

Prediction

Much of undergraduate Statistics is concerned with **estimation** of fixed (non-random) quantities such as:

μ = mean travel time of company's workers

p = proportion of American voters preferring Sarah Palin as next President

β_1 = coefficient of birthweight in a logistic regression

How about **random** quantities such as:

X = tomorrow's maximum temperature in Canberra

Y = US dollar exchange rate on 2nd May, 2011

Z = $\begin{cases} 1 & \text{if South Sydney win their next game} \\ 0 & \text{otherwise} \end{cases}$

The word **prediction** is usually used when the target is **random**.

Suppose X and Y are defined by

X = today's US dollar exchange rate

Y = tomorrow's US dollar exchange rate

and suppose we have a probabilistic model defined by the joint density function

$$f_{X,Y}(x, y) = [x, y].$$

Given that $X = 0.91$ what is a good prediction of Y ?

More formally, we want $g(\cdot)$ such that $g(X)$ 'close to' Y

In **estimation** theory $\hat{\theta}$ 'close to' θ

is often measured via mean squared error:

$$\text{MSE}(\hat{\theta}) = E\{(\hat{\theta} - \theta)^2\}.$$

In **prediction** theory we have

$$\text{MSE}\{g(X)\} = E[\{g(X) - Y\}^2].$$

i.e. choose $g(\cdot)$ to minimise

$$E[\{g(X) - Y\}^2] = \text{MSE}\{g(X)\}.$$

RESULT

The **best MSE predictor** of Y based on X is

$$g^*(X) = E(Y|X).$$

It is often called the **best predictor** (for short).

Proof to be given a little later.

Example

$$[x, y] = 21x^2y^3, \quad 0 < x < y < 1.$$

X is observed to be 0.254.

What is the best prediction of Y ?

Answer

$$\begin{aligned} [y|x] &= \frac{[x, y]}{[x]} \\ &= \frac{21x^2y^3}{\int_x^1 21x^2y^3 dy} \\ &= \frac{4y^3}{1 - x^4}, \quad 0 < x < y < 1. \end{aligned}$$

$$\begin{aligned} E(Y|X = x) &= \int_{-\infty}^{\infty} y[y|x] dy \\ &= \int_x^1 \frac{4y^4}{1 - x^4} dy \\ &= \frac{4(1 - x^5)}{5(1 - x^4)}, \quad 0 < x < 1 \end{aligned}$$

So the **best predictor** of Y is

$$g^*(X) = \frac{4(1 - X^5)}{5(1 - X^4)}.$$

If $X = 0.254$ we get

$$\frac{4(1 - 0.254^5)}{5(1 - 0.254^4)} = 0.802.$$

Our best prediction for Y is 0.802.

Proof of Best Prediction Result

We will do the proof for X and Y both continuous.

$$\begin{aligned} \text{MSE}\{g(X)\} &= E[\{Y - g(X)\}^2] \\ &= E(E[\{Y - g(X)\}^2|X]) \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \{y - g(x)\}^2 f_{Y|X}(y|x) dy \right) f_X(x) dx \\ &= \int_{-\infty}^{\infty} p(x) f_X(x) dx \end{aligned}$$

where $p(x) = \int_{-\infty}^{\infty} \{y - g(x)\}^2 f_{Y|X}(y|x) dy$.

Note that

$$p(x) \geq 0 \quad \text{for all } x \in \mathbb{R}.$$

$\text{MSE}\{g(X)\}$ is minimised by minimising $p(x)$ pointwise for $x \in \mathbb{R}$.

For example, if $x = 2.7$,

$$p(2.7) = \int_{-\infty}^{\infty} \{y - g(2.7)\}^2 f_{Y|X}(y|2.7) dy$$

Let $c = g(2.7)$. We need to choose c to minimise

$$p(2.7) = \int_{-\infty}^{\infty} (y - c)^2 f_{Y|X}(y|2.7) dy$$

$$g^*(x) = E(Y|X = x)$$

$$\begin{aligned} \frac{\partial p(2.7)}{\partial c} &= -2 \int_{-\infty}^{\infty} (y - c) f_{Y|X}(y|2.7) dy \\ &= 0 \end{aligned}$$

if and only if

$$\begin{aligned} c \int_{-\infty}^{\infty} f_{Y|X}(y|2.7) dy &= \int_{-\infty}^{\infty} y f_{Y|X}(y|2.7) dy \\ c \times 1 &= E(Y|X = 2.7) \\ c &= E(Y|X = 2.7) \\ g(2.7) &= E(Y|X = 2.7) \end{aligned}$$

Hence,

$$g^*(X) = E(Y|X).$$

◆ ◆ ◆

Repeating the same argument for general x we get