Linear Regression

Hans Dohms

26/11/2018

Table of Contents

Linear regression is commonly used to quantify the relationship between two or more variables. It is also used to adjust for confounding. In this course, we cover how to implement linear regression and adjust for confounding in practice using R.

The class notes for this course series can be found in Professor Irizarry’s freely available Introduction to Data Science book.

## Course overview

There are three major sections in this course: introduction to linear regression, linear models, and confounding.

### 1. Introduction to Linear Regression

In this section, you’ll learn the basics of linear regression through this course’s motivating example, the data-driven approach used to construct baseball teams. You’ll also learn about correlation, the correlation coefficient, stratification, and the variance explained.

### 2. Linear Models

In this section, you’ll learn about linear models. You’ll learn about least squares estimates, multivariate regression, and several useful features of R, such as tibbles, lm, do, and broom. You’ll learn how to apply regression to baseball to build a better offensive metric.

### 3. Confounding

In the final section of the course, you’ll learn about confounding and several reasons that correlation is not the same as causation, such as spurious correlation, outliers, reversing cause and effect, and confounders. You’ll also learn about Simpson’s Paradox.

## 1. Introduction to Linear Regression

In the **Introduction to Regression section**, you will learn the basics of linear regression.

After completing this section, you will be able to:

* Understand how Galton developed **linear regression**.
* Calculate and interpret the **sample correlation**.
* **Stratify** a dataset when appropriate.
* Understand what a **bivariate normal distribution** is.
* Explain what the term **variance explained** means.
* Interpret the two **regression lines**.

This section has three parts: **Baseball as a Motivating Example, Correlation**, and **Stratification and Variance Explained**. There are comprehension checks that follow most videos.

### 1.1 Motivating Example: Moneyball

As motivation for this course, we’ll go back to 2002 and try to build a baseball team with a limited budget. Note that in 2002, the Yankees payroll was almost $ 130 million, and had more than tripled the Oakland A’s $ 40 million budget. Statistics have been used in baseball since its beginnings. Note that the data set we will be using, included in the **Lahman Library**, goes back to the 19th century. For example, a summary of statistics we will describe soon, the batting average, has been used to summarize a batter’s success for decades. Other statistics such as home runs, runs batted in, and stolen bases, we’ll describe all this soon, are reported for each player in the game summaries included in the sports section of newspapers. And players are rewarded for high numbers. Although summary statistics were widely used in baseball, data analysis per se was not. These statistics were arbitrarily decided on without much thought as to whether they actually predicted, or were related to helping a team win. This all changed with Bill James. In the late 1970s, this aspiring writer and baseball fan started publishing articles describing more in-depth analysis of baseball data. He named the approach of using data to predict what outcomes best predict if a team wins **sabermetrics**. Until Billy Beane made sabermetrics the center of his baseball operations, Bill James’ work was mostly ignored by the baseball world. Today, pretty much every team uses the approach, and it has gone beyond baseball into other sports. In this course, to simplify the example we use, we’ll focus on predicting scoring runs. We will ignore pitching and fielding, although those are important as well. We will see how regression analysis can help develop strategies to build a competitive baseball team with a constrained budget. The approach can be divided into two separate data analyses. In the first, we determine which recorded player specific statistics predict runs. In the second, we examine if players were undervalued based on what our first analysis predicts.

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

summary(cars)

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

## Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.