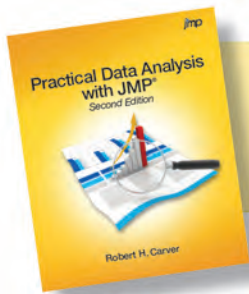


Practical Data Analysis with JMP[®] *Second Edition*



Robert H. Carver



From *Practical Data Analysis with JMP®*, Second Edition. Full book available for purchase [here](#).

Contents

About This Book	xiii
About The Author	xxiii
Chapter 1 Getting Started: Data Analysis with JMP	1
Overview.....	1
Goals of Data Analysis: Description and Inference	2
Types of Data.....	3
Starting JMP	4
A Simple Data Table.....	5
Graph Builder: An Interactive Tool to Explore Data	9
Using an Analysis Platform	12
Row States.....	14
Exporting JMP Results to a Word-Processor Document	17
Saving Your Work.....	18
Leaving JMP	19
Chapter 2 Data Sources and Structures.....	21
Overview.....	21
Populations, Processes, and Samples.....	22
Representativeness and Sampling.....	23
Simple Random Sampling.....	24
Other Types of Random Sampling.....	26
Non-Random Sampling	26
Big Data	26
Cross-Sectional and Time Series Sampling.....	27
Study Design: Experimentation, Observation, and Surveying.....	27

Experimental Data—An Example	28
Observational Data—An Example	31
Survey Data—An Example	31
Creating a Data Table	34
Raw Case Data and Summary Data	34
Application	36
Chapter 3 Describing a Single Variable	39
Overview	39
The Concept of a Distribution	40
Variable Types and Their Distributions	40
Distribution of a Categorical Variable	41
Using the Data Filter to Temporarily Narrow the Focus	43
Using the Chart Command to Graph Categorical Data	44
Using the Graph Builder to Explore Categorical Data	46
Distribution of a Quantitative Variable	47
Using the Distribution Platform for Continuous Data	48
Taking Advantage of Linked Graphs and Tables to Explore Data	51
Customizing Bars and Axes in a Histogram	52
Exploring Further with the Graph Builder	54
Summary Statistics for a Single Variable	55
Outlier Box Plots	56
Application	57
Chapter 4 Describing Two Variables at a Time	63
Overview	63
Two-by-Two: Bivariate Data	63
Describing Covariation: Two Categorical Variables	65
Describing Covariation: One Continuous, One Categorical Variable	71
Describing Covariation: Two Continuous Variables	73
More Informative Scatter Plots	78
Application	79

Chapter 5 Review of Descriptive Statistics	85
Overview.....	85
The World Development Indicators.....	86
Millennium Development Goals	86
Questions for Analysis.....	87
Applying an Analytic Framework.....	88
Data Source and Structure.....	88
Observational Units	89
Variable Definitions and Data Types.....	89
Preparation for Analysis	90
Univariate Descriptions	90
Explore Relationships with Graph Builder.....	93
Further Analysis with the Multivariate Platform.....	96
Further Analysis with Fit Y by X.....	97
Summing Up: Interpretation and Conclusions.....	99
Visualizing Multiple Relationships.....	100
Chapter 6 Elementary Probability and Discrete Distributions.....	103
Overview.....	103
The Role of Probability in Data Analysis.....	104
Elements of Probability Theory.....	104
Probability of an Event	105
Rules for Two Events	106
Assigning Probability Values	107
Contingency Tables and Probability	108
Discrete Random Variables: From Events to Numbers.....	111
Three Common Discrete Distributions	111
Integer Distribution.....	112
Binomial.....	113
Poisson	115
Simulating Random Variation with JMP	116
Discrete Distributions as Models of Real Processes	118
Application	120

Chapter 7 The Normal Model	125
Overview.....	125
Continuous Data and Probability.....	125
Density Functions.....	126
The Normal Model.....	128
Normal Calculations.....	129
Solving Cumulative Probability Problems.....	130
Solving Inverse Cumulative Problems.....	134
Checking Data for the Suitability of a Normal Model.....	136
Normal Quantile Plots.....	136
Generating Pseudo-Random Normal Data.....	140
Application.....	141
Chapter 8 Sampling and Sampling Distributions	145
Overview.....	145
Why Sample?.....	145
Methods of Sampling.....	146
Using JMP to Select a Simple Random Sample.....	147
Variability Across Samples: Sampling Distributions.....	150
Sampling Distribution of the Sample Proportion.....	150
From Simulation to Generalization.....	154
Sampling Distribution of the Sample Mean.....	156
The Central Limit Theorem.....	158
Stratification, Clustering, and Complex Sampling (optional).....	161
Application.....	164
Chapter 9 Review of Probability and Probabilistic Sampling	169
Overview.....	169
Probability Distributions and Density Functions.....	170
The Normal and t Distributions.....	170
The Usefulness of Theoretical Models.....	172
When Samples Surprise: Ordinary and Extraordinary Sampling Variability.....	174
Case 1: Sample Observations of a Categorical Variable.....	175
Case 2: Sample Observations of a Continuous Variable.....	176
Conclusion.....	179

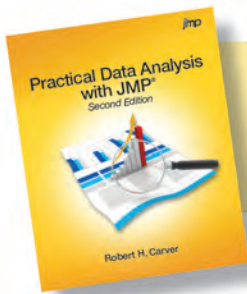
Chapter 10 Inference for a Single Categorical Variable	181
Overview.....	181
Two Inferential Tasks.....	182
Statistical Inference is Always Conditional	182
Using JMP to Conduct a Significance Test.....	183
Confidence Intervals	187
Using JMP to Estimate a Population Proportion	188
Working with Casewise Data.....	188
Working with Summary Data.....	190
A Few Words About Error.....	191
Application	191
Chapter 11 Inference for a Single Continuous Variable	197
Overview.....	197
Conditions for Inference.....	197
Using JMP to Conduct a Significance Test.....	198
More About P-Values	201
The Power of a Test.....	203
What if Conditions Aren't Satisfied?	205
Using JMP to Estimate a Population Mean	206
Matched Pairs: One Variable, Two Measurements	207
Application	209
Chapter 12 Chi-Square Tests	215
Overview.....	215
Chi-Square Goodness-of-Fit Test	216
What Are We Assuming?.....	218
Inference for Two Categorical Variables	219
Contingency Tables Revisited	219
Chi-Square Test of Independence.....	221
What Are We Assuming?.....	223
Application	224
Chapter 13 Two-Sample Inference for a Continuous Variable	227
Overview.....	227
Conditions for Inference.....	227

Using JMP to Compare Two Means.....	228
Assuming Normal Distributions or CLT.....	228
Using Sampling Weights (optional section)	231
Equal vs. Unequal Variances	232
Dealing with Non-Normal Distributions	233
Using JMP to Compare Two Variances	235
Application	237
Chapter 14 Analysis of Variance	241
Overview.....	241
What Are We Assuming?	241
One-Way ANOVA	243
Does the Sample Satisfy the Assumptions?	245
Factorial Analysis for Main Effects	248
What if Conditions Are Not Satisfied?.....	251
Including a Second Factor with Two-Way ANOVA	252
Evaluating Assumptions	254
Interaction and Main Effects	255
Application	259
Chapter 15 Simple Linear Regression Inference	265
Overview.....	265
Fitting a Line to Bivariate Continuous Data.....	266
The Simple Regression Model	269
Thinking About Linearity	270
Random Error	271
What Are We Assuming?	271
Interpreting Regression Results	272
Summary of Fit.....	272
Lack of Fit	273
Analysis of Variance	273
Parameter Estimates and t-tests	274
Testing for a Slope Other Than Zero	275
Application	278

Chapter 16 Residuals Analysis and Estimation	285
Overview.....	285
Conditions for Least Squares Estimation.....	286
Residuals Analysis	287
Linearity	289
Curvature	289
Influential Observations	291
Normality	292
Constant Variance	292
Independence.....	293
Estimation	295
Confidence Intervals for Parameters.....	296
Confidence Intervals for $Y X$	297
Prediction Intervals for $Y X$	298
Application	299
Chapter 17 Review of Univariate and Bivariate Inference	305
Overview.....	305
Research Context.....	306
One Variable at a Time.....	306
Life Expectancy by Income Group	307
Checking Assumptions	307
Conducting an ANOVA	310
Life Expectancy by GDP Per Capita	312
Summing Up	314
Chapter 18 Multiple Regression	315
Overview.....	316
The Multiple Regression Model	316
Visualizing Multiple Regression.....	316
Fitting a Model.....	319
A More Complex Model	322
Residuals Analysis in the Fit Model Platform.....	324

Collinearity	325
An Example Free of Collinearity Problems	326
An Example of Collinearity.....	329
Dealing with Collinearity	332
Evaluating Alternative Models	333
Application	335
Chapter 19 Categorical, Curvilinear, and Non-Linear Regression Models	339
Overview.....	339
Dichotomous Independent Variables.....	340
Dichotomous Dependent Variable.....	343
Whole Model Test.....	346
Parameter Estimates.....	346
Effect Likelihood Ratio Tests	346
Curvilinear and Non-Linear Relationships.....	347
Quadratic Models	347
Logarithmic Models.....	352
Application	356
Chapter 20 Basic Forecasting Techniques	361
Overview.....	361
Detecting Patterns Over Time	362
Smoothing Methods.....	365
Simple Moving Average	365
Simple Exponential Smoothing	367
Linear Exponential Smoothing (Holt's Method).....	369
Winters' Method.....	370
Trend Analysis	371
Autoregressive Models.....	373
Application	376
Chapter 21 Elements of Experimental Design	381
Overview.....	381
Why Experiment?	382
Goals of Experimental Design	382
Factors, Blocks, and Randomization	383

Multi-Factor Experiments and Factorial Designs	384
Blocking	391
Fractional Designs	393
Response Surface Designs	397
Application	400
Chapter 22 Quality Improvement	407
Overview.....	407
Processes and Variation.....	408
Control Charts	408
Charts for Individual Observations	409
Charts for Means	411
Charts for Proportions	415
Capability Analysis	418
Pareto Charts.....	421
Application	423
Appendix A Data Sources	427
Overview.....	427
Data Tables and Sources	427
Appendix B Data Management	431
Overview.....	431
Entering Data from the Keyboard.....	432
Moving Data from Excel Files into a JMP Data Table	437
Importing an Excel File from JMP.....	437
The JMP Add-in for Excel	439
Importing Data Directly from a Website	440
Combining Data from Two or More Sources	441
Bibliography	445
Index	449



From *Practical Data Analysis with JMP®*, Second Edition. Full book available for purchase [here](#).

15

Simple Linear Regression Inference

Overview	265
Fitting a Line to Bivariate Continuous Data	266
The Simple Regression Model	269
Thinking About Linearity	270
Random Error	271
What Are We Assuming?	271
Interpreting Regression Results	272
Summary of Fit	272
Lack of Fit	273
Analysis of Variance	273
Parameter Estimates and <i>t</i> -tests	274
Testing for a Slope Other Than Zero	275
Application	278

Overview

In Chapter 4, we learned to summarize two continuous variables at a time using scatterplot, correlations, and line fitting. In this chapter, we'll 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'll use is known as *linear regression analysis*. We'll devote this chapter and the three later chapters to the subject of regression.

Because Chapter 4 is now many pages back, we'll begin by reviewing some basic concepts of bivariate data and line fitting. Then, we'll discuss the fundamental model used in simple linear regression. After that, we'll discuss the crucial conditions necessary for inference, and finally, we'll 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 2005**. 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 2005. In this data table, the United Nations reports figures for 194 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.

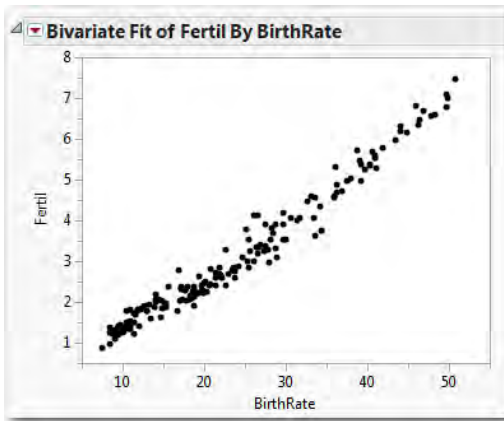
1. Open the **Birthrate 2005** data table now.

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



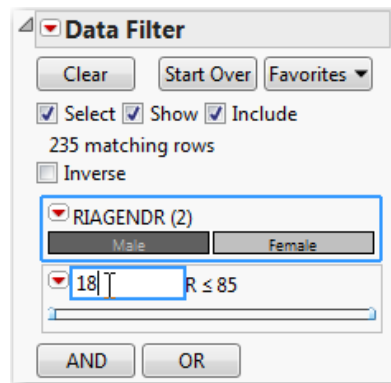
This is the same graph that we saw in Figure 4.10. Again, we note that general pattern is upward from left to right: fertility rates increase as the birth rate increases, although there are some countries that depart from the pattern. The pattern can be described as linear, although there is a mild curvature 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’ll return to a subset of the NHANES data¹, and look at two body measurement variables. Because adult body proportions are different from children and because males and females differ, we’ll restrict the first illustrative analysis to male respondents ages 18 and up. Our subset is a simple random sample of 465 observations drawn from the full NHANES data table, representing approximately 5% of the original data.

3. Open the data table called **NHANES SRS**. This table contains young and female respondents in addition to the males. To use only the males 18 years and older in our analysis, we’ll use the Data Filter.
4. Select **Rows ▶ Data Filter**.
5. While pressing the CTRL key, highlight **RIAGENDR** and **RIDAGEYR**, and click **Add**.
6. In the **Data Filter** (see Figure 15.2 after step 6 below), select the **Show** and **Include** options (the **Select** option is already selected).
7. Then click **Male** under **RIAGENDR** to include just the male subjects.
8. Finally, click the number 0 to the left **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've restricted the analysis to male respondents who are 18 years of age and older. Now we can begin the regression analysis. We'll 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'll 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.

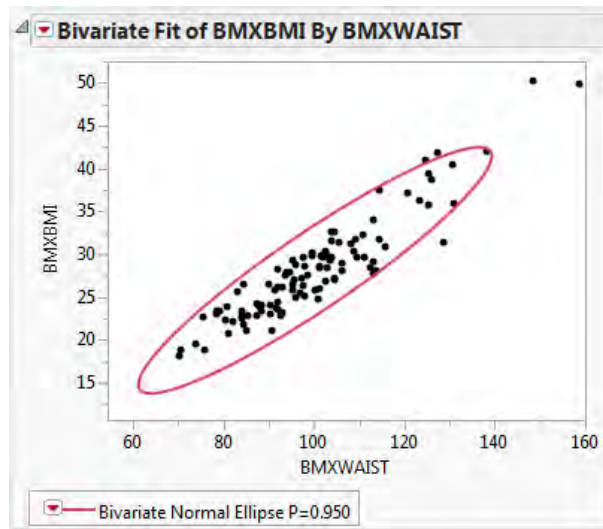
1. Select **Fit Y by X**. Cast **BMXBMI** as **Y** and **BMXWAIST** as **X** and click **OK**.

This graph (see Figure 15.3) 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.

2. 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 vs. Waist



3. 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 fairly 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 particular 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. Generally speaking, 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 on the next page. We see a green *fitted line* that approximates the upward pattern of the points.

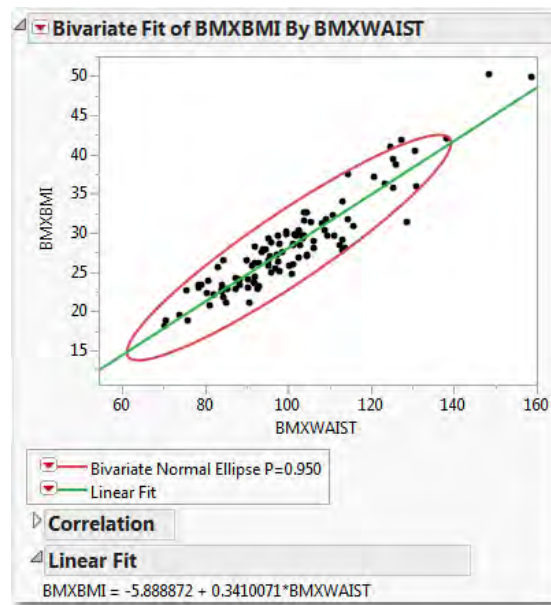
Below the graph, we find the equation of that line:

$$\text{BMXBMI} = -5.888872 + 0.3410071 * \text{BMXWAIST}$$

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 0.34 kg/m² 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.

Figure 15.4: Estimated Line of Best Fit



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 kind 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'll 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 doesn't 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 method² 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 of the implications of these conditions. Start by understanding that the following might be red flags to look for in a scatter plot with a fitted line:

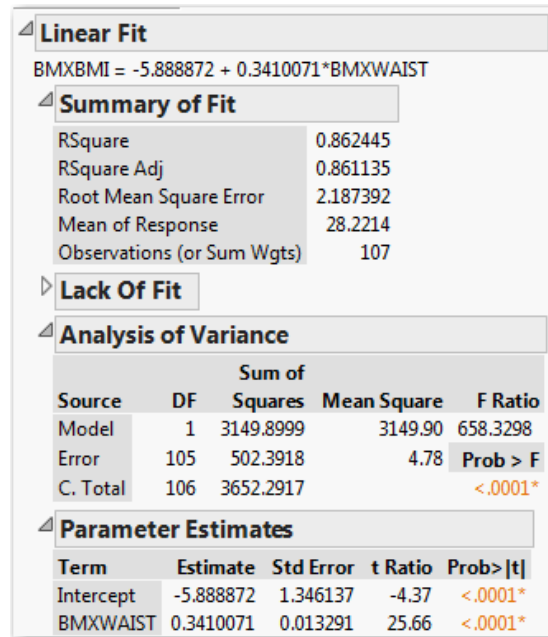
- The points seem to bend or oscillate predictably around the line.
- There are a small number of outliers that stand well apart from the mass of the points.
- The points seem snugly concentrated near one end of the line, but fan out toward the other end.
- There seem to be greater concentrations of 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'll 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've already seen the equation of the line of best fit and discussed its meaning. In this part of the chapter, we'll discuss the three other sections in order.

Figure 15.5: Regression Results



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 RSquare statistics range between 0.0 and 1.0, where 1.0 is a perfect linear fit. Just as in Chapter 4, think of

RSquare as the proportion of variation in Y that is associated with X . Here, both statistics are approximately 0.86, suggesting that a man'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 107 men.

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've 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 is to say 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 658 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 sampling error. We reject the null hypothesis that waist circumference and BMI are unrelated, and conclude that we've found a statistically significant relationship.

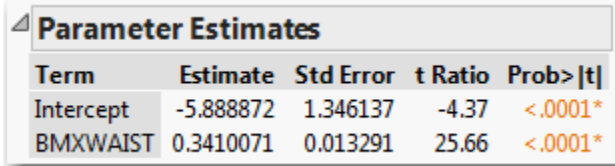
Not only can we say that the pattern describes the sample, we can say with confidence that the relationship generalizes to the entire population of men over age 17 in the United States.

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



Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-5.888872	1.346137	-4.37	<.0001*
BMXWAIST	0.3410071	0.013291	25.66	<.0001*

Because we're 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? The reason relates to the slope and 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 conclusion³ about the hypothesized independence of the response and factor variables. Here, we find that our estimated slope of 0.341 kg/m² change in BMI per 1 cm. increase in waist circumference is very convincingly different from 0: in fact, it's more than 25 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'll use it to test for a specific slope value other than 0.

We'll 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 consisting of two parts. In the first part, the composer introduces a melody—the basic tune of the piece—known formally as the exposition. After the exposition comes a second portion that elaborates upon the basic melody, developing it more fully, offering some variations, and then recapitulating or repeating the melody. Some music scholars believe that Haydn and Mozart strove for an aesthetically pleasing but asymmetric balance in the lengths of the exposition and development or recapitulation sections. More specifically, they might have divided their sonatas (deliberately or not) so that the relative lengths of the shorter and longer portions approximated the golden mean.

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 portions 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:

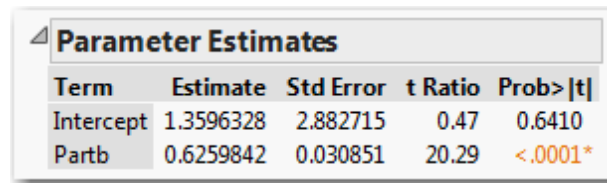
$$a = 0 + 0.61803(b)$$

So, we'll want to test the hypothesis that $\beta_1 = 0.61803$ 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



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'll want to compute the following ratio:

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

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

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

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 are able to 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.

Application

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

1. *Scenario:* Return to the **NHANES SRS** data table.
 - a. Exclude and hide respondents under age 18 and all males, leaving only adult females. 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 SRS** data table. For this analysis, we'll 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'll continue to examine the World Development Indicators data in **BirthRate 2005**. We'll 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 Rsquare 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 Rsquare 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 particular 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'll fit a random walk model to daily closing prices for McDonald's Corporation for the first six months of 2009 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're 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'll 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.

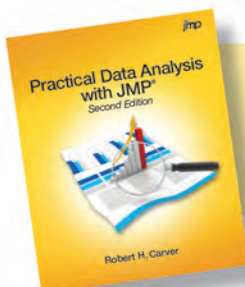
10. *Scenario:* Prof. Frank Anscombe of Yale University created an artificial data set to illustrate the hazards of applying linear regression analysis without looking at a scatterplot (Anscombe 1973). His work has been very influential, and JMP includes his illustration among the sample data tables 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'll 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'll begin by analyzing the data for one model: the Civic EX. Use the **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'll return to the World Development Indicators data in **WDI**. In this scenario, we'll 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 **Data Filter** to **Show** and **Include** only the observations for the **Year 2010**, and the **Latin America & Caribbean Region** nations. Describe the relationship you observe between access to improved sanitation and life expectancy.
 - b. Repeat the analysis for East Asia & Pacific countries in 2010.
 - c. Now do the same one additional time 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.
 - 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 Why a subsample? Some of the key concepts in this chapter deal with the way individual points scatter around a line. With a smaller number of observations, we'll be able to better visualize these concepts.
- ² 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.
- ³ They will have identical *P*-values and the F-ratio will be the square of the *t* ratio.

From *Practical Data Analysis with JMP®, Second Edition* by Robert H. Carver. Copyright © 2014, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.



From *Practical Data Analysis with JMP®, Second Edition*. Full book available for purchase [here](#).

Index

A

alpha (α) 203–205
alternative hypothesis 183–184
alternative models, evaluating 333–335
analysis
 See also specific types
 capability 418–421
 with Fit Y by X 97–99
 with multivariate platform 96–97
 trend 371–373
analysis of variance (ANOVA)
 about 241
 applications 259–264
 assumptions about 241–243
 conducting 310–311
 interpreting regression results 273
 one-way 243–251
 satisfaction of conditions 251–252
 two-way 252–259
analysis platform, using 12–14
analytics frameworks, applying 88–89
ANOVA
 See analysis of variance (ANOVA)
applications
 analysis of variance (ANOVA) 259–264
 chi-square tests 224–226
 data 36–37, 57–61
 discrete distributions 120–123
 experimental design 400–406
 forecasting techniques 376–380
 inference 191–196, 209–213, 237–240
 linear regression analysis 278–284
 multiple regression 335–338
 normal model 141–144

 probability 120–123
 quality improvement 423–426
 regression analysis 356–360
 residuals analysis 299–304
 residuals estimation 299–304
 sampling and sampling distributions 164–168
 variables 79–83
applying analytics frameworks 88–89
ARIMA (AutoRegressive Integrated Moving Average) models 373–376
assigning probability values 107–108
assumptions
 See also inference, conditions for
 about analysis of variance (ANOVA) 241–243
 evaluating 254–255
asterisk (*) 187
autocorrelation 362–363, 364
AutoRegressive Integrated Moving Average (ARIMA) models 373–376
autoregressive models 373–376
axes, customizing in histograms 52–53

B

bars, customizing in histograms 52–53
beta (β) 203–205
"Big Data" 26
binomial distribution 113–115
bivariate data 63–64
bivariate inference
 about 305–307
 life expectancy by GDP per capita 312–314

- life expectancy by income group 307–311
 - research context 306
- blocking 391–393
- blocks 383–384
- bootstrapping 206
- Box, George 373, 382–383
- box plot 47
- box-and-whiskers plot 47
- bubble plots 78–79
- C**
- capability analysis 418–421
- cases 22
- casewise data 188–189
- categorical 3
- categorical regression models 339
 - See also* regression analysis
- categorical variables
 - distributions of 41–47
 - inference for 181–195
 - inference for two 219
 - one continuous variable and one 71–73
 - sample observations of 175–176
 - two 65–71
- center of distributions 47, 48
- Central Limit Theorem (CLT) 158–161, 228–231
- central tendency, of distributions 47, 48
- Chart command 44–46
- checking data for suitability of normal model 136–140
- chi-square distribution 186, 215
- chi-square tests
 - about 215
 - applications 224–226
 - contingency tables 219–221
 - goodness-of-fit test 216–219
 - of independence 221–223
 - inference for two categorical variables 219
- Classical method, of assigning probabilities 107
- CLT (Central Limit Theorem) 158–161, 228, 231
- clustering 161–163
- collinearity
 - about 325
 - dealing with 332–333
 - example 326–332
- column properties 7
- Column Switcher 95
- columns, of data tables 3
- combining data from sources 441–444
- comparing
 - two means with JMP 228–235
 - two variances with JMP 235–236
- complement of an event 105
- complex sampling 161–163
- conditional probability 106
- conditional values 221
- conducting
 - analysis of variance (ANOVA) 310–311
 - significance testing with JMP 183–187, 198–205
- confidence band 297
- confidence intervals
 - about 187–188
 - estimating 182
 - for parameters 296–297
 - for $Y|X$ 297–298
- confidence limits 57
- constant variance 292–293
- contingency tables
 - about 219–221
 - displaying covariation in categorical variables 68–71
 - probability and 108–110
- continuous columns 3–4
- continuous data
 - fitting lines to bivariate 266–269
 - probability and 125–126
 - using Distribution platform for 47–51
- continuous variables
 - inference for single 197–213
 - one categorical variable and 71–73
 - sample observations of 176–178
 - two 73–77
 - two-sample inference for 227–240

- control charts
 - about 408–409
 - for individual observations 409–411
 - for means 411–415
 - for proportions 415–418
 - control limits 410, 412
 - correlation 77
 - covariation
 - one continuous, one categorical variable 71–73
 - two categorical variables 65–71
 - two continuous variables 73–77
 - creating
 - data tables 5–9, 34
 - pseudo-random normal data 140–141
 - cross-section 23
 - cross-sectional data 88
 - cross-sectional sampling 27
 - crosstabulation 68–71
 - CTRL key 137
 - cumulative probabilities 115, 130–134
 - curvature 289
 - curvilinear regression models 339
 - See also* regression analysis
 - curvilinear relationships 347–356
 - customizing histograms 52–53
 - cycle pattern 362
- D**
- data
 - See also* continuous data
 - applications 36–37, 57–61
 - bivariate 63–64
 - casewise 188–189
 - checking for suitability of normal model 136–140
 - combining from sources 441–444
 - cross-sectional 88
 - entering from keyboards 432–437
 - experimental 27–31
 - importing directly from websites 440–441
 - longitudinal 88
 - matched pairs of 207–209
 - moving from Excel files to JMP data tables 437–440
 - observational 31, 88
 - panel 23
 - populations 22–23
 - processes 22–23
 - raw case data 34–36
 - representativeness 23–26
 - samples and sampling 22–26
 - study design 27–34
 - summary 34–36, 190
 - survey 31–34
 - time-series 88
 - types of 3–4, 89
 - data analysis
 - goals of 2
 - role of probability in 104
 - data dictionary 32
 - Data Filter tool 43–44
 - Data Grid area 8
 - data management
 - See also* data 431
 - data sources 427–429
 - data tables
 - about 3
 - creating 5–9, 34
 - moving data from Excel files to JMP 437–440
 - degrees of freedom (DF) 217
 - density functions 126–128, 170
 - description 2
 - descriptive statistics
 - about 85–86
 - analysis with Fit Y by X 97–99
 - analysis with multivariate platform 96–97
 - applying analytics frameworks 88–89
 - data source and structure 88
 - exploring relationship with Graph Builder 93–96
 - interpretation 99
 - observational units 89
 - preparation for analysis 90

- questions for analysis 87–88
 - univariate descriptions 90–92
 - variables and data types 89
 - visualizing multiple relationships 100–101
 - World Development Indicators (WDI) 86–87
 - detecting patterns 362–365
 - DF (degrees of freedom) 217
 - dichotomous dependent variables 343–346
 - dichotomous independent variables 340–342
 - disclosure button 8–9
 - discrete distributions
 - about 103
 - applications 120–123
 - as models of real processes 118–119
 - discrete random variables
 - about 111
 - three common 111–116
 - dispersion, of distributions 47, 48
 - Distribution command 173
 - Distribution platform, for continuous data 47–51
 - "distribution-free" methods 223
 - distributions
 - See also* discrete distributions
 - binomial 113–115
 - of categorical variables 41–47
 - center of 47, 48
 - central tendency of 47, 48
 - chi-square 186, 215
 - dispersion of 47, 48
 - Hypergeometric 180
 - integer 112–113
 - non-normal 233–235
 - normal 170–172, 228–231
 - Poisson 115–116
 - probability 111, 170
 - of quantitative variables 47–57
 - theoretical discrete 111
 - of variables 40–41
 - dummy variables 340–342
 - Dunnett's method 249
- E**
- effect likelihood ratio tests 346
 - equal variances, compared with unequal variances 232
 - error 191
 - estimating
 - confidence intervals 182
 - population means with JMP 206–207
 - population proportions with JMP 188–190
 - evaluating
 - alternative models 333–335
 - assumptions 254–255
 - events
 - probability of 105
 - rules for two 106–107
 - Excel
 - JMP Add-in for 439–440
 - moving data to JMP data tables from files in 437–440
 - excluded rows 15
 - expected frequency 219
 - experimental data 27–31
 - experimental design
 - about 381–382
 - applications 400–406
 - blocks and blocking 383–384, 391–393
 - factorial designs 384–391
 - factors 383–384
 - fractional designs 393–397
 - goals of 382–383
 - multi-factor experiments 384–391
 - randomization 383–384
 - reasons for experimenting 382
 - response surface designs 397–400
 - experimental runs 384
 - exporting JMP results to word-processor documents 17–18
 - extraordinary sampling variability 174–178

F

- factor profiles 256
- factorial analysis 248–251
- factorial designs 384–391
- factors 383–384
- Fit Model platform, residuals analysis in 324–325
- Fit Y by X, analysis with 97–99
- fitted line 77
- fitting 12
- five-number summary 56
- fly ash 382
- forecasting techniques
 - about 361
 - applications 376–380
 - autoregressive models 373–376
 - detecting patterns 362–365
 - smoothing methods 365–371
 - trend analysis 371–373
- fractional designs 393–397
- frequency of values 47
- full factorial experimental design 385–391

G

- Gaussian density function 128
- generalization, simulation to 154–155
- golden mean 275
- goodness-of-fit test 216–219
- Gosset, William 207
- Grabber 52
- Graph Builder
 - about 9–12
 - exploring categorical data with 46–47
 - exploring data with 75
 - exploring relationships with 93–96
 - using 54–55
- graphing categorical data 44–46
- graphs, linked 51

H

- Hand tool 52
- Haydn, Franz Joseph 275
- Help tool 254
- heterogeneity of variance 292–293
- heteroskedasticity 292–293, 324–325
- hidden rows 15
- histograms 47, 52–53
- Holt, Charles 369
- Holt's Method 369–370
- homogeneity 235
- homogeneity of variance 292–293
- homoskedasticity 292–293
- Hypergeometric distribution 180
- hypothesis testing 182

I

- IIP (Index of Industrial Production) 362
- importing
 - data directly from websites 440–441
 - Excel files from JMP 437–439
- independence
 - about 293–295
 - chi-square tests of 221–223
- independent events 107
- Index of Industrial Production (IIP) 362
- indicator variables 340–342
- individual observations, charts for 409–411
- inference
 - See also* bivariate inference
 - See also* linear regression analysis
 - See also* univariate inference
 - about 2, 197, 227
 - applications 191–196, 209–213, 237–240
 - comparing two means with JMP 228–235
 - comparing two variances with JMP 235–236
 - conditional status of statistical 182–183
 - conditions for 197–198, 227–228
 - conducting significance testing with 183–187

- conducting significance testing with JMP 198–205
 - confidence interval estimation 182, 187–188
 - estimating population means with JMP 206–207
 - estimating population proportions with JMP 188–190
 - matched pairs 207–209
 - satisfying conditions 205–206
 - for single categorical variable 181–195
 - for single continuous variable 197–213
 - for two categorical variables 219
 - two-sample 227–240
 - influential observations 289–291
 - integer distribution 112–113
 - interaction effect 252, 255–259
 - interpretation 99
 - interpreting regression results 272–278
 - interquartile range (IQR) 57
 - inverse cumulative problems, solving 134–136
 - IQR (interquartile range) 57
 - irregular pattern 362
- J**
- Jenkins, Gwilym 373
 - jitter 10, 238
 - JMP
 - See also specific topics*
 - Add-in, for Excel 439–440
 - comparing two means with 228–235
 - comparing two variances with 235–236
 - conducting significance testing with 183–187, 198–205
 - estimating population means with 206–207
 - estimating population proportions with 188–190
 - exporting results to word-processor documents 17–18
 - leaving 19
 - selecting simple random samples with 147–150
 - simulating random variation with 116–118 starting 4–5
 - JMP Scripting Language (JSL) 150
 - joint probability 106
 - joint relative frequency 221
 - joint-frequency table 68–71
 - JSL (JMP Scripting Language) 150
- K**
- KDD (Knowledge Discovery in Databases) 427
 - key fields 442
 - Knowledge Discovery in Databases (KDD) 427
 - Kruskal-Wallis Test 234
- L**
- label property 7
 - labeled rows 15
 - Lack of Fit 273
 - least squares estimation, conditions for 286
 - leaving JMP 19
 - linear exponential smoothing (Holt's Method) 369–370
 - linear regression analysis
 - about 265
 - applications 278–284
 - assumptions of 271–272
 - fitting lines to bivariate continuous data 266–269
 - interpreting regression results 272–278
 - simple regression model 269–271
 - linearity 270, 287–289
 - linked graphs/tables 51
 - logarithmic growth 312
 - logarithmic models 352–356
 - longitudinal data 88
 - longitudinal sampling 27
 - lower fences 57

M

Mann-Whitney U Test 234
margin of error 189
matched pairs 207–209
MDGs (Millennium Development Goals) 86–87
means
 comparing two with JMP 228–235
 control charts for 411–415
metadata 6
Millennium Development Goals (MDGs) 86–87
missing data 65, 66
model specification 339
modeling types 3
modifying analysis 67
Mozart, Wolfgang Amadeus 275
multicollinearity 325
multi-factor experiments 384–391
multiple regression
 about 315
 applications 335–338
 collinearity 325–333
 evaluating alternative models 333–335
 fitting a model 319–321
 model 316, 322–323
 residuals analysis in Fit Model platform
 324–325
 visualizing 316–319
multivariate platform, analysis with 96–97
mutually exclusive events 106

N

National Health and Nutrition Examination
 Survey (NHANES) 32
nominal columns 4
non-linear regression models 339
 See also regression analysis
non-linear relationships 347–356
non-normal distributions, comparing two means
 with JMP 233–235
nonparametric equivalent test 251–252
nonparametric methods 223
non-parametric test 205

non-random sampling 26
normal density function 128
normal distributions 170–172, 228–231
normal model
 about 125, 128–129
 applications 141–144
 checking data for suitability of 136–140
 continuous data and probability 125–126
 density functions 126–128
 generating pseudo-random normal data
 140–141
 normal calculations 129–136
Normal Probability Plot (NPP) 136–140
Normal Quantile function 135
Normal Quantile Plots 136–140
normality 291–292, 314
NPP (Normal Probability Plot) 136–140
null hypothesis 184

O

observational data 31, 88
observational units 22, 89
observations 3
one-way analysis of variance (ANOVA) 243–
 251
optimization 383
ordinal columns 4
ordinary least squares estimation (OLS) 284n2
ordinary sampling variability 174–178
outlier box plots 56–57
overlap marks 245

P

panel data 23
panel studies 27
panning axes 53
parameter estimates 274, 346
parameters, confidence intervals for 296–297
Pareto charts 421–423
patterns, detecting 362–365
percentiles 55–56

Pipeline and Hazardous Materials Program
(PHMSA) 118–119

Poisson distribution 115–116

polynomial functions 347

population means, estimating with JMP 206–
207

population proportions, estimating with JMP
188–190

populations 2, 22–23

post-stratification weights 162

power of a test 203–205

predictability, of risks 23

prediction bands 298

prediction intervals, for $Y|X$ 298

Prediction Variance Profile Plot 398

primitives 101

probability and probabilistic sampling
about 103, 169
applications 120–123
assigning values 107–108
contingency tables and 108–110
continuous data and 125–126
cumulative probabilities 115, 130–134
events, probability of 105
extraordinary sampling variability 174–178
normal distributions 170–172
ordinary sampling variability 174–178
probability distributions and density
functions 170
role of in data analysis 104
t distributions 170–172
usefulness of theoretical models 172–174

probability distributions 111, 170

probability of an event ($\Pr(A)$) 105

probability theory 104–108

process capability 418–419

processes
about 22–23
in quality improvement 408

proportions, charts for 415–418

pseudo-random normal data, generating 140–
141

p-value 186–187, 191, 201–202

Q

quadratic models 347–351

quality improvement
about 407
applications 423–426
capability analysis 418–421
control charts 408–418
Pareto charts 421–423
processes 408
variation in 408

quantile 55

quantitative 3

quantitative variables, distributions of 47–57

R

random error 271

Random function 140–141

random variation, simulating with JMP 116–118

randomization 23, 383–384

Rasmussen, Marianne 104–105, 110

raw case data 34–36

red triangles 6, 99, 137

regression analysis
See also multiple regression
applications 356–360
curvilinear relationships 347–356
dichotomous dependent variable 343–346
dichotomous independent variables 340–
342
interpreting results 273
non-linear relationships 347–356

relationships
curvilinear 347–356
exploring with Graph Builder 93–96
non-linear 347–356
visualizing multiple 100–101

Relative Frequency method, of assigning
probabilities 107

re-launching analysis 67

representativeness, of data 23–26

residuals, normality in 314

residuals analysis

- about 285, 286–287
 - applications 299–304
 - conditions for least squares estimation 286
 - constant variance 292–293
 - curvature 289
 - in Fit Model platform 324–325
 - independence 293–295
 - influential observations 289–291
 - linearity 287–289
 - normality 291–292
 - residuals estimation
 - about 285, 295
 - applications 299–304
 - conditions for least squares estimation 286
 - confidence intervals for parameters 296–297
 - confidence intervals for $Y|X$ 297–298
 - prediction intervals for $Y|X$ 298
 - response combinations, to bivariate data 64
 - response surface 384
 - response surface designs 397–400
 - row states 14–17
 - Rsquare (r^2) 77
 - Run Chart 362, 409
 - Rydén, Jesper 275
- S**
- sales lift 397
 - sample mean, sampling distribution of 156–158
 - sample proportion, sampling distribution of 150–154
 - sampling and sampling distributions
 - about 22–23, 23–24, 145, 174–175
 - applications 164–168
 - Central Limit Theorem (CLT) 158–161
 - clustering 161–163
 - complex sampling 161–163
 - cross-sectional sampling 27
 - defined 2
 - methods of sampling 146–147
 - non-random 26
 - reasons for sampling 145–146
 - of sample mean 156–158
 - simple random sampling (SRS) 24–26, 146–147, 147–150
 - from simulation to generalization 154–155
 - stratification 161–163
 - time series sampling 23, 27
 - using JMP to select simple random samples 147–150
 - variability across samples 150–163
 - sampling error 23
 - sampling frame 24, 147
 - sampling variability, ordinary and extraordinary 174–178
 - sampling weights, comparing two means with JMP 231–232
 - saving 18
 - scatterplot 74–75, 78–79
 - screening 383
 - script 150
 - seasonal pattern 362
 - selected rows 15
 - session script, saving 18
 - shadowgrams 52–53, 128
 - shape, of distributions 47, 48
 - Shewhart, Walter 426n1
 - Shewhart Charts
 - See* control charts
 - shortest half bracket 57
 - sidereal period of orbit 347–348
 - significance testing
 - about 182
 - conducting with JMP 183–187, 198–205
 - simple exponential smoothing 367–369
 - Simple Moving Average 365–366
 - simple random sampling (SRS) 24–26, 146–147, 147–150
 - simple regression model 269–271
 - simulating
 - to generalization 154–155
 - random variation with JMP 116–118
 - smoothing methods
 - about 365

linear exponential smoothing (Holt's Method) 369–370
 simple exponential smoothing 367–369
 Simple Moving Average 365–366
 Winters' Method 370–371
 solving
 cumulative probability problems 130–134
 inverse cumulative problems 134–136
 split plot experiment 207
 SRS (simple random sampling) 24–26, 146–147, 147–150
 standard deviation 56
 standard error 159
 Standard Normal Distribution 128–129
 starting JMP 4–5
 stationary time-series 364
 statistics
 See descriptive statistics
 stratification 161–163
 study design 27–34
 Subjective method, of assigning probabilities 108
 summary data 34–36, 190
 Summary of Fit 272–273
 summary statistics, for single variables 55–56
 survey data 31–34

T

t distributions 159–161, 170–172
 Table variable note 7
 tables, linked 51
 See also data tables
 tails, in continuous distributions 132
 Test Means command 205
 testing, for slopes other than zero 275–278
 theoretical discrete distribution 111
 time series sampling 23, 27
 time-series data 88
 transforming the variable 312
 treatment effect 242
 trend analysis 371–373
 trend pattern 362

t-tests 274
 Tukey's HSD (Honestly Significant Difference) 249, 251–252
 two-sample inference, for continuous variables 227–240
 two-way analysis of variance 252–259
 two-way table 68–71
 Type I error 191
 Type II error 191

U

unequal variances, compared with equal variances 232
 uniform scaling option 51
 union of two events 106
 univariate descriptions 90–92
 univariate inference
 about 305–306, 306–307
 life expectancy by GDP per capita 312–314
 life expectancy by income group 307–311
 research context 306
 unusual observations, of distributions 47, 48–51
 upper fences 57

V

values
 assigning probability 107–108
 frequency of 47
 variability, across samples 150–163
 variables
 See also bivariate data
 See also categorical variables
 See also continuous variables
 about 39
 applications 79–83
 defined 3
 descriptive statistics 89
 dichotomous dependent 343–346
 dichotomous independent 340–342
 distributions of 40–41
 dummy 340–342

- indicator 340–342
 - quantitative 47–57
 - summary statistics for single 55–56
 - transforming 312
 - types of 40–41
- variance
 - heterogeneity of 292–293
 - homogeneity of 292–293
- variances, comparing two with JMP 235–236
- variation, in quality improvement 408
- visualizing
 - multiple regression 316–319
 - multiple relationships 100–101

W

- WDI (World Development Indicators) 86–87
- websites
 - data sources 427–429
 - importing data directly from 440–441
- weighting 161
- Welch's test 246
- whiskers 57
- whole model test 346
- Wilcoxon Signed Rank Test 205
- Wilson Estimator 189
- Winters, Peter 370
- Winters' Method 370–371
- word-processor documents, exporting JMP
 - results to 17–18
- World Development Indicators (WDI) 86–87

Y

- Y-hat 297
- $Y|X$, confidence intervals for 297–298
- $Y|X$, prediction intervals for 298

Z

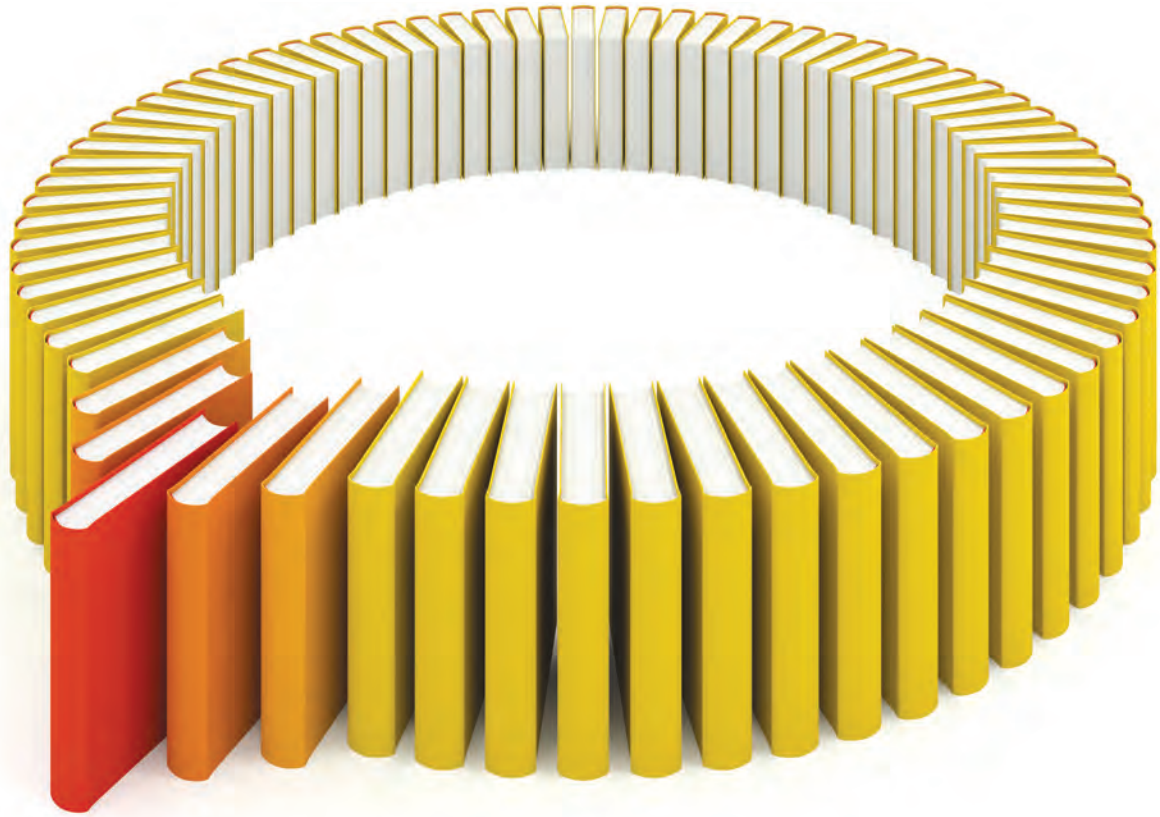
- z-scores 128–129

About The Author



Robert Carver is Professor of Business Administration at Stonehill College in Easton, Massachusetts, and Adjunct Professor at the International Business School at Brandeis University in Waltham, Massachusetts. At both institutions, he teaches courses on business analytics in addition to general management courses, and has won teaching awards at both schools. His primary research interest is statistics education. A JMP user since 2006, Carver holds an A.B. in political science from Amherst College in Amherst, Massachusetts and an M.P.P. and Ph.D. in public policy from the University of Michigan at Ann Arbor.

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



Gain Greater Insight into Your JMP[®] Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

 support.sas.com/bookstore
for additional books and resources.

