Regression Models for Data Science in R

Chapter - 1: Introduction

Charlene P. Garridos

2023-03-07

Contents

1 Introduction

1.1 Regression Models

Regression models are the *workhorse of data science*. They are the most well described, practical and theoretically understood models in statistics. A data scientist well versed in regression models will be able to solve an incredible array of problems.

The *key insight* for regression models is that *they produce highly interpretable model fits*. This is unlike machine learning algorithms, which often sacrifice interpretability for improved prediction performance or automation. These are, of course, valuable attributes in their own rights. However, the benefit of simplicity, parsimony and interpretability offered by regression models (and their close generalizations) should make them a first tool of choice for any practical problem.

1.2 Motivating Examples

1.2.1 Francis Galton's height data

Francis Galton, the 19th century polymath, can be credited with discovering regression. In his landmark paper *Regression Toward Mediocrity in Hereditary Stature* (https://galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf) he compared the heights of parents and their children. He referred to this as "regression to mediocrity" (or regression to the mean). In quantifying regression to the mean, *he invented what we would call regression*

1.2.2 Simply Statistics versus Kobe Bryant

Simply Statistics (https://simplystatistics.org/)is a blog by Jeff Leek, Roger Peng and Rafael Irizarry. It is one of the most widely read statistics blogs, written by three of the top statisticians in academics. Rafa wrote a (somewhat tongue in cheek) *post regarding ball hogging* among NBA basketball players.

key sentences:

- "Data supports the claim that if Kobe stops ball hogging the Lakers will win more"
- "Linear regression suggests that an increase of 1% in % of shots taken by Kobe results in a drop of 1.16 points (+/- 0.22) in score differential."

1.3 Summary Notes: questions for this book

Regression models are incredibly handy statistical tools. One can use them to answer all sorts of questions. Consider three of the most common tasks for regression models:

1. **Prediction** e.g.: to use the parent's heights to predict children's heights.
2. **Modeling** e.g.: to try to find a parsimonious, easily described mean relationship between parental and child heights.
3. **Covariation** e.g.: to investigate the variation in child heights that appears unrelated to parental heights (residual variation) and to quantify what impact genotype information has beyond parental height in explaining child height.

An important aspect, especially in questions 2 and 3 is assessing modeling assumptions. For example,it is important to figure out how/whether and what assumptions are needed to generalize findings beyond the data in question.

1.4 Exploratory analysis of Galtons's Data

Francis Galton (https://mathshistory.st-andrews.ac.uk/Biographies/Galton/) was a statistician who *invented the term and concepts of regression and correlation*, founded the journal *Biometrika* (https://academic.oup.com/biomet), and was the cousin of *Charles Darwin*.

The marginal (parents disregarding children and children disregarding parents) distributions:

The parental distribution is all heterosexual couples. The parental average was corrected for gender via multiplying female heights by 1.08. Remember, Galton didn't have regression to help figure out a better way to do this correction!

```
library(UsingR); data(galton); library(reshape); long <- melt(galton)
g <- ggplot(long, aes(x = value, fill = variable))
g <- g + geom_histogram(colour="black", binwidth=1)
g <- g + facet_grid(. ~ variable)
g
```

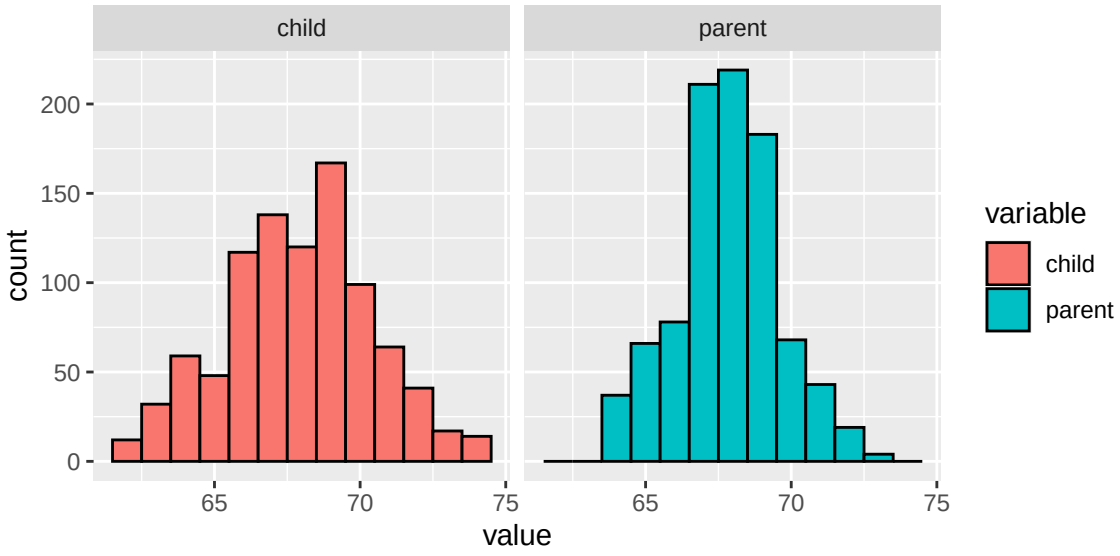

Figure 1: Plotting the galton dataset

1.4.1 Finding the middle via leat squares

Consider only the children's heights. How could one describe the "middle"? Consider one definition. Let Y_i be the height of child i for $i = 1, \ldots, n = 928$, then define the middle as the value of μ that minimizes

$$\sum_{i=1}^{n} (Y_i - \mu)^2$$

This is *physical center of mass* of the histogram.

The answer $\mu = \bar{Y}$. This is called the **least squares estimate** for μ. It is the point that minimizes the sum of the squared distances between the observed data and itself.

> **Note**, if there was no variation in the data, every value of Y_i was the same, then there would be no error around the mean. Otherwise, our estimate has to balance the fact that our estimate of μ isn't going to predict every observation perfectly.

1.5 The math (not required)

$$
\begin{aligned}
\sum_{i=1}^{n}(Y_i-\mu)^2 &= \sum_{i=1}^{n}(Y_i-\bar{Y}+\bar{Y}-\mu)^2 \\
&= \sum_{i=1}^{n}(Y_i-\bar{Y})^2 + 2\sum_{i=1}^{n}(Y_i-\bar{Y})(\bar{Y}-\mu) + \sum_{i=1}^{n}(\bar{Y}-\mu)^2 \\
&= \sum_{i=1}^{n}(Y_i-\bar{Y})^2 + 2(\bar{Y}-\mu)\sum_{i=1}^{n}(Y_i-\bar{Y}) + \sum_{i=1}^{n}(\bar{Y}-\mu)^2 \\
&= \sum_{i=1}^{n}(Y_i-\bar{Y})^2 + 2(\bar{Y}-\mu)(\sum_{i=1}^{n}Y_i - n\bar{Y}) + \sum_{i=1}^{n}(\bar{Y}-\mu)^2 \\
&= \sum_{i=1}^{n}(Y_i-\bar{Y})^2 + \sum_{i=1}^{n}(\bar{Y}-\mu)^2 \\
&\geq \sum_{i=1}^{n}(Y_i-\bar{Y})^2
\end{aligned}
$$

1.6 Comparing children's heights and their parent's heights

```
ggplot(galton, aes(x = parent, y = child)) + geom_point()
```

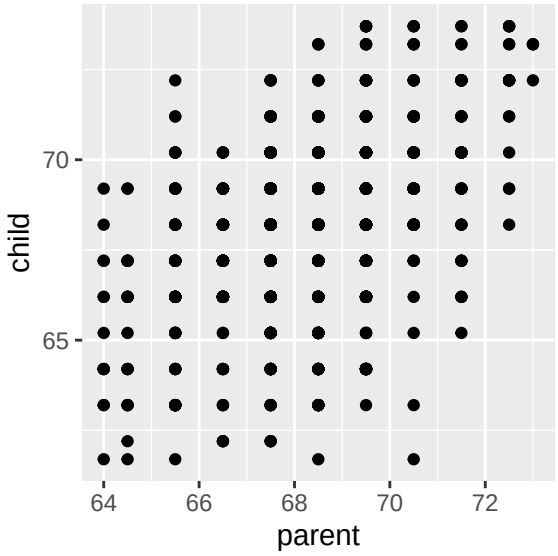

Figure 2: Plot of parents and child heights

1.7 Regression through the origin

A line requires two parameters to be specified, the *intercept and the slope.*

- Focus on the slope.
- Find the slope of the line that best fits the data. However, pick a good intercept.

- Subtract the mean from both the parent and child heights so that their subsequent means are 0.
- Find the line that goes through the origin (has intercept 0) by picking the best slope.

Suppose that X_i are the parent heights with the mean subtracted. Consider picking the slope β that minimizes

$$\sum_{i=1}^{n}(Y_i - X_i\beta)^2$$

Each $X_i\beta$ is the vertical height of a line through the origin at point X_i. Thus, $Y_i - X_i\beta$ is the vertical distance between the line at each observed X_i point (parental height) and the Y_i (child height).

> **Goal:** is exactly to use the origin as a pivot point and pick the line that minimizes the sum of the squared vertical distances of the points to the line.