Prove The Hessian Matrix Of The Log Likelihood Mfunction Of The Multinomial Logit Model Is Negative Semi-definite

Prove The Hessian Matrix Of The Log Likelihood Mfunction Of The Multinomial Logit Model Is Negative Semi-definite

Introduction

The multinomial logit (MNL) model is one of the most widely used discrete choice models in econometrics, marketing, transportation, and other social sciences. Its popularity stems from its ability to model choices among multiple alternatives based on observed characteristics and individual preferences. Central to understanding and estimating the MNL model is the log likelihood function, which encapsulates the probability of observed choices given the model parameters. To ensure the validity of maximum likelihood estimation (MLE) procedures, it is crucial to analyze the properties of the log likelihood function, particularly its curvature, characterized by the Hessian matrix.

In this article, we focus on proving that the Hessian matrix of the log likelihood function of the multinomial logit model is negative semi-definite. This property underpins the concavity of the log likelihood, guaranteeing the existence of a unique maximum and enabling efficient optimization algorithms. We will explore the mathematical foundations, derive the Hessian, and establish its negative semi-definiteness through detailed reasoning.

Background on the Multinomial Logit Model

Model Specification

The multinomial logit model predicts the probability that an individual chooses a particular alternative from a set of options. Suppose there are J alternatives, labeled $j = 1, 2, ..., J$. Let $\mathbf{x}_{i,j}$ denote the observed features associated with individual i and alternative j, and $\boldsymbol{\beta}$ the vector of parameters.

The probability that individual i chooses alternative j is modeled as:
\[
P_{i,j}(\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_{i,j})}{\sum_{k=1}^J \exp(\boldsymbol{\beta}^\top

\mathbf{x}_{i,k})}
\]

The likelihood function for the entire sample (assuming independence among individuals) is:

$$
L(\boldsymbol{\beta}) = \prod_{i=1}^N \prod_{j=1}^J P_{i,j}(\boldsymbol{\beta})^{y_{i,j}}
$$

where $y_{i,j}$ is 1 if individual i chose alternative j, and 0 otherwise.

The log likelihood function simplifies to:

$$
\ell(\boldsymbol{\beta}) = \sum_{i=1}^N \sum_{j=1}^J y_{i,j} \log P_{i,j}(\boldsymbol{\beta})
$$

Deriving the Gradient and Hessian

Gradient of the Log Likelihood

The gradient (score vector) of the log likelihood with respect to $\boldsymbol{\beta}$ is:

$$
\nabla \ell(\boldsymbol{\beta}) = \sum_{i=1}^N \left(\mathbf{x}_i^\top (\mathbf{y}_i - \mathbf{p}_i) \right)
$$

where:
- $\mathbf{y}_i$ is a vector indicating the chosen alternative for individual i.
- $\mathbf{p}_i$ is a vector of predicted choice probabilities for individual i.

Expressed component-wise, the j-th component of the gradient is:

$$
\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^N \left(\mathbf{x}_{i,j} (y_{i,j} - P_{i,j}) \right)
$$

Hessian of the Log Likelihood

The Hessian matrix, the matrix of second derivatives, is derived by differentiating the gradient:

$$
\mathbf{H}(\boldsymbol{\beta}) = \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}
$$

\]

For a single individual \(i \), the contribution to the Hessian is:
\[
\mathbf{H}_i(\boldsymbol{\beta}) = - \sum_{j=1}^J P_{i,j} \mathbf{x}_{i,j} \mathbf{x}_{i,j}^\top + \sum_{j=1}^J P_{i,j} \mathbf{x}_{i,j} \left(\sum_{k=1}^J P_{i,k} \mathbf{x}_{i,k}^\top \right)
\]

This simplifies to:
\[
\mathbf{H}_i(\boldsymbol{\beta}) = - \sum_{j=1}^J P_{i,j} \left(\mathbf{x}_{i,j} - \sum_{k=1}^J P_{i,k} \mathbf{x}_{i,k} \right) \left(\mathbf{x}_{i,j} - \sum_{k=1}^J P_{i,k} \mathbf{x}_{i,k} \right)^\top
\]

Aggregating over all individuals:
\[
\mathbf{H}(\boldsymbol{\beta}) = - \sum_{i=1}^N \mathbf{X}_i^\top \left(\mathbf{\Pi}_i - \mathbf{p}_i \mathbf{p}_i^\top \right) \mathbf{X}_i
\]
where \(\mathbf{X}_i \) stacks the feature vectors for individual \(i \), and \(\mathbf{\Pi}_i \) is a diagonal matrix with \(P_{i,j} \) on the diagonal.

Establishing Negative Semi-definiteness of the Hessian

Key Mathematical Properties

To prove the Hessian's negative semi-definiteness, we need to analyze the structure of the second derivative matrix. The key observations are:
- The Hessian is a sum over individuals of matrices of the form \(- \mathbf{A}_i \), where \(\mathbf{A}_i \) is positive semi-definite.
- The matrices \(\mathbf{A}_i \) are constructed as outer products of vectors involving the predicted probabilities and feature vectors.
- The negative sign outside the sum indicates the Hessian is a sum of negative semi-definite matrices, hence itself negative semi-definite.

Detailed Proof

1. Outer Product Representation

For each individual \(i \), define the vector:
\[
\mathbf{z}_i = \left(\mathbf{x}_{i,1} - \mathbf{\bar{x}}_i, \mathbf{x}_{i,2} - \mathbf{\bar{x}}_i, ..., \mathbf{x}_{i,J} - \mathbf{\bar{x}}_i \right)

\]
where \(\mathbf{\bar{x}}_i = \sum_{k=1}^J P_{i,k} \mathbf{x}_{i,k} \).

The contribution to the Hessian becomes:
\[
\mathbf{H}_i(\boldsymbol{\beta}) = - \mathbf{Z}_i^\top \mathbf{\Lambda}_i \mathbf{Z}_i
\]
where \(\mathbf{Z}_i \) stacks the differences \(\mathbf{x}_{i,j} - \mathbf{\bar{x}}_i \), and \(\mathbf{\Lambda}_i \) is a diagonal matrix with entries \(P_{i,j} (1 - P_{i,j}) \) along the diagonal, and off-diagonal entries \(- P_{i,j} P_{i,k} \).

2. Properties of \(\mathbf{\Lambda}_i \)

Since \(P_{i,j} \in (0,1) \), the matrix \(\mathbf{\Lambda}_i \) is a covariance-like matrix with non-negative eigenvalues, thus positive semi-definite.

3. Summation over Individuals

The overall Hessian:
\[
\mathbf{H}(\boldsymbol{\beta}) = - \sum_{i=1}^N \mathbf{Z}_i^\top \mathbf{\Lambda}_i \mathbf{Z}_i
\]
is a sum of negative semi-definite matrices because each \(\mathbf{Z}_i^\top \mathbf{\Lambda}_i \mathbf{Z}_i \) is positive semi-definite, and the negative sign ensures the sum is negative semi-definite.

4. Conclusion

Since the sum of negative semi-definite matrices is itself negative semi-definite, the Hessian matrix \(\mathbf{H}(\boldsymbol{\beta}) \) of the log likelihood function of the multinomial logit model is negative semi-definite.

Implications for Optimization and Estimation

The negative semi-definiteness of the Hessian implies that the log likelihood function is concave in \(\boldsymbol{\beta} \). This property guarantees:
- The existence of a global maximum for the likelihood function.
- The suitability of convex optimization algorithms such as Newton-Raphson.
- Stability and convergence properties of the maximum likelihood estimator.

These properties are fundamental for reliable inference and robust estimation procedures in multinomial logit models.

Summary

- The log likelihood function of the multinomial

Frequently Asked Questions

What is the significance of the Hessian matrix in the context of the multinomial logit model?

The Hessian matrix represents the second derivatives of the log-likelihood function and provides information about the convexity or concavity of the function, which is crucial for understanding the stability and optimization of the model parameters.

Why is the Hessian matrix of the log-likelihood in the multinomial logit model expected to be negative semi-definite?

Because the log-likelihood function is concave with respect to the parameters, its Hessian matrix must be negative semi-definite, ensuring that any critical point corresponds to a maximum.

How can we mathematically prove that the Hessian matrix of the multinomial logit log-likelihood is negative semi-definite?

By expressing the Hessian as a negative of a Fisher information matrix, which is positive semi-definite, we demonstrate that the Hessian is negative semi-definite due to its form as a negative of a positive semi-definite matrix.

What role does the Fisher information matrix play in proving the negative semi-definiteness of the Hessian?

The Fisher information matrix is positive semi-definite; since the Hessian of the log-likelihood is its negative, this directly implies that the Hessian is negative semi-definite.

Can you explain the structure of the Hessian matrix in the multinomial logit model that leads to its negative semi-definiteness?

The Hessian is composed of sums of outer products of certain vectors related to the model’s predicted probabilities, scaled by negative coefficients, which results in a negative semi-definite matrix.

Are there any assumptions necessary for the Hessian of the multinomial logit log-likelihood to be negative semi-definite?

Yes, assumptions include correctly specified models, differentiability of the likelihood, and that the model parameters lie within the parameter space where the likelihood function is concave.

How does the negative semi-definiteness of the Hessian ensure the convexity properties needed for maximum likelihood estimation?

Negative semi-definiteness guarantees the concavity of the log-likelihood function, ensuring that the optimization process converges to a global maximum during maximum likelihood estimation.

Is the negative semi-definiteness of the Hessian unique to the multinomial logit model, or does it generalize to other models?

While this property is characteristic of the multinomial logit model due to its specific structure, similar negative semi-definiteness properties hold for other models with concave log-likelihood functions, such as logistic regression, under appropriate conditions.

Additional Resources

Prove The Hessian Matrix Of The Log Likelihood Function Of The Multinomial Logit Model Is Negative Semi-Definite

---

Introduction

The multinomial logit (MNL) model is a cornerstone of discrete choice analysis, widely applied across economics, transportation, marketing, and social sciences. Its appeal lies in its simplicity and interpretability, primarily due to the logit link function and the probabilistic framework it employs. Central to statistical inference within the MNL model is understanding the properties of the likelihood function, particularly its curvature as encapsulated by the Hessian matrix.

A fundamental property underpinning maximum likelihood estimation (MLE) in the MNL context is that the Hessian matrix of the log likelihood function is negative semi-definite. This property guarantees that the likelihood function is concave, ensuring the existence of a unique global maximum when the model is correctly specified and the data are informative.

This article delves into the mathematical underpinnings of this property, providing a thorough proof that the Hessian matrix of the log likelihood function of the multinomial logit model is negative semi-definite. We will explore the structure of the MNL model, derive the Hessian explicitly, and then rigorously establish its negative semi-definiteness. The discussion aims to offer clarity, detail, and rigor suitable for researchers, statisticians, and practitioners interested in the theoretical foundations of discrete choice modeling.

The Multinomial Logit Model: An Overview

Model Specification

Suppose we observe data for N individuals choosing among J alternatives. For individual i, the probability of selecting alternative j is modeled as:

$$
P_{ij}(\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta})}{\sum_{k=1}^J \exp(\boldsymbol{x}_{ik}^\top \boldsymbol{\beta})}
$$

where:

- $\boldsymbol{x}_{ij}$ is a K-dimensional vector of observed attributes associated with individual i and alternative j,
- $\boldsymbol{\beta}$ is a K-dimensional parameter vector to be estimated.

The likelihood for the entire sample is:

$$
L(\boldsymbol{\beta}) = \prod_{i=1}^N \prod_{j=1}^J P_{ij}(\boldsymbol{\beta})^{y_{ij}}
$$

where $y_{ij} = 1$ if individual i chooses alternative j, and 0 otherwise.

The log likelihood function is thus:

$$
\ell(\boldsymbol{\beta}) = \sum_{i=1}^N \sum_{j=1}^J y_{ij} \log P_{ij}(\boldsymbol{\beta})
$$

\]

Mathematical Derivation of the Log Likelihood's Hessian

Expressing the Log Likelihood

Substituting the expression for \(P_{ij}(\boldsymbol{\beta}) \):

\[
\ell(\boldsymbol{\beta}) = \sum_{i=1}^N \sum_{j=1}^J y_{ij} \left(
\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} - \log \left(\sum_{k=1}^J
\exp(\boldsymbol{x}_{ik}^\top \boldsymbol{\beta}) \right) \right)
\]

which simplifies to:

\[
\ell(\boldsymbol{\beta}) = \sum_{i=1}^N \left(\sum_{j=1}^J y_{ij}
\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} - \log \left(\sum_{k=1}^J
\exp(\boldsymbol{x}_{ik}^\top \boldsymbol{\beta}) \right) \right)
\]

Define the design matrix for individual \(i \):

\[
\boldsymbol{X}_i = \begin{bmatrix}
\boldsymbol{x}_{i1} & \boldsymbol{x}_{i2} & \cdots & \boldsymbol{x}_{iJ}
\end{bmatrix}
\]

and similarly, the vector of choices:

\[
\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})^\top
\]

Gradient of the Log Likelihood

The gradient with respect to \(\boldsymbol{\beta} \) is:

\[
\nabla \ell(\boldsymbol{\beta}) = \sum_{i=1}^N \left(\sum_{j=1}^J y_{ij}
\boldsymbol{x}_{ij} - \sum_{j=1}^J P_{ij}(\boldsymbol{\beta})
\boldsymbol{x}_{ij} \right)
\]

or, in matrix form:

\[

$$
\nabla \ell(\boldsymbol{\beta}) = \sum_{i=1}^N \boldsymbol{X}_i^\top (\boldsymbol{y}_i - \boldsymbol{P}_i(\boldsymbol{\beta}))
$$

where $\boldsymbol{P}_i(\boldsymbol{\beta}) = (P_{i1}(\boldsymbol{\beta}), \ldots, P_{iJ}(\boldsymbol{\beta}))^\top$.

Hessian of the Log Likelihood

The second derivative (Hessian) matrix is:

$$
\boldsymbol{H}(\boldsymbol{\beta}) = \nabla^2 \ell(\boldsymbol{\beta}) = - \sum_{i=1}^N \boldsymbol{X}_i^\top \boldsymbol{\Omega}_i(\boldsymbol{\beta}) \boldsymbol{X}_i
$$

where $\boldsymbol{\Omega}_i(\boldsymbol{\beta})$ is a $J \times J$ matrix with entries:

$$
\omega_{ij,k} = \begin{cases}
P_{ij}(\boldsymbol{\beta}) (1 - P_{ij}(\boldsymbol{\beta})), & \text{if } j = k \\
- P_{ij}(\boldsymbol{\beta}) P_{ik}(\boldsymbol{\beta}), & \text{if } j \neq k
\end{cases}
$$

which can be succinctly written as:

$$
\boldsymbol{\Omega}_i(\boldsymbol{\beta}) = \text{Diag}(\boldsymbol{P}_i(\boldsymbol{\beta})) - \boldsymbol{P}_i(\boldsymbol{\beta}) \boldsymbol{P}_i(\boldsymbol{\beta})^\top
$$

Establishing Negative Semi-Definiteness

Core Mathematical Result

The primary goal is to demonstrate that:

$$
\boldsymbol{H}(\boldsymbol{\beta}) = - \sum_{i=1}^N \boldsymbol{X}_i^\top \boldsymbol{\Omega}_i(\boldsymbol{\beta}) \boldsymbol{X}_i
$$

is negative semi-definite. Because this is a sum over individual components, it suffices to show that each summand:

$$
\boldsymbol{X}_i^\top \boldsymbol{\Omega}_i(\boldsymbol{\beta}) \boldsymbol{X}_i
$$

is positive semi-definite, thus ensuring the entire Hessian is negative semi-definite.

Properties of $\boldsymbol{\Omega}_i$

- $\boldsymbol{\Omega}_i(\boldsymbol{\beta})$ is the multinomial variance-covariance matrix of the choice probabilities for individual i. It is well-known that this matrix is positive semi-definite.

- It can be viewed as a generalized covariance matrix of the multinomial distribution with probabilities $P_{ij}(\boldsymbol{\beta})$.

Proof of Positive Semi-Definiteness of $\boldsymbol{\Omega}_i(\boldsymbol{\beta})$

For any vector $\boldsymbol{z} \in \mathbb{R}^J$:

$$
\boldsymbol{z}^\top \boldsymbol{\Omega}_i(\boldsymbol{\beta}) \boldsymbol{z} = \sum_{j=1}^J P_{ij}(\boldsymbol{\beta}) z_j^2 - \left(\sum_{j=1}^J P_{ij}(\boldsymbol{\beta}) z_j \right)^2
$$

This is essentially the variance of a random variable Z defined on the discrete distribution with probabilities $P_{ij}(\boldsymbol{\beta})$:

$$
\operatorname{Var}(Z) = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 \ge 0
$$

since variance is always non-negative. Therefore, the quadratic form is non-negative for all $\boldsymbol{z}$, confirming that:

$$
\boldsymbol{\Omega}_i(\boldsymbol{\beta}) \succeq 0
$$

i.e., $\boldsymbol{\Omega}_i(\boldsymbol{\beta})$ is positive semi-definite.

Concluding the Negative Semi-Definiteness of the Hessian

Given the above, for any vector $\boldsymbol{v}$

Prove The Hessian Matrix Of The Log Likelihood Mfunction Of The Multinomial Logit Model Is Negative Semi Definite

Prove The Hessian Matrix Of The Log Likelihood Mfunction Of The Multinomial Logit Model Is Negative Semi Definite

Related Articles

- What+is+the+total+future+value+six+years+from+now+of+$80+received+today,+if+the+interest+rate+is+6%,+assuming+monthly+compounding?
- What Is A Boundaryless Organization? What Are The Features Of A Virtual Organization And A Network Organization?
- Glycerin Flows Through A Tube That Expands From A 1.00 Cmcross-section Area At Point 1 To A 4.00 Cm Cross-section

Back to Home