CS 70 Discrete Mathematics and Probability Theory
Fall 2024 Hug, Rao

DIS 13A

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

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

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 $\mathrm{Var}(X)$.

(c) Compute $\mathrm{cov}(X, Y)$. (*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

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

2 Continuous LLSE

Note 20

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

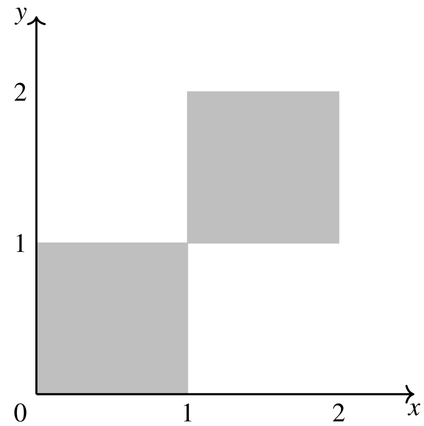

Figure 1: The joint density of (X,Y) is uniform over the shaded region.

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 $\mathbb{E}[Y \mid X]$?