Lecture 29

Simple linear regression.

29.1 Method of least squares.

Suppose that we are given a sequence of observations

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

where each observation is a pair of numbers $X_,Y_i \in \mathbb{R}$. Suppose that we want to predict variable Y as a function of X because we believe that there is some underlying relationship between Y and X and, for example, Y can be approximated by a function of X, i.e. $Y \approx f(X)$. We will consider the simplest case when $f(x)$ is a linear function of x:

$$f(x) = \beta_0 + \beta_1 x.$$

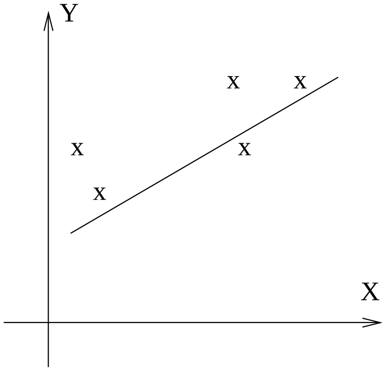

Figure 29.1: The least-squares line.

Of course, we want to find the line that fits our data best and one can define the measure of the quality of the fit in many different ways. The most common approach is to measure how Y_i is approximated by $\beta_0 + \beta_1 X_i$ in terms of the squared difference $(Y_i - (\beta_0 + \beta_1 X_i))^2$ which means that we measure the quality of approximation globally by the loss function

$$L = \sum_{i=1}^{n} (\underbrace{Y_i}_{actual} - (\underbrace{\beta_0 + \beta_1 X_i}_{estimate}))^2 \rightarrow \text{ minimize over } \beta_0, \beta_1$$

and we want to minimize it over all choices of parameters β_0, β_1. The line that minimizes this loss is called the *least-squares line*. To find the critical points we write:

$$\begin{aligned} \frac{\partial L}{\partial \beta_0} &= -\sum_{i=1}^{n} 2(Y_i - (\beta_0 + \beta_1 X_i)) = 0 \\ \frac{\partial L}{\partial \beta_1} &= -\sum_{i=1}^{n} 2(Y_i - (\beta_0 + \beta_1 X_i))X_i = 0 \end{aligned}$$

If we introduce the notations

$$\bar{X} = \frac{1}{n}\sum X_i, \ \bar{Y} = \frac{1}{n}\sum Y_i, \ \overline{X^2} = \frac{1}{n}\sum X_i^2, \ \overline{XY} = \frac{1}{n}\sum X_i Y_i$$

then the critical point conditions can be rewritten as

$$\beta_0 + \beta_1 \bar{X} = \bar{Y} \text{ and } \beta_0 \bar{X} + \beta_1 \overline{X^2} = \overline{XY}$$

and solving it for β_0 and β_1 we get

$$\beta_1 = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2} \text{ and } \beta_0 = \bar{Y} - \beta_1 \bar{X}.$$

If each X_i is a vector $X_i = (X_{i1}, \ldots, X_{ik})$ of dimension k then we can try to approximate Y_is as a linear function of the coordinates of X_i :

$$Y_i \approx f(X_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}.$$

In this case one can also minimize the square loss:

$$L = \sum (Y_i - (\beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}))^2 \rightarrow \text{ minimize over } \beta_0, \beta_1, \ldots, \beta_k$$

by taking the derivatives and solving the system of linear equations to find the parameters $\beta_0, \ldots, \beta_k$.

29.2 Simple linear regression.

First of all, when the response variable Y in a random couple (X, Y) is predicted as a function of X then one can model this situation by

$$Y = f(X) + \varepsilon$$

where the random variable ε is independent of X (it is often called *random noise*) and on average it is equal to zero: $\mathbb{E}\varepsilon = 0$. For a fixed X, the response variable Y in this model on average will be equal to $f(X)$ since

$$\mathbb{E}(Y|X) = \mathbb{E}(f(X) + \varepsilon|X) = f(X) + \mathbb{E}(\varepsilon|X) = f(X) + \mathbb{E}\varepsilon = f(X).$$

and $f(x) = \mathbb{E}(Y|X = x)$ is called the *regression function.*

Next, we will consider a *simple linear regression* model in which the regression function is linear, i.e. $f(x) = \beta_0 + \beta_1 x$, and the response variable Y is modeled as

$$Y = f(X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon,$$

where the random noise ε is assumed to have normal distribution $N(0, \sigma^2)$.

Suppose that we are given a sequence $(X_1, Y_1), \ldots, (X_n, Y_n)$ that is described by the above model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $N(0, \sigma^2)$. We have three unknown parameters - β_0, β_1 and σ^2 - and we want to estimate them using the given sample. Let us think of the points $X_1, \ldots, X_n$ as fixed and non random and deal with the randomness that comes from the noise variables ε_i. For a fixed X_i, the distribution of Y_i is equal to $N(f(X_i), \sigma^2)$ with p.d.f.

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - f(X_i))^2}{2\sigma^2}}$$

and the likelihood function of the sequence $Y_1, \ldots, Y_n$ is:

$$f(Y_1, \ldots, Y_n) = \Bigl(\frac{1}{\sqrt{2\pi}\sigma}\Bigr)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - f(X_i))^2} = \Bigl(\frac{1}{\sqrt{2\pi}\sigma}\Bigr)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2}.$$

Let us find the maximum likelihood estimates of β_0, β_1 and σ^2 that maximize this likelihood function. First of all, it is obvious that for any σ^2 we need to minimize

$$\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

over β_0, β_1 which is the same as finding the least-squares line and, therefore, the MLE for β_0 and β_1 are given by

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \text{ and } \hat{\beta}_1 = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2}.$$

Finally, to find the MLE of σ^2 we maximize the likelihood over σ^2 and get:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

Let us now compute the joint distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$. Since X_is are fixed, these estimates are written as linear combinations of Y_is which have normal distributions and, as a result, $\hat{\beta}_0$ and $\hat{\beta}_1$ will have normal distributions. All we need to do is find their means, variances and covariance. First, if we write $\hat{\beta}_1$ as

$$\hat{\beta}_1 = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2} = \frac{1}{n}\frac{\sum (X_i - \bar{X})Y_i}{\bar{X^2} - \bar{X}^2}$$

then its expectation can be computed:

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_1) &= \frac{\sum (X_i - \bar{X})\mathbb{E}Y_i}{n(\overline{X^2} - \bar{X}^2)} = \frac{\sum (X_i - \bar{X})(\beta_0 + \beta_1 X_i)}{n(\overline{X^2} - \bar{X}^2)} \\
&= \underbrace{\beta_0 \frac{\sum (X_i - \bar{X})}{n(\overline{X^2} - \bar{X}^2)}}_{=0} + \beta_1 \frac{\sum X_i (X_i - \bar{X})}{n(\overline{X^2} - \bar{X}^2)} = \beta_1 \frac{n\overline{X^2} - n\bar{X}^2}{n(\overline{X^2} - \bar{X}^2)} = \beta_1.
\end{aligned}$$

Therefore, $\hat{\beta}_1$ is unbiased estimator of β_1. The variance of $\hat{\beta}_1$ can be computed:

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}\Big(\frac{\sum (X_i - \bar{X})Y_i}{n(\overline{X^2} - \bar{X}^2)}\Big) = \sum \mathrm{Var}\Big(\frac{(X_i - \bar{X})Y_i}{n(\overline{X^2} - \bar{X}^2)}\Big) \\
&= \sum \Big(\frac{X_i - \bar{X}}{n(\overline{X^2} - \bar{X}^2)}\Big)^2 \sigma^2 = \frac{1}{n^2(\overline{X^2} - \bar{X}^2)^2} n(\overline{X^2} - \bar{X}^2)\sigma^2 \\
&= \frac{\sigma^2}{n(\overline{X^2} - \bar{X}^2)}.
\end{aligned}$$

Therefore, $\hat{\beta}_1 \sim N\Big(\beta_1, \frac{\sigma^2}{n(\overline{X^2} - \bar{X}^2)}\Big)$. A similar straightforward computations give:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \sim N\Big(\beta_0, \Big(\frac{1}{n} + \frac{\bar{X}^2}{n(\overline{X^2} - \bar{X}^2)}\Big)\sigma^2\Big)$$

and

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{X}\sigma^2}{n(\overline{X^2} - \bar{X}^2)}.$$