SIMPLE LINEAR REGRESSION

DAVAR KHOSHNEVISAN

1. Line-Fitting

Points $(x_1, y_1), \ldots, (x_n, y_n)$ are given; e.g. on a scatterplot.

> What is "the best line" that describes the relationship between the x's and the y's?

To understand this better, let us focus on a line $L(x) = \alpha + \beta x$ where $\alpha, \beta \in \mathbf{R}$ are fixed but otherwise arbitrary.

"Fitting L to the points (x_i, y_i)" means estimating y_i by $L(x_i)$ for all $i = 1, \ldots, n$. The error, e_i, in estimating y_i by $L(x_i)$ is called the $i^{\underline{th}}$ *residual*.

We choose "the best line of fit" according to the *least squares principle* of C. Gauss:[1] Minimize $\sum_{i=1}^n e_i^2$ among all possible lines of the form $L(x) = \alpha + \beta x$. This is done by doing a little calculus: For a fixed line $L(x) = \alpha + \beta x$,

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (L(x_i) - y_i)^2 = \sum_{i=1}^n (\alpha + \beta x_i - y_i)^2 . \tag{1}$$

Call this $\mathscr{H}(\alpha, \beta)$. Then we are asked to find a and b that minimize $\mathscr{H}$. But this is a two-variable calculus problem. It turns out to be enough to solve:

$$\frac{\partial \mathscr{H}(\alpha, \beta)}{\partial \alpha} = 0 \quad \text{and} \quad \frac{\partial \mathscr{H}(\alpha, \beta)}{\partial \beta} = 0. \tag{2}$$

Date: August 30, 2004.

[1] Sometimes, people have reason to use other, more "robust," principles. A common alternative in such a case is the principle of "least absolute deviation." It seeks to find a line that minimizes $\sum_{i=1}^n |e_i|$. Occasionally, this is also called the "$\mathscr{L}^1$ method"; this is to distinguish it from least squares which is also sometimes called the "$\mathscr{L}^2$ method."

First,

$$
\begin{aligned}
\frac{\partial \mathscr{H}(\alpha,\beta)}{\partial\alpha} &= \sum_{i=1}^n 2(\alpha+\beta x_i - y_i)\\
&= 2\alpha n + 2\beta\sum_{i=1}^n x_i - 2\sum_{i=1}^n y_i, \text{ and}\\
\frac{\partial \mathscr{H}(\alpha,\beta)}{\partial\beta} &= \sum_{i=1}^n 2(\alpha+\beta x_i - y_i)x_i\\
&= 2\alpha\sum_{i=1}^n x_i + 2\beta\sum_{i=1}^n x_i^2 - 2\sum_{i=1}^n x_i y_i.
\end{aligned}
\tag{3}
$$

These are called the *normal equations*. Now (2) is transformed into two equations in two unknowns [α and β]:

$$
\begin{aligned}
\alpha\overline{x} + \beta\overline{x^2} &= \overline{xy}\\
\alpha + \beta\overline{x} &= \overline{y}.
\end{aligned}
\tag{4}
$$

Multiply the second equation of (4) by $\overline{x}$, and then subtract from the first equation to find that $\beta(\overline{x^2} - (\overline{x})^2) = \overline{xy} - \overline{x}\cdot\overline{y}$. You should recognize this, in statistical terms, as $\beta\mathrm{Var}(x) = \mathrm{Cov}(x,y)$. Equivalently, $\beta = \mathrm{Cov}(x,y)/\mathrm{Var}(x) = \mathrm{Corr}(x,y)\mathrm{SD}_y/\mathrm{SD}_x$. Plug this into the second equation of (4) to find that $\alpha = \overline{y} - \beta\overline{x}$ for the computed β. To summarize,

Theorem 1.1. *The least-squares line through $(x_1,y_1),\ldots,(x_n,y_n)$ is unique and defined by*

$$
L(x) = \overline{y} + \widehat{\beta}(x-\overline{x}), \quad \textit{where} \quad \widehat{\beta} = \mathrm{Corr}(x,y)\frac{\mathrm{SD}_y}{\mathrm{SD}_x}.
\tag{5}
$$

2. The Measurement-Error Model

Let $\mathbf{Y} = (Y_1,\ldots,Y_n)$ be a random sample of n i.i.d. copies of the response variable. The *measurement-error* model posits the following:

$$
Y_i = \alpha + \beta X_i + \varepsilon_i \qquad i = 1,\ldots,n.
\tag{6}
$$

Here, $\mathbf{X} = (X_1,\ldots,X_n)$ is a non-random vector of constants—the explanatory variables—and α and β are unknown parameters. This model also assumes that the ε_i's are i.i.d. $N(0,\sigma^2)$ for an unknown parameter $\sigma > 0$. The Y_i's are random only because the ε_i's are (and not the X_i's).

According to the principle of least squares (Theorem 1.1), the best least-squares estimates of α and β are, respectively,

$$
\widehat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i - \overline{X}\cdot\overline{Y}}{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2} \quad \text{and} \quad \widehat{\alpha} = \overline{Y} - \widehat{\alpha}\overline{X}.
\tag{7}
$$

The more imortant parameter is β. For instance, consider the test,

$$
H_0 : \beta = 0 \quad \text{vs.} \quad H_1 : \beta \neq 0.
\tag{8}
$$

This is testing the hypothesis that the explanatory variable X has a (linear) effect on Y.

So we need the distribution of $\widehat{\beta}$. Note that

$$\widehat{\beta} = \frac{\sum_{i=1}^n Y_i \left(X_i - \overline{X}\right)}{\sum_{i=1}^n (X_i - \overline{X})^2} = \sum_{i=1}^n b_i Y_i, \tag{9}$$

where

$$b_i = \frac{X_i - \overline{X}}{n s_{\mathbf{X}}^2} \quad \text{for} \quad s_{\mathbf{X}}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2. \tag{10}$$

Recall that the X_i's are not random. Therefore, neither are the b_i's. Also recall,

Lemma 2.1. *If $V_1, \ldots, V_n$ are independent and $V_i \sim N(\mu_i, \sigma_i^2)$, then for all non-random $c_1, \ldots, c_n$, $\sum_{i=1}^n c_i V_i \sim N(\mu, \sigma^2)$ where $\mu = \mu_1 + \cdots + \mu_n$ and $\sigma^2 = \sigma_1^2 + \cdots + \sigma_2^2$.*

Consequently, $\widehat{\beta} \sim N(\sum_{i=1}^n b_i E[Y_i], \sum_{i=1}^n b_i^2 \mathrm{Var}(Y_i))$. But $Y_i = \alpha + \beta X_i + \varepsilon_i$. So, $E[Y_i] = \alpha + \beta X_i$, and $\mathrm{Var}(Y_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2$. It is easy to check that: (i) $\sum_{i=1}^n b_i = 0$; (ii) $\sum_{i=1}^n b_i X_i = 1$; and (iii) $\sum_{i=1}^n b_i^2 = 1/(n s_{\mathbf{X}}^2)$. This proves that

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{n s_{\mathbf{X}}^2}\right). \tag{11}$$

Therefore, $E[\widehat{\beta}] = \beta$. That is, $\widehat{\beta}$ is an unbiased estimator of β. Moreover, if we knew σ^2, then we could perform the test of hypothesis (8) at any predescribed level, say at 95%. The trouble is that we generally do not know σ^2.

Because the Y_i's have variance σ^2, we can estimate σ^2 by $s_{\mathbf{Y}}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2$. But then we need the joint distribution of $(\widehat{\beta}, s_{\mathbf{Y}}^2)$. The key to this theory is that $\widehat{\beta}$ is independent of $s_{\mathbf{Y}}^2$. We just determined the distribution of $\widehat{\beta}$, and we will see later on that the H_0-distribution of $s_{\mathbf{Y}}^2$ is essentially χ^2. The rest will be smooth sailing.

To recap, we need to accomplish two things:

(1) Derive the independence of $\widehat{\beta}$ and $s_{\mathbf{Y}}^2$; and
(2) Honestly compute the distribution of $s_{\mathbf{Y}}^2$ under H_0.

Just about all of this semester's work is concerned with accomplishing these two goals (for more general models).