4 Describing relationships between variables

This chapter provides methods that address a more involved problem of describing relationships between variables and require more computation. We start with relationships between two variables and move on to more.

4.1 Fitting a line by least squares

Goal:

We would like to use an equation to describe how a dependent (response) variable, y, changes in response to a change in one or more independendent (experimental) variable(s), x.

Line review: Recall a linear equation of the form $y = mx + b$

In statistics, we use the notation $y = \beta_0 + \beta_1 x + \epsilon$ where we assume β_0 and β_1 are unknown parameters and ϵ is some error.

The goal is to find estimates b_0 and b_1 for the parameters.

Example 4.1 (Plastic hardness)**.** Eight batches of plastic are made. From each batch one test item is molded and its hardness, y, is measured at time x. The following are the 8 measurements and times:

time	32	72	64	48	16	40	80	56
hardness	230	323	298	255	199	248	359	305

How do we find an equation for the line that best fits the data?

Definition 4.1. A *residual* is the vertical distance between the actual data point and a fitted line, $e = y - \hat{y}$.

The *principle of least squares* provides a method of choosing a “best” line to describe the data.

Definition 4.2. To apply the *principle of least squares* in the fitting of an equation for y to an n-point data set, values of the equation parameters are choen to minimize

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where $y_1, y_2, \ldots, y_n$ are the observed responses and $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are corresponding responses predicted or fitted by the equation.

We want to choose b_0 and b_1 to minimize

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$$

4.2 Fitting curves and surfaces by least squares

4.2.1 Polynomial regression

4.2.2 Multiple regression (surface fitting)