Today

Finish Linear Regression:
Best linear function prediction of $Y$ given $X$.

MMSE: Best Function that predicts $Y$ from $S$.

Conditional Expectation.

Applications to random processes.

---

**Estimation Error**

We saw that the LLSE of $Y$ given $X$ is

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

How good is this estimator?

Or what is the mean squared estimation error?

We find

$$ E[(Y - L[Y|X])^2] = E[(Y - E[Y] - \frac{\text{cov}(X,Y)}{\text{var}(X)}(X - E[X]))^2] $$

$$ = E[(Y - E[Y])^2] - 2\text{cov}(X,Y)/\text{var}(X)E[(Y - E[Y])(X - E[X])] $$

$$ + \frac{\text{cov}(X,Y)^2}{\text{var}(X)}E[(X - E[X])^2] $$

$$ = \text{var}(Y) - \text{cov}(X,Y)^2/\text{var}(X). $$

Without observations, the estimate is $E[Y]$. The error is $\text{var}(Y)$. Observing $X$ reduces the error.

---

**LLSE**

**Theorem**

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

Then,

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

**Proof 1:**

$$ E[Y|X] = E[Y] + \frac{\text{cov}(X,Y)}{\text{var}(X)}(X - E[X]) $$

Or what is the mean squared estimation error?

Since:

$$ \text{var}(Y) = E[(Y - E[Y])^2] $$

Then,

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

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**

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

Hence, $E[Y - \hat{Y}] = 0$. We want to show that $E[(Y - \hat{Y})^2] = 0$.

Note that

$$ E[(Y - \hat{Y})^2] = E[(Y - \hat{Y})(X - E[X])]. $$

because $E[(Y - \hat{Y})X] = E(Y - \hat{Y})E[X] = 0$.

Now,

$$ E[(Y - \hat{Y})(X - E[X])] $$

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

$$ = \text{cov}(X,Y) - \frac{\text{cov}(X,Y)^2}{\text{var}(X)}. $$

(Recall that $\text{cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$ and $\text{var}(X) = E[(X - E[X])^2]$.)

---

**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 $E[Y] = 0$.

Dimensions correspond to sample points, uniform sample space.

---

**Linear Regression Examples**

**Example 1:**

- **Linear Regression**
- $\theta_0 + \theta_1 X_0$
### Linear Regression Examples

**Example 2:**

We find:

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

**Example 3:**

We find:

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

**Example 4:**

We find:

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

### Summary

1. Linear Regression: \[ L[Y|X] = E[Y] + \frac{\text{cov}(X, Y)}{\text{var}(X)}(X - E[X]) \]
2. Non-Bayesian: minimize \[ \sum_a(Y_n - a - bX_n)^2 \]
3. Bayesian: minimize \[ E[(Y - a - bX)^2] \]

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

We set to zero the derivatives w.r.t.

The conditional expectation of

Hence,

The quadratic regression of

That is, the estimation error is orthogonal to all

\( Y \) \text{es.}

Quite!

Or

Simple but most convenient.

It could be that

Or that

Conditional Expectation

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

\[
\begin{align*}
0 &= E[Y - a - bX - cX^2] \\
0 &= E[(Y - a - bX - cX^2)|X] \\
0 &= E[(Y - a - bX - cX^2)|X]
\end{align*}
\]

We solve these three equations in the three unknowns \( a, b, c \).

**Note:** These equations imply that \( E[(Y - Q[Y|X])h(X)] = 0 \) for any \( h(X) = d + eX + fx^2 \). That is, the estimation error is orthogonal to all the quadratic functions of \( X \). Hence, \( Q[Y|X] \) is the projection of \( Y \) onto the space of quadratic functions of \( X \).

Quadratic Regression

Nonlinear Regression: Motivation

There are many situations where a good guess about \( Y \) given \( X \) is not linear.

E.g., (diameter of object, weight), (school years, income), (PSA level, cancer risk).

Our goal: explore estimates \( \hat{Y} = g(X) \) for nonlinear functions \( g(\cdot) \).

Deja vu, all over again?

Have we seen this before? Yes.

Is anything new? Yes.

The idea of defining \( g(x) = E[Y|X = x] \) and then \( E[Y|X] = g(X) \).

Big deal? Quite! Simple but most convenient.

Recall that \( E[Y|X = x] = a + bX \) is a function of \( X \).

This is similar: \( E[Y|X] = g(X) \) for some function \( g(\cdot) \).

In general, \( g(X) \) is not linear, i.e., not \( a + bX \). It could be that \( g(X) = a + bX + cX^2 \). Or that \( g(X) = 2\sin(4X) + \exp(-3X) \). Or something else.

Conditional Expectation

**Definition** Let \( X \) and \( Y \) be RVs on \( \Omega \). The conditional expectation of \( Y \) given \( X \) is defined as

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

where

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

**Fact**

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

**Proof:**

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

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) \( E[Yh(X)|X] = h(x)E[Y|X], \forall h(\cdot) \);

(d) \( E[h(X)E[Y|X]|X] = E[h(x)E[Y|X]], \forall h(\cdot) \);

(e) \( E[E[Y|X]] = E[Y] \).

**Proof:**

\[ \begin{align*}
(a) & \text{ Obvious} \\
(b) & \text{ E[aY + bZ|X] = aE[Y|X] + bE[Z|X] } \\
(c) & \text{ E[Yh(X)|X] = \sum_{\omega} Y(\omega)h(X(\omega))Pr[\omega|X = x] } \\
& \quad = \sum_{\omega} Y(\omega)h(x)Pr[\omega|X = x] = h(x)E[Y|X = x]
\end{align*} \]
Consequently, if you pick a red ball, we find the theorem:

**Theorem**

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

**Proof:** (continued)

(d) E[Xh(X)]E[Y|X] = ∑h(x)E[Y|x]Pr[X = x]Pr[X = x] = ∑h(x)∑yPr(y|x)Pr[X = x] = ∑h(x)E[X = x|y]Pr[X = x] = E[Xh(X)].

Let h(X) = 1 in (d).

We say that the estimation error Y − E[Y|X] is orthogonal to every function h(X) of X.

We call this the projection property. More about this later.
Application: Going Viral

Consider a social network (e.g., Twitter).

You start a rumor (e.g., Rao is bad at making copies).

You have \( d \) friends. Each of your friend retweets w.p. \( p \).

Each of your friends has \( d \) friends, etc.

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

In this example, \( d = 4 \).

Fact: Number of tweets \( X = \sum_{n=1}^{\infty} X_n \) where \( X_n \) is tweets in level \( n \).

Then, \( E[X] < \infty \) iff \( pd < 1 \).

Proof:

Given \( X_0 = k, X_{n+1} = B(k, p) \). Hence, \( E[X_{n+1}|X_0] = kpd \).

Thus, \( E[X_{n+1}|X_0] = pdX_0 \). Consequently, \( E[X_0] = (pd)^{n-1}, n > 1 \).

If \( pd < 1 \), then \( E[X_0 + \cdots + X_n] \leq (1 - pd)^{-1} \implies E[X] \leq (1 - pd)^{-1} \).

If \( pd \geq 1 \), then for all \( C \) one can find \( n \) s.t. \( E[X] \geq E[X_0 + \cdots + X_n] \geq C \).

In fact, one can show that \( pd \geq 1 \implies \Pr[X = w] > 0 \).
Application: Going Viral

An easy extension: Assume that everyone has an independent number $D_i$ of friends with $E[D_i] = d$. Then, the same fact holds.
To see this, note that given $X_n = k$, and given the numbers of friends $D_1 = d_1$, ..., $D_k = d_k$ of these $X_n$ people, one has $X_{n+1} = B(d_1 + \cdots + d_k, p)$. Hence,
$$E[X_{n+1}|X_n = k, D_1 = d_1, \ldots, D_k = d_k] = p(d_1 + \cdots + d_k).$$
Thus, $E[X_{n+1}|X_n = k, D_1, \ldots, D_k] = p(D_1 + \cdots + D_k)$.
Consequently, $E[X_{n+1}|X_n = k] = E[p(D_1 + \cdots + D_k)] = pdk$.
Finally, $E[X_{n+1}|X_n] = pdX_n$, and $E[X_{n+1}] = pdE[X_n]$.
We conclude as before.

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 $E[X_n] = \mu$ for all $n \geq 1$.
Then,
$$E[X_1 + \cdots + X_Z | Z] = \mu \cdot Z.$$

Proof:
$$E[X_1 + \cdots + X_Z | Z = k] = \mu k.$$
Thus, $E[X_1 + \cdots + X_Z | Z] = \mu Z$.
Hence, $E[X_1 + \cdots + X_Z] = E[\mu Z] = \mu E[Z]$.

CE = MMSE

Theorem CE = MMSE

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

Proof:
Let $h(X)$ be any function of $X$. Then
$$E[(Y - h(X))^2] = E[(Y - g(X) + g(X) - h(X))^2]$$
$$= E[(Y - g(X))^2] + E[(g(X) - h(X))^2]$$
$$\quad + 2E[(Y - g(X))(g(X) - h(X))];$$
But,
$$E[(Y - g(X))(g(X) - h(X))] = 0$$
by the projection property.

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

$E[Y|X]$ and $L[Y|X]$ as projections

$L[Y|X]$ is the projection of $Y$ on $\{a + bX, a, b \in \mathbb{R}\}$: LLSE

$E[Y|X]$ is the projection of $Y$ on $\{g(X), g(\cdot) : \mathbb{R} \rightarrow \mathbb{R}\}$: MMSE.

Summary

- Definition: $E[Y|X] := \sum_y y \cdot \Pr(Y = y | X = x)$
- Properties: Linearity: $Y - E[Y|X] \perp h(X); E[E[Y|X]] = E[Y]$.
- Some Applications:
  - Calculating $E[Y|X]$
  - Diluting
  - Mixing
  - Rumors
  - Wald
- MMSE: $E[Y|X]$ minimizes $E[(Y - g(X))^2]$ over all $g(\cdot)$.