Linear statistical models

1. Introduction

The goal of this course is, in rough terms, to predict a variable y, given that we have the opportunity to observe variables $x_1, \ldots, x_{p-1}$. This is a very important statistical problem. Therefore, let us spend a bit of time and examine a simple example:

Given the various vital statistics of a newborn baby, you wish to predict his height y at maturity. Examples of those "vital statistics" might be x_1 := present height, and x_2 := present weight. Or perhaps there are still more predictive variables x_3 := your height and x_4 := your spouse's height, etc.

In actual fact, a visit to your pediatrist might make it appear that this prediction problem is trivial. But that is not so [though it is nowadays fairly well understood in this particular case]. A reason the problem is nontrivial is that there is no *a priori* way to know "how" y depends on $x_1, \ldots, x_4$ [say, if we plan to use all 4 predictive variables]. In such a situation, one resorts to writing down a reasonable model for this dependence structure, and then analyzing that model. [Finally, there might be need for model verification as well.]

In this course, we study the general theory of "linear statistical models." That theory deals with the simplest possible nontrivial setting where such problems arises in various natural ways. Namely, in that theory we *posit* that y is a linear function of $(x_1, \ldots, x_4)$, possibly give or take some "noise." In other words, the theory of linear statistical models posits that

there exist unknown parameters $\beta_0, \ldots, \beta_{p-1}$ [here, $p = 5$] such that

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon, \tag{1}$$

where ε is a random variable. The problem is still not fully well defined [for instance, what should be the distribution of ε, etc.?]. But this is roughly the starting point of the theory of linear statistical models. And one begins asking natural questions such as, "how can we estimate $\beta_0, \ldots, \beta_4$?," or "can we perform inference for these parameters?" [for instance, can we test to see if y does not depend on x_1 in this model; i.e., test for $H_0 : \beta_1 = 0$?]. And so on.

We will also see, at some point, how the model can be used to improve itself. For instance, suppose we have only one predictive variable x, but believe to have a nonlinear dependence between y and x. Then we could begin by thinking about *polynomial regression*; i.e., a linear statistical model of the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_{p-1} x^{p-1} + \varepsilon.$$

Such a model fits in the general form (1) of linear statistical models, as well: We simply define new predictive variables $x_j := x^j$ for all $1 \le j < p$. One of the conclusions of this discussion is that we are studying models that are linear functions of unknown parameters $\beta_0, \ldots, \beta_{p-1}$ and not $x_1, \ldots, x_{p-1}$. This course studies statistical models with such properties. And as it turns out, not only these models are found in a great number of diverse applications, but also they have a rich mathematical structure.

2. The method of least squares

Suppose we have observed n data points in pairs: $(x_1, y_1), \ldots (x_n, y_n)$. The basic problem here is, what is the best straight line that fits this data? There is of course no unique sensible answer, because "best" might mean different things.

We will use the *method of least squares*, introduced by C.-F. Gauss. Here is how the method works: If we used the line $y = \beta_0 + \beta_1 x$ to describe how the x_i's affect the y_i's, then the error of approximation, at $x = x_i$, is $y_i - (\beta_0 + \beta_1 x_i)$; this is called the ith *residual error*. The sum of the squared residual errors is $\text{SSE} := \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$, and the method of least squares is to find the line with the smallest SSE. That is, we need to find the optimal β_0 and β_1—written $\hat{\beta}_0$ and $\hat{\beta}_1$—that solve the following optimization problem:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 . \tag{2}$$

Theorem 1 (Gauss). *The least-squares solution to* (2) *is given by*

$$\hat{\beta}_1 := \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \textit{and} \quad \hat{\beta}_0 := \bar{y} - \hat{\beta}_1 \bar{x}.$$

Proof. Define

$$L(\beta_0, \beta_1) := \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 .$$

Our goal is to minimize the function L. An inspection of the graph of L shows that L has a unique minimum; multivariable calculus then tells us that it suffices to set $\partial L/\partial \beta_j = 0$ for $j = 1, 2$ and solve. Because

$$\begin{aligned} \frac{\partial}{\partial \beta_0} L(\beta_0, \beta_1) &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) , \\ \frac{\partial}{\partial \beta_1} L(\beta_0, \beta_1) &= -2 \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) , \end{aligned}$$

a few lines of simple arithmetic finish the derivation. □

The preceding implies that, given the points $(x_1, y_1), \ldots, (x_n, y_n)$, the best line of fit through these points—in the sense of least squares—is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{3}$$

For all real numbers x and y, define x_{SU} and y_{SU} to be their respective "standardizations." That is,

$$x_{\text{SU}} := \frac{x - \bar{x}}{\sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}}, \qquad y_{\text{SU}} := \frac{y - \bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Then, (3) can be re-written in the following equivalent form:

$$\begin{aligned} y_{\text{SU}} &= \frac{\hat{\beta}_0 - \bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2}} + \frac{\hat{\beta}_1 x}{\sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{\hat{\beta}_1}{\sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2}} (x - \bar{x}) , \end{aligned}$$

the last line following from the identity $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. We re-write the preceding again:

$$y_{\text{SU}} = \frac{\hat{\beta}_1 \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2}} x_{\text{SU}}.$$

Because

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2},$$

we can re-write the best line of fit, yet another time, this time as the following easy-to-remember formula:

$$y_{\text{SU}} = r x_{\text{SU}},$$

where r [a kind of "correlation coefficient"] is defined as

$$r := \frac{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

3. Simple linear regression

Suppose $Y_1, \ldots, Y_n$ are observations from a distribution, and they satisfy

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (1 \le i \le n), \tag{4}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are [unobserved] i.i.d. $\text{N}(0, \sigma^2)$ for a fixed [possibly unknown] $\sigma > 0$. We assume that $x_1, \ldots, x_n$ are known, and seek to find the "best" β_0 and β_1.[1]

In other words, we believe that we are observing a certain linear function of the variable x at the x_i's, but our measurement [and/or modeling] contains noise sources [ε_i's] which we cannot observe.

Theorem 2 (Gauss). *Suppose $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with common distribution $\text{N}(0, \sigma^2)$, where $\sigma > 0$ is fixed. Then the maximum likelihood estimators of β_1 and β_0 are, respectively,*

$$\hat{\beta}_1 = \frac{\sum_{j=1}^n (x_j - \bar{x})(Y_j - \bar{Y})}{\sum_{j=1}^n (x_j - \bar{x})^2} \quad \textit{and} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

Therefore, based on the data $(x_1, Y_1), \ldots, (x_n, Y_n)$, we predict the y-value at $x = x_*$ to be $y_* := \hat{\beta}_0 + \hat{\beta}_1 x_*$.

Proof. Note that $Y_1, \ldots, Y_n$ are independent [though not i.i.d.], and the distribution of Y_j is $\text{N}(\beta_0 + \beta_1 x_j, \sigma^2)$. Therefore, the joint probability density function of $(Y_1, \ldots, Y_n)$ is

$$f(y_1, \ldots, y_n) := \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{j=1}^n (y_j - \beta_0 - \beta_1 x_j)^2 \right].$$

[1] In actual applications, the x_i's are often random. In such a case, we assume that the model holds after conditioning on the x_i's.

According to the MLE principle we should maximize $f(Y_1, \ldots, Y_n)$ over all choices of β_0 and β_0. But this is equivalent to minimizing $L(\beta_0, \beta_1) := \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2$. But this was exactly what we did in Theorem 1 [except for real variables $y_1, \ldots, y_n$ in place of the random variables $Y_1, \ldots, Y_n$]. □

In this way we have the following "regression equation," which uses the observed data $(x_1, Y_1), \ldots, (x_n, Y_n)$ in order to predict a y-value corresponding to $x = x_*$:

$$Y(x) = \hat{\beta}_0 + \hat{\beta}_1 x.$$

But as we shall see, this method is good not only for prediction, but also for inference. Perhaps a first important question is, "do the y's depend linearly on the x's"? Mathematically speaking, we are asking to test the hypothesis that $\beta_1 = 0$. If we could compute the distribution of $\hat{\beta}_1$, then standard methods can be used to accomplish this. We will see later on how this can be accomplished. But before we develop the theory of linear inference we need to know a few things about linear algebra and some of its probabilistic consequences.