Linear Models Lecture 7: Classic Normal Regression (small sample)

Robert Gulotty

University of Chicago

February 20, 2026

Using Distributional assumptions

- Statistical inference aims to make statements about unobserved parameters β, σ etc.
- In this lecture, we do so by imposing distributional assumptions on our population with a *parametric model*
- In particular, we will be using Maximum Likelihood Estimation (MLE)

Parametric Model

- A parametric model for X is a complete probability function that depends on an unknown parameter vector θ.
- In the continuous case, we can write it as a probability density function $f(x|\theta)$.
- E.g. If $X \sim N(\mu, \sigma^2)$, $f(x|\mu, \sigma^2) = \sigma^{-1}\phi((x-\mu)/\sigma)$, the parameters are $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.
- Recall, $\phi(z)$ is the density for the standard normal, $N(0,1) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})$

Parametric Model

- A parametric model for X is a complete probability function that depends on an unknown parameter vector θ.
- In the continuous case, we can write it as a probability density function $f(x|\theta)$.
- E.g. If $X \sim N(\mu, \sigma^2)$, $f(x|\mu, \sigma^2) = \sigma^{-1}\phi((x-\mu)/\sigma)$, the parameters are $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.
- Recall, $\phi(z)$ is the density for the standard normal, $N(0,1) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})$

Parametric Model

- A parametric model for X is a complete probability function that depends on an unknown parameter vector θ.
- In the continuous case, we can write it as a probability density function $f(x|\theta)$.
- E.g. If $X \sim N(\mu, \sigma^2)$, $f(x|\mu, \sigma^2) = \sigma^{-1}\phi((x-\mu)/\sigma)$, the parameters are $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.
- Recall, $\phi(z)$ is the density for the standard normal, $N(0,1) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})$

Specified Parametric Models

- A model is called **correctly specified** when there is a unique parameter value θ_0 such that $f(x|\theta_0) = f(x)$, the true data distribution.
- For example, if the true density is

$$f(x) = 2\exp(-2x)$$

  - the exponential model $f(x|\lambda) = \lambda^{-1}\exp(-x/\lambda)$ is a correctly specific model with $\lambda_0 = 1/2$
  - the lognormal model $f(x|\mu,\ \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}}\exp(-\frac{(\ln x-\mu)^2}{2\sigma^2})$ is mispecified, cannot equal $2\exp(-2x)$ under any parameter value.

Specified Parametric Models

- A model is called **correctly specified** when there is a unique parameter value θ_0 such that $f(x|\theta_0) = f(x)$, the true data distribution.
- For example, if the true density is

$$f(x) = 2\exp(-2x)$$

  - the exponential model $f(x|\lambda) = \lambda^{-1}\exp(-x/\lambda)$ is a correctly specific model with $\lambda_0 = 1/2$
  - the lognormal model $f(x|\mu,\ \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}}\exp(-\frac{(\ln x-\mu)^2}{2\sigma^2})$ is mispecified, cannot equal $2\exp(-2x)$ under any parameter value.

Specified Parametric Models

- A model is called **correctly specified** when there is a unique parameter value θ_0 such that $f(x|\theta_0) = f(x)$, the true data distribution.
- For example, if the true density is

$$f(x) = 2\exp(-2x)$$

  - the exponential model $f(x|\lambda) = \lambda^{-1}\exp(-x/\lambda)$ is a correctly specific model with $\lambda_0 = 1/2$
  - the lognormal model $f(x|\mu,\ \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}}\exp(-\frac{(\ln x-\mu)^2}{2\sigma^2})$ is mispecified, cannot equal $2\exp(-2x)$ under any parameter value.

Likelihoods

- Call $f(\theta|\boldsymbol{X})$ the **probability density function** of some model and parameters θ, given the data $\boldsymbol{X}$.
- If we reverse the order $f(\boldsymbol{X}|\theta)$ we have a *likelihood*, how probable is the data given the θ.
- Example: Binomial model of term lengths of candidates, $P(x,p) = \binom{n}{x} p^x (1-p)^{n-x}$.
- Suppose we have data $\boldsymbol{x} = \{1,\ 0,\ 1,\ 2,\ 0\}$.
- The Joint Likelihood of the data is:

$$\begin{aligned} P(\boldsymbol{x}|\ p) &= \binom{2}{1} p^1 (1-p)^1 \binom{2}{0} p^0 (1-p)^2 \binom{2}{1} p^1 (1-p)^1 \binom{2}{2} p^2 (1-p)^0 \binom{2}{0} p^0 (1-p)^2 \\ &= 4p^4(1-p)^6 \end{aligned}$$

Likelihoods

- Call $f(\theta|\boldsymbol{X})$ the **probability density function** of some model and parameters θ, given the data $\boldsymbol{X}$.
- If we reverse the order $f(\boldsymbol{X}|\theta)$ we have a *likelihood*, how probable is the data given the θ.
- Example: Binomial model of term lengths of candidates, $P(x,p) = \binom{n}{x}p^x(1-p)^{n-x}$.
- Suppose we have data $\boldsymbol{x} = \{1,\ 0,\ 1,\ 2,\ 0\}$.
- The Joint Likelihood of the data is:

$$P(\boldsymbol{x}|\ p) = \binom{2}{1}p^1(1-p)^1\binom{2}{0}p^0(1-p)^2\binom{2}{1}p^1(1-p)^1\binom{2}{2}p^2(1-p)^0\binom{2}{0}p^0(1-p)^2$$
$$= 4p^4(1-p)^6$$

Likelihoods

- Call $f(\theta|\boldsymbol{X})$ the **probability density function** of some model and parameters θ, given the data $\boldsymbol{X}$.
- If we reverse the order $f(\boldsymbol{X}|\theta)$ we have a *likelihood*, how probable is the data given the θ.
- Example: Binomial model of term lengths of candidates, $P(x, p) = \binom{n}{x} p^x (1-p)^{n-x}$.
- Suppose we have data $\boldsymbol{x} = \{1,\ 0,\ 1,\ 2,\ 0\}$.
- The Joint Likelihood of the data is:

$$
\begin{aligned}
P(\boldsymbol{x}|\ p) &= \binom{2}{1} p^1 (1-p)^1 \binom{2}{0} p^0 (1-p)^2 \binom{2}{1} p^1 (1-p)^1 \binom{2}{2} p^2 (1-p)^0 \binom{2}{0} p^0 (1-p)^2 \\
&= 4p^4 (1-p)^6
\end{aligned}
$$

Likelihoods

- Call $f(\theta|\boldsymbol{X})$ the **probability density function** of some model and parameters θ, given the data $\boldsymbol{X}$.
- If we reverse the order $f(\boldsymbol{X}|\theta)$ we have a *likelihood*, how probable is the data given the θ.
- Example: Binomial model of term lengths of candidates, $P(x, p) = \binom{n}{x} p^x (1-p)^{n-x}$.
- Suppose we have data $\boldsymbol{x} = \{1,\ 0,\ 1,\ 2,\ 0\}$.
- The Joint Likelihood of the data is:

$$P(\boldsymbol{x}|\ p) = \binom{2}{1} p^1 (1-p)^1 \binom{2}{0} p^0 (1-p)^2 \binom{2}{1} p^1 (1-p)^1 \binom{2}{2} p^2 (1-p)^0 \binom{2}{0} p^0 (1-p)^2$$
$$= 4p^4 (1-p)^6$$

Likelihoods

- Call $f(\theta|\boldsymbol{X})$ the **probability density function** of some model and parameters θ, given the data $\boldsymbol{X}$.
- If we reverse the order $f(\boldsymbol{X}|\theta)$ we have a *likelihood*, how probable is the data given the θ.
- Example: Binomial model of term lengths of candidates, $P(x,p) = \binom{n}{x} p^x (1-p)^{n-x}$.
- Suppose we have data $\boldsymbol{x} = \{1,\ 0,\ 1,\ 2,\ 0\}$.
- The Joint Likelihood of the data is:

$$P(\boldsymbol{x}|\ p) = \binom{2}{1} p^1(1-p)^1 \binom{2}{0} p^0(1-p)^2 \binom{2}{1} p^1(1-p)^1 \binom{2}{2} p^2(1-p)^0 \binom{2}{0} p^0(1-p)^2$$

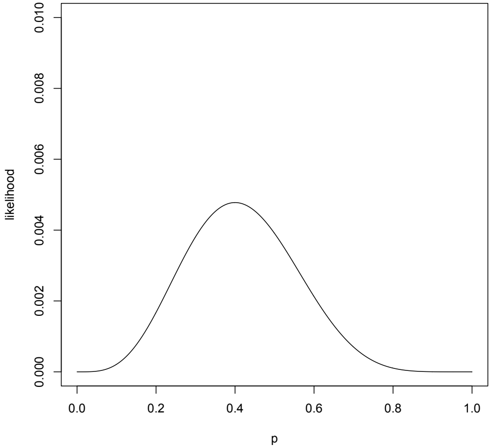
$$= 4p^4(1-p)^6$$

Likelihoods

- Call $f(\theta|\boldsymbol{X})$ the **probability density function** of some model and parameters θ, given the data $\boldsymbol{X}$.
- If we reverse the order $f(\boldsymbol{X}|\theta)$ we have a *likelihood*, how probable is the data given the θ.
- Example: Binomial model of term lengths of candidates, $P(x,p) = \binom{n}{x} p^x (1-p)^{n-x}$.
- Suppose we have data $\boldsymbol{x} = \{1,\ 0,\ 1,\ 2,\ 0\}$.
- The Joint Likelihood of the data is:

$$
\begin{aligned}
P(\boldsymbol{x}|\ p) &= \binom{2}{1} p^1 (1-p)^1 \binom{2}{0} p^0 (1-p)^2 \binom{2}{1} p^1 (1-p)^1 \binom{2}{2} p^2 (1-p)^0 \binom{2}{0} p^0 (1-p)^2 \\
&= 4p^4(1-p)^6
\end{aligned}
$$

Likelihood for Binomial x



Maximum Likelihood Estimator (from Hansen Probability)

Definition

The **maximum likelihood estimator** $\hat{\theta}$ of θ is the value that maximizes the likelihood:

$$\mathcal{L}_n(\theta) \equiv f(X_1, X_2, \ldots X_n|\theta)$$

Call $\ell_n = \sum_{i=1}^{n} \log f(X_i|\theta)$ the log-likelihood.

Maximizing the likelihood

$$
\begin{aligned}
P(\mathbf{x}|\ p) &= 4p^4(1-p)^6 \\
\frac{\partial}{\partial p} P(\mathbf{x}|\ p) &= 16p^3(1-p)^6 - 24p^4(1-p)^5 \\
&= 0 \\
3p^4(1-p)^5 &= 2p^3(1-p)^6 \\
3p &= 2(1-p) \\
5p &= 2 \\
p^* &= \frac{2}{5}
\end{aligned}
$$

Maximizing the likelihood

$$
\begin{aligned}
P(\boldsymbol{x}|\ p) &= 4p^4(1-p)^6 \\
\frac{\partial}{\partial p}P(\boldsymbol{x}|\ p) &= 16p^3(1-p)^6 - 24p^4(1-p)^5 \\
&= 0 \\
3p^4(1-p)^5 &= 2p^3(1-p)^6 \\
3p &= 2(1-p) \\
5p &= 2 \\
p^* &= \frac{2}{5}
\end{aligned}
$$

Maximizing the likelihood

$$
\begin{aligned}
P(\boldsymbol{x}|\ p) &= 4p^4(1-p)^6 \\
\frac{\partial}{\partial p}P(\boldsymbol{x}|\ p) &= 16p^3(1-p)^6 - 24p^4(1-p)^5 \\
&= 0 \\
3p^4(1-p)^5 &= 2p^3(1-p)^6 \\
3p &= 2(1-p) \\
5p &= 2 \\
p^* &= \frac{2}{5}
\end{aligned}
$$

Maximizing the likelihood

$$
\begin{aligned}
P(\mathbf{x}|\ p) &= 4p^4(1-p)^6 \\
\frac{\partial}{\partial p}P(\mathbf{x}|\ p) &= 16p^3(1-p)^6 - 24p^4(1-p)^5 \\
&= 0 \\
3p^4(1-p)^5 &= 2p^3(1-p)^6 \\
3p &= 2(1-p) \\
5p &= 2 \\
p^* &= \frac{2}{5}
\end{aligned}
$$

Maximizing the likelihood

$$
\begin{aligned}
P(\boldsymbol{x}|\ p) &= 4p^4(1-p)^6 \\
\frac{\partial}{\partial p} P(\boldsymbol{x}|\ p) &= 16p^3(1-p)^6 - 24p^4(1-p)^5 \\
&= 0 \\
3p^4(1-p)^5 &= 2p^3(1-p)^6 \\
3p &= 2(1-p) \\
5p &= 2 \\
p^* &= \frac{2}{5}
\end{aligned}
$$

Maximizing the likelihood

$$
\begin{aligned}
P(\boldsymbol{x}|\ p) &= 4p^4(1-p)^6 \\
\frac{\partial}{\partial p}P(\boldsymbol{x}|\ p) &= 16p^3(1-p)^6 - 24p^4(1-p)^5 \\
&= 0 \\
3p^4(1-p)^5 &= 2p^3(1-p)^6 \\
3p &= 2(1-p) \\
5p &= 2 \\
p^* &= \frac{2}{5}
\end{aligned}
$$

Maximizing the likelihood

$$
\begin{aligned}
P(\boldsymbol{x} \mid p) &= 4p^4(1-p)^6 \\
\frac{\partial}{\partial p} P(\boldsymbol{x} \mid p) &= 16p^3(1-p)^6 - 24p^4(1-p)^5 \\
&= 0 \\
3p^4(1-p)^5 &= 2p^3(1-p)^6 \\
3p &= 2(1-p) \\
5p &= 2 \\
p^* &= \frac{2}{5}
\end{aligned}
$$

Invariance (Useful result)

- Recall that usually, $E[g(x)] \neq g(E[x])$, that is, functions of unbiased estimators will not be unbiased.
- If $\hat{\theta}$ is the MLE of θ, then for any transformation $\beta = h(\theta)$, the MLE of β is $\hat{\beta} = h(\hat{\theta})$

Invariance Example $f(X|\lambda) = \lambda^{-1}\exp(-X/\lambda)$.

$$\mathcal{L}_n(\lambda) = \prod_{i=1}^{n} \lambda^{-1}\exp(-X_i/\lambda) = \lambda^{-n}\exp(\frac{1}{\lambda}\sum_{i=1}^{n} X_i)$$

$$\log \mathcal{L}_n(\lambda) = -n\log\lambda - \frac{1}{\lambda}\sum_{i=1}^{n} X_i$$

$$\frac{d\ell_n(\lambda)}{d\lambda} = -\frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i = n\lambda$$

The MLE is $\hat{\lambda} = \bar{X}_n$

Invariance Example $f(X|\lambda) = \lambda^{-1}\exp(-X/\lambda)$.

$$\mathcal{L}_n(\lambda) = \prod_{i=1}^{n} \lambda^{-1}\exp(-X_i/\lambda) = \lambda^{-n}\exp(\frac{1}{\lambda}\sum_{i=1}^{n} X_i)$$

$$\log \mathcal{L}_n(\lambda) = -n\log\lambda - \frac{1}{\lambda}\sum_{i=1}^{n} X_i$$

$$\frac{d\ell_n(\lambda)}{d\lambda} = -\frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i = n\lambda$$

The MLE is $\hat{\lambda} = \bar{X}_n$

Invariance Example $f(X|\lambda) = \lambda^{-1}\exp(-X/\lambda)$.

$$\mathcal{L}_n(\lambda) = \prod_{i=1}^{n} \lambda^{-1}\exp(-X_i/\lambda) = \lambda^{-n}\exp(\frac{1}{\lambda}\sum_{i=1}^{n} X_i)$$

$$\log \mathcal{L}_n(\lambda) = -n\log\lambda - \frac{1}{\lambda}\sum_{i=1}^{n} X_i$$

$$\frac{d\ell_n(\lambda)}{d\lambda} = -\frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i = n\lambda$$

The MLE is $\hat{\lambda} = \bar{X}_n$

Invariance Example $f(X|\lambda) = \lambda^{-1}\exp(-X/\lambda)$.

$$\mathcal{L}_n(\lambda) = \prod_{i=1}^{n} \lambda^{-1}\exp(-X_i/\lambda) = \lambda^{-n}\exp(\frac{1}{\lambda}\sum_{i=1}^{n} X_i)$$

$$\log \mathcal{L}_n(\lambda) = -n\log\lambda - \frac{1}{\lambda}\sum_{i=1}^{n} X_i$$

$$\frac{d\ell_n(\lambda)}{d\lambda} = -\frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i = n\lambda$$

The MLE is $\hat{\lambda} = \bar{X}_n$

Invariance Example $f(X|\lambda) = \lambda^{-1}\exp(-X/\lambda)$.

$$\mathcal{L}_n(\lambda) = \prod_{i=1}^{n} \lambda^{-1}\exp(-X_i/\lambda) = \lambda^{-n}\exp(\frac{1}{\lambda}\sum_{i=1}^{n} X_i)$$

$$\log \mathcal{L}_n(\lambda) = -n\log\lambda - \frac{1}{\lambda}\sum_{i=1}^{n} X_i$$

$$\frac{d\ell_n(\lambda)}{d\lambda} = -\frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i = n\lambda$$

The MLE is $\hat{\lambda} = \bar{X}_n$

Invariance Example $\beta = 1/\lambda$

- Set $\beta = 1/\lambda$, so $h(\lambda) = 1/\lambda$.
- The log density of this model is $\log f(x|\lambda) = \log[\beta \exp(-x\beta)] = \log \beta - x\beta$.
- The log likelihood is $n \log \beta - \beta n\bar{X}_n$
- Take the derivative with respect to β:

$$n/\hat{\beta} - n\bar{X}_n = 0$$

$$\hat{\beta} = 1/\bar{X}_n$$

Invariance Example $\beta = 1/\lambda$

- Set $\beta = 1/\lambda$, so $h(\lambda) = 1/\lambda$.
- The log density of this model is $\log f(x|\lambda) = \log[\beta \exp(-x\beta)] = \log \beta - x\beta$.
- The log likelihood is $n \log \beta - \beta n\bar{X}_n$
- Take the derivative with respect to β:

$$n/\hat{\beta} - n\bar{X}_n = 0$$

$$\hat{\beta} = 1/\bar{X}_n$$

Invariance Example $\beta = 1/\lambda$

- Set $\beta = 1/\lambda$, so $h(\lambda) = 1/\lambda$.
- The log density of this model is $\log f(x|\lambda) = \log[\beta \exp(-x\beta)] = \log \beta - x\beta$.
- The log likelihood is $n \log \beta - \beta n \bar{X}_n$
- Take the derivative with respect to β:

$$n/\hat{\beta} - n\bar{X}_n = 0$$

$$\hat{\beta} = 1/\bar{X}_n$$

Invariance Example $\beta = 1/\lambda$

- Set $\beta = 1/\lambda$, so $h(\lambda) = 1/\lambda$.
- The log density of this model is $\log f(x|\lambda) = \log[\beta \exp(-x\beta)] = \log \beta - x\beta$.
- The log likelihood is $n \log \beta - \beta n\bar{X}_n$
- Take the derivative with respect to β:

$$n/\hat{\beta} - n\bar{X}_n = 0$$

$$\hat{\beta} = 1/\bar{X}_n$$

Invariance Example $\beta = 1/\lambda$

- Set $\beta = 1/\lambda$, so $h(\lambda) = 1/\lambda$.
- The log density of this model is $\log f(x|\lambda) = \log[\beta \exp(-x\beta)] = \log \beta - x\beta$.
- The log likelihood is $n \log \beta - \beta n\bar{X}_n$
- Take the derivative with respect to β:

$$n/\hat{\beta} - n\bar{X}_n = 0$$

$$\hat{\beta} = 1/\bar{X}_n$$

Invariance Example $\beta = 1/\lambda$

- Set $\beta = 1/\lambda$, so $h(\lambda) = 1/\lambda$.
- The log density of this model is $\log f(x|\lambda) = \log[\beta \exp(-x\beta)] = \log \beta - x\beta$.
- The log likelihood is $n \log \beta - \beta n\bar{X}_n$
- Take the derivative with respect to β:

$$n/\hat{\beta} - n\bar{X}_n = 0$$

$$\hat{\beta} = 1/\bar{X}_n$$

Score: the slope of the log likelihood

- The **likelihood score** is the derivative of the log-likelihood function:

$$S_n(\theta) = \frac{\partial}{\partial\theta}\ell_n(\theta)$$

- The score is a function of θ and tells us how sensitive the log-likelihood is to the parameter, and equals zero at the optimum.
- The **efficient score** is the derivative of the log likelihood for a single observation, evaluated at $\boldsymbol{x} = X_1, X_2, \ldots X_n$ and the true parameter vector

$$S = \frac{\partial}{\partial\theta}\log f(\boldsymbol{x}|\theta_0)$$

- The efficient score fixes θ at the true value θ_0.

Score: the slope of the log likelihood

- The **likelihood score** is the derivative of the log-likelihood function:

$$S_n(\theta) = \frac{\partial}{\partial\theta}\ell_n(\theta)$$

- The score is a function of θ and tells us how sensitive the log-likelihood is to the parameter, and equals zero at the optimum.
- The **efficient score** is the derivative of the log likelihood for a single observation, evaluated at $\boldsymbol{x} = X_1, X_2, \ldots X_n$ and the true parameter vector

$$S = \frac{\partial}{\partial\theta} \log f(\boldsymbol{x}|\theta_0)$$

- The efficient score fixes θ at the true value θ_0.

Score: the slope of the log likelihood

- The **likelihood score** is the derivative of the log-likelihood function:

$$S_n(\theta) = \frac{\partial}{\partial \theta} \ell_n(\theta)$$

- The score is a function of θ and tells us how sensitive the log-likelihood is to the parameter, and equals zero at the optimum.
- The **efficient score** is the derivative of the log likelihood for a single observation, evaluated at $\boldsymbol{x} = X_1, X_2, \ldots X_n$ and the true parameter vector

$$S = \frac{\partial}{\partial \theta} \log f(\boldsymbol{x}|\theta_0)$$

- The efficient score fixes θ at the true value θ_0.

Score: the slope of the log likelihood

- The **likelihood score** is the derivative of the log-likelihood function:

$$S_n(\theta) = \frac{\partial}{\partial \theta} \ell_n(\theta)$$

- The score is a function of θ and tells us how sensitive the log-likelihood is to the parameter, and equals zero at the optimum.
- The **efficient score** is the derivative of the log likelihood for a single observation, evaluated at $\boldsymbol{x} = X_1, X_2, \ldots X_n$ and the true parameter vector

$$S = \frac{\partial}{\partial \theta} \log f(\boldsymbol{x}|\theta_0)$$

- The efficient score fixes θ at the true value θ_0.

Hessian: the curvature of the log likelihood

- The **likelihood Hessian** is the negative second derivative:

$$\mathscr{H}_n(\theta) = -\frac{\partial^2}{\partial\theta\partial\theta'}\ell_n(\theta)$$

- The Hessian matrix is used to calculate the variance.

Fisher Information

- The **Fisher information** is the variance of the efficient score (score evaluated at true θ_0).

$$\mathcal{I}_\theta = \mathbb{E}[SS']$$

- The **expected Hessian** is the expectation of the Hessian for a single observation:

$$\mathcal{H}_\theta = -\mathbb{E}[\frac{\partial^2}{\partial\theta\partial\theta'} \log f(X|\theta)]$$

- If the model is correctly specified:

$$\mathcal{I}_\theta = \mathcal{H}_\theta$$

Fisher Information

- The **Fisher information** is the variance of the efficient score (score evaluated at true θ_0).

$$\mathcal{I}_\theta = \mathbb{E}[SS']$$

- The **expected Hessian** is the expectation of the Hessian for a single observation:

$$\mathcal{H}_\theta = -\mathbb{E}[\frac{\partial^2}{\partial\theta\partial\theta'} \log f(X|\theta)]$$

- If the model is correctly specified:

$$\mathcal{I}_\theta = \mathcal{H}_\theta$$

Fisher Information

- The **Fisher information** is the variance of the efficient score (score evaluated at true θ_0).

$$\mathcal{I}_\theta = \mathbb{E}[SS']$$

- The **expected Hessian** is the expectation of the Hessian for a single observation:

$$\mathcal{H}_\theta = -\mathbb{E}[\frac{\partial^2}{\partial\theta\partial\theta'} \log f(X|\theta)]$$

- If the model is correctly specified:

$$\mathcal{I}_\theta = \mathcal{H}_\theta$$

Cramér-Rao Lower Bound

- The term efficient refers to an estimator which has minimum variance.
- If $\tilde{\theta}$ is an unbiased estimator of θ, then $var[\tilde{\theta}] \geq (n\mathcal{I}_\theta)^{-1}$

Variance Estimators

- The sample Hessian Estimator depends on calculating the second derivatives of the log-likelihood:

$$\hat{\mathcal{H}}_\theta = \frac{1}{n}\sum_{i=1}^{n} -\frac{\partial}{\partial\theta\partial\theta'} \log f(X_i|\hat{\theta})$$

$$\hat{\mathbf{V}}_1 = \hat{\mathcal{H}}_\theta^{-1}$$

- The Outer Product Estimator is based on the Fisher Information:

$$\hat{\mathcal{I}}_\theta = \frac{1}{n}\sum_{i=1}^{n}(\frac{\partial}{\partial\theta} \log f(X_i|\hat{\theta}))(\frac{\partial}{\partial\theta} \log f(X_i|\hat{\theta}))'$$

$$\hat{\mathbf{V}}_2 = \hat{\mathcal{I}}_\theta^{-1}$$

Variance Estimators

- The sample Hessian Estimator depends on calculating the second derivatives of the log-likelihood:

$$\hat{\mathcal{H}}_\theta = \frac{1}{n}\sum_{i=1}^{n} -\frac{\partial}{\partial\theta\partial\theta'} \log f(X_i|\hat{\theta})$$

$$\hat{\mathbf{V}}_1 = \hat{\mathcal{H}}_\theta^{-1}$$

- The Outer Product Estimator is based on the Fisher Information:

$$\hat{\mathcal{I}}_\theta = \frac{1}{n}\sum_{i=1}^{n}(\frac{\partial}{\partial\theta} \log f(X_i|\hat{\theta}))(\frac{\partial}{\partial\theta} \log f(X_i|\hat{\theta}))'$$

$$\hat{\mathbf{V}}_2 = \hat{\mathcal{I}}_\theta^{-1}$$

Example: Normal with known mean, unknown variance

$X \sim N(0, \theta)$, where $\theta \equiv \sigma^2$. $\mathbb{E}[X] = 0$, $\mathbb{E}[X^2] = \theta_0$, $\mathbb{E}[X^4] = 3\theta_0^2$.

The density is:

$$f(x|\theta) = \frac{1}{(2\pi\theta)^{1/2}} \exp(-\frac{x^2}{2\theta})$$

The log density is:

$$\log f(x|\theta) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\theta) - \frac{x^2}{2\theta}$$

the first and second derivatives are:

$$\frac{d}{d\theta}\log f(x|\theta) = -\frac{1}{2\theta} + \frac{x^2}{2\theta^2} = \frac{x^2 - \theta}{2\theta^2}$$

$$\frac{d^2}{d\theta^2}\log f(x|\theta) = \frac{1}{2\theta^2} - \frac{x^2}{\theta^3}$$

Example: Normal log likelihood

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\theta) - \frac{1}{2\theta}\sum_{i=1}^{n} X_i^2$$

$$\frac{d}{d\theta}\ell_n(\theta) = -\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n} X_i^2$$

The MLE is $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$

Example: Normal with known mean, unknown variance

$$\frac{d}{d\theta} \log f(x|\theta) = \frac{x^2 - \theta}{2\theta^2}$$

The efficient score $S = \frac{x^2 - \theta_0}{2\theta_0^2}$

$$var[S] = \frac{\mathbb{E}[(X^2 - \theta_0)^2]}{4\theta_0^4} = \frac{\mathbb{E}[X^4 - 2X^2\theta_0 + \theta_0^2]}{4\theta_0^4} = \frac{3\theta_0^2 - 2\theta_0^2 + \theta_0^2}{4\theta_0^4} = \frac{1}{2\theta_0^2}$$

expected Hessian

$$\mathcal{H}_\theta = \mathbb{E}[-\frac{d}{d\theta^2}\log f(X|\theta_0)] = \mathbb{E}[-\frac{d}{d\theta}\frac{x^2-\theta_0}{2\theta_0^2}] = \mathbb{E}[\frac{2x^2-\theta_0}{2\theta_0^3}] = \frac{1}{2\theta_0^2}$$

$$\mathcal{I}_\theta = var[S] = \frac{1}{2\theta_0^2} = \mathcal{H}_\theta$$

The total Fisher information for n observations is $I_n(\theta_0) = n \times \mathcal{I}_\theta = \frac{n}{2\theta_0^2}$

The variance of $\hat{\theta}$ is $\frac{1}{I_n(\theta_0)} = 2\theta_0^2/n$

So we can get the plug in estimator for the standard error as $\sqrt{2\hat{\theta}_0^2/n}$.

expected Hessian

```
for(i in 1:100000){
 x <- rnorm(100, sd=sqrt(5))
thetahat[i] <- sum(x^2)/100
diff[i] <- thetahat[i]-5
se[i] <- sqrt(2*thetahat[i]^2/100) }
mean(thetahat)
5.002307
sqrt(var(diff))
0.7102136
mean(se)
0.707433
```

Robust Variance Estimator

- Under misspecification, $\mathscr{I}_\theta \neq \mathscr{H}(\theta)$
- A consistent estimator for the variance is:

$$\hat{\boldsymbol{V}} = \hat{\mathscr{H}}^{-1}\hat{\mathscr{I}}\hat{\mathscr{H}}^{-1}$$

- This is calculated by the sandwich package.

Summary: The MLE Toolkit

We now have four key ingredients:

1. **Likelihood** $\rightarrow$ a model-based measure of how well parameters fit data.
2. **Score** $\rightarrow$ first derivative of ℓ; equals zero at the MLE.
3. **Fisher Information** $\rightarrow$ curvature of ℓ; determines how precisely we can estimate θ.
4. **Cramér-Rao bound** $\rightarrow$ no unbiased estimator can beat $(n\mathcal{I}_\theta)^{-1}$.

Next: apply this machinery to the normal linear regression model and see that OLS *is* the MLE.

Summary: The MLE Toolkit

We now have four key ingredients:

1. **Likelihood** $\rightarrow$ a model-based measure of how well parameters fit data.
2. **Score** $\rightarrow$ first derivative of ℓ; equals zero at the MLE.
3. **Fisher Information** $\rightarrow$ curvature of ℓ; determines how precisely we can estimate θ.
4. **Cramér-Rao bound** $\rightarrow$ no unbiased estimator can beat $(n\mathcal{I}_\theta)^{-1}$.

Next: apply this machinery to the normal linear regression model and see that OLS *is* the MLE.

Summary: The MLE Toolkit

We now have four key ingredients:

1. **Likelihood** $\rightarrow$ a model-based measure of how well parameters fit data.
2. **Score** $\rightarrow$ first derivative of ℓ; equals zero at the MLE.
3. **Fisher Information** $\rightarrow$ curvature of ℓ; determines how precisely we can estimate θ.
4. **Cramér-Rao bound** $\rightarrow$ no unbiased estimator can beat $(n\mathcal{I}_\theta)^{-1}$.

Next: apply this machinery to the normal linear regression model and see that OLS *is* the MLE.

Summary: The MLE Toolkit

We now have four key ingredients:

1. **Likelihood** $\rightarrow$ a model-based measure of how well parameters fit data.
2. **Score** $\rightarrow$ first derivative of ℓ; equals zero at the MLE.
3. **Fisher Information** $\rightarrow$ curvature of ℓ; determines how precisely we can estimate θ.
4. **Cramér-Rao bound** $\rightarrow$ no unbiased estimator can beat $(n\mathcal{I}_\theta)^{-1}$.

Next: apply this machinery to the normal linear regression model and see that OLS *is* the MLE.

Summary: The MLE Toolkit

We now have four key ingredients:

1. **Likelihood** $\rightarrow$ a model-based measure of how well parameters fit data.
2. **Score** $\rightarrow$ first derivative of ℓ; equals zero at the MLE.
3. **Fisher Information** $\rightarrow$ curvature of ℓ; determines how precisely we can estimate θ.
4. **Cramér-Rao bound** $\rightarrow$ no unbiased estimator can beat $(n\mathcal{I}_\theta)^{-1}$.

Next: apply this machinery to the normal linear regression model and see that OLS *is* the MLE.

Summary: The MLE Toolkit

We now have four key ingredients:

1. **Likelihood** $\to$ a model-based measure of how well parameters fit data.
2. **Score** $\to$ first derivative of ℓ; equals zero at the MLE.
3. **Fisher Information** $\to$ curvature of ℓ; determines how precisely we can estimate θ.
4. **Cramér-Rao bound** $\to$ no unbiased estimator can beat $(n\mathcal{I}_\theta)^{-1}$.

Next: apply this machinery to the normal linear regression model and see that OLS *is* the MLE.

Likelihood methods for linear model

- The Normal Regression model assumes that $y \sim N(\mu_Y, \sigma^2)$, or that $e \sim N(0, \sigma^2)$.

$$
\begin{aligned}
f(y|\boldsymbol{x}) &= \frac{1}{(2\pi\sigma^2)^{1/2}} exp\left(-\frac{1}{2\sigma^2}(y - \boldsymbol{x}'\boldsymbol{\beta})^2\right) \\
f(y_1, \dots y_n|\boldsymbol{x}_1, \dots \boldsymbol{x}_n) &= \prod_{i=1}^{n} f(y_i|x_i) \\
&= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} exp\left(-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\right) \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\right) \\
&\equiv \mathcal{L}_n(\beta, \sigma^2)
\end{aligned}
$$

Log Likelihood of the normal linear model

$$log\mathcal{L}_n(\beta,\sigma^2) = -\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i'\beta)^2 \equiv \ell_n(\beta,\sigma^2)$$

Classic Normal Regression Model

- The Classic Normal Regression Model consists of the following assumptions:
  1. $\boldsymbol{y} = \boldsymbol{X'}\beta + \boldsymbol{e}$
  2. $\boldsymbol{e}|\boldsymbol{X} \sim N(\boldsymbol{0},\ \sigma^2\boldsymbol{I})$
     - or equivalently: $\boldsymbol{y} \sim N(\boldsymbol{X}\beta,\ \sigma^2\boldsymbol{I})$
  3. rank$(\boldsymbol{X}) = K$.

  In this model, $N(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$ is the multivariate normal density function whose pdf is:

$$f(\boldsymbol{e}) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}\boldsymbol{e}'\boldsymbol{e})$$

Background on Multivariate Normal

- A multivariate normal's parameters are a vector of means and a variance covariance matrix.
- Linear functions of a multinormal vector $\boldsymbol{y}$ are also normal.
- If $\boldsymbol{z} = \boldsymbol{g} + \boldsymbol{H}\boldsymbol{y}$, where $\boldsymbol{g}$ and $\boldsymbol{H}$ are non-random, and $\boldsymbol{H}$ has full row rank, then

$$\boldsymbol{z} \sim N(\boldsymbol{g} + \boldsymbol{H}\boldsymbol{\mu},\ \boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}')$$

- In the case of MVN, independence is the same as uncorrelated.

Bivariate Normal CEF

- Given two random variables, they are distributed bivariate normal if $\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}\right)$
- Call $X^* = \frac{X-\mu_X}{\sigma_X}$, $Y^* = \frac{Y-\mu_Y}{\sigma_Y}$

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} exp\left(-\frac{1}{2}(X^*)^2\right)$$

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}\sigma_X\sqrt{2\pi}\sigma_Y}\sqrt{1-\rho^2} exp\left(-\frac{1}{2[1-\rho^2]}\left[(X^*)^2 - 2\rho(X^*)(Y^*) + (Y^*)^2\right]\right)$$

Bivariate Normal CEF

- Given two random variables, they are distributed bivariate normal if
$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}\right)$
- Call $X^* = \frac{X-\mu_X}{\sigma_X}$, $Y^* = \frac{Y-\mu_Y}{\sigma_Y}$

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} exp\left(-\frac{1}{2}(X^*)^2\right)$$

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}\sigma_X\sqrt{2\pi}\sigma_Y}\sqrt{1-\rho^2} exp\left(-\frac{1}{2[1-\rho^2]}\left[(X^*)^2 - 2\rho(X^*)(Y^*) + (Y^*)^2\right]\right)$$

Bivariate Normal CEF

- Given two random variables, they are distributed bivariate normal if $\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}\right)$
- Call $X^* = \frac{X-\mu_X}{\sigma_X}$, $Y^* = \frac{Y-\mu_Y}{\sigma_Y}$

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} exp\left(-\frac{1}{2}(X^*)^2\right)$$

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}\sigma_X\sqrt{2\pi}\sigma_Y}\sqrt{1-\rho^2} exp\left(-\frac{1}{2[1-\rho^2]}\left[(X^*)^2 - 2\rho(X^*)(Y^*) + (Y^*)^2\right]\right)$$

Bivariate Normal CEF

- Given two random variables, they are distributed bivariate normal if $\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}\right)$
- Call $X^* = \frac{X-\mu_X}{\sigma_X}$, $Y^* = \frac{Y-\mu_Y}{\sigma_Y}$

$$
\begin{aligned}
f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\
f_X(x) &= \frac{1}{\sqrt{2\pi}\sigma_X} exp\left(-\frac{1}{2}(X^*)^2\right) \\
f_{X,Y}(x,y) &= \frac{1}{\sqrt{2\pi}\sigma_X\sqrt{2\pi}\sigma_Y}\sqrt{1-\rho^2} exp\left(-\frac{1}{2[1-\rho^2]}\left[(X^*)^2 - 2\rho(X^*)(Y^*) + (Y^*)^2\right]\right)
\end{aligned}
$$

Bivariate Normal CEF: Derivation

Dividing the joint by the marginal and completing the square:

$$f_{Y|X}(y|x) \propto \exp\left(-\frac{1}{2(1-\rho^2)}(Y^* - \rho X^*)^2\right)$$
$$= \exp\left(-\frac{1}{2}\left(\frac{Y - \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X)}{\sigma_Y\sqrt{1-\rho^2}}\right)^2\right)$$

This is the kernel of a normal density with:

- Conditional mean: $\mathbb{E}[Y|X = x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) = \mu_Y + \frac{\text{Cov}(Y,X)}{\text{Var}(X)}(x - \mu_X)$
- Conditional variance: $\sigma_Y^2(1 - \rho^2)$

Key insight: Under joint normality, the CEF is linear and the regression coefficient is $\beta = \text{Cov}(Y, X)/\text{Var}(X)$. This is a classical motivation for linear regression.

Bivariate Normal CEF: Derivation

Dividing the joint by the marginal and completing the square:

$$
\begin{aligned}
f_{Y|X}(y|x) &\propto \exp\left(-\frac{1}{2(1-\rho^2)}\left(Y^* - \rho X^*\right)^2\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{Y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X)}{\sigma_Y\sqrt{1-\rho^2}}\right)^2\right)
\end{aligned}
$$

This is the kernel of a normal density with:

- Conditional mean: $\mathbb{E}[Y|X=x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X) = \mu_Y + \frac{\mathrm{Cov}(Y,X)}{\mathrm{Var}(X)}(x-\mu_X)$
- Conditional variance: $\sigma_Y^2(1-\rho^2)$

Key insight: Under joint normality, the CEF is linear and the regression coefficient is $\beta = \mathrm{Cov}(Y,X)/\mathrm{Var}(X)$. This is a classical motivation for linear regression.

Bivariate Normal CEF: Derivation

Dividing the joint by the marginal and completing the square:

$$f_{Y|X}(y|x) \propto \exp\left(-\frac{1}{2(1-\rho^2)}(Y^* - \rho X^*)^2\right)$$
$$= \exp\left(-\frac{1}{2}\left(\frac{Y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X)}{\sigma_Y\sqrt{1-\rho^2}}\right)^2\right)$$

This is the kernel of a normal density with:

- Conditional mean: $\mathbb{E}[Y|X=x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X) = \mu_Y + \frac{\text{Cov}(Y,X)}{\text{Var}(X)}(x-\mu_X)$
- Conditional variance: $\sigma_Y^2(1-\rho^2)$

Key insight: Under joint normality, the CEF is linear and the regression coefficient is $\beta = \text{Cov}(Y,X)/\text{Var}(X)$. This is a classical motivation for linear regression.

Bivariate Normal CEF: Derivation

Dividing the joint by the marginal and completing the square:

$$f_{Y|X}(y|x) \propto \exp\left(-\frac{1}{2(1-\rho^2)}\left(Y^* - \rho X^*\right)^2\right)$$
$$= \exp\left(-\frac{1}{2}\left(\frac{Y - \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X)}{\sigma_Y\sqrt{1-\rho^2}}\right)^2\right)$$

This is the kernel of a normal density with:

- Conditional mean: $\mathbb{E}[Y|X=x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) = \mu_Y + \frac{\text{Cov}(Y,X)}{\text{Var}(X)}(x - \mu_X)$
- Conditional variance: $\sigma_Y^2(1-\rho^2)$

Key insight: Under joint normality, the CEF is linear and the regression coefficient is $\beta = \text{Cov}(Y,X)/\text{Var}(X)$. This is a classical motivation for linear regression.

Roadmap: From Likelihood to Inference

- We have the normal linear model: $\boldsymbol{y}|\boldsymbol{X} \sim N(\boldsymbol{X}\beta,\ \sigma^2\boldsymbol{I})$.
- Now we solve for the MLE and show it equals OLS.
- Then we connect the score of the normal model to moment conditions — the bridge to GMM.
- Finally, the distributional results (t, χ^2, F) that make exact small-sample inference possible.

Roadmap: From Likelihood to Inference

- We have the normal linear model: $\boldsymbol{y}|\boldsymbol{X} \sim N(\boldsymbol{X}\beta,\ \sigma^2\boldsymbol{I})$.
- Now we solve for the MLE and show it equals OLS.
- Then we connect the score of the normal model to moment conditions — the bridge to GMM.
- Finally, the distributional results (t, χ^2, F) that make exact small-sample inference possible.

Roadmap: From Likelihood to Inference

- We have the normal linear model: $\boldsymbol{y}|\boldsymbol{X} \sim N(\boldsymbol{X}\beta,\ \sigma^2\boldsymbol{I})$.
- Now we solve for the MLE and show it equals OLS.
- Then we connect the score of the normal model to moment conditions — the bridge to GMM.
- Finally, the distributional results (t, χ^2, F) that make exact small-sample inference possible.

Roadmap: From Likelihood to Inference

- We have the normal linear model: $\boldsymbol{y}|\boldsymbol{X} \sim N(\boldsymbol{X}\beta,\ \sigma^2\boldsymbol{I})$.
- Now we solve for the MLE and show it equals OLS.
- Then we connect the score of the normal model to moment conditions — the bridge to GMM.
- Finally, the distributional results (t, χ^2, F) that make exact small-sample inference possible.

Solving for the MLE

- Given the likelihood, we can solve for the MLE $(\hat{\beta}_{MLE},\ \hat{\sigma}^2_{MLE})$

$$\frac{\partial}{\partial \beta}[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\beta)^2]$$

$$\frac{2}{2\hat{\sigma}^2_{MLE}}\sum_{i=1}^{n} X_i(Y_i - \mathbf{x}_i'\hat{\beta}_{MLE}) = 0$$

$$\frac{\partial}{\partial \sigma^2}[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\beta)^2]$$

$$-\frac{n}{2}\frac{1}{\hat{\sigma}^2_{mle}} + \frac{1}{2\hat{\sigma}^4_{mle}}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\hat{\beta}_{MLE})^2 = 0$$

Solving for the MLE

- Given the likelihood, we can solve for the MLE $(\hat{\beta}_{MLE}, \ \hat{\sigma}^2_{MLE})$

$$\frac{\partial}{\partial \beta}[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\beta)^2]$$

$$\frac{2}{2\hat{\sigma}^2_{MLE}}\sum_{i=1}^{n}X_i(Y_i - \mathbf{x}_i'\hat{\beta}_{MLE}) = 0$$

$$\frac{\partial}{\partial \sigma^2}[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\beta)^2]$$

$$-\frac{n}{2}\frac{1}{\hat{\sigma}^2_{mle}} + \frac{1}{2\hat{\sigma}^4_{mle}}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\hat{\beta}_{MLE})^2 = 0$$

Solving for the MLE

- Given the likelihood, we can solve for the MLE $(\hat{\beta}_{MLE}, \hat{\sigma}^2_{MLE})$

$$\frac{\partial}{\partial\beta}[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\beta)^2]$$

$$\frac{2}{2\hat{\sigma}^2_{MLE}}\sum_{i=1}^{n}X_i(Y_i - \mathbf{x}_i'\hat{\beta}_{MLE}) = 0$$

$$\frac{\partial}{\partial\sigma^2}[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\beta)^2]$$

$$-\frac{n}{2}\frac{1}{\hat{\sigma}^2_{mle}} + \frac{1}{2\hat{\sigma}^4_{mle}}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\hat{\beta}_{MLE})^2 = 0$$

Solving for the MLE

- Given the likelihood, we can solve for the MLE $(\hat{\beta}_{MLE},\ \hat{\sigma}^2_{MLE})$

$$\frac{\partial}{\partial\beta}[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\beta)^2]$$

$$\frac{2}{2\hat{\sigma}^2_{MLE}}\sum_{i=1}^{n}X_i(Y_i - \mathbf{x}_i'\hat{\beta}_{MLE}) = 0$$

$$\frac{\partial}{\partial\sigma^2}[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\beta)^2]$$

$$-\frac{n}{2}\frac{1}{\hat{\sigma}^2_{mle}} + \frac{1}{2\hat{\sigma}^4_{mle}}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\hat{\beta}_{MLE})^2 = 0$$

Overall Likelihood

- $\hat{\beta}_{MLE} = (\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i')^{-1} (\sum_{i=1}^{n} \boldsymbol{x}_i Y_i) = \hat{\beta}_{ols}$.
- $\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i' \hat{\beta}_{mle})^2 = \hat{\sigma}_{ols}$.
- $\log \mathcal{L}(\hat{\beta}_{MLE}, \ \hat{\sigma}^2_{MLE}) = -\frac{n}{2} \log(2\pi\hat{\sigma}^2_{mle}) - n/2$
- You will see the "log likelihood" reported as a measure of fit, logLik()
- The Akaike's AIC is -2log-likehood$+2n_{par}$, weighing model performance versus complexity.
- AIC gives an asymptotically unbiased estimator of the expected relative Kullback-Leibler divergence when approximating an unknown distribution and a given model.

Overall Likelihood

- $\hat{\beta}_{MLE} = (\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i')^{-1} (\sum_{i=1}^{n} \mathbf{x}_i Y_i) = \hat{\beta}_{ols}$.
- $\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i' \hat{\beta}_{mle})^2 = \hat{\sigma}_{ols}$.
- $\log \mathcal{L}(\hat{\beta}_{MLE}, \hat{\sigma}^2_{MLE}) = -\frac{n}{2} \log(2\pi \hat{\sigma}^2_{mle}) - n/2$
- You will see the "log likelihood" reported as a measure of fit, logLik()
- The Akaike's AIC is -2log-likehood$+2n_{par}$, weighing model performance versus complexity.
- AIC gives an asymptotically unbiased estimator of the expected relative Kullback-Leibler divergence when approximating an unknown distribution and a given model.

Overall Likelihood

- $\hat{\beta}_{MLE} = (\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i')^{-1} (\sum_{i=1}^{n} \mathbf{x}_i Y_i) = \hat{\beta}_{ols}$.
- $\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i' \hat{\beta}_{mle})^2 = \hat{\sigma}_{ols}$.
- $\log \mathcal{L}(\hat{\beta}_{MLE},\ \hat{\sigma}^2_{MLE}) = -\frac{n}{2} \log(2\pi \hat{\sigma}^2_{mle}) - n/2$
- You will see the "log likelihood" reported as a measure of fit, logLik()
- The Akaike's AIC is -2log-likehood$+2n_{par}$, weighing model performance versus complexity.
- AIC gives an asymptotically unbiased estimator of the expected relative Kullback-Leibler divergence when approximating an unknown distribution and a given model.

Overall Likelihood

- $\hat{\beta}_{MLE} = (\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i')^{-1}(\sum_{i=1}^{n} \mathbf{x}_i Y_i) = \hat{\beta}_{ols}$.
- $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\hat{\beta}_{mle})^2 = \hat{\sigma}_{ols}$.
- $\log \mathcal{L}(\hat{\beta}_{MLE},\ \hat{\sigma}^2_{MLE}) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2_{mle}) - n/2$
- You will see the "log likelihood" reported as a measure of fit, logLik()
- The Akaike's AIC is -2log-likehood$+2n_{par}$, weighing model performance versus complexity.
- AIC gives an asymptotically unbiased estimator of the expected relative Kullback-Leibler divergence when approximating an unknown distribution and a given model.

Overall Likelihood

- $\hat{\beta}_{MLE} = (\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i')^{-1}(\sum_{i=1}^{n} \mathbf{x}_i Y_i) = \hat{\beta}_{ols}$.
- $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\hat{\beta}_{mle})^2 = \hat{\sigma}_{ols}$.
- $\log \mathcal{L}(\hat{\beta}_{MLE},\ \hat{\sigma}^2_{MLE}) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2_{mle}) - n/2$
- You will see the "log likelihood" reported as a measure of fit, logLik()
- The Akaike's AIC is -2log-likehood$+2n_{par}$, weighing model performance versus complexity.
- AIC gives an asymptotically unbiased estimator of the expected relative Kullback-Leibler divergence when approximating an unknown distribution and a given model.

Overall Likelihood

- $\hat{\beta}_{MLE} = (\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i')^{-1} (\sum_{i=1}^{n} \mathbf{x}_i Y_i) = \hat{\beta}_{ols}$.
- $\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i' \hat{\beta}_{mle})^2 = \hat{\sigma}_{ols}$.
- $\log \mathcal{L}(\hat{\beta}_{MLE}, \ \hat{\sigma}^2_{MLE}) = -\frac{n}{2} \log(2\pi\hat{\sigma}^2_{mle}) - n/2$
- You will see the "log likelihood" reported as a measure of fit, logLik()
- The Akaike's AIC is -2log-likehood$+2n_{par}$, weighing model performance versus complexity.
- AIC gives an asymptotically unbiased estimator of the expected relative Kullback-Leibler divergence when approximating an unknown distribution and a given model.

R Example: Log-Likelihood and AIC

```
data(swiss)
mod1 <- lm(Fertility ~ Education, data = swiss)
mod2 <- lm(Fertility ~ Education + Agriculture,
           data = swiss)
mod3 <- lm(Fertility ~ Education + Agriculture
           + Catholic + Infant.Mortality, data=swiss)
# Log-likelihoods (higher = better fit)
sapply(list(mod1, mod2, mod3), logLik)
#  -168.3   -166.0   -155.3
# AIC = -2*logLik + 2*k (lower = better)
sapply(list(mod1, mod2, mod3), AIC)
#   342.5    340.0    322.6
```

AIC penalizes complexity: mod3 wins because the likelihood improvement outweighs the $2k$ penalty.

Score of the Normal Regression Model (Hansen 5.14)

The likelihood scores are the derivatives of the log-likelihood:

$$\frac{\partial}{\partial \beta}\ell_n(\beta, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n} X_i(Y_i - X_i'\beta) = \frac{1}{\sigma^2}\boldsymbol{X'e}$$

$$\frac{\partial}{\partial \sigma^2}\ell_n(\beta, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2$$

Setting the score for β to zero:

$$\frac{1}{\sigma^2}\boldsymbol{X'e} = 0 \iff \boldsymbol{X'e} = 0 \iff \boldsymbol{X'}(\boldsymbol{Y} - \boldsymbol{X}\beta) = 0$$

These are exactly the OLS normal equations! The MLE for β equals the OLS estimator because the score FOC is proportional to the least squares FOC.

Score of the Normal Regression Model (Hansen 5.14)

The likelihood scores are the derivatives of the log-likelihood:

$$\frac{\partial}{\partial \beta}\ell_n(\beta, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n} X_i(Y_i - X_i'\beta) = \frac{1}{\sigma^2}\boldsymbol{X'e}$$

$$\frac{\partial}{\partial \sigma^2}\ell_n(\beta, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2$$

Setting the score for β to zero:

$$\frac{1}{\sigma^2}\boldsymbol{X'e} = 0 \quad \Longleftrightarrow \quad \boldsymbol{X'e} = 0 \quad \Longleftrightarrow \quad \boldsymbol{X'(Y - X\beta)} = 0$$

These are exactly the OLS normal equations! The MLE for β equals the OLS estimator because the score FOC is proportional to the least squares FOC.

Score of the Normal Regression Model (Hansen 5.14)

The likelihood scores are the derivatives of the log-likelihood:

$$\frac{\partial}{\partial \beta}\ell_n(\beta, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n} X_i(Y_i - X_i'\beta) = \frac{1}{\sigma^2}\boldsymbol{X'e}$$

$$\frac{\partial}{\partial \sigma^2}\ell_n(\beta, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2$$

Setting the score for β to zero:

$$\frac{1}{\sigma^2}\boldsymbol{X'e} = 0 \quad \Longleftrightarrow \quad \boldsymbol{X'e} = 0 \quad \Longleftrightarrow \quad \boldsymbol{X'(Y - X\beta)} = 0$$

These are exactly the OLS normal equations! The MLE for β equals the OLS estimator because the score FOC is proportional to the least squares FOC.

Score of the Normal Regression Model (Hansen 5.14)

The likelihood scores are the derivatives of the log-likelihood:

$$\frac{\partial}{\partial\beta}\ell_n(\beta,\sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n} X_i(Y_i - X_i'\beta) = \frac{1}{\sigma^2}\boldsymbol{X'e}$$

$$\frac{\partial}{\partial\sigma^2}\ell_n(\beta,\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2$$

Setting the score for β to zero:

$$\frac{1}{\sigma^2}\boldsymbol{X'e} = 0 \quad \Longleftrightarrow \quad \boldsymbol{X'e} = 0 \quad \Longleftrightarrow \quad \boldsymbol{X'(Y - X\beta)} = 0$$

These are exactly the OLS normal equations! The MLE for β equals the OLS estimator because the score FOC is proportional to the least squares FOC.

Scores as Moment Conditions

The score for β evaluated at the true parameter is:

$$S_i(\beta_0) = \frac{\partial}{\partial \beta} \log f(Y_i | X_i, \beta_0, \sigma^2) = \frac{1}{\sigma^2} X_i e_i$$

The population moment condition from the CEF is:

$$\mathbb{E}[X_i e_i] = 0$$

Key observation: Setting the score to zero in the sample,

$$\frac{1}{n}\sum_{i=1}^{n} S_i(\hat{\beta}) = 0 \quad \Longleftrightarrow \quad \frac{1}{n}\sum_{i=1}^{n} X_i \hat{e}_i = 0$$

is the same as solving the sample moment condition $\frac{1}{n}\sum X_i(Y_i - X_i'\hat{\beta}) = 0$.

MLE, OLS, and method of moments all give the *same* estimator for β in the normal linear model.

Scores as Moment Conditions

The score for β evaluated at the true parameter is:

$$S_i(\beta_0) = \frac{\partial}{\partial \beta} \log f(Y_i | X_i, \beta_0, \sigma^2) = \frac{1}{\sigma^2} X_i e_i$$

The population moment condition from the CEF is:

$$\mathbb{E}[X_i e_i] = 0$$

Key observation: Setting the score to zero in the sample,

$$\frac{1}{n}\sum_{i=1}^{n} S_i(\hat{\beta}) = 0 \quad \Longleftrightarrow \quad \frac{1}{n}\sum_{i=1}^{n} X_i \hat{e}_i = 0$$

is the same as solving the sample moment condition $\frac{1}{n}\sum X_i(Y_i - X_i'\hat{\beta}) = 0$.

MLE, OLS, and method of moments all give the *same* estimator for β in the normal linear model.

Scores as Moment Conditions

The score for β evaluated at the true parameter is:

$$S_i(\beta_0) = \frac{\partial}{\partial \beta} \log f(Y_i | X_i, \beta_0, \sigma^2) = \frac{1}{\sigma^2} X_i e_i$$

The population moment condition from the CEF is:

$$\mathbb{E}[X_i e_i] = 0$$

Key observation: Setting the score to zero in the sample,

$$\frac{1}{n}\sum_{i=1}^{n} S_i(\hat{\beta}) = 0 \quad \Longleftrightarrow \quad \frac{1}{n}\sum_{i=1}^{n} X_i \hat{e}_i = 0$$

is the same as solving the sample moment condition $\frac{1}{n}\sum X_i(Y_i - X_i'\hat{\beta}) = 0$.

MLE, OLS, and method of moments all give the *same* estimator for β in the normal linear model.

Scores as Moment Conditions

The score for β evaluated at the true parameter is:

$$S_i(\beta_0) = \frac{\partial}{\partial \beta} \log f(Y_i | X_i, \beta_0, \sigma^2) = \frac{1}{\sigma^2} X_i e_i$$

The population moment condition from the CEF is:

$$\mathbb{E}[X_i e_i] = 0$$

Key observation: Setting the score to zero in the sample,

$$\frac{1}{n} \sum_{i=1}^{n} S_i(\hat{\beta}) = 0 \quad \Longleftrightarrow \quad \frac{1}{n} \sum_{i=1}^{n} X_i \hat{e}_i = 0$$

is the same as solving the sample moment condition $\frac{1}{n} \sum X_i(Y_i - X_i' \hat{\beta}) = 0$.

MLE, OLS, and method of moments all give the *same* estimator for β in the normal linear model.

Information Matrix of Normal Regression (Hansen 5.14)

The Fisher information matrix for the normal regression model is:

$$\mathcal{I} = \text{var}\begin{bmatrix} \frac{\partial}{\partial\beta}\ell(\beta,\sigma^2) \\ \frac{\partial}{\partial\sigma^2}\ell(\beta,\sigma^2) \end{bmatrix} = \begin{pmatrix} \frac{1}{\sigma^2}\boldsymbol{X'X} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2\sigma^4} \end{pmatrix}$$

- The matrix is block diagonal: estimation of β and σ^2 are independent.
- The Cramér-Rao lower bound for β is:

$$\mathcal{I}_\beta^{-1} = \sigma^2(\boldsymbol{X'X})^{-1}$$

  which is exactly the variance of $\hat{\beta}_{OLS}$.
- OLS achieves the Cramér-Rao bound — it is efficient among all unbiased estimators under normality.

Information Matrix of Normal Regression (Hansen 5.14)

The Fisher information matrix for the normal regression model is:

$$\mathcal{I} = \text{var}\begin{bmatrix} \frac{\partial}{\partial\beta}\ell(\beta,\sigma^2) \\ \frac{\partial}{\partial\sigma^2}\ell(\beta,\sigma^2) \end{bmatrix} = \begin{pmatrix} \frac{1}{\sigma^2}\boldsymbol{X'X} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2\sigma^4} \end{pmatrix}$$

- The matrix is block diagonal: estimation of β and σ^2 are independent.
- The Cramér-Rao lower bound for β is:

$$\mathcal{I}_\beta^{-1} = \sigma^2(\boldsymbol{X'X})^{-1}$$

  which is exactly the variance of $\hat{\beta}_{OLS}$.
- OLS achieves the Cramér-Rao bound — it is efficient among all unbiased estimators under normality.

Information Matrix of Normal Regression (Hansen 5.14)

The Fisher information matrix for the normal regression model is:

$$\mathcal{I} = \text{var}\begin{bmatrix} \frac{\partial}{\partial \beta}\ell(\beta, \sigma^2) \\ \frac{\partial}{\partial \sigma^2}\ell(\beta, \sigma^2) \end{bmatrix} = \begin{pmatrix} \frac{1}{\sigma^2}\boldsymbol{X'X} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2\sigma^4} \end{pmatrix}$$

- The matrix is block diagonal: estimation of β and σ^2 are independent.
- The Cramér-Rao lower bound for β is:

  $$\mathcal{I}_\beta^{-1} = \sigma^2(\boldsymbol{X'X})^{-1}$$

  which is exactly the variance of $\hat{\beta}_{OLS}$.
- OLS achieves the Cramér-Rao bound — it is efficient among all unbiased estimators under normality.

From Scores to the Sandwich (connecting to Lecture 6)

Under **correct specification** ($e \sim N(0, \sigma^2)$):

$$\mathcal{I}_\theta = \mathcal{H}_\theta \quad \Longrightarrow \quad V = \mathcal{H}^{-1} = \mathcal{I}^{-1} = \sigma^2(\boldsymbol{X'X})^{-1}$$

One formula suffices. This is the classical OLS variance.

Under **misspecification** (heteroskedasticity, non-normality):

$$\mathcal{I}_\theta \neq \mathcal{H}_\theta \quad \Longrightarrow \quad V = \mathcal{H}^{-1}\mathcal{I}\mathcal{H}^{-1}$$

- The "bread" $\mathcal{H}^{-1} \to (\boldsymbol{X'X})^{-1}$
- The "meat" $\mathcal{I} \to \boldsymbol{X'DX}$ where $D = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$
- The sandwich: $(\boldsymbol{X'X})^{-1}(\boldsymbol{X'DX})(\boldsymbol{X'X})^{-1}$

This is the heteroskedasticity-robust variance from Lecture 6 (HC0–HC3)!

The "sandwich" estimator is the likelihood-based variance under misspecification.

From Scores to the Sandwich (connecting to Lecture 6)

Under **correct specification** ($e \sim N(0, \sigma^2)$):

$$\mathcal{I}_\theta = \mathcal{H}_\theta \quad \Longrightarrow \quad V = \mathcal{H}^{-1} = \mathcal{I}^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

One formula suffices. This is the classical OLS variance.

Under **misspecification** (heteroskedasticity, non-normality):

$$\mathcal{I}_\theta \neq \mathcal{H}_\theta \quad \Longrightarrow \quad V = \mathcal{H}^{-1}\mathcal{I}\mathcal{H}^{-1}$$

- The "bread" $\mathcal{H}^{-1} \to (\boldsymbol{X}'\boldsymbol{X})^{-1}$
- The "meat" $\mathcal{I} \to \boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}$ where $D = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$
- The sandwich: $(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}$

This is the heteroskedasticity-robust variance from Lecture 6 (HC0–HC3)!

The "sandwich" estimator is the likelihood-based variance under misspecification.

From Scores to the Sandwich (connecting to Lecture 6)

Under **correct specification** ($e \sim N(0, \sigma^2)$):

$$\mathcal{I}_\theta = \mathcal{H}_\theta \quad \Longrightarrow \quad V = \mathcal{H}^{-1} = \mathcal{I}^{-1} = \sigma^2(\boldsymbol{X'X})^{-1}$$

One formula suffices. This is the classical OLS variance.

Under **misspecification** (heteroskedasticity, non-normality):

$$\mathcal{I}_\theta \neq \mathcal{H}_\theta \quad \Longrightarrow \quad V = \mathcal{H}^{-1}\mathcal{I}\mathcal{H}^{-1}$$

- The "bread" $\mathcal{H}^{-1} \to (\boldsymbol{X'X})^{-1}$
- The "meat" $\mathcal{I} \to \boldsymbol{X'DX}$ where $D = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$
- The sandwich: $(\boldsymbol{X'X})^{-1}(\boldsymbol{X'DX})(\boldsymbol{X'X})^{-1}$

This is the heteroskedasticity-robust variance from Lecture 6 (HC0–HC3)!

The "sandwich" estimator is the likelihood-based variance under misspecification.

From Scores to the Sandwich (connecting to Lecture 6)

Under **correct specification** ($e \sim N(0, \sigma^2)$):

$$\mathcal{I}_\theta = \mathcal{H}_\theta \quad \Longrightarrow \quad V = \mathcal{H}^{-1} = \mathcal{I}^{-1} = \sigma^2(\boldsymbol{X'X})^{-1}$$

One formula suffices. This is the classical OLS variance.

Under **misspecification** (heteroskedasticity, non-normality):

$$\mathcal{I}_\theta \neq \mathcal{H}_\theta \quad \Longrightarrow \quad V = \mathcal{H}^{-1}\mathcal{I}\mathcal{H}^{-1}$$

- The "bread" $\mathcal{H}^{-1} \to (\boldsymbol{X'X})^{-1}$
- The "meat" $\mathcal{I} \to \boldsymbol{X'DX}$ where $D = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$
- The sandwich: $(\boldsymbol{X'X})^{-1}(\boldsymbol{X'DX})(\boldsymbol{X'X})^{-1}$

This is the heteroskedasticity-robust variance from Lecture 6 (HC0–HC3)!
The "sandwich" estimator is the likelihood-based variance under misspecification.

Preview: From MLE Scores to GMM

Maximum likelihood solves k score equations in k unknowns: $\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(Y_i|X_i,\theta)=0$

Method of moments solves m moment conditions in k unknowns: $\frac{1}{n}\sum_{i=1}^{n}g(Y_i,X_i,\theta)=0$

- When $m=k$ and $g=$ score: GMM = MLE.
- When $m>k$: GMM uses an *optimal weighting matrix* to combine them.
- IV example: $\mathbb{E}[Z_ie_i]=0$ gives more moment conditions than parameters.

Takeaway: MLE $\subset$ GMM. Score equations are one set of moment conditions. Robust SEs arise when the likelihood is misspecified.

Preview: From MLE Scores to GMM

Maximum likelihood solves k score equations in k unknowns: $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(Y_i|X_i, \theta) = 0$

Method of moments solves m moment conditions in k unknowns: $\frac{1}{n}\sum_{i=1}^{n} g(Y_i, X_i, \theta) = 0$

- When $m = k$ and $g =$ score: GMM = MLE.
- When $m > k$: GMM uses an *optimal weighting matrix* to combine them.
- IV example: $\mathbb{E}[Z_i e_i] = 0$ gives more moment conditions than parameters.

Takeaway: MLE $\subset$ GMM. Score equations are one set of moment conditions. Robust SEs arise when the likelihood is misspecified.

Preview: From MLE Scores to GMM

Maximum likelihood solves k score equations in k unknowns: $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(Y_i|X_i, \theta) = 0$

Method of moments solves m moment conditions in k unknowns: $\frac{1}{n}\sum_{i=1}^{n} g(Y_i, X_i, \theta) = 0$

- When $m = k$ and g = score: GMM = MLE.
- When $m > k$: GMM uses an *optimal weighting matrix* to combine them.
- IV example: $\mathbb{E}[Z_i e_i] = 0$ gives more moment conditions than parameters.

Takeaway: MLE $\subset$ GMM. Score equations are one set of moment conditions. Robust SEs arise when the likelihood is misspecified.

Preview: From MLE Scores to GMM

Maximum likelihood solves k score equations in k unknowns: $\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(Y_i|X_i,\theta) = 0$

Method of moments solves m moment conditions in k unknowns: $\frac{1}{n}\sum_{i=1}^{n} g(Y_i, X_i, \theta) = 0$

- When $m = k$ and $g =$ score: GMM = MLE.
- When $m > k$: GMM uses an *optimal weighting matrix* to combine them.
- IV example: $\mathbb{E}[Z_i e_i] = 0$ gives more moment conditions than parameters.

Takeaway: MLE $\subset$ GMM. Score equations are one set of moment conditions. Robust SEs arise when the likelihood is misspecified.

Preview: From MLE Scores to GMM

Maximum likelihood solves k score equations in k unknowns: $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(Y_i|X_i,\theta) = 0$

Method of moments solves m moment conditions in k unknowns: $\frac{1}{n}\sum_{i=1}^{n} g(Y_i, X_i, \theta) = 0$

- When $m = k$ and $g =$ score: GMM $=$ MLE.
- When $m > k$: GMM uses an *optimal weighting matrix* to combine them.
- IV example: $\mathbb{E}[Z_i e_i] = 0$ gives more moment conditions than parameters.

Takeaway: MLE $\subset$ GMM. Score equations are one set of moment conditions. Robust SEs arise when the likelihood is misspecified.

Summary: The Score-Based View

Concept	Formula	Role
Score	$S_i = \frac{1}{\sigma^2} X_i e_i$	FOC $\rightarrow$ OLS normal eqns
Information	$\mathcal{I}_\beta = \frac{1}{\sigma^2} \boldsymbol{X}'\boldsymbol{X}$	Precision of $\hat{\beta}$
CR Bound	$\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$	OLS achieves it
Sandwich	$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$	Robust to misspec.
GMM link	Score = moment cond.	MLE $\subset$ GMM

Next: exact finite-sample distributions under normality — $\hat{\beta}$, s^2, t, and F.

Summary: The Score-Based View

Concept	Formula	Role
Score	$S_i = \frac{1}{\sigma^2} X_i e_i$	FOC $\rightarrow$ OLS normal eqns
Information	$\mathcal{I}_\beta = \frac{1}{\sigma^2} \boldsymbol{X}'\boldsymbol{X}$	Precision of $\hat{\beta}$
CR Bound	$\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$	OLS achieves it
Sandwich	$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$	Robust to misspec.
GMM link	Score = moment cond.	MLE $\subset$ GMM

Next: exact finite-sample distributions under normality — $\hat{\beta}$, s^2, t, and F.

Summary: The Score-Based View

Concept	**Formula**	**Role**
Score	$S_i = \frac{1}{\sigma^2} X_i e_i$	FOC $\to$ OLS normal eqns
Information	$\mathcal{I}_\beta = \frac{1}{\sigma^2} \boldsymbol{X'X}$	Precision of $\hat{\beta}$
CR Bound	$\sigma^2(\boldsymbol{X'X})^{-1}$	OLS achieves it
Sandwich	$(\boldsymbol{X'X})^{-1}\boldsymbol{X'DX}(\boldsymbol{X'X})^{-1}$	Robust to misspec.
GMM link	Score = moment cond.	MLE $\subset$ GMM

Next: exact finite-sample distributions under normality — $\hat{\beta}$, s^2, t, and F.

Summary: The Score-Based View

Concept	Formula	Role
Score	$S_i = \frac{1}{\sigma^2} X_i e_i$	FOC $\rightarrow$ OLS normal eqns
Information	$\mathcal{I}_\beta = \frac{1}{\sigma^2} \boldsymbol{X'X}$	Precision of $\hat{\beta}$
CR Bound	$\sigma^2 (\boldsymbol{X'X})^{-1}$	OLS achieves it
Sandwich	$(\boldsymbol{X'X})^{-1} \boldsymbol{X'DX} (\boldsymbol{X'X})^{-1}$	Robust to misspec.
GMM link	Score = moment cond.	MLE $\subset$ GMM

Next: exact finite-sample distributions under normality — $\hat{\beta}$, s^2, t, and F.

Summary: The Score-Based View

Concept	Formula	Role
Score	$S_i = \frac{1}{\sigma^2} X_i e_i$	FOC $\rightarrow$ OLS normal eqns
Information	$\mathcal{I}_\beta = \frac{1}{\sigma^2} \boldsymbol{X}'\boldsymbol{X}$	Precision of $\hat{\beta}$
CR Bound	$\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$	OLS achieves it
Sandwich	$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$	Robust to misspec.
GMM link	Score = moment cond.	MLE $\subset$ GMM

Next: exact finite-sample distributions under normality — $\hat{\beta}$, s^2, t, and F.

Summary: The Score-Based View

Concept	Formula	Role
Score	$S_i = \frac{1}{\sigma^2} X_i e_i$	FOC $\to$ OLS normal eqns
Information	$\mathcal{I}_\beta = \frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X}$	Precision of $\hat{\beta}$
CR Bound	$\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$	OLS achieves it
Sandwich	$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$	Robust to misspec.
GMM link	Score $=$ moment cond.	MLE $\subset$ GMM

Next: exact finite-sample distributions under normality — $\hat{\beta}$, s^2, t, and F.

Proof of normality of $\hat{\boldsymbol{\beta}}$ in CNRM

We can show that $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta},\ \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{e})$$

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$$

$$\hat{\beta} = \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$$

$$\hat{\beta} - \beta = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$$

$$\text{Rank}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \text{Rank}(\boldsymbol{X}) = K$$

So $\hat{\beta} - \beta$ is a full-row-rank linear transformation of the multinormal vector $\boldsymbol{e}$

$$\hat{\beta} - \beta \sim N(\boldsymbol{0},\ \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

Proof of normality of $\hat{\beta}$ in CNRM

We can show that $\hat{\beta} \sim N(\beta,\ \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{e})$$

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$$

$$\hat{\beta} = \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$$

$$\hat{\beta} - \beta = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$$

$$\text{Rank}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \text{Rank}(\boldsymbol{X}) = K$$

So $\hat{\beta} - \beta$ is a full-row-rank linear transformation of the multinormal vector $\boldsymbol{e}$

$$\hat{\beta} - \beta \sim N(\boldsymbol{0},\ \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

Proof of normality of $\hat{\beta}$ in CNRM

We can show that $\hat{\beta} \sim N(\beta,\ \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{e})$$

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$$

$$\hat{\beta} = \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$$

$$\hat{\beta} - \beta = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$$

$$\text{Rank}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \text{Rank}(\boldsymbol{X}) = K$$

So $\hat{\beta} - \beta$ is a full-row-rank linear transformation of the multinormal vector $\boldsymbol{e}$

$$\hat{\beta} - \beta \sim N(\boldsymbol{0},\ \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

Proof of normality of $\hat{\beta}$ in CNRM

We can show that $\hat{\beta} \sim N(\beta,\ \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{e})$$

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$$

$$\hat{\beta} = \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$$

$$\hat{\beta} - \beta = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$$

$$\text{Rank}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \text{Rank}(\boldsymbol{X}) = K$$

So $\hat{\beta} - \beta$ is a full-row-rank linear transformation of the multinormal vector $\boldsymbol{e}$

$$\hat{\beta} - \beta \sim N(\boldsymbol{0},\ \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

Proof of normality of $\hat{\beta}$ in CNRM

We can show that $\hat{\beta} \sim N(\beta,\ \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$

$$\begin{aligned}
\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\
\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{e}) \\
\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e} \\
\hat{\beta} &= \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e} \\
\hat{\beta} - \beta &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}
\end{aligned}$$

$$\text{Rank}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \text{Rank}(\boldsymbol{X}) = K$$

So $\hat{\beta} - \beta$ is a full-row-rank linear transformation of the multinormal vector $\boldsymbol{e}$

$$\hat{\beta} - \beta \sim N(\mathbf{0},\ \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

Distribution of Residuals (Hansen 5.7)

Recall $\hat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{e}$ where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

Since $\boldsymbol{e}|\boldsymbol{X} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ and $\hat{\boldsymbol{e}}$ is a linear function of $\boldsymbol{e}$:

$$\hat{\boldsymbol{e}}|\boldsymbol{X} \sim N(\boldsymbol{0},\ \sigma^2 \boldsymbol{M})$$

Key fact: $\hat{\beta}$ and $\hat{\boldsymbol{e}}$ are independent (conditional on $\boldsymbol{X}$).

Proof: Their joint covariance is

$$\mathrm{Cov}(\hat{\beta} - \beta,\ \hat{\boldsymbol{e}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \cdot \sigma^2 \boldsymbol{I} \cdot \boldsymbol{M} = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} \underbrace{\boldsymbol{X}'\boldsymbol{M}}_{=\boldsymbol{0}} = \boldsymbol{0}$$

Since they are jointly normal and uncorrelated, they are independent.

Why this matters: It means $s^2 = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}/(n-k)$ is independent of $\hat{\beta}$. This is what allows us to form the t-statistic as a ratio of independent normal and χ^2 components.

Distribution of Residuals (Hansen 5.7)

Recall $\hat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{e}$ where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

Since $\boldsymbol{e}|\boldsymbol{X} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ and $\hat{\boldsymbol{e}}$ is a linear function of $\boldsymbol{e}$:

$$\hat{\boldsymbol{e}}|\boldsymbol{X} \sim N(\boldsymbol{0},\ \sigma^2\boldsymbol{M})$$

Key fact: $\hat{\beta}$ and $\hat{\boldsymbol{e}}$ are independent (conditional on $\boldsymbol{X}$).

Proof: Their joint covariance is

$$\text{Cov}(\hat{\beta} - \beta,\ \hat{\boldsymbol{e}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \cdot \sigma^2\boldsymbol{I} \cdot \boldsymbol{M} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\underbrace{\boldsymbol{X}'\boldsymbol{M}}_{=\boldsymbol{0}} = \boldsymbol{0}$$

Since they are jointly normal and uncorrelated, they are independent.

Why this matters: It means $s^2 = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}/(n-k)$ is independent of $\hat{\beta}$. This is what allows us to form the t-statistic as a ratio of independent normal and χ^2 components.

Distribution of Residuals (Hansen 5.7)

Recall $\hat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{e}$ where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

Since $\boldsymbol{e}|\boldsymbol{X} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ and $\hat{\boldsymbol{e}}$ is a linear function of $\boldsymbol{e}$:

$$\hat{\boldsymbol{e}}|\boldsymbol{X} \sim N(\boldsymbol{0},\ \sigma^2 \boldsymbol{M})$$

Key fact: $\hat{\beta}$ and $\hat{\boldsymbol{e}}$ are independent (conditional on $\boldsymbol{X}$).

Proof: Their joint covariance is

$$\text{Cov}(\hat{\beta} - \beta,\ \hat{\boldsymbol{e}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \cdot \sigma^2 \boldsymbol{I} \cdot \boldsymbol{M} = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} \underbrace{\boldsymbol{X}'\boldsymbol{M}}_{=\boldsymbol{0}} = \boldsymbol{0}$$

Since they are jointly normal and uncorrelated, they are independent.

Why this matters: It means $s^2 = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}/(n-k)$ is independent of $\hat{\beta}$. This is what allows us to form the t-statistic as a ratio of independent normal and χ^2 components.

Distribution of residual sum of squares $s^2_{\hat{e}} = \frac{\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}}{(N-K)}$ in CNRM

Thm: Let $\boldsymbol{z} \sim N(\boldsymbol{0},\ \boldsymbol{I})$, let $\boldsymbol{A}$ be idempotent. Then $\boldsymbol{z}'\boldsymbol{A}\boldsymbol{z} \sim \chi^2(\nu)$ where $\nu = \text{Rank}(\boldsymbol{A})$.

$$\hat{\boldsymbol{e}}|\boldsymbol{X} \sim N(\boldsymbol{0},\ \sigma^2\boldsymbol{I})$$
$$\hat{\boldsymbol{e}}/\sigma|\boldsymbol{X} \sim N(\boldsymbol{0},\ \boldsymbol{I})$$

Recall

$$\begin{aligned}(N-K)s^2_{\hat{e}} &= \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} \\ &= \hat{\boldsymbol{e}}\boldsymbol{M}\hat{\boldsymbol{e}} \\ \frac{\hat{\boldsymbol{e}}'}{\sigma}\boldsymbol{M}\frac{\hat{\boldsymbol{e}}}{\sigma} &\sim \chi^2(N-K) \qquad (\boldsymbol{M} \text{ is idempotent})\end{aligned}$$

That is, in a linear regression model, $\frac{(N-K)s^2_{\hat{e}}}{\sigma^2} \sim \chi^2_{n-k}$

Distribution of residual sum of squares $s_{\hat{e}}^2 = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{(N-K)}$ in CNRM

Thm: Let $\boldsymbol{z} \sim N(\mathbf{0},\ \boldsymbol{I})$, let $\boldsymbol{A}$ be idempotent. Then $\boldsymbol{z}'\boldsymbol{A}\boldsymbol{z} \sim \chi^2(\nu)$ where $\nu = \text{Rank}(\boldsymbol{A})$.

$$\hat{\mathbf{e}}|\boldsymbol{X} \sim N(\mathbf{0},\ \sigma^2\boldsymbol{I})$$
$$\hat{\mathbf{e}}/\sigma|\boldsymbol{X} \sim N(\mathbf{0},\ \boldsymbol{I})$$

Recall

$$(N-K)s_{\hat{e}}^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}}$$
$$= \hat{\mathbf{e}}\boldsymbol{M}\hat{\mathbf{e}}$$
$$\frac{\hat{\mathbf{e}}'}{\sigma}\boldsymbol{M}\frac{\hat{\mathbf{e}}}{\sigma} \sim \chi^2(N-K) \qquad (\boldsymbol{M} \text{ is idempotent})$$

That is, in a linear regression model, $\frac{(N-K)s_{\hat{e}}^2}{\sigma^2} \sim \chi^2_{n-k}$

Distribution of residual sum of squares $s_{\hat{e}}^2 = \frac{\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}}{(N-K)}$ in CNRM

Thm: Let $\boldsymbol{z} \sim N(\boldsymbol{0},\ \boldsymbol{I})$, let $\boldsymbol{A}$ be idempotent. Then $\boldsymbol{z}'\boldsymbol{A}\boldsymbol{z} \sim \chi^2(\nu)$ where $\nu = \text{Rank}(\boldsymbol{A})$.

$$\hat{\boldsymbol{e}}|\boldsymbol{X} \sim N(\boldsymbol{0},\ \sigma^2\boldsymbol{I})$$
$$\hat{\boldsymbol{e}}/\sigma|\boldsymbol{X} \sim N(\boldsymbol{0},\ \boldsymbol{I})$$

Recall

$$\begin{aligned}(N-K)s_{\hat{e}}^2 &= \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}\\ &= \hat{\boldsymbol{e}}\boldsymbol{M}\hat{\boldsymbol{e}}\\ \frac{\hat{\boldsymbol{e}}'}{\sigma}\boldsymbol{M}\frac{\hat{\boldsymbol{e}}}{\sigma} &\sim \chi^2(N-K) && (\boldsymbol{M} \text{ is idempotent})\end{aligned}$$

That is, in a linear regression model, $\frac{(N-K)s_{\hat{e}}^2}{\sigma^2} \sim \chi^2_{n-k}$

R Example: χ^2 Distribution of s^2

```
set.seed(1); n <- 50; k <- 3; sigma2 <- 4
X <- cbind(1, matrix(rnorm(n*(k-1)), n, k-1))
beta <- c(2, -1, 0.5)
scaled_s2 <- replicate(10000, {
  y <- X %*% beta + rnorm(n, sd = sqrt(sigma2))
  ehat <- resid(lm(y ~ X - 1))
  sum(ehat^2) / sigma2   # (n-k)*s^2 / sigma^2
})
```

```
# Compare to chi-squared(n-k)
mean(scaled_s2)      # ~47 (= n-k)
var(scaled_s2)       # ~94 (= 2*(n-k))
```

Theory: $\mathbb{E}[\chi^2_{47}] = 47$, $\text{Var}[\chi^2_{47}] = 94$.

The simulation matches.

R Example: χ^2 Distribution of s^2

```
set.seed(1); n <- 50; k <- 3; sigma2 <- 4
X <- cbind(1, matrix(rnorm(n*(k-1)), n, k-1))
beta <- c(2, -1, 0.5)
scaled_s2 <- replicate(10000, {
  y <- X %*% beta + rnorm(n, sd = sqrt(sigma2))
  ehat <- resid(lm(y ~ X - 1))
  sum(ehat^2) / sigma2   # (n-k)*s^2 / sigma^2
})
```

```
# Compare to chi-squared(n-k)
mean(scaled_s2)     # ~47 (= n-k)
var(scaled_s2)      # ~94 (= 2*(n-k))
```

Theory: $\mathbb{E}[\chi^2_{47}] = 47$, $\text{Var}[\chi^2_{47}] = 94$.

The simulation matches.

Distribution of OLS Estimates

Under the classical linear model assumptions, the OLS estimator is normally distributed:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta},\ \sigma^2(\boldsymbol{X'X})^{-1})$$

Focus on the j-th coefficient:

$$\hat{\beta}_j \sim N\left(\beta_j,\ \sigma^2\left[(\boldsymbol{X'X})^{-1}\right]_{jj}\right)$$

Therefore, the standardized version is:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2\left[(\boldsymbol{X'X})^{-1}\right]_{jj}}} \sim N(0,1)$$

From Normal to t: Estimated Variance

In practice, we do not know the true error variance σ^2. The unbiased estimator is:

$$s_{\hat{e}}^2 = \frac{1}{n-k}\sum_{i=1}^{n} \hat{e}_i^2 = \frac{\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}}{n-k}$$

We plug in this estimate to standardize:

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s_{\hat{e}}^2\,[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{jj}}}$$

This is no longer standard normal, but it has a **t-distribution**.

The t Statistic and its Distribution

Recall from probability theory:

$$\frac{Z}{\sqrt{W/(n-k)}} \sim t(n-k) \quad \text{where} \quad Z \sim N(0,1),\ W \sim \chi^2(n-k),\ Z \perp W$$

In our case:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 \left[(\boldsymbol{X'X})^{-1}\right]_{jj}}} \sim N(0,1)$$

$$\frac{(n-k)s_{\hat{e}}^2}{\sigma^2} \sim \chi^2(n-k)$$

So the statistic:

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s_{\hat{e}}^2 \left[(\boldsymbol{X'X})^{-1}\right]_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\frac{(n-k)s_{\hat{e}}^2}{\sigma^2}\sigma^2 \left[(\boldsymbol{X'X})^{-1}\right]_{jj}/(n-k)}} \sim t(n-k)$$

Limits of t-statistic

- The T statistic follows the t distribution under homoskedasticity (by using s^2) and i.i.d. normality of e.
- Without normality, we can still say the OLS estimators are unbiased, but our exact distributions will not apply.
- The generic T statistic is $\frac{\hat{\beta}-\beta_0}{SE(\hat{\beta})}$.
- We will not have an exact finite-sample distribution of T when we use HC0-HC3 errors.
- Instead, we will use large sample approximations.

The F-test as a Likelihood Ratio Test (Hansen 5.13)

Consider testing $\mathbb{H}_0 : \beta_2 = 0$ in $Y = X_1'\beta_1 + X_2'\beta_2 + e$.

The likelihood ratio statistic compares the maximized log-likelihoods:

$$\text{LR} = 2\left(\ell_n(\hat{\beta}, \hat{\sigma}^2) - \ell_n(\tilde{\beta}_1, \tilde{\sigma}^2)\right) = n \log\left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right)$$

where $\tilde{\sigma}^2$ is from the restricted (null) model. This is equivalent to the F-statistic:

$$F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-k)} \sim F_{q,\,n-k} \quad \text{under } \mathbb{H}_0$$

- $q = \dim(\beta_2)$ is the number of restrictions.
- Under the null, F has an *exact* F-distribution in the normal model.
- This justifies the F-test reported by `anova()` and regression output.
- The t-test is a special case: when $q = 1$, $F = T^2$ and $F_{1,n-k} = t^2_{n-k}$.

The F-test as a Likelihood Ratio Test (Hansen 5.13)

Consider testing $\mathbb{H}_0 : \beta_2 = 0$ in $Y = X_1'\beta_1 + X_2'\beta_2 + e$.
The likelihood ratio statistic compares the maximized log-likelihoods:

$$\text{LR} = 2\left(\ell_n(\hat{\beta}, \hat{\sigma}^2) - \ell_n(\tilde{\beta}_1, \tilde{\sigma}^2)\right) = n \log\left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right)$$

where $\tilde{\sigma}^2$ is from the restricted (null) model. This is equivalent to the F-statistic:

$$F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-k)} \sim F_{q,\,n-k} \quad \text{under } \mathbb{H}_0$$

- $q = \dim(\beta_2)$ is the number of restrictions.
- Under the null, F has an *exact* F-distribution in the normal model.
- This justifies the F-test reported by `anova()` and regression output.
- The t-test is a special case: when $q = 1$, $F = T^2$ and $F_{1,n-k} = t^2_{n-k}$.

The F-test as a Likelihood Ratio Test (Hansen 5.13)

Consider testing $\mathbb{H}_0 : \beta_2 = 0$ in $Y = X_1'\beta_1 + X_2'\beta_2 + e$.

The likelihood ratio statistic compares the maximized log-likelihoods:

$$\mathrm{LR} = 2\left(\ell_n(\hat{\beta}, \hat{\sigma}^2) - \ell_n(\tilde{\beta}_1, \tilde{\sigma}^2)\right) = n \log\left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right)$$

where $\tilde{\sigma}^2$ is from the restricted (null) model. This is equivalent to the F-statistic:

$$F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-k)} \sim F_{q,\, n-k} \quad \text{under } \mathbb{H}_0$$

- $q = \dim(\beta_2)$ is the number of restrictions.
- Under the null, F has an *exact* F-distribution in the normal model.
- This justifies the F-test reported by `anova()` and regression output.
- The t-test is a special case: when $q = 1$, $F = T^2$ and $F_{1,n-k} = t^2_{n-k}$.

The F-test as a Likelihood Ratio Test (Hansen 5.13)

Consider testing $\mathbb{H}_0 : \beta_2 = 0$ in $Y = X_1'\beta_1 + X_2'\beta_2 + e$.
The likelihood ratio statistic compares the maximized log-likelihoods:

$$\text{LR} = 2\left(\ell_n(\hat{\beta}, \hat{\sigma}^2) - \ell_n(\tilde{\beta}_1, \tilde{\sigma}^2)\right) = n \log\left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right)$$

where $\tilde{\sigma}^2$ is from the restricted (null) model. This is equivalent to the F-statistic:

$$F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-k)} \sim F_{q,\, n-k} \quad \text{under } \mathbb{H}_0$$

- $q = \dim(\beta_2)$ is the number of restrictions.
- Under the null, F has an *exact* F-distribution in the normal model.
- This justifies the F-test reported by `anova()` and regression output.
- The t-test is a special case: when $q = 1$, $F = T^2$ and $F_{1,n-k} = t_{n-k}^2$.

Summary: Small-Sample Distributions

Under the classical normal regression model:

- $\hat{\beta}|\boldsymbol{X} \sim N(\beta,\ \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$ *(linear fn of normal)*
- $\hat{\mathbf{e}}|\boldsymbol{X} \sim N(\mathbf{0},\ \sigma^2\boldsymbol{M}), \quad \hat{\beta} \perp \hat{\mathbf{e}}$ *(projection)*
- $\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$ *(idempotent quadratic form)*
- $T = \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t_{n-k}$ *(normal / $\sqrt{\chi^2/df}$)*
- $F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-k)} \sim F_{q,n-k}$ *(likelihood ratio)*

These exact results hold in finite samples. Without normality, we rely on asymptotic approximations.

Summary: Small-Sample Distributions

Under the classical normal regression model:

- $\hat{\beta}|\boldsymbol{X} \sim N(\beta,\ \sigma^2(\boldsymbol{X'X})^{-1})$ *(linear fn of normal)*
- $\hat{\mathbf{e}}|\boldsymbol{X} \sim N(\mathbf{0},\ \sigma^2\boldsymbol{M}), \quad \hat{\beta} \perp \hat{\mathbf{e}}$ *(projection)*
- $\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$ *(idempotent quadratic form)*
- $T = \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t_{n-k}$ *(normal / $\sqrt{\chi^2/df}$)*
- $F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-k)} \sim F_{q,n-k}$ *(likelihood ratio)*

These exact results hold in finite samples. Without normality, we rely on asymptotic approximations.

Summary: Small-Sample Distributions

Under the classical normal regression model:

- $\hat{\beta}|\boldsymbol{X} \sim N(\beta,\ \sigma^2(\boldsymbol{X'X})^{-1})$ *(linear fn of normal)*
- $\hat{\boldsymbol{e}}|\boldsymbol{X} \sim N(\boldsymbol{0},\ \sigma^2\boldsymbol{M}),\quad \hat{\beta} \perp \hat{\boldsymbol{e}}$ *(projection)*
- $\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$ *(idempotent quadratic form)*
- $T = \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t_{n-k}$ *(normal / $\sqrt{\chi^2/df}$)*
- $F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-k)} \sim F_{q,n-k}$ *(likelihood ratio)*

These exact results hold in finite samples. Without normality, we rely on asymptotic approximations.

Summary: Small-Sample Distributions

Under the classical normal regression model:

- $\hat{\beta}|\boldsymbol{X} \sim N(\beta,\ \sigma^2(\boldsymbol{X'X})^{-1})$ *(linear fn of normal)*
- $\hat{\boldsymbol{e}}|\boldsymbol{X} \sim N(\boldsymbol{0},\ \sigma^2\boldsymbol{M}), \quad \hat{\beta} \perp \hat{\boldsymbol{e}}$ *(projection)*
- $\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$ *(idempotent quadratic form)*
- $T = \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t_{n-k}$ *(normal / $\sqrt{\chi^2/df}$)*
- $F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-k)} \sim F_{q,n-k}$ *(likelihood ratio)*

These exact results hold in finite samples. Without normality, we rely on asymptotic approximations.

Summary: Small-Sample Distributions

Under the classical normal regression model:

- $\hat{\beta}|\boldsymbol{X} \sim N(\beta,\ \sigma^2(\boldsymbol{X'X})^{-1})$ *(linear fn of normal)*
- $\hat{\mathbf{e}}|\boldsymbol{X} \sim N(\mathbf{0},\ \sigma^2\boldsymbol{M}), \quad \hat{\beta} \perp \hat{\mathbf{e}}$ *(projection)*
- $\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$ *(idempotent quadratic form)*
- $T = \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t_{n-k}$ *(normal / $\sqrt{\chi^2/df}$)*
- $F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-k)} \sim F_{q,n-k}$ *(likelihood ratio)*

These exact results hold in finite samples. Without normality, we rely on asymptotic approximations.

Summary: Small-Sample Distributions

Under the classical normal regression model:

- $\hat{\beta}|\boldsymbol{X} \sim N(\beta,\ \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$ *(linear fn of normal)*
- $\hat{\boldsymbol{e}}|\boldsymbol{X} \sim N(\boldsymbol{0},\ \sigma^2\boldsymbol{M}), \quad \hat{\beta} \perp \hat{\boldsymbol{e}}$ *(projection)*
- $\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$ *(idempotent quadratic form)*
- $T = \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t_{n-k}$ *(normal / $\sqrt{\chi^2/df}$)*
- $F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-k)} \sim F_{q,n-k}$ *(likelihood ratio)*

These exact results hold in finite samples. Without normality, we rely on asymptotic approximations.

Summary: Small-Sample Distributions

Under the classical normal regression model:

- $\hat{\beta}|\boldsymbol{X} \sim N(\beta,\ \sigma^2(\boldsymbol{X'X})^{-1})$ *(linear fn of normal)*
- $\hat{\boldsymbol{e}}|\boldsymbol{X} \sim N(\boldsymbol{0},\ \sigma^2\boldsymbol{M}), \quad \hat{\beta} \perp \hat{\boldsymbol{e}}$ *(projection)*
- $\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$ *(idempotent quadratic form)*
- $T = \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t_{n-k}$ *(normal / $\sqrt{\chi^2/df}$)*
- $F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-k)} \sim F_{q,n-k}$ *(likelihood ratio)*

These exact results hold in finite samples. Without normality, we rely on asymptotic approximations.

Confidence Intervals

- $\hat{\beta}$ is a **point estimate** for a coefficient β.
- We can instead estimate an interval, $\hat{C} = [\hat{L},\ \hat{U}]$ which contains the true value with high probability.
- An interval estimate $\hat{C}$ is called a $1-\alpha$ confidence interval when $Pr(\beta \in \hat{C}) = 1-\alpha$. The value $1-\alpha$ is called the **coverage probability**.
- The key mistake is in thinking that the above statement treats β is random and $\hat{C}$ is fixed, (the probability that β is in $\hat{C}$).
- $Pr(\beta \in \hat{C})$ is the probability that the random set $\hat{C}$ covers or contains β.

$$\hat{C} = [\hat{\beta} - c * s(\hat{\beta}),\ \hat{\beta} + c * s(\hat{\beta})]$$

Confidence Intervals

- $\hat{\beta}$ is a **point estimate** for a coefficient β.
- We can instead estimate an interval, $\hat{C} = [\hat{L},\ \hat{U}]$ which contains the true value with high probability.
- An interval estimate $\hat{C}$ is called a $1-\alpha$ confidence interval when $Pr(\beta \in \hat{C}) = 1-\alpha$. The value $1-\alpha$ is called the **coverage probability**.
- The key mistake is in thinking that the above statement treats β is random and $\hat{C}$ is fixed, (the probability that β is in $\hat{C}$).
- $Pr(\beta \in \hat{C})$ is the probability that the random set $\hat{C}$ covers or contains β.

$$\hat{C} = [\hat{\beta} - c * s(\hat{\beta}),\ \hat{\beta} + c * s(\hat{\beta})]$$

Confidence Intervals

- $\hat{\beta}$ is a **point estimate** for a coefficient β.
- We can instead estimate an interval, $\hat{C} = [\hat{L},\ \hat{U}]$ which contains the true value with high probability.
- An interval estimate $\hat{C}$ is called a $1 - \alpha$ confidence interval when $Pr(\beta \in \hat{C}) = 1 - \alpha$. The value $1 - \alpha$ is called the **coverage probability**.
- The key mistake is in thinking that the above statement treats β is random and $\hat{C}$ is fixed, (the probability that β is in $\hat{C}$).
- $Pr(\beta \in \hat{C})$ is the probability that the random set $\hat{C}$ covers or contains β.

$$\hat{C} = [\hat{\beta} - c * s(\hat{\beta}),\ \hat{\beta} + c * s(\hat{\beta})]$$

Confidence Intervals

- $\hat{\beta}$ is a **point estimate** for a coefficient β.
- We can instead estimate an interval, $\hat{C} = [\hat{L}, \; \hat{U}]$ which contains the true value with high probability.
- An interval estimate $\hat{C}$ is called a $1-\alpha$ confidence interval when $Pr(\beta \in \hat{C}) = 1-\alpha$. The value $1-\alpha$ is called the **coverage probability**.
- The key mistake is in thinking that the above statement treats β is random and $\hat{C}$ is fixed, (the probability that β is in $\hat{C}$).
- $Pr(\beta \in \hat{C})$ is the probability that the random set $\hat{C}$ covers or contains β.

$$\hat{C} = [\hat{\beta} - c * s(\hat{\beta}), \; \hat{\beta} + c * s(\hat{\beta})]$$

Confidence Intervals

- $\hat{\beta}$ is a **point estimate** for a coefficient β.
- We can instead estimate an interval, $\hat{C} = [\hat{L},\ \hat{U}]$ which contains the true value with high probability.
- An interval estimate $\hat{C}$ is called a $1 - \alpha$ confidence interval when $Pr(\beta \in \hat{C}) = 1 - \alpha$. The value $1 - \alpha$ is called the **coverage probability**.
- The key mistake is in thinking that the above statement treats β is random and $\hat{C}$ is fixed, (the probability that β is in $\hat{C}$).
- $Pr(\beta \in \hat{C})$ is the probability that the random set $\hat{C}$ covers or contains β.

$$\hat{C} = [\hat{\beta} - c * s(\hat{\beta}),\ \hat{\beta} + c * s(\hat{\beta})]$$

Confidence Intervals

- $\hat{\beta}$ is a **point estimate** for a coefficient β.
- We can instead estimate an interval, $\hat{C} = [\hat{L},\ \hat{U}]$ which contains the true value with high probability.
- An interval estimate $\hat{C}$ is called a $1-\alpha$ confidence interval when $Pr(\beta \in \hat{C}) = 1-\alpha$. The value $1-\alpha$ is called the **coverage probability**.
- The key mistake is in thinking that the above statement treats β is random and $\hat{C}$ is fixed, (the probability that β is in $\hat{C}$).
- $Pr(\beta \in \hat{C})$ is the probability that the random set $\hat{C}$ covers or contains β.

$$\hat{C} = [\hat{\beta} - c * s(\hat{\beta}),\ \hat{\beta} + c * s(\hat{\beta})]$$

Confidence Intervals in Practice

$$Pr(\beta \in \hat{C}) = Pr(-c \leq T(\beta) \leq c)$$
$$Pr(\beta \in \hat{C}) = 2F(c) - 1$$

Our goal is to set this coverage probablity equal to $1 - \alpha$, or $F(c) = 1 - \alpha/2$. If $\alpha = .05$, we solve $c = F^{-1}(1 - .05/2)$. In case of a normal, c=1.96 ≈ 2

$$\hat{C} = [\hat{\beta} - 2 * s(\hat{\beta}),\ \hat{\beta} + 2 * s(\hat{\beta})]$$

t test and p-values

- A theory is said to have *testable implications* if it can be falsified.
- For example, a theory may be false if $\beta = \beta_0$. This is called a "null hypothesis" $\mathbb{H}_0$.
- We further specify the complement of $\mathbb{H}_0$ as $\mathbb{H}_1$.
- A statistic can be informative, some realizations may be unlikely if $\mathbb{H}_0$ is true.
- Define a test statistic: $|T| = \left|\frac{\hat{\beta}-\beta_0}{s(\hat{\beta})}\right|$ and set a critical value c.

$$\text{Reject } \mathbb{H}_0 \text{ if } |T| > c$$

- A p-value indexes a test's strength of rejection of the null.
- In a normal regression model, $p = 2(1 - F_{n-k}(|T|))$

t test and p-values

- A theory is said to have *testable implications* if it can be falsified.
- For example, a theory may be false if $\beta = \beta_0$. This is called a "null hypothesis" $\mathbb{H}_0$.
- We further specify the complement of $\mathbb{H}_0$ as $\mathbb{H}_1$.
- A statistic can be informative, some realizations may be unlikely if $\mathbb{H}_0$ is true.
- Define a test statistic: $|T| = \left|\frac{\hat{\beta}-\beta_0}{s(\hat{\beta})}\right|$ and set a critical value c.

$$\text{Reject } \mathbb{H}_0 \text{ if } |T| > c$$

- A p-value indexes a test's strength of rejection of the null.
- In a normal regression model, $p = 2(1 - F_{n-k}(|T|))$

t test and p-values

- A theory is said to have *testable implications* if it can be falsified.
- For example, a theory may be false if $\beta = \beta_0$. This is called a "null hypothesis" $\mathbb{H}_0$.
- We further specify the complement of $\mathbb{H}_0$ as $\mathbb{H}_1$.
- A statistic can be informative, some realizations may be unlikely if $\mathbb{H}_0$ is true.
- Define a test statistic: $|T| = \left|\frac{\hat{\beta}-\beta_0}{s(\hat{\beta})}\right|$ and set a critical value c.

$$\text{Reject } \mathbb{H}_0 \text{ if } |T| > c$$

- A p-value indexes a test's strength of rejection of the null.
- In a normal regression model, $p = 2(1 - F_{n-k}(|T|))$

t test and p-values

- A theory is said to have *testable implications* if it can be falsified.
- For example, a theory may be false if $\beta = \beta_0$. This is called a "null hypothesis" $\mathbb{H}_0$.
- We further specify the complement of $\mathbb{H}_0$ as $\mathbb{H}_1$.
- A statistic can be informative, some realizations may be unlikely if $\mathbb{H}_0$ is true.
- Define a test statistic: $|T| = \left|\frac{\hat{\beta}-\beta_0}{s(\hat{\beta})}\right|$ and set a critical value c.

$$\text{Reject } \mathbb{H}_0 \text{ if } |T| > c$$

- A p-value indexes a test's strength of rejection of the null.
- In a normal regression model, $p = 2(1 - F_{n-k}(|T|))$

t test and p-values

- A theory is said to have *testable implications* if it can be falsified.
- For example, a theory may be false if $\beta = \beta_0$. This is called a "null hypothesis" $\mathbb{H}_0$.
- We further specify the complement of $\mathbb{H}_0$ as $\mathbb{H}_1$.
- A statistic can be informative, some realizations may be unlikely if $\mathbb{H}_0$ is true.
- Define a test statistic: $|T| = \left|\frac{\hat{\beta}-\beta_0}{s(\hat{\beta})}\right|$ and set a critical value c.

$$\text{Reject } \mathbb{H}_0 \text{ if } |T| > c$$

- A p-value indexes a test's strength of rejection of the null.
- In a normal regression model, $p = 2(1 - F_{n-k}(|T|))$

t test and p-values

- A theory is said to have *testable implications* if it can be falsified.
- For example, a theory may be false if $\beta = \beta_0$. This is called a "null hypothesis" $\mathbb{H}_0$.
- We further specify the complement of $\mathbb{H}_0$ as $\mathbb{H}_1$.
- A statistic can be informative, some realizations may be unlikely if $\mathbb{H}_0$ is true.
- Define a test statistic: $|T| = \left|\frac{\hat{\beta}-\beta_0}{s(\hat{\beta})}\right|$ and set a critical value c.

$$\text{Reject } \mathbb{H}_0 \text{ if } |T| > c$$

- A p-value indexes a test's strength of rejection of the null.
- In a normal regression model, $p = 2(1 - F_{n-k}(|T|))$

Example

	Dependent variable: Fertility
Education (%)	-0.98^{***}
	(0.15)
Agriculture (%)	-0.15^{**}
	(0.07)
Catholic (%)	0.12^{***}
	(0.03)
Infant.Mortality (%)	1.08^{***}
	(0.38)
Constant	62.10^{***}
	(9.60)
Observations	47
R^2	0.70
Adjusted R^2	0.67
Residual Std. Error	7.17 (df = 42)
F Statistic	24.42^{***} (df = 4; 42)
Note:	*p<0.1; **p<0.05; ***p<0.01

Did education share predict the average fertility across Swiss districts in 1888?

$$\hat{\beta} = -.98, \quad \beta_0 = 0, \quad se(\hat{\beta}) = 0.15, \quad Df = 47 - 5$$

$$t = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})} = \frac{-.98 - 0}{0.15} = -6.61$$

$$c = F^{-1}(1 - .05/2) = \text{qt}(1 - .025, 41) = 2.02$$

$$|t| > 2.02$$

$$p = 2(1 - F_{n-k}(|T|)) = 2 * (1 - \text{pt}(6.61, 42)) = 5.2 \times 10^{-8}$$

For each 1 percentage point of post-primary attendance, fertility rate is (.7, 1.3) percentage points lower

Example

	Dependent variable: Fertility
Education (%)	-0.98^{***}
	(0.15)
Agriculture (%)	-0.15^{**}
	(0.07)
Catholic (%)	0.12^{***}
	(0.03)
Infant.Mortality (%)	1.08^{***}
	(0.38)
Constant	62.10^{***}
	(9.60)
Observations	47
R^2	0.70
Adjusted R^2	0.67
Residual Std. Error	7.17 (df = 42)
F Statistic	24.42^{***} (df = 4; 42)
Note:	*p<0.1; **p<0.05; ***p<0.01

Did education share predict the average fertility across Swiss districts in 1888?

$$\hat{\beta} = -.98, \quad \beta_0 = 0, \quad se(\hat{\beta}) = 0.15, \quad Df = 47 - 5$$

$$t = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})} = \frac{-.98 - 0}{0.15} = -6.61$$

$$c = F^{-1}(1 - .05/2) = \text{qt}(1 - .025, 41) = 2.02$$

$$|t| > 2.02$$

$$p = 2(1 - F_{n-k}(|T|)) = 2 * (1 - \text{pt}(6.61, 42)) = 5.2 \times 10^{-8}$$

For each 1 percentage point of post-primary attendance, fertility rate is (.7, 1.3) percentage points lower

Example

	Dependent variable: Fertility
Education (%)	-0.98^{***}
	(0.15)
Agriculture (%)	-0.15^{**}
	(0.07)
Catholic (%)	0.12^{***}
	(0.03)
Infant.Mortality (%)	1.08^{***}
	(0.38)
Constant	62.10^{***}
	(9.60)
Observations	47
R^2	0.70
Adjusted R^2	0.67
Residual Std. Error	7.17 (df = 42)
F Statistic	24.42^{***} (df = 4; 42)
Note:	*p<0.1; **p<0.05; ***p<0.01

Did education share predict the average fertility across Swiss districts in 1888?

$$\hat{\beta} = -.98, \quad \beta_0 = 0, \quad se(\hat{\beta}) = 0.15, \quad Df = 47 - 5$$

$$t = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})} = \frac{-.98 - 0}{0.15} = -6.61$$

$$c = F^{-1}(1 - .05/2) = \text{qt}(1 - .025, 41) = 2.02$$

$$|t| > 2.02$$

$$p = 2(1 - F_{n-k}(|T|)) = 2 * (1 - \text{pt}(6.61, 42)) = 5.2 \times 10^{-8}$$

For each 1 percentage point of post-primary attendance, fertility rate is (.7, 1.3) percentage points lower

Example

	Dependent variable: Fertility
Education (%)	-0.98^{***}
	(0.15)
Agriculture (%)	-0.15^{**}
	(0.07)
Catholic (%)	0.12^{***}
	(0.03)
Infant.Mortality (%)	1.08^{***}
	(0.38)
Constant	62.10^{***}
	(9.60)
Observations	47
R^2	0.70
Adjusted R^2	0.67
Residual Std. Error	7.17 (df = 42)
F Statistic	24.42^{***} (df = 4; 42)
Note:	*p<0.1; **p<0.05; ***p<0.01

Did education share predict the average fertility across Swiss districts in 1888?

$$\hat{\beta} = -.98, \quad \beta_0 = 0, \quad se(\hat{\beta}) = 0.15, \quad Df = 47 - 5$$

$$t = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})} = \frac{-.98 - 0}{0.15} = -6.61$$

$$c = F^{-1}(1 - .05/2) = \mathrm{qt}(1 - .025, 41) = 2.02$$

$$|t| > 2.02$$

$$p = 2(1 - F_{n-k}(|T|)) = 2 * (1 - \mathrm{pt}(6.61, 42)) = 5.2 \times 10^{-8}$$

For each 1 percentage point of post-primary attendance, fertility rate is (.7, 1.3) percentage points lower

Example

	Dependent variable: Fertility
Education (%)	-0.98^{***}
	(0.15)
Agriculture (%)	-0.15^{**}
	(0.07)
Catholic (%)	0.12^{***}
	(0.03)
Infant.Mortality (%)	1.08^{***}
	(0.38)
Constant	62.10^{***}
	(9.60)
Observations	47
R^2	0.70
Adjusted R^2	0.67
Residual Std. Error	7.17 (df = 42)
F Statistic	24.42^{***} (df = 4; 42)
Note:	*p<0.1; **p<0.05; ***p<0.01

Did education share predict the average fertility across Swiss districts in 1888?

$$\hat{\beta} = -.98, \quad \beta_0 = 0, \quad se(\hat{\beta}) = 0.15, \quad Df = 47 - 5$$

$$t = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})} = \frac{-.98 - 0}{0.15} = -6.61$$

$$c = F^{-1}(1 - .05/2) = \text{qt}(1 - .025, 41) = 2.02$$

$$|t| > 2.02$$

$$p = 2(1 - F_{n-k}(|T|)) = 2 * (1 - \text{pt}(6.61, 42)) = 5.2 \times 10^{-8}$$

For each 1 percentage point of post-primary attendance, fertility rate is (.7, 1.3) percentage points lower

Example

	Dependent variable: Fertility
Education (%)	-0.98^{***}
	(0.15)
Agriculture (%)	-0.15^{**}
	(0.07)
Catholic (%)	0.12^{***}
	(0.03)
Infant.Mortality (%)	1.08^{***}
	(0.38)
Constant	62.10^{***}
	(9.60)
Observations	47
R^2	0.70
Adjusted R^2	0.67
Residual Std. Error	7.17 (df = 42)
F Statistic	24.42^{***} (df = 4; 42)
Note:	*p<0.1; **p<0.05; ***p<0.01

Did education share predict the average fertility across Swiss districts in 1888?

$$\hat{\beta} = -.98, \quad \beta_0 = 0, \quad se(\hat{\beta}) = 0.15, \quad Df = 47 - 5$$

$$t = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})} = \frac{-.98 - 0}{0.15} = -6.61$$

$$c = F^{-1}(1 - .05/2) = \mathtt{qt}(1 - .025, 41) = 2.02$$

$$|t| > 2.02$$

$$p = 2(1 - F_{n-k}(|T|)) = 2 * (1 - \mathtt{pt}(6.61, 42)) = 5.2 \times 10^{-8}$$

For each 1 percentage point of post-primary attendance, fertility rate is (.7, 1.3) percentage points lower

Example

	Dependent variable: Fertility
Education (%)	-0.98^{***}
	(0.15)
Agriculture (%)	-0.15^{**}
	(0.07)
Catholic (%)	0.12^{***}
	(0.03)
Infant.Mortality (%)	1.08^{***}
	(0.38)
Constant	62.10^{***}
	(9.60)
Observations	47
R^2	0.70
Adjusted R^2	0.67
Residual Std. Error	7.17 (df = 42)
F Statistic	24.42^{***} (df = 4; 42)
Note:	*p<0.1; **p<0.05; ***p<0.01

Did education share predict the average fertility across Swiss districts in 1888?

$$\hat{\beta} = -.98, \quad \beta_0 = 0, \quad se(\hat{\beta}) = 0.15, \quad Df = 47 - 5$$

$$t = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})} = \frac{-.98 - 0}{0.15} = -6.61$$

$$c = F^{-1}(1 - .05/2) = \mathtt{qt}(1 - .025, 41) = 2.02$$

$$|t| > 2.02$$

$$p = 2(1 - F_{n-k}(|T|)) = 2 * (1 - \mathtt{pt}(6.61, 42)) = 5.2 \times 10^{-8}$$

For each 1 percentage point of post-primary attendance, fertility rate is (.7, 1.3) percentage points lower

Example

	Dependent variable: Fertility
Education (%)	-0.98^{***}
	(0.15)
Agriculture (%)	-0.15^{**}
	(0.07)
Catholic (%)	0.12^{***}
	(0.03)
Infant.Mortality (%)	1.08^{***}
	(0.38)
Constant	62.10^{***}
	(9.60)
Observations	47
R^2	0.70
Adjusted R^2	0.67
Residual Std. Error	7.17 (df = 42)
F Statistic	24.42^{***} (df = 4; 42)
Note:	*p<0.1; **p<0.05; ***p<0.01

Did education share predict the average fertility across Swiss districts in 1888?

$$\hat{\beta} = -.98, \quad \beta_0 = 0, \quad se(\hat{\beta}) = 0.15, \quad Df = 47 - 5$$

$$t = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})} = \frac{-.98 - 0}{0.15} = -6.61$$

$$c = F^{-1}(1 - .05/2) = \mathtt{qt}(1 - .025, 41) = 2.02$$

$$|t| > 2.02$$

$$p = 2(1 - F_{n-k}(|T|)) = 2 * (1 - \mathtt{pt}(6.61, 42)) = 5.2 \times 10^{-8}$$

For each 1 percentage point of post-primary attendance, fertility rate is (.7, 1.3) percentage points lower

Example

	Dependent variable: Fertility
Education (%)	-0.98^{***}
	(0.15)
Agriculture (%)	-0.15^{**}
	(0.07)
Catholic (%)	0.12^{***}
	(0.03)
Infant.Mortality (%)	1.08^{***}
	(0.38)
Constant	62.10^{***}
	(9.60)
Observations	47
R^2	0.70
Adjusted R^2	0.67
Residual Std. Error	7.17 (df = 42)
F Statistic	24.42^{***} (df = 4; 42)
Note:	*p<0.1; **p<0.05; ***p<0.01

Did education share predict the average fertility across Swiss districts in 1888?

$$\hat{\beta} = -.98, \quad \beta_0 = 0, \quad se(\hat{\beta}) = 0.15, \quad Df = 47 - 5$$

$$t = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})} = \frac{-.98 - 0}{0.15} = -6.61$$

$$c = F^{-1}(1 - .05/2) = \mathtt{qt}(1 - .025, 41) = 2.02$$

$$|t| > 2.02$$

$$p = 2(1 - F_{n-k}(|T|)) = 2 * (1 - \mathtt{pt}(6.61, 42)) = 5.2 \times 10^{-8}$$

For each 1 percentage point of post-primary attendance, fertility rate is (.7, 1.3) percentage points lower