= 10$, and a correlation $\rho = 0.7$ and add two noteworthy points that show an apparently paradoxical result.

```
set.seed(421)
N <- 100
r <- 0.7
mu <- c(30, 30)
cov <- matrix(c(10,   10*r,
                10*r, 10), ncol=2)

X <- MASS::mvrnorm(N, mu, cov) |> as.data.frame()
colnames(X) <- c("x1", "x2")

# add 2 points
```

[1] If group sizes are greatly unequal **and** homogeneity of variance is violated, then the F statistic is too liberal (p values too large) when large sample variances are associated with small group sizes. Conversely, the F statistic is too conservative if large variances are associated with large group sizes.

[2] If group sizes are greatly unequal **and** homogeneity of variance is violated, then the F statistic is too liberal (p values too large) when large sample variances are associated with small group sizes. Conversely, the F statistic is too conservative if large variances are associated with large group sizes.

```
X <- rbind(X,
           data.frame(x1 = c(28, 38),
                      x2 = c(42, 35)))
```

The Mahalanobis squared distances of these points can be calculated using `heplots::Mahalanobis()`, and their corresponding hatvalues found using `hatvalues()` for any linear model using both `x1` and `x2`.

```
X <- X |>
  mutate(Dsq = heplots::Mahalanobis(X)) |>
  mutate(y = 2*x1 + 3*x2 + rnorm(nrow(X), 0, 5),
         hat = hatvalues(lm(y ~ x1 + x2)))
```

Plotting `x1` and `x2` with data ellipses shows the relation of leverage to squared distance from the mean. The blue point looks to be farther from the mean, but the red point is actually very much further by Mahalanobis squared distance, which takes the correlation into account; it thus has much greater leverage.

```
par(mar = c(4, 4, 1, 1) + 0.1)
dataEllipse(X$x1, X$x2,
            levels = c(0.40, 0.68, 0.95),
            fill = TRUE, fill.alpha = 0.05,
            col = "darkgreen",
            xlab = "X1", ylab = "X2")
points(X[1:nrow(X) > N, 1:2], pch = 16, col=c("red", "blue"), cex = 2)
X |> slice_tail(n = 2) |>      # last two rows
  points(pch = 16, col=c("red", "blue"), cex = 2)
```

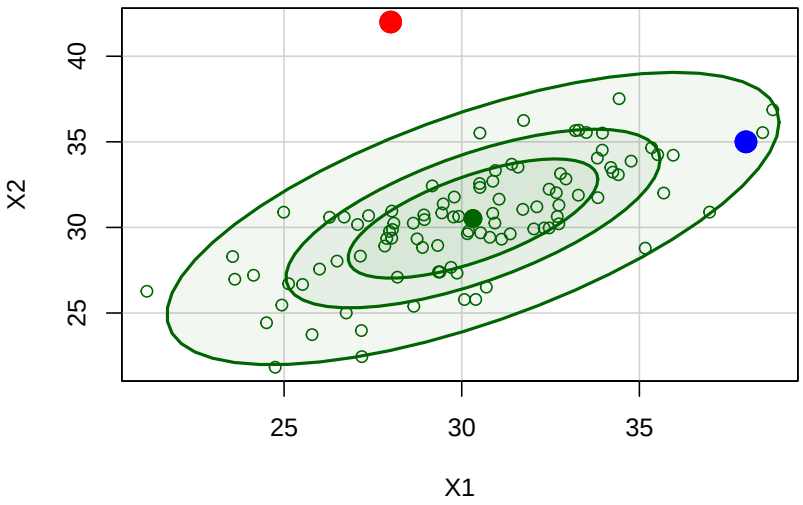

Figure 6.5: Data ellipses for a bivariate normal sample with correlation 0.7, and two additional noteworthy points. The blue point looks to be farther from the mean, but the red point is actually more than 5 times further by Mahalanobis squared distance, and thus has much greater leverage.

The fact that hatvalues are proportional to leverage can be seen by plotting one against the other. I highlight the two noteworthy points in their colors from Figure 6.5 to illustrate how much greater leverage the red point has compared to the blue point.

```
plot(hat ~ Dsq, data = X,
     cex = c(rep(1, N), rep(2, 2)),
     col = c(rep("black", N), "red", "blue"),
     pch = 16,
     ylab = "Hatvalue",
     xlab = "Mahalanobis Dsq")
```

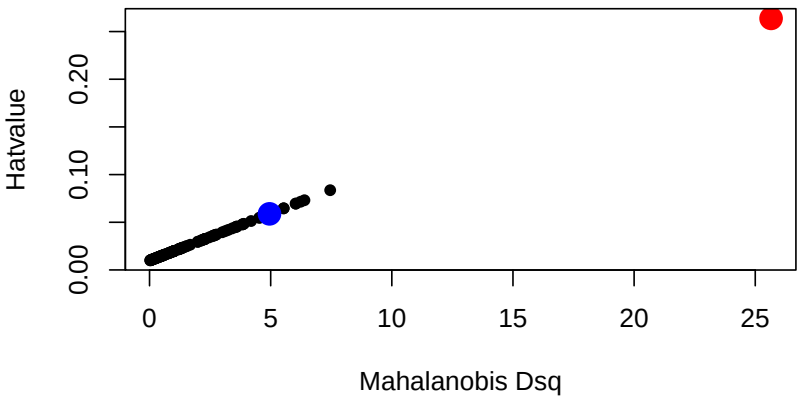

Figure 6.6: Hat values are proportional to squared Mahalanobis distances from the mean.

Look back at these two points in Figure 6.5. Can you guess how much further the red point is from the mean than the blue point? You might be surprised that its' D^2 and leverage are about five times as great!

```
X |> slice_tail(n=2)
#>   x1 x2   Dsq   y    hat
#> 1 28 42 25.65 179 0.2638
#> 2 38 35  4.95 175 0.0588
```

6.6.3 Outliers: Measuring residuals

From the discussion in Section 6.6, outliers for the response y are those observations for which the residual $e_i = y_i - \hat{y}_i$ are unusually large in magnitude. However, as demonstrated in Figure 6.3, a high-leverage point will pull the fitted line towards it, reducing its' residual and thus making them look less unusual.

The standard approach (Cook and Weisberg 1982; Hoaglin and Welsch 1978) is to consider a *deleted residual* $e_{(-i)}$, conceptually as that obtained by re-fitting the model with observation i omitted and obtaining the fitted value $\hat{y}_{(-i)}$ from the remaining $n - 1$ observations,

$$e_{(-i)} = y_i - \hat{y}_{(-i)} \ .$$

The (externally) *studentized residual* is then obtained by dividing $e_{(-i)}$ by it's estimated standard error, giving

$$e^{\star}_{(-i)} = \frac{e_{(-i)}}{\text{sd}(e_{(-i)})} = \frac{e_i}{\sqrt{\text{MSE}_{(-i)}\ (1 - h_i)}} \ .$$

This is just the ordinary residual e_i divided by a factor that increases with the residual variance but decreases with leverage. It can be shown that these studentized residuals follow a t distribution with $n - p - 2$ degrees of freedom, so a value $|e^{\star}_{(-i)}| > 2$ can be considered large enough to pay attention to. In practice for classical linear models, it is unnecessary to actually re-fit the model n times ...

6.6.4 Measuring influence

As described at the start of this section, the actual influence of a given case depends multiplicatively on its' leverage and residual. But how can we measure it?

The essential idea introduced above, is to delete the observations one at a time, each time refitting the regression model on the remaining $n\breve{}1$ observations. Then, for observation i compare the results using all n observations to those with the i^{th} observation deleted to see how much influence the observation has on the analysis.

The simplest such measure, called DFFITS, compares the predicted value for case i with what would be obtained when that observation is excluded.

$$\begin{aligned} \text{DFFITS}_i &= \frac{\hat{y}_i - \hat{y}_{(-i)}}{\sqrt{\text{MSE}_{(-i)}\ h_i}} \\ &= e^{\star}_{(-i)} \times \sqrt{\frac{h_i}{1 - h_i}} \ . \end{aligned}$$

The first equation gives the signed difference in fitted values in units of the standard deviation of that difference weighted by leverage; the second version (Belsley, Kuh, and Welsch 1980)

represents that as a product of residual and leverage. A rule of thumb is that an observation is deemed to be influential if $|\text{DFFITS}_i| > 2\sqrt{(p+1)/n}$.

```
#> Writing packages to  C:/R/Projects/Vis-MLM-quarto/bib/pkgs.txt
#> 30  packages used here:
#>  base, bayestestR, car, carData, correlation, datasets, datawizard, dplyr, easystats, effect
```

7 Collinearity & Ridge Regression

> Some of my collinearity diagnostics have large values, or small values, or whatever they are not supposed to be * What is bad? * If bad, what can I do about it?

In univariate multiple regression models, we usually hope to have high correlations between the outcome y and each of the predictors, $x_1, x_2, \ldots$, but high correlations *among* the predictors can cause problems in estimating and testing their effects. The quote above shows the a typical quandary of some researchers in trying do understand these problems and and take steps to resolve them. This chapter illustrates the problems of collinearity, describes diagnostic measures to asses its effects, and presents some novel visual tools for these purposes using the **VisCollin** package.

One class of solutions for collinearity involves *regularization methods* such as ridge regression. Another collection of graphical methods, generalized ridge trace plots, implemented in the **genridge** package, sheds further light on what is accomplished by this technique.

Packages

In this chapter we use the following packages. Load them now.

```
library(car)
library(VisCollin)
library(genridge)
library(MASS)
library(dplyr)
library(factoextra)
```

```
library(ggrepel)
library(patchwork)
```

7.1 What is collinearity?

The chapter quote above is not untypical of researchers who have read standard treatments of linear models (eg.: ???) and yet are still confused about what collinearity is, how to find its sources and how to correct them. In Friendly and Kwan (2009), we liken this problem to that of the reader of Martin Hansford's successful series of books, *Where's Waldo*. These consist of a series of full-page illustrations of hundreds of people and things and a few Waldos— a character wearing a red and white striped shirt and hat, glasses, and carrying a walking stick or other paraphernalia. Waldo was never disguised, yet the complex arrangement of misleading visual cues in the pictures made him very hard to find. Collinearity diagnostics often provide a similar puzzle.

Recall the standard classical linear model for a response variable y with a collection of predictors in $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p)$

$$\begin{aligned} \mathbf{y} &= \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \cdots + \beta_p \mathbf{x}_p + \epsilon \\ &= \mathbf{X}\beta + \epsilon \ , \end{aligned}$$

for which the ordinary least squares solution is:

$$\widehat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\ \mathbf{X}^T\mathbf{y} \ ,$$

with sampling variances and covariances $\text{Var}(\widehat{\mathbf{b}}) = \sigma^2 \times (\mathbf{X}^T\mathbf{X})^{-1}$ and σ^2 is the variance of the residuals ϵ, estimated by the mean squared error (MSE).

In the limiting case, when one x_i is *perfectly* predictable from the other xs, i.e., $R^2(x_i|\text{other } x) = 1$,

- there is no *unique* solution for the regression coefficients $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$;

- the standard errors $s(b_i)$ of the estimated coefficients are infinite and t statistics $t_i = b_i/s(b_i)$ are 0.

This extreme case reflects a situation when one or more predictors are effectively redundant, for example when you include two variables x and y and their sum $z = x + y$ in a model, or use *ipsatized* scores that sum to a constant. More generally, collinearity refers to the case when there are very high **multiple correlations** among the predictors, such as $R^2(x_i|\text{other } x) \geq 0.9$. Note that you can't tell simply by looking at the simple correlations. A large correlation r_{ij} is *sufficient* for collinearity, but not *necessary* — you can have variables x_1, x_2, x_3 for which the pairwise correlation are low, but the multiple correlation is high.

The consequences are:

- The estimated coefficients have large standard errors, $s(\hat{b_j})$. They are multiplied by the square root of the variance inflation factor, $\sqrt{\text{VIF}}$, discussed below.
- This deflates the t-statistics, $t = \hat{b_j}/s(\hat{b_j})$ by the same factor.
- Thus you may find a situation where an overall model is highly significant (large F-statistic), while no (or few) of the individual predictors are. This is a puzzlement!
- Beyond this, the least squares solution may have poor numerical accurracy (Longley 1967), because the solution depends on the determinant $|\,\mathbf{X}^T\mathbf{X}\,|$, which approaches 0 as multiple correlations increase.
- As well, recall that the coefficients $\hat{b}$ are **partial coefficients**, meaning the estimated change Δy in y when x changes by one unit Δx, but **holding all other variables constant**. Then, the model may be trying to estimate something that does not occur in the data.

7.1.1 Visualizing collinearity

Collinearity can be illustrated in data space for two predictors in terms of the stability of the regression plane for a linear model `Y = X1 + X2`. In Figure 7.1 (adapted from Fox (2016), Fig. 13.2):

(a) shows a case where X_1 and X_2 are uncorrelated as can be seen in their scatter in the horizontal plane (+ symbols). The regression plane is well-supported; a small change in Y for one observation won't make much difference.

(b) In panel (b), X_1 and X_2 have a perfect correlation, $r(x_1, x_2) = 1.0$. The regression plane is not unique; in fact there are an infinite number of planes that fit the data equally well. Note that, if all we care about is prediction (not the coefficients), we could use X_1 or X_2, or both, or any weighted sum of them in a model and get the same predicted values.

(c) Shows a typical case where there is a strong correlation between X_1 and X_2. The regression plane here is unique, but is not well determined. A small change in Y **can** make quite a difference in the fitted value or coefficients, depending on the values of X_1 and X_2. Where X_1 and X_2 are far from their near linear relation in the botom plane, you can imagine that it is easy to tilt the plane substantially by a small change in Y.

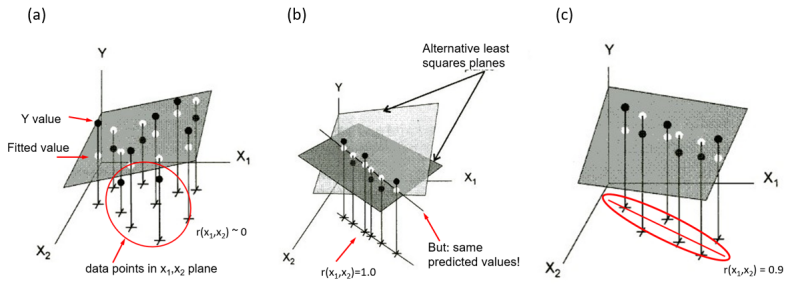

Figure 7.1: Effect of collinearity on the least squares regression plane. (a) Small correlation between predictors; (b) Perfect correlation ; (c) Very strong correlation. The black points show the data Y values, white points are the fitted values in the regression plane, and + signs represent the values of X1 and X2. *Source*: Adapted from Fox (2016), Fig. 13.2

7.1.2 Data space and β space

It is also useful to visualize collinearity by comparing the representation in **data space** with the analogous view of the confidence ellipses for coefficients in **beta space**. To do so, we generate data from a known model $y = 3x_1 + 3x_2 + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 100)$ and various true correlations between x_1 and x_2, $\rho_{12} = (0, 0.8, 0.97)$ [1].

First, we use `MASS:mvrnorm()` to construct a list of data frames `XY` with specified values for the means and covariance matrices and a corresponding list of models, `mods`.

Working file: `R/collin-data-beta.R`

```
library(MASS)
library(car)

set.seed(421)             # reproducibility
N <- 200                  # sample size
mu <- c(0, 0)             # means
s <- c(1, 1)              # standard deviations
rho <- c(0, 0.8, 0.97)    # correlations
beta <- c(3, 3)           # true coefficients

# Specify a covariance matrix, with standard deviations s[1], s[2] and correlation r
Cov <- function(s, r){
  matrix(c(s[1],    r * prod(s),
           r * prod(s), s[2]), nrow = 2, ncol = 2)
}

# Generate a dataframe of X, y for each rho
# Fit the model for each
XY <- vector(mode ="list", length = length(rho))
mods <- vector(mode ="list", length = length(rho))
for (i in seq_along(rho)) {
  r <- rho[i]
  X <- mvrnorm(N, mu, Sigma = Cov(s, r))
```

[1] If group sizes are greatly unequal **and** homogeneity of variance is violated, then the F statistic is too liberal (p values too large) when large sample variances are associated with small group sizes. Conversely, the F statistic is too conservative if large variances are associated with large group sizes.

```
  colnames(X) <- c("x1", "x2")
  y <- beta[1] * X[,1] + beta[2] * X[,2] + rnorm(N, 0, 10)

  XY[[i]] <- data.frame(X, y=y)
  mods[[i]] <- lm(y ~ x1 + x2, data=XY[[i]])
}
```

The estimated coefficients in these models are:

```
coefs <- sapply(mods, coef)
colnames(coefs) <- c("Intercept", "b1", "b2")
coefs
#>             Intercept      b1    b2
#> (Intercept)      1.01 -0.0535 0.141
#> x1               3.18  3.4719 3.053
#> x2               1.68  2.9734 2.059
```

Then, we define a function to plot the data ellipse (`car::dataEllipse()`) for each data frame and confidence ellipse (`car::dataEllipse()`) in the corresponding fitted model. In this figure, I specify the x, y limits for each plot so that the relative sizes of these ellipses are comparable, so that variance inflation can be assesed visually.

```
do_plots <- function(XY, mod, r) {
  X <- as.matrix(XY[, 1:2])
  dataEllipse(X,
              levels= 0.95,
              col = "darkgreen",
              fill = TRUE, fill.alpha = 0.05,
              xlim = c(-3, 3),
              ylim = c(-3, 3), asp = 1)
  text(0, 3, bquote(rho == .(r)), cex = 2, pos = NULL)

  confidenceEllipse(mod,
                    col = "red",
                    fill = TRUE, fill.alpha = 0.1,
                    xlab = "x1 coefficient",
                    ylab = "x2 coefficient",
                    xlim = c(-5, 10),
```

```
                    ylim = c(-5, 10),
                    asp = 1)
  points(beta[1], beta[2], pch = "+", cex=2)
  abline(v=0, h=0, lwd=2)
}

op <- par(mar = c(4,4,1,1)+0.1,
          mfcol = c(2, 3),
          cex.lab = 1.5)

for (i in seq_along(rho)) {
  do_plots(XY[[i]], mods[[i]], rho[i])
}
par(op)
```

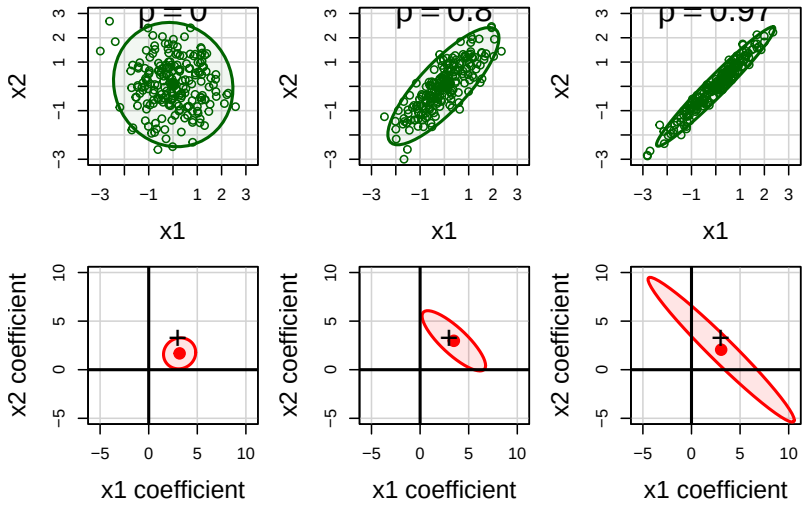

Figure 7.2: 95% Ddata ellipses for x1, x2 and the corresponding 95% confidence ellipses for their coefficients. In the confidence ellipse plots, reference lines show the value (0,0) for the null hypothesis and "+" marks the true values for the coefficients. This figure adapts an example by John Fox (2022).

Recall (Section #sec-data-beta) that the confidence ellipse for (β_1, β_2) is just a 90 degree rotation (and rescaling) of the data ellipse for (x_1, x_2): it is wide (more variance) in any direction

where the data ellipse is narrow.

The shadows of the confidence ellipses on the coordinate axes in Figure 7.2 represent the standard errors of the coefficients, and get larger with increasing ρ. This is the effect of variance inflation, described in the following section.

7.2 Measuring collinearity

7.2.1 Variance inflation factors

How can we measure the effect of collinearity? The essential idea is to compare, for each predictor the variance $s^2(\widehat{b_j})$ that the coefficient that x_j would have if it was totally unrelated to the other predictors to the actual variance it has in the given model.

For two predictors such as shown in Figure 7.2 the sampling variance of x_1 can be expressed as

$$s^2(\widehat{b_1}) = \frac{MSE}{(n-1)\, s^2(x_1)} \times \left[\frac{1}{1-r_{12}^2}\right]$$

The first term here is the variance of b_1 when the two predictors are uncorrelated. The term in brackets represents the **variance inflation factor** (Marquardt 1970), the amount by which the variance of the coefficient is multiplied as a consequence of the correlation r_{12} of the predictors. As $r_{12} \to 1$, the variances approaches infinity.

More generally, with any number of predictors, this relation has a similar form, replacing the simple correlation r_{12} with the multiple correlation predicting x_j from all others,

$$s^2(\widehat{b_j}) = \frac{MSE}{(n-1)\, s^2(x_j)} \times \left[\frac{1}{1-R^2_{j|\text{others}}}\right]$$

So, we have that the variance inflation factors are:

$$\text{VIF}_j = \frac{1}{1-R^2_{j\,|\,\text{others}}}$$

In practice, it is often easier to think in terms of the square root, $\sqrt{\text{VIF}_j}$ as the multiplier of the standard errors. The denominator, $1 - R^2_{j|\text{others}}$ is sometimes called **tolerance**, a term I don't find particularly useful.

For the cases shown in Figure 7.2 the VIFs and their square roots are:

```
vifs <- sapply(mods, car::vif)
colnames(vifs) <- paste("rho:", rho)
vifs
#>    rho: 0 rho: 0.8 rho: 0.97
#> x1      1     3.09      18.6
#> x2      1     3.09      18.6

sqrt(vifs)
#>    rho: 0 rho: 0.8 rho: 0.97
#> x1      1     1.76      4.31
#> x2      1     1.76      4.31
```

Note that when there are terms in the model with more than one df, such as education with four levels (and hence 3 df) or a polynomial term specified as `poly(x, 3)`, the standard VIF calculation gives results that vary with how those terms are coded in the model. Fox and Monette (1992) define *generalized*, GVIFs as the inflation in the squared area of the confidence ellipse for the coefficients of such terms, relative to what would be obtained with uncorrelated data. Visually, this can be seen by comparing the areas of the ellipses in the bottom row of Figure 7.2. Because the magnitude of the GVIF increases with the number of degrees of freedom for the set of parameters, Fox & Monette suggest the analog $\sqrt{\text{GVIF}^{1/2\text{df}}}$ as the measure of impact on standard errors.

Example: This example uses the `cars` data set in the `VisCollin` package containing various measures of size and performance on 406 models of automobiles from 1982. Interest is focused on predicting gas mileage, `mpg`.

```
data(cars, package = "VisCollin")
str(cars)
```

```
#> 'data.frame':    406 obs. of  10 variables:
#>  $ make    : Factor w/ 30 levels "amc","audi","bmw",..: 6 4 22 1 12 12 6 22 23 1 ...
#>  $ model   : chr  "chevelle" "skylark" "satellite" "rebel" ...
#>  $ mpg     : num  18 15 18 16 17 15 14 14 14 15 ...
#>  $ cylinder: int  8 8 8 8 8 8 8 8 8 8 ...
#>  $ engine  : num  307 350 318 304 302 429 454 440 455 390 ...
#>  $ horse   : int  130 165 150 150 140 198 220 215 225 190 ...
#>  $ weight  : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...
#>  $ accel   : num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
#>  $ year    : int  70 70 70 70 70 70 70 70 70 70 ...
#>  $ origin  : Factor w/ 3 levels "Amer","Eur","Japan": 1 1 1 1 1 1 1 1 1 1 ...
```

We fit a model predicting gas mileage (`mpg`) from the number of cylinders, engine displacement, horsepower, weight, time to accelerate from 0 – 60 mph and model year (1970–1982). Perhaps surprisingly, only `weight` and `year` appear to significantly predict gas mileage. What's going on here?

```
cars.mod <- lm (mpg ~ cylinder + engine + horse + weight + accel + year,
                data=cars)
Anova(cars.mod)
#> Anova Table (Type II tests)
#>
#> Response: mpg
#>           Sum Sq  Df F value Pr(>F)
#> cylinder      12   1    0.99   0.32
#> engine        13   1    1.09   0.30
#> horse          0   1    0.00   0.98
#> weight      1214   1  102.84 <2e-16 ***
#> accel          8   1    0.70   0.40
#> year        2419   1  204.99 <2e-16 ***
#> Residuals   4543 385
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We check the variance inflation factors, using `car::vif()`. We see that most predictors have very high VIFs, indicating moderately severe multicollinearity.

```
vif(cars.mod)
#> cylinder   engine    horse   weight    accel     year
#>    10.63    19.64     9.40    10.73     2.63     1.24

sqrt(vif(cars.mod))
#> cylinder   engine    horse   weight    accel     year
#>     3.26     4.43     3.07     3.28     1.62     1.12
```

According to $\sqrt{\text{VIF}}$, the standard error of `cylinder` has been multiplied by 3.26 and it's t-value divided by this number, compared with the case when all predictors are uncorrelated. `engine`, `horse` and `weight` suffer a similar fate.

Connection with inverse of correlation matrix

In the linear regression model with standardized predictors, the covariance matrix of the estimated intercept-excluding parameter vector $\mathbf{b}^\star$ has the simpler form,

$$\mathcal{V}(\mathbf{b}^\star) = \frac{\sigma^2}{n-1}\mathbf{R}_X^{-1} \; .$$

where $\mathbf{R}_X$ is the correlation matrix among the predictors. It can then be seen that the VIF_j are just the diagonal entries of $\mathbf{R}_X^{-1}$.
More generally, the matrix $\mathbf{R}_X^{-1} = (r^{ij})$, when standardized to a correlation matrix as $-r^{ij}/\sqrt{r^{ii}\, r^{jj}}$ gives the matrix of all partial correlations, $r_{ij\,|\,\text{others}}$. }

7.2.2 Collinearity diagnostics

OK, we now know that large VIF_j indicate predictor coefficients whose estimation is degraded due to large $R^2_{j\,|\,\text{others}}$. But for this to be useful, we need to determine:

- how many dimensions in the space of the predictors are associated with nearly collinear relations?
- which predictors are most strongly implicated in each of these?

Answers to these questions are provided using measures developed by Belsley and colleagues (Belsley, Kuh, and Welsch 1980; Belsley 1991). These measures are based on the eigenvalues $\lambda_1, \lambda_2, \ldots \lambda_p$ of the correlation matrix R_X of the predictors (preferably centered and scaled, and not including the constant term for the intercept), and the corresponding eigenvectors in the columns of $\mathbf{V}_{p \times p}$, given by the the eigen decomposition

$$\mathbf{R}_X = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T$$

By elementary matrix algebra, the eigen decomposition of $\mathbf{R}_{XX}^{-1}$ is then

$$\mathbf{R}_X^{-1} = \mathbf{V} \mathbf{\Lambda}^{-1} \mathbf{V}^T \; , \tag{7.1}$$

so, $\mathbf{R}_X$ and $\mathbf{R}_{XX}^{-1}$ have the same eigenvectors, and the eigenvalues of $\mathbf{R}_X^{-1}$ are just λ_i^{-1}. Using Equation 7.1, the variance inflation factors may be expressed as

$$\text{VIF}_j = \sum_{k=1}^{p} \frac{V_{jk}^2}{\lambda_k} \; ,$$

which shows that only the *small* eigenvalues contribute to variance inflation, but only for those predictors that have large eigenvector coefficients on those small components. These facts lead to the following diagnostic statistics for collinearity:

- **Condition indices**: The smallest of the eigenvalues, those for which $\lambda_j \approx 0$, indicate collinearity and the number of small values indicates the number of near collinear relations. Because the sum of the eigenvalues, $\Sigma \lambda_i = p$ increases with the number of predictors p, it is useful to scale them all in relation to the largest. This leads to *condition indices*, defined as $\kappa_j = \sqrt{\lambda_1 / \lambda_j}$. These have the property that the resulting numbers have common interpretations regardless of the number of predictors.
  - For completely uncorrelated predictors, all $\kappa_j = 1$.
  - $\kappa_j \to \infty$ as any $\lambda_k \to 0$.
- **Variance decomposition proportions**: Large VIFs indicate variables that are involved in *some* nearly collinear relations, but they don't indicate *which* other variable(s) each is involved with. For this purpose, Belsley, Kuh, and

> Welsch (1980) and Belsley (1991) proposed calculation of the proportions of variance of each variable associated with each principal component as a decomposition of the coefficient variance for each dimension.

These measures can be calculated using `VisCollin::colldiag()`. For the current model, the usual display contains both the condition indices and variance proportions. However, even for a small example, it is often difficult to know what numbers to pay attention to.

```
(cd <- colldiag(cars.mod, center=TRUE))
#> Condition
#> Index    Variance Decomposition Proportions
#>           cylinder engine horse weight accel year
#> 1   1.000 0.005    0.003  0.005 0.004  0.009 0.010
#> 2   2.252 0.004    0.002  0.000 0.007  0.022 0.787
#> 3   2.515 0.004    0.001  0.002 0.010  0.423 0.142
#> 4   5.660 0.309    0.014  0.306 0.087  0.063 0.005
#> 5   8.342 0.115    0.000  0.654 0.715  0.469 0.052
#> 6  10.818 0.563    0.981  0.032 0.176  0.013 0.004
```

Belsley (1991) recommends that the sources of collinearity be diagnosed (a) only for those components with large κ_j, and (b) for those components for which the variance proportion is large (say, ≥ 0.5) on *two* or more predictors. The print method for `"colldiag"` objects has a `fuzz` argument controlling this.

```
print(cd, fuzz = 0.5)
#> Condition
#> Index    Variance Decomposition Proportions
#>           cylinder engine horse weight accel year
#> 1   1.000  .        .      .     .      .     .
#> 2   2.252  .        .      .     .      .    0.787
#> 3   2.515  .        .      .     .      .     .
#> 4   5.660  .        .      .     .      .     .
#> 5   8.342  .        .     0.654 0.715   .     .
#> 6  10.818 0.563    0.981   .     .      .     .
```

The mystery is solved, if you can read that table with these recommendations in mind. There are two nearly collinear rela-

tions among the predictors, corresponding to the two smallest dimensions.

- Dimension 5 reflects the high correlation between horsepower and weight,
- Dimension 6 reflects the high correlation between number of cylinders and engine displacement.

Note that the high variance proportion for `year` (0.787) on the second component creates no problem and should be ignored because (a) the condition index is low and (b) it shares nothing with other predictors.

7.2.3 Tableplots

The default tabular display of condition indices and variance proportions from `colldiag()` is what triggered the comparison to "Where's Waldo". It suffers from the fact that the important information — (a) how many Waldos? (b) where are they hiding — is disguised by being embedded in a sea of mostly irrelevant numbers. The simple option of using a principled `fuzz` factor helps considerably, but not entirely.

The simplified tabular display above can be improved to make the patterns of collinearity more visually apparent and to signify warnings directly to the eyes. A **tableplot** (Kwan, Lu, and Friendly 2009) is a semi-graphic display that presents numerical information in a table using shapes proportional to the value in a cell and other visual attributes (shape type, color fill, and so forth) to encode other information.

For collinearity diagnostics, these show:

- the condition indices, using using *squares* whose background color is red for condition indices > 10, green for values > 5 and green otherwise, reflecting danger, warning and OK respectively. The value of the condition index is encoded within this using a white square whose side is proportional to the value (up to some maximum value, `cond.max` that fills the cell).

- Variance decomposition proportions are shown by filled *circles* whose radius is proportional to those values and are filled (by default) with shades ranging from white through pink to red. Rounded values of those diagnostics are printed in the cells.

The tableplot below (Figure 7.3) encodes all the information from the values of `colldiag()` printed above (but using `prop.col` color breaks such that variance proportions < 0.3 are shaded white). The visual message is that one should attend to collinearities with large condition indices **and** large variance proportions implicating two or more predictors.

```
tableplot(cd, title = "Tableplot of cars data", cond.max = 30 )
```

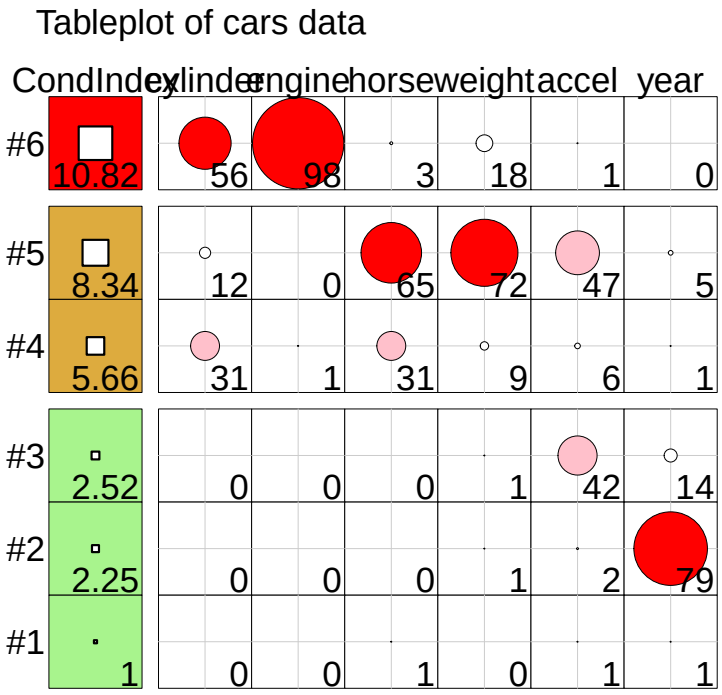

Figure 7.3: Tableplot of condition indices and variance proportions for the Cars data. In column 1, the square symbols are scaled relative to a maximum condition index of 30. In the remaining columns, variance proportions (times 100) are shown as circles scaled relative to a maximum of 100.

7.2.4 Collinearity biplots

As we have seen, the collinearity diagnostics are all functions of the eigenvalues and eigenvectors of the correlation matrix of the predictors in the regression model, or alternatively, the SVD of the $\mathbf{X}$ matrix in the linear model (excluding the constant). The standard biplot (Gabriel 1971; Gower and Hand 1996) (see: Section 4.2) can be regarded as a multivariate analog of a scatterplot, obtained by projecting a multivariate sample into a low-dimensional space (typically of 2 or 3 dimensions) accounting for the greatest variance in the data.

However the standard biplot is less useful for visualizing the relations among the predictors that lead to nearly collinear relations. Instead, biplots of the **smallest dimensions** show these relations directly, and can show other features of the data as well, such as outliers and leverage points. We use `prcomp(X, scale.=TRUE)` to obtain the PCA of the correlation matrix of the predictors:

```
cars.X <- cars |>
  select(where(is.numeric)) |>
  select(-mpg) |>
  tidyr::drop_na()
cars.pca <- prcomp(cars.X, scale. = TRUE)
cars.pca
#> Standard deviations (1, .., p=6):
#> [1] 2.070 0.911 0.809 0.367 0.245 0.189
#>
#> Rotation (n x k) = (6 x 6):
#>             PC1    PC2    PC3    PC4     PC5     PC6
#> cylinder -0.454 0.1869 -0.168  0.659 -0.2711  0.4725
#> engine   -0.467 0.1628 -0.134  0.193 -0.0109 -0.8364
#> horse    -0.462 0.0177  0.123 -0.620 -0.6123  0.1067
#> weight   -0.444 0.2598 -0.278 -0.350  0.6860  0.2539
#> accel     0.330 0.2098 -0.865 -0.143 -0.2774 -0.0337
#> year      0.237 0.9092  0.335 -0.025 -0.0624 -0.0142
```

The standard deviations above are the square roots $\sqrt{\lambda_j}$ of the eigenvalues of the correlation matrix, and are returned in the `sdev` component of the `"prcomp"` object. The eigenvectors

are returned in the `rotation` component, whose directions are arbitrary. Because we are interested in seeing the relative magnitude of variable vectors, we are free to multiply them by any constant to make them more visible in relation to the scores for the cars.

```
cars.pca$rotation <- -2.5 * cars.pca$rotation    # reflect & scale the variable vectors

ggp <- fviz_pca_biplot(
  cars.pca,
  axes = 6:5,
  geom = "point",
  col.var = "blue",
  labelsize = 5,
  pointsize = 1.5,
  arrowsize = 1.5,
  addEllipses = TRUE,
  ggtheme = ggplot2::theme_bw(base_size = 14),
  title = "Collinearity biplot for cars data")

# add point labels for outlying points
dsq <- heplots::Mahalanobis(cars.pca$x[, 6:5])
scores <- as.data.frame(cars.pca$x[, 6:5])
scores$name <- rownames(scores)

ggp + geom_text_repel(data = scores[dsq > qchisq(0.95, df = 6),],
                aes(x = PC6,
                    y = PC5,
                    label = name),
                vjust = -0.5,
                size = 5)
```

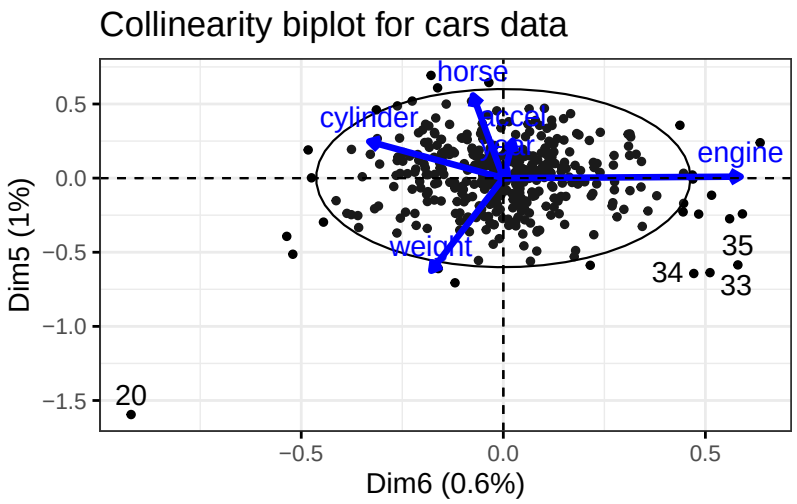

Figure 7.4: Collinearity biplot of the Cars data, showing the last two dimensions. The projections of the variable vectors on the coordinate axes are proportional to their variance proportions. To reduce graphic clutter, only the most outlying observations in predictor space are identified by case labels. An extreme outlier (case 20) appears in the lower left corner.

As with the tabular display of variance proportions, Waldo is hiding in the dimensions associated with the smallest eigenvalues (largest condition indices).
As well, it turns out that outliers in the predictor space (also high leverage observations) can often be seen as observations far from the centroid in the space of the smallest principal components.

The projections of the variable vectors in Figure 7.4 on the Dimension 5 and Dimension 6 axes are proportional to their variance proportions shown above. The relative lengths of these variable vectors can be considered to indicate the extent to which each variable contributes to collinearity for these two near-singular dimensions.

Thus, we see again that Dimension 6 is largely determined by `engine` size, with a substantial (negative) relation to `cylinder`. Dimension 5 has its' strongest relations to `weight` and `horse`.

Moreover, there is one observation, #20, that stands out as an outlier in predictor space, far from the centroid. It turns out that this vehicle, a Buick Estate wagon, is an early-year (1970) American behemoth, with an 8-cylinder, 455 cu. in, 225 horse-power engine, and able to go from 0 to 60 mph in 10 sec. (Its MPG is only slightly under-predicted from the regression model, however.)

7.3 Remedies for collinearity: What can I do?

Collinearity is often a **data** problem, for which there is no magic cure. Nevertheless there are some general guidelines and useful techniques to address this problem.

- **Pure prediction**: If we are only interested in predicting / explaining an outcome, and not the model coefficients or which are "significant", collinearity can be largely ignored. The fitted values are unaffected by collinearity, even in the case of perfect collinearity as shown in Figure 7.1 (b).
- **structural collinearity**: Sometimes collinearity results from structural relations among the variables that relate to how they have been defined.
  - For example, polynomial terms, like x, x^2, x^3 or interaction terms like $x_1, x_2, x_1 * x_2$ are necessarily correlated. A simple cure is to *center* the predictors at their means, using $x - \bar{x}, (x - \bar{x})^2, (x - \bar{x})^3$ or $(x_1 - \bar{x}_1), (x_2 - \bar{x}_2), (x_1 - \bar{x}_1) * (x_2 - \bar{x}_2)$. Centering removes the spurious ill-conditioning, thus reducing the VIFs. Note that in polynomial models, using `y ~ poly(x, 3)` to specify a cubic model generates *orthogonal* (uncorrelated) regressors, whereas in `y ~ x + I(x^2) + I(x^3)` the terms have built-in correlations.
  - When some predictors share a common cause, as in GNP or population in time-series or cross-national data, you can reduce collinearity by re-defining predictors to reflect *per capita measures*. In a related

example with sports data, when you have cumulative totals (e.g., runs, hits, homeruns in baseball) for players over years, expressing these measures as *per year* will reduce the common effect of longevity on these measures.

- **Model re-specification**:
  - Drop one or more regressors that have a high VIF if they are not deemed to be essential
  - Replace highly correlated regressors with linear combination(s) of them. For example, two related variables, x_1 and x_2 can be replaced without any loss of information by replacing them with their sum and difference, $z_1 = x_1 + x_2$ and $z_2 = x_1 - x_2$. For example, in a data set on fitness, we may have correlated predictors of resting pulse rate and pulse rate while running. Transforming these to average pulse rate and their difference gives new variables which are interpretable and less correlated.
- **Statistical remedies**:
  - Transform the predictors $\mathbf{X}$ to uncorrelated principal component scores $\mathbf{Z} = \mathbf{XV}$, and regress $\mathbf{y}$ on $\mathbf{Z}$. These will have the identical overall model fit without loss of information. A related technique is *incomplete* principal components regression, where some of the smallest dimensions (those causing collinearity) are omitted from the model. The trade-off is that it may be more difficult to interpret what the model means, but this can be countered with a biplot, showing the projections of the original variables into the reduced space of the principal components.
  - use **regularization methods** such as ridge regression and lasso, which correct for collinearity by introducing shrinking coefficients towards 0, introducing a small amount of bias, . See the genridge package and its `pkgdown` documentation for visualization methods.

– use Bayesian regression; if multicollinearity prevents a regression coefficient from being estimated precisely, then a prior on that coefficient will help to reduce its posterior variance.

Example: Centering

To illustrate the effect of centering a predictor in a polynomial model, we generate a perfect quadratic relationship, $y = x^2$ and consider the correlations of y with x and with $(x - \bar{x})^2$. The correlation of y with x is 0.97, while the correlation of y with $(x - \bar{x})^2$ is zero.

```
x <- 1:20
y1 <- x^2
y2 <- (x - mean(x))^2
XY <- data.frame(x, y1, y2)

(R <- cor(XY))
#>        x    y1    y2
#> x  1.000 0.971 0.000
#> y1 0.971 1.000 0.238
#> y2 0.000 0.238 1.000
```

The effect of centering here is remove the linear association in what is a purely quadratic relationship, as can be seen by plotting `y1` and `y2` against `x`.

```
r1 <- R[1, 2]
r2 <- R[1, 3]

gg1 <-
ggplot(XY, aes(x = x, y = y1)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", formula = y~x, linewidth = 2, se = FALSE) +
  labs(x = "X", y = "Y") +
  theme_bw(base_size = 16) +
  annotate("text", x = 5, y = 350, size = 6,
           label = paste("X Uncentered\nr =", round(r1, 3)))

gg2 <-
```

```
  ggplot(XY, aes(x = x, y = y2)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", formula = y~x, linewidth = 2, se = FALSE) +
  labs(x = "X", y = "Y") +
  theme_bw(base_size = 16) +
  annotate("text", x = 5, y = 80, size = 6,
           label = paste("X Centered\nr =", round(r2, 3)))

gg1 + gg2          # show plots side-by-side
```

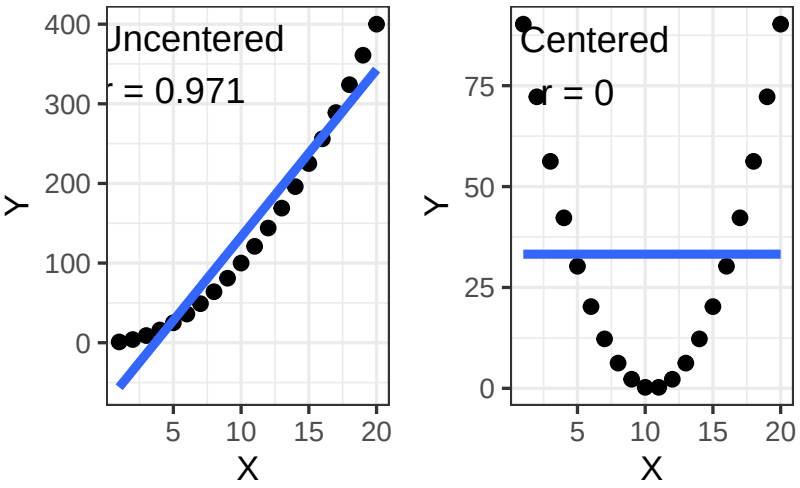

Figure 7.5: Centering a predictor removes the nessessary correlation in a quadratic regression

Example: Interactions

The data set `genridge::Acetylene` gives data from Marquardt and Snee (1975) on the `yield` of a chemical manufacturing process to produce acetylene in relation to reactor temperature (`temp`), the `ratio` of two components and the contact `time` in the reactor. A naive response surface model might suggest that yield is quadratic in time and there are potential interactions among all pairs of predictors.

```
data(Acetylene, package = "genridge")
acetyl.mod0 <- lm(yield ~ temp + ratio + time + I(time^2) +
```

```
                        temp:time + temp:ratio + time:ratio,
                    data=Acetylene)

(acetyl.vif0 <- vif(acetyl.mod0))
#>       temp      ratio       time  I(time^2)  temp:time temp:ratio
#>        383      10555      18080        564       9719       9693
#> ratio:time
#>        225
```

These results are horrible! How much does centering help? I first center all three predictors and then use `update()` to re-fit the same model using the centered data.

```
Acetylene.centered <-
  Acetylene |>
  mutate(temp = temp - mean(temp),
         time = time - mean(time),
         ratio = ratio - mean(ratio))

acetyl.mod1 <- update(acetyl.mod0, data=Acetylene.centered)
(acetyl.vif1 <- vif(acetyl.mod1))
#>       temp      ratio       time  I(time^2)  temp:time temp:ratio
#>      57.09       1.09      81.57      51.49      44.67      30.69
#> ratio:time
#>      33.33
```

This is far better, although still not great in terms of VIF. But, how much have we improved the situation by the simple act of centering the predictors? The square roots of the ratios of VIFs tell us the impact of centering on the standard errors.

```
sqrt(acetyl.vif0 / acetyl.vif1)
#>       temp      ratio       time  I(time^2)  temp:time temp:ratio
#>       2.59      98.24      14.89       3.31      14.75      17.77
#> ratio:time
#>       2.60
```

Finally, we use `poly(time, 2)` in the model for the centered data. Because there are multiple degree of freedom terms in the model, `car::vif()` calculates GVIFs here. The final column gives $\sqrt{\text{GVIF}^{1/2\text{df}}}$, the remaining effect of collinearity on the

standard errors of terms in this model.

```
acetyl.mod2 <- lm(yield ~ temp + ratio + poly(time, 2) +
                          temp:time + temp:ratio + time:ratio,
                  data=Acetylene.centered)

vif(acetyl.mod2, type = "term")
#>                   GVIF Df GVIF^(1/(2*Df))
#> temp             57.09  1            7.56
#> ratio             1.09  1            1.05
#> poly(time, 2) 1733.56  2            6.45
#> temp:time       44.67  1            6.68
#> temp:ratio      30.69  1            5.54
#> ratio:time      33.33  1            5.77
```

7.4 Ridge regression

7.4.1 What is ridge regression?

7.4.2 Univariate ridge trace plots

7.4.3 Bivariate ridge trace plots

```
#> Writing packages to  C:/R/Projects/Vis-MLM-quarto/bib/pkgs.txt
#> 17  packages used here:
#>  base, car, carData, datasets, dplyr, factoextra, genridge, ggplot2, ggrepel, graphics, grD
```

References

8 Hotelling's T^2

Just as the one- and two- sample univariate t-test is the gateway drug for understanding analysis of variance, so too Hotelling's T^2 test provides an entry point to multivariate analysis of variance. This simple case provides an entry point to understanding the collection of methods I call the **HE plot framework** for visualizing effects in multivariate linear models, which are a main focus of this book.

The essential idea is that Hotelling's T^2 provides a test of the difference in means between two groups on a *collection* of variables, $\mathbf{x} = x_1, x_2, \ldots x_p$ *simultaneously*, rather than one by one. This has the advantages that it:

- does not require p-value corrections for multiple tests (e.g., Bonferroni);
- combines the evidence from the multiple response variables, and *pools strength*, accumulating support across measures;
- clarifies how the multiple response are *jointly* related to the group effect along a single dimension, the *discriminant axis*;
- in so doing, it reduces the problem of testing differences for two (and potentially more) response variables to testing the difference on a single variable that best captures the multivariable relations.

After describing it's features, I use an example of a two-group T^2 test to illustrate the basic ideas behind multivariate tests and hypothesis error plots.

Packages

In this chapter we use the following packages. Load them now.

```
library(car)
library(heplots)
library(Hotelling)
library(ggplot2)
library(dplyr)
library(tidyr)
```

8.1 T^2 as a generalized t-test

Hotelling's T^2 (Hotelling 1931) is an analog the square of a univariate t statistic, extended to the Consider the basic one-sample t-test, where we wish to test the hypothesis that the mean $\bar{x}$ of a set of N measures on a test of basic math, with standard deviation s does not differ from an assumed mean $\mu_0 = 150$ for a population. The t statistic for testing $H_0 : \mu = \mu_0$ against the two-sided alternative, $H_0 : \mu \neq \mu_0$ is

$$t = \frac{(\bar{x} - \mu_0)}{s/\sqrt{N}} = \frac{(\bar{x} - \mu_0)\sqrt{N}}{s}$$

Squaring this gives

$$t^2 = \frac{N(\bar{x} - \mu_0)^2}{s} = N(\bar{x} - \mu_0)(s^2)^{-1}(\bar{x} - \mu_0)$$

Now consider we also have measures on a test of solving word problems for the same sample. Then, a hypothesis test for the means on basic math (BM) and word problems (WP) is the test of the means of these two variables jointly equal their separate values, say, $(150, 100)$.

$$H_0 : \mu = \mu_{\mathbf{0}} = \begin{pmatrix} \mu_{0,BM} \\ \mu_{0,WP} \end{pmatrix} = \begin{pmatrix} 150 \\ 100 \end{pmatrix}$$

Hotelling's T^2 is then the analog of t^2, with the variance-covariance matrix $\mathbf{S}$ of the scores on (BM, WP) replacing the variance of a single score. This is nothing more than the squared Mahalanobis distance between the sample mean vector

$(\bar{x}_{BM}, \bar{x}_{WP})^T$ and the hypothesized means μ_0, in the metric of $\mathbf{S}$, as shown in Figure 8.1.

$$\begin{aligned} T^2 &= N(\bar{\mathbf{x}} - \mu_0)^T \, \mathbf{S}^{-1} \, (\bar{\mathbf{x}} - \mu_0) \\ &= N D_M^2(\bar{\mathbf{x}}, \mu_0) \end{aligned}$$

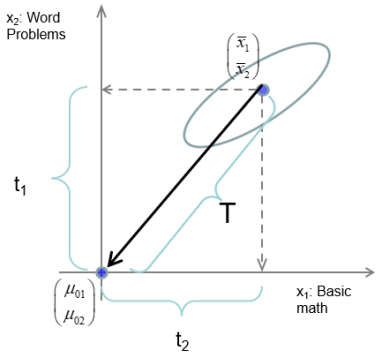

Figure 8.1: Hotelling's T^2 statistic as the squared distance between the sample means and hypothesized means relative to the variance-covariance matrix. *Source*: Author

8.2 T^2 properties

Aside from it's elegant geometric interpretation Hotelling's T^2 has simple properties that aid in understanding the extension to more complex multivariate tests.

- **Maximum** t^2 : Consider constructing a new variable w as a linear combination of the scores in a matrix $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}]$ with weights $\mathbf{a}$,

  $$w = a_1\mathbf{x_1} + a_2\mathbf{x_2} + \cdots + a_p\mathbf{x_p} = \mathbf{Xa}$$

  Hotelling's T^2 is then the maximum value of a univariate $t^2(\mathbf{a})$ over all possible choices of the weights in $\mathbf{a}$. In this way, Hotellings test reduces a multivariate problem to a univariate one.

- **Eigenvalue** : Hotelling showed that T^2 is the one non-zero eigenvalue (latent root) λ of the matrix $\mathbf{Q}_H = N(\bar{\mathbf{x}} - \mu_0)^T(\bar{\mathbf{x}} - \mu_0)$ relative to $\mathbf{Q}_E = \mathbf{S}$ that solves the equation

$$(\mathbf{Q}_H - \lambda \mathbf{Q}_E)\mathbf{a} = 0 \tag{8.1}$$

  In more complex MANOVA problems, there are more than one non-zero latent roots, $\lambda_1, \lambda_2, \ldots \lambda_s$, and test statistics are functions of these.

- **Eigenvector** : The corresponding eigenvector is $\mathbf{a} = \mathbf{S}^{-1}(\bar{\mathbf{x}} - \mu_0)$. These are the (raw) *discriminant coefficients*, giving the relative contribution of each variable to T^2.

- **Critical values** : For a single response, the square of a t statistic with $N-1$ degrees of freedom is an $F(1, N-1)$ statistic. But we chose $\mathbf{a}$ to give the *maximum* $t^2(\mathbf{a})$; this can be taken into account with a transformation of T^2 to give an **exact** F test with the correct sampling distribution:

$$F^* = \frac{N-p}{p(N-1)} T^2 \;\sim\; F(p, N-p) \tag{8.2}$$

- **Invariance under linear transformation** : Just as a univariate t-test is unchanged if we apply a linear transformation to the variable, $x \to ax + b$, T^2 is invariant under all linear (*affine*) transformations,

$$\mathbf{x}_{p\times 1} \to \mathbf{C}_{p\times p}\mathbf{x} + \mathbf{b}$$

  So, you get the same results if you convert penguins flipper lengths from millimeters to centimeters or inches. The same is true for all MANOVA tests.

- **Two-sample tests** : With minor variations in notation, everything above applies to the more usual test of equality of multivariate means in a two sample test of $H_0 : \mu_1 = \mu_2$.

$$T^2 = N(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \, \mathbf{S}_p^{-1} \, (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

  where $\mathbf{S}_p$ is the pooled within-sample variance covariance matrix.

Example

The data set `heplots::mathscore` gives (fictitious) scores on a test of basic math skills (BM) and solving word problems (WP) for two groups of $N = 6$ students in an algebra course, each taught by different instructors.

```
data(mathscore, package = "heplots")
str(mathscore)
#> 'data.frame':    12 obs. of  3 variables:
#>  $ group: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 2 2 2 2 ...
#>  $ BM   : int  190 170 180 200 150 180 160 190 150 160 ...
#>  $ WP   : int  90 80 80 120 60 70 120 150 90 130 ...
```

You can carry out the test that the means for both variables are jointly equal using either `Hotelling::hotelling.test()` (Curran and Hersh 2021) or `car::Anova()`,

```
hotelling.test(cbind(BM, WP) ~ group, data=mathscore) |> print()
#> Test stat:  64.174
#> Numerator df:  2
#> Denominator df:  9
#> P-value:  0.0001213

math.mod <- lm(cbind(BM, WP) ~ group, data=mathscore)
Anova(math.mod)
#>
#> Type II MANOVA Tests: Pillai test statistic
#>       Df test stat approx F num Df den Df  Pr(>F)
#> group  1     0.865     28.9      2      9 0.00012 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What's wrong with just doing the two t-tests (or equivalent F-test with `lm()`)?

```
Anova(mod1 <- lm(BM ~ group, data=mathscore))
#> Anova Table (Type II tests)
#>
#> Response: BM
```

```
#>           Sum Sq Df F value Pr(>F)
#> group       1302  1    4.24  0.066 .
#> Residuals   3071 10
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Anova(mod2 <- lm(WP ~ group, data=mathscore))
#> Anova Table (Type II tests)
#>
#> Response: WP
#>           Sum Sq Df F value Pr(>F)
#> group       4408  1    10.4  0.009 **
#> Residuals   4217 10
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this, we might conclude that the two groups do *not* differ significantly on Basic Math but strongly differ on Word problems. But the two univariate tests do not take the correlation among the mean differences into account.

To see the differences between the groups on both variables together, we draw their data (68%) ellipses, using `heplots::covEllpses()`

```
colors <- c("darkgreen", "blue")
covEllipses(mathscore[,c("BM", "WP")], mathscore$group,
            pooled=FALSE,
            col = colors,
            fill = TRUE,
            fill.alpha = 0.05,
            cex = 2, cex.lab = 1.5,
            asp = 1,
            xlab="Basic math", ylab="Word problems")
# plot points
pch <- ifelse(mathscore$group==1, 15, 16)
col <- ifelse(mathscore$group==1, colors[1], colors[2])
points(mathscore[,2:3], pch=pch, col=col, cex=1.25)
```

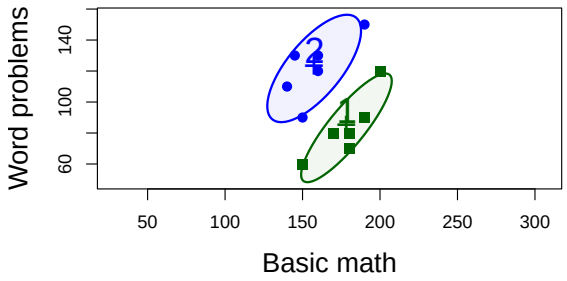

Figure 8.2: Data ellipses for the `mathscore` data, enclosing approximately 68% of the observations in each group

We can see that:

- Group 1 > Group 2 on Basic Math, but worse on Word Problems
- Group 2 > Group 1 on Word Problems, but worse on Basic Math
- Within each group, those who do better on Basic Math also do better on Word Problems

We can also see why the univariate test, at least for Basic math is non-significant: the scores for the two groups overlap considerably on the horizontal axis. They are slightly better separated along the vertical axis for word problems. The plot also reveals why Hotelling's T^2 reveals such a strongly significant result: the two groups are very widely separated along an approximately 45^o line between them.

A relatively simple interpretation is that the groups don't really differ in overall math ability, but perhaps the instructor in Group 1 put more focus on basic math skills, while the instructor for Group 2 placed greater emphasis on solving word problems.

In Hotelling's T^2, the "size" of the difference between the means (labeled "1" and "2") is assessed relative to the pooled within-group covariance matrix $\mathbf{S}_p$, which is just a size-weighted average of the two within-sample matrices, $\mathbf{S}_1$ and $\mathbf{S}_2$,

$$\mathbf{S}_p = [(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2]/(n_1 + n_2 - 2)$$

Visually, imagine sliding the the separate data ellipses to the grand mean, $(\bar{x}_{\text{BM}}, \bar{x}_{\text{WP}})$ and finding their combined data ellipse. This is just the data ellipse of the sample of deviations of the scores from their group means, or that of the residuals from the model `lm(cbind(BM, WP) ~ group, data=mathscore)`

To see this, we plot $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_p$ together,

```
covEllipses(mathscore[,c("BM", "WP")], mathscore$group,
            col = c(colors, "red"),
            fill = c(FALSE, FALSE, TRUE),
            fill.alpha = 0.3,
            cex = 2, cex.lab = 1.5,
            asp = 1,
            xlab="Basic math", ylab="Word problems")
```

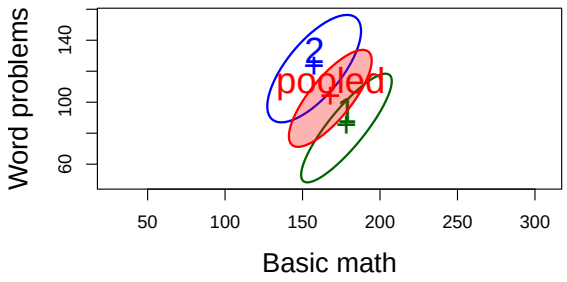

Figure 8.3: Data ellipses and the pooled covariance matrix `mathscore` data.

One of the assumptions of the T^2 test (and of MANOVA) is that the within-group variance covariance matrices, $\mathbf{S}_1$ and $\mathbf{S}_2$, are the same. In Figure 8.3, you can see how the shapes of $\mathbf{S}_1$ and $\mathbf{S}_2$ are very similar, differing in that the variance of word Problems is slightly greater for group 2. In Chapter XX we take of the topic of visualizing tests of this assumption, based on Box's M-test.

8.3 HE plot and discriminant axis

As we describe in detail in Chapter XX, all the information relevant to the T^2 test and MANOVA can be captured in the

remarkably simple *Hypothesis Error* plot, which shows the relative size of two data ellipses,

- **H**: the data ellipse of the *fitted* values, which are just the group means on the two variables, $\bar{\mathbf{x}}$, corresponding to $\mathbf{Q}_H$ in Equation 8.1. In case of T^2, the **H** matrix is of rank 1, so the "ellipse" plots as a line.

```
# calculate H directly
fit <- fitted(math.mod)
xbar <- colMeans(mathscore[,2:3])
N <- nrow(mathscore)
crossprod(fit) - N * outer(xbar, xbar)
#>       BM    WP
#> BM  1302 -2396
#> WP -2396  4408

# same as: SSP for group effect from Anova
math.aov <- Anova(math.mod)
(H <- math.aov$SSP)
#> $group
#>       BM    WP
#> BM  1302 -2396
#> WP -2396  4408
```

- **E**: the data ellipse of the *residuals*, the deviations of the scores from the group means, $\mathbf{x} - \bar{\mathbf{x}}$, corresponding to $\mathbf{Q}_E$.

```
# calculate E directly
resids <- residuals(math.mod)
crossprod(resids)
#>      BM   WP
#> BM 3071 2808
#> WP 2808 4217

# same as: SSPE from Anova
(E <- math.aov$SSPE)
#>      BM   WP
#> BM 3071 2808
#> WP 2808 4217
```

8.3.1 `heplot()`

`heplots::heplot()` takes the model object, extracts the **H** and **E** matrices (from `summary(Anova(math.mod))`) and plots them. There are many options to control the details.

```
heplot(math.mod,
       fill=TRUE, lwd = 3,
       asp = 1,
       cex=2, cex.lab=1.8,
       xlab="Basic math", ylab="Word problems")
```

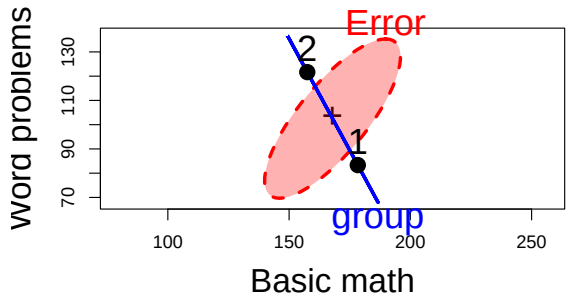

Figure 8.4: Hypothesis error plot of the `mathscore` data. The line through the group means is the H ellipse, which plots as a line here. The red ellipse labeled 'Error' represents the pooled within-group covariance matrix.

But the HE plot offers more:

- A visual test of significance: the **H** ellipse is scaled so that it projects *anywhere* outside the **E** ellipse, if and only if the test is significant at a given α level ($\alpha = 0.05$ by default)
- The **H** ellipse, which appears as a line, goes through the means of the two groups. This is also the *discriminant axis*, the direction in the space of the variables which maximally discriminates between the groups. That is, if we project the data points onto this line, we get the linear combination w which has the maximum possible univariate t^2.

You can see how the HE plot relates to the plots of the separate data ellipses by overlaying them in a single figure. We also plot the scores on the discriminant axis, by using this small function to find the orthogonal projection of a point $\mathbf{a}$ on the line joining two points, $\mathbf{p}_1$ and $\mathbf{p}_2$, which in math is $\mathbf{p}_1 + \frac{\mathbf{d}^T(\mathbf{a}-\mathbf{p}_1)}{\mathbf{d}^T\mathbf{d}}$, letting $\mathbf{d} = \mathbf{p}_1 - \mathbf{p}_2$.

```
dot <- function(x, y) sum(x*y)
project_on <- function(a, p1, p2) {
    a <- as.numeric(a)
    p1 <- as.numeric(p1)
    p2 <- as.numeric(p2)
    dot <- function(x,y) sum( x * y)
    t <- dot(p2-p1, a-p1) / dot(p2-p1, p2-p1)
    C <- p1 + t*(p2-p1)
    C
}
```

Then, we run the same code as before to plot the data ellipses, and follow this with a call to `heplot()` using the option `add=TRUE` which adds to an existing plot. Following this, we find the group means and draw lines projecting the points on the line between them.

```
covEllipses(mathscore[,c("BM", "WP")], mathscore$group,
            pooled=FALSE,
            col = colors,
            cex=2, cex.lab=1.5,
            asp=1,
            xlab="Basic math", ylab="Word problems"
            )
pch <- ifelse(mathscore$group==1, 15, 16)
col <- ifelse(mathscore$group==1, "red", "blue")
points(mathscore[,2:3], pch=pch, col=col, cex=1.25)

# overlay with HEplot (add = TRUE)
heplot(math.mod,
       fill=TRUE,
       cex=2, cex.lab=1.8,
       fill.alpha=0.2, lwd=c(1,3),
```

```
        add = TRUE,
        error.ellipse=TRUE)

# find group means
means <- mathscore |>
  group_by(group) |>
  summarize(BM = mean(BM), WP = mean(WP))

for(i in 1:nrow(mathscore)) {
    gp <- mathscore$group[i]
    pt <- project_on( mathscore[i, 2:3], means[1, 2:3], means[2, 2:3])
    segments(mathscore[i, "BM"], mathscore[i, "WP"], pt[1], pt[2], lwd = 1.2)
}
```

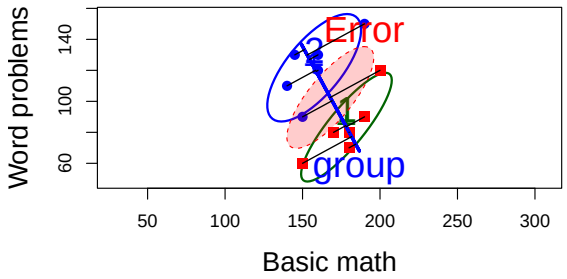

Figure 8.5: HE plot overlaid on top of the within-group data ellipses, with lines showing the projection of each point on the discriminant axis.

8.4 Discriminant analysis

Discriminant analysis for two-group designs or for one-way MANOVA essentially turns the problem around: Instead of asking whether the mean vectors for two or more groups are equal, discriminant analysis tries to find the linear combination w of the response variables that has the greatest separation among the groups, allowing cases to be best classified. It was developed by Fisher (1936) as a solution to the biological taxonomy problem of developing a rule to classify instances of flowers—in his famous case, Iris flowers—into known species (*I. setosa*, *I. versicolor*, *I. virginica*) on the basis of multiple measurements

(length and width of their sepals and petals).

```
(math.lda <- MASS::lda(group ~ ., data=mathscore))
#> Call:
#> lda(group ~ ., data = mathscore)
#>
#> Prior probabilities of groups:
#>   1   2
#> 0.5 0.5
#>
#> Group means:
#>    BM    WP
#> 1 178  83.3
#> 2 158 121.7
#>
#> Coefficients of linear discriminants:
#>        LD1
#> BM -0.0835
#> WP  0.0753
```

The coefficients give $w = -0.84\text{BM} + 0.75\text{WP}$. This is exactly the direction given by the line for the **H** ellipse in Figure 8.5.

To round this out, we can calculate the discriminant scores by multiplying the matrix $\mathbf{X}$ by the vector $\mathbf{a} = \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ of the discriminant weights.

```
math.lda$scaling
#>        LD1
#> BM -0.0835
#> WP  0.0753

scores <- cbind(group = mathscore$group,
                as.matrix(mathscore[, 2:3]) %*% math.lda$scaling) |>
  as.data.frame()
scores |>
  group_by(group) |>
  slice(1:3)
#> # A tibble: 6 x 2
#> # Groups:   group [2]
#>   group   LD1
```

```
#>   <dbl> <dbl>
#> 1     1 -9.09
#> 2     1 -8.17
#> 3     1 -9.01
#> 4     2 -4.33
#> 5     2 -4.58
#> 6     2 -5.75
```

Then a t-test on these scores gives Hotelling's T, accessed via the `statistic` component of `t.test()`

```
t <- t.test(LD1 ~ group, data=scores)$statistic
c(t, T2 =t^2)
#>     t  T2.t
#> -8.01 64.17
```

Finally, it is instructive to compare violin plots for the three measures, `BM`, `WP` and `LD1`. To do this with `ggplot2` requires reshaping the data from wide to long format so the plots can be faceted.

```
scores <- mathscore |>
  bind_cols(LD1 = scores[, "LD1"])

scores |>
  tidyr::gather(key = "measure", value ="Score", BM:LD1) |>
  mutate(measure = factor(measure, levels = c("BM", "WP", "LD1"))) |>
  ggplot(aes(x = group, y = Score, color = group, fill = group)) +
    geom_violin(alpha = 0.2) +
    geom_jitter(width = .2, size = 2) +
    facet_wrap( ~ measure, scales = "free", labeller = label_both) +
    scale_fill_manual(values = c("darkgreen", "blue")) +
    scale_color_manual(values = c("darkgreen", "blue")) +
    theme_bw(base_size = 14) +
    theme(legend.position = "none")
```

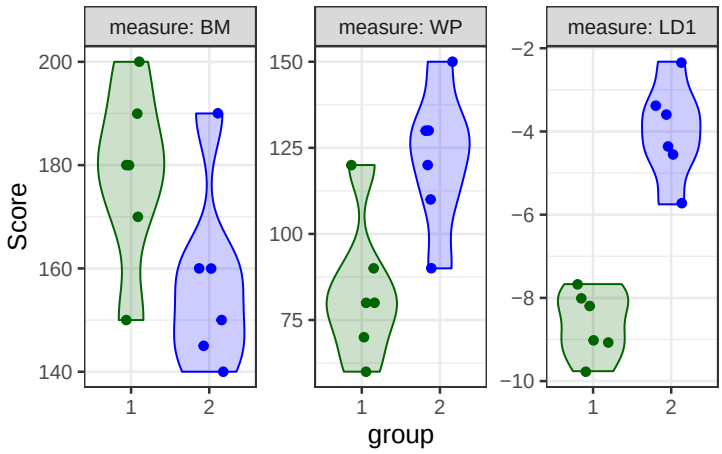

Figure 8.6: Violin plots comparing group 1 and 2 for the two observed measures and the linear discriminant score.

You can readily see how well the groups are separated on the discriminant axes, relative to the two individual variables.

8.5 Exercises

1. The value of Hotelling's T^2 found by `hotelling.test()` is 64.17. The value of the equivalent F statistic found by `Anova()` is 28.9. Verify that Equation 8.2 gives this result.

```
#> Writing packages to  C:/R/Projects/Vis-MLM-quarto/bib/pkgs.txt
#> 16  packages used here:
#>  base, broom, car, carData, corpcor, datasets, dplyr, ggplot2, graphics, grDevices, heplots
```

References

9 Visualizing Multivariate Models

Packages

In this chapter we use the following packages. Load them now

```
library(car)
library(heplots)
library(ggplot2)
library(dplyr)
library(tidyr)
```

9.1 HE plot framework

Chapter Chapter 8 illustrated the basic ideas of the framework for visualizing multivariate linear models in the context of a simple two group design, using Hotelling's T^2. These are illustrated in Figure 9.1.

- In data space, each group is summarized by its data ellipse, representing the means and covariances.
- Variation against the hypothesis of equal means can be seen by the **H** ellipse in the HE plot, representing the data ellipse of the fitted values. Error variance is shown in the **E** ellipse, representing the pooled within-group covariance matrix, $\mathbf{S}_p$ and the data ellipse of the residuals from the model.
- The MANOVA (or Hotelling's $T^2) is formally equivalent to a discriminant analysis, predicting group membership from the response variables. This effectively projects the p-dimensional space of the predictors into the smaller

canonical space that shows the greatest differences among the groups.

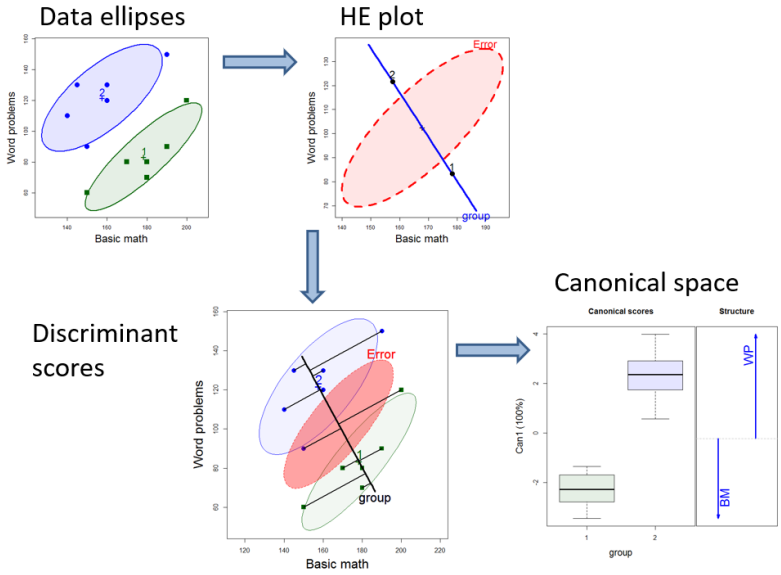

Figure 9.1: The Hypothesis Error plot framework. *Source*: author

For more complex models such as MANOVA with multiple factors or multivariate multivariate regression, there is one **H** ellipse for each term in the model. ...

9.1.1 HE plot details

9.2 Canonical discriminant analysis

```
#> Writing packages to  C:/R/Projects/Vis-MLM-quarto/bib/pkgs.txt
#> 14  packages used here:
#>  base, broom, car, carData, datasets, dplyr, ggplot2, graphics, grDevices, heplots, methods
```

10 Brief review of the multivariate linear model

The general multivariate linear model (MLM) can be understood as an extension of the univariate linear model, with the main difference being that there are multiple response variables instead of just one.

In this context, there are multiple techniques that can be applied depending on the structure of the variables at hand. For instance, with one or more continuous predictors and multiple response variables, one could use multivariate regression to obtain estimates useful for prediction. Instead, if the predictors are categorical, multivariate analysis of variance (MANOVA) can be applied to test for differences between groups. Again, this is akin to multiple regression and ANOVA in the univariate context – the same underlying model is utilized, but the tests for terms in the model are multivariate ones for the collection of all response variables, rather than univariate ones for a single response.

In each of these cases, the underlying MLM is given most compactly using the matrix equation,

$$\mathbf{Y}_{n\times p} = \mathbf{X}_{n\times q}\,\mathbf{B}_{q\times p} + \mathbf{U}_{n\times p} \ ,$$

where

- $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p)$ is the matrix of n observations on p responses;
- $\mathbf{X}$ is the model matrix with columns for q regressors, which typically includes an initial column of 1s for the intercept;
- $\mathbf{B}$ is a matrix of regression coefficients, one column for each response variable; and $\mathbf{U}$ is a matrix of errors.

The structure of the model matrix $\mathbf{X}$ is the same as the univariate linear model, and may contain, therefore,

- quantitative predictors, such as `age`, `income`, years of `education`
- transformed predictors like $\sqrt{\text{age}}$ or $\log$ income
- polynomial terms: age^2, $\text{age}^3, \ldots$ (using `poly(age, k)` in R)
- categorical predictors ("factors"), such as treatment (Control, Drug A, drug B), or sex; internally a factor with `k` levels is transformed to `k-1` dummy (0, 1) variables, representing comparisons with a reference level, typically the first.
- interaction terms, involving either quantitative or categorical predictors, e.g., `age * sex`, `treatment * sex`.

10.1 ANOVA -> MANOVA

10.2 MRA -> MMRA

10.3 ANCOVA -> MANCOVA

10.4 Repeated measures designs

11 Case studies

Packages

In this chapter we use the following packages. Load them now

```
library(car)
library(heplots)
library(candisc)
library(ggplot2)
library(dplyr)
library(tidyr)
library(corrgram)
```

This chapter presents some complete analyses of data sets that will be prominent in the book. Some of this material may later be moved to earlier chapters.

11.1 Neuro- and Social-cognitive measures in psychiatric groups

A Ph.D. dissertation by Laura Hartman (2016) at York University was designed to evaluate whether and how clinical patients diagnosed (on the DSM-IV) as schizophrenic or with schizoaffective disorder could be distinguished from each other and from a normal, control sample on collections of standardized tests in the following domains:

- **Neuro-cognitive**: processing speed, attention, verbal learning, visual learning and problem solving;
- **Social-cognitive**: managing emotions, theory of mind, externalizing, personalizing bias.

The study is an important contribution to clinical research because the two diagnostic categories are subtly different and their symptoms often overlap. Yet, they're very different and often require different treatments. A key difference between schizoaffective disorder and schizophrenia is the prominence of mood disorder involving bipolar, manic and depressive moods. With schizoaffective disorder, mood disorders are front and center. With schizophrenia, that is not a dominant part of the disorder, but psychotic ideation (hearing voices, seeing imaginary people) is.

11.1.1 Research questions

This example is concerned with the following substantitive questions:

- To what extent can patients diagnosed as schizophrenic or with schizoaffective disorder be distinguished from each other and from a normal control sample using a well-validated, comprehensive neurocognitive battery specifically designed for individuals with psychosis (Heinrichs et al. 2015) ?
- If the groups differ, do any of the cognitive domains show particularly larger or smaller differences among these groups?
- Do the neurocognitive measures discriminate among the in the same or different ways? If different, how many separate aspects or dimensions are distinguished?

Apart from the research interest, it could aid diagnosis and treatment if these similar mental disorders could be distinguished by tests in the cognitive domain.

11.1.2 Data

The clinical sample comprised 116 male and female patients who met the following criteria: 1) a diagnosis of schizophrenia ($n = 70$) or schizoaffective disorder ($n = 46$) confirmed by the Structured Clinical Interview for DSM-IV-TR Axis I Disorders;

2) were outpatients; 3) a history free of developmental or learning disability; 4) age 18-65; 5) a history free of neurological or endocrine disorder; and 6) no concurrent diagnosis of substance use disorder. Non-psychiatric control participants ($n = 146$) were screened for medical and psychiatric illness and history of substance abuse and were recruited from three outpatient clinics.

```
data(NeuroCog, package="heplots")
glimpse(NeuroCog)
#> Rows: 242
#> Columns: 10
#> $ Dx        <fct> Schizophrenia, Schizophrenia, Schizophrenia, Sch~
#> $ Speed     <int> 19, 8, 14, 7, 21, 31, -1, 17, 7, 37, 30, 26, 32,~
#> $ Attention <int> 9, 25, 23, 18, 9, 10, 8, 20, 30, 15, 27, 20, 23,~
#> $ Memory    <int> 19, 15, 15, 14, 35, 26, 3, 27, 26, 17, 28, 22, 2~
#> $ Verbal    <int> 33, 28, 20, 34, 28, 29, 20, 30, 26, 33, 34, 33, ~
#> $ Visual    <int> 24, 24, 13, 16, 29, 21, 12, 32, 27, 21, 19, 18, ~
#> $ ProbSolv  <int> 39, 40, 32, 31, 45, 33, 29, 29, 30, 33, 30, 39, ~
#> $ SocialCog <int> 28, 37, 24, 36, 28, 28, 28, 44, 39, 24, 32, 36, ~
#> $ Age       <int> 44, 26, 55, 53, 51, 21, 53, 56, 48, 46, 48, 31, ~
#> $ Sex       <fct> Female, Male, Female, Male, Male, Male, Male, Fe~
```

The diagnostic classification variable is called `Dx` in the data set. To facilitate answering questions regarding group differences, the following contrasts were applied: the first column compares the control group to the average of the diagnosed groups, the second compares the schizophrenia group against the schizoaffective group.

```
contrasts(NeuroCog$Dx)
#>                 [,1] [,2]
#> Schizophrenia   -0.5    1
#> Schizoaffective -0.5   -1
#> Control          1.0    0
```

In this analysis, we ignore the `SocialCog` variable. The primary focus is on the variables `Attention : ProbSolv`.

11.1.3 A first look

As always, plot the data first! We want an overview of the distributions of the variables to see the centers, spread, shape and possible outliers for each group on each variable.

The plot below combines the use of boxplots and violin plots to give an informative display. As we saw earlier (Chapter XXX), doing this with `ggplot2` requires reshaping the data to long format.

```
# Reshape from wide to long
NC_long <- NeuroCog |>
  dplyr::select(-SocialCog, -Age, -Sex) |>
  tidyr::gather(key = response, value = "value", Speed:ProbSolv)
# view 3 observations per group and measure
NC_long |>
  group_by(Dx) |>
  sample_n(3) |> ungroup()
#> # A tibble: 9 x 3
#>   Dx              response  value
#>   <fct>           <chr>     <int>
#> 1 Schizophrenia   Speed        39
#> 2 Schizophrenia   Visual       21
#> 3 Schizophrenia   Memory       40
#> 4 Schizoaffective ProbSolv     40
#> 5 Schizoaffective Speed        25
#> 6 Schizoaffective Verbal       48
#> 7 Control         Speed        33
#> 8 Control         ProbSolv     43
#> 9 Control         Attention    37
```

In the plot, we take care to adjust the transparency (`alpha`) values for the points, violin plots and boxplots so that all can be seen. Options for `geom_boxplot()` are used to give these greater visual prominence.

```
ggplot(NC_long, aes(x=Dx, y=value, fill=Dx)) +
  geom_jitter(shape=16, alpha=0.8, size=1, width=0.2) +
  geom_violin(alpha = 0.1) +
  geom_boxplot(width=0.5, alpha=0.4,
```

```
              outlier.alpha=1, outlier.size = 3, outlier.color = "red") +
  scale_x_discrete(labels = c("Schizo", "SchizAff", "Control")) +
  facet_wrap(~response, scales = "free_y", as.table = FALSE) +
  theme_bw() +
  theme(legend.position="bottom",
        axis.title = element_text(size = rel(1.2)),
        axis.text  = element_text(face = "bold"),
        strip.text = element_text(size = rel(1.2)))
```

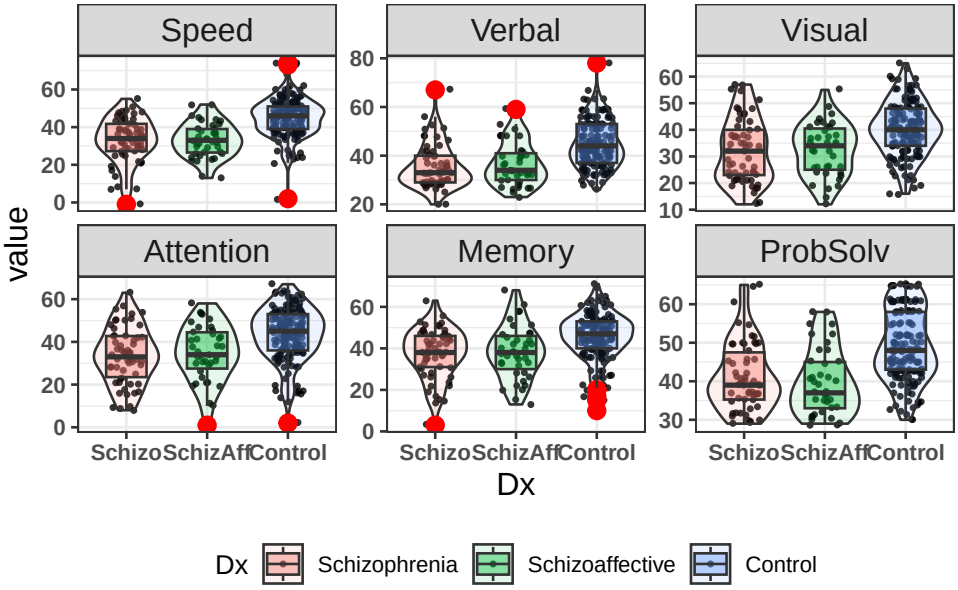

Figure 11.1: Boxplots and violin plots of the `NeuroCog` data.

We can see that the control participants score higher on all measures, but there is no consistent pattern of medians for the two patient groups. But these univariate summaries do not inform about the relations among variables.

11.1.4 Bivariate views

Corrgram

A corrgram (Friendly 2002) provides a useful reconnaisance plot of the bivariate correlations in the data set. It sup-

presses details, and allows focus on the overall pattern. The `corrgram::corrgram()` function has the ability to enhance perception by permuting the variables in the order of their variable vectors in a biplot, so more highly correlated variables are adjacent in the plot, and example of *effect ordering* for data displays (Friendly and Kwan 2003).

The plot below includes all variables except for `Dx` group. There are a number of `panel.*` functions for choosing how the correlation for each pair is rendered.

```
NeuroCog |>
  select(-Dx) |>
  corrgram(order = TRUE,
           diag.panel = panel.density,
           upper.panel = panel.pie)
```

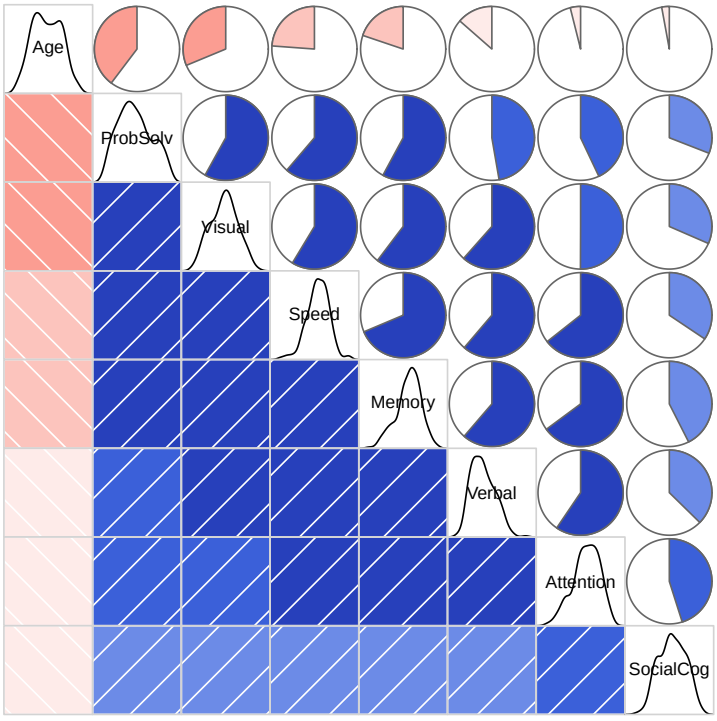

Figure 11.2: corrgram of the `NeuroCog` data. The upper and lower triangles use two different ways of encoding the value of the correlation for each pair of variables.

In this plot you can see that adjacent variables are more highly correlated than those more widely separated. The diagonal panels show that most variables are reasonably symmetric in their distributions. `Age`, not included in this analysis is negatively correlated with the others: older participants tend to do less well on these tests.

Scatterplot matrix

A scatterplot matrix gives a more detailed overview of all pairwise relations. The plot below suppresses the data points and summarizes the relation using data ellipses and regression lines. The model syntax, `~ Speed + ... |Dx`, treats `Dx` as a conditioning variable (similar to the use of the `color` aestheic in

ggplot2) giving a separate data ellipse and regression line for each group. (The legend is suppressed here. The groups are Schizophrenic, SchizoAffective, Normal.)

```
scatterplotMatrix(~ Speed + Attention + Memory + Verbal + Visual + ProbSolv | Dx,
  data=NeuroCog,
  plot.points = FALSE,
  smooth = FALSE,
  legend = FALSE,
  col = scales::hue_pal()(3),
  ellipse=list(levels=0.68))
```

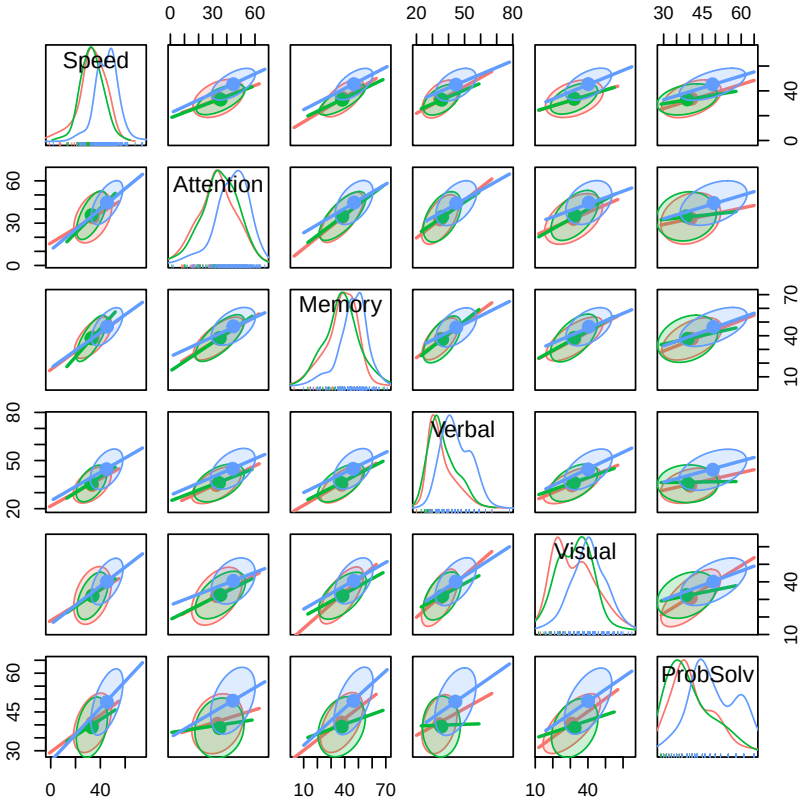

Figure 11.3: Scatterplot matrix of the NeuroCog data. Points are suppressed here, focusing on the data ellipses and regression lines. Colors for the groups: Schizophrenic (red), SchizoAffective (green), Normal (blue)

In this figure, we can see that the regression lines have similar slopes and similar data ellipses for the groups, though with a few exceptions.

TODO: Should we add biplot here?

11.2 Fitting the MLM

We proceed to fit the one-way MANOVA model.

```
NC.mlm <- lm(cbind(Speed, Attention, Memory, Verbal, Visual, ProbSolv) ~ Dx,
             data=NeuroCog)
Anova(NC.mlm)
#>
#> Type II MANOVA Tests: Pillai test statistic
#>    Df test stat approx F num Df den Df  Pr(>F)
#> Dx  2     0.299     6.89     12    470 1.6e-11 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The first research question is captured by the contrasts for the `Dx` factor shown above. We can test these with `car::linearHypothesis()`. The contrast `Dx1` for control vs. the diagnosed groups is highly significant,

```
# control vs. patients
print(linearHypothesis(NC.mlm, "Dx1"), SSP=FALSE)
#>
#> Multivariate Tests:
#>                  Df test stat approx F num Df den Df  Pr(>F)
#> Pillai            1     0.289     15.9      6    234 2.8e-15 ***
#> Wilks             1     0.711     15.9      6    234 2.8e-15 ***
#> Hotelling-Lawley  1     0.407     15.9      6    234 2.8e-15 ***
#> Roy               1     0.407     15.9      6    234 2.8e-15 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

but the second contrast, `Dx2`, comparing the schizophrenic and schizoaffective group, is not.

```
# Schizo vs SchizAff
print(linearHypothesis(NC.mlm, "Dx2"), SSP=FALSE)
#>
#> Multivariate Tests:
#>                  Df test stat approx F num Df den Df Pr(>F)
#> Pillai            1     0.006    0.249      6    234   0.96
#> Wilks             1     0.994    0.249      6    234   0.96
#> Hotelling-Lawley  1     0.006    0.249      6    234   0.96
#> Roy               1     0.006    0.249      6    234   0.96
```

11.2.1 HE plot

So the question becomes: how to understand these results. `heplot()` shows the visualization of the multivariate model in the space of two response variables (the first two by default). The result (Figure 11.4) tells a very simple story: The control group performs higher on higher measures than the diagnosed groups, which do not differ between themselves.

(For technical reasons, to abbreviate the group labels in the plot, we need to `update()` the MLM model after the labels are reassigned.)

```
# abbreviate levels for plots
NeuroCog$Dx <- factor(NeuroCog$Dx,
                      labels = c("Schiz", "SchAff", "Contr"))
NC.mlm <- update(NC.mlm)
```

```
op <- par(mar=c(5,4,1,1)+.1)
heplot(NC.mlm,
       fill=TRUE, fill.alpha=0.1,
       cex.lab=1.3, cex=1.25)
par(op)
```

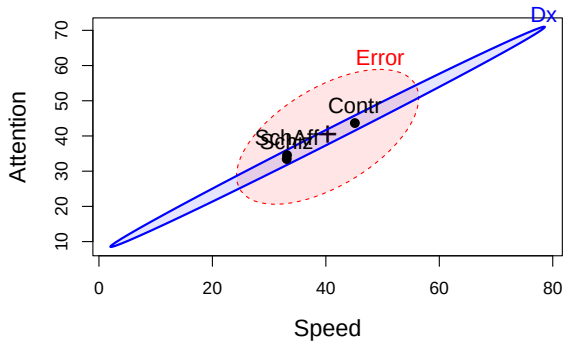

Figure 11.4: HE plot of Speed and Attention in the MLM for the `NeuroCog` data. The labeled points show the means of the groups on the two variables. The blue H ellipse for groups indicates the strong positive correlation of the group means.

This pattern is consistent across all of the response variables, as we see from a plot of `pairs(NC.mlm)`:

```
pairs(NC.mlm,
      fill=TRUE, fill.alpha=0.1,
      var.cex=2)
```

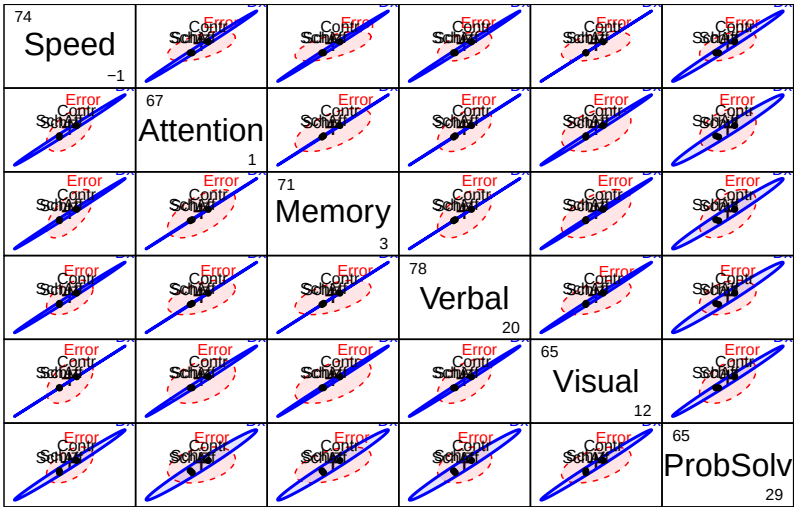

Figure 11.5: HE plot matrix of the MLM for `NeuroCog` data.

11.2.2 Canonical space

We can gain further insight, and a simplified plot showing all the response variables by projecting the MANOVA into the canonical space, which is entirely 2-dimensional (because $df_h = 2$). However, the output from `candisc()` shows that 98.5% of the mean differences among groups can be accounted for in one canonical dimension. ::: {.cell layout-align="center"}

```
NC.can <- candisc(NC.mlm)
NC.can
#>
#> Canonical Discriminant Analysis for Dx:
#>
#>    CanRsq Eigenvalue Difference Percent Cumulative
#> 1 0.29295    0.41433      0.408    98.5       98.5
#> 2 0.00625    0.00629      0.408     1.5      100.0
#>
#> Test of H0: The canonical correlations in the
#> current row and all that follow are zero
#>
#>   LR test stat approx F numDF denDF Pr(> F)
#> 1        0.703     7.53    12   468   9e-13 ***
#> 2        0.994     0.30     5   235    0.91
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

:::

Figure 11.6 is the result of the `plot()` method for class `"candisc"` objects, that is, the result of calling `plot(NC.can, ...)`. It plots the two canonical scores, $\mathbf{Z}_{n\times 2}$ for the subjects, together with data ellipses for each of the three groups.

```
pos <- c(4, 1, 4, 4, 1, 3)
col <- c("red", "darkgreen", "blue")
op <- par(mar=c(5,4,1,1)+.1)
plot(NC.can,
     ellipse=TRUE,
     rev.axes=c(TRUE,FALSE),
```

```
      pch=c(7,9,10),
      var.cex=1.2, cex.lab=1.5, var.lwd=2,  scale=4.5,
      col=col,
        var.col="black", var.pos=pos,
      prefix="Canonical dimension ")
par(op)
```

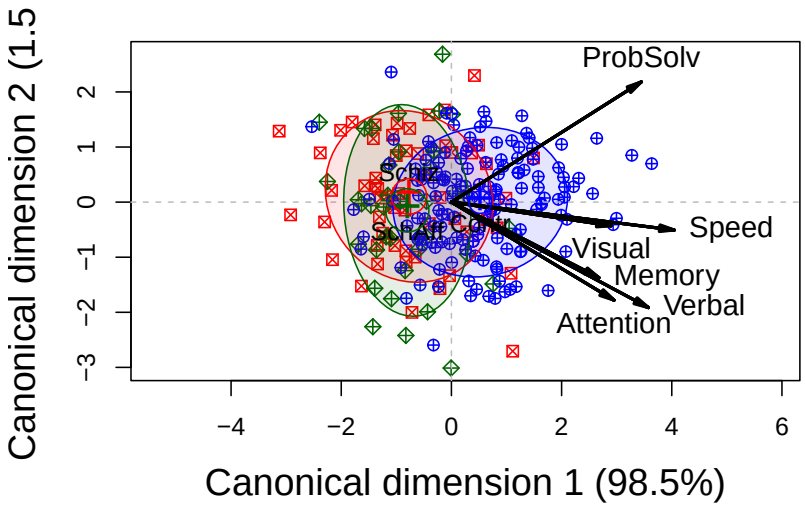

Figure 11.6: Canonical discriminant plot for the `NeuroCog` data MANOVA. Scores on the two canonical dimensions are plotted, together with 68% data ellipses for each group.

The interpretation of Figure 11.6 is again fairly straightforward. As noted earlier (REF???), the projections of the variable vectors in this plot on the coordinate axes are proportional to the correlations of the responses with the canonical scores. From this, we see that the normal group differs from the two patient groups, having higher scores on all the neurocognitive variables, most of which are highyl correlated. The problem solving measure is slightly different, and this, compared to the cluster of memory, verbal and attention, is what distinguishes the schizophrenic group from the schizoaffectives.

The separation of the groups is essentially one-dimensional, with the control group higher on all measures. Moreover, the variables

processing speed and visual memory are the purest measures of this dimension, but all variables contribute positively. The second canonical dimension accounts for only 1.5% of group mean differences and is non-significant (by a likelihood ratio test). Yet, if we were to interpret it, we would note that the schizophrenia group is slightly higher on this dimension, scoring better in problem solving and slightly worse on working memory, attention, and verbal learning tasks.

Summary

This analysis gives a very simple description of the data, in relation to the research questions posed earlier:

- On the basis of these neurocognitive tests, the schizophrenic and schizoaffective groups do not differ significantly overall, but these groups differ greatly from the normal controls.
- All cognitive domains distinguish the groups in the same direction, with the greatest differences shown for the variables most closely aligned with the horizontal axis in Figure 11.6.

11.3 Social cognitive measures

The social cognitive measures were designed to tap various aspects of the perception and cognitive processing of emotions of others. Emotion perception was assessed using a Managing Emotions score from the MCCB. A "theory of mind" (`ToM`) score assessed ability to read the emotions of others from photographs of the eye region of male and female faces. Two other measures, externalizing bias (`ExtBias`) and personalizing bias (`PersBias`) were calculated from a scale measuring the degree to which individuals attribute internal, personal or situational causal attributions to positive and negative social events.

The analysis of the `SocialCog` data proceeds in a similar way: first we fit the MANOVA model, then test the overall differences among groups using `Anova()`. We find that the overall multivariate test is again significant,

```
data(SocialCog, package="heplots")
SC.mlm <-  lm(cbind(MgeEmotions,ToM, ExtBias, PersBias) ~ Dx,
              data=SocialCog)
Anova(SC.mlm)
#>
#> Type II MANOVA Tests: Pillai test statistic
#>    Df test stat approx F num Df den Df  Pr(>F)
#> Dx  2     0.212     3.97      8    268 0.00018 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing the same two contrasts using `linearHypothesis()` (results not shown), w e find that the overall multivariate test is again significant, but now *both* contrasts are significant (Dx1: $F(4, 133) = 5.21, p < 0.001$; Dx2: $F(4, 133) = 2.49, p = 0.0461$), the test for `Dx2` just barely.

```
# control vs. patients
print(linearHypothesis(SC.mlm, "Dx1"), SSP=FALSE)
# Schizo vs. SchizAff
print(linearHypothesis(SC.mlm, "Dx2"), SSP=FALSE)
```

These results are important, because, if they are reliable and make sense substantively, they imply that patients with schizophrenia and schizoaffective diagnoses *can* be distinguished by their performance on tasks assessing social perception and cognition. This was potentially a new finding in the literature on schizophrenia.

As we did above, it is useful to visualize the nature of these differences among groups with HE plots for the `SC.mlm` model. Each contrast has a corresponding **H** ellipse, which we can show in the plot using the `hypotheses` argument. With a single degree of freedom, these degenerate ellipses plot as lines.

```
op <- par(mar=c(5,4,1,1)+.1)
heplot(SC.mlm,
       hypotheses=list("Dx1"="Dx1", "Dx2"="Dx2"),
       fill=TRUE, fill.alpha=.1,
       cex.lab=1.5, cex=1.2)
```

```
par(op)
```

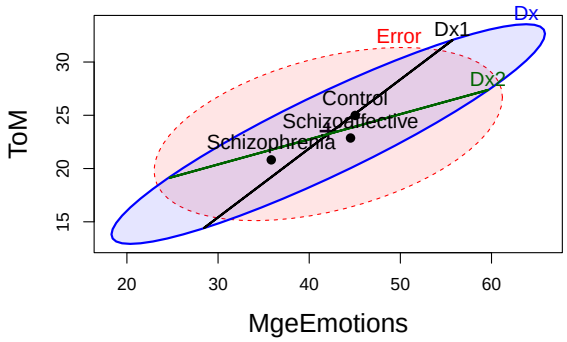

Figure 11.7: HE plot of Speed and Attention in the MLM for the `SocialCog` data. The labeled points show the means of the groups on the two variables. The lines for Dx1 and Dx2 show the tests of the contrasts among groups.

It can be seen that the three group means are approximately equally spaced on the `ToM` measure, whereas for `MgeEmotions`, the control and schizoaffective groups are quite similar, and both are higher than the schizophrenic group. This ordering of the three groups was somewhat similar for the other responses, as we could see in a `pairs(SC.mlm)` plot.

11.3.1 Model checking

Normally, we would continue this analysis, and consider other HE and canonical discriminant plots to further interpret the results, in particular the relations of the cognitive measures to group differences, or perhaps an analysis of the relationships between the neuro- and social-cognitive measures. We don't pursue this here for reasons of length, but this example actually has a more important lesson to demonstrate.

Before beginning the MANOVA analyses, extensive data screening was done by the client using SPSS, in which all the response *and* predictor variables were checked for univariate normality

and multivariate normality (MVN) for both sets. This traditional approach yielded a huge amount of tabular output and no graphs, and did not indicate any major violation of assumptions.[1]

A simple visual test of MVN and the possible presence of multivariate outliers is related to the theory of the data ellipse: Under MVN, the squared Mahalanobis distances $D^2_M(\mathbf{y}) = (\mathbf{y} - \bar{\mathbf{y}})' \mathbf{S}^{-1} (\mathbf{y} - \bar{\mathbf{y}})$ should follow a χ^2_p distribution. Thus, a quantile-quantile plot of the ordered D^2_M values vs. corresponding quantiles of the χ^2 distribution should approximate a straight line (Cox 1968; Healy 1968). Note that this should be applied to the *residuals* from the model – `residuals(SC.mlm)` – and not to the response variables directly.

`heplots::cqplot()` implements this for `"mlm"` objects Calling this function for the model `SC.mlm` produces (**fig:SC-cqplot?**). It is immediately apparent that there is one extreme multivariate outlier; three other points are identified, but the remaining observations are nearly within the 95% confidence envelope (using a robust MVE estimate of $\mathbf{S}$).

```
op <- par(mar=c(5,4,1,1)+.1)
cqplot(SC.mlm, method="mve",
       id.n=4,
       main="",
       cex.lab=1.25)
par(op)
```

[1] Actually, multivariate normality of the predictors in $\mathbf{X}$ is not required in the MLM. This assumption applies only to the conditional values $\mathbf{Y} \mid \mathbf{X}$, i.e., that the errors $\mathbf{u}'_i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ with constant covariance matrix. Moreover, the widely used MVN test statistics, such as Mardia's (1970) test based on multivariate skewness and kurtosis are known to be quite sensitive to mild departures in kurtosis (Mardia 1974) which do not threaten the validity of the multivariate tests.

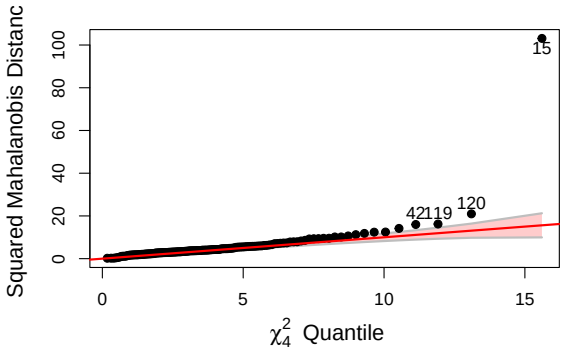

Figure 11.8: Chi-square quantile-quantile plot for residuals from the model `SC.mlm`. The confidence band gives a point-wise 95% envelope, providing information about uncertainty. One extreme multivariate outlier is highlighted.

Further checking revealed that this was a data entry error where one case (15) in the schizophrenia group had a score of -33 recorded on the `ExtBias` measure, whose valid range was (-10, +10). In R, it is very easy to re-fit a model to a subset of observations (rather than modifying the data set itself) using `update()`. The result of the overall Anova and the test of `Dx1` were unchanged; however, the multivariate test for the most interesting contrast `Dx2` comparing the schizophrenia and schizoaffective groups became non-significant at the $\alpha = 0.05$ level ($F(4, 133) = 2.18, p = 0.0742$).

```
SC.mlm1 <- update(SC.mlm,
                  subset=rownames(SocialCog)!="15")

Anova(SC.mlm1)
print(linearHypothesis(SC.mlm1, "Dx1"), SSP=FALSE)
print(linearHypothesis(SC.mlm1, "Dx2"), SSP=FALSE)
```

11.3.2 Canonical HE plot

This outcome creates a bit of a quandry for further analysis (do univariate follow-up tests? try a robust model?) and reporting (what to claim about the `Dx2` contrast?) that we don't explore

here. Rather, we proceed to attempt to interpret the MLM with the aid of canonical analysis and a canonical HE plot. The canonical analysis of the model `SC.mlm1` now shows that both canonical dimensions are significant, and account for 83.9% and 16.1% of between group mean differences respectively.

```
SC.can1 <- candisc(SC.mlm1)
SC.can1
#>
#> Canonical Discriminant Analysis for Dx:
#>
#>   CanRsq Eigenvalue Difference Percent Cumulative
#> 1 0.1645     0.1969      0.159    83.9       83.9
#> 2 0.0364     0.0378      0.159    16.1      100.0
#>
#> Test of H0: The canonical correlations in the
#> current row and all that follow are zero
#>
#>   LR test stat approx F numDF denDF Pr(> F)
#> 1        0.805     3.78     8   264 0.00032 ***
#> 2        0.964     1.68     3   133 0.17537
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
op <- par(mar=c(5,4,1,1)+.1)
heplot(SC.can1,
    fill=TRUE, fill.alpha=.1,
    hypotheses=list("Dx1"="Dx1", "Dx2"="Dx2"),
    lwd = c(1, 2, 3, 3),
    col=c("red", "blue", "darkgreen", "darkgreen"),
    var.lwd=2,
    var.col="black",
    label.pos=c(3,1),
    var.cex=1.2,
    cex=1.25, cex.lab=1.2,
    scale=2.8,
    prefix="Canonical dimension ")
par(op)
```

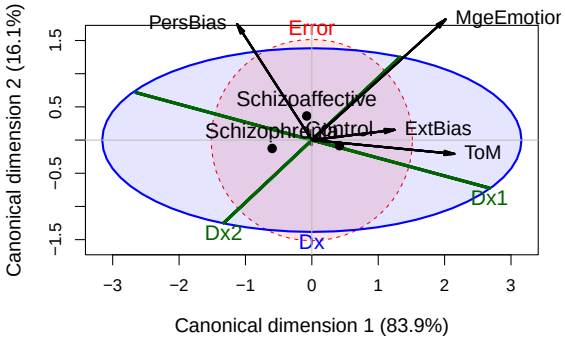

Figure 11.9: Canonical HE plot for the corrected `SocialCog` MANOVA. The variable vectors show the correlations of the responses with the canonical variables. The embedded green lines show the projections of the **H** ellipses for the contrasts `Dx1` and `Dx2` in canonical space.

The HE plot version of this canonical plot is shown in Figure 11.9. Because the `heplot()` method for a `"candisc"` object refits the original model to the **Z** canonical scores, it is easy to also project other linear hypotheses into this space. Note that in this view, both the `Dx1` and `Dx2` contrasts project outside **E** ellipse.[2].

This canonical HE plot has a very simple description:

- Dimension 1 orders the groups from control to schizoaffective to schizophrenia, while dimension 2 separates the schizoaffective group from the others;
- Externalizing bias and theory of mind contributes most to the first dimension, while personal bias and managing emotions are more aligned with the second; and,
- The relations of the two contrasts to group differences and to the response variables can be easily read from this plot.

```
#cat("Packages used here:\n")
write_pkgs(file = .pkg_file)
```

[2]The direct application of significance tests to canonical scores probably requires some adjustment because these are computed to have the optimal between-group discrimination.

```
#> 16  packages used here:
#>  base, broom, candisc, car, carData, corrgram, datasets, dplyr, ggplot2, graphics, grDevi
```

References

12 Visualizing Tests for Equality of Covariance Matrices

> *To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port.* — G. E. P. Box (1953)

This chapter concerns the extension of tests of homogeneity of variance from the classical univariate ANOVA setting to the analogous multivariate (MANOVA) setting. Such tests are a routine but important aspect of data analysis, as particular violations can drastically impact model estimates (Lix and Keselman 1996).

We provide some answers to the following questions:

- **Visualization**: How can we visualize differences among group variances and covariance matrices, perhaps in a way that is analogous to what is done to visualize differences among group means? As will be illustrated, differences among covariance matrices can be comprised of spread in overall size ("scatter") and shape ("orientation"). These can be seen in data space with data ellipses, particularly if the data is centered by shifting all groups to the grand mean,
- **Low-D views**: When there are more than a few response variables, what low-dimensional views can show the most interesting properties related to the equality of covariance matrices? Projecting the data into the space of the principal components serves well again here. Surprisingly,

we will see that the small dimensions contain useful information about differences among the group covariance matrices.

- **Other statistics**: Box's M-test is most widely used. Are there other worthwhile test statistics? We will see that graphics methods suggest alternatives.

The following subsections provide a capsule summary of the issues in this topic. Most of the discussion is couched in terms of a one-way design for simplicity, but the same ideas can apply to two-way (and higher) designs, where a "group" factor is defined as the product combination (interaction) of two or more factor variables. When there are also numeric covariates, this topic can also be extended to the multivariate analysis of covaraiance (MANCOVA) setting. This can be accomplished by applying these techniques to the residuals from predictions by the covariates alone.

Packages

In this chapter we use the following packages. Load them now

```
library(car)
library(heplots)
library(candisc)
library(ggplot2)
library(dplyr)
library(tidyr)
```

12.1 Homogeneity of Variance in Univariate ANOVA

In classical (Gaussian) univariate ANOVA models, the main interest is typically on tests of mean differences in a response y according to one or more factors. The validity of the typical F test, however, relies on the assumption of *homogeneity of variance*: all groups have the same (or similar) variance,

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_g^2 \; .$$

It turns out that the F test for differences in means is relatively robust to violation of this assumption (Harwell et al. 1992), as long as the group sizes are roughly equal.[1]

A variety of classical test statistics for homogeneity of variance are available, including Hartley's F_{max} (Hartley 1950), Cochran's C (Cochran 1941),and Bartlett's test (Bartlett 1937), but these have been found to have terrible statistical properties (Rogan and Keselman 1977), which prompted Box's famous quote.

Levene (1960) introduced a different form of test, based on the simple idea that when variances are equal across groups, the average absolute values of differences between the observations and group means will also be equal, i.e., substituting an L_1 norm for the L_2 norm of variance. In a one-way design, this is equivalent to a test of group differences in the means of the auxilliary variable $z_{ij} = |y_{ij} - \bar{y}_i|$.

More robust versions of this test were proposed by Brown and Forsythe (1974). These tests substitute the group mean by either the group median or a trimmed mean in the ANOVA of the absolute deviations, and should be almost always preferred to Levene's version. See Conover, Johnson, and Johnson (1981) for an early review and Gastwirth, Gel, and Miao (2009) for a general discussion of these tests. In what follows, we refer to this class of tests as "Levene-type" tests and suggest a multivariate extension described below (**?@sec-mlevene**).

12.2 Homogeneity of variance in MANOVA

In the MANOVA context, the main emphasis, of course, is on differences among mean vectors, testing

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_g \ .$$

[1] If group sizes are greatly unequal **and** homogeneity of variance is violated, then the F statistic is too liberal (p values too large) when large sample variances are associated with small group sizes. Conversely, the F statistic is too conservative if large variances are associated with large group sizes.

However, the standard test statistics (Wilks' Lambda, Hotelling-Lawley trace, Pillai-Bartlett trace, Roy's maximum root) rely upon the analogous assumption that the within-group covariance matrices for all groups are equal,

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_g \ .$$

Insert pairs covEllipses for penguins data

To preview the main example, Figure 12.1 shows data ellipses for the main size variables in the `palmerpenguins::penguins` data.

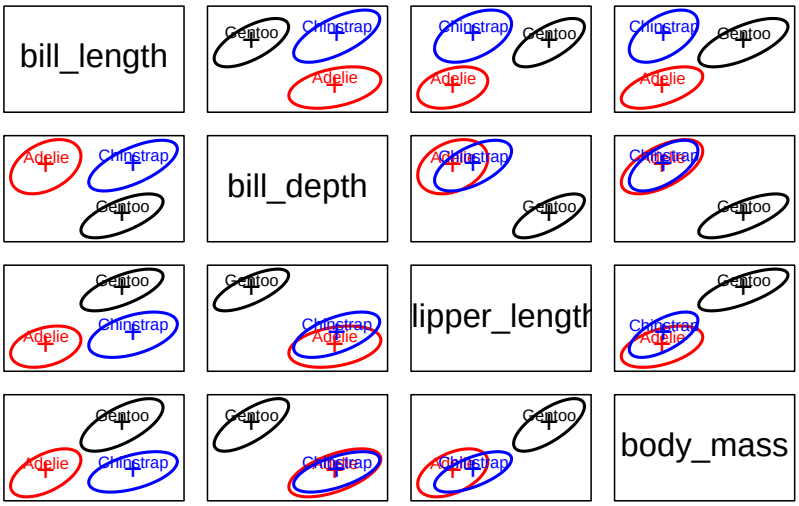

Figure 12.1: Data ellipses for the penguins data.

They covariance ellipses look pretty similar in size, shape and orientation. But what does Box's M test (described below) say? As you can see, it concludes strongly against the null hypothesis.

```
boxM(cbind(bill_length, bill_depth, flipper_length, body_mass) ~ species, data=peng)
#>
#>  Box's M-test for Homogeneity of Covariance Matrices
#>
#> data:  Y
#> Chi-Sq (approx.) = 75, df = 20, p-value = 3e-08
```

12.3 Assessing heterogeneity of covariance matrices: Box's M test

Box (1949) proposed the following likelihood-ratio test (LRT) statistic for testing the hypothesis of equal covariance matrices,

$$M = (N-g) \ln \mid \mathbf{S}_p \mid \; - \sum_{i=1}^{g} (n_i - 1) \ln \mid \mathbf{S}_i \mid \;\; ,$$

{eq-boxm}

where $N = \sum n_i$ is the total sample size and $\mathbf{S}_p = (N-g)^{-1} \sum_{i=1}^{g} (n_i - 1)\mathbf{S}_i$ is the pooled covariance matrix. M can thus be thought of as a ratio of the determinant of the pooled $\mathbf{S}_p$ to the geometric mean of the determinants of the separate $\mathbf{S}_i$.

In practice, there are various transformations of the value of M to yield a test statistic with an approximately known distribution (Timm 1975). Roughly speaking, when each $n_i > 20$, a χ^2 approximation is often used; otherwise an F approximation is known to be more accurate.

Asymptotically, $-2\ln(M)$ has a χ^2 distribution. The χ^2 approximation due to Box (1949, 1950) is that

$$X^2 = -2(1-c_1)\ln(M) \quad \sim \quad \chi^2_{df}$$

with $df = (g-1)p(p+1)/2$ degrees of freedom, and a bias correction constant:

$$c_1 = \left(\sum_i \frac{1}{n_i - 1} - \frac{1}{N-g} \right) \frac{2p^2 + 3p - 1}{6(p+1)(g-1)} \; .$$

In this form, Bartlett's test for equality of variances in the univariate case is the special case of Box's M when there is only one response variable, so Bartlett's test is sometimes used as univariate follow-up to determine which response variables show heterogeneity of variance.

Yet, like its univariate counterpart, Box's test is well-known to be highly sensitive to violation of (multivariate) normality and the presence of outliers. For example, Tiku and Balakrishnan

(1984) concluded from simulation studies that the normal-theory LRT provides poor control of Type I error under even modest departures from normality. O'Brien (1992) proposed some robust alternatives, and showed that Box's normal theory approximation suffered both in controlling the null size of the test and in power. Zhang and Boos (1992) also carried out simulation studies with similar conclusions and used bootstrap methods to obtain corrected critical values.

12.4 Visualizing heterogeneity

The goal of this chapter is to use the above background as a platform for discussing approaches to visualizing and testing the heterogeneity of covariance matrices in multivariate designs. While researchers often rely on a single number to determine if their data have met a particular threshold, such compression will often obscure interesting information, particularly when a test concludes that differences exist, and one is left to wonder "why?''. It is within this context where, again, visualizations often reign supreme. In fact, we find it somewhat surprising that this issue has not been addressed before graphically in any systematic way. **TODO: cut this down**

In what follows, we propose three visualization-based approaches to questions of heterogeneity of covariance in MANOVA designs:

(a) direct visualization of the information in the $\mathbf{S}_i$ and $\mathbf{S}_p$ using *data ellipsoids* to show size and shape as minimal schematic summaries;

(b) a simple dotplot of the components of Box's M test: the log determinants of the $\mathbf{S}_i$ together with that of the pooled $\mathbf{S}_p$. Extensions of these simple plots raise the question of whether measures of heterogeneity other than that captured in Box's test might also be useful; and,

(c) the connection between Levene-type tests and an ANOVA (of centered absolute differences) suggests a parallel with a multivariate extension of Levene-type tests and a MANOVA. We explore this with a version of

Hypothesis-Error (HE) plots we have found useful for visualizing mean differences in MANOVA designs.

```
#> Writing packages to  C:/R/Projects/Vis-MLM-quarto/bib/pkgs.txt
#> 15  packages used here:
#>  base, broom, candisc, car, carData, datasets, dplyr, ggplot2, graphics, grDevices, heplots
```

12.5 References

13 Summary

This will grow into a summary of the book. But for now, I leave it as the most beautiful of identities, Euler's $e^{\pi i} + 1 = 0$

```
exp(pi * 0i)
```

```
[1] 1+0i
```

References

Abbott, Edwin A. 1884. *Flatland: A Romance of Many Dimensions*. Cutchogue, NY: Buccaneer Books.

Adler, Daniel, and Duncan Murdoch. 2023. *Rgl: 3D Visualization Using OpenGL*. https://CRAN.R-project.org/package=rgl.

Anscombe, F. J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27: 17–21.

Bartlett, M. S. 1937. "Properties of Sufficiency and Statistical Tests." *Proceedings of the Royal Society of London. Series A* 160 (901): 268–82. https://doi.org/10.2307/96803.

Belsley, David A. 1991. *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. New York, NY: Wiley.

Belsley, David A., E. Kuh, and Roy E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley; Sons.

Biecek, Przemyslaw, Hubert Baniecki, Mateusz Krzyzinski, and Dianne Cook. 2023. "Performance Is Not Enough: A Story of the Rashomon's Quartet," February. https://arxiv.org/abs/2302.13356.

Box, G. E. P. 1949. "A General Distribution Theory for a Class of Likelihood Criteria." *Biometrika* 36 (3-4): 317–46. https://doi.org/10.1093/biomet/36.3-4.317.

———. 1950. "Problems in the Analysis of Growth and Wear Curves." *Biometrics* 6: 362–89.

———. 1953. "Non-Normality and Tests on Variances." *Biometrika* 40 (3/4): 318–35. https://doi.org/10.2307/2333350.

Brown, Morton B., and Alan B. Forsythe. 1974. "Robust Tests for Equality of Variances." *Journal of the American Statistical Association* 69 (346): 364–67. https://doi.org/10.1080/01621459.1974.10482955.

Cajori, Florian. 1926. "Origins of Fourth Dimension Concepts." *The American Mathematical Monthly* 33 (8): 397–406. https:

//doi.org/10.1080/00029890.1926.11986607.

Cochran, W. G. 1941. "The Distribution of the Largest of a Set of Estimated Variances as a Fraction of Their Total." *Annals of Eugenics* 11 (1): 47–52. https://doi.org/10.1111/j.1469-1809.1941.tb02271.x.

Conover, W. J., Mark E. Johnson, and Myrle M. Johnson. 1981. "A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data." *Technometrics* 23 (4): 351–61. https://doi.org/10.1080/00401706.1981.10487680.

Cook, R. D., and S. Weisberg. 1982. *Residuals and Influence in Regression*. New York: Chapman; Hall.

Cotton, R. 2013. *Learning R*. Sebastopol, CA: O'Reilly Media.

Cox, D. R. 1968. "Notes on Some Aspects of Regression Analysis." *Journal of the Royal Statistical Society Series A* 131: 265–79.

Curran, James, and Taylor Hersh. 2021. *Hotelling: Hotelling's t^2 Test and Variants*. https://CRAN.R-project.org/package=Hotelling.

Davies, Rhian, Steph Locke, and Lucy D'Agostino McGowan. 2022. *datasauRus: Datasets from the Datasaurus Dozen*. https://CRAN.R-project.org/package=datasauRus.

Davis, C. 1990. "Body Image and Weight Preoccupation: A Comparison Between Exercising and Non-Exercising Women." *Appetite* 16 (1): 84. https://doi.org/10.1016/0195-6663(91)90115-9.

Dempster, A. P. 1969. *Elements of Continuous Multivariate Analysis*. Reading, MA: Addison-Wesley.

Duncan, O. D. 1961. "A Socioeconomic Index for All Occupations." In *Occupations and Social Status*, edited by Jr. A. J. Reiss, P. K. Hatt O. D. Duncan, and C. C. North. New York: The Free Press.

Farquhar, A. B., and H. Farquhar. 1891. *Economic and Industrial Delusions: A Discourse of the Case for Protection*. New York: Putnam.

Fisher, R. A. 1936. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7 (2): 179–88. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Fox, John. 2016. *Applied Regression Analysis and Generalized Linear Models*. Third edition. Los Angeles: SAGE.

———. 2020. *Regression Diagnostics*. 2nd ed. SAGE Publications, Inc. https://doi.org/10.4135/9781071878651.

Fox, John, and Georges Monette. 1992. "Generalized Collinearity Diagnostics." *Journal of the American Statistical Association* 87 (417): 178–83.

Fox, John, and Sandford Weisberg. 2018. *An r Companion to Applied Regression*. Third. Thousand Oaks CA: SAGE Publications. https://books.google.ca/books?id=uPNrDwAAQBAJ.

Fox, John, Sanford Weisberg, and Brad Price. 2023. *Car: Companion to Applied Regression*. https://CRAN.R-project.org/package=car.

Friendly, Michael. 2002. "Corrgrams: Exploratory Displays for Correlation Matrices." *The American Statistician* 56 (4): 316–24. http://datavis.ca/papers/corrgram.pdf.

———. 2007. "HE Plots for Multivariate General Linear Models." *Journal of Computational and Graphical Statistics* 16 (2): 421–44. https://doi.org/10.1198/106186007X208407.

———. 2008. "The Golden Age of Statistical Graphics." *Statistical Science* 23 (4): 502–35. https://doi.org/10.1214/08-STS268.

Friendly, Michael, and E. Kwan. 2003. "Effect Ordering for Data Displays." *Computational Statistics and Data Analysis* 43 (4): 509–39. http://authors.elsevier.com/sd/article/S0167947302002906.

Friendly, Michael, and Ernest Kwan. 2009. "Where's Waldo: Visualizing Collinearity Diagnostics." *The American Statistician* 63 (1): 56–65. https://doi.org/10.1198/tast.2009.0012.

Friendly, Michael, and David Meyer. 2016. *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data*. Boca Raton, FL: Chapman & Hall/CRC.

Friendly, Michael, Georges Monette, and John Fox. 2013. "Elliptical Insights: Understanding Statistical Methods Through Elliptical Geometry." *Statistical Science* 28 (1): 1–39. https://doi.org/10.1214/12-STS402.

Friendly, Michael, and Howard Wainer. 2021. *A History of Data Visualization and Graphic Communication*. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/9780674259034.

Funkhouser, H. Gray. 1937. "Historical Development of the Graphical Representation of Statistical Data." *Osiris* 3 (1): 269–405. http://tinyurl.com/32ema9.

Gabriel, K. R. 1971. "The Biplot Graphic Display of Matrices with Application to Principal Components Analysis." *Biometrics* 58 (3): 453–67.

Galton, Francis. 1886. "Regression Towards Mediocrity in Hereditary Stature." *Journal of the Anthropological Institute* 15: 246–63. http://www.jstor.org/cgi-bin/jstor/viewitem/09595295/dm995266/99p0374f/0.

Gastwirth, Joseph L., Yulia R. Gel, and Weiwen Miao. 2009. "The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice." *Statistical Science* 24 (3): 343–60. https://doi.org/10.1214/09-STS301.

Gelman, Andrew, Jessica Hullman, and Lauren Kennedy. 2023. "Causal Quartets: Different Ways to Attain the Same Average Treatment Effect." http://www.stat.columbia.edu/~gelman/research/unpublished/causal_quartets.pdf.

Gower, J. C., and D. J. Hand. 1996. *Biplots*. London: Chapman & Hall.

Guerry, André-Michel. 1833. *Essai Sur La Statistique Morale de La France*. Paris: Crochard.

Hartley, H. O. 1950. "The Use of Range in Analysis of Variance." *Biometrika* 37 (3–4): 271–80. https://doi.org/10.1093/biomet/37.3-4.271.

Hartman, L. I. 2016. "Schizophrenia and Schizoaffective Disorder: One Condition or Two?" PhD dissertation, York University.

Harwell, M. R., E. N. Rubinstein, W. S. Hayes, and C. C. Olds. 1992. "Summarizing Monte Carlo Results in Methodological Research: The One- and Two-Factor Fixed Effects ANOVA Cases." *Journal of Educational and Behavioral Statistics* 17 (4): 315–39. https://doi.org/10.3102/10769986017004315.

Healy, M. J. R. 1968. "Multivariate Normal Plotting." *Journal of the Royal Statistical Society Series C* 17 (2): 157–61.

Heinrichs, R. Walter, Farena Pinnock, Eva Muharib, Leah Hartman, Joel Goldberg, and Stephanie McDermid Vaz. 2015. "Neurocognitive Normality in Schizophrenia Revisited." *Schizophrenia Research: Cognition* 2 (4): 227–32. https://doi.org/10.1016/j.scog.2015.09.001.

Herschel, John F. W. 1833. "On the Investigation of the Orbits

of Revolving Double Stars: Being a Supplement to a Paper Entitled "Micrometrical Measures of 364 Double Stars"." *Memoirs of the Royal Astronomical Society* 5: 171–222.

Hoaglin, David C., and Roy E. Welsch. 1978. "The Hat Matrix in Regression and ANOVA." *The American Statistician* 32 (1): 17–22. https://doi.org/10.1080/00031305.1978.10479237.

Hotelling, Harold. 1931. "The Generalization of Student's Ratio." *The Annals of Mathematical Statistics* 2 (3): 360–78. https://doi.org/10.1214/aoms/1177732979.

Kwan, Ernest, Irene R. R. Lu, and Michael Friendly. 2009. "Tableplot: A New Tool for Assessing Precise Predictions." *Zeitschrift für Psychologie / Journal of Psychology* 217 (1): 38–48. https://doi.org/10.1027/0044-3409.217.1.38.

Levene, Howard. 1960. "Robust Tests for Equality of Variances." In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, edited by Ingram Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, and H. B. Mann, 278–92. Stanford, Calif: Stanford University Press.

Lix, J. M., L. M. Keselman, and H. J. Keselman. 1996. "Consequences of Assumption Violations Revisited: A Quantitative Review of Alternatives to the One-Way Analysis of Variance F Test." *Review of Educational Research* 66 (4): 579–619. https://doi.org/10.3102/00346543066004579.

Longley, James W. 1967. "An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User." *Journal of the American Statistical Association* 62: 819–41. https://doi.org/https://www.tandfonline.com/doi/abs/10.1080/01621459.1967.10500896.

Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. 2021. "performance: An R Package for Assessment, Comparison and Testing of Statistical Models." *Journal of Open Source Software* 6 (60): 3139. https://doi.org/10.21105/joss.03139.

Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, Brenton M. Wiernik, and Dominique Makowski. 2022. *Easystats: Framework for Easy Statistical Modeling, Visualization, and Reporting. CRAN*. https://easystats.github.io/easystats/.

Mardia, K. V. 1970. "Measures of Multivariate Skewness and Kurtosis with Applications." *Biometrika* 57 (3): 519–30. https://doi.org/http://dx.doi.org/10.2307/2334770.

———. 1974. "Applications of Some Measures of Multivariate Skewness and Kurtosis in Testing Normality and Robustness Studies." *Sankhya: The Indian Journal of Statistics, Series B* 36 (2): 115–28. http://www.jstor.org/stable/25051892.

Marquardt, Donald W. 1970. "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation." *Technometrics* 12: 591–612.

Marquardt, Donald W., and Ronald D. Snee. 1975. "Ridge Regression in Practice." *The American Statistician* 29 (1): 3–20. https://doi.org/10.1080/00031305.1975.10479105.

Matejka, Justin, and George Fitzmaurice. 2017. "Same Stats, Different Graphs." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3025453.3025912.

Matloff, Norman. 2011. *The Art of R Programming: A Tour of Statistical Software Design*. San Francisco, CA: No Starch Press.

Monette, Georges. 1990. "Geometry of Multiple Regression and Interactive 3-D Graphics." In *Modern Methods of Data Analysis*, edited by J. Fox and S. Long, 209–56. Beverly Hills, CA: SAGE Publications.

O'Brien, Peter C. 1992. "Robust Procedures for Testing Equality of Covariance Matrices." *Biometrics* 48 (3): 819–27. http://www.jstor.org/stable/2532347.

Pearson, Karl. 1896. "Contributions to the Mathematical Theory of Evolution—III, Regression, Heredity and Panmixia." *Philosophical Transactions of the Royal Society of London*, A, 187: 253–318.

Playfair, William. 1786. *Commercial and Political Atlas: Representing, by Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England, During the Whole of the Eighteenth Century*. London: Debrett; Robinson;; Sewell. http://ucpj.uchicago.edu/Isis/journal/demo/v000n000/000000/000000.fg4.html.

———. 1801. *Statistical Breviary; Shewing, on a Principle Entirely New, the Resources of Every State and Kingdom in Europe*. London: Wallis.

Rogan, J. C., and H. J. Keselman. 1977. "Is the ANOVA f-Test Robust to Variance Heterogeneity When Sample Sizes Are Equal?: An Investigation via a Coefficient of Variation." *American Educational Research Journal* 14 (4): 493–98.

https://doi.org/10.3102/00028312014004493.
Teetor, Paul. 2011. *R cookbook*. Sebastopol, CA: O'Reilly Media.
Tiku, M. L., and N. Balakrishnan. 1984. "Testing Equality of Population Variances the Robust Way." *Communications in Statistics - Theory and Methods* 13 (17): 2143–59. https://doi.org/10.1080/03610928408828818.
Timm, N. H. 1975. *Multivariate Analysis with Applications in Education and Psychology*. Belmont, CA: Wadsworth (Brooks/Cole).
Wickham, Hadley. 2014. *Advanced R*. Boca Raton, FL: Chapman and Hall/CRC.
Xie, Yihui. 2021. *Animation: A Gallery of Animations in Statistics and Utilities to Create Animations*. https://yihui.org/animation/.
Zhang, Ji, and Dennis D. Boos. 1992. "Bootstrap Critical Values for Testing Homogeneity of Covariance Matrices." *Journal of the American Statistical Association* 87 (418): 425–29. http://www.jstor.org/stable/2290273.

Package used