Section: Module 7

Misspecified models / Logistic Regression

Kentaro Nakamura

GOV 2003

March 13th, 2026

Today's Agenda

- Asymptotics for MLE
  - Delta Method
  - Efficiency (Cramer-Rao inequality)
  - Misspecified models (Consistency / Asymptotic Normality)

- Logistic Regression
  - Latent variable interpretation / random utility model
  - Ordinal Logistic Regression
  - Multinomial Logistic Regression
  - Probit

- Extra: Theory of Generalized Linear Models (GLM)
  - Exponential Family
  - Link function

Delta Method

Theorem (Delta Method)

suppose that your estimator $\hat{\theta}_n$ satisfies the asymptotic normality

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Then, for any function g that is differentiable, we have

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, \sigma^2 \cdot [g'(\theta)]^2)$$

- **Interpretation**: if $\hat{\theta}_n$ is (approximately) normal, then any *smooth* function of it is also (approximately) normal.
  - This is how we get standard errors for nonlinear quantities of interest.
- You can make it multivariate
  - If $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and $g : \mathbb{R}^k \to \mathbb{R}$ is differentiable, then

$$\sqrt{n}\{g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})\} \xrightarrow{d} \mathcal{N}\Big(0,\ \nabla g(\boldsymbol{\theta})^\top \boldsymbol{\Sigma}\, \nabla g(\boldsymbol{\theta})\Big).$$

Example of Delta Method: Odds Ratio

- Consider the logistic regression

$$\Pr(Y = 1 \mid X) = \frac{1}{1 + \exp(-X\beta)} \Leftrightarrow \log\left(\frac{\Pr(Y = 1 \mid X)}{1 - \Pr(Y = 1 \mid X)}\right) = X\beta$$

- To explore the meaning of coefficient, consider the following

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_i \quad \text{and} \quad \log\left(\frac{p'}{1-p'}\right) = \beta_0 + \beta_1(X_i + 1)$$

- This tells us that

$$\beta_1 = \log\left(\frac{p'}{1-p'}\right) - \log\left(\frac{p}{1-p}\right) \Rightarrow \exp(\beta_1) = \left(\frac{p'}{1-p'}\right) \Big/ \left(\frac{p}{1-p}\right)$$

  which is called **odds ratio**

  - The exponentiated coefficient gives us odds ratio
  - This is a nonlinear transformation of $\hat{\beta}_1$.

- Delta method with $g(t) = \exp(t)$ and $g'(t) = \exp(t)$:

$$\mathbb{V}(\widehat{\text{OR}_1}) \approx \{\exp(\beta_1)\}^2 \mathbb{V}(\hat{\beta}_1) \quad \Rightarrow \quad \widehat{\text{SE}}(\widehat{\text{OR}_1}) \approx \widehat{\text{OR}_1}\, \widehat{\text{SE}}(\hat{\beta}_1).$$

Example of Delta Method: Predicted Probability

- Delta method can be used to estimate the predicted probability
- Now, we have the asymptotic normality

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

- Predicted probability is $\Pr(\widehat{Y_i = 1} \mid X_i) = \frac{1}{1+\exp(-X_i^\top \hat{\beta})}$. The delta method says that

$$\sqrt{n}(g(\hat{\beta}) - g(\beta)) \xrightarrow{d} \mathcal{N}(0, \nabla g(\beta)^\top \Sigma \nabla(\beta))$$

with $g(\beta) = \sigma(x^\top \beta)$

  - Note that X_i is considered fixed

- Predicted probability is easier to interpret

Proof of Delta Method

- By the first order Taylor expansion, for some θ_n^* between θ and $\hat{\theta}_n$,

$$g(\hat{\theta}_n) = g(\theta) + g'(\theta_n^*)(\hat{\theta}_n - \theta)$$

- Rearranging this, we get

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) = \frac{g'(\theta_n^*)}{g'(\theta)} g'(\theta)\sqrt{n}(\hat{\theta}_n - \theta)$$

- Then, by continuous mapping theorem, we have $g'(\theta_n^*) \xrightarrow{p} g'(\theta)$, and thus $\frac{g'(\theta_n^*)}{g'(\theta)} \xrightarrow{p} 1$.
- By using Slutsky, we get

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} \mathcal{N}\left(0, \frac{(g'(\theta))^2}{I(\theta)}\right)$$

Asymptotic Efficiency: Cramer-Rao's Inequality

Theorem (Cramér–Rao Lower Bound (CRLB))

Assume standard regularity conditions (so that differentiation and integration can be exchanged). For any ***unbiased*** *estimator* $\tilde{\theta}_n$ *of* θ_0,

$$\mathbb{V}(\tilde{\theta}_n) \geq \frac{1}{I_n(\theta_0)} = \frac{1}{n\, I_1(\theta_0)}.$$

- **Implication**: Correctly specified MLE achieves the minimum variance among all the asymptotically unbiased estimator
  - This is because under correct specification, we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}\left(0,\ I_1(\theta_0)^{-1}\right).$$

Proof of Asymptotic Efficiency (1)

- Let $\tilde{\theta}_n = \tilde{\theta}_n(\boldsymbol{Y})$ be any estimator with finite variance.
- The definition of expectation is given by

$$\mathbb{E}_\theta[\tilde{\theta}_n] = \int \tilde{\theta}_n(\boldsymbol{y})\, f_\theta(\boldsymbol{y})\, d\boldsymbol{y}.$$

- Differentiate with respect to θ, we get

$$\begin{aligned}
\frac{\partial}{\partial\theta}\mathbb{E}_\theta[\tilde{\theta}_n] &= \int \tilde{\theta}_n(\boldsymbol{y})\, \frac{\partial}{\partial\theta} f_\theta(\boldsymbol{y})\, d\boldsymbol{y} \\
&= \int \tilde{\theta}_n(\boldsymbol{y})\, f_\theta(\boldsymbol{y})\, \frac{\partial}{\partial\theta} \log f_\theta(\boldsymbol{y})\, d\boldsymbol{y} \\
&= \mathbb{E}_\theta\left[\tilde{\theta}_n\, s_n(\theta)\right].
\end{aligned}$$

- Note that the second equality is by

$$\frac{\partial}{\partial\theta} \log f_\theta(\boldsymbol{y}) = \frac{1}{f_\theta(\boldsymbol{y})} \times \frac{\partial f_\theta(\boldsymbol{y})}{\partial\theta}$$

Proof of Asymptotic Efficiency (2)

- As $\mathbb{E}_\theta[s_n(\theta)] = 0$,

$$\mathrm{Cov}_\theta(\tilde{\theta}_n, s_n(\theta)) = \mathbb{E}_\theta[\tilde{\theta}_n s_n(\theta)] - \mathbb{E}_\theta[\tilde{\theta}_n] \underbrace{\mathbb{E}_\theta[s_n(\theta)]}_{=0} = \mathbb{E}_\theta[\tilde{\theta}_n s_n(\theta)].$$

- Therefore, from the previous slide,

$$\frac{\partial}{\partial \theta} \mathbb{E}_\theta[\tilde{\theta}_n] = \mathrm{Cov}_\theta(\tilde{\theta}_n, s_n(\theta)).$$

- Applying the covariance inequality, we get

$$\mathrm{Cov}_\theta(\tilde{\theta}_n, s_n(\theta))^2 \leq \mathbb{V}_\theta(\tilde{\theta}_n)\, \mathbb{V}_\theta(s_n(\theta)).$$

Proof of Asymptotic Efficiency (3)

- Because $\mathbb{E}_\theta[s_n(\theta)] = 0$,

$$\mathbb{V}_\theta(s_n(\theta)) = \mathbb{E}_\theta[s_n(\theta)^2] = I_n(\theta).$$

- Based on the previous slide, we get

$$\left\{ \frac{\partial}{\partial\theta}\mathbb{E}_\theta[\tilde{\theta}_n] \right\}^2 \leq \mathbb{V}_\theta(\tilde{\theta}_n)\, I_n(\theta) \quad \Rightarrow \quad \mathbb{V}_\theta(\tilde{\theta}_n) \geq \frac{\left\{ \frac{\partial}{\partial\theta}\mathbb{E}_\theta[\tilde{\theta}_n] \right\}^2}{I_n(\theta)}.$$

- If $\tilde{\theta}_n$ is unbiased, $\mathbb{E}_\theta[\tilde{\theta}_n] = \theta$, so the derivative is 1 and

$$\mathbb{V}(\tilde{\theta}_n) \geq \frac{1}{I_n(\theta)}.$$

KL-divergence

Definition (KL-Divergence)

Let P and Q be probability distributions with densities (or pmfs) p and q. The **Kullback–Leibler divergence** (KL-divergence) is defined as

$$D_{\mathrm{KL}}(P \,||\, Q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx \quad \text{(replace } \int \text{ by } \sum \text{ for discrete } x\text{)}.$$

- $D_{\mathrm{KL}}(P \,||\, Q) \geq 0$ and equals 0 iff $P = Q$ (a.s.).
- This is why (under correct specification) maximizing expected log-likelihood identifies the true parameter.[1]

[1] This requires that whenever $q(x) = 0$, $p(x) = 0$ (which is said that P must be **absolute continuous** with respect to Q)

MLE of the misspecified model: Consistency (1)

- Everything so far assumed **correct specification**
  - I.e., there exists $\theta_0 \in \Theta$ such that f_{θ_0} is true density
- In applied work, models are often only approximations.
  - True model: $Y_i \overset{i.i.d.}{\sim} g(\cdot)$ (unknown)
  - Working (misspecified) model: $Y_i \overset{i.i.d.}{\sim} f(\cdot \mid \theta)$
- We still compute the MLE by

$$\hat{\theta}_n := \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log f(Y_i \mid \theta).$$

  and by LLN, we still have

$$\frac{1}{n} \sum_{i=1}^{n} \log f(Y_i \mid \theta) \xrightarrow{p} \mathbb{E}_g[\log f(Y_i \mid \theta)]$$

- By the same argument of consistency of MLE, we have

$$\hat{\theta}_n \xrightarrow{p} \theta_0^\star \quad \text{where} \quad \theta_0^\star := \arg\max_{\theta \in \Theta} \mathbb{E}_g\left[\log f(Y_i \mid \theta)\right]$$

MLE of the misspecified model: Consistency (2)

- Now, notice that

$$\begin{aligned}
\theta_0^\star &:= \arg\max_{\theta\in\Theta} \ \mathbb{E}_g\left[\log f(Y_i \mid \theta)\right] \\
&= \arg\min_{\theta\in\Theta} \ \left(-\mathbb{E}_g\left[\log f(Y_i \mid \theta)\right]\right) \\
&= \arg\min_{\theta\in\Theta} \ \left(\mathbb{E}_g[\log g(Y_i)] - \mathbb{E}_g\left[\log f(Y_i \mid \theta)\right]\right) \\
&= \arg\min_{\theta\in\Theta} \ \mathbb{E}_g\left[\log \frac{g(Y_i)}{f(Y_i \mid \theta)}\right]
\end{aligned}$$

- **Claim**: $\theta_0^\star$ is not the true parameter (i.e., $\theta_0^\star \neq \theta_0$), but this is the best approximation within the model class
  - Specifically, it minimizes the KL divergence

$$\theta_0^\star = \arg\min_{\theta\in\Theta} D_{\mathrm{KL}}\left(g \,||\, f(\cdot \mid \theta)\right) = \arg\min_{\theta\in\Theta} \ \mathbb{E}_g\left[\log \frac{g(Y_i)}{f(Y_i \mid \theta)}\right].$$

Asymptotic Normality of the misspecified model (1)

- What about asymptotic normality for misspecified model?
- By Taylor expansion, we have

$$S_n(\hat{\theta}_n) = S_n(\theta_0^*) + S_n'(\theta_0^\star)(\hat{\theta}_n - \theta_0^\star) + \frac{1}{2}S_n''(\theta_n^*)(\hat{\theta}_n - \theta_0^\star)^2$$

where θ_n^* is between $\hat{\theta}_n$ and $\theta_0^\star$.

- Then,

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_0^\star) &= \frac{\sqrt{n}S_n(\theta_0^\star)}{-S_n'(\theta_0^\star) - \frac{1}{2}S_n''(\theta_n^*)(\hat{\theta}_n - \theta_0^\star)} \\ &\approx -\left(S_n'(\theta_0^\star)\right)^{-1}\sqrt{n}S_n(\theta_0^\star)\end{aligned}$$

- Recall that $\frac{1}{2}S_n''(\theta_n^*)(\hat{\theta}_n - \theta_0^\star) \xrightarrow{p} 0$ by Slutsky.

Asymptotic Normality of the misspecified model (2)

- Now, by LLN,

$$-S_n'(\theta_0^\star) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f_\theta(Y_i)\Big|_{\theta=\theta_0^\star} \xrightarrow{p} \mathbb{E}_g\left[-\frac{\partial^2}{\partial\theta^2} \log f_{\theta^\star}(Y_i)\Big|_{\theta=\theta_0^\star}\right]$$

- By CLT,

$$\sqrt{n}S_n(\theta_0^\star) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}_g\left[\left(\frac{\partial}{\partial\theta^2} \log f_\theta(Y_i)\right)^2\right]\right)$$

- Therefore, the limiting distribution would be

$$\sqrt{n}(\hat{\theta}_n - \theta_0^\star) \xrightarrow{d} \mathcal{N}(0, H^{-1}\Sigma H^{-1})$$

where

$$\Sigma = \mathbb{E}_g\left[\left(\frac{\partial}{\partial\theta^2} \log f_\theta(Y_i)\right)^2\right], \qquad H = \mathbb{E}_g\left[-\frac{\partial^2}{\partial\theta^2} \log f_{\theta_0^\star}(Y_i)\right]$$

  - If $g = f$, they become equal by information equality.

Logistic Regression

- **Logistic Regression**: Suppose that $Y_i \in \{0, 1\}$. Then,

$$Y_i \sim \text{Bernoulli}(\pi(\boldsymbol{X}_i)), \quad \pi(\boldsymbol{X}_i) = \frac{\exp(\boldsymbol{X}_i^\top \beta)}{1 + \exp(\boldsymbol{X}_i^\top \beta)} = \sigma(\boldsymbol{X}_i^\top \beta)$$

  where $\sigma(x) = \frac{1}{1+\exp(-x))}$ is called **sigmoid function**.

- $\text{logit}(p) = \log(\frac{p}{1-p})$ is called **logit function**, and you can show that logit is inverse function of sigmoid (Appendix)

- This logit function is not arbitrary; this is derived naturally from theory of Generalized Linear Models (GLM)

Latent Variable Interpretation

- Logistic regression is interpreted through latent variable Y_i^*

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases}, \quad Y_i^* = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \epsilon_i$$

- Latent variable Y_i^* is interpreted as individual utility to choose choice 1

- This is conceptually important to understand logit / probit
  - For logistic regression, you assume $\epsilon_i \sim \mathrm{Logistic}$ (the density is $\exp(-\epsilon_i)/\{1+\exp(-\epsilon_i)\}^2$)
  - For probit, you assume $\epsilon_i \sim \mathcal{N}(0,1)$

MLE of Logistic Regression

- The likelihood function is

$$L(\beta) = \prod_{i=1}^{n} \pi(\boldsymbol{X}_i)^{Y_i} (1 - \pi(\boldsymbol{X}_i))^{1-Y_i}$$

  and thus the logistic regression is

$$\ell(\beta) = \sum_{i=1}^{n} \Big(Y_i \log \pi(\boldsymbol{X}_i) + (1 - Y_i) \log(1 - \pi(\boldsymbol{X}_i)) \Big)$$

  - In machine learning, $-\frac{1}{n}\ell(\beta)$ is called **binary cross entropy**

- The score function is

$$\frac{\partial}{\partial \beta} \ell(\beta) = \sum_{i=1}^{n} \boldsymbol{X}_i (Y_i - \pi(\boldsymbol{X}_i))$$

- However, $\pi(\boldsymbol{X}_i)$ is nonlinear in β, and thus we cannot solve the score equation algebraically for β

Optimization for Logistic Regression

- Instead, we use **Newton-Raphson Algorithm** to obtain β
  - We know that the solution satisfies $S_n(\beta) = \frac{\partial}{\partial \boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) = 0$.

- By first order Taylor expansion, with $\tilde{\beta} \in (\beta, \beta^{(t)})$,

$$0 = S_n(\beta) = S_n(\beta^{(t)}) + S_n'(\tilde{\beta})(\beta - \beta^{(t)})$$

- Newton-Raphson update of the parameter is given by

$$\underbrace{\beta^{(t+1)}}_{\text{New Param}} = \underbrace{\beta^{(t)}}_{\text{Current}} - \underbrace{S_n'(\beta^{(t)})^{-1}}_{(\text{Curviture})^{-1}} \underbrace{S_n(\beta^{(t)})}_{\text{gradient}}$$

Ordinal Logistic Regression (1)

- Now, consider the latent outcome model

$$Y_i^* = X_i^\top \beta + \epsilon_i, \quad \epsilon_i \sim \text{Logistic}(0, 1)$$

$$Y_i = \begin{cases} 1 & Y_i^* \leq \tau_1 \\ 2 & \tau_1 \leq Y_i^* \leq \tau_2 \\ \vdots & \\ J & Y_i^* \geq \tau_{J-1} \end{cases}$$

- Then, as the CDF of logistic distribution is $F_\epsilon(z) = \frac{1}{1+\exp(-z)}$,

$$\begin{aligned} \Pr(Y_i \leq j \mid X_i) &= \Pr(Y_i^* \leq \tau_j \mid X_i) = \Pr(X_i^\top \beta + \epsilon_i \leq \tau_j \mid X_i) \\ &= \Pr(\epsilon_i \leq \tau_j - X_i^\top \beta) \\ &= \frac{1}{1 + \exp(-\{\tau_j - X_i^\top \beta\})} \quad (\because \epsilon_i \sim \text{Logistic}(0, 1)) \end{aligned}$$

Ordinal Logistic Regression (2)

- Therefore,

$$\Pr(Y_i = j \mid X_j) = \frac{1}{1 + \exp(-\{\tau_j - X_i^\top \beta\})} - \frac{1}{1 + \exp(-\{\tau_{j-1} - X_i^\top \beta\})}$$

or equivalently,

$$\log \frac{\Pr(Y_i \leq j \mid X_i)}{\Pr(Y_i \geq j \mid X_i)} = \tau_j - X_i^\top \beta$$

- **Proportional Odds Assumption**
  - The effect of X_i is constant across thresholds

Multinomial Logistic Regression (1)

- Suppose an individual chooses among J alternatives, and the utility from alternative j is

$$U_{ij} = X_i^\top \beta_j + \epsilon_{ij}$$

and they choose the alternative with the highest utility (i.e., $Y_i = \arg\max_j U_{ij}$)

  - Notice that each category j has different coefficient β_j (i.e., relaxing proportional odds assumption)

- Now,

$$\begin{aligned}
\Pr(Y_i = j \mid X_i) &= \Pr(U_{ij} \geq U_{ik} \ \forall k \mid X_i) \\
&= \Pr(\epsilon_{ik} \leq \epsilon_{ij} + X_i^\top(\beta_j - \beta_k) \ \ \forall k \mid X_i) \\
&= \int \Pr(\epsilon_{ik} \leq \epsilon_{ij} + X_i^\top(\beta_j - \beta_k) \ \ \forall k \mid \epsilon_{ij} = \epsilon, X_i) f_{\epsilon_{ij}}(\epsilon) d\epsilon
\end{aligned}$$

where $\Pr(\epsilon_{ik} \leq \epsilon_{ij} + X_i^\top(\beta_j - \beta_k) \ \ \forall k \mid \epsilon_{ij} = \epsilon, X_i)$ is CDF of ϵ_{ik}.

Multinomial Logistic Regression (2)

- **Independence of Irrelevant Alternatives (IIA)** assumption
  - $\epsilon_{i1}, \cdots, \epsilon_{iJ}$ are independent
  - Under this assumption,

$$\frac{\Pr(Y_i = j \mid X_i)}{\Pr(Y_i = k \mid X_i)} = \frac{\exp(X_i^\top \beta_j)}{\exp(X_i^\top \beta_k)} = \exp(X_i^\top (\beta_j - \beta_k))$$

    so adding/removing other choices does not affect the ratio

- In addition, suppose that the error follows **Gumbel distribution**, which PDF and CDF are given by

$$F(\epsilon) = \exp(-\exp(-\epsilon)), \quad f(\epsilon) = \exp(-\epsilon)\exp(-\exp(-\epsilon))$$

- Using the Gumbel distribution, by some algebra (in Appendix),

$$\Pr(Y_i = j \mid X_i) = \frac{\exp(X_i^\top \beta_j)}{\sum_k \exp(X_i^\top \beta_k)}$$

  - This is called **softmax function** (i.e., $\mathrm{softmax}(z_j) = \frac{\exp(z_j)}{\sum_k \exp(z_k)}$)

Probit

- We can instead assume that errors are from multivariate normal distribution (called **probit**)

- Consider the binary case, and suppose

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases}, \quad Y_i^* = X_i^\top \beta + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1)$$

- Then,

$$\begin{aligned} \Pr(Y_i = 1 \mid X_i) &= \Pr(X_i^\top \beta + \epsilon_i > 0 \mid X_i) = \Pr(\epsilon_i > -X_i^\top \beta \mid X_i) \\ &= \Pr(\epsilon < X_i^\top \beta \mid X_i) = \Phi(X_i^\top \beta) \end{aligned}$$

  where $\Phi(\cdot)$ is the CDF of normal distribution
  - Third equality is by symmetricity of normal distribution.

Extra: Generalized Linear Models (GLM)

- Actually, logistic regression is one of models in Generalized Linear Model (GLM)

- **Generalized Linear Model** (GLM) is a broad class of statistical models that extends ordinary linear regression so that we can model non-Gaussian outcomes

- Key Structure: apply **link function** $g(\cdot)$ and use the linear model

$$g(\mathbb{E}[Y_i \mid X_i]) = X_i^\top \beta \quad \text{(equivalently, } \mathbb{E}[Y_i \mid X_i] = g^{-1}(X_i\beta)\text{)}$$

  - Example: $g(\mu) = \mu$ is linear regression
  - Example 2: $g(\mu) = \log \frac{\mu}{1-\mu}$ is logistic regression
  - Example 3: $g(\mu) = \log \mu$ is poission regression

- This link function is deeply related to exponential family

Exponential Family: Definition

Definition (Exponential Family)

A distribution belongs to the **exponential family** if its density can be written as

$$f(y_i \mid \theta, \phi) = \exp\left(\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i, \phi)\right)$$

where θ is called **natural parameter**, ϕ is called dispersion parameter, and $b(\theta)$ is a cumulant function.

- To show whether the distribution belongs to exponential family, you just need to transform the distribution
  - I show the example in next page

Why exponential family?

- Exponential family is important for GLM as it gives us the link function.

Lemma (Properties of GLM)

If Y follows the exponential family, then

$$\mathbb{E}_\theta[Y] = b'(\theta),$$

where $b'(\theta)$ is the derivative of cumulant function.

- As a result, we have

$$\theta = (b')^{-1}\mathbb{E}_\theta[Y]$$

  which means that for exponential family, the link function always exists (because $g(\mu) = (b')^{-1}(\mu)$ is the one)

- Such link function is called **canonical link function**

Example (1): Bernoulli distribution

- Suppose that $y_i \in \{0, 1\}$ follows Bernollui distribution, which PMF is written as

$$f(y_i) = p_i^y (1-p)^{1-y_i}$$

- Then,

$$f(y_i) = p_i^y (1-p)^{1-y_i} = \exp(y_i \log p + (1-y_i)\log(1-p))$$
$$= \exp\left(y_i \log \frac{p}{1-p} + \log(1-p)\right)$$

- This means that natural parameter is $\theta = \log \frac{p}{1-p}$
  - This means that $p = \frac{\theta}{1-\theta}$
- Therefore, PMF is written as

$$f(y_i) = \exp(y_i\theta - \log(1+\exp(\theta))), \quad \theta = \log \frac{p}{1-p}$$

and thus cumulant function is $b(\theta) = \log(1+\exp(\theta))$

- The derivative of cumulant function thus gives us the logistic link.

Example (2): Poisson distribution

- Suppose that $Y_i \sim \text{Poission}(\lambda)$; i.e.,

$$\Pr(Y_i = y) = \frac{\exp(-\lambda)\lambda^y}{y!}$$

- This gives us

$$\Pr(Y_i = y) = \exp(y \log \lambda - \lambda - \log(y!))$$

  which means that

$$\text{natural parameter} : \theta = \log \lambda$$
$$\text{cumulant function} : b(\theta) = \exp(\theta)$$

- The derivative of cumulant function is

$$b'(\theta) = \exp(\theta)$$

  and thus the canonical link function is $g(\mu) = \log \mu$.

Extra: Nested Multinomial Logit Model (1)

- One way to relax IIA assumption is to use **Nested Multinomial Logit Model**
  - Key idea: group similar alternatives into nests.
  - Example: first choose a party, then choose a candidate within that party.

- **Setup**: suppose that there are j choices, but they are nested in m categories.
  - Let $m(j)$ be the category that choice j belongs to.
  - Utility of choosing choice j for unit i is written as

$$U_{ij} = X_i^\top \beta_j + \eta_{i,m(j)} + \nu_{ij}$$

- The common component $\eta_{i,m(j)}$ induces the correlation among alternatives in the same nest.

Extra: Nested Multinomial Logit Model (2)

- Let $\lambda_m \in (0, 1]$ be the dissimilarity parameter for nest m.
  - Smaller λ_m means stronger within-nest correlation.
- The conditional probability of choosing option j within nest m is

$$\Pr(Y_i = j \mid m(j) = m, X_i) = \frac{\exp(V_{ij}/\lambda_m)}{\sum_{k \in B_m} \exp(V_{ik}/\lambda_m)}.$$

- The probability of choosing nest m is

$$\Pr(m \mid X_i) = \frac{\left\{\sum_{k \in B_m} \exp(V_{ik}/\lambda_m)\right\}^{\lambda_m}}{\sum_{r=1}^{M} \left\{\sum_{k \in B_r} \exp(V_{ik}/\lambda_r)\right\}^{\lambda_r}}.$$

- Therefore,

$$\Pr(Y_i = j \mid X_i) = \Pr(Y_i = j \mid m(j), X_i) \Pr(m(j) \mid X_i).$$

  - If all $\lambda_m = 1$, the model reduces to standard multinomial logit.

Appendix: Logit and Sigmoid

- First, compute $\mathrm{logit}(\sigma(x))$. Notice that

$$\frac{\sigma(x)}{1-\sigma(x)} = \frac{\frac{1}{1+\exp(-x)}}{\frac{\exp(-x)}{1+\exp(-x)}} = \exp(x)$$

and thus

$$\mathrm{logit}(\sigma(x)) = \log\left(\frac{\sigma(x)}{1-\sigma(x)}\right) = \log\exp(x) = x$$

- Second, compute $\sigma(\mathrm{logit}(x))$. Now, notice that

$$\exp\left(-\log\left(\frac{x}{1-x}\right)\right) = \frac{1-x}{x}$$

and thus

$$\sigma(\mathrm{logit}(x)) = \frac{1}{1+\exp\left(-\log\left(\frac{x}{1-x}\right)\right)} = \frac{1}{1+\frac{1-x}{x}} = x$$

Appendix: Proof of Properties of GLM (1)

- Now,

$$1 = \int f(y_i \mid \theta, \phi) dy = \int \exp\left(\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i, \phi) \right) dy$$

$$\Rightarrow 0 = \int \frac{\partial}{\partial\theta} \exp\left(\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i, \phi) \right) dy$$

- Then,

$$\frac{\partial}{\partial\theta} \log f(y_i \mid \theta, \phi) = \frac{1}{a(\phi)}(y - b'(\theta))$$

and thus

$$\frac{\partial}{\partial\theta} f(y_i \mid \theta, \phi) = f(y_i \mid \theta, \phi) \cdot \frac{1}{a(\phi)}(y - b'(\theta))$$

Appendix: Proof of Properties of GLM (2)

- Therefore,

$$
\begin{aligned}
0 &= \int f(y_i \mid \theta, \phi) \cdot \frac{1}{a(\phi)}(y - b'(\theta)) \\
&= \frac{1}{a(\phi)} \int y f(y_i \mid \theta, \phi) dy - b'(\theta) \int f(y_i \mid \theta, \phi) dy
\end{aligned}
$$

- This means that

$$
\begin{aligned}
0 &= \frac{1}{a(\phi)}(\mathbb{E}_\theta[y_i] - b'(\theta) \cdot 1) \\
&\Rightarrow \mathbb{E}_\theta[y_i] = b'(\theta)
\end{aligned}
$$

Appendix: Deriving Multionominal Logit (1)

- Let $V_{ij} = X_i^\top \beta_j$ and suppose the errors are i.i.d. Gumbel. Then

$$\Pr(Y_i = j \mid X_i) = \int \prod_{k \neq j} F(\epsilon + V_{ij} - V_{ik})\, f(\epsilon)\, d\epsilon,$$

where

$$F(u) = \exp\{-\exp(-u)\}, \qquad f(u) = \exp(-u)\exp\{-\exp(-u)\}.$$

- Plugging in F and f,

$$\Pr(Y_i = j \mid X_i) = \int \exp(-\epsilon)\exp\left[-\exp(-\epsilon)\sum_{k=1}^{J}\exp(V_{ik} - V_{ij})\right] d\epsilon.$$

Appendix: Deriving Multionominal Logit (2)

- Use the substitution $t = \exp(-\epsilon)$, so $dt = -\exp(-\epsilon)\, d\epsilon$. Then

$$\begin{aligned}\Pr(Y_i = j \mid X_i) &= \int_0^{\infty} \exp\left[-t \sum_{k=1}^{J} \exp(V_{ik} - V_{ij})\right] dt \\ &= \frac{1}{\sum_{k=1}^{J} \exp(V_{ik} - V_{ij})}\end{aligned}$$

- Therefore,

$$\Pr(Y_i = j \mid X_i) = \frac{\exp(V_{ij})}{\sum_{k=1}^{J} \exp(V_{ik})}.$$