CS70: Lecture 33.

Let's Guess!
Dollar or not with equal probability?
Guess how much you get!
Guess a 1/2! The expected value.
Win X, 100 times.
How much will you win the 101st.
Guess average!

Let's Guess!
How much does random person weigh?
Guess the expected value!
How much does professor Rao weigh?
Remember: I am pretty tall!

Knowing that I am tall should you guess he is heavier than expected!

CS70: Lecture 33

Previously: Single variable.
When do you get an accurate measure of a random variable.
Predictor: Expectation.
Accuracy: Variance.
Want to find expectation? Poll.
Sampling: Many trials and average.
Accuracy: Chebyshev. Chernoff.

Today:

What does the value of one variable tell you about another?
Exact: Conditional probability among all events.
Summary: Covariance.
Predictor: Linear function.
Bayesion: Best linear estimator from covariance, and expectations.
Sampling: Linear regression from set of samples.

CS70: Lecture 33

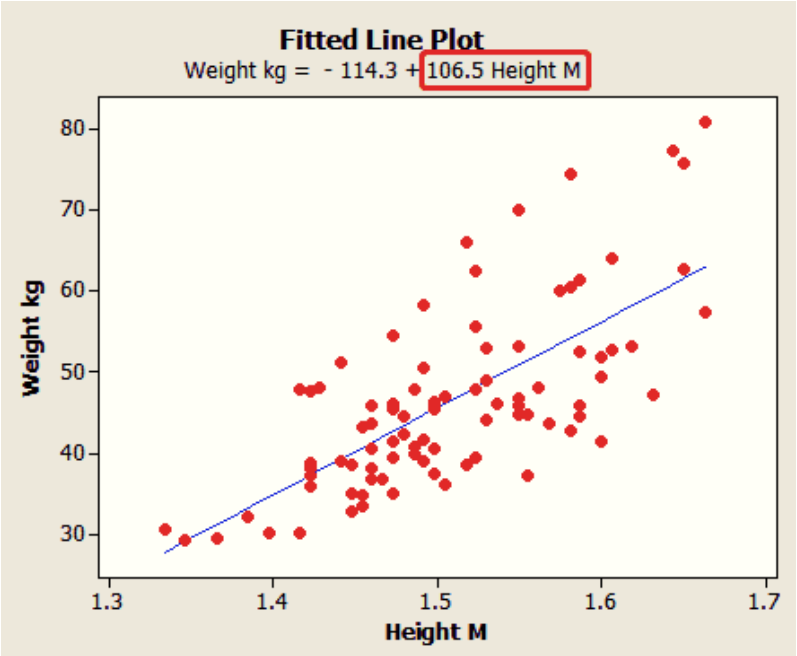
Linear Regression

1. Examples
2. History
3. Multiple Random variables
4. Linear Regression
5. Derivation
6. More examples

Illustrative Example

Example 1: 100 people.

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

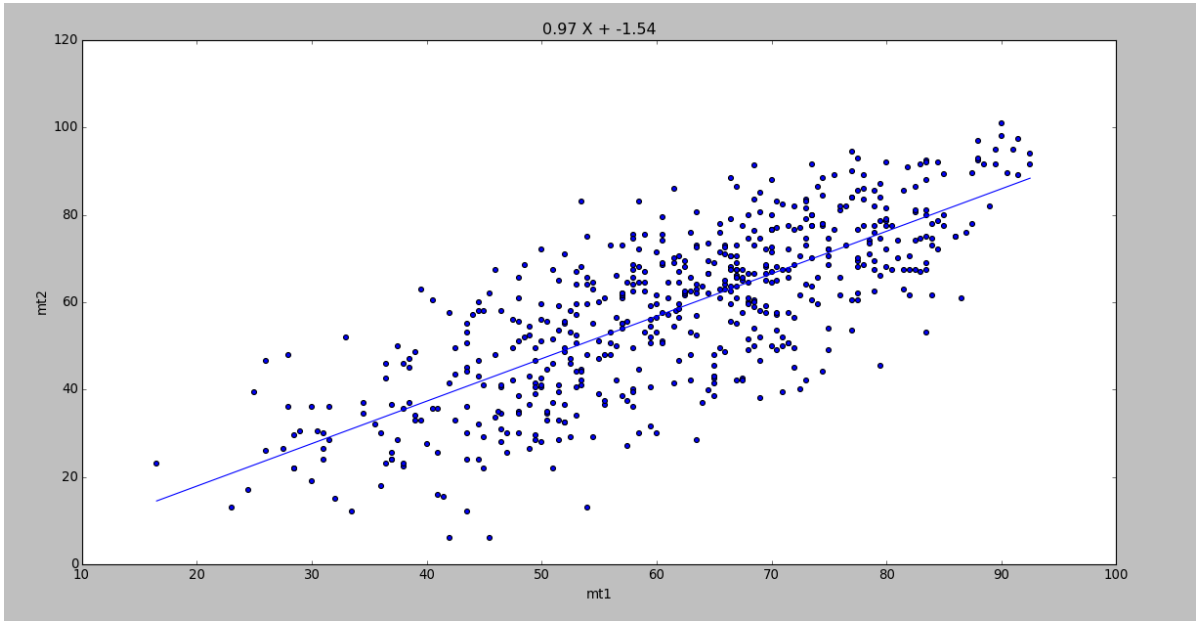


The blue line is $Y = -114.3 + 106.5X$. (X in meters, Y in kg.)

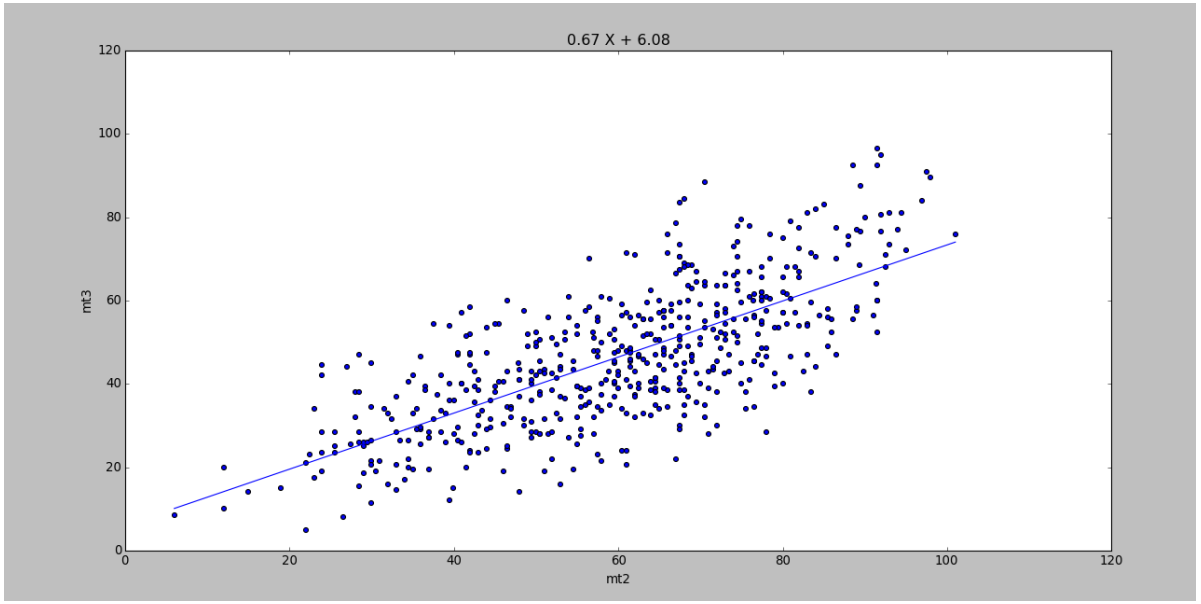
Best linear fit: Linear Regression.

Should you really use a linear function? Cubic, maybe.
Then *logHeight* and log Weight is linear.

Painful Example



Midterm 1 v Midterm 2.
$Y = .97X - 1.54$

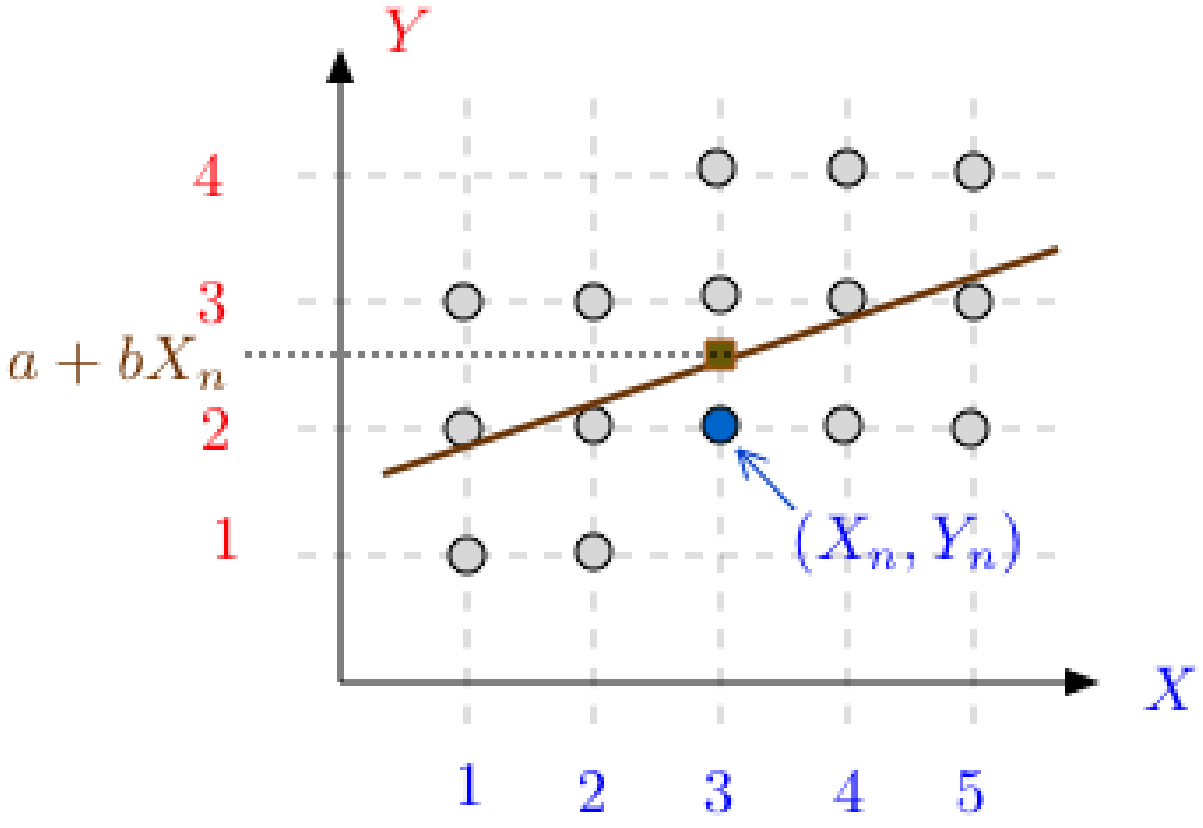


Midterm 2 v Midterm 3
$Y = .67X + 6.08$

Illustrative Example: sample space.

Example 3: 15 people.

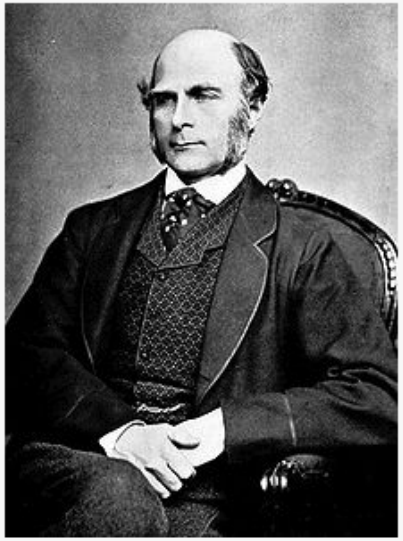
We look at two attributes: (X_n, Y_n) of person n, for $n = 1, \ldots, 15$:



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

History

Sir Francis Galton



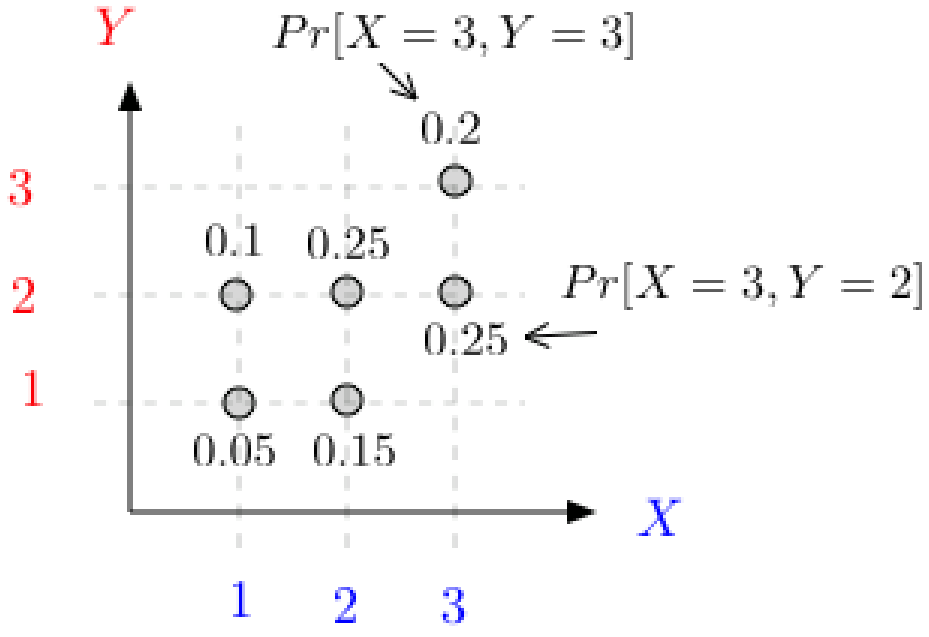
Born	16 February 1822 Birmingham, England
Died	17 January 1911 (aged 88) Haslemere, Surrey, England

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.

Multiple Random Variables



The pair (X, Y) takes 6 different values with the probabilities shown. This figure specifies the joint distribution of X and Y.

Questions: Where is Ω? What are $X(\omega)$ and $Y(\omega)$?

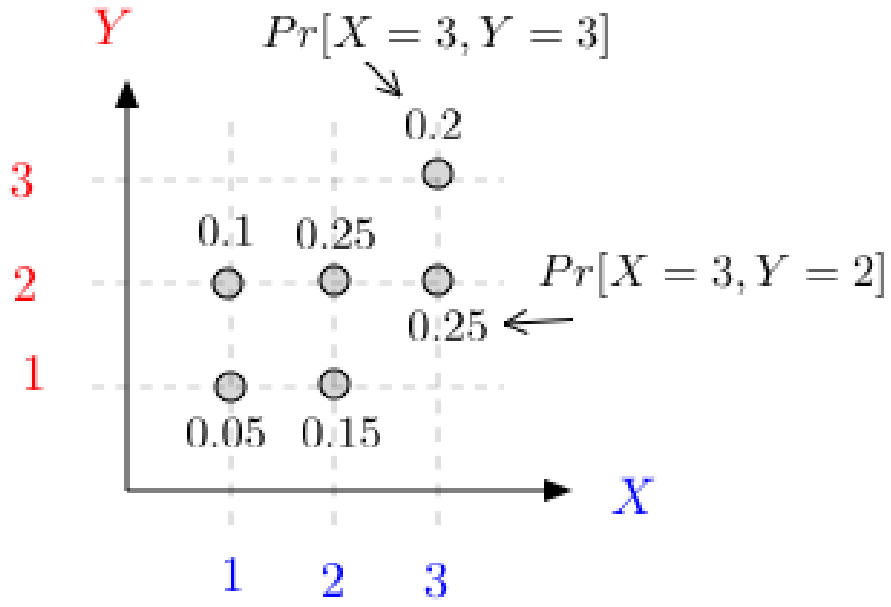
Answer: For instance, let Ω be the set of values of (X, Y) and assign them the corresponding probabilities.
This is the "canonical" probability space.

Definitions

Definitions Let X and Y be RVs on Ω.

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

Marginal and Conditional



- $Pr[X = 1] = 0.05 + 0.1 = 0.15$; $Pr[X = 2] = 0.4$; $Pr[X = 3] = 0.45$.
- This is the marginal distribution of X:
  $Pr[X = x] = \sum_y Pr[X = x, Y = y]$.
- $Pr[Y = 1|X = 1] = Pr[X = 1, Y = 1]/Pr[X = 1] = 0.05/0.15 = 1/3$.
- This is the conditional distribution of Y given $X = 1$:
  $Pr[Y = y|X = x] = Pr[X = x, Y = y]/Pr[X = x]$.

Quick question: Are X and Y independent?

Covariance

d **Definition** The covariance of X and Y is

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

Fact

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

Quick Question: For indpendent X and Y,

$$cov(X,Y) =$$

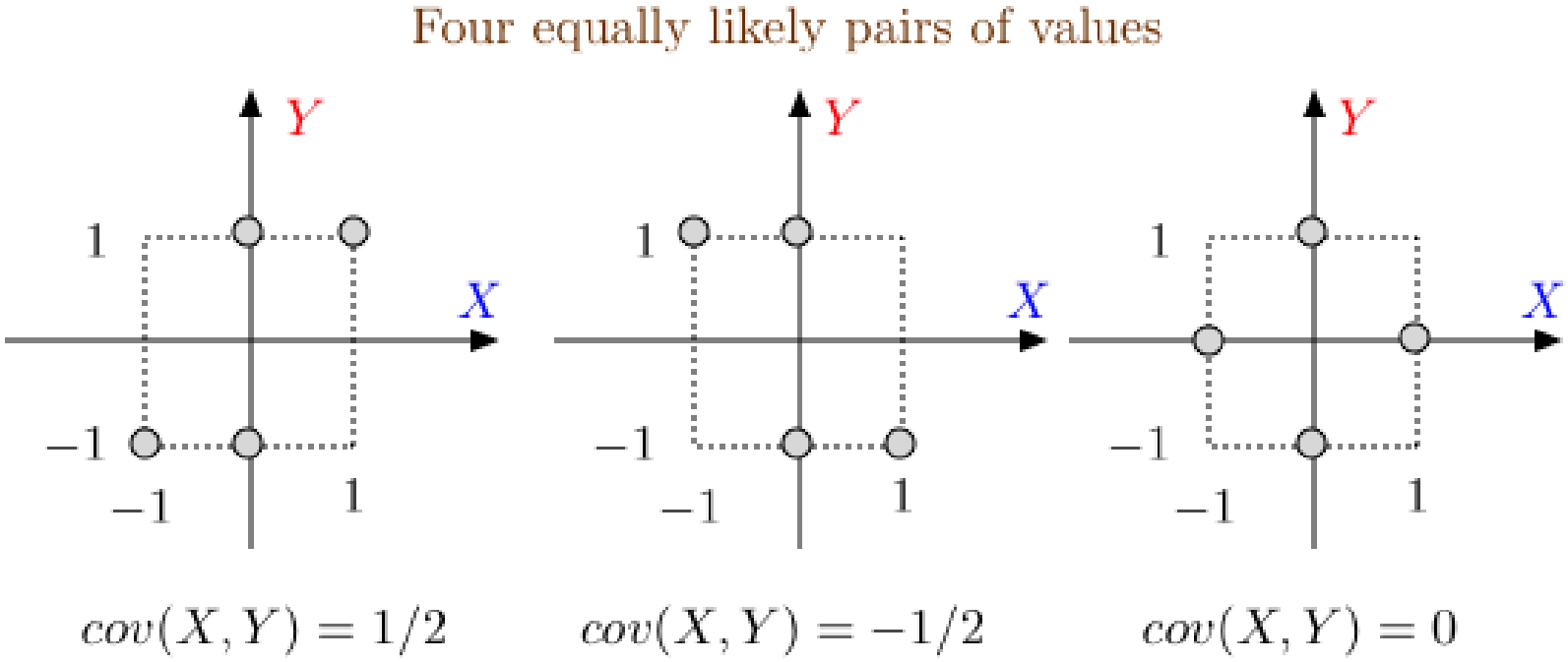
? 1 ? 0?

Proof:

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

$\square$

Examples of Covariance

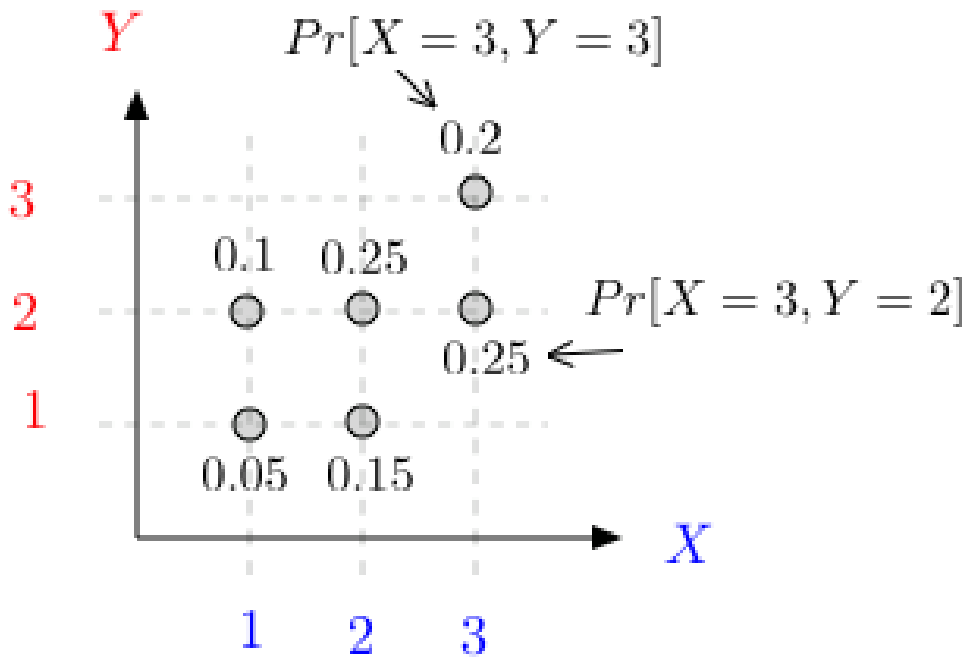


Note that $E[X] = 0$ and $E[Y] = 0$ in these examples. Then $cov(X, Y) = E[XY]$.

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

When $cov(X, Y) < 0$, when X is larger, Y tends to be smaller.

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

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,

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

$\square$

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.

Single Variable: Average minimizes squared distance to sample points.

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

$$E[(Y - a - bX)^2].$$

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.

Single Variable: $E(X)$ minimizes expected squared error.

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,

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

If $cov(X,Y) = 0$, what do you predict for Y? $E(Y)$!
Make sense? Sure.
Independent!

If $cov(X,Y)$ is positive, and $X > E(X)$, is $\hat{Y} \geq E(Y)$? Sure.
Make sense? Sure.
Taller $\rightarrow$ Heavier!

If $cov(X,Y)$ is negative, and $X > E(X)$, is $\hat{Y} \geq E(Y)$? No! $\hat{Y} \leq E(Y)$
Make sense? Sure.
Heavier $\rightarrow$ Slower!

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

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

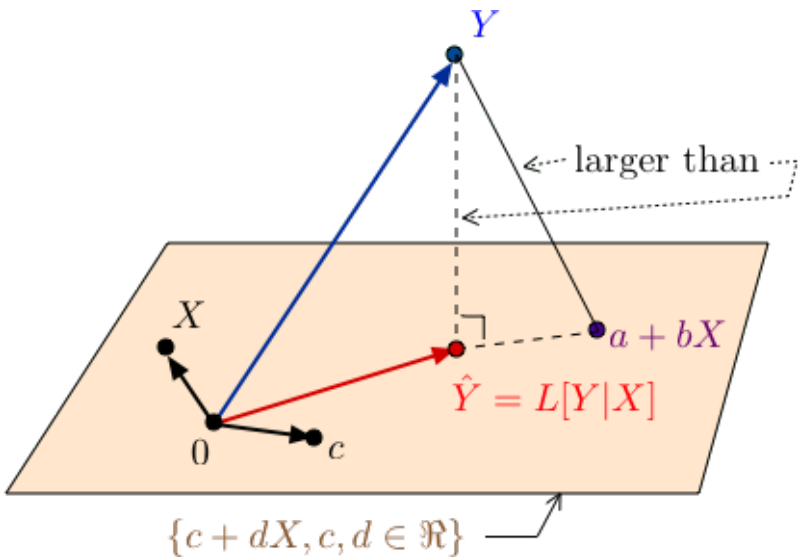
Now,

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

$^{(*)}$ 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:



X, Y vectors where X_i, Y_i is outcome.
c is a constant vector.

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 $\{c + dX, c, d \in \Re\}$.

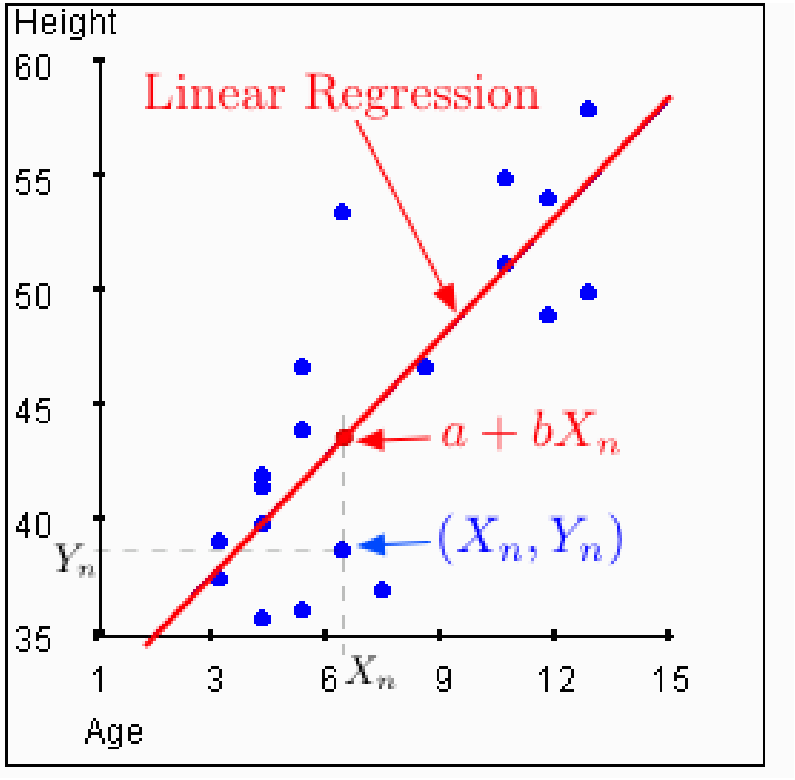
Consequently, $Y - \hat{Y} \perp \hat{Y} - a - bX$. Pythagoras then says that $\hat{Y}$ is closer to Y than $a + bX$.

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

Note: this picture corresponds to uniform probability space.

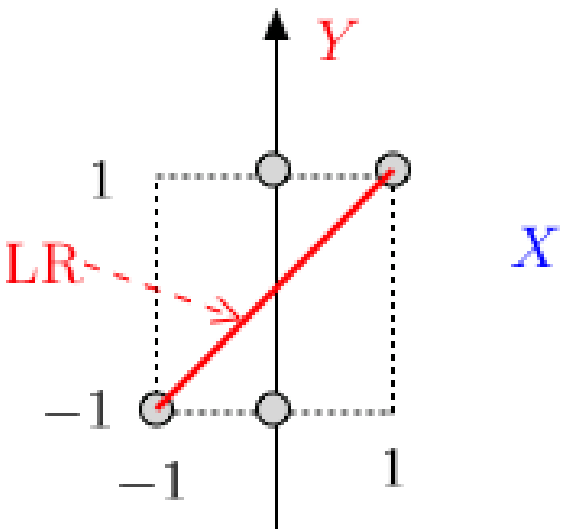
Linear Regression Examples

Example 1:



Linear Regression Examples

Example 2:

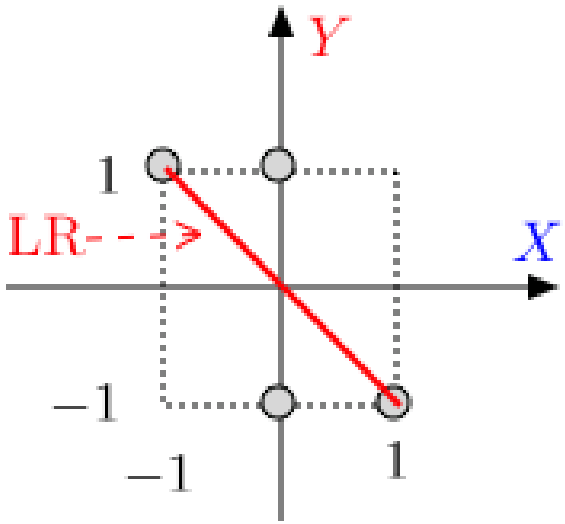


We find:

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

Linear Regression Examples

Example 3:

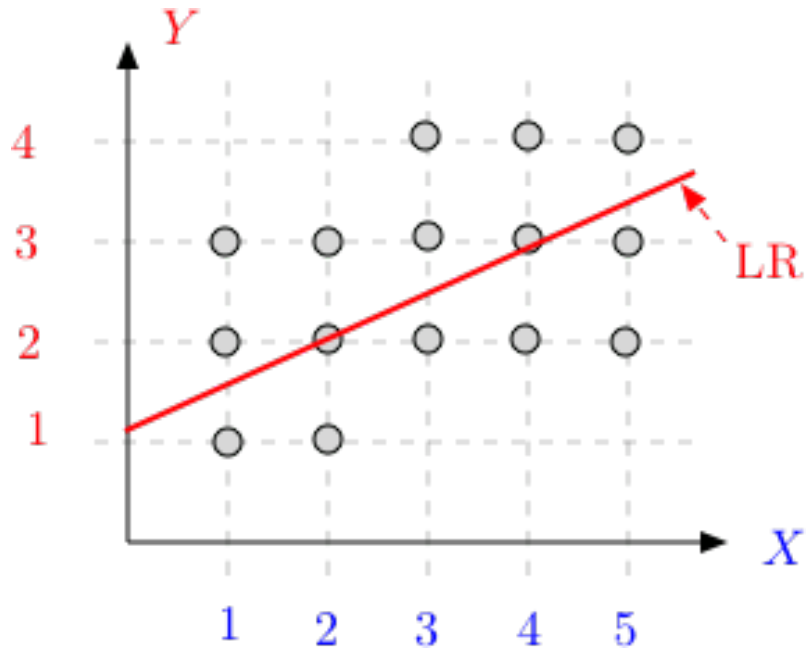


We find:

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

Linear Regression Examples

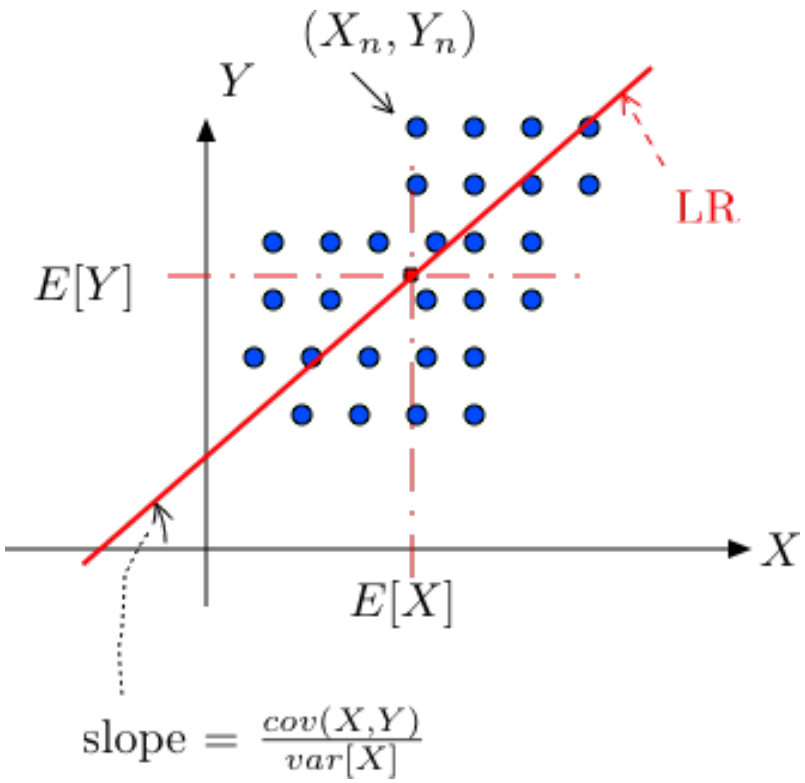
Example 4:



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

Summary

$\boxed{\text{Linear Regression}}$

1. Multiple Random variables: X, Y with $Pr[X = x, Y = y]$.
2. Marginal & conditional probabilities
3. Linear Regression: $L[Y|X] = E[Y] + \frac{cov(X,Y)}{var(X)}(X - E[X])$
4. Non-Bayesian: minimize $\sum_n (Y_n - a - bX_n)^2$
5. Bayesian: minimize $E[(Y - a - bX)^2]$