Linear Regression
Linear Regression

1. Preamble
2. Motivation for LR
3. History of LR
4. Linear Regression
5. Derivation
6. More examples
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 $a = E[Y]$.

Proof: Let $\hat{Y} := Y - E[Y]$. Then, $E[\hat{Y}] = 0$.

So, $E[\hat{Y}^2] = 0$, $\forall c$.


with $c = E[Y] - a = E[\hat{Y}]^2 + 2E[\hat{Y} c] + c^2$

$= E[\hat{Y}^2] + 0 + c^2$

$\geq E[\hat{Y}^2]$.

Hence, $E[(Y - a)^2] \geq E[(Y - E[Y])^2]$, $\forall a$. 
The best guess about $Y$, 

$$
E[Y] = a = E[Y].
$$

Proof:

Let $\hat{Y} = Y - E[Y]$. Then,

$$
E[\hat{Y}] = 0.
$$

So,

$$
E[\hat{Y}^c] = 0, \forall c.
$$

Now,

$$
$$

Hence,

$$
E[(Y - a)^2] \geq E[(Y - E[Y])^2], \forall a.
$$
The best guess about $Y$, if we know only the distribution of $Y$, is $\mathbb{E}[Y]$. More precisely, the value of $a$ that minimizes $\mathbb{E}[(Y - a)^2]$ is $a = \mathbb{E}[Y]$. Proof: Let $\hat{Y} = Y - \mathbb{E}[Y]$. Then, $\mathbb{E}[\hat{Y}] = 0$. So, $\mathbb{E}[\hat{Y}^2] = 0$, $\forall c$. Now, $\mathbb{E}[(Y - a)^2] = \mathbb{E}[(Y - \mathbb{E}[Y] + \mathbb{E}[Y] - a)^2] = \mathbb{E}[\hat{Y}^2 + 2\hat{Y}c + c^2] = \mathbb{E}[\hat{Y}^2] + 2\mathbb{E}[\hat{Y}c] + c^2 \geq \mathbb{E}[\hat{Y}^2]$. Hence, $\mathbb{E}[(Y - a)^2] \geq \mathbb{E}[(Y - \mathbb{E}[Y])^2], \forall a$. 
The best guess about $Y$, if we know only the distribution of $Y$, is $E[Y]$. 
Linear Regression: Preamble

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
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 $a = E[Y]$. 

Proof:


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 $a = E[Y]$.

**Proof:**
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 $a = E[Y]$.

**Proof:**
Let $\hat{Y} := Y - E[Y]$. 
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 $a = E[Y]$.

**Proof:**

Let $\hat{Y} := Y - E[Y]$. Then, $E[\hat{Y}] = 0$. 
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 $a = E[Y]$.

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

**Proof:**

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

Linear Regression: Preamble

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 $a = E[Y]$.

**Proof:**

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

$$= E[(\hat{Y} + c)^2]$$
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 $a = E[Y]$.

**Proof:**

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

\[
\]
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 $a = E[Y]$.

**Proof:**

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

\[
\]
Linear Regression: Preamble

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 $a = E[Y]$. 

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


$$= E[(\hat{Y} + c)^2] \text{ with } c = E[Y] - a$$

$$= E[\hat{Y}^2 + 2\hat{Y}c + c^2] = E[\hat{Y}^2] + 2E[\hat{Y}c] + c^2$$
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 $a = E[Y]$.

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

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

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

\[
\]

\[
= E[\hat{Y}^2 + 2\hat{Y}c + c^2] = E[\hat{Y}^2] + 2E[\hat{Y}c] + c^2
\]

\[
= E[\hat{Y}^2] + 0 + c^2 \geq E[\hat{Y}^2].
\]
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 $a = E[Y]$.

**Proof:**

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

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

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

□
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: Preamble
Thus, if we want to guess the value of $Y$, we choose $E[Y]$. 

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

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 $Y$. 
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 $Y$?
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 $Y$. 
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 $X$. 
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: $g(X) = a + bX$. 
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?
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.
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 on $X$.

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 \((X_n, Y_n) = (\text{height, weight})\) of person \(n\), for \(n = 1, \ldots, 100\):

\[
E[Y] = -114.3 + 106.5X.
\]  

(\(X\) in meters, \(Y\) in kg.) Best linear fit: Linear Regression.
Example 1: 100 people.
Example 1: 100 people.
Let \((X_n, Y_n) = (\text{height, weight})\) of person \(n\), for \(n = 1, \ldots, 100\):
Linear Regression: Motivation

Example 1: 100 people.
Let \((X_n, Y_n) = \text{(height, weight)}\) of person \(n\), for \(n = 1, \ldots, 100\):
Linear Regression: Motivation

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.)
Linear Regression: Motivation

Example 1: 100 people.
Let \((X_n, Y_n) = \text{height, weight}\) of person \(n\), for \(n = 1, \ldots, 100\):

\[
E[Y] = -114.3 + 106.5X. \quad (X \text{ in meters, } Y \text{ in kg.})
\]

Best linear fit: Linear Regression.
Motivation

Example 2: 15 people.
Motivation

Example 2: 15 people.

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

Example 2: 15 people.

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

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

$$\text{cov}(X, Y) := E[(X - E[X])(Y - E[Y])].$$
Covariance

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

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

**Fact**

$$\text{cov}(X, Y) = E[XY] - E[X]E[Y].$$
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:**

**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:**

$$
$$
Covariance

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

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

**Fact**

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

**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. $X$ 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.

Four equally likely pairs of values

\[ cov(X, Y) = \frac{1}{2} \] \[ cov(X, Y) = -\frac{1}{2} \] \[ cov(X, Y) = 0 \]
Examples of Covariance

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

Four equally likely pairs of values

$\text{cov}(X, Y) = 1/2$  $\text{cov}(X, Y) = -1/2$  $\text{cov}(X, Y) = 0$
Examples of Covariance

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

When $\text{cov}(X, Y) > 0$, the RVs $X$ and $Y$ tend to be large or small together.
Examples of Covariance

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

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

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

When $\text{cov}(X, Y) = 0$, we say that $X$ and $Y$ are uncorrelated.
Examples of Covariance

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

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

When $\text{cov}(X, Y) < 0$, when $X$ is larger, $Y$ tends to be smaller.
Notes that $E[X] = 0$ and $E[Y] = 0$ in these examples. Then $\text{cov}(X, Y) = E[XY]$.

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

When $\text{cov}(X, Y) < 0$, when $X$ is larger, $Y$ tends to be smaller. $X$ and $Y$ are said to be negatively correlated.
Examples of Covariance

Four equally likely pairs of values

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

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

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

When $\text{cov}(X, Y) = 0$, we say that $X$ and $Y$ are uncorrelated.
Examples of Covariance

\[ E[X] = 1 \times 0.15 + 2 \times 0.4 + 3 \times 0.45 = 1.9 \]

\[ E[X^2] = 1^2 \times 0.15 + 2^2 \times 0.4 + 3^2 \times 0.45 = 5.8 \]

\[ E[Y] = 1 \times 0.2 + 2 \times 0.6 + 3 \times 0.2 = 2 \]

\[ E[XY] = \sum \text{xypr} = 1 \times 0.05 + 2 \times 0.25 + 3 \times 0.25 + 4 \times 0.15 + 5 \times 0.25 = 4.85 \]

\[ \text{cov}(X,Y) = E[XY] - E[X]E[Y] = 4.85 - 1.9 \times 2 = 2.1 \]

\[ \text{var}(X) = E[X^2] - (E[X])^2 = 5.8 - 1.9^2 = 2.19 \]
Examples of Covariance

\[ E[X] = 1 \times 0.15 + 2 \times 0.4 + 3 \times 0.45 = 1.9 \]
Examples of Covariance

\[ E[X] = 1 \times 0.15 + 2 \times 0.4 + 3 \times 0.45 = 1.9 \]
\[ E[X^2] = 1^2 \times 0.15 + 2^2 \times 0.4 + 3^2 \times 0.45 = 5.8 \]
Examples of Covariance

\[ E[X] = 1 \times 0.15 + 2 \times 0.4 + 3 \times 0.45 = 1.9 \]
\[ E[X^2] = 1^2 \times 0.15 + 2^2 \times 0.4 + 3^2 \times 0.45 = 5.8 \]
\[ E[Y] = 1 \times 0.2 + 2 \times 0.6 + 3 \times 0.2 = 2 \]
Examples of Covariance

\[ E[X] = 1 \times 0.15 + 2 \times 0.4 + 3 \times 0.45 = 1.9 \]

\[ E[X^2] = 1^2 \times 0.15 + 2^2 \times 0.4 + 3^2 \times 0.45 = 5.8 \]

\[ E[Y] = 1 \times 0.2 + 2 \times 0.6 + 3 \times 0.2 = 2 \]

\[ E[XY] = 1 \times 0.05 + 1 \times 2 \times 0.1 + \ldots + 3 \times 3 \times 0.2 = 4.85 \]
Examples of Covariance

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

Properties of Covariance

\[
\]

Fact
(a) \(\text{var}[X] = \text{cov}(X, X)\)
Properties of Covariance


**Fact**
(a) \( \text{var}[X] = \text{cov}(X, X) \)
(b) \( X, Y \) independent \( \Rightarrow \) \( \text{cov}(X, Y) = \)
Properties of Covariance


Fact
(a) \( \text{var}[X] = \text{cov}(X, X) \)
(b) \( X, Y \) independent \( \Rightarrow \) \( \text{cov}(X, Y) = 0 \)
Properties of Covariance


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

**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,
\[ \text{cov}(aX + bY, cU + dV) = E[(aX + bY)(cU + dV)] \]
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,

\[
\text{cov}(aX + bY, cU + dV) = E[(aX + bY)(cU + dV)]
= ac \cdot E[UX] + ad \cdot E[XV] + bc \cdot E[UY] + bd \cdot E[YV]
\]
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) \( \text{cov}(aX + bY, cU + dV) = ac.\text{cov}(X, U) + ad.\text{cov}(X, V) \)
\[ + bc.\text{cov}(Y, U) + bd.\text{cov}(Y, V). \]

**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,
\[ \text{cov}(aX + bY, cU + dV) = E[(aX + bY)(cU + dV)] \]
\[ = ac.\text{cov}(X, U) + ad.\text{cov}(X, V) + bc.\text{cov}(Y, U) + bd.\text{cov}(Y, V). \]
Linear Regression: Non-Bayesian

Definition
Given the samples \( \{(X_n, Y_n), n = 1, \ldots, N\} \),
Linear Regression: Non-Bayesian

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

\[
\hat{Y} = a + bX
\]
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

\[ \sum_{n=1}^{N} (Y_n - a - bX_n)^2. \]
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 \( X_n \).
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 \( (Y_n - \hat{Y}_n)^2 \).
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.
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?
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!
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.
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.
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.
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

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

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.
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

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

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

Note: This is a Bayesian formulation: there is a 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 $(Y - \hat{Y})^2$. 
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.
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.
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!
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

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

where one assumes that \((X, Y) = (X_n, Y_n), \) w.p. 1 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!
Observe that

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

where one assumes that

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

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

Proof 1:

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

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.
Theorem

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

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

Proof 1:

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

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

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

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

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

$$E[(Y - \hat{Y})(c + dX)] = 0.$$
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:
$Y - \hat{Y} = (Y - E[Y]) - \frac{\text{cov}(X, Y)}{\text{var}(X)}(X - E[X])$. 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, $E[(Y - \hat{Y})(\hat{Y} - a - bX)] = 0, \forall a, b$. 

**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 $\alpha, \beta$,
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:
$Y - \hat{Y} = (Y - E[Y]) - \frac{\text{cov}(X, Y)}{\text{var}[X]}(X - E[X])$. 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 $c, d$. 

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 $\alpha, \beta$, so that $\hat{Y} - a - bX = c + dX$ for some $c, d$. Now,

$$E[(Y - a - bX)^2] = E[(Y - \hat{Y} + \hat{Y} - a - bX)^2]$$
**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 $\alpha, \beta$, so that $\hat{Y} - a - bX = c + dX$ for some $c, d$. Now,

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

$$= E[(Y - \hat{Y})^2] + E[(\hat{Y} - a - bX)^2] + 0$$
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:
\[
Y - \hat{Y} = (Y - E[Y]) - \frac{\text{cov}(X, Y)}{\text{var}[X]}(X - E[X]).
\]
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,
\[
E[(Y - a - bX)^2] = E[(Y - \hat{Y} + \hat{Y} - a - bX)^2] \\
= E[(Y - \hat{Y})^2] + E[(\hat{Y} - a - bX)^2] + 0 \geq E[(Y - \hat{Y})^2].
\]
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:
$Y - \hat{Y} = (Y - E[Y]) - \frac{\text{cov}(X, Y)}{\text{var}(X)}(X - E[X])$. 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 $(a, b)$. 
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 $\alpha, \beta$, 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

\[ Y - \hat{Y} = (Y - E[Y]) - \frac{\text{cov}(X,Y)}{\text{var}[X]} (X - E[X]). \]
A Bit of Algebra

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

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

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

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

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

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

\(*\) Recall that \( \text{cov}(X, Y) = E[(X - E[X])(Y - E[Y])] \) and \( \text{var}[X] = E[(X - E[X])^2] \).
We saw that the LLSE of $Y$ given $X$ is

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

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

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

How good is this estimator?
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?

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

$$E[(Y - L[Y|X])^2] = E[(Y - E[Y] - (cov(X, Y)/var(X))(X - E[X]))^2]$$
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

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

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


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

$$+ (cov(X, Y)/var(X))^2 E[(X - E[X])^2]$$

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


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

$$+ (cov(X, Y)/var(X))^2 E[(X - E[X])^2]$$

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

Without observations, the estimate is $E[Y] = 0$. 
Estimation Error

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

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

How good is this estimator? That is, what is the mean squared estimation error?

We find

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

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

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

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

Without observations, the estimate is $E[\,Y\,] = 0$. The error is $\text{var}(\,Y\,)$. 
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


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

$$+ (cov(X, Y)/var(X))^2 E[(X - E[X])^2]$$

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

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

We saw that

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

and
\[ E[|Y - L[Y|X]|^2] = \text{var}(Y) - \frac{\text{cov}(X, Y)^2}{\text{var}(X)}. \]
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 \( E[X] = 0, E[Y] = 0 \):
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 \( E[X] = 0, E[Y] = 0 \):
Linear Regression Examples

Example 1:
Example 1:
Linear Regression Examples

Example 2:

\[
E[X] = 0; \\
E[Y] = 0; \\
E[X^2] = 1/2; \\
E[XY] = 1/2; \\
\text{var}[X] = E[X^2] - E[X]^2 = 1/2; \\
\text{cov}(X, Y) = E[XY] - E[X]E[Y] = 1/2; \\
\]

\[
\hat{Y} = E[Y] + \frac{\text{cov}(X, Y)}{\text{var}[X]} (X - E[X]) = X.
\]
Example 2:

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

Example 2:

We find:

\[ E[X] = \]

\[ E[Y] = \]

\[ E[X^2] = \frac{1}{2} \]

\[ E[XY] = \frac{1}{2} \]

\[ \text{var}[X] = E[X^2] - E[X]^2 = \frac{1}{2} \]

\[ \text{cov}(X, Y) = E[XY] - E[X]E[Y] = \frac{1}{2} \]

\[ \hat{Y} = E[Y] + \frac{\text{cov}(X, Y)}{\text{var}[X]} (X - E[X]) \]
Example 2:

We find:

\[ E[X] = 0; \]
Linear Regression Examples

Example 2:

We find:

\[ E[X] = 0; E[Y] = \]
Linear Regression Examples

Example 2:

We find:

\[ E[X] = 0; E[Y] = 0; \]
Linear Regression Examples

Example 2:

We find:

\[ E[X] = 0; \ E[Y] = 0; \ E[X^2] = \]

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

\[ \text{cov}(X, Y) = E[XY] - E[X]E[Y] = \]

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

Example 2:

We find:

\[ E[X] = 0; \ E[Y] = 0; \ E[X^2] = \frac{1}{2}; \]
Linear Regression Examples

Example 2:

We find:

\[ E[X] = 0; \ E[Y] = 0; \ E[X^2] = 1/2; \ E[XY] = \]
Example 2:

We find:

\[
E[X] = 0; \ E[Y] = 0; \ E[X^2] = 1/2; \ E[XY] = 1/2;
\]
Linear Regression Examples

Example 2:

We find:

\[ E[X] = 0; \ E[Y] = 0; \ E[X^2] = 1/2; \ E[XY] = 1/2; \]

\[ \text{var}[X] = E[X^2] - E[X]^2 = \]
Linear Regression Examples

Example 2:

We find:

\[ E[X] = 0; \ E[Y] = 0; \ E[X^2] = 1/2; \ E[XY] = 1/2; \]

\[ \text{var}[X] = E[X^2] - E[X]^2 = 1/2; \]
Example 2:

We find:

\[ E[X] = 0; \quad E[Y] = 0; \quad E[X^2] = 1/2; \quad E[XY] = 1/2; \]

\[ var[X] = E[X^2] - E[X]^2 = 1/2; \quad cov(X, Y) = E[XY] - E[X]E[Y] = \]
Linear Regression Examples

Example 2:

We find:

\[ E[X] = 0; \ E[Y] = 0; \ E[X^2] = 1/2; \ E[XY] = 1/2; \]
\[ var[X] = E[X^2] - E[X]^2 = 1/2; \ cov(X, Y) = E[XY] - E[X]E[Y] = 1/2; \]
Linear Regression Examples

Example 2:

We find:

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

\[ E[X] = 0; \]
\[ E[Y] = 0; \]
\[ E[X^2] = \frac{1}{2}; \]
\[ E[XY] = -\frac{1}{2}; \]

\[ \text{var}[X] = E[X^2] - E[X]^2 = \frac{1}{2}; \]
\[ \text{cov}(X,Y) = E[XY] - E[X]E[Y] = -\frac{1}{2}; \]

\[ \hat{Y} = E[Y] + \frac{\text{cov}(X,Y)}{\text{var}[X]} (X - E[X]) = -X. \]
Example 3:
Linear Regression Examples

Example 3:

We find:

\[ E[X] = \]

\[ E[Y] = \]

\[ E[X^2] = \frac{1}{2} \]

\[ E[XY] = -\frac{1}{2} \]

\[ \text{var}[X] = E[X^2] - E[X]^2 = \frac{1}{2} \]

\[ \text{cov}(X, Y) = E[XY] - E[X]E[Y] = -\frac{1}{2} \]

\[ \hat{Y} = E[Y] + \text{cov}(X, Y) \text{var}[X] (X - E[X]) = -X \]
Linear Regression Examples

Example 3:

We find:

\[ E[X] = 0; \]

\[ E[Y] = 0; \]

\[ E[X^2] = \frac{1}{2}; \]

\[ E[XY] = -\frac{1}{2}; \]

\[ \text{Var}[X] = E[X^2] - (E[X])^2 = \frac{1}{2}; \]

\[ \text{Cov}(X,Y) = E[XY] - E[X]E[Y] = -\frac{1}{2}; \]

LR:

\[ \hat{Y} = E[Y] + \text{Cov}(X,Y) \text{Var}[X] (X - E[X]) = -X. \]
Linear Regression Examples

Example 3:

We find:

$$E[X] = 0; E[Y] =$$
Example 3:

We find:

\[ E[X] = 0; E[Y] = 0; \]
Linear Regression Examples

Example 3:

We find:

\[ E[X] = 0; \ E[Y] = 0; \ E[X^2] = \]
Linear Regression Examples

Example 3:

We find:

\[ E[X] = 0; E[Y] = 0; E[X^2] = 1/2; \]
Linear Regression Examples

Example 3:

We find:

\[ E[X] = 0; \ E[Y] = 0; \ E[X^2] = 1/2; \ E[XY] = \]
Example 3:

We find:

\[ E[X] = 0; \ E[Y] = 0; \ E[X^2] = 1/2; \ E[XY] = -1/2; \]
Linear Regression Examples

Example 3:

We find:

\[ E[X] = 0; E[Y] = 0; E[X^2] = 1/2; E[XY] = -1/2; \]
\[ var[X] = E[X^2] - E[X]^2 = \]
Linear Regression Examples

Example 3:

We find:

\[ E[X] = 0; E[Y] = 0; E[X^2] = 1/2; E[XY] = -1/2; \]
\[ \text{var}[X] = E[X^2] - E[X]^2 = 1/2; \]
Example 3:

We find:

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

Example 3:

We find:

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

Example 3:

We find:

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

\[ E[X] = 3; \]
\[ E[Y] = 2.5; \]
\[ E[X^2] = \left(\frac{3}{15}\right)(1^2 + 2^2 + 3^2 + 4^2 + 5^2) = 11; \]
\[ E[XY] = \left(\frac{1}{15}\right)(1 \times 1 + 1 \times 2 + \cdots + 5 \times 4) = 8.4; \]
\[ \text{var}[X] = 11 - 9 = 2; \]
\[ \text{cov}(X, Y) = 8.4 - 3 \times 2.5 = 0.9; \]

LR:

\[ \hat{Y} = 2.5 + 0.9(2) = 1.15 + 0.45X. \]
Linear Regression Examples

Example 4:

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

Example 4:

We find:

\[ E[X] = 3 \]

\[ E[Y] = 2.5 \]

\[ E[X^2] = \left( \frac{3}{15} \right) (1^2 + 2^2 + 3^2 + 4^2 + 5^2) = 11 \]

\[ E[XY] = \left( \frac{1}{15} \right) (1 \times 1 + 1 \times 2 + \cdots + 5 \times 4) = 8.4 \]

\[ \text{var}[X] = 11 - 9 = 2 \]

\[ \text{cov}(X, Y) = 8.4 - 3 \times 2.5 = 0.9 \]

\[ \hat{Y} = 2.5 + 0.9(X - 3) = 1.15 + 0.45X \]
Linear Regression Examples

Example 4:

We find:

\[ E[X] = 3; \]

\[ E[Y] = 2.5; \]

\[ E[X^2] = \left( \frac{3}{15} \right) (1^2 + 2^2 + 3^2 + 4^2 + 5^2) = 11; \]

\[ E[XY] = \left( \frac{1}{15} \right) (1 \times 1 + 1 \times 2 + \cdots + 5 \times 4) = 8.4; \]

\[ \text{var}[X] = E[X^2] - E[X]^2 = 11 - 9 = 2; \]

\[ \text{cov}(X, Y) = E[XY] - E[X] \cdot E[Y] = 8.4 - 3 \times 2.5 = 0.9; \]

\[ \hat{Y} = 2.5 + 0.9 \cdot (X - 3) = 1.15 + 0.945X. \]
Linear Regression Examples

Example 4:

We find:

\[ E[X] = 3; E[Y] = \]

\[ E[X^2] = \frac{3}{15} (1^2 + 2^2 + 3^2 + 4^2 + 5^2) = 11; \]

\[ E[XY] = \frac{1}{15} (1 \times 1 + 1 \times 2 + \cdots + 5 \times 4) = 8.4; \]

\[ \text{var}[X] = 11 - 9 = 2; \]

\[ \text{cov}(X, Y) = 8.4 - 3 \times 2.5 = 0.9; \]

\[ \hat{Y} = 2.5 + 0.9(2) = 1.15 + 0.45X. \]
Linear Regression Examples

Example 4:

We find:

\[ E[X] = 3; \ E[Y] = 2.5; \]
Linear Regression Examples

Example 4:

We find:

\[ E[X] = 3; \quad E[Y] = 2.5; \quad E[X^2] = \left(\frac{3}{15}\right)(1 + 2^2 + 3^2 + 4^2 + 5^2) = 11; \]
Linear Regression Examples

Example 4:

We find:

\[ E[X] = 3; \ E[Y] = 2.5; \ E[X^2] = (3/15)(1 + 2^2 + 3^2 + 4^2 + 5^2) = 11; \]
\[ E[XY] = (1/15)(1 \times 1 + 1 \times 2 + \cdots + 5 \times 4) = 8.4; \]
Linear Regression Examples

Example 4:

We find:

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

Example 4:

We find:

\[ E[X] = 3; \ E[Y] = 2.5; \ E[X^2] = (3/15)(1 + 2^2 + 3^2 + 4^2 + 5^2) = 11; \]
\[ E[XY] = (1/15)(1 \times 1 + 1 \times 2 + \cdots + 5 \times 4) = 8.4; \]
\[ var[X] = 11 - 9 = 2; \ cov(X, Y) = 8.4 - 3 \times 2.5 = 0.9; \]
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 \((X_n, Y_n)\). Its slope is \(\frac{\text{cov}(X,Y)}{\text{var}[X]}\).
Note that

- the LR line goes through \((E[X], E[Y])\)
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. Linear Regression:
   \[ Y | X = E[Y] + \text{cov}(X,Y) \cdot \text{var}(X) \cdot (X - E[X]) \]

2. Non-Bayesian: minimize
   \[ \sum_{n} (Y_n - a - bX_n)^2 \]

3. Bayesian: minimize
   \[ E[(Y - a - bX)^2] \]
Summary

1. Linear Regression: \( L[Y|X] = E[Y] + \frac{\text{cov}(X,Y)}{\text{var}(X)} (X - E[X]) \)
Summary

Linear Regression

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

2. Non-Bayesian: minimize $\sum_n (Y_n - a - bX_n)^2$
Summary

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