MATH 637: Mathematical Techniques in Data Science
Consistency of Linear Regression

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

February 24, 2020

Distribution of regression coefficients

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

Distribution of regression coefficients

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

Assumptions:

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

Distribution of regression coefficients

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

Assumptions:

1. $Y_i = \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i \qquad (\epsilon_i = \text{error})$.
   In other words:
   $$Y = X\beta + \epsilon.$$
   ($\beta = (\beta_1, \ldots \beta_p)$ is a **fixed** unknown vector)

Distribution of regression coefficients

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

Assumptions:

1. $Y_i = \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i \qquad (\epsilon_i = \text{error})$.
   In other words:
   $$Y = X\beta + \epsilon.$$
   ($\beta = (\beta_1, \ldots \beta_p)$ is a **fixed** unknown vector)
2. x_{ij} are non-random. ϵ_i are random.

Distribution of regression coefficients

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

Assumptions:

1. $Y_i = \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i \qquad (\epsilon_i = \text{error})$.
   In other words:
   $$Y = X\beta + \epsilon.$$
   ($\beta = (\beta_1, \ldots \beta_p)$ is a **fixed** unknown vector)
2. x_{ij} are non-random. ϵ_i are random.
3. ϵ_i are independent $N(0, \sigma^2)$.

We have

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

Distribution of regression coefficients

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

1. $Y_i = \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i \qquad (\epsilon_i = \text{error})$.
   In other words:
   $$Y = X\beta + \epsilon.$$
   ($\beta = (\beta_1, \ldots \beta_p)$ is a **fixed** unknown vector)
2. x_{ij} are non-random. ϵ_i are random.
3. ϵ_i are independent $N(0, \sigma^2)$.

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

What is the distribution of $\hat{\beta}$?

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

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

Multivariate normal distribution

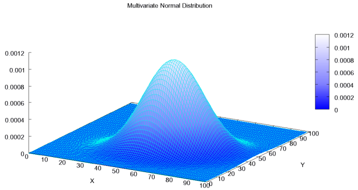
Recall: $X = (X_1, \ldots, X_p) \sim N(\mu, \Sigma)$ where

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



Multivariate normal distribution

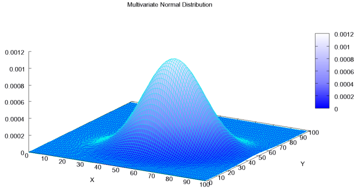
Recall: $X = (X_1, \dots, X_p) \sim N(\mu, \Sigma)$ where

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

if

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

Bivariate case:



We have

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

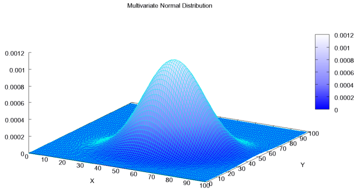
Recall: $X = (X_1, \dots, X_p) \sim N(\mu, \Sigma)$ where

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

if

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

Bivariate case:



We have

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

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

$$Y \sim N(c + B\mu, B\Sigma B^T).$$

Distribution of the regression coefficients (cont.)

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

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

Distribution of the regression coefficients (cont.)

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

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

Therefore,

$$\hat{\beta} = (X^T X)^{-1} X^T Y \sim N(\beta, \sigma^2 (X^T X)^{-1}).$$

Distribution of the regression coefficients (cont.)

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

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

Therefore,

$$\hat{\beta} = (X^T X)^{-1} X^T Y \sim N(\beta, \sigma^2 (X^T X)^{-1}).$$

In particular,

$$E(\hat{\beta}) = \beta.$$

Thus, $\hat{\beta}$ is **unbiased**.

Statistical consistency of least squares

- We saw that $E(\hat{\beta}) = \beta$.

Statistical consistency of least squares

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

Statistical consistency of least squares

- 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 θ is said to be **consistent** if $\theta_n \to \theta$ in probability ($\theta_n \xrightarrow{p} \theta$) as $n \to \infty$.

Statistical consistency of least squares

- 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 θ 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 $\epsilon > 0$,

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

Statistical consistency of least squares

- 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 θ 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 $\epsilon > 0$,

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

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

Statistical consistency of least squares

- 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 θ 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 $\epsilon > 0$,

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

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...)

Statistical consistency of least squares (cont.)

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
$\mathbf{x}_i := (x_{i,1}, \dots, x_{i,n}) \in \mathbb{R}^p \qquad (i = 1, \dots, n)$.

Statistical consistency of least squares (cont.)

Observations: $y = (y_i) \in \mathbb{R}^n$, $X = (x_{ij}) \in \mathbb{R}^{n \times p}$. Let
$\mathbf{x}_i := (x_{i,1}, \dots, x_{i,n}) \in \mathbb{R}^p \qquad (i = 1, \dots, n)$.

We will **assume**:

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

Statistical consistency of least squares (cont.)

Observations: $y = (y_i) \in \mathbb{R}^n$, $X = (x_{ij}) \in \mathbb{R}^{n\times p}$. Let
$\mathbf{x}_i := (x_{i,1}, \dots, x_{i,n}) \in \mathbb{R}^p \qquad (i = 1, \dots, n)$.

We will **assume**:

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

Under these assumptions, we have the following theorem.

Theorem: Let $\hat{\beta}_n = (X^T X)^{-1} X^T y$. Then, under the above assumptions, we have

$$\hat{\beta}_n \xrightarrow{p} \beta.$$

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

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

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

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

Proof of the theorem

We have

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

Using Cauchy–Schwarz,

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

Proof of the theorem

We have

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

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

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

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 $E(|x_{ij} y_i|) < \infty$.

By the weak law of large numbers, we obtain

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \xrightarrow{p} E(\mathbf{x}_i \mathbf{x}_i^T) = Q,$$

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i y_i \xrightarrow{p} E(\mathbf{x}_i y_i).$$

Proof of the theorem (cont.)

Using the continuous mapping theorem, we obtain

$$\hat{\beta}_n \xrightarrow{p} E(\mathbf{x}_i \mathbf{x}_i^T)^{-1} E(\mathbf{x}_i y_i).$$

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

Using the continuous mapping theorem, we obtain

$$\hat{\beta}_n \xrightarrow{p} E(\mathbf{x}_i \mathbf{x}_i^T)^{-1} E(\mathbf{x}_i y_i).$$

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

Recall: $y_i = \mathbf{x}_i^T \beta + \epsilon_i$. So

$$\mathbf{x}_i y_i = \mathbf{x}_i \mathbf{x}_i^T \beta + \mathbf{x}_i \epsilon_i.$$

Proof of the theorem (cont.)

Using the continuous mapping theorem, we obtain

$$\hat{\beta}_n \xrightarrow{p} E(\mathbf{x}_i \mathbf{x}_i^T)^{-1} E(\mathbf{x}_i y_i).$$

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

Recall: $y_i = \mathbf{x}_i^T \beta + \epsilon_i$. So

$$\mathbf{x}_i y_i = \mathbf{x}_i \mathbf{x}_i^T \beta + \mathbf{x}_i \epsilon_i.$$

Taking expectations,

$$E(\mathbf{x}_i y_i) = E(\mathbf{x}_i \mathbf{x}_i^T)\beta + E(\mathbf{x}_i \epsilon_i).$$

Using the continuous mapping theorem, we obtain

$$\hat{\beta}_n \xrightarrow{p} E(\mathbf{x}_i \mathbf{x}_i^T)^{-1} E(\mathbf{x}_i y_i).$$

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

Recall: $y_i = \mathbf{x}_i^T \beta + \epsilon_i$. So

$$\mathbf{x}_i y_i = \mathbf{x}_i \mathbf{x}_i^T \beta + \mathbf{x}_i \epsilon_i.$$

Taking expectations,

$$E(\mathbf{x}_i y_i) = E(\mathbf{x}_i \mathbf{x}_i^T)\beta + E(\mathbf{x}_i \epsilon_i).$$

Note that $E(\mathbf{x}_i \epsilon_i) = 0$ since $\mathbf{x}_i$ and ϵ_i are independent by assumption.

Proof of the theorem (cont.)

Using the continuous mapping theorem, we obtain

$$\hat{\beta}_n \xrightarrow{p} E(\mathbf{x}_i \mathbf{x}_i^T)^{-1} E(\mathbf{x}_i y_i).$$

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

Recall: $y_i = \mathbf{x}_i^T \beta + \epsilon_i$. So

$$\mathbf{x}_i y_i = \mathbf{x}_i \mathbf{x}_i^T \beta + \mathbf{x}_i \epsilon_i.$$

Taking expectations,

$$E(\mathbf{x}_i y_i) = E(\mathbf{x}_i \mathbf{x}_i^T)\beta + E(\mathbf{x}_i \epsilon_i).$$

Note that $E(\mathbf{x}_i \epsilon_i) = 0$ since $\mathbf{x}_i$ and ϵ_i are independent by assumption.

We conclude that

$$\beta = E(\mathbf{x}_i \mathbf{x}_i^T)^{-1} E(\mathbf{x}_i y_i)$$

and so $\hat{\beta}_n \xrightarrow{p} \beta$. □