Today

Estimation.

MMSE: Best Function that predicts X from Y.

Conditional Expectation.

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

Applications to random processes.

Estimation: cs70 style

Given distribution for Y.

What is the distribution?
Probability "mass" function: $Pr[Y=y]$.

What should we guess for the value of Y, before hand?

That is what number $\hat{Y}$ should we predict for Y?

Estimation: Expectation and Mean Squared Error.

Given distribution (probability mass function): $Pr[Y=y]$.

"Best" guess about Y, is $E[Y]$.
If "best" is Mean Squared Error.

More precisely, the value of a that minimizes $E[(Y-a)^2]$ is $a=E[Y]$.

Proof:

Let $\hat{Y} := Y - E[Y]$.
Then, $E[\hat{Y}] = E[Y - E[Y]] = E[Y] - E[Y] = 0$.
So, $E[\hat{Y}c] = 0, \forall c$. Now,

$$\begin{aligned} E[(Y-a)^2] &= E[(Y - E[Y] + E[Y] - a)^2] \\ &= E[(\hat{Y}+c)^2] \text{ with } c = E[Y]-a \\ &= E[\hat{Y}^2 + 2\hat{Y}c + c^2] = E[\hat{Y}^2] + 2E[\hat{Y}c] + c^2 \\ &= E[\hat{Y}^2] + 0 + c^2 \geq E[\hat{Y}^2]. \end{aligned}$$

Hence, $E[(Y-a)^2] \geq E[(Y-E[Y])^2], \forall a$. □

Estimation: Preamble

Thus, best guess, $\hat{Y}$, for the value of Y, is $E[Y]$.

Now assume we make some observation X related to Y.

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

Review

Definitions Let X and Y be RVs on Ω.

- Distribution: $Pr[Y=y]$
- Joint Distribution: $Pr[X=x, Y=y]$
- Marginal Distribution: $Pr[X=x] = \sum_y Pr[X=x, Y=y]$
- Conditional Distribution: $Pr[Y=y|X=x] = \frac{Pr[X=x,Y=y]}{Pr[X=x]}$

What is $\sum_{x,y} Pr[X=x, Y-y]$? 1.

What is $\sum_x Pr[X=x]$? 1

What is $\sum_y Pr[X=x, Y=y]$? $Pr[X=x]$.

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

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

Fact

$$E[Y|X=x] = \sum_\omega Y(\omega) \times Pr[\omega|X=x].$$

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

What is "X = x"? An event. In the above? The event $A = 'X=x'$.

Note: $E[Y|X]$ is a function on values for X that gives a number.

Today: we view as a predicted value for Y.

Properties of CE

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

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

Proof:

(a) Obvious and $Pr[Y=y|X=x] = Pr[Y=y]$
](b)Linearity of expectation in sample space.
(c) $E[Yh(X)|X=x] = \sum_\omega Y(\omega)h(X(\omega))Pr[\omega|X=x]$

$$= \sum_\omega Y(\omega)h(x)Pr[\omega|X=x] = h(x)E[Y|X=x]$$

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) $E[Yh(X)|X] = h(X)E[Y|X], \forall h(\cdot)$;
(d) $E[h(X)E[Y|X]] = E[h(X)Y], \forall h(\cdot)$;
(e) $E[E[Y|X]] = E[Y]$.

Proof: (continued)

(d) $E[h(X)E[Y|X]] = \sum_x h(x)E[Y|X=x]Pr[X=x]$

$$= \sum_x h(x)\sum_y y \times Pr[Y=y|X=x]Pr[X=x]$$
$$= \sum_x h(x)\sum_y y \times Pr[X=x, y=y]$$
$$= \sum_{x,y} h(x)y \times Pr[X=x, y=y] = E[h(X)Y].$$

□

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) $E[h(X)E[Y|X]] = E[h(X)Y], \forall h(\cdot)$;
(e) $E[E[Y|X]] = E[Y]$.

Proof: (continued) (e) Let $h(X) = 1$ in (d).

□

Properties of CE

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) $E[h(X)E[Y|X]] = E[h(X)Y], \forall h(\cdot)$;
(e) $E[E[Y|X]] = E[Y]$.

Note that (d) says that

$$E[(Y - E[Y|X])h(X)|X] = 0.$$

Note: one view is that the estimation error $Y - E[Y|X]$ is orthogonal to every function $h(X)$ of X.

This the projection property.

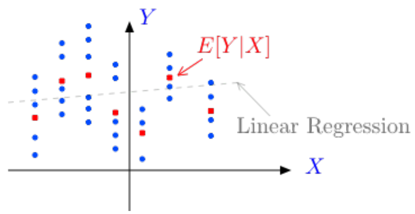
It gives that $E[Y|X]$ is best estimator for Y given X.

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 CE = MMSE

$g(X) := E[Y|X]$ is the function of X that minimizes $E[(Y - g(X))^2]$.
Proof: Recall: Expectation of r.v. mimimizes 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

$$\begin{aligned} E[(Y - h(X))^2] &= E[(Y - g(X) + g(X) - h(X))^2] \\ &= E[(Y - g(X))^2] + E[(g(X) - h(X))^2] \\ &\quad +2E[(Y - g(X))(g(X) - h(X))]. \end{aligned}$$

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

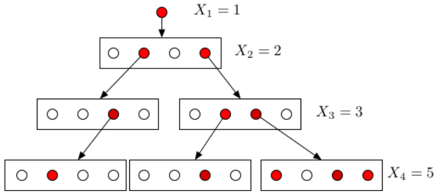
Consider a social network (e.g., Twitter).

You start a rumor (e.g., Rao is not funny.)

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

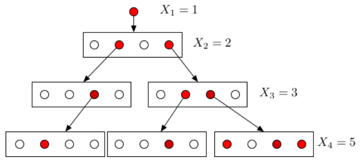
Each of your friends has d friends, etc.

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



In this example, $d = 4$.

Application: Going Viral



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 $X_n = k$, $X_{n+1} = B(kd, p)$. Hence, $E[X_{n+1}|X_n = k] = kpd$.

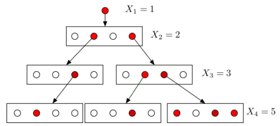
Thus, $E[X_{n+1}|X_n] = pdX_n$. Consequently, $E[X_n] = (pd)^{n-1}, n \geq 1$.

If $pd < 1$, then $E[X_1 + \cdots + X_n] \leq (1-pd)^{-1} \implies E[X] \leq (1-pd)^{-1}$.

If $pd \geq 1$, then for all C one can find n s.t.
$E[X] \geq E[X_1 + \cdots + X_n] \geq C$. □

In fact, one can show that $pd \geq 1 \implies Pr[X = \infty] > 0$.

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 $D_1 = d_1, \ldots, D_k = d_k$ 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,

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

Proof:

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

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

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

Summary

Conditional Expectation

- Definition: $E[Y|X] := \sum_y yPr[Y = y|X = x]$
- Properties: $E[Y - E[Y|X]h(X)|X] = 0$; $E[E[Y|X]] = E[Y]$
- Applications:
  - Viral Propagation.
  - Wald
- MMSE: $E[Y|X]$ minimizes $E[(Y - g(X))^2]$ over all $g(\cdot)$

Linear Estimation: Preamble

Best MMSE, $\hat{Y}$, the value of Y, we choose $E[Y]$.

Given 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 $\hat{Y}(X) = g(X)$ of the observation to estimate Y.

The "right" function is $E[X|Y]$.

A simpler function?

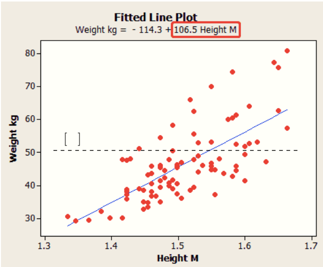
"Simplest" function is linear: $g(X) = a + bX$.

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

Linear Regression: Motivation

Example 1: 100 people.

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



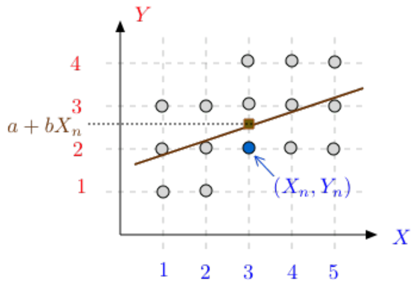
The blue line is $Y = -114.3 + 106.5X$. (X in meters, Y in kg.)

Best linear fit: Linear Regression.

Motivation

Example 2: 15 people.

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



The line $Y = a + bX$ is the linear regression.

LLSE

$LLSE[Y|X]$ - best guess for Y given X.

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:

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

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

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

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

Then, $E[(Y - \hat{Y})(\hat{Y} - a - bX)] = 0, \forall a, b$. Now,

$$E[(Y - a - bX)^2] = E[(Y - \hat{Y} + \hat{Y} - a - bX)^2]$$
$$= E[(Y - \hat{Y})^2] + E[(\hat{Y} - a - bX)^2] + 0 \geq E[(Y - \hat{Y})^2].$$

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

Thus $\hat{Y}$ is the LLSE. $\square$

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

Note that

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

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

Now,

$$E[(Y - \hat{Y})(X - E[X])]$$
$$= E[(Y - E[Y])(X - E[X])] - \frac{cov(X,Y)}{var[X]} E[(X - E[X])(X - E[X])]$$
$$=^{(*)} cov(X,Y) - \frac{cov(X,Y)}{var[X]} var[X] = 0. \quad \square$$

(*) Recall that $cov(X,Y) = E[(X - E[X])(Y - E[Y])]$ and $var[X] = E[(X - E[X])^2]$.

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

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

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

Estimation Error: A Picture

We saw that

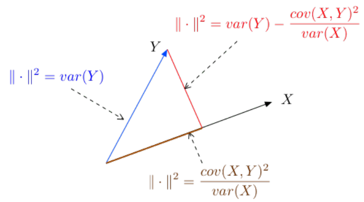
$$L[Y|X] = \hat{Y} = E[Y] + \frac{cov(X,Y)}{var(X)}(X - E[X])$$

and

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

Here is a picture when $E[X] = 0, E[Y] = 0$:

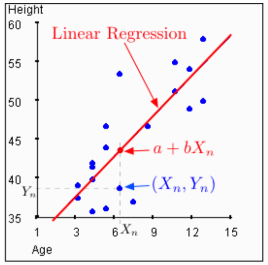
Dimensions correspond to sample points, uniform sample space.



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

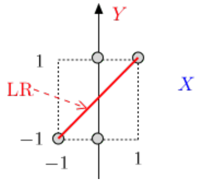
Linear Regression Examples

Example 1:



Linear Regression Examples

Example 2:

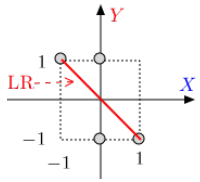


We find:

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

Linear Regression Examples

Example 3:

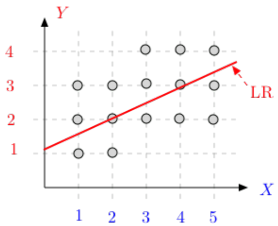


We find:

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

Linear Regression Examples

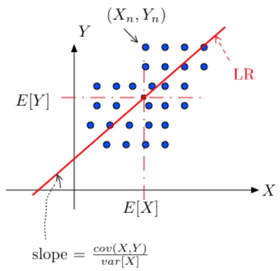
Example 4:



We find:

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

LR: Another Figure



Note that

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

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

$$0 = E[Y - a - bX - cX^2] = E[Y] - a - bE[X] - cE[X^2]$$
$$0 = E[(Y - a - bX - cX^2)X] = E[XY] - a - bE[X^2] - cE[X^3]$$
$$0 = E[(Y - a - bX - cX^2)X^2] = E[X^2Y] - aE[X^2] - bE[X^3] - cE[X^4]$$

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

Note on pedagogy.

We used the projection property to verify MMSE and LLSE.

MMSE: $E[h(X)(Y - E(Y|X))] = 0$ implies $E[Y|X]$ is best predictor given X.

LLSE: $E[L(X)(Y - LLSE(Y|X))] = 0$ implise $LLSE(Y|X)$ is best linear predictor given X.

We used calculus to do best Quadratic prediction.

Notes: use calculus to prove optimaliaty of $E[Y|X]$ and $LLSE[Y|X]$.

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

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.