Computational
Social Science

Logistic Regression Fundamentals .I

Roberto Cerina

20.02.2024

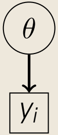
University of Amsterdam

(Refresher) The Problem of Statistical Inference

❶ Observe data from a sample of n units (e.g. individuals):
$y_i, \; \forall i \in \{1, \ldots, n\}$

❷ Posit a theory (model) for how the data was generate (Data Generating Process – DGP):
$y_i \sim f^\star$

❸ Describe the DGP in terms of some well defined *model parameters*:
$f^\star = f(\theta)$



Note: This representation is called a 'Directed Acyclic Graph' (DAG)

(Refresher)The Problem of Statistical Inference

❹ Estimate value of unknown parameter $\theta \rightarrow$ choose the 'most compatible with observed data' (Maximum Likelihood Estimate):
$\hat{\theta}_{MLE} = \arg\max_{\theta} \mathcal{L}(\boldsymbol{y} \mid \theta)$

❺ From the *MLE* we can derive the Empirical Posterior Distribution of θ
$\theta \sim g(\hat{\theta}_{MLE})$

❻ From this distribution, we can sample *plausible* values of the parameter θ, and make statements about its nature, accounting for uncertainty:
$\theta_s \sim g(\hat{\theta}_{MLE})$ – draw S plausible values of θ;
$\hat{\Pr}(\theta > 0) = \frac{1}{n} \sum_s^n (\theta_s > 0)$ – count how many are > 0 to see if 'significant' (for example ...)

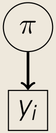
The ‘Homogeneous Probability’ Bernoulli Model

$$y_i \sim \text{Bernoulli}(\pi) \qquad \forall\, y_i \in \{0, 1\},\ i \in \{1, ..., n\}$$

$\pi = \Pr(y_i = 1)$

- the probability that an event happens.

The ‘Homogeneous Probability’ Bernoulli Model



Note:

- ‘square’ nodes indicate ‘observed / known values’;
- ‘circular’ nodes indicate ‘unknown parameters’;
- ‘solid’ arrows indicate ‘stochastic’ relationships – i.e. subject to random variability;
- ‘dotted’ arrows indicate ‘deterministic’ relationships – i.e. a given input will always provide the same output;

Estimation (Point Estimate)

Define a loss function, and minimise !

- Likelihood: $\mathcal{L}_i(\boldsymbol{y} \mid \pi)$
- for an observation i: $\mathcal{L}_i = \pi^{y_i}(1-\pi)^{1-y_i}$
- for the entire sample: $\boldsymbol{\mathcal{L}} = \Pi_i^n \mathcal{L}_i$
- log-likelihood: $\log(\boldsymbol{\mathcal{L}}) = \log(\Pi_i^n \mathcal{L}_i) = \sum_i^n \log(\mathcal{L}_i) \leftarrow$ this we want to maximise
- negative log-likelihood: $L = -\sum_i^n \log(\mathcal{L}_i) = \sum_i^n -(y_i \log(\pi) + (1-y_i)\log(1-\pi))$
  - a.k.a. as 'Binary Cross-Entropy' or 'Log-Loss for Binary Classification

✏ This has an analytical solution:
$\hat{\pi}_{MLE} = \frac{1}{n}\sum_i^n y_i = \bar{y} \leftarrow$ the sample mean is the MLE of π.

Uncertainty

- by asymptotic normality ($n \to \infty$) / Central Limit Theorem / Laplace approximation:

  $\pi \sim N\left(\hat{\pi}_{MLE}, \hat{\sigma}_{\pi}^{2}\right)$

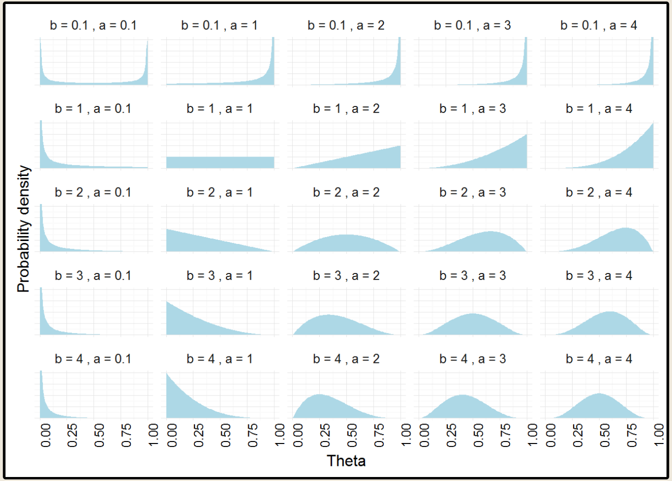
  $\hat{\sigma}_{\pi}^{2} = \frac{\hat{\pi}_{MLE}(1-\hat{\pi}_{MLE})}{n}$

- This is a good approximation of the true posterior distribution when:
  1. n is large enough;
  2. the sample is reasonably balanced.
- When the approximation fails, we risk producing samples estimates outside $[0, 1]$...

Uncertainty

- Alternative distribution by Bayesian empirical posterior distribution:
  $\pi \sim \text{Beta}(\alpha = \sum_i^n y_i, \beta = n - \sum_i^n y_i)$;
  - α: n. of observable 'successes' (events which happened – $y_i = 1$);
  - β: n. of observable 'failures' (events which happened – $y_i = 0$);

Beta Distribution



[0] `https://rpubs.com/RRisto/betadist`

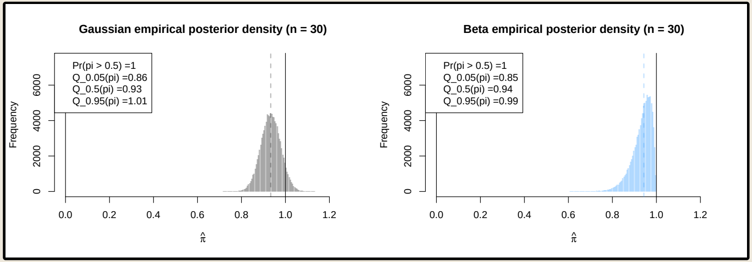
Simulation-Based (Monte-Carlo) Inference

- What statements can be made about π ?
- draw S values from the empirical posterior:

  $\pi_S \sim N\left(\hat{\pi}_{MLE}, \frac{\hat{\pi}_{MLE}(1-\hat{\pi}_{MLE})}{n}\right)$;

  $\pi_S \sim \text{Beta}(\alpha = \sum_i^n y_i, \beta = n - \sum_i^n y_i)$;

  $\pi_{1:S} = \{0.98, 0.98, 0.87, 0.940.95, 0.88, 0.94...\}$;
- Plot the distribution of π and infer its properties:
  - ✏ Classification rule: what % of the simulated values of π are larger than 0.5 ?

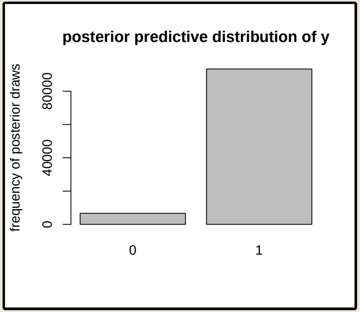


Prediction

Posterior Predictive Distribution:

❶ simulate S 'new' values from π according to its empirical posterior distribution:
$\tilde{\pi}_s \sim \text{Beta}(\alpha = \sum_i^n y_i, \beta = n - \sum_i^n y_i)$

❷ simulate S new values $\tilde{y}$ according to its likelihood, conditional on the simulated values of π:
$\tilde{y}_s \sim \text{Bernoulli}(\tilde{\pi}_s)$



The ‘Heterogeneous Probability’ Bernoulli Model

$$
\begin{aligned}
y_i \sim & \text{Bernoulli}(\pi_i) \qquad\qquad \forall\, y_i \in \{0,1\},\ i \in \{1,\dots,n\}, \\
\pi_i = & f(\mu_i) \\
\mu_i = & \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \\
\pi_i = & f(\mu_i) = \frac{\exp(\mu_i)}{1+\exp(\mu_i)} \rightarrow \text{inverse-logit link}
\end{aligned}
$$

✏ μ_i is ‘squeezed’ to a probability scale to generate π:

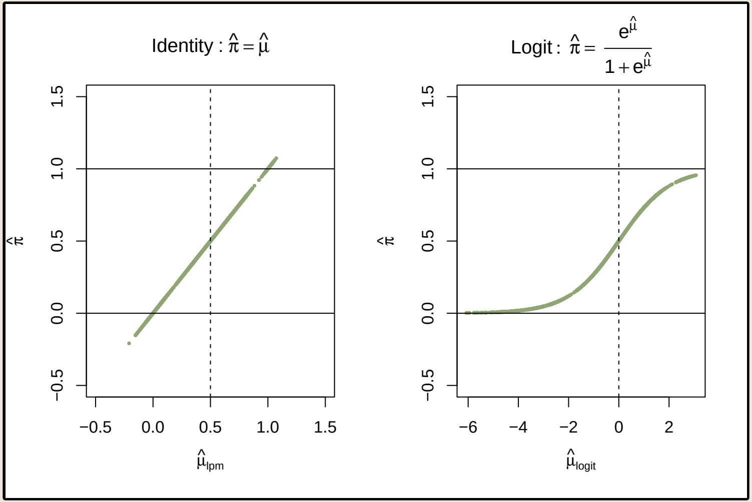
❶ $\pi \in [0,1]$;

❷ a change in the covariates of one-unit **cannot** generate an increase or decrease in probability larger than 1 ;

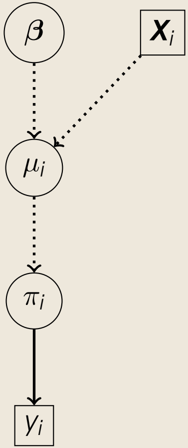
❸ change of one unit in one covariate has **heterogeneous** effects:

- $\rightarrow$ as μ gets closer to the ‘tails’ (closer to 0 or 1 in terms of the distribution of π), the impact of x on π exponentially decreases ;
- $\rightarrow$ the impact of any x is maximised when μ is around 0 (closer to 0.5 in terms of the distribution of π).

Modeling Probabilities: Linear v. Logistic



The ‘Heterogeneous Probability’ Bernoulli Model



Estimation (Point Estimate)

Define a loss function, and minimise !

- Likelihood: $\mathcal{L}_i(\boldsymbol{y} \mid \beta_0, ..., \beta_p)$
- for an observation i: $\mathcal{L}_i = \left(\frac{\exp(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})}{1+\exp(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})}\right)^{y_i}\left(1-\frac{\exp(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})}{1+\exp(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})}\right)^{1-y_i}$
- minimise: $L = -\sum_i^n \log(\mathcal{L}_i)$

✏ This **does not have an analytical solution !**

Uncertainty

- by asymptotic normality ($n \to \infty$) / Laplace approximation:

$$\boldsymbol{\beta} \sim MN(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}})$$

  - software will estimate $\hat{\boldsymbol{\beta}}$ via optimisation...
  - use estimate $\hat{\boldsymbol{\Sigma}}$ by plugging in $\hat{\boldsymbol{\beta}}$ into the *Hessian* matrix...

The Hessian Matrix

$$\hat{\boldsymbol{\Sigma}} = (-\boldsymbol{H})^{-1}$$

$$\boldsymbol{H} = \left.\frac{\partial^2 \log \boldsymbol{\mathcal{L}}(\boldsymbol{y} \mid \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \qquad H_{kj} = \frac{\partial^2 \log \boldsymbol{\mathcal{L}}(\boldsymbol{y} \mid \boldsymbol{\beta})}{\partial \beta_k \partial \beta_j}$$

The Hessian Matrix

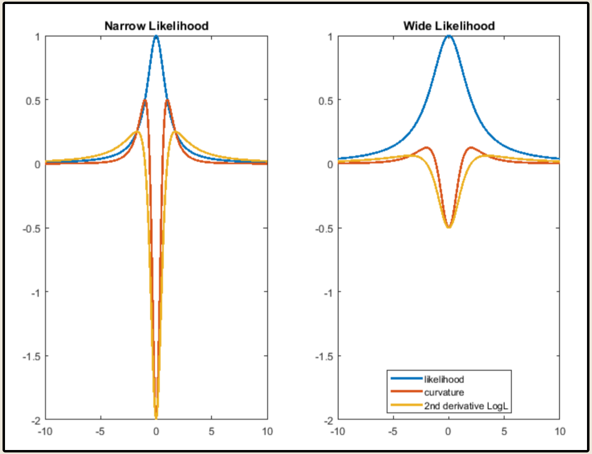
- For a function $f(\mathbf{x})$, with $\mathbf{x}$ being a vector of parameters, the Hessian matrix H is defined as:

$$H_{kj} = \frac{\partial^2 f}{\partial x_k \partial x_j}$$

- Represents *curvature* of that function with respect to its variables...
  - i.e. whether the rate of change of the function is increasing or decreasing with respect to a change in a variable of interest.

- At its peak (i.e. evaluated at the MLE) the curvature is negative ...
  - $\rightarrow$ multiply it by -1 to use it as a *positive definite* covariance matrix;

The Hessian Matrix



[0]https://stats.stackexchange.com/questions/289190/theoretical-motivation-for-using-log-likelihood-vs-likelihood

The Hessian Matrix

✏ Why does the curvature of the log-likelihood indicate the covariance of the empirical posterior?

- When H is evaluated at the MLE, the log-likelihood will be at its ‘peak’...
- H then tells us about the ‘tightness’ or ‘concavity’ of the peak of the log-likelihood function...
- A more negative second derivative (indicating a steeper and tighter peak) suggests that the parameter can be estimated more precisely, as small changes in β lead to large decreases in likelihood, pinpointing the maximum more distinctly...

Simulation-Based (Monte-Carlo) Inference

- What statements can be made about each β_j ?
- draw S simulated values from the *marginal distribution* of β_j:

$$\beta_j^s \sim N(\hat{\beta}_j, \hat{\sigma}^2_{\beta_j}) \qquad \forall s \in \{1, ..., S\}$$

$\hat{\sigma}^2_{\beta_j}$ is simply the j^{th} diagonal element of Σ.

✏ Monte Carlo estimates from these simulated values can reveal significance, and intervals, much like in the univariate case.

Note:

✏ The marginal distribution (which is univariate normal) is sufficient to make inference about a single coefficient amongst those in $\boldsymbol{\beta}$.

✏ If we wanted to make contemporaneous inference about the value of every beta, the *joint distribution* (the multivariate normal) would be necessary.

Predicting Probability / Risk

❶ define a set of L 'new subjects' characterised by design vector:
$\tilde{x}_l, \forall l \in \{1, ..., L\}$;

❷ simulate S 'new' values from $\boldsymbol{\beta}$ according to its empirical (joint) posterior distribution:
$\boldsymbol{\beta}^s \sim N(\hat{\boldsymbol{\beta}}_{MLE}, \hat{\boldsymbol{\Sigma}})$;

❸ Calculate μ_l for each simulation round:
$\tilde{\mu}_l^s = \beta_0^s + \beta_1^s \tilde{x}_{l1} + ... + \beta_p^s \tilde{x}_{lp}$

❹ Calculate π_l for each simulated μ_l:
$\tilde{\pi}_l^s = \frac{\exp(\tilde{\mu}_l^s)}{1+\exp(\tilde{\mu}_l^s)}$

- You can then use Monte Carlo methods to make inference about predictions – the MC median will typically be 'your best guess'.

Predicting Class

❺ simulate S new values $\tilde{y}$ according to its likelihood, conditional on the simulated values of $\tilde{\pi}$:
$\tilde{y}_l^s \sim \text{Bernoulli}(\tilde{\pi}_l^s)$;

- the MC mode will typically be 'most likely class'.