Confidence Intervals; Linear Regression
Confidence Intervals; Linear Regression

1. Review
2. Confidence Intervals
3. Motivation for LR
4. History of LR
5. Linear Regression
6. Derivation
7. More examples
Review: Probability Ideas Map - Details

- **Probability**
- **Space**
- **Cond. Prob.**
- **BR Ind.**
- **RVs**
- **E[X]**
- **Lin. Mon.**
- **E[XY] =**
- **var[∑] = ∑var[ ]**
- **Cheby**
- **WLLN**
- **MAP, MLE**

**definitions**

**blue: results**

\[
\text{var}(X_1 + \cdots + X_n) = \frac{1}{n} \text{var}(X_1) + \cdots + \text{var}(X_n)
\]

\[
\text{Pr}[|X - \mu| < \varepsilon] \leq \frac{2}{\varepsilon^2}
\]

\[
\text{Pr}[A_m | B] = p_m q_m
\]

\[
\text{arg max} p_m q_m = \text{arg max} q_m
\]

\[
X = X_1 + \cdots + X_n
\]

\[
X \sim \mu
\]
Review: Probability Ideas Map - Details

- Probability Space
  - Cond. Prob.
  - RVs
- Cond. Prob.
  - BR
  - Ind.
  - MAP, MLE
- MAP, MLE
  - WLLN
    - $X \approx \mu$
    - $X = \frac{X_1 + \cdots + X_n}{n}$
    - $\Pr[|X - \mu| \geq \epsilon] \leq \frac{\sigma^2}{\epsilon^2}$
    - $\arg\max p_m q_m$
    - $\arg\max q_m$
- WLLN
  - Cheby
    - $\text{Var}[\Sigma] = \Sigma \text{Var}[\ ]$
  - Lin.
  - Mon.
    - $E[X]$
    - $E[XY] = \text{Lin.}$
    - $\text{Var}[\sum] = \sum \text{Var}[\ ]$
- Lin.
  - Mon.
    - $E[X]$
Review: Probability Ideas Map - Today
Review: Probability Ideas Map - Today

- **Probability Space**
- Cond. Prob.
- RVs
  - E[X]
  - Var[∑] = ∑Var[
- E[XY] =
- Lin.
- Mon.
- BR
- Ind.
- LLSE, LR
- MAP, MLE
- WLLN
- CI
- Cheby

**Labels**
- Black: definitions
- Blue: results
- Red: today
Confidence Intervals: Example

Flip a coin \( n \) times. Let \( A_n \) be the fraction of Heads.

We know that \( p := \Pr[H] \approx A_n \) for \( n \) large (WLLN).

Can we find \( a \) such that \( \Pr[p \in [A_n - a, A_n + a]] \geq 95\% \)?

If so, we say that \([A_n - a, A_n + a]\) is a 95\% Confidence Interval for \( p \).

Using Chebyshev, we will see that \( a = 2\frac{1}{\sqrt{n}} \) works.

Thus \([A_n - 2\frac{1}{\sqrt{n}}, A_n + 2\frac{1}{\sqrt{n}}]\) is a 95\%-CI for \( p \).

Example: If \( n = 1500 \), then \( \Pr[p \in [A_n - 0.05, A_n + 0.05]] \geq 95\% \).

In fact, we will see later that \( a = \frac{1}{\sqrt{n}} \) works, so that with \( n = 1000 \) one has \( \Pr[p \in [A_n - 0.02, A_n + 0.02]] \geq 95\% \).
Confidence Intervals: Example

- Flip a coin \( n \) times.
Confidence Intervals: Example

- Flip a coin \( n \) times. Let \( A_n \) be the fraction of \( Hs \).
Confidence Intervals: Example

- Flip a coin $n$ times. Let $A_n$ be the fraction of $H$s.
- We know that $p := Pr[H] \approx A_n$ for $n$ large (WLLN).
Confidence Intervals: Example

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

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

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

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

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

\[
[A_n - a, A_n + a] \text{ is a } 95\%- \text{ Confidence Interval for } p.
\]

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

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

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

\[ [A_n - a, A_n + a] \text{ is a } 95\%-\text{ Confidence Interval for } p. \]

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

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

Example: If \( n = 1500 \), then \( Pr[p \in [A_n - 0.05, A_n + 0.05]] \geq 95\% \).
Confidence Intervals: Example

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

$$[A_n - a, A_n + a]$$

is a 95\%- Confidence Interval for $p$.

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

$$[A_n - \frac{2.25}{\sqrt{n}}, A_n + \frac{2.25}{\sqrt{n}}]$$

is a 95\%-CI for $p$.

Example: If $n = 1500$, then $Pr[p \in [A_n - 0.05, A_n + 0.05]] \geq 95\%$.

In fact, we will see later that $a = \frac{1}{\sqrt{n}}$ works,
Confidence Intervals: Example

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

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

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

Example: If \( n = 1500 \), then \( Pr[p \in [A_n - 0.05, A_n + 0.05]] \geq 95\% \).

In fact, we will see later that \( a = \frac{1}{\sqrt{n}} \) works, so that with \( n = 1,500 \) one has \( Pr[p \in [A_n - 0.02, A_n + 0.02]] \geq 95\% \).
Theorem:
Let $X_n$ be i.i.d. with mean $\mu$ and variance $\sigma^2$.

Define $A_n = X_1 + \cdots + X_n$.

Then,

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

Thus,

$$\left[ A_n - 4.5\sigma \sqrt{n}, A_n + 4.5\sigma \sqrt{n} \right]$$ 

is a 95\% CI for $\mu$.

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

Then $\mu = E[ X_n ] = p := \Pr[ H ]$.

Also, $\sigma^2 = \text{var}(X_n) = p(1 - p) \leq 1/4$.

Hence,

$$\left[ A_n - 4.5\frac{1}{2} \sqrt{n}, A_n + 4.5\frac{1}{2} \sqrt{n} \right]$$ 

is a 95\% CI for $p$. 
Theorem: Let $X_n$ be i.i.d. with mean $\mu$ and variance $\sigma^2$. Then,

$$\Pr[A_n - 4.5 \sqrt{n} < \mu < A_n + 4.5 \sqrt{n}] \geq 95\%.$$ 

Thus, 

$$[A_n - 4.5 \sqrt{n}, A_n + 4.5 \sqrt{n}]$$ 

is a 95\% CI for $\mu$. 

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

$$\mu = E[X_n] = p := \Pr[H].$$ 

Also, 

$$\sigma^2 = \text{var}(X_n) = p(1-p) \leq \frac{1}{4}.$$ 

Hence, 

$$[A_n - 4.5 \frac{1}{\sqrt{2n}}, A_n + 4.5 \frac{1}{\sqrt{2n}}]$$ 

is a 95\% CI for $p$. 
Theorem:
Let $X_n$ be i.i.d. with mean $\mu$ and variance $\sigma^2$. Define $A_n = \frac{X_1 + \ldots + X_n}{n}$. 
Confidence Intervals: Result

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

$$Pr[\mu \in [A_n - 4.5 \frac{\sigma}{\sqrt{n}}, A_n + 4.5 \frac{\sigma}{\sqrt{n}}]] \geq 95\%.$$
Theorem:
Let $X_n$ be i.i.d. with mean $\mu$ and variance $\sigma^2$. Define $A_n = \frac{X_1 + \cdots + X_n}{n}$. Then,

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

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

Example:
Let $X_n = 1\{\text{coin } n \text{ yields } H\}$. Then $\mu = E[X_n] = p := \Pr[H]$. Also, $\sigma^2 = \text{var}(X_n) = p(1 - p) \leq \frac{1}{4}$. Hence, $[A_n - 4.5 \frac{\sigma}{\sqrt{n}}, A_n + 4.5 \frac{\sigma}{\sqrt{n}}]$ is a 95%-CI for $p$. 

Confidence Intervals: Result
Confidence Intervals: Result

**Theorem:**
Let $X_n$ be i.i.d. with mean $\mu$ and variance $\sigma^2$.
Define $A_n = \frac{X_1 + \ldots + X_n}{n}$. Then,

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

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

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

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

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

**Example:** Let $X_n = 1\{\text{coin } n \text{ yields } H\}$. 
Confidence Intervals: Result

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

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

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

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

$$\mu = E[X_n] = p := Pr[H].$$
Confidence Intervals: Result

**Theorem:**
Let $X_n$ be i.i.d. with mean $\mu$ and variance $\sigma^2$. Define $A_n = \frac{X_1 + \ldots + X_n}{n}$. Then,

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

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

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

$$\mu = E[X_n] = p := Pr[H].$$

Also, $\sigma^2 = var(X_n) = \ldots$
Confidence Intervals: Result

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Proof:

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

From Chebyshev:

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

Thus,

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

Hence,

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

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

From Chebyshev:

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

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

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

Confidence Interval: Analysis
Confidence Interval: Analysis

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

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

From Chebyshev:

\[
Pr[|A_n - \mu| \geq 4.5\sigma/\sqrt{n}] \leq \frac{\text{var}(A_n)}{[4.5\sigma/\sqrt{n}]^2}
\]
Confidence Interval: Analysis

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

From Chebyshev:

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

$$\leq \frac{\sigma^2 / n}{20\sigma^2 / n}$$
Confidence Interval: Analysis

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

From Chebyshev:

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

$$\leq \frac{\sigma^2 / n}{20\sigma^2 / n} = 5\%.$$
Confidence Interval: Analysis

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

From Chebyshev:

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

$$\leq \frac{\sigma^2 / n}{20\sigma^2 / n} = 5\%.$$  

Thus,

$$Pr[|A_n - \mu| \leq 4.5\sigma / \sqrt{n}] \geq 95\%.$$
Proof:
We prove the theorem, i.e., that $A_n \pm 4.5\sigma/\sqrt{n}$ is a 95%-CI for $\mu$. From Chebyshev:

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

$$\leq \frac{\sigma^2/n}{20\sigma^2/n} = 5\%.$$

Thus,

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

Hence,

$$Pr[\mu \in [A_n - 4.5\sigma/\sqrt{n}, A_n + 4.5\sigma/\sqrt{n}]] \geq 95\%.$$
Recall that 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 \( a = \mathbb{E}[Y] \).

Let's review one proof of that fact.

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

Then, \( \mathbb{E}[\hat{Y}] = 0 \).

So, \( \mathbb{E}[\hat{Y}c] = 0, \forall c \).

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

with \( c = \mathbb{E}[Y] - a \).

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

Hence, \( \mathbb{E}[(Y - a)^2] \geq \mathbb{E}[(Y - \mathbb{E}[Y])^2], \forall a \).
Recall that the best guess about \( Y \),
Recall that 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 $a = \mathbb{E}[Y]$. Let's review one proof of that fact.

Let $\hat{Y} := Y - \mathbb{E}[Y]$. Then, $\mathbb{E}[\hat{Y}] = 0$. So, $\mathbb{E}[\hat{Y}^2] = 0$, $\forall c$. Now, $\mathbb{E}[(Y - a)^2] = \mathbb{E}[(Y - \mathbb{E}[Y] + \mathbb{E}[Y] - a)^2] = \mathbb{E}[\hat{Y}^2 + 2\hat{Y}c + c^2] = \mathbb{E}[\hat{Y}^2] + 0 + c^2 \geq \mathbb{E}[\hat{Y}^2]$, $\forall a$. Hence, $\mathbb{E}[(Y - a)^2] \geq \mathbb{E}[(Y - \mathbb{E}[Y])^2]$, $\forall a$. 

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

More precisely, the value of $a$ that minimizes $E[(Y - a)^2]$ is
Recall that the best guess about $Y$, if we know only the distribution of $Y$, is $E[Y]$.

More precisely, the value of $a$ that minimizes $E[(Y - a)^2]$ is $a = E[Y]$.
Recall that the best guess about $Y$, if we know only the distribution of $Y$, is $E[Y]$.

More precisely, the value of $a$ that minimizes $E[(Y - a)^2]$ is $a = E[Y]$. Let’s review one proof of that fact.
Recall that the best guess about $Y$, if we know only the distribution of $Y$, is $E[Y]$.

More precisely, the value of $a$ that minimizes $E[(Y - a)^2]$ is $a = E[Y]$. Let’s review one proof of that fact.

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

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

Hence,
\[
E[(Y - a)^2] \geq E[\hat{Y}^2], \quad \forall a.
\]
Recall that the best guess about $Y$, if we know only the distribution of $Y$, is $E[Y]$.

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

Let’s review one proof of that fact.

Let $\hat{Y} := Y - E[Y]$. Then, $E[\hat{Y}] = 0$. 
Recall that the best guess about $Y$, if we know only the distribution of $Y$, is $E[Y]$. More precisely, the value of $a$ that minimizes $E[(Y - a)^2]$ is $a = E[Y]$. Let’s review one proof of that fact.

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

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

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

Let’s review one proof of that fact.

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

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

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

Let’s review one proof of that fact.

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


$$= E[(\hat{Y} + c)^2]$$
Recall that the best guess about $Y$, if we know only the distribution of $Y$, is $E[Y]$.

More precisely, the value of $a$ that minimizes $E[(Y - a)^2]$ is $a = E[Y]$. Let's review one proof of that fact.

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


$$= E[(\hat{Y} + c)^2] \text{ with } c = E[Y] - a$$
Recall that the best guess about $Y$, if we know only the distribution of $Y$, is $E[Y]$.

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

Let’s review one proof of that fact.

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

$$= E[(\hat{Y} + c)^2] \text{ with } c = E[Y] - a$$
$$= E[\hat{Y}^2 + 2\hat{Y}c + c^2]$$
Recall that the best guess about $Y$, if we know only the distribution of $Y$, is $E[Y]$.

More precisely, the value of $a$ that minimizes $E[(Y - a)^2]$ is $a = E[Y]$. Let’s review one proof of that fact.

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

$$
$$
Linear Regression: Preamble

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

More precisely, the value of $a$ that minimizes $E[(Y - a)^2]$ is $a = E[Y]$. Let’s review one proof of that fact.

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

\[
= E[(\hat{Y} + c)^2] \text{ with } c = E[Y] - a \\
= E[\hat{Y}^2 + 2\hat{Y}c + c^2] = E[\hat{Y}^2] + 2E[\hat{Y}c] + c^2 \\
= E[\hat{Y}^2] + 0 + c^2
\]
Recall that the best guess about $Y$, if we know only the distribution of $Y$, is $E[Y]$.

More precisely, the value of $a$ that minimizes $E[(Y - a)^2]$ is $a = E[Y]$. Let’s review one proof of that fact.

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


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

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

$$= E[\hat{Y}^2] + 0 + c^2 \geq E[\hat{Y}^2].$$
Recall that the best guess about $Y$, if we know only the distribution of $Y$, is $E[Y]$. More precisely, the value of $a$ that minimizes $E[(Y - a)^2]$ is $a = E[Y]$. Let’s review one proof of that fact.

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

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

Hence, $E[(Y - a)^2] \geq E[(Y - E[Y])^2], \forall a.$
Here is a picture that summarizes the calculation.

\[
E[Y] = a \cdot (Pythagoras) \cdot E[\hat{Y}] = 0
\]

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

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

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

Here is a picture that summarizes the calculation.
Here is a picture that summarizes the calculation.

\[ \hat{Y} = Y - E[Y] \]
\[ c = E[Y] - a \]

\[ E[\hat{Y}c] = 0 \Leftrightarrow \hat{Y} \perp c \]

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

(Pythagoras)
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
Linear Regression: Preamble

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 $Y$? 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 of \( X \).

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

What is the best linear function? That is our next topic. A bit later, we will consider a general function \( g(X) \).
Linear Regression: Motivation

Example 1: 100 people.

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

$$E[Y] = \mu_Y, \quad X \in \text{meters}, \quad Y \in \text{kg}.$$
Linear Regression: Motivation

Example 1: 100 people.

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

$E[Y] = Y = -114.3 + 106.5X.$ (X in meters, Y in kg.)
Example 1: 100 people.
Let \((X_n, Y_n) = (\text{height, weight})\) of person \(n\), for \(n = 1, \ldots, 100\):

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

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

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

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

\(X\) in meters, \(Y\) in kg.

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

Let \((X_n, Y_n) = (\text{height}, \text{weight})\) 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 \((X_n, Y_n) = (\text{height}, \text{weight})\) 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.
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.
Example 2: 15 people.

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

The line \(Y = a + bX\) is the linear regression.
History

Galton produced over 340 papers and books. He created the statistical concept of correlation. In an effort to reach a wider audience, Galton worked on a novel entitled Kantsaywhere. The novel described a utopia organized by a eugenic religion, designed to breed fitter and smarter humans. The lesson is that smart people can also be stupid.
Galton produced over 340 papers and books. He created the statistical concept of correlation.
Galton produced over 340 papers and books. He created the statistical concept of correlation.

In an effort to reach a wider audience, Galton worked on a novel entitled Kantsaywhere.
Galton produced over 340 papers and books. He created the statistical concept of correlation.

In an effort to reach a wider audience, Galton worked on a novel entitled Kantsaywhere. The novel described a utopia organized by a eugenic religion, designed to breed fitter and smarter humans.
Galton produced over 340 papers and books. He created the statistical concept of correlation.

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

The lesson is that smart people can also be stupid.
Covariance

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

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

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

**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:**


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

□
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.

\begin{align*}
\text{Four equally likely pairs of values} \\
&\begin{align*}
&\begin{array}{c}
&1 \\
&1 \\
&-1 \\
&-1 \\
\end{array}
&\begin{array}{c}
&1 \\
&1 \\
&-1 \\
&-1 \\
\end{array}
&\begin{array}{c}
&1 \\
&1 \\
&-1 \\
&-1 \\
\end{array}
&\begin{array}{c}
&1 \\
&1 \\
&-1 \\
&-1 \\
\end{array}
\end{align*}
\end{align*}

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

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

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

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

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

$$E[XY] = 1 \times 0.05 + 2 \times 0.25 + 3 \times 0.25 + 3 \times 3 \times 0.2 = 4.85$$

Covariance:

$$\text{cov}(X,Y) = E[XY] - E[X]E[Y] = 1.05 - 1.9 \times 2 = -3.85$$

Variance:

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

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

\[
E[X] = 1 \times 0.15 + 2 \times 0.4 + 3 \times 0.45 = 1.9
\]
\[
E[X^2] = 1^2 \times 0.15 + 2^2 \times 0.4 + 3^2 \times 0.45 = 5.8
\]
\[
E[Y] = 1 \times 0.2 + 2 \times 0.6 + 3 \times 0.2 = 2
\]
\[
E[XY] = 1 \times 0.05 + 1 \times 2 \times 0.1 + \cdots + 3 \times 3 \times 0.2 = 4.85
\]
\[
cov(X, Y) = E[XY] - E[X]E[Y] = 1.05
\]
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) \( \text{cov}(aX + bY, cU + dV) = ac \cdot \text{cov}(X, U) + ad \cdot \text{cov}(X, V) \)
  \[ + bc \cdot \text{cov}(Y, U) + bd \cdot \text{cov}(Y, V). \]

**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) $\var[X] = \text{cov}(X, X)$
(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,

$$\text{cov}(aX + bY, cU + dV) = E[(aX + bY)(cU + dV)] = ac.\text{E}[XU] + ad.\text{E}[XV] + bc.\text{E}[YU] + bd.\text{E}[YV]$$
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:**
(a)-(b)-(c) are obvious.
(d) In view of (c), one can subtract the means and assume that the RVs are zero-mean. Then,
\[ \text{cov}(aX + bY, cU + dV) = E[(aX + bY)(cU + dV)] \]
\[ = ac\cdot E[XU] + ad\cdot E[XV] + bc\cdot E[YU] + bd\cdot E[YV] \]
\[ = ac\cdot \text{cov}(X, U) + ad\cdot \text{cov}(X, V) + bc\cdot \text{cov}(Y, U) + bd\cdot \text{cov}(Y, V). \]
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.

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.
The **Linear Least Squares Estimate** is a method used to estimate the parameters of a linear model. Given two random variables $X$ and $Y$ with known distribution $\text{Pr}[X = x, Y = y]$, 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? The 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 $Pr[X = x, Y = y]$, 

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

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

$$\hat{Y} = a + bX =: L[Y|X]$$
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 $X$. 
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 $(Y - \hat{Y})^2$. 

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.
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?
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.
Non-Bayesian or Uniform?

Observe that
\[
\sum_{n=1}^{N} (Y_n - a - bX_n)^2 = E\left[(Y - a - bX)^2\right]
\]
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!
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!
Theorem

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

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

Proof 1:

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

Hence,

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

Also,

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

after a bit of algebra. (See next slide.)

Hence, by combining the two equalities,

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

Then,

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

Indeed:

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

so that

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

for some \( c, d \).

Now,

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

This shows that

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

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.

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:

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

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

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

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

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

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

This shows that $E[(Y - \hat{Y})^2] \leq E[(Y - a - bX)^2], \forall (a, b)$. Thus $\hat{Y}$ is the LLSE.
Theorem
Consider two RVs $X, Y$ with a given distribution $Pr[X = x, Y = y]$. Then,
\[
L[Y|X] = \hat{Y} = E[Y] + \frac{\text{cov}(X, Y)}{\text{var}(X)}(X - E[X]).
\]
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 $\alpha, \beta$, 

LLSE
**Theorem**

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

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

**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 $c, d$. 
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,

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

**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,

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

$$= E[(Y - \hat{Y})^2] + E[(\hat{Y} - a - bX)^2] + 0$$
**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,
$$L[Y | X] = \hat{Y} = E[Y] + \frac{\text{cov}(X, Y)}{\text{var}(X)} (X - E[X]).$$

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 $(a, b)$.
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 \( \text{var}[X] = E[(X - E[X])^2]. \)
The following picture explains the algebra:

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

We also saw that
\[
E[(Y - \hat{Y})X] = 0.
\]
In the picture, this says that
\[
Y - \hat{Y} \perp X.
\]
Hence,
\[
Y - \hat{Y}
\]
is orthogonal to the plane \(\{c + dX, c, d \in \mathbb{R}\}\).

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

That is,
\[
\hat{Y}
\]
is the projection of \(Y\) onto the plane.
We saw that $E[Y - \hat{Y}] = 0$. In the picture, this says that $Y - \hat{Y} \perp c$, for any $c$.

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

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

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

That is, $\hat{Y}$ is the projection of $Y$ onto the plane.
We saw that $E[Y - \hat{Y}] = 0$. 

A picture

The following picture explains the algebra:

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

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

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

Consequently, $Y - \hat{Y} \perp \hat{Y} - a - bX$. Pythagoras then says that $\hat{Y}$ is closer to $Y$ than $a + bX$. That is, $\hat{Y}$ is the projection of $Y$ onto the plane.
We saw that \( E[Y - \hat{Y}] = 0 \). In the picture, this says that \( Y - \hat{Y} \perp c \), for any \( c \).

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

Hence, \( Y - \hat{Y} \) is orthogonal to the plane \( \{c + dX, c, d \in \mathbb{R}\} \).
A picture

The following picture explains the algebra:

We saw that $E[Y - \hat{Y}] = 0$. In the picture, this says that $Y - \hat{Y} \perp c$, for any $c$. We also saw that $E[(Y - \hat{Y})X] = 0$. In the picture, this says that $Y - \hat{Y} \perp X$. Hence, $Y - \hat{Y}$ is orthogonal to the plane $\{c + dX, c, d \in \mathbb{R}\}$.

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

$\hat{Y} = L[Y|X]$
The following picture explains the algebra:

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

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

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

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

That is, $\hat{Y}$ is the projection of $Y$ onto the plane.
**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 2:**
First assume that $E[X] = 0$ and $E[Y] = 0$. Then,

$$g(a, b) := E[(Y - a - bX)^2] = E[Y^2] + a^2 + b^2E[X^2] - 2aE[Y] - 2bE[XY] + 2abE[X].$$

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

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

$$0 = \frac{\partial}{\partial b} g(a, b) = 2bE[X^2] - 2E[XY] \Rightarrow b = \frac{E[XY]}{E[X^2]} = \frac{\text{cov}(X, Y)}{\text{var}(X)}.$$
**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 2:**
First assume that $E[X] = 0$ and $E[Y] = 0$. 

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

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

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

$$g(a, b) := E[(Y - a - bX)^2]$$
Theorem
Consider two RVs $X$, $Y$ with a given distribution $Pr[X = x, Y = y]$. Then,

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

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

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

$$= E[Y^2 + a^2 + b^2X^2 - 2aY - 2bXY + 2abX]$$
Theorem
Consider two RVs $X, Y$ with a given distribution $Pr[X = x, Y = y]$. Then,

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

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

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

$$= E[Y^2 + a^2 + b^2X^2 - 2aY - 2bXY + 2abX]$$

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

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

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

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

$$= E[Y^2 + a^2 + b^2 X^2 - 2aY - 2bXY + 2abX]$$


$$= a^2 + E[Y^2] + b^2 E[X^2] - 2bE[XY].$$
**Theorem**
Consider two RVs $X, Y$ with a given distribution $Pr[X = x, Y = y]$. Then,

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

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

$$g(a, b) := E[(Y - a - bX)^2]$$
$$= E[Y^2 + a^2 + b^2 X^2 - 2aY - 2bXY + 2abX]$$
$$= a^2 + E[Y^2] + b^2 E[X^2] - 2bE[XY].$$

We set the derivatives of $g$ w.r.t. $a$ and $b$ equal to zero.
**Theorem**
Consider two RVs $X, Y$ with a given distribution $Pr[X = x, Y = y]$. Then,

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

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

$$g(a, b) := E[(Y - a - bX)^2]$$
$$= E[Y^2 + a^2 + b^2X^2 - 2aY - 2bXY + 2abX]$$
$$= a^2 + E[Y^2] + b^2 E[X^2] - 2bE[XY].$$

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

$$0 = \frac{\partial}{\partial a} g(a, b) = 2a$$
**Theorem**
Consider two RVs $X, Y$ with a given distribution $Pr[X = x, Y = y]$. Then,

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

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

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

$$= E[Y^2 + a^2 + b^2 X^2 - 2aY - 2bXY + 2abX]$$


$$= a^2 + E[Y^2] + b^2 E[X^2] - 2bE[XY].$$

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

$$0 = \frac{\partial}{\partial a} g(a, b) = 2a \Rightarrow a = 0.$$
**Theorem**
Consider two RVs $X$, $Y$ with a given distribution $Pr[X = x, Y = y]$.
Then,

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

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

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

$$= E[Y^2 + a^2 + b^2X^2 - 2aY - 2bXY + 2abX]$$


$$= a^2 + E[Y^2] + b^2 E[X^2] - 2bE[XY].$$

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

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

$$0 = \frac{\partial}{\partial b} g(a, b) = 2bE[X^2] - 2E[XY]$$
**Theorem**
Consider two RVs $X, Y$ with a given distribution $Pr[X = x, Y = y]$. Then,

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

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


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

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

$$0 = \frac{\partial}{\partial b} g(a, b) = 2bE[X^2] - 2E[XY] \Rightarrow b = E[XY]/E[X^2] = cov(X, Y)/var(X).$$
**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 2:**
First assume that $E[X] = 0$ and $E[Y] = 0$. Then,


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

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

$$0 = \frac{\partial}{\partial b} g(a, b) = 2bE[X^2] - 2E[XY] \Rightarrow b = E[XY]/E[X^2] = \text{cov}(X, Y)/\text{var}(X).$$
**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 2:**

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


with $c = a - E[Y] + bE[X]$.

From the first part, we know that the best values of $c$ and $b$ are $c = 0$ and $b = \frac{cov(X, Y)}{var(X)} = \frac{cov(X, Y)}{var(X)} / \frac{var(X)}{var(X)}$.

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

Hence, $a + bX = E[Y] - bE[X] + bX = E[Y] + b(X - E[X]).$
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 2:
In the general case (i.e., when $E[X]$ and $E[Y]$ may be nonzero),
Theorem
Consider two RVs $X, Y$ with a given distribution $Pr[X = x, Y = y]$. Then,

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

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

$$Y - a - bX = Y - E[Y] - (a - E[Y]) - b(X - E[X]) + bE[X]$$
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 2:
In the general case (i.e., when $E[X]$ and $E[Y]$ may be nonzero),

$$Y - a - bX = Y - E[Y] - (a - E[Y]) - b(X - E[X]) + bE[X]$$

$$= Y - E[Y] - (a - E[Y] + bE[X]) - b(X - E[X]).$$
**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 2:**
In the general case (i.e., when $E[X]$ and $E[Y]$ may be nonzero),

$$Y - a - bX = Y - E[Y] - (a - E[Y]) - b(X - E[X]) + bE[X]$$

$$= Y - E[Y] - (a - E[Y] + bE[X]) - b(X - E[X])$$

$$= Y - E[Y] - c - b(X - E[X])$$
Theorem
Consider two RVs $X, Y$ with a given distribution $Pr[X = x, Y = y]$. Then,

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

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

$$Y - a - bX = Y - E[Y] - (a - E[Y]) - b(X - E[X]) + bE[X]$$
$$= Y - E[Y] - (a - E[Y] + bE[X]) - b(X - E[X])$$
$$= Y - E[Y] - c - b(X - E[X])$$

with $c = a - E[Y] + bE[X]$. 
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 2:
In the general case (i.e., when $E[X]$ and $E[Y]$ may be nonzero),

$$Y - a - bX = Y - E[Y] - (a - E[Y]) - b(X - E[X]) + bE[X]$$

$$= Y - E[Y] - (a - E[Y] + bE[X]) - b(X - E[X])$$

$$= Y - E[Y] - c - b(X - E[X])$$

with $c = a - E[Y] + bE[X]$.
From the first part, we know that the best values of $c$ and $b$ are
**Theorem**
Consider two RVs $X$, $Y$ with a given distribution $Pr[X = x, Y = y]$. Then,

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

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

$$Y - a - bX = Y - E[Y] - (a - E[Y]) - b(X - E[X]) + bE[X]$$
$$= Y - E[Y] - (a - E[Y] + bE[X]) - b(X - E[X])$$
$$= Y - E[Y] - c - b(X - E[X])$$

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

$c = 0$ and $b = cov(X - E[X], Y - E[Y])/var(X - E[X]) = cov(X, Y)/var(X)$.
Theorem
Consider two RVs $X, Y$ with a given distribution $Pr[X = x, Y = y]$. Then,

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

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

$$Y - a - bX = Y - E[Y] - (a - E[Y]) - b(X - E[X]) + bE[X]$$

$$= Y - E[Y] - (a - E[Y] + bE[X]) - b(X - E[X])$$

$$= Y - E[Y] - c - b(X - E[X])$$

with $c = a - E[Y] + bE[X]$.

From the first part, we know that the best values of $c$ and $b$ are

$$c = 0 \text{ and } b = \frac{cov(X - E[X], Y - E[Y])}{var(X - E[X])} = \frac{cov(X, Y)}{var(X)}.$$
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 2:
In the general case (i.e., when $E[X]$ and $E[Y]$ may be nonzero),

$$Y - a - bX = Y - E[Y] - (a - E[Y]) - b(X - E[X]) + bE[X]$$
$$= Y - E[Y] - (a - E[Y] + bE[X]) - b(X - E[X])$$
$$= Y - E[Y] - c - b(X - E[X])$$

with $c = a - E[Y] + bE[X]$.

From the first part, we know that the best values of $c$ and $b$ are

$$c = 0 \text{ and } b = \frac{\text{cov}(X - E[X], Y - E[Y])}{\text{var}(X - E[X])} = \frac{\text{cov}(X, Y)}{\text{var}(X)}.$$  

Thus, $0 = c = a - E[Y] + bE[X]$, so that $a = E[Y] - bE[X]$. 
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 2:
In the general case (i.e., when $E[X]$ and $E[Y]$ may be nonzero),

$$Y - a - bX = Y - E[Y] - (a - E[Y]) - b(X - E[X]) + bE[X]$$
$$= Y - E[Y] - (a - E[Y] + bE[X]) - b(X - E[X])$$
$$= Y - E[Y] - c - b(X - E[X])$$

with $c = a - E[Y] + bE[X]$.

From the first part, we know that the best values of $c$ and $b$ are

$$c = 0 \text{ and } b = \frac{cov(X - E[X], Y - E[Y])}{var(X - E[X])} = \frac{cov(X, Y)}{var(X)}.$$

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

$$a + bX = E[Y] - bE[X] + bX = E[Y] + b(X - E[X])$$
**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 2:**
In the general case (i.e., when $E[X]$ and $E[Y]$ may be nonzero),

$$Y - a - bX = Y - E[Y] - (a - E[Y]) - b(X - E[X]) + bE[X]$$

$$= Y - E[Y] - (a - E[Y] + bE[X]) - b(X - E[X])$$

$$= Y - E[Y] - c - b(X - E[X])$$

with $c = a - E[Y] + bE[X]$.

From the first part, we know that the best values of $c$ and $b$ are

$$c = 0$$

and

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

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

$$a + bX = E[Y] - bE[X] + bX = E[Y] + b(X - E[X])$$

$$= E[Y] + \frac{cov(X, Y)}{var(X)}(X - E[X]).$$
We saw that the LLSE of $Y$ given $X$ is

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

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


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

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

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

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

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

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

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

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

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 $var(Y)$.
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? That is, what is the mean squared estimation error?

We find


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

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

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

Without observations, the estimate is $E[Y] = 0$. The error is $\text{var}(Y)$. Observing $X$ reduces the error.
We saw that

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

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

and

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

Example 1:
Linear Regression Examples

Example 2:

We find:

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

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

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

Example 2:

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

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

We find:

\[ E[X] = \]

\[ E[Y] = \]

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

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

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

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

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

We find:

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

Example 2:

We find:

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

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

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

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

Example 2:

We find:

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

Example 2:

We find:

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

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

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

Example 2:

We find:

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

Example 2:

We find:

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

We find:

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

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

\[ \text{var}[X] = E[X^2] - E[X]^2 = 1/2; \]
Linear Regression Examples

Example 2:

We find:

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

Example 2:

We find:

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

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

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

**Example 2:**

We find:

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

\[
\]

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

We find:

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

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

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

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

Example 3:

We find:

\[ E[X] = \]

\[ E[Y] = \]

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

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

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

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

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

Example 3:

We find:

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

Example 3:

We find:

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

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:

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

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

Example 3:

We find:

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

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

Example 3:

We find:

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

Example 4:

We find:

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

\[ \hat{Y} = 2.5 + 0.92(X - 3) = 1.15 + 0.45X. \]
Linear Regression Examples

Example 4:

\[
\begin{align*}
E[X] &= 3; \\
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; \\

\hat{Y} &= 2.5 + 0.292(\bar{X} - 3) = 1.15 + 0.45X.
\end{align*}
\]
Linear Regression Examples

Example 4:

We find:

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

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

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

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

\[ \hat{Y} = 2.5 + 0.92(X - 3) = 1.15 + 0.45X. \]
Linear Regression Examples

Example 4:

We find:

\[ E[X] = 3; \]

\[ E[Y] = 2.5; \]

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

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

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

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

\[ \hat{Y} = 2.5 + 0.292 \times (X - 3) = 1.15 + 0.45X. \]
We find:

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

Example 4:

We find:

\[ E[X] = 3; \ E[Y] = 2.5; \]
We find:

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

Example 4:

We find:

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

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

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

Example 4:

We find:

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

Example 4:

We find:

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

1. 95% Confidence Interval for $\mu$:
   \[ A \pm 1.96 \times \frac{\sigma}{\sqrt{n}} \]

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
   \[ E[(Y - a - bX)^2] \]
Summary

Confidence Interval; Linear Regression

1. 95%-Confidence Interval for $\mu$: $A_n \pm 4.5\sigma/\sqrt{n}$
Summary

Confidence Interval; Linear Regression

1. 95%-Confidence Interval for $\mu$: $A_n \pm 4.5\sigma / \sqrt{n}$
2. Linear Regression: $L[Y|X] = E[Y] + \frac{\text{cov}(X,Y)}{\text{var}(X)} (X - E[X])$
Summary

1. 95%-Confidence Interval for $\mu$: $A_n \pm 4.5\sigma/\sqrt{n}$
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

Confidence Interval; Linear Regression

1. 95%-Confidence Interval for $\mu$: $A_\pm 4.5\sigma/\sqrt{n}$
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 $E[(Y - a - bX)^2]$