# Likelihood with an incorrect model

Patrick Breheny

December 8, 2025

# Introduction

- In this final week of class, we're going to look at what happens to the likelihood when you fit a model, but the true data-generating mechanism doesn't match that model (this is known as *model misspecification*)
- Today, we look at this from a theoretical perspective: what happens when the likelihood is not correct?
- Our final class will focus on using these insights for practical purposes and developing methods that work even when the model is incorrectly specified or not fully specified

# All models are wrong. . .

- In the middle part of this course, we showed that the MLE has many attractive properties: it is consistent, asymptotically normal, and efficient
- All of those statements, however, are based on the assumption that the true distribution of the data lies within the family of probability distributions parameterized by our model
- In the real world, this is almost certainly never going to be the case
- So, what happens when we're wrong? How sensitive is the likelihood to our model being correct?

# Terminology and notation

For this lecture (and the next), we are going to make a distinction between two probability distributions:

- **True distribution:** Also known as the data generation mechanism; we will denote this distribution $P_*$ and use $\mathbb{E}_*$ to denote expectations taken relative to this distribution:

  $$\mathbb{E}_* g(X) = \int g(x)\, dP_*(x);$$

  all expectations will be taken relative to this true distribution
- **Model distribution:** This is what we're assuming when we calculate the log-likelihood $\ell$, score $\mathbf{u}$, and information $\boldsymbol{\mathcal{I}}$

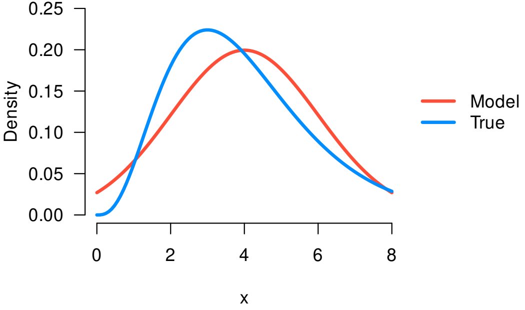
# Terminology and notation (cont'd)

In other words, our notation mostly stays the same, but:

- There is no $\boldsymbol{\theta}^*$ anymore — the true distribution of the data may have a completely different structure
- There is no $\mathscr{I}$ anymore either — the variance of the score still exists, but we no longer have $\mathbb{V}\mathbf{u} = \mathbb{E}\boldsymbol{\mathcal{I}}$, so I will avoid calling anything the Fisher information because its interpretation is unclear

# Example

- For example, suppose that $X \sim \mathrm{Gamma}(4, 1)$, but we assume a normal distribution
- Clearly, the MLE will still converge, and in fact, converge to a distribution with the correct mean and variance, but obviously not to a gamma distribution:



# Population log-likelihood

- Can we define the target that $\hat{\boldsymbol{\theta}}$ converges to?
- Let's begin by noting that by the LLN,

$$\frac{1}{n}\sum \ell(\boldsymbol{\theta} \,|\, x_i) \xrightarrow{\text{P}} \mathbb{E}_*\ell(\boldsymbol{\theta} \,|\, X)$$

- The quantity on the right is known as the *expected log-likelihood*, or *population log-likelihood*; note that it is determined by *both* the true distribution and the assumed model

# Argmax theorem

- We can then define $\boldsymbol{\theta}_*$ as the parameter that maximizes the population log-likelihood:

$$\boldsymbol{\theta}_* = \arg\max \mathbb{E}_* \ell(\boldsymbol{\theta} \,|\, X)$$

- **Argmax theorem:** Suppose $\mathbb{E}_* \ell(\boldsymbol{\theta} \,|\, X)$ exists for all $\boldsymbol{\theta}$ in a compact parameter space $\boldsymbol{\Theta}$. If
    1. $\frac{1}{n} \sum \ell(\boldsymbol{\theta} \,|\, x_i) \xrightarrow{\text{P}} \mathbb{E}_* \ell(\boldsymbol{\theta} \,|\, X)$ uniformly for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,
    2. $\mathbb{E}_* \ell(\boldsymbol{\theta} \,|\, X)$ is continuous and has a unique maximizer $\boldsymbol{\theta}_*$,

  then

$$\hat{\boldsymbol{\theta}} \xrightarrow{\text{P}} \boldsymbol{\theta}_*.$$

# Technical remarks

- In order for the convergence to hold uniformly, two conditions must hold:
    - The likelihood $\ell(\boldsymbol{\theta} \,|\, x)$ must be a continuous function of $\boldsymbol{\theta}$ for each x
    - $|\ell(\boldsymbol{\theta} \,|\, x)| \leq h(x)$ for all $x$ and $\boldsymbol{\theta}$, where $h$ is integrable with respect to $P_*$
- The argmax theorem is stated for compact parameter spaces, but in practice we only need the estimator to fall in a compact region with high probability

# Connection to Kullback-Liebler

- Note that maximizing $\mathbb{E}_*\ell(\boldsymbol{\theta}\,|\,X)$ is the same thing as minimizing
$$\mathbb{E}_* \log p_*(X) - \mathbb{E}_* \log p_{\boldsymbol{\theta}}(X)$$
- In other words, maximizing the likelihood is equivalent to finding the model that is closest to the true distribution in the sense of Kullback-Liebler "distance"
- This distribution is known as the *KL projection* onto the model space

# . . . but some are useful?

- Is the KL projection useful, scientifically?
- Maybe? For example, in the Gamma case, we still get consistent estimates of the mean and variance
- Maybe not? But we don't get consistent estimates of, say, the skewness or kurtosis
- Of course, this doesn't mean that we still have an *efficient* estimate (in the Gamma example, our estimate of the mean is 12% less efficient if we use a normal model instead of the true Gamma model)

# Distribution of the MLE under misspecification

- We've now seen how model misspecification affects the consistency of the MLE; what about its asymptotic normality?
- **Theorem:** Let $x_i \overset{\text{iid}}{\sim} P_*$ and $\boldsymbol{\theta}_*$ denote the unique maximizer of the population log-likelihood. If
  1. The likelihood of the assumed model is continuously differentiable up to third order with bounded third derivatives in a neighborhood of $\boldsymbol{\theta}_*$
  2. $\mathbb{E}_*\nabla^2\ell(\boldsymbol{\theta}_*)$ exists and is nonsingular,
  3. The conditions of the argmax theorem are met,

  then the MLE $\hat{\boldsymbol{\theta}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*) \xrightarrow{\text{d}} \text{N}(\mathbf{0}, \mathbf{B}^{-1}\mathbf{M}\mathbf{B}^{-1}),$$

  where $\mathbf{B} = -\mathbb{E}_*\nabla^2\ell(\boldsymbol{\theta}_* \,|\, X)$ and $\mathbf{M} = \mathbb{V}_*\mathbf{u}(\boldsymbol{\theta}_* \,|\, X)$

# Using this result for inference

- This result is not merely of theoretical interest — it suggests an alternative way to estimate standard errors for maximum likelihood
- Note that if the model is correctly specified, then $\mathbf{B} = \mathbf{M} = \boldsymbol{\mathscr{I}}$, and we have the usual MLE result
- However, provided that we can estimate $\mathbf{B}$ and $\mathbf{M}$, we now have an alternative: to use $\mathbf{B}^{-1}\mathbf{M}\mathbf{B}^{-1}$ instead of the information when carrying out tests and constructing confidence intervals

# The sandwich estimator

- The quantity $\mathbf{B}^{-1}\mathbf{M}\mathbf{B}^{-1}$ goes by many different names:
  - Huber–White estimator
  - Robust variance
  - Empirical variance
  - Godambe information
- But nowadays is usually referred to as the *sandwich estimator* (the "meat" $\mathbf{M}$ is sandwiched between two slices of "bread" $\mathbf{B}$)
- *Note: The notation $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ is more common in theoretical works*

# Methods that utilize sandwich estimators

- Robust (sandwich) estimators of the variance are formed in a wide variety of situations where it is desirable for the standard errors to be protected against certain forms of model misspecification:
    - Generalized estimating equations
    - Robust regression
    - Pseudo-likelihood
    - Composite likelihood
    - Survey-weighted estimators
    - Generalized method of moments
- We will explore some of these methods on Wednesday

# Conceptual understanding of the sandwich

- Conceptually, the bread and meat are measuring different aspects of the information
- The bread measures the sensitivity of the score to the parameters:
$$\mathbf{B} = -\mathbb{E}_* \partial \mathbf{u} / \partial \boldsymbol{\theta}$$
- The meat measures the randomness of the score:
$$\mathbf{M} = \mathbb{E}_* \{\mathbf{u}(\boldsymbol{\theta} \mid X) \mathbf{u}(\boldsymbol{\theta} \mid X)^\top\}$$
- Increased sensitivity causes the variance of our estimate to decrease, while increased randomness causes it to increase:
$$\mathbb{V}\hat{\boldsymbol{\theta}} = \mathbf{B}^{-1} \mathbf{M} \mathbf{B}^{-1}$$

# The bread

- Estimating the bread is straightforward — the mean of the Hessian is almost always used:

$$\hat{\mathbf{B}} = -\frac{1}{n}\sum_{i=1}^{n} \nabla^2 \ell(\boldsymbol{\theta} \,|\, x_i)$$

# The meat

- Estimating the meat, on the other hand, is more challenging as it involves a variance
- If the data are independent (even if they're not iid), then we can simply use the sample variance of the score:

$$\hat{\mathbf{M}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{u}(\boldsymbol{\theta}\,|\,x_i)\mathbf{u}(\boldsymbol{\theta}\,|\,x_i)^\top$$

- However, if the data are no longer independent, some care needs to be taken (time series data must account for autocorrelation, longitudinal data must average across subjects, etc.)

# Arguments for robust inference

- Many statisticians have argued that these robust standard errors should be the default for all inference
- If the model is correct, the two estimates coincide (asymptotically), and if the model is wrong, you have a safety net if you, say, specify the variance incorrectly
- Their argument is basically: Using the information (the non-robust SE) is unnecessary and potentially harmful; robust SEs cost nothing and protect against a wide class of model misspecification

# Arguments against robust inference

- At the same time, other statisticians have criticized the widespread use of sandwich estimators
- The core critique is: If the model is wrong, inference about $\boldsymbol{\theta}_*$ may be meaningless — why should we care about asymptotically correct standard errors for a parameter that is only defined through misspecification?*
- David Freedman (2006): "It remains unclear why applied workers should care about the variance of an estimator for the wrong parameter."

# Fix the model, not the SE

- There isn't a clear right or wrong answer here — both arguments are strong — but it is important to remember that robust SEs don't fix everything
- Using the robust SE is a convenient shortcut, but it shouldn't replace model checking, influence diagnostics, and inspection of residuals
- Ultimately, fixing the underlying problems with the model will always be better that merely adjusting the standard errors

# Partially specified models

- The analysis of misspecified models is easily extended to the question of partially specified models
- In particular, the two theorems we have covered today provide the foundation for an alternative approach to modeling in which we don't specify the likelihood — we only specify the score equations
- This idea goes by a few different names in the statistical literature:
  - *Estimating equations*
  - *Quasi-likelihood*
  - "M-estimation" (because it's kind of like an MLE)

# Why estimating equations?

- Why would we want to only specify part of the model?
- Typically, because it makes inference modular: you can separate modeling the mean (which you probably care about) from modeling the variance (which you might not)

# Definition

- Specifically, given data $x_1, \ldots, x_n$, suppose we intend to estimate parameters $\boldsymbol{\theta}$ by solving the estimating equation

$$\sum_i \boldsymbol{\psi}(\boldsymbol{\theta} \mid y_i) = \mathbf{0},$$

  where $\boldsymbol{\psi} : \mathbb{R}^d \to \mathbb{R}^d$ is a known function
- Perhaps $\boldsymbol{\psi}$ is the score function of some likelihood, but we are not bothering to specify that likelihood

# Modifications for estimating equations

The concepts and proofs are almost exactly the same as what we've just seen, with the following minor changes:

- The "true" parameter $\boldsymbol{\theta}_*$ is now defined as the solution to

$$\mathbb{E}\boldsymbol{\psi}(\boldsymbol{\theta} \,|\, X) = \mathbf{0}$$

- The bread and meat of the sandwich estimator are now
  - $\mathbf{B} = -\mathbb{E}\nabla\boldsymbol{\psi}(\boldsymbol{\theta} \,|\, X)$
  - $\mathbf{M} = \mathbb{E}\{\boldsymbol{\psi}(\boldsymbol{\theta} \,|\, X)\boldsymbol{\psi}(\boldsymbol{\theta} \,|\, X)^\top\}$

# Modifications for estimating equations (cont'd)

- Given these changes and corresponding modifications to the regularity conditions, we again have $\hat{\boldsymbol{\theta}} \xrightarrow{\mathrm{P}} \boldsymbol{\theta}_*$ and

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, \mathbf{B}^{-1}\mathbf{M}\mathbf{B}^{-1})$$

- We have seen today that sandwich estimators provide the correct variance, even when the model is misspecified
- Next lecture, we will see examples of estimating equations, quasi-likelihood, pseudo-likelihood, and composite likelihood, all of which will utilize the theory we developed today