MS&E 226: Fundamentals of Data Science

Lecture 9: Standard errors and confidence intervals

Ramesh Johari

Recall: The challenge

To quantify uncertainty in our estimates, we need to understand the sampling distribution of our estimator.

But we can't actually compute the sampling distribution, because:

- We would need to resample from the population model.
- But we don't know the true parameters that define the population model.
- We only have one dataset.

Addressing the challenge

In this lecture, we show how we can characterize the sampling distribution for three important, commonly used estimators: sample means; OLS linear regression; and logistic regression.

A recurring theme will be that in the cases we study:

- the sampling distribution is well approximated by a *normal distribution*,
- centered at the true parameter, and
- with a variance we can estimate from data.

The estimated variance gives us estimated *standard errors*, which we will use to build *confidence intervals*.

The sample mean

Example: Sample mean of flight delays

Recall our flight delays example:

We have a population of 2,926,854 valid flights in 2024, with:

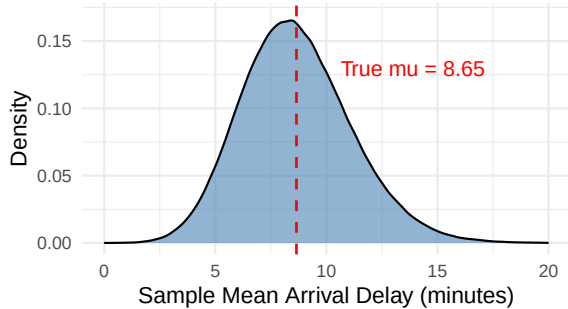
- Population mean: $\mu = 8.65$ minutes
- Population standard deviation: $\sigma = 55.47$ minutes

We draw a sample of n flights, and use the sample mean $\overline{Y}$ to estimate μ.

Question: What is the sampling distribution of $\overline{Y}$?

Sampling distribution of the sample mean

Here is the sampling distribution when $n = 500$ (with 1,000,000 parallel universes):



Key observation: *For large n, the sampling distribution appears to be approximately normal!*

The Central Limit Theorem

The Central Limit Theorem (CLT) formalizes this observation:

Theorem (Central Limit Theorem)

Suppose $Y_1, \ldots, Y_n$ are independent, identically distributed random variables with mean μ and variance σ^2.

Then as $n \to \infty$, the standardized sample mean converges in distribution to a standard normal distribution:

$$\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

(Here $\mathcal{N}(a, b^2)$ denotes a normal distribution with mean a and variance b^2.)

Equivalently: $\overline{Y} \approx \mathcal{N}(\mu, \sigma^2/n)$ for large n.

Using the CLT

The CLT tells us that for large n:

- The sample mean $\overline{Y}$ is approximately normally distributed.
- The mean of this normal distribution is μ (the true population mean).
- The variance of this normal distribution is σ^2/n.
- So the (approximate) standard error for large n is $\text{SE} = \sigma/\sqrt{n}$.

Using the CLT

The CLT tells us that for large n:

- The sample mean $\overline{Y}$ is approximately normally distributed.
- The mean of this normal distribution is μ (the true population mean).
- The variance of this normal distribution is σ^2/n.
- So the (approximate) standard error for large n is $\text{SE} = \sigma/\sqrt{n}$.

This is remarkable: *No matter what the distribution of the individual Y_i is*, the sample mean is approximately normal for large n!

Estimating the standard error

The CLT tells us the asymptotic standard error is $\text{SE} = \sigma/\sqrt{n}$.

But there's a problem: *We don't know σ!*

Solution: We can estimate σ from the data using the *sample standard deviation*:

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}.$$

Then we estimate the standard error as:

$$\widehat{\text{SE}} = \frac{\hat{\sigma}}{\sqrt{n}}.$$

This is called the *estimated standard error*.

Using the estimated standard error

For large n, we can use $\widehat{\text{SE}}$ in place of SE:

$$\overline{Y} \approx \mathcal{N}\left(\mu, \frac{\hat{\sigma}^2}{n}\right).$$

Next we will show how to use this approximate sampling distribution to construct *confidence intervals*.

Confidence intervals for the population mean

Confidence intervals

Now that we have the estimated standard error $\widehat{\text{SE}}$, we can construct *confidence intervals* for μ.

For large n:

Confidence intervals

Now that we have the estimated standard error $\widehat{\text{SE}}$, we can construct *confidence intervals* for μ.

For large n:

- The sampling distribution of $\overline{Y}$ is approximately normal with *mean* μ, and *standard deviation* $\widehat{\text{SE}}$.

Confidence intervals

Now that we have the estimated standard error $\widehat{\text{SE}}$, we can construct *confidence intervals* for μ.

For large n:

- The sampling distribution of $\overline{Y}$ is approximately normal with *mean* μ, and *standard deviation* $\widehat{\text{SE}}$.
- A normal distribution has $\approx 95\%$ of its mass within 1.96 standard deviations of the mean.

Confidence intervals

Now that we have the estimated standard error $\widehat{\text{SE}}$, we can construct *confidence intervals* for μ.

For large n:

- The sampling distribution of $\overline{Y}$ is approximately normal with *mean* μ, and *standard deviation* $\widehat{\text{SE}}$.
- A normal distribution has $\approx 95\%$ of its mass within 1.96 standard deviations of the mean.
- Therefore, in 95% of our "universes", $\overline{Y}$ will be within $1.96\ \widehat{\text{SE}}$ of the true value of μ:

$$\mu - 1.96\widehat{\text{SE}} \leq \overline{Y} \leq \mu + 1.96\widehat{\text{SE}}.$$

Confidence intervals

Now that we have the estimated standard error $\widehat{\text{SE}}$, we can construct *confidence intervals* for μ.

For large n:

- The sampling distribution of $\overline{Y}$ is approximately normal with *mean* μ, and *standard deviation* $\widehat{\text{SE}}$.
- A normal distribution has $\approx 95\%$ of its mass within 1.96 standard deviations of the mean.
- Therefore, in 95% of our "universes", $\overline{Y}$ will be within $1.96\ \widehat{\text{SE}}$ of the true value of μ:

$$\mu - 1.96\widehat{\text{SE}} \leq \overline{Y} \leq \mu + 1.96\widehat{\text{SE}}.$$

- In other words: in 95% of our universes:

$$\overline{Y} - 1.96\ \widehat{\text{SE}} \leq \mu \leq \overline{Y} + 1.96\ \widehat{\text{SE}}.$$

Confidence intervals

We refer to $[\overline{Y} - 1.96\ \widehat{\text{SE}}, \overline{Y} + 1.96\ \widehat{\text{SE}}]$ as a *95% confidence interval* for μ.

More generally, let z_α be the value such that $P(Z \leq z_\alpha) = 1 - \alpha$ for a $\mathcal{N}(0,1)$ random variable Z. Then:

$$[\overline{Y} - z_{\alpha/2}\widehat{\text{SE}}, \overline{Y} + z_{\alpha/2}\widehat{\text{SE}}]$$

is a $1 - \alpha$ confidence interval for μ.

In R, you can get z_α using the `qnorm` function. When $\alpha = 0.05$, then $z_{\alpha/2} \approx 1.96$.

Confidence intervals: What's random?

Important observation: Note that *the interval is random, and μ is fixed*!

In particular, often you will hear: "There is a 95% chance that the true μ is between $\overline{Y} - 1.96\widehat{\text{SE}}$ and $\overline{Y} + 1.96\widehat{\text{SE}}$."

But when this statement is made, *it's the endpoints of the interval that are random,* not μ.

Confidence intervals: “Coverage”

The confidence level $1 - \alpha$ is called the *coverage* of the confidence interval.

As we change α to be *larger* (i.e., lower coverage or “confidence”), the confidence interval gets *narrower*.

Further, for a given α, confidence intervals can be enlarged while still having at least $1 - \alpha$ coverage; so the goal is always to construct the smallest interval possible that has the desired coverage.

Finally, note that other approaches to building $1 - \alpha$ confidence intervals are possible, that may yield asymmetric intervals.

Ordinary least squares

The linear population model: Three assumptions

Recall these assumptions:

(A1) The population model is *linear*, i.e., there are *parameters* $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ such that for every observation:

$$Y = \sum_{j=1}^{p} \beta_j X_j + \epsilon = \vec{X}\boldsymbol{\beta} + \epsilon,$$

where Y is the outcome, $X_1, \ldots, X_p$ are the associated features, and ϵ is an *error* random variable.

(A2) Observations in the sample are independently drawn from the population model, and all features from (A1) are observed. The sample consists of n observations $\mathbf{Y}$ and associated features $\mathbf{X}$ ($n \times p$ matrix).

(A3) All errors have $\mathbb{E}[\epsilon|\vec{X}] = 0$, and the same variance $\mathrm{Var}(\epsilon|\vec{X}) = \sigma^2$.

Another assumption: Normal errors

Now we'll add a fourth assumption to that group:

(A4) The errors are *normally distributed conditional on the features*:
$\epsilon_i \mid \mathbf{X}_i \sim \mathcal{N}(0, \sigma^2)$ for each i.

Under the assumptions (A1)-(A4), we showed previously that linear regression via OLS produces coefficients $\hat{\boldsymbol{\beta}}$ that are *maximum likelihood estimates of the true coefficients* $\boldsymbol{\beta}$.

(A1)-(A3) may already be implausible, and it's reasonable to be skeptical of (A4) too...but let's see where it leads us first.

The true SE for linear regression

Under (A1)-(A4), it can be shown that *for any* n, conditional on the feature matrix $\mathbf{X}$, the OLS coefficient vector $\hat{\boldsymbol{\beta}}$:

1. is normally distributed,
2. with mean $\boldsymbol{\beta}$ (the true coefficient vector in the population model), i.e., it is *unbiased*, and
3. with covariance matrix $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$.[1]

(Recall that $\mathbf{X}$ denotes the $n \times p$ feature matrix.)
Thus the squared standard error of $\hat{\beta}_j$ is the j'th diagonal entry of the covariance matrix:

$$\text{SE}_j^2 = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}.$$

[1] This result is derived using a similar analysis to the bias-variance decomposition for OLS.

Estimating σ^2 [$*$]

An important detail here is that we don't directly use the MLE $\hat{\sigma}^2_{\text{MLE}}$ to estimate σ^2.

Recall that $\hat{\sigma}^2_{\text{MLE}}$ is the average sum of squared residuals:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2.$$

But this is *not unbiased* as an estimator of σ^2. In fact it can be shown that:

$$\mathbb{E}_{\mathbf{Y}}[\hat{\sigma}^2_{\text{MLE}}|\boldsymbol{\beta}, \sigma^2, \mathbf{X}] = \frac{n-p}{n}\sigma^2.$$

Estimating σ^2 [$*$]

In other words, $\hat{\sigma}^2_{\text{MLE}}$ *underestimates* the true error variance.

This is because the MLE solution $\hat{\boldsymbol{\beta}}$ was *chosen* to minimize squared error on the sample data. We need to account for this "favorable selection" of the variance estimate by "reinflating" it. An unbiased estimate of σ^2 is:

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2.$$

The quantity $n - p$ is called the *degrees of freedom* (DoF). (It can be shown this is a good estimate when n is large.)

Conventional $\widehat{\text{SE}}$ and confidence intervals

Putting things together, the conventional estimate of the squared standard error under (A1)-(A4) is:

$$\widehat{\text{SE}}_j^2 = \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}_{jj}.$$

Thus if (A1)-(A4) hold, then an approximate 95% confidence interval for β_j is:

$$[\hat{\beta}_j - 1.96\widehat{\text{SE}}_j, \hat{\beta}_j + 1.96\widehat{\text{SE}}_j].$$

Housing example

Default output from `summary(lm(...))`:

```
Call:
lm(formula = price ~ 1 + livingArea + bedrooms, data = sh)
...
Coefficients:
             Estimate Std. Error ...
(Intercept)  36667.895   6610.293 ...
livingArea     125.405      3.527 ...
bedrooms    -14196.769   2675.159 ...
...
```

Thus an approximate 95% confidence interval for `livingArea` coefficient is $[118.49, 132.32]$.

Relaxing assumptions

Relaxing assumptions

Our derivation in the previous section of the conventional regression table output *assumes (A1)-(A4) hold*.

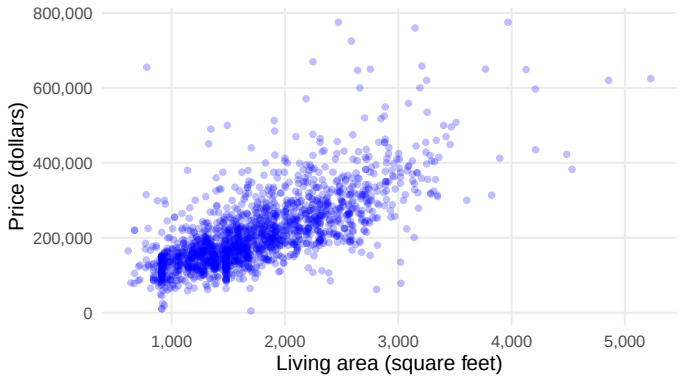
In this section, we discuss how we can relax some of these assumptions while still successfully estimating standard errors and confidence intervals.

We will:

- Eliminate (A4), i.e., no longer require errors to be normally distributed; and
- Relax (A3) so that error variance does not need to be constant.

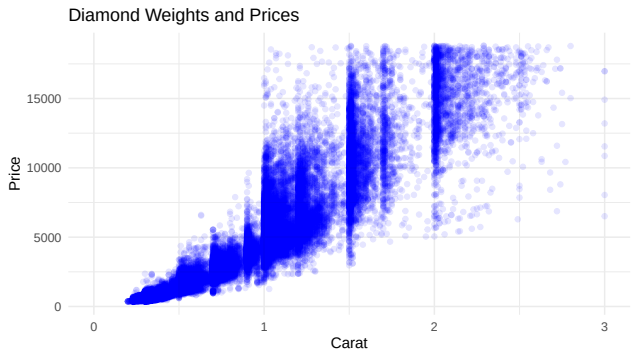
Nonconstant error variance: House prices

The spread in house prices suggests that error variance may not be constant as `livingArea` changes.



Nonconstant error variance: Diamond prices

The spread in diamond prices also increases with diamond size:



M-estimators and asymptotic normality

The key observation is that sample means, OLS coefficients, and logistic regression are special cases of *M-estimators*.

The "M" stands for "maximum likelihood type"; informally, M-estimators are estimators that are obtained by solving loss minimization problems (like minimizing the negative log likelihood). (See appendix for full details on M-estimators.)

M-estimators and asymptotic normality

The key observation is that sample means, OLS coefficients, and logistic regression are special cases of *M-estimators*.

The "M" stands for "maximum likelihood type"; informally, M-estimators are estimators that are obtained by solving loss minimization problems (like minimizing the negative log likelihood). (See appendix for full details on M-estimators.)

For M-estimators, under appropriate regularity conditions, for large n, the sampling distribution of an estimate $\hat{\theta}$ of a population parameter θ is:

1. approximately normal;
2. centered at the true population parameter θ;
3. with a variance V/n (i.e., decreasing proportionally to $1/n$) that can be estimated from data.

M-estimators and robust standard errors

In the next few slides, we discuss how this approach applies to OLS.

In particular, applying the M-estimation approach yields an estimated standard error $\widehat{\text{SE}} = \sqrt{\hat{V}/n}$.

This estimate is called a *robust standard error*.

OLS: Relaxed assumptions

Consider the following assumptions:

(A1) *Linear population model*: $Y = \vec{X}\boldsymbol{\beta} + \epsilon$.

(A2) Observations $(\vec{X}_i, Y_i)$ are independent draws from the same population, with all features from (A1) observed.

(A3′) *Conditional zero mean*: $\mathbb{E}[\epsilon \mid \vec{X}] = 0$.

Note (A3′) is weaker than (A3) since constant error variance is no longer assumed; and (A4) – normally distributed errors – is entirely removed.

Robust standard errors for OLS

Since OLS is an M-estimator, we can show under (A1)-(A3') (together with appropriate regularity conditions) that for large n, conditional on the feature matrix $\mathbf{X}$, the OLS coefficient vector $\hat{\boldsymbol{\beta}}$ is:

1. approximately normal;
2. centered at the true coefficient vector $\boldsymbol{\beta}$;
3. with a variance that shrinks inversely with n, and that can be estimated from data.

The resulting standard error is the OLS robust standard error, and can be estimated using the `sandwich` package in R.

Note that the computation of OLS coefficients remains the same – all that we have changed is how we estimate their uncertainty.

"Appropriate regularity conditions" [$*$]

For OLS, the key additional conditions behind the large-sample theory are:

- *Fixed number of features*: The number of features p stays fixed as the sample size n grows.
- *Sufficiently light tails*: Features and errors have sufficiently many finite moments. Finite fourth moments are one sufficient condition.
- *Identifiable coefficients*: There is enough variation in the population features to distinguish the coefficients; no feature is an exact linear combination of the others.

These conditions let us apply the law of large numbers and the CLT, and consistently estimate the variance of the OLS coefficients.

The appendix gives precise mathematical statements.

Robust standard errors in R: Housing example

Here we use HC3, a finite-sample adjustment to the robust standard error. (HC stands for *heteroskedasticity-consistent*.)

```
fm <- lm(price ~ livingArea + bedrooms, data = sh)
V <- sandwich::vcovHC(fm, type = "HC3")
lmtest::coeftest(fm, vcov. = V)
```

```
             Estimate Std. Error ...
(Intercept)  36667.895   6962.756 ...
livingArea     125.405      4.611 ...
bedrooms    -14196.769   2912.782 ...
```

Summary: Standard errors for OLS

The key takeaways are that:

- The *conventional* errors in an OLS regression table are not robust standard errors; in particular, they assume constant variance (homoskedasticity), and are typically derived using the assumption of normal errors.
- When *you* compute standard errors for OLS, you should always use robust standard errors, as they are valid under weaker assumptions (in particular, non-normal errors, and non-constant variance, or heteroskedasticity).
- More generally, additional variants of robust standard errors are available for data with clustered observations, or time-series data with autocorrelation (outside the scope of this class).

Housing example: A robust confidence interval

The coefficient on `livingArea` is 125.405, with HC3 standard error 4.611.

This is the same estimated coefficient as in the conventional table output, but the robust standard error is larger than the conventional standard error (which was 3.527).
The approximate 95% confidence interval using the robust standard error is thus wider: $[116.367, 134.443]$.

Maximum likelihood estimators

MLEs are M-estimators

Maximum likelihood estimators also belong to the M-estimator family.

With a correctly specified model, independent sampling, and suitable regularity conditions, each estimated parameter has an approximately normal sampling distribution for large n.

We apply this result to logistic regression, keeping its population model and independent sampling assumptions.

Example: Logistic regression

Assumptions:

(B1) *Population model*: For a given covariate vector $\vec{X}$, there are *parameters* $\beta_0, \ldots, \beta_p$ such that:

$$P(Y = 1|\vec{X}) = \frac{\exp(\vec{X}\boldsymbol{\beta})}{1 + \exp(\vec{X}\boldsymbol{\beta})} = 1 - \mathbb{P}(Y = 0|\vec{X}),$$

where Y is the associated (binary) outcome.

(Note: $\vec{X}\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.)

(B2) The sample data $\mathbf{X}, \mathbf{Y}$ consists of n independent draws from the population model.

Logistic regression: Example

R reports estimated standard errors as part of the logistic regression table.

An example with the CORIS dataset:

```
Call:
glm(formula = chd ~ ., family = "binomial", data = coris)
...
Coefficients:
             Estimate Std. Error ...
(Intercept) -0.878545   0.123218 ...
sbp          0.133308   0.117452 ...
tobacco      0.364578   0.122187 ...
ldl          0.360181   0.123554 ...
...
```

Logistic regression: Example

In this logistic regression, the coefficient on `ldl` is 0.360, with $\widehat{SE} = 0.124$.

Interpretation: For large n, under assumptions (B1)-(B2), the sampling distribution of the estimated coefficient $\hat{\beta}_{\text{ldl}}$ is approximately normal, with:

$$\hat{\beta}_{\text{ldl}} \approx \mathcal{N}\left(\beta_{\text{ldl}}, 0.124^2\right).$$

Here β_{ldl} is the *true* coefficient under (B1)-(B2).

We can build confidence intervals as before: a 95% confidence interval for this coefficient is: $[0.117, 0.603]$.

More on MLEs: Fisher information [$*$]

In fact, for MLEs, we can be more explicit in characterizing the variance $\mathbf{V}$ that appears in asymptotic normality for M-estimators.

Under correct specification and the MLE regularity conditions, for large n:

$$\hat{\boldsymbol{\theta}} \approx \mathcal{N}\left(\boldsymbol{\theta}^*, \frac{1}{n}(\mathcal{I}(\boldsymbol{\theta}^*))^{-1}\right),$$

where $\mathcal{I}(\boldsymbol{\theta}^*)$ is the *Fisher information matrix* at the true parameter $\boldsymbol{\theta}^*$.

Further, it can be shown that $\mathcal{I}(\boldsymbol{\theta}^*)$ can be estimated from the single data sample you have. (See Appendix.)

More on MLEs: Asymptotic efficiency [$*$]

If you actually *believe* the parametric population model specified by the likelihood is correct, then MLEs have an additional important property (see Appendix):

Asymptotic efficiency: Under suitable regularity conditions, among regular, "consistent" estimators of the true $\boldsymbol{\theta}^*$ (i.e., estimators that converge in probability to the true $\boldsymbol{\theta}^*$ as $n \to \infty$), the MLE has the *smallest* asymptotic variance. (See appendix.)

So *if* you really believe the model (a big "if"!),
this is a strong theoretical justification for using the MLE.

Discussion

What we've seen

We've used the perspective of M-estimation to provide a unified approach to standard errors and confidence intervals across sample means, OLS linear regression, and logistic regression.

We found that for such estimators, for large n:

- the sampling distribution is well approximated by a *normal distribution*,
- centered at the true parameter, and
- with a variance we can estimate from data.

While powerful, there are also significant limits...

Limitation (1): Asymptotics

Throughout our discussion we've talked about *asymptotic* normality, which requires "large" sample size n.

How "large" depends on context, and in particular, typically depends on how many features or parameters are being estimated.

When the number of features p is large, the sample size n needed for accurate estimation and inference is much larger as well.

(The setting where p grows together with n is a practically important and advanced topic in statistical inference referred to as *high-dimensional* statistics.)

Limitation (2): "Parameters"

Constructing a confidence interval for a parameter requires accepting that a *true parameter exists*.

For example, if we don't even believe house prices are linear in features, then there is no "true coefficient" for `livingArea`.

So for example, assumptions (A1) for OLS and (B1) for logistic regression (existence of coefficients in the population model) are essential to what we've presented.

A note on interpretability

Even when we don't necessarily believe assumptions such as (A1) or (B1) hold, it can be very useful to still "pretend" such assumptions hold, because they lead to *interpretable* parameters:

We can interpret a coefficient as informing us about the *association* between a feature and the outcome (holding other features constant).

Limitation (3): Beyond M-estimators

The normal approximation and standard error formulas we derived do not apply to other learning algorithms in general:

- *Lasso and ridge*: M-estimation theory does not apply directly in the presence of regularization.
- *Trees and random forests*: M-estimation does not apply to these learning algorithms that use discrete optimization.
- *Neural networks*: Different weight vectors can represent the same prediction function, so individual weight parameters need not even be uniquely *identifiable*.

Appendix: M-estimation [$*$]

A broader perspective: M-estimators [$*$]

The sample mean, OLS, and logistic regression belong to the broader class of estimators called *M-estimators*.[2]

Informally, an M-estimator $\hat{\boldsymbol{\theta}}$ has three ingredients:

1. For each possible parameter vector $\boldsymbol{\theta}$ and outcome $\boldsymbol{Y}$, a *loss function* $\ell(\boldsymbol{Y}; \boldsymbol{\theta})$ is defined.
2. The estimator $\hat{\boldsymbol{\theta}}$ is constructed by minimizing the *average empirical loss* on the training data: compute the loss at each training sample, then find the $\hat{\boldsymbol{\theta}}$ that minimizes the average training loss.
3. The target true parameter — denoted $\boldsymbol{\theta}^*$ — is the *unique* minimizer of the *expected population loss*, i.e., the average of $\ell(\boldsymbol{Y}; \boldsymbol{\theta})$ over the population distribution.

   The requirement of uniqueness is called *identifiability*.

[2] The "M" stands for "maximum likelihood-type" estimators; we'll see why shortly.

Formal definition: M-estimators [$*$]

An *M-estimator* $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}^*$ is defined via a *loss function* $\ell(\boldsymbol{Y}; \boldsymbol{\theta})$ with two properties:

1. $\hat{\boldsymbol{\theta}}$ minimizes the *average empirical loss*:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\tilde{\boldsymbol{\theta}}} \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{Y}_i, \tilde{\boldsymbol{\theta}});$$

2. The true parameter $\boldsymbol{\theta}^*$ *uniquely* minimizes the *expected population loss*:

$$\boldsymbol{\theta}^* = \arg\min_{\tilde{\boldsymbol{\theta}}} \mathbb{E}_{\boldsymbol{Y}}[\ell(\boldsymbol{Y}; \tilde{\boldsymbol{\theta}})].$$

The uniqueness requirement is called *identifiability*.

Example: Sample mean as an M-estimator [$*$]

(1) We showed previously that the sample mean minimizes the mean squared error loss (MSE):

$$\bar{Y} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \theta)^2.$$

Here the loss function is the squared error $\ell(Y_i, \theta) = (Y_i - \theta)^2$.

(2) We also showed previously that:

$$\mu = \arg\min_{\theta} \mathbb{E}_Y[(Y - \theta)^2] = \arg\min_{\theta} \mathbb{E}_Y[\ell(Y, \theta)],$$

i.e., the population mean minimizes expected loss across the population.

Example: OLS as an M-estimator (1) [$*$]

The OLS coefficients $\hat{\boldsymbol{\beta}}$ minimize mean squared error on the training data:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i \boldsymbol{\beta})^2.$$

So the loss, given features $\vec{X}$, is the squared error $\ell(Y; \vec{X}, \boldsymbol{\beta}) = (Y - \vec{X}\boldsymbol{\beta})^2$.
(Note that here we consider a loss function for the outcome *given* features.)

Example: OLS as an M-estimator (2) [$*$]

In Lecture 7, we had shown that in general, the conditional expectation $f^*(\vec{X}) = \mathbb{E}_Y[Y|\vec{X}]$ minimizes expected population squared error $\mathbb{E}_Y[(Y - f(\vec{X}))^2|\vec{X}]$.

Under (A1) and the conditional-zero-mean part of (A3), the conditional expectation is:

$$f^*(\vec{X}) = \sum_{j=1}^{p} X_j \beta_j = \vec{X}\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ are the *true* parameters in the population model.

Thus the true $\boldsymbol{\beta}$ minimizes expected population squared error. This minimizer is unique when the features identify the coefficients. Constant error variance is not needed.

Other M-estimators [$*$]

M-estimation is a very broad and useful approach to statistical inference!

For example:

- The *sample median* can be shown to be the M-estimator for the *population median*, with loss function $\ell(Y;\theta) = |Y - \theta|$ (the absolute deviation);
- *p-quantile regression* (see problem set) can be shown to be the M-estimator for *conditional population p-quantiles*, with loss function given features $\vec{X}$:

$$\ell_p(Y;\boldsymbol{\beta},\vec{X}) = p[Y - \vec{X}\boldsymbol{\beta}]^+ + (1-p)[\vec{X}\boldsymbol{\beta} - Y]^+.$$

Asymptotic normality of M-estimators [$*$]

Why are M-estimators useful?

With independent observations and sufficient mathematical "regularity" conditions,
we can show that for large n, an M-estimator $\hat{\boldsymbol{\theta}}$ has a sampling distribution:

1. that is approximately normal;
2. centered at $\boldsymbol{\theta}^*$ (the corresponding population parameter vector);
3. with a covariance matrix $\mathbf{V}/n$.

Further, it can be shown that the covariance matrix $\mathbf{V}$ can be estimated from the single data sample that you observe.

The following slides give the mathematical details.

M-estimation theory [$*$]

We now develop the theory, with technical details on:

1. Regularity conditions for asymptotic normality of M-estimators.
2. Derivation of the asymptotic variance formula for M-estimators.
3. Estimation of the asymptotic variance in practice.

For simplicity, we start in a setting where observations Y are real-valued (e.g., arrival delays), and the parameter θ of interest is a scalar (e.g., the population mean), then generalize to vector-valued parameters.

Note: This material is advanced technically; it is optional and intended for students interested in the theoretical foundations.

Setup: M-estimator definition [$*$]

An M-estimator $\hat{\theta}$ is obtained:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i; \theta),$$

where:

- $\ell(Y; \theta)$ is a loss function; and
- $Y_1, \ldots, Y_n$ are independent samples from the (unknown) population distribution.

Let $L(\theta) = \mathbb{E}_Y[\ell(Y; \theta)]$ denote the *population loss*, and let:

$$\theta^* = \arg\min_{\theta} L(\theta).$$

be the *true parameter* (the minimizer of the population loss).

Regularity conditions – part 1 [$*$]

We start with two regularity conditions:

1. *Identifiability*: The population loss stays bounded away from its minimum outside any fixed neighborhood of θ^*. For every $\varepsilon > 0$,

$$\inf_{|\theta - \theta^*| \geq \varepsilon} [L(\theta) - L(\theta^*)] > 0.$$

2. *Uniform law of large numbers*: As $n \to \infty$, the empirical average loss $(1/n) \sum_{i=1}^{n} \ell(Y_i; \theta)$ converges *uniformly in* θ to the expected population loss $\mathbb{E}_Y[\ell(Y; \theta)]$.

Consistency of M-estimators [∗]

Under the previous two regularity conditions, our estimator is *consistent*: Consistency means that as our sample size n grows, our estimator $\hat{\theta}$ converges in probability to the true parameter θ^*.

Theorem (Consistency)

Under the previous two regularity conditions, the M-estimator is consistent:

$$\hat{\theta} \xrightarrow{p} \theta^* \quad \text{as } n \to \infty.$$

This is a prerequisite for asymptotic normality, which describes the distribution of $\hat{\theta}$ *around* θ^*.

Regularity conditions – part 2 [$*$]

We now add four more regularity conditions:

3. *Smoothness*: θ^* is interior to the parameter space. The loss $\ell(y;\theta)$ is twice continuously differentiable in θ for all y, and throughout a neighborhood of θ^*, there exists a function $M(Y)$ such that:
$$|\ell''(Y;\theta)| \le M(Y), \qquad \text{with } \mathbb{E}[M(Y)] < \infty.$$
4. *Interchange*: We can interchange expectations and derivatives. Formally:
$$L'(\theta) = \mathbb{E}[\ell'(Y;\theta)], \quad L''(\theta) = \mathbb{E}[\ell''(Y;\theta)].$$
Here $\ell'(Y;\theta) = \frac{\partial}{\partial\theta}\ell(Y;\theta)$ and $\ell''(Y;\theta) = \frac{\partial^2}{\partial\theta^2}\ell(Y;\theta)$.
5. *Non-degeneracy*: The second derivative of population loss at θ^* is positive:
$$H = L''(\theta^*) = \mathbb{E}[\ell''(Y;\theta^*)] > 0.$$
6. *Finite variance*: The variance of the derivative is finite:
$$J = \mathrm{Var}[\ell'(Y;\theta^*)] = \mathbb{E}[(\ell'(Y;\theta^*))^2] - (L'(\theta^*))^2.$$
(Note: By first-order optimality, $L'(\theta^*) = 0$, so $J = \mathbb{E}[(\ell'(Y;\theta^*))^2]$.)

Asymptotic normality theorem [$*$]

Theorem (Asymptotic normality of M-estimators)

Under the regularity conditions stated above, as $n \to \infty$ the following convergence in distribution holds:

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, V),$$

where the asymptotic variance is:

$$V = \frac{J}{H^2}.$$

This is often called the *sandwich formula* for the asymptotic variance.

Equivalently: $\hat{\theta} \approx \mathcal{N}(\theta^*, V/n)$ for large n.

Derivation: Key idea [$*$]

The proof uses a Taylor expansion of the first-order condition.

Since $\hat{\theta}$ minimizes the empirical loss, the derivative at $\hat{\theta}$ is zero:

$$\frac{1}{n}\sum_{i=1}^{n} \ell'(Y_i; \hat{\theta}) = 0.$$

Taylor expanding around θ^*:

$$\frac{1}{n}\sum_{i=1}^{n} \ell'(Y_i; \theta^*) + \frac{1}{n}\sum_{i=1}^{n} \ell''(Y_i; \tilde{\theta})(\hat{\theta} - \theta^*) = 0,$$

where $\tilde{\theta}$ is between $\hat{\theta}$ and θ^* (by the mean value theorem).

Derivation: Key idea (continued) [∗]

Rearranging:

$$\sqrt{n}(\hat{\theta} - \theta^*) = -\left[\frac{1}{n}\sum_{i=1}^{n} \ell''(Y_i; \tilde{\theta})\right]^{-1} \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \ell'(Y_i; \theta^*)\right].$$

Now apply:

- *Law of large numbers*: $\frac{1}{n}\sum_{i=1}^{n} \ell''(Y_i; \tilde{\theta}) \to \boldsymbol{H}$ (by the local uniform law of large numbers and consistency of $\tilde{\theta}$).
- *Central limit theorem*: $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \ell'(Y_i; \theta^*) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{J})$.

Combining these (using Slutsky's theorem):

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, \frac{\boldsymbol{J}}{\boldsymbol{H}^2}\right).$$

Estimating the asymptotic variance [$*$]

To construct confidence intervals, we need to estimate $V = J/H^2$.

We use the *plug-in principle*: Replace population quantities with sample analogs.

1. Estimate H using the *empirical second derivative*:

$$\hat{H} = \frac{1}{n} \sum_{i=1}^{n} \ell''(Y_i; \hat{\theta}).$$

2. Estimate J using the *empirical variance of derivatives*:

$$\hat{J} = \frac{1}{n} \sum_{i=1}^{n} (\ell'(Y_i; \hat{\theta}))^2.$$

(We can do this since $L'(\theta^*) = 0$, so $\hat{L}'(\hat{\theta}) \approx 0$.)

Sandwich estimator [$*$]

The estimated asymptotic variance is:

$$\hat{V} = \frac{\hat{J}}{\hat{H}^2}.$$

This is called the *sandwich estimator*. It uses independent observations; for OLS it yields the *Huber-White (HC0)* covariance estimator.

The resulting estimated standard error (also called a *robust* standard error) for $\hat{\theta}$ is:

$$\widehat{\mathrm{SE}} = \sqrt{\frac{\hat{V}}{n}}.$$

Generalization: Vector parameters [$*$]

All of the theory above extends to the case where $\boldsymbol{\theta}$ is a vector of parameters.

The key generalizations are:

- Instead of derivatives ℓ' and ℓ'', we use *gradients* $\nabla_{\boldsymbol{\theta}} \ell$ (a vector) and *Hessians* $\nabla^2_{\boldsymbol{\theta}} \ell$ (a matrix).
- Instead of scalars H and J, we have *matrices*:

$$\mathbf{H} = \mathbb{E}[\nabla^2_{\boldsymbol{\theta}} \ell(Y; \boldsymbol{\theta}^*)],$$
$$\mathbf{J} = \mathbb{E}[(\nabla_{\boldsymbol{\theta}} \ell(Y; \boldsymbol{\theta}^*))(\nabla_{\boldsymbol{\theta}} \ell(Y; \boldsymbol{\theta}^*))^\top].$$

- The sandwich formula becomes: $\mathbf{V} = \mathbf{H}^{-1}\mathbf{J}\mathbf{H}^{-1}$ (a matrix).

Consistency (vector case) [∗]

Similarly to the scalar case, under appropriate corresponding regularity conditions (identifiability and uniform law of large numbers), the vector M-estimator is consistent (proof omitted).

Theorem (Consistency)

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^* \quad \text{as } n \to \infty.$$

This ensures our vector of estimated parameters $\hat{\boldsymbol{\theta}}$ converges to the true vector $\boldsymbol{\theta}^*$ as the sample size grows.

Generalization: Vector parameters [$*$]

The asymptotic normality result becomes:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}),$$

where $\mathbf{V} = \mathbf{H}^{-1}\mathbf{J}\mathbf{H}^{-1}$ is the asymptotic covariance matrix.

For inference on individual components $\hat{\theta}_j$:

- The asymptotic variance is $[\mathbf{V}]_{jj}/n$ (the j-th diagonal entry of $\mathbf{V}/n$).
- The standard error is $\text{SE}_j = \sqrt{[\mathbf{V}]_{jj}/n}$.
- A 95% confidence interval is: $[\hat{\theta}_j - 1.96\text{SE}_j, \hat{\theta}_j + 1.96\text{SE}_j]$.

Estimating the asymptotic variance (vector) [$*$]

For independent observations, we use the plug-in principle to estimate the sandwich formula $\mathbf{V} = \mathbf{H}^{-1}\mathbf{J}\mathbf{H}^{-1}$.

1. Estimate $\mathbf{H}$ using the *empirical average Hessian*:

$$\hat{\mathbf{H}} = \frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \ell(Y_i; \hat{\boldsymbol{\theta}}).$$

2. Estimate $\mathbf{J}$ using the *empirical outer product of gradients*:

$$\hat{\mathbf{J}} = \frac{1}{n}\sum_{i=1}^{n} (\nabla_{\boldsymbol{\theta}} \ell(Y_i; \hat{\boldsymbol{\theta}}))(\nabla_{\boldsymbol{\theta}} \ell(Y_i; \hat{\boldsymbol{\theta}}))^\top.$$

The estimated covariance of $\hat{\boldsymbol{\theta}}$ is

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\theta}}) = \frac{\hat{\mathbf{V}}}{n} = \frac{1}{n}\hat{\mathbf{H}}^{-1}\hat{\mathbf{J}}\hat{\mathbf{H}}^{-1}.$$

The estimated standard error for a single parameter $\hat{\theta}_j$ is $\widehat{\mathrm{SE}}_j = \sqrt{[\hat{\mathbf{V}}]_{jj}/n}$.

Example: OLS linear regression [$*$]

For OLS, $\boldsymbol{\theta} = \boldsymbol{\beta}$ and the loss is $\ell(Y; \boldsymbol{\beta}, \vec{X}) = (Y - \vec{X}\boldsymbol{\beta})^2$. Here $\vec{X}$ is a feature vector; we view it as a $1 \times p$ *row* vector.

The gradient (a $p \times 1$ vector) at $(\vec{X}, Y)$ is:

$$\nabla_{\boldsymbol{\beta}} \ell(Y; \boldsymbol{\beta}, \vec{X}) = -2(Y - \vec{X}\boldsymbol{\beta})\vec{X}^\top.$$

The Hessian (a $p \times p$ matrix) is:

$$\nabla^2_{\boldsymbol{\beta}} \ell(Y; \boldsymbol{\beta}, \vec{X}) = 2\vec{X}^\top \vec{X}.$$

OLS asymptotics: Sufficient conditions [$*$]

For the OLS normal approximation and heteroskedasticity-consistent (HC) standard errors, one sufficient set of conditions is:

- *Fixed dimension*: p is fixed as $n \to \infty$.
- *Independent sampling*: $(\mathbf{X}_i, Y_i)$ are i.i.d., with $Y_i = \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i$ and $\mathbb{E}[\epsilon_i \mid \mathbf{X}_i] = 0$.
- *Finite moments*: $\mathbb{E}[\|\mathbf{X}_i\|^4] < \infty$ and $\mathbb{E}[\epsilon_i^4] < \infty$; hence $\mathbb{E}[\epsilon_i^2\|\mathbf{X}_i\|^2] < \infty$.
- *Full rank*: $\mathbf{Q} = \mathbb{E}[\mathbf{X}_i^\top \mathbf{X}_i]$ is positive definite: $\mathbf{a}^\top \mathbf{Q}\mathbf{a} > 0$ for every $\mathbf{a} \neq \mathbf{0}$.

Then $\mathbf{X}^\top\mathbf{X}/n \xrightarrow{p} \mathbf{Q}$, so the sample feature matrix has full column rank with probability tending to one. Error variance may depend on the features.

OLS with heteroskedastic errors [$*$]

Using the OLS gradient and Hessian, the population matrices are:

$$\mathbf{H} = 2\mathbb{E}[\mathbf{X}_i^\top \mathbf{X}_i] = 2\mathbf{Q},$$
$$\mathbf{J} = 4\mathbb{E}[\epsilon_i^2 \mathbf{X}_i^\top \mathbf{X}_i].$$

Thus the asymptotic covariance matrix is:

$$\mathbf{V} = \mathbf{H}^{-1}\mathbf{J}\mathbf{H}^{-1} = \mathbf{Q}^{-1}\mathbb{E}[\epsilon_i^2 \mathbf{X}_i^\top \mathbf{X}_i]\mathbf{Q}^{-1}.$$

No normality or constant-variance assumption is needed. For a nondegenerate normal interval for β_j, require $[\mathbf{V}]_{jj} > 0$.

OLS sandwich: The Huber-White (HC0) estimator [$*$]

Let $r_i = Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}$ be the fitted residual. The sample analogs are:

$$\hat{\mathbf{H}} = \frac{2}{n}\mathbf{X}^\top\mathbf{X}, \qquad \hat{\mathbf{J}} = \frac{4}{n}\sum_{i=1}^{n} r_i^2\mathbf{X}_i^\top\mathbf{X}_i.$$

Substituting into the sandwich gives:

$$\begin{aligned}\widehat{\mathrm{Cov}}_{\mathrm{HC0}}(\hat{\boldsymbol{\beta}}) &= \frac{1}{n}\hat{\mathbf{H}}^{-1}\hat{\mathbf{J}}\hat{\mathbf{H}}^{-1} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\left(\sum_{i=1}^{n} r_i^2\mathbf{X}_i^\top\mathbf{X}_i\right)(\mathbf{X}^\top\mathbf{X})^{-1}.\end{aligned}$$

Robust standard errors are the square roots of the diagonal entries.

HC3: A finite-sample adjustment [$*$]

HC0 uses the squared fitted residual r_i^2 to estimate the error variance for observation i.

HC3 adjusts these residuals using the *leverage* (from the *hat matrix*):

$$h_{ii} = \mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top.$$

In the HC0 formula, replace each residual weight by:

$$r_i^2 \quad \longrightarrow \quad \frac{r_i^2}{(1-h_{ii})^2}.$$

This inflates the residual weights, especially for high-leverage observations. HC0 and HC3 have the same large-sample target under the usual regularity conditions.

Constant variance as a special case [$*$]

If we also assume $\mathrm{Var}(\epsilon_i \mid \mathbf{X}_i) = \sigma^2$, then:

$$\mathbf{J} = 4\sigma^2\mathbf{Q},$$
$$\mathbf{V} = (2\mathbf{Q})^{-1}(4\sigma^2\mathbf{Q})(2\mathbf{Q})^{-1} = \sigma^2\mathbf{Q}^{-1}.$$

Estimating $\mathbf{Q}$ by $\mathbf{X}^\top\mathbf{X}/n$ and σ^2 by $\hat{\sigma}^2 = \sum_i r_i^2/(n-p)$ gives the conventional covariance estimate:

$$\widehat{\mathrm{Cov}}_{\mathrm{conv}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}.$$

Normality is not needed for this large-sample result; constant variance is. Robust standard errors remain valid in this special case too.

Robust standard errors in more general settings [∗]

The robust standard error approach is quite flexible for a variety of aspects of data encountered in practice:

- More general robust standard errors can be computed even if there is correlation structure within *clusters* of observations; these are called *cluster-robust* standard errors.
- In time series data, robust standard errors can be developed that account for *autocorrelation* of the time series; these are called *Newey-West* robust standard errors.

Valid application of these robust standard errors also requires corresponding regularity conditions; we omit the details here.

MLEs are M-estimators [$*$]

Earlier in the class we learned about *maximum likelihood estimation* (MLE).

Let's recall the basic idea, now written out in the notation of parametric estimation:

1. *Distributional assumption*: "Pretend" that the population model is a distribution $f(\boldsymbol{Y}; \boldsymbol{\theta})$ with a known structure, but unknown parameters $\boldsymbol{\theta}$.
2. *Likelihood computation*: For each choice of parameters $\boldsymbol{\theta}$, compute the chance of seeing the training data $\mathbf{Y}$, given a particular value of $\boldsymbol{\theta}$.
3. *Optimization*: Pick the parameter values $\hat{\boldsymbol{\theta}}$ that maximize the likelihood.

MLEs are M-estimators [$*$]

Recall that the *log likelihood function* (LLF) is: $\log f(\boldsymbol{Y}; \boldsymbol{\theta})$.

A key fact: MLEs are M-estimators, with loss function
$\ell(\boldsymbol{Y}; \boldsymbol{\theta}) = -\text{LLF} = -\log f(\boldsymbol{Y}; \boldsymbol{\theta})$. (The following slides show this connection.)

The asymptotic normality theory applies to *regular* MLEs under correct model specification and the required regularity conditions.

A key example: *logistic regression*.

MLEs: Loss function [$*$]

For MLEs, the loss function is $\ell(\boldsymbol{Y}; \boldsymbol{\theta}) = -\log f(\boldsymbol{Y}; \boldsymbol{\theta})$ (negative log-likelihood).

The first condition for an M-estimator is immediately satisfied:

Since the MLE $\hat{\boldsymbol{\theta}}$ maximizes likelihood, it *minimizes the empirical average loss* with ℓ defined as above.

MLEs: Identification [$*$]

To show MLE is an M-estimator, we must also verify the second condition:

The true parameter $\boldsymbol{\theta}^*$ *uniquely* minimizes the *expected population loss* $\boldsymbol{L}(\boldsymbol{\theta})$.

Proof:

Consider the difference $\boldsymbol{L}(\boldsymbol{\theta}) - \boldsymbol{L}(\boldsymbol{\theta}^*)$:

$$\begin{aligned} \boldsymbol{L}(\boldsymbol{\theta}) - \boldsymbol{L}(\boldsymbol{\theta}^*) &= \mathbb{E}_{\boldsymbol{\theta}^*}[-\log f(\boldsymbol{Y};\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}^*}[-\log f(\boldsymbol{Y};\boldsymbol{\theta}^*)] \\ &= \mathbb{E}_{\boldsymbol{\theta}^*}\left[\log f(\boldsymbol{Y};\boldsymbol{\theta}^*) - \log f(\boldsymbol{Y};\boldsymbol{\theta})\right] \\ &= \mathbb{E}_{\boldsymbol{\theta}^*}\left[\log \frac{f(\boldsymbol{Y};\boldsymbol{\theta}^*)}{f(\boldsymbol{Y};\boldsymbol{\theta})}\right] \end{aligned}$$

This quantity is the *Kullback-Leibler (KL) divergence* $\boldsymbol{D_{KL}}(f_{\boldsymbol{\theta}^*} || f_{\boldsymbol{\theta}})$.

MLEs: Identification [$*$]

Proof: (Continued)

By *Jensen's Inequality*, since $-\log(x)$ is convex:

$$\mathbb{E}_{\theta^*}\left[-\log\frac{f(Y;\boldsymbol{\theta})}{f(Y;\boldsymbol{\theta}^*)}\right] \geq -\log\left(\mathbb{E}_{\theta^*}\left[\frac{f(Y;\boldsymbol{\theta})}{f(Y;\boldsymbol{\theta}^*)}\right]\right)$$

The term inside the $\log$ is:

$$\mathbb{E}_{\theta^*}\left[\frac{f(Y;\boldsymbol{\theta})}{f(Y;\boldsymbol{\theta}^*)}\right] = \int f(y;\boldsymbol{\theta}^*)\left(\frac{f(y;\boldsymbol{\theta})}{f(y;\boldsymbol{\theta}^*)}\right)dy = \int f(y;\boldsymbol{\theta})dy = 1$$

Therefore:

$$D_{KL}(f_{\theta^*}||f_{\theta}) = -\mathbb{E}_{\theta^*}\left[\log\frac{f(Y;\boldsymbol{\theta})}{f(Y;\boldsymbol{\theta}^*)}\right] \geq -\log(1) = 0$$

So, $L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) \geq 0$, which means $L(\boldsymbol{\theta}^*) \leq L(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$. Assuming different parameter values give different distributions, the inequality is strict whenever $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$, establishing uniqueness.

MLEs: Fisher information matrix [$*$]

For MLEs, the loss function is $\ell(\boldsymbol{Y}; \boldsymbol{\theta}) = -\log f(\boldsymbol{Y}; \boldsymbol{\theta})$ (negative log-likelihood).

Under the assumption that the parametric model is *correctly specified* (i.e., $\boldsymbol{Y}$ truly follows $f(y; \boldsymbol{\theta}^*)$), a special relationship holds:

The *Fisher information matrix* is:

$$\mathcal{I}(\boldsymbol{\theta}^*) = -\mathbb{E}\left[\nabla^2_{\boldsymbol{\theta}} \log f(\boldsymbol{Y}; \boldsymbol{\theta}^*)\right] = \mathbb{E}\left[(\nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{Y}; \boldsymbol{\theta}^*))(\nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{Y}; \boldsymbol{\theta}^*))^\mathsf{T}\right].$$

This is the *information matrix equality*, and it implies:

$$\mathbf{H} = \mathbf{J} = \mathcal{I}(\boldsymbol{\theta}^*).$$

(Recall $\mathbf{H} = \mathbb{E}[\nabla^2_{\boldsymbol{\theta}} \ell]$ and $\mathbf{J} = \mathbb{E}[(\nabla_{\boldsymbol{\theta}} \ell)(\nabla_{\boldsymbol{\theta}} \ell)^\mathsf{T}]$.)

MLEs: Asymptotic normality [$*$]

Therefore, for MLEs under correct specification, the sandwich formula for $\mathbf{V}$ simplifies:

$$\mathbf{V} = \mathbf{H}^{-1}\mathbf{J}\mathbf{H}^{-1} = (\mathcal{I}(\boldsymbol{\theta}^*))^{-1}(\mathcal{I}(\boldsymbol{\theta}^*))(\mathcal{I}(\boldsymbol{\theta}^*))^{-1} = (\mathcal{I}(\boldsymbol{\theta}^*))^{-1}.$$

So the asymptotic distribution simplifies to:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, (\mathcal{I}(\boldsymbol{\theta}^*))^{-1}\right).$$

MLEs: Asymptotic efficiency [$*$]

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, (\mathcal{I}(\boldsymbol{\theta}^*))^{-1}\right)$$

This is the classical result for MLEs, and it's the basis for asymptotic efficiency:

The Cramér-Rao lower bound states that the covariance matrix of any unbiased estimator, $\mathbf{C}$, must satisfy $\mathbf{C} - \frac{1}{n}(\mathcal{I}(\boldsymbol{\theta}^*))^{-1}$ being positive semidefinite.

For scalar θ, this means that asymptotically in n, the variance of any unbiased estimator is at least as large as the variance of the MLE.