Regression Intro

Note 20

Estimation: In estimation, we have an unknown random variable Y that we want to estimate. Y may also depend on another random variable X that we know. In the simplest case, we don't incorporate any information about X when creating our estimate $\hat{Y}$ and just estimate Y with a constant. Our choice of constant will minimize the **mean squared error**, $\mathbb{E}[(Y-\hat{Y})^2]$. This minimum occurs at

$$\hat{Y} = \mathbb{E}[Y].$$

If we want to incorporate X into our estimate, we can model $Y = g(X)$ and try to find the best $\hat{Y}$ such that the mean squared error $\mathbb{E}[(Y-\hat{Y})^2 \mid X]$ is again minimized. This occurs at

$$\hat{Y} = \mathbb{E}[Y \mid X].$$

We call this the **minimum mean squared estimate** (MMSE) of Y given X.

Since finding the conditional expectation is often very difficult, we compromise by estimating with a *linear function*: $\hat{Y} = aX + b$. Here, we want to minimize $\mathbb{E}[(Y-aX-b)^2 \mid X]$, which has a minimum at

$$\hat{Y} = \mathbb{E}[Y] + \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}(X - \mathbb{E}[X]) :- \text{LLSE}[Y \mid X].$$

This is known as the **linear least squares estimate** (LLSE) of Y given X.

1 LLSE

Note 20

We have two bags of balls. The fractions of red balls and blue balls in bag A are $2/3$ and $1/3$ respectively. The fractions of red balls and blue balls in bag B are $1/2$ and $1/2$ respectively. Someone gives you one of the bags (unmarked) uniformly at random. You then draw 6 balls from that same bag with replacement. Let X_i be the indicator random variable that ball i is red. Now, let us define $X = \sum_{1 \le i \le 3} X_i$ and $Y = \sum_{4 \le i \le 6} X_i$.

(a) Compute $\mathbb{E}[X]$ and $\mathbb{E}[Y]$.

(b) Compute $\operatorname{Var}(X)$.

(c) Compute $\operatorname{cov}(X,Y)$. (*Hint*: Recall that covariance is bilinear.)

(d) Now, we are going to try and predict Y from a value of X. Compute $L(Y \mid X)$, the best linear estimator of Y given X. Recall that

$$L(Y \mid X) = \mathbb{E}[Y] + \frac{\operatorname{cov}(X,Y)}{\operatorname{Var}(X)}\big(X - \mathbb{E}[X]\big).$$

Solution: Although the indicator random variables are not independent, we can still apply linearity of expectation. By symmetry, we also know that each indicator follows the same distribution.

(a)

$$\mathbb{E}[X] = \mathbb{E}[Y] = 3 \cdot \mathbb{E}[X_1] = 3 \cdot \mathbb{P}[X_1 = 1] = 3 \cdot \left(\frac{1}{2} \cdot \frac{2}{3} + \frac{1}{2} \cdot \frac{1}{2}\right) = \frac{7}{4}.$$

(b)

$$\begin{aligned}
\operatorname{Var}(X) &= \operatorname{cov}\left(\sum_{1 \le i \le 3} X_i, \sum_{1 \le j \le 3} X_j\right) \\
&= 3 \cdot \operatorname{Var}(X_1) + 6 \cdot \operatorname{cov}(X_1, X_2) \\
&= 3\big(\mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2\big) + 6 \cdot \frac{1}{144} \\
&= 3\left[\frac{7}{12} - \left(\frac{7}{12}\right)^2\right] + 6 \cdot \frac{1}{144} = \frac{111}{144}.
\end{aligned}$$

(c)

$$\begin{aligned}
\operatorname{cov}(X,Y) &= \operatorname{cov}\left(\sum_{1 \le i \le 3} X_i, \sum_{4 \le j \le 6} X_j\right) \\
&= 9 \cdot \operatorname{cov}(X_1, X_4) \\
&= 9 \cdot \big(\mathbb{E}[X_1 X_4] - \mathbb{E}[X_1] \cdot \mathbb{E}[X_4]\big) \\
&= 9 \cdot \Big(\mathbb{P}[X_1 = 1, X_4 = 1] - \mathbb{P}[X_1 = 1]^2\Big) \\
&= 9 \cdot \left(\left[\frac{1}{2} \cdot \left(\frac{2}{3}\right)^2 + \frac{1}{2} \cdot \left(\frac{1}{2}\right)^2\right] - \left[\frac{1}{2} \cdot \left(\frac{2}{3}\right) + \frac{1}{2} \cdot \left(\frac{1}{2}\right)\right]^2\right) = \frac{9}{144}.
\end{aligned}$$

(d)

$$L(Y \mid X) = \frac{7}{4} + \frac{9}{111}\left(X - \frac{7}{4}\right) = \frac{3}{37}X + \frac{119}{74}.$$

2 Continuous LLSE

Note 20

Suppose that X and Y are uniformly distributed on the shaded region in the figure below.

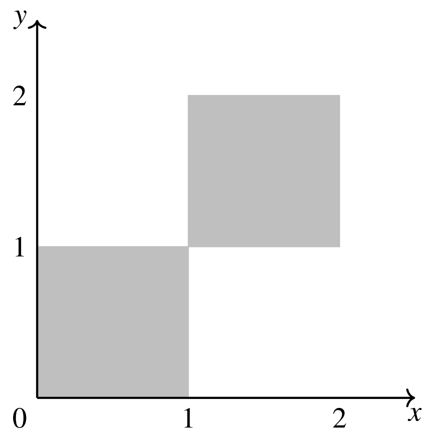

Figure 1: The joint density of (X,Y) is uniform over the shaded region.

That is, X and Y have the joint distribution:

$$f_{X,Y}(x,y) = \begin{cases} 1/2, & 0 \le x \le 1,\ 0 \le y \le 1 \\ 1/2, & 1 \le x \le 2,\ 1 \le y \le 2 \end{cases}$$

(a) Do you expect X and Y to be positively correlated, negatively correlated, or neither?

(b) Compute the marginal distribution of X.

(c) Compute $L[Y \mid X]$, the best linear estimator of Y given X.

(d) What is $\mathbb{E}[Y \mid X]$?

Solution:

(a) Positively correlated, because high values of Y correspond to high values of X.

(b) Intuitively, if we slice the joint distribution at any $x \in [0,2]$, then the probability is the same, so we should expect X to be uniformly distributed on $[0,2]$. We verify this by explicit computation:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y = 1\{0 \le x \le 1\} \int_0^1 \frac{1}{2}\,\mathrm{d}y + 1\{1 \le x \le 2\} \int_1^2 \frac{1}{2}\,\mathrm{d}y \\ &= \frac{1}{2}\mathbf{1}\{0 \le x \le 2\} \end{aligned}$$

(c) $\mathbb{E}[X] = \mathbb{E}[Y] = 1$ by symmetry. Since X is uniform on $[0,2]$, $\text{Var}(X) = 4 \cdot 1/12 = 1/3$ (since the variance of a $U[0,1]$ random variable is $1/12$). We compute the covariance:

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_0^1\int_0^1 xy \cdot \frac{1}{2}\,\mathrm{d}x\,\mathrm{d}y + \int_1^2\int_1^2 xy \cdot \frac{1}{2}\,\mathrm{d}x\,\mathrm{d}y$$
$$= \frac{1}{2}\left(\int_0^1 x\,\mathrm{d}x \int_0^1 y\,\mathrm{d}y + \int_1^2 x\,\mathrm{d}x \int_1^2 y\,\mathrm{d}y\right) = \frac{1}{2}\left(\frac{1}{4} + \frac{9}{4}\right) = \frac{5}{4}$$

So $\text{cov}(X,Y) = 5/4 - 1 \cdot 1 = 1/4$. The LLSE is

$$\begin{aligned} L[Y \mid X] &= \frac{\text{cov}(X,Y)}{\text{Var}(X)}(X - \mathbb{E}[X]) + \mathbb{E}[Y] \\ &= \frac{1/4}{1/3}(X-1) + 1 \\ &= \frac{3}{4}X + \frac{1}{4} \end{aligned}$$

(d) The easiest way to solve this is to look at the picture of the joint density, and draw horizontal lines through middles of each of the two squares. Intuitively, $\mathbb{E}[Y \mid X]$ means "for each slice of $X = x$, what is the best guess of Y"? Slightly more formally, one can argue that conditioned on $X = x$ for $0 < x < 1$, $Y \sim U[0,1]$, so $\mathbb{E}[Y \mid X = x] = 1/2$ in this region. Conditioned on $X = x$ for $1 < x < 2$, $Y \sim U[1,2]$, so $\mathbb{E}[Y \mid X = x] = 3/2$ in this region. See Figure 2.

$$\mathbb{E}[Y \mid X = x] = \begin{cases} 1/2, & 0 \le x \le 1 \\ 3/2, & 1 \le x \le 2 \end{cases}$$

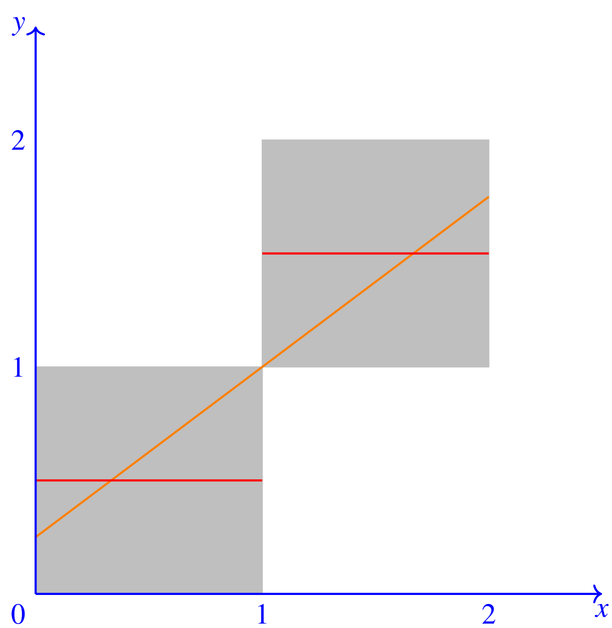

Figure 2: $L[Y \mid X]$ is the orange line. $\mathbb{E}[Y \mid X]$ is the red function.