**Conditional Expectation**

Definition Let X and Y be RVs on Ω. The conditional expectation of Y given X is defined as

\[ E[Y|X] = g(X) \]

where \( g(x) = E[Y|X=x] = \sum_y y \Pr[Y=y|X=x] \).

Fact \( E[Y|X=x] = \sum_y Y(\omega) \Pr[\omega|X=x] \).

Proof: \( E[Y|X=x] = E[Y|A] \) with \( A = \{ \omega : X(\omega) = x \} \).

---

**Properties of CE**

\[ E[Y|X=x] = \sum_y y \Pr[Y=y|X=x] \]

Theorem
(a) \( X, Y \) independent \( \Rightarrow E[Y|X] = E[Y] \);
(b) \( E[aY + bZ|X] = aE[Y|X] + bE[Z|X] \);
(c) \( E[Yh(X)|X] = h(X)E[Y|X] \), \( \forall h \);
(d) \( E[h(X)E[Y|X]] = E[h(X)Y] \), \( \forall h \);
(e) \( E[E[Y|X]] = E[Y] \).

Proof:
(a), (b) Obvious
(c) \( E[Yh(X)|X=x] = \sum_{\omega} Y(\omega) h(X(\omega)) \Pr[\omega|X=x] \)
\[ = \sum_{\omega} Y(\omega) h(X(\omega)) \Pr[\omega|X=x] = h(x)E[Y|X=x] \]
Properties of CE

\[ E[Y|X = x] = \sum_y yPr[Y = y|X = x] \]

**Theorem**

(a) \( X, Y \) independent \( \Rightarrow E[Y|X] = E[Y] \);
(b) \( E[aY + bZ|X] = aE[Y|X] + bE[Z|X] \);
(c) \( E[h(X)|X] = h(X)E[Y|X|Y]h(X) \);
(d) \( E[h(X)]E[Y|X] = E[h(X)Y]\); \( h(X) \)
(e) \( E[Y|X] = E[Y] \).

**Proof:**

(d) \( E[h(X)|X] = \sum h(x)E[Y|X = x]Pr[X = x] \)
\[ = \sum h(x)\sum yPr[y|X = x]Pr[X = x] \]
\[ = \sum h(x)\sum yPr[y|X = x,y = y] \]
\[ = \sum h(x)yPr[X = x,y = y] = E[h(X)Y]. \]

**CE = MMSE (Minimum Mean Squared Estimator)**

**Theorem**

\( E[Y|X] \) is the ‘best’ guess about \( Y \) based on \( X \).
Specifically, it is the function \( g(X) \) of \( X \) that

minimizes \( E[(Y - g(X))^2] \).

**CE = MMSE**

**Theorem**

\( g(X) := E[Y|X] \) is the function of \( X \) that minimizes \( E[(Y - g(X))^2] \).

**Proof:** Recall: Expectation of r.v. minimizes mean squared error.
Sample space \( X = x \): so \( E[Y|X = x] \) minimizes mean squared error.

**Proof:**

Let \( h(X) \) be any function of \( X \). Then

\[ E[(Y - h(X))^2] = E[(Y - g(X) + g(X) - h(X))^2] \]
\[ = E[(Y - g(X))^2] + E[(g(X) - h(X))^2] \]
\[ + 2E[(Y - g(X))(g(X) - h(X))]. \]

But,

\[ E[(Y - g(X))(g(X) - h(X))] = 0 \]

by the projection property.

Thus, \( E[(Y - h(X))^2] \geq E[(Y - g(X))^2] \).

**Application: Going Viral**

Consider a social network (e.g., Twitter).
You start a rumor (e.g., Rao is not funny.)
You have friends. Each of your friends retweets w.p. \( p \).
Each of your friends has \( d \) friends, etc.

Does the rumor spread? Does it die out (mercifully)?

In this example, \( d = 4 \).
That is our next topic.

Hence, one has

**Number of tweets**

Then, the same fact holds.

Application: Going Viral

Application: Going Viral

**Application: Wald’s Identity**

Here is an extension of an identity we used in the last slide.

**Theorem Wald’s Identity**

Assume that $X_1, X_2, \ldots$ and $Z$ are independent, where $Z$ takes values in $\{0, 1, 2, \ldots\}$ and $E[X_1] = \mu$ for all $n \geq 1$.

Then,

$$E[X_1 + \cdots + X_2] = \mu E[Z].$$

**Proof:**

$E[X_1 + \cdots + X_2|Z=k] = \mu k.$

Thus, $E[X_1 + \cdots + X_2] = \mu Z$.

Hence, $E[X_1 + \cdots + X_2] = E[\mu Z] = \mu E[Z].$

**Summary**

**Conditional Expectation**

- Definition: $E[Y|X] := \sum_y yPr[Y=y|X=x]$
- Applications:
  - Viral Propagation.
  - Wald
- **MMSE:** $E[Y|X]$ minimizes $E[(Y - g(X))^2]$ over all $g(·)$

**Linear Estimation: Preamble**

Thus, if we want to guess the value of $Y$, we choose $E[Y].$

Now assume we make some observation $X$ related to $Y$.

How do we use that observation to improve our guess about $Y$?

The idea is to use a function $g(X)$ of the observation to estimate $Y$.

The “right” function is $E[Y|X]$.

A simpler function?

“Simplest” function is linear: $g(X) = a + bX$.

What is the best linear function? That is our next topic.

**Linear Regression: Motivation**

Example 1: 100 people.

Let $(X_n, Y_n) = (\text{height, weight})$ of person $n$, for $n = 1, \ldots, 100$:

![Plot](image-url)
**Motivation**

Example 2: 15 people.

We look at two attributes: \((X_n, Y_n)\) of person \(n\), for \(n = 1, \ldots, 15\):

**LLSE**

**Theorem**

Consider two RVs \(X, Y\) with a given distribution \(Pr[X = x, Y = y]\).

Then,

\[
L(Y|X) = \hat{Y} = E[Y] + \frac{cov(X,Y)}{var(X)} (X - E[X]).
\]

**Proof 1:**

\[
Y - \hat{Y} = (Y - E[Y]) - \frac{cov(X,Y)}{var(X)} (X - E[X]), \quad E[Y - \hat{Y}] = 0 \text{ by linearity.}
\]

Also, \(E[Y - \hat{Y}|X] = 0\), after a bit of algebra. (See next slide.)

Combine brown inequalities: \(E[Y - \hat{Y}|X](c + dX) = 0\) for any \(c, d\).

Since: \(\hat{Y} = \alpha + \beta X\) for some \(\alpha, \beta\), so \(\exists c, d\) s.t. \(\hat{Y} - a - bX = c + dX\).

Then, \(E[(Y - \hat{Y})|X] = E[(Y - a - bX)^2] = \frac{cov(X,Y)}{var(X)} + \frac{cov(X,Y)}{var(X)} (X - E[X])\).

This shows that \(E[(Y - \hat{Y})^2] \leq E[(Y - a - bX)^2]\) for all \((a, b)\).

Thus \(\hat{Y}\) is the LLSE.

**A Bit of Algebra**

\[
Y - \hat{Y} = (Y - E[Y]) - \frac{cov(X,Y)}{var(X)} (X - E[X]).
\]

Hence, \(E[Y - \hat{Y}] = 0\). We want to show that \(E[(Y - \hat{Y})X] = 0\).

**Estimation Error**

We saw that the LLSE of \(Y\) given \(X\) is

\[
L(Y|X) = \hat{Y} = E[Y] + \frac{cov(X,Y)}{var(X)} (X - E[X]).
\]

How good is this estimator? Or what is the mean squared estimation error?

We find

\[
E[\hat{Y}^2] = E[Y^2] - \frac{cov^2(X,Y)}{var(X)} + \frac{cov^2(X,Y)}{var(X)} (X - E[X])^2.
\]

Without observations, the estimate is \(E[Y]\). The error is \(var(Y)\). Observing \(X\) reduces the error.
Estimation Error: A Picture
We saw that
\[ L[Y|X] = \hat{Y} = E[Y] + \frac{\text{cov}(X,Y)}{\text{var}(X)} (X - E[X]) \]
and
\[ E[Y - L[Y|X]]^2 = \text{var}(Y) - \frac{\text{cov}(X,Y)^2}{\text{var}(X)}. \]
Here is a picture when \( E[X] = 0, E[Y] = 0 \):
Dimensions correspond to sample points, uniform sample space.

Vector \( Y \) at dimension \( \omega \) is \( \frac{1}{\sqrt{\Omega}} Y(\omega) \)

Linear Regression Examples
Example 1:
We find:
\[ E[X] = 0; E[Y] = 0; E[X^2] = 1/2; E[XY] = 1/2; \]
\[ \text{var}[X] = E[X^2] - E[X]^2 = 1/2; \text{cov}(X,Y) = E[XY] - E[X]E[Y] = 1/2; \]
\[ \text{LR: } \hat{Y} = E[Y] + \text{cov}(X,Y) \]
\[ \text{var}[X] (X - E[X]) = X. \]

Linear Regression Examples
Example 3:
We find:
\[ E[X] = 0; E[Y] = 0; E[X^2] = 1/2; E[XY] = -1/2; \]
\[ \text{var}[X] = E[X^2] - E[X]^2 = 1/2; \text{cov}(X,Y) = E[XY] - E[X]E[Y] = -1/2; \]
\[ \text{LR: } \hat{Y} = E[Y] + \text{cov}(X,Y) \]
\[ \text{var}[X] (X - E[X]) = -X. \]

Linear Regression Examples
Example 4:
We find:
\[ E[X] = 3; E[Y] = 2.5; E[X^2] = (3/15)(1 + 2^2 + 3^2 + 4^2 + 5^2) = 8.4; \]
\[ \text{var}[X] = 11 - 9 = 2; \text{cov}(X,Y) = 8.4 - 3 \times 2.5 = 0.9; \]
\[ \text{LR: } \hat{Y} = 2.5 + \frac{0.9}{2} (X - 3) = 1.15 + 0.45X. \]

Linear Regression Examples
Example 2:
We find:
\[ E[X] = 0; E[Y] = 0; E[X^2] = 1/2; E[XY] = 1/2; \]
\[ \text{var}[X] = E[X^2] - E[X]^2 = 1/2; \text{cov}(X,Y) = E[XY] - E[X]E[Y] = 1/2; \]
\[ \text{LR: } \hat{Y} = E[Y] + \frac{\text{cov}(X,Y)}{\text{var}[X]} (X - E[X]) = X. \]

Note that
\[ \text{the LR line goes through } (E[X], E[Y]) \]
\[ \text{its slope is } \frac{\text{cov}(X,Y)}{\text{var}[X]}. \]
Quadratic Regression

Let \( X, Y \) be two random variables defined on the same probability space.

**Definition:** The quadratic regression of \( Y \) over \( X \) is the random variable
\[
Q[Y|X] = a + bX + cX^2
\]
where \( a, b, c \) are chosen to minimize \( E[(Y - a - bX - cX^2)^2] \).

**Derivation:** We set to zero the derivatives w.r.t. \( a, b, c \). We get
\[
\begin{align*}
0 &= E[Y - a - bX - cX^2] - a - bE[X] - cE[X^2] \\
0 &= E[(Y - a - bX - cX^2)X] - a - bE[X^2] - cE[X^3] \\
\end{align*}
\]
We solve these three equations in the three unknowns \( a, b, c \).

Summary

**Linear Regression**

Mean Squared: \( E[Y] \) is best mean squared estimator for \( Y \).

MMSE: \( E[Y|X] \) is best mean squared estimator for \( Y \) given \( X \).

Linear Regression: \( L[Y|X] = E[Y] + \frac{\text{cov}(X,Y)}{\text{var}(X)}(X - E[X]) \)

Can do other forms of functions as well, e.g., quadratic.

Warning: assumes you know distribution.

Sample Points “are” distribution in this class.

Statistics: Fix the assumption above.