Linear Regression (LR)
Linear Regression (LR)

1. Motivation for Linear Regression (LR)
2. Minimum Mean Squared Error: Discussion
3. Covariance: Definition and Properties
4. Linear Regression (LR): Non-Bayesian vs. Bayesian (LLSE)
5. Derivation and Illustration
Linear Regression: Motivation

Example 1: 100 people.

Let $(x_n, y_n) = (\text{height}, \text{weight})$ of person $n$, for $n = 1, ..., 100$:

$$E[y] = \beta_0 + \beta_1 x.$$ 

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

Best linear fit: Linear Regression.
Example 1: 100 people.

\[
\begin{align*}
E[Y] &= -114.3 + 106.5 X \\
\end{align*}
\]

(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 \(n = 1, \ldots, 100\):
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 \(n = 1, \ldots, 100\):

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 \(n = 1, \ldots, 100\):

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.

The line $Y = a + bX$ is the linear regression.
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.
If we want to guess the value of a random variable \( Y \), and know nothing more than its distribution, what's our best guess?

Depends on how we measure the 'goodness' of our guess. Say we use the expected squared error between \( Y \) and our guess as the "error" measure. Then?

Answer is:

\[
E[Y]
\]

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

\[
E[Y]
\]

Proof:

Let \( \hat{Y} := Y - E[Y] \).

Then,

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

So,

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

Now,

\[
E[(Y - a)^2] = E[(Y - E[Y] + E[Y] - a)^2] = E[\hat{Y}^2 + 2c\hat{Y} + c^2] = E[\hat{Y}^2] + 0 + c^2 \geq E[\hat{Y}^2],
\]

Hence,

\[
E[(Y - a)^2] \geq E[(Y - E[Y])^2], \quad \forall a.
\]
If we want to guess the value of a random variable $Y$, and know nothing more than its distribution, what’s our best guess?

Depends on how we measure the 'goodness' of our guess. Say we use the expected squared error between $Y$ and our guess as the "error" measure. Then?

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

If we want to guess the value of a random variable $Y$, and know nothing more than its distribution, what’s our best guess?

Depends on how we measure the ‘goodness’ of our guess.
If we want to guess the value of a random variable $Y$, and know nothing more than its distribution, what’s our best guess?

Depends on how we measure the ‘goodness’ of our guess.

Say we use the expected squared error between $Y$ and our guess as the “error” measure. Then?

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

Proof:

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

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

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


Hence, \[ E[(Y - a)^2] \geq E[(Y - E[Y])^2], \forall a. \]
Linear Regression: Discussion

If we want to guess the value of a random variable $Y$, and know nothing more than its distribution, what’s our best guess? Depends on how we measure the ‘goodness’ of our guess.

Say we use the **expected squared error** between $Y$ and our guess as the “error” measure. Then? Answer is:
If we want to guess the value of a random variable \( Y \), and know nothing more than its distribution, what’s our best guess?

Depends on how we measure the ‘goodness’ of our guess.

Say we use the **expected squared error between \( Y \) and our guess** as the “error” measure. Then? Answer is: \( E[Y] \).
Linear Regression: Discussion

If we want to guess the value of a random variable $Y$, and know nothing more than its distribution, what’s our best guess? Depends on how we measure the ‘goodness’ of our guess.

Say we use the **expected squared error between $Y$ and our guess** as the “error” measure. Then? Answer is: $E[Y]$.

More precisely, the value of $a$ that minimizes $E[(Y - a)^2]$ is
Linear Regression: Discussion

If we want to guess the value of a random variable \( Y \), and know nothing more than its distribution, what’s our best guess? Depends on how we measure the ‘goodness’ of our guess.

Say we use the **expected squared error between \( Y \) and our guess** as the “error” measure. Then? Answer is: \( E[Y] \).

More precisely, the value of \( a \) that minimizes \( E[(Y - a)^2] \) is \( a = E[Y] \).
If we want to guess the value of a random variable $Y$, and know nothing more than its distribution, what’s our best guess?

Depends on how we measure the ‘goodness’ of our guess.

Say we use the **expected squared error between $Y$ and our guess** as the “error” measure. Then? Answer 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, $E[\hat{Y}^2] = 0$, $\forall c$.


Hence, $E[(Y - a)^2] \geq E[\hat{Y}^2]$.
Linear Regression: Discussion

If we want to guess the value of a random variable $Y$, and know nothing more than its distribution, what’s our best guess?

Depends on how we measure the ‘goodness’ of our guess.

Say we use the expected squared error between $Y$ and our guess as the “error” measure. Then? Answer is: $E[Y]$.

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

Proof:

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

If we want to guess the value of a random variable $Y$, and know nothing more than its distribution, what’s our best guess? Depends on how we measure the ‘goodness’ of our guess.

Say we use the **expected squared error between $Y$ and our guess** as the “error” measure. Then? Answer 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, $E[\hat{Y}] = 0$. 
If we want to guess the value of a random variable $Y$, and know nothing more than its distribution, what’s our best guess?

Depends on how we measure the ‘goodness’ of our guess.

Say we use the **expected squared error between $Y$ and our guess** as the “error” measure. Then? Answer 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, $E[\hat{Y}c] = 0, \forall c$. 

**Expected squared error**
Linear Regression: Discussion

If we want to guess the value of a random variable $Y$, and know nothing more than its distribution, what’s our best guess?

Depends on how we measure the ‘goodness’ of our guess.

Say we use the **expected squared error between $Y$ and our guess** as the “error” measure. Then? Answer 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,

If we want to guess the value of a random variable $Y$, and know nothing more than its distribution, what's our best guess?

Depends on how we measure the ‘goodness’ of our guess.

Say we use the **expected squared error** between $Y$ and our guess as the "error" measure. Then? Answer 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,

$$= E[(\hat{Y} + c)^2]$$
Linear Regression: Discussion

If we want to guess the value of a random variable $Y$, and know nothing more than its distribution, what’s our best guess?

Depends on how we measure the ‘goodness’ of our guess.

Say we use the **expected squared error between $Y$ and our guess** as the “error” measure. Then? Answer 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,


$$= E[(\hat{Y} + c)^2] \text{ with } c = E[Y] - a$$
If we want to guess the value of a random variable $Y$, and know nothing more than its distribution, what’s our best guess? Depends on how we measure the ‘goodness’ of our guess. Say we use the **expected squared error between $Y$ and our guess** as the “error” measure. Then? Answer 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,

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

If we want to guess the value of a random variable $Y$, and know nothing more than its distribution, what’s our best guess?

Depends on how we measure the ‘goodness’ of our guess.

Say we use the **expected squared error between $Y$ and our guess** as the “error” measure. Then? Answer 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,


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

$$= E[\hat{Y}^2 + 2\hat{Y}c + c^2] = E[\hat{Y}^2] + 2E[\hat{Y}c] + c^2$$
Linear Regression: Discussion

If we want to guess the value of a random variable \( Y \), and know nothing more than its distribution, what’s our best guess?

Depends on how we measure the ‘goodness’ of our guess.

Say we use the **expected squared error between \( Y \) and our guess** as the “error” measure. Then? Answer 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,

\[
= E[(\hat{Y} + c)^2] \text{ with } c = E[Y] - a \\
= E[\hat{Y}^2 + 2\hat{Y}c + c^2] = E[\hat{Y}^2] + 2E[\hat{Y}c] + c^2 \\
= E[\hat{Y}^2] + 0 + c^2
\]
Linear Regression: Discussion

If we want to guess the value of a random variable $Y$, and know nothing more than its distribution, what’s our best guess?

Depends on how we measure the ‘goodness’ of our guess.

Say we use the **expected squared error between $Y$ and our guess** as the “error” measure. Then? Answer 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,

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

If we want to guess the value of a random variable $Y$, and know nothing more than its distribution, what’s our best guess?

Depends on how we measure the ‘goodness’ of our guess.

Say we use the **expected squared error between $Y$ and our guess** as the “error” measure. Then? Answer 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,

\[
\]

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

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

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 $Y$?

Idea: use a function $g(X)$ of the observation to estimate $Y$.

The simplest $g(X)$ 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.

(We can also consider a general function $g(X)$. Any guess on what is the best function to use? Answer: $E[Y|X]$.)
Thus, if we want to guess the value of $Y$, we choose $E[Y]$. 
Thus, if we want to guess the value of $Y$, we choose $E[Y]$. Now assume we make some observation $X$ related to $Y$. 
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 $Y$? Idea: 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 $Y$? Idea: use a function $g(X)$ of the observation to estimate $Y$. The simplest $g(X)$ 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. (We can also consider a general function $g(X)$. Any guess on what is the best function to use? Answer: $E[Y|X]$.)
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 \( Y \)?
Idea: use a function \( g(X) \) of the observation to estimate \( Y \).
The simplest \( g(X) \) is a constant that does not depend on \( X \).
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 $Y$? Idea: use a function $g(X)$ of the observation to estimate $Y$. The simplest $g(X)$ 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?
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 $Y$? Idea: use a function $g(X)$ of the observation to estimate $Y$. The simplest $g(X)$ 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.
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 $Y$? Idea: use a function $g(X)$ of the observation to estimate $Y$. The simplest $g(X)$ 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. (We can also consider a general function $g(X)$. Any guess on what is the best function to use?)
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 $Y$? Idea: use a function $g(X)$ of the observation to estimate $Y$. The simplest $g(X)$ 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. (We can also consider a general function $g(X)$. Any guess on what is the best function to use? Answer:
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 $Y$? Idea: use a function $g(X)$ of the observation to estimate $Y$. The simplest $g(X)$ 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. (We can also consider a general function $g(X)$. Any guess on what is the best function to use? Answer: $E[Y|X]$.)
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 $Y$? Idea: use a function $g(X)$ of the observation to estimate $Y$. The simplest $g(X)$ 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. (We can also consider a general function $g(X)$. Any guess on what is the best function to use? Answer: $E[Y|X]$.)
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

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

**Fact**

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

**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:

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

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.

Four equally likely pairs of values:

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

Four equally likely pairs of values

$cov(X, Y) = 1/2$  \hspace{1cm} $cov(X, Y) = -1/2$  \hspace{1cm} $cov(X, Y) = 0$

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

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.
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.
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.
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

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

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

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

$$E[XY] = 1 \times 0.05 + 1 \times 0.25 + 3 \times 0.25 + 3 \times 0.25 = 0.55 + 0.25 + 0.75 + 0.75 = 2.35$$

$$cov(X, Y) = E[XY] - E[X]E[Y] = 2.35 - 1.9 \times 2 = 0.55$$

$$var[X] = E[X^2] - (E[X])^2 = 5.8 - 1.9^2 = 0.29$$
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

\[E[X] = 1 \times 0.15 + 2 \times 0.4 + 3 \times 0.45 = 1.9\]
\[E[X^2] = 1^2 \times 0.15 + 2^2 \times 0.4 + 3^2 \times 0.45 = 5.8\]
\[E[Y] = 1 \times 0.2 + 2 \times 0.6 + 3 \times 0.2 = 2\]
\[E[XY] = 1 \times 0.05 + 1 \times 2 \times 0.1 + \cdots + 3 \times 3 \times 0.2 = 4.85\]
Examples of Covariance

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

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

Properties of Covariance


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


Fact
(a) $var[X] = cov(X, X)$
(b) $X, Y$ independent $\Rightarrow cov(X, Y) =$
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:
Prove (a),(b),(c) yourself to check your understanding.
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:**
Prove (a),(b),(c) yourself to check your understanding.
(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) \( \text{cov}(aX + bY, cU + dV) = ac \cdot \text{cov}(X, U) + ad \cdot \text{cov}(X, V) \)
\[ + bc \cdot \text{cov}(Y, U) + bd \cdot \text{cov}(Y, V). \]

**Proof:**
Prove (a),(b),(c) yourself to check your understanding.
(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) \( \text{cov}(aX + bY, cU + dV) = ac \cdot \text{cov}(X, U) + ad \cdot \text{cov}(X, V) \)
\[ \quad + bc \cdot \text{cov}(Y, U) + bd \cdot \text{cov}(Y, V). \]

Proof:
Prove (a),(b),(c) yourself to check your understanding.
(d) In view of (c), one can subtract the means and assume that the RVs are zero-mean. Then,
\[ \text{cov}(aX + bY, cU + dV) = E[(aX + bY)(cU + dV)] \]
\[ = ac \cdot E[XU] + ad \cdot E[XV] + bc \cdot E[YU] + bd \cdot E[YV] \]
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:

Prove (a),(b),(c) yourself to check your understanding.
(d) In view of (c), one can subtract the means and assume that the RVs are zero-mean. Then,

\[
\text{cov}(aX + bY, cU + dV) = E[(aX + bY)(cU + dV)]
\]
\[
= ac \cdot E[XU] + ad \cdot E[XV] + bc \cdot E[YU] + bd \cdot E[YV]
\]
\[
= ac \cdot \text{cov}(X, U) + ad \cdot \text{cov}(X, V) + bc \cdot \text{cov}(Y, U) + bd \cdot \text{cov}(Y, V).
\]
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 \( X_n \).
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 \( (Y_n - \hat{Y}_n)^2 \).
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.

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

Note: This is a Bayesian formulation: there is a prior.
Linear Least Squares Estimate

**Definition**
Given two RVs $X$ and $Y$ with known distribution $Pr[X = x, Y = y]$,

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

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.

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

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

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

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.
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.

Note: This is a Bayesian formulation: there is a prior.
Linear Regression: Example

Example 1: 100 people.

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

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

(\(X\) in meters, \(Y\) in kg.) Best linear fit: Linear Regression.
Linear Regression: Example

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

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

The blue line is \(Y = -114.3 + 106.5X\). (\(X\) in meters, \(Y\) in kg.) Best linear fit: Linear Regression.
Observe that
\[ \sum_{n=1}^{N} (Y_n - a - bX_n)^2 = E[(Y_n - a - bX_n)^2] \]
where one assumes that \((X_n, Y_n) = (X_n, Y_n)\), 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!
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

$$(X, Y) = (X_n, Y_n), \text{ w.p. } \frac{1}{N} \text{ for } n = 1, \ldots, N.$$
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!
Consider two RVs \( X, Y \) with a given distribution \( \Pr[X=x, Y=y] \). Then,

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

**Proof 1:**

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

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 equalities,

\[
E[(Y - \hat{Y})(c + dX)] = 0.
\]

Then,

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

Indeed:

\[
\hat{Y} = \alpha + \beta X
\]

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

\[
E[(Y - \hat{Y})^2] \leq E[(Y - a - bX)^2], \quad \forall (a, b).
\]

Thus \( \hat{Y} \) is the LLSE.
Theorem

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

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

Proof 1:

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

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 equalities,

$$E[(Y - \hat{Y})(c + dX)] = 0.$$

Then,

$$E[(Y - \hat{Y})(\hat{Y} - a - bX)] = 0,$$

$\forall a, b$.

Indeed:

$$\hat{Y} = \alpha + \beta X$$

so that

$$\hat{Y} - a - bX = c + dX$$

for some $c, d$.

Now,

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

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,

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

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

Proof 1:
$Y - \hat{Y} = (Y - E[Y]) - \frac{\text{cov}(X, Y)}{\text{var}[X]}(X - E[X]).$ Hence, $E[Y - \hat{Y}] = 0$. 
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,

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

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.
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:**

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

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

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 $\alpha, \beta$. 
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 $c, d$. 
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,
\[
E[(Y - a - bX)^2] = E[(Y - \hat{Y} + \hat{Y} - a - bX)^2]
\]
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,

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

$$= E[(Y - \hat{Y})^2] + E[(\hat{Y} - a - bX)^2] + 0$$
Theorem
Consider two RVs $X$, $Y$ with a given distribution $Pr[X = x, Y = y]$. Then,

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

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,

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

$$= E[(Y - \hat{Y})^2] + E[(\hat{Y} - a - bX)^2] + 0 \geq E[(Y - \hat{Y})^2].$$
Theorem
Consider two RVs $X$, $Y$ with a given distribution $Pr[X = x, Y = y]$. Then,

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

Proof 1:
$Y - \hat{Y} = (Y - E[Y]) - \frac{cov(X, Y)}{\text{var}[X]}(X - E[X])$. 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,

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

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

\((*)\) Recall that \( \text{cov}(X, Y) = E[(X - E[X])(Y - E[Y])] \) and \( \text{var}[X] = E[(X - E[X])^2]. \)
Note that the LR line goes through $(X, Y)$, its slope is $\frac{\text{cov}(X, Y)}{\text{var}[X]}$. 

- $\text{cov}(X, Y)$ denotes the covariance between $X$ and $Y$.
- $\text{var}[X]$ denotes the variance of $X$. 

In the context of linear regression, the slope of the regression line is directly related to the covariance and variance of the variables involved.
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**


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

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

Summary


Linear Regression
Summary

Linear Regression

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

Linear Regression

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

Linear Regression

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