

Prediction + miscellaneous

1 Prediction

(Source: Amemiya, ch. 4)

Best Linear Predictor: a motivation for linear univariate regression

Consider two random variables X and Y . What is the “best” predictor of Y , among all the possible linear functions of X ?

“Best” linear predictor minimizes the mean squared error of prediction:

$$\min_{\alpha, \beta} E(Y - \alpha - \beta X)^2. \quad (1)$$

The first-order conditions are:

$$\text{For } \alpha: 2\alpha - 2EY + 2\beta EX = 0$$

$$\text{For } \beta: 2\beta EX^2 - 2EXY + 2\alpha EX = 0.$$

Solving:

$$\begin{aligned} \beta^* &= \frac{Cov(X, Y)}{VX} \\ \alpha^* &= EY - \beta^* EX \end{aligned} \quad (2)$$

These are the coefficients obtained in a linear regression of Y on a single variable X .

Important: β^* does not measure the change in Y caused by a change in X . Difference between prediction and causation. The latter implies a “counterfactual” change in X .

Let $\hat{Y} \equiv \alpha^* + \beta^* X$ denote a “fitted value” of Y , and $U \equiv Y - \hat{Y}$ denote the “residual” or prediction error:

- $EU = 0$
- $V\hat{Y} = (\beta^*)^2 VX = (Cov(X, Y))^2 / VX = \rho_{XY}^2 VY$
- $VU = VY + (\beta^*)^2 VX - 2\beta^* Cov(X, Y) = VY - (Cov(X, Y))^2 / VX = (1 - \rho_{XY}^2) VY$

Hence, the b.l.p. accounts for a ρ_{XY}^2 proportion of the variance in Y ; in this sense, the correlation measures the linear relationship between Y and X .

Also note that

$$\begin{aligned}
Cov(\hat{Y}, U) &= Cov(\hat{Y}, Y - \hat{Y}) \\
&= E[(\hat{Y} - E\hat{Y})(Y - \hat{Y} - EY + E\hat{Y})] \\
&= E[(\hat{Y} - E\hat{Y})(Y - EY) - (\hat{Y} - E\hat{Y})(\hat{Y} - E\hat{Y})] \\
&= Cov(\hat{Y}, Y) - V\hat{Y} \\
&= E[(\alpha^* + \beta^*X - \alpha^* - \beta^*EX)(Y - EY)] - V\hat{Y} \tag{3} \\
&= \beta^*E[(X - EX)(Y - EY)] - V\hat{Y} \\
&= \beta^*Cov(X, Y) - V\hat{Y} \\
&= Cov^2(X, Y)/VX - Cov^2(X, Y)/VX \\
&= 0.
\end{aligned}$$

Similarly, $Cov(X, U) = 0$.

Hence, for any random variable X , the random variable Y can be written as the sum of a part which is a linear function of X , and a part which is uncorrelated with X . This decomposition of Y is done when you regress Y on X .

Finally, note that (obviously) the BLP of the BLP – that is, the best linear predictor given X of the BLP of Y given X – is just the BLP itself. There is no gain, in predicting Y , by iterating the procedure.

■■■

Predictive vs. Causal Model

Next assume that the “true” model describing the generation of the Y process is:

$$Y = \alpha^0 + \beta^0 X + \epsilon, \quad E\epsilon = 0. \tag{4}$$

What we mean by true model is that this is a causal model in the sense that a one-unit increase in X would raise Y by β^0 units. (In the previous section, we just assume that Y, X move jointly together, so there is no sense in which changes in X “cause” changes in Y .)

Question: under what assumptions does the best linear predictor of Y on X recover the true model; ie. $\alpha^* = \alpha^0$, and $\beta^* = \beta^0$?

- For α^* :

$$\begin{aligned}\alpha^* &= EY - \beta^* EX \\ &= \alpha^0 + \beta^0 EX + E\epsilon - \beta^* EX\end{aligned}$$

which is equal to α^0 if $\beta^0 = \beta^*$.

- For β^* :

$$\begin{aligned}\beta^* &= \frac{Cov(\alpha^0 + \beta^0 X + \epsilon, X)}{Var X} \\ &= \frac{1}{Var X} \cdot \{E[X(\alpha^0 + \beta^0 X + \epsilon)] - EX \cdot E[\alpha^0 + \beta^0 X + \epsilon]\} \\ &= \frac{1}{Var X} \cdot \{\alpha^0 EX + \beta^0 EX^2 + E[\epsilon X] - \alpha^0 EX - \beta^0 [EX]^2 - EXE\epsilon\} \\ &= \frac{1}{Var X} \cdot \{\beta^0 [EX^2 - (EX)^2] + E[\epsilon X]\}\end{aligned}$$

which is equal to β^0 if

$$E[\epsilon X] = 0.$$

This is an “exogeneity” assumption, that (roughly) X and the disturbance term ϵ are uncorrelated. Under this assumption, the best linear predictors from the problem (1) coincide with the true values of α^0 , β^0 . Such an exogeneity assumption also plays a key role to ensure that the estimates from least-squares regressions can be interpreted “causally”.



Best prediction

Generalize above results to general (not just linear) prediction.

What if we don't restrict ourselves to linear function of X ? What general function of X is optimal predictor of Y ?

$$\min_{\phi(\cdot)} E[Y - \phi(X)]^2.$$

Note:

$$\begin{aligned}& E[Y - \phi(X)]^2 \\ &= E[(Y - E(Y|X)) + (E(Y|X) - \phi(X))]^2 \\ &= E(Y - E(Y|X))^2 + 2E(Y - E(Y|X))(E(Y|X) - \phi(X)) + E(E(Y|X) - \phi(X))^2.\end{aligned}\tag{5}$$

The middle term is

$$\begin{aligned}
& E_{X,Y} (Y - E(Y|X)) (E(Y|X) - \phi(X)) \\
&= E_X E_{Y|X} (Y - E(Y|X)) (E(Y|X) - \phi(X)) \\
&= E_X [E_{Y|X} (Y E(Y|X)) - E(Y|X)^2 - \phi(X) E_{Y|X} Y + \phi(X) E(Y|X)] \\
&= E_X [E(Y|X)^2 - E(Y|X)^2 - \phi(X) E(Y|X) + \phi(X) E(Y|X)] = 0.
\end{aligned}$$

Hence, Eq. (5) is clearly minimized when

$$\phi(X) = E(Y|X) :$$

the conditional expectation of Y given X is the best predictor.

Define the residual $U \equiv Y - E(Y|X)$. As in the b.l.p. case, $Cov(U, E(Y|X)) = 0$:

$$\begin{aligned}
EU &= E_X E_{Y|X} (Y - E(Y|X)) = E_X 0 = 0 \\
E[UE(Y|X)] &= E_X [E(Y|X) E_{Y|X} U] = E_X [E(Y|X) \cdot 0] = 0.
\end{aligned}$$

Also, by similar calculations, $Cov(U, X) = 0$.

Note how useful the law of iterated expectations is.

■■■

Both the b.l.p. and b.p. are examples of **projections** of Y onto spaces of functions of X . More precisely:

Definition: a projection of a random variable Y onto a space \mathcal{S} is the element $\hat{S} \in \mathcal{S}$ which minimizes

$$\min_{S \in \mathcal{S}} E(Y - S)^2.$$

Projection theorem: Let \mathcal{S} be a linear space of random variables with finite second moments. Then \hat{S} is the projection of Y onto \mathcal{S} if and only if $\hat{S} \in \mathcal{S}$ and the orthogonality condition is satisfied:

$$ES(Y - \hat{S}) = 0, \quad \forall S \in \mathcal{S}. \quad (6)$$

The projection \hat{S} is unique.

Proof: For any S and \hat{S} in \mathcal{S} :

$$E(Y - S)^2 = E(Y - \hat{S})^2 + 2E(Y - \hat{S})(\hat{S} - S) + E(\hat{S} - S)^2.$$

If \hat{S} satisfies the orthogonality condition, then the middle term is zero, and we conclude that $E(Y - S)^2 \geq E(Y - \hat{S})^2$ with strict inequality unless $E(\hat{S} - S)^2 = 0$. Thus the orthogonality condition implies that \hat{S} is a projection, and also that it is unique.

Conversely, assume that \hat{S} is a projection. For any number α ,

$$E(Y - S - \alpha\hat{S})^2 - E(Y - \hat{S})^2 = -2\alpha E(Y - \hat{S})S + \alpha^2 ES^2.$$

Because \hat{S} is a projection, this expression is nonnegative for every α (note that the vector $S + \alpha\hat{S} \in \mathcal{S}$). But the parabola $-2\alpha E(Y - \hat{S})S + \alpha^2 ES^2$ is nonnegative iff $E(Y - \hat{S})S = 0$ – the orthogonality condition – is satisfied. ■

■■■

Since the projection of Y is an element of \mathcal{S} , there is no gain from iterating the projection. The projection of the projection is just the projection itself. That is, projections are *idempotent* operators; idempotency is even a defining feature of a projection.

■■■■■

Miscellaneous stuff

(Source: Amemiya, ch. 5)

1a. Bernoulli distribution

A *Bernoulli* random variable Y (with parameter p) is $= \begin{cases} 1 & \text{with prob } p \\ 0 & \text{with prob } 1 - p \end{cases}$

The characteristic function of a Bernoulli random variable Y is:

$$\begin{aligned} \phi_Y(t) &= \mathbb{E}[\exp(itY)] = \exp(it \cdot 1)p + \exp(it \cdot 0)(1 - p) \\ &= 1 - p + p \exp(it). \end{aligned}$$

1b. Binomial distribution

Consider Y_1, Y_2, \dots, Y_n be i.i.d. Bernoulli variables with parameter p . Then the random variable $X = \sum_i Y_i$ is a *binomial* random variable. We write $X \sim B(n, p)$.

Example: X is number of heads in n coin tosses.

For the binomial random variable we have:

$$\begin{aligned} P(X = k) &= \binom{n}{k} p^k (1 - p)^{n-k}, \text{ where } \binom{n}{k} = \frac{n!}{(n-k)! k!} \\ EX &= np \\ VX &= np(1 - p) \end{aligned}$$

The characteristic function of a Binomial random variable $X(n, p)$ is just the n -fold product of the CF for Bernoulli:

$$\begin{aligned}\phi_X(t) &= \mathbb{E}[\exp(itX)] = \mathbb{E}\left[\exp(it \sum_i Y_i)\right] = \prod_{i=1}^n \mathbb{E}[\exp(itY_i)] \\ &= [\phi_Y(t)]^n = [1 - p + p \exp(it)]^n\end{aligned}$$

1c. Poisson distribution

Consider a binomial random variable $X(n, p)$ where $p = \lambda/n$ for fixed $\lambda > 0$. Then as $n \rightarrow \infty$,

$$\phi_{X_n}(t) \rightarrow \exp(\lambda(\exp(it) - 1)).$$

(We will prove this later.)

This is the characteristic function of a *Poisson* random variable X with parameter $\lambda > 0$. X is discrete and takes non-negative values $0, 1, 2, 3, \dots$

- pmf: $P(X = k) = \lambda^k e^{-\lambda} / k!$, for $k = 0, 1, 2, \dots$.
- cdf can be computed by summation.
- $\mathbb{E}X = \text{Var}(X) = \lambda$.

The Poisson distribution is used to model discrete frequencies of independent events (think binomial) in a given time interval. On average, λ events occur during an interval.

The corresponding distribution of waiting times follows an *exponential distribution* with pdf $f(t; \lambda) = \lambda \exp(-\lambda t)$ and cdf $F(t; \lambda) = 1 - \exp(-\lambda t)$. $\mathbb{E}t = 1/\lambda$ and $\text{Var}(t) = 1/\lambda^2$. On average, waiting time is $1/\lambda$.

2. Normal distribution

- Univariate normal (Gaussian) density:

$$X \sim N(\mu, \sigma^2) \Leftrightarrow f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right]$$

$\mathbb{E}X = \mu$, $\text{Var}X = \sigma^2$. Normal density is symmetric and bell-shaped around μ .

- Let $X \sim N(\mu, \sigma^2)$ and let $Y = a + bX$. Then $Y \sim N(a + b\mu, b^2\sigma^2)$.

Corollary: $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$, the *standard normal* distribution, with density $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$.

Some interpretation: $\phi(x) \propto \exp\left(-\frac{x^2}{2}\right)$, with the constant of proportionality equal to $\int \exp\left(-\frac{x^2}{2}\right) dx = \sqrt{2\pi}$ so that the density integrates to 1.

- Characteristic function

$$\phi_{N(0,1)}(t) = e^{-t^2/2}.$$

Correspondingly,

$$\phi_{N(\mu, \sigma^2)}(t) = e^{-i\mu t} \phi_{N(0,1)}(\sigma t) = e^{-i\mu t + \frac{1}{2}\sigma^2 t^2}.$$

- Multivariate normal random variables: $\vec{X} = (X_1, X_2, \dots, X_n)'$ is multivariate normal with mean vector $\vec{\mu}$ and covariance matrix Σ , denoted $\vec{X} \sim N(\vec{\mu}, \Sigma)$.

Joint density is

$$f(\vec{x}) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu})' \Sigma^{-1} (\vec{x} - \vec{\mu})\right].$$

- Any individual element of \vec{X} , say X_i , has a marginal normal density, with mean μ_i (i -th element of $\vec{\mu}$) and variance σ_i^2 (i -th diagonal element of Σ). We will establish a partial converse below.

Any subvector of \vec{X} is also multivariate normal, with mean and variance given by the corresponding subvector and submatrix of $\vec{\mu}$ and Σ , respectively.

- Consider any partition of \vec{X} into (\vec{Y}', \vec{Z}') . Partition Σ conformably as $\begin{bmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{bmatrix}$.

Then the conditional distribution of $\vec{Y}|\vec{Z}$ is also multivariate normal, with mean and variance

$$\begin{aligned} E(\vec{Y}|\vec{Z}) &= E\vec{Y} + \Sigma_{YZ}\Sigma_{ZZ}^{-1}(\vec{Z} - E\vec{Z}) \\ V(\vec{Y}|\vec{Z}) &= \Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}. \end{aligned}$$

3. Gamma distribution:² useful distribution which arises in deriving distribution theory for test statistics

¹This uses $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$, another of “Euler’s greatest hits”.

²For derivations of results stated here, see S. Wilks, *Mathematical Statistics*, §7.5.

- Density:

$$f(y) = \frac{y^{\mu-1}e^{-y}}{\Gamma(\mu)}, \quad y > 0, \mu > 0.$$

- $\Gamma(x)$ denotes the Gamma function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$.^{***}

*** The Gamma function has some interesting features.

- $\Gamma(x) = (x-1)\Gamma(x-1)$. From this, it follows that
- For integers n , $\Gamma(n) = (n-1)!$. Initial value: $\Gamma(1) = 1$.
Given this feature, you can think of Gamma function as an “extension” of the integer-valued map $x \mapsto (x-1)!$ to the non-negative reals.
- For non-integers $x > 1$, $\Gamma(x) = (x-1)(x-2)\cdots\delta\Gamma(\delta)$ where $0 < \delta < 1$
- For half-integer values (which arise frequently), the above recursion can be used, with the initial value $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.
We also have $\Gamma(n + \frac{1}{2}) = \sqrt{\pi} \frac{(2n)!}{2^{2n}n!}$.

- Moments:

$$E(x^r) = \frac{1}{\Gamma(\mu)} \int_0^\infty x^{\mu+r-1}e^{-x}dx = \Gamma(\mu+r)/\Gamma(\mu).$$

From this, you see $EX = \mu$ and $EX^2 = \mu^2 + \mu$, implying $VarX = \mu$.

- Characteristic function:

$$\phi(t) = (1 - it)^{-\mu}$$

- For the “quadratic form”

$$Q(\vec{x}) \equiv (\vec{x} - \vec{\mu})'\Sigma^{-1}(\vec{x} - \vec{\mu})$$

where \vec{x} is from the k -variate normal distribution, we have that

$$\frac{1}{2}Q(\vec{x}) \sim \text{Gamma}(k/2).$$

4. Chi-squared distribution:

- Let $x \sim \text{Gamma}(\mu)$. Then $z = 2x$ has the Chi-squared distribution with 2μ degrees of freedom.
- Let $k = 2\mu$ (where k is an integer). Then density:

$$f(z; k) = \frac{1}{2^{k/2}\Gamma(\frac{k}{2})} z^{\frac{k}{2}-1} e^{-\frac{z}{2}}, \quad z > 0.$$

- Properties of χ_k^2 stem from properties of Gamma distribution. In particular:

- $Q(\vec{x}) \sim 2 \cdot \text{Gamma}(\frac{k}{2}) = \chi_k^2$
- An important corollary: for X_i i.i.d. standard normal, the sum

$$\sum_i^k X_i^2 \sim \chi_k^2.$$

Aha! Didn't we say before that a vector of random variables with marginal Gaussian distributions need not be multivariate normal? Recall the Cramer-Wold device: a multivariate distribution is completely characterized by the distributions of the (scalar) linear combinations among its components. Obviously, all linear combinations among the components of a multivariate normal are distributed (univariate) normal. For an i.i.d. vector of univariate normal random variables, every linear combination is also univariate normal. Hence this vector is distributed multivariate normal. That is, *a vector of mutually independent normal random variables is jointly multivariate normal.*

- Characteristic function:

$$\phi_{\chi_k^2}(t) = [1 - 2it]^{-k/2}$$

- $E(\chi_k^2) = k, \quad \text{Var}(\chi_k^2) = 2k.$

■■■

■■■

Misc: Inequalities and Identities

(Source: CB, 4.7)

Functional inequalities

Definition: a function $g(x)$ is *convex* iff $g(\lambda x + (1 - \lambda)y) \leq \lambda g(x) + (1 - \lambda)g(y)$ for $0 < \lambda < 1$.

Graphically: a straight line connecting any two points of the convex function lies above the function.

Jensen's Inequality: For any random variable X , if $g(X)$ is convex, then $Eg(X) \geq g(EX)$.

Example: $X \sim U[0, 1]$. Then $EX = 1/2$.

For $g(X) = X^2$, $Eg(X) = 1/3$, and $g(EX) = g(1/2) = 1/4$.

■■■

Numerical inequalities

Hoelder's Inequality: X, Y are two random variables and choose $p, q > 1$ such that $1/p + 1/q = 1$. Then

$$|EXY| \leq E|XY| \leq (E|X|^p)^{1/p} (E|Y|^q)^{1/q}.$$

The *Cauchy-Schwartz inequality* is the case where $p = q = 2$:

$$|EXY| \leq E|XY| \leq (E|X|^2)^{1/2} (E|Y|^2)^{1/2}$$

or

$$[EXY]^2 \leq EX^2 EY^2.$$

Application: correlation $|\rho_{X,Y}| \leq 1$.

$$\begin{aligned} |E(X - \mu_X) \cdot (Y - \mu_Y)| &\leq \sqrt{E(X - \mu_X)^2} \cdot \sqrt{E(Y - \mu_Y)^2} \\ \Rightarrow [Cov(X, Y)]^2 &\leq \sigma_X^2 \sigma_Y^2 \\ \Rightarrow \rho_{X,Y}^2 &\leq 1 \Rightarrow -1 \leq \rho_{X,Y} \leq 1. \end{aligned}$$

For random variables X, Y such that $EX = EY = 0$, then the CS inequality can be expressed as

$$Cov(X, Y)^2 \leq EX \cdot EY.$$

■■■

Probability Inequalities

Markov/Chebyshev Inequality: Let X be a random variable and $g(X)$ a non-negative function. Then, for any $\epsilon > 0$ and $p > 0$:

$$P(g(X) \geq \epsilon) \leq \frac{E[g(X)]^p}{\epsilon^p}.$$

Proof: $\epsilon^p (P(g(X) \geq \epsilon)) = \epsilon^p \int_{g(x) \geq \epsilon} dF(x) \leq \int_{g(x) \geq \epsilon} g(x)^p dF(x) \leq \int g(x)^p dF(x) = E[g(X)]^p.$

The case of $p = 2$ is *Chebyshev's inequality*.

• **Example 1:**

$X \sim U[0, 1]$, $g(X) = X^2$, $Eg(X) = 1/3$, $Eg(X)^2 = 1/5$.

How often should $g(X)$ exceed $4/5$? (This should be a relatively rare event, because $Eg(X) = 1/3$.)

Chebyshev's inequality says: $P(X^2 \geq 4/5) \leq \frac{1/5}{16/25} = 5/16 \approx 0.31$ — an upper bound on the probability.

How close is this bound? In this case, we can directly calculate $P(X^2 \geq 4/5)$: Note that $P(X^2 \leq 4/5) = P(X \leq \sqrt{4/5}) = \sqrt{4/5} \approx 0.89$. Hence, $P(X^2 \geq 4/5) = 1 - P(X^2 \leq 4/5) \approx 0.11 \ll 0.31$.

In this case, Chebyshev's inequality gives a pretty conservative bound.

• **Example 2:**

Consider the random variables X_1, X_2, \dots, X_n , which are all drawn independently from the same distribution F_X , with $EX = \mu$ and $VX = \sigma_X^2$.

Consider the random variable $\bar{X} \equiv \frac{1}{n} \sum_i X_i$, the *sample mean* of these X 's. Note that $E\bar{X} = \mu$ and $V\bar{X} = \frac{1}{n}\sigma^2$.

How often should \bar{X} lie outside a neighborhood of the population mean μ ? By Chebyshev's inequality:

$$P(|\bar{X} - \mu| \geq \epsilon) \leq E(\bar{X} - \mu)^2 / \epsilon^2.$$

Now $E(\bar{X} - \mu)^2 / \epsilon^2 = V\bar{X} / \epsilon^2 = \sigma^2 / (n\epsilon^2)$, which tends to zero as $n \rightarrow \infty$. This implies that

$$\lim_{n \rightarrow \infty} P(|\bar{X} - \mu| \geq \epsilon) = 0,$$

which is the (*Weak*) *Law of Large Numbers*.



One useful limit: $\lim_{n \rightarrow \infty} (1 + \frac{k}{n})^n = \exp(k)$.

Define $Y_n \equiv (1 + \frac{k}{n})^n$. Then, using Taylor's theorem,

$$\begin{aligned} \log Y_n &= n \log \left(1 + \frac{k}{n} \right) \\ &\approx n \cdot \left[\log 1 + \frac{k}{n} + o\left(\frac{k}{n}\right) \right] = \left[0 + k + n \cdot o\left(\frac{k}{n}\right) \right] \\ &\rightarrow_{n \rightarrow \infty} k. \end{aligned}$$

Then, $Y_n \rightarrow \exp(k)$.

As an application, we have the following result:

Theorem: Consider a binomial random variable $X(n, p)$ where $p = \lambda/n$ for fixed $\lambda > 0$. Then as $n \rightarrow \infty$,

$$\phi_{X_n}(t) \rightarrow \exp(\lambda(\exp(it) - 1)).$$

Proof: By the above, we have

$$\phi_{X_n}(t) = [1 + \lambda/n \cdot (\exp(it) - 1)]^n \xrightarrow{n \rightarrow \infty} \exp(\lambda(\exp(it) - 1)).$$