Modern Applied Statistics
MAST30027

Shromann Majumder

1 General Linear Models

1.1 Binomial Regression

Its main assumptions is that Y_i follows a Binomial Distribution and p_i has a relationship with the design constants and thier respective β parameters through a **Link Function**. take the inverse of g to make g^{-1} and use that for the *bin*'s p_i

$$Y_i \sim bin(m_i, p_i = g^{-1}(\eta_i = X_i^T\beta))$$

- **Link Function** to link p_i with x_i and β_i

$$g(p_i) = \eta_i = X_i^T\beta = \sum_{j=1}^{n} \beta_{ij}x_{ij}$$

1. logit:

$$\log \frac{p}{1-p}$$

2. complementary log-log:

$$\log(-\log(1-p))$$

3. probit:

$$\Phi^{-1}(p)$$

- **Log-Likelihood** to estimate β's values

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \log \Pr(Y_i = y_i) \\ &= c + \sum_{i=1}^{n} y_i \log(g^{-1}(\eta_i)) + (m_i - y_i)\log(1 - g^{-1}(\eta_i)) \end{aligned}$$

This has no closed form solution. So numerical search is needed. R uses `optim`, which is a greedy search algorithm. So multiple initial values are tested to avoid getting stuck in local optimums. Theres also `glm`, `predict`.

- **Aymptotic properties MLE** to find CI's of β estimates.

$$\hat{\theta}_{MLE} = \arg\min_{\theta} [l(\theta; y_{observed}) = f_i(\cdot;\theta) = f(\cdot; x_i, \theta)]$$

MLE's asymptotic properties:

1. Asymptotically Consistent:

$$n \to \infty,\ \hat{\theta} \to \theta^\star$$

2. Asymptotically Normal:

$$\hat{\theta} = N(\theta^\star, \mathcal{I}(\theta^\star)^{-1})$$

Observed Information: depends on the *observed* y I doubt many understand this well, but in summary: the hessain matrix is filled with second-order partial derivatives to describe the curvature of log-likelihood w.r.t θ.

$$\mathcal{J}(\theta) = -H_{l\theta} = -\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T}$$

Fischer's Information: depends on the $r.v\ Y$

$$\mathcal{I} = E[\mathcal{J}(\theta; Y)]$$

Binomial Regression with **2 parameters**, has the $\mathcal{I}$ of the form

$$\mathcal{I}(\beta) = \begin{bmatrix} \sum_{i=1}^{n} m_i p_i(1-p_i) & \sum_{i=1}^{n} m_i x_i p_i(1-p_i) \\ \sum_{i=1}^{n} m_i x_i p_i(1-p_i) & \sum_{i=1}^{n} m_i x_i^2 p_i(1-p_i) \end{bmatrix}$$

3. Asymptotically Efficient: If the above two conditions are met, then $\hat{\theta}_{MLE}$ is asymptotically unbiased estimator with smallest variance $\mathcal{I}(\theta^{\star})^{-1}$

- **Wald CI for** $t^T\theta$, when calculating CI through Asymptotic Normality We know the following,

$$\hat{\theta} \approx N(\theta^{\star}, \mathcal{I}(\theta^{\star})^{-1})$$

However, $\theta^{\star}$ (the true value of θ) is unknown. So we approximate $\mathcal{I}(\theta^{\star})^{-1}$ using $\mathcal{I}(\hat{\theta})^{-1}$. Resulting in following statements and the $100(1-\alpha)\%$ confidence interval.

$$\hat{\theta} \approx N(\theta^{\star}, \mathcal{I}(\hat{\theta})^{-1}) \implies \mathbf{t}^T\hat{\theta} \approx N(\mathbf{t}^T\theta^{\star}, \mathbf{t}^T\mathcal{I}(\hat{\theta})^{-1}\mathbf{t})$$

$$\mathbf{t}^T\hat{\theta} \pm z_\alpha\sqrt{\mathbf{t}^T\mathcal{I}(\hat{\theta})^{-1}\mathbf{t}},\ z_\alpha = \Phi^{-1}(1-\alpha/2)$$

If $\mathbf{t}$ is a standard basis vector for its dimention. Then we can obtain the CI for each invidual parameter: $\theta_1, \ldots$. In that case, the approximate CI would be

$$\hat{\theta}_i \pm z_\alpha\sqrt{[\mathcal{I}(\hat{\theta})^{-1}]_{i,i}}$$

But if we don't have $\mathcal{I}$, then just use $\mathcal{J}$ as.

$$\mathcal{I}(\hat{\theta})^{-1} \approx \mathcal{J}(\hat{\theta}; y)^{-1}$$

The steps behind calculating CI's is rather recursive. The way to go about it is to follow the steps below

1. Calculate CI for $\eta = \mathbf{t}^T\hat{\theta}$: (η_l, η_u) , $\mathbf{t}^T$ can be a possible covariate matrix and thus used to calculate the CI, or better known as the confidence region for $\mathbf{t}^T\hat{\theta}$.
2. Calculate CI for $p = g^{-1}(\eta)$: $(g^{-1}(\eta_l), g^{-1}(\eta_u))$,this p is from the $Y_i \sim bin(m_i, p_i)$

- **log likelihood ratio CI**, is better than *Wald CI* as

1. Wald CI does $2 \times$ CIs to get the CI of p, where 'log likelihood ratio CI' does 1.
2. likelihood ratio CI holds for smaller sample size.

We begin with the following

$$2l(\hat{\theta}) - 2l(\theta^{\star}) \sim \chi^2_k$$

$\cdot$ k is number of columns of $\theta^{\star}$, thus the log likelihood ratio CI is defined as

$$\{\theta : 2l(\hat{\theta}) - 2l(\theta^{\star}) \leq \chi^2_k(1-\alpha)\}$$

$\cdot$ $\chi^2_k(1-\alpha)$ is the $100(1-\alpha)\%$ point for χ^2_k distribution

- **MLE: regularity conditions**, what we need for MLE to actually work

1. log-likelihood function (l) is smooth, i.e thrid-order derivatives w.r.t to θ exists and continious.
2. thrid-order derivatives of l w.r.t to θ have bounded expectations.
3. support of Y_i does not depend on θ.
4. the domain of θ is finite dimentional and does not depend on Y_i.
5. $\theta^{\star}$ is not on the boundry of its domain.