Gregory Kordas
Last update: April 5, 2023

LECTURE 4
JOINT, CONDITIONAL AND MARGINAL PROBABILITY DISTRIBUTIONS

Abstract: We define joint, conditional and marginal distributions and their moments, and discuss optimal prediction under square and absolute loss.

1. Introduction

Univariate distributions describe the randomness in a single random variable X. But in applications we are interested in the interdependence between random variables X and Y, say. If X is the *income* of a household and Y is its *savings rate* (the proportion of its income that it saves), we are interested in the *joint distribution* of (X, Y). Similarly, if X is the *price* of a product and Y the *quantity demanded* of that product, we are interested in their joint distribution.

Given a pair of random variables (X, Y) we will consider five different distributions:

(1) the *joint* distribution of X and Y, denoted by $f_{X,Y}(x, y)$;
(2) the *marginal* distribution of X, denote by $f_X(x)$;
(3) the *marginal* distribution of Y, denoted by $f_Y(y)$;
(4) the *conditional* distribution of X given Y, denote by $f_{X|Y}(x|y)$; and
(5) the *conditional* distribution of Y given X, denote by $f_{Y|X}(y|x)$.

If Y is the variable of interest (e.g. the quantity demanded of a product) and X is an explanatory variable (e.g. price), we will be mostly interested in $f_{Y|X}(y|x)$. Given the joint distribution $f_{X,Y}(x, y)$, we can compute the rest of the distributions by the appropriate averaging (integration), so we are justified in regarding the joint distribution as the "mother" distribution.

The analysis generalizes easily to a vector of random variables $(X_1, X_2, ..., X_n)$. In what follows, we will present the bivariate case first, and consider the general n-tuple case later.

2. Bivariate Distributions

2.1. The Discrete Case

Let (X, Y) be a pair of discrete random variables. We define their *joint pmf* $f(x, y)$ by

$$f_{X,Y}(x, y) = \Pr(X = x, Y = y),$$

where $\Pr(X = x, Y = y)$ is the joint probability of X taking the value x and Y taking the value y. Clearly, $f_{X,Y}(x, y) \geq 0$ and

$$\sum_x \sum_y f_{X,Y}(x, y) = 1,$$

where the summation is over all possible values of x and y.

Example 1. (*The Trinomial distribution*). Consider an experiment with 3 possible outcomes $\Omega = \{A, B, C\}$, and let p $(0 \leq p \leq 1)$ be the probability of outcome A, q $(0 \leq q \leq 1, p + q \leq 1)$ be the probability of outcome B, and $1 - p - q$ be the probability of outcome C. Let X be 1 if outcome A occurs and zero otherwise, and let $Y = 1$ if outcome B occurs and zero otherwise. Then the joint probability of x outcomes A and y outcomes B in n independent trials, is given by

$$f_{X,Y}(x, y) = \frac{n!}{x!y!(n - x - y)!} p^x q^y (1 - p - q)^{n-x-y},$$

for $x = 0, 1, ..., n$, and $y = 0, 1, ..., n - x$ (note the restriction on the support of Y). This is the $Trinomial(n, p, q)$ pmf and it is, of course, an extension of the $Binomial(n, p)$ distribution. That it is a proper pmf follows from the fact that $0 \leq f(x, y) \leq 1$, and that

$$\sum_{x=0}^{n} \sum_{y=0}^{n-x} \frac{n!}{x!y!(n - x - y)!} p^x q^y (1 - p - q)^{n-x-y} = 1.$$

■

2.2. The Continuous Case

Consider now a pair of continuous random variables (X, Y) and define their *joint pdf* as the function $f_{X,Y}(x, y)$ for which we can write

$$\Pr(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y) dy dx.$$

Clearly $f_{X,Y}(x, y) \geq 0$ and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = 1.$$

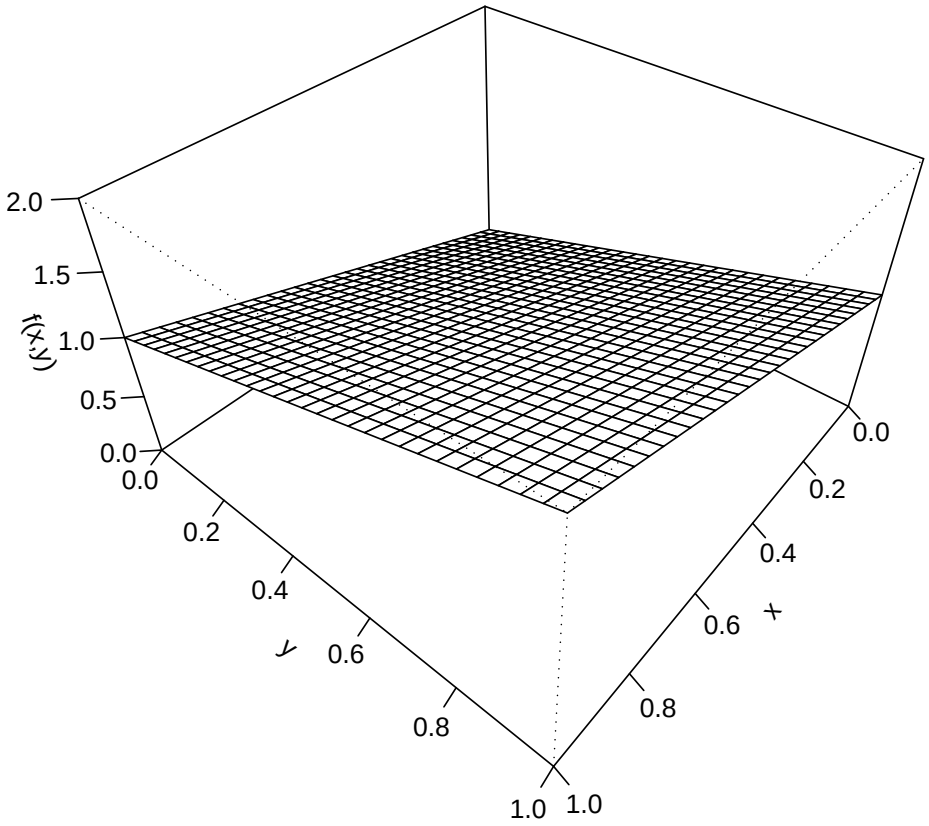


FIGURE 1. The Roof distribution

In what follows we will also be making use of the *joint cdf* $F_{X,Y}(x,y)$ defined by

$$F_{X,Y}(x,y) = \Pr(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(s,t)dtds.$$

Example 2. (*The Roof distribution*). Let (X,Y) have joint pdf given by

$$f_{X,Y}(x,y) = x + y, \qquad x \in [0,1], y \in [0,1].$$

Clearly $f_{X,Y}(x,y) \geq 0$ and

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = \int_{0}^{1}\int_{0}^{1}(x+y)dydx = 1.$$

■

3. Marginal Distributions

3.1. Discrete Case

Given the joint pmf of X and Y we can recover the *marginal* or *unconditional* pmf of X by summing $f_{X,Y}(x,y)$ over all possible values of Y, and the marginal pmf of Y by summing $f_{X,Y}(x,y)$ over all possible values of X. More explicitly, given the joint pmf $f_{X,Y}(x,y)$, the marginals of X, and Y are given by

$$f_X(x) = \sum_y f_{X,Y}(x,y), \quad \text{and} \quad f_Y(y) = \sum_x f_{X,Y}(x,y),$$

where again summation is performed over all possible values. Clearly, $f_X(x)$ and $f_Y(y)$ are proper pmf's, since they are both positive, and they sum to 1 since $f_{X,Y}(x,y)$ sums to 1.

Example 3.(*The Trinomial distribution*) Consider again the trinomial pmf and assume that we are interested in the marginal pmf of X. It can be shown that

$$\begin{aligned} f_X(x) &= \sum_{y=0}^{n-x} \frac{n!}{x!y!(n-x-y)!} p^x q^y (1-p-q)^{n-x-y} \\ &= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \end{aligned}$$

for $x = 0, ..., n$, which is, of course, a $Binomial(n,p)$ pmf. ■

3.2. Continuous Case

In the case where (X,Y) is a pair of continuous random variables, the marginal pdfs of X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy, \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx.$$

Example 4. (*The Roof Distribution*) Consider again the Roof distribution. The marginal pdf of X is given

$$f_X(x) = \int_{-\infty}^{\infty} (x+y)dy = x + \frac{1}{2}, \qquad x \in [0,1].$$

The marginal pdf of Y is similar. ■

4. Conditional Distributions

4.1. Discrete Case

Assume now that we are interested in the probability of *conditional events*, i.e., the probability that Y takes the value y, if X is known to be equal to x. If A is the event $Y = y$ and B is the event $X = x$, then the conditional probability of A given B is given by

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

Noting that, in the discrete case, $\Pr(A \cap B) = f_{X,Y}(x, y)$ and $\Pr(B) = f_Y(y)$ we can write,

$$\Pr(A|B) = \frac{f_{X,Y}(x, y)}{f_X(x)} \equiv f_{Y|X}(y|x),$$

where $f_{Y|X}(y|x)$ is the *conditional pmf* of Y given $X = x$. Analogously, the conditional pmf of X given $Y = y$ is

$$f_{X|Y}(x|y) \equiv \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Example 5. For the trinomial distribution, the conditional distribution of of $Y|X$ is

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{(n-x)!}{y!(n-x-y)!} \left(\frac{p_2}{1-p_1}\right)^y \left(\frac{p_3}{1-p_1}\right)^{n-x-y}, \quad y = 0, 1, ..., n-x.$$

The conditional distribution of $X|Y$ is analogous. ■

4.2. Continuous Case

The definition of conditional pdf's in the continuous case is exactly the same, but to show this we need to work a bit harder,

$$\begin{aligned}
f_{Y|X}(y|x) &= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \Pr(Y \leq y | x \leq X \leq x + \varepsilon) \\
&= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \frac{\Pr(Y \leq y,\ x \leq X \leq x + \varepsilon)}{\Pr(x < X \leq x + \varepsilon)} \\
&= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \frac{F_{X,Y}(x + \varepsilon, y) - F_{X,Y}(x, y)}{F_X(x + \varepsilon) - F_X(x)} \\
&= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \frac{[F_{X,Y}(x + \varepsilon, y) - F_{X,Y}(x, y)]/\varepsilon}{[F_X(x + \varepsilon) - F_X(x)]/\varepsilon} \\
&= \frac{\partial}{\partial y} \frac{\partial F_{X,Y}(x, y)/\partial x}{f_X(x)} \\
&= \frac{f_{X,Y}(x, y)}{f_X(x)},
\end{aligned}$$

provided, of course, that $f_X(x) \neq 0$. The definition of $f_{X|Y}(x|y)$ is completely analogous. That conditional pdf's are proper densities follows from the fact that $f_{Y|X}(y|x) \geq 0$, and

$$\int_{-\infty}^{\infty} f_{Y|X}(y|x)dy = \frac{1}{f_X(x)} \int_{-\infty}^{\infty} f_{Y,X}(y,x)dy = \frac{f_X(x)}{f_X(x)} = 1.$$

Example 6. For the Roof distribution,

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{x+y}{x+\frac{1}{2}}, \qquad x \in [0,1].$$

■

5. Joint, Marginal, and Conditional Moments

Given a pair of, say continuous, random variables X and Y, we have defined 5 kinds of distributions: the joint, two conditional, and two marginal distributions. The moments of these distributions are called *joint*, *conditional* and *marginal moments*, respectively.

Starting from the joint distribution of X and Y, we may define the (r,s)-th *joint raw moments* by

$$m_{X,Y}^{r,s} = E_{X,Y}(X^r Y^s) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^r y^s f_{X,Y}(x,y)dydx.$$

Similarly, the (r,s)-th *joint cental moments* are defined by

$$\mu_{X,Y}^{r,s} = E_{X,Y}[(X-\mu_X)^r (Y-\mu_Y)^s] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_X)^r (y-\mu_Y)^s f_{X,Y}(x,y)dydx.$$

A particularly important joint central moment is the *covariance* of X and Y, given by

$$C(X,Y) \equiv \sigma_{XY} = \mu_{X,Y}^{1,1}.$$

The *marginal moments* are special cases of the joint moments: setting $s = 0$ in the expressions above, we obtain the raw and central marginal moments of X, while setting $r = 0$ we obtain the raw and central marginal moments of Y. For example, $\mu_X = \mu_{X,Y}^{1,0}$, and $\mu_Y = \mu_{X,Y}^{0,1}$. An often used measure of (linear) association between X, Y is the correlation coefficient given by

$$\rho_{X,Y} = \frac{C(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}.$$

We can finally define the raw and central moments of the conditional distributions. The r-th *conditional raw moment* of $Y|X$ is given by

$$m_{Y|X}^r = E_{Y|X}(Y^r) = \int_{-\infty}^{\infty} y^r f_{Y|X}(y|x)dy,$$

while the r-th *conditional central moment* of $Y|X$ is given by

$$\mu^r_{Y|X} = E_{Y|X}[(Y - \mu_{Y|X})^r] = \int_{-\infty}^{\infty} (y - \mu_{Y|X})^r f_{Y|X}(y|x)dy,$$

where $\mu_{Y|X} = m^1_{Y|X}$ is the conditional mean of Y given $X = x$ (the conditional mean is the mean of the conditional distribution at $X = x$).

Theorem 1. (Law of Iterated Expectations). *Given a pair of random variables* (X, Y)*, and a functional* $Z = h(X, Y)$

(i) $E_Y(Y) = E_X[E_{Y|X}(Y|X)]$;
(ii) $E_{X,Y}(Z) = E_X[E_{Y|X}(Z|X)]$.

Proof: We only show (i), (ii) can be verified analogously.

$$\begin{aligned}
E_X[E_{Y|X}(Y|X)] &= \int_{-\infty}^{\infty} \mu_{Y|X} f_X(x)dx \\
&= \int_{-\infty}^{\infty} \Big[\int_{-\infty}^{\infty} y f_{Y|X}(y|X)dy\Big] f_X(x)dx \\
&= \int_{-\infty}^{\infty} \Big[\int_{-\infty}^{\infty} y \frac{f_{X,Y}(x,y)}{f_X(x)}dy\Big] f_X(x)dx \\
&= \int_{-\infty}^{\infty} y\Big[\int_{-\infty}^{\infty} f_{X,Y}(x,y)dx\Big]dy \\
&= \int_{-\infty}^{\infty} y f_Y(y)dy \\
&= E_Y(Y).
\end{aligned}$$

■

The *law of iterated expectations*, also known as the *law of total expectation*, or *the tower rule of expectations*, is often a source of confusion, mainly because of the sloppy way that it is often stated and proved. For example, we often see part (i) of the above theorem stated as

$$E(Y) = E(E(Y|X)).$$

The confusion comes from not stating explicitly with respect to which distribution the expectations are taken over. Recall there are 5 distributions here: (1) the joint distribution $f_{X,Y}$, (2) the distribution of X given Y, $f_{X|Y}$, (3) the distribution of Y given X, $f_{Y|X}$, (4) the marginal distribution of X, f_X, and (5) the marginal distribution of Y, f_Y.

Now let us see how to make sense of the last equation. The lhs, $E(Y)$, is the expectation of a random variable and thus a *number* (not a function). We should write the left hand side (lhs) as

$E_Y(Y)$ to remind us that we integrate y with respect to (wrt) the measure $dF_Y(y) = f_Y(y)\,dy$. The inner expectation on the right hand side (rhs) $E(Y|X)$ is taken wrt the conditional distribution of Y given X, $f_{Y|X}\,dy$, and as a result it is a (measurable) *function* of X (not a number), say $g(X)$. Thus instead of $E(Y|X) = g(X)$ we should have written $E_{Y|X}(Y|X) = g(X)$. You may object that up to now things were obvious and we shouldn't worry too much about this, but see what happens now. The last (outer) expectation on the rhs, namely $E(g)$, is where the confusion is often created. With regard to which distribution should we take this expectation? The answer is that, since g is a function of X, we take the expectation wrt to the marginal of X, $f_X\,dx$, i.e. we should write $E_X(g(X))$. This is also a number (not a function), which matches the lhs which is a number. Taking everything together, we have

$$E_Y(Y) = E_X(E_{Y|X}(Y|X)).$$

Compare this to the confusing statement above.

Now we can see the importance of the theorem too. Note what this "law" says: It says that

$$E_Y(Y) = E_X[g(X)] \quad \text{if and only if} \quad g(X) = E_{Y|X}(Y|X).$$

This is a very special property of conditional expectations that is NOT shared by other measures of location in general. For example, if M denotes the median, in general

$$M_Y(Y) \neq M_X[M_{Y|X}(Y|X)] \neq M_X[E_{Y|X}(Y|X)], ...$$

and all the other perturbations of E and M, unless *all* distributions involved are symmetric, in which case, their means and medians coincide.

We finally consider another theorem that is also of fundamental importance.

THEOREM 2. (Analysis of Variance). *Given a pair of random variables* (X, Y)*,*

$$\sigma_Y^2 = V_X(\mu_{Y|X}) + E_X(\sigma_{Y|X}^2).$$

Proof: Easy to verify directly. ∎

This theorem says that the total variation of Y, σ_Y^2, can be decomposed into the variation explained by X, $V_X(\mu_{Y|X})$, and a residual variation that cannot be explained by X, $E_X(\sigma_{Y|X}^2)$. We will discuss this is in great detail when we discuss the notion of the *regression* of Y on X.

6. Best Prediction Under Square Loss

As we have already seen the unconditional mean μ_Y is the *best predictor* of Y in terms of square loss, i.e.,

$$\mu_Y = \underset{c\in\mathbb{R}}{\operatorname{argmin}}\ E_Y(Y-c)^2 = \int_{-\infty}^{\infty}(y-c)^2 f_Y(y)dy.$$

It is not very difficult to imagine that the conditional mean $\mu_{Y|X}$ should have a similar property in terms of conditional square loss. Indeed, the conditional mean is the *best predictor* of $Y|X$ under square loss, i.e.,

$$\mu_{Y|X} = \underset{h:\mathbb{R}\mapsto\mathbb{R}}{\operatorname{argmin}}\ E_{Y|X}(Y-h(X))^2 = \int_{-\infty}^{\infty}(y-h(x))^2 f_{Y|X}(y|x)dy,$$

where the minimization is carried out over all real functions $h:\mathbb{R}\mapsto\mathbb{R}$. This is an important optimality property that justifies the preoccupation of econometricians with the estimation of conditional mean functions.

Example 7.(*The Trinomial Distribution*) For the trinomial distribution, the CEF of Y|X is given by

$$\begin{aligned}\mu_{Y|X} &= \sum_{y=0}^{n-x} y\frac{(n-x)!}{y!(n-x-y)!}\left(\frac{p_2}{1-p_1}\right)^y\left(\frac{p_3}{1-p_1}\right)^{n-x-y}\\ &= (n-x)\left(\frac{p_2}{1-p_1}\right),\end{aligned}$$

which is linear in x. ■

Example 8.(*The Roof Distribution*) For the Roof distribution,

$$\mu_{Y|X} = \int_0^1 y\frac{x+y}{x+\frac{1}{2}}dy = \frac{3x+2}{6x+3}, \qquad 0\le x\le 1.$$

■

Now let $\varepsilon = Y-\mu_{Y|X}$, and note that for this random variable, $E_{Y|X}(\varepsilon|X)=0$, and $V_{Y|X}(\varepsilon|X)=\sigma^2_{Y|X}$. It follows that we can decompose Y as

$$Y = \mu_{Y|X}+\varepsilon, \quad E(\varepsilon|X)=0.$$

This is called a *conditional expectation regression model*, and since Y can always be decomposed like that, the regression model is always meaningful. The "catch" is that, in order for it to be

well-specified, we need to know the *true* functional form of $\mu_{Y|X}$. For example, for the Roof distribution we have

$$Y = \frac{3X+2}{6X+3} + \varepsilon, \quad E(\varepsilon|X) = 0.$$

In applied work, however, the true conditional mean function is almost always unknown, and it is common to assume a simple linear functional form for it.

7. Best Linear Prediction Under Square Loss

Assume now that the true conditional mean function is unknown to us and that we are venturing to assume a linear form for it, i.e., we assume that $\mu_{Y|X} = \ell\mu_{Y|X} \equiv \alpha + \beta X$, where ℓ stands for "linear", and α, β are real constants. Under this assumption our regression model is given by

$$Y = \alpha + \beta X + \varepsilon, \quad E(\varepsilon|X) = 0.$$

This is a *linear conditional expectation regression model*, and $\ell\mu_{Y|X}$ is called a *linear (conditional) predictor*. The next theorem derives α and β that yield the *best linear predictor* (BLP) under square loss.

Theorem 3. *Given a pair of random variables X and Y,*

$$(\alpha^* = \mu_Y - \beta^*\mu_X, \quad \beta^* = \frac{\sigma_{XY}}{\sigma_X^2}) = \underset{a,b\in\mathbb{R}^2}{\operatorname{argmin}}\ E_{Y|X}(Y - a - bX)^2.$$

Proof: Let $\varepsilon = Y - (a + bX)$, so the objective can be written as $E_{Y|X}(\varepsilon^2)$. Differentiating and setting the derivatives equal to zero we obtain,

$$\begin{aligned}
\frac{\partial E_{Y|X}(\varepsilon^2)}{\partial a} = E_{Y|X}\Big(\frac{\partial \varepsilon^2}{\partial a}\Big) = 2E_{Y|X}\Big(\varepsilon\frac{\partial \varepsilon}{\partial a}\Big) = -2E_{Y|X}(\varepsilon) &= 0 \\
\frac{\partial E_{Y|X}(\varepsilon^2)}{\partial b} = E_{Y|X}\Big(\frac{\partial \varepsilon^2}{\partial b}\Big) = 2E_{Y|X}\Big(\varepsilon\frac{\partial \varepsilon}{\partial b}\Big) = -2E_{Y|X}(X\varepsilon) &= 0.
\end{aligned}$$

From these first order conditions we get $E_{Y|X}(\varepsilon) = 0$ and $E_{Y|X}(X\varepsilon) = 0$, which together are equivalent to $E_{Y|X}(\varepsilon) = 0$ and $C(X, \varepsilon) \equiv \sigma_{X\varepsilon} = 0$. Substituting for ε we have

$$\sigma_{X\varepsilon} = 0 \Rightarrow \sigma_{X(Y-\alpha^*-\beta^*X)} = 0 \Rightarrow \sigma_{XY} = \beta^*\sigma_X^2 \Rightarrow \beta^* = \sigma_{XY}/\sigma_X^2,$$

and

$$\mu_\varepsilon = 0 \Rightarrow \mu_{Y-\alpha^*-\beta^*X} = 0 \Rightarrow \mu_Y = \alpha^* + \beta^*\mu_X \Rightarrow \alpha^* = \mu_Y - \beta^*\mu_X.$$

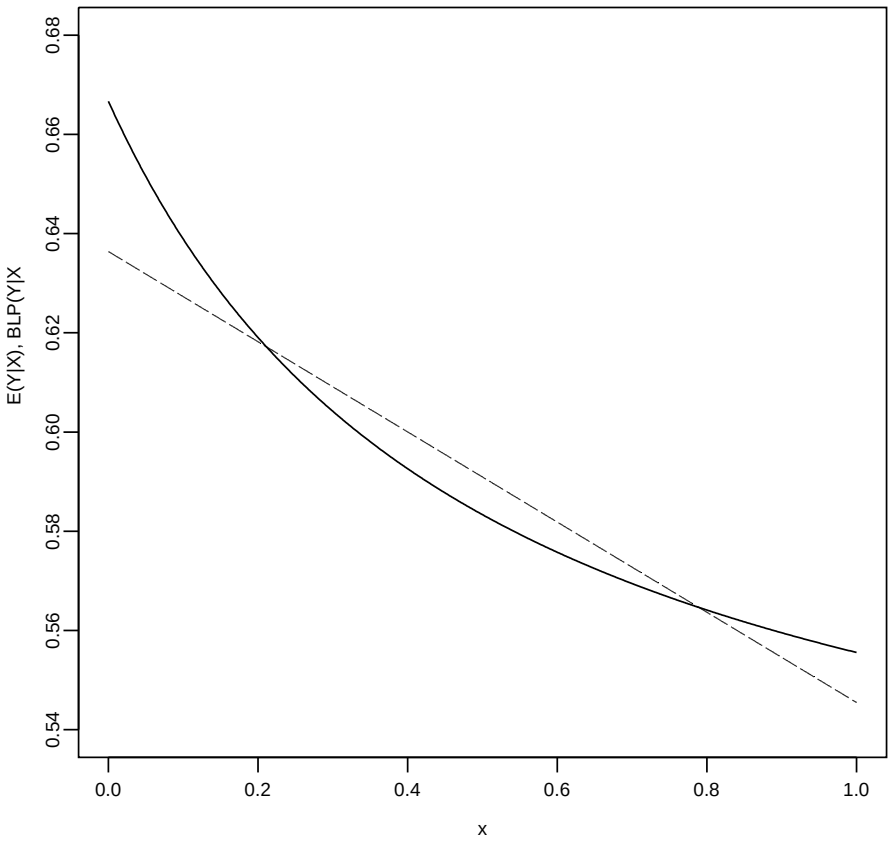
■

The optimal intercept and slope coefficients are seen to be the familiar Least Squares (LS) solutions. If the true CEF is linear in X, the BLP and the CEF coincide. But if the CEF is

nonlinear in X, then the BLP is the *best linear approximation to the CEF on the support of* X. In fact we can quantify the accuracy of the approximation by computing the *relative efficiency* ratio

$$RE = \frac{E_{Y,X}\left[\left(Y - \ell\mu_{Y|X}\right)^2\right]}{E_{Y,X}\left[\left(Y - \mu_{Y|X}\right)^2\right]} \geq 1,$$

that measures the *relative variance of the residuals* of the BLP and the CEF. Since the expectations are taken w.r.t. to the joint distribution of X and Y, the denomenator is simply the unconditional variance of Y, σ_Y^2.



Example 9. The CEF for the Roof distribution is nonlinear in X, so the CEF and the BLP are not the same. Using Theorem 3, we find that the BLP is given by

$$\ell\mu_{Y|X} = \frac{7}{11} - \frac{1}{11}X = \frac{7 - X}{11} \qquad 0 \leq X \leq 1,$$

which is indeed different from the CEF

$$\mu_{Y|X} = \frac{2+3X}{3+6X}, \qquad 0 \leq X \leq 1.$$

Plotting the two functions, however, we see that $\ell\mu_{Y|X}$ is quite informative about $\mu_{Y|X}$, as it is the best linear approximation to it on the support of X. ■

8. Best Prediction Under Absolute Loss

The mean is not the only interesting measure of location (or cental tendency). Other measures commonly used are the *median* and the rest of the *quantiles*. Like the mean, these measures are also "optimal" under appropriate (and reasonable) loss functions.

The *median* $Q_Y(.5)$ of a (continuous) random variable Y is the value for which

$$\Pr[Y \leq Q_Y(.5)] \equiv F_Y(Q_Y(.5)) = 0.5,$$

or

$$Q_Y(.5) = F^{-1}(0.5).$$

Other quantiles are defined similarly, with the τ-th quantile $Q_Y(\tau)$ of Y defined by

$$\Pr[Y \leq Q_Y(\tau)] \equiv F_Y(Q_Y(\tau)) = \tau, \quad \tau \in (0,1),$$

or,

$$Q_Y(\tau) = F^{-1}(\tau), \quad \tau \in (0,1).$$

The function $Q_Y(\tau)$ yields the quantiles of Y, and is called the *quantile function* of Y. Like the mean, the median is also a best predictor: it is best under *absolute loss*.

Theorem 4. *The median $Q_Y(.5)$ is the best predictor of Y under absolute loss, i.e.,*

$$Q_Y(.5) = \underset{c\in\mathbb{R}}{\operatorname{argmin}}\ E_Y|Y-c| = \int_{-\infty}^{\infty} |y-c| f_Y(y)dy.$$

Similarly, the conditional median $Q_{Y|X}(.5)$ is the best predictor of $Y|X$ under absolute loss, i.e.,

$$Q_{Y|X}(.5) = \underset{h:\mathbb{R}\mapsto\mathbb{R}}{\operatorname{argmin}}\ E_{Y|X}|Y-h(X)| = \int_{-\infty}^{\infty} |y-h(x)| f_{Y|X}(y)dy.$$

Proof: Fix x and let $F_{Y|X}$ be the conditional d.f. of $Y|X$. By definition, the median of $F_{Y|X}$ is the real number

$$m = \min c : F_{Y|X}(-\infty, c] \geq 1/2$$

where $F_{Y|X}(-\infty, c]$ is the probability that Y is in the interval $(-\infty, c]$.

To prove that the median is a best predictor under absolute loss, compare the expected loss at m with that at any $c < m$. We find

$$\begin{aligned}
&\int |y-c| dF_{Y|X} - \int |y-m| dF_{Y|X} = \int [|y-c| - |y-m|] dF_{Y|X} \\
= &\int_{(-\infty,c]} (c-m) dF_{Y|X} + \int_{(c,m)} [2y-(c+m)] dF_{Y|X} + \int_{[m,\infty)} (m-c) dF_{Y|X} \\
\geq &(c-m)F_{Y|X}(-\infty,c] + (c-m)F_{Y|X}(c,m) + (m-c)F_{Y|X}[m,\infty) \\
= &-(m-c)F_{Y|X}(-\infty,m) + (m-c)F_{Y|X}[m,\infty) \\
= &(m-c)\{F_{Y|X}[m,\infty) - F_{Y|X}(-\infty,m)\}.
\end{aligned}$$

By the definition of the median, $F_{Y|X}(-\infty,c] < 1/2$ for all $c < m$. Hence, $F_{Y|X}(-\infty,m) \leq 1/2$, so

$$(m-c)\{F_{Y|X}[m,\infty) - F_{Y|X}(-\infty,m)\} \geq 0.$$

Now compare the expected loss at m with that at any $c > m$.

$$\begin{aligned}
&\int |y-c| dF_{Y|X} - \int |y-m| dF_{Y|X} = \int [|y-c| - |y-m|] dF_{Y|X} \\
= &\int_{(-\infty,m]} (c-m) dF_{Y|X} + \int_{(m,c)} [(c+m)-2y] dF_{Y|X} + \int_{[c,\infty)} (m-c) dF_{Y|X} \\
\geq &(c-m)F_{Y|X}(-\infty,m] + (m-c)F_{Y|X}(m,c) + (m-c)F_{Y|X}[c,\infty) \\
= &(c-m)F_{Y|X}(-\infty,m) + (c-m)F_{Y|X}[m,\infty) \\
= &(c-m)\{F_{Y|X}(-\infty,m) - F_{Y|X}[m,\infty)\}.
\end{aligned}$$

By the definition of the median, $F_{Y|X}(c,\infty) \leq 1/2$ for all $c > m$. Hence, $F_{Y|X}(m,\infty) \leq 1/2$, so

$$(c-m)\{F_{Y|X}(-\infty,m) - F_{Y|X}[m,\infty)\} \geq 0.$$

Pulling the two results for $c > m$ and $c < m$ together, we conclude that the expected loss is minimized at the median. ■

The rest of the quantiles may also be shown to be optimal predictors under *asymmetric absolute loss*. The idea here is that if over-prediction is more costly than under-prediction, a quantile below the median would be optimal, while in the opposite scenario, the optimal quantile would be above the median. To formalize this thought, define the *asymmetric absolute loss function*

$$\rho_\tau(u) = (\tau - I\{u < 0\})u.$$

This function generalizes the absolute loss function $|u|$, which, ignoring the multiplicative constant, is a special case for $\tau = .5$,

$$\rho_{.5}(u) = (.5 - I\{u < 0\})u = \frac{1}{2}|u|.$$

The following theorem generalizes Theorem 4, and can be proven in a similar way.

THEOREM 5. *The τ-th quantile $Q_Y(\tau)$ is the best predictor of Y under asymmetric absolute loss, i.e.,*

$$Q_Y(\tau) = \underset{c\in\mathbb{R}}{\operatorname{argmin}}\ E_Y \rho_\tau(Y - c) = \int_{-\infty}^{\infty} \rho_\tau(y - c) f_Y(y) dy.$$

Similarly, the τ-th conditional quantile $Q_{Y|X}(\tau)$ is the best predictor of $Y|X$ under asymmetric absolute loss, i.e.,

$$Q_{Y|X}(\tau) = \underset{h:\mathbb{R}\mapsto\mathbb{R}}{\operatorname{argmin}}\ E_{Y|X}\, \rho_\tau(Y - h(X)) = \int_{-\infty}^{\infty} \rho_\tau(y - h(x)) f_{Y|X}(y) dy.$$

Now let $\varepsilon = Y - Q_{Y|X}(\tau)$ be the deviations of Y from its τ-th conditional quantile, and observe that the τ-th conditional quantile of this variable is zero, i.e., $Q_{\varepsilon|X}(\tau) = 0$. Given $\tau \in (0, 1)$, we may now decompose Y as

$$Y = Q_{Y|X}(\tau) + \varepsilon, \quad Q_{\varepsilon|X}(\tau) = 0.$$

This is called a *quantile regression model*. The *median regression model* is a very important special case, which, in a slightly simplified notation, may be written as,

$$Y = \text{Med}(Y|X) + \varepsilon, \quad \text{Med}(\varepsilon|X) = 0.$$

As with mean regression, the true functional form of $Q_{Y|X}(\tau)$ is often unknown in applications. If we assume the linear form $Q_{Y|X}(\tau) = \ell Q_{Y|X}(\tau) \equiv \alpha(\tau) + \beta(\tau)X$, we obtain a *linear quantile regression model* given by

$$Y = \alpha(\tau) + \beta(\tau)X + \varepsilon, \quad Q_{\varepsilon|X}(\tau) = 0.$$

It is interesting to note that the regression parameters $\alpha(\tau)$ and $\beta(\tau)$ are free to vary across τ. This means that X is free to affect different conditional quantiles of Y in different ways, providing a very rich model that yields many interesting empirical results. We will return to this at a later point where we will discuss quantile regression in more detail.

9. Mean vs. Median Regressions

The mean and median regressions of Y on X both express the "central tendency" of Y on X. It is important, however, to recognize that these two regression functions generally do not coincide. For example, one could be a linear function of X and the other not.

Example 10. Let X be a Bernoulli (dummy) random variable. The mean and median regressions of Y on X can be written as the linear functions

$$E(Y|X) = E(Y|X=0) + [E(Y|X=1) - E(Y|X=0)]X$$

and

$$\text{Med}(Y|X) = \text{Med}(Y|X=0) + [\text{Med}(Y|X=1) - \text{Med}(Y|X=0)]X.$$

The slope parameters of the two regression functions are $[E(Y|X=1) - E(Y|X=0)]$ and $[\text{Med}(Y|X=1) - \text{Med}(Y|X=0)]$, respectively. It is clear that these differences of means and medians need not be the same, or even have the same sign. ■

It follows that prediction under square and absolute loss can yield very different conclusions about the "central tendency" of Y as a function of X. There is, however, a special situation in which the two functions coincide. This happens when $F_{Y|x}$ is symmetric *for all* $x \in \text{support}(X)$, i.e., when the family of conditional cf's $\{F_{Y|x}, x \in \text{support}(X)\}$ are symmetric. This is a very restrictive assumption, but if it happens to be true, the mean and the median regressions do coincide. In terms of our example above, this would happen if $F_{Y|0}$ and $F_{Y|1}$ are both symmetric, for example normal.

10. Independence

A pair of random variables X and Y are *stochastically independent* if and only if their joint density is equal to the product of their marginals, i.e.

$$f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

When this happens

$$f_{Y|X} = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_X(x)f_Y(y)}{f_X(x)} = f_Y(y),$$

and similarly, $f_{X|Y}(x|y) = f_X(x)$. It is clear that stochastic independence is a symmetric relation. However, it is not transitive, i.e., it is possible for a set of r.v.'s to be pairwise independent but dependend if taken all together.

Example 11.(Berstein). Let X_1, X_2, X_3 have the joint p.d.f.

$$\begin{aligned} f_{123}(x_1, x_2, x_3) &= \tfrac{1}{4}, \quad (x_1, x_2, x_3) \in \{(1,0,0), (0,1,0), (0,0,1), (1,1,1)\}, \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

The joint p.d.f. of X_i and X_j, $i \neq j$, is

$$\begin{aligned} f_{ij}(x_1, x_2, x_3) &= \tfrac{1}{4}, \quad (x_i, x_j) \in \{(0,0), (1,0), (0,1), (1,1)\}, \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

whereas the marginal p.d.f. of X_i is

■

The random variable Y is said to be *mean independent* of X if

$$\mu_{Y|X} = \mu_Y.$$

Unlike stochastic independence, mean independence is not a symmetric relation: if Y is mean independent of X, then X may or may not be mean independent of Y. It is easy to see that stochastic independence implies mean independence, since under stochastic independence,

$$\mu_{Y|X} = \int y f_{Y|X}(y|x) dy = \int y f_Y(y) dy = \mu_Y.$$

The reverse, however, is true only in special cases, so stochastic independence is a stronger condition than mean independence. Furthermore, mean independence implies zero correlation, but, again, the reverse is not always true. To see this, note that under mean independence,

$$C(X, Y) = C[X, \mu_{Y|X}] = C[X, \mu_Y] = 0,$$

since μ_Y is a constant (show that $C(X, Y) = C[X, \mu_{Y|X}]$ as an exercise). Uncorrelatedness, however, does not imply mean independence since $C[X, \mu_{Y|X}] = 0$ can happen even if $\mu_{Y|X} \neq \mu_Y$. Summarizing,

$$\textsf{stochastic independence} \Rightarrow \textsf{mean independence} \Rightarrow \textsf{uncorrelatedness},$$

but the reverse implications may be false.

In a similar fashion, two random variables are said to be τ-quantile independent if

$$Q_{Y|X}(\tau) = Q_Y(\tau), \quad \text{and} \quad Q_{X|Y}(\tau) = Q_X(\tau).$$

Again, stochastic independence implies quantile independence

$$Q_{Y|X}(\tau) \equiv F^{-1}_{Y|X}(\tau) = F^{-1}_Y(\tau) \equiv Q_Y(\tau),$$

but the reverse is not always true. Note that when X and Y are stochastically independent, then they are also τ-quantile independent for all $\tau \in (0, 1)$, but it is possible that, for two quantiles τ_1 and τ_2, X and Y are τ_1-quantile independent but τ_2-quantile dependent.

> When a 12th century youth fell in love he did not take three paces backward, gaze in to her eyes, and tell her she was too beautiful to live. He said he would step outside and see about it. And if, when he got out, he met a man and broke his head – the others man's head, I mean – then that proved that his – the first fellow's – girl was a pretty girl. But if the other fellow broke his head – not his own, you know, but the other fellow's – the other fellow to the second fellow, that is, because of course the other fellow would only be the other fellow to him, not the first fellow who – well, if he broke his head, then his girl – not the other fellow's, but the fellow who was the – – Look here, if x broke y's head, then x's girl was a pretty girl, but if y broke x's head, then x's girl wasn't a pretty girl, but y's girl was.
>
> — *Jerome K. Jerome*
>
> *Idle Thoughts of an Idle Man, 1889, pp.58-59*