Linear predictors and the law of large numbers

Linear predictors.

Let X and Y be dependent random variables. We want to use the method of least squares to find a linear function of Y to approximate X. The first step is finding t which minimizes

$$\begin{aligned}\operatorname{Var}(X - tY) &= \operatorname{Cov}(X - tY, X - tY) = \operatorname{Cov}(X, X) - 2t\operatorname{Cov}(X, Y) + t^2\operatorname{Cov}(Y, Y) \\ &= \operatorname{Var}(X) - 2t\operatorname{Cov}(X, Y) + t^2\operatorname{Var}(Y).\end{aligned} \tag{1}$$

Since $C - 2tB + At^2$ is minimized when $t = B/A$, we conclude that $\operatorname{Var}(X - tY)$ is minimized when

$$t = \frac{\operatorname{Cov}(X, Y)}{\operatorname{Var}(Y)}. \tag{2}$$

The second step is adding a constant to adjust the expected value to be correct, i.e. finding C such that

$$E[tY + C] = E[X], \tag{3}$$

or in other words

$$C = E[X] - tE[Y].$$

Thus the linear function of Y which best approximates X is

$$X \approx tY + C, \qquad t = \frac{\operatorname{Cov}(X, Y)}{\operatorname{Var}(Y)}, \qquad C = E[X] - tE[Y].$$

The random variable $tY + C$ is the *best linear predictor* of X using Y.

Example.

Let X be a uniform random variable on $[0, 2b]$, and let

$$I = \begin{cases} 1, & \text{if } X \geq b, \\ 0, & \text{if } X < b. \end{cases}$$

Then I is a Bernoulli random variable with $p = 1/2$, so $E[I] = 1/2$ and $\operatorname{Var}(I) = 1/4$. Moreover, $E[X] = b$, and $E(XI) = \int_b^{2b} \frac{x}{2b}\, dx = 3b/4$, so $\operatorname{Cov}(X, I) = b/4$, and we get

$$t = b, \qquad C = b/2, \qquad X \approx bI + \frac{b}{2} = \begin{cases} \frac{3b}{2}, & \text{if } X \geq b, \\ \frac{b}{2}, & \text{if } X < b. \end{cases}$$

Exercise.

Let X be a standard normal random variable, and let

$$I = \begin{cases} 1, & \text{if } X \geq 0, \\ 0, & \text{if } X < 0. \end{cases}$$

Find the best linear predictor of X using I.

Kiril Datchev, December 6, 2025. Questions, comments, and corrections are gratefully received at kdatchev@purdue.edu.

Accuracy.

By (3), the best linear predictor $tY + C$ has the same mean as X. We measure its accuracy by inserting the value of t from (2) into the variance equation (1) to obtain

$$\operatorname{Var}(X - tY) = \operatorname{Var}(X) - \frac{\operatorname{Cov}(X,Y)^2}{\operatorname{Var}(Y)}. \tag{4}$$

Thus, in the example above, we have

$$\operatorname{Var}(X - tI) = \frac{b^2}{3} - \frac{b^2}{4} = \frac{b^2}{12},$$

which matches the variance of a uniform random variable on an interval of length b.

Inserting (4) into $\operatorname{Var}(X - tY) \geq 0$ and rearranging yields the *Cauchy–Schwarz inequality*:

$$\operatorname{Cov}(X,Y)^2 \leq \operatorname{Var}(X) \operatorname{Var}(Y),$$

which allows us to control the covariance of two variables in terms of their individual variances.

Law of large numbers.

We can use the Cauchy–Schwarz inequality to prove versions of the law of large numbers for dependent variables. For example, let $X_1, X_2, \ldots$ be a sequence of random variables such that $\operatorname{Var}(X_j) \leq \sigma^2$ for all j. Then by Cauchy–Schwarz we also have

$$\operatorname{Cov}(X_j, X_k) \leq \sigma^2 \text{ for all } j \text{ and } k. \tag{5}$$

Suppose now that there is a constant L such that

$$\text{if } |j-k| \geq L \text{ then } \operatorname{Cov}(X_j, X_k) = 0; \tag{6}$$

in other words, there may be a correlation between variables which are nearby but not between variables which are far away. Then the law of large numbers states that

$$\lim_{n\to\infty} P\Big(\Big|\frac{X_1 + \cdots + X_n}{n} - \mu_n\Big| > \varepsilon\Big) = 0, \tag{7}$$

for every $\varepsilon > 0$, where

$$\mu_n = E\Big(\frac{X_1 + \cdots + X_n}{n}\Big).$$

To prove (7), write

$$\begin{aligned} P\Big(\Big|\frac{X_1 + \cdots + X_n}{n} - \mu_n\Big| > \varepsilon\Big) &\leq \frac{1}{\varepsilon^2} \operatorname{Var}\Big(\frac{X_1 + \cdots + X_n}{n}\Big) \\ &= \frac{1}{n^2\varepsilon^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \operatorname{Cov}(X_j, X_k) \\ &\leq \frac{1}{n^2\varepsilon^2} (2L+1) n \sigma^2, \end{aligned}$$

where in the first step we used Chebyshev's inequality, in the second we expanded in the manner of (1), and in the third we used the fact that, for each of the n values of j, at most $2L+1$ summands are nonzero by (6), and each nonzero summand is $\leq \sigma^2$ by (5). Hence

$$0 \leq P\Big(\Big|\frac{X_1 + \cdots + X_n}{n} - \mu_n\Big| > \varepsilon\Big) \leq \frac{(2L+1)\sigma^2}{n\varepsilon^2},$$

so (7) follows by the squeeze theorem.