Gov 51: Linear Regression

Matthew Blackwell

Harvard University

Predicting my weight

- I've been tracking my physical activity and weight for a few years now.
- Can we use my activity to predict my weight on a day-to-day basis?

Name	Description
`date`	date of measurements
`active.calories`	calories burned
`steps`	number of steps taken (in 1,000s)
`weight`	weight (lbs)
`steps.lag`	steps on day before (in 1,000s)
`calories.lag`	calories burned on day before

Predicting using bivariate relationship

- Goal: what's our best guess about Y_i if we know what X_i is?
  - what's our best guess about my weight this morning if I know how many steps I took yesterday?
- Terminology:
  - **Dependent/outcome variable**: what we want to predict (weight).
  - **Independent/explanatory variable**: what we're using to predict (steps).

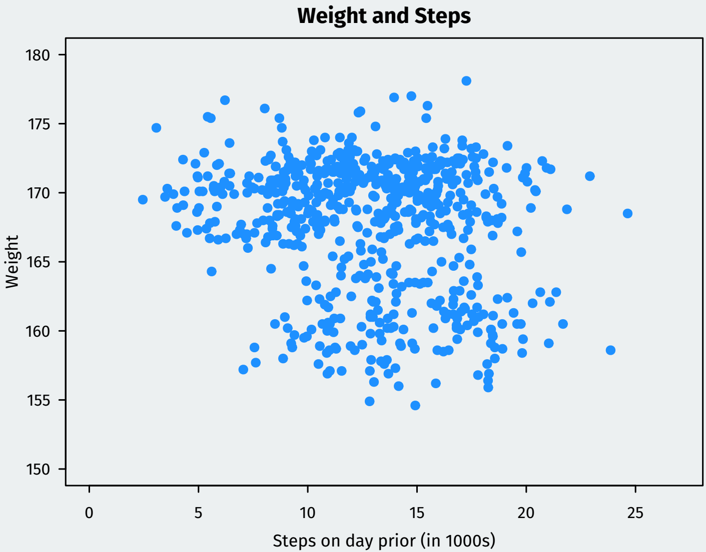
Weight data

- Load the data:

```
health <- read.csv("data/health.csv")
health <- na.omit(health)
```

- Plot the data:

```
plot(health$steps.lag, health$weight, pch = 19,
     col =  "dodgerblue",
     xlim = c(0, 27), ylim = c(150, 180),
     xlab = "Steps on day prior (in 1000s)",
     ylab = "Weight",
     main = "Weight and Steps")
```


Weight and Steps
Weight
180
175
170
165
160
155
150
0
5
10
15
20
25
Steps on day prior (in 1000s)

Using a line to predict

- Prediction: for any value of X, what's the best guess about Y?
  - Need a function $y = f(x)$ that maps values of X into predictions.
  - **Machine learning**: fancy ways to determine $f(x)$
- Simplest possible way to relate two variables: a line.

$$y = mx + b$$

- Problem: for any line we draw, not all the data is on the line.
  - Some points will be above the line, some below.
  - Need a way to account for **chance variation** away from the line.

Linear regression model

- Model for the line of best fit:

$$Y_i = \underbrace{\alpha}_{\text{intercept}} + \underbrace{\beta}_{\text{slope}} \cdot X_i + \underbrace{\epsilon_i}_{\text{error term}}$$

- **Coefficients/parameters** (α, β): true unknown intercept/slope of the line of best fit.
- **Chance error** ϵ_i: accounts for the fact that the line doesn't perfectly fit the data.
  - Each observation allowed to be off the regression line.
  - Chance errors are 0 on average.
- Useful fiction: this model represents the **data generating process**
  - George Box: "all models are wrong, some are useful"

Interpreting the regression line

$$Y_i = \alpha + \beta \cdot X_i + \epsilon_i$$

- **Intercept** α: average value of Y when X is 0
    - Average weight when I take 0 steps the day prior.
- **Slope** β: average change in Y when X increases by one unit.
    - Average decrease in weight for each additional 1,000 steps.
- But we don't know α or β. How can we estimate them? Next time...