Observations $Y = (y_i) \in \mathbb{R}^n$, $X = (x_{ij}) \in R^{n \times p}$. 

Assumptions:

1. $Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i$ ($\epsilon_i$ = error).

2. $x_{ij}$ are non-random.

3. $\epsilon_i$ are independent $N(0, \sigma^2)$.

We have $\hat{\beta} = (X^T X)^{-1} X^T Y$. What is the distribution of $\hat{\beta}$?
Observations $Y = (y_i) \in \mathbb{R}^n$, $X = (x_{ij}) \in R^{n \times p}$.

Assumptions:

1. $Y_i = \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i \quad (\epsilon_i = \text{error})$. 

What is the distribution of $\hat{\beta}$?
Observations $Y = (y_i) \in \mathbb{R}^n$, $X = (x_{ij}) \in \mathbb{R}^{n \times p}$.

Assumptions:

1. $Y_i = \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i$ \hspace{1cm} ($\epsilon_i = \text{error}$).

In other words:

$$Y = X \beta + \epsilon.$$ 

$(\beta = (\beta_1, \ldots, \beta_p)$ is a fixed unknown vector$)$
Observations $Y = (y_i) \in \mathbb{R}^n$, $X = (x_{ij}) \in R^{n \times p}$.

Assumptions:

1. $Y_i = \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i$ $\quad (\epsilon_i = \text{error})$.

In other words:

$$Y = X\beta + \epsilon.$$  

($\beta = (\beta_1, \ldots, \beta_p)$ is a \textbf{fixed unknown vector})

2. $x_{ij}$ are non-random. $\epsilon_i$ are random.
Observations $Y = (y_i) \in \mathbb{R}^n$, $X = (x_{ij}) \in R^{n \times p}$.

Assumptions:

1. $Y_i = \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i$ \hspace{1cm} ($\epsilon_i =$ error).
   
   In other words:

   $$Y = X \beta + \epsilon.$$ 

   ($\beta = (\beta_1, \ldots, \beta_p)$ is a \textbf{fixed} unknown vector)

2. $x_{ij}$ are non-random. $\epsilon_i$ are random.

3. $\epsilon_i$ are independent $N(0, \sigma^2)$.

We have

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$
Observations $Y = (y_i) \in \mathbb{R}^n$, $X = (x_{ij}) \in \mathbb{R}^{n \times p}$.

Assumptions:

1. $Y_i = \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i$ \quad ($\epsilon_i = \text{error}$).

   In other words:
   
   $$Y = X\beta + \epsilon.$$

   ($\beta = (\beta_1, \ldots, \beta_p)$ is a **fixed** unknown vector)

2. $x_{ij}$ are non-random. $\epsilon_i$ are random.

3. $\epsilon_i$ are independent $N(0, \sigma^2)$.

We have

$$\hat{\beta} = (X^T X)^{-1}X^T Y.$$

What is the distribution of $\hat{\beta}$?
Recall: $X = (X_1, \ldots, X_p) \sim N(\mu, \Sigma)$ where

- $\mu \in \mathbb{R}^p$,
- $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{p \times p}$ is positive definite,

if

$$P(X \in A) = \frac{1}{\sqrt{(2\pi)^p \det \Sigma}} \int_A e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \, dx_1 \ldots dx_p.$$
Recall: $X = (X_1, \ldots, X_p) \sim N(\mu, \Sigma)$ where
- $\mu \in \mathbb{R}^p$,
- $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{p \times p}$ is positive definite,

if

$$P(X \in A) = \frac{1}{\sqrt{(2\pi)^p \det \Sigma}} \int_A e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \, dx_1 \ldots dx_p.$$ 

Bivariate case:
Multivariate normal distribution

Recall: $X = (X_1, \ldots, X_p) \sim N(\mu, \Sigma)$ where

- $\mu \in \mathbb{R}^p$,
- $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{p \times p}$ is positive definite,

if

$$P(X \in A) = \frac{1}{\sqrt{(2\pi)^p \det \Sigma}} \int_A e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \, dx_1 \ldots dx_p.$$ 

Bivariate case:

We have

$$E(X) = \mu, \quad \text{Cov}(X_i, X_j) = \sigma_{ij}.$$
Multivariate normal distribution

Recall: $X = (X_1, \ldots, X_p) \sim N(\mu, \Sigma)$ where

- $\mu \in \mathbb{R}^p$,
- $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{p \times p}$ is positive definite,

if

$$P(X \in A) = \frac{1}{\sqrt{(2\pi)^p \det \Sigma}} \int_A e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \, dx_1 \ldots dx_p.$$ 

Bivariate case:

We have

$$E(X) = \mu, \quad \text{Cov}(X_i, X_j) = \sigma_{ij}.$$ 

If $Y = c + BX$, where $c \in \mathbb{R}^p$ and $B \in \mathbb{R}^{m \times p}$, then

$$Y \sim N(c + B\mu, B\Sigma B^T).$$
Back to our problem: $Y = X\beta + \epsilon$ where $\epsilon_i$ are iid $N(0, \sigma^2)$. We have

$$Y \sim N(X\beta, \sigma^2 I).$$
Back to our problem: \( Y = X\beta + \epsilon \) where \( \epsilon_i \) are iid \( N(0, \sigma^2) \). We have

\[
Y \sim N(X\beta, \sigma^2 I).
\]

Therefore,

\[
\hat{\beta} = (X^T X)^{-1} X^T Y \sim N(\beta, \sigma^2 (X^T X)^{-1}).
\]
Back to our problem: \( Y = X\beta + \epsilon \) where \( \epsilon_i \) are iid \( N(0, \sigma^2) \). We have

\[
Y \sim N(X\beta, \sigma^2 I).
\]

Therefore,

\[
\hat{\beta} = (X^TX)^{-1}X^TY \sim N(\beta, \sigma^2(X^TX)^{-1}).
\]

In particular,

\[
E(\hat{\beta}) = \beta.
\]

Thus, \( \hat{\beta} \) is unbiased.
We saw that $E(\hat{\beta}) = \beta$. 
We saw that $E(\hat{\beta}) = \beta$.

What happens as the sample size $n$ goes to infinity? We expect $\hat{\beta} = \hat{\beta}(n) \to \beta$. 
We saw that $E(\hat{\beta}) = \beta$.

What happens as the sample size $n$ goes to infinity? We expect $\hat{\beta} = \hat{\beta}(n) \rightarrow \beta$.

A sequence of estimators $\{\theta_n\}_{n=1}^{\infty}$ of a parameter $\theta$ is said to be consistent if $\theta_n \rightarrow \theta$ in probability ($\theta_n \overset{p}{\rightarrow} \theta$) as $n \rightarrow \infty$. 

(Without any assumptions, nothing prevents the observations to be all the same for example...)
We saw that $E(\hat{\beta}) = \beta$.

What happens as the sample size $n$ goes to infinity? We expect $\hat{\beta} = \hat{\beta}(n) \to \beta$.

A sequence of estimators $\{\theta_n\}_{n=1}^{\infty}$ of a parameter $\theta$ is said to be **consistent** if $\theta_n \to \theta$ in probability ($\theta_n \xrightarrow{p} \theta$) as $n \to \infty$.

(Recall: $\theta_n \xrightarrow{p} \theta$ if for every $\epsilon > 0$,

$$\lim_{n \to \infty} P(|\theta_n - \theta| \geq \epsilon) = 0.$$
We saw that $E(\hat{\beta}) = \beta$.

What happens as the sample size $n$ goes to infinity? We expect $\hat{\beta} = \hat{\beta}(n) \rightarrow \beta$.

A sequence of estimators $\{\theta_n\}_{n=1}^{\infty}$ of a parameter $\theta$ is said to be consistent if $\theta_n \rightarrow \theta$ in probability ($\theta_n \overset{p}{\rightarrow} \theta$) as $n \rightarrow \infty$.

(Recall: $\theta_n \overset{p}{\rightarrow} \theta$ if for every $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|\theta_n - \theta| \geq \epsilon) = 0.$$)

In order to prove that $\hat{\beta}_n$ (estimator with $n$ samples) is consistent, we will make some assumptions on the data generating model.
We saw that $E(\hat{\beta}) = \beta$.

What happens as the sample size $n$ goes to infinity? We expect $\hat{\beta} = \hat{\beta}(n) \to \beta$.

A sequence of estimators $\{\theta_n\}_{n=1}^{\infty}$ of a parameter $\theta$ is said to be **consistent** if $\theta_n \to \theta$ in probability ($\theta_n \overset{p}{\to} \theta$) as $n \to \infty$.

(Recall: $\theta_n \overset{p}{\to} \theta$ if for every $\epsilon > 0$,

$$\lim_{n \to \infty} P(|\theta_n - \theta| \geq \epsilon) = 0.$$

In order to prove that $\hat{\beta}_n$ (estimator with $n$ samples) is consistent, we will make some assumptions on the *data generating model*.

(Without any assumptions, nothing prevents the observations to be all the same for example...)
Observations: $y = (y_i) \in \mathbb{R}^n$, $X = (x_{ij}) \in \mathbb{R}^{n \times p}$. 
Observations: $y = (y_i) \in \mathbb{R}^n$, $X = (x_{ij}) \in \mathbb{R}^{n \times p}$. Let $x_i := (x_{i,1}, \ldots, x_{i,n}) \in \mathbb{R}^p$ \hspace{1em} (i = 1, \ldots, n).
Observations: \( y = (y_i) \in \mathbb{R}^n, X = (x_{ij}) \in \mathbb{R}^{n \times p}. \) Let \( x_i := (x_{i,1}, \ldots, x_{i,n}) \in \mathbb{R}^p \quad (i = 1, \ldots, n). \)

We will assume:

1. \((x_i)_{i=1}^n\) are iid random vectors.
2. \(y_i = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \epsilon_i\) where \(\epsilon_i\) are iid \(N(0, \sigma^2)\).
3. The error \(\epsilon_i\) is independent of \(x_i\).
4. \(E x_{ij}^2 < \infty\) (finite second moment).
5. \(Q = E(x_i x_i^T) \in \mathbb{R}^{p \times p}\) is invertible.
Observations: \( y = (y_i) \in \mathbb{R}^n, \ X = (x_{ij}) \in \mathbb{R}^{n \times p} \). Let \( x_i := (x_{i,1}, \ldots, x_{i,n}) \in \mathbb{R}^p \quad (i = 1, \ldots, n) \).

We will assume:

1. \((x_i)_{i=1}^n\) are iid random vectors.
2. \(y_i = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \epsilon_i\) where \(\epsilon_i\) are iid \(\mathcal{N}(0, \sigma^2)\).
3. The error \(\epsilon_i\) is independent of \(x_i\).
4. \(E x_{i,j}^2 < \infty\) (finite second moment).
5. \(Q = E(x_i x_i^T) \in \mathbb{R}^{p \times p}\) is invertible.

Under these assumptions, we have the following theorem.

**Theorem:** Let \(\hat{\beta}_n = (X^TX)^{-1}X^T y\). Then, under the above assumptions, we have
\[
\hat{\beta}_n \overset{p}{\to} \beta.
\]
Recall:

**Weak law of large numbers:** Let $(X_i)_{i=1}^\infty$ be iid random variables with finite first moment $E(|X_i|) < \infty$. Let $\mu := E(X_i)$. Then

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{p} \mu.$$
Background for the proof

Recall:

**Weak law of large numbers:** Let \((X_i)_{i=1}^{\infty}\) be iid random variables with finite first moment \(E(|X_i|) < \infty\). Let \(\mu := E(X_i)\). Then

\[
\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{p} \mu.
\]

**Continuous mapping theorem:** Let \(S, S'\) be metric spaces. Suppose \((X_i)_{i=1}^{\infty}\) are \(S\)-valued random variables such that \(X_i \xrightarrow{p} X\). Let \(g : S \to S'\). Denote by \(D_g\) the set of points in \(S\) where \(g\) is discontinuous and suppose \(P(X \in D_g) = 0\). Then \(g(X_n) \xrightarrow{p} g(X)\).
Proof of the theorem

We have

\[ \hat{\beta} = (X^T X)^{-1} X^T y = \left( \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} x_i y_i \right). \]
Proof of the theorem

We have

$$\hat{\beta} = (X^T X)^{-1} X^T y = \left( \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} x_i y_i \right).$$

Using Cauchy–Schwarz,

$$E(|x_{ij} x_{ik}|) \leq \left( E(x_{ij}^2) E(x_{ik}^2) \right)^{1/2} < \infty.$$
Proof of the theorem

We have

\[ \hat{\beta} = (X^T X)^{-1} X^T y = \left( \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} x_i y_i \right). \]

Using Cauchy–Schwarz,

\[ E(|x_{ij} x_{ik}|) \leq (E(x_{ij}^2) E(x_{ik}^2))^{1/2} < \infty. \]

In a similar way, we prove that \( E(|x_{ij} y_i|) < \infty \).
Proof of the theorem

We have

\[ \hat{\beta} = (X^T X)^{-1} X^T y = \left( \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} x_i y_i \right). \]

Using Cauchy–Schwarz,

\[ E(|x_{ij} x_{ik}|) \leq \left( E(x_{ij}^2) E(x_{ik}^2) \right)^{1/2} < \infty. \]

In a similar way, we prove that \( E(|x_{ij} y_i|) < \infty. \)

By the weak law of large numbers, we obtain

\[ \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \xrightarrow{p} E(x_i x_i^T) = Q, \]

\[ \frac{1}{n} \sum_{i=1}^{n} x_i y_i \xrightarrow{p} E(x_i y_i). \]
Using the continuous mapping theorem, we obtain

$$\hat{\beta}_n \xrightarrow{p} E(x_ix_i^T)^{-1}E(x_iy_i).$$

(Define $g : \mathbb{R}^{p \times p} \times \mathbb{R}^p \to \mathbb{R}^p$ by $g(A, b) = A^{-1}b$.)
Proof of the theorem (cont.)

Using the continuous mapping theorem, we obtain

\[ \hat{\beta}_n \overset{p}{\to} E(x_i x_i^T)^{-1} E(x_i y_i). \]

(Define \( g : \mathbb{R}^{p \times p} \times \mathbb{R}^p \to \mathbb{R}^p \) by \( g(A, b) = A^{-1}b \).)

Recall: \( y_i = x_i^T \beta + \epsilon_i \). So

\[ x_i y_i = x_i x_i^T \beta + x_i \epsilon_i. \]
Using the continuous mapping theorem, we obtain

\[ \hat{\beta}_n \xrightarrow{p} E(x_i x_i^T)^{-1}E(x_i y_i). \]

(Define \( g : \mathbb{R}^{p \times p} \times \mathbb{R}^p \rightarrow \mathbb{R}^p \) by \( g(A, b) = A^{-1}b \).)

Recall: \( y_i = x_i^T \beta + \epsilon_i \). So

\[ x_i y_i = x_i x_i^T \beta + x_i \epsilon_i. \]

Taking expectations,

\[ E(x_i y_i) = E(x_i x_i^T) \beta + E(x_i \epsilon_i). \]
Proof of the theorem (cont.)

Using the continuous mapping theorem, we obtain

$$\hat{\beta}_n \overset{p}{\to} E(x_i x_i^T)^{-1}E(x_i y_i).$$

(Define $g : \mathbb{R}^{p \times p} \times \mathbb{R}^p \to \mathbb{R}^p$ by $g(A, b) = A^{-1}b$.)

Recall: $y_i = x_i^T \beta + \epsilon_i$. So

$$x_i y_i = x_i x_i^T \beta + x_i \epsilon_i.$$  

Taking expectations,

$$E(x_i y_i) = E(x_i x_i^T) \beta + E(x_i \epsilon_i).$$

Note that $E(x_i \epsilon_i) = 0$ since $x_i$ and $\epsilon_i$ are independent by assumption.
Proof of the theorem (cont.)

Using the continuous mapping theorem, we obtain

\[ \hat{\beta}_n \xrightarrow{p} E(x_i x_i^T)^{-1} E(x_i y_i). \]

(Define \( g : \mathbb{R}^{p \times p} \times \mathbb{R}^p \to \mathbb{R}^p \) by \( g(A, b) = A^{-1}b \).)

Recall: \( y_i = x_i^T \beta + \epsilon_i \). So

\[ x_i y_i = x_i x_i^T \beta + x_i \epsilon_i. \]

Taking expectations,

\[ E(x_i y_i) = E(x_i x_i^T) \beta + E(x_i \epsilon_i). \]

Note that \( E(x_i \epsilon_i) = 0 \) since \( x_i \) and \( \epsilon_i \) are independent by assumption.

We conclude that

\[ \beta = E(x_i x_i^T)^{-1} E(x_i y_i) \]

and so \( \hat{\beta}_n \xrightarrow{p} \beta \).