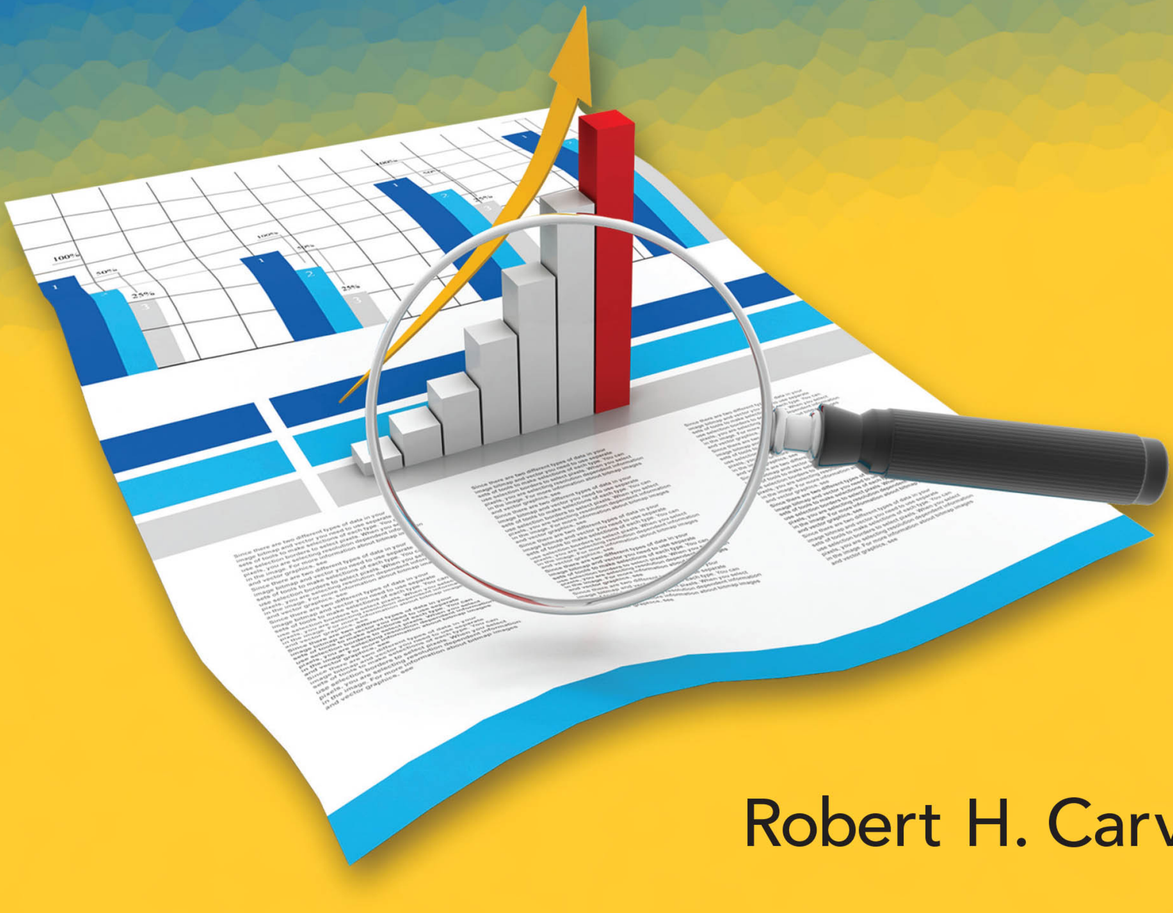
Practical Data Analysis with JMP®

Third Edition



Robert H. Carver

The correct bibliographic citation for this manual is as follows: Carver, Robert. 2019. *Practical Data Analysis with JMP®, Third Edition*. Cary, NC: SAS Institute Inc.

Practical Data Analysis with JMP®, Third Edition

ISBN 978-1-64295-614-6 (Hardcover)
ISBN 978-1-64295-610-8 (Paperback)
ISBN 978-1-64295-611-5 (Web PDF)
ISBN 978-1-64295-612-2 (EPUB)
ISBN 978-1-64295-613-9 (Kindle)

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

October 2019

Contents

About This Book

What Does This Book Cover?

Purpose: Learning to Reason Statistically

We live in a world of uncertainty. Today more than ever before, we have vast resources of data available to shed light on crucial questions. But at the same time, the sheer volume and complexity of the "data deluge" can distract and overwhelm us. The goal of applied statistical analysis is to work with data to calibrate, cope with, and sometimes reduce uncertainty. Business decisions, public policies, scientific research, and news reporting are all shaped by statistical analysis and reasoning. Statistical thinking is an essential part of the boom in "big data analytics" in numerous professions. This book will help you use and discriminate among some fundamental techniques of analysis, and it will also help you engage in statistical thinking by analyzing real problems. You will come to see statistical investigations as an iterative process and will gain experience in the major phases of that process.

To be an effective analyst or consumer of other people's analyses, you must know how to use these techniques, when to use them, and how to communicate their implications. Knowing how to use these techniques involves mastery of computer software like JMP. Knowing when to use these techniques requires an understanding of the theory underlying the techniques and practice with applications of the theory. Knowing how to effectively communicate with consumers of an analysis or with other analysts requires a clear understanding of the theory and techniques, as well as clarity of expression, directed toward one's audience.

There was a time when a first course in statistics emphasized abstract theory, laborious computation, and small sets of artificial data—but not practical data analysis or interpretation. Those days are thankfully past, and now we can address all three of the skill sets just cited.

Scope and Structure of This Book

As a discipline, statistics is large and growing; the same is true of JMP. One paperback book must limit its scope, and the content boundaries of this book are set intentionally along several dimensions.

First, this book provides considerable training in the basic functions of JMP 15. JMP is a full-featured, highly interactive, visual, and comprehensive package. The book assumes that you have the software at your school or office. The software's capabilities extend far beyond an introductory course, and this book makes no attempt to "cover" the entire program. The book introduces students to its major platforms and essential features and should leave students with sufficient background and confidence to continue exploring on their own. Fortunately, the Help system and accompanying manuals are quite extensive, as are the learning resources available online at http://www.jmp.com.

Second, the chapters largely follow a traditional sequence, making the book compatible with many current texts. As such, instructors and students will find it easy to use the book as a companion volume in an introductory course. Chapters are organized around core statistical concepts rather than software commands, menus, or features. Several chapters include topics that some instructors might view as "advanced"—typically when the output from JMP makes it a natural extension of a more elementary topic. This is one way in which software can redefine the boundaries of introductory statistics.

Third, nearly all the data sets in the book are real and are drawn from those disciplines whose practitioners are the primary users of JMP software. Inasmuch as most undergraduate programs now require coursework in statistics, the examples span major areas in which statistical analysis is an important path to knowledge. Those areas include engineering, life sciences, business, and economics.

Fourth, each chapter invites students to practice the habits of thought that are essential to statistical reasoning. Long after readers forget the details of a particular procedure or the options available in a specific JMP analysis platform, this book may continue to resonate with valuable lessons about variability, uncertainty, and the logic of inference.

Each chapter concludes with a set of "Application Scenarios," which lay out a problem-solving or investigative context that is in turn supported by a data table. Each scenario includes a set of questions that implicitly require the application of the techniques and concepts presented in the chapter.

New in the Third Edition

This edition preserves much of the content and approach of the earlier editions, while updating examples and introducing new JMP features. As in the second edition, there are three review chapters (Chapters 5, 9, and 17) that pause to recap concepts and techniques. One of the perennial challenges in learning statistics is that it is easy to lose sight of major themes as a course progresses through a series of seemingly disconnected techniques and topics. Some readers should find the review chapters to be helpful in this respect. The review chapters share a single large data set of World Development Indicators, published by the World Bank.

The scope and sequence of chapters is basically the same as the prior edition. There is some additional new material about the importance of documenting one's work with an eye toward reproducibility of analyses, as well as production of presentation-ready reporting. The second edition was based on JMP 11, and since that time, platforms have been added or modified, and some functionality has relocated in the menu system. This edition captures those changes.

Some of the updated data tables are considerably larger than their counterparts in earlier editions. This creates the opportunity to demonstrate methods for meaningful graphs when data density and overplotting become issues. I also use some of the larger data tables to introduce machine learning practices like partitioning a data set into training and validation sets.

JMP Projects are introduced in Chapter 2 and used throughout the book. Projects are a way to organize, preserve, and document multiple analyses using multiple data tables. They naturally support a logical and reproducible workflow. Using projects is a way for newcomers to establish good habits and for JMP veterans to be more efficient.

Other additions and amendments include:

- Early introduction of more data types, Header Graphs, and JMP Public.
- Expanded use of Subset, Global and Local Data Filters and Animate. In the prior editions, for example, the set of data tables included some subsets of larger tables. Because data preparation is such an important part of the analytical cycle, readers learn to perform filtering and subsetting functions on their own.
- The Recode command has evolved since JMP 11, as have the lessons using Recode. Readers will learn why and how to recode a column.
- In the Regression chapters, coverage of the Profiler has expanded, and I have added the Partition Platform to the discussion of variable selection. The Fit Curve platform also makes its first appearance, as do temporary variable transformations.
- For JMP Pro users, there is a brief treatment of the Formula Depot to facilitate comparison of models.
- In Chapter 21 on Design of Experiments, we meet Definitive Screening Designs.
- In Chapter 22, Variability Charts have been added.
- Simulators and calculators previously supplied as JSL scripts in earlier editions have been bundled among JMP's teaching demonstrations in the Help system. The text now reflects this very useful change.

Is This Book for You?

Intended Audience

This book is intended to supplement an introductory college-level statistics course with real investigations of some important and engaging problems. Each chapter presents a set of self-paced exercises to help students learn the skills of quantitative reasoning by performing the types of analyses that typically form the core of a first course in applied statistics. Students can learn and practice the software skills outside of class. Instructors can devote class time to statistics and statistical reasoning, rather than to rudimentary software instruction. Both students and teachers can direct their energies to the practice of data analysis in ways that inform students' understanding of the world through investigations of problems that matter in various fields of study.

Though written with undergraduate and beginning graduate students in mind, some practitioners might find the book helpful on the job and are well-advised to read the book selectively to address current tasks or projects. Chapters 1 and 2 form a good starting point before reading later sections. Appendix B (online for this edition) covers several data management topics that might be helpful for readers who undertake projects involving disparate data sources.

Prerequisites

No prior statistical knowledge is presumed. A basic grounding in algebra and some familiarity with the Mac OS or Windows environment are all you need in advance. An open, curious mind is also helpful.

A Message for Instructors

I assume that most teachers view class time as a scarce resource. One of my goals in writing this book was to strive for clarity throughout so that students can be expected to work through the book on their own and learn through their encounters with the examples and exercises. This book may be especially welcome for instructors using an inverted, or flipped, classroom approach.

Instructors might selectively use exercises as in-class demonstrations or group activities, interspersing instruction or discussion with computer work. More often, the chapters and scenarios can serve as homework exercises or assignments, either to prepare for other work, to acquire skills and understanding, or to demonstrate progress and mastery. Finally, some instructors might want to assign a chapter in connection with an independent analysis project. Several of the data tables contain additional variables that are not used within chapters. These variables might form the basis for original analyses or explorations.

The bibliography may also aid instructors seeking additional data sources or background material for exercises and assignments. Tips for classroom use of JMP are also available at the book's website, accessible through the author's page at support.sas.com/carver.

A Message for Students

Remember that the primary goal of this book is to help you understand the concepts and techniques of statistical analysis. JMP provides an ideal software environment to do just that. Naturally, each chapter is "about" the software and at times you will find yourself focusing on the details of a JMP analysis platform and its options. If you become entangled in the specifics of a problem, step back and try to refocus on the main statistical ideas rather than software issues.

This book should augment, but not replace, your primary textbook or your classroom time. To get the maximum benefit from the book, work mindfully and carefully. Read through a chapter before you sit down at the computer. Each chapter will require approximately 30 minutes of computer time; work at your own pace and take your time. Remember that variability is omnipresent, so expect that the time you need to complete a chapter may be more or less than 30 minutes.

The Application Scenarios at the end of each chapter are designed to reinforce and extend what you have learned in the chapter. The questions in this section are designed to challenge you.

Sometimes, it is obvious how to proceed with your analysis; sometimes, you will need to think a bit before you issue your first command. The idea is to engage in statistical thinking, integrating what you have learned throughout your course. There is much more to data analysis than finding a numerical answer, and these questions provide an opportunity to do realistic analysis. Because the examples use real data, don't expect to find neat "pat" results; computations won't typically come out to nice round numbers.

JMP is a large program designed for diverse user needs. Many of the features of the software are beyond the scope of an introductory course, and therefore this book does not discuss them. However, if you are curious or adventurous, you should explore the menus and Help system as well as the JMP website. You might find a quicker, more intuitive, or more interesting way to approach a problem. For most of the topics addressed in the book, you will see an introduction. There is almost always more to know.

What Should You Know about the Examples?

Real statistical investigations begin with pressing, important, or interesting questions, rather than with a set of techniques. Researchers do not begin a study by saying "Today is a good day to compute some standard deviations." Instead, they pose questions that can be pursued by analyzing data and follow a relatively straightforward protocol to refine the question, generate or gather suitable data, apply appropriate methods, and interpret their findings. The chapters in this book present questions that I hope you will find interesting, and then rely on the data tables provided to search for answers. The questions and analyses become progressively more challenging through the book.

Software Used to Develop the Book's Content

The book was developed using pre-production versions of JMP15 Pro. The essential examples work with JMP. Whenever a section illustrates JMP Pro functionality, that fact is clearly announced.

Example Data

As previously noted, each of the data tables referenced within the book contains real data, much of it downloaded from public websites. There are 45 different data tables, most of which have been updated for this edition. Readers should download all of the JMP data tables via the author page at support.sas.com/carver. Appendix A describes each file and its source. Many of the tables include columns (variables) in addition to those featured in exercises and examples. These variables might be useful for projects or other assignments.

Where Are the Exercise Solutions?

Solutions to the scenario questions are available via the author page at support.sas.com/carver. Instructors who adopt the book will be able to access all solutions. Students and other readers can find solutions to the even-numbered problems at the same site.

Thanks and Acknowledgments

This first edition of this book began at the urging of Curt Hinrichs, the Academic Program Manager for JMP. This led to conversations with Julie Palmieri, Editor-in-Chief at SAS Press at the time, after which the project started to take shape. I have had the great good fortune to work with a different editor for each edition: Stephenie Joyner, Sian Roberts, and, most recently, Catherine Connolly have kept this little trolley on the tracks.

At SAS Press, so many people have contributed to the planning and execution of the book in its development. For this edition, Sian Roberts is now Publisher. Suzanne Morgen handled the copyediting, Denise Jones the production, Robert Harris the cover design, and Missy Hannah the marketing effort.

Earlier editions were shaped and tended by Shelley Sessoms, Stacey Hamilton, Shelly Goodin, Mary Beth Steinbach, Cindy Puryear, Brenna Leath, Brad Kellam, Candy Farrell, Patrice Cherry, and Jennifer Dilley. My enduring thanks go to them all.

Many other professionals at JMP have influenced and informed the content of this book at critical points along the way. I am very grateful to John Sall, Xan Gregg, Jon Weisz, Brad Jones, Brady Brady, Jonathan Gatlin, Jeff Perkinson, Ian Cox, Chuck Pirrello, Brian Corcoran, Christopher Gotwalt, Curt Hinrichs, Mia Stephens, Volker Kraft, Julian Parris, Ruth Hummel, Kathleen Watts, Mary Loveless, Gail Massari, Lori Harris, Holly McGill, Peng Liu, and Eric Hill for encouraging me, answering my questions, setting me straight, and listening to my thoughts. To this group I send a special shout-out to JMP Senior Systems Engineer Rob Lievense, who has been a consistent advocate and supporter of this work.

I am especially thankful for the care and attention of those people who have reviewed this and the prior editions. Technical reviews of the current edition were provided by Mark Bailey, Duane Hayes, and Kristen Bradford. Mark has been the constant among reviewers, having made invaluable recommendations to all three editions. Performing double duty on the first two editions were Tonya Mauldin and Sue Walsh. Fang Chen, Paul Marovich, and Volker Kraft rounded out the many superb reviewers. Collectively, their critiques tightened and improved this book, and whatever deficiencies that may remain are entirely mine.

Naturally, the completion of a book requires time, space, and an amenable environment. I want to express public thanks to three institutions that provided facilities, time, and atmospherics suitable for steady work on this project. My home institution, Stonehill College, was exceptionally supportive, particularly through the efforts of Provost Joe Favazza and my chairperson, Debra Salvucci, and Department Administrative Assistant Carolyn McGuinness. Colleagues Dick Gariepy and Michael Salé generously tested several chapters and problems in their classrooms, and Jan Harrison and Susan Wall of our IT Department eased several technical aspects of this project as well.

Colleagues and students at the International Business School at Brandeis University sharpened my pedagogy and inspired numerous examples found in the book. During a sabbatical leave from Stonehill, Babson College was good enough to offer a visiting position and a wonderful place to write the first edition. For that opportunity, thanks go to Provost Shahid Ansari, former chairperson Norean Radke Sharpe, then-chair Steve Ericksen, and colleagues John McKenzie and George Recck.

During the summer of 2013, the Stonehill Undergraduate Research Experience (SURE) program provided a grant to support this work with time, space, and finances. Carolyn Moodie (Class of 2015) was a superior and self-directed research collaborator, assisting in the critical phases of problem formulation, data identification, data cleaning and exploratory analysis. Carolyn also brought a keen eye to editorial tasks, and willingly gave her feedback on which topics student readers might find engaging. Thanks also to Bonnie Troupe for her skillful administration of the SURE program. During the spring and summer of 2013, Stonehill students Dan Doherty, Erin Hollander, and Tate Molaghan also pitched in with editorial and research assistance.

Special acknowledgment also goes to former Stonehill students from BUS207 (Intermediate Statistics) who “road tested” several chapters, and very considerable thanks to three students who assisted greatly in shaping prose and examples, as well as developing solutions to scenario problems: Frank Groccia, Dan Bouchard, and Matt Arey. Later students in BUS206 (Quantitative Analysis for Business) at Stonehill also class-tested several chapters and exercises.

Several of the data tables came through the gracious permission of their original authors and compilers. I gratefully acknowledge the permission granted by my good friend George Aronson for the Maine SW table; by Prof. Max A. Little for the Parkinson’s disease vocal data; by Prof. Jesper Rydén for the Sonatas data table (from which the Haydn and Mozart tables were extracted); by Prof. John Holcomb for the North Carolina birth weight data; and by Prof. I-Cheng Yeh for the Concrete table and the two subsets from that data.

In recent years, my thoughts about what is important in statistics education have been radically reshaped by colleagues in the ISOSTAT listserv and the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) and the U.S. Conference on Teaching Statistics that CAUSE organizes every two years. The May 2013 CAUSE-sponsored workshop “Teaching the Statistical Investigation Process with Randomization-Based Inference” given by Beth Chance, Allan Rossman, and Nathan Tintle influenced some of the changes in my presentation of inference. Over an even longer period, our local group of New England Isolated Statisticians and the great work of the ASA’s Section on Statistics Education influence me daily in the classroom and at the keyboard.

Finally, it is a pleasure to thank my family. My sons, Sam and Ben, keep me modest and regularly provide inspiration and insight. My wife Donna—partner, friend, wordsmith extraordinaire—has my love and thanks for unflagging encouragement, support, and warmth. This book is dedicated to them.

We Want to Hear from You

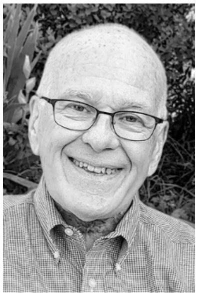
SAS Press books are written *by* SAS Users *for* SAS Users. We welcome your participation in their development and your feedback on SAS Press books that you are using. Please visit sas.com/books to do the following:

- Sign up to review a book
- Recommend a topic
- Request information on how to become a SAS Press author
- Provide feedback on a book

Do you have questions about a SAS Press book that you are reading? Contact the author through saspress@sas.com or https://support.sas.com/author_feedback.

SAS has many resources to help you find answers and expand your knowledge. If you need additional help, see our list of resources: sas.com/books.

About The Author



Robert Carver is Professor Emeritus of Business Administration at Stonehill College in Easton, Massachusetts, and he recently retired as a senior lecturer at the Brandeis University International Business School in Waltham, Massachusetts. At both institutions, he was instrumental in establishing programs in Data Science and Business Analytics, taught courses on quantitative methods in addition to general management courses, and won teaching awards at both schools. His primary research interest is statistics education. A JMP user since 2006, Carver holds an AB in political science from Amherst College in Amherst, Massachusetts, and an MPP and PhD in public policy from the University of Michigan at Ann Arbor.

Learn more about this author by visiting his author page at support.sas.com/carver. There you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more.

Chapter 15: Simple Linear Regression Inference

Overview

In Chapter 4, we learned to summarize two continuous variables at a time using scatterplot, correlations, and line fitting. In this chapter, we will return to that subject, this time with the object of generalizing from the patterns in sample data in order to draw conclusions about an entire population. The main statistical tool that we will use is known as *linear regression analysis.* We will devote this chapter and three later chapters to the subject of regression.

Because Chapter 4 is now many pages back, we will begin by reviewing some basic concepts of bivariate data and line fitting. Then, we will discuss the fundamental model used in simple linear regression. After that, we will discuss the crucial conditions necessary for inference, and finally, we will see how to interpret the results of a regression analysis.

Fitting a Line to Bivariate Continuous Data

We introduced regression in Chapter 4 using the data table **Birthrate 2017**. This data table contains several columns related to the variation in the birth rate and the risks related to childbirth around the world as of 2017. In this data table, the United Nations reports figures for 193 countries. Let’s briefly revisit that data now to review some basic concepts, focusing on two measures of the frequency of births in different nations.

In the next several chapters, we will expand our investigation of regression analysis and will return to the illustrative data tables several times. For this reason, you should work through the examples in a JMP Project. By creating and saving a project in this chapter, your workflow in the coming chapters will be more efficient. At this point, create a new project called Chap_15.

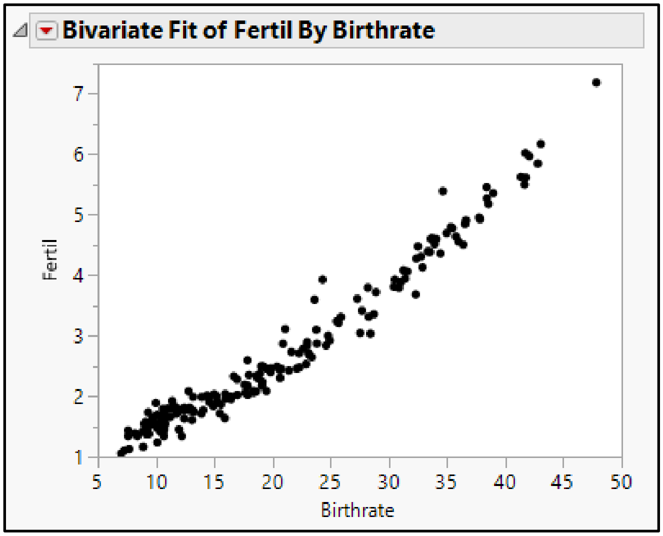
1. Open the **Birthrate 2017** data table now. If any row states were previously set and saved, clear them now.

As we did in Chapter 4, let's look at the columns labeled **BirthRate** and **Fertil**. A country's annual birth rate is defined as the number of live births per 1,000 people in the country. The fertility rate is the mean number of children that would be born to a woman during her lifetime. We plotted these two variables in Chapter 4; let us do that again now.

2. Select **Analyze ► Fit Y by X**. Cast **Fertil** as **Y** and **BirthRate** as **X** and click **OK**.

Your results will look like those shown in Figure 15.1.

Figure 15.1: Relationship Between Birth Rate and Fertility Rate



This is the same graph that we saw in Figure 4.13. Again, we note that general pattern is upward from left to right: fertility rates increase as the birth rate increases, although there are some countries that depart from the pattern. The pattern can be described as linear, although there is a mild curvature at the lower left. We also see that a large number of countries are concentrated in the lower left, with low birth rates and relatively low maternal mortality.

In Chapter 4, we illustrated the technique of line-fitting using these two columns. Because these two columns really represent two ways of thinking about a single construct ("how many babies?”), let us turn to a different example to expand our study of simple linear regression analysis.

We will return to the NHANES data and look at two body measurement variables. Because adult body proportions are different from children and because males and females differ, we will restrict the first illustrative analysis to male respondents ages 18 and up.

3. Open the data table called **NHANES 2016**. If row states were saved from an earlier session, clear them now. In the coming chapters, we will work through several illustrations using NHANES data.

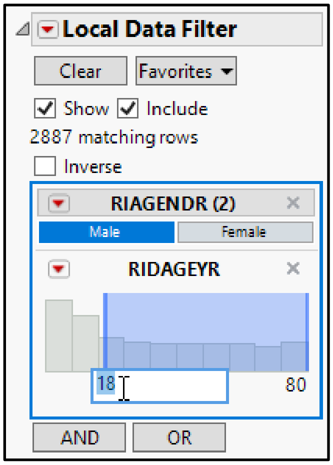
This table contains respondents of all ages, male and female. To isolate adult males, we might be inclined to apply a data filter to the entire table, but because we will want to use other respondents in later examples, let’s wait and apply a local data filter within the context of the first regression analysis. The global data filter temporarily affects the whole data table; if you resave the data table, the row states are saved with it. A local data filter affects the rows used within a single analysis report but does not alter the underlying data table.

Now we can begin the regression analysis. We will examine the relationship between waist circumference and body mass index, or BMI, which is the ratio of a person’s weight to the square of height. In the data table, waist measurements are in centimeters, and BMI is kilograms per square meter. In this analysis, we will see if there is a predictable relationship between men’s waist measurements and their BMIs.

We begin the analysis as we have done so often, using the **Fit Y by X** platform.

4. Select **Fit Y by X**. Cast **BMXBMI** as **Y** and **BMXWAIST** as **X** and click **OK.** The scatterplot contains points for the entire NHANES sample. Let’s filter the rows to restrict the analysis to adult males.
5. Click the red triangle next to **Bivariate Fit** and select **Local Data Filter.**
6. While pressing the Ctrl key, highlight **RIAGENDR** and **RIDAGEYR**, and click the dark plus sign.
7. In the **Data Filter** (see Figure 15.2), make sure that the **Show** and **Include** options are checked.
8. Then click **Male** under **RIAGENDR** to include just the male subjects.
9. Finally, click the number 0 on the left of **RIDAGEYR** and replace it with 18. This sets the lower bound for **RIDAGEYR** to be just 18 years. We want to select any respondent who is a male age 18 or older.

Figure 15.2: Selection Criteria for Males Age 18 and Older



We have restricted the analysis to the 2,887 rows of male respondents who are 18 years of age and older. You might have noticed that the resulting graph has far fewer points and very clearly shows a positive linear trend, but the plot is difficult to decipher because there are so many points crowded together, with many of them plotted on top of each other. As very large data tables have become more common, we need strategies for making interpretable graphs with densely packed points. Do the following to make the graph more readable.

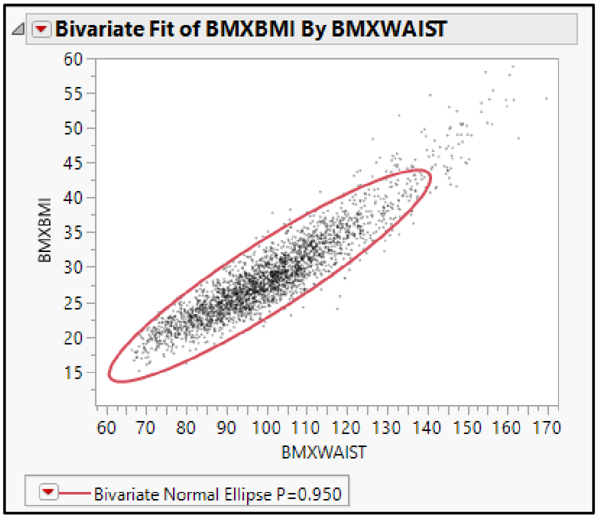
10. Right-click anywhere in the plot region, and select **Marker Size ▶ 0, Dot**. This makes each point much smaller, reducing the overlap.
11. Right-click again and select **Transparency**. Within the dialog box, set the transparency parameter to 0.4. This makes dot partially transparent, so that overplotted points appear darker than single points.

This graph (see Figure 15.3, which reflects the next two numbered steps as well) illustrates the first thing that we want to look for when planning to conduct a linear regression analysis—we see a general linear trend in the data. Think of stretching an elliptical elastic band around the cloud of points; that would result in a long and narrow ellipse lying at a slant, which would contain most, if not all, of the points. In fact, we can use JMP to overlay such an ellipse on the graph.

12. Click the red triangle next to **Bivariate Fit** and select **Density Ellipse ▶ 0.95**.

The resulting ellipse appears incomplete because of the default axis settings on our graph. We can customize the axes to show the entire ellipse using the grabber to shift the axes.

Figure 15.3: A Linear Pattern of BMI Versus Waist



13. Move the grabber tool near the origin on the vertical axis and slide upward until you see a hash mark below 15 appear on the Y axis. Do the same on the horizontal axis until the waist value of 60 cm appears on the X axis.

This graph is a typical candidate for linear regression analysis. Nearly all of the points lie all along the same sloped axis in the same pattern, with consistent scatter. Before running the regression, let's step back for a moment and consider the fundamental regression model.

The Simple Regression Model

When we fit a line to a set of points, we do so with a model in mind and with a provisional idea about how we came to observe the points in our sample. The reasoning goes like this. We speculate or hypothesize that there is a linear relationship between Y and X such that whenever X increases by one unit (centimeters of waist circumference, in this case), then Y changes, on average, by a constant amount. For any specific individual, the observed value of Y could deviate from the general pattern.

Algebraically, the model looks like this:

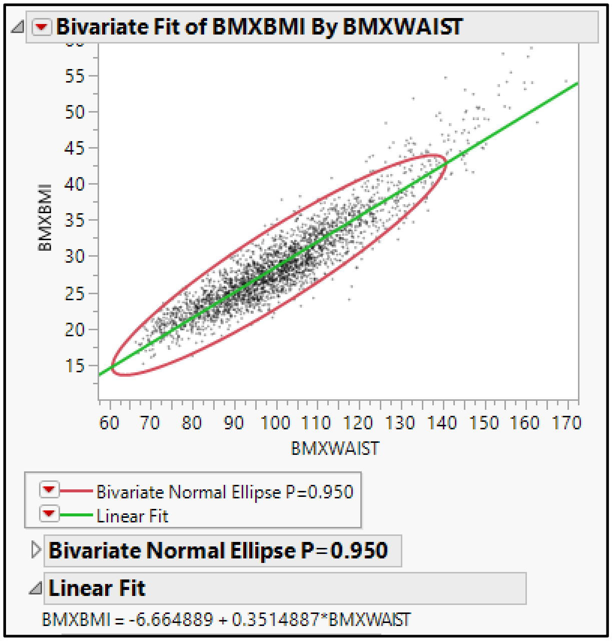
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where Y_i and X_i are the observed values for one respondent, β_0 and β_1 are the intercept and slope of the underlying (but unknown) relationship, and ε_i is the amount by which an individual's BMI departs from the usual pattern. We envision ε_i as purely random noise. In short, we can express each observed value of Y_i as partially reflecting the underlying linear pattern, and partially reflecting a random deviation from the pattern. Look again at Figure 15.3. Can you visualize each point as lying in the vicinity of a line? Let’s use JMP to estimate the location of such a line.

1. Click the red triangle next to **Bivariate Fit** and select **Fit Line**.

Now your results will look like Figure 15.4 below. We see a green *fitted line* that approximates the upward pattern of the points.

Figure 15.4: Estimated Line of Best Fit


Bivariate Fit of BMXBMI By BMXWAIST
55
50
45
40
35
30
25
20
15
BMXBMI
60
70
80
90
100
110
120
130
140
150
160
170
BMXWAIST
Bivariate Normal Ellipse P=0.950
Linear Fit
Bivariate Normal Ellipse P=0.950
Linear Fit

BMXBMI = -6.664889 + 0.3514887*BMXWAIST

Below the graph, we find the equation of that line:

BMXBMI = -6.664889 + 0.3514887*BMXWAIST

The slope of this line describes how these two variables co-vary. If we imagine two groups of men whose waist circumferences differ by 1 centimeter, the group with the larger waists would average BMIs that are approximately 0.35 kg/m^2 higher. As we learned in Chapter 4, this equation summarizes the relationship among the points in this sample. Before learning about the inferences that we might draw from this, let’s refine our understanding of the two chunks of the model: the linear relationship and the random deviations.

Thinking About Linearity

If two variables have a linear relationship, their scatterplot forms a line or at least suggests a linear pattern. In this example, our variables have a *positive* relationship: as X increases, Y increases. In another case, the relationship might be negative, with Y decreasing as X increases.

But what does it mean to say that two variables have a *linear* relationship? What type of underlying dynamic generates a linear pattern of dots?

As noted earlier, linearity involves a constant change in Y each time X changes by one unit. Y might rise or fall, but the key feature of a linear relationship is that the shifts in Y do not accelerate or diminish at different levels of X. If we plan to generalize from our sample, it is important to ask if it is reasonable to expect Y to vary in this particular way as we move through the domain of realistically possible X values.

Random Error

The regression model also posits that empirical observations tend to deviate from the linear pattern, and that the deviations are themselves a random variable. We will have considerably more to say about the random deviations in Chapter 16, but it is very useful at the outset to understand this aspect of the regression model.

Linear regression analysis does not demand that all points line up perfectly, or that the two continuous variables have a very close (or “strong”) association. On the other hand, if groups of observations systematically depart from the general linear pattern, we should ask if the deviations are truly random, or if there is some other factor to consider as we untangle the relationship between Y and X.

What Are We Assuming?

The preceding discussion outlines the conditions under which we can generalize using regression analysis. First, we need a logical or theoretical reason to anticipate that Y and X have a linear relationship. Second, the default method[1] that we use to estimate the line of best fit works reliably. We know that the method works reliably when the random errors, ε_i, satisfy four conditions:

- They are normally distributed.
- They have a mean value of 0.
- They have a constant variance, σ^2, regardless of the value of X.
- They are independent across observations.

At this early stage in the presentation of this technique, it might be difficult to grasp all the implications of these conditions. Start by understanding that the following might be red flags in a scatter plot with a fitted line:

- The points seem to bend or oscillate predictably around the line.
- There are a small number of outliers that stand well apart from the mass of the points.

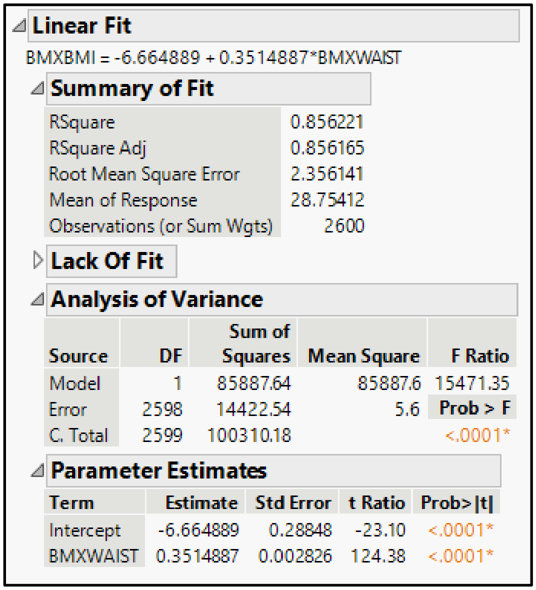
- The points seem snugly concentrated near one end of the line but fan out toward the other end.
- There seem to be greater concentrations of points distant from the line, but not so many points concentrated near the line.

In this example, none of these trouble signs is present. In the next chapter, we will learn more about looking for problems with the important conditions for inference. For now, let's proceed assuming that the sample satisfies all of the conditions.

Interpreting Regression Results

There are four major sections in the results panel for the linear fit (see Figure 15.5), three of which are fully disclosed by default. We have already seen the equation of the line of best fit and discussed its meaning. In this part of the chapter, we will discuss the three other sections in order.

Figure 15.5: Regression Results



```
Linear Fit
BMXBMI = -6.664889 + 0.3514887*BMXWAIST

  Summary of Fit
  RSquare                       0.856221
  RSquare Adj                   0.856165
  Root Mean Square Error        2.356141
  Mean of Response              28.75412
  Observations (or Sum Wgts)        2600

  Lack Of Fit

  Analysis of Variance
                        Sum of
  Source      DF       Squares   Mean Square    F Ratio
  Model        1      85887.64       85887.6   15471.35
  Error     2598      14422.54           5.6   Prob > F
  C. Total  2599     100310.18                  <.0001*

  Parameter Estimates
  Term          Estimate   Std Error   t Ratio   Prob>|t|
  Intercept    -6.664889     0.28848    -23.10    <.0001*
  BMXWAIST     0.3514887    0.002826    124.38    <.0001*
```

Summary of Fit

Under the heading **Summary of Fit**, we find five statistics that describe the fit between the data and the model.

- **RSquare** and **RSquare Adj** both summarize the strength of the linear relationship between the two continuous variables. The R-square statistics range between 0.0 and 1.0, where 1.0 is a perfect linear fit. Just as in Chapter 4, think of R square as the proportion of variation in Y that is associated with X. Here, both statistics are approximately 0.86, suggesting that a male's waist measurement could be a very good predictor of his BMI.

- **Root Mean Square Error** (RMSE) is a measure of the dispersion of the points from the estimated line. Think of it as the sample standard deviation of the random noise term, ε. When points are tightly clustered near the line, this statistic is relatively small. When points are widely scattered from the line, the statistic is relatively large. Comparing the RMSE to the mean of the response variable (next statistic) is one way to assess its relative magnitude.
- **Mean of Response** is just the sample mean value of Y.
- **Observations** is the sample size. In this table, we have complete waist and BMI data for 2600 men. Although we selected 2,887 rows with the data filter, 287 of the rows were missing a value for one or both of the columns in the regression model.

Lack of Fit

The next heading is **Lack of Fit**, but this panel is initially minimized in this case. Lack of fit tests typically are considered topics for more advanced statistics courses, so we only mention them here without further comment.

Analysis of Variance

These ANOVA results should look familiar if you have just completed Chapter 14. In the context of regression, ANOVA gives us an overall test of significance for the regression model. In a one-way ANOVA, we hypothesized that the mean of a response variable was the same across several categories. In regression, we hypothesize that the mean of the response variable is the same regardless of X—that Y does not vary in tandem with X.

We read the table just as we did in the previous chapter, focusing on the F-ratio and the corresponding *P*-value. Here F is over 15.471 and the *P*-value is smaller than 0.0001. This probability is so small that it is highly unlikely that the computed F-ratio came about through the error associated with random sampling. We reject the null hypothesis that waist circumference and BMI are unrelated, and conclude that we have found a statistically significant relationship.

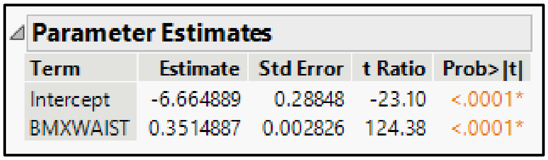
Not only can we say that the pattern describes the sample, we can say with confidence that the relationship generalizes to the entire population of males over age 17 in the United States.

Parameter Estimates and t Tests

The final panel in the results provides the estimated intercept and slope of the regression line and the individual t tests for each. The slope and intercept are sometimes called the *coefficients* in the regression equation, and we treat them as the *parameters* of the linear regression model.

In Figure 15.6, we reproduce the parameter estimate panel, which contains five columns. The first two columns—**Term** and **Estimate**—are the estimated intercept and slope that we saw earlier in the equation of the regression line.

Figure 15.6: Parameter Estimates



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-6.664889	0.28848	-23.10	<.0001*
BMXWAIST	0.3514887	0.002826	124.38	<.0001*

Because we are using a sample of the full population, our estimates are subject to sampling error. The **Std Error** column estimates the variability attributable to sampling. The **t Ratio** and **Prob>|t|** columns show the results of a two-sided test of the null hypothesis that a parameter is truly equal to 0.

Why do we test the hypotheses that the intercept and slope equal zero? Think about what a zero slope represents. If X and Y are genuinely independent and unrelated, then changes in the value of X have no influence or bearing on the values of Y. In other words, the slope of a line of best fit for two such variables should be zero. For this reason, we always want to look closely at the significance test for the slope. Depending on the study and the meaning of the data, the test for the intercept may or may not have practical importance to us.

In a simple linear regression, the ANOVA and t test results for the slope will always lead to the same conclusion[2] about the hypothesized independence of the response and factor variables. Here, we find that our estimated slope of 0.351 kg/m^2 change in BMI per 1 cm increase in waist circumference is very convincingly different from 0: in fact, it’s more than 124 standard errors away from 0. It’s inconceivable that such an observed difference is the coincidental result of random sampling.

Testing for a Slope Other Than Zero

In some investigations, we might begin with a theoretical model that specifies a value for the slope or the intercept. In that case, we come to the analysis with hypothesized values of either β_0 or β_1 or both, and we want to test those values. The **Fit Y by X** platform does not accommodate such significance tests, but the **Fit Model** platform does. We used **Fit Model** in the prior chapter to perform a two-way ANOVA. In this example, we will use it to test for a specific slope value other than 0.

We will illustrate with an example from the field of classical music, drawn from an article by Prof. Jesper Rydén of Uppsala University in Sweden (Rydén 2007). The article focuses on piano sonatas by Franz Joseph Haydn (1732–1809) and Wolfgang Amadeus Mozart (1756–1791) and investigates the idea that these two composers incorporated the *golden mean* within their compositions. A sonata is a form of instrumental music that follows a standard structure. In the first part, known formally as the exposition, the composer introduces a melodic theme in one key and then a second longer theme in a related but different key. After the exposition comes a second portion called development, which elaborates upon the basic melodies, developing them more fully, offering some variations, Finally, both themes recur in the recapitulation section, but this time, both tunes are in the same key as the opening of the piece.

Some music scholars believe that Haydn and Mozart strove for an aesthetically pleasing but asymmetric balance in the lengths of the melodic themes. More specifically, these scholars hypothesize that the composers might have structured their sonatas (deliberately or not) so that the relative lengths of the two themes approximated the golden mean. We will refer to the first, shorter theme as a and the second theme as b.

The golden mean (sometimes called the golden ratio), characterized and studied in the West at least since the ancient Greeks, refers to the division of a line into a shorter segment *a*, and a longer segment *b*, such that the ratio of *a:b* equals the ratio of *b:(a+b).* Equivalently,

$$\frac{a}{b} = \frac{b}{(a+b)} = \phi \approx 0.61803.$$

We have a data table called **Mozart** containing the lengths, in musical measures, of the shorter and longer themes of 29 Mozart sonatas. If, in fact, Mozart was aiming for the golden ratio in these compositions, then we should find a linear trend in the data. Moreover, it should be characterized by this line:

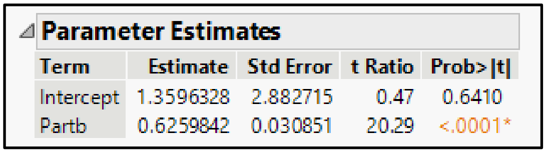
$$\text{a} = 0 + 0.61803(\text{b})$$

So, we will want to test the hypothesis that β_1 = 0.61803 rather than 0.

1. Open the data table called **Mozart**.
2. Select **Analyze ▶ Fit Model**. Select **Parta** as **Y**, then add **Partb** as the only model effect, and run the model.

Both the graph and the **Summary of Fit** indicate a strong linear relationship between the two parts of these sonatas. Figure 15.7 shows the parameter estimates panel from the results.

Figure 15.7: Estimates for Mozart Data



⊿ **Parameter Estimates**

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1.3596328	2.882715	0.47	0.6410
Partb	0.6259842	0.030851	20.29	<.0001*

Rounding the estimates slightly we can write an estimated line as *Parta = 1.3596 + 0.626(Partb)*. On its face, this does not seem to match the proposed equation above. However, let's look at the t tests. The estimated intercept is not significantly different from 0, so we cannot conclude that the intercept is other than 0. The hypothesized intercept of 0 is still credible.

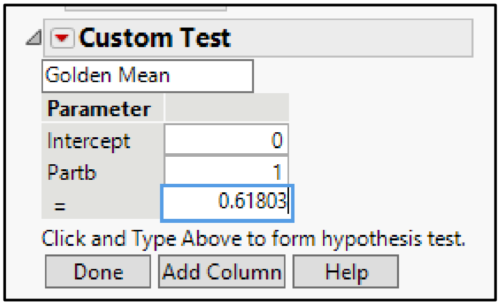
Now look at the results for the slope. The estimated slope is 0.6259842, and its standard error is 0.030851. The reported *t*-ratio of about 20 standard errors implicitly compares the estimated slope to a hypothesized value of 0. To compare it to a different hypothesized value, we will want to compute the following ratio:

$$\frac{estimate - hypothesized}{std.error} = \frac{0.6259842 - 0.61803}{0.030851} = 0.2578$$

We can have JMP compute this ratio and its corresponding *P*-value as follows:

3. Click the red triangle next to **Response Parta** and select **Estimates ▶ Custom Test**. Scroll to the bottom of the results report where you will see a panel like the one shown in Figure 15.8.
4. The upper white rectangle is an editable field for adding a title; type **Golden Mean** in the box.
5. In the box next to **Partb**, change the **0** to a **1** to indicate that we want to test the coefficient of **Partb**.
6. Finally, enter the hypothesized value of the golden mean, **.61803** in the box next to **=**, and click the **Done** button.

Figure 15.8: Specifying the Column and Hypothesized Value



Custom Test

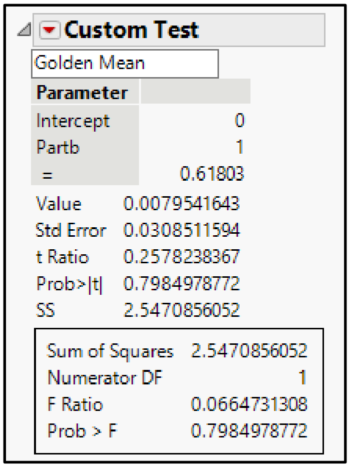
Golden Mean

Parameter	
Intercept	0
Partb	1
=	0.61803

Click and Type Above to form hypothesis test.

The **Custom Test** panel now becomes a results panel, presenting both a t test and an F test, as shown in Figure 15.9. As our earlier calculation showed, the estimated slope is less than 0.26 standard errors from the hypothesized value, which is very close. Based on the large *P*-value of 0.798, we fail to reject the null hypothesis that the slope equals the golden mean.

Figure 15.9: Custom Test Results



Custom Test

Golden Mean

Parameter	
Intercept	0
Partb	1
=	0.61803

Value	0.0079541643
Std Error	0.0308511594
t Ratio	0.2578238367
Prob>|t|	0.7984978772
SS	2.5470856052

Sum of Squares	2.5470856052
Numerator DF	1
F Ratio	0.0664731308
Prob > F	0.7984978772

In other words, the golden mean theory is credible. As always, we cannot prove a null hypothesis, so this analysis does not definitively establish that Mozart's sonatas conform to the golden mean. This is an important distinction in the logic of statistical testing—our tests can discredit a null hypothesis with a high degree of confidence, but we cannot confirm a null hypothesis. What we can say is that we have put a hypothesis to the test, and it is still plausible.

Before ending this session, be sure to save your project. If prompted to save unsaved documents within the project, save them all.

Application

Now that you have completed all of the activities in this chapter, use the concepts and techniques that you have learned to respond to these questions.

1. *Scenario:* Return to the **NHANES 2016** data table.
   a. Using the Data Filter, show and include female respondents ages 18 and older. Perform a regression analysis for BMI and waist circumference for adult women and report your findings and conclusions.
   b. Is waist measurement a better predictor (in other words, a better fit) of BMI for men or for women?
   c. Perform one additional regression analysis, this time looking only at respondents under the age of 17. Summarize your findings.

2. *Scenario:* High blood pressure continues to be a leading health problem in the United States. In this problem, continue to use the **NHANES 2016** data table. For this analysis, we will focus on just the following variables:
   - **RIAGENDR:** Respondent’s gender
   - **RIDAGEYR**: Respondent’s age in years
   - **BMXWT**: Respondent’s weight in kilograms
   - **BPXPLS**: Respondent’s resting pulse rate
   - **BPXSY1**: Respondent’s systolic blood pressure (“top” number in BP reading)
   - **BPXD1**: Respondent’s diastolic blood pressure (“bottom” number in BP reading)

   a. Investigate a possible linear relationship of systolic blood pressure versus age. What, specifically, tends to happen to blood pressure as people age? Would you say there is a strong linear relationship?
   b. Perform a regression analysis of systolic and diastolic blood pressure. Explain fully what you have found.
   c. Create a scatterplot of systolic blood pressure and pulse rate. One might suspect that higher pulse rate is associated with higher blood pressure. Does the analysis bear out this suspicion?

3. *Scenario:* We will continue to examine the World Development Indicators data in **BirthRate 2017**. We will broaden our analysis to work with other variables in that file:
   - **MortUnder5**: Deaths, children under 5 years per 1,000 live births
   - **MortInfant**: Deaths, infants per 1,000 live births

   a. Create a scatterplot for **MortUnder5** and **MortInfant**. Report the equation of the fitted line and the R-square value and explain what you have found.
4. *Scenario:* How do the prices of used cars vary according to the mileage of the cars? Our data table **Used Cars** contains observational data about the listed prices of three popular compact car models in three different metropolitan areas in the U.S. All of the cars are two years old.

   a. Create a scatterplot of price versus mileage. Report the equation of the fitted line and the R-square value and explain what you have found.
5. *Scenario*: Stock market analysts are always on the lookout for profitable opportunities and for signs of weakness in publicly traded stocks. Market analysts make extensive use of regression models in their work, and one of the simplest ones is known as the *random* (or *drunkard's*) *walk* model. Simply put, the model hypothesizes that over a relatively short period of time, the price of a share of stock is a random deviation from its price on the prior day. If Y_t represents the price at time *t*, then $Y_t = Y_{t-1} + \varepsilon$. In this problem, you will fit a random walk model to daily closing prices for McDonald's Corporation for the first six months of 2019 and decide how well the random walk model fits. The data table is called **MCD**.

   a. Create a scatterplot with the daily closing price on the vertical axis and the prior day's closing price on the horizontal. Comment on what you see in this graph.

   b. Fit a line to the scatterplot and test the credibility of the random walk model. Report on your findings.
6. *Scenario*: Franz Joseph Haydn was a successful and well-established composer when the young Mozart burst upon the cultural scene. Haydn wrote more than twice as many piano sonatas as Mozart. Use the data table **Haydn** to perform a parallel analysis to the one we did for Mozart.

   a. Report fully on your findings from a regression analysis of **Parta** versus **Partb**.

   b. How does the fit of this model compare to the fit using the data from Mozart?

7. *Scenario*: Throughout the animal kingdom, animals require sleep, and there is extensive variation in the number of hours in a day that different animals sleep. The data table called **Sleeping Animals** contains information for more than 60 mammalian species, including the average number of hours per day of total sleep. This will be the response column in this problem.
    a. Estimate a linear regression model using gestation as the factor. Gestation is the mean number of days that females of these species carry their young before giving birth. Report on your results and comment on the extent to which gestational period is a good predictor of sleep hours.
    b. Now perform a similar analysis using brain weight as the factor. Report fully on your results and comment on the potential usefulness of this model.

8. *Scenario*: For many years, it has been understood that tobacco use leads to health problems related to the heart and lungs. The **Tobacco Use** data table contains recent data about the prevalence of tobacco use and of certain diseases around the world.
    a. Using cancer mortality (**CancerMort**) as the response variable and the prevalence of tobacco use in both sexes (**TobaccoUse**), run a regression analysis to decide whether total tobacco use in a country is a predictor of the number of deaths from cancer annually in that country.
    b. Using cardiovascular mortality (**CVMort**) as the response variable and the prevalence of tobacco use in both sexes (**TobaccoUse**), run a regression analysis to decide whether total tobacco use in a country is a predictor of the number of deaths from cardiovascular disease annually in that country.
    c. Review your findings in the earlier two parts. In this example, we are using aggregated data from entire nations rather than individual data about individual patients. Can you think of any ways in which this fact could explain the somewhat surprising results?

9. *Scenario*: In Chapter 2, our first illustration of experimental data involved a study of the compressive strength of concrete. In this scenario, we look at a set of observations all taken at 28 days (4 weeks) after the concrete was initially formulated. The data table is **Concrete28**. The response variable is the **Compressive Strength** column, and we will examine the relationship between that variable and two candidate factor variables.
    a. Use **Cement** as the factor and run a regression. Report on your findings in detail. Explain what this slope tells you about the impact of adding more cement to a concrete mixture.
    b. Use **Water** as the factor and run a regression. Report on your findings in detail. Explain what this slope tells you about the impact of adding more water to a concrete mixture.

10. *Scenario*: Prof. Frank Anscombe of Yale University created an artificial data set to illustrate the hazards of applying linear regression analysis without looking at a scatterplot (Anscombe 1973). His work has been very influential, and JMP includes his illustration among the sample data tables packaged with the software. You'll find **Anscombe** both in this book's data tables and in the JMP sample data tables. Open it now.
    a. In the upper left panel of the data table, you will see a red triangle next to the words **The Quartet**. Click the triangle, and select **Run Script**. This produces four regression analyses corresponding to four pairs of response and predictor variables. Examine the results closely, and write a brief response comparing the regressions. What do you conclude about this quartet of models?
    b. Now return to the results and click the red triangle next to **Bivariate Fit of Y1 By X1**; select **Show Points** and re-interpret this regression in the context of the revised scatterplot.
    c. Now reveal the points in the other three graphs. Is the linear model equally appropriate in all four cases?

11. *Scenario*: Many cities in the U.S. have active used car markets. Typically, the asking price for a used car varies by model, age, mileage, and features. The data table called **Used Cars** contains asking prices (**Price**) and mileage (**Miles**) for three popular budget models; all cars were two years old at the time the data were gathered, and we have data from three U.S. metropolitan areas. All prices are in dollars. In this analysis, **Price** is the response and **Miles** is the factor.
    a. Because the car model is an important consideration, we will begin by analyzing the data for one model: the Civic EX. Use a **Local Data Filter** to isolate the Civic EX data for analysis. Run a regression. How much does the asking price decline, on average, per mile driven? What would be a mean asking price for a two-year-old Civic EX that had never been driven? Comment on the statistical significance and goodness-of-fit of this model.
    b. Repeat the previous step using the Corolla LE data.
    c. Repeat one more time using the PT Cruiser data.
    d. Finally, compare the three models. For which set of data does the model fit best? Explain your thinking. For which car model are you most confident about the estimated slope?

12. *Scenario:* We will return to the World Development Indicators data in **WDI**. In this scenario, we will investigate the relationship between access to improved sanitation (the percent of the population with access to sewers and the like) and life expectancy. The response column is **life_exp** and the factor is **sani_acc**.
    a. Use the **Local Data Filter** to **Show** and **Include** only the observations for the **Year 2015**, and the **Latin America & Caribbean Region** nations. Describe the relationship that you observe between access to improved sanitation and life expectancy.
    b. Repeat the analysis for **East Asia & Pacific** countries in 2010.
    c. Repeat the same analysis for the countries located in **Sub-Saharan Africa**.
    d. How do the three regression models compare? What might explain the differences in the models?

13. *Scenario:* The data table called **USA Counties** contains a wide variety of measures for every county in the United States based on the 2010 U.S. Census.
    a. Run a regression casting **sales_per_capita** (retail sales dollars per person, 2007) as **Y** and **per_capita_income** as **X**. Write a short paragraph explaining why county-wide retail sales might vary with per capita income, and report on the strengths and weaknesses of this regression model.

[1] Like all statistical software, JMP uses a default method to line-fitting that is known as *ordinary least squares estimation*, or *OLS.* A full discussion of OLS is well beyond the scope of this book, but it’s worth noting that these assumptions refer to OLS in particular, not to regression in general.

[2] They will have identical *P*-values and the F-ratio will be the square of the *t* ratio.