NMAI059 – Probability and Statistics 1

Mykhaylo Tyomkyn

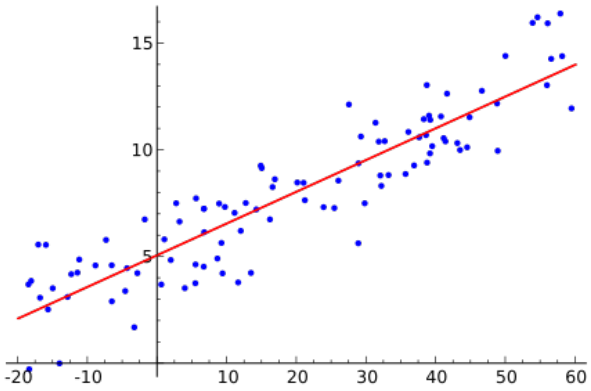
Lecture 14 - Linear regression.

Given n data points (x_i, y_i): $i = 1, \dots n$, suppose we are tasked with inferring a *linear* relationship between the two underlying random variables X and Y. That is, we suspect that (informally)

$$Y = \theta_0 + \theta_1 X + \text{“Noise”},$$

and we need to estimate θ_0 and θ_1. One common way of doing so the *Least squares method*: take $\hat{\theta}_0$ and $\hat{\theta}_1$ achieving the minimum[1]

$$\min_{\theta_0,\theta_1} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2. \tag{1}$$



This is not only good intuitively, but also has the following theoretical explanation. It is a common scenario that the “noise” is Gaussian. That is, we have

$$Y = \theta_0 + \theta_1 X + W,$$

where $W \sim N(0, \sigma^2)$ for some (typically unknown) $\sigma > 0$ and X and W are independent. The corresponding likelihood function is

$$L(x, y, \theta_0, \theta_1, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)^2\right). \tag{2}$$

Taking logs, we see that maximizing L amounts (for any σ) to minimizing $\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)^2$, which is exactly (1).

[1]Image: Sewaqu, Public domain, via Wikimedia Commons `https://commons.wikimedia.org/w/index.php?curid=11967659`

In order to solve (1), we take partial derivatives and equate them to 0. With $f(\theta_0, \theta_1) = \sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2$ we obtain

$$\frac{\partial f}{\partial \theta_0} = 2n\theta_0 + 2\theta_1(\sum_{i=1}^n x_i) - 2\sum_{i=1}^n y_i = 0,$$

resulting in

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}, \tag{3}$$

where $\bar{x} = (x_1 + \cdots + x_n)/n$ and $\bar{y} = (y_1 + \cdots + y_n)/n$. Using this and taking the partial derivative with respect to θ_1 yields

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{4}$$

Let us now give a broader interpretation of the rationale behind the least squares method and the resulting estimators. Assume, as before, that

$$Y = \theta_0 + \theta_1 X + W,$$

where $\mathbb{E}(W) = 0$ and W is independent of X, but let us no longer assume that W is Gaussian. Then, taking expectations,

$$\mathbb{E}(Y) = \theta_0 + \theta_1 \mathbb{E}(X) + 0,$$

so

$$\theta_0 = \mathbb{E}(Y) - \theta_1 \mathbb{E}(X). \tag{5}$$

So, heuristically, estimating $\mathbb{E}(X)$, $\mathbb{E}(Y)$ and θ_1 by $\bar{X}_n$, $\bar{Y}_n$ and $\hat{\theta}_1$ (an estimator for θ_1, to be defined), respectively, would make (3) a plausible estimator of θ_0.

What about $\hat{\theta}_1$? We claim that

$$\theta_1 = \frac{Cov(X,Y)}{Var(X)} = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\mathbb{E}(X^2) - \mathbb{E}(X)^2}. \tag{6}$$

Indeed, we may assume[2] that $\mathbb{E}(X) = 0$. Recalling that also $\mathbb{E}(W) = 0$, we are claiming that

$$\theta_1 = \frac{\mathbb{E}(XY)}{\mathbb{E}(X^2)}. \tag{7}$$

On the other hand, we know that

$$XY = \theta_0 X + \theta_1 X^2 + XW,$$

so

$$\mathbb{E}(XY) = \theta_0 \mathbb{E}(X) + \theta_1 \mathbb{E}(X^2) + \mathbb{E}(X)\mathbb{E}(W) \overset{(5)}{=} \theta_1 \mathbb{E}(X^2),$$

implying (7). So, estimating $Cov(X,Y)$ and $Var(X)$ in 6 by their empirical counterparts $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ and $\sum_{i=1}^n (x_i - \bar{x})^2$, respectively, results in the estimator $\hat{\theta}_1$ as in (4).

[2]Can you see why? Hint: $Cov(X + c, Y) = Cov(X, Y)$ and $Var(X + c) = Var(X)$ for any constant c.

Let us now go back to the Gaussian noise model: $W \sim \mathcal{N}(0, \sigma^2)$. The MLE for the "noise" σ is obtained by from (2):

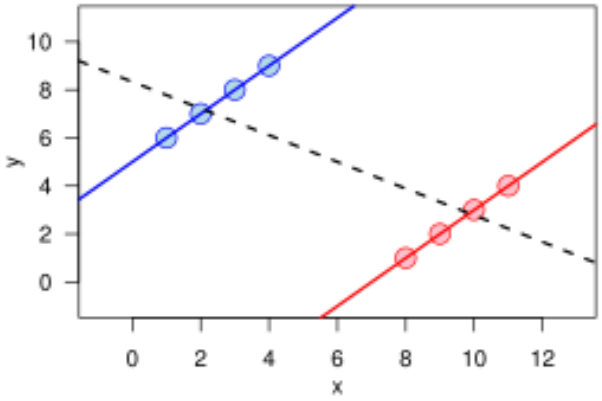
$$\frac{\partial \log L}{\partial \sigma} = \left(\frac{1}{\sigma^3} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2 \right) - \frac{n}{\sigma} = 0.$$

This gives

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i)^2},$$

where $\hat{\theta}_0$ and $\hat{\theta}_1$ are the least squares estimators from (3) and (4).

Linear regression is to be applied with caution. Simpson's paradox: an attempt to infer a linear dependency between X: time spent studying for an exam, and Y: the exam grade, may result in a negative slope (the less you study the better you do in the exam) if one accidentally mixes two classes of different difficulties. [3]



[3]Image: Schutz, Public domain, via Wikimedia Commons `https://commons.wikimedia.org/w/index.php?curid=2240877`