1 Statistical Models and Conditional Expectation

"essentially, all models are wrong, but some are useful"

George E.P. Box

Mathematical model

A **mathematical model** is a description of a system using mathematical concepts and language.

- Bacteria growth $\frac{dy}{dt} = ky$, ..., $y = \underbrace{A_0 e^{kt}}_{\text{Not random}}$
- $Y = \underbrace{f(x)}_{\text{Not random}} + \varepsilon$ random

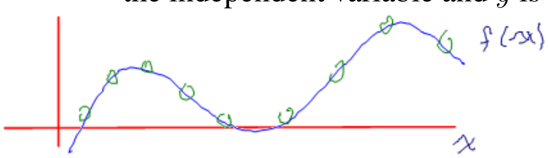
Statistical model

A **statistical model** embodies a set of assumptions concerning the generation of the observed data, and similar data from a larger population.

Simulation

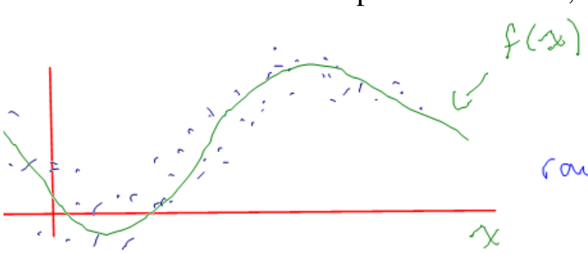
Relations between variables

A **functional relation** between two variables is expressed by a mathematical formula. If x is the independent variable and y is the dependent variable, then a function relation is of the form:



$$y = f(x).$$

A **statistical relation**, unlike a functional relation, is not a perfect one. In general, the observations for a statistical relation do not fall directly on the curve of relationship. This is commonly expressed as a functional relation coupled with a random error ϵ. If x is the independent variable and Y is the dependent variable, then a statistical relation *often* takes the form:



$$Y = f(x) + \epsilon.$$

random   Not random   random

A statistical relation is also commonly expressed in terms of **conditional expectation**. That is, for random variables Y and X,

$$Y = \underbrace{E[Y|X = x]}_{f(x)} + \epsilon.$$

1.1 Conditional Expectation

Goal: The goal of this section is to motivate conditional expectation:

$$E[Y|X = x]$$

Consider an example from probability theory.

Example 1

Consider rolling two fair six sided dice (D_1 & D_2) and recording the sum of the faces and the maximum of the faces. Define two random variables $Y = D_1 + D_2$ and $X = \max\{D_1, D_2\}$. The joint probability distribution $P(X = x, Y = y)$ of these two random variables is:

1 and

$X \diagdown Y$	2	3	4	5	6	7	8	9	10	11	12	$P(X = x)$
1	1/36	0	0	0	0	0	0	0	0	0	0	1/36
2	0	2/36	1/36	0	0	0	0	0	0	0	0	3/36
3	0	0	2/36	2/36	1/36	0	0	0	0	0	0	5/36
4	0	0	0	2/36	2/36	2/36	1/36	0	0	0	0	7/36
5	0	0	0	0	2/36	2/36	2/36	2/36	1/36	0	0	9/36
6	0	0	0	0	0	2/36	2/36	2/36	2/36	2/36	1/36	11/36
$P(Y = y)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	1

DEFINITION 1.1 The **conditional probability mass function** of $Y|X = x$ is defined by

$$p(y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)},$$

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

where $P(X = x, Y = y)$ is the joint distribution of X and Y and $P(X = x)$ is the marginal distribution of X. Note: $P(X = x) > 0$ for all x.

Example 1 continued

$$P(8|X = 4) = \frac{1}{36} \div \frac{7}{36} = \frac{1}{7}$$

$$P(8|X = 5) = \frac{2}{9}$$

DEFINITION 1.2 The conditional expectation of $Y|X = x$ is defined by

$$E[Y|X = x] = \sum_y y * p(y|X = x).$$

Note: We can also define conditional variance Var$[Y|X = x]$ analogously.

Note:

Note:

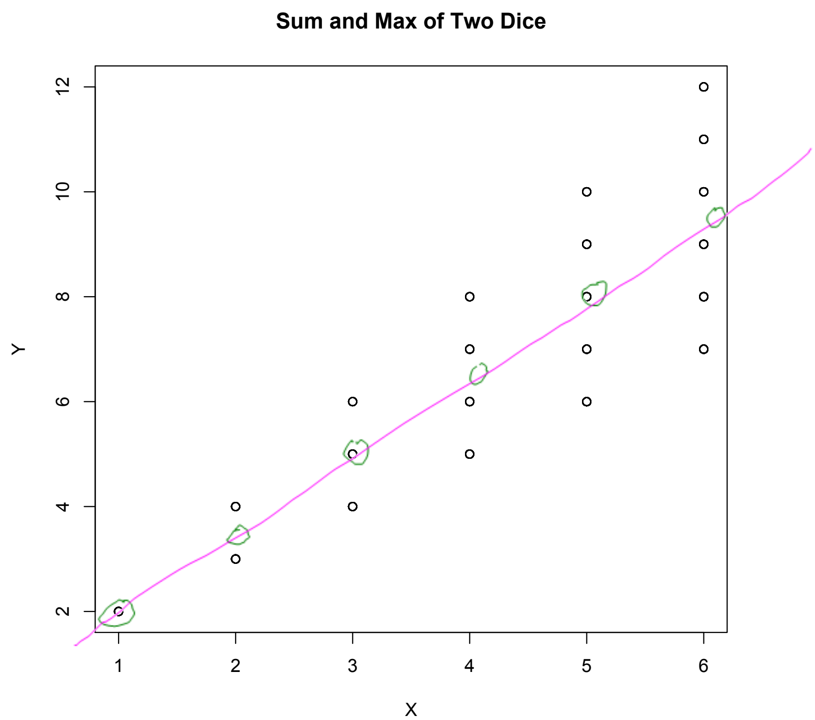
Example 1 continued

Find the conditional expectation of $Y|X = x$. Note: There will be six values corresponding to $X = 1, 2, 3, 4, 5, 6$.

$$E[Y|X = 1] = 2 * \frac{1/36}{1/36} + 3 * \frac{0}{1/36} + 4 * \frac{0}{1/36} + \cdots + 12 * \frac{0}{1/36} = 2$$
$$E[Y|X = 2] = 2 * \frac{0}{3/36} + 3 * \frac{2/36}{3/36} + 4 * \frac{1/36}{3/36} + \cdots + 12 * \frac{0}{3/36} = \frac{10}{3}$$
$$E[Y|X = 3] = 2 * \frac{0}{5/36} + 3 * \frac{0}{5/36} + 4 * \frac{2/36}{5/36} + \cdots + 12 * \frac{0}{5/36} = \frac{24}{5}$$

$$E[Y|X=4] = \ldots = \frac{44}{7}$$

x	1	2	3	4	5	6
$E[Y\|X = x]$	2	10/3	24/5	44/7	70/9	102/11



Criticize this motivating example:

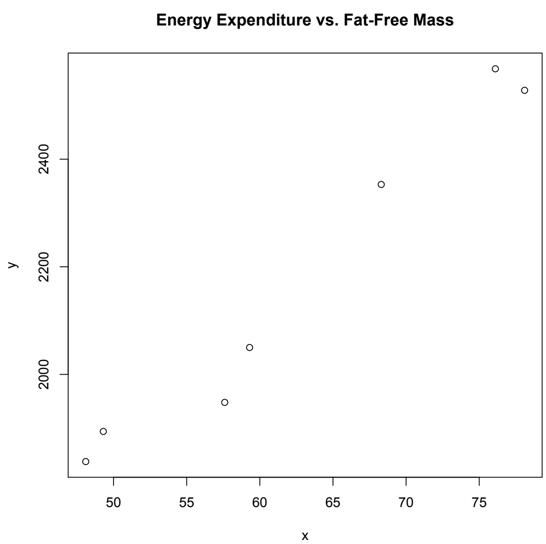
- The "real world" does not behave like rolling dice
- In the above example, we know $E[Y \mid X = x]$. We do not need to estimate the mean.
- Regression analysis generally assumes cont. Y.

Example 2

Realistic Example:

To investigate the dependence of energy expenditure on body build, researches used underwater weighing techniques to determine the fat-free body mass for each of seven men. They also measured the total 24-hour energy expenditure for each man during conditions of quiet sedentary activity. The results are shown in the table.

Subject	1	2	3	4	5	6	7
x	49.3	59.3	68.3	48.1	57.61	78.1	76.1
y	1,894	2,050	2,353	1,838	1,948	2,528	2,568



Questions:

i. How do we compute the conditional expectation $E[Y|X = x]$?

ii. What type of functional form will $E[Y|X = x]$ take on?

iii. Are the variables X and Y discrete or continuous?

iv. What probability distributions govern the behavior of X and Y?

v. Should X be thought of as fixed? (non-random)

vi. Is our model correct? How off are we?

i) To estimate $E[Y|X=x]$, we need to choose a model.

ii) Some increasing function. Maybe linear?

iii) Both X and Y are continuous.

iv) should we assume normality? Can't have negative numbers.

v) In this example, X is random.

Note: In a clinical trial, the experimenter has control over X (dosage), X is fixed

vi) A Model is never correct.

Computing $E[Y|X=x]$

- Suppose X and Y are jointly normally distributed.
- We use the bivariate normal distribution to model this relationship.

Continuous random variables

DEFINITION 1.3 Let X and Y be two continuous random variables. The **conditional probability density function** of $Y|X = x$ is defined by

$$f(y|X = x) = \frac{f(x, y)}{f_X(x)},$$

where $f(x, y)$ is the joint density of X and Y and $f_X(x)$ is the marginal density of X. Note: $f_X(x) > 0$ for all x.

DEFINITION 1.4 Let X and Y be two continuous random variables and let $f(y|X = x)$ be the conditional density function of $Y|X = x$. The conditional expectation of $Y|X = x$ is defined by

$$E[Y|X = x] = \int y * f(y|X = x)dy.$$

Note: we can also define $Var(Y|X=x)$

1.2 Bivariate Normal Distribution

DEFINITION 1.5 The **bivariate normal distribution** of random vector (X, Y) has probability density function defined by

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_1\sigma_2} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\},$$

with

$$-\infty < x < \infty \qquad -\infty < y < \infty.$$

Note: $\iint f(x,y)\,dx\,dy = 1$

Note: The Bivariate Normal distribution is a special case of a multivariate normal.

PROPOSITION 1.1 *If (X, Y) is a random vector from the bivariate normal distribution, then the conditional expectation and variance of Y given $X = x$ are*

$$E[Y|X = x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) = \beta_0 + \beta_1 x \tag{1.1}$$

linear

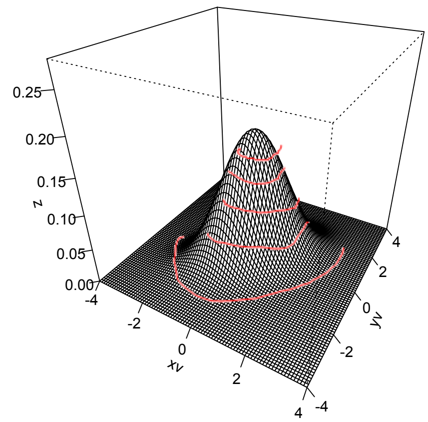
and

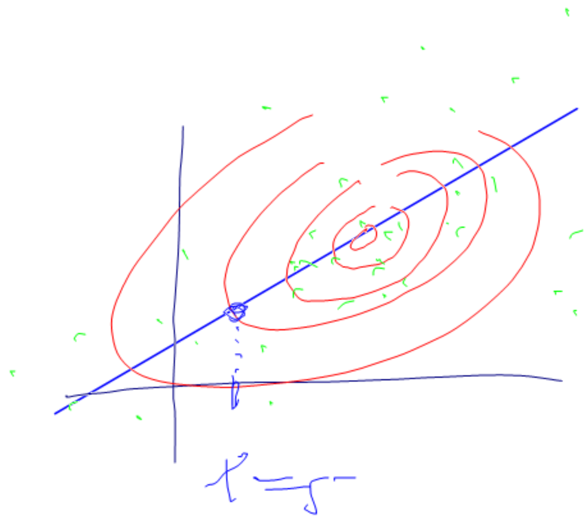
$$\left(\mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X\right) + \left(\rho\frac{\sigma_Y}{\sigma_X}\right)\cdot x$$

$$Var[Y|X = x] = \underbrace{\sigma_Y^2(1 - \rho^2)}_{\text{constant}}. \tag{1.2}$$

$\mu_X \quad \mu_Y \quad \sigma_X^2 \quad \sigma_Y^2$

Figure 1: Bivariate Normal with parameters $\mu_1 = 0,\ \mu_2 = 0,\ \sigma_1^2 = 1,\ \sigma_2^2 = 1,\ \rho = .25$





Summary:

1. $E[Y|X = x]$ is a linear function (assuming X, Y are bivariate normal).

2. The goal is to estimate the parameters β_0 and β_1.

3. By method of guess

$$\hat{\beta}_1 = \hat{\rho}\,\frac{\hat{\sigma}_Y}{\hat{\sigma}_X} = r \cdot \frac{s_Y}{s_X} \qquad \hat{\beta}_0 = \hat{\mu}_Y - \hat{\beta}_1 \hat{\mu}_X = \bar{Y} - \hat{\beta}_1 \bar{X}$$