ECON5128 MRes Econometrics 2

Week 7: Misspecification and Maximum Likelihood

Santiago Montoya-Blandón

February 25, 2025

Lecture outline

Models

- Econometrics aims to understand economic phenomena using data
- We use *models* as abstractions to highlight key features without having to specify all real-life complex economic interactions
- Models are then *approximations* and always incorrect to some degree

Models

- Econometrics aims to understand economic phenomena using data
- We use *models* as abstractions to highlight key features without having to specify all real-life complex economic interactions
- Models are then *approximations* and always incorrect to some degree
- Models are in a sense also necessary: learning requires assumptions
- No free lunch theorem for (machine) learning: there is no algorithm that can learn *all* data generating processes (DGPs)
- In optimization terms: no algorithm beats random search if we care about *all possible objectives*
- Intuitively: there is no model that works the best in all scenarios

Objective Functions

- We summarize the value of a given action in the objective function
- Also called cost, criterion, loss, utility, etc.
- In econometrics, usually focus on *statistical* measures:
  - Mean squared error
  - Mean absolute error
- Objective functions *should be* application-specific when possible:
  - Money
  - Consumption
  - Quality-adjusted life years

Objective Functions: Example

- Consider a simple problem: summarize the distribution of random variable Y according to some scalar measure
- That is, provide a number that best describes the distribution

Question for you

What would you choose?

Objective Functions: Example

- Consider a simple problem: summarize the distribution of random variable Y according to some scalar measure
- That is, provide a number that best describes the distribution

Question for you

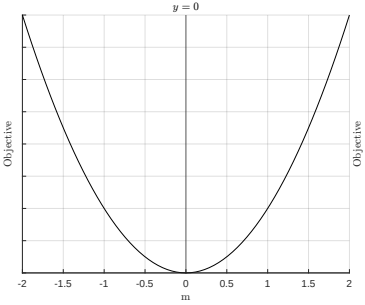
What would you choose?

- As we saw: answer depends on what you mean by *best*
- Different summaries are optimal according to different measures
- Formally:
  - $y \in \mathbb{R}$ is a realization from Y
  - $m \in \mathbb{R}$ represents a scalar measure
  - $\mathbb{E}[L(m, Y)]$ is the expected loss function of using m to summarize Y

Objective Functions: Example

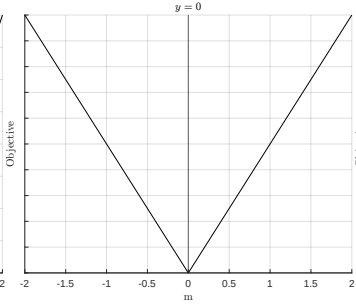
Solutions to $m^* = \arg\min_{m \in \mathbb{R}} \mathbb{E}[L(m, Y)]$ for different objectives (losses) L:

Figure: Squared loss



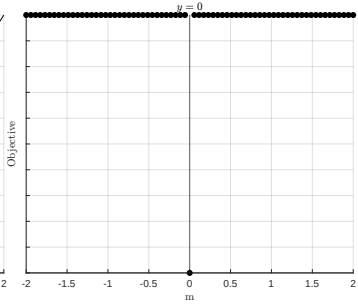
$$L(m, y) = (y - m)^2$$
$$m^* = \mathbb{E}[Y]$$

Figure: Absolute loss



$$L(m, y) = |y - m|$$
$$m^* = \text{median}(Y)$$

Figure: Zero-one loss



$$L(m, y) = \mathbb{I}(y \neq m)$$
$$m^* = \text{mode}(Y)$$

Conditional Mean

- In a modelling context: outcome y and covariates x (also called controls, independent variables, features, etc.)
- A model is a mapping from features to outcome: $f : x \mapsto y$
- Without extra information, we use Mean Squared Error as the objective:

$$\text{MSE}(f) = \mathbb{E}[L(f(x), Y)] = \mathbb{E}[(Y - f(x))^2]$$

- The *conditional mean* turns out to be the optimal predictor in the MSE sense

$$m(x) \coloneqq \mathbb{E}[Y \mid X = x] = \arg\min_{f \in \mathcal{F}} \mathbb{E}[(Y - f(x))^2]$$

- The class of models $\mathcal{F}$ includes all square-integrable functions

Conditional Mean: Example

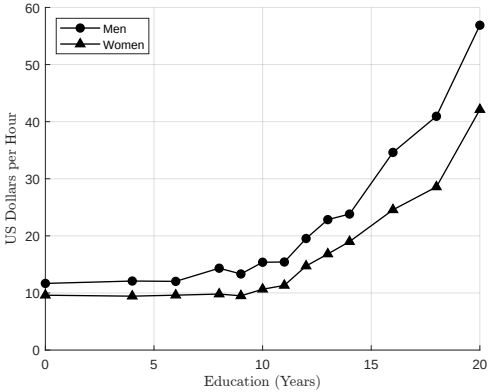


Figure: Conditional expectation of wage as function of education (Hansen, 2022)

Conditional Mean

- No better predictor exists than the conditional mean function (CEF) under MSE as the objective
- CEF is unfeasible to calculate exactly as it depends on the underlying distribution of the data
- Even if this distribution was known, the CEF is difficult to compute when dimensionality and support of covariates X is large

Question for you

Since the conditional mean function is unknown and unfeasible to calculate, what do we do?

Lecture outline

CEF Misspecification

- Note: setup thus far is a *population* problem
- That is, this is not an issue of estimation accuracy, rather an issue of ignorance about true DGP
- Econometricians' answer has been to approximate the CEF using a model
- In practice, we sample observations from the DGP to estimate the model
- *CEF Misspecification*: what happens if our model for the CEF is wrong?

CEF Misspecification

- Note: setup thus far is a *population* problem
- That is, this is not an issue of estimation accuracy, rather an issue of ignorance about true DGP
- Econometricians' answer has been to approximate the CEF using a model
- In practice, we sample observations from the DGP to estimate the model
- *CEF Misspecification*: what happens if our model for the CEF is wrong?
- *Misspecification*: what happens if the assumptions you impose for your model are not satisfied?
- Answer can vary depending on setting and statistical technique used

Best Linear Approximation

- Simplest approximation to the CEF one can consider is a linear one
- Linear predictor: $f(x) = f(x; b) = x^\top b$
- Linear predictor is completely determined by the coefficients $b \in \mathbb{R}^p$
- Mean-squared error as objective: $S_0(b) = \mathbb{E}[(Y - X^\top b)^2]$
- Coefficients yielding *best linear predictor* under MSE?

$$\beta^* := \arg\min_{b \in \mathbb{R}^p} S_0(b)$$

Best Linear Approximation

- Simplest approximation to the CEF one can consider is a linear one
- Linear predictor: $f(x) = f(x; b) = x^\top b$
- Linear predictor is completely determined by the coefficients $b \in \mathbb{R}^p$
- Mean-squared error as objective: $S_0(b) = \mathbb{E}[(Y - X^\top b)^2]$
- Coefficients yielding *best linear predictor* under MSE?

$$\beta^* := \arg\min_{b \in \mathbb{R}^p} S_0(b)$$

- Under regularity conditions (moments and no perfect collinearity), this problem has a closed-form solution

$$\beta^* = \mathbb{E}[XX^\top]^{-1}\,\mathbb{E}[XY] \quad \text{and} \quad f(x; \beta^*) = x^\top \beta^*$$

Best Linear Approximation

- β^* provides the best *linear projection* of the outcome Y onto the space spanned by the covariates X
- Best linear predictor $x^\top \beta^*$ is also the best linear approximation to the CEF
- That is, β^* also solves

$$\beta^* = \arg\min_{b \in \mathbb{R}^p} \mathbb{E}[(m(X) - X^\top b)^2]$$

- This motivates the popularity of linear models: they provide the best linear approximation to the unfeasible CEF problem

Best Linear Approximation

- β^* provides the best *linear projection* of the outcome Y onto the space spanned by the covariates X
- Best linear predictor $x^\top \beta^*$ is also the best linear approximation to the CEF
- That is, β^* also solves

$$\beta^* = \arg\min_{b \in \mathbb{R}^p} \mathbb{E}[(m(X) - X^\top b)^2]$$

- This motivates the popularity of linear models: they provide the best linear approximation to the unfeasible CEF problem
- *Correct specification*: $\exists \beta_0 \in \mathbb{R}^p$ such that $f(x, \beta_0) = x^\top \beta_0 = \mathbb{E}[Y \mid X = x]$ for almost all realizations x
- *Misspecification*: No such "true" β_0 exists

Identification

- For *identification*, we only need to assume X and Y to be square-integrable and for $\mathbb{E}[XX^\top]^{-1}$ to exist
- But what is identified in this scenario?
- Let $e := Y - m(x)$ be the CEF error from using the optimal predictor
- Under *correct specification*: $m(x) = x^\top \beta_0$ (true model is linear)
- This means we can write: $Y = m(X) + e = X^\top \beta_0 + e$
- As β^* provides the best linear approximation, in this case we have

$$\begin{aligned}\beta^* &= \arg\min_{b \in \mathbb{R}^p} \mathbb{E}[(X^\top \beta_0 - X^\top b)^2] \\ &= \arg\min_{b \in \mathbb{R}^p} (\beta_0 - b)^\top \mathbb{E}[XX^\top](\beta_0 - b) = \beta_0\end{aligned}$$

Identification

- Solution $\beta^* = \mathbb{E}[XX^\top]^{-1}\,\mathbb{E}[XY]$ is the *minimizer* of the population problem
- β^* is also an example of a *pseudo-true parameter*
- Under correct specification, there is a true β_0 and $\beta^* = \beta_0$
- Under misspecification, no such true value exists, and all we can hope to recover is β^*
- Also suspiciously similar to the ordinary least squares (OLS) coefficient...
- Remember we can include non-linear transformations of X as covariates without breaking the linearity of the predictor

Linear Projection: Example

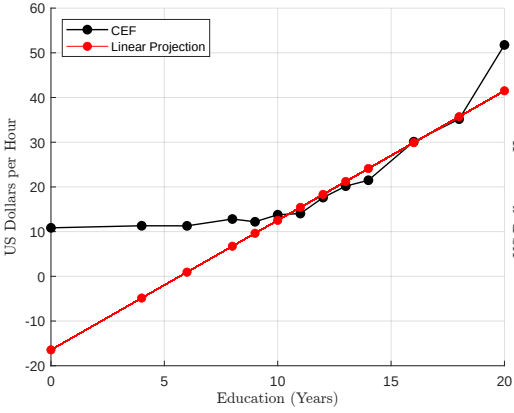


Figure: Linear projection

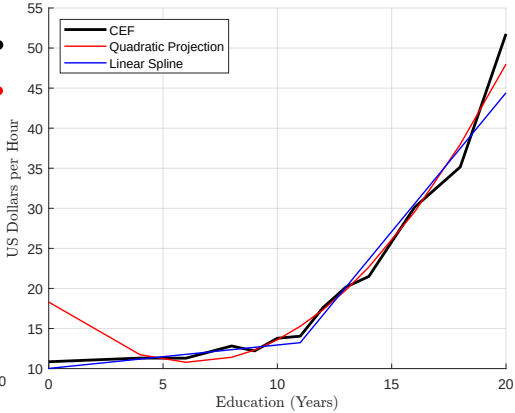


Figure: Linear projection with nonlinear covariates

Lecture outline

Sample Problem

- We have just shown β^* minimizes the *population* objective $S_0(b)$
- In practice, we do not usually have access to the population, and instead only have access to a sample $\{(y_1, x_1), \ldots, (y_n, x_n)\}$
- By analogy, we can replace the expectation in the population objective with a sample average to obtain

$$S_n(b) := \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top b)^2$$

- This defines the *sample objective* function $S_n(b)$
- By the law of large numbers: $S_n(b) \xrightarrow{p} S_0(b)$ for all $b \in \mathbb{R}^p$

Ordinary Least Squares Estimator

- The solution to the sample problem is simply the OLS estimator:

$$\hat{\beta}_n := \arg\min_{b \in \mathbb{R}^p} S_n(b) = \left(\sum_{i=1}^{n} x_i x_i^\top \right)^{-1} \left(\sum_{i=1}^{n} x_i y_i \right)$$

- As β^* minimizes the sum of squared population disturbances, $\hat{\beta}_n$ minimizes the sum of squared residuals
- As $x^\top \beta^*$ provided the best linear projection of Y onto X, $\hat{y} = x^\top \hat{\beta}_n$ provides the *best linear fit* of the observed $(y_1, \ldots, y_n)$ onto $(X_1, \ldots, X_n)$

Ordinary Least Squares: Example

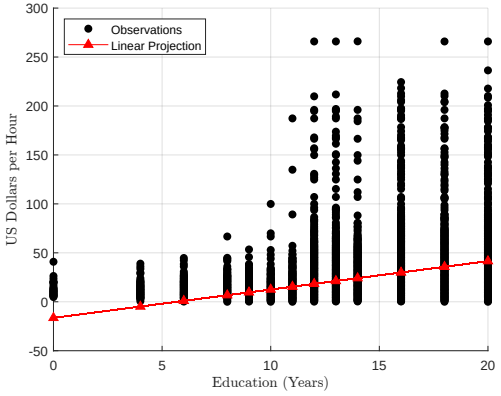


Figure: Line of best fit through wages and education

Ordinary Least Squares: Consistency

- Assuming similar regularity conditions (existence of moments and no perfect multicollinearity in the sample), OLS is consistent
- But what is OLS consistent to?

Ordinary Least Squares: Consistency

- Assuming similar regularity conditions (existence of moments and no perfect multicollinearity in the sample), OLS is consistent
- But what is OLS consistent to?
- Intuition: as sample objective converges to population objective, minimizer of sample problem should converge to minimizer of population problem
- This means we should expect under general conditions: $\hat{\beta}_n \xrightarrow{p} \beta^*$
- As $\beta^* = \beta_0$ under correct specification, and β_0 is undefined under misspecification, OLS is always consistent to the *pseudo-true* β^*!
- Conclusions:
  1. OLS recovers best linear projection under general assumptions
  2. β^* reflects a true causal effect or not depending on *model assumptions*

Lecture outline

Population Problem

- It turns out that similar ideas hold for a non-linear version of least squares
- Recall the CEF $m(x) = \mathbb{E}[Y \mid X = x]$ is the optimal predictor of Y given X
- We covered linear approximations to the CEF in the previous section
- Expand to non-linear predictors: $f(x; \theta)$ is non-linear and depends on a low-dimensional vector of parameters θ
- Continue focusing on MSE as objective of interest:

$$S_0(\theta) = \mathbb{E}[(Y - f(X; \theta))^2]$$

- This defines the population objective for non-linear least squares (NLS)
- Correct specification: $\exists \theta_0 \in \Theta \subset \mathbb{R}^p$ such that $f(x; \theta_0) = m(x)$
- Misspecification: no such "true" θ_0 exists

Sample Problem

- By analogy, we can obtain the sample problem for NLS as

$$S_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i;\theta)]^2$$

- We define the population and sample minimizers as

$$\theta^* := \arg\min_{\theta\in\Theta} S_0(\theta) \quad \text{and} \quad \hat{\theta}_n := \arg\min_{\theta\in\Theta} S_n(\theta)$$

- The NLS estimator $\hat{\theta}_n$ no longer has a closed-form solution and we need to use numerical techniques to minimize the objective
- By the LLN, we still have $S_n(\theta) \xrightarrow{p} S_0(\theta)$ for all $\theta \in \Theta$

Identification

- The population problem now provides the *best non-linear approximation* to the CEF offered by the non-linear function $f(x;\theta)$

$$\theta^* := \arg\min_{\theta\in\Theta} \mathbb{E}\{[m(X) - f(X;\theta)]^2\}$$

- $f(x;\theta^*)$ is the *best non-linear predictor using* f with MSE as criterion
- θ^* is a pseudo-true parameter as $\theta^* = \theta_0$ under correct specification:

$$\theta^* := \arg\min_{\theta\in\Theta} \mathbb{E}\{[f(X;\theta_0) - f(X;\theta)]^2\} = \theta_0$$

- Need to assume global (local) identification to guarantee uniqueness of θ^* in the general case

Consistency

- Under conditions guaranteeing the existence of a minimum for the sample objective and regularity conditions, NLS is consistent
- As the sample objective converges to population objective, we expect the minimizer of the sample to be consistent to the minimizer of the population
- Again, this means we should expect: $\hat{\theta}_n \xrightarrow{p} \theta^*$ as $n \to \infty$
- Under correct specification, this guarantees consistency to true value θ_0

Consistency

- Under conditions guaranteeing the existence of a minimum for the sample objective and regularity conditions, NLS is consistent
- As the sample objective converges to population objective, we expect the minimizer of the sample to be consistent to the minimizer of the population
- Again, this means we should expect: $\hat{\theta}_n \xrightarrow{p} \theta^*$ as $n \to \infty$
- Under correct specification, this guarantees consistency to true value θ_0
- In contrast to OLS, inference is different in the correctly specified case versus the misspecified case (larger variance due to misspecification)
- Inference depends on whether $\mathbb{E}[m(X) - f(X;\theta^*)] = 0$ holds or not; i.e., based on *model assumptions*
- We explore effects on inference next in a more standard context

Asymptotic Normality

- We have focused on identification of and estimator consistency to the pseudo-true θ^*
- Asymptotic normality holds under similar caveats and additional regularity conditions as those used to prove consistency
- That is, we now have: $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, V)$
- Asymptotic variance V can potentially depend on θ^* and model assumptions
- We do not discuss general estimation of asymptotic variances, but similar arguments could be used to show consistency of these estimators

Lecture outline

Motivation

- Likelihood-based analysis requires a full specification of the data-generating process (DGP) for data $y = (y_1, \ldots, y_n)$ based on parameters θ
- The true DGP is given by $h(y)$
- A *model* formally corresponds to $\mathcal{F} = \{f(y \mid \theta) : \theta \in \Theta \subset \mathbb{R}^p\}$, a family of distributions indexed by p-dimensional parameters θ
- We will assume dimension of parameters is *small* ($p \ll n$)
- *Correct specification* of the likelihood: $\exists \theta_0 \in \Theta$ such that $f(y \mid \theta_0) = h(y)$
- Intuitively: chosen model captures all features of the data, h should be contained in specified family $\mathcal{F}$

Sample Problem

- Assume data form an independent random sample (cross-sectional sampling)
- Likelihood function: joint distribution as a function of parameters
  $\mathcal{L}(\theta \mid y_1, \ldots, y_n) = f(y_1, \ldots, y_n \mid \theta)$

Sample Problem

- Assume data form an independent random sample (cross-sectional sampling)
- Likelihood function: joint distribution as a function of parameters
  $\mathcal{L}(\theta \mid y_1, \ldots, y_n) = f(y_1, \ldots, y_n \mid \theta)$
- Joint likelihood under independence:

$$\mathcal{L}(\theta \mid y) = \prod_{i=1}^{n} f(y_i \mid \theta)$$

- Focus on log-likelihood function (simple monotone transformation)

$$\ell_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \log f(y_i \mid \theta)$$

Sample Problem

- Idea of MLE: choose θ to maximize likelihood of replicating original data y
- Recall $\max h(x) = \min -h(x)$ for any scalar-valued function h. We turn the maximization into a minimization problem
- Final objective includes an additional term unrelated to θ:

$$D_n(\theta) \coloneqq -\frac{1}{n}\sum_{i=1}^{n} \log f(y_i \mid \theta) + \frac{1}{n}\sum_{i=1}^{n} \log g(y_i)$$

- Finally define the *maximum likelihood estimator*:

$$\hat{\theta}_n \coloneqq \arg\min_{\theta \in \Theta} D_n(\theta)$$

- Estimator unaffected by scaling: $\hat{\theta}_n = \arg\max_{\theta \in \Theta} \ell_n(\theta)$
- Requires numerical optimization of the likelihood in most cases

Digression: Two important identities

- Before we define the population problem, we provide two important identities
- Score: $s_n(\theta) \coloneqq d\ell_n(\theta)/d\theta$ and Hessian: $A_n(\theta) \coloneqq d^2\ell_n(\theta)/d\theta d\theta^\top$
- Notation for expectation under a distribution: $\mathbb{E}_g[h(y)] = \int h(y)g(y)dy$
- Given any density integrates to 1, we have: $\int f(y \mid \theta)dy = 1$
- Taking derivatives of this identity yields

$$\mathbb{E}_f\left[\frac{d\ell_n(\theta)}{d\theta}\right] = 0 \qquad \text{(Zero expected score)}$$

$$\text{Var}_f\left[\frac{d\ell_n(\theta)}{d\theta}\right] = -\mathbb{E}_f\left[\frac{d^2\ell_n(\theta)}{d\theta d\theta^\top}\right] \qquad \text{(Information Equality)}$$

Digression: Two important identities

- Before we define the population problem, we provide two important identities
- Score: $s_n(\theta) \coloneqq d\ell_n(\theta)/d\theta$ and Hessian: $A_n(\theta) \coloneqq d^2\ell_n(\theta)/d\theta d\theta^\top$
- Notation for expectation under a distribution: $\mathbb{E}_g[h(y)] = \int h(y)g(y)dy$
- Given any density integrates to 1, we have: $\int f(y \mid \theta)dy = 1$
- Taking derivatives of this identity yields

$$\mathbb{E}_f\left[\frac{d\ell_n(\theta)}{d\theta}\right] = 0 \qquad \text{(Zero expected score)}$$

$$\text{Var}_f\left[\frac{d\ell_n(\theta)}{d\theta}\right] = -\mathbb{E}_f\left[\frac{d^2\ell_n(\theta)}{d\theta d\theta^\top}\right] \qquad \text{(Information Equality)}$$

- First identity looks suspiciously like a moment condition...
- Second identity: both side are expected Fisher information

Digression: Two important identities

- Second identity: Variance of score = outer product of score = $-$Hessian

$$\text{Var}_f\left[\frac{d\ell_n(\theta_0)}{d\theta}\right] = \mathbb{E}_f\left[\left(\frac{d\ell_n(\theta_0)}{d\theta}\right)\left(\frac{d\ell_n(\theta_0)}{d\theta}\right)^\top\right] = -\mathbb{E}_f\left[\frac{d^2\ell_n(\theta_0)}{d\theta d\theta^\top}\right]$$

Digression: Two important identities

- Second identity: Variance of score = outer product of score = −Hessian

$$\text{Var}_f\left[\frac{d\ell_n(\theta_0)}{d\theta}\right] = \mathbb{E}_f\left[\left(\frac{d\ell_n(\theta_0)}{d\theta}\right)\left(\frac{d\ell_n(\theta_0)}{d\theta}\right)^\top\right] = -\,\mathbb{E}_f\left[\frac{d^2\ell_n(\theta_0)}{d\theta d\theta^\top}\right]$$

- However, *identities do not hold under misspecification*; i.e., when

$$f(y \mid \theta) \neq g(y) \text{ for all } \theta \in \Theta$$

- Outer product of score $\neq$ −Hessian (under misspecification)
- We can only recover an analog of the first identity:

$$\mathbb{E}_g\left[\frac{d\ell_n(\hat{\theta}_n)}{d\theta}\right] = 0$$

Population Problem

- Using the LLN, motivates the following population objective

$$D_0(\theta) \coloneqq -\mathbb{E}_g[\log f(Y_i \mid \theta)] + \mathbb{E}_g[\log g(Y_i)]$$

- Re-write this expression slightly as

$$D_0(\theta) \coloneqq \mathbb{E}_g\{\log[g(Y_i)/f(Y_i \mid \theta)]\}$$

- This expression is the *Kullback-Leibler divergence* between distributions f and g, a measure of our ignorance about the true structure
- Define the minimizer of the population criterion:

$$\theta^* \coloneqq \arg\min_{\theta \in \Theta} D_0(\theta)$$

- Maximum likelihood minimizes our ignorance with respect to the truth

Identification

- The population problem provides the *best distributional approximation* to the true distribution in the Kullback-Leibler sense
- $f(y \mid \hat{\theta}_n)$ is the *best distributional predictor* of $g(y)$ in class $\mathcal{F}$

Identification

- The population problem provides the *best distributional approximation* to the true distribution in the Kullback-Leibler sense
- $f(y \mid \hat{\theta}_n)$ is the *best distributional predictor* of $g(y)$ in class $\mathcal{F}$
- All identifying information comes from the likelihood and not g: $dD_0(\theta)/d\theta = -\mathbb{E}_g[d \log f(Y_i \mid \theta)/d\theta]$
- θ^* is a pseudo-true parameter as $\theta^* = \theta_0$ under correct specification:

$$\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_g\{\log f(Y_i \mid \theta_0) - \log f(Y_i \mid \theta)\} = \theta_0$$

- Need to assume global (local) identification to guarantee uniqueness of θ^* in the general case

Consistency

- We constructed the problem such that: $D_n(\theta) \xrightarrow{p} D_0(\theta)$ for all $\theta \in \theta$
- As before in non-linear GMM, this is not sufficient to guarantee the minimizer of the sample converges to the minimizer of the population
- Thus, we again introduce assumptions to strengthen pointwise convergence to uniform convergence:

$$\sup_{\theta \in \Theta} |D_n(\theta) - D_0(\theta)| \xrightarrow{p} 0$$

- This guarantees the required consistency: $\hat{\theta}_n \xrightarrow{p} \theta^*$
- Under correct specification, this is the standard consistency result to θ_0

Asymptotic Normality

- Impose regularity conditions on the likelihood and its derivatives
- Mean-value expansion of score function $s_n(\hat{\theta}_n)$ around θ^*:

$$0 = s_n(\hat{\theta}_n) = s_n(\theta^*) + A_n(\bar{\theta})(\hat{\theta}_n - \theta^*)$$

- $\bar{\theta}$ is some value of the parameters between $\hat{\theta}_n$ and θ^*
- Decomposition in terms of $\sqrt{n}(\hat{\theta}_n - \theta^*)$:

$$\sqrt{n}(\hat{\theta}_n - \theta^*) = [-A_n(\bar{\theta})]^{-1} \frac{1}{\sqrt{n}} s_n(\theta^*)$$

- Similar decomposition into two terms: one that converges in probability, the other converges in distribution to a normal

Asymptotic Normality

- By the LLN and consistency: $A_n(\bar{\theta}) \xrightarrow{p} \mathbb{E}_g[d^2 \log f(Y_i \mid \theta^*)/d\theta d\theta^\top] := A_0$
- By an appropriate central limit theorem (CLT): $\frac{1}{\sqrt{n}} s_n(\theta^*) \xrightarrow{d} \mathcal{N}(0, V_0)$
- Putting it together using Slutsky's theorem, this means

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, A_0^{-1} V_0 A_0^{-1})$$

Asymptotic Normality

- By the LLN and consistency: $A_n(\bar{\theta}) \xrightarrow{p} \mathbb{E}_g[d^2 \log f(Y_i \mid \theta^*)/d\theta d\theta^\top] := A_0$
- By an appropriate central limit theorem (CLT): $\frac{1}{\sqrt{n}} s_n(\theta^*) \xrightarrow{d} \mathcal{N}(0, V_0)$
- Putting it together using Slutsky's theorem, this means

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, A_0^{-1} V_0 A_0^{-1})$$

- Variance of asymptotic normal is the variance of the score:

$$V_0 := \mathrm{Var}_g \left[\frac{d \log f(Y_i \mid \theta^*)}{d\theta} \right] = \mathbb{E}_g \left[\left(\frac{d \log f(Y_i \mid \theta^*)}{d\theta} \right) \left(\frac{d \log f(Y_i \mid \theta^*)}{d\theta} \right)^\top \right]$$

Variance Estimation

- Having established asymptotic normality of the maximum likelihood estimator $\hat{\theta}_n$, we visit the estimation of its asymptotic variance
- Asymptotic variance: $A_0^{-1} V_0 A_0^{-1}$
- Estimator: $A_n(\hat{\theta}_n)^{-1} V_n(\hat{\theta}_n) A_n(\hat{\theta}_n)^{-1}$
- $A_n(\hat{\theta}_n)$ is the sample Hessian at the MLE estimate
- Second term is sample outer product of score (with appropriate scaling)

$$V_n(\hat{\theta}_n) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial \log f(y_i \mid \hat{\theta}_n)}{\partial \theta} \right) \cdot \left(\frac{\partial \log f(y_i \mid \hat{\theta}_n)}{\partial \theta} \right)^{\top}$$

Variance Estimation: Correct Specification

- In asymptotic variance, V_0 is "sandwiched" between two copies of $-A_0^{-1}$
- Under correct specification, we can rely on the information equality identity
- $V_0 = -A_0$ and both equal Fisher information matrix
- Asymptotic variance-covariance matrix reduces to $-A_0^{-1}$
- Both $-A_n(\hat{\theta}_n)^{-1}$ and $V_n(\hat{\theta}_n)$ are valid (consistent) estimators of asymptotic variance, though with different small-sample performance
- In the misspecified world, we need to compute both
- White (1982) proposed a misspecification test based on $H_0 : V_0 + A_0 = 0$