S&DS 241 Lecture 22

Conditional mean, best estimates

B-H: 9.2,9.3,9.5

Setup

Consider random variables X (observed) and Y (unobserved)

- The goal is to estimate/predict Y on the basis of X.

Setup

Consider random variables X (observed) and Y (unobserved)

- The goal is to estimate/predict Y on the basis of X.
- Example:
  - Observe the temperature in New Haven, estimate the temperature in New York (easy) or that in Tokyo (hard)

Setup

Consider random variables X (observed) and Y (unobserved)

- The goal is to estimate/predict Y on the basis of X.
- Example:
  - Observe the temperature in New Haven, estimate the temperature in New York (easy) or that in Tokyo (hard)
  - Observe the stock price of GOOG in the past, predict the price tomorrow.

Setup

Consider random variables X (observed) and Y (unobserved)

- The goal is to estimate/predict Y on the basis of X.
- Example:
  - Observe the temperature in New Haven, estimate the temperature in New York (easy) or that in Tokyo (hard)
  - Observe the stock price of GOOG in the past, predict the price tomorrow.
- Let the estimate by $\hat{Y}$. A common way to measure the quality of the estimate:

$$\text{Mean Squared Error (MSE)} = E[(Y - \hat{Y})^2]$$

Setup

Consider random variables X (observed) and Y (unobserved)

- The goal is to estimate/predict Y on the basis of X.
- Example:
  - Observe the temperature in New Haven, estimate the temperature in New York (easy) or that in Tokyo (hard)
  - Observe the stock price of GOOG in the past, predict the price tomorrow.
- Let the estimate by $\hat{Y}$. A common way to measure the quality of the estimate:

$$\text{Mean Squared Error (MSE)} = E[(Y - \hat{Y})^2]$$

- We want to find the best rule for estimate Y as a function of X that minimizes the MSE.

Setup

Consider random variables X (observed) and Y (unobserved)

- The goal is to estimate/predict Y on the basis of X.
- Example:
  - Observe the temperature in New Haven, estimate the temperature in New York (easy) or that in Tokyo (hard)
  - Observe the stock price of GOOG in the past, predict the price tomorrow.
- Let the estimate by $\hat{Y}$. A common way to measure the quality of the estimate:

$$\text{Mean Squared Error (MSE)} = E[(Y - \hat{Y})^2]$$

- We want to find the best rule for estimate Y as a function of X that minimizes the MSE.
- Throughout the lecture we consider continuous RVs (X, Y). But everything works for discrete RVs, with f_{XY} replaced by p_{XY} and $f_{Y|X}$ by $p_{Y|X}$ etc.

Warm-up: X is absent

No data is available

- Suppose X is not observed. We need to make a blind guess about Y.

No data is available

- Suppose X is not observed. We need to make a blind guess about Y.
- Let the estimate by δ, which is a constant. Then

$$\begin{aligned}\text{MSE} = E[(Y-\delta)^2] &= \int (\delta - y)^2 f_Y(y) dy \\ &= (E(Y-\delta))^2 + \text{Var}(Y-\delta) = (EY-\delta)^2 + \text{Var}(Y) \geq \text{Var}(Y)\end{aligned}$$

- So the MSE is at least the variance, achieved by

$$\delta = E(Y)$$

i.e., the best blind guess under the MSE criterion is just the expectation of Y

No data is available

- Suppose X is not observed. We need to make a blind guess about Y.
- Let the estimate by δ, which is a constant. Then

$$\begin{aligned}\text{MSE} = E[(Y-\delta)^2] &= \int (\delta - y)^2 f_Y(y) dy \\ &= (E(Y-\delta))^2 + \text{Var}(Y-\delta) = (EY-\delta)^2 + \text{Var}(Y) \geq \text{Var}(Y)\end{aligned}$$

- So the MSE is at least the variance, achieved by

$$\delta = E(Y)$$

  i.e., the best blind guess under the MSE criterion is just the expectation of Y
- What if X is observed?

No data is available

- Suppose X is not observed. We need to make a blind guess about Y.
- Let the estimate by δ, which is a constant. Then

$$\begin{aligned}\text{MSE} = E[(Y-\delta)^2] &= \int (\delta - y)^2 f_Y(y) dy \\ &= (E(Y-\delta))^2 + \text{Var}(Y-\delta) = (EY-\delta)^2 + \text{Var}(Y) \geq \text{Var}(Y)\end{aligned}$$

- So the MSE is at least the variance, achieved by

$$\delta = E(Y)$$

  i.e., the best blind guess under the MSE criterion is just the expectation of Y
- What if X is observed?
  - Estimate by conditional expectation of Y given X!

Best estimate

Conditional expectation

Consider continuous RVs (X, Y). Recall:

- Joint PDF: $f_{XY}(x, y)$
- Marginal PDF: $f_X(x) = \int f_{XY}(x, y)dy$
- Conditional PDF:

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

Conditional expectation

Consider continuous RVs (X, Y). Recall:

- Joint PDF: $f_{XY}(x, y)$
- Marginal PDF: $f_X(x) = \int f_{XY}(x, y)dy$
- Conditional PDF:

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

- Conditional expectation (of Y given $X = x$):

$$E(Y|X = x) \triangleq \int y f_{Y|X}(y|x)dy$$

Note that this is a function of the value x, because the conditional PDF of Y depends on the value of $X = x$ and so does its mean.

Conditional expectation

Consider continuous RVs (X, Y). Recall:

- Joint PDF: $f_{XY}(x, y)$
- Marginal PDF: $f_X(x) = \int f_{XY}(x, y)dy$
- Conditional PDF:

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

- Conditional expectation (of Y given $X = x$):

$$E(Y|X = x) \triangleq \int y f_{Y|X}(y|x)dy$$

Note that this is a function of the value x, because the conditional PDF of Y depends on the value of $X = x$ and so does its mean.

- Conditional variance (of Y given $X = x$):

$$\mathrm{Var}(Y|X = x) \triangleq \int y^2 f_{Y|X}(y|x)dy - E(Y|X = x)^2$$

which is also a function of x.

Best estimate

Observe $X = x$, estimate Y by $\delta(x)$. Goal: minimize

$$\text{MSE} = E[(Y - \delta(X))^2] = \iint (\delta(x) - y)^2 \underbrace{f_{XY}(x,y)}_{f_X(x) f_{Y|X}(y|x)} dxdy$$

$$= \int \underbrace{\left(\int (\delta(x) - y)^2 f_{Y|X}(y|x) dy \right)}_{} f_X(x) dx$$

Best estimate

Observe $X = x$, estimate Y by $\delta(x)$. Goal: minimize

$$\text{MSE} = E[(Y - \delta(X))^2] = \iint (\delta(x) - y)^2 \underbrace{f_{XY}(x,y)}_{f_X(x) f_{Y|X}(y|x)} dxdy$$

$$= \int \underbrace{\left(\int (\delta(x) - y)^2 f_{Y|X}(y|x) dy \right)}_{\text{we have solved this problem before!}} f_X(x) dx$$

Best estimate

Observe $X = x$, estimate Y by $\delta(x)$. Goal: minimize

$$\text{MSE} = E[(Y - \delta(X))^2] = \iint (\delta(x) - y)^2 \underbrace{f_{XY}(x,y)}_{f_X(x) f_{Y|X}(y|x)} dxdy$$

$$= \int \underbrace{\left(\int (\delta(x) - y)^2 f_{Y|X}(y|x) dy \right)}_{\text{we have solved this problem before!}} f_X(x) dx$$

- For each x, the best rule is

$$\delta(x) = E(Y|X = x) = \int y f_{Y|X}(y|x) dy$$

Best estimate

Observe $X = x$, estimate Y by $\delta(x)$. Goal: minimize

$$\text{MSE} = E[(Y - \delta(X))^2] = \iint (\delta(x) - y)^2 \underbrace{f_{XY}(x,y)}_{f_X(x) f_{Y|X}(y|x)} dxdy$$

$$= \int \underbrace{\left(\int (\delta(x) - y)^2 f_{Y|X}(y|x) dy \right)}_{\text{we have solved this problem before!}} f_X(x) dx$$

- For each x, the best rule is

$$\delta(x) = E(Y|X = x) = \int y f_{Y|X}(y|x) dy$$

- Intuition:
  - Without observing X, the PDF of Y is $f_Y(y)$ and the best estimate is the unconditional mean $E(Y)$;
  - Upon observing $X = x$, the PDF of Y becomes $f_{Y|X}(y|x)$ and the best estimate is the conditional mean $E(Y|X = x)$.

Property of conditional expectation

- Recall

$$E(Y|X=x) = \int y f_{Y|X}(y|x) dy$$

  is a function of x. Let's call it $g(x)$.
- The notation $E(Y|X)$ is understood as the random variable $g(X)$.

Property of conditional expectation

- Recall

$$E(Y|X=x) = \int y f_{Y|X}(y|x)dy$$

  is a function of x. Let's call it $g(x)$.
- The notation $E(Y|X)$ is understood as the random variable $g(X)$.

Law of total expectation (aka "tower property" of expectation)

"Expectation of conditional mean = unconditional mean"

$$\boxed{E(E(Y|X)) = E(Y)}$$

Proof:

$$\begin{aligned} E(E(Y|X)) &\overset{\text{LOTUS}}{=} \int E(Y|X=x) f_X(x)dx \\ &= \iint y f_{Y|X}(y|x) f_X(x)dxdy \\ &= \int \underbrace{\left(\int f_{XY}(x,y)dx \right)}_{f_Y(y)} ydy = E(Y) \end{aligned}$$

Conditional variance and best MSE

- Recall

$$\mathrm{Var}(Y|X=x) \triangleq \int y^2 f_{Y|X}(y|x)dy - E(Y|X=x)^2$$

  is a function of x. Again, $\mathrm{Var}(Y|X)$ is understood as a random variable.

- Using the best estimate $\delta(X) = E(Y|X)$, the minimum MSE is

$$\text{MSE} = E(Y - E(Y|X))^2 = E(\mathrm{Var}(Y|X))$$

When does X help?

- The best blind guess $\delta = E(Y)$ achieves MSE $= \mathrm{Var}(Y)$.
- The best estimate $\delta(X) = E(Y|X)$ achieves MSE $= E(\mathrm{Var}(Y|X))$.

When does X help?

- The best blind guess $\delta = E(Y)$ achieves MSE $= \mathrm{Var}(Y)$.
- The best estimate $\delta(X) = E(Y|X)$ achieves MSE $= E(\mathrm{Var}(Y|X))$.

What have we gained from observing X?

- Since $\mathrm{Var}(Y|X = x) = E(Y^2|X = x) - (E(Y|X = x))^2$, we have $E(\mathrm{Var}(Y|X)) = E(E(Y^2|X) - (E(Y|X))^2) = E(Y^2) - E(E(Y|X)^2)$.

When does X help?

- The best blind guess $\delta = E(Y)$ achieves MSE $= \mathrm{Var}(Y)$.
- The best estimate $\delta(X) = E(Y|X)$ achieves MSE $= E(\mathrm{Var}(Y|X))$.

What have we gained from observing X?

- Since $\mathrm{Var}(Y|X = x) = E(Y^2|X = x) - (E(Y|X = x))^2$, we have $E(\mathrm{Var}(Y|X)) = E(E(Y^2|X) - (E(Y|X))^2) = E(Y^2) - E(E(Y|X)^2)$.
- Therefore the improvement of MSE (variance reduction) is:

$$\mathrm{Var}(Y) - E(\mathrm{Var}(Y|X)) = E(E(Y|X)^2) - E(Y)^2 = \mathrm{Var}(E(Y|X)) \geq 0$$

- The observation X is useless iff $E(Y|X) = E(Y)$ (trivial estimate)

When does X help?

- The best blind guess $\delta = E(Y)$ achieves MSE $= \mathrm{Var}(Y)$.
- The best estimate $\delta(X) = E(Y|X)$ achieves MSE $= E(\mathrm{Var}(Y|X))$.

What have we gained from observing X?

- Since $\mathrm{Var}(Y|X = x) = E(Y^2|X = x) - (E(Y|X = x))^2$, we have $E(\mathrm{Var}(Y|X)) = E(E(Y^2|X) - (E(Y|X))^2) = E(Y^2) - E(E(Y|X)^2)$.
- Therefore the improvement of MSE (variance reduction) is:

$$\mathrm{Var}(Y) - E(\mathrm{Var}(Y|X)) = E(E(Y|X)^2) - E(Y)^2 = \mathrm{Var}(E(Y|X)) \geq 0$$

- The observation X is useless iff $E(Y|X) = E(Y)$ (trivial estimate)
  - For instance, when X and Y are independent.
  - Even when X and Y are dependent, it's still possible that X does not help reducing the variance of Y, e.g., X, Y are uniform on disk.

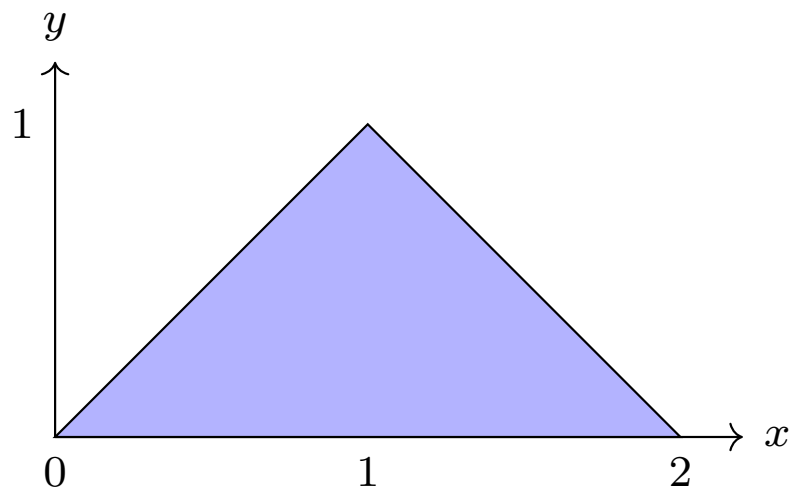
Summary

- Best (minimizing MSE) estimate of Y given X: $E(Y|X)$
- Best MSE:

$$\begin{aligned} E((Y - E(Y|X))^2) &= E(\mathrm{Var}(Y|X)) \\ &= E(Y^2) - E(E(Y|X)^2) \\ &= \mathrm{Var}(Y) - \mathrm{Var}(E(Y|X)) \end{aligned}$$

Example

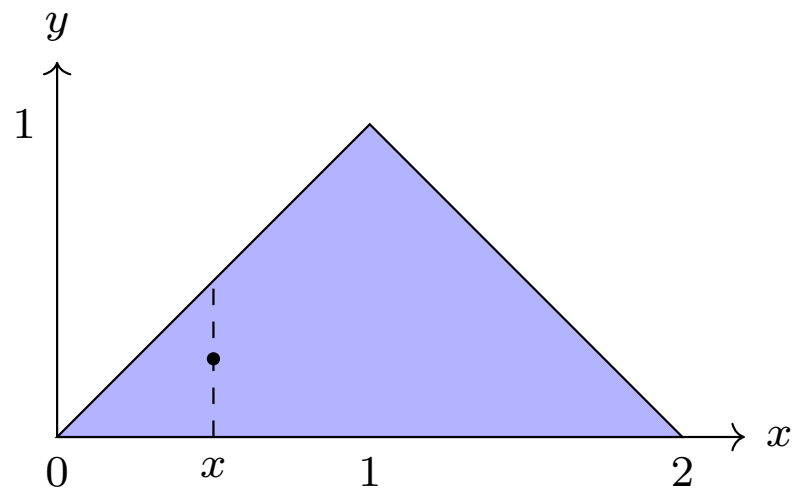
Let (X, Y) be uniformly distributed over the triangle:



What is the best estimate of Y given X?

Example

Let (X, Y) be uniformly distributed over the triangle:

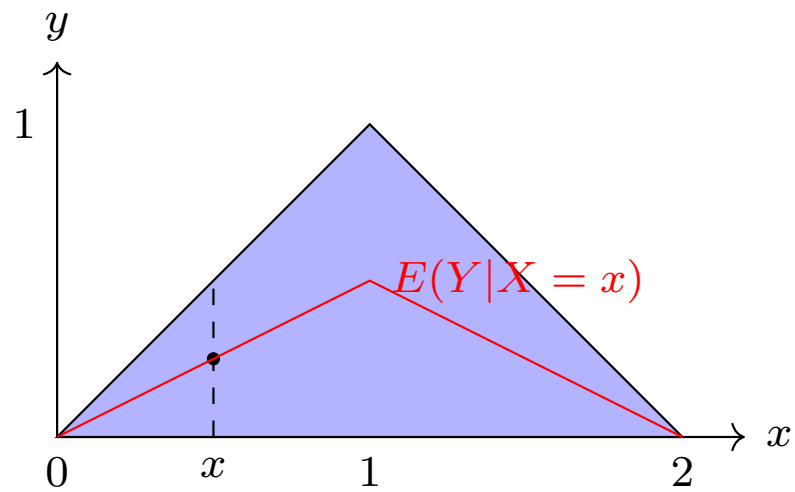


What is the best estimate of Y given X?

- Note that conditioned on $X = x$, Y is uniformly distributed over the vertical slice (Lec 18), with mean at the center

Example

Let (X, Y) be uniformly distributed over the triangle:



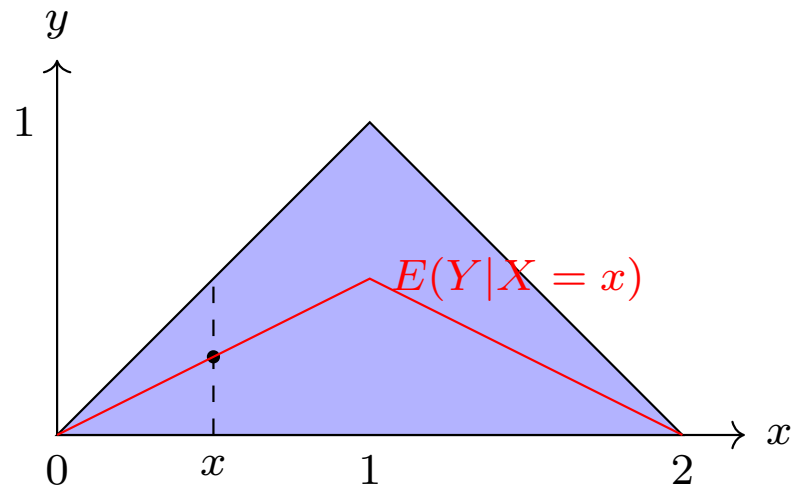
What is the best estimate of Y given X?

- Note that conditioned on $X = x$, Y is uniformly distributed over the vertical slice (Lec 18), with mean at the center
- So $E(Y|X = x)$ is given by the red curve:

$$E(Y|X = x) = \begin{cases} x/2 & x \in [0, 1] \\ 1 - x/2 & x \in [1, 2] \end{cases}$$

Example

Let (X, Y) be uniformly distributed over the triangle:



What is the best estimate of Y given X?

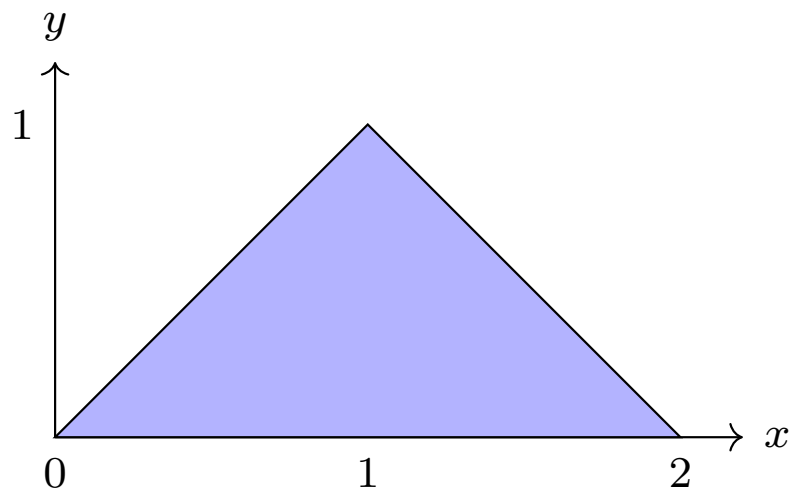
- Note that conditioned on $X = x$, Y is uniformly distributed over the vertical slice (Lec 18), with mean at the center
- So $E(Y|X = x)$ is given by the red curve:

$$E(Y|X = x) = \begin{cases} x/2 & x \in [0, 1] \\ 1 - x/2 & x \in [1, 2] \end{cases}$$

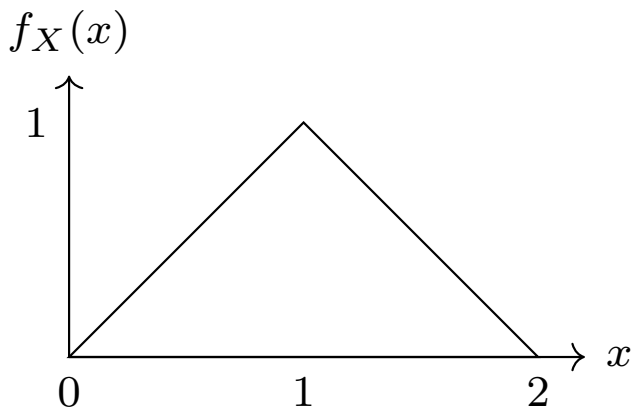
- Next let's evaluate the estimation error.

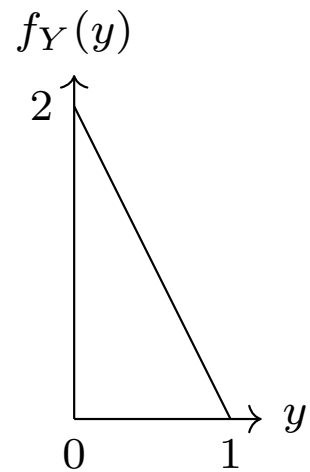
Example

Let (X, Y) be uniformly distributed over the triangle:



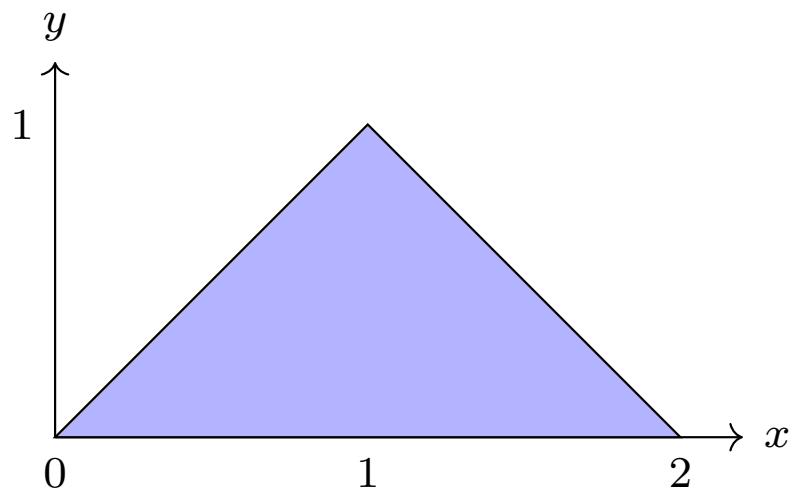
Find marginals:



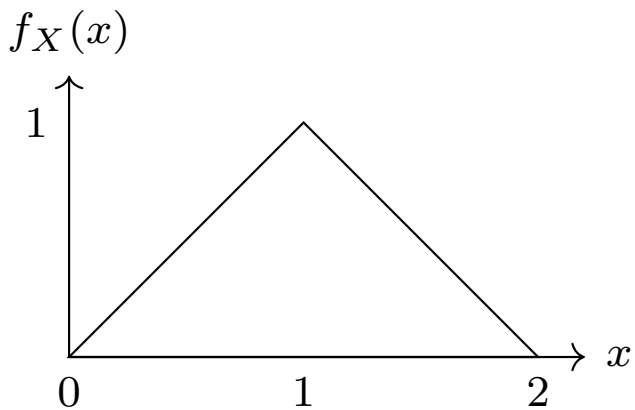


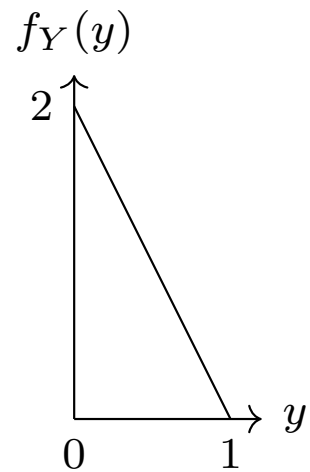
Example

Let (X, Y) be uniformly distributed over the triangle:



Find marginals:





Then $E(Y) = \int_0^1 (2-2y)y dy = 1/3$, $\mathrm{Var}(Y) = 1/18$

Example

- Given $X = x$,

$$Y \sim \begin{cases} \text{Unif}(0, x) & x \in (0, 1) \\ \text{Unif}(0, 2 - x) & x \in (1, 2) \end{cases}$$

Example

- Given $X = x$,

$$Y \sim \begin{cases} \text{Unif}(0, x) & x \in (0, 1) \\ \text{Unif}(0, 2 - x) & x \in (1, 2) \end{cases}$$

- Therefore (Lec 14)

$$\text{Var}(Y|X = x) = \begin{cases} \frac{x^2}{12} & x \in (0, 1) \\ \frac{(2-x)^2}{12} & x \in (1, 2) \end{cases}$$

Example

- Given $X = x$,

$$Y \sim \begin{cases} \text{Unif}(0, x) & x \in (0, 1) \\ \text{Unif}(0, 2 - x) & x \in (1, 2) \end{cases}$$

- Therefore (Lec 14)

$$\text{Var}(Y|X = x) = \begin{cases} \frac{x^2}{12} & x \in (0, 1) \\ \frac{(2-x)^2}{12} & x \in (1, 2) \end{cases}$$

- Averaging over X, best MSE is given by

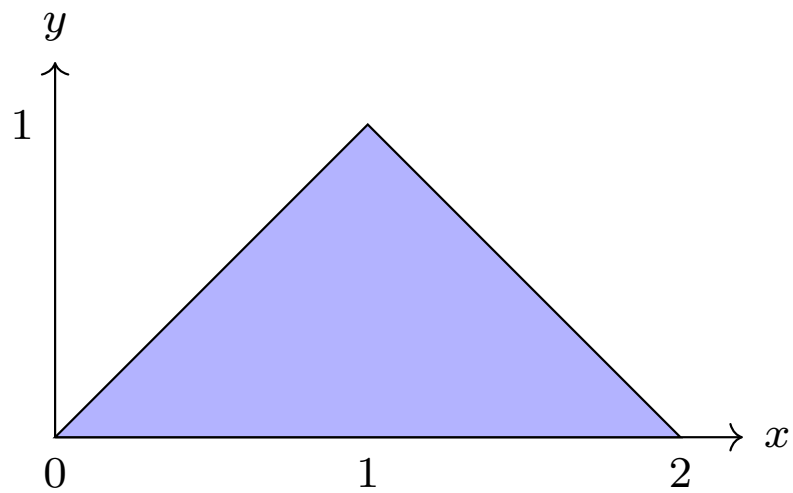
$$\begin{aligned} E(\text{Var}(Y|X)) &= \int \text{Var}(Y|X = x) f_X(x) dx \\ &= \int_0^1 \frac{x^2}{12} x dx + \int_1^2 \frac{(2-x)^2}{12}(2 - x) dx = \boxed{\frac{1}{24}} \\ &< \text{Var}(Y) = 1/18 \end{aligned}$$

  Therefore X is useful in predicting Y!

- Alternative solution: Best MSE $= E(Y^2) - E(E(Y|X)^2)$.

Example

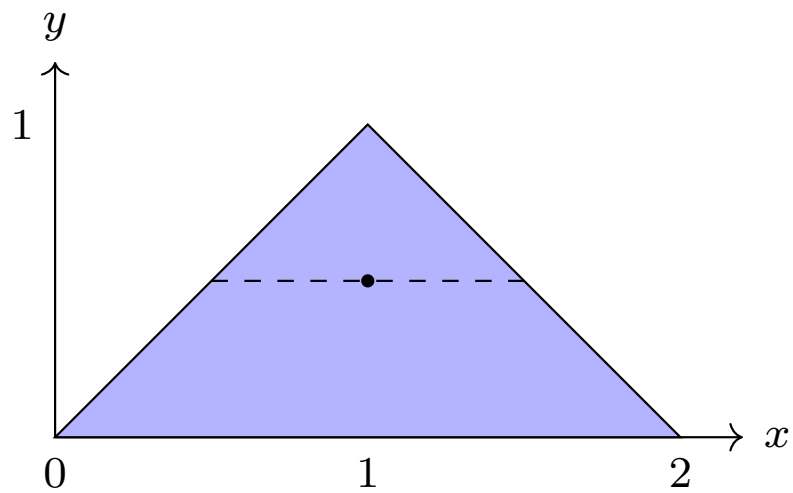
Let (X, Y) be uniformly distributed over the triangle:



How to predict X based on Y?

Example

Let (X, Y) be uniformly distributed over the triangle:

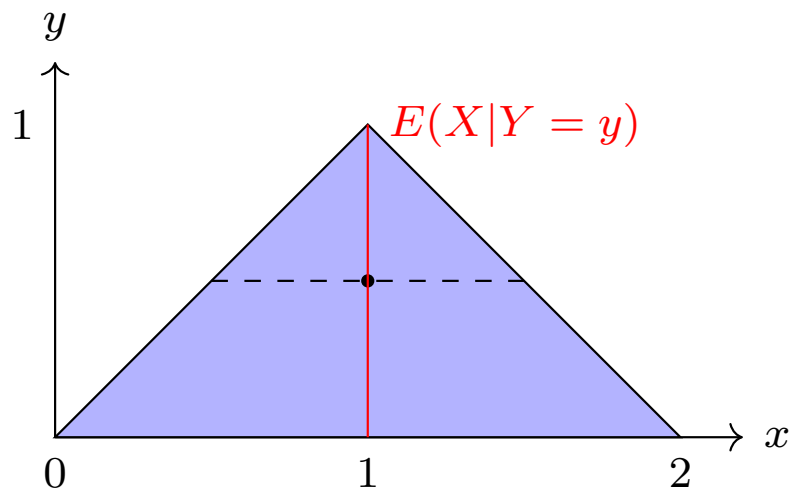


How to predict X based on Y?

$$E(X|Y) = E(X) = 1$$

Example

Let (X, Y) be uniformly distributed over the triangle:

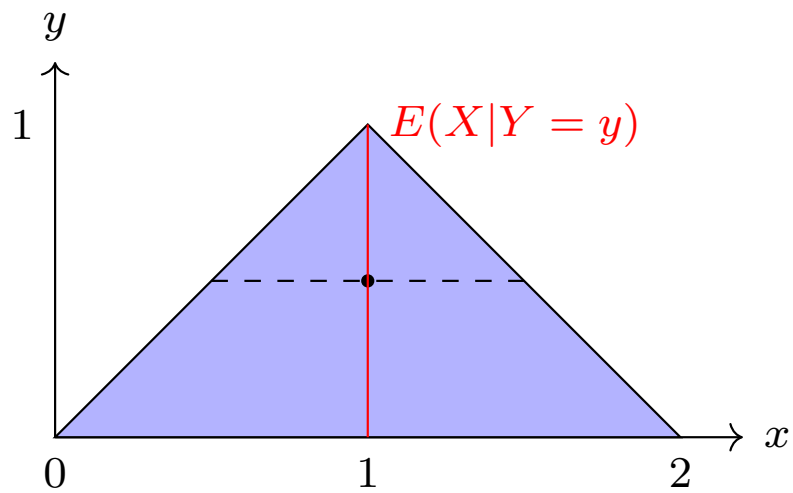


How to predict X based on Y?

$$E(X|Y) = E(X) = 1$$

Example

Let (X, Y) be uniformly distributed over the triangle:



How to predict X based on Y?

$$E(X|Y) = E(X) = 1$$

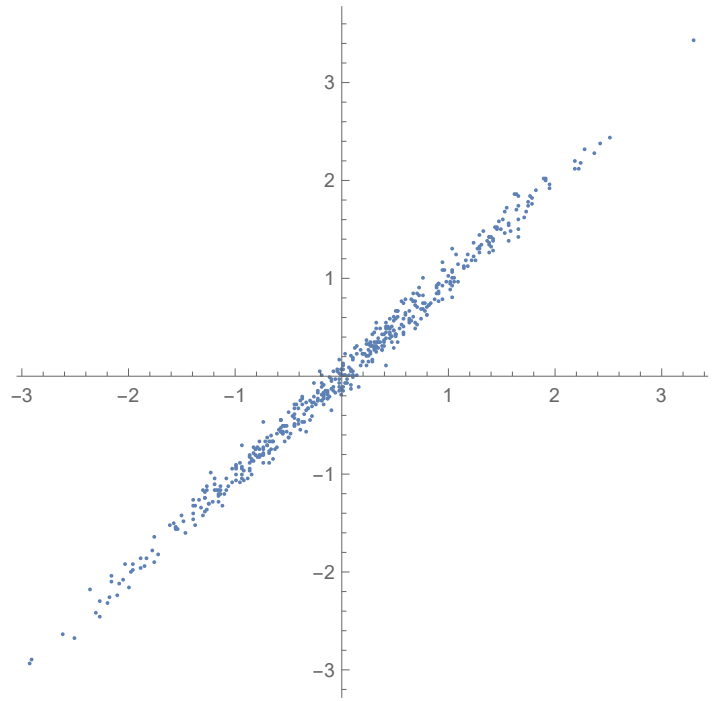
So Y is not helpful for predicting X (under the MSE criterion), even when they are dependent.

Best linear estimate

Linear estimate

Two reasons why a linear rule for estimating Y using X is desirable:

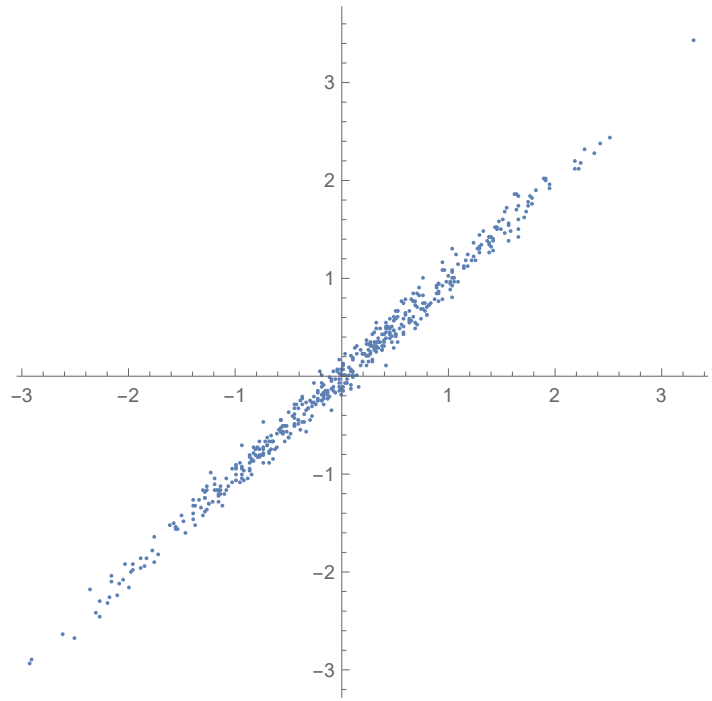
- Linear estimate is simple and interpretable:



Linear estimate

Two reasons why a linear rule for estimating Y using X is desirable:

- Linear estimate is simple and interpretable:

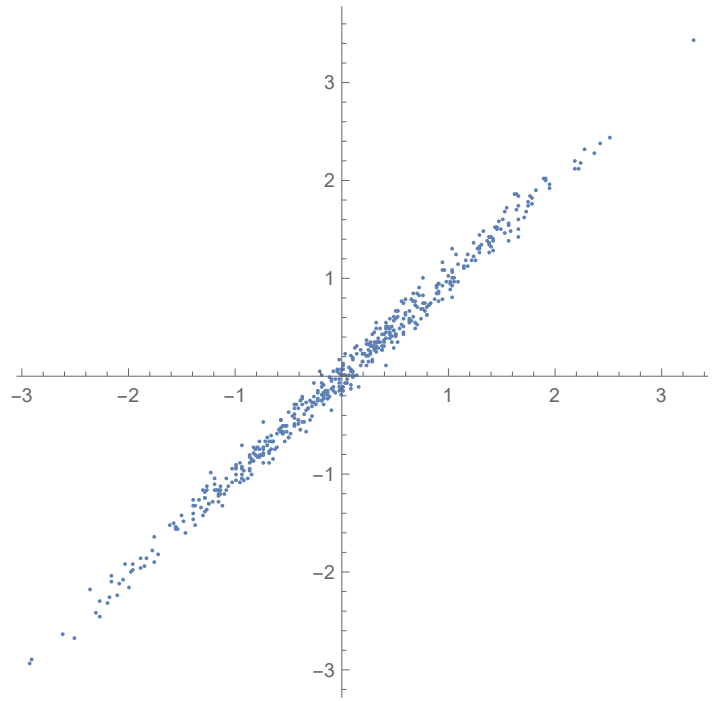


- Evaluating the best estimate requires knowing the conditional or the joint PDF, which might not be available.

Linear estimate

Two reasons why a linear rule for estimating Y using X is desirable:

- Linear estimate is simple and interpretable:



- Evaluating the best estimate requires knowing the conditional or the joint PDF, which might not be available.
  - It turns out for best linear estimate, we only need mean, variance and correlation coefficient!

Best linear estimate

Estimator: $\delta(X) = aX + b$. Next optimize over slope a and intercept b:

$$\text{MSE} = E(Y - aX - b)^2$$

Best linear estimate

Estimator: $\delta(X) = aX + b$. Next optimize over slope a and intercept b:

$$\begin{aligned} \text{MSE} &= E(Y - aX - b)^2 \\ &= \text{Var}(Y - aX) \end{aligned} \qquad b = \mu_Y - a\mu_X$$

Best linear estimate

Estimator: $\delta(X) = aX + b$. Next optimize over slope a and intercept b:

$$\begin{aligned}\text{MSE} &= E(Y - aX - b)^2 \\ &= \text{Var}(Y - aX) \qquad\qquad b = \mu_Y - a\mu_X \\ &= a^2\text{Var}(X) + \text{Var}(Y) - 2a\text{Cov}(X, Y)\end{aligned}$$

Best linear estimate

Estimator: $\delta(X) = aX + b$. Next optimize over slope a and intercept b:

$$\begin{aligned}
\text{MSE} &= E(Y - aX - b)^2 \\
&= \text{Var}(Y - aX) \qquad\qquad b = \mu_Y - a\mu_X \\
&= a^2\text{Var}(X) + \text{Var}(Y) - 2a\text{Cov}(X, Y) \\
&= a^2\sigma_X^2 + \sigma_Y^2 - 2a\rho(X, Y)\sigma_X\sigma_Y
\end{aligned}$$

Best linear estimate

Estimator: $\delta(X) = aX + b$. Next optimize over slope a and intercept b:

$$\begin{aligned}
\text{MSE} &= E(Y - aX - b)^2 \\
&= \text{Var}(Y - aX) \qquad\qquad b = \mu_Y - a\mu_X \\
&= a^2\text{Var}(X) + \text{Var}(Y) - 2a\text{Cov}(X, Y) \\
&= a^2\sigma_X^2 + \sigma_Y^2 - 2a\rho(X, Y)\sigma_X\sigma_Y \\
&= (a\sigma_X - \sigma_Y\rho)^2 + (1 - \rho^2)\sigma_Y^2
\end{aligned}$$

Best linear estimate

Estimator: $\delta(X) = aX + b$. Next optimize over slope a and intercept b:

$$\begin{aligned}
\text{MSE} &= E(Y - aX - b)^2 \\
&= \text{Var}(Y - aX) \qquad\qquad b = \mu_Y - a\mu_X \\
&= a^2\text{Var}(X) + \text{Var}(Y) - 2a\text{Cov}(X, Y) \\
&= a^2\sigma_X^2 + \sigma_Y^2 - 2a\rho(X, Y)\sigma_X\sigma_Y \\
&= (a\sigma_X - \sigma_Y\rho)^2 + (1 - \rho^2)\sigma_Y^2
\end{aligned}$$

Best coefficients:

$$\begin{cases} a = \frac{\sigma_Y\rho(X,Y)}{\sigma_X} = \frac{\text{Cov}(X,Y)}{\sigma_X^2} \\ b = \mu_Y - a\mu_X \end{cases}$$

Best linear estimate:

$$\boxed{aX + b = \mu_Y + \frac{X - \mu_X}{\sigma_X}\sigma_Y\rho(X, Y)}$$

achieves the minimum MSE among all linear estimators: $(1 - \rho^2)\sigma_Y^2$

When is linear estimation useful

- Linear estimator is useful if X and Y are correlated $(\rho \neq 0)$ and perfect if $\rho = \pm 1$
- Linear estimator is useless if X and Y are uncorrelated $(\rho = 0)$

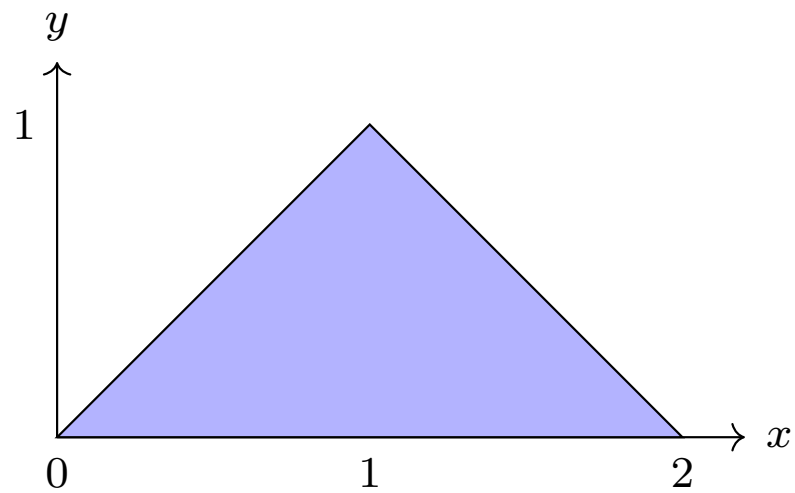
When is linear estimation useful

- Linear estimator is useful if X and Y are correlated ($\rho \neq 0$) and perfect if $\rho = \pm 1$
- Linear estimator is useless if X and Y are uncorrelated ($\rho = 0$)
- Example: $X \sim N(0, 1)$ and $Y = X^2$. Then
  - Linear estimator is trivial: since $\rho = 0$, the best linear estimate is $E(Y) = 1$ and X does not help
  - Best (non-linear) estimator: $E(Y|X) = X^2 = Y$ which is perfect.

  This example again demonstrates the limitation of linear regression

Example

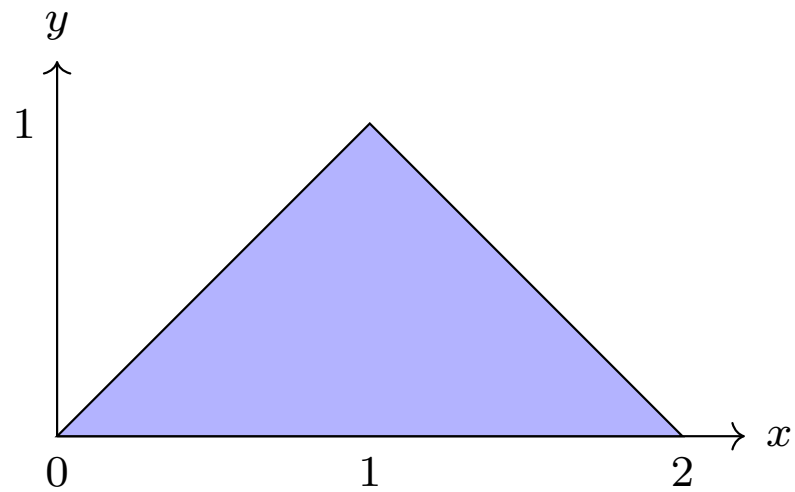
Let (X, Y) be uniformly distributed over the triangle:



What is the best linear estimate of Y given X?

Example

Let (X, Y) be uniformly distributed over the triangle:



What is the best linear estimate of Y given X?

- Find covariance:

$$E(XY) = \iint xyf_{XY}(x,y)dxdy = \int_0^1 ydy \int_y^{2-y} xdx$$

$$= \int_0^1 ydy(2-2y) = \frac{1}{3} = E(X)E(Y)$$

So $\mathrm{Cov}(X, Y) = 0$ and X and Y are uncorrelated, and linear estimate is useless.

Example: dice

Roll a fair die for n times, observe X=number of ⚀, how to predict Y=number of ⚅?

Example: dice

Roll a fair die for n times, observe X=number of ⚀, how to predict Y=number of ⚅?

- Example: suppose $n = 60$, observe ⚀ 20 times, it's reasonable to guess

Example: dice

Roll a fair die for n times, observe X=number of ⚀, how to predict Y=number of ⚅?

- Example: suppose $n = 60$, observe ⚀ 20 times, it's reasonable to guess ⚅ appear 8 times.

Example: dice

Roll a fair die for n times, observe X=number of ⚀, how to predict Y=number of ⚅?

- Example: suppose $n = 60$, observe ⚀ 20 times, it's reasonable to guess ⚅ appear 8 times.
- Let's find the best linear estimate:

$$\mu_Y + \frac{X - \mu_X}{\sigma_X}\sigma_Y \rho(X, Y)$$

where

- $\mu_X = \mu_Y = n/6$, $\sigma_X = \sigma_Y$ by symmetry
- Lec 21: $\rho(X, Y) = -1/5$

Example: dice

Roll a fair die for n times, observe X=number of ⚀, how to predict Y=number of ⚅?

- Example: suppose $n = 60$, observe ⚀ 20 times, it's reasonable to guess ⚅ appear 8 times.
- Let's find the best linear estimate:

$$\mu_Y + \frac{X - \mu_X}{\sigma_X}\sigma_Y \rho(X, Y)$$

where

- $\mu_X = \mu_Y = n/6$, $\sigma_X = \sigma_Y$ by symmetry
- Lec 21: $\rho(X, Y) = -1/5$

So best linear estimate:

$$\frac{n}{6} + \left(x - \frac{n}{6}\right)\left(-\frac{1}{5}\right) = \frac{n - x}{5}$$

Makes sense: because the other five outcomes are equally likely.