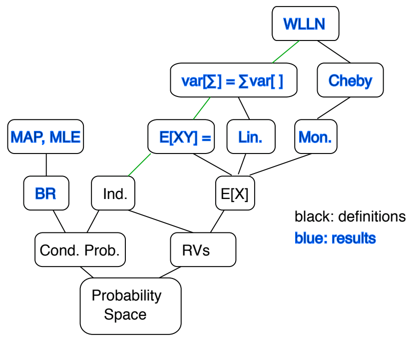
CS70: Lecture 22.

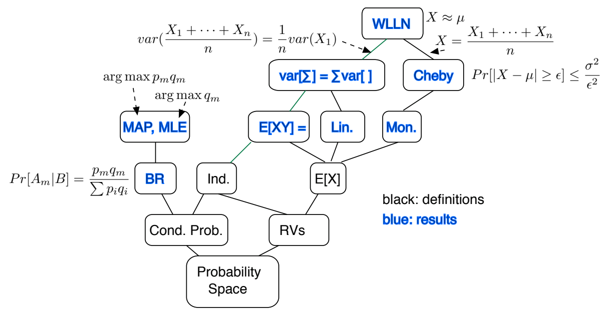
Confidence Intervals; Linear Regression

1. Review
2. Confidence Intervals
3. Motivation for LR
4. History of LR
5. Linear Regression
6. Derivation
7. More examples

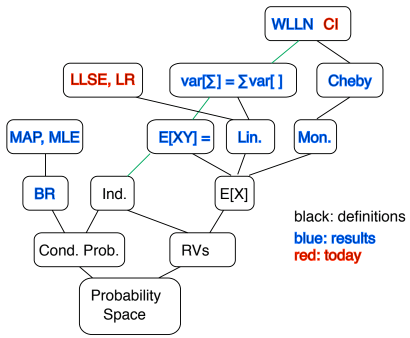
Review: Probability Ideas Map



Review: Probability Ideas Map - Details



Review: Probability Ideas Map - Today



Confidence Intervals: Example

- Flip a coin n times. Let A_n be the fraction of Hs.
- We know that $p := Pr[H] \approx A_n$ for n large (WLLN).
- Can we find a such that $Pr[p \in [A_n - a, A_n + a]] \geq 95\%$?
- If so, we say that

  $[A_n - a, A_n + a]$ is a 95%- Confidence Interval for p.

Using Chebyshev, we will see that $a = 2.25\frac{1}{\sqrt{n}}$ works. Thus

$$[A_n - \frac{2.25}{\sqrt{n}}, A_n + \frac{2.25}{\sqrt{n}}] \text{ is a 95\%-CI for } p.$$

Example: If $n = 1500$, then $Pr[p \in [A_n - 0.06, A_n + 0.06]] \geq 95\%$.

In fact, we will see later that $a = \frac{1}{\sqrt{n}}$ works, so that with $n = 1,500$ one has $Pr[p \in [A_n - 0.03, A_n + 0.03]] \geq 95\%$.

Confidence Intervals: Result

Theorem:
Let X_n be i.i.d. with mean μ and variance σ^2.
Define $A_n = \frac{X_1 + \cdots + X_n}{n}$. Then,

$$Pr[\mu \in [A_n - 4.5\frac{\sigma}{\sqrt{n}}, A_n + 4.5\frac{\sigma}{\sqrt{n}}]] \geq 95\%.$$

Thus, $[A_n - 4.5\frac{\sigma}{\sqrt{n}}, A_n + 4.5\frac{\sigma}{\sqrt{n}}]$ is a 95%-CI for μ.

Example: Let $X_n = 1\{$ coin n yields $H\}$. Then

$$\mu = E[X_n] = p := Pr[H]. \text{ Also, } \sigma^2 = var(X_n) = p(1-p) \leq \frac{1}{4}.$$

Hence, $[A_n - 4.5\frac{1/2}{\sqrt{n}}, A_n + 4.5\frac{1/2}{\sqrt{n}}]$ is a 95%-CI for p.

Confidence Interval: Analysis

Proof:
We prove the theorem, i.e., that $A_n \pm 4.5\sigma/\sqrt{n}$ is a 95%-CI for μ.

From Chebyshev:

$$Pr[|A_n - \mu| \geq 4.5\sigma/\sqrt{n}] \leq \frac{var(A_n)}{[4.5\sigma/\sqrt{n}]^2}$$
$$\leq \frac{\sigma^2/n}{20\sigma^2/n} = 5\%.$$

Thus,
$$Pr[|A_n - \mu| \leq 4.5\sigma/\sqrt{n}] \geq 95\%.$$

Hence,
$$Pr[\mu \in [A_n - 4.5\sigma/\sqrt{n}, A_n + 4.5\sigma/\sqrt{n}]] \geq 95\%.$$

$\square$

Linear Regression: Preamble

Recall that the best guess about Y, if we know only the distribution of Y, is $E[Y]$.

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

Let's review one proof of that fact.

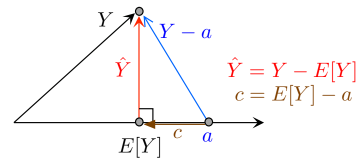
Let $\hat{Y} := Y - E[Y]$. Then, $E[\hat{Y}] = 0$. So, $E[\hat{Y}c] = 0, \forall c$. Now,

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

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

Linear Regression: Preamble

Here is a picture that summarizes the calculation.



$$E[\hat{Y}c] = 0 \Leftrightarrow \hat{Y} \perp c$$
$$\begin{aligned}
E[(Y-a)^2] &= E[(\hat{Y} + c)^2] \\
&= E[\hat{Y}^2 + 2c\hat{Y} + c^2] \\
&= E[\hat{Y}^2] + c^2
\end{aligned}$$
(Pythagoras)

Linear Regression: 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 simplest function $g(X)$ is a constant that does not depend of X.

The next simplest function is linear: $g(X) = a + bX$.

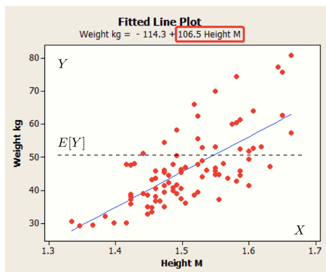
What is the best linear function? That is our next topic.

A bit later, we will consider a general function $g(X)$.

Linear Regression: Motivation

Example 1: 100 people.

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



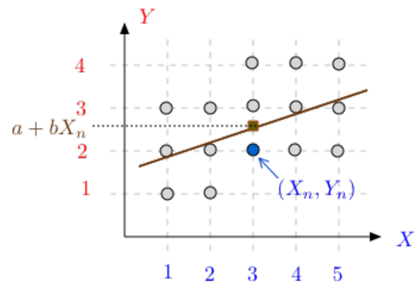
The blue line is $Y = -114.3 + 106.5X$. (X in meters, Y in kg.)

Best linear fit: Linear Regression.

Motivation

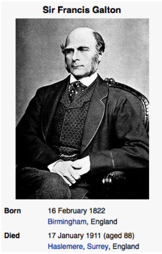
Example 2: 15 people.

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.

History



Galton produced over 340 papers and books. He created the statistical concept of correlation.

In an effort to reach a wider audience, Galton worked on a novel entitled Kantsaywhere. The novel described a utopia organized by a eugenic religion, designed to breed fitter and smarter humans.

The lesson is that smart people can also be stupid.

Covariance

Definition The covariance of X and Y is

$$cov(X,Y) := E[(X - E[X])(Y - E[Y])].$$

Fact

$$cov(X,Y) = E[XY] - E[X]E[Y].$$

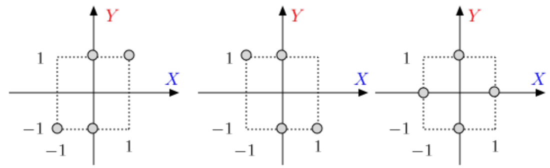
Proof:

$$\begin{aligned} E[(X-E[X])(Y-E[Y])] &= E[XY - E[X]Y - XE[Y] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]. \end{aligned}$$

□

Examples of Covariance

Four equally likely pairs of values



$cov(X,Y) = 1/2$ $\quad$ $cov(X,Y) = -1/2$ $\quad$ $cov(X,Y) = 0$

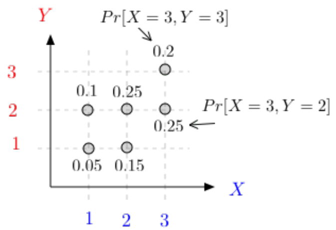
Note that $E[X] = 0$ and $E[Y] = 0$ in these examples. Then $cov(X,Y) = E[XY]$.

When $cov(X,Y) > 0$, the RVs X and Y tend to be large or small together. X and Y are said to be positively correlated.

When $cov(X,Y) < 0$, when X is larger, Y tends to be smaller. X and Y are said to be negatively correlated.

When $cov(X,Y) = 0$, we say that X and Y are uncorrelated.

Examples of Covariance



$$E[X] = 1 \times 0.15 + 2 \times 0.4 + 3 \times 0.45 = 1.9$$
$$E[X^2] = 1^2 \times 0.15 + 2^2 \times 0.4 + 3^2 \times 0.45 = 5.8$$
$$E[Y] = 1 \times 0.2 + 2 \times 0.6 + 3 \times 0.2 = 2$$
$$E[XY] = 1 \times 0.05 + 1 \times 2 \times 0.1 + \cdots + 3 \times 3 \times 0.2 = 4.85$$
$$cov(X,Y) = E[XY] - E[X]E[Y] = 1.05$$
$$var[X] = E[X^2] - E[X]^2 = 2.19.$$

Properties of Covariance

$$cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y].$$

Fact

(a) $var[X] = cov(X,X)$

(b) X, Y independent $\Rightarrow cov(X,Y) = 0$

(c) $cov(a + X, b + Y) = cov(X,Y)$

(d) $cov(aX + bY, cU + dV) = ac.cov(X,U) + ad.cov(X,V) + bc.cov(Y,U) + bd.cov(Y,V)$.

Proof:

(a)-(b)-(c) are obvious.

(d) In view of (c), one can subtract the means and assume that the RVs are zero-mean. Then,

$$\begin{aligned} cov(aX + bY, cU + dV) &= E[(aX + bY)(cU + dV)] \\ &= ac.E[XU] + ad.E[XV] + bc.E[YU] + bd.E[YV] \\ &= ac.cov(X,U) + ad.cov(X,V) + bc.cov(Y,U) + bd.cov(Y,V). \end{aligned}$$

□

Linear Regression: Non-Bayesian

Definition

Given the samples $\{(X_n, Y_n), n = 1, \ldots, N\}$, the Linear Regression of Y over X is

$$\hat{Y} = a + bX$$

where (a,b) minimize

$$\sum_{n=1}^{N} (Y_n - a - bX_n)^2.$$

Thus, $\hat{Y}_n = a + bX_n$ is our guess about Y_n given X_n. The squared error is $(Y_n - \hat{Y}_n)^2$. The LR minimizes the sum of the squared errors.

Why the squares and not the absolute values? Main justification: much easier!

Note: This is a non-Bayesian formulation: there is no prior.

Linear Least Squares Estimate

Definition
Given two RVs X and Y with known distribution $Pr[X = x, Y = y]$, the Linear Least Squares Estimate of Y given X is

$$\hat{Y} = a + bX =: L[Y|X]$$

where (a, b) minimize

$$g(a, b) := E[(Y - a - bX)^2].$$

Thus, $\hat{Y} = a + bX$ is our guess about Y given X. The squared error is $(Y - \hat{Y})^2$. The LLSE minimizes the expected value of the squared error.

Why the squares and not the absolute values? Main justification: much easier!

Note: This is a Bayesian formulation: there is a prior.

LR: Non-Bayesian or Uniform?

Observe that

$$\frac{1}{N}\sum_{n=1}^{N}(Y_n - a - bX_n)^2 = E[(Y - a - bX)^2]$$

where one assumes that

$$(X, Y) = (X_n, Y_n), \text{ w.p. } \frac{1}{N} \text{ for } n = 1, \ldots, N.$$

That is, the non-Bayesian LR is equivalent to the Bayesian LLSE that assumes that (X, Y) is uniform on the set of observed samples.

Thus, we can study the two cases LR and LLSE in one shot.

However, the interpretations are different!

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

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

Hence, by combining the two brown equalities, $E[(Y - \hat{Y})(c + dX)] = 0$. Then, $E[(Y - \hat{Y})(\hat{Y} - a - bX)] = 0, \forall a, b$.
Indeed: $\hat{Y} = \alpha + \beta X$ for some α, β, so that $\hat{Y} - a - bX = c + dX$ for some c, d. Now,

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

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

Note that

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

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

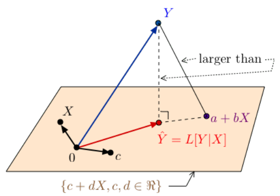
Now,

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

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

A picture

The following picture explains the algebra:



We saw that $E[Y - \hat{Y}] = 0$. In the picture, this says that $Y - \hat{Y} \perp c$, for any c.

We also saw that $E[(Y - \hat{Y})X] = 0$. In the picture, this says that $Y - \hat{Y} \perp X$.

Hence, $Y - \hat{Y}$ is orthogonal to the plane $\{c + dX, c, d \in \Re\}$.

Consequently, $Y - \hat{Y} \perp \hat{Y} - a - bX$. Pythagoras then says that Y is closer to $\hat{Y}$ than $a + bX$.

That is, $\hat{Y}$ is the projection of Y onto the plane.

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 2:
First assume that $E[X] = 0$ and $E[Y] = 0$. Then,

$$\begin{aligned} g(a, b) &:= E[(Y - a - bX)^2] \\ &= E[Y^2 + a^2 + b^2X^2 - 2aY - 2bXY + 2abX] \\ &= a^2 + E[Y^2] + b^2E[X^2] - 2aE[Y] - 2bE[XY] + 2abE[X] \\ &= a^2 + E[Y^2] + b^2E[X^2] - 2bE[XY]. \end{aligned}$$

We set the derivatives of g w.r.t. a and b equal to zero.

$$0 = \frac{\partial}{\partial a} g(a, b) = 2a \Rightarrow a = 0.$$

$$0 = \frac{\partial}{\partial b} g(a, b) = 2bE[X^2] - 2E[XY]$$

$$\Rightarrow b = E[XY]/E[X^2] = cov(X, Y)/var(X).$$

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 2:
In the general case (i.e., when $E[X]$ and $E[Y]$ may be nonzero),

$$\begin{aligned} Y - a - bX &= Y - E[Y] - (a - E[Y]) - b(X - E[X]) + bE[X] \\ &= Y - E[Y] - (a - E[Y] + bE[X]) - b(X - E[X]) \\ &= Y - E[Y] - c - b(X - E[X]) \end{aligned}$$

with $c = a - E[Y] + bE[X]$.
From the first part, we know that the best values of c and b are

$c = 0$ and $b = cov(X - E[X], Y - E[Y])/var(X - E[X]) = cov(X,Y)/var(X)$.

Thus, $0 = c = a - E[Y] + bE[X]$, so that $a = E[Y] - bE[X]$. Hence,

$$\begin{aligned} a + bX &= E[Y] - bE[X] + bX = E[Y] + b(X - E[X]) \\ &= E[Y] + \frac{cov(X,Y)}{var(X)}(X - E[X]). \end{aligned}$$

$\square$

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? That is, what is the mean squared estimation error?

We find

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

Without observations, the estimate is $E[Y] = 0$. The error is $var(Y)$. Observing X reduces the error.

Estimation Error: A Picture

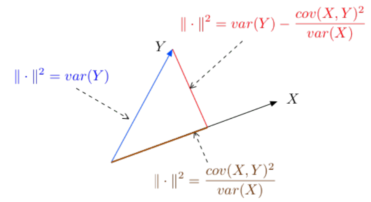
We saw that

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

and

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

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



LLSE and LR

Consider the non-Bayesian case: sample $(X_1, Y_1), \ldots, (X_K, Y_K)$.

Then,

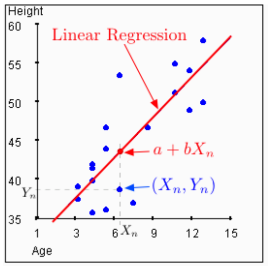
$$L[Y|X] = E[Y] + \frac{cov(X,Y)}{var(X)}(X - E[X]).$$

Here,

$$\begin{aligned} E[X] &= \frac{1}{K}\sum_{k=1}^{K} X_k \\ E[Y] &= \frac{1}{K}\sum_{k=1}^{K} Y_k \\ E[X^2] &= \frac{1}{K}\sum_{k=1}^{K} X_k^2 \\ E[XY] &= \frac{1}{K}\sum_{k=1}^{K} X_k Y_k \\ cov(X,Y) &= E[XY] - E[X]E[Y] \\ var(X) &= E[X^2] - E[X]^2. \end{aligned}$$

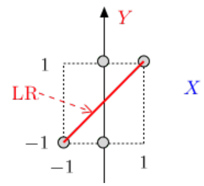
Linear Regression Examples

Example 1:



Linear Regression Examples

Example 2:

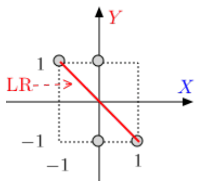


We find:

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

Linear Regression Examples

Example 3:



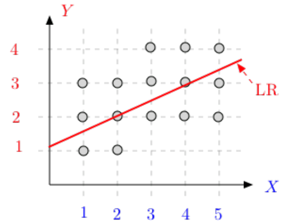
We find:

$E[X]=0; E[Y]=0; E[X^2]=1/2; E[XY]=-1/2;$
$var[X]=E[X^2]-E[X]^2=1/2; cov(X,Y)=E[XY]-E[X]E[Y]=-1/2;$

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

Linear Regression Examples

Example 4:

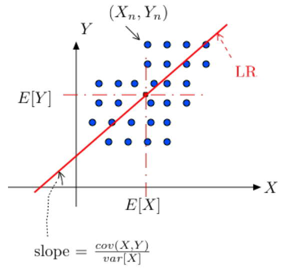


We find:

$E[X]=3; E[Y]=2.5; E[X^2]=(3/15)(1+2^2+3^2+4^2+5^2)=11;$
$E[XY]=(1/15)(1\times 1+1\times 2+\cdots+5\times 4)=8.4;$
$var[X]=11-9=2; cov(X,Y)=8.4-3\times 2.5=0.9;$

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 $\frac{cov(X,Y)}{var(X)}$.

Summary

Confidence Interval; Linear Regression

1. 95%-Confidence Interval for μ: $A_n \pm 4.5\sigma/\sqrt{n}$
2. Linear Regression: $L[Y|X]=E[Y]+\frac{cov(X,Y)}{var(X)}(X-E[X])$
3. Non-Bayesian: minimize $\sum_n (Y_n-a-bX_n)^2$
4. Bayesian: minimize $E[(Y-a-bX)^2]$