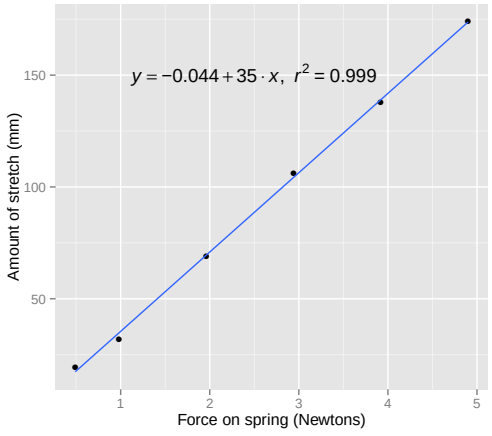
Chapter 3

Linear Regression

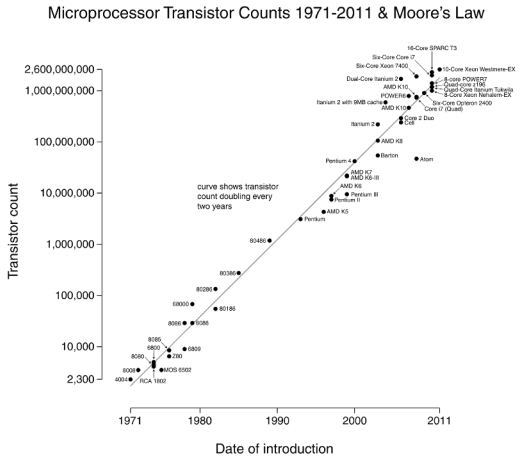
Once we've acquired data with multiple variables, one very important question is how the variables are related. For example, we could ask for the relationship between people's weights and heights, or study time and test scores, or two animal populations. **Regression** is a set of techniques for estimating relationships, and we'll focus on them for the next two chapters.

In this chapter, we'll focus on finding one of the simplest type of relationship: linear. This process is unsurprisingly called **linear regression**, and it has many applications. For example, we can relate the force for stretching a spring and the distance that the spring stretches (Hooke's law, shown in Figure 3.1a), or explain how many transistors the semiconductor industry can pack into a circuit over time (Moore's law, shown in Figure 3.1b).

Despite its simplicity, linear regression is an incredibly powerful tool for analyzing data. While we'll focus on the basics in this chapter, the next chapter will show how just a few small tweaks and extensions can enable more complex analyses.



(a) In classical mechanics, one could empirically verify Hooke's law by dangling a mass with a spring and seeing how much the spring is stretched.



(b) In the semiconductor industry, Moore's law is an observation that the number of transistors on an integrated circuit doubles roughly every two years.

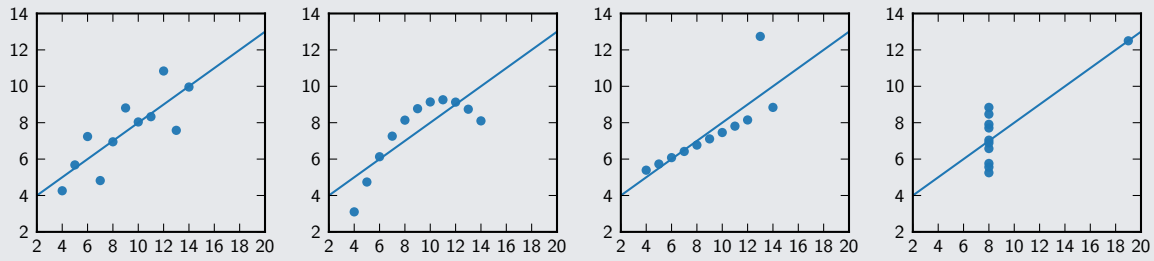
Figure 3.1: Examples of where a line fit explains physical phenomena and engineering feats.[1]

[1]The Moore's law image is by Wgsimon (own work) [CC-BY-SA-3.0 or GFDL], via Wikimedia Commons.

But just because fitting a line is easy doesn't mean that it always makes sense. Let's take another look at Anscombe's quartet to underscore this point.

Example: Anscombe's Quartet Revisited

Recall Anscombe's Quartet: 4 datasets with very similar statistical properties under a simple quantitative analysis, but that look very different. Here they are again, but this time with linear regression lines fitted to each one:



For all 4 of them, the slope of the regression line is 0.500 (to three decimal places) and the intercept is 3.00 (to two decimal places). This just goes to show: visualizing data can often reveal patterns that are hidden by pure numeric analysis!

We begin with **simple linear regression** in which there are only two variables of interest (e.g., weight and height, or force used and distance stretched). After developing intuition for this setting, we'll then turn our attention to **multiple linear regression**, where there are more variables.

Disclaimer: While some of the equations in this chapter might be a little intimidating, it's important to keep in mind that as a user of statistics, the most important thing is to understand their uses and limitations. Toward this end, make sure not to get bogged down in the details of the equations, but instead focus on understanding how they fit in to the big picture.

3.1 Simple linear regression

We're going to fit a line $y = \beta_0 + \beta_1 x$ to our data. Here, x is called the **independent variable** or **predictor variable**, and y is called the **dependent variable** or **response variable**.

Before we talk about how to do the fit, let's take a closer look at the important quantities from the fit:

- β_1 is the slope of the line: this is one of the most important quantities in any linear regression analysis. A value very close to 0 indicates little to no relationship; large positive or negative values indicate large positive or negative relationships, respectively. For our Hooke's law example earlier, the slope is the spring constant[2].

[2]Since the spring constant k is defined as $F = -kx$ (where F is the force and x is the stretch), the slope in Figure 3.1a is actually the inverse of the spring constant.

- β_0 is the intercept of the line.

In order to actually fit a line, we'll start with a way to quantify how good a line is. We'll then use this to fit the "best" line we can.

One way to quantify a line's "goodness" is to propose a probabilistic model that generates data from lines. Then the "best" line is the one for which data generated from the line is "most likely". This is a commonly used technique in statistics: proposing a probabilistic model and using the probability of data to evaluate how good a particular model is. Let's make this more concrete.

A probabilistic model for linearly related data

We observe paired data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where we assume that as a function of x_i, each y_i is generated by using some true underlying line $y = \beta_0 + \beta_1 x$ that we evaluate at x_i, and then adding some Gaussian noise. Formally,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \tag{3.1}$$

Here, the noise ε_i represents the fact that our data won't fit the model perfectly. We'll model ε_i as being Gaussian: $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Note that the intercept β_0, the slope β_1, and the noise variance σ^2 are all treated as fixed (i.e., deterministic) but unknown quantities.

Solving for the fit: least-squares regression

Assuming that this is actually how the data $(x_1, y_1), \ldots, (x_n, y_n)$ we observe are generated, then it turns out that we can find the line for which the probability of the data is highest by solving the following optimization problem[3]:

$$\min_{\beta_0, \beta_1} : \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i) \right]^2, \tag{3.2}$$

where $\min_{\beta_0, \beta_1}$ means "minimize over β_0 and β_1". This is known as the **least-squares linear regression problem**. Given a set of points, the solution is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2} \tag{3.3}$$

$$= r \frac{s_y}{s_x}, \tag{3.4}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{3.5}$$

[3] This is an important point: the assumption of Gaussian noise leads to squared error as our minimization criterion. We'll see more regression techniques later that use different distributions and therefore different cost functions.

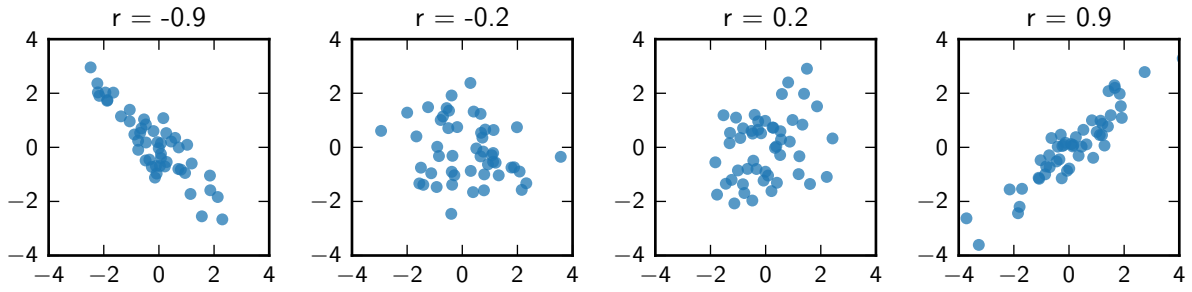


Figure 3.2: An illustration of correlation strength. Each plot shows data with a particular correlation coefficient r. Values farther than 0 (outside) indicate a stronger relationship than values closer to 0 (inside). Negative values (left) indicate an inverse relationship, while positive values (right) indicate a direct relationship.

where $\bar{x}$, $\bar{y}$, s_x and s_y are the sample means and standard deviations for x values and y values, respectively, and r is the **correlation coefficient**, defined as

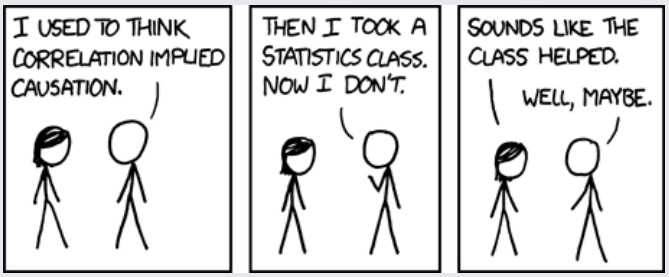
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right). \tag{3.6}$$

By examining the second equation for the estimated slope $\hat{\beta}_1$, we see that since sample standard deviations s_x and s_y are positive quantities, the correlation coefficient r, which is always between -1 and 1, measures how much x is related to y and whether the trend is positive or negative. Figure 3.2 illustrates different correlation strengths.

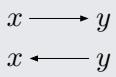
The square of the correlation coefficient r^2 will always be positive and is called the **coefficient of determination**. As we'll see later, this also is equal to the proportion of the total variability that's explained by a linear model.

As an extremely crucial remark, correlation does not imply causation! We devote the entire next page to this point, which is one of the most common sources of error in interpreting statistics.

Example: Correlation and Causation



Just because there's a strong correlation between two variables, there isn't necessarily a causal relationship between them. For example, drowning deaths and ice-cream sales are strongly correlated, but that's because both are affected by the season (summer vs. winter). In general, there are several possible cases, as illustrated below:

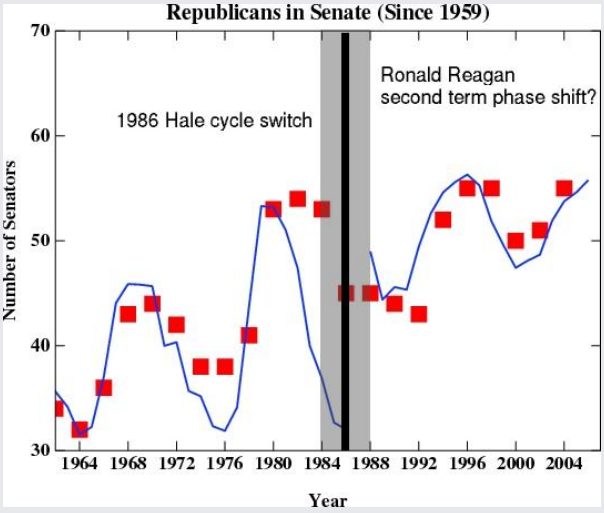


(a) **Causal link**: Even if there is a causal link between x and y, correlation alone cannot tell us whether y causes x or x causes y.

(b) **Hidden Cause**: A hidden variable z causes both x and y, creating the correlation.

(c) **Confounding Factor**: A hidden variable z and x both affect y, so the results also depend on the value of z.

(d) **Coincidence**: The correlation just happened by chance (e.g. the strong correlation between sun cycles and number of Republicans in Congress, as shown below).



(e) The number of Republican senators in congress (red) and the sunspot number (blue, before 1986)/inverted sunspot number (blue, after 1986). This figure comes from http://www.realclimate.org/index.php/archives/2007/05/fun-with-correlations/.

Figure 3.3: Different explanations for correlation between two variables. In this diagram, arrows represent causation.