Intro Regression

Maria Tackett

Latest update: 2020-08-23

Contents

Welcome to Intro Regression!

The content in this book was originally developed for STA 210: Regression Analysis at Duke University.The computing aspects of the assignments are written using the `tidyverse` syntax in R; however, the assignments can be adapted to fit the computing language of your choice. All of the files are available in the Intro Regression GitHub repo.

This book is under development and will be periodically upated with new material. Please email me (maria.tackett@duke.edu) if you have any questions, feedback, or suggestions. I would also love to hear about your experience if you use any of the content in your course.

Chapter 1

Getting Started

1.1 How to use this book

Each chapter of this book is a topic that may be covered in an intermediate-level regression analysis course. The topics are arranged based on the way they were taught in STA 210: Regression Analysis; however, the assignments do not have to be used in the order they are presented. Feel free to use the text and adapt it to fit the needs of your course.

Each chapter includes several sections of assignments and supplmental notes about the mathematical details. Each section begins with one of the codes below to help you determine the type of assignment or note in that section:

- **COMP**: These assignments focus on the computing skills needed to conduct regression analysis. They were originally designed to be completed in groups in a weekly lab/discussion session; however, they can be also be used for homework assignments or in-class work days. Because the emphasis is computing, they include a lot of step-by-step instruction.
- **IN-CLASS**: Assignments to be completed as short in-class activities. Most of the code is already written, so students mostly run the code and interpret the output. Students may also need to fill in short lines of code.
- **HW**: Focus on putting together conceptual knowledge and computing skills. Most homework assingments include two parts: (1) *Concepts & Computations* - guided short-answer exercises that focus on conceptual knowledge and short computational tasks, (2) *Data Analysis* - open-ended question where students perform a complete regression analysis and write results as a narrative.
- **NOTES**: Supplemental notes providing more mathematical details. To

fully understand the notes, the reader should be familiar with basic concepts in linear algebra.

1.2 Review: Intro Statistics and R

The primary audience for this text is students who have completed an introductory statistics course. It is assumed that students are familiar with the concept of statitical inference. This text is also written assuming students have had some exposure to R and the tidyverse syntax. (There is one "Intro to R" assingment included; however, this assignment is not a comprehensive introduction to R.) The following are suggested texts to review statistical concepts and computing:

- *OpenIntro Statistics*
- *Modern Dive*
- *R for Data Science*

Chapter 2

Simple Linear Regression

2.1 Getting started

Putting text here

2.2 Foundation

Putting text here

2.3 Inference

Putting text here

2.4 Prediction

2.5 Checking conditions

2.6 Partioning variability

2.7 Derivation for slope and intercept

This document contains the mathematical details for deriving the least-squares estimates for slope (β_1) and intercept (β_0). We obtain the estimates, $\hat{\beta}_1$ and $\hat{\beta}_0$ by finding the values that minimize the sum of squared residuals (2.1).

$$SSR = \sum_{i=1}^{n}[y_i - \hat{y}_i]^2 = [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 = [y_i - (\hat{\beta}_0 - \hat{\beta}_1 x_i]^2 \tag{2.1}$$

Recall that we can find the values of $\hat{\beta}_1$ and $\hat{\beta}_0$ that minimize (2.1) by taking the partial derivatives of (2.1) and setting them to 0. Thus, the values of $\hat{\beta}_1$ and $\hat{\beta}_0$ that minimize the respective partial derivative also minimize the sum of squared residuals. The partial derivatives are

$$\begin{aligned} \frac{\partial \text{SSR}}{\partial \hat{\beta}_1} &= -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ \frac{\partial \text{SSR}}{\partial \hat{\beta}_0} &= -2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \end{aligned} \tag{2.2}$$

Let's begin by deriving $\hat{\beta}_0$.

$$
\begin{aligned}
\frac{\partial \text{SSR}}{\partial \hat{\beta}_0} &= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\
&\Rightarrow -\sum_{i=1}^{n}(y_i + \hat{\beta}_0 + \hat{\beta}_1 x_i) = 0 \\
&\Rightarrow -\sum_{i=1}^{n} y_i + n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \\
&\Rightarrow n\hat{\beta}_0 = \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i \\
&\Rightarrow \hat{\beta}_0 = \frac{1}{n}\Big(\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i\Big) \\
&\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
\end{aligned}
\tag{2.3}
$$

Now, we can derive $\hat{\beta}_1$ using the $\hat{\beta}_0$ we just derived

$$
\begin{aligned}
&\frac{\partial \text{SSR}}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\
&\Rightarrow -\sum_{i=1}^{n} x_i y_i + \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \\
(\text{Fill in } \hat{\beta}_0) &\Rightarrow -\sum_{i=1}^{n} x_i y_i + (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \\
&\Rightarrow (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \\
&\Rightarrow \bar{y} \sum_{i=1}^{n} x_i - \hat{\beta}_1 \bar{x} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \\
&\Rightarrow n\bar{y}\bar{x} - \hat{\beta}_1 n\bar{x}^2 + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \\
&\Rightarrow \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 - \hat{\beta}_1 n\bar{x}^2 = \sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x} \\
&\Rightarrow \hat{\beta}_1 \Big(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \Big) = \sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x} \\
&\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}}{\sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2}
\end{aligned}
\tag{2.4}
$$

To write $\hat{\beta}_1$ in a form that's more recognizable, we will use the following:

$$
\sum x_i y_i - n\bar{y}\bar{x} = \sum (x - \bar{x})(y - \bar{y}) = (n-1)\text{Cov}(x, y) \tag{2.5}
$$

$$
\sum x_i^2 - n\bar{x}^2 - \sum (x - \bar{x})^2 = (n-1)s_x^2 \tag{2.6}
$$

where $\text{Cov}(x, y)$ is the covariance of x and y, and s_x^2 is the sample variance of x (s_x is the sample standard deviation).

Thus, applying (2.5) and (2.6), we have

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\
&= \frac{\sum_{i=1}^{n} (x-\bar{x})(y-\bar{y})}{\sum_{i=1}^{n} (x-\bar{x})^2} \\
&= \frac{(n-1)\text{Cov}(x,y)}{(n-1)s_x^2} \\
&= \frac{\text{Cov}(x,y)}{s_x^2}
\end{aligned} \tag{2.7}$$

The correlation between x and y is $r = \frac{\text{Cov}(x,y)}{s_x s_y}$. Thus, $\text{Cov}(x,y) = r s_x s_y$. Plugging this into (2.7), we have

$$\hat{\beta}_1 = \frac{\text{Cov}(x,y)}{s_x^2} = r\frac{s_y s_x}{s_x^2} = r\frac{s_y}{s_x} \tag{2.8}$$

Chapter 3

Analysis of Variance

Chapter 4

Multiple Linear Regression

Chapter 5

Model Selection

Chapter 6

Logistic Regression

Chapter 7

Multinomial Logistic Regression

Chapter 8

Special Topics

Chapter 9

Data Sets

Below is a list of the datasets used in this book. More details about each dataset are coming soon.

- advertising.csv
- airbnb_basic.csv
- airbnb_details.csv
- beer.csv
- bikeshare.csv
- evals_mod.csv
- fivethirtyeight-recent-grads.R
- framingham.csv
- gss2016.csv
- KingCountyHouses.csv
- movies.csv
- recent-grads.csv
- sesame.csv
- sis.csv
- spotify.csv
- test_songs.csv