MLE Under Missepcification

Kullback-Leibler (KL) Divergence and MLE Properties #flashcard

- Denote the true density as $p_0(y)$ and the density under our model specification as $f(y;\theta)$
- The **Kullback-Leibler (KL) Divergence** measures the difference between the true expected log-likelihood and the expected log-likelihood under misspecification:

$$\begin{aligned} KL(p_0; f_\theta) &= \mathbb{E}\left[\log\left(\frac{p_0(y)}{f(y;\theta)}\right)\right] \\ &= \underbrace{\mathbb{E}\left[\log p_0(y)\right]}_{\text{Ture exp log likelihood, const wrt } \theta} - \underbrace{\mathbb{E}\left[\log f(y;\theta)\right]}_{\text{Exp log likelihood under our model specification}} \end{aligned}$$

- Therefore:

$$\text{Minimise KL Divergence} \iff \text{Maximise Model Likelihood}$$

and

$$\hat{\theta}_{MLE} \rightarrow^p \begin{cases} \theta_0 = \theta & \text{(true parameter)} & \text{under correc specification:} f_0 = p_0 \\ \theta_0 & \text{(parameter minmises KL)} & \text{under misspecification} \end{cases}$$

- Note that KL Divergence:
  - is not symmetric $KL(p_0; f_\theta) \neq KL(f_\theta; o_0)$
  - ≥ 0 and is 0 iff $f_\theta = p_0$

(Unconditional) MLE Consistency and Asymptotic Distribution under Misspecification #flashcard

- Suppose that $y_1, \dots, y_N \sim^{iid} p_0$ and our (possibly) misspecified model has density $f(y;\theta)$. Then, under mild regularity conditions:
- $\hat{\theta}_{MLE}$ consistently estimates the **pseudo-true** value θ_0, defined as the minimiser of KL divergence over the parameter space :

$$\hat{\theta}_{MLE} \rightarrow^p \theta_0 = arg\min_\theta KL(p_0, f_\theta)$$

- Asymptotic Distribution:

$$\sqrt{N}\left(\hat{\theta}_{MLE} - \theta\right) \rightarrow^d N\left(0, J^{-1}KJ^{-1}\right)$$

where

$$\begin{cases} J & = -\mathbb{E}\left[\frac{\partial^2 \log f(y_i;\theta_0)}{\partial\theta\partial\theta^T}\right] & \text{essian Matrix} \\ K & = Var\left[\frac{\partial \log f(y_i;\theta_0)}{\partial\theta}\right] & \text{Var(Score)} \end{cases}$$

- Sample Analogues:

$$\begin{aligned} \hat{\theta}_{MLE} &\sim^a N\left(\theta_0, \frac{\hat{J}^{-1}\hat{K}\hat{J}^{-1}}{N}\right) \\ &\sim^a N\left(\theta_0, \left(\sum_{i=1}^N \frac{\partial^2 \log f(y_i;\hat{\theta})}{\partial\theta\partial\theta^T}\right)^{-1} \left\{\sum_{i=1}^N \left[\frac{\partial \log f(y_i;\hat{\theta})}{\partial\theta}\right]\left[\frac{\partial \log f(y_i;\hat{\theta})}{\partial\theta}\right]^T\right\} \left(\sum_{i=1}^N \frac{\partial^2 \log f(y_i;\hat{\theta})}{\partial\theta\partial\theta^T}\right)^{-1}\right) \end{aligned}$$

where $\hat{J}^{-1}\hat{K}\hat{J}^{-1}$ is called the **robust asymptotic variance matrix estimator** and

$$\begin{cases} \hat{J} &= -\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 \log f(y_i;\theta_0)}{\partial\theta\partial\theta^T} \\ \hat{K} &= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial \log f(y_i;\theta_0)}{\partial\theta}\right]\left[\frac{\partial \log f(y_i;\theta_0)}{\partial\theta}\right]^T \end{cases}$$

(Unconditional) MLE Consistency and Asymptotic Distribution under Correct Specification #flashcard

- Suppose that $y_1, \dots, y_N \sim^{iid} f(y;\theta) = p_0$ (correct specification). Then, under mild regularity conditions:
- $\hat{\theta}_{MLE}$ consistently estimates the true model parameter θ_0:

$$\hat{\theta}_{MLE} \to^p \theta_0 \text{ s.t. } f(y;\theta_0) = p_0$$

- Asymptotic distribution:

$$\sqrt{N}\left(\hat{\theta}_{MLE} - \theta\right) \sim^a N\left(0, J^{-1}\right)$$

- Information Matrix Equality:

$$J = K \iff J^{-1}KJ^{-1} = J^{-1}$$

  where:

$$\begin{cases} J &= -\mathbb{E}\left[\frac{\partial^2 \log f(y_i;\theta_0)}{\partial\theta\partial\theta^T}\right] & \text{essian Matrix} \\ K &= Var\left[\frac{\partial \log f(y_i;\theta_0)}{\partial\theta}\right] & \text{Var(Score)} \end{cases}$$

- Sample Analogues for Asymp Dist:

$$\begin{aligned} \hat{\theta}_{MLE} &\sim^a N\left(\theta_0, \frac{\hat{J}^{-1}}{N}\right) \\ &\sim^a N\left(\theta_0, -\left(\sum_{i=1}^{N}\frac{\partial^2 \log f(y_i;\hat{\theta})}{\partial\theta\partial\theta^T}\right)^{-1}\right) \end{aligned}$$

  where

$$\hat{J} = -\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 \log f(y_i;\hat{\theta})}{\partial\theta\partial\theta^T}$$

Models for Count Data: Conditional MLE and Poisson Regression

Minimum MSE Predictor and Minimum MSE Linear Predictor #flashcard

- **Minimum MSE Predictor**: the min MSE predictor is the conditional mean:

$$arg\min_{\psi(.)}\mathbb{E}\left[(y-\psi(X))^2\right] = \mu(X) \equiv \mathbb{E}\left[y|X\right]$$

- **Minimum MSE Linear Predictor**: the min MSE linear predictor is the OLS predictor:

$$arg\min_{\text{linear }\psi(.)}\mathbb{E}\left[(y-\psi(X))^2\right] = X(X^TX)^{-1}X^Ty$$

Unconditional MLE and Conditional MLE #flashcard

- **Unconditional MLE** specifies the unconditional distribution of a random vector y with a unconditional density function:

$$y \sim f(y;\theta)$$

- The population estimator is:

$$\hat{\theta}_{MLE} \equiv arg \max_{\theta \in} \mathbb{E}\left[\log f(y_i;\theta)\right]$$

- The sample analogue is:

$$\hat{\theta}_{MLE} \equiv arg \max_{\theta \in} \frac{1}{N}\sum_{i=1}^{N} \log f(y_i;\theta)$$

- **Conditional MLE** specifies the conditional distribution of a random vector y with a conditional density function:

$$y|x \sim f(y,x;\theta)$$

- This abstracts away the distribution $f(x;\theta)$ which we are not interested in.
- The population estimator is:

$$\hat{\theta}_{MLE} \equiv arg \max_{\theta \in} \mathbb{E}\left[\log f(y_i|x_i;\theta)\right]$$

- The sample analogue is:

$$\hat{\theta}_{MLE} \equiv arg \max_{\theta \in} \frac{1}{N}\sum_{i=1}^{N} \log f(y_i|x_i;\theta)$$

- Asymptotic variance is the same as unconditional MLE with $f(y;\theta)$ replaced by $f(y|x;\theta)$

Poisson Regression #flashcard

- Assumption:

$$y_i|x_i \sim \text{Poisson}(\mathbb{E}\left[y_i|x_i\right]) \text{ and } \mathbb{E}\left[y_i|x_i\right] = \exp(x_i^T\beta)$$

equivalently:

$$f(y_i|x_i) = \frac{\left(\exp(x_i^T\beta)\right)^{y_i} \exp\left(-\exp(x_i^T\beta)\right)}{y_i!}$$

- Log likelihood:

$$l_i(\beta) = \log f(y_i|x_i;\beta) = y_i x_i^T\beta - \exp(x_i^T\beta) - \log(y_i!)$$

- Consistency: if we correctly $\mathbb{E}\left[y_i|x_i\right]$ is indeed $\exp(x_i^T\beta_0)$ then $\hat{\beta}_{Possion} \rightarrow^p \beta_0$ even if the conditional distribution is not Poisson.
- Avar under different assumptions:
- Strongest - Poisson Assumption: the conditional distribution is indeed Poisson ($Var[y_i|x_i] = \mathbb{E}\left[y_i|x_i\right] = \exp(x_i^T\beta)$), then:

$$\sqrt{N}\left(\hat{\beta}_{Possion} - \beta_0\right) \sim^a N\left(0, \hat{J}^{-1}\right)$$

- Mild - Quasi-Poisson Assumption: the conditional distribution is a scaled Poisson ($Var[y_i|x_i] = \sigma^2\mathbb{E}\left[y_i|x_i\right] = \sigma^2\exp(x_i^T\beta)$), then:

$$\sqrt{N}\left(\hat{\beta}_{Possion} - \beta_0\right) \sim^a N\left(0, \hat{\sigma}^2\hat{J}^{-1}\right)$$

where $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{\hat{u}^2}{\exp(x_i^T\hat{\beta})}$
- Weakest - No assumption on $Var[y_i|x_i]$, then use the robust variance estimator:

$$\sqrt{N}\left(\hat{\beta}_{Possion} - \beta_0\right) \sim^a N\left(0, \hat{J}^{-1}\hat{K}\hat{J}^{-1}\right)$$

- Where:

$$\hat{J} = \frac{1}{N} \sum_{i=1}^{N} \exp(x_i^T \hat{\beta}) x_i x_i^T$$

and

$$\hat{K} = \frac{1}{N} \sum_{i=1}^{N} \hat{u}_i^2 x_i x_i^T$$

- Efficiency:
- If the true conditional distribution is Poisson $\implies$ Poisson MLE is efficient
- If the true conditional distribution is Quasi-Poisson $\implies$ Poisson QMLE is still more efficient than NLLS and various other models

Partial Effects and Average Partial Effects #flashcard

- **Partial Effects**: for continuous x_j, the partial effect of x_j is $\frac{\partial}{\partial x_j}\mathbb{E}\left[y|X\right]$; for discrete x_j, the partial effect is the difference in $\mathbb{E}\left[y|X\right]$ for 2 different values of x_j.
  - In general, the partial effects will be different for different values of X.
- **Average Partial Effects** is the expectation of the partial effects over the distribution of X.

Models for Binary Outcomes

Problems of Linear Probability Model in Binary Outcome Modelling #flashcard

- Out-of-bound predictions: a linear probability model can produce prediction outside $[0, 1]$
- Constant marginal effects
- Heteroskedasticity

Index Models for Binary Outcome (Logit, Probit) #flashcard

- Index Model: we use a **index function** $G(.)$ on the **linear index** $x^T\beta$ to model the outcome probability:

$$p(x) \equiv \mathbb{E}\left[y|x\right] = G(x^T\beta)$$

  where $G(.)$ has the following properties:
  - $0 \leq G(.) \leq 1$
  - differentiable and strictly increasing
  - $\lim_{z\to\infty} G(z) = 1, \lim_{z\to-\infty} G(z) = 0$
- The typical choices of $G(.)$ are CDFs:
  - **Logit**:

$$G(.) = (.) \implies p(x) \equiv \mathbb{E}\left[y|x\right] = (x^T\beta) = \frac{\exp(x^T\beta)}{1 + \exp(x^T\beta)}$$

  - **Probit**:

$$G(.) = \Phi(.) \implies \int_{-\infty}^{x^T\beta} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$$

- Partial Effects:

$$\frac{\partial \mathbb{E}\left[y|x\right]}{\partial x_j} = \frac{\partial G(x^T\beta)}{\partial x_j} = G(x^T\beta)\beta_j$$

which depends on x

- Relative Partial Effects:

$$\frac{\frac{\partial \mathbb{E}[y|x]}{\partial x_j}}{\frac{\partial \mathbb{E}[y|x]}{\partial x_k}} = \frac{G(x^T\beta)\beta_j}{G(x^T\beta)\beta_k} = \frac{\beta_j}{\beta_k}$$

does not depend on x

- Partial Effect

$$\frac{\partial}{\partial x_j} G(\mathbf{x}_i'\boldsymbol{\beta}) = g(\mathbf{x}_i'\boldsymbol{\beta})\beta_j$$

Estimated Partial Effect

$$\frac{\partial}{\partial x_j} G(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}) = g(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})\widehat{\beta}_j$$

Average Partial Effect

$$\mathbb{E}\left[\frac{\partial}{\partial x_j} G(\mathbf{x}_i'\boldsymbol{\beta})\right] = \mathbb{E}[g(\mathbf{x}_i'\boldsymbol{\beta})]\beta_j = \int g(x_i^T\beta)\, dF(x)\, \beta_j$$

Estimated Average Partial Effect

$$\left[\frac{1}{N}\sum_{i=1}^{N} g(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})\right]\widehat{\beta}_j$$

- Conditional Likelihood

$$\prod \left\{ [p(y_i)]^{\mathbb{1}\{y=y_i\}} \right\}$$

$$f(y_i|\mathbf{x}_i,\boldsymbol{\beta}) = \begin{cases} 1 - G(\mathbf{x}_i'\boldsymbol{\beta}) & \text{if } y_i = 0 \\ G(\mathbf{x}_i'\boldsymbol{\beta}) & \text{if } y_i = 1 \end{cases} \iff \underline{f(y_i|\mathbf{x}_i,\boldsymbol{\beta}) = G(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}\left[1 - G(\mathbf{x}_i'\boldsymbol{\beta})\right]^{1-}}$$

Take log

Conditional Log-Likelihood

$$\ell_i(\boldsymbol{\beta}) \equiv \log f(y_i|\mathbf{x}_i,\boldsymbol{\beta}) = \underline{y_i \log\left[G(\mathbf{x}_i'\boldsymbol{\beta})\right] + (1 - y_i)\log\left[1 - G(\mathbf{x}_i'\boldsymbol{\beta})\right]}$$

Sample

$$\widehat{\boldsymbol{\beta}} \equiv \underset{\boldsymbol{\beta}\in\boldsymbol{\Theta}}{\arg\max} \frac{1}{N}\sum_{i=1}^{N} \ell_i(\boldsymbol{\beta})$$

Population

$$\boldsymbol{\beta}_o \equiv \underset{\boldsymbol{\beta}\in\boldsymbol{\Theta}}{\arg\max}\, \mathbb{E}\left[\ell(\boldsymbol{\beta})\right]$$

$\boxed{\text{Correct specification: } \mathbb{E}(y|\mathbf{x}) = p(\mathbf{x}) = G(\mathbf{x}'\boldsymbol{\beta}_o). \text{ Otherwise } \boldsymbol{\beta}_o = \text{KL-minimizer.}}$

- Possibly Mis-specified Model

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) \to_d \mathcal{N}(\mathbf{0}, \mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1}) \text{ where } \mathbf{J} = -\mathbb{E}\left[\mathbf{H}_i(\boldsymbol{\beta}_o)\right] \text{ and } \mathbf{K} = \mathbb{E}\left[\mathbf{s}_i(\boldsymbol{\beta}_o)\mathbf{s}_i(\boldsymbol{\beta}_o)'\right]$$

Correct Specification

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) \to_d \mathcal{N}(\mathbf{0}, \mathbf{J}^{-1}) \text{ where } \mathbf{J} = -\mathbb{E}\left[\mathbf{H}_i(\boldsymbol{\beta}_o)\right]$$

-

Under Correct Specifications

Asymptotic Distribution

$$\sqrt{N}(\widehat{\beta} - \beta_o) \to_d \mathcal{N}\left(\mathbf{0}, \mathbf{J}^{-1}\right), \quad \mathbf{J}^{-1} = \mathbb{E}\left\{ \frac{g(\mathbf{x}_i'\boldsymbol{\beta_o})^2 \mathbf{x}_i \mathbf{x}_i'}{G(\mathbf{x}_i'\boldsymbol{\beta_o})\left\{1 - G(\mathbf{x}_i'\boldsymbol{\beta_o})\right\}} \right\}^{-1}$$

Consistent Estimator

$$\widehat{\mathbf{J}}^{-1} \equiv \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{g(\mathbf{x}_i'\widehat{\beta})^2 \mathbf{x}_i \mathbf{x}_i'}{G(\mathbf{x}_i'\widehat{\beta})\left[1 - G(\mathbf{x}_i'\widehat{\beta})\right]} \right\}^{-1}$$

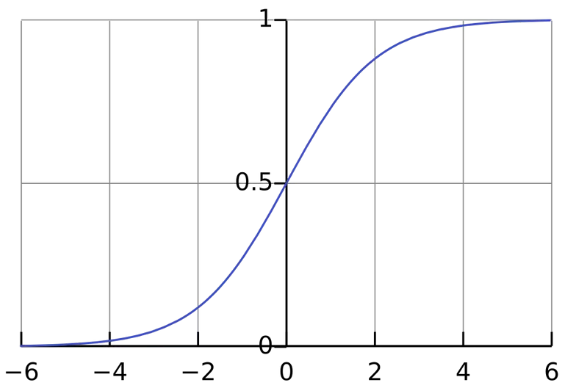
Notes

- Assumes correct specification, i.e. $p(\mathbf{x}) = \mathbb{E}(y|\mathbf{x}) = G(\mathbf{x}'\beta_o)$
- In contrast, *robust* variance matrix $\mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1}$ is complicated, but R can do it.

Logistic Function is Symmetric #flashcard

- Logistic function is symmetric around 0:

$$(-k) = 1 - (k)$$



Random Utility Models

Random Utility Model and Multinomial Logit #flashcard

- Observables:
  - x_{nj} - attributes of each alternative (e.g. product characteristics)
  - s_n - attributes of the decision-maker (e.g. demographics)
  - Individual choices (but not corresponding utilities)
- Specify:
  - a function $V_{nj}(x_{nj}, s_n)$ relating attributes x_{nj} of each alternative j and attributes s_n
  - error term $\epsilon_{nj} = U_{nj} - V_{nj}$ is the difference between true utility U_{nj} and modelled utility V_{nj} assumed to follow a random distribution
- Choice probabilities:

$$P_{ni} \equiv P(U_{ni} > U_{nj} \,\forall\, j \neq i) = P(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj} \,\forall\, v \neq i)$$

  with this we can write the joint log-likelihood and estimate using MLE

Identification in RUT #flashcard

- A parameter is **identified** if it could be *uniquely* determined by observing the whole population of data from which our sample is drawn.
- In RUT MLE estimation:
  - Only difference in utility matter for choices.
  - The scale of utility is irrelevant.
- $\implies$
  - Absolute level of utility is not identified: if there are J alternatives, we can only set the intercept of one option to be 0 and identify the rest $J-1$ intercepts:

$$V_{ij} = \alpha_j + x_{ij}^T \beta, \quad \text{set } \alpha_1 = 0$$

  - Features *invariant across options* will be jointly identified $\implies$ we will have to normalise coefficient for one base option to be 0:

$$V_{ij} = \alpha_j + x_i^T \beta_j, \quad \text{set } \beta_1 = 0$$

    - Alternatively, without normalising the base group, we can be uniquely identify features invariant across options when interacting with alternative-specific variables or dummies:

$$V_{ij} = \beta_1 \cdot \text{Cost}_{ij} + \beta_2 \cdot \text{Cost}_{ij} \cdot Income_i$$

  - Features *varying across options* can be uniquely identified without the need for normalising:

$$V_{ij} = \beta_1 \cdot \text{TravelTime}_{ij} + \beta_2 \cdot \text{Cost}_{ij}$$

    - However, if we want to have option-specific coefficients, we will have to normalise a base group since they are jointly identified:

$$V_{ij} = \alpha_j + x_{ij}^T \beta_j, \quad \text{set } \beta_1 = 0$$

- Summary

Feature Type	Can Be Identified?	Condition
Absolute utility levels	❌	Normalize 1 intercept ($\alpha_1 = 0$)
Variables invariant across alternatives	❌	Unless interacted with alt-specific dummy or normalising the base group $\beta_1 = 0$
Variables varying across alternatives	✅	Identified with generic β
Alt-specific coefficients on varying vars	✅ (jointly)	Normalise one group (e.g. $\beta_1 = 0$)

Feature Type	Can Be Identified?	Condition
Absolute utility levels	❌	Normalize 1 intercept ($\alpha_1 = 0$)
Variables invariant across alternatives	❌	Unless interacted with alt-specific dummy or normalising the base group $\beta_1 = 0$
Variables varying across alternatives	✅	Identified with generic β
Alt-specific coefficients on varying vars	✅ (jointly)	Normalise one group (e.g. $\beta_1 = 0$)

Multinomial/Conditional/Mixed Logit Model #flashcard

- Based on our Random Utility Model, if $\epsilon_{n1}, \ldots, \epsilon_{nJ} \sim^{iid}$ Gumbel / Type 1 Extreme Value distribution $\implies$ **Logit**:

$$P_{ni} = \frac{\exp(V_{ni})}{\sum_{j=1}^{J} \exp(V_{nj})}$$

- **Multinomial Logit**: we only include attributes that are fixed across alternatives in the utility:

$$V_{nj} = s_{n\,j}^T$$

  we typically set $_1 = 0$ for the base group 1 and identify the difference:$_j - _0 = _j$
- **Conditional Logit**: we only include attributes that vary across alternatives (e.g. prices):

$$V_{nj} = x_{nj}^T \beta$$

  note that the parameters β are the same across alternatives
- **Mixed Logit**: we include both types of attributes in the utility:

$$V_{nj} = s_{n\,j}^T + x_{nj}^T \beta$$

Interpreting Multinomial Logit Coefficients #flashcard

- In multinomial logit ($V_{nj} = s_{n\,j}^T$), we specify a base group 1 where $_1 = 0$:

$$_1 = 0 \implies \exp(s_{n1}) = \exp(0) = 1$$

- Therefore:

$$\begin{aligned} \log\left(\frac{P_{ni}}{P_{n1}}\right) &= \log\left(\frac{\exp(s_{ni})}{\sum_{j=1}^{J} \exp(s_{nj})} \times \frac{\sum_{j=1}^{J} \exp(s_{nj})}{\exp(s_{n1})}\right) \\ &= \log\left(\frac{\exp(s_{ni})}{\cancelto{1}{\exp(s_{n1})}}\right) \\ &= \log(\exp(s_{ni})) \\ &= s_{ni} \end{aligned}$$

- i.e. $_i$ measures the marginal effect of s_n on the relative probability of choosing the alternative i compared to the base group measured on the log-scale.

Interpreting Conditional Logit Coefficients #flashcard

- In the conditional logit model ($V_{nj} = x_{nj}^T \beta$), attributes x_{nj} are specific to a particular alternative j. Thus, partial effects are much simpler for conditional logit than multinomial logit:

$$\begin{cases} \frac{\partial P_{nj}}{\partial x_{nj}} & = P_{nj}(1 - P_{nj})\beta \quad \text{wn Partial Effect} \\ \frac{\partial P_{nj}}{\partial x_{ni}} & = -P_{nj} P_{ni} \beta \quad \text{Corss Attribute Effect} \end{cases}$$

  - We can see that, for one particular attribute, the own partial effect and cross partial effect are in different directions.

Independence of Irrelevant Alternatives (IIA) Property for Logit #flashcard

- In logit models, the ratio between choice probabilities is:

$$\frac{P_{ni}}{P_{nj}} = \exp(V_{ni} - V_{nj})$$

- In other words, the relative probability of choosing i versus j only depends on the representative utilities for i and j, irrelevant to any 3rd alternative.

- IIA property is the consequence of assuming the error terms are iid.