

# Statistical Modeling

A FRESH APPROACH

DANIEL T. KAPLAN  
MACALESTER COLLEGE

For review purposes only

Copyright (c) 2009 by Daniel Kaplan.

All rights reserved. No part of the material protected by this copyright notice may be reproduced or utilized in any form, electronic or mechanical, including photocopying, recording, or scanning, or stored on any information storage or retrieval system, without written permission from the author.

**Cover Photo:** The trunk of a scribbly gum eucalyptus tree on Fraser Island in Queensland, Australia. The scribbly gum moth lays its eggs between the old and new bark layers. The larvae burrow between the bark layers, leaving a winding tunnel that is revealed when the old bark falls away. [Photo credit: the author.]

To my wife, Maya.

For review purposes only

For review purposes only

# Contents

<b>1</b>	<b>Statistical Models</b>	<b>3</b>
1.1	Models and their Purposes . . . . .	6
1.2	Observation and Knowledge . . . . .	7
1.3	The Main Points of this Book . . . . .	10
1.4	Introduction to Computation with R . . . . .	11
<b>2</b>	<b>Data: Cases, Variables, Samples</b>	<b>31</b>
2.1	Kinds of Variables . . . . .	32
2.2	Data Frames and the Unit of Analysis . . . . .	34
2.3	Populations and Samples . . . . .	35
2.4	Longitudinal and Cross-Sectional Samples . . . . .	37
2.5	Computational Technique . . . . .	39
<b>3</b>	<b>Describing Variation</b>	<b>43</b>
3.1	Coverage Intervals . . . . .	45
3.2	The Variance and Standard Deviation . . . . .	47
3.3	Displaying Variation . . . . .	51
3.4	Normal and Non-normal Distributions . . . . .	55
3.5	Categorical Variables . . . . .	56
3.6	Computational Technique . . . . .	60
<b>4</b>	<b>The Language of Models</b>	<b>67</b>
4.1	Models as Functions . . . . .	68
4.2	Multiple Explanatory Variables . . . . .	70
4.3	Reading a Model . . . . .	73
4.4	Choices in Model Design . . . . .	75
4.5	Model Terms . . . . .	76
4.6	Standard Notation for Describing Model Design . . . . .	83
4.7	Computational Technique . . . . .	84

<b>5</b>	<b>Model Formulas and Coefficients</b>	<b>93</b>
5.1	The Linear Model Formula . . . . .	94
5.2	Linear Models with Multiple Terms . . . . .	95
5.3	Formulas with Categorical Variables . . . . .	95
5.4	Model Coefficients Describe Relationships . . . . .	97
5.5	Model Values and Residuals . . . . .	98
5.6	Coefficients of Basic Model Designs . . . . .	98
5.7	Coefficients have Units . . . . .	102
5.8	Untangling Explanatory Variables . . . . .	103
5.9	Why Linear Models? . . . . .	106
5.10	Computational Technique . . . . .	110
<b>6</b>	<b>Fitting Models to Data</b>	<b>113</b>
6.1	The Least Squares Criterion . . . . .	113
6.2	Partitioning Variation . . . . .	117
6.3	Redundancy . . . . .	119
6.4	Computational Technique . . . . .	122
<b>7</b>	<b>Measuring Correlation</b>	<b>127</b>
7.1	Properties of $R^2$ . . . . .	127
7.2	Simple Correlation . . . . .	129
7.3	Nested Models . . . . .	132
7.4	Toward Statistical Models . . . . .	134
7.5	Computational Technique . . . . .	137
<b>8</b>	<b>Total and Partial Relationships</b>	<b>141</b>
8.1	Total and Partial Relationships . . . . .	142
8.2	Explicitly Holding Covariates Constant . . . . .	153
8.3	Adjustment and Truth . . . . .	154
8.4	Computational Technique . . . . .	156
<b>9</b>	<b>Model Vectors</b>	<b>163</b>
9.1	Vectors . . . . .	164
9.2	Vectors as Collections of Numbers . . . . .	166
9.3	Model Vectors . . . . .	167
9.4	Model Vectors in the Linear Formula . . . . .	169
9.5	Model Vectors and Redundancy . . . . .	170
9.6	Geometry by Arithmetic . . . . .	171
9.7	Computational Technique . . . . .	173
<b>10</b>	<b>Statistical Geometry</b>	<b>177</b>
10.1	Case Space and Variable Space . . . . .	178
10.2	Subspaces and Geometrical Operations . . . . .	179
10.3	The Model Triangle . . . . .	185
10.4	Simple Statistics: Geometrically . . . . .	186
10.5	Drawing Vector Diagrams . . . . .	189
10.6	Computational Technique . . . . .	191

<b>11 Geometry with Multiple Vectors</b>	<b>195</b>
11.1 Finding Coefficients . . . . .	198
11.2 The Coefficient of Determination . . . . .	199
11.3 Collinearity . . . . .	200
11.4 Redundancy . . . . .	203
11.5 Multi-Collinearity . . . . .	203
11.6 Higher Dimensions . . . . .	205
11.7 Computational Technique . . . . .	206
<b>12 Modeling Randomness</b>	<b>207</b>
12.1 Describing Pure Randomness . . . . .	207
12.2 Settings for Probability Models . . . . .	210
12.3 Models of Counts . . . . .	210
12.4 Common Probability Calculations . . . . .	213
12.5 Continuous Probability Models . . . . .	214
12.6 The Normal Distribution . . . . .	217
12.7 Computational Technique . . . . .	219
12.8 Simulations . . . . .	224
<b>13 Geometry of Random Vectors</b>	<b>227</b>
13.1 Random Angles . . . . .	227
13.2 Random Models . . . . .	230
13.3 Random Walks . . . . .	231
<b>14 Confidence in Models</b>	<b>235</b>
14.1 The Sampling Distribution . . . . .	237
14.2 Standard Errors and the Regression Report . . . . .	240
14.3 Confidence Intervals . . . . .	242
14.4 Interpreting the Confidence Interval . . . . .	244
14.5 Confidence in Predictions . . . . .	245
14.6 Finding the Resampling Distribution . . . . .	248
14.7 Confidence and Collinearity . . . . .	258
14.8 Confidence and Bias . . . . .	260
14.9 Computational Technique . . . . .	261
<b>15 The Logic of Hypothesis Testing</b>	<b>269</b>
15.1 An Example of a Hypothesis Test . . . . .	271
15.2 Inductive and Deductive Reasoning . . . . .	271
15.3 The Null Hypothesis . . . . .	275
15.4 The p-value . . . . .	276
15.5 Rejecting by Mistake . . . . .	277
15.6 Failing to Reject . . . . .	279
15.7 A Glossary of Hypothesis Testing . . . . .	281
15.8 Computational Technique . . . . .	282

<b>16 Hypothesis Testing on Whole Models</b>	<b>287</b>
16.1 The Permutation Test . . . . .	288
16.2 $R^2$ and the F Statistic . . . . .	290
16.3 The ANOVA Report . . . . .	296
16.4 Visualizing p-values . . . . .	298
16.5 Tests of Simple Models . . . . .	300
16.6 Interpreting the p-value . . . . .	303
16.7 Computational Technique . . . . .	307
<b>17 Hypothesis Testing on Parts of Models</b>	<b>317</b>
17.1 The Term-by-Term ANOVA Table . . . . .	318
17.2 Covariates Soak Up Variance . . . . .	319
17.3 Measuring the Sum of Squares . . . . .	322
17.4 ANOVA, Collinearity, and Multi-Collinearity . . . . .	325
17.5 Hypothesis Tests on Single Coefficients . . . . .	334
17.6 Non-Parametric Statistics . . . . .	337
17.7 Sample Size and Power . . . . .	338
17.8 Computational Technique . . . . .	341
<b>18 Models of Yes/No Variables</b>	<b>345</b>
18.1 The 0-1 Encoding . . . . .	345
18.2 Inference on Logistic Models . . . . .	351
18.3 Model Probabilities . . . . .	353
18.4 Computational Technique . . . . .	356
<b>19 Causation</b>	<b>361</b>
19.1 Interpreting Models Causally . . . . .	362
19.2 Causation and Correlation . . . . .	364
19.3 Hypothetical Causal Networks . . . . .	367
19.4 Networks and Covariates . . . . .	369
<b>20 Experiment</b>	<b>383</b>
20.1 Experiments . . . . .	383
20.2 When Experiments are Impossible . . . . .	394
20.3 Conclusion . . . . .	403
<b>Further Readings &amp; Bibliography</b>	<b>405</b>
<b>Datasets</b>	<b>412</b>
Dataset Index . . . . .	419
R Operator Index . . . . .	421
Concept Index . . . . .	426



# Preface

The purpose of this book is to provide an introduction to statistics that gives readers a sufficient mastery of statistical concepts, methods, and computations to apply them to authentic systems. By “authentic,” I mean the sort of multivariable systems often encountered when working in the natural or social sciences, commerce, government, law, or any of the many contexts in which data are collected with an eye to understanding how things work or to making predictions about what will happen.

The world is complex and uncertain. We deal with the complexity and uncertainty with a variety of strategies including the scientific method and the discipline of statistics.

Statistics helps to deal with uncertainty, quantifying it so that you can assess how reliable — how likely to be repeatable — your findings are. The scientific method helps to deal with complexity: reduce systems to simpler components, define and measure quantities carefully, do experiments in which some conditions are held constant but others are varied systematically.

Beyond helping to quantify uncertainty and reliability, statistics provides another great insight of which most people are unaware. When dealing with systems involving multiple influences, it is possible and best to deal with those influences simultaneously. By appropriate data collection and analysis, the confusing tangle of influences can sometimes be straightened out (and it is possible to know when the attempt gives ambiguous and unreliable results). In other words, statistics goes hand-in-hand with the scientific method when it comes to dealing with complexity and understanding how systems work.

The statistical methods that can accomplish this are often considered advanced: multiple regression, analysis of covariance, logistic regression, etc. With appropriate software, any method is accessible in the sense of being able to produce a summary report on the computer. But a method is useful only when the user has a way to understand whether the method is appropriate for the situation, what the method is telling about the data, and what the method is not capable of revealing. Computer scientist Richard Hamming (1915-1998) said: “The purpose of computing is insight, not numbers.” Without a solid understanding of the theory that underlies a method, the numbers generated by the computer may not give insight.

The methods of statistics can give tremendous insight, particularly the so-called advanced methods. For this reason, these methods need to be accessible

both computationally and theoretically to the widest possible audience. Historically, the audience has been limited because few people have the algebraic skills needed to approach the methods in the way they are usually presented. But there are many paths to understanding and I have undertaken to find one that takes the greatest advantage of the actual skills that most people already have in abundance.

In trying to meet that challenge, I have made many unconventional choices. Theory becomes simpler when there is a unified framework for treating many aspects of statistics, so I have chosen to present just about everything in the context of models: descriptive statistics as well as inference as well as experimental design.

Another choice is the use of simple geometry to present theory. The underlying geometrical concepts are elementary: lengths, angles, projection, the Pythagorean theorem. Much of the statistical theory can be displayed in a two-dimensional sketch; once in a while three-dimensional visualization is needed. When I have presented the geometrical ideas in this manner to professionals in research groups or at statistics conferences and workshops, I find uniformly that people gain significant insight into their work. A common reaction is, "It can't be that easy." Yes, it can be and it should be.

Also unconventional, but hardly innovative, is the use of resampling and randomization to motivate the concepts of statistical inference. George Cobb [1] cogently describes the logic of statistical inference as the three-Rs: "Randomize, Repeat, Reject." In a decade of teaching statistics, I have found that students can understand this algorithmic logic much better than the derivations of algebraic formulas for means and standard deviations of sampling distributions.

As you might expect from the preceding comments, algebraic notation and formulas are strongly de-emphasized in this book. I find that most people are not skilled in interpreting them and extracting meaning from them. In any event, formulas are no longer needed as a way to describe calculations since statistical work is now done on the computer.

And then there is software. Some people think that statistics should be taught without computers in order to help develop conceptual understanding. Others think that it is silly to ignore a technology that is universally used in practice and greatly expands our capabilities. Both points of view are right.

The main body of this book is presented in a way that makes little or no reference to software; the statistical concepts are paramount. But each chapter has a section on computational technique that shows how to get things done and aims to give the reader concrete skills in the analysis of data. An extensive set of exercises and classroom activities is published in workbook form, via the Internet. The computer is an effective teaching tool, so many of the exercises and activities make heavy use of it.

The software used is R, a modern, powerful and freely available system for statistical computations and graphics. The book assumes that you know nothing at all about scientific software and, accordingly, introduces R from basics. If you have experience with statistics, you probably already have a preferred software package. So long as that software will fit linear models with multiple

explanatory variables and produce a more-or-less standard regression report, it can be used to follow this book. That said, I strongly encourage you to think about learning and using R. You can learn it easily by following the examples and can be doing productive statistics very quickly. Not only will you easily be able to fit models and get reports, but you can use R to explore ideas such as re-sampling and randomization. If you now use “educational” software, learning R will give you a professional-level tool for use in the future.

For many instructors, this book can support a nice *second* course in statistics — a follow-up to a conventional first introductory course. Increasingly, such a course is needed as more and more young people encounter basic statistical ideas in grade school and many of the topics of the conventional university course are absorbed into the high-school curriculum. At Macalester College, where I developed this book, mainstream students of biology, economics, political science, and so on use this book for their *first* statistics course. Accordingly, the book is written to be self-contained, making no assumption that readers have had any previous formal study in statistics.

Thanks and acknowledgments ...

I have been fortunate to have the assistance and support of many people. Some of the colleagues who have played important roles are David Bressoud, George Cobb, Dan Flath, Tom Halverson, Gary Krueger, Weiwen Miao, Phil Poronnik, Victor Addona, Karen Saxe, Michael Schneider, and Libby Shoop. Critical institutional support was given by Brian Rosenberg, Jan Serie, Dan Hornbach, Helen Warren, and Diane Michelfelder at Macalester and Mercedes Talley at the Keck Foundation.

I received encouragement from many in the statistics education community, including George Cobb, Joan Garfield, Dick De Veaux, Bob delMas, Julie Legler, Milo Schield, Paul Alper, Andy Zieffler, Sharon Lane-Getaz, Katie Makar, Michael Bulmer, Nick Horton, Frank Shaw, and the participants in the monthly “Stat Chat” sessions. Helpful suggestions came from from Simon Blomberg, Dominic Hyde, Michael Lavine, Erik Larson, Julie Dolan, and Kendrick Brown. Michael Edwards helped with proofreading. Nick Trefethen and Dave Saville provided important insights about the geometry of fitting linear models.

Thanks also go to the hundred or so students at Macalester College who enrolled in the early, experimental sessions of Math 155 where many of the ideas in this book were first tried out. Among those students, I want to acknowledge particular help from Alan Eisinger, Caroline Ettinger, Bernd Verst, Wes Hart, Sami Saqer, and Michael Snively.

Crucial early support for this project was provided by a grant from the Howard Hughes Medical Institute. An important Keck Foundation grant was crucial to the continuing refinement of the approach and the writing of this book.

Finally, my thanks and love to my wife, Maya, and daughters, Tamar, Liat, and Netta, who endured the many, many hours during which I was preoccupied by some or another statistics-related enthusiasm, challenge, or difficulty.

For review purposes only

For review purposes only

# Chapter 1

## Statistical Models

*All models are wrong. Some models are useful.* — George Box

*Art is a lie that tells the truth.* — Pablo Picasso

This book is about **statistical modeling**, two words that themselves require some definition.

“Modeling” (note the “ing” ending) is a process of asking questions. “Statistical” refers in part to data — the statistical models you will construct will be rooted in data. But it refers also to a distinctively modern idea: that you can measure what you *don’t* know and that doing so contributes to your understanding.

There is a saying, “A person with a watch knows the time. A person with two watches is never sure.” The statistical point of view is that it’s better not to be sure. With two watches you can see how they disagree with each other. This provides an idea of how precise the watches are. You don’t know the time exactly, but knowing the precision tells you something about what you don’t know. The non-statistical certainty of the person with a single watch is merely an uninformed self-confidence: the single watch provides no indication of what the person doesn’t know.

The physicist Ernest Rutherford (1871-1937) famously said, “If your experiment needs statistics, you ought to have done a better experiment.” In other words, if you can make a good enough watch, you need only one: no statistics. This is bad advice. Statistics never hurts. A person with two watches that agree perfectly not only knows the time, but has evidence that the watches are working at high precision. Sensibly, the official world time is based on an average of many atomic clocks. The individual clocks are fantastically precise; the point of averaging is to know when one or more of the clocks is drifting out of precision for some reason.

Why “statistical modeling” and not simply “statistics” or “data analysis?” Many people imagine that data speak for themselves and that the purpose of statistics is to extract the information that the data carry. Such people see data

For review purposes only

analysis as an objective process in which the researcher should, ideally, have no influence. This can be true when very simple issues are involved, for instance how precise is the average of the atomic clocks used to set official time or what is the difference in time between two events. But many questions are much more complicated; they involve many variables and you don't necessarily know what is doing what to what.[2]

The conclusions you reach from data depend on the specific questions you ask. Like it or not, the researcher plays an active and creative role in constructing and interrogating data. This means that the process involves some subjectivity. But this is not the same as saying anything goes. Statistical methods allow you to make objective statements about how the data answer your questions. In particular, the methods help you to know if the data show anything at all.

The word “modeling” highlights that your goals, your beliefs, and your current state of knowledge all influence your analysis of data. The core of the scientific method is the formation of hypotheses that can be tested and perhaps refuted by experiment or observation. Similarly, in statistical modeling you examine your data to see whether they are consistent with the hypotheses that frame your understanding of the system under study.

**Example 1.1: Grades** A woman is applying to law school. The schools she applies to ask for her class rank, which is based on the average of her college course grades.

A simple statistical issue concerns the precision of the grade-point average. This isn't a question of whether the average was correctly computed or whether the grades were accurately recorded. Instead, imagine that you could send two essentially identical students to essentially identical schools. Their grade-point averages might well differ, reflecting perhaps the grading practices of their different instructors or slightly different choices of subjects or random events such as illness or mishaps or the scheduling of classes. One way to think about this is that the students' grades are to some extent random, contingent on factors that are unknown or perhaps irrelevant to the students' capabilities.

How do you measure the extent to which the grades are random? There is no practical way to create “identical” students and observe how their grades differ. But you can look at the variation in a single student's grades — from class to class — and use this as an indication of the size of the random influence in each grade. From this, you can calculate the likely range of the random influences on the overall grade-point average.

Statistical models let you go further in interpreting grades. It's a common belief that there are easy- and hard-grading teachers and that a grade reflects not just the student's work but the teacher's attitude and practices. Statistical modeling provides a way to use data on grades to see whether teachers grade differently and to correct for these differences between teachers. Doing this involves some subtlety, for example taking into account the possibility that strong students take different courses than weaker students.

**Example 1.2: Nitrogen Fixing Plants** Biologist Michael Anderson studies how plants fix nitrogen in the soil. All plants need nitrogen to grow. Since nitrogen is the primary component of air, there is plenty around. But it's hard for plants to get nitrogen from the air; they get it instead from the soil. Some plants, like alder and soybean, support nitrogen-fixing bacteria in nodules on the plant roots. The plant creates a hospitable environment for the bacteria; the bacteria, by fixing nitrogen in the soil, create a good environment for the plant. In a word, symbiosis.

Anderson is interested in how genetic variation in the bacteria influences the success with which they fix nitrogen. One can imagine using this information to breed plants and bacteria that are more effective at fixing nitrogen and thereby reducing the need for agricultural fertilizer.

Anderson has an promising early result. His extensive field studies indicate that different genotypes of bacteria fix nitrogen at different rates. Unfortunately, the situation is confusing since the different genotypes tend to populate different areas with different amounts of soil moisture, different soil temperatures, and so on. How can he untangle the relative influences of the genotype and the other environmental factors in order to decide whether the variation in genotype is genuinely important and worth further study?

---

**Example 1.3: Sex Discrimination** A large trucking firm is being audited by the government to see if the firm pays wages in a racially or sexually discriminatory way. The audit finds wage discrepancies between men and women for "office and clerical workers" but not for other job classifications such as technicians, supervisors, sales personnel, or "skilled craftworkers." It finds no discrepancies based on race.

A simple statistical question is whether the observed difference in average wages for men and women office and clerical workers is based on enough data to be reliable. In answering this question, it actually makes a difference what other groups the government auditors looked at when deciding to focus on sex discrimination in office and clerical workers.

Further complicating matters are the other factors that contribute to people's wages: the kind of job they have, their skill level, their experience. Statistical models can be used to quantify how these various factors contribute and to see whether they account for some or all of the wage discrepancy associated with sex. For instance, it turns out that men on average tend to have more job experience than women, and some or all of the men's higher average wages might be due to this.

Models can help you decide whether this potential explanation is plausible. For instance, if you see that both men's and women's wages increase with experience in the same way, you might be more inclined to believe that job experience is a legitimate factor rather than just a mask for discrimination.

---

## 1.1 Models and their Purposes

Many of the toys you played with as a child are models: dolls, balsa-wood airplanes with wind-up propellers, wooden blocks, model trains. But so are many serious objects of the adult world: architectural plans, bank statements, train schedules, the results of medical diagnostic tests, the signals transmitted by a telephone, the equations of physics, the genetic sequences used by biologists. There are too many to list.

What all models have in common is this:

*A model is a representation for a particular purpose.*

A model might be a physical object or it might be an idea, but it always stands for something else: it's a representation. Dolls stand for babies and other creatures, architectural plans stand for buildings and bridges, a white blood-cell count stands for the function of the immune system.

When you create a model, you have (or ought to have) a purpose in mind. Toys are created for the entertainment and (sometimes) edification of children. The various kinds of toys — dolls, blocks, model airplanes and trains — have a form that serves this purpose. Unlike the things they represent, the toy versions are small, safe, and inexpensive.

Models always leave things out and get some things — many things — wrong. Architectural plans are not houses; you can't live in them. But they are easy to transport, copy, and modify. That's the point. Telephone signals — unlike the physical sound waves that they represent — can be transported over long distances and even stored. A train schedule tells you something important but it obviously doesn't reproduce every aspect of the trains it describes; it doesn't carry passengers.

Statistical models revolve around data. But even so, they are first and foremost models. They are created for a purpose. The intended use of a model should shape the appropriate form of the model and determines the sorts of data that can properly be used to build the model.

There are three main uses for statistical models. They are closely related, but distinct enough to be worth enumerating.

**Description.** Sometimes you want to describe the range or typical values of a quantity. For example, what's a "normal" white blood cell count? Sometimes you want to describe the relationship between things. Example: What's the relationship between the price of gasoline and consumption by automobiles?

**Classification or prediction.** You often have information about some observable traits, qualities, or attributes of a system you observe and want to draw conclusions about other things that you can't directly observe. For instance, you know a patient's white blood-cell count and other laboratory measurements and want to diagnose the patient's illness.

**Anticipating the consequences of interventions.** Here, you intend to do something: you are not merely an observer but an active participant in the sys-

For review purposes only



tem. For example, people involved in setting or debating public policy have to deal with questions like these: To what extent will increasing the tax on gasoline reduce consumption? To what extent will paying teachers more increase student performance?

The appropriate form of a model depends on the purpose. For example, a model that diagnoses a patient as ill based on an observation of a high number of white blood cells can be sensible and useful. But that same model would give absurd predictions about intervention: Do you really think that lowering the white blood cell count by bleeding a patient will make the patient better?

To anticipate correctly the effects of an intervention you need to get the direction of cause and effect correct in your models. But for a model used for classification or prediction, it may be unnecessary to represent causation correctly. Instead, other issues, e.g. the reliability of data, can be the most important. One of the thorniest issues in statistical modeling — with tremendous consequences for science, medicine, government, and commerce — is how you can legitimately draw conclusions about interventions from models based on data collected without performing these interventions.

## 1.2 Observation and Knowledge

How do you know what you know? How did you find it out? How can you find out what you don't yet know? These are questions that philosophers have addressed for thousands of years. The views that they have expressed are complicated and contradictory.

From the earliest times in philosophy, there has been a difficult relationship between knowledge and observation. Sometimes philosophers see your knowledge as emerging from your observations of the world, sometimes they emphasize that the way you see the world is rooted in your innate knowledge: the things that are obvious to you.

This tension plays out on the pages of newspapers as they report the controversies of the day. Does the death penalty deter crime? Does increased screening for cancer reduce mortality? Will paying teachers more improve student outcomes?

Consider the simple, obvious argument for why severe punishment deters crime. Punishments are things that people don't like. People avoid what they don't like. If crime leads to punishment, then people will avoid committing crime.

Each statement in this argument seems perfectly reasonable, but none of them is particularly rooted in observations of actual and potential criminals. It's artificial — a learned skill — to base knowledge such as "people avoid punishment" on observation. It might be that this knowledge was formed by our own experiences, but usually the only explanation you can give is something like, "that's been my experience" or give one or two anecdotes.

When observations contradict opinions — opinions are what you think you know — people often stick with their opinions. Put yourself in the place of

someone who believes that the death penalty really does deter crime. You are presented with accurate data showing that when a neighboring state eliminated the death penalty, crime did not increase. So do you change your views on the matter? Possibly, but possibly not. A skeptic can argue that it's not just punishment but also other factors that influence the crime rate, for instance the availability of jobs. Perhaps it was that a generally improving economic condition in the other state kept the crime rate steady even at a time when society is imposing lighter punishments.

It's difficult to use observation to inform knowledge because relationships are complicated and involve multiple factors. It isn't at all obvious how people can discover or demonstrate causal relationships through observation. Suppose one school district pays teachers well and another pays them poorly. You observe that the first district has better student outcomes than the second. Can you legitimately conclude that teacher pay accounts for the difference? Perhaps something else is at work: greater overall family wealth in the first district (which is what enabled them to pay teachers more), better facilities, smaller classes, and so on.

Historian Robert Hughes concisely summarized the difficulty of trying to use observation to discover causal relationships. In describing the extensive use of hanging in 18th and 19th century England, he wrote, "One cannot say whether public hanging did terrify people away from crime. Nor can anyone do so, until we can count crimes that were never committed." [3, p.35] To know whether hanging did deter crime, you would need to observe a **counterfactual**, something that didn't actually happen: the crimes in a world without hanging. You can't observe counterfactuals. So you need somehow to generate observations that give you data on what happens for different levels of the causal variable.

A modern idea is the **controlled experiment**. In its simplest ideal form, a controlled experiment involves changing one thing — teacher pay, for example — while *holding everything else constant*: family wealth, facilities, etc.

The experimental approach to gaining knowledge has had great success for example in medicine and science. For many people, experiment is the essence of science. But experiments are hard to perform and sometimes not possible at all. How do you "hold everything else constant?" Partly for this reason, you rarely see reports of experiments when you read the newspaper, unless the article happens to be about a scientific discovery.

Scientists pride themselves on recording their observations carefully and systematically in lab notebooks. Laboratories are filled with high-precision instrumentation. The quest for precision culminates perhaps in the physicist's fundamental quantities: the speed of light is reported to be  $299,792,500 \pm 1000$  meters per second, the mass of the electron reported as  $9.10938215 \pm 0.00000045 \times 10^{-31}$  kg. Each of these is precise to about 50 parts in a billion. Contrast this extreme precision with the humble speed measurements from a policeman's radar gun (perhaps a couple of miles or kilometers per hour — one part in 50) or the weight indicated on a bathroom scale (give or take a kilogram or a couple of pounds — about one part in 100 for an adult).

All such observations and measures are the stuff of **data**, the records of ob-

servations. Observations do not become data by virtue of high precision or expensive instrumentation or the use of metric rather than traditional units. For many purposes, data of low precision is used. An ecologist's count of the number of mating pairs of birds in a territory is limited by the ability to find nests. A national census of a country's population, conducted by the government can be precise to only a couple of percent. The physicist counting neutrinos in huge observatories buried under mountains to shield them from extraneous events waits for months for her results. These results are precise to only one part in two.

The precision that is needed in data depends on the purpose for which the data will be used. The important question for the person using the data is whether the precision, whatever it be, is adequate for the purpose at hand. To answer this question, you need to know how to measure precision and how to compare this to a standard reflecting the needs of your task. The scientist with expensive instrumentation and the framer of social policy both need to deal with data in similar ways to understand and interpret the precision of their results.

It's common for people to believe that conclusions drawn from data apply to certain areas — science, economics, medicine — but aren't terribly useful in other areas. In teaching, for example, almost all decisions are based on "experience" rather than observation. Indeed, there is often strong resistance to making formal observations of student progress as interfering with the teaching process.

This book is based on the idea that techniques for drawing valid conclusions from observations — data — are valuable for two groups of people. The first group is scientists and others who routinely need to use statistical methods to analyze experimental and other data.

The second group is everybody else. All of us need to draw conclusions from our experiences, even if we're not in a laboratory. It's better to learn how to do this in valid ways, and to understand the limitations of these ways, than to rely on an informal, unstated process of opinion formation. It may turn out that in any particular area of interest there are no useful data. In such situations, you won't be able to use the techniques. But at least you will know what you're missing. You may be inspired to figure out how to supply it or to recognize it when it does come along, and you'll be aware of when others are misusing data.

As you will see, the manner in which the data are collected plays a central role in what sorts of conclusions can be legitimately made; data do not always speak for themselves. You will also see that strongly supported statements about causation are difficult to make. Often, all you can do is point to an "association" or a "correlation," a weaker form of statement.

Statistics is sometimes loosely described as the "science of data." This description is apt, particularly when it covers both the collection and analysis of data, but it does not mean much until you understand what data are. That's the subject of the next chapter.

### 1.3 The Main Points of this Book

Statistics is about variation. Describing and interpreting variation is a major goal of statistics.

You can create empirical, mathematical descriptions not only of a single trait or variable but also of the relationships between two or more traits. Empirical means based on measurements, data, observations.

Models let you split variation into components: “explained” versus “unexplained.” How to measure the size of these components and how to compare them to one another is a central aspect of statistical methodology. Indeed, this provides a definition of statistics:

*Statistics is the explanation of variation in the context of what remains unexplained.*

By collecting data in ways that require care but are quite feasible, you can estimate how reliable your descriptions are, e.g., whether it’s plausible that you should see similar relationships if you collected new data. This notion of reliability is very narrow and there are some issues that depend critically on the context in which the data were collected and the correctness of assumptions that you make about how the world works.

Relationships between pairs of traits can be studied in isolation only in special circumstances. In general, to get valid results it is necessary to study entire systems of traits simultaneously. Failure to do so can easily lead to conclusions that are grossly misleading.

Descriptions of relationships are often **subjective** — they depend on choices that you, the modeler, make. These choices are generally rooted in your own beliefs about how the world works, or the theories accepted as plausible within some community of inquiry.

If data are collected properly, you can get an indication of whether the data are consistent or inconsistent with your subjective beliefs or — and this is important — whether you don’t have enough data to tell either way.

Models can also be used to check out the sensitivity of your conclusions to different beliefs. People who disagree in their views of how the world works often may not be able to reconcile their differences based on data, but they will be able to decide objectively whether their own or the other party’s beliefs are reasonable given the data.

Notwithstanding everything said above about the strong link between your prior, subjective beliefs and the conclusions you draw from data, by collecting data in a certain context — experiments — you can dramatically simplify the interpretation of the results. It’s actually possible to remove the dependence on identified subjective beliefs by intervening in the system under study experimentally.

This book takes a different approach than most statistics texts. Many people want statistics to be presented as a kind of automatic, algorithmic way to process data. People look for mathematical certainty in their conclusions. After all, there are right-or-wrong answers to the mathematical calculations that peo-

ple (or computers) perform in statistics. Why shouldn't there be right-or-wrong answers to the conclusions that people draw about the world?

The answer is that there can be, but only when you are dealing with narrow circumstances that may not apply to the situations you want to study. An insistence on certainty and provable correctness often results in irrelevancy.

The point of view taken in this book is that it is better to be useful than to be provably certain. The objective is to introduce methods and ideas that can help you deal with drawing conclusions about the real world from data. The methods and ideas are meant to guide your reasoning; even if the conclusions you draw are not guaranteed by proof to be correct, they can still be more useful than the alternative, which is the conclusions that you draw without data, or the conclusions you draw from simplistic methods that don't honor the complexity of the real system.

## 1.4 Introduction to Computation with R

Modern statistics is done on the computer. There was a time, 60 years ago and before, when computation could only be done by hand or using balky mechanical calculators. The methods of applied statistics developed during this time reflected what could be done using such calculators, not necessarily what was best for illuminating the system under study. These methods took on a life of their own — they became the operational definition of statistics. They continue to be taught today, using electronic calculators or personal computers or even just using paper and pencil. For the old statistical methods, computers are merely a labor saving device.

But not for modern statistics. The statistical methods at the core of this book cannot be applied in a authentic and realistic way without powerful computers. Thirty years ago, many of the methods could not be done at all unless you had access to the resources of a government agency or a large university. But with the revolutionary advances in computer hardware and numerical algorithms over the last half-century, modern statistical calculations can be performed on an ordinary home computer or laptop. (Even a cell phone has phenomenal computational power, often besting the mainframes of thirty years ago.) Hardware and software today pose no limitation; they are readily available.

Each chapter of this book includes a section on computational technique. Many readers will be completely new to the use of computers for scientific and statistical work, so the first chapters cover the foundations, techniques that are useful for many different aspects of computation. Working through the early chapters is essential for developing the skills that will be used later in actual statistical work. It will take a few hours, but this investment will pay off handsomely.

Chances are, you use a computer almost every day: for email, word-processing, managing your music or your photograph collection, perhaps even using a spreadsheet program for accounting. The software you use for such activities makes

it easy to get started. Possibly you have never even looked at an instruction manual or used the “help” features on your computer.

When you use a word processor or email, the bulk of what you enter into the computer — the content of your documents and email — is without meaning to the computer. This is not at all to say that it is meaningless. Your documents and letters are intended for human readers; most of the work you do is directed so that the recipients can understand them. But the computer doesn’t need to understand what you write in order to format it, print it, or transmit it over the Internet. Indeed, the computer would be equally effective at handling random text generated by typing monkeys.

When doing scientific and statistical computing, things are different. What you enter into the computer is instructions to the computer to perform calculations and re-arrangements of data. Those instructions have to be comprehensible to the computer. If they make no sense or if they are inconsistent or ill formed, the computer won’t be able to carry out your instructions. Worse, if the instructions make sense in some formal way but don’t convey your actual intentions, the computer will perform some operation but the result will mislead you.

The difficulty with using software for mathematics and statistics is in making sure that your instructions make sense and do what you want them to do. This difficulty is not a matter of bad software design; it’s intrinsic to the problem of communicating your intentions to the computer. The same difficulty would arise in word processing if the computer had to make sense of your writing, rejecting it when a claim is unconvincing or when a sentence is ambiguous. Statistical computing pioneer John Chambers refers to the “Prime Directive” of software[4]: “to program in such a way that computations can be understood and trusted.”

Much of the design of packages for scientific and statistical work is oriented around the difficulty of communicating intentions. A popular approach is based on the computer mouse: the program provides a list of possible operations — like the keys on a calculator — and lets the user choose which operation to apply to some selected data. This style of user interface is employed, for example, in spreadsheet software, letting users add up columns of numbers, make graphs, etc. The reason this style is popular is that it can make things extremely easy ... so long as the operation that you want has been included in the software. But things get very hard if you need to construct your own operation and it can be difficult to understand or trust the operations performed by others.

Another style of scientific computation — the one used in this book — is based on language. Rather than selecting an option with a mouse, you construct a **command** that conveys both the operation that you want and the data to which you want to apply that operation. There are dramatic advantages to this language-based style of computation:

- It lets you **connect** computations to one another, so that the output of one operation can become the input to another.
- It lets you **repeat** the operation on new or modified data, allowing you to

automate tedious tasks and, importantly, to verify the correctness of your computations on data where you already know the answer.

- It lets you **accumulate** the results of previous operations, treating those results as new data.
- It lets you **document** concisely what was done so that you can demonstrate that what you said you did is what you actually did. In this way, you or others can repeat the analysis later if necessary to confirm your results.
- It lets you **modify** the computation in a controlled way to correct it or to vary some aspect of it while holding other aspects exactly the same as in the original.

In order to use the language-based approach, you will need to learn a few principles of the language itself: some vocabulary, some syntax, some grammar. This is much, much easier for the computer language than for a natural language like English or Chinese; it will take you only a couple of hours before you are fluent enough to do useful computations. In addition to letting you perform scientific computations in ways that use the computer and your own time and effort effectively, the principles that you will learn are broadly applicable to many computer systems and can provide significant insight even to how to use mouse-based interfaces.

### 1.4.1 The R Environment

The software package used in this book is called R. The R package provides an environment for doing statistical and scientific computation at a professional level. It was designed for statistics work, but suitable for other forms of scientific calculations and the creation of high-quality scientific graphics.[5]

There are several other major software packages widely used in statistics. Among the leaders are SPSS, SAS, and STATA. Each of them provides the computational power needed for statistical modeling. Each has its own advantages and its own devoted group of users.

One reason for the choice of R is that it offers a command-based computing environment. That makes it much easier to write about computing and also reveals better the structure of the computing process.[6] Also nice is that R is available for free and works on the major types of computers, e.g., Windows, Macintosh, and Unix/Linux. You can get information about how to install R on your computer at [www.r-project.org](http://www.r-project.org).

In making your own choice, the most important thing is this: *choose something!* Readers who are familiar with SPSS, SAS, or STATA can use the information in each chapter's computational technique section to help them identify the facilities to look for in those packages.

Another package that's often used in statistical work is the spreadsheet program Excel. This package has its own advantages. It's effective for entering data and has nice facilities for formatting tables. The visual layout of the data seems to be intuitive to many people. Many businesses use Excel and it's widely taught

in high schools. Unfortunately, it's very difficult to use for statistical analyses of any sophistication. Indeed, even some very elementary tasks such as making a histogram are difficult to do in Excel and the results are usually unsatisfactory from a graphical point of view. Worse, Excel is very hard to use reliably. There are lots of opportunities to make mistakes that will go undetected; Excel encourages bad programming practices that make software unreliable.

### 1.4.2 Setting R Up

You can skip this section if you want to jump ahead to read about how R works.

Better, though, if you try out the commands as they are shown, using the R software on your own computer.

To do this, you will need to start the R software. If R is not already installed on your computer (this is likely if you are using your own computer and have never used R before), the first step is to install it.

If you are used to installing software, you will find R follows the usual pattern.

1. Use a web browser to go to [www.r-project.org](http://www.r-project.org). Select the Download/CRAN link on the left of the page. This will bring you to a list of download sites.
2. Choose one in your own region of the world. This will bring up a page with a choice of Linux, Mac OS X, or Windows. Choose whichever is appropriate for your computer.
3. For Windows: choose the link for the "base" distribution, and then download the link that looks like "R-2.9.0-win32.exe." The name may be slightly different, depending on what new versions have been released. Run this program, accepting the defaults.

For Macintosh, follow the link that looks like "R-2.9.0.dmg" and follow the instructions. Again, the name may be slightly different, depending on what new versions have been released.

If you are using Linux, you probably don't need any instructions on how to install software.

Once R is installed, you start it like most programs, by clicking on an icon. On a Windows computer, this will be under the "start" button, on a Macintosh this will be in the applications folder.

When you start R, a new window appears on your screen. It will look something like Figure 1.1. You're ready to go!

Actually, there is one more thing that you can do that will make things easier later on, when you start to analyze the data sets that go along with this book. Download one file from a web site and put it in some convenient place on your computer, for instance your "desktop" or a folder that you make for this book. The file is at

[www.maclester.edu/~kaplan/ISM/ISM.Rdata](http://www.maclester.edu/~kaplan/ISM/ISM.Rdata)

This file contains various data sets that you will be using.



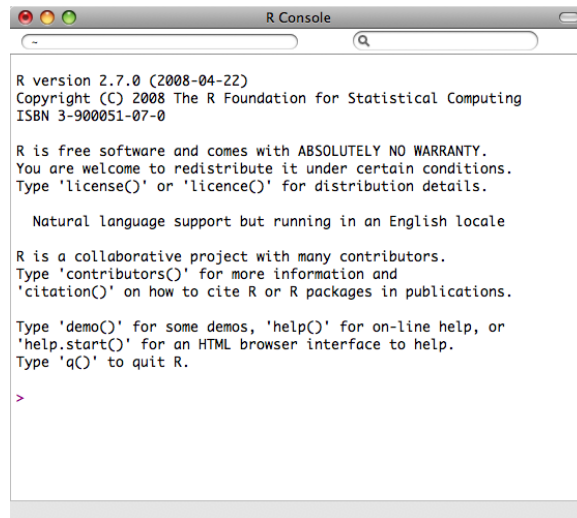


Figure 1.1: The R command window.

Once you have downloaded the file, called `ISM.Rdata`, double-clicking on it in the ordinary way will start R or load the data into an already started session of R.

### 1.4.3 Invoking an Operation

People often think of computers as *doing* things: sending email, playing music, storing files. Your job in using a computer is to tell the computer *what* to do. There are many different words used to refer to the “what”: a procedure, a task, a function, a routine, and so on. I’ll use the word **computation**. Admittedly, this is a bit circular, but it is easy to remember: computers perform computations.

Complex computations are built up from simpler computations. This may seem obvious, but it is a powerful idea. An **algorithm** is just a description of a computation in terms of other computations that you already know how to perform. To help distinguish between the computation as a whole and the simpler parts, it is helpful to introduce a new word: an **operator** performs a computation.

It’s helpful to think of the computation carried out by an operator as involving four parts:

1. The name of the operator
2. The input arguments
3. The output value
4. Side effects

A typical operation takes one or more **input arguments** and uses the information in these to produce an **output value**. Along the way, the computer might

take some action: display a graph, store a file, make a sound, etc. These actions are called **side effects**.


To tell the computer to perform a computation — call this **invoking an operation** or giving a **command** — you need to provide the name and the input arguments in a specific format. The computer then returns the output value. For example, the command `sqrt(25)` invokes the square root operator (named `sqrt`) on the argument 25. The output from the computation will, of course, be 5.

The syntax for invoking an operation consists of the operator's name, followed by round parentheses. The input arguments go inside the parentheses.

The software program that you use to invoke operators is called an **interpreter**. (The interpreter is the program you are running when you start R.) You enter your commands as a dialog between you and the interpreter. To start, the interpreter prints a prompt, after which you type your command:

PROMPT →  `sqrt(25)` ← COMMAND

When you press “Enter,” the interpreter reads your command and performs the computation. For commands such as this one, the interpreter will print the output value from the computation:

 `sqrt(25)`

OUTPUT MARKER → `[1]`    5 ← OUTPUT VALUE

NEXT PROMPT → 

The dialog continues as the interpreter prints another prompt and waits for your further command.

To save space, I'll usually show just the give-and-take from one round of the dialog:

```
> sqrt(25)
[1] 5
```

(Go ahead! Type `sqrt(25)` after the prompt in the R interpreter, press “enter,” and see what happens.)

Often, operations involve more than one argument. The various arguments are separated by commas. For example, here is an operation named `seq` that produces a sequence of numbers:

```
> seq(3, 10)
[1] 3 4 5 6 7 8 9 10
```

The first argument tells where to start the sequence, the second tells where to end it.

The order of the arguments is important. Here is the sequence produced when 10 is the first argument and 3 the second:

```
> seq(10, 3)
[1] 10 9 8 7 6 5 4 3
```

For some operators, particularly those that have many input arguments, some of the arguments can be referred to by name rather than position. This is particularly useful when the named argument has a sensible default value. For example, the `seq` operator can be instructed how big a jump to take between successive items in the sequence. This is accomplished using an argument named `by`:

```
> seq(3,10,by=2)
[1] 3 5 7 9
```

Depending on the circumstances, all four parts of a operation need not be present. For example, the `date` operation returns the current time and date; no input arguments are needed.

```
> date()
[1] "Wed Apr 16 06:18:06 2008"
```

Note that even though there are no arguments, the parentheses are still used. Think of the pair of parentheses as meaning, “Do this.”

### Naming and Storing Values

Often the value returned by an operation will be used later on. Values can be stored for later use with the **assignment operator**. This has a different syntax that reminds the user that a value is being stored. Here’s an example of a simple assignment:

```
> x = 16
```

This command has stored the value 16 under the name `x`. The syntax is always the same: an equal sign (=) with a name on the left and a value on the right.

Such stored values are called **objects**. Making an assignment to an object defines the object. Once an object has been defined, it can be referred to and used in later computations.

Notice that an assignment operation does not return a value or display a value. Its sole purpose is to have the side effects of defining the object and thereby storing a value under the object’s name.

To refer to the value stored in the object, just use the object’s name itself. For instance:

```
> x
[1] 16
```

Doing a computation on the value stored in an object is much the same:

```
> sqrt(x)
[1] 4
```

You can create as many objects as you like and give them names that remind you of their purpose. Some examples: `wilma`, `ages`, `temp`, `dog.houses`, `foo3`. There are some rules for object names:

- Use only letters and numbers and the two punctuation marks “dot” (.) and “underscore” (\_).
- Do NOT use spaces anywhere in the name.
- A number or underscore cannot be the first character in the name.
- Capital letters are treated as distinct from lower-case letters. The objects named `wilma` and `Wilma` are different.

For the sake of readability, keep object names short. But if you really must have an object named something like `agesOfChildrenFromTheClinicalTrial`, feel free.

Objects can store all sorts of things, for example a sequence of numbers:

```
> x = seq(1,7)
```

When you assign a new value to an existing object, as just done to `x`, the former value of that object is erased from the computer memory. The former value of `x` was 16, but after the above assignment command it is

```
> x
[1] 1 2 3 4 5 6 7
```

The value of an object is changed only *via* the assignment operator. Using an object in a computation does not change the value. For example, suppose you invoke the square-root operator on `x`:

```
> sqrt(x)
[1] 1.00 1.41 1.73 2.00 2.24 2.45 2.65
```

The square roots have been returned as a value, but this doesn't change the value of `x`:

```
> x
[1] 1 2 3 4 5 6 7
```

If you want to change the value of `x`, you need to use the assignment operator:

```
> x = sqrt(x)
> x
[1] 1.00 1.41 1.73 2.00 2.24 2.45 2.65
```

### Connecting Computations

The brilliant thing about organizing operators in terms of input arguments and output values is that the output of one operator can be used as an input to another. This lets complicated computations be built out of simpler ones.

For example, suppose you have a list of 10000 voters in a precinct and you want to select a random sample of 20 of them for a survey. The `seq` operator can

**Aside. 1.1** Assignment vs Algebra

An assignment command like `x=sqrt(x)` can be confusing to people who are used to algebraic notation. In algebra, the equal sign describes a relationship between the left and right sides. So,  $x = \sqrt{x}$  tells us about how the quantity  $x$  and the quantity  $\sqrt{x}$  are related. Students are usually trained to “solve” such relationships, going through a series of algebraic steps to find values for  $x$  that are consistent with the mathematical statement. (For  $x = \sqrt{x}$ , the solutions are  $x = 0$  and  $x = 1$ .) In contrast, the assignment command `x=sqrt(x)` is a way of replacing the previous values stored in `x` with new values that are the square root of the old ones.

be used to generate a set of 10000 choices. The `sample` operator can be used to select some of these choices at random.

One way to connect the computations is by using objects to store the intermediate outputs.

```
> choices = seq(1,10000)
> sample( choices, 20 )
[1] 5970 8476 9340 8266 6909
[6] 3692 8979 1640 4266 5580
[11] 1208 6141 4973 5575 8498
[16] 1001  923 3246 4194 2126
```

You can also pass the output of an operator *directly* as an argument to another operator. Here’s another way to accomplish exactly the same thing as the above.

```
> sample( seq(1,10000), 20 )
```

**Numbers and Arithmetic**

The language has a concise notation for arithmetic that looks very much like the traditional one:

```
> 7+2
[1] 9
> 3*4
[1] 12
> 5/2
[1] 2.5
> 3-8
[1] -5
> -3
[1] -3
> 5^2
[1] 25
```

Arithmetic operators, like any other operators, can be connected to form more complicated computations. For instance,

```
> 8+4/2
[1] 10
```

To a human reader, the command  $8+4/2$  might seem ambiguous. Is it intended to be  $(8+4)/2$  or  $8+(4/2)$ ? The computer uses unambiguous rules to interpret the expression, but it's a good idea for you to use parenthesis so that you can make sure that what you intend is what the computer carries out:

```
> (8+4)/2
[1] 6
```

Traditional mathematical notation uses superscripts and radicals to indicate exponentials and roots, e.g.,  $3^2$  or  $\sqrt{3}$  or  $\sqrt[3]{8}$ . This special typography doesn't work well with an ordinary keyboard, so R and most other computer languages uses a different notation:

```
> 3^2
[1] 9
> sqrt(3)
[1] 1.73
> 8^(1/3)
[1] 2
```

There is a large set of mathematical functions: exponentials, logs, trigonometric and inverse trigonometric functions, etc. Some examples:

Traditional	Computer
$e^2$	<code>exp(2)</code>
$\log_e(100)$	<code>log(100)</code>
$\log_{10}(100)$	<code>log10(100)</code>
$\log_2(100)$	<code>log2(100)</code>
$\cos(\frac{\pi}{2})$	<code>cos(pi/2)</code>
$\sin(\frac{\pi}{2})$	<code>sin(pi/2)</code>
$\tan(\frac{\pi}{2})$	<code>tan(pi/2)</code>
$\cos^{-1}(-1)$	<code>acos(-1)</code>

Numbers can be written in **scientific notation**. For example, the “universal gravitational constant” that describes the gravitational attraction between masses is  $6.67428 \times 10^{-11}$  (with units meters-cubed per kilogram per second squared). In the computer notation, this would be written `G=6.67428e-11`. The Avogadro constant, which gives the number of atoms in a mole, is  $6.02214179 \times 10^{23}$  per mole, or `6.02214178e23`.

The computer language does not directly support the recording of units. This is unfortunate, since in the real world numbers often have units and the units matter. For example, in 1999 the Mars Climate Orbiter crashed into Mars because the design engineers specified the engine's thrust in units of pounds, while the guidance engineers thought the units were newtons.

Computer arithmetic is accurate and reliable, but it often involves very slight rounding of numbers. Ordinarily, this is not noticeable. However, it can become

apparent in some calculations that produce results that are zero. For example, mathematically  $\sin(\pi) = 0$ , however the computer does not duplicate this mathematical relationship exactly:

```
> sin(pi)
[1] 1.22e-16
```

Whether a number like this is properly interpreted as “close to zero,” depends on the context and, for quantities that have units, on the units themselves. For instance, the unit “parsec” is used in astronomy in reporting distances between stars. The closest star to the sun is Proxima, at a distance of 1.3 parsecs. A distance of  $1.22 \times 10^{-16}$  parsecs is tiny in astronomy but translates to about 2.5 meters — not so small on the human scale.

In statistics, many calculations relate to probabilities which are always in the range 0 to 1. On this scale, 1.22e-16 is very close to zero.

There are two “special” numbers. Inf stands for  $\infty$ , as in

```
> 1/0
[1] Inf
```

NaN stands for “not a number,” and is the result when a numerical operation isn’t defined, for instance

```
> 0/0
[1] NaN
> sqrt(-9)
[1] NaN
```

---

#### Aside. 1.2 Complex Numbers

---

Mathematically oriented readers will wonder why R should have any trouble with a computation like  $\sqrt{-9}$ ; the result is the imaginary number  $3i$ . R works with complex numbers, but you have to tell the system that this is what you want to do. To calculate  $\sqrt{-9}$ , use `sqrt(-9+0i)`.

---

### Types of Objects

Most of the examples used so far have dealt with numbers. But computers work with other kinds of information as well: text, photographs, sounds, sets of data, and so on. The word **type** is used to refer to the kind of information.

Modern computer languages support a great variety of types. There are four types that will be most important here:

**numeric** The numbers of the sort already encountered.

**character** Text data.

**logical** Answers to yes/no questions.

**data frames** Collections of data more or less in the form of a spreadsheet table.

It's important to know about the types of data because operators expect their input arguments to be of specific types. When you use the wrong type of input, the computer might not be able to process your command.

### Character Data

You indicate character data to the computer by enclosing the text in double quotation marks. For example:

```
> filename = "swimmers.csv"
```

There is something a bit subtle going on in the above command, so look at it carefully. The purpose of the command is to create an object, named `filename`, that stores a little bit of text data. Notice that the name of the object is not put in quotes, but the text characters are.

Whenever you refer to an object name, make sure that you don't use quotes, for example:

```
> filename  
[1] "swimmers.csv"
```

If you make a command with the object name in quotes, it won't be treated as referring to an object. Instead, it will merely mean the text itself:

```
> "filename"  
[1] "filename"
```

Similarly, if you omit the quotation marks from around text, the computer will treat it as if it were an object name and will look for the object of that name. For instance, the following command directs the computer to look up the value contained in an object named `swimmers.csv` and insert that value into the object `filename`.

```
> filename = swimmers.csv  
Error: object "swimmers.csv" not found
```

As it happens, there was no object named `swimmers.csv` because it had not been defined by any previous assignment command. So, the computer generated an error.

For the most part, you will not need to use very many operators on text data; you just need to remember to include text, such as file names, in quotation marks. Sometimes, you will want to convert non-text items to text in order to display them in graphs. There is a special operator, `as.character` for doing this:

```
> as.character(3)  
[1] "3"
```



The quotes in the output show that it is character type rather than numeric type. This isn't terribly important to the human reader, but the computer regards "3" as a different thing than 3. For instance, you can do arithmetic on numbers but not on characters.

```
> 3 + 2
[1] 5
> as.character(3) + 2
Error in as.character(3) + 2 :
  non-numeric argument to binary operator
```

### Data Frames

A data frame is a collection of values arranged as a table. For example, here is part of a data frame that records an experiment on the uptake of carbon dioxide by the grass species *Echinochloa crus-galli*.

Plant	Type	Treatment	conc	uptake
Mc1	Mississippi	chilled	350	18.9
Mc3	Mississippi	chilled	250	17.9
Mn3	Mississippi	nonchilled	95	11.3
Mn3	Mississippi	nonchilled	350	27.9
Qn1	Quebec	nonchilled	500	35.3
Qc2	Quebec	chilled	500	38.6
Qc3	Quebec	chilled	175	21.0
Qn3	Quebec	nonchilled	500	42.9
Qn3	Quebec	nonchilled	175	32.4

... and so on for 84 lines altogether

A data frame is a kind of tabular organization of data. In this example, it records several variables: the geographic origin of the plant (Mississippi or Quebec), and whether the plant had been chilled overnight before the uptake measurement was made, the ambient atmospheric CO<sub>2</sub> level and the uptake of CO<sub>2</sub> by the plant.

Each of the components of the data frame could be stored by an object of character type or of numerical type, for instance, `Treatment` is character and `conc` is numerical. The data frame brings the various components together in one place, facilitating processing and analysis of the data.

The information for a data frame is often stored in a spreadsheet file and read into R for analysis. Until you learn how to read in such files, you can use some of the built-in data frames intended for example purposes. If you want to follow along the examples with the CO<sub>2</sub> data frame, use this command to create an object named `C02` that contains the data frame:

```
> data(C02)
```

### Columns of Data Frames

Perhaps the most common operation on a data frame is to refer to the values in a single column. This can be done using a special syntax involving the \$ sign. To refer to the conc column in the C02 data frame, you would use the command C02\$conc. To refer to the Treatment column, use C02\$Treatment. Think of this style of reference as analogous to naming a person with a first name and a last name: the name of the data frame object comes first and the variable name second, separated by the \$, something like Einstein\$Albert.

Each component is just like an ordinary object and can be used in any way you would use an object:

```
> length(C02$conc)
[1] 84
> mean(C02$conc)
[1] 435
> max(C02$uptake)
[1] 45.5
> table( C02$Treatment)
nonchilled    chilled
          42         42
```

### Logical Data and Logical Operators

Many computations involve selections of subsets of data that meet some criterion. For example, in studying the health of newborn babies, you might want to focus only on those below a certain birthweight or perhaps those babies whose mother smoked during pregnancy. The question of whether the case satisfies the criterion boils down to a yes-or-no answer.

Logic is the study of valid inference; it is intimately tied up with the idea of truth versus falsehood. In computer languages, **logical data** refers to a type of data that can represent whether something is true or false. To illustrate, consider the simple sequence 1 to 7

```
> x = seq(1,7)
```

Now a simple question about the values in x: Are any of them greater than  $\pi$ ? Here's how you can ask that question:

```
> x > pi
[1] FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE
```

A computer command like `x>pi` is not the same as an algebraic statement like  $x > \pi$ . The algebraic statement describes the relationship between  $x$  and  $\pi$ , namely that  $x$  is greater than  $\pi$ . The computer command asks a question: Is the value of  $x$  greater than the value of  $\pi$ ? Asking this question invokes a computation; the returned value is the answer to the question, either TRUE or FALSE.

Here are some of the operators for asking such questions:

<code>x&lt;y</code>	Is x less than y?
<code>x&lt;=y</code>	Is x less than or equal to y?
<code>x&gt;y</code>	Is x greater than y?
<code>x&gt;=y</code>	Is x greater than or equal to y?
<code>x==y</code>	Is x equal to y?
<code>x!=y</code>	Is x unequal to y?

Notice the double equal signs in `x==y`. A single equal sign would be the assignment operator.

Mostly, these comparison operators apply to numbers. The `==` and `!=` operators also apply to character strings. To illustrate, I'll define two objects `v` and `w`:

```
> v = "hello"
> w = seq(1,15,by=3)
> w
 1  4  7 10 13
> w < 12
[1] TRUE TRUE TRUE TRUE FALSE
> w > 7
[1] FALSE FALSE FALSE TRUE TRUE
> w != 4
[1] TRUE FALSE TRUE TRUE TRUE
> v == "goodbye"
[1] FALSE
> v != "hi"
[1] TRUE
> v == "hello"
[1] TRUE
> v == "Hello"
[1] FALSE
```

The `FALSE` in the last line results from taking into account the difference between upper-case and lower-case letters.

Sometimes you need to combine more than one logical result. For example, to ask, "Is `w` between 7 and 12?" involves combining two separate questions: "Is `w` greater than 7 AND is it less than 12?" In the computer language, this question would be stated `w>7&w<12`.

```
> w > 7 & w < 12
[1] FALSE FALSE FALSE TRUE FALSE
```

There are also logical operators for "or" and "not". For instance, you might ask whether `w` is less than 5 OR greater than 9:

```
> w < 5 | w > 9
[1] TRUE TRUE FALSE TRUE TRUE
```

The "not" operator just flips `TRUE` and `FALSE`:

```
> !(w < 5 | w > 9)
[1] FALSE FALSE TRUE FALSE FALSE
```

As this example illustrates, you can group logical operations in just the same way as arithmetic operations.

### Missing Data

When recording data from an experiment or an observational study, it sometimes happens that a particular measurement can't be made or is lost or is otherwise unavailable. In R, such **missing data** can be recorded with the special code `NA`. As you might expect, arithmetic and other operations on missing data can't be sensibly performed: giving `NA` as an input produces `NA` as an output.

```
> 7 == NA
[1] NA
> NA == 'Hello'
[1] NA
> NA == NA
[1] NA
```

In order to test whether there is missing data, a special operator `is.na` can be used:

```
> is.na(NA)
[1] TRUE
```

### Collections

R can work with collections of numbers and character strings. Some operators work on each item in the collection, while others combine the items together in some way. To illustrate, I'll define three small collections, `x`, `y`, and `fruits`:

```
> x = seq(1,7)
> x
[1] 1 2 3 4 5 6 7

> y = c(7,8,9)
> y
[1] 7 8 9

> fruits = c("apple","berry","cherry")
> fruits
[1] "apple" "berry" "cherry"
```

The `c` operator used in defining `y` and `fruits` is useful for creating a small collection “by hand.” Often, however, collections will be created by reading in data from a file or using some other operator. I'll introduce these as needed for specific tasks.

Arithmetic and comparison operators often work item-by-item on the collection. For example:

```
> x + 100
[1] 101 102 103 104 105 106 107
> sqrt(y)
[1] 2.645751 2.828427 3.000000
> fruits == "cherry"
[1] FALSE FALSE  TRUE
```

If the operator involves two collections, they have to be the same size, or R will reuse the smaller collection to match the size of the larger one:

```
> x == y
[1] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
```

Warning message:

```
In x == y : longer object length is not a
multiple of shorter object length
```

The warning message is displayed when some aspect of the computation is deemed suspect or odd. Pay attention to such messages since they may signal that the computation the interpreter carried out is not the one you intended.

It's usually obvious what sorts of operators will combine the items of the collection rather than working on them item by item. Here are some examples:

```
> x
[1] 1 2 3 4 5 6 7
> y
[1] 7 8 9
> mean(x)
[1] 4
> median(y)
[1] 8
> min(x)
[1] 1
> max(x)
[1] 7
> sum(x)
[1] 28
> any( fruits == "cherry")
[1] TRUE
> all( fruits == "cherry")
[1] FALSE
```

The length operator tells how many items there are in the collection:

```
> length(x)
[1] 7
> length(y)
```

```
[1] 3
> length(fruits)
[1] 3
```

When there are too many items in a collection to display conveniently in one line, the R interpreter will break up the display over multiple lines.

```
> seq(3, 19)
[1] 3 4 5 6 7 8 9 10 11
[10] 12 13 14 15 16 17 18 19
```

At the start of each line, the number in brackets tells the index of the item that starts that line. In the above, for instance, the item 3 is displayed following a [1] because 3 is the first item in the collection. Similarly, the [10] indicates that the item 12 is the tenth item in the collection. The brackets are just for display purposes; they are not part of the collection itself.

### Defining your own operators

Occasionally, you may need to define your own operators. This is convenient if you need to repeat an operation many times or if you need to define a mathematical function.

It's important to keep in mind the difference between an operator and a command. A command is an instruction to perform a particular computation on a particular input argument or set of input arguments. The input arguments always have to be values, though of course you can refer to the value by giving the name of an object that has already been assigned a value.

In contrast, in defining an operator, you can treat the arguments abstractly; just a name without a value having been assigned. To illustrate, here is a command that creates the mathematical function  $f(x) = 3x^2 + 2$  and stores it in an operator named `f`:

```
> f = function(x) { 3*x^2 + 2 }
```

Once you have defined the function, you can invoke it in the standard way. For example:

```
> f(3)
[1] 29
> f(10)
[1] 302
```

There are some novel features to the syntax used to define a new operator. First, the arguments to `function` aren't treated as values but as pure names. Second, the contents of the curly braces `{` and `}` — the function contents — are the commands that will be evaluated when the function is invoked.

It doesn't matter what names you use in the function contents so long as they match the names used in the arguments to `function`. For example, here is another operator, called `g`, that will perform exactly the same computation as `f` when invoked:

```
> g = function(marge) { 3*marge^2 + 2 }
```

When you invoke an operator, the interpreter carries out several steps. Consider the invocation

```
> g(7)
```

In carrying out this command, the interpreter will:

1. Temporarily define or redefine an object `marge` that has the value 7.
2. Execute the function contents.
3. Return the value of these contents as the return value of the command.
4. Discard the definition or redefinition in (1).

Operators can have more than one argument. For instance, here is an operator `hypotenuse` that computes the length of the hypotenuse of a right triangle given the lengths of the legs

```
> hypotenuse = function(a,b) { sqrt( a^2 + b^2 ) }
```

When programmers create new operators that they expect to use on many different occasions, they put the commands to define the operators into a text file called a **source file**. This file can be read into R using a special operator, called `source` that causes the commands to be executed, thereby defining the new operators.

### Saving and Documenting Your Work

Up to now, the commands used as examples have been simple one-line statements. As you work further, you will build more elaborate computations by combining simpler ones. It will become important to be able to document what you have done, providing a record so that others can confirm your results and so that you and others can modify your work as needed.

One way in which a record is created of your interaction with the computer is the dialog in the interpreter console itself. In some ways this is analogous to a document created by a word processor: for example, you can copy the contents and paste it into another document.

But the idea of dialog-as-document is flawed. For example, in a word-processor, when you correct a mistake the old version is erased. But in the R dialog, to fix a mistake you give a new command — the old, mistaken command is still there in the dialog.

You should keep in mind that there are several different components, some of which are more appropriate than others for your documentation.

**Your Commands** The commands that you execute are what defines the computation being performed. These commands themselves are a valuable form of documentation.

**The objects you create** These objects, and the values that are stored in them, reflect the **state** of the computation. If you want to pick up on your work where you left off, you can save these objects. This is called “saving the **workspace**.”

**Side effects** This refers to the output printed by the interpreter and plots. Sometimes you will want to include this in your documentation, but usually just select elements.

#### 1.4.4 Using Customizations to R

One of the features that makes R so powerful is that new commands can easily be added to the system. This makes it possible, for instance, to customize the software to make routine tasks easier. Such customizations have been written specifically for the people following this book. To use them — and you will need them in later chapters — you should download a file from the web:

`www.macalester.edu/~kaplan/ISM/ISM.Rdata`

This file contains various data sets that you be using.

Once you have downloaded the file, called `ISM.Rdata`, double-click on it in the ordinary way to start R or to load the data into an already started session of R.

For review purposes only



# Chapter 2

## Data: Cases, Variables, Samples

*The tendency of the casual mind is to pick out or stumble upon a sample which supports or defies its prejudices, and then to make it the representative of a whole class. — Walter Lippmann (1889-1974)*

The word “data” is plural. This is appropriate. Statistics is, at its heart, about variability: how things differ from one another.

Figure 2.1 shows a small collection of sea shells gathered during an idle quarter hour sitting at one spot on the beach on Malololailai island in Fiji. All

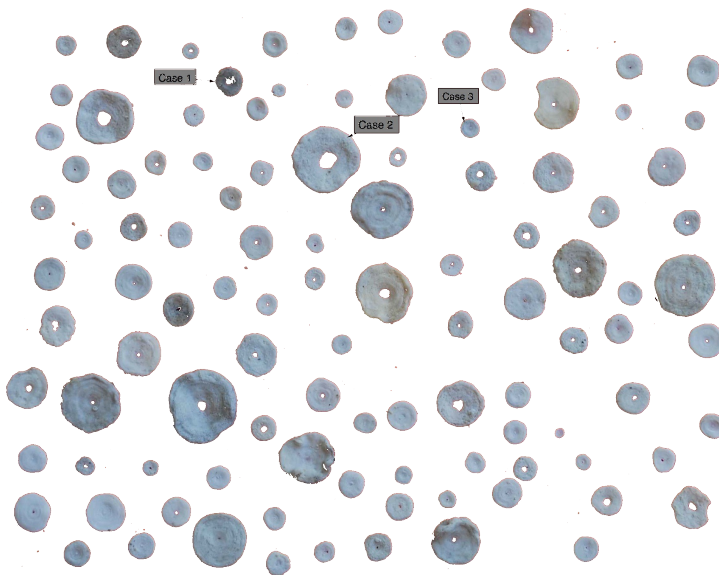


Figure 2.1: A collection of 103 sea shells. (The photo is printed to 3/4 scale.)

For review purposes only

the shells in the collection are similar: small disk-shaped shells with a hole in the center. But the shells also differ from one another in overall size and weight, in color, in smoothness, in the size of the hole, etc.

Any data set is something like the shell collection. It consists of **cases**: the objects in the collection. Each case has one or more attributes or qualities, called **variables**. This word “variable” emphasizes that it is differences or variation that is often of primary interest.

Usually, there are many possible variables. The researcher chooses those that are of interest, often drawing on detailed knowledge of the system that is under study. The researcher measures or observes the value of each variable for each case. The result is a **table**, also known as a **data frame**: a sort of spreadsheet. Within the data frame, each row refers to one case, each column to one variable.

A data frame for the sea shell collection might look like this:

Case	diameter	weight	color	hole
1	4.3 mm	0.010 mg	dark	medium
2	12.0 mm	0.050 mg	light	very large
3	3.8 mm	0.005 mg	light	none
and so on — there are 103 cases altogether				

Each individual shell has a case number that identifies it; three of these are shown in the figure. These case numbers are arbitrary; the position of each case in the table — first, second, tenth, and so on — is of no significance. The point of assigning an identifying case number is just to make it easier to refer to individual cases later on.

If the data frame had been a collection of people rather than shells, the person’s name or another identifying label could be used to identify the case.

There are many sorts of data that are recorded in different formats. For example, photographic images are stored as arrays of pixels without any reference to cases and variables. But the data frame format is very general and can be used to arrange all sorts of data, even if it is not the format always used in practice. Even data that seems at first not to be suitable to arrange as a data frame is often stored this way, for example the title/artist/genre/album organization found in music MP3 players or the geographic features (place locations and names, boundaries, rivers, etc.) found in geographic information systems (GIS).

## 2.1 Kinds of Variables

Most people tend to think of data as numeric, but variables can also be descriptions, as the sea shell collection illustrates. The two basic types of data are:

**Quantitative:** Naturally represented by a number, for instance diameter, weight, temperature, age, and so on.

**Categorical:** A description that can be put simply into words or categories, for instance male versus female or red vs green vs yellow, and so on. The

value for each case is selected from a fixed set of possibilities. This set of possibilities are the **levels** of the categorical variable.

Categorical variables show up in many different guises. For example, a data frame holding information about courses at a college might have a variable *subject* with levels *biology*, *chemistry*, *dance*, *economics*, and so on. The variable *semester* could have levels *Fall2008*, *Spring2009*, *Fall2009*, and so on. Even the *instructor* is a categorical variable. The instructor's name or ID number could be used as the level of the variable.

Quantitative variables are numerical but they often have units attached to them. In the shell data frame, the variable *diameter* has been recorded in millimeters, while *weight* is given in milligrams. The usual practice is to treat quantitative variables as a pure number, without units being given explicitly. The information about the units — for instance that *diameter* is specified in millimeters — is kept in a separate place called a **code book**.

The code book contains a short description of each variable. For instance, the code book for the shell data might look like this:

Code book for shells from Malololailai island, collected on January 12, 2008.

*diameter*: the diameter of the disk in millimeters.

*weight*: the shell weight in milligrams.

*color*: a subjective description of the color. Levels: *light*, *medium*, and *dark*.

*hole*: the size of the inner hole. Levels: *none*, *small*, *large*, very *large*.

On the computer, a data frame is usually stored as a spreadsheet file, while the corresponding code book can be stored separately as a text file.

Sometimes quantitative information is represented by categories as in the *hole* variable for the sea shells. For instance, a data frame holding a variable *income* might naturally be stored as a number in, say, dollars per year. Alternatively, the variable might be treated categorically with levels of, say, “less than \$10,000 per year,” “between \$10,000 and \$20,000,” and so on. Almost always, the genuinely quantitative form is to be preferred. It conveys more information even if it is not correct to the last digit.

The distinction between quantitative and categorical data is essential, but there are other distinctions that can be helpful even if they are less important.

Some categorical variables have levels that have a *natural* order. For example, a categorical variable for temperature might have levels such as “cold,” “warm,” “hot,” “very hot.” Variables like this are called **ordinal**. Opinion surveys often ask for a choice from an ordered set such as this: *strongly disagree*, *disagree*, *no opinion*, *agree*, *strongly agree*.

For the most part, this book will treat ordinal variables like any other form of categorical variable. But it's worthwhile to pay attention to the natural ordering of an ordinal variable. Indeed, sometimes it can be appropriate to treat an ordinal variable as if it were quantitative.

## 2.2 Data Frames and the Unit of Analysis

When collecting and organizing data, it's important to be clear about what is a case. For the sea shells, this is pretty obvious; each individual shell is an individual case. But in many situations, it's not so clear.

A key idea is the **unit of analysis**. Suppose, for instance, that you want to study the link between teacher pay, class size, and the performance of school children. There are all sorts of possibilities for how to analyze the data you collect. You might decide to compare different schools, looking at the average class size, average teacher pay, and average student performance in each school. Here, the unit of analysis is the school.

Or perhaps rather than averaging over all the classes in one school, you want to compare different classes, looking at the performance of the students in each class separately. The unit of analysis here is the class.

You might even decide to look at the performance of individual students, investigating how their performance is linked to the individual student's family income or the education of the student's parents. Perhaps even include the salary of student's teacher, the size of the student's class, and so on. The unit of analysis here is the individual student.

What's the difference between a unit of analysis and a case? A case is a row in a data frame. In many studies, you need to bring together different data frames, each of which may have a different notion of case. Returning to the teacher's pay example, one can easily imagine at least three different data frames being involved, with each frame storing data at a different level:

1. A frame with each class being a case and variables such as the size of the class, the school in which the class is taught, etc.
2. A frame with each teacher being a case and variables such as the teacher's salary, years of experience, advanced training, etc.
3. A frame with each student being a case and variables such as the student's test scores, the class that the student is in, and the student's family income, and parent education.

Once you choose the unit of analysis, you combine information from the different data frames to carry out the data analysis, generating a single data frame in which cases are your chosen unit of analysis. The choice of the unit of analysis can be determined by many things, such as the availability of data. As a general rule, it's best to make the unit of analysis as small as possible. But there can be obstacles to doing this. You might find, for instance, that for privacy reasons (or less legitimate reasons of secrecy) the school district is unwilling to release the data at the individual student level, or even to release data on individual classes.

In the past, limitations in data analysis techniques and computational power provided a reason to use a coarse unit of analysis. Only a small amount of data could be handled effectively, so the unit of analysis was made large. For example, rather than using tens of thousand of individual students as the unit of

analysis, a few dozen schools might be used instead. Nowadays these reasons are obsolete. The methods that will be covered in this book allow for a very fine unit of analysis. Standard personal computers have plenty of power to perform the required calculations.

## 2.3 Populations and Samples

A data frame is a collection, but a collection of what? Two important statistical terms are “population” and “sample.” A **population** is the set of all the possible objects or units which might have been included in the collection. The root of the word “population” refers to people, and often one works with data frames in which the cases are indeed individual people. The statistical concept is broader; one might have a population of sea shells, a population of houses, a population of events such as earthquakes or coin flips.

A **sample** is a selection of cases from the population. The **sample size** is the number of cases in the sample. For the shells in Figure 2.1, the sample size is  $n = 103$ .

A **census** is a sample that contains the entire population. The most familiar sort of census is the kind to count the people living in a country. The United States and the United Kingdom have a census every ten years. Countries such as Canada, Australia, and New Zealand hold a census every five years.

Almost always, the sample is just a small fraction of the population. There are good reasons for this. It can be expensive or damaging to take a sample: Imagine a biologist who tried to use all the laboratory rats in the world for his or her work! Still, when you draw a sample, it is generally because you are interested in finding out something about the population rather than just the sample at hand. That is, you want the sample to be genuinely representative of the population. (In some fields, the ability to draw conclusions from a sample that can be generalized is referred to as **external validity** or **transferability**.)

The process by which the sample is taken is important because it controls what sorts of conclusions can legitimately be drawn from the sample. One of the most important ideas of statistics is that a sample will be representative of the population if the sample is collected at random. In a **simple random sample**, each member of the population is equally likely to be included in the sample.

Ironically, taking a random sample, even from a single spot on the beach, requires organization and planning. The sea shells were collected haphazardly, but this is not a genuinely random sample. The bigger shells are much easier to see and pick up than the very small ones, so there is reason to think that the small shells are under-represented: the collection doesn't have as big a proportion of them as in the population. To make the sample genuinely random, you need to have access in some way to the entire population so that you can pick any member with equal probability. For instance, if you want a sample of students at a particular university, you can get a list of all the students from the university registrar and use a computer to pick randomly from the list. Such a list of the entire set of possible cases is called a **sampling frame**.

In a sense, the sampling frame is the *definition* of the population for the purpose of drawing conclusions from the sample. For instance, a researcher studying cancer treatments might take the sampling frame to be the list of all the patients who visit a particular clinic in a specified month. A random sample from that sampling frame can reasonably be assumed to represent that particular population, but not necessarily the population of all cancer patients.

You should always be careful to define your sampling frame precisely. If you decide to sample university students by picking randomly from those who enter the front door of the library, you will get a sample that might not be typical for *all* university students. There's nothing wrong with using the library students for your sample, but you need to be aware that your sample will be representative of just the library students, not necessarily all students.

When sampling at random, use formal random processes. For example, if you are sampling students who walk into the library, you can flip a coin to decide whether to include that student in your sample. When your sampling frame is in the form of a list, it's wise to use a computer random number generator to select the cases to include in the sample.

A **convenience sample** is one where the sampling frame is defined mainly in a way that makes it easy for the researcher. For example, during lectures I often sample from the set of students in my class. These students — the ones who take statistics courses from me — are not necessarily representative of all university students. It might be fine to take a convenience sample in a quick, informal, preliminary study. But don't make the mistake of assuming that the convenience sample is representative of the population. Even if you believe it yourself, how will you convince the people who are skeptical about your results?

When cases are selected in an informal way, it's possible for the researcher to introduce a non-representativeness or **sampling bias**. For example, in deciding which students to interview who walk into the library, you might consciously or subconsciously select those who seem most approachable or who don't seem to be in a hurry.

There are many possible sources of sampling bias. In surveys, sampling bias can come from non-response or self-selection. Perhaps some of the students who you selected randomly from the people entering the library have declined to participate in your survey. This **non-response** can make your sample non-representative. Or, perhaps some people who you didn't pick at random have walked up to you to see what you are up to and want to be surveyed themselves. Such **self-selected** people are often different from people who you would pick at random.

In a famous example of self-selection bias, the newspaper columnist Ann Landers asked her readers, "If you had it to do over again, would you have children?" Over 70% of the respondents who wrote in said "No." This result is utterly different from what was found in surveys on the same question done with a random sampling methodology: more than 90% said "Yes." Presumably the people who bothered to write were people who had had a particularly bad experience as parents whereas the randomly selected parents are representative of the whole population. Or, as Ann Landers wrote, "[T]he hurt, angry, and

disenchanted tend to write more readily than the contented ....” (See [7].)

Non-response is often a factor in political polls, where people don’t like to express views that they think will be unpopular.

It’s hard to take a genuinely random sample. But if you don’t, you have no guarantee that your sample is representative. Do what you can to define your sampling frame precisely and to make your selections as randomly as possible from that frame. By using formal selection procedures (e.g., coin flips, computer random number generators) you have the opportunity to convince skeptics who might otherwise wonder what hidden biases were introduced by informal selections. If you believe that your imperfect selection may have introduced biases — for example the suspected under-representation of small shells in my collection — be up-front and honest about it. In surveys, you should keep track of the non-response rate and include that in whatever report you make of your study.

**Example 2.1: Struggling for a Random Sample** Good researchers take great effort to secure a random sample. One evening I received a phone call at home from the state health department. They were conducting a survey of access to health care, in particular how often people have illnesses for which they don’t get treatment. The person on the other end of the phone told me that they were dialing numbers randomly, checked to make sure that I live in the area of interest, and asked how many adults over age 18 live in the household. “Two,” I replied, “Me and my wife.” The researcher asked me to hold a minute while she generated a random number. Then the researcher asked to speak to my wife. “She isn’t home right now, but I’ll be happy to help you,” I offered. No deal.

The sampling frame was adults over age 18 who live in a particular area. Once the researcher had made a random selection, as she did after asking how many adults are in my household, she wasn’t going to accept any substitutes. It took three follow-up phone calls over a few days — at least that’s how many I answered, who knows how many I wasn’t home for — before the researcher was able to contact my wife. The researcher declined to interview me in order to avoid self-selection bias and worked hard to contact my wife — the randomly selected member of our household — in order to avoid non-response bias.

## 2.4 Longitudinal and Cross-Sectional Samples

Data are often collected to study the links between different traits. For example, the data in the following table are a small part of a larger data set of the speeds of runners in a ten-mile race held in Washington, D.C. in 2004. The variable net gives the time from the start line to the finish line, in seconds. Such data might be used to study the link between age and speed, for example to find out at what age people run the fastest and how much they slow down as they age beyond that.

For review purposes only



state	net	age	sex
DC	6382	23	F
VA	5080	26	F
DC	4742	27	M
Kenya	2962	27	M
DC	6291	29	F
MD	6405	32	M
VA	6608	34	F
DC	5921	37	M
MD	5549	41	F
MD	5486	46	F
VA	8374	53	F
PA	6026	53	M
VA	5526	60	M
VA	5585	61	M
VA	5931	65	M

... and so on.

This sample is a **cross section**, a snapshot of the population that includes people of different ages. Each person is included only once.

Another type of sample is **longitudinal**, where the cases are tracked over time, each person being included more than once in the data frame. A longitudinal data set for the runners might look like this:

state	net	age	sex	year
DC	6382	23	F	2004
DC	6516	24	F	2005
DC	6493	25	F	2006
DC	6526	26	F	2007
DC	6571	27	F	2008
MD	6405	32	M	2004
MD	6819	34	M	2006
MD	6753	35	M	2007

... and so on.

If your concern is to understand how individual change as they age, it's best to collect data that show such change in individuals. Using cross-sectional data to study a longitudinal problem is risky. Suppose, as seems likely, that younger runners who are slow tend to drop out of racing as they age, so the older runners who do participate are those who tend to be faster. This could bias your estimate of how running speed changes with age.

For review purposes only



## 2.5 Computational Technique

### 2.5.1 Reading and Writing Data

Data used in statistical modeling are usually organized into tables, often created using spreadsheet software. Most people presume that the same software used to create a table of data should be used to display and analyze it. This is part of the reason for the popularity of spreadsheet programs such as Excel.

For statistical modeling, it's helpful to take another approach that strictly separates the processes of data collection and of data analysis: use one program to create data files and another program to analyze the data stored in those files. By doing this, one guarantees that the original data are not modified accidentally in the process of analyzing them. This also makes it possible to perform many different analyses of the data; modelers often create and compare many different models of the same data.

The spreadsheet programs that can create data files use a variety of different formats. Many of these formats are proprietary and include various features that make it difficult for any other software to read the file. A good, simple, general purpose format supported by spreadsheet software and by statistical software is called the **comma separated value** or **CSV** format.

The next sections describe how to read data from a CSV file into R and how to use a spreadsheet program to create new data.

#### Reading CSV Files into R

For the data sets associated with this book and its exercises, an easy way to import data into R is with the `ISMdata` operator that's included in the extensions to R found in the `ISM.Rdata` workspace file. (See Section 1.4.4 on page 30. You must load the workspace file before you can use `ISMdata`.)

`ISMdata` lets you refer to a file by a short name in quotes without worrying about where it is located. It knows where to locate the data files used with this book, whether they be on your computer or on the web. For example, the data set `hdd-minneapolis.csv` is stored on the Internet. You can read it into an object named `hdd` with this statement:

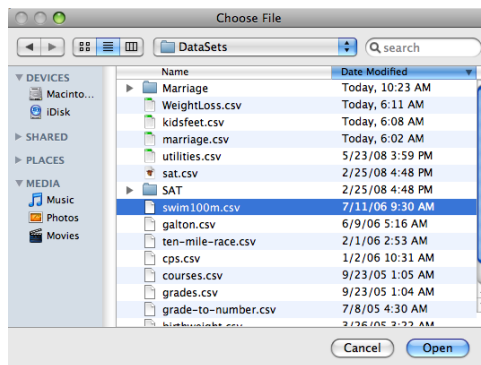
```
> hdd = ISMdata("hdd-minneapolis.csv")
Not in library. Trying to find it on the web ...
File was read from the web.
```

If you do not have an Internet connection, then `ISMdata` will be able to locate only files on your computer.

It is possible to use `ISMdata` to read in your own data that you have created and stored in CSV files. To do this, you need to tell `ISMdata` where the data is located. The easiest way to do this is to use a mouse-based file navigator. To do this, invoke `ISMdata` with no input argument:

```
> swim = ISMdata()
```

Since there was no character string file name given as an argument, `ISMdata` brings up a file navigator to let you select the file interactively:



Selecting a file interactively.

Selecting and opening the desired file will cause it to be read into R as a data frame. Make sure to choose the correct CSV file. If you choose some other sort of file by accident, `ISMdata` will struggle to read it in, displaying various junk on your screen.

The statement above will cause the resulting data frame to be called `swim`. But be careful. If you accidentally choose a different file (e.g., `kidsfeet.csv`) the data from that file will be read in and stored under the name `swim`, even though they have nothing to do with swimming.

[Optional] Although `ISMdata` will work well for most purposes, some users might prefer the flexibility offered by the built-in R operators for reading in files. These include `read.csv`, `scan`, and a variety of operators for importing files from other packages, such as the `read.spss` operator in the “foreign” library package.

For obvious reasons, the most of the examples used in this book are based on data that has already been collected and stored in CSV files. In your own investigations, you will generally need to create your own data sets. Instructions and suggestions for creating CSV files and codebooks are given in the exercises.

## 2.5.2 Simple Operations with Data Frames

To illustrate, here are data on world-record swimming times:

```
> swim = ISMdata("swim100m.csv")
```

This data set, stored in the object named `swim`, has three variables:

```
> names(swim)
[1] "year" "time" "sex"
```

Year refers to the calendar year in which the record was set; time is the record time itself, in seconds; sex records whether the record is for men or women.



```
> head(swim)
  year time sex
1 1905 65.8  M
2 1908 65.6  M
3 1910 62.8  M
... and so on.
```

### Numerical Operations

There are two basic kinds of variables: quantitative and categorical. R treats these variables differently, as it should since operations that make perfect sense for a quantitative variable (such as the sum or the mean or median) make little sense for categorical variables. For instance, it makes sense to compute

```
> max(swim$year)
[1] 2004
```

But it does not make sense to compute the mean sex:

```
> max(swim$sex)
Error in Summary.factor
  max not meaningful for factors
```

The word **factor** is how R refers to categorical variables. Once you know that, the second line of the error message makes more sense.

### Adding a New Variable

Sometimes you will compute a new quantity from the variables and you want to treat this as a new variable. You can do this by assignment to the data frame, using the \$ and giving a new name for the variable. As a trivial example, here is how to convert the swim time from seconds to minutes:

```
> swim$minutes = swim$time / 60
```

Once this has been done, the new variable appears just like the old ones:

```
> names(swim)
[1] "year"    "time"    "sex"     "minutes"
```

You could also, if you want, redefine an existing variable, for instance:

```
> swim$time = swim$time / 60
```

Such assignment operations do not change the original file from which the data were read, only the data frame in the current session of R.

### Extracting Subsets of Data

Selecting a subset of cases is done with the `subset` operator. For instance, here's how to create a data frame with just the women's records:

```
> women = subset( swim, sex=='F' )
```

The `subset` operator takes two arguments: the first is a data frame from which to extract the subset. The second is a logical (TRUE/FALSE) criterion for each case, saying whether to include it.

Notice that in this example, the name `sex` was used, rather than the full `swim$sex`. The `subset` operator allows the shorthand since the first argument sets a context for evaluating any names in the second argument. Other operators also allow this sort of shorthand.

The `subset` operator creates a new data frame which you can assign to a name; it does not modify the original data frame. You can have as many data frames as you want in an R session, so there is little reason to modify the original. But if you want to, you can do it by re-assignment:

```
> swim = subset( swim, sex=='F' )
```

After this command, the male records are no longer in `swim`. If you want them back, you have to re-read the original data file.

### 2.5.3 Sampling

Suppose you have a sampling frame with 1000 cases, arranged as a spreadsheet with one row for each case. You want to select a random set of 20 of these. The `shuffle` operator lets you pick random members of a set.

A common use for `shuffle` is to pick random cases from a data frame, just as you might do when sampling randomly from a sampling frame.

```
> shuffle(swim, 5)
  year  time sex
55 1976 55.65  F
39 1924 72.20  F
12 1944 55.90  M
18 1964 52.90  M
46 1936 64.60  F
```

The results returned by `shuffle` will never contain the same case more than once, just as if you were dealing cards from a shuffled deck. In contrast, `resample` replaces each case after it is dealt so that it can appear more than once in the result. This is called **sampling with replacement**. This will be useful later on when studying the statistical properties of the sampling process. For example, `resample` allows you to generate a sample of any size, even one that is bigger than the data set from which the sample is being drawn.

# Chapter 3

## Describing Variation

*Variation itself is nature's only irreducible essence. Variation is the hard reality, not a set of imperfect measures for a central tendency. Means and medians are the abstractions. — Stephen Jay Gould*

A statistical model partitions variation into parts. People describe the partitioning in different ways depending on their purposes and the conventions of the field in which they work: explained variation versus unexplained variation; described variation versus undescribed; predicted variation versus unpredicted; signal versus noise; common versus individual. This chapter describes ways to quantify variation in a single variable. Once you can quantify variation, you can describe how models divide it up.

To start, consider a familiar situation: the variation in human heights. Everyone is familiar with height and how heights vary from person to person, so variation in height provides a nice example to compare intuition with the formal descriptions of statistics. Perhaps for this reason, height was an important topic for early statisticians. In the 1880s, Francis Galton, one of the pioneers of statistics, collected data on the heights of about 900 adult children and their parents in London. Figure 3.1 shows part of his notebook.

Galton was interested in studying the relationship between a full-grown child's height and his or her mother's and father's height. One way to study this relationship is to build a model that accounts for the child's height — the **response variable** — by one or more **explanatory variables** such as mother's height or father's height or the child's sex. In later chapters you will construct such models. For now, though, the objective is to describe and quantify how the response variable varies across children without taking into consideration any explanatory variables.

Galton's height measurements are from about 200 more-or-less normal families in the city of London. In itself, this is an interesting choice. One might think that the place to start would be with exceptionally short or tall people, looking at what factors are associated with these extremes.

For review purposes only

	Father	Mother	Sons in order of height	Daughters in order of height.
1	18.5	7.0	13.2	9.2, 9.0, 9.0
2	15.5	6.5	13.5, 12.5	5.5, 5.5
3	15.0	about 4.0	11.0	8.0
4	15.0	4.0	10.5, 8.5	7.0, 4.5, 3.0
5	15.0	1.5	12.0, 9.0, 8.0	6.5, 2.5, 2.5

Figure 3.1: Part of Francis Galton's notebook recording the heights of parents and their adult children. [8]

Adults range in height from a couple of feet to about nine feet. (See Figure 3.2.) One way to describe variation is by the **range of extremes**: an interval that includes every case from the smallest to the largest.

What's nice about describing variation through the extremes is that the range includes every case. But there are disadvantages. First, you usually don't always have a **census**, measurements on the entire population. Instead of a census, you typically have only a **sample**, a subset of the population. Usually only a small proportion of the population is included in a sample. For example, of the millions of people in London, Galton's sample included only 900 people. From such a sample, Galton would have had no reason to believe the either the tallest or shortest person in London, let alone the world, happened to be included.

A second disadvantage of using the extremes is that it can give a picture that is untypical. The vast majority of adults are between  $4\frac{1}{2}$  feet and 7 feet tall. Indeed, an even narrower range — say 5 to  $6\frac{1}{2}$  feet — would give you a very good idea of typical variation in heights. Giving a comprehensive range — 2 to 9 feet — would be misleading in important ways, even if it were literally correct.

A third disadvantage of using the extremes is that even a single case can have a strong influence on your description. There are about six billion people on earth. The discovery of even a single, exceptional 12-foot person would cause you substantially to alter your description of heights even though the population — minus the one new case — remains unchanged.

It's natural for people to think about variation in terms of records and extremes. People are used to drawing conclusions from stories that are newsworthy and exceptional; indeed, such anecdotes are mostly what you read and hear about. In statistics, however, the focus is usually on the unexceptional cases, the typical cases. With such a focus, it's important to think about **typical variation** rather than extreme variation.



Figure 3.2: Some extremes of height. Angus McAskill (1825-1863) and Charles Sherwood Stratton (1838-1883). McAskill was 7 feet 9 inches tall. Stratton, also known as Tom Thumb, was 2 feet 6 inches in height.

### 3.1 Coverage Intervals

One way to describe typical variation is to specify a fraction of the cases that are regarded as typical and then give the **coverage interval** or range that includes that fraction of the cases.

Imagine arranging all of the people in Galton's sample of 900 into order, the way a school-teacher might, from shortest to tallest. Now walk down the line, starting at the shortest, counting heads. At some point you will reach the person at position 225. This position is special because one quarter of the 900 people in line are shorter and three quarters are taller. The height of the person at position 225 — 64.0 inches — is the 25th **percentile** of the height variable in the sample.

Continue down the line until you reach the person at position 675. The height of this person — 69.7 inches — is the 75th percentile of height in the sample.

The range from 25th to 75th percentile is the **50-percent coverage interval**: 64.0 to 69.7 inches in Galton's data. Within this interval is 50% of the cases.

For most purposes, a 50% coverage interval excludes too much; half the cases are outside of the interval.

Scientific practice has established a convention, the **95-percent coverage interval** that is more inclusive than the 50% interval but not so tied to the extremes as the 100% interval. The 95% coverage interval includes all but 5% of the cases, excluding the shortest 2.5% and the tallest 2.5%.

To calculate the 95% coverage interval, find the 2.5 percentile and the 97.5

Position in list $k$	Height (inches)	# of previous cases $k - 1$	Percentile
1	59.0	0	0
2	61.5	1	10
3	62.0	2	20
4	63.0	3	30
5	64.7	4	40
6	65.5	5	50
7	68.0	6	60
8	69.0	7	70
9	72.0	8	80
10	72.0	9	90
11	72.0	10	100

Table 3.1: Finding sample percentiles by sorting and counting.

percentile. In Galton's height data, these are 60 inches and 73 inches, respectively.

There is nothing magical about using the coverage fractions 50% or 95%. They are just conventions that make it easier to communicate clearly. But they are important and widely used conventions.

To illustrate in more detail the process of finding coverage intervals, let's look at Galton's data. Looking through all 900 cases would be tedious, so the example involves just a few cases, the heights of 11 randomly selected people:

72.0 62.0 68.0 65.5 63.0 64.7 72.0 69.0 61.5 72.0 59.0

Table 3.1 puts these 11 cases into sorted order and gives for each position in the sorted list the corresponding percentile. For any given height in the table, you can look up the percentile of that height, effectively the fraction of cases that came previously in the list. In translating the position in the sorted list to a percentile, convention puts the smallest case at the 0th percentile and the largest case at the 100th percentile. For the  $k$ th position in the sorted list of  $n$  cases, the percentile is taken to be  $(k - 1)/(n - 1)$ .

With  $n = 11$  cases in the sample, the sorted cases themselves stand for the 0th, 10th, 20th, ..., 90th, and 100th percentiles. If you want to calculate a percentile that falls in between these values, you (or the software) interpolate between the samples. For instance, the 75th percentile would, for  $n = 11$ , be taken as half-way between the values of the 70th and 80th percentile cases.

Coverage intervals are found from the tabulated percentiles. The 50% coverage interval runs from the 25th to the 75th percentile. Those exact percentiles happen not be in the table, but you can estimate them. Take the 25th percentile to be half way between the 20th and 30th percentiles: 62.5 inches. Similarly, take the 75th percentile to be half way between the 70th and 80th percentiles: 70.5 inches.

Thus, the 50% coverage interval for this small subset of  $N = 11$  cases is 62.5



to 70.5 inches. For the complete set of  $N = 900$  cases, the interval is 64 to 69 inches — not exactly the same, but not too much different. In general you will find that the larger the sample, the closer the estimated values will be to what you would have found from a census of the entire population.

This small  $N = 11$  subset of Galton's data illustrates a potential difficulty of a 95% coverage interval: The values of the 2.5th and 97.5th percentiles in a small data set depend heavily on the extreme cases, since interpolation is needed to find the percentiles. In larger data sets, this is not so much of a problem.

A reasonable person might object that the 0th percentile of the sample is probably not the 0th percentile of the population; a small sample almost certainly does not contain the shortest person in the population. There is no good way to know whether the population extremes will be close to the sample extremes and therefore you cannot demonstrate that the estimates of the extremes based on the sample are valid for the population. The ability to draw demonstrably valid inferences from sample to population is one of the reasons to use a 50% or 95% coverage interval rather than the range of extremes.

Different fields have varying conventions for dividing groups into parts. In the various literatures, one will read about **quintiles** (division into 5 equally sized groups, common in giving economic data), **stanines** (division into 9 unevenly sized groups, common in education testing), and so on.

In general-purpose statistics, it's conventional to divide into four groups: **quartiles**. The dividing point between the first and second quartiles is the 25th percentile. For this reason, the 25th percentile is often called the "first quartile." Similarly, the 75th percentile is called the "third quartile."

The most famous percentile is the 50th: the **median**, which is the value that half of the cases are above and half below. The median gives a good representation of a typical value. In this sense, it is much like the **mean**: the average of all the values.

Neither the median nor the mean describes the variation. For example, knowing that the median of Galton's sample of heights is 66 inches does not give you any indication of what is a typical range of heights. In the next section, however, you'll see how the mean can be used to set up an important way of measuring variation: the typical distance of cases from the mean.

## 3.2 The Variance and Standard Deviation

The 95% and 50% coverage intervals are important descriptions of variation. For constructing models, however, another format for describing variation is more widely used: the variance and standard deviation.

To set the background, imagine that you have been asked to describe a variable in terms of a *single* number that typifies the group. Think of this as a very simple model, one that treats all the cases as exactly the same. For human heights, for instance, a reasonable model is that people are about 5 feet 8 inches tall (68 inches).

**Aside. 3.1** What's Normal?

You've just been told that a friend has hypernatremia. Sounds serious. So you look it up on the web and find out that it has to do with the level of sodium in the blood:

For adults and older children, the normal range [of sodium] is 137 to 145 millimoles per liter. For infants, the normal range is somewhat lower, up to one or two millimoles per liter from the adult range. As soon as a person has more than 145 millimoles per liter of blood serum, then he has hypernatremia, which is too great a concentration of sodium in his blood serum. If a person's serum sodium level falls below 137 millimoles per liter, then they have hyponatremia, which is too low a concentration of sodium in his blood serum. [From <http://www.ndif.org/faqs>.]

You wonder, what do they mean by “normal range?” Do they mean that your body stops functioning properly once sodium gets above 145 millimoles per liter? How much above? Is 145.1 too large for the body to work properly? Is 144.9 fine?

Or do they mean a coverage interval? But what kind of coverage interval: 50%, 80%, 95%, 99%, or somethings else? And in what population? Healthy people? People who go for blood tests? Hospitalized people?

If 137 to 145 were a 95% coverage interval for healthy people, then 19 of 20 healthy people would fall in the 137 to 145 range. Of course this would mean that 1 out of 20 healthy people would be out of the normal range. Depending on how prevalent sickness is, it might even mean that most of the people outside of the normal range are actually healthy.

People frequently confuse “normal” in the sense of “inside a 95% coverage interval” with “normal” in the sense of “functions properly.” It doesn't help that publications often don't make clear what they mean by normal. In looking at the literature, the definition of hypernatremia as being above 145 millimoles of sodium per liter appears in many places. Evidently, a sodium level above 145 is very uncommon in healthy people who drink normal amounts of water, but it's hard to find out from the literature just how uncommon it is.

If it seems strange to model every case as the same, you're right. Typically you will be interested in relating one variable to others. To model height, depending on your purposes, you might consider explanatory variables such as age and sex as pediatricians do when monitoring a child's growth. Galton, who was interested in what is now called genetics, modeled child's height (when grown to adulthood) using sex and the parents' heights as explanatory variables. One can also imagine using economic status, nutrition, childhood illnesses, participation in sports, etc. as explanatory variables. The next chapter will introduce such models. At this point, though, the descriptions involve only a single variable — there are no other variables that might distinguish one case from another. So the only possible model is that all cases are the same.

Any given individual case will likely deviate from the single model value. The value of an individual can be written in a way that emphasizes the common

For review purposes only



Figure 3.3: The heights of nine brothers were recorded on a tintype photograph in the 1880s. The version here has been annotated to show the mean height and the individual deviations from the mean.

model value shared by all the cases and the deviation from that value of each individual:

$$\text{individual case} = \text{model value} + \text{deviation of that case.}$$

Writing things in this way partitions the description of the individual into two pieces. One piece reflects the information contained in the model. The other piece — the deviation of each individual — reflects how the individual is different from the model value.

As the single model value of the variable, it's reasonable to choose the mean. For Galton's height data, the mean is 66.7 inches. Another equally good choice would be the *median*, 66.5 inches. (Chapter 6 deals with the issue of how to choose the "best" model value, and how to define "best.")

Once you have chosen the model value, you can find how much each case deviates from the model. For instance, Figure 3.3 shows a group of nine brothers. Each brother's height differs somewhat from the model value; that difference is the deviation for that person.

In the more interesting models of later chapters, the model values will differ from case to case and so part of the variation among individual cases will be captured by the model. But here the model value is the same for all cases, so the deviations encompass all of the variation.

The word "deviation" has a negative connotation: a departure from an accepted norm or behavior. Indeed, in the mid-1800s, early in the history of statistics, it was widely believed that "normal" was best quantified as a single model value. The deviation from the model value was seen as a kind of mistake or imperfection. Another, related word from the early days of statistics is "error." Nowadays, though, people have a better understanding that a range of behaviors is normal; normal is not just a single value or behavior.

The word **residual** provides a more neutral term than "deviation" or "error" to describe how each individual differs from the model value. It refers to what is left over when the model value is taken away from the individual case. The word "deviation" survives in statistics in a technical terms such as **standard deviation**, and **deviance**. Similarly, "error" shows up in some technical terms such as **standard error**.

Height (inches)	Model Value (inches)	Residual (inches)	Square-Residual (inches <sup>2</sup> )
72.0	66.25	5.75	33.06
62.0	66.25	-4.25	17.06
68.0	66.25	1.75	3.05
65.5	66.25	-0.75	0.56
63.0	66.25	-3.25	10.56
64.7	66.25	-1.55	2.40
72.0	66.25	5.75	33.06
69.0	66.25	2.75	7.56
61.5	66.25	-4.75	22.56
72.0	66.25	5.75	33.06
59.0	66.25	-7.25	52.56
Sum of Squares: 216.5			

Table 3.2: Calculation of the sum of squares of the residuals for the subset of  $N = 11$  cases from Galton's data used in Table 3.1. (For this small subset, the mean height is 66.25 inches, somewhat different from the mean of 66.7 inches in the complete set.)

Any of the measures from Section 3.1 can be used to describe the variation in the residuals. The range of extremes is from  $-10.7$  inches to  $12.3$  inches. These numbers describe how much shorter the shortest person is from the model value and how much taller the tallest. Alternatively, you could use the 50% interval ( $-2.7$  to  $2.8$  inches) or the 95% interval ( $-6.7$  to  $6.5$  inches). Each of these is a valid description.

In practice, instead of the coverage intervals, a very simple, powerful, and perhaps unexpected measure is used: the **mean square** of the residuals. To produce this measure, add up the square of the residuals for the individual cases. This gives the **sum of squares** of the residuals as shown in Table 3.2. Such sums of squares will show up over and over again in statistical modeling.

It may seem strange to square the residuals. Squaring the residuals changes their units. For the height variable, the residuals are naturally in inches (or centimeters or meters or any other unit of length). The sum of squares, however is in units of square-inches, a bizarre unit for relating information about height. A good reason to compute the square is to emphasize the interest in *how far* each individual is from the mean. If you just added up the raw residuals, negative residuals (for those cases below the mean) would tend to cancel out positive residuals (for those cases above the mean). By squaring, all the negative residuals become positive square-residuals. It would also be reasonable to do this by taking an absolute value rather than squaring, but the operation of squaring captures in a special way important properties of the potential randomness of the residuals.

Finding the sum of squares is an intermediate step to calculating an important measure of variation: the **mean square**. The mean square is intended to report a typical square residual for each case. The obvious way to get this is to divide the sum of squares by the number of cases,  $N$ . This is not exactly what

is done by statisticians. For reasons that you will see later, they divide by  $N - 1$  instead. (When  $N$  is big, there is no practical difference between using  $N$  and  $N - 1$ .) For the subset of the Galton data in Table 3.2, the mean square residual is 21.65 square-inches.

Later in this book, mean squares will be applied to all sorts of models in various ways. But for the very simple situation here — the every-case-the-same model — the mean square has a special name that is universally used: the **variance**.

The variance provides a compact description of how far cases are, on average, from the mean. As such, it is a simple measure of variation.

It is undeniable that the unfamiliar units of the variance — squares of the natural units — make it hard to interpret. There is a simple cure, however: take the square root of the variance. This is called, infamously to many students of statistics, the **standard deviation**. For the Galton data in Table 3.2, the standard deviation is 4.65 inches (the square-root of 21.65 square-inches).

The term “standard deviation” might be better understood if “standard” were replaced by a more contemporary equivalent word: “typical.” The standard deviation is a measure of how far cases typically deviate from the mean of the group.

Historically, part of the appeal of using the standard deviation to describe variation comes from the ease of calculating it using arithmetic. Percentiles require more tedious sorting operations. Nowadays, computers make it easy to calculate 50% or 95% intervals directly from data. Still, standard deviations remain an important way of describing variation both because of historical convention and, more important for this book, because of its connection to concepts of modeling and randomness encountered in later chapters.

### 3.3 Displaying Variation

One of the best ways to depict variation is graphically. This takes advantage of the human capability to capture complicated patterns at a glance and to discern detail. There are several standard formats for presenting variation; each has its own advantages and disadvantages.

A simple but effective display plots out each case as a point on a number line. This **rug plot** is effectively a graphical listing of the values of the variable, case-by-case.

From the rug plot of Galton’s heights in Figure 3.4, you can see that the shortest person is about 56 inches tall, the tallest about 79 inches. you can also see that there is a greater density in the middle of the range. Unfortunately, the rug plot also suppresses some important features of the data. There is one tick at each height value for which there is a measurement, but that one tick might correspond to many individuals. Indeed, in Galton’s data, heights were rounded off to a quarter of an inch, so many different individuals have exactly the same recorded height. This is why there are regular gaps between the tick marks.

for review purposes only

DATA FILE  
galton.csv

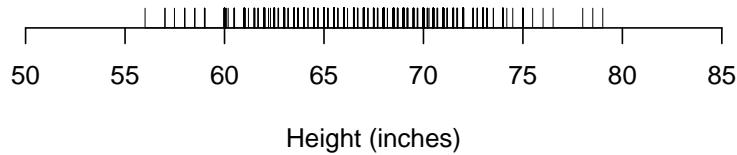


Figure 3.4: A rug plot of Galton's height data. Each case is one tick mark.

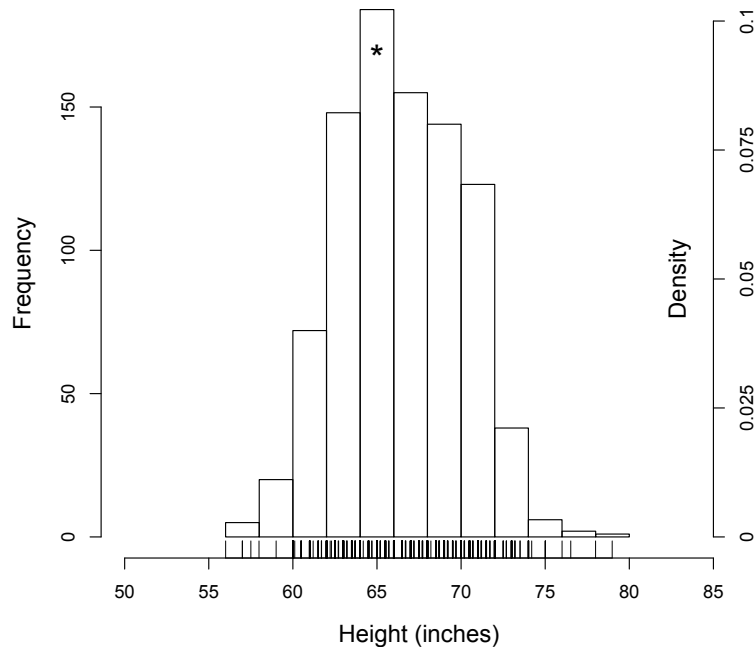


Figure 3.5: A histogram of Galton's height data. A rug plot is underneath. There are two axes: one ("Frequency") arranged so that the height of the bars reflects the number of cases that fall into the respective bins, the other ("Density" or "Relative Frequency") so that the area of the bars reflects the proportion of cases that fall into the bins.

Another simple display, the **histogram**, avoids this overlap by using two different axes. As in the rug plot, one of the axes shows the variable. The other axis displays the number of individual cases within set ranges, or "bins," of the variable.

There are two widely used ways to mark the vertical axis in a histogram. The simplest is called **frequency** (or **absolute frequency**) and is arranged so that each bar's height gives the direct count of individuals that fall into the cor-

responding bin. For example, consider the histogram shown in Figure 3.5 and, in particular, the bar marked with a \* that covers the bin from 64 to 66 inches. The bar's height on the frequency axis is 184, meaning that there are 184 individuals between 64 and 66 inches. Adding up the heights of all the bars gives the total number of cases being plotted in the histogram: 898 cases in this example.

The other format for the vertical axis of a histogram is called **density** (or **relative frequency**). In the density format, each bar's *area* shows the proportion of cases that fall into the bin. The density format makes it easy to compare histograms from different size samples.

It takes a bit of thought to understand the units of measuring density. In general, the word “density” refers to an amount per unit length, or area, or volume. For example, the mass-volume density of water is the mass of water per unit of volume, typically given as kilograms per liter. The sort of density used for histograms is the fraction-of-cases per unit of the binning variable.

To see how this plays out in Figure 3.5, consider again the bin for heights between 64 and 66 inches, marked in the figure with a \*. As shown by the frequency axis, there are 184 individual cases that fall into that bin. Since there are 898 cases altogether, the fraction of cases in the bin is  $\frac{184}{898} = 0.205$ . These cases are spread out over a range of 2 inches (64 to 66) in the variable. So, the density will be 0.205 per 2 inches, or 0.1025 per inch.

The units of a fraction density will always be the reciprocal of the units of the variable that is being studied. Since Galton's height data was measured in inches, the units of density are inches<sup>-1</sup>.

When calibrating a histogram in terms of density, it is the area of the bar, not the bar height, that gives the fraction of cases spanned by the bar. Adding together the areas of all the bar will always give an area of 1, since all of the cases fall into one bar or another.

### 3.3.1 Shapes of Distributions

The histogram of Galton's height data displays a bell-shaped pattern that is typical of many variables: the individuals are distributed with most near the center and fewer and fewer near the edges. The pattern is so common that a mathematical idealization of it is called the **normal distribution**.

The boundary points between bins in a histogram are somewhat arbitrary. It is not unheard of for writers to modify the divisions in order to make one bar higher; the careful reader should be alert to such shenanigans. In addition, the many divisions between bars make the plot somewhat busy graphically.

A modern alternative to the histogram is the density plot, an example of which is shown in Figure 3.6. This is very much like the histogram, but dispenses with the discrete bins. The resulting curve follows the same overall shape as the histogram, but is smoother. This is less distracting to the eye. The vertical axis of the density plot is always calibrated in density format and so the area under the curve is always 1.

Still another useful graphical format is a **box plot**, which shows the minimum, first quartile, median, third quartile, and maximum. Figure 3.7 shows a

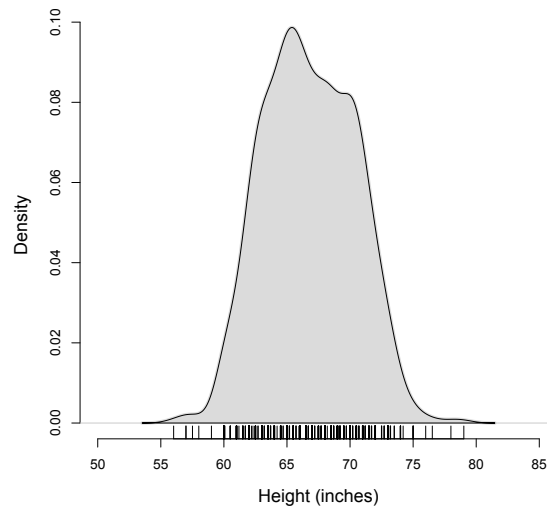


Figure 3.6: A density plot of Galton's height data.

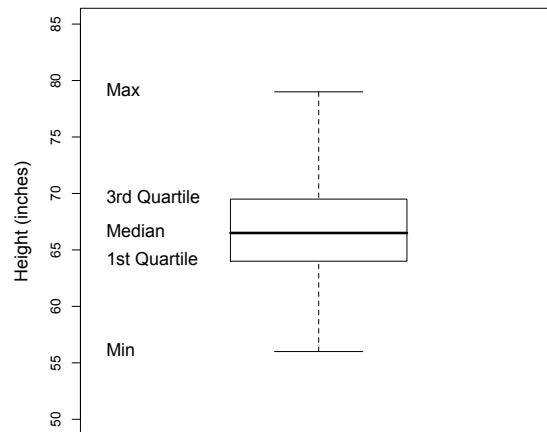


Figure 3.7: A box plot of Galton's height data.

box plot of the Galton height data. Note that the variable itself is being plotted on the vertical axis, contrasting with the other graphical depictions where the variable has been plotted on the horizontal axis.

The stylized display of the box plot is most useful when comparing the distribution of a variable across two or more different groups. For example, Figure



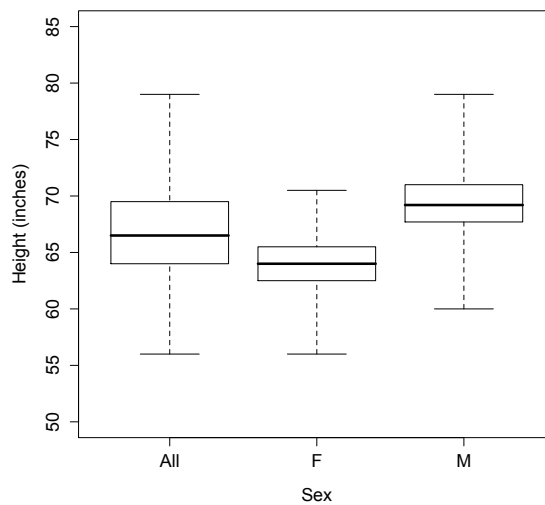


Figure 3.8: A box plot of height versus sex.

3.8 shows a box plot of the height data, breaking out separately the males and females. This plot is showing two variables; height is plotted as the response variable, with sex as an explanatory variable. The box plot shows that the variation among the males or among the females — as measured say by the interquartile interval or the min-to-max range — is less than the variation of the data when the person's sex is ignored. That is, some of the variation in the height variable is accounted for by the explanatory variable sex. Such partitioning of variation between the response and explanatory variables will be a major theme of modeling.

### 3.4 Normal and Non-normal Distributions

The symmetrical bell-shaped distribution seen, for example, in the Galton height data is so common that it is called “normal.” It's important to remember that despite the name “normal,” many variables display a very different pattern.

For example, consider a very mundane variable: the monthly natural gas utility bill for the author's house. The distribution (plotted in Figure 3.11) shows the main concentration of bills near about \$20 per month. But this is by no means typical. For many months — winter months — the bill is much higher. The overall shape is called **right skew** because the tail on the right side of the

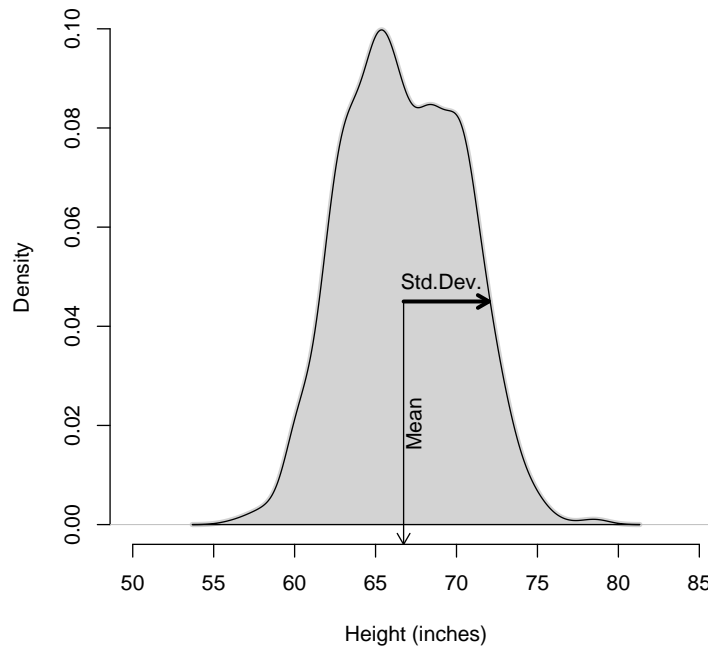


Figure 3.9: Eye-balling the mean and standard deviation from a density plot. The mean value is at the center of a bell-shaped distribution. The standard deviation is roughly the half-width at half-height.

peak is much longer than the tail on the left side.

There is nothing abnormal or strange about these data, even if the distribution does not have the so-called “normal” shape. In most months, a typical value for the gas bill is \$25 to \$50, but there are a few winter months where much, much more gas is consumed for heating.

Measurements such as the mean or standard deviation are most meaningful when the underlying data are normally distributed, or close to it. For strongly skewed data, or for data containing **outliers**, the median can offer a better indication than the mean of a typical value. For such data, rather than using the standard deviation to quantify spread, it may be preferable to use measures such as the **interquartile interval**, or **IQR** for short, which gives the length of the 50% coverage interval.

### 3.5 Categorical Variables

All of the measures of variation encountered so far are for quantitative variables; the kind of variable for which it’s meaningful to average or sort.

**Aside. 3.2** Back of the Envelope: Center and Spread

One way to interpret the mean and standard deviation of a variable is in terms of the **center** and **spread** of the distribution. This is most straightforward when the distribution of the variable is symmetric and roughly bell-shaped. In this case, the value of the mean falls at the center of the distribution, as shown in Figure 3.9.

The standard deviation measures the spread of the distribution. Arithmetically, the standard deviation is set by the square root of the variance which is, in turn, the mean square of the deviations from the mean. Graphically, there is a simple and useful approximation for a bell-shaped distribution: the standard deviation is roughly the half-width at half-height of the distribution as shown in Figure 3.9.

Both the mean and the standard deviation have units: the same as the units of the variable itself. When estimating the numerical values of the mean and standard deviation from the graph, use the scale of the variable shown on the horizontal axis.

Such “eye-ball” estimates are rough. The half-width-at-half-height method is merely a rule of thumb. For the Galton height data plotted in Figure 3.9, the rule overestimates the standard deviation: the actual standard deviation is about 3.6, not 5 as the figure suggests. There is no point in eye-balling an estimate when you can do the calculation. But it often happens that instead of having access to the actual data, you have only a graphical summary.

There is also a graphical interpretation of coverage intervals in terms of density plots. The 95% interval supports 95% of the area under the distribution as shown in Figure 3.10. Similarly, the 50%-interval supports half the area under the distribution.

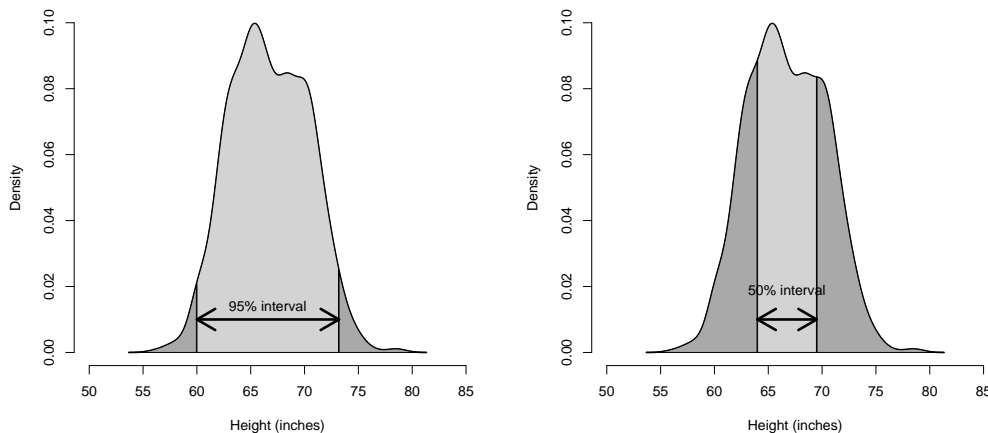


Figure 3.10: Coverage intervals shown on density plots. Left: The 95% coverage interval supports the central 95% of the area under the density curve; the dark-shaded area at the tails is only 5% of the total area. Right: The 50% coverage interval supports the central 50% of the area, with 25% of the area on each side.

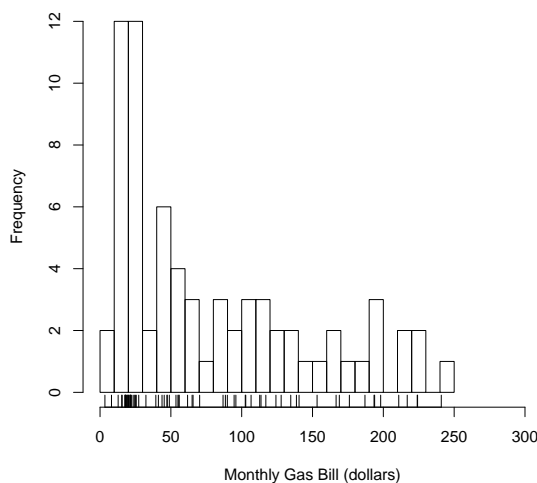
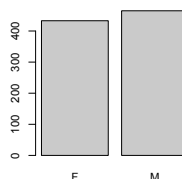


Figure 3.11: Histogram of the monthly gas bill from the utility data. The distribution is skewed to the right.

DATA FILE  
galton.csv

Categorical variables are a different story. One thing to do is to display the variation using tables and bar charts. For example, Galton’s height data has a categorical variable sex with levels F and M. Either a table or a chart are effective ways to show how many cases there are for each level.

Female	Male
433	465



Describing categorical variation quantitatively is a more difficult matter. Recall that for quantitative variables, one can define a “typical” value, e.g., the mean or the median. It makes sense to measure residuals as how far an individual value is from the typical value, and to quantify variation as the average size of a residual. This led to the variance and the standard deviation as measures of variation.

For a categorical variable like sex, concepts of mean or median or distance or size don’t apply. Neither F nor M is typical and in-between doesn’t really make sense. Even if you decided that, say, M is typical — there are somewhat more of them in Galton’s data — how can you say what the “distance” is between F and M?

Statistics textbooks hardly ever give a quantitative measure of variation in categorical variables. That’s understandable for the reasons described above. But it’s also a shame because many of the methods used in statistical modeling

For review purposes only

rely on quantifying variation in order to evaluate the quality of models.

There are quantitative notions of variation in categorical variables. You won't have use for them directly, but they are used in the software that fits models of categorical variables.

Solely for the purposes of illustration, here is one measure of variation in categorical variables with two levels, say F and M as in sex in Galton's data. Two-level variables are important in lots of settings, for example diagnostic models of whether or not a patient has cancer or models that predict whether a college applicant will be accepted.

Imagine that altogether you have  $N$  cases, with  $k$  being level F and  $N - k$  being level M. (In Galton's data,  $N$  is 898, while  $k = 433$  and  $N - k = 465$ .)

A simple measure of variation in two-level categorical variables is called, somewhat awkwardly, the **unlikeability**. This is

$$\text{unlikeability} = 2 \frac{k}{N} \frac{N - k}{N}.$$

Or, more simply, if you write the proportion of level F as  $p_F$ , and therefore the proportion of level M as  $1 - p_F$ ,

$$\text{unlikeability} = 2p_F(1 - p_F).$$

For instance, in the variable `sex`,  $p_F = \frac{433}{898} = 0.482$ . The unlikeability is therefore  $2 \times 0.482 \times .518 = 0.4993$ .

Some things to notice about unlikeability: If all of the cases are the same (e.g., all are F), then the unlikeability is zero — no variation at all. The highest possible level of unlikeability occurs when there are equal numbers in each level; this gives an unlikeability of 0.5.

Where does the unlikeability come from? One way to think about unlikeability is as a kind of numerical trick. Pretend that the level F is 1 and the level M is 0 — turn the categorical variable into a quantitative variable. With this transformation, `sex` looks like 0 1 1 1 0 0 1 1 0 1 0 and so on. Now you can quantify variation in the ordinary way, using the variance. It turns out that the unlikeability is exactly twice the variance.

Here's another way to think about unlikeability: although you can't really say *how far* F is from M, you can certainly see the difference. Pick two of the cases at random from your sample. The unlikeability is the probability that the two cases will be different: one an M and the other an F.

There is a much more profound interpretation of unlikeability that is introduced in Chapter 18, which covers modeling two-level categorical variables. For now, just keep in mind that you can measure variation for any kind of variable and use that measure to calculate how much of the variation is being captured by your models.

## 3.6 Computational Technique



To illustrate computer techniques for describing variability, consider the data that Galton collected on the heights of adult children and their parents. The file `galton.csv` stores these data in a modern, case/variable format.

```
> galton = ISMdata("galton.csv")
```

### 3.6.1 Simple Statistical Calculations

Simple numerical descriptions are easy to compute. Here are the mean, median, standard deviation and variance of the children's heights (in inches).

```
> mean( galton$height )
[1] 66.76069
> median( galton$height )
[1] 66.5
> sd( galton$height )
[1] 3.582918
> var( galton$height )
[1] 12.83730
```

Notice that the variance (`var`) is just the square of the standard deviation (`sd`). In principle, it's unnecessary to have both operators. Having both is merely a convenience.

A percentile tells where a given value falls in a distribution. For example, a height of 63 inches is on the short side in Galton's data:

```
> pdata( 63, galton$height )
[1] 0.1915367
```

Only about 19% of the cases have a height less than or equal to 63 inches. The `pdata` operator takes one or more values as a first argument and finds where they fall in the distribution of values in the second argument.

A quantile refers to the same sort of calculation, but inverted. Instead of giving a value in the same units as the distribution, you give a probability: a number between 0 and 1. The `qdata` operator then calculates the value whose percentile would be that value:

```
> qdata( .20, galton$height )
20%
63.5
```

Remember that the probability is given as a number between 0 and 1, so use 0.50 to indicate that you want the value which falls at the 50th percentile.

- The 25th and 75th percentile in a single command — in other words, the 50 percent coverage interval:

```
> qdata(c(0.25, 0.75), galton$height )
      25%   75%
64.0  69.7
```

- The 2.5th and 97.5th percentile — in other words, the 95 percent coverage interval:

```
> qdata(c(0.025, 0.975), galton$height )
      2.5%  97.5%
      60    73
```

The interquartile range is the width of the 50 percent coverage interval:

```
> IQR(galton$height)
[1] 5.7
```

Some other useful operators are `min`, `max`, and `range`.

For convenience, the summary operator gives a quick description of a quantitative variable:

```
> summary(galton$height)
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 56.00   64.00   66.50   66.76   69.70   79.00
```

In exercises in later chapters, you will need to compute the sum of squares of quantitative variables and of residuals. This is done by connecting simple computations: squaring then summing.

```
> sum( galton$height^2 )
[1] 4013892
```

In all of these examples, the operator has been applied directly to a variable in a data frame. Of course, any of these operators can be applied to any set of numbers. For example, here is the mean of the numbers 1, 2, 3,  $\dots$ , 100 created by `seq`:

```
> mean( seq(1,100))
[1] 50.5
```

### 3.6.2 Residuals and Sums of Squares

The residual from the mean can be computed like this:

```
> resids = galton$height - mean(galton$height)
```

Remember that each case has its own residual. There are 898 cases in the Galton data, so there are 898 residuals.

```
> resid
[1] 6.43931 2.43931 2.23931 2.23931
[5] 6.73931 5.73931 -1.26069 -1.26069
... and so on ...
[893] 1.93931 1.73931 0.93931 -2.76069
[897] -3.26069 -3.76069
```

The sum of squares is

```
> sum(resids^2)
[1] 11515.06
```

### 3.6.3 Simple Statistical Graphics

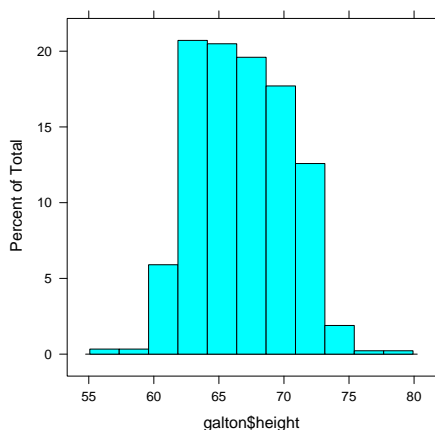
There are several basic types of statistical graphics to display the distribution of a variable: histograms, density plots, and boxplots. These are easily mastered by example.

Technical note: These graphics are implemented by the “lattice” package in R. This is loaded automatically with the `ISM.Rdata` workspace, so you don’t have to worry about it. But if you don’t use the `ISM.Rdata` workspace, you will need to load `lattice` manually with the command `library(lattice)`.

#### Histograms

Constructing a histogram involves dividing the range of a variable up into bins and counting how many cases fall into each bin. This is done in an almost entirely automatic way:

```
> histogram( galton$height )
```

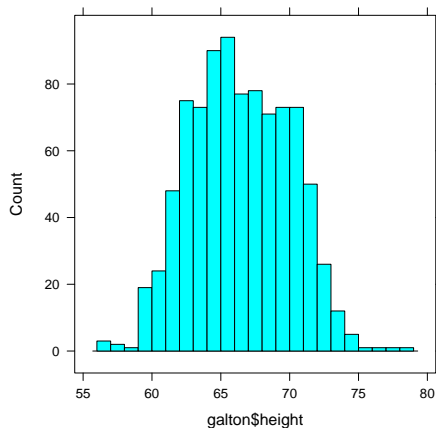


When constructing a histogram, R makes an automatic but sensible choice of the number of bins. If you like, you can control this yourself. For instance:

```
> histogram( galton$height, breaks=25 )
```

For review purposes only





The horizontal axis of the histogram is always in the units of the variable. For the histograms above, the horizontal axis is in “inches” because that is the unit of the `galton$height` variable.

The vertical axis is conventionally drawn in one of three ways: controlled by an optional argument named `type`.

**Absolute Frequency or Counts** A simple count of the number of cases that falls into each bin. This mode is set with `type="count"` as in

```
> histogram( galton$height, type="count")
```

**Relative Frequency** The vertical axis is scaled so that the height of the bar give the proportion of cases that fall into the bin. This is the default.

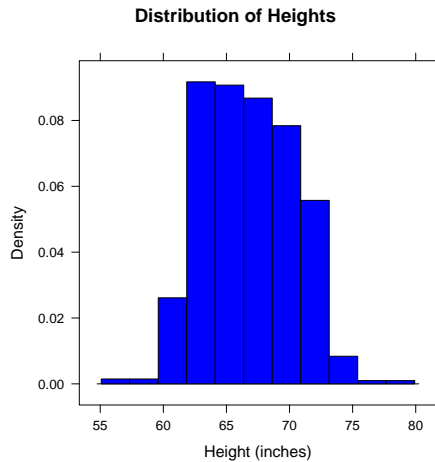
**Density** The vertical axis *area* of the bar gives the relative proportion of cases that fall into the bin. Set `type="density"` as in `histogram(galton$height, type="density")`.

In a density plot, areas can be interpreted as probabilities and the area under the entire histogram is equal to 1.

Other useful optional arguments set the labels for the axes and the graph as a whole and color the bars. For example,

```
> histogram(galton$height, type="density",
  xlab="Height (inches)",
  main="Distribution of Heights",
  col="blue")
```

For review purposes only

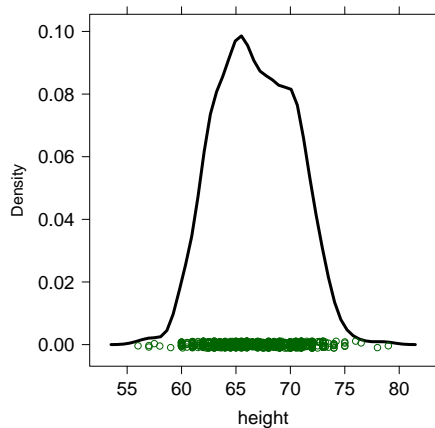


The above command is so long that it has been broken into several lines for display purposes. R ignores the line breaks, holding off on executing the command until it sees the final closing parentheses. Notice the use of quotation marks to delimit the labels and names like "blue".

### Density Plots

A **density plot** avoids the need to create bins and plots out the distribution as a continuous curve. Making a density plot involves two operators. The **density** operator performs the basic computation which is then displayed using either the plot or the lines operator. For example:

```
> densityplot( galton$height )
```



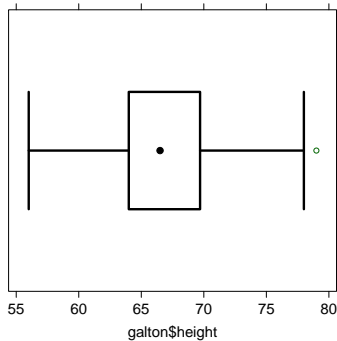
If you want to suppress the rug-like plotting of points at the bottom of the graph, use `densityplot(galton$height, plot.points=FALSE)`.

For review purposes only

**Box-and-Whisker Plots**

Box-and-whisker plots are made with the `bwplot` command:

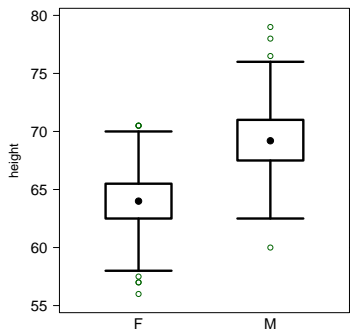
```
> bwplot( galton$height )
```



The median is represented by the heavy dot in the middle. Outliers, if any, are marked by dots outside the whiskers.

The real power of the box-and-whisker plot is for comparing distributions. This will be raised again more systematically in later chapters, but just to illustrate, here is how to compare the heights of males and females:

```
> bwplot( height ~ sex, data=galton )
```

**3.6.4 Displays of Categorical Variables**

For categorical variables, it makes no sense to compute descriptive statistics such as the mean, standard deviation, or variance. Instead, look at the number of cases at each level of the variable.

```
> table( galton$sex )
```

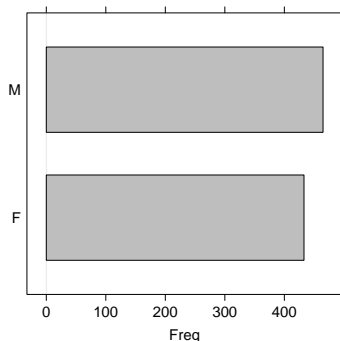
```
  F  M
433 465
```

By processing such a table with the `prop.table` operator, you can calculate the proportion of cases at each level.

```
> prop.table( table( galton$sex ) )
      F      M
0.4822 0.5178
```

The barchart operator will produce graphics from tables.

```
> barchart( table(galton$sex) )
```



### 3.6.5 Outliers

The `outlier` operator uses the same rule of thumb as in `bwplot` to identify which cases are outliers. It returns a logical variable, `TRUE` or `FALSE` for each case.

```
> outlier(galton$height)
[1] FALSE FALSE FALSE FALSE FALSE
[6] FALSE FALSE FALSE FALSE FALSE
... and so on.
```

The direct output of `outlier` is rarely what you want. Typically you will want to use `outlier` to help in counting how many outliers there are or to look at the outlier cases themselves, or to extract cases that are not outliers:

```
> table( outlier( galton$height ) )
FALSE TRUE
 897    1
> subset( galton, outlier(galton$height) )
  family father mother sex height nkids
289    72    70    65  M    79    7
> cleaned = subset( galton, !outlier(galton$height) )
```

There was just one outlier (according to the rule of thumb). The object named `cleaned` contains those cases that were not outliers with respect to height. Other cases might be outliers with respect to mother or father — the `outlier` program looks at only one variable. (The `!` is the logical operator meaning “not,” so `!outlier(galton$height)` refers to the cases that are not outliers with respect to height.) That it’s easy to remove outliers does not mean that you should do so without careful thought.

# Chapter 4

## The Language of Models

*I do not believe in things. I believe only in their relationships.* — Georges Braque (1882-1963, cubist painter)

*Mathematicians do not study objects, but relations among objects.* — Henri Poincaré (1854-1912, mathematician)

One October I received a bill from the utility company: \$52 for natural gas for the month. According to the bill, the average outdoor temperature during October was 49°F (9.5°C).

I had particular interest in this bill because two months earlier, in August, I replaced the old furnace in our house with a new, expensive, high-efficiency furnace that's supposed to be saving gas and money. This bill is the first one of the heating season and I want to know whether the new furnace is working.

To be able to answer such questions, I keep track of my monthly utility bills. The gasbill variables shows a lot of variation: from \$3.42 to \$240.90 per month. The 50% coverage interval is \$21.40 to \$117.90. The mean bill is \$77.85. Judging from this, the \$52 October bill is low. It looks like the new furnace is working!

Perhaps that conclusion is premature. My new furnace is just one factor that can influence the gas bill. Some others: the weather, the price of natural gas (which fluctuates strongly from season to season and year to year), the thermostat setting, whether there were windows or doors left open, how much gas we use for cooking, etc. Of all the factors, I think the weather and the price of natural gas are the most important. I'd like to take these into account when judging whether the \$52 bill is high or low.

Rather than looking at the bill in dollars, it makes sense to look at the quantity of gas used. After all, the new furnace can only reduce the amount of gas used, it doesn't influence the price of gas. According to the utility bill, the usage in October was 65 ccf. ("ccf" means cubic feet, one way to measure the quantity of gas.) The variable ccf gives the historical values from past bills: the range is 0

For review purposes only

DATA FILE  
utilities.csv

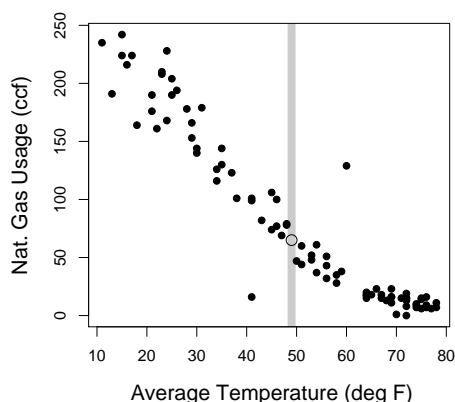


Figure 4.1: Monthly natural gas usage versus average outdoor temperature during the month.

to 242 ccf, so 65 ccf seems perfectly reasonable, but it's higher than the median monthly usage, which is 51 ccf. Still, this doesn't take into account the weather.

Now it is time to build a model: a representation of the utility data for the purpose of telling whether 65 ccf is low or high given the temperature. The variable temperature contains the average temperature during the billing period.

A simple graph of ccf versus temperature will suffice. (See Figure 4.1.) The open point shows the October bill (49 deg. and 65 ccf). The gray line indicates which points to look at for comparison, those from months near 49 degrees.

The graph suggests that 65 ccf is more or less what to expect for a month where the average temperature is 49 degrees. The new furnace doesn't seem to make much of a difference.

## 4.1 Models as Functions

Figure 4.1 gives a pretty good idea of the relationship between gas usage and temperature.

The concept of a **function** is very important when thinking about relationships. A function is a mathematical concept: the relationship between an **output** and one or more **inputs**. One way to talk about a function is that you plug in the inputs and receive back the output. For example, the formula  $y = 3x + 7$  can be read as a function with input  $x$  and output  $y$ . Plug in a value of  $x$  and receive the output  $y$ . So, when  $x$  is 5, the output  $y$  is 22.

One way to represent a function is with a formula, but there are other ways as well, for example graphs and tables. Figure 4.2 shows a function representing the relationship between gas usage and temperature. The function is much simpler than the data. In the data, there is a scatter of usage levels at each tem-

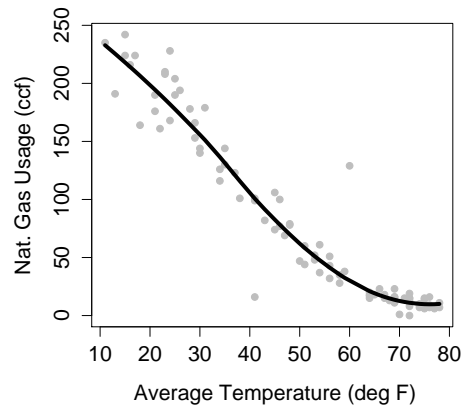


Figure 4.2: A model of natural gas usage versus outdoor temperature.

perature. But in the function there is only one output value for each input value.

Some vocabulary will help to describe how to represent relationships with functions.

- The **response variable** is the variable whose behavior or variation you are trying to understand. On a graph, the response variable is conventionally plotted on the vertical axis.
- The **explanatory variables** are the other variables that you want to use to explain the variation in the response. Figure 4.1, shows just one explanatory variable, temperature. It's plotted on the horizontal axis.
- **Conditioning on explanatory variables** means taking the value of the explanatory variables into account when looking at the response variables. When in Figure 4.1 you looked at the gas usage for those months with a temperature near  $49^\circ$ , you were conditioning gas usage on temperature.
- The **model value** is the output of a function. The function — called the **model function** — has been arranged to take the explanatory variables as inputs and return as output a typical value of the response variable. That is, the model function gives the typical value of the response variable *conditioning on* the explanatory variables. The function shown in Figure 4.2 is a model function. It gives the typical value of gas usage conditioned on the temperature. For instance, at  $49^\circ$ , the typical usage is 65 ccf. At  $20^\circ$ , the typical usage is much higher, about 200 ccf.
- The **residuals** show how far each case is from its model value. For example, one of the cases plotted in Figure 4.2 is a month where the temperature was  $13^\circ$  and the gas usage was 191 ccf. When the input is  $13^\circ$ , the model

function gives an output of 228 ccf. So, for that case, the residual is  $191 - 228 = -37$  ccf. Residuals are always “actual value minus model value.”

Graphically, the residual for each case tells how far above the model function that case is. A negative residual means that the case is below the model function.

The idea of a function is fundamental to understanding statistical models. Whether the function is represented by a formula or a graph, the function takes one or more inputs and produces an output. In the statistical models in this book, that output is the model value, a “typical” or “ideal” value of the response variable at given levels of the inputs. The inputs are the values explanatory variables.

The model function describes how the typical value of the response variable depends on the explanatory variables. The output of the model function varies along with the explanatory variables. For instance, when temperature is low, the model value of gas usage is high. When temperature is high, the model value of gas usage is low. The idea of “depends on” is very important. In some fields such as economics, the term **dependent variable** is used instead of “response variable.” Other phrases are used for this notion of “depends on,” so you may hear statements such as these: “the value of the response *given* the explanatory variables,” or “the value of the response *conditioned on* the explanatory variables.”

The model function describes a relationship. If you plug in values for the explanatory variables for a given case, you get the model value for that case. The model value is usually different from one case to another, at least so long as the values of the explanatory variables are different. When two cases have exactly the same values of the explanatory values, they will have exactly the same model value even though the actual response value might be different for the two cases.

The residuals tell how each case differs from its model value. Both the model values and the residuals are important. The model values tell what’s typical or average. The residuals tell how far from typical an individual case is likely to be. This might remind you of the mean and standard deviation.

As already said, models partition the variation in the response variable. Some of the variability is explained by the model, the remainder is unexplained. The model values capture the “deterministic” or “explained” part of the variability of the response variable from case to case. The residuals represent the “random” or “unexplained” part of the variability of the response variable.

## 4.2 Model Functions with Multiple Explanatory Variables

Historically, women tended to be paid less than men. To some extent, this reflected the division of jobs along sex lines and limited range of jobs that were open to women — secretarial, nursing, school teaching, etc. But often there was

For review purposes only



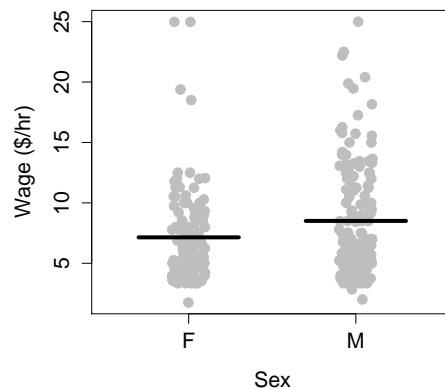


Figure 4.3: Hourly wages versus sex from the Current Population Survey data of 1985.

simple discrimination; an attitude that women’s work wasn’t as valuable or that women shouldn’t be in the workplace. Over the past thirty or forty years, the situation is changing. Training and jobs that were once rarely available to women — police work, management, medicine, law, science — are now open to them.

Surveys consistently show that women tend to earn less than men, a “wage gap.” To illustrate, consider data from one such survey, the Current Population Survey (CPS) from 1985. In the survey data, each case is one person. The variables are the person’s hourly wages at the time of the survey, age, sex, marital status, the sector of the economy in which they work, etc.

One aspect of these data is displayed by plotting wage versus sex, as in Figure 4.3. The model plotted along with the data show that typical wages for men are higher than for women.

The situation is somewhat complex since the workforce reflected in the 1985 data is a mixture of people who were raised in the older system and those who were emerging in a more modern system. A woman’s situation can depend strongly on when she was born. This is reflected in the data by the age variable.

There are other factors as well. The roles and burdens of women in family life remain much more traditional than their roles in the economy. Perhaps marital status ought to be taken into account. In fact, there are all sorts of variables that you might want to include — job type, race, location, etc.

A statistical model can include multiple explanatory variables, all at once. To illustrate, consider explaining wage using the worker’s age, sex, and marital status.

In a typical graph of data, the vertical axis stands for the response variable and the horizontal axis for the explanatory variable. But what do you do when there is more than one explanatory variable? One approach, when some of the

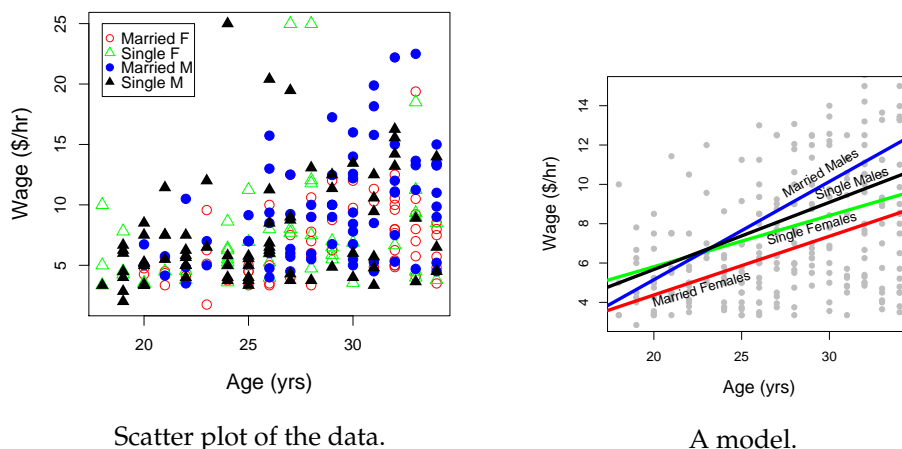


Figure 4.4: Hourly wages modeled by age, sex, and marital status.

explanatory variables are categorical, is to use differing symbols or colors to represent the differing levels of the categories. The left panel in Figure 4.4 shows wages versus age, sex, and marital status plotted out this way.

The first thing that might be evident from the scatter plot is that not very much is obvious from the data on their own. There does seem to be a slight increase in wages with age, but the cases are scattered all over the place.

The model, shown in the right panel of the figure, simplifies things. The relationships shown in the model are much clearer. You can see that wages tend to increase with age, and they do so differently for men and for women and differently for married people and single people.

Models can sometimes reveal patterns that are not evident in a graph of the data. This is a great advantage of modeling over simple visual inspection of data. There is a real risk, however, that a model is imposing structure that is not really there on the scatter of data, just as people imagine animal shapes in the stars. A skeptical approach is always warranted. Much of the second half of the book is about ways to judge whether the structure suggested by a model is justified by the data.

There is a tendency for those who first encounter models to fix attention on the clarity of the model and ignore the variation around the model. This is a mistake. Keep in mind the definition of statistics offered in the first chapter:

*Statistics is the explanation of variation in the context of what remains unexplained.*

The scatter of the wage data around the model is very large; this is a very important part of the story. The scatter suggests that there might be other factors that account for large parts of person-to-person variability in wages, or perhaps just that randomness plays a big role.

Compare the broad scatter of wages to the rather tight way the gas usage in Figure 4.1 is modeled by average temperature. The wage model is explaining only a small part of the variation in wages, the gas-usage model explains a very large part of the variation in those data.

Adding more explanatory variables to a model can sometimes usefully reduce the size of the scatter around the model. If you included the worker's level of education, job classification, age, or years working in their present occupation, the unexplained scatter might be reduced. Even when there is a good explanation for the scatter, if the explanatory variables behind this explanation are not included in the model, the scatter due to them will appear as unexplained variation. (In the gas-usage data, it's likely that wind velocity, electricity usage that supplements gas-generated heat, and the amount of gas used for cooking and hot water would explain a lot of the scatter. But these potential explanatory variables were not measured.)

### 4.3 Reading a Model

There are two distinct ways that you can read a model.

**Read out the model value.** Plug in specific values for the explanatory variables and read out the resulting model value. For the model in Figure 4.1, an input temperature of  $35^{\circ}$  produces an output gas usage of 125 ccf. For the model in Figure 4.4 a single, 30-year old female has a model value of \$8.10 per hour. (Remember, the model is based on data from 1985!)

**Characterize the relationship described by the model.** In contrast to reading out a model value for some specific values of the explanatory variables, here interest is in the overall relationship: how gas usage depends on temperature; how wages depend on sex or marital status or age.

Reading out the model value is useful when you want to make a prediction (What would the gas usage be if the temperature were  $10^{\circ}$ ?) or when you want to compare the actual value of the response variable to what the model says is a typical value. (Is the gas usage in the  $49^{\circ}$  month lower than expected, perhaps due to my new furnace?).

Characterizing the relationship is useful when you want to make statements about broad patterns that go beyond individual cases. Is there really a connection between marital status and wage? Which way does it go?

The "shape" of the model function tells you about such broad relationships. Reading the shape from a graph of the model is not difficult.

For a quantitative explanatory variable, e.g., temperature or age, the model form is a continuous curve or line. An extremely important aspect of this curve is its slope. For the model of gas usage in Figure 4.2, the slope is down to the right: a negative slope. This means that as temperature increases, the gas usage goes down. In contrast, for the model of wages in Figure 4.4, the slope is up to the right: a positive slope. This means that as age increases, the wage goes up.

The slope is measured in the usual way: rise over run. The numerical size of the slope is a measure of the strength of the relationship, the sign tells which way the relationship goes. Some examples: For the gas usage model in Figure 4.2 in winter-like temperatures, the slope is about  $-4$  ccf/degree. This means that gas usage can be expected to go down by 4 ccf for every degree of temperature increase. For the model of wages in Figure 4.4, the slope for single females is about 0.20 dollars-per-hour/year: for every year older a single female is, wages typically go up by 20 cents-per-hour.

Slopes have units. These are always the units of the response variable divided by the units of the explanatory variable. In the wage model, the response has units of dollars-per-hour while the explanatory variable age has units of years. Thus the units of the slope are dollars-per-hour/year.

For categorical variables, slopes don't apply. Instead, the pattern can be described in terms of *differences*. In the model where wage is explained only by sex, the difference between typical wages for males and females is 2.12 dollars per hour.

When there is more than one explanatory variable, there will be a distinct slope or difference associated with each.

When describing models, the words used can carry implications that go beyond what is justified by the model itself. Saying "the difference between typical wages" is pretty neutral: a description of the pattern. But consider this statement: "Typical wages go up by 20 cents per hour for every year of age." There's an implication here of **causation**, that as a person ages his or her wage will go up. That might in fact be true, but the data on which the model is based were not collected in a way to support that claim. Those data don't trace people as they age; they are not longitudinal data. Instead, the data are a snapshot of different people at different ages: cross-sectional data. It's dangerous to draw conclusions about changes over time from cross-sectional data of the sort in the CPS data set. Perhaps people's wages stay the same over time but that the people who were hired a long time ago tended to start at higher wages than the people who have just been hired.

Consider this statement: "A man's wage rises when he gets married." The model in Figure 4.4 is consistent with this statement; it shows that a married man's typical wage is higher than an unmarried man's typical wage. But does marriage cause a higher wage? It's possible that this is true, but that conclusion isn't justified from the data. There are other possible explanations for the link between marital status and wage. Perhaps the men who earn higher wages are more attractive candidates for marriage. It might not be that marriage causes higher wages but that higher wages cause marriage.

To draw conclusions about causation, it's important to collect data in an appropriate way. For instance, if you are interested in the effect of marriage on wages, you might want to collect data from individuals both before and after marriage and compare their change in wages to that over the same time period in individuals who don't marry. The strongest statements about causation require something more: that the condition be imposed experimentally, picking the people who are to get married at random. Such an experiment is hardly

possible when it comes to marriage.

## 4.4 Choices in Model Design

The suitability of a model for its intended purpose depends on choices that the modeler makes. There are three fundamental choices:

1. The data.
2. The response variable.
3. The explanatory variables.

### 4.4.1 The Data

How were the data collected? Are they a random sample from a relevant sampling frame? Are they part of an experiment in which one or more variables were intentionally manipulated by the experimenter, or are they observational data? Are the relevant variables being measured? (This includes those that may not be directly of interest but which have a strong influence on the response.) Are the variables being measured in a meaningful way?

Unfortunately, work on building models often starts only after the data have been collected. Consulting statisticians often have a researcher approach them with a data set and ask for help. Regrettably, the answer is often, “It’s too late. You needed to talk to me **before** you collected your data.” Start thinking about your models while you are still planning your data collection.

When you are confronted with a situation where your data are not suitable, you need to be honest and realistic about the limitations of the conclusions you can draw. The issues involved will be discussed starting in Chapter 8.

### 4.4.2 The Response Variable

The appropriate choice of a response variable for a model is often obvious. The response variable should be the thing that you want to predict, or the thing whose variability you want to understand. Often, it is something that you think is the effect produced by some other cause.

For example, in examining the relationship between gas usage and outdoor temperature, it seems clear that gas usage should be the response: temperature is a major determinant of gas usage. But suppose that the modeler wanted to be able to measure outdoor temperature from the amount of gas used. Then it would make sense to take temperature as the response variable.

Similarly, wages make sense as a response variable when you are interested in how wages vary from person to person depending on traits such as age, experience, and so on. But suppose that a sociologist was interested in assessing the influence of income on personal choices such as marriage. Then the marital status might be a suitable response variable, and wage would be an explanatory variable.

For review purposes only

Most of the modeling techniques in this book require that the response variable be quantitative. The main reason is that there are straightforward ways to measure variation in a quantitative variable and measuring variation is key to assessing the reliability of models. There are, however, methods for building models with a categorical response variable. (One of them, logistic regression, is the subject of Chapter 18.)

### 4.4.3 The Explanatory Variables

Choices of explanatory variables are richer and much more nuanced. Much of this book concerns ways to tell if an explanatory variable ought to be included in a model.

That said, some of the things that shape the choice of explanatory variables are obvious. Do you want to study sex-related differences in wage? Then sex had better be an explanatory variable. Is temperature a major determinant of the usage of natural gas? Then it makes sense to include it as an explanatory variable.

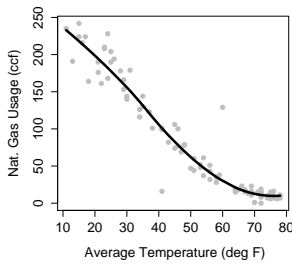
You will see situations where including an explanatory variable hurts the model, so it is important to be careful. (This will be discussed in Chapter 14.) A much more common mistake is to leave out explanatory variables. Unfortunately, few people learn the techniques for handling multiple explanatory variables and so your task will often need to go beyond modeling to include explaining how this is done.

When designing a model, you should think hard about what are potential explanatory variables and be prepared to include them in a model along with the variables that are of direct interest.

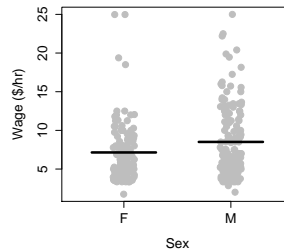
## 4.5 Model Terms

Once the modeler has selected explanatory variables, a choice must be made about **model terms**.

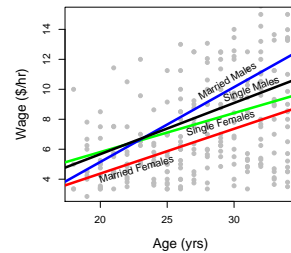
Notice that the models in the preceding examples have graphs of different shapes. The gas-usage model is a gentle curve, the wage-vs-sex model is just two values, and the more elaborate wage model is four lines with different slopes and intercepts.



Gas usage vs  
temperature



Wages vs sex



Wages vs age, sex, and  
marital status.

The modeler determines the shape of the model through his or her choice of **model terms**. The basic idea of a model term is that explanatory variables can be included in a model in more than one way. Each kind of term describes a different way to include a variable in the model.

You need to learn to describe models using model terms for several reasons. First, you will communicate in this language with the computers that you will use to perform the calculations for models. Second, when there is more than one or two explanatory variables, it's hard to visualize the model function with a graph: knowing the language of model terms will help you “see” the shape of the function even when you can't graph it. Third, model terms are the way to talk about “parts” of models. In evaluating a model, statistical methods can be used to take the model apart and describe the contribution of each part. This analysis — the word “analysis” literally means to loosen apart — helps the modeler to decide which parts are worth keeping.

There are just a few basic kinds of models terms. They are:

1. the intercept term
2. main terms
3. interaction terms
4. transformation terms

Models almost always include the intercept term and a main term for each of the explanatory variables. (There are rare exceptions.) Transformation and interaction terms can be added to create more expressive or flexible shapes.

To form an understanding of how different kinds of terms contribute to the overall “shape” of the model function, it's best to look at the different shape functions that result from including different model terms. The next section illustrates this. Several differently shaped models are constructed of the same data plotted in Figure 4.5. The data are the record time (in seconds) for the 100 meter freestyle race along with the year in which the record was set and the sex of the swimmer. The response variable will be time, the explanatory variables will be year and sex.

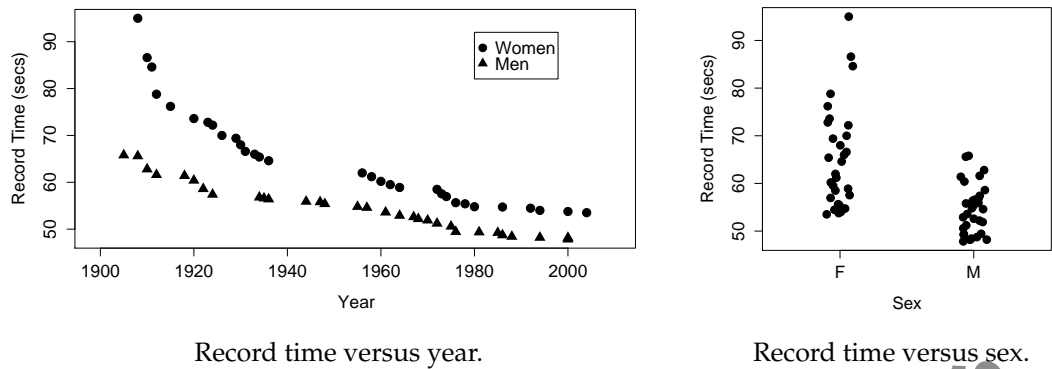


Figure 4.5: World record swimming times in the 100 meter freestyle.

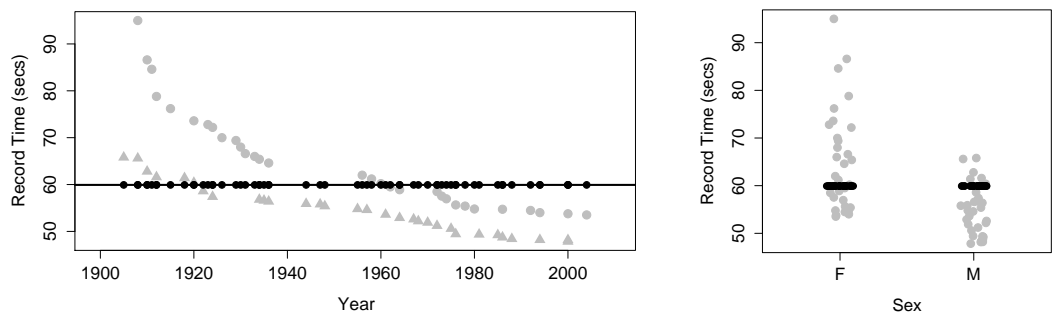
Figure 4.5 shows some obvious patterns, seen most clearly in the plot of time versus year. The record time has been going down over the years. This is natural, since setting a new record means beating the time of the previous record. There's also a clear difference between the men's and women's records; men's record times are faster than women's, although the difference has decreased markedly over the years.

The following models may or may not reflect these patterns, depending on which model terms are included.

#### 4.5.1 The Intercept Term (and no other terms)

The **intercept term** is included in almost every statistical model. The intercept term is a bit strange because it isn't something you measure; it isn't a variable. (The term "intercept" will make sense when model formulas are introduced in the next chapter.)

The figure below shows the swimming data with a simple model consisting only of the intercept term.



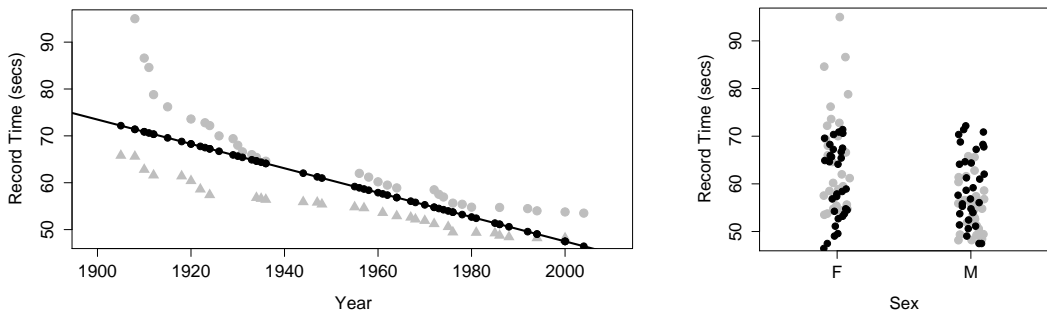


to create model variation from case to case, you would need to include at least one explanatory variable in the model.

### 4.5.2 Intercept and Main Terms

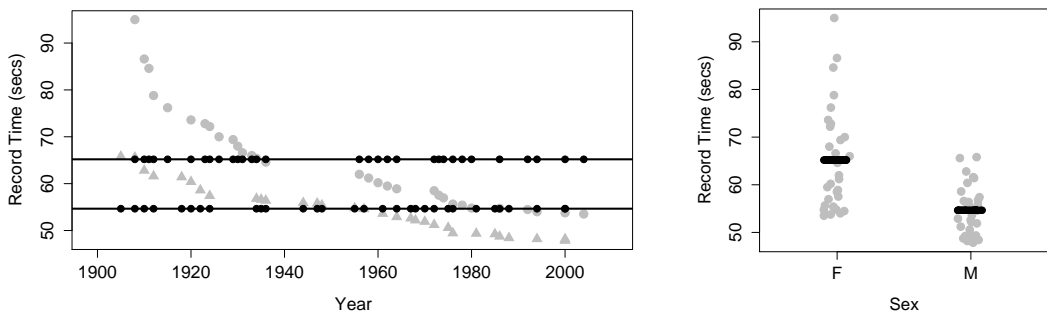
The most basic and common way to include an explanatory variable is as a main effect. Almost all models include the intercept term and a main term for each of the explanatory variables. The figures below show three different models each of this form:

**The intercept and a main term from year.** This produces model values that vary with year, but show no difference between the sexes. This is because sex has not been included in the model.



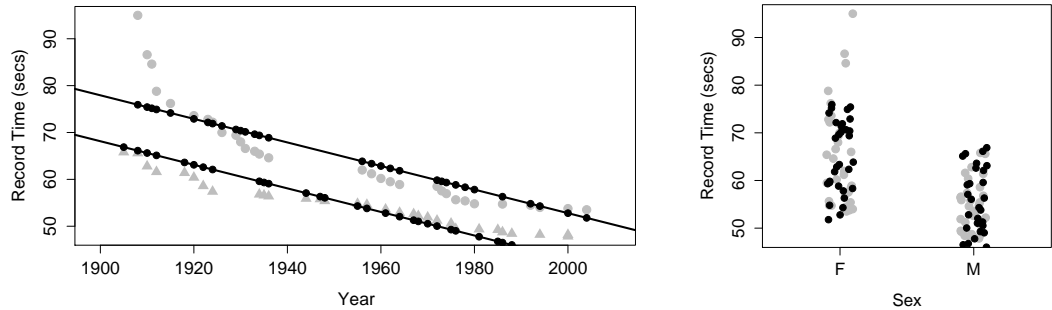
The model values have been plotted out as small black dots. The model pattern is evident in the left graph: swim time versus year. But in the right graph — swim time versus sex — it seems to be all scrambled. Don't be confused by this. The right-hand graph doesn't include year as a variable, so the dependence of the model values on year is not at all evident from that graph. Still, each of the model value dots in the left graph occurs in the right graph at exactly the same *vertical* coordinate.

**The intercept and a main term from sex.** This produces different model values for each level of sex.



There is no model variation with year because year has not been included in the model.

**The intercept and main terms from sex and from year.** This model gives dependence on both sex and year.

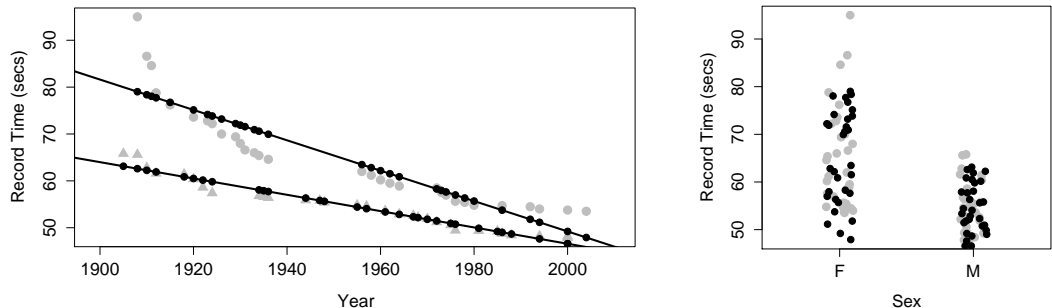


Note that the model values form two parallel lines in the graph of time versus year: one line for each sex.

### 4.5.3 Interaction Terms

Interaction terms combine two other terms, typically two main terms. An interaction term can describe how one explanatory variable modulates the role of another explanatory variable in modeling the relationship of both with the response variable.

In the graph, including the interaction term between sex and year produces a model with two non-parallel lines for time versus year. (The model also has the main terms for both sex and year and the intercept term, as always.)



One way to think about the meaning of the interaction term in this model is that it describes how the effect of sex changes with year. Looking at the model, you can see how the difference between the sexes changes over the years; the difference is getting smaller. Without the interaction term, the model values would be two *parallel* lines; the difference between the lines wouldn't be able to change over the years.

Another, equivalent way to put things is that the interaction term describes how the effect of year changes with sex. The effect of year on the response is reflected by the slope of the model line. Looking at the model, you can see that the slope is different depending on sex: steeper for women than men.

For most people, it's surprising that one term — the interaction between sex and year — can describe both how the effect of year is modulated by sex, and how the effect of sex is modulated by year. But these are just two ways of looking at the same thing.

---

**Aside. 4.1** Interaction terms and partial derivatives

---

The mathematically oriented reader may recall that one way to describe the effect of one variable on another is a partial derivative: the derivative of the response variable with respect to the explanatory variable. The interaction — how one explanatory variable modulates the effect of another on the response variable — corresponds to a mixed second-order partial derivative. Writing the response as  $z$  and the explanatory variables as  $x$  and  $y$ , the interaction corresponds to  $\frac{\partial^2 z}{\partial x \partial y}$  which is exactly equal to  $\frac{\partial^2 z}{\partial y \partial x}$ . That is, the way that  $x$  modulates the effect of  $y$  on  $z$  is the same thing as the way that  $y$  modulates the effect of  $x$  on  $z$ .

---

A common misconception about interaction terms is that they describe how one explanatory variable affects another explanatory variable. Don't fall into this error. Model terms are always about how the response variable depends on the explanatory variables, not how explanatory variables depend on one another. An interaction term between two variables describes how two explanatory variables combine jointly to influence the response variable.

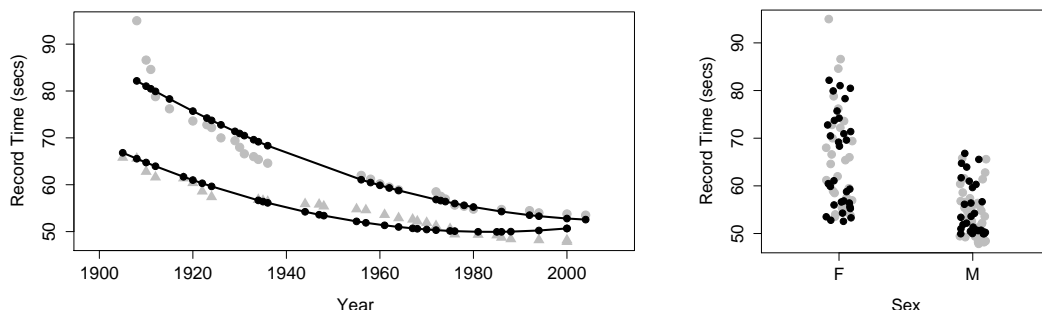
Once people learn about interaction terms, they are tempted to include them everywhere. After all, it's natural to think that world record swimming times would depend differently on year for women than for men. Of course wages might depend differently on age for men and women! Regretably, the uncritical use of interaction terms can lead to poor models. The problem is not the logic of interaction, the problem is in the data. As you will see in Chapter 11, interaction terms can introduce a problem called **multi-collinearity** which can reduce the reliability of models. Fortunately, it's not hard to detect multi-collinearity and to drop interaction terms if they are causing a problem. The model diagnostics that will be introduced in later chapters will make it possible to play safely with interaction terms.

#### 4.5.4 Transformation Terms

A **transformation term** is a modification of another term using some mathematical transformation. Transformation terms only apply to quantitative variables. Some common transformations are  $x^2$  or  $\sqrt{x}$  or  $\log x$ , where the quantitative explanatory variable is  $x$ .

A transformation term allows the model to have a dependence on  $x$  that is not a straight line. The graph shows a model that includes these terms: an

intercept, main effects of sex and year, an interaction between sex and year, and a year-squared transformation term.



Adding in the year-squared term provides some curvature to the model function.

---

**Aside. 4.2** Are swimmers slowing down?

---

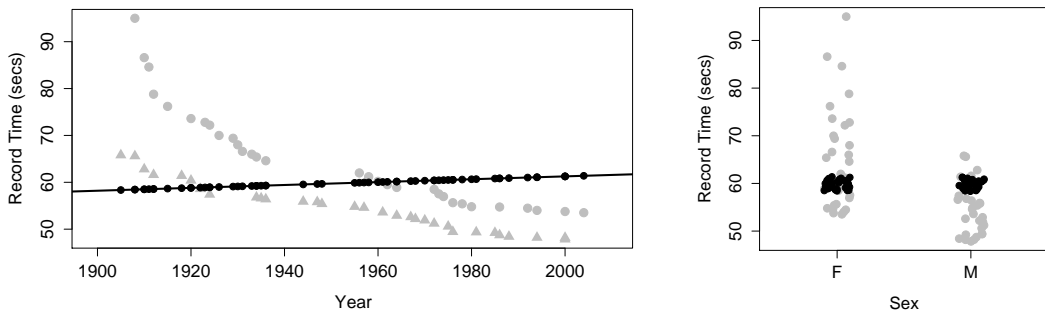
Look carefully at the model with a year-squared transformation term. You may notice that, according to the model, world record times for men have been getting worse since about year 1990. This is, of course, nonsense. Records can't get worse. A new record is set only when an old record is beaten. The model doesn't know this common sense about records — the model terms allow the model to curve in a certain way and the model curves in exactly that way. What you probably want out of a model of world records is a slight curve that's constrained never to slope upward. There is no elementary way to do this. Indeed, it is an unresolved problem in statistics how best to include in a model additional knowledge that you might have such as "world records can't get worse with time."

---

#### 4.5.5 Main Effects without the Intercept

It's possible to construct a model with main terms but no intercept terms. If the explanatory variables are all quantitative, this is almost always a mistake. The figure, which plots the model function for swim time modeled by age with no intercept term, shows why.

For review purposes



The model function is sloping slightly upward rather than falling as the data clearly indicate. This is because, without an intercept term, the model line is forced to go through the origin. The line is sloping upward so that it will show a time of zero in the hypothetical year zero! Silly. It's no wonder that the model function fails to look anything like the data.

Never leave out the intercept unless you have a very good reason. Indeed, statistical software typically includes the intercept term by default. You have to go out of your way to tell the software to exclude the intercept.

## 4.6 Standard Notation for Describing Model Design

There is a concise notation for specifying the choices made in a model design, that is, which is the response variable, what are the explanatory variables, and what model terms to use. This notation, introduced originally in [9], will be used throughout the rest of this book and you will use it in working with computers.

To illustrate, here is the notation for some of the models looked at earlier in this chapter:

- $\text{ccf} \sim 1 + \text{temperature}$
- $\text{wage} \sim 1 + \text{sex}$
- $\text{time} \sim 1 + \text{year} + \text{sex} + \text{year}:\text{sex}$

The  $\sim$  symbol (pronounced “tilde”) divides each statement into two parts. On the left of the tilde is the name of the response variable. On the right is a list of model terms. When there is more than one model term, as is typically the case, the terms are separated by a  $+$  sign.

The examples show three types of model terms:

1. The symbol 1 stands for the intercept term.
2. A variable name (e.g., sex or temperature) stands for using that variable in a main term.
3. An interaction term is written as two names separated by a colon, for instance year:sex.

Although this notation looks like arithmetic or algebra, IT IS NOT. The plus sign does not mean arithmetic addition, it simply is the divider mark between terms. In English, one uses a comma to mark the divider as in “rock, paper, and scissors.” The modeling notation uses + instead: “rock + paper + scissors.” So, in the modeling notation  $1 + \text{age}$  does NOT mean “arithmetically add 1 to the age.” Instead, it means “two model terms: the intercept and age as a main term.”

Similarly, don’t confuse the tilde with an algebraic equal sign. The model statement is not an equation. So the statement  $\text{wage} \sim 1 + \text{age}$  does *not* mean “wage equals 1 plus age.” Instead it means, “wage is the response variable and there are two model terms: the intercept and age as a main term.”

As concise as the modeling notation is, it’s used so much that they like to use some shorthand. Two main points will cover most of what you will do:

- You don’t have to type the 1 term; it will be included by default. So,  $\text{wage} \sim \text{age}$  is the same thing as  $\text{wage} \sim 1 + \text{age}$ .

On those very rare occasions when you might want to insist that there be no intercept term, you can indicate this with a minus sign:  $\text{wage} \sim \text{age} - 1$ .

- Almost always, when you include an interaction term between two variables, you will also include the main terms for those variables. The \* sign can be used as shorthand. The model  $\text{wage} \sim 1 + \text{sex} + \text{age} + \text{sex}:\text{age}$  can be written simply as  $\text{wage} \sim \text{sex} * \text{age}$ .

## 4.7 Computational Technique

At the core of the language of modeling is the notation that uses the tilde character (~) to identify the response variable and the explanatory variables. This notation is incorporated into many of operators that you will use.

To illustrate the computer commands for modeling and graphically displaying relationships between variables, use the utilities data set:

DATA FILE  
utilities.csv

```
> utils = ISMdata("utilities.csv")
```

The examples make particular use of these variables

- *ccf* — the natural gas usage in cubic feet during the billing period.
- *month* — the month coded as 1 to 12 for January to December.
- *temp* — the average temperature during the billing period.

DATA FILE  
cps.csv

Another example uses Current Population Survey wage data:

```
> cps = ISMdata("cps.csv")
```

and focuses on the variables *wage*, *sex*, and *sector*.

For review purposes only

### 4.7.1 Bi-variate Plots

The basic idea of a bi-variate (two variable) plot is to examine one variable as it relates to another. The conventional format is to plot the response variable on the vertical axis and an explanatory variable on the horizontal axis.

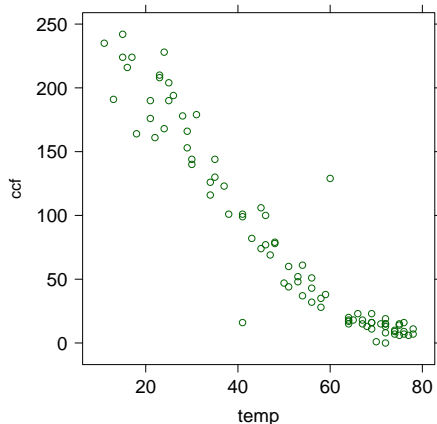
#### Quantitative Explanatory Variable

When the explanatory variable is quantitative, a scatter-plot is an appropriate graphical format. In the scatter plot, each case is a single point.

The basic computer operator for making scatter plots is `xyplot`:

```
> xyplot( ccf ~ temp, data=utils)
```

The first argument is a model formula written using the `~` modeling notation. This formula, `ccf ~ temp` is pronounced “ccf versus temperature.”



In order to keep the model notation concise, the model formula has left out the name of the data frame to which the variables belong. Instead, the frame is specified in the `data` argument. Since `data` has been set to be `utils`, the formula `ccf~temp` is effectively translated to `utils$ccf~utils$temp`.

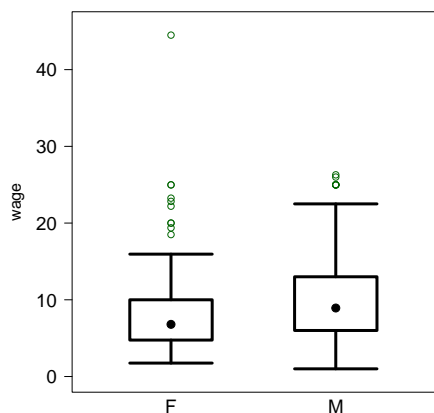
You can specify the labels by hand, if you like. For example,

```
> xyplot( ccf ~ temp, data=utils,
  xlab="Temperature (deg F)",
  ylab="Natural Gas Usage (ccf)")
```

#### Categorical Explanatory Variable

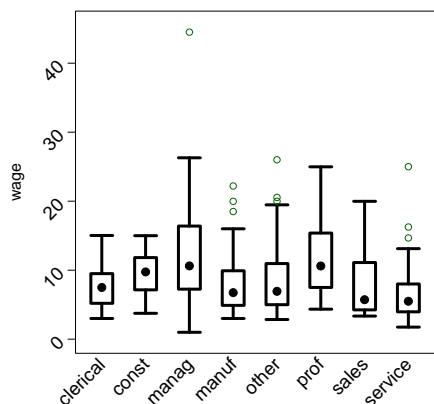
When the explanatory variable is categorical, an appropriate format of display is the box-and-whiskers plot, made with the `bwplot` operator. Here, for example, is the wage versus sex from the Current Population Survey:

```
> bwplot( wage ~ sex, data=cps)
```



When there are many levels, and when the names of the levels are long, it can become hard to read the labels on the graph. An effective solution is to rotate the labels, perhaps by 45 degrees. Here is `wage ~ sector`

```
> bwplot(wage~sector, data=cps, scales=list(rot=45) )
```



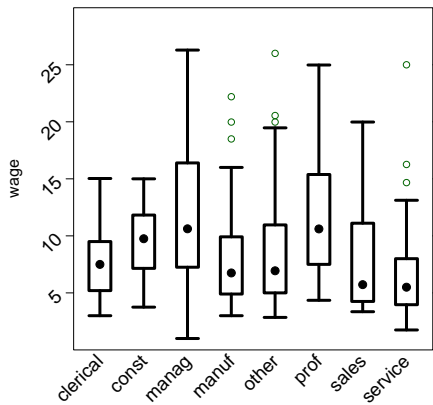
Admittedly, the argument `scales=list(rot=45)` is obscure. When you need to use it, just copy it from this example.

Notice that the outliers are setting the overall vertical scale for the graph and obscuring the detail at typical wage levels. You can use the `ylim` argument to set the scale of the y-axis however you want. For example:

```
> bwplot(wage~sector, data=cps, scales=list(rot=45),
        ylim=c(0,30) )
```

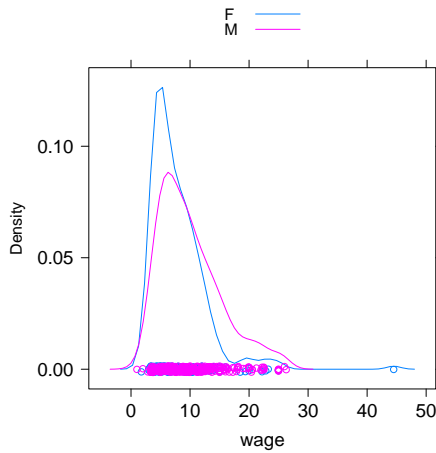
For review purposes only





You can also make side-by-side density plots which show more detail than the box-and-whisker plots. For instance:

```
> densityplot( ~ wage, groups=sex, data=cps, auto.key=TRUE )
```



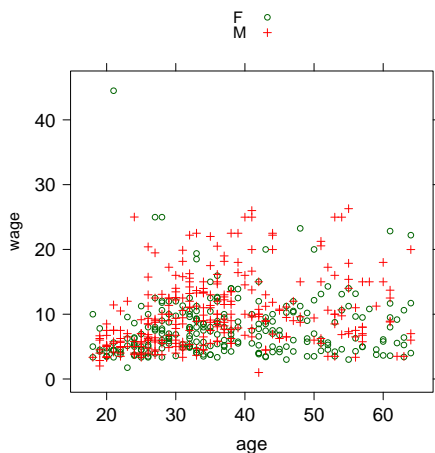
It seems a bit odd to have the notation `~wage` — nothing is in the role of the response variable. This idiosyncratic notation perhaps is meant to reflect that `wage` is on the horizontal axis.

### Multiple Explanatory Variables

The two-dimensional nature of paper or the computer screen lends itself well to displaying two variables: a response versus a single explanatory variable. Sometimes it is important to be able to add an additional explanatory variable. The graphics system gives a variety of options in this regard:

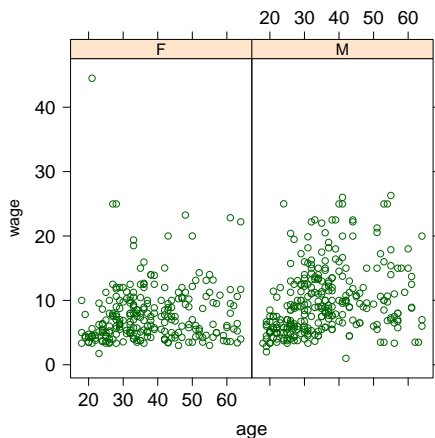
**Coding the additional explanatory variable** using color or symbol shapes. This is done by using the `groups` argument set to the name of the additional explanatory variable. For example:

```
> xyplot(wage ~ age, groups=sex, data=cps,
  auto.key=TRUE)
```



**Splitting the plot** into the groups defined by the additional explanatory variable. This is done by including the additional variable in the model formula using a `|` separator. For example:)

```
> xyplot(wage ~ age | sex, data=cps)
```



#### 4.7.2 Fitting Models and Finding Model Values

The `lm` operator (short for “Linear Model”) will translate a model design into fitted model values. It does this by “fitting” the model to data, a process that will be explained in later chapters. For now, focus on how to use `lm` to compute the fitted model values.

The `lm` operator uses the same model language as in the book. To illustrate, consider the world-record swim-times data :

```
> swim = ISMdata("swim100m.csv")
```

To construct the model `time ~ 1` for the swim data:

```
> mod1 = lm( time ~ 1, data=swim)
```

Here the model has been given a name, `mod1`, so that you can refer to it later. You can use any name you like, so long as it is valid in R.

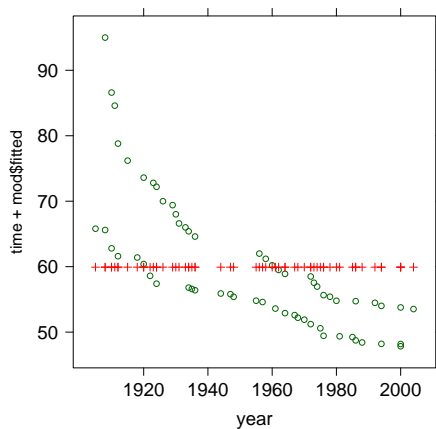
Once the model has been constructed, the fitted values can be found using the fitted operator:

```
> fitted(mod1)
      1      2      3      4      5      6      7      8
59.92 59.92 59.92 59.92 59.92 59.92 59.92 59.92
      9     10     11     12     13     14     15     16
59.92 59.92 59.92 59.92 59.92 59.92 59.92 59.92
... and so on for 62 cases altogether.
```

There is an individual fitted model value for each case. Of course, in this model all the model values are exactly the same since the model `time ~ 1` treats all the cases as exactly the same.

In later chapters you'll see how to analyze the model values, make predictions from the model, and assess the contribution of each model term. For now, just look at the model values by plotting them out along with the data used. I'll plot out both the data values and the model values versus year just to emphasize that the model values are the same for every case:

```
> xyplot( time + fitted(mod1) ~ year, data=swim)
```



Pay careful attention to the syntax used in the above command. There are two quantities to the left of the `~`. This is not part of the modeling language, where there is always a single response variable. Instead, it is a kind of shorthand, telling `xyplot` that it should plot out *both* of the quantities on the left side against the quantity on the right side. Of course, if you wanted to plot just the model values, without the actual data, you could specify the formula as `fitted(mod1)~year`.

Here are more interesting models:

```
mod2 = lm( time ~ 1+year, data=swim)
```

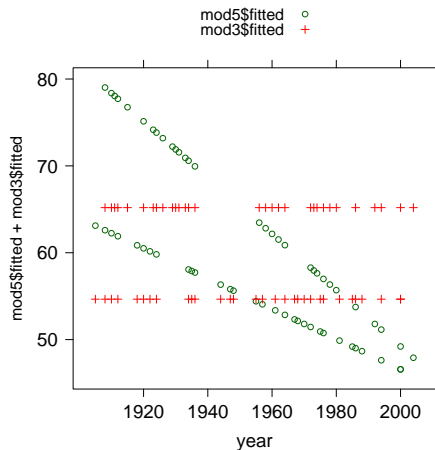


For review purposes only

```
mod3 = lm( time ~ 1+sex, data=swim)
mod4 = lm( time ~ 1+sex+year, data=swim)
mod5 = lm( time ~ 1+year+sex+year:sex, data=swim)
```

You can, if you like, compare the fitted values from different models on one plot:

```
> xyplot( fitted(mod5) + fitted(mod3) ~ year, data=swim,
  auto.key=TRUE)
```



### Shorthand notation for modeling

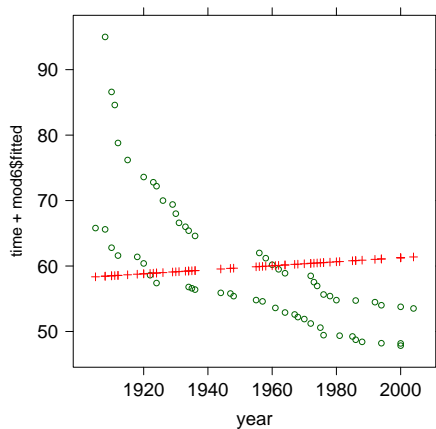
The intercept term is almost always included in models. For this reason, it's included by default even if you don't have the 1 term explicitly in the design of your model. For example, `mod2` could have been constructed with this statement:

```
mod2 = lm( time ~ year, data=swim )
```

### Suppressing the Intercept Term

You will rarely have to do it, but if you want to *exclude* the intercept term from a model, you use the notation `-1` in the model formula:

```
mod6 = lm( time ~ year-1, data=swim)
xyplot( time + fitted(mod6) ~ year, data=swim)
```



The model shown in the graph is obviously a poor fit to the data. This is because the intercept has been left out. The line indicated by the fitted model values therefore has to run through the origin:  $\text{time}=0$  when  $\text{year}=0$ .

### Interactions and Main Effects

Typically a model that includes an interaction term between two variables will include the main terms from those variables too. As a shorthand for this, the modeling language has a `*` symbol. So, the formula `time~year+sex+year:sex` can also be written `time~year*sex`.

### Transformation Terms

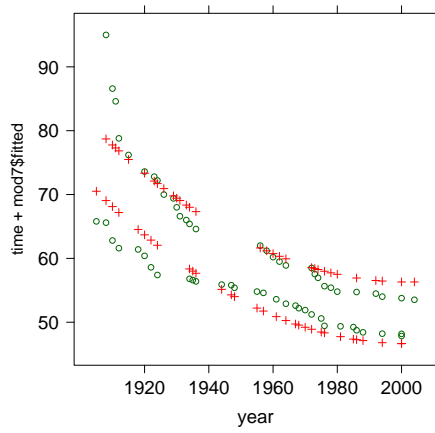
Transformation terms such as squares can also be included in the model formula. To mark the quantity clearly as a single term, it's best to wrap the term with `I()` as follows:

```
> mod7 = lm( time ~ year + I(year^2) + sex, data=swim)
```

Another way to accomplish this, for polynomials, is to use the operator `poly` as in the model formula `time~poly(year, 2)+sex`.

Here's a plot of the result:

```
> xyplot( time + fitted(mod7) ~ year, data=swim)
```



For review purposes only

# Chapter 5

## Model Formulas and Coefficients

*All economical and practical wisdom is an extension or variation of the following arithmetical formula:  $2 + 2 = 4$ . Every philosophical proposition has the more general character of the expression  $a + b = c$ . We are mere operatives, empirics, and egotists, until we learn to think in letters instead of figures. — Oliver Wendell Holmes (1841-1935)*

The previous chapter presents models as graphs. The response variable is plotted on the vertical axis and one of the explanatory variables on the horizontal axis. Such a visual depiction of a model is extremely useful, when it can be done. The relationship between the response and explanatory variables can be seen as slopes or differences; interactions can be seen as differences between slopes; the match or mismatch between the data and the model can be visualized easily.

A graph is also a useful mode of communicating; many people have the graph-reading skills needed to interpret the sorts of models in the previous chapter, or, at least, to get an impression of what the models are about.

Presenting models via a graph is, however, very limiting. The sorts of models that can be graphed effectively have only one or two explanatory variables, whereas often models need many more explanatory variables. (There are models with thousands of explanatory variables, but even a model with only three or four explanatory variables can be impossible to graph in an understandable way.) Even when a model is simple enough to be graphed, it's helpful to be able to quantify the relationships. Such quantification becomes crucial when you want to characterize the reliability of a model or draw conclusions about the strength of the evidence to support the claim of a relationship shown by a model.

This chapter introduces important ways to present models in a non-graphical way as formulas and as coefficients.

For review purposes only

## 5.1 The Linear Model Formula

Everyone who takes high-school algebra encounters this equation describing a straight-line relationship:

$$y = mx + b$$

The equation is so familiar to many people that they automatically make the following associations:  $x$  and  $y$  are the variables,  $m$  is the slope of the line,  $b$  is the  $y$ -intercept — the place the line crosses the  $y$  axis.

The straight-line equation is fundamental to statistical modeling, although the nomenclature is a little different. To illustrate, consider this model of the relationship between an adult's height and the height of his or her mother, based on Galton's height data:



$$\text{Model Values} \quad \text{height} = 46.7 + 0.313 \text{ mother}$$

This is a model represented not as a graph but as a **model formula**. Just reading the model in the same way as  $y = mx + b$  shows that the intercept is 46.7 and the slope is 0.313. That is, if you make a graph of the model values of the response variable height (of the adult child) against the values of the explanatory variable mother, you will see a straight line with that slope and that intercept.

To find the model value given by the formula, just plug in numerical values for the explanatory variable. So, according to the formula, a mother who is 65 inches tall will have children with a typical height of  $46.7 + 0.313 \times 65$ , giving 67.05 inches.

You can also interpret the model in terms of the relationship between the child's height and the mother's height. Comparing two mothers who differ in height by one inch, their children would typically differ in height by 0.313 inches.

In the design language, the model is specified like this:

$$\text{height} \sim 1 + \text{mother}$$

There is a simple correspondence between the model design and the model formula. The model formula takes each of the terms in the model design and multiplies it by a number. The intercept term is multiplied by 46.7 and the mother term is multiplied by 0.313. Such numbers have a generic name: **model coefficients**. Instead of calling 0.313 the "slope," it's called the "coefficient on the term mother." Similarly, 46.7 could be the "coefficient on the intercept term," but it seems more natural just to call it the intercept.

Where did these precise values for the coefficients come from? A process called **fitting the model to the data** finds coefficients that bring the model values from the formula to correspond as closely as possible to the response values in the data. As a result, the coefficients that result from fitting a given model design will depend on the data to which the model is fitted. Chapter 6 describes the fitting process in more detail.

For review purposes only



## 5.2 Linear Models with Multiple Terms

It's easy to generalize the linear model formula to include more than one explanatory variable or additional terms such as interactions. For each term, just add a new component to the formula. For example, suppose you want to model the child's height as a function of **both** the mother's and the father's height. As a model design, take

$$\text{height} \sim 1 + \text{mother} + \text{father}$$

Here is the model formula for this model design fitted to Galton's data:

$$\begin{array}{l} \text{Model} \\ \text{Values} \end{array} \quad \text{height} = 22.3 + 0.283 \text{ mother} + 0.380 \text{ father}$$

As before, each of the terms has its own coefficient. This same pattern of terms and coefficients can be extended to include as many variables as you like.

Interaction terms also fit into this framework. Suppose you want to fit a model with an interaction between the father's height and the mother's height. The model design is

$$\text{height} \sim 1 + \text{mother} + \text{father} + \text{father}:\text{mother}$$

For Galton's data, the formula that corresponds to this design is

$$\begin{array}{l} \text{Model} \\ \text{Values} \end{array} \quad \text{height} = 132.3 - 1.43 \text{ mother} - 1.21 \text{ father} + 0.0247 \text{ father} \times \text{mother}$$

The formula follows the same pattern as before: a coefficient multiplying the value of each term. So, to find the model value of height for a child whose mother is 65 inches tall and whose father is 67 inches tall, multiply the coefficient by the value of corresponding term:  $132.3 - 1.43 \times 65 - 1.21 \times 67 + 0.0247 \times 65 \times 67$  giving 66.12 inches. The term  $\text{father} \times \text{mother}$  may look a little odd, but it just means to multiply the mother's height by the father's height.

## 5.3 Formulas with Categorical Variables

Since quantitative variables are numbers, they can be reflected in a model formula in a natural, direct way: just multiply the value of the model term by the coefficient on that term.

Categorical variables are a little different. They don't have numerical values. It doesn't mean anything to multiply a category name by a coefficient.

In order to include a categorical variable in a model formula, a small translation is necessary. Suppose, for example, that you want to model the child's height by both its father's and mother's height and also the sex of the child. The model design is nothing new: just include a model term for sex.

$$\text{height} \sim 1 + \text{mother} + \text{father} + \text{sex}$$

---

**Aside. 5.1** Interpreting Interaction Terms

---

The numerics of interaction terms are easy. You just have to remember to multiply the coefficient by the product of all the variables in the term. Multiply the coefficient times the values of all the variables in the interaction term. The meaning of interaction terms is somewhat more difficult. Many people initially mistake interaction terms to refer to a relationship between two variables. For instance, they would (wrongly) think that an interaction between mother's and father's heights means that, say, tall mothers tend to be married to tall fathers. Actually, interaction terms do not describe the relationship between the two variables, they are about *three* variables: how one explanatory variable modulates the effect of another on the response.

A model like  $\text{height} \sim 1 + \text{mother} + \text{father}$  captures some of how a child's height varies with the height of either parent. The coefficient on mother (when fitting this model to Galton's data) was 0.283, indicating that an extra inch of the mother's height is associated with an extra 0.283 inches in the child. This coefficient on mother doesn't depend on the father's height; the model provides no room for it to do so.

Suppose that the relationship between the mother's height and her child's height is potentiated by the father's height. This would mean that if the father is very tall, then the mother has even more influence on the child's height than for a short father. That's an interaction.

To see this sort of effect in a model, you have to include an interaction term, as in the model  $\text{height} \sim 1 + \text{mother} + \text{father} + \text{mother}:\text{father}$ . The coefficients from fitting this model to Galton's data allow you to compare what happens when the father is very short (say, 60 inches) to when the father is very tall (say, 75 inches). With a short father, an extra inch in the mother is associated with an extra 0.05 inches in the child. With a tall father, an extra inch in the mother is associated with an extra 0.42 inches in the child.

The interaction term can also be read the other way: the relationship between a father's height and the child's height is greater when the mother is taller.

Of course, this assumes that the model coefficients can be taken at face value as reliable. Later chapters will deal with how to evaluate the strength of the evidence for a model. It will turn out that Galton's data do not provide good evidence for an interaction between mother's and father's heights in determining the child's height.

---

The corresponding model formula for this design on the Galton data is

$$\begin{array}{l} \text{Model} \\ \text{Values} \end{array} \quad \text{height} = 15.3 + 0.322 \text{ mother} + 0.406 \text{ father} + 5.23 \text{ sexM}$$

Interpret the quantitative terms in the ordinary way. The new term, 5.23  $\text{sexM}$ , means “add 5.23 whenever the case has level M on sex.” This is a very roundabout way of saying that males tend to be 5.23 inches taller than females, according to the model.

Another way to think about the meaning of  $\text{sexM}$  is that it is a new quantitative variable, called an **indicator variable**, that has the numeric value 1 when the case is level M in variable sex, and numeric value 0 otherwise.

A categorical variable will have one indicator variable for each level of the variable. Thus, a variable language with levels Chinese, English, French, German, Hindi, etc. has a separate indicator variable for each level.

For the sex variable, there are two levels: F and M. But notice that there is only one coefficient for sex, the one for  $\text{sexM}$ . Get used to this. Whenever there is an intercept term in a model, at least one of the indicator variables from any categorical variable will be left out. The level that is omitted is a **reference level**. Chapter 6 explains why things are done this way.

## 5.4 Model Coefficients Describe Relationships

Model coefficients describe the strength of relationships. Consider again this model of height:

$$\begin{array}{l} \text{Model} \\ \text{Values} \end{array} \quad \text{height} = 15.3 + 0.322 \text{ mother} + 0.406 \text{ father} + 5.23 \text{ sexM}$$

The coefficient 5.23 on  $\text{sexM}$  indicates that males are typically 5.23 inches taller than females (who are the reference group). Thus, the coefficient indicates the strength of the relationship between sex and height.

The coefficient 0.322 on the mother main term means that two mothers who differ in height by 1 inch will have children whose typical height differs by 0.322 inches. This means that there is indeed a relationship between the mother’s height and the child’s. If the coefficient were bigger, the relationship would be stronger.

The sign of a coefficient tells which way the relationship goes. If the coefficient on  $\text{sexM}$  had been  $-5.23$ , it would mean that males are typically shorter than females. If the coefficient on mother had been  $-0.322$ , then taller mothers would have shorter children.

When a variable is included in more than one model term, the individual coefficients can no longer be read directly as indicating how that variable contributes to the relationship. Instead, one needs to compare the model values under different settings of the explanatory variables. To see why, consider again the above formula. The model includes each variable only as a main term, so

For review purposes only

the result will be the same as you could read off directly from the coefficient. For instance, to see the strength of the relationship between mother's height and the child's height, pick values for the inputs, calculate the resulting model value, and then compare this to the model value for a slightly different value of mother's heights. Suppose that you pick sex as M, mother as 65 inches, and father as 68 inches. This results in a model value of  $15.3 + 0.322 \times 65 + 0.406 \times 68 + 5.23 \times 1$  or 69.0744 inches. Now change the value for mother to 66; the resulting model value is different: 69.3959 inches. The difference between them is the typical height difference associated with a 1-inch difference in mother's height: 0.3215 inches, just what you already saw from the coefficient on mother.

Those familiar with calculus may recognize the connection to the *partial derivative* in this approach of comparing model values due to a change of inputs. Chapter 8 will consider the issues in more detail, particularly the matter of **holding constant** other variables or **adjusting** for other variables.

## 5.5 Model Values and Residuals

If you plug in the values of the explanatory variables from your data, the model formula gives, as an output, the model values. It's rarely the case that the model values are an exact match with the actual response variable in your data. The difference is the *residual*.

For any one case in the data, the residual is always defined in terms of the actual response variable and the model value that arises when that case's values for the explanatory variables are given as the input to the model formula. For instance, if you are modeling height, the residuals are

$$\text{height} = \text{Model Values} \text{ height} + \text{residuals}$$

The residuals are always defined in terms of a particular data set and a particular model. Each case's residual would likely change if the model were altered or if another data set were used to fit the model.

## 5.6 Coefficients of Basic Model Designs

The presentation of a linear model as a set of coefficients is a compact shorthand for the complete model formula. It often happens that for some purposes, interest focuses on a single coefficient of interest. In order to help you to interpret coefficients correctly, it is helpful to see how they relate to some basic model designs. Often, the interpretation for more complicated designs is not too different.

To illustrate how basic model designs apply generally, I will use A to stand for a generic response variable, B to stand for a quantitative explanatory variable, and G for a categorical explanatory variable. As always, 1 refers to the intercept term.

### Model $A \sim 1$

The model  $A \sim 1$  is the simplest of all. There are no explanatory variables; the only term is the intercept. Think of this model as saying that all the cases are the same. In fitting the model, you are looking for a single value that is as close as possible to each of the values in  $A$ .

The coefficient from this model is the mean of  $A$ . The model values are the same for every case: the mean of all the samples. This is sometimes called the **grand mean** to distinguish it from group means, the mean of  $A$  for different groups.

**Example 5.1:** The mean height of all the cases in Galton's height data — the "grand mean" — is the coefficient on the model  $\text{height} \sim 1$ . Fitting this model to Galton's data gives this coefficient:

Model Term	Coefficient
Intercept	66.7607

Thus, the mean height is 66.76 inches.

**Example 5.2:** The mean wage earned by all of the people in the Current Population Survey is given by the coefficient of the model  $\text{wage} \sim 1$ . Fitting the model to the CPS gives:

Model Term	Coefficient
Intercept	8.96

Thus, the mean wage is \$8.96 per hour.

### Model $A \sim 1+G$

The model  $A \sim 1+G$  is also very simple. The categorical variable  $G$  can be thought of as dividing the data into groups, one group for each level of  $G$ . There is a separate model value for each of the groups.

The model values are the group-wise means: separate means of  $A$  for the cases in each group. The model coefficients, however, are not exactly these group-wise means. Instead, the coefficient of the intercept term is the mean of one group, which can be called the **reference group** or **reference level**. Each of the other coefficients is the *difference* between its group's mean and the mean of the reference group.

**Example 5.3:** Calculate the group-wise means of the heights of men and women in Galton's data by fitting the model  $\text{height} \sim \text{sex}$ .

Model Term	Coefficient
Intercept	64.11
sex <b>M</b>	5.12

The mean height for the reference group, women, is 64.11 inches. Men are taller by 5.12 inches. In the standard form of a model report, the identity of the reference group is not stated explicitly. You have to figure it out from which levels of the variable are missing.

By suppressing the intercept term, you change the meaning of the remaining coefficients; they become simple group-wise means rather than the difference of the mean from a reference group's mean. Here's the report from fitting the model  $\text{height} \sim \text{sex} - 1$ .

Model Term	Coefficient
sex <b>F</b>	64.11
sex <b>M</b>	69.23

It might seem obvious that this simple form is to be preferred, since you can just read off the means without doing any arithmetic on the coefficients. That can be the case sometimes, but almost always you will want to include the intercept term in the model. The reasons for this will become clearer when hypothesis testing is introduced in later chapters.

**Example 5.4:** Calculate group-wise means of wages in the different sectors of the economy by fitting the model  $\text{wage} \sim \text{sector}$ :

Model Term	Coefficient
Intercept	7.42
sector <b>const</b>	2.08
sector <b>manag</b>	4.69
sector <b>manuf</b>	0.61
sector <b>other</b>	1.08
sector <b>prof</b>	4.52
sector <b>sales</b>	0.17
sector <b>service</b>	-0.89

There are eight levels of the sector variable, so the model has 8 coefficients. The coefficient of the intercept term gives the group-wise mean of the reference group. The reference group is the one level that isn't listed explicitly in the other coefficients; it turns out to be the clerical sector. So, the mean wage of clerical workers is \$7.42 per hour. The other coefficients give the *difference* between the mean of the reference group and the means of other groups. For example, workers in the construction sector make, on average \$2.08 per hour more than clerical workers. Similarly, service sector workers make 89 cents per hour *less* than clerical workers.

**Model  $A \sim 1+B$** 

Model  $A \sim 1+B$  is the basic straight line relationship. The two coefficients are the intercept and the slope of the line. The slope tells what change in A corresponds to a one-unit change in B.

**Example 5.5:** The model  $\text{wage} \sim 1+\text{educ}$  shows how workers wages are different for people with different amounts of education (as measured by years in school). Fitting this model to the Current Population Survey data gives the following coefficients:

Model Term	Coefficient
Intercept	-0.69
educ	0.74

According to this model, a one-year increase in the amount of education that a worker received is associated with a 74 cents per hour increase in wages. (Remember, these data are from 1985.)

It may seem odd that the intercept coefficient is negative. Nobody is paid a negative wage. Keep in mind that the intercept of a straight line  $y = mx + b$  refers to the value of  $y$  when  $x = 0$ . The intercept coefficient  $-0.69$ , tells the typical wage for workers with zero years of education. There are no workers in the data set with zero years; only three workers have less than five years. In this data set, the intercept is an **extrapolation** outside of the range of the data.

**Model  $A \sim 1 + G + B$** 

The model  $A \sim 1 + G + B$  gives a straight-line relationship between A and B, but allows different lines for each group defined by G. The lines are different only in their intercepts; all of the lines have the same slope.

The coefficient labeled “intercept” is the intercept of the line for the reference group. The coefficients on the various levels of categorical variable G reflect how the intercepts of the lines from those groups differ from the reference group’s intercept.

The coefficient on B gives the slope. Since all the lines have the same slope, a single coefficient will do the job.

**Example 5.6:** Wages versus educational level for the different sexes:  $\text{wage} \sim 1 + \text{educ} + \text{sex}$  The coefficients are

Model Term	Coefficient
Intercept	-1.92773
sexM	2.27
educ	0.74

The educ coefficient tells how education is associated with wages. The sexM coefficient says that men tend to make \$2.27 more an hour than women when

comparing men and women with the same amount of education. Chapter 8 will explain why the inclusion of the `educ` term allows this comparison *at the same level* of education.

Note that the model  $A \sim 1 + G + B$  is the same as the model  $A \sim 1 + B + G$ . The order of model terms doesn't make a difference.

### Model $A \sim 1 + G + B + G:B$

The model  $A \sim 1 + G + B + G:B$  is also a straight-line model, but now the different groups defined by `G` can have different slopes and different intercepts. (The interaction term `G:B` says how the slopes differ for the different group. That is, thinking of the slope as effect of `B` on `A`, the interaction term `G:B` tells how the effect of `B` is modulated by different levels of `G`.)

**Example 5.7:** Here is a model describing wages versus educational level, separately for the different sexes:  $\text{wage} \sim 1 + \text{sex} + \text{educ} + \text{sex}:\text{educ}$

Model Term	Coefficient
Intercept	-3.10
<code>sexM</code>	4.20
<code>educ</code>	0.83
<code>sexM:educ</code>	-0.15

Interpreting these coefficients: For women, an extra year of education is associated with an increase of wages of 83 cents per hour. For men, the relationship is weaker: an increase in education of one year is associated with only an increase of wages of only  $0.831 - 0.148$  or 68 cents per hour.

## 5.7 Coefficients have Units

A common convention is to write down coefficients and model formulas without being explicit about the units of variables and coefficients. This convention is unfortunate. Although leaving out the units leads to neater tables and simpler-looking formulas, the units are fundamental to interpreting the coefficients. Ignoring the units can mislead severely.

To illustrate how units come into things, consider the model design  $\text{wages} \sim 1 + \text{educ} + \text{sex}$ . Fitting this model design to the Current Population Survey gives this model formula:

Model Values  $\text{wage} = -1.93 + 0.742 \text{educ} + 2.27 \text{sexM}$

A first glance at this formula might suggest that `sex` is more strongly related than `educ` to `wage`. After all, the coefficient on `educ` is much smaller than the

For review purposes only



coefficient on  $\text{sexM}$ . But this interpretation is invalid, since it doesn't take into account the units of the variables or the coefficients.

The response variable,  $\text{wage}$ , has units of dollars-per-hour. The explanatory variable  $\text{educ}$  has units of years. The explanatory variable  $\text{sex}$  is categorical and has no units; the indicator variables for  $\text{sex}$  are just zeros and ones: pure numbers with no units.

The coefficients have the units needed to transform the quantity that they multiply into the units of the response variable. So, the coefficient on  $\text{educ}$  has units of "dollars-per-hour per year." This sounds very strange at first, but remember that the coefficient will multiply a quantity that has units of years, so the product will be in units of dollars-per-hour, just like the  $\text{wage}$  variable. In this formula, the units of the intercept coefficient and the coefficient on  $\text{sexM}$  both have units of dollars-per-hour, because they multiply something with no units and need to produce a result in terms of dollars-per-hour.

A person who compares 0.742 dollars-per-hour per year with 2.27 dollars per hour is comparing apples and oranges, or, in the metric-system equivalent, comparing meters and kilograms. If the people collecting the data had decided to measure education in months rather than in years, the coefficient would have been a measly  $0.742/12 = 0.0618$  even though the relationship between education and wages would have been exactly the same.

---

**Aside. 5.2 Comparing Coefficients**

---

If you want to compare the two coefficients, say the coefficient on  $\text{educ}$  and the coefficient on  $\text{sexM}$ , you have to put them on a common footing. It's not always obvious how to do this, because the coefficients have different units. So you have to be clever.

One approach that might work here is to find the number of years of education that produces a similar size effect of education on wages as is seen with  $\text{sexM}$ . The answer turns out to be about 3 years. (To see this, note that the wage gain associated with  $\text{sexM}$ , 2.27 dollars per hour to the wage gain associated with three years of education,  $3 \times 0.742 = 2.21$  dollars per hour.)

Thus, according to the model, being a male is equivalent in wage gains to an increase of 3 years in education.

---

## 5.8 Untangling Explanatory Variables

One of the advantages of using formulas to describe models is the way they facilitate using multiple explanatory variables. Many people assume that they can study relationships one variable at a time, even when they know there are influences from multiple factors. Underlying this assumption is a belief that influences add up in a simple way. Indeed model formulas without interaction terms actually do simply add up the contributions of each variable.

But this does not mean that an explanatory variable can be considered in isolation from other explanatory variables. There is something else going on

that makes it important to consider the explanatory variables not one at a time but simultaneously, at the same time.

In many situations, the explanatory variables are themselves related to one another. As a result, variables can to some extent stand for one another. An effect attributed to one variable might equally well be assigned to some other variable.

Due to the relationships between explanatory variables, you need to untangle them from one another. The way this is done is to use the variables together in a model, rather than in isolation. The way the tangling shows up is in the way the coefficient on a variable will change when another variable is added to the model or taken away from the model. That is, model coefficients on a variable tend to depend on the context set by other variables in the model.

To illustrate, consider this criticism of spending on public education in the United States from a respected political essayist:

*The 10 states with the lowest per pupil spending included four — North Dakota, South Dakota, Tennessee, Utah — among the 10 states with the top SAT scores. Only one of the 10 states with the highest per pupil expenditures — Wisconsin — was among the 10 states with the highest SAT scores. New Jersey has the highest per pupil expenditures, an astonishing \$10,561, which teachers' unions elsewhere try to use as a negotiating benchmark. New Jersey's rank regarding SAT scores? Thirty-ninth... The fact that the quality of schools... [fails to correlate] with education appropriations will have no effect on the teacher unions' insistence that money is the crucial variable. — George F. Will, (September 12, 1993), "Meaningless Money Factor," The Washington Post, C7. Quoted in [10].*

The response variable here is the score on a standardized test taken by many students finishing high school: the SAT. The explanatory variable — there is only one — is the level of school spending. But even though the essayist implies that spending is *not* the “crucial variable,” he doesn’t include any other variable in the analysis. In part, this is because the method of analysis — pointing to individual cases to illustrate a trend — doesn’t allow the simultaneous consideration of multiple explanatory variables. (Another flaw with the informal method of comparing cases is that it fails to quantify the strength of the effect: Just how much negative influence does spending have on performance? Or is the claim that there is no connection between spending and performance?)

You can confirm the claims in the essay by modeling. The SAT dataset contains state-by-state information from the mid-1990s on *per capita* yearly school expenditures in thousands of dollars, average statewide SAT scores, average teachers' salaries, and other variables.[10]

The analysis in the essay corresponds to a simple model:  $\text{sat} \sim 1 + \text{expend}$ . Fitting the model to the state-by-state data gives a model formula,

$$\text{Model Values} \quad \text{sat} = 1089 - 20.9 \text{ expend}$$

DATA FILE  
sat.csv

For review purposes only

The formula is consistent with the claim made in the essay; the coefficient on *expend* is negative. According to the model, an increase in expenditures by \$1000 per capita is associated with a 21 point decrease in the SAT score. That's not very good news for people who think society should be spending more on schools: the schools that spend the least per capita have the highest average SAT scores. (A 21 point decrease in the SAT doesn't mean much for an individual student, but as an average over tens of thousands of students, it's a pretty big deal.)

Perhaps expenditures is the wrong thing to look at — you might be studying administrative inefficiency or even corruption. Better to look at teachers' salaries. Here's the model  $\text{sat} \sim 1 + \text{salary}$ , where *salary* is the average annual salary of public school teachers in \$1000s:

$$\begin{array}{l} \text{Model} \\ \text{Values} \end{array} \quad \text{sat} = 1159 - 5.54 \text{ salary}$$

The essay's claim is still supported. Higher salaries are associated with lower average SAT scores! But maybe states with high salaries manage to pay well because they overcrowd classrooms. So, look at the average student/teacher ratio in each state:  $\text{sat} \sim 1 + \text{ratio}$

$$\begin{array}{l} \text{Model} \\ \text{Values} \end{array} \quad \text{sat} = 921 + 2.68 \text{ ratio}$$

Finally, a positive coefficient! That means . . . larger classes are associated with higher SAT scores.

All this goes against the conventional wisdom that holds that higher spending, higher salaries, and smaller classes will be associated with better performance.

At this point, many advocates for more spending, higher salaries, and smaller classes will explain that you can't measure the quality of an education with a standardized test, that the relationship between a student and a teacher is too complicated to be quantified, that students should be educated as complete human beings and not as test-taking machines, and so on. Perhaps.

Whatever the criticisms of standardized tests, the tests have the great benefit of allowing comparisons across different conditions: different states, different curricula, etc. If there is a problem with the tests, it isn't standardization itself but the material that is on the test. Absent a criticism of that material, rejections of standardized testing ought to be treated with considerable skepticism.

But there is something wrong with using the SAT as a test, even if the content of the test is good. What's wrong is that the test isn't required for all students. Depending on the state, a larger or smaller fraction of students will take the SAT. In states where very few students take the SAT, those students who do are the ones bound for out-of-state colleges, in other words, the high performers.

What's more, the states which spend the least on education tend to have the fewest students who take the SAT. That is, the fraction of students taking the SAT is entangled with expenditures and other explanatory variables.

To untangle the variables, they have to be included simultaneously in a model. That is, in addition to *expend* or *salary* or *ratio*, the model needs to take into account the fraction of students who take the SAT (variable *frac*).

Here are three models that attempt to untangle the influences of *frac* and the spending-related variables:

Model Values  $\text{sat} = 994 + 12.29 \text{ expend} - 2.85 \text{ frac}$

Model Values  $\text{sat} = 988 + 2.18 \text{ salary} - 2.78 \text{ frac}$

Model Values  $\text{sat} = 1119 - 3.73 \text{ ratio} - 2.55 \text{ frac}$

In all three models, including `frac` has completely altered the relationship between performance and the principal explanatory variable of interest, be it `expend`, `salary`, or `ratio`. Not only is the coefficient different, it is different in sign.

Chapter 8 discusses why adding `frac` to the model can be interpreted as an attempt to examine the other variables while holding `frac` constant, as if you compared only states with similar values of `frac`.

The situation seen here, where adding a new explanatory variable (e.g., `frac`) changes the sign of the coefficient on another variable (e.g., `expend`, `salary`, `ratio`) is called **Simpson's paradox**.

Simpson's Paradox is an extreme version of a common situation: that the coefficient on an explanatory variable can depend on what other explanatory variables have been included in the model. In other words, the role of an explanatory variable can depend, sometimes strongly, on the context set by other explanatory variables. You can't look at explanatory variables in isolation; you have to interpret them in context.

There is nothing magical about Simpson's Paradox or the dependence of model coefficients on context. It appears paradoxical only when the details of model fitting are hidden in a black box of software. You will open the black box in Chapter 11, which will help you to understand when and why Simpson's Paradox occurs and how to anticipate it.

But which is the right model? What's the right context? Do `expend`, `salary`, and `ratio` have a positive role in school performance as the second set of models indicate, or should you believe the first set of models? This is an important question and one that comes up often in statistical modeling. At one level, the answer is that you need to be aware that context matters. At another level, you should always check to see if your conclusions would be altered by including or excluding some other explanatory variable. At a still higher level, the choice of which variables to include or exclude needs to be related to the modeler's ideas about what causes what.

## 5.9 Why Linear Models?

Many people are uncomfortable with using linear models to describe potentially complicated relationships. The process seems a bit unnatural: specify the model terms, fit the model to the data, get the coefficients. How do you know that a model fit in this way will give a realistic match to the data? Coefficients seem an overly abstract way to describe a relationship. Modeling without graphing seems like dancing in the dark; it's nice to be able to see your partner.

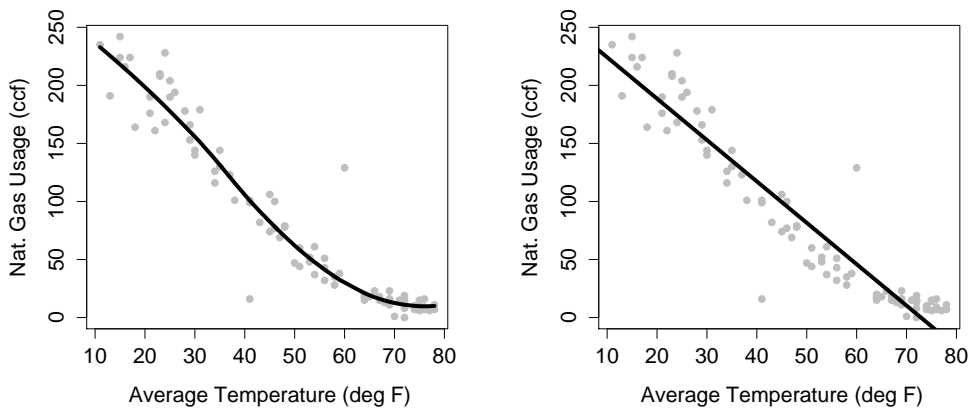


Figure 5.1: Two models of natural gas usage versus outdoor temperature.

Return to the example of world record swim times plotted in Figure 4.5 on page 78. There's a clear curvature to the relationship between record time and year. Without graphing the data, how would you have known whether to put in a transformation term? How would you know that you should use sex as an explanatory variable? But when you do graph the data, you can see easily that something is wrong with a straight-line model like  $\text{time} \sim 1 + \text{year}$ .

However, the advantages of graphing are obvious only in retrospect, once you have found a suitable graph that is informative. Why graph world record time against year? Why not graph it versus body weight of the swimmer or the latitude of the pool in which the record was broken?

Researchers decide to collect some variables and not others based on their knowledge and understanding of the system under study. People know, for example, that world records can only get better over the years. This may seem utterly obvious but it is nonetheless a bit of expert knowledge. It is expert knowledge that makes *year* an obvious explanatory variable. In general, expert knowledge comes from experience and education rather than data. It's common knowledge, for example, that in many sports, separate world records are kept for men and women. This feature of the system, obvious to experts, makes *sex* a sensible choice as an explanatory variable.

When people use a graph to look at how well a model fits data, they tend to look for a satisfyingly snug replication of the details. In Chapter 4 a model was shown of the relationship between monthly natural gas usage in a home and the average outdoor temperature during the month. Figure 5.1 shows that model along with a straight-line model. Which do you prefer? Look at the model in the left panel. That's a nice looking model! The somewhat irregular curve seems like a natural shape, not a rigid and artificial straight line like the model in the right panel.

Yet how much better is the curved model than the straight-line model? The

straight-line model captures an important part of the relationship between natural gas usage and outdoor temperature: that colder temperatures lead to more usage. The overall slopes of the two models are very similar. The biggest discrepancy comes at warm temperatures, above 65°F. But people who know about home heating can tell you that above 65°F, homes don't need to be heated. At those temperatures gas usage is due only to cooking and water heating: a very different mechanism. To study heating, you should use only data for fairly low temperatures, say below 65°F. To study two different mechanisms — heating at low temperatures and no heating at higher temperatures — you should perhaps construct two different models, perhaps by including an interaction between temperature as a quantitative variable and a categorical variable that indicates whether the temperature is above 65°.

Humans are powerful pattern recognition machines. People can easily pick out faces in a crowd, but they can also pick out faces in a cloud or on the moon. The downside to using human criteria to judge how well a model fits data is the risk that you will see patterns that aren't warranted by the data.

To avoid this problem, you can use formal measures of how well a model fits the data, measures based on the size of residuals and that take into account chance variations in shape. Much of the rest of the book is devoted to such measures.

With the formal measures of fit, a modeler has available a strategy for finding effective models. First, fit a model that matches, perhaps roughly, what you know about the system. The straight-line model in the right panel of Figure 5.1 is a good example of this. Check how good the fit is. Then, try refining the model, adding detail by including curvy transformation terms. Check the fit again and see if the improvement in the fit goes beyond what would be expected from chance.

A reasonable strategy is to start with model designs that include only main terms for your explanatory variables (and, of course, the intercept term, which is to be included in almost every model). Then add in some interaction terms, see if they improve the fit. Finally, you can try transformation terms to capture more detail.

Most of the time you will find that the crucial decision is which explanatory variables to include. Since it's difficult to graph relationships with multiple explanatory variables, the benefits of making a snug fit to a single variable are illusory.

Often the relationship between an explanatory variable and a response is rather loose. People, as good as they are at recognizing patterns, aren't effective at combining lots of cases to draw conclusions of overall patterns. People can have trouble seeing for forest for the trees. Figure 5.2 shows Galton's height data: child's height plotted against mother's height. It's hard to see anything more than a vague relationship in the cloud of data, but it turns out that there is sufficient data here to justify a claim of a pretty precise relationship. Use your human skills to decide which of the two models in Figure 5.2 is better. Pretty hard to decide, isn't it! Decisions like this are where the formal methods of

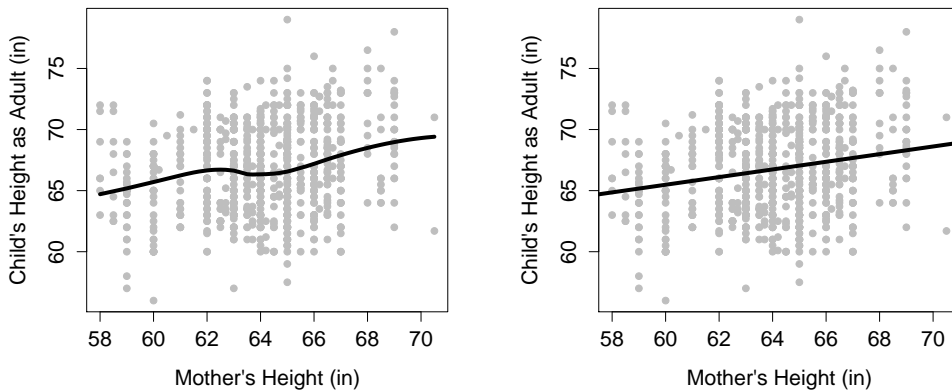


Figure 5.2: Two models of child's height as an adult versus mother's height.

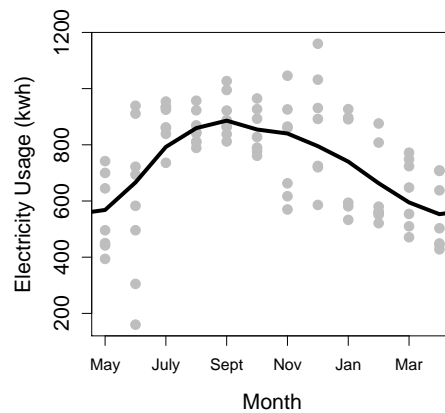


Figure 5.3: Electricity usage versus Month has a relationship that can't be captured by a straight-line model.

linear models pay off. Start with the straight-line terms then see if elaboration is warranted.

There are, however, some situations when you can anticipate that straight-line model terms will not do the job. For example, consider electricity use in a house that is heated electrically in the winter and cooled electrically in the summer. The relationship between electricity and temperature can be expected to be V-shaped — heavy use both for very cold temperatures when the heat is on and for very warm temperatures when air conditioning is in use. Or consider the relationship between college grades and participation in extra-curricular activities such as school sports, performances, the newspaper, etc. There's reason

to believe that some participation in extra-curricular activities is associated with higher grades. This might be because students who are doing well have the confidence to spend time on extra-curriculars. But there's also reason to think that very heavy participation takes away from the time students need to study. So the overall relationship between grades and participation might be  $\Lambda$ -shaped. Relationships that have a V- or  $\Lambda$ -shape won't be effectively captured by straight-line models; transformation terms and interaction terms will be required.

## 5.10 Computational Technique

The `lm` operator finds model coefficients. To illustrate, here's a pair of statements that read in a data frame and fit a model to it:

```
> swim = ISMdata("swim100m.csv")
> mod = lm( time ~ year + sex, data=swim)
```

DATA FILE  
swim100m.csv

The first argument to `lm` is a model design, the second is the data frame.

The object created by `lm` — here given the name `mod` — contains a variety of information about the model. To access the coefficients themselves, use the `coef` operator applied to the model:

```
> coef(mod)
(Intercept)      year      sexM
  555.7168    -0.2515    -9.7980
```

As shorthand to display the coefficients, just type the name of the object that is storing the model:

```
> mod
```

Call:

```
lm(formula = time ~ year + sex, data = swim)
```

Coefficients:

```
(Intercept)      year      sexM
  555.717    -0.251    -9.798
```

A more detailed report can be gotten with the summary operator. This gives additional statistical information that will be used in later chapters:

```
> summary(mod)
```

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 555.7168    33.7999   16.44 < 2e-16
year        -0.2515     0.0173  -14.52 < 2e-16
sexM         -9.7980     1.0129   -9.67 8.8e-14
```

For review purposes only



From time to time in the exercises, you will be asked to calculate model values “by hand.” This is accomplished by multiplying the coefficients by the appropriate values and adding them up. For example, the model value for a male swimmer in 2010 would be:

```
> 555.7 - 0.2515*2010 - 9.798
[1] 40.39
```

Notice that the “value” used to multiply the intercept is always 1, and the “value” used for a categorical level is either 0 or 1 depending on whether there is a match with the level. In this example, since the swimmer in question was male, the value of `sexM` is 1. If the swimmer had been female, the value for `sexM` would have been 0.

When a model includes interaction terms, the interaction coefficients need to be multiplied by all the values involved in the interaction. For example, here is a model with an interaction between year and sex:

```
> mod2 = lm( time ~ year*sex, data=swim)
> coef(mod2)
(Intercept)      year      sexM  year:sexM
   697.3012   -0.3240  -302.4638    0.1499
> 697.3 -0.3240*2010 - 302.5 +0.1499*2010
[1] 44.86
```

The `year:sexM` coefficient is being multiplied by the year (2010) and the value of `sexM`, which is 1 for this male swimmer.

### 5.10.1 Other Useful Operators

**cross** will combine two categorical variables into a single variable. For example, in the Current Population Survey data, the variable `sex` has levels F and M, while the variable `race` has levels W and NW. Crossing the two variables combines them; the new variable has four levels: F.NW, M.NW, F.W, M.W:

```
> cross(cps$sex, cps$race)
[1] M.W M.W F.W F.W M.W F.W F.W M.W M.W
[10] F.W M.W M.W M.W M.W M.W M.W M.W M.W
[19] M.NW M.W F.NW M.NW F.W F.W M.NW F.W F.W
... and so on.
Levels: F.NW M.NW F.W M.W
```

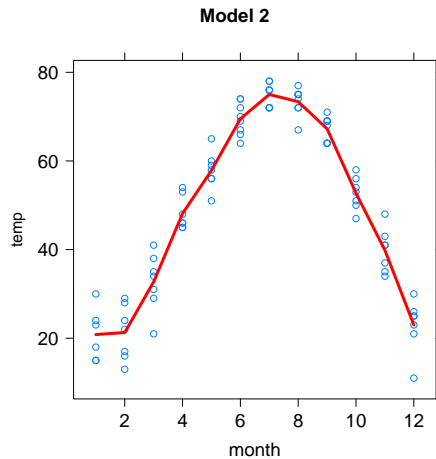
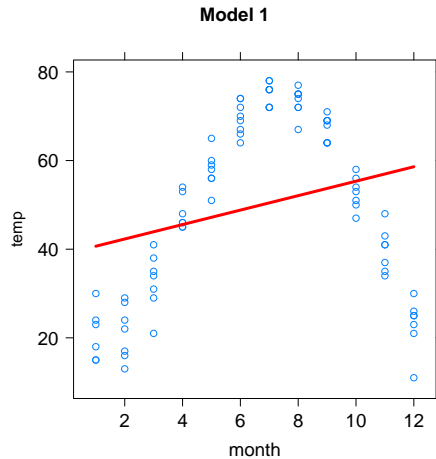
**as.factor** will convert a quantitative variable to a categorical variable. This is useful when a quantity like month has been coded as a number, say 1 for January and 2 for February, etc. but you do not want models to treat it as such.

To illustrate, consider two different models of the usage temperature versus month:

```
utils = ISMdata("utilities.csv")
mod1 = lm( temp ~ month, data=utils)
mod2 = lm( temp ~ as.factor(month), data=utils)
```

DATA FILE  
utilities.csv

Here are the graphs of those models:



In the first model, month is treated quantitatively, so the model term month produces a straight-line relationship that does not correspond well to the data.

In the second model, month is treated categorically, allowing a more complicated model relationship.

For review purposes only

# Chapter 6

## Fitting Models to Data

*With four parameters, I can fit an elephant; with five I can make it wiggle its trunk. — John von Neumann*

*The purpose of models is not to fit the data but to sharpen the questions. — Samuel Karlin*

The selection of variables and terms is part of the modeler’s art. It is a creative process informed by your goals, your understanding of how the system you are studying works, and the data you can collect. Later chapters will deal with how to evaluate the success of your creation as well as general principles for designing effective models that capture the salient aspects of your system.

This chapter is about how to find the coefficients. The goal is to find the “best” coefficients — to fit the model to the data. The outcome of fitting is the coefficients that capture as well as possible the variation in the response variable.

Fitting a model — once you have specified the design — is an entirely automatic process that requires no human decision making. It’s ideally suited to computers and, in practice, you will always use the computer to find the best fit. Nevertheless, it’s important to understand the way in which the computer finds the coefficients. Much of the logic of interpreting models is based on how firmly the coefficients are tied to the data, to what extent the data dictate precise values of the coefficients. The fitting process reveals this. In addition, depending on your data and your model design, you may introduce ambiguities into your model that make the model less reliable than it could be. These ambiguities, which stem from **redundancy**, are revealed by the fitting process.

### 6.1 The Least Squares Criterion

Fitting a model is somewhat like buying clothes. You choose the kind of clothing that you want: the cut, style, and color. There are usually several or many different items of this kind, and you have to pick the one that matches your body

For review purposes only

appropriately. In this analogy, the style of clothing is the model design — you choose this. Your body shape is like the data — you're pretty much stuck with what you've got.

There is a system of sizes of clothing: the waist, hip, and bust sizes, inseam, sleeve length, neck circumference, foot length, etc. By analogy, these are the model coefficients. When you are in the fitting room of a store, you try on different items of clothing with a range of values of the coefficients. Perhaps you bring in pants with waist sizes 31 through 33 and leg sizes 29 and 30. For each of the pairs of pants that you try on, you get a sense of the overall fit, although this judgment can be subjective. By trying on all the possible sizes, you can find the item that gives the best overall match to your body shape.

Similarly, in fitting a model, the computer tries different values for the model coefficients. Unlike clothing, however, there is a very simple, entirely objective criterion for judging the overall fit called the **least squares** criterion. To illustrate, consider the following very simple data set of age and height in children:

age (years)	height (cm)
5	80
7	100
10	110

As the model design, take  $\text{height} \sim 1 + \text{age}$ . To highlight the link between the model formula and the data, here's a way to write the model formula that includes not just the names of the variables, but the actual column of data for each variable. Each of these columns is called a **vector**. Note that the intercept vector is just a column of 1s.

$$\begin{array}{c} \text{Model} \\ \text{Values} \end{array} \text{ height} = c_1 \begin{array}{c} \text{intercept} \\ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \end{array} + c_2 \begin{array}{c} \text{age} \\ \begin{bmatrix} 5 \\ 7 \\ 10 \end{bmatrix} \end{array} .$$

The goal of fitting the model is to find the best numerical values for the two coefficients  $c_1$  and  $c_2$ .

To start, just guess something:  $c_1 = 5$  and  $c_2 = 10$ . Just a guess. Plugging in these guesses to the model formula gives a model value of height for each case:

$$\begin{array}{c} \text{Model} \\ \text{Values} \end{array} \text{ height} = 5 \begin{array}{c} \text{intercept} \\ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \end{array} + 10 \begin{array}{c} \text{age} \\ \begin{bmatrix} 5 \\ 7 \\ 10 \end{bmatrix} \end{array} = \begin{bmatrix} 55 \\ 75 \\ 105 \end{bmatrix} .$$

Arithmetic with vectors is done in the ordinary way. Coefficients multiply each element of the vector, so that  $10 \begin{bmatrix} 5 \\ 7 \\ 10 \end{bmatrix}$  gives  $\begin{bmatrix} 50 \\ 70 \\ 100 \end{bmatrix}$ . To add or subtract two vectors means to add or subtract the corresponding elements:  $\begin{bmatrix} 5 \\ 5 \\ 5 \end{bmatrix} + \begin{bmatrix} 50 \\ 70 \\ 100 \end{bmatrix}$  gives

$$\begin{bmatrix} 55 \\ 75 \\ 105 \end{bmatrix} .$$

The residuals tell you how the response values (that is, the actual heights)

differ from the model value of height. For this model, the residuals are

$$\begin{array}{c} \text{height} \\ \left[ \begin{array}{c} 80 \\ 100 \\ 110 \end{array} \right] \end{array} - \begin{array}{c} \text{Model height} \\ \left[ \begin{array}{c} 55 \\ 77 \\ 105 \end{array} \right] \end{array} = \begin{array}{c} \text{residuals} \\ \left[ \begin{array}{c} 25 \\ 23 \\ 5 \end{array} \right] \end{array}.$$

The smaller the residuals, the better the model values match the actual response values.

Try another guess of the coefficients, say  $c_1 = 40$  and  $c_2 = 8$ . Plugging in these guesses ...

$$\begin{array}{c} \text{Model} \\ \text{Values} \end{array} \text{ height} = 40 \begin{array}{c} \text{intercept} \\ \left[ \begin{array}{c} 1 \\ 1 \\ 1 \end{array} \right] \end{array} + 8 \begin{array}{c} \text{age} \\ \left[ \begin{array}{c} 5 \\ 7 \\ 10 \end{array} \right] \end{array} = \begin{array}{c} \text{Model height} \\ \left[ \begin{array}{c} 80 \\ 96 \\ 120 \end{array} \right] \end{array}$$

The residuals from the new coefficients are

$$\text{residuals} = \left[ \begin{array}{c} 80 \\ 100 \\ 110 \end{array} \right] - \left[ \begin{array}{c} 80 \\ 96 \\ 120 \end{array} \right] = \left[ \begin{array}{c} 0 \\ 4 \\ -10 \end{array} \right].$$

Which of the two model formulas is better? For the first case — the 5-year old who is 80 cm tall — the second formula is exactly right, but the first formula understates the height by 25 cm. Similarly, the second formula is better for the second case. But for the third case, the residual from the first formula is only 5 cm, while the second formula gives a bigger residual:  $-10$  cm.

It is not uncommon in comparing two formulas to find, as in this model, one formula is better for some cases and the other formula is better for other cases. This is not so different from clothing, where one item may be too snug in the waist but just right in the leg, and another item may be perfect in the waist but too long in the leg.

What's needed is a way to judge the *overall* fit. In modeling, the overall fit of a model formula is measured by a single number: the sum of squares of the residuals. For the first formula, this is  $25^2 + 23^2 + 5^2 = 1179$ . For the second formula, the sum of squares of the residuals is  $0^2 + 4^2 + (-10)^2 = 116$ . By this measure, the second formula gives a much better fit.

To find the **best** fit, just keep trying different formulas until one is found that gives the least sum of square residuals: in short, the **least squares**. The process isn't nearly so laborious as it might seem. There are systematic ways to find the coefficients that give the least sum of square residuals, just as there are systematic ways to find clothing that fits without trying on every item in the store. The details of these systematic methods are not important here. Later chapters explore the geometry of fitting, making it easy to see how a computer can quickly find the coefficients that give the best fitting model formula.

Using the least squares criterion, a computer will quickly find the best coefficients for the height vs age model to be  $c_1 = 54.211$  and  $c_2 = 5.789$ . Plugging in these coefficients to the model formula produces fitted model values of  $\left[ \begin{array}{c} 83.158 \\ 94.737 \\ 112.105 \end{array} \right]$ , residuals of  $\left[ \begin{array}{c} -3.158 \\ 5.263 \\ -2.105 \end{array} \right]$ , and therefore a sum of square residuals of 42.105, clearly much better than either previous two model formulas.

How do you know that the coefficients that the computer gives are indeed

the best? You can try varying them in any way whatsoever. Whatever change you make, you will discover that the sum of square residuals is never going to be less than it was for the coefficients that the computer provided.

The smallest possible sum of square residuals is zero. This can occur only when the model formula gives the response values exactly, that is, when the fitted model values are an exact match with the response values. Because models typically do not capture all of the variation in the response variable, usually you do not see a zero sum of square residuals. When you do, however, it may well be a sign that there is something wrong, that your model is too detailed.

It's important to keep in mind that the coefficients found using the least squares criterion are the best possible but only in a fairly narrow sense. They are the best *given* the design of the model and *given* the data used for fitting. If the data change (for example, you collect new cases), or if you change the model design, the best coefficients will likely change as well. It makes sense to compare the sum of square residuals for two model formulas only when the two formulas have the same design and are fitted with the same data. So don't try to use the sum of square residuals to decide which of two different designs are better. Tools to do that will be introduced in later chapters.

Why the sum of square residuals? Why not some other criterion? The answers are in part mathematical, in part statistical, and in part historical.

In the late 18th century, three different criteria were competing, each of which made sense:

1. The least squares criterion, which seeks to minimize the sum of square residuals.
2. The least absolute value criterion. That is, rather than the sum of square residuals, one looks at the sum of absolute values of the residuals.
3. Make the absolute value of the residual as small as possible for the single worst case.

All of these criteria make sense in the following ways. Each tries to make the residuals small, that is, to make the fitted model values match closely to the response values. Each treats a positive residual in the same way as a negative residual; a mismatch is a mismatch regardless of whether the fitted values are too high or too low. The first and second criteria each combine all of the cases in determining the best fit.

The third criterion is hardly ever used because it allows one or two cases to dominate the fitting process — only the two most extreme residuals count in evaluating the model. Think of the third criterion as a dead end in the historical evolution of statistical practice.

Computationally, the least squares criterion leads to simpler procedures than the least absolute value criterion. In the 18th century, when computing was done by hand, this was a very important issue. It's no longer so, and there can be good statistical reasons to favor a least absolute value criterion when it is thought that the data may contain outliers.

The least squares criterion is best justified when variables have a bell-shaped distribution. Insofar as this is the situation — and it often is — the least squares criterion is arguably better than least absolute value.

Another key advantage of the least squares criterion is in interpretation, as you'll see in the next section.

## 6.2 Partitioning Variation

A model is intended to explain the variation in the response variable using the variation in the explanatory variables. It's helpful in quantifying the success of this to be able to measure how much variation there is in the response variable, how much has been explained, and how much remains unexplained. This is a partitioning of variation into parts.

You might think that such a partitioning would always be possible and would always make sense, but in fact it depends on how one chooses to measure variation. To illustrate, suppose you and your friend earn \$50 by door-to-door bagel deliveries. You decide to partition this: you keep \$30 and your friend gets \$20. (Perhaps you worked harder than your friend.) Whichever way you partition the money between you and your friend, the total amount will always equal exactly the sum of the amounts that you and your friend get. Obvious.

But suppose you decided — bizarrely, admittedly — to measure money not in the ordinary sense but as “square root dollars.” The two of you start out with  $\sqrt{\$50} = 7.07$  square-root dollars. After splitting it up, You have  $\sqrt{\$30} = 5.48$  square dollars and your friend gets  $\sqrt{\$20} = 4.47$  square-root dollars. Notice that the partitioning doesn't work:  $7.07 \neq 5.48 + 4.47$ . That's one reason why it's natural to stick with counting money in an ordinary way, and not with silly square-root dollars.

Now consider how variation in the response variable is broken into parts by a model, partitioned into the variation in the fitted model values and the variation in the residuals.

One way to measure the variation is with the standard deviation. (See Chapter 3.) The table below gives the standard deviations of the response variable, the fitted model values, and the residuals for the best-fitting model (found by the computer) for height versus age:

Source	Standard Deviation
Fitted Model values	14.570
+ Residuals	4.588
≠ Response Variable (height)	15.275

A little arithmetic shows that the partitioning does not work:  $15.275 \neq 14.570 + 4.588$ .

The problem is that measuring variability with a standard deviation is very much like using “square-root dollars.”

If you want the partitioning to work, you have to measure variability in the right way. One of the consequences of using the least squares criterion to fit models is that there is actually a simple way to measure variability that does produce a meaningful partitioning: measure variability with the variance.

	Source	Variance
	Fitted Model values	212.28
+	Residuals	21.05
=	Response Variable (height)	233.33

Now the partitioning works:  $233.33 = 212.28 + 21.05$ .

The variance — and its square root, the standard deviation — measure deviation from a central value. Although the standard deviation has nicer units, the variance is best from a deeper point of view: it provides a natural partitioning of variation.

Another sensible choice is the sum of squares, which also permits the partitioning of the response variable into parts: the fitted model values and the residuals.

	Source	Sum of Squares
	Fitted Model values	28457.89
+	Residuals	42.11
=	Response Variable (height)	28500.00

Again, the partitioning works:  $28500.00 = 28457.89 + 42.11$ .

Perhaps this use of squares — the variance, the sum of squares — reminds you of another kind of partitioning. The **Pythagorean theorem** says that the length of the hypotenuse of a right triangle and the lengths of the two other sides are related like this:  $A^2 = B^2 + C^2$ . This is a kind of partitioning: the square length of the hypotenuse can be partitioned into two other square lengths: those of the legs.

In later chapters you'll see that there is a close connection between the right-triangle geometry, the least-squares criterion for fitting, and the partitioning of variation.

**Example 6.1: Residuals in Global Temperature** In studying human-induced global climate change, one issue is the extent to which fluctuations in global temperatures reflect naturally occurring climate variability. A strategy for addressing this is to model global temperature using as explanatory variables measurements of year-to-year natural variability in some climate systems.

Figure 6.1 shows a record of global mean temperature from the late 1800s through this century from Thompson et al. [11]. The global mean temperature shows a slow increase of the sort described by the phrase “global warming.” But there is also a sharp dip in the middle. Might this be due to short term fluctuations, such as those due to the El Niño/Southern Oscillation (ENSO) or the “cold oceans-warm land” (COWL) system in the Northern Hemisphere?



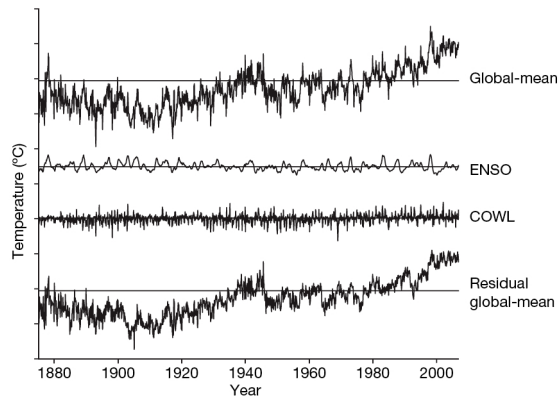


Figure 6.1: Global mean temperatures (top) partitioned into variation associated with natural climate variation (“ENSO” and “COWL”) and residual variation (at the bottom). From Thompson *et al.* [11]

In the figure, the global mean temperature data has been partitioned into three components: (1) variation associated with ENSO, (2) variation associated with COWL, and (3) the residuals. The residuals show a long-term warming trend, but they also show a sharp short-term discontinuity in 1945 when the residuals suddenly drop. This suggests that the drop in 1945 is not due to ENSO or COWL.

Thompson *et al.* investigated the 1945 discontinuity, in part by considering additional explanatory variables such as timing of large volcanic eruptions. They traced the 1945 discontinuity to an interesting form of sampling bias. During World War II, the fraction of sea-surface measurements made by US ships increased steadily, up to about 80% of all measurements in August 1945. Then there was a sudden drop to about 50%, as measurements from UK ships increased. US ships tended to make measurements from engine-room intakes; these are biased warmer than the ambient sea water. UK ships tended to use uninsulated buckets to make measurements; these are biased cold. When the mix of measurements shifted from the US to the UK, so did the bias in the measurements.

Residuals are not necessarily random or impossible to explain, they are just the part of the variation in your response variable that is not explained by your model.

### 6.3 Redundancy

Data sets often contain variables that are closely related. For instance, consider the following data on prices of used cars (the Ford Taurus model), the year the

car was built, and the number of miles the car has been driven.

Price	Year	Mileage
14997	2008	22613
8995	2007	53771
7990	2006	36050
18990	2008	25896

... and so on.

Although the year and mileage variables are different, they are related: older cars typically have higher mileage than newer cars. A group of students thought to see how price depends on year and mileage. After collecting a sample of 635 cars advertised on the Internet, they fit a model  $\text{price} \sim \text{mileage} + \text{year}$ . The coefficients came out to be:

price $\sim$ mileage + year		
Term	Coefficient	Units
Intercept	-1136311.500	dollars
mileage	-0.069	dollars/mile
year	573.500	dollars/year

The coefficient on mileage is straightforward enough; according to the model the price of these cars typically decreases by 6.9 cents per mile. But the intercept seems strange: very big and negative. Of course this is because the intercept tells the model value at zero mileage in the year zero! They checked the plausibility of the model by plugging in values for a 2006 car with 50,000 miles:  $-1136312 - 0.069 \times 50000 + 2006 \times 573.50$  giving \$10679.50 — a reasonable sounding price for such a car.

The students thought that the coefficients would make more sense if they were presented in terms of the age of the car rather than the model year. Since their data was from 2009, they calculated the age by taking  $2009 - \text{year}$ . Then they fit the model  $\text{price} \sim \text{mileage} + \text{age}$ .

price $\sim$ mileage + age		
Term	Coefficient	Units
Intercept	15850.000	dollars
mileage	-0.069	dollars/mile
age	-573.500	dollars/year

The mileage coefficient is unchanged, but now the intercept term looks much more like a real price: the price of a hypothetical brand new car with zero miles. The age coefficient makes sense; typically the cars decrease in price by \$573.50 per year of age. Again, the students calculated the model for a 2006 car (thus age 3 years in 2009) with 50,000 miles: \$10679.50, the same as before.

Surprised that two different sets of coefficients could give the same model value, the students computed the fitted model value for all the cars in their data set for both models. Exactly the same for every car.

It makes sense that the two models should give the same model values. They are based on exactly the same information. The only difference is that in one model the age is given by the model year of the car, in the other by the age itself.

For review purposes only

The students decided to experiment. “What happens if we include both age and year in the model?” They started with the model  $\text{price} \sim \text{mileage} + \text{year} + \text{age}$ . The software responded by giving coefficients for the first three model terms — the same intercept, mileage and year coefficients as in the first table — but reported the age coefficient as “NA,” not available. Why?

Actually, there is no mathematical reason why the computer could not have given coefficients for all four model terms. For example, any of the following sets of coefficients would give exactly the same model values for *any* inputs of year and mileage, with age being appropriately calculated from year.

Term	Set 1	Set 2	Set 3	Set 4
Intercept	15850.000	-185050.000	-734511.500	-1136311.500
mileage	-0.069	-0.069	-0.069	0.069
year	0.000	100.000	373.500	573.500
age	-573.500	-473.500	-200.000	0.000

The overall pattern is that the coefficient on age minus the coefficient on year has to be  $-573.5$ . This works because age and year are basically the same thing since  $\text{age} = 2009 - \text{year}$ .

The dual identity of age and year is **redundancy**. It arises whenever an explanatory model vector can be modeled exactly in terms of the other explanatory vectors. Here, age can be computed by 2009 times the intercept plus  $-1$  times the year.

The problem with redundancy is that it creates ambiguity. Which of the four sets of coefficients in the table above is the right one? Whenever there is redundancy, there is no unique set of coefficients that gives the best fit of the model to the data.

In order to avoid this ambiguity, modeling software is written to spot any redundancy and drop the redundant model vectors from the model. One possibility would be to report a coefficient of zero for the redundant vector. But this could be misleading. For instance, the user might interpret the zero coefficient on Set 1 above as meaning that year isn’t associated with price. But that would be wrong; year is a very important determinant of price, it’s just that all the relationship is represented by the coefficient on age in Set 1. Rather than report a coefficient of zero, it’s helpful when software reports the redundancy by displaying a NA. This makes it clear that it is redundancy that is the issue and not that a coefficient is genuinely zero.

### Example 6.2: Almost redundant

When model vectors are redundant, modeling software can spot the situation and do the right thing. But sometimes, the redundancy is only approximate. In such cases it’s important for you to be aware of the potential for problems.

In collecting the Current Population Survey wage data, interviewers asked people their age and the number of years of education they had. From this, they computed the number of years of experience. They presumed that kids spend six years before starting school. So a 40-year old with 12 years of education would have 22 years of experience:  $6 + 12 + 22 = 40$ . This creates redundancy

for review purposes only

DATA FILE  
cps.csv

of experience with age and education. Ordinarily, the computer could identify such a situation. However, there is a mistake in the CPS data. Case 350 is an 18-year old woman with 16 years of education. This seems hard to believe since very few people start their education at age 2. Presumably, either the age or education were mis-recorded. But either way, this small mistake in one case means that the redundancy among experience, age, and education is not exact. It also means that standard software doesn't spot the problem when all three of these variables are included in a model. As a result, coefficients on these variables are highly unreliable.

The situation of approximate redundancy is called **multi-collinearity**. With multi-collinearity, as opposed to exact redundancy, there is a unique least squares fit of the model to the data. However, this uniqueness is hardly a blessing. What breaks the tie to produce a unique best fitting set of coefficients are the small deviations from exact redundancy. These are often just a matter of random noise, arithmetic round-off, or mistakes as with case 350 in the Current Population Survey data. As a result, the choice of the winning set of coefficients is to some extent arbitrary and potentially misleading.

The costs of multi-collinearity can be measured when fitting a model. (See section 14.7.) When the costs of multi-collinearity are too high, the modeler often chooses to take terms out of the model.

## 6.4 Computational Technique

Using the `lm` software is mainly a matter of familiarity with the model design language. Computing the fitted model values and the residuals is done with the `fitted` and `resid`. These operators take a model as an input. To illustrate:

```
swim = ISMdata("swim100m.csv")
mod1 = lm( time ~ year + sex, data=swim)
```

Once you have constructed the model, you can use `fitted` and `resid`:

```
> fitted(mod1)
  1      2      3      4      5      6      7      8
66.88 66.13 65.62 65.12 63.61 63.11 62.61 62.10
... and so on ...
 55     56     57     58     59     60     61     62
58.82 58.32 57.82 56.31 54.80 54.30 52.79 51.78
> resid(mod1)
  1      2      3      4      5      6
-1.081 -0.526 -2.823 -3.520 -2.212 -2.709
... and so on.
```



Sometimes it's helpful to look for outliers in the residuals, and to plot the residuals versus the fitted model values or versus explanatory variables. For instance:

```
> bwplot( as.numeric(resid(mod1)) )
> subset(swim, outlier(resid(mod1)))
   year time sex
32 1908 95.0  F
33 1910 86.6  F
34 1911 84.6  F
> xyplot( resid(mod1) ~ fitted(mod1) )
> xyplot( resid(mod1) ~ year, data=swim)
```

(Technical note: The `as.numeric` operator translates the residuals to a format that `bwplot` will work with.)

### 6.4.1 Sums of Squares

Computations can be performed on the fitted model values and the residuals, just like any other quantity:

```
> mean( fitted(mod1))
[1] 59.92
> var( resid(mod1))
[1] 15.34
> sd( resid(mod1))
[1] 3.917
> summary( resid(mod1))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-4.70  -2.70   -0.597  2.98e-16  1.28    19.1
```

Sums of squares are very important in statistics. Here's how to calculate them for the response values, the fitted model values, and the residuals:

```
> sum( swim$time^2 )
[1] 228635
> sum( fitted(mod1)^2 )
[1] 227699
> sum( resid(mod1)^2 )
[1] 935.8
```

The partitioning of variation by models is seen by the way the sum of squares of the fitted and the residuals add up to the sum of squares of the response:

```
> 227699 + 935.8
[1] 228635
```

Don't forget the squaring stage of the operation! The sum of the residuals (without squaring) is very different from the sum of squares of the residuals:

```
> sum( resid(mod1) )
[1] 1.849e-14
> sum( resid(mod1)^2 )
[1] 935.8
```

Take care in reading numbers formatted like 1.849e-14. The notation stands for  $1.849 \times 10^{-14}$ . That number, 0.00000000000001849, is effectively zero compared to the residuals themselves!

## 6.4.2 Redundancy

The `lm` operator will automatically detect redundancy and deal with it by leaving the redundant terms out of the model.

To see how redundancy is handled, here is an example with a constructed redundant variable in the swimming dataset. The following statement adds a new variable to the dataframe counting how many years after the end of World War II each record was established:

```
> swim$afterwar = swim$year - 1945
```

Here is a model that doesn't involve redundancy

```
> mod1 = lm( time ~ year + sex, data=swim)
> coef(mod1)
(Intercept)      year      sexM
  555.7168    -0.2515    -9.7980
```

When the redundant variable is added in, `lm` successfully detects the redundancy and handles it. This is indicated by a coefficient of NA on the redundant variable.

```
> mod2 = lm( time ~ year + sex + afterwar, data=swim)
> coef(mod2)
(Intercept)      year      sexM  afterwar
  555.7168    -0.2515    -9.7980         NA
```

In the absence of redundancy, the model coefficients don't depend on the order in which the model terms are specified. But this is not the case when there is redundancy, since any redundancy is blamed on the later variables. For instance, here `afterwar` has been put first in the explanatory terms, so `lm` identifies `year` as the redundant variable:

```
> mod3 = lm( time ~ afterwar + year + sex, data=swim)
> coef(mod3)
(Intercept)  afterwar      year      sexM
  66.6199    -0.2515         NA    -9.7980
```

Even though the coefficients are different, the fitted model values and the residuals are exactly the same (to within computer round-off) regardless of the order of the model terms.

```
> fitted(mod2)
      1      2      3      4      5      6
66.88 66.13 65.62 65.12 63.61 63.11 and so on.
> fitted(mod3)
      1      2      3      4      5      6
66.88 66.13 65.62 65.12 63.61 63.11 and so on.
```

Note that whenever you use a categorical variable and an intercept term in a model, there is a redundancy. This is not shown explicitly. For example, here is a model with no intercept term, and both levels of the categorical variable sex show up with coefficients:

```
> lm( time ~ sex - 1, data=swim)
Coefficients:
sexF  sexM
65.2  54.7
```

If the intercept term is included (as it is by default unless - 1 is used in the model formula), one of the levels is simply dropped in the report:

```
> lm( time ~ sex, data=swim)
Coefficients:
(Intercept)      sexM
      65.2      -10.5
```

Remember that this coefficient report implicitly involves a redundancy. If the software had been designed differently, the report might look like this:

```
(Intercept)      sexF      sexM
      65.2      NA      -10.5
```

### 6.4.3 Technical Notes: Missing Data

Occasionally, you may have a data frame with missing data. The `lm` program will handle this sensibly by excluding those cases where one or more of the variables used in the model are missing. Those cases with missing data will show up in the lists of residuals and fitted values as NAs. (This depends on the setting of the `na.action` argument to `lm`. In the ISM library, this has been set to `na.exclude`.)

The NAs can cause a problem in calculations involving the residuals or the fitted values. For instance, suppose you have constructed a model called `mymodel` fitted to a data frame with missing data. Here is the sum of squares of the fitted values:

```
> sum( fitted(mymodel)^2 )
[1] NA
```

Any NAs in the fitted values propagate through to the overall sum.

There are a couple of ways to deal with this:

```
> sum( fitted(mymodel)^2, na.rm=TRUE )  
[1] 683434.3  
> sum( na.omit(fitted(mymodel))^2 )  
[1] 683434.3
```

You can also delete any cases with missing data from the data frame *before* fitting the model:

```
> cleaned = na.omit(swim)
```

Be careful, however, since this will exclude any cases that have any missing variables, even if those variables are not involved in the model.

For review purposes only



# Measuring Correlation

*The invalid assumption that correlation implies cause is probably among the two or three most serious and common errors of human reasoning.* — Stephen Jay Gould

*“Co-relation or correlation of structure” is a phrase much used ... but I am not aware of any previous attempt to define it clearly, to trace its mode of action in detail, or to show how to measure its degree.’* — Francis Galton, 1888

The last chapter described how the variance of the response variable can be divided into two parts: the variance of the fitted model values and the variance of the residuals. This partitioning is at the heart of a statistical model; the more of the variation that’s accounted for by the model, the less is left in the residuals.

Because of the partitioning, an effective way to summarize a model is the proportion of the total variation in the response variable that is accounted for by the model. This description is called the  $R^2$  (“**R-Squared**”) of the model. It is a ratio:

$$R^2 = \frac{\text{variance of fitted model values}}{\text{variance of response values}}.$$

Another name for  $R^2$  is the **coefficient of determination**, but this is not a coefficient in the same sense used to refer to a multiplier in a model formula.

## 7.1 Properties of $R^2$

$R^2$  has a nice property that makes it easy to interpret: its value is always between zero and one. When  $R^2 = 0$ , the model accounts for none of the variance of the response values: the model is useless. When  $R^2 = 1$ , the model captures all

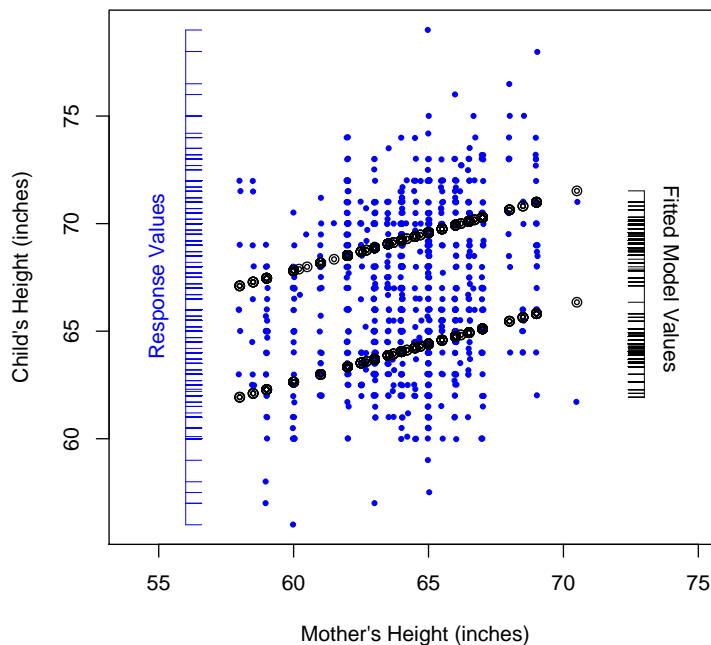


Figure 7.1: The small dots show child's height versus mother's height from Galton's data. The bigger symbols show the fitted model values from the linear model  $\text{height} \sim 1 + \text{mother} + \text{sex}$  that includes both mother's height and child's sex as explanatory variables.

of the variance of the response values: the model values are exactly on target. Typically,  $R^2$  falls somewhere between zero and one, meaning that the model accounts for some, but not all, of the variance in the response values.

The history of  $R^2$  can be traced to a paper[12] presented to the Royal Society in 1888 by Francis Galton. It is fitting to illustrate  $R^2$  with some of Galton's data: measurements of the heights of parents and their children.

Figure 7.1 shows a scatter plot of the children's height versus mother's height. Also plotted are the fitted model values for the model  $\text{height} \sim \text{mother} + \text{sex}$ . The model values show up as two parallel lines, one for males and one for females. The slope of the lines shows how the child's height typically varies with the mother's height.

Two rug plots have been added to the scatter plot in order to show the distribution of the response values (child's height) and the fitted model values (the output of the model). The rug plots are positioned vertically because each of them relates to the response variable height, which is plotted on the vertical

DATA FILE  
galton.csv

For review purposes only

For review purposes only

# Chapter 8

## Total and Partial Relationships

*I do not feel obliged to believe that the same God who has endowed us with sense, reason, and intellect has intended us to forgo their use. —Galileo Galilei*

One of the most important ideas in science is **experiment**. In a simple, ideal form of an experiment, you cause one explanatory factor to vary and, holding all the other conditions constant, observe the result on some other variable. A famous story of such an experiment involves Galileo Galilei (1564-1642) dropping balls of different masses but equal diameter from the Leaning Tower of Pisa.<sup>1</sup> Would a heavy ball fall faster than a light ball, as theorized by Aristotle 2000 years previously? The quantity that Galileo varied was the weight of the ball, the quantity he observed was how fast the balls fell, the conditions he held constant were the height of the fall and the diameter of the balls. The experimental method of dropping balls side by side also holds constant the atmospheric conditions: temperature, humidity, wind, air density, etc.

Today, Galileo's experiment seems obvious. But not at the time. In the history of science, Galileo's work was a landmark: he put *observation* at the fore, rather than the beliefs passed down from authority. Aristotle's ancient theory, still considered authoritative in Galileo's time, was that heavier objects fall faster.

The ideal of "holding all other conditions constant" is not always so simple as with dropping balls from a tower. Consider an experiment to test the effect of a blood-pressure drug. Take two groups of people, give the people in one group the drug and give nothing to the other group. Observe how blood pressure changes in the two groups. The factor being caused to vary is whether or not a person gets the drug. But what is being held constant? Presumably the researcher took care to make the two groups as similar as possible: similar med-

<sup>1</sup>The picturesque story of balls dropped from the Tower of Pisa may not be true. Galileo did record experiments done by rolling balls down ramps.

ical conditions and histories, similar weights, similar ages. But “similar” is not “constant.”

For non-experimentalists — people who study data collected through observation, without doing an experiment — a central question is whether there is a way to mimic “holding all other conditions constant.” For example, suppose you observe the academic performance of students, some taught in large classes and some in small classes, some taught by well-paid teachers and some taught by poorly-paid teachers, some coming from families with positive parental involvement and some not, and so on. Is there a way to analyze data so that you can separate the influences of these different factors, examining one factor while, through analysis if not through experiment, holding the others constant?

In this chapter you’ll see how models can be used to examine data as if some variables were being held constant. Perhaps the most important message of the chapter is that there is no point hiding your head in the sand; simply ignoring a variable is not at all the same thing as holding that variable constant. By including multiple variables in a model you make it possible to interpret that model in terms of holding the variables constant. But there is no methodological magic at work here. The results of modeling can be misleading if the model does not reflect the reality of what is happening in the system under study. Understanding how and when models can be used effectively, and when they can be misleading, will be a major theme of the remainder of the book.

## 8.1 Total and Partial Relationships

The common phrase “all other things being equal” is an important qualifier in describing relationships. To illustrate: A simple claim in economics is that a high price for a commodity reduces the demand. For example increasing the price of heating fuel will reduce demand as people turn down thermostats in order to save money. But the claim can be considered obvious only with the qualifier *all other things being equal*. For instance, the fuel price might have increased because winter weather has increased the demand for heating compared to summer. Thus, higher prices may be associated with higher demand. Unless you hold other variables constant — e.g., weather conditions — increased price may not in fact be associated with lower demand.

In fields such as economics, the Latin equivalent of “all other things being equal” is sometimes used: **ceteris paribus**. So, the economics claim would be, “higher prices are associated with lower demand, *ceteris paribus*.”

Although the phrase “all other things being equal” has a logical simplicity, it’s impractical to implement “all.” Instead of the blanket “all other things,” it’s helpful to be able to consider just “some other things” to be held constant, being explicit about what those things are. Other phrases along these lines are “taking into account ...” and “controlling for ....” Such phrases apply when you want to examine the relationship between two variables, but there are additional variables that may be coming into play. The additional variables are sometimes called **covariates** or **confounders**.

# Chapter 9

## Model Vectors

*By the aid of symbolism, we can make transitions in reasoning almost mechanically by the eye, which otherwise would require the higher faculties of the brain. ... Civilization advances by extending the number of important operations which we can perform without thinking about them.*

— Alfred North Whitehead (1861-1947)

Many people are disappointed to trade the graphical display of models (as in Chapter 4) for model formulas and coefficients (as in Chapter 5). The formulas give an added ability to model complicated relationships involving multiple explanatory variables, but at the cost of losing touch. Without graphics, people can't apply their powerful human facility to visualize or make use of our cognitive strengths in recognizing spatial relationships.

The link between reasoning and our ability to perceive things in space is powerful and often unappreciated. It shows up, for instance, in the physical and spatial metaphors used to describe reasoning, learning, and understanding: turn it over, see it from a new perspective, get a grasp on it, take it apart, illuminate, pick up, take in, compare side-by-side, run through, sort out, put it together, have it at one's fingertips ....

This chapter introduces the tools and concepts needed to think about complicated, multivariate models using our powerful spatial intuition. The spatial/geometrical approach will make it easier to visualize and intuit the process of fitting, the sources of redundancy, and the meaning of correlation. The approach will also prepare you to apply statistical reasoning to your models: setting the explanations of our models in the context of what remains unexplained in the data.

The approach is founded on basic geometry: lengths, angles, directions, etc. But it is not the sort of scientific and statistical graphics you are used to: scatterplots of individual cases, histograms, etc. So be prepared to drop some of your preconceptions about the types of graphs that can display relationships.

For review purposes only

## 9.1 Vectors

A **vector** is a mathematical idea that is deeply rooted in everyday physical experience. A vector is simply something that specifies a **length** and a **direction**. It lives in a space of some **dimension**.


You can draw vectors on a piece of paper. The paper is the two-dimensional space that the vectors inhabit. The convention is to draw a vector as an arrow, as in Figure 9.1. Note that each vector has a length that you can measure with a ruler. Each vector points in some direction.

Figure 9.1 shows vectors in different spaces: a one-dimensional space, a two-dimensional space, and a three-dimensional space. On a piece of paper, only vectors in a 2-dimensional space can be drawn in a natural way. But vectors can exist in three dimensions: point a pencil — it has a length and a direction. Drawing such vectors on paper requires tricks of perspective and adding a triple-coordinate axis in order to orient you.

You can also think of vectors in one-dimension, as if they were directions on a railroad track. Drawing such vectors requires laying down a track along the paper — a thin line — to emphasize that the vectors live only in a small part of the paper.

The mathematical notion of vectors is abstract: they don't have to be drawn or embodied as a pencil. One of the aspects of abstraction concerns position. Obviously, each of the vectors drawn in Figure 9.1 has a position where it has been drawn. But position is not an attribute of a mathematical vector; the only quantities that matter are length and direction.

If two vectors have the same length and direction, they are the same vector. Within each panel of Figure 9.1 there are two vectors with the same label. These two vectors are exactly the same — they have the same length and direction even though they have been drawn in different places.

You can position a mathematical vector wherever it is convenient for your purposes. For example, you can ask what is the **angle** between two vectors. To answer this, pick up the vectors — maintaining their direction — and position them tail to tail, like this:  Then you can measure the angle with a protractor.

You can draw vectors as arrows, but a good way to think about them is in terms of movement: as steps. A vector is an instruction to take a step. Suppose one vector indicates a movement 1 meter to the north-east, another vector a movement 3 meters to the south. Such instructions make sense even without needing to specify where you are right now: position doesn't enter into it. There are two simple mathematical operations on such movements — scaling and addition — as shown in Figure 9.2.

**Scale a vector.** Make a change to a vector's length to make it longer or shorter, but keep the direction the same. Suppose the vector says, "move one meter to the north-east." Scaling the vector by a factor of 3 means to make it 3 times longer: "move three meters to the north-east." Scaling the vector by a factor of 0.5 means to make it half as long: "move  $\frac{1}{2}$  meter to the north-east." A negative scale means to step backwards, so a scale of  $-2$  means

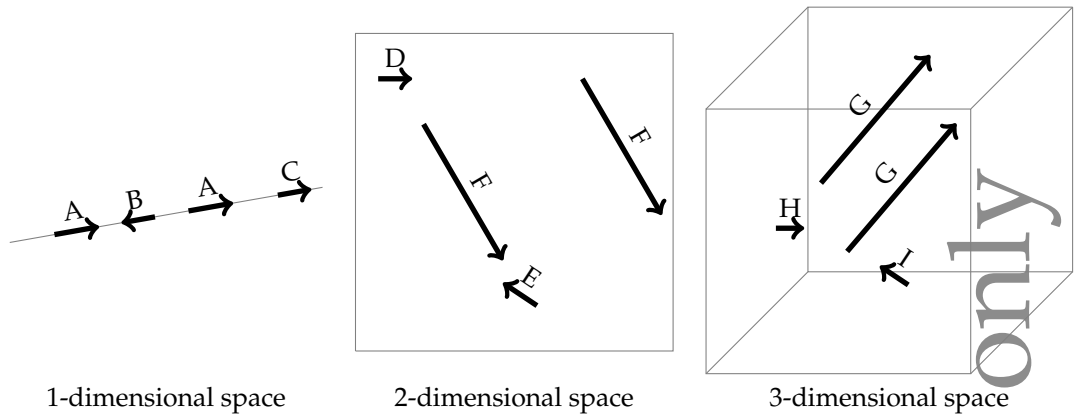


Figure 9.1: Some vectors in different spaces.

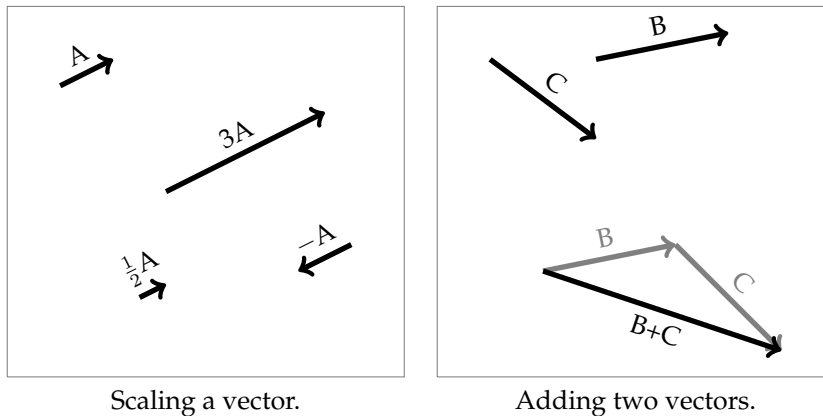


Figure 9.2: Scaling and adding vectors.

to move 2 meters backwards while facing north-east, effectively moving 2 meters to the south-west.

**Add two vectors.** Pick a place to start. Take one step according to the length and direction of the first vector. Then, from the point you ended up after the step, take another step according to the length and direction of the second vector. The overall distance and direction from the place you started to the place you ended up is the vector that results from adding up the two vectors.

Mathematics extends the notion of movement to high dimensions. Talking about 1- or 2- or 3-dimensional objects makes intuitive sense. An arrow drawn on a piece of paper inhabits a 2-dimensional space; a pencil lives in a three-dimensional space. But what is a 4-dimensional space, or a space that is 100-dimensional? The key to understanding is to consider length and direction abstractly without tying them to a physical embodiment like an arrow.



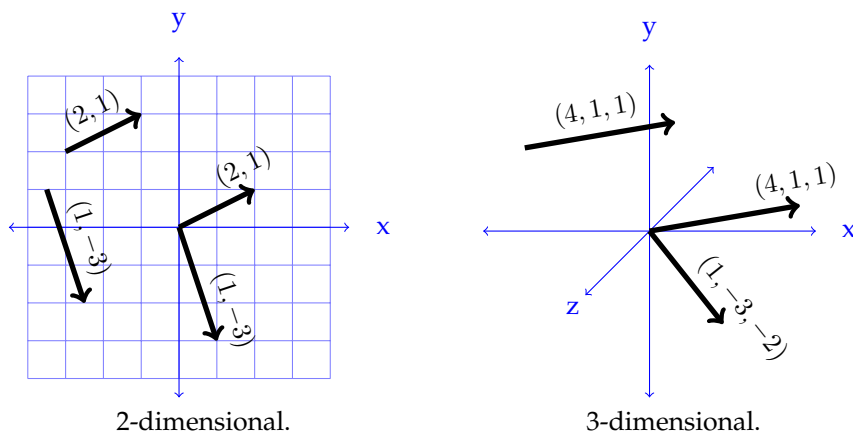


Figure 9.3: Vectors as coordinates.

For vectors in 2-dimensional space, you can specify a length and direction with coordinates on the Cartesian plane. (See Figure 9.3.) Most people are used to using the Cartesian plane to plot out position, but it works just as well for direction and length.

Figure 9.3 shows some vectors specified as coordinates. For example,  $(2, 1)$  means move two units to the right and 1 unit up, altogether a length of 2.236 units directed to the east-north-east.<sup>1</sup> Where you end up — the position of the arrow head — depends on where you started. The vector, however, isn't the position, it's the step.

You can use the same idea of Cartesian coordinates to think about 3-dimensional vectors. Three coordinates are required — one along each axis. In geography, the coordinates system might be the east-west axis, the north-south axis, and the up-down axis. In a room, a convenient coordinate axis can be given by the three lines emerging from the corner: (1) the line where the two walls meet, (2) the line where one wall and the floor meet, and (3) the line where the other wall and the floor meet. What's key is that three numbers are needed to specify a coordinate in 3-dimensional space.

The generalization to 4- and higher-dimensional spaces is straightforward. A 4-dimensional vector is represented by 4 numbers, a 5-dimensional vector by 5 numbers, and a 100-dimensional vector by 100 numbers. Although you can't see a 100-dimensional vector in the same way that you can see a pencil, it works perfectly well to think about it using your 2- and 3-dimensional capabilities.

## 9.2 Vectors as Collections of Numbers

The direct connection of vectors and statistical models comes from the representation of a vector as a collection of numbers, just as a quantitative variable or an

<sup>1</sup>2.236 comes from the Pythagorean theorem: the movements to the right and up are the legs of a right triangle and the vector described is the hypotenuse of that triangle with length  $\sqrt{2^2 + 1^2}$ .

indicator variable is a collection of numbers. In this direct sense, vectors merely organize the calculations involved in fitting models. The indirect, geometrical connection with models is richer and will help in interpreting and understanding models.

As a collection of numbers, a vector is written conventionally as a column of numbers, for instance  $\begin{bmatrix} 5 \\ 9 \\ 2 \\ 3 \end{bmatrix}$  is a vector in 4-dimensional space.

The two vector operations of scaling and vector addition have a simple arithmetic form.

To perform **vector addition**, add the corresponding terms of each vector. For instance

$$\begin{bmatrix} 5 \\ 9 \\ 2 \\ 3 \end{bmatrix} + \begin{bmatrix} 6 \\ 8 \\ 103 \\ -7 \end{bmatrix} \text{ gives } \begin{bmatrix} 11 \\ 17 \\ 105 \\ -4 \end{bmatrix}.$$

Of course, the two vectors being added have to be the same dimension, that is, they have to have the same number of components, 4 in this case.

**Scaling** a vector amounts to multiplication. For instance, to scale a vector to three times its length, multiply each component in the vector by 3:

$$3 \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \text{ gives } \begin{bmatrix} 3 \\ 0 \\ 3 \\ 3 \end{bmatrix} \quad 3 \begin{bmatrix} 5 \\ 9 \\ 2 \\ 3 \end{bmatrix} \text{ gives } \begin{bmatrix} 15 \\ 27 \\ 6 \\ 9 \end{bmatrix}.$$

Notice that both vector addition and vector scaling produce an output that is a vector of the same dimension as the input vector or vectors.

A **linear combination** of vectors involves scaling the vectors and adding up the result. For instance, consider the two vectors  $\begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}$  and  $\begin{bmatrix} 2 \\ 4 \\ -1 \\ 8 \end{bmatrix}$ . Here is a linear combination of them:

$$3 \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} - 6 \begin{bmatrix} 2 \\ 4 \\ -1 \\ 8 \end{bmatrix} \text{ gives } \begin{bmatrix} -9 \\ -24 \\ 9 \\ -45 \end{bmatrix}.$$

Such combinations are at the heart of statistical modeling.

## 9.3 Model Vectors

Recall the distinction between variables and model terms that you've already seen. A modeler chooses the explanatory variables of interest and then needs to specify which terms involving these variables are to be included in a model design.

There is another step in the process that happens automatically: the translation from model terms to **model vectors**. This step doesn't involve any decision making or choice by the modeler, but mastering it is important in understanding and interpreting models.

In the translation each model term becomes one or more model vectors according to these rules:

1. The **intercept term** becomes a vector of all ones.

2. A **quantitative variable**, or a transformation of a quantitative variable, is already a collection of numbers, so it is already a vector.
3. A **categorical variable** becomes a set of vectors, the indicator vectors for each level of the variable.
4. An **interaction** between two terms becomes a vector by multiplying together the vectors from each of the terms. If one of the terms involves a categorical variable and is therefore a set of vectors, multiply each member of the set by the vector or vectors from the other terms.

DATA FILE  
cps.csv

**Example 9.1: Variables, Model Terms, and Model Vectors** Consider this very small subsample from the Current Population Survey wage data.

wage	educ	sex	married	age	sector
12.00	12	M	Married	32	manuf
8.00	12	F	Married	33	service
16.26	12	M	Single	32	service
13.65	16	M	Married	33	prof
8.50	17	M	Single	26	clerical

The categorical variable `sex` has two levels: F and M. It therefore is a set of two model vectors, the indicators for F and M respectively:

$$\begin{array}{c} \text{sexF} \\ \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{array} \quad \text{and} \quad \begin{array}{c} \text{sexM} \\ \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \end{array} .$$

The little label over each vector reminds you where the vector comes from. The label is not a part of the vector — the vector is just the collection of numbers.

The categorical variable `married` has two vectors, one for each of its levels Married and Single. The indicator vectors are:

$$\begin{array}{c} \text{Married} \\ \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \end{array} \quad \text{and} \quad \begin{array}{c} \text{Single} \\ \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \end{array}$$

The variable `sector` has eight levels — clerical, construction, manufacturing, professional, service, management, sales, and “other” — but only four of them show up in this small subset. The eight indicator vectors are:

$$\begin{array}{c} \text{clerical} \\ \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{constr} \\ \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{manuf} \\ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{prof} \\ \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{service} \\ \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{manag} \\ \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{array} \quad \begin{array}{c} \text{sales} \\ \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{array} \quad \text{and} \quad \begin{array}{c} \text{other} \\ \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{array} .$$

Notice that some of these vectors are all zeros. That’s because none of the five cases in this subsample happened to take on the levels for construction, management, sales, or “other.”

The vectors from the quantitative variables `wage`, `educ`, `age` are simply reiterations of the values of those variables.

For review purposes only

$$\begin{array}{ccc} \text{wage} & \text{educ} & \text{age} \\ \begin{bmatrix} 12.00 \\ 8.00 \\ 16.26 \\ 13.65 \\ 8.50 \end{bmatrix} & \begin{bmatrix} 12 \\ 12 \\ 12 \\ 16 \\ 17 \end{bmatrix} & \begin{bmatrix} 32 \\ 33 \\ 32 \\ 33 \\ 26 \end{bmatrix} \end{array}$$

An interaction between two terms is the pairwise product of the vectors from those terms. Here, for example, is the interaction between the two quantitative terms, education and age:

$$\begin{array}{ccc} \text{educ:age} & \text{educ} & \text{age} \\ \begin{bmatrix} 384 \\ 396 \\ 384 \\ 528 \\ 442 \end{bmatrix} & \text{is } \begin{bmatrix} 12 \\ 12 \\ 12 \\ 16 \\ 17 \end{bmatrix} & \times \begin{bmatrix} 32 \\ 33 \\ 32 \\ 33 \\ 26 \end{bmatrix} \end{array}$$

The interaction between a categorical variable and a quantitative variable involves multiplying each of the vectors from the categorical variable by the vector from the quantitative variable. Here is the interaction of age and married:

$$\begin{array}{ccc} \text{age:Married} & \text{age} & \text{Married} \\ \begin{bmatrix} 32 \\ 33 \\ 0 \\ 33 \\ 0 \end{bmatrix} & = \begin{bmatrix} 32 \\ 33 \\ 32 \\ 33 \\ 26 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} & \text{and} \end{array}$$

$$\begin{array}{ccc} \text{age:Single} & \text{age} & \text{Single} \\ \begin{bmatrix} 0 \\ 0 \\ 32 \\ 0 \\ 26 \end{bmatrix} & = \begin{bmatrix} 32 \\ 33 \\ 32 \\ 33 \\ 26 \end{bmatrix} \times \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \end{array}$$

Interactions between two categorical variables involve all the combinations of vectors from the variables. So, the interaction of sex and married produces four vectors, since each sex and married has two. The four vectors stand for Female Married, Female Single, Male Married, and Male Single:

$$\begin{array}{cccc} \text{Female} & \text{Female} & \text{Male} & \text{Male} \\ \text{Married} & \text{Single} & \text{Married} & \text{Single} \\ \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \end{array}$$

As it happens, there are no single females in the small subsample, so the second vector is all zeros.

The interaction of sector with its eight levels and sex with its two levels, produces a set of 16 vectors, more than it's appropriate to print here.

Finally, don't forget the intercept vector, which consists of a 1 for every case. For the five cases in the small data set, the intercept vector is

$$\begin{array}{c} \text{Intercept} \\ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \end{array}$$


---

## 9.4 Model Vectors in the Linear Formula

The model formula for the linear model has a very simple structure in terms of model vectors: each coefficient multiplies a single vector — there's one coefficient for each vector. In constructing a linear model, you need to provide a list of model terms: the model design. The computer then translates this list of terms into a set of model vectors. (In mathematics, a set of vectors, all with the same dimension, is called a **matrix**.) The software then finds the coefficient for each vector.

To illustrate, consider the following model design fitted to a small subset of the Current Population Survey wage data:

$$\text{wage} \sim 1 + \text{age} + \text{sex}$$

There are four vectors associated with these three model terms: the intercept vector, the vector for age, and the two vectors associated with the two levels of sex:

$$\begin{matrix} \text{Intercept} \\ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \end{matrix}, \begin{matrix} \text{age} \\ \begin{bmatrix} 32 \\ 33 \\ 32 \\ 33 \\ 26 \end{bmatrix} \end{matrix}, \begin{matrix} \text{sexM} \\ \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} \end{matrix}, \text{ and } \begin{matrix} \text{sexF} \\ \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{matrix}$$

To describe the model formula, the software only needs to tell the coefficient on each vector. A typical report from statistical software looks like this:

	Estimate	Std. Error	t value	p value
(Intercept)	-7.21	9.04	-0.80	0.509
age	0.64	0.37	1.76	0.220
sexM	-3.71	3.30	-1.12	0.378

The actual coefficients are in the column labelled “Estimate.” The other columns provide information about the reliability and the strength of evidence for each coefficient provided by the data. (The interpretation of these columns will be introduced in later chapters.)

There is one coefficient for each vector. As usual, one vector from a categorical term is dropped as being redundant. In this case, the sexF vector has been dropped.

The model values result from multiplying each vector by its coefficient, e.g.,

$$-7.21 \begin{matrix} \text{Intercept} \\ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \end{matrix} + 0.64 \begin{matrix} \text{age} \\ \begin{bmatrix} 32 \\ 33 \\ 32 \\ 33 \\ 26 \end{bmatrix} \end{matrix} - 3.71 \begin{matrix} \text{sexM} \\ \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} \end{matrix} \text{ giving } \begin{matrix} \text{Model Values} \\ \begin{bmatrix} 13.43 \\ 10.37 \\ 13.43 \\ 14.08 \\ 9.56 \end{bmatrix} \end{matrix}.$$

In other words, the model values are a **linear combination** of the model vectors.

## 9.5 Model Vectors and Redundancy

As a rule, there is one model coefficient for each model vector. The exception comes when one or more of the model vectors are **redundant**. In that case, the coefficients on the redundant vectors are dropped.

The most common situation where redundancy comes into play is when a model includes a categorical explanatory variable. It’s always the case that one of the indicator vectors from the categorical variable will be dropped as redundant.

A model vector is redundant when that vector can be written as a linear combination of other model vectors. For categorical variables, it’s easy to see

why this happens. Recall the indicator vectors for the sex variable:

$$\begin{matrix} \text{sexM} \\ \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{matrix} \quad \text{and} \quad \begin{matrix} \text{sexF} \\ \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} \end{matrix}.$$

Consider the linear combination which comes from multiplying each of these vectors by 1 and then adding them up. Since the indicator vectors have 1s in complementary places, adding up all the indicators produces a vector of all ones.

$$1 \begin{matrix} \text{sexM} \\ \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{matrix} + 1 \begin{matrix} \text{sexF} \\ \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} \end{matrix} = \begin{matrix} \text{Intercept} \\ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \end{matrix}.$$

The vector of all ones is, of course, the intercept vector. Thus, there is always redundancy between the set of indicator vectors for a categorical variable and the intercept vector.

## 9.6 Geometry by Arithmetic

It's fine to talk about 4- and 5-dimensional vectors, but how do you measure their geometrical properties, for instance their lengths or the angle between two vectors. After all, you can't sneak a ruler or a protractor into some hypothetical 5-dimensional space.

The trick is to perform the measurement by manipulating the numbers that represent the vector. Consider the vector labeled (2, 1) in Figure 9.3. The length of this ordinary, two-dimensional vector can be measured with a ruler. Or, you can use the Pythagorean theorem to calculate the length:  $\sqrt{2^2 + 1^2}$ . The formula for the length is very simple: square each of the coordinates, add up the squares, and take the square-root of the sum.

For dimensions higher than 2, simply *define* the length of a vector in the same way: the square root of the sum of squares of the coordinates.

An operation called the **dot product** simplifies the notation. The dot product takes two vectors as inputs, and produces a single number as an output. That number is found by multiplying the corresponding coordinates of the two vectors and adding up the results. For example

$$\begin{bmatrix} 3 \\ 1 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ -3 \\ 2 \end{bmatrix} = 12 - 3 + 0 = 9.$$

You write the dot product between vectors A and B as  $A \cdot B$ .

The length of a vector A is just the square root of A dotted with itself. Or, using mathematical notation where  $\|A\|$  stands for the length of A,

$$\|A\| = \sqrt{A \cdot A}$$

**Example 9.2: Vector Length** Find the length of  $\begin{bmatrix} 1 \\ -4 \\ 2 \\ 3 \end{bmatrix}$ .

$$\sqrt{\begin{bmatrix} 1 \\ -4 \\ 2 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ -4 \\ 2 \\ 3 \end{bmatrix}} = \sqrt{1 + 16 + 4 + 9} = \sqrt{30} = 5.477.$$

Angles can also be computed with dot products. The formula for the angle  $\theta$  between vectors A and B is

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}.$$

You hardly ever need to use this formula, but it is important that you know that angles can be found by arithmetic operations.

**Example 9.3: Angles Between Vectors** Find the angle between  $\begin{bmatrix} 1 \\ -4 \\ 2 \\ 3 \end{bmatrix}$  and  $\begin{bmatrix} 0 \\ 3 \\ -2 \\ 1 \end{bmatrix}$ .

This involves three simple calculations:

- $\mathbf{A} \cdot \mathbf{B} = \begin{bmatrix} 1 \\ -4 \\ 2 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 3 \\ -2 \\ 1 \end{bmatrix} = -13$
- $\|\mathbf{A}\| = \sqrt{\begin{bmatrix} 1 \\ -4 \\ 2 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ -4 \\ 2 \\ 3 \end{bmatrix}} = \sqrt{30}$
- $\|\mathbf{B}\| = \sqrt{\begin{bmatrix} 0 \\ 3 \\ -2 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 3 \\ -2 \\ 1 \end{bmatrix}} = \sqrt{14}$

Putting these together into the overall formula gives

$$\cos(\theta) = \frac{-13}{\sqrt{30}\sqrt{14}} = -0.6343.$$

Finding  $\theta$  itself involves inverting the cosine. It turns out that  $\cos(129.37^\circ) = -0.6343$ , so  $\theta = 129.37^\circ$  (or, equivalently, 2.258 radians).

What's remarkable here is that angles and lengths can be found entirely through arithmetic. This allows you to generalize the notions of angles and lengths to vectors in high-dimensional spaces.

Angles and lengths play an important role in the interpretation of statistical models. But it's inconvenient to have to take square-roots or to invert the cosine function. For this reason, statistics is often written in terms of **square lengths** and **cosines of angles**. The shorthand name for a square length in statistics is

a **sum of squares**, reflecting that the square-length of a vector is calculated by squaring each of the components and adding them up. There is no standard shorthand name for the cosine of an angle, but in the setting where the cosine of an angle most frequently appears in statistics, it is the **correlation coefficient**.

## 9.7 Computational Technique

In your ordinary statistical work, you will not often compute explicitly with vectors in the ways talked about in this chapter. The geometrical calculations are contained implicitly within other operators such as `lm`. Still, in order to work with the geometrical concepts in the exercises, you need to know how to calculate the geometrical quantities numerically.

To illustrate, I'll use the kids' feet dataset :

```
> feet = ISMdata("kidsfeet.csv")
```

There are  $n = 39$  cases in this data frame, so the vectors are 39-dimensional. Of course it seems impossible to visualize a 39-dimensional space, but the calculations are the same for 39-dimensions as they would be for 2-dimensional space or 2000-dimensional space.

### 9.7.1 Length and Dimension

The length operator in R can be used to compute the dimension of a vector; `length` simply counts how many numbers there are in a vector:

```
> length(feet$length)
[1] 39
```

The use of the name "length" for this operator can be confusing, since the geometrical length is different from the count of components returned by the `length` operator.

To find the geometrical length of a vector, use the Pythagorean formula: the square root of the sum of squares.

```
> sqrt( sum( feet$length^2 ))
[1] 154.6
```

Sometimes it's more convenient to talk about the "square-length" of a vector, which is just the same as the above, but without the square root:

```
> sum( feet$length^2 )
[1] 23904
```

Remember that the values are squared *before* they are summed. It's easy to make a mistake and put the square *after* the summation but this would give a wrong result.

Keep in mind that nothing about these computation requires the use of the `$` sign. It's only because the vectors used in the examples happen to be variables



in a data frame that the examples use \$. If you had a vector with a different name, just use that.

```
> myvec = c(1,2,3,4)
> sqrt(sum( myvec^2 ))
[1] 5.477
```

### 9.7.2 Dot products and angles

A fundamental computation on two vectors is the dot product: the sum of the pairwise products of the components of the two vectors.

```
> sum( feet$length * feet$width )
[1] 8687
```

The angle between two vectors is important enough that the book software defines an operator:

```
> angle( feet$length, feet$width )
[1] 0.04602
```

The angle is reported in **radians**, since most other operators (such as cos) take their inputs in radians. People, however, are usually more comfortable with degrees. You can convert radians to degrees by multiplying by  $180/\pi$ :

```
> angle( feet$length, feet$width )*180/pi
[1] 2.637
```

For convenience, the angle operator can be instructed to output values in degrees.

```
> angle( feet$length, feet$width, degrees=TRUE )
[1] 2.637
```

Make sure not to pass this to another operator such as cos:

```
> cos(angle( feet$length, feet$width ))
[1] 0.999 # Correct
> cos(angle( feet$length, feet$width, degrees=TRUE ))
[1] -0.8751 # WRONG, WRONG, WRONG
```

### 9.7.3 Scaling and linear combinations

Scaling and addition of vectors is done in the usual way. For example, here is a linear combination of 3 times the length vector and  $-2$  times the width vector:

```
> foo = 3*feet$length - 2*feet$width
```

Quantitative variables can be computed on directly. It's different for categorical variables. A categorical variable does not correspond to a single vector. Instead, there is one indicator vector for each level of the variable. To construct the indicator vector, use the == comparison operator for the particular level whose vector you want. For example, here is the indicator vector for sexB:

```
> boyvector = (feet$sex == "B")
> boyvector
[1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE
... and so on ...
[33] FALSE TRUE TRUE FALSE FALSE FALSE FALSE
```

Even though these are printed as logical values, when you do arithmetic on them the value TRUE is treated as 1 and FALSE as 0:

```
> 3*boyvector
[1] 3 3 3 3 3 3 3 0 0 3 3 3 3 3 0 0 0 0 0 3 3 0
[24] 0 0 3 0 3 3 3 0 0 0 3 3 0 0 0 0
```

For review purposes only

For review purposes only

# Chapter 10

## Statistical Geometry

*I made straight for the ship, roused up the men  
to get aboard and cast off at the stern.  
They scrambled to their places by the rowlocks  
and all in line dipped oars in the grey sea.  
But soon an off-shore breeze blew to our liking —  
a canvas-bellying breeze, a lusty shipmate  
sent by the singing nymph with sunbright hair.  
So we made fast the braces, and we rested,  
letting the wind and steersman work the ship.  
— The Odyssey, Book XII. Translation by Robert Fitzgerald.*

This chapter introduces a way to think about model fitting without the arithmetic involved in minimizing sums of squares of residuals. Surprisingly, one can do statistical calculations with just a ruler and protractor: by drawing straight lines, measuring lengths and measuring angles. The intention, however, is not to replace the computer, which is capable of much more precise and rapid calculations. Rather, the point is to help develop a concise picture of the operations that the computer is implementing so that you can understand and anticipate the output that will be provided by the computer.

Underlying the geometry are two main metaphors. The first is that fitting a model corresponds to making a journey. The second metaphor has to do with the terrain on which the journey is to be made: spaces in which model vectors specify directions and destinations.

In the epic poems of the ancient Greeks, long voyages were often by sea. A ship's course was set by the wind and the winds themselves, favorable or unfavorable, were directed by the gods. In Homer's *Odyssey*, the hero Odysseus is being kept from his home by an opposing wind from hostile Poseidon. Athena, the gray-eyed daughter of Zeus, arranges a favorable wind that carries Odysseus's raft toward his destination. But "toward" is not "to." Odysseus follows the wind as far as advantageous, but then needs to take matters into his own hands

For review purposes only

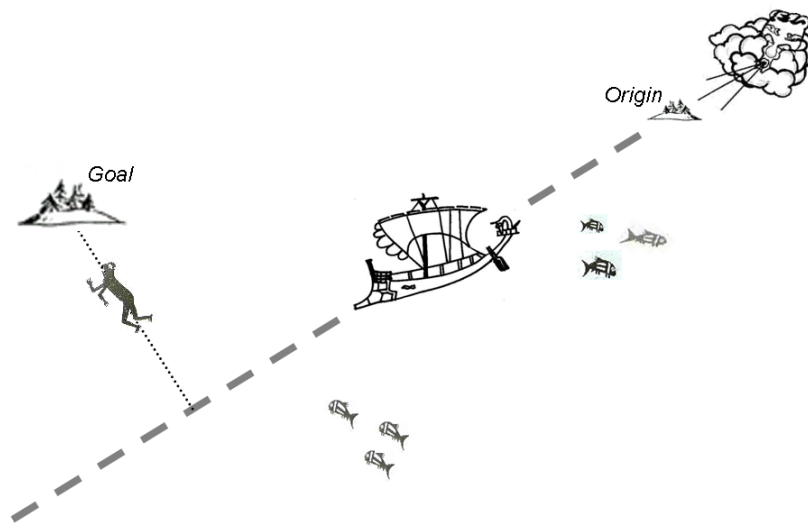


Figure 10.1: An ancient voyager could be carried downwind, but must complete his trip by swimming.

and swim. A sensible voyager tries to make the swimming leg of the trip short. Clever Odysseus jumps ship at the point where the ship passes as close as possible to the goal and then swims perpendicularly to the ship's course, directly toward the goal.

Fitting a statistical model is analogous. The modeler's explanatory variables are the winds, blowing where they will. The response variable is the goal. But the modeler does not cross the sea. Instead a more abstract mathematical space is being traversed: variable space.

## 10.1 Case Space and Variable Space

The scatter plot is a familiar graphical format for displaying data. Chapter 4 shows many of them. In a scatter plot, each case in the data frame is plotted as a single point. The whole data sample produces a scattering of points reflecting the case-to-case variation.

This chapter works with a different sort of plot, one that shows model vectors. When drawing model vectors, you are turning the scatter plot inside out. Instead of a point being a case and a variable being an axis, each variable is a vector and each case is an axis.

Both graphical configurations are useful. It's important not to confuse one with the other. To avoid confusion, it helps to have some nomenclature to refer

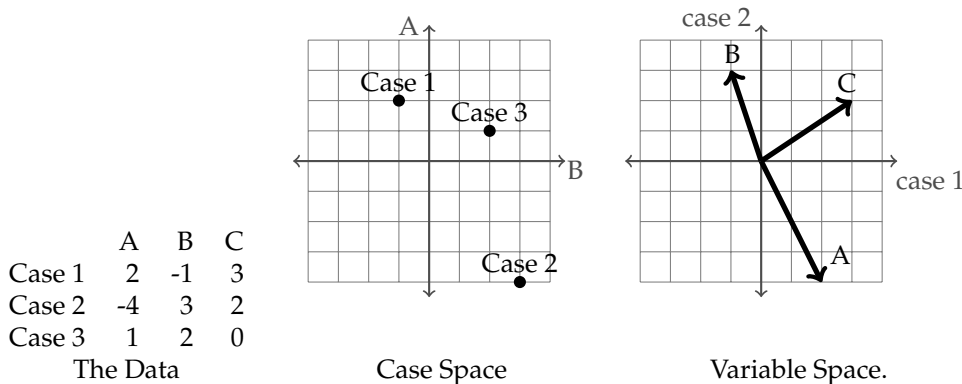


Figure 10.2: A small data set shown in both case- and variable-space formats. Case space can show only two variables at a time. Variable-space can show only two cases at a time.

to the two different configurations:

**Case Space.** The space in which the usual scatter plot is made. Each point is a single case. Each axis is a variable. It's called "case space" because the axes define a space for plotting the cases.

**Variable Space.** The space in which model vectors are plotted. Each vector is a variable. Each case is a coordinate axis. It's called "variable space" because the coordinate axes define the space that the variables are drawn in.

Figure 10.2 shows a very small dataset as it would be plotted in both case space and variable space.

A variable-space plot drawn on paper can show only  $n = 2$  cases, not enough data to be interesting statistically. The value of variable space is not in displaying data but in diagramming the geometrical relationships that underlie statistical modeling. Simple concepts that one can visualize in two or three dimensions — lengths, angles, and so on — give a framework for understanding the operations of statistical modeling.

The geometrical relationships that people capture visually can also be revealed by appropriate arithmetic computations — vector addition, vector scaling, dot products. The computer programs that perform model fitting are actually doing the geometry, but using numbers rather than pictures. This enables the computer to work with large samples, much larger than the  $n = 2$  cases that can be graphed on paper.

## 10.2 Subspaces and Geometrical Operations

A **subspace** is a part of the entire space. But it's not just any arbitrary part; it has a very specific definition that is related to vectors. Each vector has a direction

For review purposes only

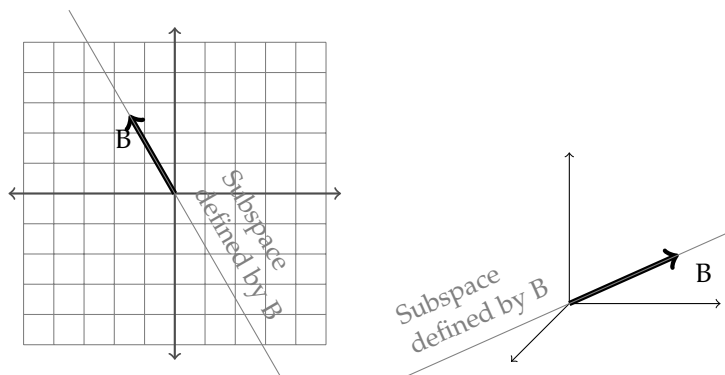


Figure 10.3: A vector and its subspace in  $n = 2$  (left) and  $n = 3$  (right) dimensional spaces.

and a length. The subspace defined by a vector is all the points in space that can be reached by moving from the origin in the direction of that vector. That is, the subspace is all the points that can be reached by scaling a vector. This is illustrated for  $n = 2$  and  $n = 3$  in Figure 10.3. The subspace extends out indefinitely far.

When dealing with more than one vector, each vector has its own subspace, as shown in Figure 10.4. Different vectors define different subspaces, although the subspaces from two aligned vectors will be identical.

Collections of two or more vectors also define subspaces: all the points that can be reached by moving from the origin first in the direction of one vector, then turning and moving in the direction of the other (and so on if there are more than two vectors). That is, the subspace defined by a set of vectors is all the points that can be reached by linear combinations of the vectors. A collection of two vectors defines a subspace that is a plane — unless the two vectors happen to be exactly aligned, in which case the subspace is just a line.

There are three geometrical operations that you will need for fitting models to data: projection, finding coefficients, and centering.

The operation of **projecting** a vector  $A$  onto a subspace defined by  $B$  finds the single point in the subspace of  $B$  that is as close as possible to  $A$ . As notation, write the projection of  $A$  onto  $B$  as  $A_{\parallel B}$ . Figure 10.5 shows an example. The parallel sign ( $\parallel$ ) in the notation is intended as a reminder that the projection of  $A$  onto  $B$  is a vector in the  $B$  subspace. So, the vector  $A_{\parallel B}$  is always aligned with  $B$ .

Projection of  $A$  onto  $B$  corresponds to fitting the model  $A \sim B$ . In fitting, as described in Chapter 6, you are looking for the coefficient  $c$  in the model formula  $\text{Model Values } A = cB$ , that is, multiply vector  $B$  by the number  $c$ . This is just vector scaling, so the quantity  $cB$  lies somewhere in the subspace defined by  $B$ . When you fit the model, you are finding the point in that subspace that is the

# Chapter 11

## Geometry with Multiple Vectors

*Philosophy is written in this grand book - the universe - which stands continuously open to our gaze. But the book cannot be understood unless one first learns to comprehend the language and interpret the characters in which it is written. It is written in the language of mathematics, and its characters are triangles, circles, and other geometrical figures, without which it is humanly impossible to understand a single word of it; without these one is wandering about in a dark labyrinth. — Galileo Galilei (1564-1642)*

Consider the data on SAT scores introduced in Chapter 5 and modeled with two explanatory variables: the fraction of students taking the test and the level of expenditures on education. This model involves three vectors: `sat`, `expend`, and `frac`.

Since any two vectors lie in a plane, you can easily draw a realistic representation of any two model vectors. For instance, Figure 11.1 shows the relationship between `frac` and `expend`. The angle between the vectors is based on the correlation between the two variables:  $r = 0.59$  so  $\theta = 53.7$  degrees.

The variance of `frac` is much larger than the variance of `expend`; they have completely different units. Since variation is shown by the length of the vectors, the lengths of `frac` and `expend` in the figure are very different. To mark the relative directions clearly, the figure shows the subspaces defined by `expend` and by `frac`.

Where to draw the `sat` vector? In general, three vectors don't lie on a plane, so it would be misleading to draw `sat` in the plane. But, seeing as this book is printed on paper, there isn't much choice — the drawing has to be done on the page! A key insight is given by a model formula derived by fitting the model

For review purposes only



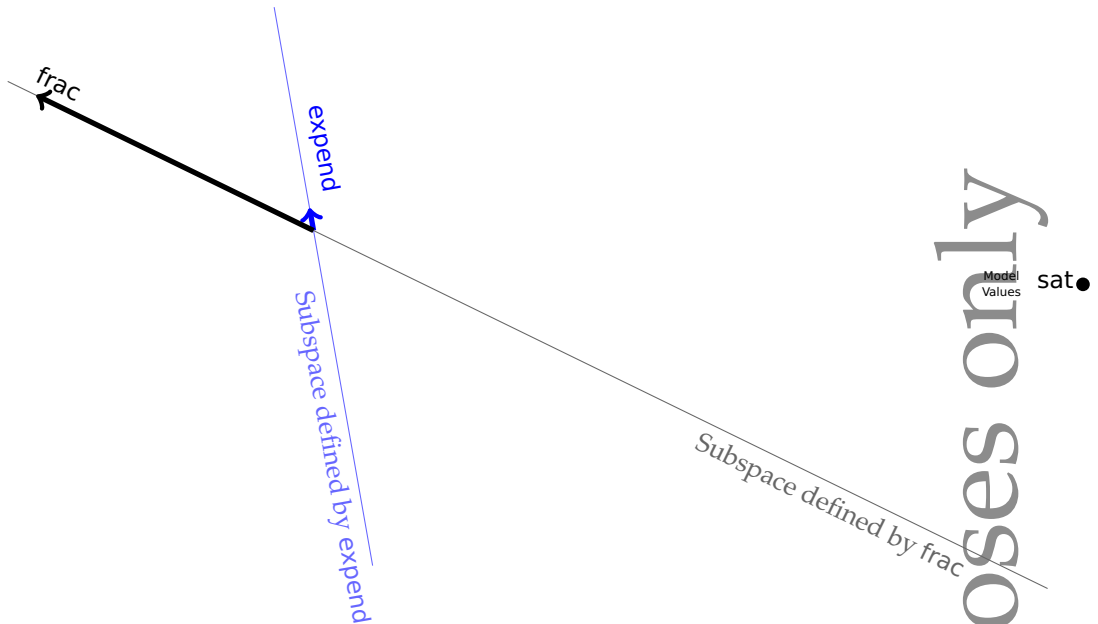


Figure 11.1: The fitted model values of `sat` from the model  $\text{sat} \sim \text{frac} + \text{expend}$  lie in the plane defined by `frac` and `expend`.

$\text{sat} \sim 1 + \text{frac} + \text{expend}$ :

$$\text{Model Values } \text{sat} = 994 - 2.85 \text{ frac} + 12.3 \text{ expend.}$$

The model values of `sat` are a linear combination of `frac` and `expend`, so they can be plotted in the plane defined by `frac` and `sat`. (The intercept is ignored in these diagrams. See Aside 11.1.)

The model values of `sat` are the dot in the far right of the figure. You can verify that this is drawn in the right place by following the directions in the model formula: walk 2.85 `frac` steps backwards and then take 12.3 `expend` steps. You will arrive at the `sat` point.

Where is the `sat` point itself? It's out of the plane, hovering in the air above the `sat` point. The vector that connects the hovering `sat` point and the `sat` point in the plane is the residual vector. The residual is perpendicular to the plane.

The general situation for a model  $A \sim B+C$  is shown in Figure 11.2. The two explanatory vectors  $B$  and  $C$  define a plane. The vector  $A$  will not in general be in that plane, but the fitted model values  $A_{\parallel(B,C)}$  will be on the plane. Indeed, any linear combination of  $B$  and  $C$  is a point in the  $(B,C)$ -plane.

The idea of a model triangle still holds. Since the point  $A_{\parallel(B,C)}$  is the closest point in the  $(B,C)$  plane to  $A$ , the residual vector  $A_{\perp(B,C)}$  is perpendicular to the plane. Vector  $A$  is the hypotenuse of the triangle. The two vectors  $A_{\parallel(B,C)}$  and

DATA FILE  
sat.csv

For review purposes only

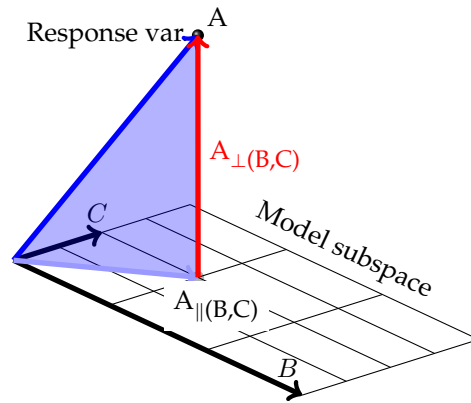


Figure 11.2: Projecting  $A$  onto the subspace defined by a set of two model vectors,  $B$  and  $C$ . The model triangle is shaded.

$A_{\perp}(B,C)$  are the legs and meet at a right angle.

No matter how many model vectors there are, the residual will be orthogonal to every one of them.

Take a look again at Figure 11.1. The actual sat is floating in the air above the Model Values sat point. By finding the length of the residual vector, you can figure out how far above the plane the sat vector is. It turns out to be about 4 cm when you scale things to the picture.

Hold your fingertip 4 cm directly above the Model Values sat point in Figure 11.1. Now find the point on the subspace defined by `expend` that is as close as possible to your finger. You have just fitted the model  $\text{sat} \sim 1 + \text{expend}$ . If you did it right, you found a point very near the first “d” in “Subspace defined by `expend`.”

Now do the same thing, but rather than starting from the point sat in the air, start from the Model Values sat point on the plane and project onto the same subspace, the one defined by `expend`. You should get exactly the same result, ending up near the “d.”

The point of all this is that you can get the same result in two different ways: project a vector (such as `sat`) directly onto a subspace, or project it onto a plane containing the subspace and then project that point onto the subspace. The reason has to do with the residual being perpendicular to the plane.

This means that in thinking about modeling a response variable with two explanatory variables, you only need to draw a plane. Rather than drawing the response variable itself as a vector, you can draw it as the projection down onto the plane defined by the two explanatory variables.

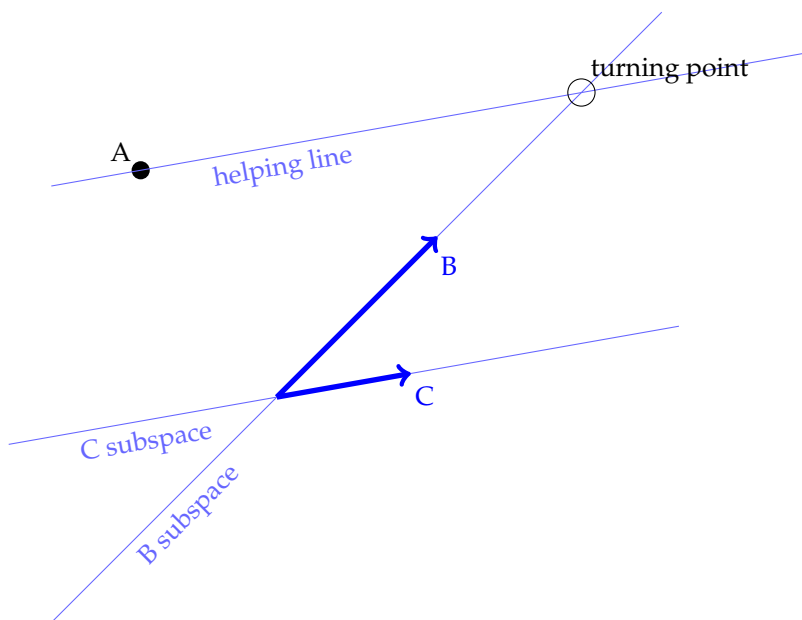


Figure 11.3: Finding coefficients with two explanatory vectors.

## 11.1 Finding Coefficients

To fit a model involving multiple explanatory terms, first project the response variable onto the subspace defined by those terms. So, for the model  $A \sim B + C$ , the first step is to project  $A$  onto the plane spanned by  $B$  and  $C$ . Figure 11.3 shows an example where the projection has already been done.

The next step is to find the coefficients on  $B$  and  $C$ . In thinking about this geometrically, imagine that you are finding how far to walk along  $B$  so that you can reach a point where you can turn in the direction of  $C$ . After turning, head in the  $C$  direction to reach the goal  $A$ .

The hard part is finding the turning point. One way to figure this out is to draw a helping line: a pathway parallel to  $C$  but passing through  $A$ . The place where this pathway crosses the  $B$  subspace tells where you will make the turn. In the figure, the turning point is marked with a circle. Finding the coefficient on  $B$  is a matter of scaling vector  $B$  to get from the origin to the turning point. This is a little less than 2 in the figure. With a ruler, by measuring the total path length along the  $B$  subspace to the turning point and dividing by the length of  $B$ , you would get 1.9. The coefficient on  $C$  is the number of steps of length  $C$  to take along the helping line in order to get from the turning point to  $A$ . This is a little more than 3 steps, in the negative direction — opposite to the way the  $C$  arrow points. With a ruler, measuring the path length along the helping line and dividing by the length of  $C$ , you would get a more precise number,  $-3.3$ , with the negative sign indicating the direction relative to  $C$ .

# Chapter 12

## Modeling Randomness

*The race is not to the swift, nor the battle to the strong, neither yet bread to the wise, nor yet riches to men of understanding, nor yet favor to men of skill; but time and chance happeneth to them all. — Ecclesiastes*

Until now, emphasis has been on the deterministic description of variation: how explanatory variables can account for the variation in the response. Little attention has been paid to the residual other than to minimize it in the fitting process.

It's time now to take the residual more seriously. It has its own story to tell. By listening carefully, the modeler gains insight into even the deterministic part of the model. Keep in mind the definition of statistics offered in Chapter 1:

Statistics is the explanation of variability **in the context of what remains unexplained.**

The next two chapters develop concepts and techniques for dealing with “what remains unexplained.” In later chapters these concepts will be used when interpreting the deterministic part of models.

### 12.1 Describing Pure Randomness

Consider an **event** whose **outcome** is completely random, for instance, the flip of a coin. How to describe such an event? Even though the outcome is random, there is still structure to it. With a coin, for instance, the outcome must be “heads” or “tails” — it can't be “rain.” So, at least part of the description should say what are the possible outcomes: H or T for a coin flip. This is called the **outcome set**. (The outcome set is conventionally called the **sample space** of the event. This terminology can be confusing since it has little to do with the sort of sampling encountered in the collection of data nor with the sorts of spaces that vectors live in.)

For review purposes only

For a coin flip, one imagines that the two outcomes are equally likely. This is usually specified as a **probability**, a number between zero and one. Zero means “impossible.” One means “certain.”

A **probability model** assigns a probability to each member of the outcome set. For a coin flip, the accepted probability model is 0.5 for H and 0.5 for T — each outcome is equally likely.

What makes a coin flip purely random is *not* that the probability model assigns equal probabilities to each outcome. If coins worked differently, an appropriate probability model might be 0.6 for H and 0.4 for T. The reason the flip is purely random is that the probability model contains all the information; there are no explanatory variables that account for the outcomes in any way.

**Example 12.1: Rolling a Die** The outcome set is the possibilities 1, 2, 3, 4, 5, and 6. The accepted probability model is to assign probability  $\frac{1}{6}$  to each of the outcomes. They are all equally likely.

Now suppose that the die is “loaded.” This is done by drilling into the dots to place weights in them. In such a situation, the heavier sides are more likely to face down. Since 6 is the heaviest side, the most likely outcome would be a 1. (Opposite sides of a die are arranged to add to seven, so 1 is opposite 6, 2 opposite 5, and 3 opposite 4.) Similarly, 5 is considerably heavier than 2, so a 2 is more likely than a 5. An appropriate probability model is this:

Outcome	1	2	3	4	5	6
Probability	0.28	0.22	0.18	0.16	0.10	0.06

One view of probabilities is that they describe how often outcomes occur. For example, if you conduct 100 trials of a coin flip, you should expect to get something like 50 heads. According to the **frequentist** view of probability, you should base a probability model of a coin on the relative proportion of times that heads or tails comes up in a very large number of trials.

Another view of probabilities is that they encode the modeler’s assumptions and beliefs. This view gives everyone a license to talk about things in terms of probabilities, even those things for which there is only one possible trial, for instance current events in the world. To a **subjectivist**, it can be meaningful to think about current international events and conclude, “there’s a one-quarter chance that this dispute will turn into a war.” Or, “the probability that there will be an economic recession next year is only 5 percent.”

**Example 12.2: The Chance of Rain.** Tomorrow’s weather forecast calls for a 10% chance of rain. Even though this forecast doesn’t tell you what the outcome will be, it’s useful; it contains information. For instance, you might use the forecast in making a decision not to cancel your picnic plans.

# Chapter 13

## Geometry of Random Vectors

*Oh, many a shaft at random sent  
Finds mark the archer little meant!  
And many a word at random spoken  
May soothe, or wound, a heart that's broken!* — Walter Scott (1771-1832)

In the next several chapters, models will be interpreted using ideas of randomness. Random simulations of residuals will be generated to explore the repeatability of model coefficients. Explanatory variables will be replaced by random variables to see whether the genuine variable is any better than junk.

This chapter introduces some of the geometry of randomness — for instance, what happens when model vectors are generated at random and used in fitting models. Some of it will contradict your intuition. For example, even though a random vector is equally likely to point in any direction, when you compare a random vector to another vector, they are very likely to meet at an angle near 90 degrees.

### 13.1 Random Angles

A statistician from the National Union of Transcendent Science (NUTS) approaches you. He claims to have mathematical proof that everything in the universe is linked. He discovered this when he decided to model the attendance count at the last two NUTS meetings. Taking this as his response variable  $A$ , he chose as an explanatory variable the last two stock market closing prices,  $B$ . In constructing the model  $A \sim B$ , he found a non-zero coefficient on  $B$ . But if there had been no relationship between NUTS attendance and the stock market, the coefficient would be expected to be zero.

Then he went on to try many other possible  $B$ : the distance to Mars and Jupiter at the time of the NUTS meetings; the altitudes of the two highest mountains in his home state; the number of times his dog barked during the NUTS

For review purposes only

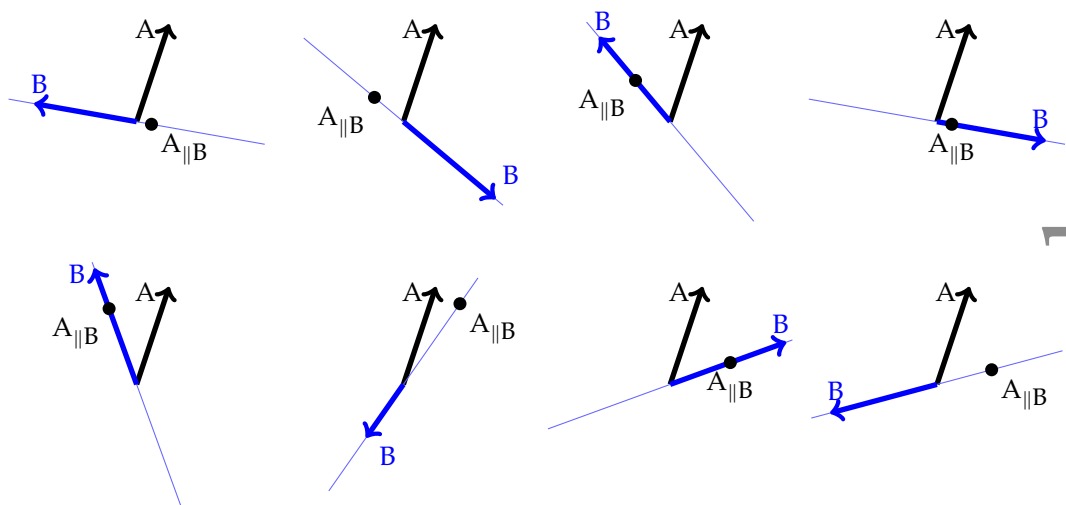


Figure 13.1: The model  $A \sim B$  for several random  $B$ . The fitted model values point  $A_{\parallel B}$  is not usually at the origin, so the coefficient on  $B$  is not usually zero.

meetings. Hundreds of different  $B$  variables. In almost every case, the coefficient on  $B$  in  $A \sim B$  was non-zero. “This proves that everything is linked!” he says enthusiastically.

You patiently try to explain things to the NUTS guy. The coefficient on  $B$  tells a lot about how  $A$  and  $B$  are aligned as vectors, but it doesn’t necessarily tell very much about how the underlying variables are connected. What he proved is not that almost everything is linked, but that even random vectors can be aligned.

Figure 13.1 illustrates the situation. Whatever direction  $B$  points in, it’s likely that the fitted model values  $A_{\parallel B}$  will lie at some distance along  $B$  and so the coefficient on  $B$  will be non-zero. The only situation in which the coefficient will be zero is when  $B$  is exactly perpendicular to  $A$ .

It’s correct that, in the absence of a link, the expected value of the coefficient on  $B$  will be zero. But the expected value tells only part of the story about the distribution of the coefficient on  $B$ . The spread around the expected value is also important.

It’s easier to think about things in terms of the angle  $\theta$  between  $A$  and  $B$ . There’s only one situation where the coefficient will be zero, when  $\theta = 90^\circ$ . But there are many situations where  $\theta$  is different from  $90^\circ$ . So it’s much more likely that the coefficient will be non-zero than it will be exactly zero.

Picking a random direction in 2-dimensional space amounts to picking a random point anywhere on a circle, as shown in Figure 13.2. Since the angles are evenly spaced, each angle is equally likely to be chosen. Thus, the distribution of angles between  $A$  and a random vector  $B$  is uniform.

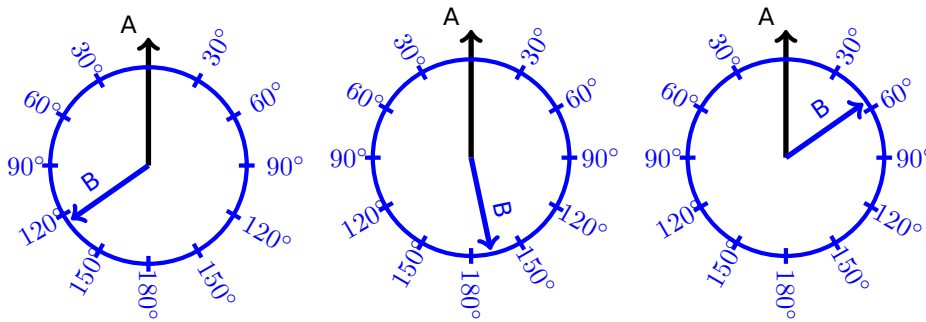


Figure 13.2: In  $n = 2$  dimensions, picking a random direction corresponds to picking a random point on the circle. The angle between A and B is equally likely to be anything between 0 and  $180^\circ$ .

This idea that all angles are equally likely matches well with most people's intuition. Unfortunately, that intuition is misleading when it comes to angles in higher-dimensional spaces.

Consider the situation in 3-dimensional space. Picking a random direction can be done by picking a random point on a sphere. Figure 13.3 shows a vector A piercing the pole of a sphere. The angle of a vector B going through some other point on the sphere can be read off from the circle of latitude on the sphere. For example, the vector B shown in the figure points to the circle of latitude marked  $45^\circ$ , so it is at an angle of  $45^\circ$  to A.

The probability that a randomly chosen B will point to any particular circle of latitude is proportional to the circumference of that circle. From the figure, you can see that the  $15^\circ$  circle is short compared to the  $30^\circ$  circle, which is itself short compared to the  $60^\circ$  or  $90^\circ$  circles. That is, a randomly selected point on a sphere is much more likely to be near the equator (the  $90^\circ$  circle) than it is to be near either the north or south pole.

This pattern where angles near  $90^\circ$  are much more likely than other angles gets even stronger as the dimension of the space increases. Figure 13.4 shows the probability of angles between random vectors for a variety of dimensions  $n$ .

One way to summarize the distribution of angles is with the mean and standard deviation of the distribution. The mean angle is  $90^\circ$  for any  $n$  — on average, random vectors are perpendicular. This is why the NUTS guy was reasonable in thinking that, on average, the coefficient on B in his model  $A \sim B$  will be zero if there is no relationship between A and B. But “on average” isn't the same as “always.” To know how far from zero a coefficient is likely to be when A is unrelated to B, you have to know the standard deviation of the distribution.

The standard deviation of the angle between random vectors depends on the dimension  $n$  of the space. You can see this in Figure 13.4: the distributions become narrower for larger  $n$ . The precise form of the standard deviation is particularly simple if you consider not the angle directly, but the cosine of the angle — the quantity that's relevant when dealing with model coefficients.



# Chapter 14

## Confidence in Models

*To know one's ignorance is the best part of knowledge. — Lao-Tse*

If you are a skilled modeler, you try to arrange things so that your model coefficients are random numbers.

That statement may sound silly, but before you jump to conclusions it's important to understand where the randomness comes from and why it's a good thing. Then you can use the tools in the previous two chapters to deal with the randomness and interpret it.

Recall the steps in building a statistical model:

1. Collect data. This is the hardest part, often involving great effort and expense.
2. Design your model, choosing the response variable, the explanatory variables, and the model terms. (It's sensible to have the design in mind *before* you collect your data, so you know what data are needed.)
3. Fit the model design to your data.

Step (3) is entirely deterministic. Given the model design and the data to which the model is to be fitted, fitting is an automatic process that will give the same results every time and on every computer. There is no randomness there. (There is some choice in choosing which to remove from a set of vectors containing redundancy, but that is a choice, not randomness. In any event, the fitted model values are not affected by this choice.)

Step (2) appears, at first glance, to leave space for chance. After all, when explanatory terms are collinear, as they often are, the fitted coefficients on any term can depend strongly on which other terms have been included in the model. The coefficients depend on the modeler's choices. But this means only that the coefficients reflect the beliefs of the modeler. If those beliefs aren't random, then the randomness of coefficients doesn't stem from the modeler.

For review purposes only

It's step (1) that introduces the randomness. In collecting data, the sample cases are selected from a population. As described in Chapter 2, it's advantageous to make a random selection from the population; this helps to make the sample representative of the population.

Saying something is random means that it is uncertain, that if the process were repeated again the result might be different. When a sample is selected at random, the particular sample that is produced is just one of a set of possibilities. One can imagine other possibilities that might have come about from the luck of the draw.

Insofar as the sample is random, the coefficients that come from fitting the model design to the sample are also random. The randomness of model coefficients means that the coefficients that come from a model design fitted to any particular data set are not likely to be an exact match to what you would get if the model design were fitted to the *entire population*. After all, the randomly selected sample is unlikely to be an exact match to the entire population.

It's helpful to know how close the results from the sample are to what would have been obtained if the sample had been the entire population. Ultimately, the only way to know this for sure is to create a sample that is the entire population. Usually this is impractical and often it is impossible.

But there is an approach that will give insight, even if it does not give certainty. To start, imagine that the sample were actually a census: a sample that contains the entire population. Repeating the study with a new sample would give exactly the same result because the new sample would be the same as the old one; it's the same population.

When the sample is not the entire population, repeating the study won't give the same result every time because the sample will include different members of the population. If the results vary wildly from one repeat to another, you have reason to think that the results are not a reliable indication of the population. If the results vary only a small amount from one repeat to another, then there's reason to think that you have closely approximated the results that you would have gotten if the sample had been the entire population. The repeatability of the process indicates how well the modeler knows the coefficients, or, in a word, the **precision** of the coefficients.

Knowing the precision of coefficients is key in drawing conclusions from them. Consider Galton's problem in studying whether height is a heritable trait. Had Galton known about modeling, he might have constructed a model like  $\text{height} \sim 1 + \text{mother} + \text{father} + \text{sex}$ . A relationship between the mother's height and her child's height should show up in the coefficient on mother. If there is a relationship, that coefficient should be non-zero.

Fitting the model to Galton's data, the coefficient is 0.32. Is this non-zero? Yes, for this particular set of data. But how might it have been different if Galton repeated his sampling, selecting a new set of cases from the population? How precise is the coefficient? Until this is known, it's mere bravado to say that a result of 0.32 means anything.

This chapter introduces methods that can be used to estimate the precision of coefficients from data. A standard format for presenting this estimation is the

**confidence interval.** For instance, from Galton's data, the estimated coefficient on mother is  $0.32 \pm 0.06$ , giving confidence that a different sample would also show a non-zero coefficient.

It's important to contrast precision with **accuracy**. Precision is about repeatability. Accuracy is about how the result matches the real world. Ultimately, accuracy is what the modeler wants. But the results of a model always depend on the choices the modeler makes, e.g., which explanatory variables to include, how to choose a sample, etc. The results can be accurate only if the modeler makes good choices. Knowing whether this is the case depends on knowing how the world really works, and this is what you are seeking to find out in the first place. Accuracy is elusive knowledge. Precision will have to suffice.

## 14.1 The Sampling Distribution

In principle, the way to see how much variation in model coefficients is introduced by the sampling process is to repeat the process of sampling and model fitting. The coefficients from each repetition can be collected; their distribution is called the **sampling distribution**. This is illustrated schematically in Figure 14.1, which shows just three random samples selected from the population. In reality, a much larger number of samples should be used, not just three.

To illustrate, consider the ten-mile race dataset. This is a census containing the running times for all 8636 registered participants in the Cherry Blossom Ten Mile race held in Washington, D.C. in April 2005.

The variable `net` records the start-line to finish-line time of each runner. There are also variables `age` and `sex`. Any model would do to illustrate the sampling distribution. Try

$$\text{net} \sim 1 + \text{age} + \text{sex}$$

Just for reference, here are the coefficients when the model is fit to the entire population:

	Intercept	age	sexM
Population	5540	16.9	-727

The coefficients indicate that runners who are one year older tend to take about 17 seconds longer to run the 10 miles.

Of course, if you knew the coefficients that fit the whole population, you would hardly need to collect a sample! But the purpose here is to demonstrate the effects of a random sampling process. The table below gives the coefficients from several sampling draws; each sample has  $n = 100$  cases randomly selected from the population. That is, each sample simulates the situation where someone has randomly selected  $n = 100$  cases.

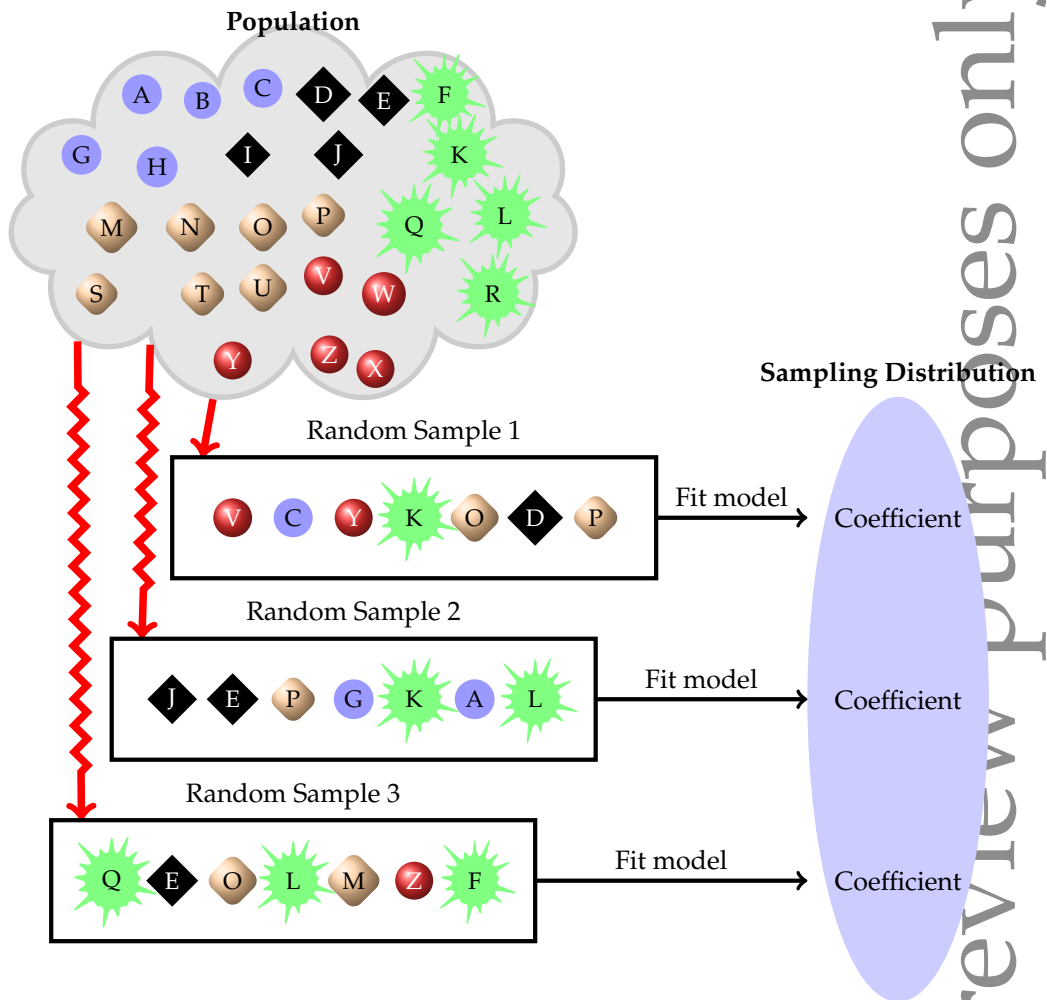


Figure 14.1: The sampling distribution reflects the variation in model coefficients from one random sample to another.

For review purposes only

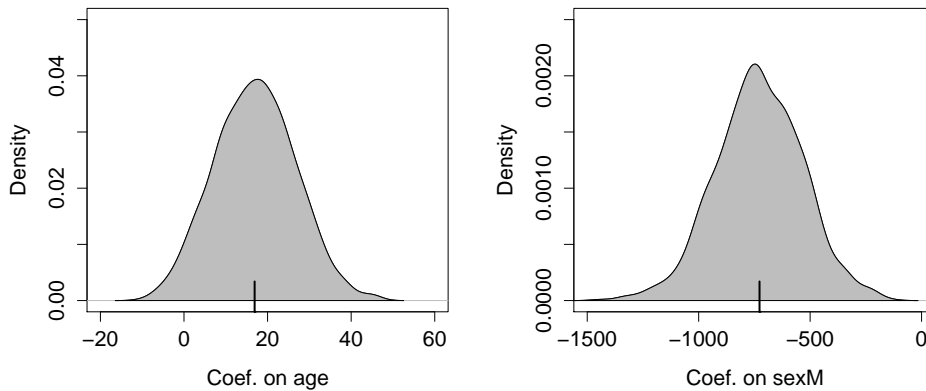


Figure 14.2: Sampling distributions for the age and sex coefficients for a sample size of  $n = 100$ . The bell-shaped distributions are centered on the population value (shown as a tick).

	Intercept	age	sexM
Sample 1	5570	12.6	-867
Sample 2	4630	31.2	-430
Sample 3	5980	2.1	-765
Sample 4	5850	5.0	-772
Sample 5	5420	16.5	-1010
Sample 6	5950	3.1	-583
Sample 7	4770	33.7	-779
Sample 8	5420	11.9	-551

Judging from these few samples, there is a lot of variability in the coefficients. For example, the age coefficient ranges from 2.1 to 33.7 in just these few samples. Figure 14.2 shows the distribution of coefficients for the age and sex coefficients for 1000 repeats of the sampling process. Each of the samples has 100 cases randomly selected from the population. These distributions, which reflect the randomness of the sampling process, are called **sampling distributions**. It's very common for sampling distributions to be bell-shaped and to be centered on the population value.

The width of the sampling distribution shows the reliability or repeatability of the coefficients. The two most common ways to summarize the width are the standard deviation and the 95% coverage interval.

An important item to add to your vocabulary is the “standard error.”

Sampling distributions, like other distributions, have standard deviations. The term **standard error** is used to refer to the standard deviation of a sampling distribution.

For the `age` coefficient, the standard error is about 9.5 seconds per year (keeping in mind the units of the data), and the standard error in `sexM` is about 192 seconds. These standard errors were calculated by drawing 10,000 repeats of samples of size  $n = 100$ , and fitting the model to each of the 10000 samples. This is easy to do on the computer, but practically impossible in actual field work.

Every model coefficient has its own standard error which indicates the precision of the coefficient. The size of the standard error depends on several things:

- The quality of the data. The precision with which individual measurements of variables is made, or errors in those measurements, translates through to the standard errors on the model coefficients.
- The quality of the model. The size of standard errors is proportional to the size of the residuals, so a model that produces small residuals tends to have small standard errors. However, collinearity or multi-collinearity among the explanatory variables tends to inflate standard errors.
- The sample size  $n$ . Standard errors tend to get tighter the more data is used to fit the model. A simple relationship holds very widely; it's worth remembering this rule:

The standard error typically gets smaller as the sample size  $n$  increases, but slowly; it's proportional to  $1/\sqrt{n}$ . This means, for instance, that to make the standard error 10 times smaller, you need a dataset that is 100 times larger.

## 14.2 Standard Errors and the Regression Report

Finding the standard error of a model coefficient would be straightforward if one could follow the procedure above: repeatedly draw new samples from the population, fit the model to each new sample, and collect the resulting coefficients from each sample to produce the sampling distribution. This process is impractical, however, because of the expense of collecting new samples.

Fortunately, there can be enough information in the original sample to make a reasonable guess about the sampling distribution. How to make this guess is the subject of Section 14.6. For now, it suffices to say that the guess is based on an approximation to the sampling distribution called the **resampling distribution**.

A conventional report from software, often called a **regression report**, provides an estimate of the standard error. Here is the regression report from the model `net ~ 1 + age + sex` fitted to a sample of size  $n = 100$  from the running data:

	Estimate	Std. Error	t value	Pr(> t )
Intercept	5110	340.0	14.89	0.0000
age	29	9.1	3.21	0.0018
sexM	-860	188.0	-4.57	0.0000

The coefficient itself is in the column labelled “Estimate.” The next column, labelled “Std. Error,” gives the standard error of each coefficient. You can ignore the last two columns of the regression report for now. They present the information from the first two columns in a different format.

#### Example 14.1: The standard error of the mean

Even the simple model  $A \sim 1$  involves a sampling distribution. The coefficient from this model is, of course, the sample mean of  $A$  and so the standard error of the coefficient is called the **standard error of the mean**.

To illustrate, consider the widespread process of calculating a student’s “grade-point average.” Where grades are given as letters, these are converted to numbers (i.e.,  $A=4$ ,  $B=3$ , and so on) and the numbers are averaged. Here is an example of part of the transcript, for the student whose ID is 31509.

DATA FILE  
grades.csv

sessionID	grade	sid	grpt	dept	level	sem	enroll	iid
C1959	B+	S31509	3.33	Q	300	FA2001	13	i323
C2213	A-	S31509	3.66	i	100	SP2002	27	i209
C2308	A	S31509	4.00	O	300	SP2002	18	i293
C2344	C+	S31509	2.33	C	100	FA2001	28	i140
C2493	S	S31509		n	300	FA2002	10	i500
C2562	A-	S31509	3.66	Q	300	FA2002	22	i327
C2585	A	S31509	4.00	O	200	FA2002	19	i310
C2737	A	S31509	4.00	q	200	SP2003	11	i364
C2764	S	S31509		j	100	SP2003	54	i447
C2851	A	S31509	4.00	O	300	SP2003	14	i308
C2928	B+	S31509	3.33	O	300	FA2003	22	i316
C3036	B+	S31509	3.33	q	300	FA2003	21	i363
C3443	A	S31509	4.00	O	200	FA2004	17	i300

(The course name and other identifying information such as the department and instructor have been coded for confidentiality.)

The student’s grade-point average is just the mean of the fourth column: 3.60. It would be nice to know how precise this number is.

Calculating a standard error of the mean is easy. The regression report on the model  $\text{gradepoint} \sim 1$  gives:

	Estimate	Std. Error	t value	Pr(> t )
Intercept	3.6036	0.1549	23.27	0.0000

There is actually a simple formula for the standard error of the mean:

$$\text{standard error} = \frac{\text{std. dev. of } A}{\sqrt{n}}.$$

This will give the same result as the regression report.

For review purposes only

Formulas like this are useful when making back-of-the-envelope calculations, for instance when trying to interpret “statistics” given in newspaper articles. If you can make a reasonable guess about the standard deviation of the variable, you can often estimate the standard error from data given in the article. Journalists often give information about means and sample sizes, but rarely about standard errors. Reporter James Surowiecki writes, “As many studies have shown, people don’t have an intuitive understanding of things like margins of error and random sampling; they prefer to focus on a single number, even if it’s falsely precise, and so end up overemphasizing [that] number.”[23]

Notice in the formula that the standard error depends on the sample size  $n$ : it is proportional to  $1/\sqrt{n}$ .

### 14.3 Confidence Intervals

The regression report gives the standard error explicitly, but it’s common in many fields to report the precision of a coefficient in another way, as a **confidence interval**.

A confidence interval is a little report about a coefficient that is written like this: “ $15.8 \pm 17.8$  with 95% confidence.” The report involves three components:

**Point Estimate** The center of the confidence interval: 15.8 here. Read this directly from the regression report.

**Margin of Error** The half-width of the confidence interval: 17.8 here. This is two times the standard error from the regression report.

**Confidence Level** The percentage of the coverage interval. This is typically 95%. Since people get tired of saying the same thing over and over again, they often omit the “with 95% confidence” part of the report. This can be dangerous, since sometimes people use confidence levels other than 95%.

The purpose of multiplying the standard error by two is to make the confidence interval approximate a 95% coverage interval of the resampling distribution. (For more information about this, see section 14.9.1.)

**Example 14.2: The GPA Confidence Interval** The grade-point average in the above example was 3.60 with a standard error of 0.155. This translates to a confidence interval of  $3.60 \pm 0.31$ .

Calculating the confidence interval from the regression report is very simple: you just need to remember to multiply the standard error by 2.

**Example 14.3: Wage discrimination in trucking?** Section 8.1.2 (page 148) looked at data from a trucking company to see how earnings differ between men



and women. It's time to revisit that example, using confidence intervals to get an idea of whether the data clearly point to the existence of a wage difference.

The model  $\text{earnings} \sim \text{sex}$  ascribed all differences between men and women to their sex itself. Here is the regression report:

	Estimate	Std. Error	t value	Pr(> t )
Intercept	35501	2163	16.41	0.0000
sexM	4735	2605	1.82	0.0714

The estimated difference in earnings between men and women is  $\$4700 \pm \$2600$  — not at all precise.

Another model can be used to take into account the worker's experience, using age as a proxy:

	Estimate	Std. Error	t value	Pr(> t )
Intercept	14970	3912	3.83	0.0002
sexM	2354	2338	1.01	0.3200
age	594	99	6.02	0.0000

Earnings go up by  $\$600 \pm \$200$  for each additional year of age. The model suggests that part of the difference between the earnings of men and women at this trucking company is due to their age: women tend to be younger than men. The confidence interval on the earnings difference is very broad —  $\$2350 \pm \$2350$  — so broad that the sample doesn't provide much evidence for any difference at all.

One issue is whether the age dependence of earnings is just a mask for discrimination. To check out this possibility, fit another model that looks at the age dependence separately for men and women:

	Estimate	Std. Error	t value	Pr(> t )
Intercept	17178	8026	2.14	0.0343
sexM	-443	9174	-0.05	0.9615
age	530	225	2.35	0.0203
sexM:age	79	251	0.32	0.7530

The coefficient on the interaction term between age and sex is  $79 \pm 251$  — no reason at all to think that it's different from zero. So, evidently, both women and men show the same increase in earnings with age.

Notice that in this last model the coefficient on sex itself has reversed sign from the previous models: Simpson's paradox. Of course the confidence interval is so broad —  $-\$443 \pm \$9174$  — that the sex coefficient is not distinguishable from zero. This huge inflation in the width of the confidence interval is the result of the collinearity between age and sex and their interaction term. As discussed in Section 14.7, sometimes it is necessary to leave out model terms in order to get reliable results.



**Example 14.4: SAT Scores and Spending, revisited** Example 8.1.3 used data from 50 US states to try to see how teacher salaries and student-teacher ratios are related to test scores. The model used was  $\text{sat} \sim \text{salary} + \text{ratio} + \text{frac}$ , where  $\text{frac}$  is a covariate — what fraction of students in each state took the SAT. To interpret the results properly, it's important to know the confidence interval of the coefficients:

	Estimate	Std. Error	t value	Pr(> t )
Intercept	1057.9	44.3	23.86	0.0000
salary	2.5	1.0	2.54	0.0145
ratio	-4.6	2.1	-2.19	0.0339
frac	-2.9	0.3	-12.76	0.0000

The confidence interval on salary is  $2.5 \pm 2.0$ , leaving little doubt that the data support the idea that higher salaries are associated with higher test scores. Higher student-faculty ratios seem to be associated with lower test scores. You can see this because the confidence interval on ratio,  $-4.6 \pm 4.2$ , doesn't cover zero. The important role of the covariate  $\text{frac}$  is also confirmed by its tight confidence interval away from zero.

Collecting more data would allow more precise estimates to be made. For instance, collecting data over 4 years would reduce the standard errors in half, although what the estimates would be with the new data cannot be known until the analysis is done.

---

**Aside. 14.1** Confidence intervals for very small samples.

When you have data with a very small  $n$ , say  $n < 20$ , a multiplier of 2 can be misleading. The reason has to do with how well a small sample can be assumed to represent the population. The correct multiplier depends on the difference between the sample size  $n$  and the number of model coefficients  $m$ :

$n - m$	1	2	3	5	10	15
Multiplier for 95% confidence level	12.7	4.3	3.2	2.6	2.2	2.1

For example, if you fit the running time versus age and sex model to 4 cases, the appropriate multiplier should be 12.7, not even close to 2.

---

## 14.4 Interpreting the Confidence Interval

Take a typical confidence interval, perhaps something like " $17 \pm 6$  with 95% confidence." Calculating the confidence interval is easy. Interpreting it is hard.

A 95% confidence interval is intended to reflect a 95% coverage interval of the sampling distribution, as approximated by the resampling distribution. So,

For review purposes only

it's tempting to say something like this, "The true coefficient will be in the range 11 to 23 with 95 percent probability." One question this raises is what "true" means. Statisticians are more comfortable talking about **population parameters** — the value of the model if it could be fit to the entire population — than "truth." Those who interpret probabilities according to relative frequencies can point out that unlike the coefficients from random samples, the population parameter is not actually random, so you shouldn't talk about the probability of it being this or that.

Another tempting statement is, "If I repeated our study with a different random sample, the new result would be within  $\pm 6$  of the original result." But that statement isn't correct mathematically, unless your point estimate happens to align perfectly with the population parameter — and there's no reason to think this is the case.

Treat the confidence interval just as an indication of the precision of the measurement. If you do a study that gets a coefficient of  $17 \pm 6$  and someone else does a study that gives  $23 \pm 5$ , then there is little reason to think that the two studies are inconsistent. On the other hand, if your study gives  $17 \pm 2$  and the other study is  $23 \pm 1$ , then something seems to be going on.

Now return back to the first interpretation offered of the interval  $17 \pm 6$  with 95% confidence: "The true coefficient will be in the range 11 to 23 with 95 percent probability." Taking "true" to mean "population parameter," you can get around the frequentist's objection if you consider probability from a different angle. It's not the population parameter that's random; it is your study that is a random sample from all the possible studies that could have been done. From this perspective, restate the interpretation like this:

Of all the studies that have computed 95% confidence intervals properly, 95% of them will have captured the population parameter relevant to their study within their confidence interval.

Why use a 95% confidence level? Why not 100%? Because a 100% confidence interval would be too broad to be useful. In theory, 100% confidence intervals tend to look like  $-\infty$  to  $\infty$  — that doesn't give any information. Certainty comes at the cost of ignorance.

The 95% confidence level is standard in contemporary science; it's a convention. For that reason alone, it is a good idea to use 95% so that the people reading your work will tend to interpret things right.

## 14.5 Confidence in Predictions

When a model is used for making a prediction, the coefficients themselves aren't of direct interest; it's the prediction that counts. The logic of confidence intervals can be extended to prediction. The idea is to take the precision of the coefficients and propagate that through the model formula.

It's important to distinguish two kinds of prediction confidence intervals. One kind is the interval on the model value itself; this reflects the uncertainty in



the coefficients as reflected by the standard errors. The second kind — the prediction interval — is the interval of the likely outcome according to the model. This outcome interval incorporates the uncertainty due to the precision of the coefficients and also due to the residuals of actual values around the model value.

To illustrate, consider making a prediction of a child's adult height when you know the heights of the mother and father and the child's sex. Using Galton's data from the 19th century, a simple and appropriate model is  $\text{height} \sim \text{sex} + \text{mother} + \text{father}$ .

	Estimate	Std. Error	t value	Pr(> t )
Intercept	15.3448	2.7470	5.59	0.0000
sexM	5.2260	0.1440	36.29	0.0000
mother	0.3215	0.0313	10.28	0.0000
father	0.4060	0.0292	13.90	0.0000

Now consider a hypothetical man — call him Bill — whose mother is 67 inches tall and whose father is 69 inches tall. According to the model formula, Bill's predicted height is  $15.3448 + 5.226 + 0.3215 \times 67 + 0.406 \times 69 = 70.13$  inches.

Calculating a confidence interval on the model value is more involved and usually done using software. For Bill, the 95% confidence interval on this model value is  $70.13 \pm 0.27$ : precise to about a quarter of an inch. This high precision reflects the small standard errors of the coefficients which arise in turn from the large amount of data used to fit the model.

It is wrong to interpret this interval as saying something about the actual range of heights of men like Bill, that is, men whose mother is 67 inches and father 69 inches. The model-value confidence interval should not be interpreted as saying that 95 percent of such men will have heights in the interval  $70.13 \pm 0.27$ . Instead, this interval means that if you were to fit a model based on the entire population — not just the 898 cases in Galton's data — the model you would fit would likely produce a model value for Bill close to 70.13. In other words, the model-value confidence interval is not so much about the uncertainty in Bill's height as in what the model has to say about the average for men like Bill.

If you are interested in what the model has to say about the uncertainty in Bill's height, you need to ask a different question and compute a different confidence interval. The prediction confidence interval takes into account the spread of the cases around their model values: the residuals. For the model given above, the standard deviation of the residuals is 2.15 inches — a typical person varies from the model value by that amount.

The prediction confidence interval takes into account this case-by-case residual to give an indication of the range of heights into which the actual value is likely to fall. For men like Bill, the 95% prediction interval is  $70.13 \pm 4.24$  inches. This is much larger than the interval on the model values, and reflects mainly the size of the residual of actual cases around their model values.

**Example 14.5: Catastrophe in Grand Forks** In April 1997, there was massive flooding on the Red River in Minnesota and North Dakota, states in the north-

For review purposes only

ern US, due to record setting winter snows. The towns of Grand Forks and East Grand Forks were endangered and the story was in the news. I remember hearing a news report saying that the dikes in Grand Forks could protect against a flood level of 50 feet and that the National Weather Service predictions were for the river to reach a maximum of 47.5 to 49 feet. To the reporter and the city planners, this was good news. The city had never been better prepared and the preparations were paying off. To me, even knowing nothing about the area, the report was a sign of trouble. What kind of confidence interval was this 47.5 to 49? No confidence level was reported. Was it at a 50% level, was it at a 95% level? Was it a confidence interval on the model values alone or did it include the residuals from the model values? Did it take into account the extrapolation involved in handling record-setting conditions? Nothing in the news stories gave any insight into how precise previous predictions had been.

In the event, the floods reached 54.11 feet in Grand Forks, overtopping the dikes and inundating both towns. Damage was estimated to be more than \$1 billion, a huge amount given the small population of the area. In the aftermath of the flood, the mayor of East Grand Forks said, understandably, “They [the National Weather Service] missed it, and they not only missed it, they blew it big.” The Grand Forks city engineer lamented, “with proper advance notice we could have protected the city to almost any elevation . . . if we had known, I’m sure that we could have protected a majority of the city.”

But all the necessary information was available at the time. The forecast would have been accurate if a proper prediction confidence interval had been given. It turns out that the quoted 47.5 to 49 foot interval was not a confidence interval at all — it was the range of predictions from a model under two different scenarios, with and without ongoing precipitation.

Looking back on the history of predictions from the National Weather Service, the typical residual was about 11% of the prediction. Thus, a reasonable 95% confidence interval might have been  $\pm 22\%$ , or, translated to feet,  $48 \pm 10$  feet. Had this interval been presented, the towns might have been better prepared for the actual level of 54.11 feet, well within the confidence band.

Whether a 95% confidence level is appropriate for disaster planning is an open question and reflects the balance between the costs of preparation and the potential damage. If you plan using a 95% level, the upper boundary of the interval will be exceeded something like 2.5% of the time. This might be acceptable, or it might not be.

What shouldn’t be controversial is that confidence intervals need to come with a clear statement of what they mean. For disaster planning, a model-value confidence interval is not so useful — it’s about the quality of the model rather than the uncertainty in the actual outcome.

[Much of this example is drawn from the account by Roger Pielke [24].]

For review purposes only

## 14.6 Finding the Resampling Distribution

When you fit a model to a dataset consisting of random samples from a population, the resulting coefficients will be somewhat random. The sampling distribution describes this variation.

One can in principle find the sampling distribution by doing the work of actually drawing many random samples from the population. In practice, there is little need to do this. That's fortunate, since samples are typically collected with great difficulty and expense and there is often no real possibility of collecting much more just to find the sampling distribution.

Instead, the same sample to which the model was fitted is used to construct a different but closely related distribution, called the **resampling distribution**, which approximates the important properties of the sampling distribution, particularly the standard error or the 95% coverage interval.

This section shows two methods to construct resampling distributions: resampling and simulation. These methods are rooted in simple notions of randomness. Both of the methods can and are used in practice. For linear models, it's possible to build a simple theory of standard errors and to give formulas for them. Such formulas are used by statistical software in constructing the regression report.

### 14.6.1 Resampling

Here's a simple idea: use the sample itself to stand for the population. The sample is already in hand, in the form of a data frame, so it's easy to draw cases of it. Such new samples, taken from your original sample, not from the population, are called **resamples**. Sampling from the sample.

Figure 14.3 illustrates how resampling works. There is just one sample drawn (with the concordant expense) from the real population. Thereafter, the sample itself is used as a stand-in for the population and new samples are drawn from the sample.

Will such resamples capture the sampling variation that would be expected if you were genuinely drawing new samples from the population? An objection might come to mind: If you draw  $n$  cases out of a sample consisting of  $n$  cases, the resample will look exactly like the sample itself. No variability. As you'll see, this problem is easily overcome by **sampling with replacement**.

To illustrate, consider a very small sample of size  $n = 5$ , although in practice one uses resampling only for samples that are much bigger, say  $n > 20$ .

state	net	age	sex
MD	5701	53	M
VA	5610	28	F
DC	5621	24	F
MA	5541	45	F
DC	4726	34	M

DATA FILE  
ten-mile-  
race.csv

For review purposes only

Taking this sample as exactly representative of the population, imagine that the population looks like this:

state	net	age	sex
MD	5701	53	M
MD	5701	53	M
MD	5701	53	M
and so on ...			
VA	5610	28	F
VA	5610	28	F
VA	5610	28	F
and so on ...			
DC	5621	24	F
DC	5621	24	F
DC	5621	24	F
and so on ...			
MA	5541	45	F
MA	5541	45	F
MA	5541	45	F
and so on ...			
DC	4726	34	M
DC	4726	34	M
DC	4726	34	M
and so on ...			

The simulated population looks just like the sample, but every case in the sample is repeated over and over again. Thus, the population is very large, but the cases in it look just like the sample.

The resampling process is arranged to make it seem that the sample itself is infinite in size. This is accomplished by sampling with replacement: whenever a case is drawn from the sample to put in a resample, the case is put back so that it is available to be used again. This is not something you would do when collecting the original sample; in sampling (as opposed to resampling) you don't use a case more than once.

The resamples in Figure 14.3 may seem a bit odd. They often repeat cases and omit cases. And, of course, any case in the population that was not included in the sample cannot be included in any of the resamples. Even so, the resamples do the job; they simulate the variation in model coefficients introduced by the process of random sampling.

It's important to emphasize what the resamples do not and cannot do: they don't construct the sampling distribution. The resamples merely show what the sampling distribution would look like *if the population looked like your sample*. The center of the resampling distribution from any given sample is generally not aligned exactly with the center of the sampling distribution. However, in practice, the width of the resampling distribution is a good match to the width

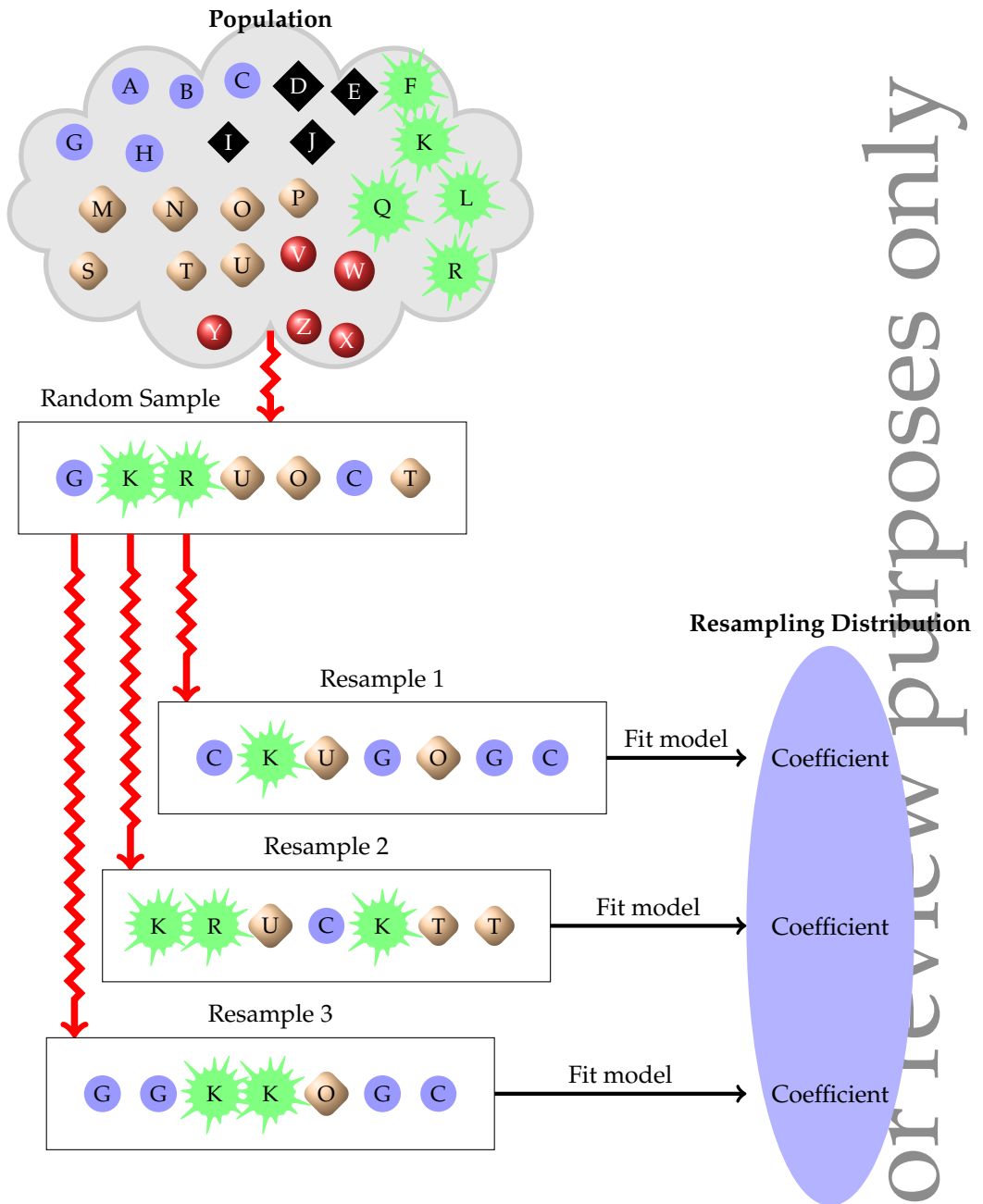


Figure 14.3: Resampling draws randomly from the sample to create new samples.



of the sampling distribution. The resampling distribution is adequate for the purpose of finding standard errors and margins of error.

Figure 14.4 shows an experiment to demonstrate this. At the top of the figure is the sampling distribution on the *age* coefficient in the running model. This was found by drawing 1000 genuine repeated samples of size  $n = 100$  from the population. The remaining panels show resampling distributions, each panel based on a single genuine sample of size  $n = 100$  from the population and then drawing 1000 resamples from that sample. The resampling distribution varies from sample to sample. Sometimes it's to the left of the sampling distribution, sometimes to the right, occasionally well aligned. Consistently, however, the resampling distributions have a standard error that is a close match to the standard error of the sampling distribution.

The process of finding confidence intervals directly via resampling is called **bootstrapping**.

### 14.6.2 Randomizing the Residuals

Models partition variation in the response variable into two parts. In the past, these parts have been called the “explained” and “unexplained,” or the “model values” and the “residuals,” or the “modeled” or “unmodeled” parts. For this discussion, consider them the *deterministic* and the *random* parts.

The notation used has focussed on the deterministic part of the model. So, a model formula has been written like

$$\begin{array}{c} \text{Model} \\ \text{Values} \end{array} \quad \text{net} = 5540 + 16.9 \text{ age} - 727 \text{ sexM}.$$

The above formula doesn't include the residuals. Of course, the residuals are defined as the difference between the model values and the actual values of the response variable. So the response values can be written

$$\text{net} = 5540 + 16.9 \text{ age} - 727 \text{ sexM} + \mathcal{E}$$

where  $\mathcal{E}$  is the vector of random numbers, one random number for each case.

This process is illustrated in Figure 14.5, which depicts the situation when the response variable *A* is deterministically set by another variable *B*, a population coefficient which can be written  $c_B$ , and a random component  $\mathcal{E}$  as

$$A = Bc_B + \mathcal{E}.$$

Conventionally, the random component is called the **error**. The word comes from the Latin *errare*, meaning “to wander” or “to stray.” After following the determinist path to  $Bc_B$ , the system strays off the path to *A*.

Each case comes with its own random component, so one way to think about random sampling is it is in effect picking a random set of numbers to add to the cases in the sample. If a different sample had been picked, the random numbers would have been different.

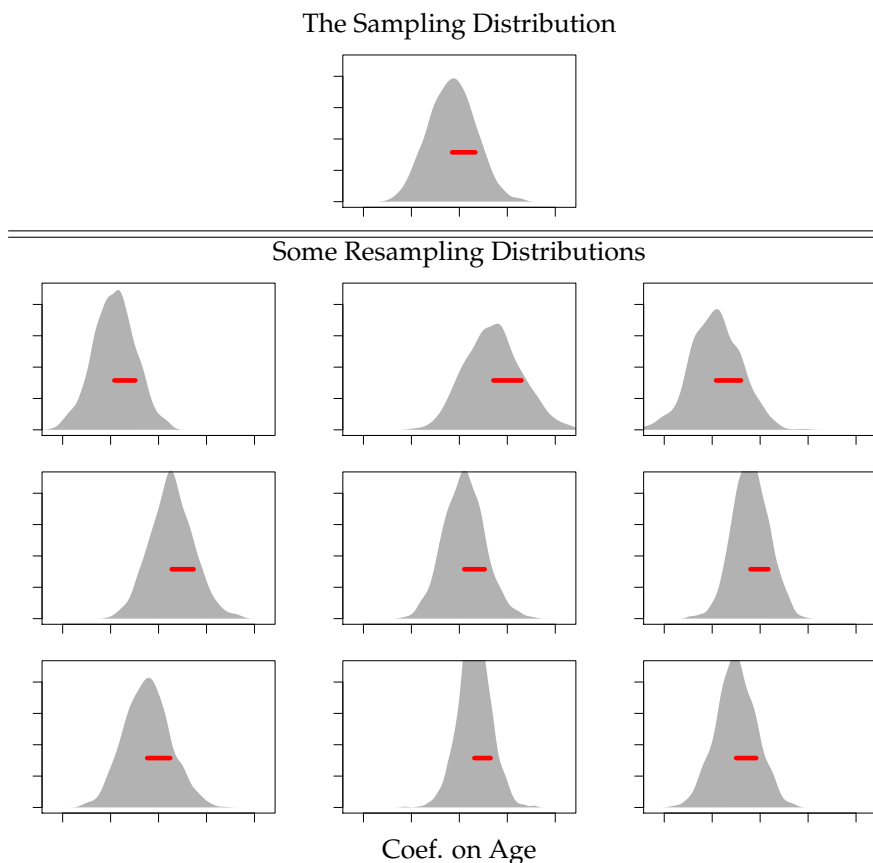


Figure 14.4: Top: Sampling distribution for the coefficient on age for samples of size  $n = 100$ . The standard error is shown as bar. Below: Several resampling distributions, each from one sample of size  $n = 100$ . The position of the resampling distributions doesn't always match the sampling distribution, but the standard error is often a good match.

Fitting a model to a sample tries to infer the value of the coefficient from measurements of A and B by projecting A onto B. If the random component  $\mathcal{E}$  had been perpendicular to B, this projection would give exactly the right answer. But  $\mathcal{E}$  is random, so it won't necessarily be perpendicular to B. (However, if the sample size  $n$  is large,  $\mathcal{E}$  will almost certainly be close to perpendicular to B.) Because of the randomness, the projection  $A_{\parallel B}$  won't be a perfect match to the deterministic mechanism. That's why the sample doesn't generally give the same coefficients that would be found from the entire population.

To create a resampling distribution take the fitted model values  $A_{\parallel B}$  as the deterministic part of the mechanism and add in random values.

But how big should those random values be? A clue is given by the size of the residuals estimated when the model was fit to the data: make the simulated  $\mathcal{E}$  just like the residuals.

# Chapter 15

## The Logic of Hypothesis Testing

*Extraordinary claims demand extraordinary evidence.* — Carl Sagan

A **hypothesis test** is a standard format for assessing statistical evidence. It is ubiquitous in scientific literature, most often appearing in the form of statements of **statistical significance** and notations like “ $p < 0.01$ ” that pepper scientific journals.

Hypothesis testing involves a substantial technical vocabulary: null hypotheses, alternative hypotheses, test statistics, significance, power, p-values, and so on. The last section of this chapter lists the terms and gives definitions.

The technical aspects of hypothesis testing arise because it is a highly formal and quite artificial way of reasoning. This isn’t a criticism. Hypothesis testing is this way because the “natural” forms of reasoning are inappropriate. To illustrate why, consider an example.

The stock market’s ups and downs are reported each working day. Some people make money by investing in the market, some people lose. Is there reason to believe that there is a trend in the market that goes beyond the random-seeming daily ups and downs?

Figure 15.1 shows the closing price of the Dow Jones Industrial Average stock index for a period of about 10 years when stocks were considered a good investment. It’s evident that the price is going up and down in an irregular way, like a random walk. But it’s also true that the price at the end of the period is much higher than the price at the start of the period.

Is there a trend or is this just a random walk? It’s undeniable that there are fluctuations that look something like a random walk, but is there a trend buried under the fluctuations?

As phrased, the question contrasts two different possible hypotheses. The first is that the market is a pure random walk. The second is that the market has

For review purposes only

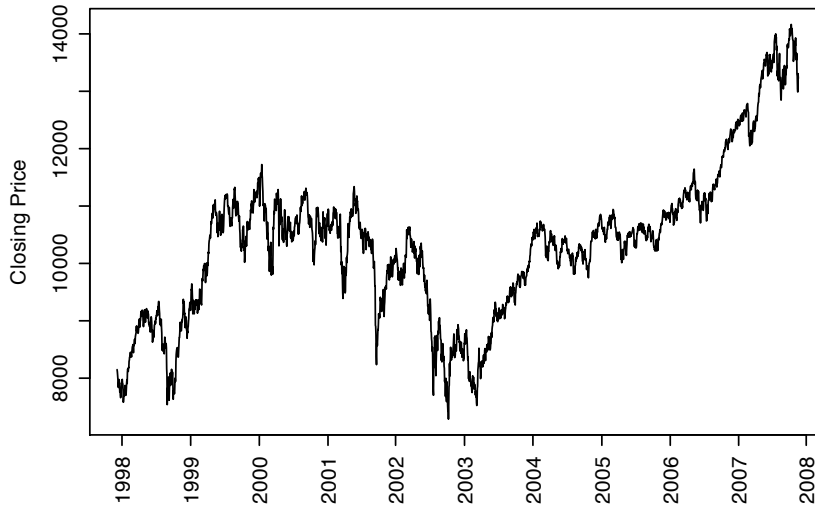


Figure 15.1: The closing price of the DJIA each day over 2500 trading days — a roughly 10 year period from the close on Dec. 5, 1997 to the close on Nov. 14, 2007, about a year before the "great recession" of 2008/9.

a systematic trend in addition to the random walk.

The natural question to ask is this: Which hypothesis is right?

Each of the hypotheses is actually a model: a representation of the world for a particular purpose. But each of the models is an incomplete representation of the world, so each is wrong.

It's tempting to rephrase the question slightly to avoid the simplistic idea of right versus wrong models: Which hypothesis is a better approximation to the real world? That's a nice question, but how to answer it in practice? To say how each hypothesis differs from the real world, you need to know already what the real world is like: Is there a trend in stock prices or not? That approach won't take you anywhere.

Another idea: Which hypothesis gives a better match to the data? This seems a simple matter: fit each of the models to the data and see which one gives the better fit. But recall that even junk model terms can lead to smaller residuals. In the case of the stock market data, it happens that the model that includes a trend will almost always give smaller residuals than the pure random walk model, even if the data really do come from a pure random walk.

The logic of hypothesis testing avoids these problems. The basic idea is to avoid having to reason about the real world by setting up a hypothetical world that is completely understood. The observed patterns of the data are then compared to what would be generated in the hypothetical world. If they don't match, then there is reason to doubt that the data support the hypothesis.

## 15.1 An Example of a Hypothesis Test

To illustrate the basic structure of a hypothesis test, here is one using the stock-market data.

The **test statistic** is a number that is calculated from the data and summarizes the observed patterns of the data. A test statistic might be a model coefficient or an  $R^2$  value or something else. For the stock market data, it's sensible to use as the test statistic the start-to-end dollar difference<sup>1</sup> in prices over the 2500-day period. The observed value of this test statistic is \$5074.80 — the DJIA stocks went up by this amount over the 10-year period.

This test statistic can be used to test the hypothesis that the stock market is a random walk. (The reasons to choose the random walk hypothesis instead of the trend hypothesis will be discussed later.)

In a random-walk world the start-to-end price difference would be random. As described in Section 13.3 the price difference is a random variable with a mean of 0 and standard deviation of  $s\sqrt{n}$ , where  $s$  is the typical daily fluctuation. Since the data cover 2500 days, it's safe to set  $n = 2500$ . But what should the parameter  $s$  be? It was not specified by the hypothesis. Such an unknown parameter is called a **nuisance parameter**. You need to know it in order to say what would happen in the hypothetical world, but the hypothesis doesn't state it explicitly.

If the hypothesis were true, you would have a way to find  $s$ : measure it from your data. This can be done by taking the standard deviation of day-to-day price changes. For the stock market data, it's \$106.70. That is, the DJIA typically went up or down by about 100 dollars per day for the 10-year period covered by the data.

This gives a complete description of what the test statistic should look like in the hypothesized world: the start-to-end price difference in dollars will be a normal distribution with mean 0 and standard deviation of  $106.70 \times \sqrt{2500} = 5335$  dollars. This distribution is drawn in Figure 15.2. Also shown is a tick mark at the observed value of the test statistic, \$5074.80. It seems obvious from the figure that the observed value is quite plausible as an outcome in the world of the hypothesis. In other words, the hypothesis is consistent with the data.

It's tempting to use this result to say, perhaps, "the observations support the hypothesis." In actuality, however, the permitted conclusion is stiff and unnatural:

*We fail to reject the hypothesis.*

## 15.2 Inductive and Deductive Reasoning

Hypothesis testing involves a combination of two different styles of reasoning: deduction and induction. In the deductive part, the hypothesis tester makes an

<sup>1</sup>Another way to describe the change in stock prices is by the *proportional increase*, which is 62.3% for the DJIA over the period in Figure 15.1: a rate of 5% per year when compounded. Economists usually prefer to study the proportional change rather than the dollar change.

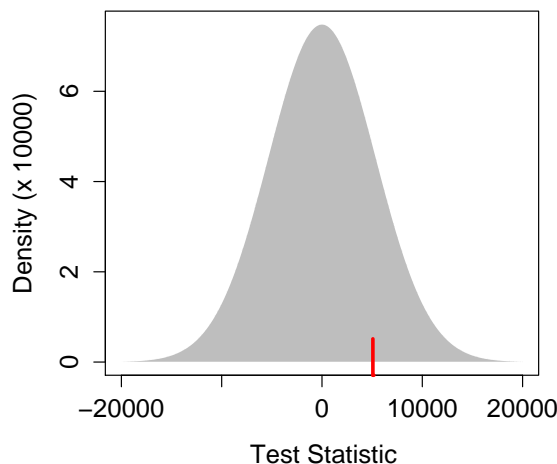


Figure 15.2: The distribution of start-to-end differences in stock price under the hypothesis that day-to-day changes in price are random with no trend. The value observed in the data, \$5074.80, is marked with a tick.

assumption about how the world works and draws out, deductively, the consequences of this assumption: what the observed value of the test statistic should be if the hypothesis is true. For instance, the hypothesis that stock prices are a random walk was translated into a statement of the probability distribution of the start-to-end price difference.

In the inductive part of a hypothesis test, the tester compares the actual observations to the deduced consequences of the assumptions and decides whether the observations are consistent with them.

### 15.2.1 Deductive Reasoning

**Deductive reasoning** involves a series of rules that bring you from given assumptions to the consequences of those assumptions. For example, here is a form of deductive reasoning called a **syllogism**:

**Assumption 1** No healthy food is fattening.

**Assumption 2** All cakes are fattening.

**Conclusion** No cakes are healthy.

The actual assumptions involved here are questionable, but the pattern of logic is correct. If the assumptions were right, the conclusion would be right also.

Deductive reasoning is the dominant form in mathematics. It is at the core of mathematical proofs and lies behind the sorts of manipulations used in algebra. For example, the equation  $3x + 2 = 8$  is a kind of assumption. Another assumption, known to be true for numbers, is that subtracting the same amount from both sides of an equation preserves the equality. So you can subtract 2 from both sides to get  $3x = 6$ . The deductive process continues — divide both sides by 3 — to get a new statement,  $x = 2$ , that is a logical consequence of the initial assumption. Of course, if the assumption  $3x + 2 = 8$  was wrong, then the conclusion  $x = 2$  would be wrong too.

The **contrapositive** is a way of recasting an assumption in a new form that will be true so long as the original assumption is true. For example, suppose the original assumption is, “My car is red.” Another way to state this assumption is as a statement of implication, an if-then statement:

**Assumption** If it is my car, then it is red.

To form the contrapositive, you re-arrange the assumption to produce another statement:

**Contrapositive** If it is **not** red, then it is **not** my car.

Any assumption of the form “if [statement 1] then [statement 2]” has a contrapositive. In the example, statement 1 is “it is my car.” Statement 2 is “it is red.” The contrapositive looks like this:

**Contrapositive** If [negate statement 2] then [negate statement 1]

The contrapositive is, like algebraic manipulation, a re-arrangement: reverse and negate. Reversing means switching the order of the two statements in the if-then structure. Negating a statement means saying the opposite. The negation of “it is red” is “it is not red.” The negation of “it is my car” is “it is not my car.” (It would be wrong to say that the negation of “it is my car” is “it is your car.” Clearly it’s true that if it is your car, then it is not my car. But there are many ways that the car can be not mine and yet not be yours. There are, after all, many other people in the world than you and me!)

Contrapositives often make intuitive sense to people. That is, people can see that a contrapositive statement is correct even if they don’t know the name of the logical re-arrangement. For instance, here is a variety of ways of re-arranging the two clauses in the assumption, “If that is my car, then it is red.” Some of the arrangements are logically correct, and some aren’t.

Original Assumption: *If it is my car, then it is red.*

— Negate first statement: *If it is not my car, then it is red.*

Wrong. Other people can have cars that are not red.

— Negate only second statement: *If it is my car, then it is not red.*

Wrong. The statement contradicts the original assumption that my car is red.

— Negate both statements: *If it is not my car, then it is not red.*

Wrong. Other people can have red cars.

— Reverse statements: *If it is red, then it is my car.*

Wrong. Apples are red and they are not my car. Even if “it” is a car, not every red car is mine.

— Reverse and negate first: *If it is red, then it is not my car.*

Wrong. My car is red.

— Reverse and negate second: *If it is not red, then it is my car.*

Wrong. Oranges are not red, and they are not my car.

— Reverse and negate both — the contrapositive: *If it is not red, then it is not my car.*

Correct.

### 15.2.2 Inductive Reasoning

In contrast to deductive reasoning, **inductive reasoning** involves generalizing or extrapolating from a set of observations to conclusions. An observation is not an assumption: it is something we see or otherwise perceive. For instance, you can go to Australia and see that kangaroos hop on two legs. Every kangaroo you see is hopping on two legs. You conclude, inductively, that all kangaroos hop on two legs.

Inductive conclusions are not necessarily correct. There might be one-legged kangaroos. That you haven’t seen them doesn’t mean they can’t exist. Indeed, Europeans believed that all swans are white until explorers discovered that there are black swans in Australia.

Suppose you conduct an experiment involving 100 people with fever. You give each of them aspirin and observe that in all 100 the fever is reduced. Are you entitled to conclude that giving aspirin to a person with fever will reduce the fever? Not really. How do you know that there are no people who do not respond to aspirin and who just happened not to be included in your study group?

Perhaps you’re tempted to hedge by weakening your conclusion: “Giving aspirin to a person with fever will reduce the fever most of the time.” This seems reasonable, but it is still not necessarily true. Perhaps the people in your study had a special form of fever-producing illness and that most people with fever have a different form.



By the standards of deductive reasoning, inductive reasoning does not work. No reasonable person can argue about the deductive, contrapositive reasoning concerning the red car. But reasonable people can very well find fault with the conclusions drawn from the study of aspirin.

Here's the difficulty. If you stick to valid deductive reasoning, you will draw conclusions that are correct given that your assumptions are correct. But how can you know if your assumptions are correct? How can you make sure that your assumptions adequately reflect the real world? At a practical level, most knowledge of the world comes from observations and induction.

The philosopher David Hume noted the everyday inductive "fact" that food nourishes us, a conclusion drawn from everyday observations that people who eat are nourished and people who do not eat waste away. Being inductive, the conclusion is suspect. Still, it would be a foolish person who refuses to eat for want of a deductive proof of the benefits of food.

Inductive reasoning may not provide a proof, but it is nevertheless useful.

### 15.3 The Null Hypothesis

A key aspect of hypothesis testing is the choice of the hypothesis to test. The stock market example involved testing the random-walk hypothesis rather than the trend hypothesis. Why? After all, the hypothesis of a trend is more interesting than the random-walk hypothesis; it's more likely to be useful if true.

It might seem obvious that the hypothesis you should test is the hypothesis that you are most interested in. But this is wrong.

In a hypothesis test one *assumes* that the hypothesis to be tested is true and draws out the consequences of that assumption in a deductive process. This can be written as an if-then statement:

If hypothesis  $H$  is true, then the test statistic  $S$  will be drawn from a probability distribution  $P$ .

For example, in the stock market test, the assumption that the prices are a random walk led to the conclusion that the test statistic — the start-to-end price difference — would be a draw from a normal distribution with mean 0 and standard deviation 5335 dollars.

The inductive part of the test involves comparing the observed value of the test statistic  $S$  to the distribution  $P$ . There are two possible outcomes of this comparison:

**Agreement**  $S$  is a plausible outcome from  $P$ .

**Disagreement**  $S$  is not a plausible outcome from  $P$ .

Suppose the outcome is agreement between  $S$  and  $P$ . What can be concluded? Not much. Recall the statement "If it is my car, then it is red." An observation of a red car does not legitimately lead to the conclusion that the car is mine. For an if-then statement to be applicable to observations, one needs to observe the if-part of the statement, not the then-part.

An outcome of disagreement gives a more interesting result, because the contrapositive gives logical traction to the observation; “If it is not red, then it is not my car.” Seeing “not red” implies “not my car.” Similarly, seeing that S is not a plausible outcome from P, tells you that H is not a plausible possibility. In such a situation, you can legitimately say, “I reject the hypothesis.”

Ironically, in the case of observing agreement between S and P, the only permissible statement is, “I fail to reject the hypothesis.” You certainly aren’t entitled to say that the evidence causes you to accept the hypothesis.

This is an emotionally unsatisfying situation. If your observations are consistent with your hypothesis, you certainly want to accept the hypothesis. But that is not an acceptable conclusion when performing a formal hypothesis test. There are only two permissible conclusions from a formal hypothesis test:

- I reject the hypothesis.
- I fail to reject the hypothesis.

In choosing a hypothesis to test, you need to keep in mind two criteria.

**Criterion 1** The only possible interesting outcome of a hypothesis test is “I reject the hypothesis.” So make sure to pick a hypothesis that it will be interesting to reject.

The role of the hypothesis is to be refuted or nullified, so it is called the **null hypothesis**.

What sorts of statements are interesting to reject? Often these take the form of the **conventional wisdom** or of **no effect**.

For example, in comparing two fever-reducing drugs, an appropriate null hypothesis is that the two drugs have the same effect. If you reject the null, you can say that they don’t have the same effect. But if you fail to reject the null, you’re in much the same position as before you started the study.

Failing to reject the null may mean that the null is true, but it equally well may mean only that your work was not adequate: not enough data, not a clever enough experiment, etc. Rejecting the null can reasonably be taken to indicate that the null hypothesis is false, but failing to reject the null tells you very little.

**Criterion 2** To perform the deductive stage of the test, you need to be able to calculate the range of likely outcomes of the test statistic. This means that the hypothesis needs to be specific.

The assumption that stock prices are a random walk has very definite consequences for how big a start-to-end change you can expect to see. On the other hand, the assumption “there is a trend” leaves open the question of how big the trend is. It’s not specific enough to be able to figure out the consequences.

## 15.4 The p-value

One of the consequences of randomness is that there isn’t a completely clean way to say whether the observations fail to match the consequences of the null

hypothesis. In principle, this is a problem even with simple statements like “the car is red.” There is a continuous range of colors and at some point one needs to make a decision about how orange the car can be before it stops being red.

Figure 15.2 shows the probability distribution for the start-to-end stock price change under the null hypothesis that stock prices are a random walk. The observed value of the test statistic, \$5074.80, falls under the tall part of the curve — it’s a plausible outcome of a random draw from the probability distribution.

The conventional way to measure the plausibility of an outcome is by a **p-value**. The p-value of an observation is always calculated with reference to a probability distribution derived from the null hypothesis.

P-values are closely related to percentiles. The observed value \$5074.80 falls at the 83rd percentile of the distribution. This is plausible because it’s in the middle of the distribution. An observation that’s at or beyond the extremes of the distribution is implausible. This would correspond to either very high percentiles or very low percentiles. Being at the 83rd percentile implies that 17 percent of draws would be even more extreme, falling even further to the right than \$5074.80.

The p-value is the fraction of possible draws from the distribution that are as extreme or more extreme than the observed value. If the concern is only with values bigger than \$5074.80, then the p-value is 0.17.

This p-value of 0.17 is called a **one-tailed** p-value, since it considers only events that are extreme to one side of the distribution. Of course, an observation might also be extreme in the other way. For the stock prices, this would be a large negative start-to-end change, corresponding to a downward trend in stock prices. To take such possible draws into account, you double the p-value. Thus, the p-value for the observed start-to-end stock price change (under the null hypothesis of a random walk) is  $2 \times 0.17 = 0.34$ .

A small p-value indicates that the actual value of the test statistic is quite surprising as an outcome from the null hypothesis. A large p-value means that the test statistic value is run of the mill, not surprising, not enough to satisfy the “if” part of the contrapositive.

The convention in hypothesis testing is to consider the observation as being implausible when the p-value is less than 0.05. In the stock market example, the p-value is larger than 0.05, so the outcome is to fail to reject the null hypothesis that stock prices are a random walk with no trend.

## 15.5 Rejecting by Mistake

The p-value for the hypothesis test of the possible trend in stock-price was 0.34, not small enough to justify rejecting the null hypothesis that stock prices are a random walk with no trend. A smaller p-value, one less than 0.05 by convention, would have led to rejection of the null. The small p-value would have indicated that the observed value of the test statistic was implausible in a world where the null hypothesis is true.

For review purposes only

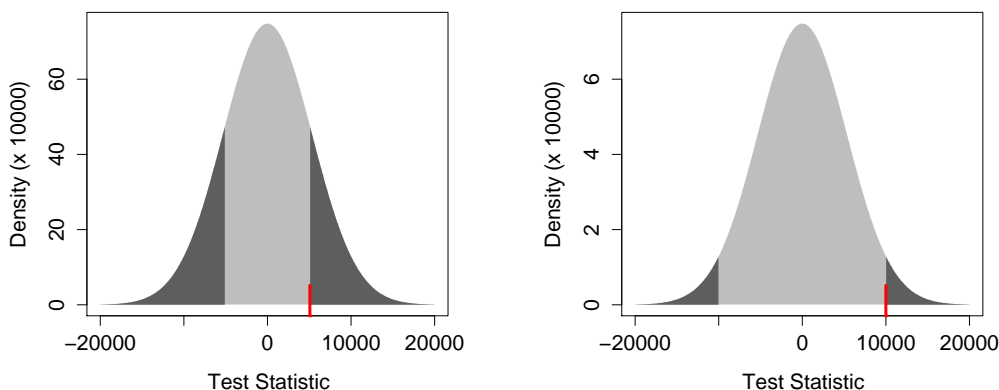


Figure 15.3: The p-value is a probability of observing a test statistic as extreme as the actual value, or more extreme. Like percentiles, this corresponds to the area under the probability distribution for observations assuming that the null hypothesis is true. Left: the shaded area corresponds to the p-value for the actual observation of a start-to-end change of \$5074.80. This p-value is 0.34. Right: If the observation had instead been \$10,000.00, the p-value would have been smaller: about 0.06.

Now turn this around. Suppose the null hypothesis really were true; suppose stock prices really are a random walk with no trend. In such a world, it's still possible to see an implausible value of the test statistic. But, if the null hypothesis is true, then seeing an implausible value is misleading; rejecting the null is a mistake. This sort of mistake is called a **Type I error**.

Such mistakes are not uncommon. In a world where the null is true — the only sort of world where you can falsely reject the null — they will happen 5% of the time so long as the threshold for rejecting the null is a p-value of 0.05.

The way to avoid such mistakes is to lower the p-value threshold for rejecting the null. Lowering it to, say, 0.01, would make it harder to mistakenly reject the null. On the other hand, it would also make it harder to correctly reject the null in a world where the null ought to be rejected.

The threshold value of the p-value below which the null should be rejected is a probability: the probability of rejecting the null in a world where the null hypothesis is true. This probability is called the **significance level** of the test.

It's important to remember that the significance level is a **conditional probability**. It is the probability of rejecting the null in a world where the null hypothesis is actually true. Of course that's a hypothetical world, not necessarily the real world.

## 15.6 Failing to Reject

In the stock-price example, the large p-value of 0.34 led to a failure to reject the null hypothesis that stock prices are a random walk. Such a failure doesn't mean that the null hypothesis is true, although it's encouraging news to people who want to believe that the null hypothesis is true.

You never get to “accept the null” because there are reasons why, even if the null were wrong, it might *not* have been rejected:

- You might have been unlucky. The randomness of the sample might have obscured your being able to see the trend in stock prices.
- You might not have had enough data. Perhaps the trend is small and can't easily be seen.
- Your test statistic might not be sensitive to the ways in which the system differs from the null hypothesis. For instance, suppose that there is an small average tendency for each day's activity on the stock market to undo the previous day's change: the walk isn't exactly random. Looking for large values of the start-to-end price difference will not reveal this violation of the null. A more sensitive test statistic would be the correlation between price changes on successive days.

A helpful idea in hypothesis testing is the **alternative hypothesis**: the pet idea of what the world is like if the null hypothesis is wrong. The alternative hypothesis plays the role of the thing that you would like to prove. In the hypothesis-testing drama, this is a very small role, since the only possible outcomes of a hypothesis test are (1) reject the null and (2) fail to reject the null. The alternative hypothesis is not directly addressed by the outcome of a hypothesis test.

The role of the alternative hypothesis is to guide you in interpreting the results if you do fail to reject the null. The alternative hypothesis is also helpful in deciding how much data to collect.

To illustrate, suppose that the stock market really does have a trend hidden inside the random day-to-day fluctuations with a standard deviation of \$106.70. Imagine that the trend is \$2 per day: a pet hypothesis. If this were true, the start-to-end change in the price over  $N = 2500$  days would be a draw from a normal distribution with mean  $\$2 \times 2500 = \$5000$  and standard deviation  $\$106.70 \times \sqrt{2500} = \$5335$ .

Suppose the world really were like the alternative hypothesis. What is the probability that, in such a world, you would end up failing to reject the null hypothesis? (Such a mistake, where you fail to reject the null in a world where the alternative is actually true, is called a **Type II error**.)

The probability of rejecting the null in a world where the alternative is true is called the **power** of the hypothesis test. Of course, if the alternative is true, then it's completely appropriate to reject the null, so a large power is desirable.

**Aside. 15.1** Calculating a Power

Here are the steps in calculating the power of the hypothesis test of stock market prices. The null hypothesis is that prices are a pure random walk. The alternative hypothesis is that they have a trend of \$2 per day.

1. Go back to the null hypothesis world and find the thresholds for the test statistic that would cause you to reject the null hypothesis. Referring to Figure 15.3, you can see that a test statistic of \$10,000.00 would have produced a p-value of 0.061, close to the rejection threshold. So if the test statistic were a little larger, you would have reached the threshold for rejection. The exact threshold value turns out to be \$10,456.41, which is the 97.5th percentile for the null hypothesis distribution (the normal distribution with mean zero and standard deviation of \$5335). So, you would have rejected the null at a significance level of 0.05 if the test statistic had been bigger than \$10,456.41.
2. Now return to the alternative hypothesis world. In this world, what is the probability that the test statistic would have been bigger than \$10,456.41? This is a straightforward probability calculation, since the distribution of the test statistic in the alternative hypothesis world is normal with mean \$5000 and standard deviation \$5335. Doing the calculation gives a probability of 0.15.

Section 17.7 discusses power calculations for models.

A power calculation involves considering both the null and alternative hypotheses. Aside 15.1 shows the logic applied to the stock-market question. It results in a power of 15%.

The power of 15% for the stock market test means that even if the pet theory of the \$2 trend were correct, there is only a 15% chance of rejecting the null. In other words, the study is quite weak.

When the power is small, failure to reject the null can reasonably be interpreted as a failure in the modeler (or in the data collection or in the experiment). The study has given very little information.

Just because the power is small is no reason to doubt the null hypothesis. Instead, you should think about how to conduct a better, more powerful study.

One way a study can be made more powerful is to increase the sample size. Rather than studying the trend over  $N = 2500$  days, perhaps it should have been studied over twice or three times as long. To illustrate, consider what would happen in a study with  $N = 5000$  days of data.

**Null Hypothesis** In the world where the null is true, the start-to-end change would be normal with mean \$0 and standard deviation  $\$106.70 \times \sqrt{5000} = \$7544.83$ . The threshold for rejection will be a start-to-end change of \$14,787.66 or bigger.

**Alternative Hypothesis** In the alternative hypothesis world, the start-to-end change in price will be normal with mean  $\$2 \times 5000 = \$10,000$  and standard

deviation  $\$106.70 \times \sqrt{5000} = \$7544.83$ . The probability of the start-to-end change being above the threshold for rejecting the null hypothesis is 26%.

So, even a sample size of  $n = 5000$  doesn't give a very powerful study: there is little reason to think anyone could reject the null even if there is a trend in stock prices.

The logic of the power calculation can be used to decide how big a study is needed. Repeating the power calculation for different values of  $n$  gives the following powers:

$n$	Power
2500	15%
5000	26%
10000	47%
20000	76%
30000	90%

Reliably detecting a \$2 per day trend requires a lot of data. For instance  $n = 20000$  is about 80 years of data. This long historical period is probably not relevant to today's investor. Indeed, it's just about all the data that is actually available: the DJIA was started in 1928.

When the power is small for realistic amounts of data, the phenomenon you are seeking to find may be undetectable.

## 15.7 A Glossary of Hypothesis Testing

**Null Hypothesis** A statement about the world that you are interested to disprove. The null is almost always something that is clearly relevant and not controversial: that the conventional wisdom is true or that there is no relationship between variables. Examples: "The drug has no influence on blood pressure." "Smaller classes do not improve school performance."

The allowed outcomes of the hypothesis test relate only to the null:

- Reject the null hypothesis.
- Fail to reject the null hypothesis.

**Alternative Hypothesis** A statement about the world that motivates your study and stands in contrast to the null hypothesis. "The drug will reduce blood pressure by 5 mmHg on average." "Decreasing class size from 30 to 25 will improve test scores by 3%."

The outcome of the hypothesis test is not informative about the alternative. The importance of the alternative is in setting up the study: choosing a relevant test statistic and collecting enough data.

**Test Statistic** The number that you use to summarize your study. This might be the sample mean, a model coefficient, or some other number. Later

chapters will give several examples of test statistics that are particularly appropriate for modeling.

**Type I Error** A wrong outcome of the hypothesis test of a particular type. Suppose the null hypothesis were really true. If you rejected it, this would be an error: a type I error.

**Type II Error** A wrong outcome of a different sort. Suppose the alternative hypothesis were really true. In this situation, failing to reject the null would be an error: a type II error.

**Significance Level** A conditional probability. In the world where the null hypothesis is true, the significance is the probability of making a type I error. Typically, hypothesis tests are set up so that the significance level will be less than 1 in 20, that is, less than 0.05. One of the things that makes hypothesis testing confusing is that you do not know whether the null hypothesis is correct; it is merely assumed to be correct for the purposes of the deductive phase of the test. So you can't say what is the probability of a type I error. Instead, the significance level is the probability of a type I error *assuming* that the null hypothesis is correct.

Ideally, the significance level would be zero. In practice, one accepts the risk of making a type I error in order to reduce the risk of making a type II error.

**p-value** This is the usual way of presenting the result of the hypothesis test. It is a number that summarizes how atypical the observed value of the test statistic would be in a world where the null hypothesis is true. The convention for rejecting the null hypothesis is  $p < 0.05$ .

The p-value is closely related to the significance level. It is sometimes called the **achieved significance level**.

**Power** This is a conditional probability. But unlike the significance, the condition is that the alternative hypothesis is true. The power is the probability that, in the world where the alternative is true, you will reject the null. Ideally, the power should be 100%, so that if the alternative really were true the null hypothesis would certainly be rejected. In practice, the power is less than this and sometimes much less.

In science, there is an accepted threshold for the p-value: 0.05. But, somewhat strangely, there is no standard threshold for the power. When you see a study which failed to reject the null, it is helpful to know what the power of the study was. If the power was small, then failing to reject the null is not informative.

## 15.8 Computational Technique

The computational techniques described here relate to finding p-values and illustrating how power can be estimated.

For review purposes only



### 15.8.1 Computing p-values

The p-value always involves the position of an observed value of the test statistic within a sampling distribution. Once you have the observed value, finding the p-value involves three steps:

1. Finding the sampling distribution under the null hypothesis.
2. Finding the percentile of the observed value of the test statistic within the sampling distribution.
3. Translating the percentile to a p-value.

Later chapters will deal with ways to find the sampling distribution of model coefficients and other test statistics (such as  $R^2$ ). Here I'll assume that you already know the form of the sampling distribution either as a probability model or as a sample of values from the distribution.

To illustrate, consider the example of the start-to-end difference in the stock price. Under the null hypothesis that there is no systematic trend in prices, the sampling distribution described in the chapter is that the price difference will have a mean of zero and a standard deviation of 5335 dollars. The observed value of the test statistic was \$5074.80.

Since the sampling distribution in this case is a normal distribution, the percentile of the observed value in the sampling distribution can be found using the `pnorm` operator:

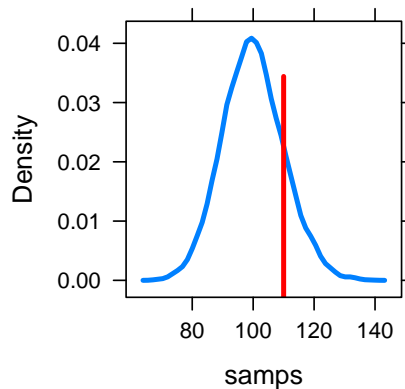
```
> pnorm( 5074.80, mean=0, sd=5335)
[1] 0.8293
```

This is not yet a p-value. The value 0.8293 says that the sampling distribution will generate a value that is less than the observed value (5074.80) on 82.9 percent of trials. Thus, there is a  $100 - 82.9 = 17.1$  percent chance that a randomly generated value would be greater than the observed value. The p-value is the chance that the randomly generated value would be *more extreme* than the observed value. "More extreme" might mean "bigger than" or "less than," depending on the context. In order to judge, it helps to have a picture of the situation. There happens to be one in figure 15.3.

In this case, a value *bigger* than 5074.80 will be more extreme, so the p-value is 17.1 percent. This is a one-tailed p-value, since it considers only one way to be more extreme. The two-tailed p-value includes the possibility that the test statistic might have been extreme but on the other side of the sampling distribution. The two-tailed p-value is, for the nicely symmetric normal distribution, twice the one-tailed value:  $2 \times 17.1 = 34.2$  percent, or 0.342.

It can be helpful to sketch your own pictures of sampling distributions in order to judge which tails of the distribution should be included. Here are the commands for making a simple plot to compare the sampling density to the observed value of the test statistic. In the example, the sampling density will be poisson with rate parameter `lambda=100` and the observed value will be 110.

```
samps = rpois( 10000, lambda=100 )
densityplot( samps, panel=hypothesis.test.panel,
             observed=110,
             plot.points=FALSE,lwd=3 )
```



From the graph, you can see that, in this case, “more extreme than” means “greater than.” In making your own plots, you would need to change the first line (`samps=...`) to match your own sampling distribution and the `observed=` value to match your own observation of the test statistic in the actual data.

In later chapters, you’ll see situations where you have only a sample from the sampling distribution. Translating this into a percentile can be done with `table` and `prop.table`. To illustrate, I’ll use simulation to generate a sample from the stock market price-change distribution.

Recall that the null hypothesis model of the stock market was a simple random walk in prices with no systematic trend. The standard deviation of price changes on each day was estimated from the data to be \$106.70. Over a period of  $N = 2500$  days the random walk can be simulated as the sum of 2500 normal random numbers with a mean of zero (no trend) and a standard deviation of \$106.70. Here’s one trial:

```
> sum( rnorm( 2500, mean=0, sd=106.70) )
[1] -3996
```

The simulated market went down by \$3996 over the 2500 days.

To draw a sizeable sample from this random process, use the `do` operator.

```
> samps = do(500)*sum( rnorm(2500,mean=0,sd=106.70) )
```

This command produces a sample of size  $n = 500$  from the random process — as if there were 500 different copies of the world in which the null hypothesis was active.

Now to the main point. Given a sample like this from the null hypothesis, the probability of the process generating a value that’s smaller than the observed test statistic can be found like this:

```
> pdata( 5074.80, samps)
[1] 0.834
```

This is more or less the same as what was found using the theoretically derived distribution and `pnorm`.

Power calculations involve finding rejection thresholds. To do this, you need to find the quantile at the appropriate level of the sampling distribution. For a 5% significance level, the appropriate quantile levels for a two-tailed test are 0.025 and 0.975. To find these from the sampling distribution, use the appropriate probability model or the samples:

```
> qnorm( c(.025, .975), mean=0, sd=5335)
[1] -10456 10456
> qdata( c(.025, .975), samps )
      2.5%      97.5%
-10369.931  9794.801
```

As always, the approach based on a sample from the distribution is subject to random fluctuations due to the small size of the sample. These are particularly acute when looking at quantiles near the extremes of the distributions. When the sampling distribution is normally shaped, as it often is, a better estimate can be had by taking the standard deviation, multiplying it out to reach the 95% coverage interval:

```
> sd(samps)*2
[1] 10511
```

For review purposes only

For review purposes only

# Chapter 16

## Hypothesis Testing on Whole Models

*A wise man ... proportions his belief to the evidence.* — David Hume  
(1711 – 1776)

Fitted models describe patterns in samples. Modelers interpret these patterns as indicating relationships between variables in the population from which the sample was drawn. But there is another possibility. Just as the constellations in the night sky are the product of human imagination applied to the random scattering of stars within range of sight, so the patterns indicated by a model might be the result of accidental alignments in the sample.

Deciding how seriously to take the patterns identified by a model is a problem that involves judgment. Are the patterns consistent with well established understanding of how the system works? Are the patterns corroborated by other sources of data? Are the model results sensitive to trivial changes in the model design?

Before undertaking that judgment, it helps to apply a much simpler standard of evidence. The conventional interpretation of a model such as  $A \sim B+C+\dots$  is that the variables on the right side of the modeler's tilde explain the response variable on the left side. The first question to ask is whether the explanation provided by the model is stronger than the "explanation" that would be arrived at if the variables on the right side were random — explanatory variables in name only without any real connection with the response variable  $A$ .

It's important to remember that in a hypothesis test, the null hypothesis is about the **population**. The null hypothesis claims that in the population the explanatory variables are unlinked to the response variable. Such a hypothesis does not rule out the possibility that, in the sample, the explanatory variables are aligned with the response variable. The hypothesis merely claims that any such alignment is accidental, due to the randomness of the sampling process.

For review purposes only

## 16.1 The Permutation Test

The null hypothesis is that the explanatory variables are unlinked with the response variable. One way to see how big a test statistic will be in a world where the null hypothesis holds true is to randomize the explanatory variables in the sample to destroy any relationship between them and the response variable. To illustrate how this can be done in a way that stays true to the sample, consider a small data set:

A	B	C
3	37.1	M
4	17.4	M
5	26.8	F
7	44.3	F
5	19.7	F

Imagine that the table has been cut into horizontal slips with one case on each slip. The response variable — say, A — has been written to the left of a dotted line. The explanatory variables B and C are on the right of the dotted line, like this:

A=3	...	B=37.1	C=M
A=4	...	B=17.4	C=M
A=5	...	B=26.8	C=F
A=7	...	B=44.3	C=F
A=5	...	B=19.7	C=F

To randomize the cases, tear each sheet along the dotted line. Place the right sides — the explanatory variables — on a table in their original order. Then, randomly shuffle the left halves — the response variable — and attach each to a right half.

A=4	⚡	B=37.1	C=M
A=5	⚡	B=17.4	C=M
A=5	⚡	B=26.8	C=F
A=3	⚡	B=44.3	C=F
A=7	⚡	B=19.7	C=F

None of the cases in the shuffle are genuine cases, except possibly by chance. Yet each of the shuffled explanatory variables is true to its distribution in the original sample and the relationships among explanatory variables — collinearity and multi-collinearity — are also authentic.

Each possible order for the left halves of the cards is called a **permutation**. A hypothesis test conducted in this way is called a **permutation test**.

For review purposes only

The logic of a permutation test is straightforward. To set up the test, you need to choose a test statistic that reflects some aspect of the system of interest to you.

Here are the steps involved in permutation test:

- Step 1 Calculate the value of the test statistic on the original data.
- Step 2 Permute the data and calculate the test statistic again. Repeat this many times, collecting the results. This gives the distribution of the test statistic under the null hypothesis.
- Step 3 Read off the p-value as the fraction of the results in (2) that are more extreme than the value in (1).

To illustrate, consider a model of heights from Galton's data and a question Galton didn't consider: Does the number of children in a family help explain the eventual adult height of the children? Perhaps in families with large numbers of children, there is competition over food, so children don't grow so well. Or, perhaps having a large number of children is a sign of economic success, and the children of successful families have more to eat.

The regression report indicates that larger family size is associated with shorter children:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	67.800	0.296	228.96	0.0000
nkids	<b>-0.169</b>	0.044	-3.83	0.0001

For every additional sibling, the family's children are shorter by about 0.17 inches on average. The confidence interval is  $-0.169 \pm 0.088$ .

Now for the permutation test, using the coefficient on nkids as the test statistic:

- Step 1 Calculate the test statistic on the data without any shuffling. As shown above, the coefficient on nkids is  $-0.169$ .
- Step 2 Permute and re-calculate the test statistic, many times. One set of 10000 trials gave values of  $-0.040$ ,  $0.037$ ,  $-0.094$ ,  $-0.069$ ,  $0.062$ ,  $-0.045$ , and so on. The distribution is shown in Figure 16.1.
- Step 3 The p-value is fraction of times that the values from (2) are more extreme than the value of  $-0.169$  from the unshuffled data. As Figure 16.1 shows, few of the permutations produced an nkid coefficient anywhere near  $-0.169$ . The p-value is very small,  $p < 0.001$ .

Conclusion, the number of kids in a family accounts for somewhat more of the children's heights than is likely to occur with a random explanatory variable.

The idea of a permutation test is almost a century old. It was proposed originally by the brilliant statistician Ronald Fisher (1890-1960). Permutation tests

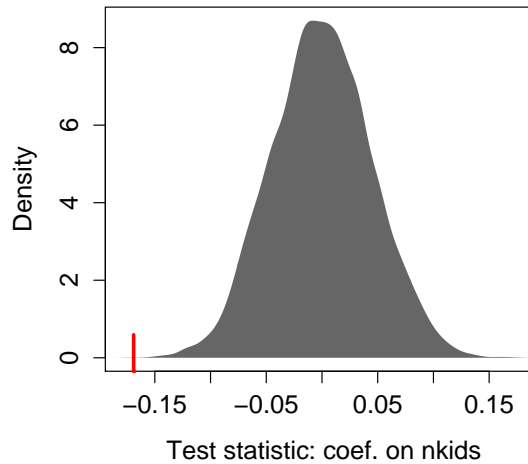


Figure 16.1: Distribution of the coefficient on nkids from the model  $\text{height} \sim 1 + \text{nkids}$  from many permutation trials. Tick mark: the coefficient  $-0.169$  from the unshuffled data.

were infeasible for even moderately sized data sets until the 1970s when inexpensive computation became a reality. In Fisher’s day, when computing was expensive, permutation tests were treated as a theoretical notion and actual calculations were performed using algebraic formulas. Such formulas could be derived for a narrow range of test statistics such as the sample mean, differences between group means, and the coefficient of determination  $R^2$ . Fisher himself derived the sampling distributions of these test statistics.[25]



Ronald Fisher

## 16.2 $R^2$ and the F Statistic

The coefficient of determination  $R^2$  measures what fraction of the variance of the response variable is “explained” or “accounted for” or — to put it simply — “modeled” by the explanatory variables.  $R^2$  is a comparison of two quantities: the variance of the fitted model values to the variance of the response variable.



$R^2$  is a single number that puts the explanation in the context of what remains unexplained. It's a good test statistic for a hypothesis test.

Using  $R^2$  as the test statistic in a permutation test would be simple enough. There are advantages, however, to thinking about things in terms of a closely related statistic invented by Fisher and named in honor of him: the **F statistic**.

Like  $R^2$ , the F statistic compares the size of the fitted model values to the size of the residuals. But the notion of "size" is somewhat different. Rather than measuring size directly by the variance or by the sum of squares, the F statistic takes into account the number of model vectors.

To see where F comes from, consider a special sort of random walk: the **random model walk**. In a regular random walk (Chapter 13), each new step is taken in a random direction. In a random model walk, each "step" consists of adding a new random explanatory term to a model. The "position" is measured as the  $R^2$  from the model.

The starting point of the random model walk is the simple model  $A \sim 1$  with just  $m = 1$  model vector. This model always produces  $R^2 = 0$  because the all-cases-the-same model can't account for any variance. Taking a "step" means adding a random model vector,  $x_1$ , giving the model  $A \sim 1 + x_1$ . Each new step adds a new random vector to the model:

$m$	Model
1	$A \sim 1$
2	$A \sim 1 + x_1$
3	$A \sim 1 + x_1 + x_2$
4	$A \sim 1 + x_1 + x_2 + x_3$
$\vdots$	
$n$	$A \sim 1 + x_1 + x_2 + x_3 + \cdots + x_{n-1}$

Figure 16.2 shows  $R^2$  versus  $m$  for several random model walks in data with  $n = 50$  cases. Each successive step adds in its own individual random explanatory term.

All the random walks start at  $R^2 = 0$  for  $m = 1$ . All of them reach  $R^2 = 1$  when  $m = n$ . Adding any more vectors beyond  $m = n$  simply creates redundancy;  $R^2 = 1$  is the best that can be done.

Notice that each step increases  $R^2$  — none of the random walks goes down in value as  $m$  gets bigger.

The  $R^2$  from a fitted model gives a single point on the model walk that divides the overall walk into two segments, as shown in Figure 16.3. The slope of each of the two segments has a straightforward interpretation. The slope of the segmented labeled "Model" describes the rate at which  $R^2$  is increased by a typical model vector. The slope can be calculated as  $R^2/(m - 1)$ .

The slope of the segment labeled "Residuals" describes how adding a random vector to the model would increase  $R^2$ . From the figure, you can see that a typical model vector increases  $R^2$  much faster than a typical random vector. Numerically, the slope is  $(1 - R^2)/(m - n)$ .

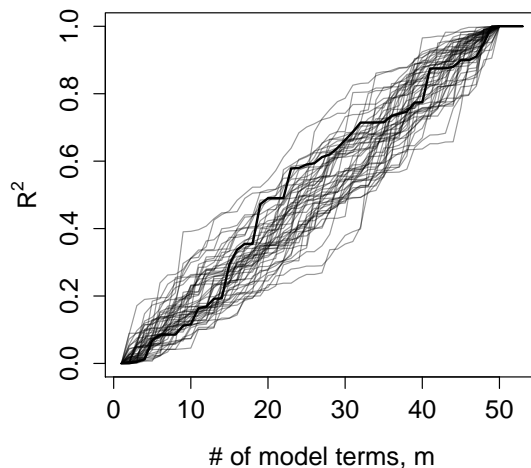


Figure 16.2: Random model walks for  $n = 50$  cases. The “position” of the walk is  $R^2$ . This is plotted versus number of model vectors  $m$ . The heavy line shows one simulation. The light lines show other simulations. All of the simulations reach  $R^2 = 1$  when  $m = n$ .

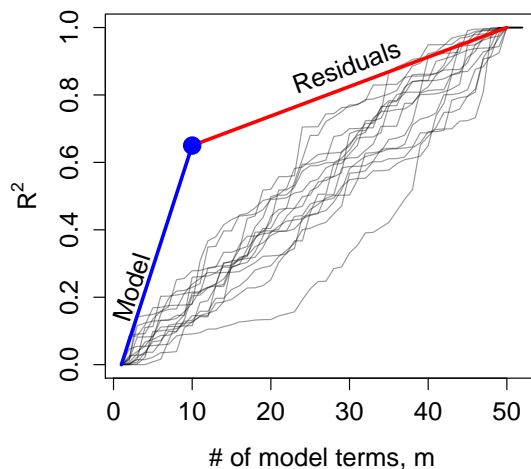


Figure 16.3: The  $R^2$  from a fitted model defines a model walk with two segments. The F statistic is the ratio of the slopes of these segments. Here a model with  $m = 10$  has produced  $R^2 = 0.65$  when fitted to a data set with  $n = 50$  cases.

# Chapter 17

## Hypothesis Testing on Parts of Models

*Everything should be made as simple as possible, but not simpler. — Albert Einstein*

It often happens in studying a system that a single explanatory variable is of direct interest to the modeler. Other explanatory variables may be important in the way the system works, but they are not of primary interest to the modeler. As in previous chapters, call the explanatory variable of interest simply the “explanatory variable.” The other explanatory variables, the ones that aren’t of direct interest, are the **covariates**. The covariates are ordinary variables. Their designation as covariates reflects the interests of the modeler rather than any intrinsic property of the variables themselves.

The source of the interest in the explanatory variable might be to discover whether it is relevant to a model of a response variable. For example, in studying people’s heights, one expects that several covariates, for instance the person’s sex or the height of their father or mother, play a role. It would be interesting to find out whether the number of siblings a person has, *nkids* in Galton’s dataset, is linked to the height. Studying *nkids* in isolation might be misleading since the other variables are reasonably regarded as influencing height. But doing a whole-model hypothesis test on a model that includes both *nkids* and the covariates would be uninformative: Of course one can account for a significant amount of the variation in height using sex and mother and father. (See Example 16.3 on page 297.) The question about *nkids* is whether it contributes something to the explanation that goes beyond the covariates. To answer this question, one needs a way of assessing the contribution of *nkids* on its own, but in the context set by the other variables.

Sometimes you may choose to focus on a single explanatory variable because a decision may hang on the role of that variable. Keep in mind that role of the explanatory variable of interest can depend on the context set by covariates.

For review purposes only

Those covariates might enhance, mask, or reverse the role of the variable of interest. For instance, it's legitimate that wages might vary according to the type of job and the worker's level of experience and education. It's also possible that wages might vary according to sex or race insofar as those variables happen to be correlated with the covariates, that is, correlated with the type of job or level of education and experience. But, if wages depend on the worker's sex or race *even taking into account* the legitimate factors, that's a different story and suggests a more sinister mechanism at work.

This chapter is about conducting hypothesis tests on a single explanatory variable in the context set by covariates. In talking about general principles, the text will refer to models of the form  $A \sim B+C$  or  $A \sim B+C+D$  where  $A$  is the response variable,  $B$  is the variable of interest, and  $C$  and  $D$  are the variables that you aren't directly interested in: the covariates.

Recall the definition of statistics offered in Chapter 1:

*Statistics is the explanation of variation in the context of what remains unexplained.*

It's important to include covariates in a hypothesis test because they influence both aspects of this definition:

**Explanation of variation** When the covariates are correlated with the explanatory variable, as they often are, including the covariates in a model will change the coefficients on the explanatory variable.

**What remains unexplained** Including covariates raises the  $R^2$  of a model. Or, to put it another way, including covariates reduces the size of the residuals. This can make the explanatory variable look better: smaller residuals generally mean smaller standard errors and higher F statistics. This is not an accounting trick, it genuinely reflects how much of the response variable remains unexplained.

## 17.1 The Term-by-Term ANOVA Table

The ANOVA table introduced in the Chapter 16 provides a framework for conducting hypothesis tests that include covariates. The whole-model ANOVA table lets you compare the "size" of the explained part of the model with the "size" of the residuals. Here is a whole-model report for the model  $\text{height} \sim \text{nkids} + \text{sex} + \text{father} + \text{mother}$  fit to Galton's data:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Model terms	4	7377.9	1845.0	401	0.0000
Residuals	893	4137.1	4.6		

The square length of the residual vector is 4137.1, the square length of the fitted model vector is 7377.9. The F test takes into account the number of vectors involved in each — four explanatory vectors (neglecting the intercept term), leaving 893 residual degrees of freedom. The F value, 401, is the ratio of the mean

DATA FILE  
galton.csv

For review purposes only

square of the model terms to the residuals. Since 401 is much, much larger than 1 (which is the expected value when the model terms are just random vectors), it's appropriate to reject the null hypothesis; the p-value is effectively zero since F is so huge.

A term-by-term ANOVA report is much the same, but the report doesn't just partition variation into "modeled" and "residual," it goes further by partitioning the modeled variation among the different explanatory terms. There is one row in the report for each individual model term.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
nkids	1	185.5	185.5	40.0	0.0000
sex	1	5766.3	5766.3	1244.7	0.0000
father	1	937.6	937.6	202.4	0.0000
mother	1	488.5	488.5	105.5	0.0000
Residuals	893	4137.1	4.6		

Notice that the "Residuals" row is exactly the same in the two ANOVA reports. This is because that row just describes the square length of the residuals — it's the same model with the same residuals being analyzed in the two ANOVA reports. Less obvious, perhaps, is that the sum of squares of the individual model terms adds up across all the terms to give exactly the sum of squares from the whole-model ANOVA report:  $185.5 + 5766.3 + 937.6 + 488.5 = 7377.9$ . The same is true for the degrees of freedom of the model terms:  $1 + 1 + 1 + 1 = 4$ .

What's new in the term-by-term ANOVA report is that there is a separate mean square, F value, and p-value for each model term. These are calculated in the familiar way: the mean square for any term is just the sum of squares for that term divided by the degrees of freedom for that term. Similarly, the F value for each term is the mean square for that term divided by the mean square of the residuals. For instance, for nkids, the F value is  $185.5/4.6 = 40$ . This is translated to a p-value in exactly the same way as in Chapter 16 (using an F distribution with 1 degree of freedom in the numerator and 893 in the denominator — values that come from the corresponding rows of the ANOVA report).

## 17.2 Covariates Soak Up Variance

An important role of covariates is to account for variance in the response. This reduces the size of residuals. Effectively, the covariates reduce the amount of randomness attributed to the system and thereby make it easier to see patterns in the data. To illustrate, consider a simple question relating to marriage: Do men tend to be older than women when they get married?

The marriage-license dataset contains information on the ages of brides and grooms at the time of their marriage as well as other data on education, race, number of previous marriages, and so on. These data were collected from the on-line marriage license records made available by Mobile County, Alabama in the US. Here's an excerpt:

For review purposes only



# Chapter 18

## Models of Yes/No Variables

Sometimes the variation that you want to model occurs in a categorical variable. For instance, you might want to know what dietary factors are related to whether a person develops cancer. Or perhaps how income, personality, and education influence what kind of job a person takes.

The techniques studied up until now are intended to model *quantitative* response variables, those where the value is a number, for example height, wage, swimming times, or foot size. Categorical variables have been used only in the role of explanatory variables.

There are good technical reasons why it's easier to build and interpret models with quantitative response variables. For one, a residual from such a model is simply a number. The size of a typical residual can be described using a variance or a mean square. In contrast, consider a model that tries to predict whether a person will become an businessman or an engineer or a farmer or a lawyer. There are many different ways for the model to be wrong, e.g., it predicts farmer when the outcome is lawyer, or engineer when the outcome is businessman. The "residual" is not a number, so how do you measure how large it is?

This chapter introduces one technique for building and evaluating models where the response variable is categorical. The technique handles a special case, where the categorical variable has only two levels. Generically, these can be called **yes/no** variables, but the two levels could be anything you like: alive/dead, success/failure, engineer/not-engineer, and so on.

### 18.1 The 0-1 Encoding

The wonderful thing about yes/no variables is that they are always effectively quantitative; they can be naturally encoded as 0 for no and 1 for yes. Given this encoding, you could if you want just use the standard linear modeling approach of quantitative response variables. That's what I'll do at first, showing what's good and what's bad about this approach before moving on to a better method.

To illustrate using the linear modeling approach, consider some data that

For review purposes only

DATA FILE  
whickham.csv

relate smoking and mortality. The table below gives a few of the cases from a data frame where women were interviewed about whether they smoked.

Case	Outcome	Smoker	Age
1	Alive	Yes	23
2	Alive	Yes	18
3	Dead	Yes	71
4	Alive	No	67
5	Alive	No	64
6	Alive	Yes	38

... and so on for 1314 cases altogether.

Of course, all the women were alive when interviewed! The outcome variable records whether each woman was still alive 20 years later, during a follow-up study. For instance, case 3 was 71 years old at the time of the interview. Twenty years later, she was no longer alive.

Outcome is the Yes/No variable I'm interested in understanding; it will be the response variable and smoker and age will be the explanatory variables. I'll encode "Alive" as 1 and "Dead" as 0, although it would work just as well to do things the other way around.

The simplest model of the outcome is all-cases-the-same:  $\text{outcome} \sim 1$ . Here is the regression report from this model:

	Estimate	Std. Error	t value	Pr(> t )
Intercept	0.7192	0.0124	57.99	0.0000

The intercept coefficient in the model  $\text{outcome} \sim 1$  has a simple interpretation; it's the mean of the response variable. Because of the 0-1 coding, the mean is just the fraction of cases where the person was alive. That is, the coefficient from this model is just the probability that a random case drawn from the sample has an outcome of "Alive."

At first that might seem a waste of modeling software, since you could get exactly the same result by counting the number of "Alive" cases. But notice that there is a bonus to the modeling approach — there is a standard error given that tells how precise is the estimate:  $0.719 \pm 0.024$  with 95% confidence.

The linear modeling approach also works sensibly with an explanatory variable. Here's the regression report on the model  $\text{outcome} \sim \text{smoker}$ :

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.6858	0.0166	41.40	0.0000
smokerYes	0.0754	0.0249	3.03	0.0025

You can interpret the coefficients from this model in a simple way — the intercept is group-wise means of outcome for the non-smokers. The other coefficient gives the difference in the means between the smokers and the non-smokers. Since outcome is a 0-1 variable encoding whether the person was alive,

For review purposes only

the coefficients indicate the proportion of people in each group who were alive at the time of the follow-up study. In other words, it's reasonable to interpret the intercept as the probability that a non-smokers was alive, and the other coefficient as the difference in probability of being alive between the non-smokers and smokers.

Again, you could have found this result in a simpler way: just count the fraction of smokers and of non-smokers who were alive.

The advantage of the modeling approach is that it produces a standard error for each coefficient and a p-value. The intercept says that  $0.686 \pm 0.033$  of non-smokers were alive. The `smokerYes` coefficient says that an additional  $0.075 \pm 0.05$  of smokers were alive.

This result might be surprising, since most people expect that mortality is higher among smokers than non-smokers. But the confidence interval does not include 0 or any negative number. Correspondingly, the p-value in the report is small: the null hypothesis that smoker is unrelated to outcome can be rejected.

Perhaps it's obvious that a proper model of mortality and smoking should adjust for the age of the person. After all, age is strongly related to mortality and smoking might be related to age. Here's the regression report from the model `outcome ~ smoker+age`:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.4726	0.0301	48.92	0.0000
smokerYes	0.0105	0.0196	0.54	0.5927
age	<b>-0.0162</b>	0.0006	-28.95	0.0000

It seems that the inclusion of `age` has had the anticipated effect. According to the coefficient  $-0.016$ , there is a negative relationship between age and being alive. The effect of `smoker` has been greatly reduced; the p-value 0.59 is much too large to reject the null hypothesis. These data in fact give little or no evidence that smoking is associated with mortality. (It's well established that smoking increases mortality, but the number of people involved in this study, combined with the relatively young ages of the smokers, means that the power of the study was small.)

There is also a mathematical issue. Consider the fitted model value for a 20-year old smoker:  $1.47 + 0.010 - 0.016 \times 20 = 1.16$ . This value, 1.16, can't be interpreted as a probability that a 20-year old smoker would still be alive at the time of the follow-up study (when she would have been 40). Probabilities must always be between zero and one.

The top panel of Figure 18.1 shows the fitted model values for the linear model along with the actual data. (Since the outcome variable is coded as 0 or 1, it's been jittered slightly up and down so that the density of the individual cases shows up better. Smokers are at the bottom of each band, plotted as small triangles.) The data show clearly that older people are more likely to have died. Less obvious in the figure is that the very old people tended not to smoke.



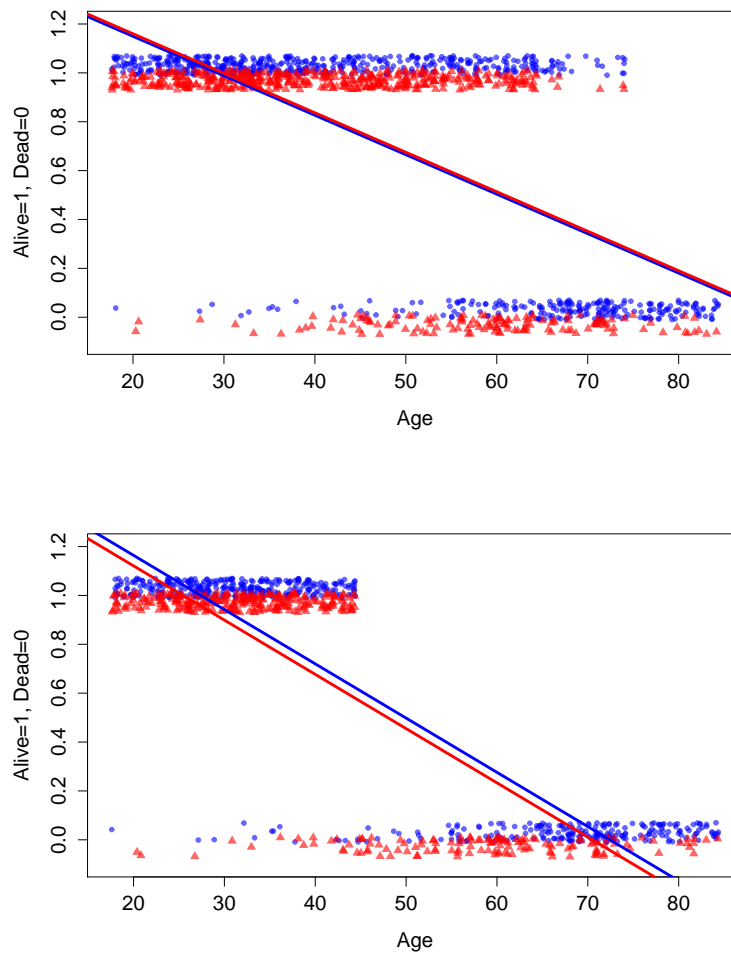


Figure 18.1: Top: The smoking/mortality data along with a linear model. Bottom: A thought experiment which eliminates the “Alive” cases above age 45.

For review purposes only

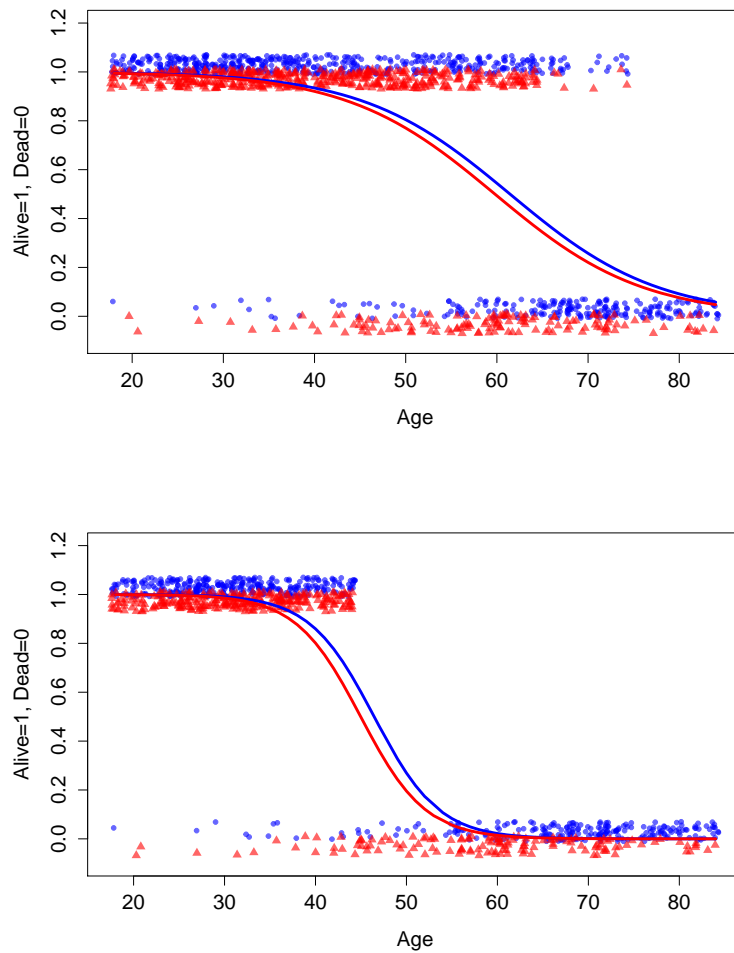


Figure 18.2: Logistic forms for outcome  $\sim$  age + smoker. Top: The smoking/mortality data. Bottom: The thought experiment in which the “Alive” cases above age 45 were eliminated.

For review purposes only

The problem with the linear model is that it is too rigid, too straight. In order for the model to reflect that the large majority of people in their 40s and 50s were alive, the model overstates survival in the 20-year olds. This isn't actually a problem for the 20-year olds — the model values clearly indicate that they are very likely to be alive. But what isn't known is the extent to which the line has been pulled down to come close to 1 for the 20-year olds, thereby lowering the model values for the middle-aged folks.

To illustrate how misleading the straight-line model can be, I'll conduct a small thought experiment and delete all the "alive" cases above the age of 45. The resulting data and the best-fitting linear model are shown in the bottom panel of Figure 18.1.

The thought-experiment model is completely wrong for people in their 50s. Even though there are no cases whatsoever where such people are alive, the model value is around 50%.

What's needed is a more flexible model — something not quite so rigid as a line, something that can bend to stay in the 0-to-1 bounds of legitimate probabilities. There are many ways that this could be accomplished. The one that is most widely used is exceptionally effective, clever, and perhaps unexpected. It consists of a two-stage process:

1. Construct a model value in the standard linear way — multiplication of coefficients times model vectors. The model outcome  $\sim \text{age} + \text{smoker}$  would involve the familiar terms: an intercept, a vector for age, and the indicator vector for smokerYes. The output of the linear formula — I'll call it  $Y$  — is not a probability but rather a **link value**: an intermediary of the fitting process.
2. Rather than taking the link value  $Y$  as the model value, transform  $Y$  into another quantity  $P$  that it is bounded by 0 and 1. A number of different transformations could do the job, but the most widely used one is called the **logistic transformation**:  $P = \frac{e^Y}{1+e^Y}$ .

This type of model is called a **logistic regression** model. The choice of the best coefficients — the model fit — is based not directly on how well the link values  $Y$  match the data but on how the probability values  $P$  match.

Figure 18.2 illustrates the logistic form on the smoking/mortality data. The logistic transformation lets the model fit the outcomes for the very young and the very old, while maintaining flexibility in the middle to match the data there.

The two-stage approach to logistic regression makes it straightforward to add explanatory terms in a model. Whatever terms there are in a model — main terms, interaction terms, nonlinear terms — the logistic transformation guarantees that the model values  $P$  will fall nicely in the 0-to-1 scale.

Figure 18.3 shows how the link value  $Y$  is translated into the 0-to-1 scale of  $P$  for different "shapes" of models.

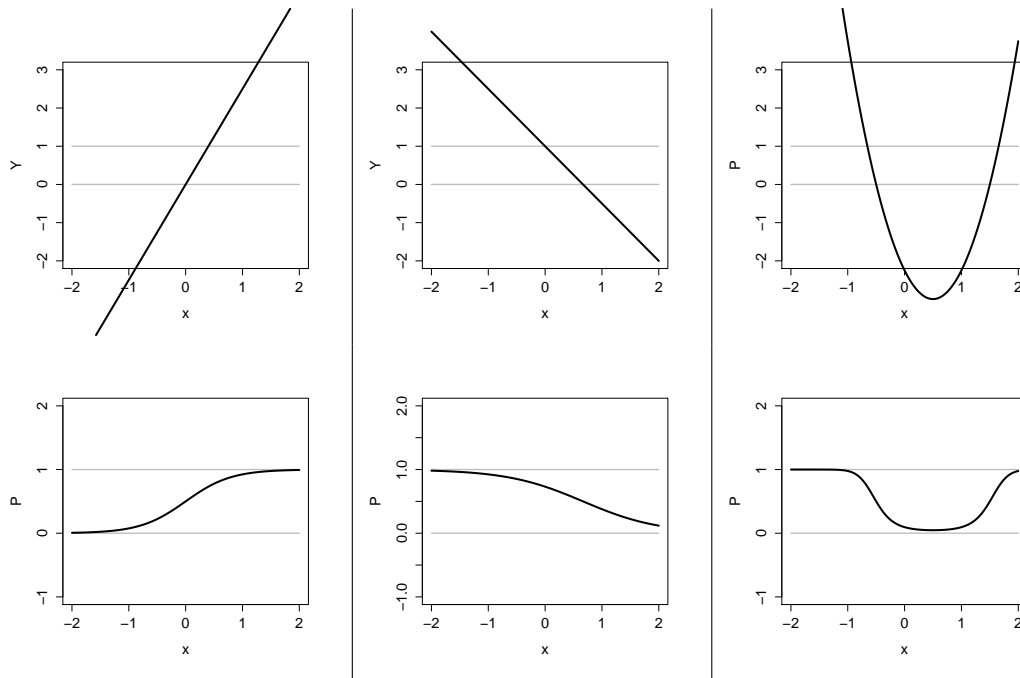


Figure 18.3: Comparing the link values  $Y$  with the probability values  $P$  for three different models, each with an explanatory variable  $x$ . Top: Link values. Bottom: Probability values.

## 18.2 Inference on Logistic Models

The interpretation of logistic models follows the same logic as in linear models. Each case has a model probability value. For example, the model probability value for a 64-year old non-smoker is 0.4215, while for a 64-year old smoker it is 0.3726. These model values are in the form of probabilities; the 64-year old non-smoker had a 42% chance of being alive at the time of the follow-up interview.

There is a regression report for logistic models. Here it is for outcome  $\sim$  age+smoker:

	Estimate	Std. Error	z value	Pr(> z )
Intercept	<b>7.5992</b>	0.4412	17.22	0.0000
age	<b>-0.1237</b>	0.0072	-17.23	0.0000
smokerYes	-0.2047	0.1684	-1.22	0.2242

The coefficients refer to the link values  $Y$ , so for the 64-year old non-smoker the value is  $Y = 7.59922 - 0.12368 \times 64 = -0.3163$ . This value is not a probability; it's negative, so how could it be? The model value is the output of applying the logistic transform to  $Y$ , that is,  $\frac{e^{-0.3163}}{1 + e^{-0.3163}} = 0.4215$ .

The logistic regression report includes standard errors and p-values, just as in the ordinary linear regression report. Even if there were no regression report, you could generate the information by resampling, using bootstrapping to

find confidence intervals and doing hypothesis tests by permutation tests or by resampling the explanatory variable.

In the regression report for the model above, you can see that the null hypothesis that age is unrelated to outcome can be rejected. (That's no surprise, given the obvious tie between age and mortality.) On the other hand, these data provide only very weak evidence that there is a difference between smokers and non-smokers; the p-value is 0.22.

An ANOVA-type report can be made for logistic regression. This involves the same sort of reasoning as in linear regression: fitting sequences of nested models and examining how the residuals are reduced as more explanatory terms are added to the model.

It's time to discuss the residuals from logistic regression. Obviously, since there are response values (the 0-1 encoding of the response variable) and there are model values, there must be residuals. The obvious way to define the residuals is as the difference between the response values and the model values, just as is done for linear models. If this were done, the model fit could be chosen as that which minimizes the sum of square residuals.

Although it would be possible to fit a logistic-transform model in this way, it is not the accepted technique. To see the problem with the sum-of-square-residuals approach, think about the process of comparing two candidate models. The table below gives a sketch of two models to be compared during a fitting process. The first model has one set of model values, and the second model has another set of model values.

Case	Model Values		Observed Value
	First Model	Second Model	
1	0.8	1.0	1
2	0.1	0.0	1
3	0.9	1.0	1
4	0.5	0.0	0
Sum. Sq. Resid.	1.11	1.00	

Which model is better? The sum of square residuals for the first model is  $(1 - 0.8)^2 + (1 - 0.1)^2 + (1 - 0.9)^2 + (0 - 0.5)^2$ , which works out to be 1.11. For the second model, the sum of square residuals is  $(1 - 1.0)^2 + (1 - 0.0)^2 + (1 - 1.0)^2 + (0 - 0.0)^2$ , or 1.00. Conclusion: the second model is to be preferred.

However, despite the sum of squares, there is a good reason to prefer the first model. For Case 2, the Second Model gives a model value of 0 — this says that it's *impossible* to have an outcome of 1. But, in fact, the observed value was 1; according to the Second Model the impossible has occurred! The First Model has a model value of 0.1 for Case 2, suggesting it is merely *unlikely* that the outcome could be 1.

The actual criterion used to fit logistic models penalizes heavily — infinitely in fact — candidate models that model as impossible events that can happen. The criterion for fitting is called the **likelihood** and is defined to be the probability of the observed values according to the probability model of the candidate.

When the observed value is 1, the likelihood of the single observation is just the model probability. When the observed value is 0, the likelihood is 1 minus the model probability. The overall likelihood is the product of the individual likelihoods of each case. So, according to the First Model, the likelihood of the observations 1, 1, 1, 0 is  $0.8 \times 0.1 \times 0.9 \times (1 - 0.5) = 0.036$ . According to the Second Model, the likelihood is  $1 \times 0 \times 1 \times (1 - 0) = 0$ . The First Model wins.

It might seem that the likelihood criterion used in logistic regression is completely different from the sum-of-square-residuals criterion used in linear regression. It turns out that they both fit into a common framework called **maximum likelihood estimation**. A key insight is that one can define a probability model for describing residuals and, from this, compute the likelihood of the observed data given the model.

In the maximum likelihood framework, the equivalent of the sum of squares of the residuals is a quantity called the **deviance**. Just as the linear regression report gives the sum of squares of the residual, the logistic regression report gives the deviance. To illustrate, here is the deviance part of the logistic regression report for the `outcome ~ age+smoker` model:

```
Null deviance: 1560.32  on 1313  degrees of freedom
Residual deviance:  945.02  on 1311  degrees of freedom
```

The **null deviance** refers to the simple model `outcome ~ 1`: it's analogous to the sum of squares of the residuals from that simple model. The reported degrees of freedom, 1313, is the sample size  $n$  minus the number of coefficients  $m$  in the model. That's  $m = 1$  because the model `outcome ~ 1` has a single coefficient. For the smoker/mortality data in the example, the sample size is  $n = 1314$ . The line labelled "Residual deviance" reports the deviance from the full model: `outcome ~ age+smoker`. The full model has three coefficients altogether: the intercept, `age`, and `smokerYes`, leaving  $1314 - 3 = 1311$  degrees of freedom in the deviance.

According to the report, the `age` and `smokerYes` vectors reduced the deviance from 1560.32 to 945.02. The deviance is constructed in a way so that a random, junky explanatory vector would, on average, consume a proportion of the deviance equal to 1 over the degrees of freedom. Thus, if you constructed the `outcome ~ age+smoker + junk`, where `junk` is a random, junky term, you would expect the deviance to be reduced by a fraction  $1/1311$ .

To perform a hypothesis test, compare the actual amount by which a model term reduces the deviance to the amount expected for random terms. This is analogous to the F test, but involves a different probability distribution called the  $\chi^2$  (chi-squared). The end result, as with the F test, is a p-value which can be interpreted in the conventional way, just as you do in linear regression.

## 18.3 Model Probabilities

The link values  $Y$  in a logistic regression are ordinary numbers that can range from  $-\infty$  to  $\infty$ . The logistic transform converts  $Y$  to numbers  $P$  on the scale 0 to 1, which can be interpreted as probabilities. I'll call the values  $P$  **model**

**probabilities** to distinguish them from the sorts of model values found in ordinary linear models. The model probabilities describe the chance that the Yes/No response variable takes on the level “Yes.”

To be more precise,  $P$  is a **conditional probability**. That is, it describes the probability of a “Yes” outcome conditioned on the explanatory variables in the model. In other words, the model probabilities are probabilities *given* the value of the explanatory variables. For example, in the model  $\text{outcome} \sim \text{age} + \text{smoker}$ , the link value for a 64-year old non-smoker is  $Y = -0.3163$  corresponding to a model probability  $P = 0.4215$ . According to the model, this is the probability that a person who was 64 and a non-smoker was still alive at the time of the follow-up interview. That is to say, the probability is 0.4215 of being alive at the follow-up study conditioned on being age 64 and non-smoking at the time of the original interview. Change the values of the explanatory variables — look at a 65-year old smoker, for instance — and the model probability changes.

There is another set of factors that conditions the model probabilities: the situation that applied to the selection of cases for the data frame. For instance, fitting the model  $\text{outcome} \sim 1$  to the smoking/mortality data gives a model probability for “Alive” of  $P = 0.672$ . It would not be fair to say that this is the probability that a person will still be alive at the time of the follow-up interview. Instead, it is the probability that a person will still be alive *given* that they were in the sample of data found in the data frame. Only if that sample is representative of a broader population is it fair to treat that probability as applying to that population. If the overall population doesn’t match the sample, the probability from the model fitted to the sample won’t necessarily match the probability of “Alive” in the overall population.

Often, the interest is to apply the results of the model to a broad population that is not similar to the sample. This can sometimes be done, but care must be taken.

To illustrate, consider a study done by researchers in Queensland, Australia on the possible link between cancer of the prostate gland and diet. Pan-fried and grilled meats are a source of carcinogenic compounds such as heterocyclic amines and polycyclic aromatic hydrocarbons. Some studies have found a link between eating meats cooked “well-done” and prostate cancer.[31] The Queensland researchers[32] interviewed more than 300 men with prostate cancer to find out how much meat of various types they eat and how they typically have their meat cooked. They also interviewed about 200 men without prostate cancer to serve as controls. They modeled whether or not each man has prostate cancer (variable `pcancer`) using both age and intensity of meat consumption as the explanatory variables. Then, to quantify the effect of meat consumption, they compared high-intensity eaters to low-intensity eaters; the model probability at the 10th percentile of intensity compared to the model probability at the 90th percentile. For example, for a 60-year old man, the model probabilities are 69.8% for low-meat intensity eaters, and 82.1% for high-meat intensity eaters.

It would be a mistake to interpret these numbers as the probabilities that a 60-year old man will have prostate cancer. The prevalence of prostate cancer

is much lower. (According to one source, the prevalence of clinically evident prostate cancer is less than 10% over a lifetime, so many fewer than 10% of 60-year old men have clinically evident prostate cancer.[33])

The sample was collected in a way that intentionally overstates the prevalence of prostate cancer. More than half of the sample was selected specifically because the person had prostate cancer. Given this, the sample is not representative of the overall population. In order to make a sample that matches the overall population, there would have had to be many more non-cancer controls in the sample.

However, there's reason to believe that the sample reflects in a reasonably accurate way the diet practices of those men with prostate cancer and also of those men without prostate cancer. As such, the sample can give information that might be relevant to the overall population, namely how meat consumption could be linked to prostate cancer. For example, the model indicates that a high-intensity eater has a higher chance of having prostate cancer than a low-intensity eater. Comparing these probabilities — 82.1% and 69.8% — might give some insight.

There is an effective way to compare the probabilities fitted to the sample so that the results can be generalized to apply to the overall population. It will work so long as the different groups in the sample — here, the men with prostate cancer and the non-cancer controls — are each representative of that same group in the overall population.

It's tempting to compare the probabilities directly as a ratio to say how much high-intensity eating increases the risk of prostate cancer compared to low-intensity eating. This ratio of model probabilities is  $0.821/0.698$ , about 1.19. The interpretation of this is that high-intensity eating increases the risk of cancer by 19%.

Unfortunately, because the presence of prostate cancer in the sample doesn't match the prevalence in the overall population, the ratio of model probabilities usually will not match that which would be found from a sample that does represent the overall population.

Rather than taking the ratio of model probabilities, it turns out to be better to consider the **odds ratio**. Even though the sample doesn't match the overall population, the odds ratio will (so long as the two groups individually accurately reflect the population in each of those groups).

**Odds** are just another way of talking about probability  $P$ . Rather than being a number between 0 and 1, an odds is a number between 0 and  $\infty$ . The odds is calculated as  $P/(1 - P)$ . For example, the odds that corresponds to a probability  $P = 50\%$  is  $0.50/(1 - 0.50) = 1$ . Everyday language works well here: the odds are fifty-fifty.

The odds corresponding to the low-intensity eater's probability of  $P = 69.8\%$  is  $0.698/0.302 = 2.31$ . Similarly, the odds corresponding to high-intensity eater's  $P = 82.1\%$  is  $0.821/0.179 = 4.60$ . The odds ratio compares the two odds:  $4.60/2.31 = 1.99$ , about two-to-one.

Why use a ratio of odds rather than a simple ratio of probabilities? Because the sample was constructed in a way that doesn't accurately represent the prevalence of cancer in the population. The ratio of probabilities would reflect the



prevalence of cancer in the sample: an artifact of the way the sample was collected. The odds ratio compares each group — cancer and non-cancer — to itself and doesn't depend on the prevalence in the sample.

**Example 18.1: Log-odds ratios of Prostate Cancer** One of the nice features of the logistic transformation is that the link values  $Y$  can be used directly to read off the logarithm of odds ratios.

In the prostate-cancer data, the coefficients on the model  $\text{pcancer} \sim \text{age} + \text{intensity}$  are these:

	Estimate	Std. Error	z value	Pr(> z )
Intercept	4.274	0.821	5.203	0.0000
Age	-0.057	0.012	-4.894	0.0000
Intensity	<b>0.172</b>	0.058	2.961	0.0031

These coefficients can be used to calculate the link value  $Y$  which corresponds to a log odds. For example, in comparing men at 10th percentile of intensity to those at the 90th percentile, you multiply the intensity coefficient by the difference in intensities. The bottom 10th percentile is 0 intensity and the top 10th percentile is an intensity of 4. So the difference in  $Y$  score for the two percentiles is  $0.1722 \times (4 - 0) = 0.6888$ . This value, 0.6888, is the log odds ratio. To translate this to an odds ratio, you need to undo the log. That is, calculate  $e^{0.6888} = 1.99$ . So, the odds ratio for the risk of prostate cancer in high-intensity eaters versus low-intensity eaters is approximately 2.

Note that the model coefficient on age is negative. This suggests that the risk of prostate cancer *goes down* as people age. This is wrong biologically as is known from other epidemiological work on prostate cancer, but it does reflect that the sample was constructed rather than randomly collected. Perhaps it's better to say that it reflects a weakness in the way the sample was constructed: the prostate cancer cases tend to be younger than the non-cancer controls. If greater care had been used in selecting the control cases, they would have matched exactly the age distribution in the cancer cases. Including age in the model is an attempt to adjust for the problems in the sample — the idea is that including age allows the model's dependence on intensity to be treated *as if* age were held constant.

## 18.4 Computational Technique

Fitting logistic models uses many of the same ideas as in linear models.

### 18.4.1 Fitting Logistic Models

The `glm` operator fits logistic models. (It also fits other kinds of models, but that's another story.) `glm` takes model design and data arguments that are iden-

tical to their counterparts in `lm`. Here's an example using the smoking/mortality data:

DATA FILE  
whickham.csv

```
> whickham = ISMdata("whickham.csv")
> mod = glm( outcome ~ age + smoker, data=whickham,
             family="binomial")
```

The last argument, `family="binomial"`, simply specifies to `glm` that the logistic transformation should be used. (`glm` is short for Generalized Linear Modeling, a broad label that covers logistic regression as well as other types of models involving links and transformations.)

The regression report is produced with the summary operator, which recognizes that the model was fit logistically and does the right thing:

```
> summary(mod)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.599221    0.441231 -17.223  <2e-16
age          0.123683    0.007177  17.233  <2e-16
smokerYes    0.204699    0.168422   1.215    0.224
```

```
Null deviance: 1560.32 on 1313 degrees of freedom
Residual deviance: 945.02 on 1311 degrees of freedom
```

Keep in mind that the coefficients refer to the intermediate model values  $Y$ . The probability  $P$  will be  $e^Y / (1 + e^Y)$ .

In fitting a logistic model, it's crucial that the response variable be categorical, with two levels. It happens that in the `whickham` data, the outcome variable fits the bill: the levels are `Alive` and `Dead`.

The `glm` software will automatically recode the response variable as 0/1. The question is, which level gets mapped to 1? In some sense, it makes no difference since there are only two levels. But if you're talking about the probability of dying, it's nice not to mistake that for the probability of staying alive. So make sure that you know which level in the response variable corresponds to 1: it's the second level.

Here is an easy way to make sure which level has been coded as "Yes". First, fit the all-cases-the-same model, `outcome ~ 1`. The fitted model value  $P$  from this model will be the proportion of cases for which the outcome was "Yes."

```
> mod2 = glm( outcome ~ 1, data=whickham, family="binomial")
> summary(mod)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.94039    0.06139  -15.32  <2e-16
> exp(-.94039)/(1+exp(-.94039))
[1] 0.280822
```

So, 28% of the cases were "Yes." But which of the two levels is "Yes?" Find out just by counting and taking a proportion:

For review purposes only

```
> prop.table(table(whickham$outcome))
      Alive      Dead
0.7191781 0.2808219
```

Evidently, by default, “Yes” means Dead.

If you want to dictate which of the two levels is going to be encoded as 1, you can use a comparison operation to do so:

```
> mod3 = glm( outcome=="Alive" ~ 1, data=whickham,
             family="binomial")
```

In this model, “Yes” means Alive.

### 18.4.2 Fitted Model Values

Logistic regression involves two different kinds of fitted values: the intermediate “link” value  $Y$  and the probability  $P$ . The fitted operator returns the probabilities:

```
> probs = fitted(mod)
      1      2      3      4
0.010458680 0.005662422 0.800110184 0.665421999
      5      6      7      8
0.578471493 0.063295027 0.138383748 0.858234119
... and so on ...
    1309    1310    1311    1312
0.008539232 0.044548261 0.028813443 0.008185469
    1313    1314
0.129003853 0.089194331
```

There is one fitted probability value for each case.

The link values can be gotten via the predict operator

```
> predict(mod, type="link")
      1      2      3      4
-4.54980922 -5.16822509 1.38698315 0.68755138
      5      6      7      8
0.31650186 -2.69456160 -1.82877938 1.80069995
... and so on.
```

Notice that the link values are not necessarily between zero and one.

The predict operator can also be used to calculate the probability values.

```
> predict(mod, type="response")
      1      2      3      4
0.010458680 0.005662422 0.800110184 0.665421999
      5      6      7      8
0.578471493 0.063295027 0.138383748 0.858234119
... and so on.
```

This is particularly useful when you want to use predict to find the model values for inputs other than that original data frame used for fitting.

For review purposes only

### 18.4.3 Analysis of Variance

The same basic logic used in analysis of variance applies to logistic regression, although the quantity being broken down into parts is not the sum of squares of the residuals but, rather, the deviance.

The anova software will take apart a logistic model, term by term, using the order specified in the model.

```
> anova(mod, test="Chisq")
```

Analysis of Deviance Table

	Df	Deviance	Resid. Df	Resid. Dev	P(> Chi )
NULL			1313	1560.32	
age	1	613.81	1312	946.51	1.659e-135
smoker	1	1.49	1311	945.02	0.22

Notice the second argument, `test="Chisq"`, which instructs `anova` to calculate a p-value for each term. This involves a slightly different test than the F test used in linear-model ANOVA.

The format of the ANOVA table for logistic regression is somewhat different from that used in linear models, but the concepts of degrees of freedom and partitioning still apply. The basic idea is to ask whether the reduction in deviance accomplished with the model terms is greater than what would be expected if random terms were used instead.

For review purposes only

For review purposes only

# Chapter 19

## Causation

*If the issues at hand involve responsibilities or decisions or plans, causal reasoning is necessary.* — Edward Tufte

*Knowing what causes what makes a big difference in how we act. If the rooster's crow causes the sun to rise we could make the night shorter by waking up our rooster earlier and make him crow - say by telling him the latest rooster joke.* — Judea Pearl

Starting in the 1860s, Europeans expanded rapidly westward in the United States, settling grasslands in the Great Plains. Early migrants had passed through these semi-arid plains, called the Great American Desert, on the way to habitable territories further west. But unusually heavy rainfall in the 1860s and 1870s supported new migrants who homesteaded on the plains rather than passing through.

The homesteaders were encouraged by a theory that the act of farming would increase rainfall. In the phrase of the day, "Rain follows the plow."

The theory that farming leads to rainfall was supported by evidence. As farming spread, measurements of rainfall were increasing. Buffalo grass, a species well adapted to dry conditions, was retreating. Other grasses more dependent on moisture were advancing. The exact mechanism of the increasing rainfall was uncertain, but many explanations were available. In a phrase-making book published in 1881, Charles Dana Wilber wrote,

*Suppose now that a new army of frontier farmers ... could, acting in concert, turn over the prairie sod, and after deep plowing and receiving the rain and moisture, present a new surface of green, growing crops instead of the dry, hard-baked earth covered with sparse buffalo grass. No one can question or doubt the inevitable effect of this cool condensing surface upon the moisture in the atmosphere as it moves over by the Western winds. A reduction in temperature must at once occur, accompanied by the usual phenomena of showers. The chief agency in this transformation is agriculture. To be more concise. Rain follows the plow. [34, p. 68]*

For review purposes only

Seen in the light of subsequent developments, this theory seems hollow. Many homesteaders were wiped out by drought. The most famous of these, the Dust Bowl of the 1930s, rendered huge areas of US and Canadian prairie useless for agriculture and led to the displacement of hundreds of thousands of families.

Wilber was correct in seeing the association between rainfall and farming, but wrong in his interpretation of the causal connection between them. With a modern perspective, you can see clearly that Wilber got it backwards; there was indeed an association between farming and rainfall, but it was the increased rainfall of the 1870s that led to the growth of farming. When the rains failed, so did the farms.

The subject of this chapter is the ways in which data and statistical models can and cannot appropriately be used to support claims of causation. When are you entitled to interpret a model as signifying a causal relationship? How can you collect and process data so that such an interpretation is justified? How can you decide which covariates to include or exclude in order to reveal causal links?

The answers to these questions are subtle. Model results can be interpreted only in the context of the researcher's beliefs and prior knowledge about how the system operates. And there are advantages when the researcher becomes a participant, not just collecting observations but actively intervening in the system under study as an experimentalist.

## 19.1 Interpreting Models Causally

Interpreting statistical models in terms of causation is done for a purpose. It is well to keep that purpose in mind so you can apply appropriate standards of evidence. When causation is an issue, typically you have in mind some intervention that you are considering performing. You want to use your models to estimate what will be the effect of that intervention.

For example, suppose you are a government health official considering the approval of flecainide, a drug intended for the treatment of overly fast heart rhythms such as atrial fibrillation. Your interest is improving patient outcomes, perhaps as measured by survival time. The purpose of your statistical models is to determine whether prescribing flecainide to patients is likely to lead to improved outcomes and how much improvement will typically be achieved.

Or suppose you are the principal of a new school. You have to decide how much to pay teachers but you have to stay within your budget. You have three options: pay relatively high salaries but make classes large, pay standard salaries and make classes the standard size, pay low salaries and make classes small. Your interest is in the effective education of your pupils, perhaps as measured by standardized test scores. The purpose of your statistical models is to determine whether the salary/class-size options will differ in their effects and by how much.

Or suppose you are a judge hearing a case involving a worker's claim of sex discrimination against her employer. You need to decide whether to find for the

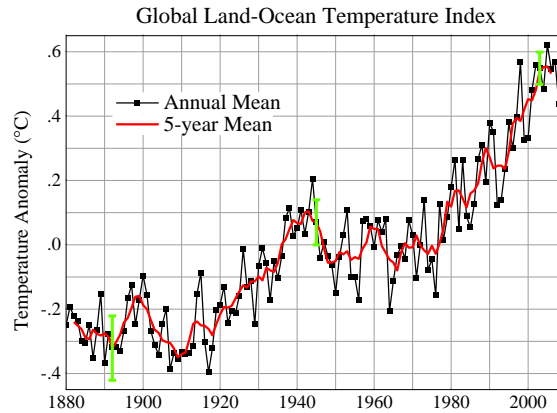


Figure 19.1: Global temperature since 1880 as reported by NASA. <http://data.giss.nasa.gov/gistemp/graphs/> accessed on July 7, 2009.

worker or the employer. This situation is somewhat different. In the education or drug examples, you planned to take action to change the variable of interest — reduce class sizes or give patients a drug — in order to produce a better outcome. But you can't change the worker's sex to avoid the discrimination. Even if you could, what's past is past. In this situation, you are dealing with a **counterfactual**: you want to find out how pay or working conditions would have been different if you changed the worker's sex and left everything else the same. This is obviously a hypothetical question. But that doesn't mean it isn't a useful one to answer or that it isn't important to answer the question correctly.

### Example 19.1: Greenhouse Gases and Global Warming

Global warming is in the news every day. Warming is not a recent trend, but one that extends over decades, perhaps even a century or more, as shown in Figure 19.1. The cause of the increasing temperatures is thought to be the increased atmospheric concentration of greenhouse gases such as CO<sub>2</sub> and methane. These gases have been emitted at a rapidly growing rate with population growth and industrialization.

In the face of growing consensus about the problem, skeptics note that such temperature data provides support but not proof for claims about greenhouse-gas induced global warming. They point out that climate is not steady; it changes over periods of decades, over centuries, and over much longer periods of time. And it's not just CO<sub>2</sub> that's been increasing over the last century; lots of other variables are correlated with temperature. Why not blame global warming, if it does exist, on them?

Imagine that you were analyzing the data in Figure 19.1 without any idea of a possible mechanism of global temperature change. The data look something



like a random walk; a sensible null hypothesis might be exactly that. The typical year-to-year change in temperature is about  $0.05^\circ$ , so the expected drift from a random walk over the 130 year period depicted in the graph is about  $0.6^\circ$ , roughly the same as that observed.

The data in the graph do not themselves provide a compelling basis to reject the null hypothesis that global temperatures change in a random way. However, it's important to understand that climatologists have proposed a physical **mechanism** that relates  $\text{CO}_2$  and methane concentration in the atmosphere to global climate change. At the core of this mechanism is the increased absorption of infra-red radiation by greenhouse gases. This core is not seriously in doubt: it's solidly established by laboratory measurements. The translation of that absorption mechanism into global climate consequences is somewhat less solid. It's based on computer models of the physics of the atmosphere and the ocean. These models have increased in sophistication over the last couple of decades to incorporate more detail: the formation of clouds, the thermohaline circulation in the oceans, etc. It's the increasing confidence in the models that drives scientific support for the theory that greenhouse gases cause global warming.

It's not data like Figure 19.1 that lead to the conclusion that  $\text{CO}_2$  is causing global warming. It's the data insofar as they support models of mechanisms.

---

## 19.2 Causation and Correlation

It's often said, "Correlation is not causation." True enough. But it's an odd thing to say, like saying, "A movie is not a train."

In the earliest days of the cinema, a Lumière brothers film showing a train arriving at a station caused viewers to rise to their feet as if the train were real. Reportedly, some panicked from fear of being run over.[35, p. 222]

Modern viewers would not be fooled; we know that a movie train is incapable of causing us harm. The Lumière movie authentically represented a real

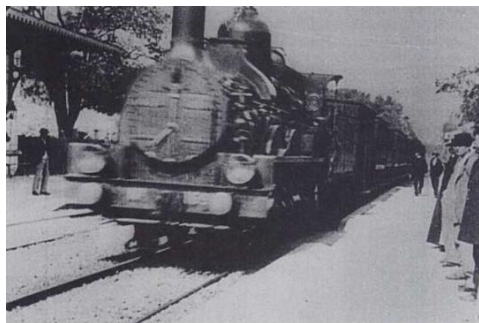


Figure 19.2: A frame from the 1895 film, "L'Arrivée d'un train en gare de La Ciotat"

train — the movie is a kind of model, a representation for a purpose — but the representation is not the mechanical reality of the train itself. Similarly, correlation is a representation of the relationship between variables. It captures some aspects of that relationship, but it is not the relationship itself and it doesn't fully reflect the mechanical realities, whatever they may be, of the real relationships that drive the system.

Correlation, along with the closely related idea of model coefficients, is a concept that applies to data and variables. The correlation between two variables depends on the data set, and how the data were collected: what sampling frame was used, whether the sample was randomly taken from the sampling frame, etc.

In contrast, causation refers to the influence that components of a system exert on one another. I write “component” rather than “variable” because the variables that are measured are not necessarily the active components themselves. For example, the score on an IQ test is not intelligence itself, but a reflection of intelligence.

As a metaphor for the differences between correlation and causation, consider a chain hanging from supports. (Figure 19.3.) Each link of the chain is mechanically connected to its two neighbors. The chain as a whole is a collection of such local connections. Its shape is set by these mechanical connections together with outside forces: the supports, the wind, etc. The overall system — both internal connections and outside forces — determines the global shape of

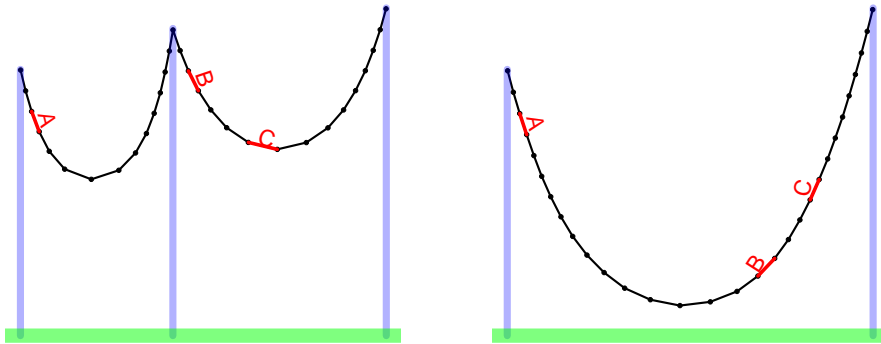


Figure 19.3: A metaphor for causation and correlation: the links of a chain with external supports. It's the same chain in both pictures but supported differently. The relationship between link orientations — links A and B are aligned in the picture on the left, but not in the picture on the right — depends both on the mechanical connections between links and on the external forces at work. So the alignment itself (correlation) is not a good signal for the mechanical connections (causation).

the chain. That overall shape sets the correlation between any two links in the chain, whether they be neighboring or distant.

In this metaphor, the shape of the chain is analogous to correlation; the mechanical connections between neighboring links is causation.

One way to understand the shape of the chain is to study its overall shape: the correlations in it. But be careful; the lessons you learn may not apply in different circumstances. For example, if you change the location of the supports, or add a new support in the middle, the shape of the chain can change completely as can the correlation between components of the chain.

Another way to understand the shape of the chain is to look at the local relationships between components: the causal connections. This does not directly tell you the overall shape, but you can use those local connections to figure out the global shape in whatever circumstances may apply. It's important to know about the mechanism so that you can anticipate the response to actions you take: actions that might change the overall shape of the system. It's also important for reasoning about counterfactuals: what would have happened had the situation been different (as in the sex discrimination example above) even if you have no way actually to make the system different. You can't change the plaintiff's sex, but you can play out the consequences of doing so through the links of causal connections.

One of the themes of this book has been that correlation, as measured by the correlation coefficient between two variables, is a severely limited way to describe relationships. Instead, the book has emphasized the use of model coefficients. These allow you to incorporate additional variables — covariates — into your interpretation of the relationship between two variables. Model coefficients provide more flexibility and nuance in describing relationships than does correlation. They make it possible, for instance, to think about the relationship between variables A and B *while adjusting for* variable C.

Even so, model coefficients describe the global properties of your data. If you want to use them to examine the local, mechanistic connections, there is more work to be done. Presumably, what people mean in saying “correlation is not causation” is that correlation is not on its own compelling evidence for causation.

An example of the difference between local causal connections and global correlations comes from political scientists studying campaign spending. Analysis of data on election results and campaign spending in US Congressional elections shows that increased spending by those running for re-election — incumbents — is associated with lower vote percentages. This finding is counter-intuitive. Can it really be that an incumbent's campaign spending causes the incumbent to lose votes? Or is it that incumbents spend money in elections that are closely contested for other reasons? When the incumbent's election is a sure thing, there is no need to spend money on the campaign. So the negative correlation between spending and votes, although genuine, is really the result of external forces shaping the election and not the mechanism by which campaign spending affects the outcome.[36]

## 19.3 Hypothetical Causal Networks

In order to think about how data can be used as evidence for causal connections, it helps to have a notation for describing local connections. The notation I will use involves simple, schematic diagrams. Each diagram depicts a **hypothetical causal network**: a theory about how the system works. The diagrams consist of **nodes** and **links**. Each node stands for a variable or a component of the system. The nodes are connected by links that show the connections between them. A one-way arrow refers to a causal mechanistic connection. To illustrate, Figure 19.4 shows a hypothetical causal network for campaign spending by an incumbent.

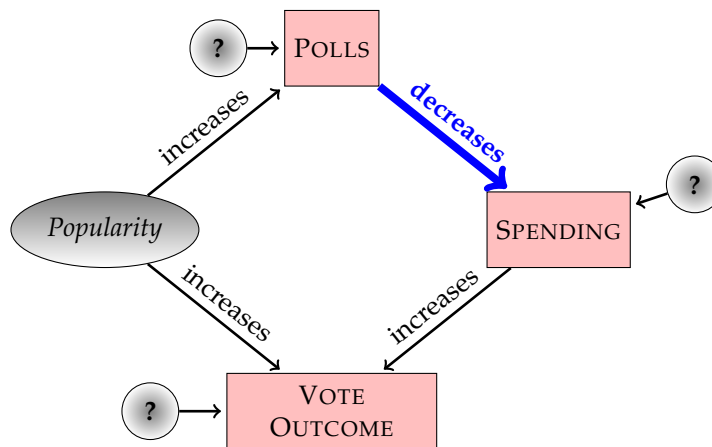



Figure 19.4: A hypothetical causal network describing how campaign spending by an incumbent candidate for political office is related to the vote outcome.

The hypothetical causal network in Figure 19.4 consists of four main components: spending and the vote outcome are the two of primary interest, but the incumbent candidate's popularity and the pre-election poll results are also included. According to the network, an incumbent's popularity influences both the vote outcome and the pre-election polls. The polls indicate how close the election is and this shapes the candidate's spending decisions. The amount spent influences the vote total.

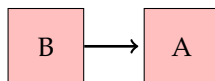
A complete description of the system would describe how the various influences impinging on each node shape the value of the quantity or condition represented by the node. This could be done with a model equation, or less completely by saying whether the connection is positive or negative (as in the above diagram). For now, however, focus on the topology of the network: what's connected to what and which way the connection runs.

Nodes in the diagrams are drawn in two shapes. A square node refers to a variable that can be measured: poll results, spending, vote outcomes. Round nodes are for unmeasured quantities. For example, it seems reasonable to think

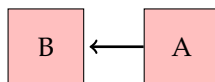
about a candidate's popularity, but how to measure this outside of the context set by a campaign? So, in the diagram, popularity itself is not directly measured. Instead, there are poll results and the vote outcome itself. Specific but unmeasured variables such as "popularity" are sometimes called **latent variables**.

Often the round, unmeasured nodes will be drawn , which stands for the idea the *something* is involved, but no description is being given about what that something is; perhaps it's just random noise.

The links connecting nodes indicate causal influence. Note that every line has an arrow that tells which way causation works. The diagram

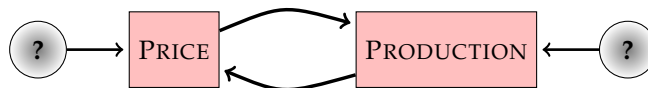



means that B causally influences A. The diagram





is the opposite: A is causally influencing B.

It's possible to have links running both ways between nodes. For instance, A causes B and B causes A. This is drawn as two different links. Such two way causation produces loops in the diagrams, but it is not necessarily illogical circular reasoning. In economics, for instance, it's conventional to believe that price influences production and that production influences price.



Why two causal links? When some outside event intervenes to change production, price is affected. For example, when a factory is closed due to a fire, production will fall and price will go up. If the outside event changes price — for instance, the government introduces price restrictions — production will change in response. Such outside influences are called **exogenous**. The  stands for an unknown exogenous input.

A hypothetical causal network is a model: a representation of the connections between components of the system. Typically, it is incomplete, not attempting to represent all aspects of the system in detail. When an exogenous influence is marked as , the modeler is saying, "I don't care to try to represent this in detail." But even so, by marking an influence with , the modeler is making an affirmative claim that the influence, whatever it be, is not itself caused by any of the other nodes in the system: it's exogenous. In contrast, nodes with a causal input — one or more links pointing to them — are **endogenous**, meaning that they are determined at least in part by other components of the system.

Often modelers decide not to represent all the links and components that might causally connect two nodes, but still want to show that there is a connection. Such **non-causal links** are drawn as double-headed dashed lines as in Figure 19.5. For instance, in many occupations there is a correlation between age and sex: older workers tend to be male, but the population of younger workers

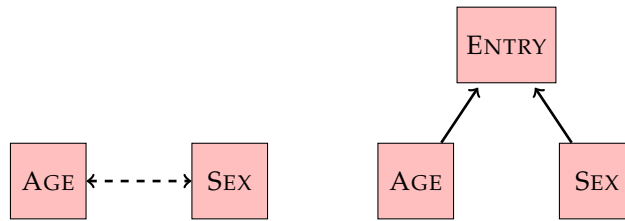


Figure 19.5: A non-causal link and its expansion into a diagram with causal links.

is more balanced. It would be silly to claim a direct causal link from age to sex: a person’s sex doesn’t change as they age! Similarly, sex doesn’t determine age. Instead, women historically were restricted in their professional options. There is an additional variable — entry into the occupation — that is determined by both the person’s age and sex. The use of a non-causal link to indicate the connection between age and sex allows the connection to be displayed without including the additional variable.

It’s important not to forget the word “hypothetical” in the name of these diagrams. A hypothetical causal network depicts a state of belief. That belief might be strongly informed by evidence, or it might be speculative. Some links in the network might be well understood and broadly accepted. Some links, or the absence of some links, might be controversial and disputed by other people.

## 19.4 Networks and Covariates

Often, you are interested in only some of the connections in a causal network; you want to use measured data to help you determine if a connection exists and to describe how strong it is. For example, politicians are interested to know how much campaign spending will increase the vote result. Knowing this would let them decide how much money they should try to raise and spend in an election campaign.

It’s wrong to expect to be able to study just the variables in which you have a direct interest. As you have seen, the inclusion of covariates in a model can affect the coefficients on the variables of interest.

For instance, even if the direct connection between spending and vote outcome is positive, it can well happen that using data to fit a model  $\text{vote outcome} \sim \text{spending}$  will produce a negative coefficient on spending.

There are three basic techniques that can be used to collect and analyze data in order to draw appropriate conclusions about causal links.

**Experiment**, that is, intervene in the system to set or influence certain variables and then examine how your intervention relates to the observed outcomes.

**Include covariates** in order to adjust for other variables.

**Exclude covariates** in order to prevent those variables from unduly influencing your results.

# Chapter 20

## Experiment

*The true method of knowledge is experiment.* — William Blake (1757-1827)

*A theory is something nobody believes, except the person who made it. An experiment is something everybody believes, except the person who made it.*  
— Albert Einstein (1879-1955)

### 20.1 Experiments

The word “experiment” is used in everyday speech to mean *a course of action tentatively adopted without being sure of the eventual outcome*. [26] You can experiment with a new cookie recipe; experiment by trying a new bus route for getting to work; experiment by planting some new flowers to see how they look.

Experimentation is at the core of the scientific method. Albert Einstein said, “[The] development of Western Science is based on two great achievements — the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance).” [38, p.1]

Scientific experimentation is, as Einstein said, systematic. In the everyday sense of experimentation, to try out a new cookie recipe, you cook up a batch and try them out. You might think that the experiment becomes scientific when you measure the ingredients carefully, set the oven temperature and timing precisely, and record all your steps in detail. Those are certainly good things to do, but a crucial aspect of systematic experimentation is **comparison**.

Two distinct forms of comparison are important in an experiment. One form of comparison is highlighted by the phrase “**controlled experiment**”. A controlled experiment involves a comparison between two interventions: the one you are interested in trying out and another one, called a **control**, that provides a baseline or basis for comparison. In the cookie experiment, for instance, you

For review purposes only

would compare two different recipes: your usual recipe (the control) and the new recipe (called the **treatment**). Why? So that you can put the results for the new recipe in an appropriate context. Perhaps the reason you wanted to try a new recipe is that you were particularly hungry. If so, it's likely that you will like the new recipe even if were not so good as your usual one. By comparing the two recipes side by side, a controlled experiment helps to reduce the effects of factors such as your state of hunger. (The word "controlled" also suggests carefully maintained experimental conditions: measurements, ingredients, temperature, etc. While such care is important in successful experiments, the phrase "controlled experiment" is really about the comparison of two or more different interventions.)

Another form of comparison in an experiment concerns intrinsic variability. The different interventions you compare in a controlled experiment may be affected by factors other than your intervention. When these factors are unknown, they appear as **random variation**. In order to know if there is a difference between your interventions — between the treatment and the control — you have to do more than just compare treatment and control. You have to compare the observed treatment-vs-control difference to that which would be expected to occur at random. The technical means of carrying out such a comparison is, for instance, the F statistic in ANOVA. When you carry out an experiment, it's important to establish the conditions that will enable such a comparison to be carried out.

A crucial aspect of experimentation involves causation. An experiment does not necessarily eliminate the need to consider the structure of a hypothetical causal network when selecting covariates. However, the experiment actually changes the structure of the network and in so doing can radically simplify the system. As a result, experiments can be much easier to interpret, and much less subject to controversy in interpretation than an observational study.

### 20.1.1 Experimental Variables and Experimental Units

In an experiment, you, the experimenter, intervene in the system.

- You choose the **experimental units**, the material or people or animals to which the control and treatment will be applied. When the "unit" is a person, a more polite term is **experiment subject**.
- You set the value of one or more variables — the **experimental variables** — for each experimental unit.

The experimenter then observes or measures the value of one or more variables for each experimental unit: a response variable and perhaps some covariates.

The choice of the experimental variables should be based on your understanding of the system under study and the questions that you want to answer. Often, the choice is straightforward and obvious; sometimes it's clever and ingenious. For example, suppose you want to study the possibility that the habit of regularly taking a tablet of aspirin can reduce the probability of death from stroke or heart attack in older people. The experimental variable will be whether



## Further Readings & Bibliography

The reader interested in learning more about statistics will do well to start with some history. David Salsburg's *The Lady Tasting Tea* [52] gives an entertaining and readable survey of 20th century statistics and statisticians. Stephen Stigler's *Statistics on the Table* [53] reaches back a century further in history and has a more conceptual focus. Ian Hacking's *The Emergence of Probability* goes back, as the title says, to the emergence of the concept of probability in the seventeenth century.

**Chapter 1. Overview.** For essays about the application of statistics in context and in practice, the various editions of *Statistics: A Guide to the Unknown* are hard to beat.[54, 55] Each *Guides* essays is oriented around a particular social, political, medical, or scientific issue. In contrast, most statistics textbooks ([42, 56, 57, 58] are among the good choices) are arranged around topics in statistical methodology. The cited books cover a complete gamut of statistical topics — sampling, descriptive statistics, statistical tests — but emphasize the statistics of single variables or the relationship between pairs of variables. They build on high-school algebra and do not require college-level calculus. At a somewhat more advanced mathematical level, the well-titled *Statistical Sleuth* introduces multivariate regression. At a still higher level, for those who are comfortable with college-level linear algebra, an excellent multivariate modeling text is [59]. For classroom learning activities in introductory statistics, [60, 61] provide many ideas.

An excellent, non-statistical introduction to constructing models, with a strong emphasis on models as “purposeful representations,” is [62].

Many materials about the R package [5] are available at the project web site: [www.r-project.org](http://www.r-project.org). Good references for R (and the language it is based on, S) are [63, 64, 65]. Some introductions to statistics using R are [66, 67, 68].

**Chapters 2 & 3. Data and Description.** Introductory statistics books (e.g. [42, 56, 57, 58]) provide a good introduction to methods and issues in collecting data and describing variables numerically and graphically. For advanced

modeling methods with units of analysis at different levels, see [69].

The series of books by Tufte [70, 71] has been rightfully influential in showing how to present data graphically (and beautifully). For technical displays of statistical data, see Cleveland's *Visualizing Data*. [72]

Important criticisms of the emphasis on conventional descriptions of center and spread are provided by the best-selling books by Gould [73] and Taleb [74].

The elementary but classic *How to Lie with Statistics* [75] has many examples of the use of quantitative information in a way that misleads; it's not so much about statistics as it is about lying. [76]

The somewhat obscure statistic "unlikeability" is described in [77] and [78].

**Chapters 4 - 8. Modeling and Correlation.** The modeling notation is described in an influential advanced book by Chambers and Hastie. [65] Box, Hunter, & Hunter give good reasons to take fitted models with a grain of salt. [2, pp. 397-407]

Stigler provides an account of the invention of correlation. [79] For a variety of different ways to interpret correlation, see [80].

Gawande writes for a general reader about why simple model formulas can be more useful for prediction and diagnosis than human judgment. [81]

**Chapters 9 - 13. Geometry.** The idea of using geometry to inform statistical reasoning started perhaps with Ronald Fisher at the start of the 20th century. [82] It was unfortunately seen as obscure and statistics pedagogy was rooted in algebra. Fisher's daughter, Joan Fisher Box, writes about the reaction of the famous statistician William Gosset:

.... Gosset found the geometric representation always highly mysterious; but he accepted the mystification, and jokingly would remark, "I take it that whatever it is follows at once 'obviously' from a consideration of  $n$ -dimensional geometry," or demand of Fisher whether a certain equation came out of  $n$ -dimensional space "or what is much the same thing, your head." [83]

Starting in the 1980s, David Saville and Graham Wood have been writing about the uses of geometry in teaching statistics. [84, 85, 86]. Other articles appear from time to time, e.g., [87, 88]. The classic book *Statistics for Experimenters* [2] contains a thread using geometry to complement algebraic explanations. See also [89]. Farebrother describes mechanical analogies for fitting [90] and gives an early history [91].

The idea of a random walk has been important in science. Robert Brown's observation of the movement of pollen grains in 1827 is reported in [20]. One of the famous trio of papers from Albert Einstein's 1905 *annus mirabilis* is about Brownian motion, [22]; Perrin [21] won a Nobel prize for his work confirming Einstein's theory experimentally. The accessible book by Berg [92] reveals many aspects of random walks in biology. There are also best-selling accounts, e.g., [93, 94, 95].

**Chapters 14-18. Statistical Inference.** Descriptions of the classic tests — t-tests and so on — are the bread and butter of statistics textbooks. A discussion of the search for significance and how this can render p-values uninterpretable

is the subject of an editorial by Freedman.[17] Other papers on multiple comparisons and inconsistency are [27, 96]; important critiques of statistical inference are [97, 98]. Good and Hardin's book, *Common Errors in Statistics (and How to Avoid Them)*, is a useful reference.[99]

The work of Bradley Efron introduced randomization methods such as bootstrapping to many social and natural scientists and statistical workers. [100, 101, 102] An early attempt to introduce randomization into teaching statistics at an introductory level is [103]. Hardly any introductory statistics textbook even mentions resampling or bootstrapping, though thankfully there are signs this is starting to change.

Saville and Wood make a case for the angle between vectors as a primary statistic. [104, 105] Figure 16.7 is inspired by Fig. 10.A.2.b in [2] and similar figures in Appendix D of [86].

**Chapters 19 & 20. Causation and Experiment.** The work of Judea Pearl on causation is the inspiration for the material on hypothetical causal networks. A book-length treatment is [37]; the last chapter should perhaps be read first as an introduction. Many of Pearl's papers are available on the web. The "gentle introduction" in [106] is a good place to start.

Box, Hunter, and Hunter [2] is a well regarded survey of experimental design, artfully combining theory and practice. Cobb [107] provides a carefully graded path through experimental design and analysis accessible; it's sophisticated but accessible to a non-mathematical reader.

For background on instrumental variables, see [46, 108]. For matched sampling: [109, 110, 111, 112]. The propensity-score method has been tested in situations where there were parallel studies — one experimental and one observational: informed skepticism is called for. [113, 114, 115, 116] The best-selling book *Freakonomics* [51] entertainingly reports many clever attempts to extract information about causation from observational data.

## References

- [1] George Cobb. The introductory statistics course: a ptolemaic curriculum? *Technology Innovations in Statistics Education*, 1(1), 2007.
- [2] George E.B. Box, J. Stuart Hunter, and William G. Hunter. *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley, 2nd edition, 2005.
- [3] Robert Hughes. *The Fatal Shore: the epic of Australia's founding*. Vintage, 1988.
- [4] John M Chambers. *Software for data analysis: Programming with R*. Springer, 2008.
- [5] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.
- [6] Daniel T Kaplan. Computing and introductory statistics. *Technology Innovations in Statistics Education*, 1(1):Article 5, 2007. <http://www.escholarship.org/uc/item/3088k195>.
- [7] David Bellhouse. Ann landers survey on parenthood. web pamphlet: [www.stats.uwo.ca/faculty/bellhouse/stat353annlanders.pdf](http://www.stats.uwo.ca/faculty/bellhouse/stat353annlanders.pdf), 2008. accessed Jan. 2008.

- [8] James A. Hanley. "transmuting" women into men: Galton's family data on human stature. *American Statistician*, 58(3):237–243, 2004.
- [9] John M Chambers and Trevor J Hastie. *Statistical Models*. Wadsworth, 1992.
- [10] Deborah Lynn Guber. Getting what you pay for: the debate over equity in public school expenditures. *Journal of Statistics Education*, 7(2), 1999.
- [11] DWJ Thompson. A large discontinuity in the mid-twentieth century in observed global-mean surface temperature. *Nature*, 453:646–649, 29 May 2008.
- [12] Francis Galton. Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, pages 135–145, 1888.
- [13] College Board. Research report no. 2008-5: Validity of the sat for predicting first-year college grade point average. Technical report, College Board, 2008.
- [14] A.A. Bailey and P.L. Hurd. Finger length ratio (2d:4d) correlates with physical aggression in men but not in women. *Biol. Psychol.*, 68(3):215–222, 2005.
- [15] Rachel Carson. *Silent Spring*. Houghton, 1962.
- [16] John Tierney. Fateful voice of a generation still drowns out real science. New York Times, June 5 2007.
- [17] David A Freedman. Editorial: Oasis or mirage? *Chance*, 21(1):59–61, 2008.
- [18] <http://www.infoplease.com/ipa/a0883976.html>, 2008.
- [19] E W Montroll and M F Schlesinger. Maximum entropy formalism, fractals, scaling phenomena, and  $1/f$  noise: a tale of tails. *Journal of Statistical Physics*, 32(209), 1983.
- [20] Robert Brown. *Miscellaneous Botanical Works*, volume 1. London, Royal Society, 1866. quoted in MP Crosland (1971) *The Science of Matter*.
- [21] M. Jean Perrin. *Brownian Movement and Molecular Reality*. Taylor and Francis, London, 1910.
- [22] Albert Einstein. Investigations on the theory of the brownian movement. In A.D. Cowper, editor, *Collected Papers*, volume 2. Dover Publications, 1956.
- [23] James Surowiecki. Running numbers. *The New Yorker*, page 29, Jan. 21 2008.
- [24] Roger A. Pielke Jr. Who decides? forecasts and responsibilities in the 1997 red river flood. *Applied Behavioral Science Review*, 7(2):83–101, 1999.
- [25] Ronald A. Fisher. On a distribution yielding the error functions of several well-known statistics. *Prod. Int. Cong. Math., Toronto*, 2:805–813, 1924.
- [26] New oxford american dictionary.
- [27] David J. Saville. Multiple comparison procedures: the practical solution. *The American Statistician*, 44(2):174–180, 1990.
- [28] Office of Federal Contract Compliance Programs. *Federal Contract Compliance Manual*, chapter Chapter III: Onsite Review Procedures. US Department of Labor, 1993, revised 2002.
- [29] Richard Shavelson Stephen Klein, Roger Benjamin. The collegiate learning assessment: Facts and fantasies. Technical report, Council for Aid to Education, [http://www.cae.org/content/pro\\_collegiate\\_reports\\_publications.htm](http://www.cae.org/content/pro_collegiate_reports_publications.htm), 2007.
- [30] Paul Basken. Test touted as 2 studies question its value. *The Chronical of Higher Education*, 54(39):A1, June 6 2008.
- [31] S. Koutros et al. Meat and meat mutagens and risk of prostate cancer in the agricultural health study. *Cancer epidemiology biomarkers and prevention*, 17:80–87, 2008.

For review purposes only

- [32] Rod Minchin. personal communication. 2008.
- [33] <http://www.prostateline.com/prostatelinehcp>. Web site, 2008.
- [34] Charles Dana Wilber. *The Great Valleys and Prairies of Nebraska and the Northwest*. Daily Republican Print, Omaha, Neb., 3rd edition, 1881.
- [35] Leonard J. Schmidt, Brooke Warner, and Peter A. Levine. *Panic: Origins, Insight, and Treatment*. North Atlantic Books, 2002.
- [36] Gary C. Jacobson. The effects of campaign spending in congressional elections. *The American Political Science Review*, 72(2):469–491, 1978.
- [37] Judea Pearl. *Causation: Models, Reasoning, and Inference*. Princeton Univ. Press, 2000.
- [38] Daniel J. Boorstin. *Cleopatra's Nose: Essays on the Unexpected*. Vintage, 1995.
- [39] A Hrobjartsson and PC Gotzsche. Is the placebo powerless? update of a systematic review with 52 new randomized trials comparing placebo with no treatment. *Journal of Internal Medicine*, 256(2):91–100, 2004.
- [40] BE Wampold, T Minami, SC Tierney, TW Baskin, and KS Bhati. The placebo is powerful: estimating placebo effects in medicine and psychotherapy from randomized clinical trials. *Journal of Clinical Psychology*, 61(7):835–854, 2005.
- [41] A Hrobjartsson and M Norup. The use of placebo interventions in medical practice — a national questionnaire survey of danish clinicians. *Evaluation and the Health Professions*, 26(2):153–165, 2003.
- [42] David Freedman, Roger Purves, and Robert Pisani. *Statistics*. WW Norton, 4th edition, 2007.
- [43] J Bruce Moseley et al. A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *New England Journal of Medicine*, 347:81–88, 2002.
- [44] TB Freeman. Use of placebo surgery in controlled trials of a cellular-based therapy for parkinson's disease. *New England Journal of Medicine*, 341(13):988–992, 1999.
- [45] CL Campbell, S Smyth, G Montalescot, and S R Steinhubl. Aspirin dose for the prevention of cardiovascular disease: a systematic review. *Journal of the American Medical Association*, 18:2018–2024, 2007.
- [46] Joshua D. Angrist and Alan B. Krueger. Instrumental variables and the search for identification: from supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85, 2001.
- [47] David A. Freedman. Statistical models and shoe leather. In Peter Marsden, editor, *Sociological Methodology 1991*. American Sociological Association, 1991.
- [48] Joshua D. Angrist and Alan B. Krueger. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Econometrics*, 106(4):979–1014, 1991.
- [49] Gary C. Jacobson. The effects of campaign spending in house elections: new evidence for old arguments. *American Journal of Political Science*, 34(2):334–362, 1990.
- [50] Donald P Green and Jonathan S Krasno. Salvation for the spendthrift incumbent: Reestimating the effects of campaign spending in house elections. *American Journal of Political Science*, 32(4):884–907, 1988.
- [51] Steven D Levitt and Stephen J Dubner. *Freakonomics*. William Morrow, 2005.
- [52] David Salsburg. *The Lady Tasting Tea: How statistics revolutionized science in the twentieth century*. W.H. Freeman, New York, 2001.

For review purposes only

- [53] Stephen M Stigler. *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard Univ. Press, 1999.
- [54] Roxy Peck, George Casella, George Cobb, Roger Hoerl, Deborah Nolan, Robert Starbuck, and Hal Stern. *Statistics: A Guide to the Unknown*. Duxbury Press, 4th edition, 2005.
- [55] Judith M. Tanur, Frederick Mosteller, William H. Kruskal, Erich L. Lehmann, Richard F. Link, Richard S. Pieters, and Gerald R. Rising. *Statistics: A Guide to the Unknown*. Duxbury Press, 3rd edition, 1989.
- [56] David S Moore, George P McCabe, and Bruce Craig. *Introduction to the Practice of Statistics*. W.H. Freeman, 6th edition, 2007.
- [57] Anne E Watkins, Richard L Scheaffer, and George W Cobb. *Statistics in Action: Understanding a World of Data*. Key College Publishing, 2004.
- [58] David E Bock, Paul F Velleman, and Richard D De Veaux. *Stats: Modeling the World*. Addison Wesley, 2nd edition, 2006.
- [59] David A. Feedman. *Statistical Models: Theory and Practice*. Cambridge University Press, 2005.
- [60] Allan J Rossman and Beth Chance. *Workshop Statistics: Discovery with Data*. Wiley, 2nd edition, 2008.
- [61] Andrew Gelman and Deborah Nolan. *Teaching Statistics: A Bag of Tricks*. Oxford Univ. Press, 2002.
- [62] Anthony M. Starfield, Karl A. Smith, and Andrew L. Blelock. *How to model it*. McGraw-Hill, 1990.
- [63] William N Venables and Brian D Ripley. *Modern Applied Statistics with S*. Springer, 4th edition, 2002.
- [64] Phil Spector. *Data Manipulation with R*. Springer, 2008.
- [65] John M Chambers and Trevor J Hastie. *Statistical Models in S*. Chapman & Hall, 1992.
- [66] Peter Dalgaard. *Introductory Statistics with R*. Springer, 2nd edition, 2008.
- [67] John Verzani. *Using R for Introductory Statistics*. Chapman & Hall, 2005.
- [68] Michael J Crawley. *Statistics: An Introduction using R*. Wiley, 2005.
- [69] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multi-level/hierarchical models*. Cambridge Univ. Press, 2007.
- [70] Edward R Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 2nd edition, 2001.
- [71] Edward R Tufte. *Beautiful Evidence*. Graphics Press, 2006.
- [72] William S Cleveland. *Visualizing Data*. Hobart Press, 1993.
- [73] Stephen J. Gould. *The Mismeasure of Man*. W.W. Norton, 1996.
- [74] Nassim Taleb. *The Black Swan*. Random House, 2007.
- [75] Darell Huff. *How to Lie with Statistics*. Norton, 1954. illustrated by Irving Geis.
- [76] J. Michael Steele. Darell huff and fifty years of how to lie with statistics. *Statistical Science*, 20(5):205–209, 2005.
- [77] Gary D. Kader and Mike Perry. Variability for categorical variables. *Journal of Statistics Education*, 15(2), 2007. <http://www.amstat.org/publications/jse/v15n2/kader.html>.

For review purposes only

- [78] Alan Agresti. *Categorical Data Analysis*. John Wiley and Sons, 1990.
- [79] Stephen M Stigler. Francis galton's account of the invention of correlation. *Statistical Science*, 4(2):73–79, 1989.
- [80] Joseph L Rodger and W Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.
- [81] Atul Gawande. *Complications: A Surgeon's Notes on an Imperfect Science*. Metropolitan/Henry Holt, 2002.
- [82] David G. Herr. On the history of the use of geometry in the general linear model. *The American Statistician*, 34(1):43–47, 1980. see 2682995.pdf.
- [83] Joan Fisher Box. Guinness, gosset, fisher, and small samples. *Statistical Science*, 2(1):45–52, 1987.
- [84] David J. Saville and Graham R. Wood. A method for teaching statistics using n-dimensional geometry. *The American Statistician*, 40(3):205–214, 1986.
- [85] David J. Saville and Graham R. Wood. *Statistical Methods: The Geometric Approach*. Springer, 1997.
- [86] David J. Saville and Graham R. Wood. *Statistical Methods: A Geometric Primer*. Springer, 1996.
- [87] Johan Bring. A geometric approach to compare variables in a regression model. *The American Statistician*, 50(1):57–62, 1996. see 2685045.pdf.
- [88] David Bock and Paul F. Velleman. Why variances add — and why it matters. In Gail F. Burrill and Portia C. Elliot, editors, *Thinking and reasoning with data and chance*. National Council of Teachers of Mathematics, 2006.
- [89] Michael J. Wichura. *The Coordinate-free Approach to Linear Models*. Cambridge University Press, 2006.
- [90] Richard W. Farebrother. *Visualizing Statistical Models and Concepts*. Marcel Dekker, 2002.
- [91] Richard W. Farebrother. *Fitting Linear Relationship: A History of the Calculus of Observations 1750-1900*. Springer, 1999.
- [92] Howard C Berg. *Random Walks in Biology*. Princeton Univ. Press, 1993.
- [93] Burton G Malkiel. *A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investment*. W.W. Norton, 2007.
- [94] Leonard Mlodinow. *The Drunkard's Walks: How Randomness Rules Our Lives*. Pantheon, 2008.
- [95] Nassim Nicholas Taleb. *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*. Random House, 2008.
- [96] David J. Saville. Basic statistics and the inconsistency of multiple comparison procedures. *Canadian Journal of Experimental Psychology*, 57(3):167–175, 2003.
- [97] Charmont Wang. *Sense and Nonsense of Statistical Inference*. CRC, 1992.
- [98] Richard A. Berk. *Regression analysis: A constructive critique*. Sage publications, 2004.
- [99] Phillip I. Good and James W. Hardin. *Common Errors in Statistics (and How to Avoid Them)*. Wiley Interscience, 2003.
- [100] Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7:1–26, 1979.

- [101] Bradley Efron. Computers and the theory of statistics: Thinking the unthinkable. *SIAM Review*, 21:460–480, 1979.
- [102] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.
- [103] Julian L Simon. *Resampling: The New Statistics*. Resampling Stats, 1974–1992.
- [104] Graham R Wood and David J Saville. The ubiquitous angle. *Journal Royal Statistical Society A*, 168(1):95–107, 2005.
- [105] Dave Saville and Graham Wood. The geometry of the p-value. In *Collaborations, Designs, and Explorations: A festschrift for Peter Johnstone*, pages 99–111. 2006. ISBN 0-478-20909-6.
- [106] Judea Pearl. Causal inference in statistics: A gentle introduction. In *Computing Science and Statistics, Proceedings of Interface '01*, volume 33, 2001.
- [107] George W Cobb. *Introduction to Design and Analysis of Experiments*. Springer, 1998.
- [108] Sander Greenland. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29:722–729, 2000.
- [109] D.B. Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74:318–328, 1979.
- [110] P.R. Rosenbaum and D.B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39:33–38, 1985.
- [111] P.R. Rosenbaum. *Observational Studies*. Springer-Verlag, 2nd ed. edition, 2002.
- [112] B.B. Hansen. Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, 99(467):609:618, 2004.
- [113] K Arceneaux, A S Gerber, and D P Green. Comparing experimental and matching methods using a large-scale voter mobilization experiment. *Political Analysis*, 14:37–62, 2006.
- [114] D A Lawlor, G DaveySmith, D Kundu, and et al. Those confounded vitamins: what can we learn from the differences between observational vs randomised trial evidence. *The Lancet*, 363:1724–1727, 2004.
- [115] R Kunz and A D Oxman. The unpredictability paradox: review of empirical comparisons of randomised and non-randomized clinical trials. *British Medical Journal*, 317:1185–1190, 1998.
- [116] William R Shadish, M H Clark, and Peter M Steiner. Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *J. American Statistical Association*, 103(484):1334–1356, 2008.

## Datasets

### College Grade Database

There are three tables in this data set: `grades.csv`, `courses.csv`, and `grade-to-number.csv`.



For review purposes only



grades.csv has one row for each course taken by a 2005 graduate in his or her college career. To preserve confidentiality, the data are a 50% random sampling of the complete listing. The variables are:

- sid — A unique ID for each student.
- sessionNo — A unique ID for each course session. If a course has multiple sections, that is, different groups of students who meet at different times, each section will have a unique session number.
- grade — The student's letter grade in that session: A, A-, B+, and so on through D-. Other codes: I=incomplete, NC=failed/no credit, S=passed (for courses taken on a pass/fail basis), AU=audit (taken not for credit). The numerical equivalents for grade-point calculations are given in the file grade-to-number.csv, described below.

courses.csv has one row for each course session listed in the grades table. Since there are many students in each session, the number of courses is much smaller than the number of cases in the grades table. For each course session, the variables are:

- sessionID — The unique session ID corresponding to the entries in the grades table.
- iid — A unique ID for each instructor.
- dept — The name of the department in which the course was offered. To preserve confidential information, a single-letter code is used for each department.
- sem — The semester in which the course session took place.
- enroll — The number of students enrolled in the course.
- level — A number indicating the "level" of the course. 100-level courses are the most elementary, 200 is more advanced, then 300, then 400. Courses marked with 600 are research courses, internships, and so on.

To preserve confidentiality, the courses table contains a random sample of the courses offered, and includes only courses with an enrollment of 10 or greater. (For this reason, there are very few 600-level courses listed.)

grade-to-number.csv contains the translation from a letter grade to a number.

The coding and subsampling done was done for reasons of confidentiality. The version of the data set retained for internal study at the institution retains all the data and proper identifying labels.

## Galton's Height Data

**Context** In the 1880's, Francis Galton was developing ways to quantify the heritability of traits. As part of this work, he collected data on the heights of adult children and their parents.

**Variables** Each child is one case.

- height the child's height (as an adult) in inches.
- sex the child's sex: F or M
- mother the mother's height in inches
- father the father's height in inches
- nkids the number of adult children in the family, or, at least, the number whose heights Galton recorded.
- family a numerical code to identify the members of one family.

Entries were deleted for those children whose heights were not recorded numerically by Galton, who sometimes used entries such as "tall", "short", "idiotic", "deformed" and so on.

**Sources** The data were transcribed by J.A. Hanley [8], who has published them at <http://www.medicine.mcgill.ca/epidemiology/hanley/galton/>.

## Kids' Foot Measurements

DATA FILE  
kidsfeet.csv

**Context** These data were collected by a statistician, Mary C. Meyer, in a fourth grade classroom in Ann Arbor, MI, in October 1997. They are a convenience sample — the kids who were in the fourth grade.

Quoted from the source: "From a very young age, shoes for boys tend to be wider than shoes for girls. Is this because boys have wider feet, or because it is assumed that girls, even in elementary school, are willing to sacrifice comfort for fashion? To assess the former, a statistician measures kids' feet."

**Variables** Each case is one child.

- name — the child's first name.
- birthmonth — the month of birth
- birthyear — the year of birth
- length — Length of longer foot (cm)
- width — Width of longer foot (cm), measured at widest part of foot
- sex — boy or girl
- biggerfoot — Which foot was longer (right or left)
- domhand — Right- or left-handedness

**Sources** Mary C. Meyer (2006) "Wider Shoes for Wider Feet?" *Journal of Statistics Education* 14(1), [www.amstat.org/publications/jse/v14n1/datasets.meyer.html](http://www.amstat.org/publications/jse/v14n1/datasets.meyer.html)

## Marriage Licenses

DATA FILE  
marriage.csv

**Context** Marriage records from the Mobile County, Alabama, probate court. Records were picked quasi-randomly by Alan Eisinger.

**Variables** Each case is one person getting married. Both the bride and groom are included in the data set.

- BookpageID: The book and page in the county register on which the marriage is recorded. Used as a unique identifier of the marriage.
- Appdate: The date on which the application was filed, in m/d/y format.
- Ceremonydate: The date on which the marriage ceremony took place, in m/d/y format.
- Delay: The number of days between the application and the ceremony.
- Officialtitle: The listed title of the official who conducted the marriage.
- Person: Which of the bride or groom is represented (byperson only)
- Dob: The date of birth of the person, in m/d/y format.
- Age: The age of the person, in years and fractions of year.
- Race: The race of the person, as listed on the application.
- Prevcount: The number of previous marriages of the person, as listed on the application.
- Prevconc: The way the last marriage ended, as listed on the application.
- Hs: The number of years High School education, as listed on the application.

For review purposes only

- College: The number of years College education, as listed on the application. Where no number was listed, this field was left blank, unless less than 12 years High School was reported, in which case it was entered as 0.
- Dayofbirth: The day of birth, as a number from 1 to 365 counting from January 1.
- Sign: The astrological sign, calculated by using dayofbirth. May not correctly sort people directly on the borders between signs. This variable is not part of the original record.

**Sources** The records were collected through <http://www.mobilecounty.org/probatecourt/recordssearch.htm>

## SATs by State

**Context** These data were assembled for a statistics education journal article on the link between SAT scores and measures of educational expenditure.[10]

DATA FILE  
sat.csv

### Variables

- State — Name of state (in quotation marks)
- expend — Expenditure per pupil in average daily attendance in public elementary and secondary schools, 1994-95 (in thousands of dollars)
- ratio — Average pupil/teacher ratio in public elementary and secondary schools, Fall 1994
- salary — Estimated average annual salary of teachers in public elementary and secondary schools, 1994-95 (in thousands of dollars)
- frac — Percentage of all eligible students taking the SAT, 1994-95
- verbal — Average verbal SAT score, 1994-95
- math — Average math SAT score, 1994-95
- sat — Average total score on the SAT, 1994-95

**Sources** From Deborah Lynn Guber (1999) "Getting What You Pay For: The Debate Over Equity in Public School Expenditures" *Journal of Statistics Education* 7(2) available from <http://www.amstat.org/publications/jse/secure/v7n2/datasets.guber.cfm>

## Smoking and Death

**Context** Data on age, smoking, and mortality from a one-in-six survey of the electoral roll in Whickham, a mixed urban and rural district near Newcastle upon Tyne, in the UK. The survey was conducted in 1972-1974 to study heart disease and thyroid disease. A follow-up on those in the survey was conducted twenty years later. This dataset contains a subset of the survey sample: women who were classified as current smokers or as never having smoked.

DATA FILE  
whickham.csv

**Variables** Each case is one woman. Three variables are measured:

- Smoker — Whether the woman identified herself as a smoker during the original survey.
- Outcome — Whether the woman was alive or not at the time of the follow up survey. (Needless to say, only living women were included in the original survey.)
- Age — The woman's age at the time of the first survey.

**Sources** DR Appleton, JM French, MPJ Vanderpump, *American Statistician* 50:4, 1996, pp.340-341. I have synthesized the data set from the summary description tables given in that paper.

## Swimming World Records

**Context** World record times in the 100m freestyle swimming race.

### Variables

- time the record time, in seconds
- sex whether this is the women's or the men's record.
- year the year in which the record was set.

**Sources** Record information is widely available on the Internet. A history is available on [http://en.wikipedia.org/wiki/World\\_record\\_progression\\_100m\\_freestyle](http://en.wikipedia.org/wiki/World_record_progression_100m_freestyle)

## Ten-Mile Race

**Context** The Cherry Blossom 10 Mile Run is a road race held in Washington, D.C. in April each year. (The name comes from the famous cherry trees that are in bloom in April in Washington.) The results of this race are published. The file contains the results from the 2005 race.

DATA FILE  
ten-mile-  
race.csv

**Variables** Each case is one runner who completed the race course.

- sex —
- age —
- time — The official time from starting gun to the finish line.
- net — The recorded time from when the runner crossed the starting line to when the runner crossed the finish line. This is generally less than the official time because of the large number of runners in the race: it takes time to reach the starting line after the gun has gone off.
- state — The state of residence of the runner. This can indicate how far the runner travelled to participate in the race. Many runners come from areas close to DC in Maryland (MD) and Virginia (VA).

Another dataset, "running-longitudinal.csv," contains the results the races from 1999 to 2008 for those runners who ran in multiple years. In that dataset, each case is one runner in one year, so there are multiple cases for each runner; 41,248 cases altogether with 14,434 different runners. The variable id identifies each runner uniquely across all years. The variable nruns tells how many different time each runner participated.

**Sources** Data are from <http://www.cherryblossom.org/>

## Trucking Jobs

**Context** A dataset from a mid-western US trucking company on annual earnings of its employees in 2007. Datasets like this are used in audits by the Federal Government to look for signs of discrimination.

DATA FILE  
truckingjobs.  
csv

For review purposes only

**Variables** Each case is one employee.

- sex: M or F
- earnings: Annual earnings, in dollars. Hourly wages have been converted to a full-time basis.
- age in years
- title the job title
- hiredyears how long the employee has been working for the company.

**Sources** For reasons of confidentiality, the source of these data cannot be revealed.

## Utility Bills

**Context** Gas and electricity bills on the author's house in Saint Paul, Minnesota 55116.

DATA FILE  
utilities.csv

### Variables

- month — The number of the month on which the meter was read
- day — The day of the month on which the meter was read
- year — The year.
- temp — The average temperature during the billing period.
- kwh — The number of kilowatt hours of electricity used.
- ccf — Cubic feet of natural gas used
- thermsPerDay — The utilities calculation of therms of energy in the natural gas per day of the billing period. A therm is a measure of energy equivalent to 10,000 British Thermal Units (BTU). One BTU raises the temperature of one pound of water by one degree fahrenheit.
- dur — How many days in the billing period.
- totalbill — The total utility bill: gas + electricity
- gasbill — The amount of the gas bill during that period
- electricbill — The amount of the electric bill during that period

Note that the meter might not always have been read accurately. Records for 5/30/2000 and 3/25/2000 are outliers easily explained by having too low a first reading followed by a correct reading, giving an apparently heavy gas usage during the second period.

A similar, but smaller version of this may have happened with 1/28/2000 and 2/26/2000.

Notes indicate when changes were made to the house, for example replacing the furnace and water heater, both of which are fueled by natural gas. Cooking is fueled by both electricity (oven, microwave, appliances) and gas (stove-top, outdoor grill).

**Sources** Monthly utility bills from XCEL Energy received by the author.

## Used Car Prices

Two different files, used-hondas.csv and used-fords.csv

**Context** Prices of used Honda Accord LXs as advertised on cars.com in October 2007.  
Prices of used Ford Tauruses as advertised on cars.com in February 2009.

DATA FILE  
used-hondas.csv

For review purposes only

### Variables

- Price — The price, in US dollars, asked for the car.
- Year — The car's model year.
- Mileage — The number of miles the car has been driven.
- Location — The city where the car is located.
- Color — The car's body color.
- Age — The age of the car: 2007 – Year for the Hondas or 2009 – Year for the Fords.

DATA FILE  
used-fords.  
csv

**Sources** These data were compiled by Macalester College students for a class project. Aleksander Azarnov collected the Honda data; Elise delMas, Emiliano Urbina, and Candace Groth collected the Ford data.

### Wages from the Current Population Survey

**Context** The Current Population Survey (CPS) is used to supplement census information between census years. These data consist of a random sample of 534 persons from the CPS, with information on wages and other characteristics of the workers, including sex, number of years of education, years of work experience, occupational status, region of residence and union membership. (Quoted from [http://lib.stat.cmu.edu/datasets/CPS\\_85\\_Wages](http://lib.stat.cmu.edu/datasets/CPS_85_Wages).)

DATA FILE  
cps.csv

Questions:

- Are wages related to these other characteristics?
- Is there a gender gap in wages?

**Variables** Each case is one person.

- educ Number of years of education.
- south Indicator variable for living in a southern region: S=lives in south, NS=doesn't
- sex M or F
- exper Number of years of work experience. This was inferred from age and education
- union Indicator variable for union membership.
- wage Wage (dollars per hour).
- age Age (years).
- race Race: White (W), Nonwhite (NW).
- sector Sector of the economy: clerical, const, management, manufacturing, professional, sales, service, other
- married Marital Status: Married or Single

**Sources** Data are from [http://lib.stat.cmu.edu/datasets/CPS\\_85\\_Wages](http://lib.stat.cmu.edu/datasets/CPS_85_Wages) which sites as the original source, Berndt, ER. *The Practice of Econometrics* 1991. Addison-Wesley. The data file cps.csv is recoded from the original, which had entirely numerical codes.

Data suggested by Prof. Naomi Altman.

### Dataset Index

For review purposes only

## Dataset Index

College courses, 413  
College grades, 241, 412  
Current Population Survey, 321, 332, 335  
  
Galton's height data, 51, 58, 60, 94, 99, 128, 133, 190, 246, 289, 307, 318, 413  
Grade point, 413  
  
Kids' foot measurements, 158, 173, 266, 414  
  
Marriage licenses, 293, 319, 414  
  
SATs by state, 104, 152, 189, 196, 205, 244, 415  
Swimming world records, 40, 77, 89, 110, 122, 137, 261, 263, 341  
  
Ten-mile race, 38, 237, 248, 416  
Trucking Jobs, 148, 243, 416  
  
Used car prices, Ford Taurus, 120, 418  
Used car prices, Honda Accord, 145, 417  
Utility bills, 55, 67, 84, 112, 130, 417  
  
Wages from CPS, 71, 84, 99, 101, 121, 137, 138, 156, 168, 192, 204, 258, 310, 321, 418  
Whickham smoking data, 346, 357, 415

For review purposes only

## R Operator Index

For review purposes only



## R Operator Index

- ==, 24
- =, 17
- \$, 24
- acos, 192
- angle\*, 174
- anova, 341
  - in logistic regression, 359
- arithmetic, 19
  - on collections, 26
- as.character, 22
- as.factor, 111
- as.numeric, 123
- auto.key=, 87
- barchart, 66
- bwplot, 53
- bwplot, 65, 85
- c, 26
- coef, 110, 191
- collections
  - c, 26
- columns
  - in data frame, 24
- comparison
  - on collections, 26
  - operators, 25
- Confidence Intervals, 261
  - confint, 262
  - interval=
    - in predict, 266
- confint, 262
- conversion
  - numeric to character, 23
- cor, 137, 192
- cos, 192
- cross\*, 111
- CSV files
  - reading, 39
- Data
  - access variables with \$, 24
  - as.factor, 111
  - as.numeric, 123
  - cross\*, 111
  - data= in lm, 85
  - edit, 159
  - importing to R, 39
  - ISMdata\*, 39
  - is.na, 26
  - levels, 161
  - missing (NA), 26
  - na.omit, 126
  - na.rm=, 125
  - outlier\*, 66
  - read.csv, 39
  - small collections with c, 26
  - subset, 42
  - subtracting out mean, 192
  - summary, 61
- data= in lm, 85
- data, 23
- date, 17
- density, 53
- densityplot, 64, 87, 284
- Descriptive Statistics
  - cor, 137, 192
  - mean, 60
  - median, 60
  - outlier\*, 123
  - pdata, 60
  - qdata, 60
  - sd, 60
  - summary, 61
  - table, 24, 65
  - var, 137
- do\*, 264
- \$ and predict, 161
- dot product
  - computing, 174
- edit, 159, 160
- equations\*, 225
- exp, 20
- FALSE, 25
- fitted, 89, 122, 158, 191

For review purposes only

- in logistic regression, 358
- function, 28, 264
- Geometry
  - acos, 192
  - angle\*, 174
  - cos, 192
  - dot product, 174
  - projecting with `lm`, 191
  - radians & degrees, 192
  - square length, 173
  - sum of squares, 61
- `glm`, 356
- histogram, 52, 62
- Hypothesis Testing
  - anova, 341
  - `hypothesis.test.panel*`, 284
  - `pf`, 310
  - rank, 343
  - `resample*`, 308
  - `shuffle*`, 308
  - summary
    - for linear models, 309, 312
    - for logistic regression, 357
- `hypothesis.test.panel*`, 284
- I for transformation terms, 91
- Importing data to R, 39
- Inf, 21
- interval=
  - in predict, 266
- IQR, 56
- `is.na`, 26
- ISM.Rdata, 14, 30
- ISMdata, 39
- key in graphics, `auto.key=`, 87
- labels
  - rotating, 86
- length, 27
- length
  - vector dimension, 173
- min, 24
- levels, 161
- `lm`, 88, 110, 122, 191
- log, 20
- max, 24, 61
- mean
  - subtracting out, 192
- mean, 24, 60
- median, 60
- min, 61
- 1, 90
- missing data, 125
  - `is.na`, 26
- Modeling
  - `lm`, 191
  - : and \*, 90
  - `coef`, 110, 191
  - `data=` in `lm`, 85
  - fitted, 89, 122, 191
    - in logistic regression, 358
  - `glm`, 356
  - I for transformation terms, 91
  - `lm`, 88, 110, 122
  - predict, 266
    - in logistic regression, 358
  - predict and \$, 161
  - `resid`, 122, 191
  - sd of residuals, 138
  - summary
    - for logistic regression, 357
    - suppressing the intercept, 90
- regression report with summary, 110
- NA, missing data, 26
- `na.omit`, 126
- `na.rm=`, 125
- NaN, 21
- numerical comparison, 25
- outlier\*, 66, 123
- `pdata*`, 60, 285
- percentile
  - p-functions, 224
- `pf`, 310
- Plotting & Graphics
  - axis limits, 86
  - `barchart`, 66
  - `bwplot`, 65, 85
  - `densityplot`, 64, 87
  - histogram, 62

- rotating labels, 86
- rug, 51
- xlab= & ylab=, 85
- xyplot, 85
- predict, 160
- predict, 160, 266
  - in logistic regression, 358
- Probability
  - p operators, 224
  - q operators, 223, 224
  - random generators, 221
- qdata\*, 60, 285
- qr, 206
- quantile, 60
- quantiles
  - q-functions, 223
- R Syntax
  - ==, 24
  - arithmetic, 19
  - assignment =, 17
  - character text, 23
  - function, 28
  - Inf & NaN, 21
  - TRUE & FALSE, 25
- R System
  - ISM.Rdata workspace, 14, 30
  - saving your work, 29
  - source, 29
- radians to degrees, 192
- range, 61
- rank, 343
- rbinom, 221
- rchisq, 221
- read.csv, 39
- reading CSV files, 39
- Regression report with summary, 110
- resample\*, 42, 219, 308
- Resampling
  - resample\*, 42, 219, 308
  - shuffle\*, 42, 308
- resid, 122, 191
- rexp, 221
- rf, 221
- rlnorm, 221
- rnorm, 221
- rpois, 221
- rt, 221
- rug, 51
- run.sim\*, 225
- runif, 221
- Saving your work in R, 29
- scatter plots, making, 85
- sd, 60
  - of residuals, 138
- seq, 16
- shuffle\*, 42, 308
- Simulations
  - coin flip, 219
  - equations\*, 225
  - of heights, 225
  - run.sim\*, 225
- source, 29
- sqrt, 20
- square length
  - computing, 173
- subset, 42
- sum, 27
- sum of squares, 61
- summary, 61
  - for lm, 110
  - for linear models, 309, 312
  - for logistic regression, 357
- svd, 206
- table, 24, 65
- trigonometry, 20
- TRUE, 25
- var, 60, 137
- xlab= for axis label, 85
- xlim= for axis limits, 86
- xyplot, 85
- ylab= for axis label, 85
- ylim= for axis limits, 86

Items in typewriter font are R operators or arguments.

- 2 A \* identifies functions defined in the ISM.Rdata workspace, not standard R.

For review purposes only

## Concept Index

For review purposes only

## Concept Index

- absolute frequency, 52, 53
- accuracy, 237, 260
  - computer arith, 20
- achieved significance level, *see* p value
- adjustment, 98, 148, 150, 154
  - and t-test, 314
  - computation, 156
  - faulty, 156
- aggression, 133, 306
- algebraic notation
  - vs computer, 19
  - vs model notation, 84
- algorithm, 15
- alternative hypothesis, 279, 339
  - definition, 281
- analysis
  - unit of, 34
- analysis of covariance, 321
- analysis of variance, *see* ANOVA
- analysis of variance table, 311
- Anderson, Michael, 5
- angle, 164
  - and correlation, 187
  - between random vectors, 227
  - between vectors, 164
  - computing, 174
  - converting radians to degrees, 192
  - degrees and radians, 174
  - vector, 172
- ANOVA
  - and collinearity, 325, 335
  - and sample size, 338
  - and sum of squares, 322
  - compare two models, 342
  - covariates, 314
  - one-way, 310
  - order dependence, 323
  - report, 296
  - table, 296, 311, 318
  - term-by-term, 319
  - vs regression report, 335
- antibiotics, 387
- aprotinin (drug), 141
- arithmetic
  - accuracy, 20
  - on collections, 26
  - R notation, 19
  - with vectors, 114
- aspirin, 392, 394
- assignment
  - computer value, 17
- astrology, 293, 295
- athletic ability, 371
- Avogadro constant, 20
- axis (in graphics)
  - limits in graphics, 86
  - labels, 85
- backdoor pathway, 375
- balanced assignment, 394
- bar charts, 66
- Bayesian, 209
- bell shaped distribution, *see* normal
- bias
  - coefficients, 260
  - non-response, 36
  - sampling, 36
  - self-selection, 36
- binomial model, 210
- Blake, William, 383
- blind experiment, 387
- block pathways, 375
- blocking in experimental design, 393
- blood, 48
- Bonferroni correction, 304, 307, 335
- boolean, *see* logical
- bootstrapping, 251, 265, 352
  - residuals, 251
- box and whisker plots, *see* box plots
- box plot, 53
  - drawing, 65, 85
  - outliers, 65
- Box, George, 3
- Braque, Georges, 67
- Brown, Robert, 231
- Brownian motion, 231

For review purposes only

- campaign spending, 367, 370, 375, 402
- cancer, 354
- car prices, 120
- Cartesian coordinates, 165
- case space, 178
- cases, 32
- categorical variable, 56, 168
  - crossing, 111
  - from quantitative, 111
  - in model formula, 95
  - levels, 56
  - reference level, 97
  - unlikeability, 59
- causal loops, 374
- causation, 74, 362
  - and covariates, 375
  - and observational studies, 394
  - and regression, 335
- cause and effect, 7
- census, 35, 44
  - precision, 9
  - vs sample, 43
- center of distribution, 57
- centering, 192
- ceteris paribus, 142
- Chambers, John, 12
- character string
  - comparison, 24
  - object type, 22
  - object type), 23
- chi-squared distribution, 353
- classification
  - using models for, 6
- climate change, 118, 119
- Cobb, George, x
- code book, 33
- coef. of determination, 127, 199
  - $R^2$ , 127
  - and F, 290
  - computing, 137, 192
  - in variable space, 199
- coefficients, 93–110
  - as slopes, 101
  - comparing with different units, 103
  - computing best fitting, 110
  - fitting, 113
  - interpreted in context, 104
  - NA (redundancy), 124
  - sign, 97
  - Simpson's paradox, 106
  - t-test, 334
  - uncertainty, 139
  - units, 102
- coin flip, 207
  - simulation, 219
- Collegiate Learning Assessment (test), 336
- collinearity, 200–202, 258
  - and ANOVA, 325
  - and correlation, 129
  - and interaction terms, 205
  - and p-value, 335
  - and Simpson's paradox, 201
  - and standard errors, 254, 257
  - effects of, 201
  - measuring, 328
- columns
  - in data frames, 24
- comma separated value, 39
- command, 12
- comparing model designs, 116
- comparison operators, 24
- complex numbers, 21
- compliance
  - experimental, 392
- computation, 15
  - connecting, 18
  - foundations, 11
- conditional probability, 278, 354
- conditioning
  - on explanatory vars., 69, 70
- confidence interval, 237, 242
  - computation, 262
  - conf. level, 247
  - for prediction, 266
  - for small samples, 244
  - on model value, 246
  - on prediction, 246
- confidence level, 247
- confounding, 142
  - latent variables, 398
- contrapositive, 273

- control, 383
  - experiment, 8, 383
  - negative, 389
- convenience sample, 36
- conventional wisdom, 276
- correlating pathway, 373, 375
- correlation
  - accidental, 387
  - and nonlinear relationships, 130
  - coefficient, 127–137, 173
    - computing, 137, 192
  - collinearity, 129
  - destroying, 397
  - experimental variables, 386
  - in variable space, 187
  - sign of, 129
  - vs causation, 364
- cosine, 172
- counterfactual, 8, 363, 366
- counting
  - and table operator, 24
- counts
  - categorical variable, 56
  - computing, 65
- covariate, 142, 153–154, 317
  - adjustment for, 148
  - ANOVA, 314
  - causation, 375
  - in hypothesis tests, 318
  - leaving out, 156
  - propensity score, 402
- coverage interval, 45
  - computing, 222
  - level, 222
  - quantiles, 60
- cross-sectional, 38, 337
  - sampling, 37, 74
- crossing categorical variables, 111
- CSV, 39
- Current Pop. Survey, 321, 332, 335
- curvature
  - and transformation terms, 82
- Darwin, Charles, 298
- Darwin, Erasmus, 297
- data, 8
  - and causation, 74
  - character, 22
  - character), 23
  - missing, 26
  - selecting subsets, 24
- data collection
  - planning for, 75
- Data files
  - on Internet, 39
  - reading with ISMdata, 39
- data frame, 23–34
  - adding a new variable, 41
  - columns of, 24
  - object type, 22
  - taking subsets, 42
- deductive reasoning, 272
- degrees of freedom, 203
  - and subspace, 203
  - in denominator, 293
  - in numerator, 293
- degrees vs radians, 174
- denominator
  - deg. freedom, 293
- density, 53, 64
  - units, 53
- density plot, 64
  - drawing, 64
  - vs histogram, 53
- dependent variable, 70
  - see response var., 70
- derivative
  - partial, 81, 98
- description
  - using models for, 6
- descriptive statistics
  - computing, 60–61
- design of models, 75
- deviance, 49, 353
- deviation, 49
- diagnosis, 7
- dimension, 164
- direction, 164
- discrimination, 378
  - sex, 5
  - wage, 5
  - wages, 71
- distribution

For review purposes only



- bell-shaped, 53
- box plot, 53
- center and spread, 57
- normal, 53
- shape of, 53–56
- skew, 56
- dividing into groups, 155
- dose-response, 389
- dot product, 171
  - computing, 174
- double blind, 387
- Dow Jones Ind. Avg., 281
- drawing vector diagrams, 195
- dummy variable, *see* indicator variable
- earthquakes, 216
- Ecclesiastes, 207
- efficiency, 155
- Einstein, Albert, 232, 317, 383
- El Niño Southern Oscillation, 119
- electron mass, 8
- empirical, 134
- endogenous, 368
- equal-variance t-test, 313, 314
- equiprobability, 219
- error, 251
  - obsolete term., 49
- ethics, 385
- evidence
  - strength, 134, 135
- Excel, 13, 39
- exogenous, 225, 368
- experiment, 141, 156
  - vs observation, 141
  - and blocking, 392, 393
  - and comparison, 383
  - and orthogonality, 392
  - balanced assignment, 394
  - blind, 387
  - design, 331, 390
  - homogeneity, 386
  - intent, 392, 395
  - intervention, 391
  - randomization, 393
  - subject, 384
  - unit, 384, 386
  - units, 384
  - variable, 384
  - vs adjustment, 156
- experimentation, 370
- explanatory variables, 43, 69, 75, 76
  - conditioning on, 69
- exponential prob., 216
- external validity, 35
- extrapolation, 101
- extremes
  - disadvantages of, 43
- F
  - and  $R^2$ , 290, 297
  - deg. freedom, 293
  - formula, 291
  - shape of distr., 294
  - significance, 305
  - statistic, 291
- factor, 41
- fail to reject the null, 280
- Fiji, 31
- finger length, 133, 306
- Fisher, Ronald, 289, 291
- fitted model values
  - computing, 89, 122
- fitting, 94
  - criteria, 116
  - lm operator, 88
  - software, 122
- flooding, 246
- Freedman
  - David, 304
- frequency, 52, 53
- frequentist, 208
- function, 68
- Galileo Galilei, 141, 195
- Galton, Francis, 43, 298
- gaussian distribution, *see* normal
- genetics of height, 297
- Geometry
  - degrees vs radians, 174
- global warming, 363
- Gould, Stephen J., 43
- grade-point average, 241
- Grand Forks, ND, 246
- grand mean, 99

- hypothesis test, 302
- graphics
  - axis labels, 85
  - axis limits, 86
  - limits of, 93
  - multiple variables, 87
  - rotating labels, 86
  - statistical, 62–66
  - two-variable, 85
- gravity, 20
- greenhouse gases, 363
- group means, 99
- groupwise slopes, 102
- Hamming, Richard, ix
- heart surgery, 141
- height
  - Galton's data, 43
  - genetics, 297
- hiring discrimination, 305
- histogram, 52
  - bin boundaries, 53
  - drawing, 62
  - vs density plot, 53
- holding constant, 98, 153
- Holmes, Oliver Wendell, 93
- home ownership, 202
- Homer, 177
- Horace, 258
- Hughes, Robert, 8
- Hume, David, 275, 287
- hypernatremia, 48
- hypothesis testing, 269
  - alternative hypothesis, 279
  - covariates, 318
  - glossary, 281
  - grand mean, 302
  - group means, 301
  - in regression report, 309
  - logic, 269–281
  - multiple comparisons, 303
  - non-parametric, 337
  - null hypothesis, 275
  - one-tailed test, 277
  - outcomes, 281
  - p value, 276
  - power, 280
  - sample size, 280
  - significance level, 278
  - slopes, 301
  - test statistic, 275
  - Type I error, 278
  - Type II error, 279
- hypothetical causal network, 367, 369
- imaginary numbers, 21
- in-sample, 304
- incomes, 216
- independent, 209
- independent variable
  - see explanatory var., 70
- indicator variable, 97, 168
  - arithmetic, 175
- inductive
  - vs deductive, 404
- inductive reasoning, 274
- infinity
  - INF on computer, 21
- instrumental variable, 397, 398, 400
- intent to treat, 395, 396
- interaction, 168
  - multi-collinearity, 81
  - symmetry between variables, 80
- interaction term, 77, 80, 95
  - as vectors, 169
  - computing model values, 111
  - in ANOVA, 342
  - interpreting, 96
  - multi-collinearity, 205
  - partial derivatives, 81
  - slopes, 102
- intercept, 77
  - grand mean, 99
  - groupwise, 101
  - importance of including, 83
  - suppressing, 82, 90
  - term, 78, 167
  - vector, 114, 167
    - in variable space, 200
- Internet
  - data files on, 39
- interpreter, 16
- interquartile interval, *see* IQR
- intervention

For review purposes only

- and causation, 362
  - using models for, 7
- inverse percentiles, 213
- invoking an operator, 29
- IQ, 215, 371
- IQR, 56
- ISM.Rdata, 14, 30
- Karlin, Samuel, 113
- Landers, Ann, 36
- Lao-Tse, 235
- latent variable, 368, 398
  - proxy, 391
- least absolute value residual, 116
- least squares, 114–116
- least worst-case residual, 116
- left skew, 56
- length, 164
  - Pythagorean formula, 173
  - vector, 172
- levels, 33
  - categorical variable, 56
- Levitt, Steven, 402
- light, speed of, 8
- likelihood, 352
- linear combination, 167, 170
- linear dependence, 203
  - and redundancy, 203
- linear models, 95
  - as approximation, 106
  - vs logistic, 347
- link value, 350
  - in logistic regression, 358
- Lippman, Walter, 31
- log-odds ratio, 356
- logical
  - comparison operator, 24
  - data type, 24
  - object type, 21
  - operators, 24–26
- logistic regression, 76, 350
  - fitting, 356
  - propensity score, 402
- logistic transformation, 350
- lognormal probability model, 215
- longitudinal, 38, 337
  - sampling, 37, 74
- Lumiere Brothers, 364
- Macalester College, 1
- main terms, 77, 79
- margin of error, 260
- marriage, 293
  - license, 319
- Mars Climate Orbiter, 20
- matched sampling, 397, 398
- matrix, 169
- maximum likelihood estimation, 353
- mean, 47, 215
  - estim. by eye, 57
  - geometrically, 183
  - groupwise, 99
  - intercept coef., 99
  - standard error, 241
  - subtracting out, 192
- mean square, 50, 297
- median, 47
  - box plot, 53
  - robust against outliers, 56
- missing data, 26, 125
- model
  - causation, 362
  - definition, 6
  - random, 230
  - uses, 6
- model coefficient, 94
  - in variable space, 198
  - slope, 73, 94
  - standard error, 239
- model design, 75
  - comparing, 116
  - multi-collinearity, 122
  - notation, 83
  - process, 113
  - strategy for, 108
- model fitting, 113
  - geometry, 177
  - lm operator, 88
- model formula, 94
  - categorical variable, 95
  - coefficients, 93–110
- model function, 69
- model notation

- vs algebra, 84
- model probabilities, 354
- model terms, 76–83
  - “shape”, 77
  - interaction, 95
- model triangle, 185
- model value, 69
  - calculating, 111
  - mean, 99
- model vectors, 167
- monotonicity, 82
- motion, Brownian, 231
- multi-collinearity, 81, 122, 203
  - and ANOVA, 325
  - approximate redundancy, 121
  - correcting, 122
- multiple tests
  - Bonferroni correction, 335
- mutatis mutandis, 144
- NA, missing data, 26
- NAN, 21
- National Weather Service, 246
- natural gas, 67
- negative control, 389
- nested models, 133
  - and  $R^2$ , 132
- net distance
  - random walk, 229, 232
- neutrinos, 9
- nitrogen fixing plants, 5
- Nobel Prize, 232
- non-causal links, 368
- non-correlating pathway, 373, 375
- non-parametric statistics, 337, 338, 343
- non-response, 36
- nonlinear relationships
  - and correlation, 130
- normal
  - distribution, 53, 55
  - not necessarily the norm, 55
  - obsolete term., 49
  - probability model, 215
- notation for model design, 83
- nuisance parameter, 271
- null deviance, 353
- null hypothesis, 275, 276
  - and population, 287
  - definition, 281
  - fail to reject, 280
  - plotting distr., 284
- numbers
  - complex, 21
  - INF and NaN, 21
- numerator
  - deg. freedom, 293
- object
  - basic types in R, 21
- objectivity, 4
- objects, in R, 17
- odds, 355
- odds ratio, 355, 356
- Odysseus, 177
- OFCCP, 305
- one-sample t-test, 302, 314
- one-tailed, 277
- one-way analysis of variance, 300, 310
- order dependence in ANOVA, 323
- ordinal, 33
- orthogonality, 188
  - creating, 392
  - in experiment, 393
  - randomization, 393
  - through instrumental variables, 399
  - through matched sampling, 398, 402
- out-of-sample, 304
- outcome set, 207
- outlier, 56
  - and non-parametrics, 343
  - and rank transform, 338
  - in box plots, 65
  - removing, 66
- p value, 135, 276, 277
  - adjusting, 304
  - computing, 283
  - definition, 282
  - from F statistic, 297, 310
  - multiple tests, 303
  - two-tailed, 283
  - visualizing, 298
- paired t-test, 315, 321

For review purposes only

- panel data, 402
- parameters, 210
  - binomial model, 211
  - exponential probability model, 216
  - F distribution, 294
  - lognormal probability model, 215
  - normal probability model, 215
  - poisson model, 212
  - uniform distribution, 214
- parsec, 21
- partial derivative, 81, 98
- partial derivatives, 148
- partial difference
  - computing, 158–161
- partial relationship, 141, 143, 146–148
- partitioning
  - Pythagorean theorem, 118
  - sums of squares, 123
  - variation, 43, 70, 117
    - correlation coefficient, 127
- pathway, 373
  - backdoor, 375
  - blocking, 377, 379, 391
  - experimental intervention, 391
  - rules for blocking, 375
  - sampling variable, 377
- Pearl, Judea, 375
- percentile, 45, 213
  - and coverage interval, 45
  - box plot, 53
  - computing, 46, 224
  - quantile operator, 60
- permutation test, 288
  - computing, 308
  - history, 289
- perpendicular, 185
- Perrin, M. Jean, 232
- Picasso, Pablo, 3
- Pisa, Leaning Tower, 141
- placebo, 386, 387
- placebo effect, 387
- Plotting & Graphics
  - adding a key, 87
- Poincaré, Henri, 67
- Poisson, 212
- poisson model, 212
- polio vaccine, 387
- population, 35
  - and null hypothesis, 287
  - parameters, 245
- positive control, 389
- power, 279, 338, 340, 390
  - and sample size, 280
  - computing, 280
  - definition, 282
  - experimental design, 390
  - F statistic, 341
  - sample size, 338
- precision, 3, 8, 236, 260
- prediction, 138
  - confidence interval, 246, 266
  - from models, 73
  - interval, 246
  - using models for, 6
- prevalence, 355
- probability, 208
  - subjective, 209
  - with linear models, 347
- probability calculus, 209
- probability model, 208
  - binomial, 210
  - exponential, 216
  - lognormal, 215
  - normal, 215
  - parameters, 210
  - poisson, 212
  - sampling from, 220
  - uniform, 214
- probability values
  - in logistic regression, 358
- projecting (a vector), 180
- propensity score, 402
- prospective studies, 304
- prostate cancer, 354, 356
- protractor, 189
- Proxima, 21
- proxy, 391
- Pythagorean theorem, 118, 186
  - vectors, 171
- quantiles, 213, 222
  - and coverage intervals, 60
  - computing, 223

- quantitative variable, 168
  - dividing into groups, 153
  - to categorical, 111
- quartile, 47
- quintiles, 47
- R software
  - arithmetic notation, 19
  - installation, 14
  - mathematical functions, 20
  - names of objects, 17
  - operator, 15
- $R^2$ , 127
  - and F, 297
  - nested models, 132
  - unitless, 129
- race
  - and home ownership, 202
  - and wages, 322, 335
- radians, 174
- random angles, 227
- random assignment, 392, 395
- random models, 230
- random number generation, 37
- random sample, 37
- random variable, 210
- random variation, 384
- random vectors
  - p value, 298
- random walk, 231
  - model walk, 291
  - and F, 291
- randomization, 393
- range of extremes, 44
- rank transform, 338
- rate, poisson model, 212
- rates and partial relationships, 145
- recurrent networks, 374
- redundancy, 113, 121, 170, 203
  - and linear dependence, 203
  - and NA, 124
  - fitted model values, 124
  - in categorical variables, 125
- reference group, 99
- reference level, 97, 99
- regression
  - and causation, 335
  - regression report, 240
    - computing, 110, 122
    - hypothesis testing, 309
- relationship
  - as function, 68
  - partial vs total, 141
- relative frequency, 53
- relevance, 385
- repeated measures, 321
- replication, 389
- representative sample, 35
- resampling, 219, 248
  - distribution, 240, 248
  - with replacement, 42
- residual, 49, 69, 70, 98, 182
  - and prediction, 138
  - and standard error, 257
  - as vector, 182
  - computing, 122
  - estimating size of, 253
  - least squares, 115
  - outliers, 123
  - randomizing, 251
  - standard deviation, 187
  - standard error, 139
  - sum of squares, 116
  - too small, 116
  - typical size, 138
  - vector, 115
- response variable, 43, 69, 70, 75
  - choice of, 75
- retrospective studies, 304
- right skew, 55, 56
- robust statistics
  - IQR, 56
  - median, 56
- roulette, 218
- rounding, 20
- rug plot, 51
- Rutherford, Ernest, 3
- Sagan, Carl, 269
- Salk vaccine, 387
- sample, 35, 44
  - representative, 35
- sample size, 35
  - and standard error, 257

For review purposes only

- hypothesis testing, 280
  - power, 341, 390
- sample space, 207
- sampling
  - with replacement, 42
  - cross-sectional, 37
  - distribution, 237, 239
  - frame, 35
  - longitudinal, 37
  - random, 37
  - variable, 377
  - with replacement, 42, 248, 249
  - with resample, 42
  - with shuffle, 42
- sampling bias, 36
- SAS, 13
- SAT, 104, 195, 200, 244, 337
- scale a vector, 164
- scaling a vector, 167
- scatter plot, 72, 85
  - and case space, 178
- school spending, 104
- scientific notation, 20, 124
- Scott, Sir Walter, 227
- screening tests, 305, 331
- sea shells, 31
- self-selection bias, 36
- Seneca, 258
- sex discrimination, 5
- sham surgery, 387
- side effects, 16, 30
- sign
  - of coefficients, 97
- significance level, 278
  - definition, 282
- significance vs substance, 305
- simple random sample, 35
- Simpson's paradox, 106, 151, 202, 370
  - and collinearity, 201
- simulation, 224, 370
  - heights, 225
  - run . sim\*, 225
- skew, 56
- slope
  - and interaction, 102
  - and interaction terms, 80
  - and model coefficients, 73
  - coefficients, 101
  - groupwise, 102
  - model coefficients, 94
- smoking, 346
- sodium, 48
- software
  - prime directive, 12
- source file, 29
- space (vector), 165
- spread of distribution, 57
- spreadsheet software, 39
- SPSS, 13
- square length, 172
  - computing, 173
  - random walk, 232
- standard deviation, 47–51, 215
  - doesn't partition variation, 117
  - estim. by eye, 57
  - geometrically, 183
  - in variable space, 186
  - of residuals, 138
- standard error, 49, 239
  - formula for, 256
  - of residuals, 139
  - of the mean, 241
- stanines, 47
- STATA, 13
- state, 30
- statistical graphics, 62–66
- statistical modeling, 3
- statistical significance, 269
- statistics
  - a definition, 10
- stock prices, 233, 281
- straight-line model, 94, 101
- strength
  - of evidence, 135
  - of relationship, 135
- structural equation modeling, 332
- subjective, 10
- subjective probability, 208, 209
- subjectivity, 4
- subsets of data frames, 42
- subspace, 179–185
  - and linear dependence, 203

- degrees of freedom, 203
- projecting into, 180
- sum of squares, 50, 173, 322
  - computing, 61, 123
  - partitioning, 123
  - residual, 116
- surgery
  - sham, 387
- syllogism, 272
- syntax
  - function definition, 28
  - in R, 16
- t-test, 300
  - and adjustment, 314
  - computing, 314
  - on coefficients, 334
  - one-sample, 302, 314
  - paired, 315
  - unequal variance, 301, 313
- table, 32
  - constructing in R, 24
  - of counts, 65
- test statistic, 271, 275
  - definition, 281
- time, atomic clocks, 3
- total distance
  - random walk, 232
- total relationship, 141, 144
- transferability, 35
- transformation term, 77, 81
- transformation terms, 81, 91
- treatment, 384
- trial, 263
- Tufte, Edward, 361
- two-sample t-test, 311
- two-stage least squares, 402
- Type I error, 278
  - definition, 282
- Type II error, 279
  - definition, 282
- type of object, 21
- typical variation, 44
- unlikeability, 59
- unequal variance t-test, 300, 313, 314
- uniform probability model, 214
- unit of analysis, 34
- units
  - $R^2$  is unitless, 129
  - coefficients, 102
  - density, 53
- used cars, 120
- utility bill data, 67
- variable, 32
  - adding to a data frame, 41
  - latent, 398
  - removing outliers, 66
  - types, 32–33
- variable space, 178
  - and coef. of determination, 199
  - and model coefficients, 198
- variance, 47–51
  - partitioning, 127
  - Pythagorean theorem, 118
  - sqrt of standard deviation, 60
- variation
  - components, 10
  - histogram, 52
  - partitioning, 43, 117
  - partitioning by ANOVA, 296
  - rug plot, 51
- vector, 114, 164–173
  - addition, 165, 167
  - and correlation, 187
  - arithmetic, 114
  - diagrams, 189
  - drawing, 195
  - dot product, 171
  - for categorical variables, 168
  - interaction terms, 169
  - length, 172
  - projection, 180
  - quantitative variables, 168
  - redundancy, 170, 203
  - residual, 182
  - scaling, 164
  - space, 165
    - dimension, 164
  - variable space, 178
- von Neuman, John, 113
- wage, 321, 332, 335

For review purposes only



- adjustment for other vars, 258
- and race, 322, 335
- discrimination, 5
- gap, 71
- Wilber, Charles Dana, 361
- Will, George, 104
- workspace, 30
  
- y intercept, 94
- yes/no, 345
  
- z-score, 217
- zero, 124

For review purposes only

For review purposes only