Chapter-1

Katlyn H. Degamo

2023-03-06

Introduction to regression

Regression models

- Regression Models Course class as part of the Data Science Specialization, and this class is code taught by, Jeff Leek, Roger Peng and Brian Caffo, at the Johns Hopkins University Department of Bio statistics.
- Regression is probably the most fundamental topic for the data scientist.

A famous motivating example

Francis Galton's heigth data

- Francis Galton, the 19th century polymath, can be credited with discovering regression.
- His landmark paper Regression Toward Mediocrity in Hereditary Stature.
- 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.

Simply Statistics versus Kobe Bryant

- Simply Statistics 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 post regarding a ball hogging.

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:
  - Prediction
  - Modeling
  - Covariation
- Click me

Exploratory analysis of Galston's Data

- This data was created by Francis Galton in 1885.
- Galton was a statistician who invented the term and concepts of regression and correlation, and was the cousin of Charles Darwin.
- 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!

Finding the middle via least 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 $i = 1, ..., n = 928$, 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. You might have guessed that 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.
- Click me

The math (not required)

- Why is the sample average the least squares estimate for μ? It's surprisingly easy to show. Perhaps more surprising is how generally these results can be extended.

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

.

- Click me

Comparing children's heigths and their parent's height

- Looking at either the parents or children on their own isn't interesting. We're interested in how the relate to each other.
- Suppose that X_i are the parents heights and Y_i are the children's height.

Regression through the origin

- A line requires two parameters to be specified, the intercept and the slope.
- We want to find the slope of the line that best fits the data.
- Let's subtract the mean from both the parent and child heights so that their subsequent means are 0.
- Consider picking the slope β that minimizes

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

  - $X_i\beta$ is the vertical height of a line through the origin at a point X_i.
  - $Y_i - X_i\beta$ is the vertical distance between the line at each observed X_i and Y_i.

- Click me
- Book