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!
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.
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 \((X_n, Y_n) = \text{(height, weight)}\) 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 \(\log\text{Height}\) and logWeight is linear.
Painful Example

Midterm 1 v Midterm 2.
\[ Y = 0.97X - 1.54 \]

Midterm 2 v Midterm 3
\[ Y = 0.67X + 6.08 \]
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.
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 \(\Omega\)? What are \(X(\omega)\) and \(Y(\omega)\)?

Answer: For instance, let \(\Omega\) be the set of values of \((X, Y)\) and assign them the corresponding probabilities. This is the “canonical” probability space.
Definitions Let $X$ and $Y$ be RVs on $\Omega$.

- **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

**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 independent $X$ and $Y$,

$$cov(X, Y) = \ ? 1 \ ? 0?$$

**Proof:**


$$= E[XY] - E[X]E[Y].$$
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

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


**Fact**
(a) \( \text{var}[X] = \text{cov}(X, X) \)
(b) \( X, Y \) independent \( \Rightarrow \text{cov}(X, Y) = 0 \)
(c) \( \text{cov}(a + X, b + Y) = \text{cov}(X, Y) \)
(d) \( \text{cov}(aX + bY, cU + dV) = ac \cdot \text{cov}(X, U) + ad \cdot \text{cov}(X, V) + bc \cdot \text{cov}(Y, U) + bd \cdot \text{cov}(Y, V) \).

**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,

\[ \text{cov}(aX + bY, cU + dV) = E[(aX + bY)(cU + dV)] \]
\[ = ac \cdot E[XU] + ad \cdot E[XV] + bc \cdot E[YU] + bd \cdot E[YV] \]
\[ = ac \cdot \text{cov}(X, U) + ad \cdot \text{cov}(X, V) + bc \cdot \text{cov}(Y, U) + bd \cdot \text{cov}(Y, V). \]
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!
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!
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 $\alpha, \beta$, so that $\hat{Y} - a - bX = c + dX$ for some $c, d$. 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.
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 \( E[(Y - \hat{Y})X] = 0. \)

Note that

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

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

Now,

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

\[ = (\ast) \text{cov}(X, Y) - \frac{\text{cov}(X, Y)}{\text{var}[X]} \text{var}[X] = 0. \]

\((\ast)\) Recall that \( \text{cov}(X, Y) = E[(X - E[X])(Y - E[Y])] \) and \( \text{var}[X] = E[(X - E[X])^2]. \)
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 $\{c + dX, c, d \in \mathbb{R}\}$.

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.

$X, Y$ vectors where $X_i, Y_i$ is outcome.
$c$ is a constant vector.
Example 1:
Linear Regression Examples

Example 2:

We find:

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

Example 3:

We find:

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

Example 4:

We find:

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

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

Linear Regression

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