

REGRESSION ANALYSIS

The background of the cover is a dark blue gradient. Overlaid on this are several abstract elements: a series of vertical blue lines of varying heights, a white line graph with circular markers, and a blue line graph with circular markers. The white line graph starts at the bottom left, rises to a peak, falls, and then rises again. The blue line graph starts at the bottom left, rises steadily, and then levels off. The overall effect is one of data analysis and statistical modeling.

An Intuitive Guide
for Using and
Interpreting Linear Models

JIM FROST, MS

Regression Analysis

AN INTUITIVE GUIDE FOR USING
AND INTERPRETING LINEAR MODELS



Jim Frost

Copyright © 2019 by Jim Frost.

All rights reserved. No part of this publication may be reproduced, distributed or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, without the prior written permission of the publisher, except in the case of brief quotations embodied in critical reviews and certain other noncommercial uses permitted by copyright law.

To contact the author, please email: statisticsbyjim@gmail.com.

Visit the author's website at statisticsbyjim.com.

Ordering Information:

Quantity sales. Special discounts are available on quantity purchases by educators. For details, contact the email address above.

Regression Analysis / Jim Frost. —1st ed.

Contents

My Approach to Teaching Regression and Statistics	13
Correlation and an Introduction to Regression	16
Graph Your Data to Find Correlations	17
Interpret the Pearson's Correlation Coefficient	18
Graphs for Different Correlations	19
Discussion about the Correlation Scatterplots	22
Pearson's Correlation Coefficient Measures Linear Relationships.....	23
Hypothesis Test for Correlations	24
Interpreting our Height and Weight Correlation Example	24
Correlation Does Not Imply Causation	25
How Strong of a Correlation is Considered Good?	25
Common Themes with Regression	26
Taking Correlation to the Next Level with Regression	27
Fundamental Terms and Goals of Regression	27

Regression Analyzes a Wide Variety of Relationships	29
Using Regression to Control Independent Variables ..	31
An Introduction to Regression Output.....	32
Review and Next Steps.....	33
Regression Basics and How it Works.....	35
Data Considerations for OLS.....	36
How OLS Fits the Best Line.....	37
Implications of Minimizing SSE.....	42
Other Types of Sums of Squares.....	43
Displaying a Regression Model on a Fitted Line Plot..	45
Importance of Staying Close to Your Data.....	46
Review and Next Steps.....	48
The Chapters/Sections below are only in the full ebook.	
Interpreting Main Effects and Significance	49
Regression Notation	50
Fitting Models is an Iterative Process.....	51
Three Types of Effects in Regression Models	52
Main Effects for Continuous Variables with Straight-Line Relationships.....	53
Recoding Your Continuous Independent Variables....	61
Main Effects of Categorical Variables.....	64

Variables that Blur the Lines Between Continuous and Categorical.....	74
Constant (Y Intercept).....	76
Review and Next Steps.....	84
Fitting Curvature.....	85
Example Curvature	86
Use Main Effects Plots to Graph Curvature in Multiple Regression	88
Why You Need to Fit Curves in a Regression Model..	90
The Difference between Linear and Nonlinear Models	92
Finding the Best Way to Model Curvature	95
Another Curve Fitting Example.....	107
Review and Next Steps.....	112
Interaction Effects.....	113
Example of Interaction Effects with Categorical Independent Variables	114
How to Interpret Interaction Effects.....	115
Overlooking Interaction Effects is Dangerous!	117
Example of an Interaction Effect with Continuous Independent Variables	118
Important Considerations for Interaction Effects	120

Common Questions about Interaction Effects	121
Review and Next Steps.....	126
Goodness-of-Fit	127
Assessing the Goodness-of-Fit.....	127
R-squared.....	128
Visual Representation of R-squared.....	129
R-squared has Limitations	131
Are Low R-squared Values Always a Problem?	131
Are High R-squared Values Always Great?	132
R-squared Is Not Always Straightforward	134
Adjusted R-Squared and Predicted R-Squared.....	134
A Caution about the Problems of Chasing a High R-squared.....	140
Standard Error of the Regression vs. R-squared	141
The F-test of Overall Significance	146
Review and Next Steps.....	148
Specify Your Model.....	150
The Importance of Graphing Your Data	151
Statistical Methods for Model Specification	153
Real World Complications in the Model Specification Process.....	155
Practical Recommendations for Model Specification	156

Omitted variable Bias.....	158
Automated Variable Selection Procedures: Stepwise Regression and Best Subsets Regression	169
Which is Better, Stepwise Regression or Best Subsets Regression?.....	176
Review and Next Steps.....	181
Problematic Methods of Specifying Your Model.....	183
Using Data Dredging and Significance to Pick Models	184
Overfitting Regression Models	190
Review and Next Steps.....	195
Checking Assumptions and Fixing Problems.....	197
Check Your Residual Plots to Ensure Trustworthy Regression Results!.....	198
The Seven Classical OLS Assumptions	203
Heteroscedasticity	213
Multicollinearity.....	222
Unusual Observations.....	232
Using Data Transformations to Fix Problems	244
Cheat Sheet for Detecting and Solving OLS Problems	250
Using Regression to Make Predictions	255

The Difference Between Explanatory Models and Predictive Models	256
The Regression Approach for Predictions.....	258
Example Scenario for Regression Predictions.....	259
Finding a Good Regression Model for Predictions....	259
The Illusion of Predictability	266
Different Example of Using Prediction Intervals	274
Tips, Common Questions, and Concerns.....	279
Five Regression Analysis Tips to Avoid Common Problems.....	280
Identifying the Most Important Independent Variables in Regression Models.....	285
Comparing Regression Lines with Hypothesis Tests	291
How High Does R-squared Need to Be?.....	296
Five Reasons Why Your R-squared can be Too High	301
How to Interpret Regression Models that have Significant Variables but a Low R-squared	306
Choosing the Correct Type of Regression.....	315
Regression Analysis with Continuous Dependent Variables.....	316
Regression Analysis with Categorical Dependent Variables.....	319

Regression Analysis with Count Dependent Variables	321
Examples of Other Types of Regression	323
Using Log-Log Plots to Determine Whether Size Matters	323
Binary Logistic Regression: Statistical Analysis of the Republican Establishment Split	330
References.....	337
About the Author.....	338

*To Carmen and Morgan who made this book possible through
their encouragement and support.*

The best thing about being a statistician is that you get to play
in everyone's backyard.

—John Tukey

INTRODUCTION



My Approach to Teaching Regression and Statistics

NOTE: This sample contains only the introduction and first two chapters. Please buy the full ebook for all the content listed in the Table of Contents. You can buy it in [My Store](#).

I love statistics and analyzing data! I also love talking and writing about it. I was a researcher at a major university. Then, I spent over a decade working at a major statistical software company. During my time at the statistical software company, I learned how to present statistics in a manner that makes it more intuitive. I want you to understand the essential concepts, practices, and knowledge for regression analysis so you can analyze your data confidently. That's the goal of my book.

In this book, you'll learn many facets of regression analysis including the following:

- How regression works and when to use it.

- Selecting the correct type of regression analysis.
- Specifying the best model.
- Interpreting the results.
- Assessing the fit of the model.
- Generating predictions and evaluating their precision.
- Checking the assumptions.
- Examples of different types of regression analyses.

I'll help you intuitively understand regression analysis by focusing on concepts and graphs rather than equations and formulas. I use regular, everyday language so you can grasp the fundamentals of regression analysis at a deeper level. I'll provide practical tips for performing your analysis. You will learn how to interpret the results while being confident that you're conducting the analysis correctly. You'll be able to trust your results because you'll know that you're performing regression properly and know how to detect and correct problems.

Regardless of your background, I will take you through how to perform regression analysis. Students, career changers, and even current analysts looking to take your skills to the next level, this book has absolutely everything you need to know for regression analysis.

I've literally received thousands of requests from aspiring data scientists for guidance in performing regression analysis. This book is my answer - years of knowledge and thousands of hours of hard work distilled into a thorough, practical guide for performing regression analysis.

You'll notice that there are not many equations in this book. After all, you should let your statistical software handle the calculations so you don't get bogged down in the calculations and can instead focus on understanding your results. Instead, I focus on the concepts and practices that you'll need to know to perform the analysis and interpret the results correctly. I'll use more graphs than equations!

Don't get me wrong. Equations are important. Equations are the framework that makes the magic, but the truly fascinating aspects are what it all means. I want you to learn the true essence of regression analysis. If you need the equations, you'll find them in most textbooks.

Please note that throughout this book I use Minitab statistical software. However, this book is not about teaching particular software but rather how to perform regression analysis. All common statistical software packages should be able to perform the analyses that I show. There is nothing in here that is unique to Minitab.



Correlation and an Introduction to Regression

Before we tackle regression analysis, we need to understand correlation. In fact, I've described regression analysis as taking correlation to the next level! Many of the practices and concepts surrounding correlation also apply to regression analysis. It's also a simpler analysis that is a more familiar subject for many. Bear with me because the correlation topics in this section apply to regression analysis as well. It's a great place to start!

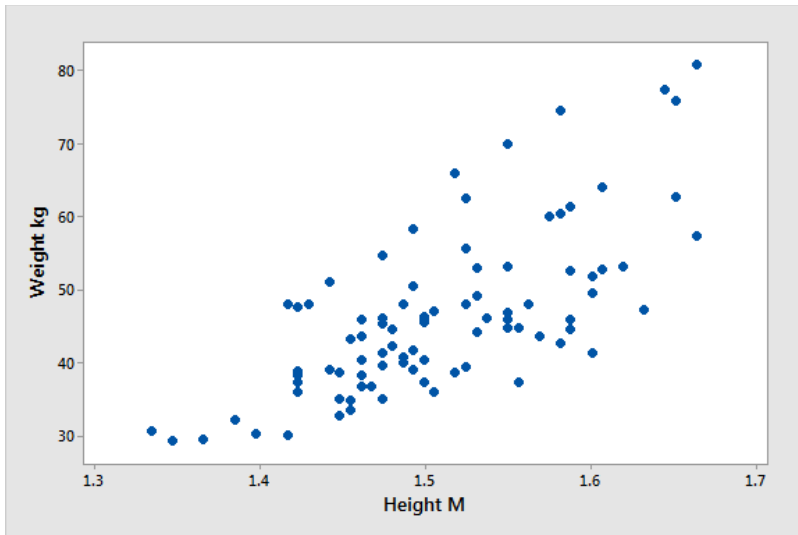
A correlation between variables indicates that as one variable changes in value, the other variable tends to change in a specific direction. Understanding that relationship is useful because we can use the value of one variable to predict the value of the other variable. For example, height and weight are correlated—as height increases, weight also tends to increase. Consequently, if we observe an individual who is unusually tall, we can predict that his weight is also above the average. In statistics, correlation is a quantitative assessment that measures both the direction and the strength of this tendency to vary together.

There are different types of correlation that you can use for different kinds of data. In this chapter, I cover the most common type of correlation—Pearson’s correlation coefficient.

Before we get into the numbers, let’s graph some data first so we can understand the concept behind what we are measuring.

Graph Your Data to Find Correlations

Scatterplots are a great way to check quickly for relationships between pairs of continuous data. The scatterplot below displays the height and weight of pre-teenage girls. Each dot on the graph represents an individual girl and her combination of height and weight. These data are real data that I collected during an experiment. We’ll return to this dataset multiple times throughout this book. Here is the CSV dataset if you want to try it yourself: [HeightWeight](#).



At a glance, you can see that there is a relationship between height and weight. As height increases, weight also tends to increase. However, it’s not a perfect relationship. If you look at a specific height, say 1.5 meters, you can see that there is a range of weights associated with it. You can also find short people who weigh more than taller people.

However, the general tendency that height and weight increase together is unquestionably present.

Pearson's correlation takes all of the data points on this graph and represents them with a single summary statistic. In this case, the statistical output below indicates that the correlation is 0.705.

Correlation: Height M, Weight kg

```
Pearson correlation of Height M and Weight kg = 0.705  
P-Value = 0.000
```

What do the correlation and p-value mean? We'll interpret the output soon. First, let's look at a range of possible correlation values so we can understand how our height and weight example fits in.

Interpret the Pearson's Correlation Coefficient

Pearson's correlation coefficient is represented by the Greek letter rho (ρ) for the population parameter and r for a sample statistic. This coefficient is a single number that measures both the strength and direction of the linear relationship between two continuous variables. Values can range from -1 to +1.

- **Strength:** The greater the absolute value of the coefficient, the stronger the relationship.
 - The extreme values of -1 and 1 indicate a perfectly linear relationship where a change in one variable is accompanied by a perfectly consistent change in the other. For these relationships, all of the data points fall on a line. In practice, you won't see either type of perfect relationship.
 - A coefficient of zero represents no linear relationship. As one variable increases, there is no tendency in the other variable to either increase or decrease.
 - When the value is in-between 0 and +1/-1, there is a relationship, but the points don't all fall on a line. As r

approaches -1 or 1 , the strength of the relationship increases and the data points tend to fall closer to a line.

- **Direction:** The coefficient sign represents the direction of the relationship.
 - Positive coefficients indicate that when the value of one variable increases, the value of the other variable also tends to increase. Positive relationships produce an upward slope on a scatterplot.
 - Negative coefficients represent cases when the value of one variable increases, the value of the other variable tends to decrease. Negative relationships produce a downward slope.

Examples of Positive and Negative Correlations

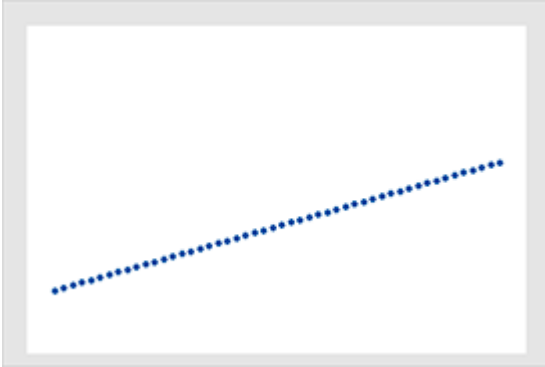
An example of a positive correlation is the relationship between the speed of a wind turbine and the amount of energy it produces. As the turbine speed increases, electricity production also increases.

An example of a negative correlation is the relationship between outdoor temperature and heating costs. As the temperature increases, heating costs decrease.

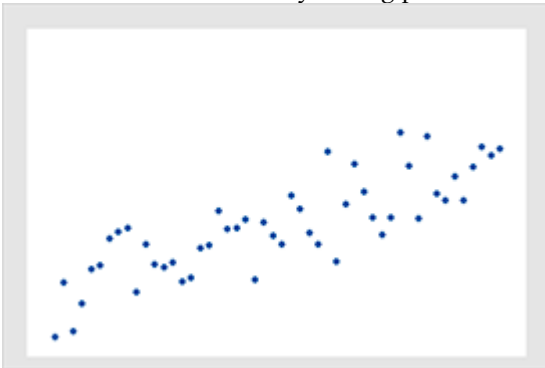
Graphs for Different Correlations

Graphs always help bring concepts to life. The scatterplots below represent a spectrum of different relationships. I've held the horizontal and vertical scales of the scatterplots constant to allow for valid comparisons between them.

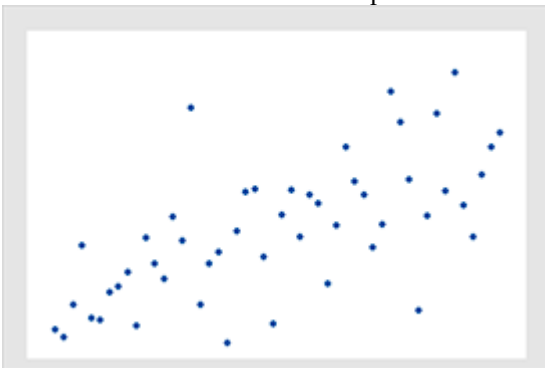
Correlation = +1: A perfect positive relationship.



Correlation = 0.8: A fairly strong positive relationship.



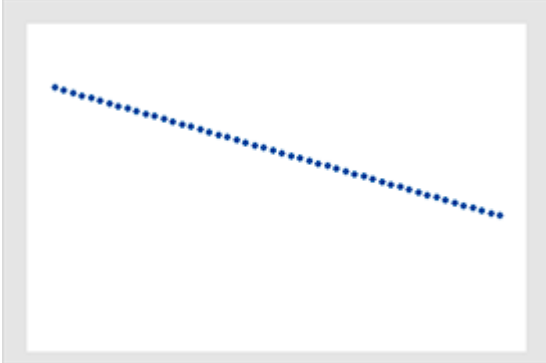
Correlation = 0.6: A moderate positive relationship.



Correlation = 0: No relationship. As one value increases, there is no tendency for the other value to change in a specific direction.



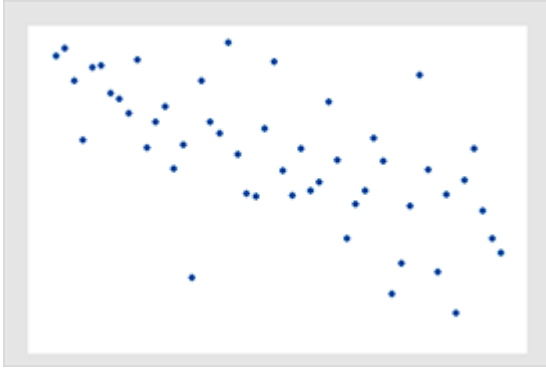
Correlation = -1: A perfect negative relationship.



Correlation = -0.8: A fairly strong negative relationship.



Correlation = -0.6: A moderate negative relationship.



Discussion about the Correlation Scatterplots

For the scatterplots above, I created one positive relationship between the variables and one negative relationship between the variables. Then, I varied only the amount of dispersion between the data points and the line that defines the relationship. That process illustrates how correlation measures the strength of the relationship. The stronger the relationship, the closer the data points fall to the line. I didn't include plots for weaker correlations that are closer to zero than 0.6 and -0.6 because they start to look like blobs of dots and it's hard to see the relationship.

A common misinterpretation is that a negative correlation coefficient indicates there is no relationship between a pair of variables. After all, a negative correlation sounds suspiciously like no relationship. However, the scatterplots for the negative correlations display real relationships. For negative relationships, high values of one variable are associated with low values of another variable. For example, there is a negative correlation between school absences and grades. As the number of absences increases, the grades decrease.

Earlier I mentioned how crucial it is to graph your data to understand them better. However, a quantitative assessment of the relationship does have an advantage. Graphs are a great way to visualize the data, but the scaling can exaggerate or weaken the appearance of a

relationship. Additionally, the automatic scaling in most statistical software tends to make all data look similar.

Fortunately, Pearson's correlation coefficient is unaffected by scaling issues. Consequently, a statistical assessment is better for determining the precise strength of the relationship.

Graphs and the relevant statistical measures often work better in tandem.

Pearson's Correlation Coefficient Measures Linear Relationships

Pearson's correlation measures only *linear* relationships. Consequently, if your data contain a curvilinear relationship, the correlation coefficient will not detect it. For example, the correlation for the data in the scatterplot below is zero. However, there is a relationship between the two variables—it's just not linear.



This example illustrates another reason to graph your data! Just because the coefficient is near zero, it doesn't necessarily indicate that there is no relationship.

Hypothesis Test for Correlations

Correlations have a hypothesis test. As with any hypothesis test, this test takes sample data and evaluates two mutually exclusive statements about the population from which the sample was drawn. For Pearson correlations, the two hypotheses are the following:

- **Null hypothesis:** There is no linear relationship between the two variables. $\rho = 0$.
- **Alternative hypothesis:** There is a linear relationship between the two variables. $\rho \neq 0$.

A correlation of zero indicates that no linear relationship exists. If your p-value is less than your significance level, the sample contains sufficient evidence to reject the null hypothesis and conclude that the correlation does not equal zero. In other words, the sample data support the notion that the relationship exists in the population.

Interpreting our Height and Weight Correlation Example

Now that we have seen a range of positive and negative relationships, let's see how our correlation of 0.705 fits in. We know that it's a positive relationship. As height increases, weight tends to increase. Regarding the strength of the relationship, the graph shows that it's not a very strong relationship where the data points tightly hug a line. However, it's not an entirely amorphous blob with a very low correlation. It's somewhere in between. That description matches our moderate correlation of 0.705.

For the hypothesis test, our p-value equals 0.000. This p-value is less than any reasonable significance level. Consequently, we can reject the null hypothesis and conclude that the relationship is statistically significant. The sample data provide sufficient evidence to conclude that the relationship between height and weight exists in the population of preteen girls.

Correlation Does Not Imply Causation

I'm sure you've heard this expression before, and it is a crucial warning. Correlation between two variables indicates that changes in one variable are associated with changes in the other variable. However, correlation does not mean that the changes in one variable actually *cause* the changes in the other variable.

Sometimes it is clear that there is a causal relationship. For the height and weight data, it makes sense that adding more vertical structure to a body *causes* the total mass to increase. Or, increasing the wattage of lightbulbs *causes* the light output to increase.

However, in other cases, a causal relationship is not possible. For example, ice cream sales and shark attacks are positively correlated. Clearly, selling more ice cream does not cause shark attacks (or vice versa). Instead, a third variable, outdoor temperatures, causes changes in the other two variables. Higher temperatures increase both sales of ice cream and the number of swimmers in the ocean, which creates the apparent relationship between ice cream sales and shark attacks.

In statistics, you typically need to perform a randomized, controlled experiment to determine that a relationship is causal rather than merely correlation.

How Strong of a Correlation is Considered Good?

What is a good correlation? How high should it be? These are commonly asked questions. I have seen several schemes that attempt to classify correlations as strong, medium, and weak.

However, there is only one correct answer. The correlation coefficient should accurately reflect the strength of the relationship. Take a look at the correlation between the height and weight data, 0.705. It's not a very strong relationship, but it accurately represents our data.

An accurate representation is the best-case scenario for using a statistic to describe an entire dataset.

The strength of any relationship naturally depends on the specific pair of variables. Some research questions involve weaker relationships than other subject areas. Case in point, humans are hard to predict. Studies that assess relationships involving human behavior tend to have correlations weaker than ± 0.6 .

However, if you analyze two variables in a physical process, and have very precise measurements, you might expect correlations near +1 or -1. There is no one-size fits all best answer for how strong a relationship should be. The correct correlation value depends on your study area. We run into this same issue in regression analysis.

Common Themes with Regression

Understanding correlation is a good place to start learning regression. In fact, there are several themes that I touch upon in this section that show up throughout this book.

For instance, analysts naturally want to fit models that explain more and more of the variability in the data. And, they come up with classification schemes for how well the model fits the data. However, there is a natural amount of variability that the model can't explain just as there was in the height and weight correlation example. Regression models can be forced to go past this natural boundary, but bad things happen. Throughout this book, be aware of the tension between trying to explain as much variability as possible and ensuring that you don't go too far. This issue pops up multiple times!

Additionally, for regression analysis, you'll need to use statistical measures in conjunction with graphs just like we did with correlation. This combination provides you the best understanding of your data and the analytical results.

Taking Correlation to the Next Level with Regression

Wouldn't it be nice if instead of just describing the strength of the relationship between height and weight, we could define the relationship itself using an equation? Regression analysis does just that by finding the line and corresponding equation that provides the best fit to our dataset. We can use that equation to understand how much weight increases with each additional unit of height and to make predictions for specific heights.

Regression analysis allows us to expand on correlation in other ways. If we have more variables that explain changes in weight, we can include them in the model and potentially improve our predictions. And, if the relationship is curved, we can still fit a regression model to the data.

Additionally, a form of the Pearson correlation coefficient shows up in regression analysis. R-squared is a primary measure of how well a regression model fits the data. This statistic represents the percentage of variation in one variable that other variables explain. For a pair of variables, R-squared is simply the square of the Pearson's correlation coefficient. For example, squaring the height-weight correlation coefficient of 0.705 produces an R-squared of 0.497, or 49.7%. In other words, height explains about half the variability of weight in preteen girls.

But we're getting ahead of ourselves. I'll cover R-squared in much more detail in both chapters 2 and 4.

Fundamental Terms and Goals of Regression

The first questions you have are probably: When should I use regression analysis? And, why? Let's dig right into these questions! In this section, I explain the capabilities of regression analysis, the types of relationships it can assess, how it controls the variables, and generally

why I love it! You'll learn when you should consider using regression analysis.

As a statistician, I should probably tell you that I love all statistical analyses equally—like parents with their kids. But, shhh, I have secret! Regression analysis is my favorite because it provides tremendous flexibility and it is useful in so many different circumstances.

You might run across unfamiliar terms. Don't worry. I'll cover all of them throughout this book! The upcoming section provides a preview for things you'll learn later in the book. For now, let's define several basics—the fundamental types of variables that you'll include in your regression analysis and your primary goals for using regression analysis.

Dependent Variables

The dependent variable is a variable that you want to explain or predict using the model. The values of this variable *depend* on other variables. It's also known as the response variable, outcome variable, and it is commonly denoted using a Y. Traditionally, analysts graph dependent variables and the vertical, or Y, axis.

Independent Variables

Independent variables are the variables that you include in the model to explain or predict changes in the dependent variable. In controlled experiments, independent variables are systematically set and changed by the researchers. However, in observational studies, values of the independent variables are not set by researchers but rather observed. These variables are also known as predictor variables, input variables, and are commonly denoted using Xs. On graphs, analysts place independent variables on the horizontal, or X, axis.

Simple versus Multiple Regression

When you include one independent variable in the model, you are performing simple regression. For more than one independent

variable, it is multiple regression. Despite the different names, it's really the same analysis with the same interpretations and assumptions.

Goals of Regression Analysis

Regression analysis mathematically describes the relationships between independent variables and a dependent variable. Use regression for two primary goals:

- To understand the relationships between these variables. How do changes in the independent variables relate to changes in the dependent variable?
- To predict the dependent variable by entering values for the independent variables into the regression equation.

Example of a Regression Analysis

Suppose a researcher studies the relationship between wattage and the output from a light bulb. In this study, light output is the dependent variable because it depends on the wattage. Wattage is the independent variable.

After performing the regression analysis, the researcher will understand the nature of the relationship between these two variables. Is this relationship statistically significant? What effect does wattage have on light output? For a given wattage, how much light output does the model predict?

Specifically, the regression equation describes the mean change in light output for every increase of one watt. P-values indicate whether the relationship is statistically significant. And, the researcher can enter wattage values into the equation to predict light output.

Regression Analyzes a Wide Variety of Relationships

Use regression analysis to describe the relationships between a set of independent variables and the dependent variable. Regression

analysis produces a regression equation where the coefficients represent the relationship between each independent variable and the dependent variable. You can also use the equation to make predictions.

Regression analysis can handle many things. For example, you can use regression analysis to do the following:

- Model multiple independent variables
- Include continuous and categorical variables
- Model linear and curvilinear relationships
- Assess interaction terms to determine whether the effect of one independent variable depends on the value of another variable

These capabilities are all cool, but they don't include an almost magical ability. Regression analysis can unscramble very intricate problems where the variables are entangled like spaghetti. For example, imagine you're a researcher studying any of the following:

- Do socio-economic status and race affect educational achievement?
- Do education and IQ affect earnings?
- Do exercise habits and diet effect weight?
- Are drinking coffee and smoking cigarettes related to mortality risk?
- Does a particular exercise intervention have an impact on bone density that is a distinct effect from other physical activities?

More on the last two examples later!

All these research questions have entwined independent variables that can influence the dependent variables. How do you untangle a web of related variables? Which variables are statistically significant and what role does each one play? Regression comes to the rescue because you can use it for all of these scenarios!

Using Regression to Control Independent Variables

As I mentioned, regression analysis describes how the changes in each independent variable are related to changes in the dependent variable. Crucially, regression also statistically controls every variable in your model.

What does controlling for a variable mean?

Typically, research studies need to isolate the role of each variable they are assessing. For example, I participated in an exercise intervention study where our goal was to determine whether the exercise intervention increased the subjects' bone mineral density. We needed to isolate the role of the exercise intervention from everything else that can impact bone mineral density, which ranges from diet to other physical activity.

Regression analysis does this by estimating the effect that changing one independent variable has on the dependent variable while holding all the other independent variables constant. This process allows you to understand the role of each independent variable without worrying about the other variables in the model. Again, you want to isolate the effect of each variable.

How do you control the other variables in regression?

A beautiful aspect of regression analysis is that you hold the other independent variables constant by merely including them in your model! Let's look at this in action with an example.

A recent study analyzed the effect of coffee consumption on mortality. The first results indicated that higher coffee intake is related to a higher risk of death. However, coffee drinkers frequently smoke, and the researchers did not include smoking in their initial model. After they included smoking in the model, the regression results indicated that coffee intake lowers the risk of mortality while smoking increases it. This model isolates the role of each variable while holding the other

variable constant. You can assess the effect of coffee intake while controlling for smoking. Conveniently, you're also controlling for coffee intake when looking at the effect of smoking.

Note that the study also illustrates how excluding a relevant variable can produce misleading results. Omitting an important variable causes it to be uncontrolled, and it can bias the results for the variables that you do include in the model. In the example above, the first model without smoking could not control for this important variable, which forced the model to include the effect of smoking in another variable (coffee consumption).

This warning is particularly applicable for observational studies where the effects of omitted variables might be unbalanced. On the other hand, the randomization process in a true experiment tends to distribute the effects of these variables equally, which lessens omitted variable bias. You'll learn about this form of bias in detail in chapter 7.

An Introduction to Regression Output

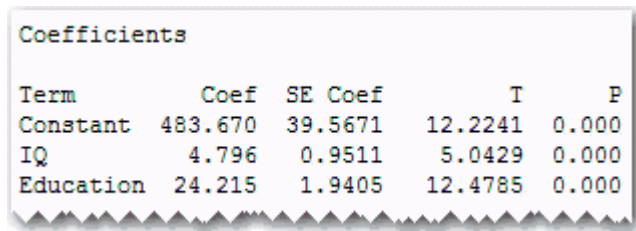
It's time to get our feet wet and interpret regression output. The best way to understand the value of regression analysis is to see an example. In Chapter 3, I cover all of these statistics in much greater detail. For now, you just need to understand the type of information that regression analysis provides.

P-values and coefficients are the key regression output. Collectively, these statistics indicate whether the variables are statistically significant and describe the relationships between the independent variables and the dependent variable.

Low p-values (typically < 0.05) indicate that the independent variable is statistically significant. Regression analysis is a form of inferential statistics. Consequently, the p-values help determine whether the relationships that you observe in your sample also exist in the larger population.

The coefficients for the independent variables represent the average change in the dependent variable given a one-unit change in the independent variable (IV) while controlling the other IVs.

For instance, if your dependent variable is income and your independent variables include IQ and education (among other relevant variables), you might see output like this:



The image shows a screenshot of a regression analysis output window titled "Coefficients". It contains a table with five columns: Term, Coef, SE Coef, T, and P. The rows represent the Constant, IQ, and Education variables. The Constant has a coefficient of 483.670, SE of 39.5671, T of 12.2241, and P of 0.000. IQ has a coefficient of 4.796, SE of 0.9511, T of 5.0429, and P of 0.000. Education has a coefficient of 24.215, SE of 1.9405, T of 12.4785, and P of 0.000.

Term	Coef	SE Coef	T	P
Constant	483.670	39.5671	12.2241	0.000
IQ	4.796	0.9511	5.0429	0.000
Education	24.215	1.9405	12.4785	0.000

The low p-values indicate that both education and IQ are statistically significant. The coefficient for IQ (4.796) indicates that each additional IQ point increases your income by an average of approximately \$4.80 while controlling everything else in the model. Furthermore, the education coefficient (24.215) indicates that an additional year of education increases average earnings by \$24.22 while holding the other variables constant.

Using regression analysis gives you the ability to separate the effects of complicated research questions. You can disentangle the spaghetti noodles by modeling and controlling all relevant variables, and then assess the role that each one plays.

We'll cover how to interpret regression analysis in much more detail in later chapters!

Review and Next Steps

In this chapter, we covered correlation between variables because it's such a good lead-in for regression. Correlation provides you with a

look at some of the fundamental issues we'll address in regression analysis itself—different types of trends in the data and the variability around those trends.

Then, you learned about regression's fundamental goals, its capabilities, and why you'd use it for your study. You can use regression models to describe the relationship between each independent variable and the dependent variable. You can also enter values into the regression equation to predict the mean of the dependent variable. We even took a quick peek at some example regression output and interpreted it.

Finally, we saw how regression analysis controls, or holds constant, all the variables you include in the model. This feature allows you to isolate the role of each independent variable.

This chapter serves as an introduction to all the above. We'll revisit all these concepts throughout this book. Next, you'll learn how least squares regression fits the best line through a dataset.



Regression Basics and How it Works

There are many different types of regression analysis procedures. This book focuses on linear regression analysis, specifically ordinary least squares (OLS). Analysts use this type most frequently. Typically, they'll look towards least squares regression first, and then use other types only when there are issues that prevent them from using OLS.

Even when you need to use a different variety of regression, understanding linear regression is crucial. Much of the knowledge about fitting models, interpreting the results, and checking assumptions for linear models that you will learn throughout this book also apply in some fashion to other types of regression analysis. In short, this book provides a broad foundation on the core type of regression, and it's also informative about using more specialized types of regression.

In later chapters, we'll cover possible reasons for using other kinds of regression analysis. I'll ensure that you know when you should consider a specialized type of analysis, and give you pointers about which alternatives to consider for various issues.

We'll start by covering some basic data requirements. Don't confuse these with the analysis assumptions. I discuss those in chapter 9. These data requirements help ensure that you are putting good data into the analysis. You know that old expression, "garbage in, garbage out?" Let's avoid that!

Data Considerations for OLS

To help ensure that your results are valid for OLS linear regression, consider the following principles while collecting data, performing the analysis, and interpreting the results.

The independent variables can be either continuous or categorical.

- Continuous variables can take on almost any numeric value and can be meaningfully divided into smaller increments, including fractional and decimal values. You often measure a continuous variable on a scale. For example, when you measure height, weight, and temperature, you have continuous data.
- Categorical variables have values that you can put into a countable number of distinct groups based on a characteristic. Categorical variables are also called qualitative variables or attribute variables. For example, college major is a categorical variable that can have values such as psychology, political science, engineering, biology, etc.

The dependent variable should be continuous. If it's not continuous, you will most likely need to use a different type of regression analysis (chapter 12) because your model is unlikely to satisfy the OLS assumptions and can produce results that you can't trust.

Use best practices while collecting your data. The following are some points to consider:

- Confirm that the data represent your population of interest.
- Collect a sufficient amount of data that allows you to fit a model which is appropriately complex for the subject area (chapter 8) and provides the necessary precision for the coefficients and predictions (chapters 3 and 10).
- Measure all variables with the highest accuracy and precision possible.
- Record data in the order you collect it. This process helps you assess an assumption about correlations between adjacent residuals (chapter 9).

Now, let's see how OLS regression goes beyond correlation and produces an equation for the line that best fits a dataset.

How OLS Fits the Best Line

Regression explains the variation in the dependent variable using variation in the independent variables. In other words, it predicts the dependent variable for a given set of independent variables.

Let's start with some basic terms that I'll use throughout this book. While I strive to explain regression analysis in an intuitive manner using everyday English, I do use proper statistical terminology. Doing so will help you if you're following along with a college statistics course or need to communicate with professionals about your model.

Observed and Fitted Values

Observed values of the dependent variable are the values of the dependent variable that you record during your study or experiment along with the values of the independent variables. These values are denoted using Y .

Fitted values are the values that the model predicts for the dependent variable using the independent variables. If you input values for the independent variables into the regression equation, you obtain the fitted value. Predicted values and fitted values are synonyms.

An observed value is one that exists in the real world while your model generates the fitted/predicted value for that observation.

Standard notation uses \hat{y} to denote fitted values, which you pronounce as Y-hat. In general, hatted values indicate they are a model's estimate for the corresponding non-hatted values.

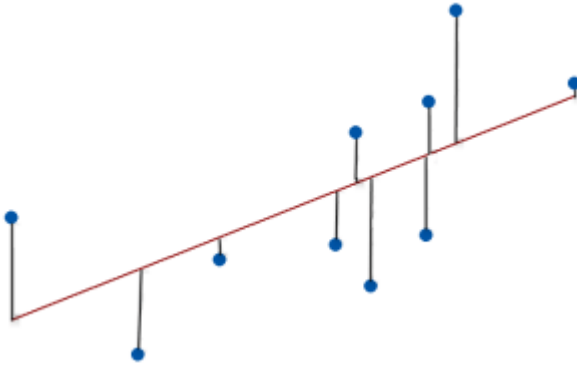
Residuals: Difference between Observed and Fitted Values

Regression analysis predicts the dependent variable. For every observed value of the dependent variable, the regression model calculates a corresponding fitted value. To understand how well your model fits the data, you need to assess the differences between the observed values and the fitted values. These differences represent the error in the model. No model is perfect. The observed and fitted values will never exactly match. However, models can be good enough to be useful.

This difference is known as a residual, and you'll be learning a lot about them in this book. A residual is the distance between an observed value and the corresponding fitted value. To calculate the difference mathematically, it's simple subtraction:

Residual = Observed value – Fitted value.

Graphically, residuals are the vertical distances between the observed values and the fitted values. On the graph, the line represents the fitted values from the regression model. We call this line . . . the fitted line! The lines that connect the data points to the fitted line represent the residuals.



The length of the line is the value of the residual. The equation below shows how to calculate the residuals, or error, for the i^{th} observation:

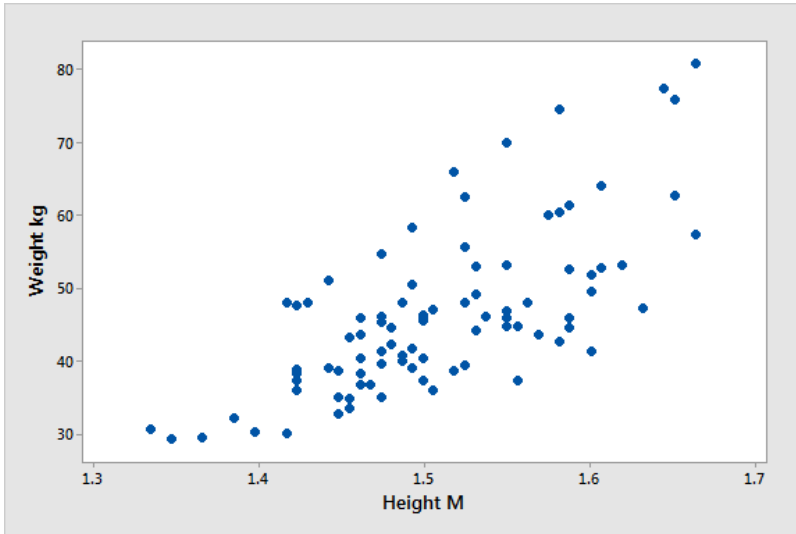
$$e_i = y_i - \hat{y}_i$$

It makes sense, right? You want to minimize the distance between the observed values and the fitted values. For a good model, the residuals should be relatively small and unbiased. In statistics, bias indicates that estimates are systematically too high or too low.

If the residuals become too large or biased, the model is no longer useful. Consequently, these differences play a vital role during both the model estimation process and later when you assess the quality of the model.

Using the Sum of the Squared Errors (SSE) to Find the Best Line

Let's go back to the height and weight dataset for which we calculated the correlation.



The goal of regression analysis is to draw a line through these data points that minimizes the overall distance of the points from the line. How would you draw the best fitting straight line through this cloud of points?

You could draw many different potential lines. Some observations will fit the model better or worse than other points, and that will vary based on the line that you draw. Which measure would you use to quantify how well the line fits all of the data points? Using what you learned above, you know that you want to minimize the residuals. And, it should be a measure that factors in the difference for all of the points. We need a summary statistic for the entire dataset.

Perhaps the average distance or residual value? If your model has many residuals with values near +10 and -10, that averages to approximately zero distance. However, another model with many residuals near +1 and -1 also averages out to be nearly zero. Obviously, you'd prefer the model with smaller distances. Unfortunately, using the average residual doesn't distinguish between these models.

You can't merely sum the residuals because the positive and negative values will cancel each other out even when they tend to be relatively large. Instead, OLS regression squares those residuals so they're always positive. In this manner, the process can add them up without canceling each other out.

This process produces squared errors (residuals). First, we obtain the residuals between the observed and fitted values using simple subtraction, and then we just square them. Simple! A data point with a residual of 3 will have a squared error of 9. A residual of -4 produces a squared error of 16.

Then, the ordinary least squares procedure sums these squared errors, as shown in the equation below:

$$\sum (y - \hat{y})^2$$

OLS draws the line that minimizes the sum of squared errors (SSE). Hopefully, you're gaining an appreciation for why the procedure is named ordinary *least squares*!

SSE is a measure of variability. As the points spread out further from the fitted line, SSE increases. Because the calculations use squared differences, the variance is in squared units rather than the original units of the data. While higher values indicate greater variability, there is no intuitive interpretation of specific values. However, for a given data set, smaller SSE values signal that the observations fall closer to the fitted values. OLS minimizes this value, which means you're getting the best possible line.

In textbooks, you'll find equations for how OLS derives the line that minimizes SSE. Statistical software packages use these equations to solve for the solution directly. However, I'm not going to cover those equations. Instead, it's crucial for you to understand the concepts of

residuals and how the procedure minimizes the SSE. If you were to draw any line other than the one that OLS produces, the SSE would increase—which indicates that the distances between the observed and fitted values are growing, and the model is not as good.

Implications of Minimizing SSE

OLS minimizes the SSE. This fact has several important implications.

First, because OLS calculates squared errors using residuals, the model fitting process ultimately ties back to the residuals very strongly. Residuals are the underlying foundation for how least squares regression fits the model. Consequently, understanding the properties of the residuals for your model is vital. They play an enormous role in determining whether your model is good or not. You'll hear so much about them throughout this book. In fact, chapter 9 focuses on them. So, I won't say much more here. For now, just know that you want relatively small and unbiased residuals (positive and negative are equally likely) that don't display patterns when you graph them.

Second, the fact that the OLS procedure squares the residuals has significant ramifications. It makes the model susceptible to outliers and unusual observations. To understand why, consider the following set of residuals: {1 2 3}. Imagine most of your residuals are in this range. These residuals produce the following squared errors: {1 4 9}. Now, imagine that one observation has a residual of 6, which yields a squared error of 36. Compare the magnitude of most squared errors (1 – 9) to that of the unusual observation (36).

To minimize the squared errors, OLS factors in that unusual observation much more heavily than the other data points. The result is that an individual outlier can exert a strong influence over the entire model and, by itself, dramatically change the results. Chapter 9 discusses this problem in greater detail and how to detect and resolve it. For now, be aware that OLS is susceptible to outliers!

Other Types of Sums of Squares

You learned about the error sum of squares above, but there are several different types of sums of squares in OLS. We won't focus on the others as much as the SSE, but you should understand what they measure and how they're related:

Sums of Squares	Measures	Calculation
Sum of Squared Errors (SSE)	Overall variability of the distance between the data points and fitted values.	Sum of squared residuals. $\sum (y - \hat{y})^2$
Regression Sum of Squares (RSS)	The amount of additional variability your model explains compared to a model that contains no variables and uses only the mean to predict the dependent variable.	Sum of the squared distances between the fitted values and the mean of the dependent variable (\bar{y}). $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$
Total Sum of Squares (TSS)	Overall variability of the dependent variable around its mean.	Sum of the squared distances between the observed values and the mean of the dependent variable. $\sum_{i=1}^n (y_i - \bar{y})^2$

These three sums of squares have the following mathematical relationship:

$$\text{RSS} + \text{SSE} = \text{TSS}$$

Understanding this relationship is fairly straight forward.

- RSS represents the variability that your model explains. Higher is usually good.
- SSE represents the variability that your model does not explain. Smaller is usually good.
- TSS represents the variability inherent in your dependent variable.

Or, Explained Variability + Unexplained Variability = Total Variability

For the same dataset, as you fit better models, RSS increases and SSE decreases by an exactly corresponding amount. RSS cannot be greater than TSS while SSE cannot be less than zero.

Additionally, if you take RSS / TSS , you'll obtain the percentage of the variability of the dependent variable around its mean that your model explains. This statistic is R-squared!

Based on the mathematical relationship shown above, you know that R-squared can range from 0 – 100%. Zero indicates that the model accounts for none of the variability in the dependent variable around its mean. 100% signifies that the model explains all of that variability.

Keep in mind that these sums of squares all measure variability. You might hear about models and variables accounting for variability, and that harkens back to these measures of variability.

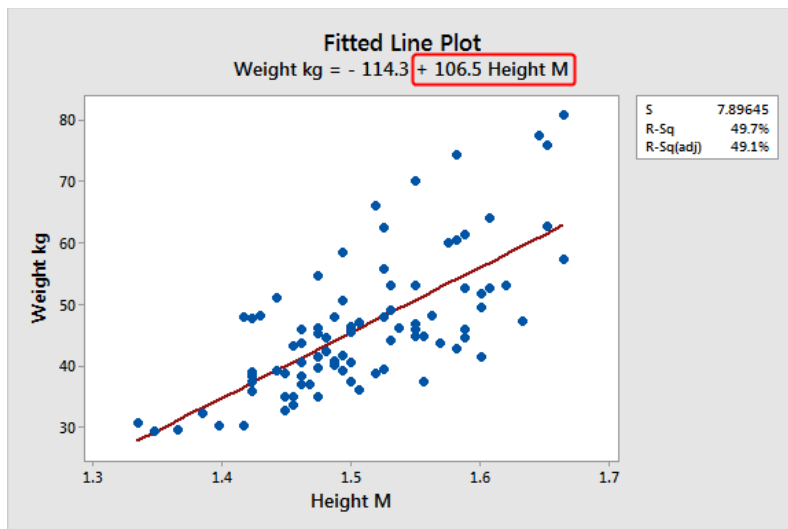
We'll talk about R-squared in much greater detail in chapter 4, which helps you determine how well your model fits the data. However, in

that chapter, I discuss it more from the conceptual standpoint and what it means for your model. I also focus on various problems with R-squared and alternative measures that address these problems. For now, my goal is for you to understand the mathematical derivation of this useful statistic.

Note: Some texts use RSS to refer to residual sums of squares (which we're calling SSE) rather than regression sums of squares. Be aware of this potentially confusing use of terminology!

Displaying a Regression Model on a Fitted Line Plot

Let's again return to our height and weight data. I'll fit the ordinary least squares model and display it in a fitted line plot. You can use this model to estimate the effect of height on weight. You can also enter height values to predict the corresponding weight. Here is the CSV dataset: [HeightWeight](#).



This graph shows all the observations together with a line that represents the fitted relationship. As is traditional, the Y-axis displays the dependent variable, which is weight. The X-axis shows the independent variable, which is height. The line is the fitted line. If you enter the

full range of height values that are on the X-axis into the regression equation that the chart displays, you will obtain the line shown on the graph. This line produces a smaller SSE than any other line you can draw through these observations.

Visually, we see that the fitted line has a positive slope that corresponds to the positive correlation we obtained earlier. The line follows the data points, which indicates that the model fits the data. The slope of the line equals the coefficient that I circled. This coefficient indicates how much mean weight tends to increase as we increase height. We can also enter a height value into the equation and obtain a prediction for the mean weight.

Each point on the fitted line represents the mean weight for a given height. However, like any mean, there is variability around the mean. Notice how there is a spread of data points around the line. You can assess this variability by picking a spot on the line and observing the range of data points above and below that point. Finally, the vertical distance between each data point and the line is the residual for that observation.

Importance of Staying Close to Your Data

It's easy to get lost in the large volume of statistical output that regression produces. All of the numerical statistical measures can cause you to lose touch with your data. However, ensuring that your model adequately represents the data, and determining what the results mean, requires that you stay close to the data. Graphs can help you meet this challenge!

I love using fitted line plots to illustrate regression concepts. In my mission to make regression analysis ideas more intuitive, fitted line plots are one of my primary tools. I'll summarize the concepts that fitted line plots illustrate below, but I'll come back to each one later in the book to explore them in more detail.

Fitted line plots are great for showing the following:

- The regression coefficient in the equation corresponds to the slope of the line. What does it mean?
- For different models, the data points vary around the line to a greater or lesser extent, which reflects the precision of the predictions and goodness-of-fit statistics, like R-squared. We'll explore this in more detail because the implications of this precision are often forgotten. How precise are your model's predictions?
- Does the fitted line fit curvature that is present in the data? For now, we're fitting a straight line, but that might not always be the case! Fitted line plots make curvature unmistakable.

As fantastic as fitted line plots are, they can only show simple regression models, which contain only one independent variable. Fitted line plots use two axes—one for the dependent variable and the other for the independent variable. Consequently, fitted line plots are great for displaying simple regression models on a screen or printed on paper. However, each additional independent variable requires another axis or physical dimension. With two independent variables, we can use a 3D representation for it. Although, that's beyond my abilities for this book. With three independent variables, we'd need a four-dimensional plot. That's not going to happen!

If you have a simple regression model, I highly recommend creating a fitted line plot for it and assessing the bullet points above. You'll obtain an excellent overview of how your model fits the data because they're graphed together. However, for multiple regression, we can't use fitted line plots to obtain that overview. For those cases, I'll show you other methods throughout this book for answering those questions. Sometimes these methods will be statistical measures, but whenever possible I'll show you special types of graphs because they bring it to life. These graphical tools include main effects plots, interaction plots, and various residual plots.

Review and Next Steps

In this chapter, I explained how learning about ordinary least squares linear regression provides an excellent foundation for learning about regression analysis. Not only is it the most frequently used type of regression, but your knowledge of OLS will help inform your usage of other types of regression. I showed you some foundational data considerations to keep in mind so you can avoid the problem of “garbage in, garbage out!”

You learned the basics of how OLS minimizes the sums of squared errors (SSE) to produce the best fitting line for your dataset. And, how SSE fits it in with two other sums of squares, regression sums of squares (RSS) and total sums of squares (TSS). In the process, you even got a sneak peek at R-squared (RSS / TSS)!

Then, we explored the height-weight regression model using a fitted line plot.

From here, we’ll move on to learning how to interpret the different types of effects for continuous and categorical independent variables, the constant, what statistical significance indicates in this context, and determining significance.

END OF FREE SAMPLE

NOTE: This sample contains only the introduction and first two chapters. Please buy the full ebook for all the content listed in the Table of Contents. You can buy it in [My Store](#).