Chapter-1

Laika Jean Paculba

2023-03-09

Chapter 1: Introduction

Regression models

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

Francis Galton

- a 19th century polymath, he is a statistician who invented the term and concepts of regression and correlation.

Francis Galton's height data

- **Regression Toward mediocrity in Hereditary Stature** - his landmark paper, he compared the heights of parents and their children. He was particularly interested in the idea that the children of tall parents tended to be tall also, but a little shorter than their parents. Children of short parents tended to be short, but not quite short as their parents. He referred to this as **"regression to mediocrity"** (or regression to the mean). In quantifying regression to the mean, he invented what we could call regression.

Simply Statistics

- a blog by Jeff Leek, Roger Peng and Rafael Irizarry. One of the most widely read statistics blogs, written by three of the top statisticians in academics.

Galton's Data

- created in 1885
- You may need to run install.packages("UsingR") if the UsingR library is not installed.
- The marginal (parents disregarding children and children disregarding parents) distributions first.
  - parental distribution is all heterosexual couples.
  - parental average was corrected for gender via multiplying female heights by 1.08.
  - over plotting is an issue from discretion.

```
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
```

```
library(manipulate)
myHist <- function(mu){
mse <- mean((galton$child - mu)^2)
g <- ggplot(galton, aes(x = child)) + geom_histogram(fill = "salmon", colour = "black", binwidth=1)
g <- g + geom_vline(xintercept = mu, size = 3)
g <- g + ggtitle(paste("mu = ", mu, ", MSE = ", round(mse, 2), sep = ""))
g
}
manipulate(myHist(mu), mu = slider(62, 74, step = 0.5))
```

```
g <- ggplot(galton, aes(x = child)) + geom_histogram(fill = "salmon", colour = "black", binwidth=1)
g <- g + geom_vline(xintercept = mean(galton$child), size = 3)
g
```

Finding the middle via least squares:

- Consider only the children's height.
  - 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 $\mu = \overline{Y}$ (least squares estimate for μ)

Comparing children's heights and their parent's height

- Looking at either the parents or children on their own isn't interesting. We're interested in how they relate to each other

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

- The over plotting is clearly hiding some data.
- Size of point represents number of points at that (X, Y) combination (See the Rmd file for the code).

```
library(dplyr)
y <- galton$child - mean(galton$child)
x <- galton$parent - mean(galton$parent)
freqData <- as.data.frame(table(x, y))
names(freqData) <- c("child", "parent", "freq")
freqData$child <- as.numeric(as.character(freqData$child))
freqData$parent <- as.numeric(as.character(freqData$parent))
myPlot <- function(beta){
g <- ggplot(filter(freqData, freq > 0), aes(x = parent, y = child))
g <- g + scale_size(range = c(2, 20), guide = "none" )
g <- g + geom_point(colour="grey50", aes(size = freq+20), show.legend = FALSE)
g <- g + geom_point(aes(colour=freq, size = freq))
g <- g + scale_colour_gradient(low = "lightblue", high="white")
g <- g + geom_abline(intercept = 0, slope = beta, size = 3)
mse <- mean( (y - beta * x) ^2 )
g <- g + ggtitle(paste("beta = ", beta, "mse = ", round(mse, 3)))
g
}
manipulate(myPlot(beta), beta = slider(0.6, 1.2, step = 0.02))
```

Regression through the origin

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

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

  - 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)
- Our 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
- Use RStudio's manipulate function to experiment.
- Subtract the means so that the origin is the mean of the parent and children heights.

The solution

```
lm(I(child - mean(child))~ I(parent - mean(parent)) - 1, data = galton)

Call:
lm(formula = I(child - mean(child)) ~ I(parent - mean(parent)) -
    1, data = galton)

Coefficients:
I(parent - mean(parent))
                   0.646
```

The results suggest that for every 1 inch increase in the parents' height, we estimate a 0.646 inch increase in the child's height