

PROSES INFERENSI PADA MODEL LOGIT

Oleh: Agus Rusgiyono
Program Studi Statistika FMIPA UNDIP

Abstracts

Let $\{ Y_1, Y_2, Y_3, \dots, Y_n \}$ represent the response on a nominal random variable of Bernoulli distribution, with $P[Y_j = 1] = p$, $P[Y_j = 0] = 1 - p$ where p is a parameter with unknown value. Problems of estimating used smallest square methods in linear regression model can overcome with used maximum likelihood method in logistic regression..

Suppose $\ln(L_n(p)) = \left(\sum_{j=1}^n Y_j \right) \ln(p) + \left(n - \sum_{j=1}^n Y_j \right) \ln(1 - p)$ is

maksimum likelihood estimators of \hat{p} . In case can be obtained from first condition, $\ln(L_n(p))$ to be maximum at point $p = \hat{p}$ then be obtained $\hat{p} = \bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j$ and

that is unbiased estimator because $E(\hat{p}) = \frac{1}{n} \sum_{j=1}^n E(Y_j) = p$

To be test hypothesis that $p = p_0$, with a large sample size used fact that $\frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1 - p_0)}} \cong N[0,1]$

Keywords : Estimator, unbiased estimator, test statistic

1. PENDAHULUAN

Diketahui $\{ Y_1, Y_2, Y_3, \dots, Y_n \}$ adalah sampel acak dari distribusi Bernoulli, dengan dua nilai yang mungkin bagi Y yaitu 0 atau 1. Misalkan $P[Y_j = 1] = p$ dan $P[Y_j = 0] = 1 - p$ dengan p sebuah nilai yang tidak diketahui.

Sebagai ilustrasi, pandang kejadian jenis kelamin anak sapi yang dilahirkan jantan atau betina. Untuk meyakinkan bahwa probabilitas jenis kelamin anak-anak sapi yang dilahirkan jantan atau betina tidak sama, artinya $p = p_0 \neq 0,5$ Maka untuk setiap kelahiran ke j dikaitkan dengan nilai $Y_j = 1$ jika anak sapi yang lahir adalah betina dan $Y_j = 0$ bila anak sapi lahir jantan.

Untuk mengetahui besarnya probabilitas nilai $P[Y_j = 1] = p$ terhadap berbagai nilai dari variabel penjelas $X_i, i = 1, 2, 3, \dots, k$ maka penggunaan metode kuadrat terkecil pada model regresi linier biasa menimbulkan suatu masalah baik mengenai daerah hasil dari nilai variabel respon Y yang dapat diluar $\{0,1\}$ ataupun asumsi kenormalan galat dan varian yang harus konstan. Sehingga diperlukan fungsi yang menghubungkan variabel respon dengan variabel penjelas yakni
$$\log \frac{p(X)}{1 - p(X)} = a + b_i X_i \quad i = 1, 2, 3, \dots, k$$

Y_j ini berdistribusi Bernoulli dengan parameter p_i dengan fungsi densitas probabilitasnya $f(y_i, p_i) = (1 - p_i) \exp \left[y_i \log \left(\frac{p_i}{1 - p_i} \right) \right]$.

$\left(\frac{p_i}{1 - p_i} \right)$ disebut odds ratio dari p , menyatakan peluang terjadinya $Y=1$ dibanding peristiwa $Y=0$.

Penaksiran parameter p_i ini menggunakan metode maksimum likelihood

$$\begin{aligned} L_n(p) &= f(Y_1|p) \cdot f(Y_2|p) \cdot f(Y_3|p) \dots f(Y_n|p) = \prod_{j=1}^n f(Y_j|p) \\ &= \prod_{j=1}^n p^{Y_j} (1 - p)^{1 - Y_j} \\ &= p^{\sum_{j=1}^n Y_j} (1 - p)^{n - \sum_{j=1}^n Y_j} \end{aligned}$$

bila $p = p_0$ merupakan probabilitas bersama dari $\{ Y_1, Y_2, Y_3, \dots, Y_n \}$.

2. PENAKSIR MAKSIMUM LIKELIHOOD UNTUK PROPORSI p

Sebuah fungsi probabilitas terkait masalah yang tersebut pada sub bab pendahuluan di atas dapat didefinisikan sebagai:

$$\begin{aligned} f(y|p_0) = P[Y_j = y] &= p_0^y (1 - p_0)^{1 - y} = p_0 \text{ jika } y = 1 \\ &= 1 - p_0 \text{ jika } y = 0 \end{aligned}$$

Fungsi likelihood dalam kasus ini didefinisikan sebagai ;

$$\begin{aligned} L_n(p) &= f(Y_1|p) \cdot f(Y_2|p) \cdot f(Y_3|p) \dots f(Y_n|p) = \prod_{j=1}^n f(Y_j|p) \\ &= \prod_{j=1}^n p^{Y_j} (1 - p)^{1 - Y_j} \\ &= p^{\sum_{j=1}^n Y_j} (1 - p)^{n - \sum_{j=1}^n Y_j} \end{aligned}$$

merupakan probabilitas bersama dari $\{ Y_1, Y_2, Y_3, \dots, Y_n \}$., bila $p = p_0$

Dasar pemikiran taksiran maksimum likelihood sekarang adalah memilih p sedemikian hingga probabilitas dari sampel yang diambil maksimal.

Karena memaksimalkan $L_n(p)$ equivalent dengan memaksimalkan

$$\ln(L_n(p)) = \left(\sum_{j=1}^n Y_j \right) \ln(p) + \left(n - \sum_{j=1}^n Y_j \right) \ln(1 - p)$$

penaksir maksimum likelihood \hat{p} dalam kasus ini dapat diperoleh dari syarat pertama untuk memaksimalkan $\ln(Ln(p))$ di titik $p = \hat{p}$, sehingga didapat :

$$0 = \frac{d \ln(Ln(\hat{p}))}{d\hat{p}} = \left(\sum_{j=1}^n Y_j \right) \frac{1}{\hat{p}} - \left(n - \sum_{j=1}^n Y_j \right) \frac{1}{1-\hat{p}} = n \left(\frac{\bar{Y}}{\hat{p}} - \frac{1-\bar{Y}}{1-\hat{p}} \right)$$

$$\text{jadi } \hat{p} = \bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j$$

Perhatikan bahwa ini adalah estimator tak bias ^[3] karena :

$$E(\hat{p}) = \frac{1}{n} \sum_{j=1}^n E(Y_j) = p_0$$

lebih dari itu dari teorema limit pusat didapat :

$$\sqrt{n}(\hat{p} - p_0) = \frac{1}{\sqrt{n}} \sum_{j=1}^n (Y_j - p_0) \cong N[0, \sigma_0^2], \text{ berlaku untuk } n \text{ besar}$$

dimana :

$$\begin{aligned} \sigma_0^2 &= \text{Var}(Y_j) = E[(Y_j - p_0)^2] \\ &= (1 - p_0)^2 p_0 + (-p_0)^2 (1 - p_0) \\ &= p_0(1 - p_0) \end{aligned}$$

$$\text{jadi untuk ukuran sampel besar berlaku : } \frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1 - p_0)}} \cong N[0,1]$$

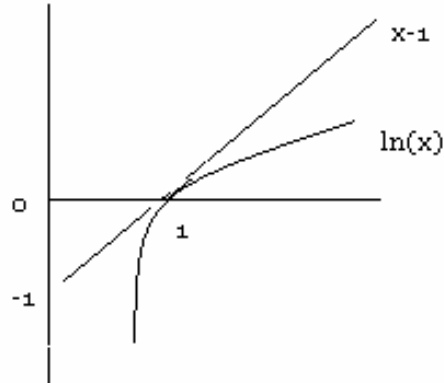
Hasil ini dapat digunakan untuk menguji hipotesis tentang p_0

Dalam hal ini dibawah hipotesis null bahwa probabilitas kelahiran sapi jantan dan

betina sama, $p_0 = 0,5$ didapat : $2\sqrt{n}(\hat{p} - 0,5) = \frac{\sqrt{n}(\hat{p} - 0,5)}{\sqrt{0,5 \times 0,5}} \cong N[0,1]$, sehingga :

$2\sqrt{n}(\hat{p} - 0,5)$ adalah statistic uji normal standard dari uji hipotesis null tentang $p_0 = 0,5$

Sebuah alasan formal bagi penaksiran maksimum likelihood didasarkan pada fakta bahwa $0 < x < 1$ dan $x > 1$, $\ln(x) < x - 1$. Hal ini diilustrasikan dalam gambar sbb :



Gambar 1. Fungsi $\ln(x) \leq x - 1$

Pertidaksamaan , $\ln(x) < x - 1$ terpenuhi dengan baik untuk $x \neq 1$, dan $\ln(1) = 1 - 1 = 0$

Konsekuensinya :

$$\ln\left(\frac{f(Y_j|p)}{f(Y_j|p_0)}\right) \leq \frac{f(Y_j|p)}{f(Y_j|p_0)} - 1$$

Dengan mengambil nilai harapan matematisnya diperoleh :

$$\begin{aligned} E\left[\ln\left(\frac{f(Y_j|p)}{f(Y_j|p_0)}\right)\right] &\leq E\left[\frac{f(Y_j|p)}{f(Y_j|p_0)}\right] - 1 \\ &= \frac{f(1|p)}{f(1|p_0)}P[Y_j = 1] + \frac{f(0|p)}{f(0|p_0)}P[Y_j = 0] - 1 \\ &= \frac{p}{p_0}p_0 + \frac{1-p}{1-p_0}(1-p_0) - 1 \\ &= p + 1 - p - 1 = 0 \end{aligned}$$

selanjutnya didapat :

$$E[\ln(f(Y_j|p))] - E[\ln(f(Y_j|p_0))] = E\left[\ln\left(\frac{f(Y_j|p)}{f(Y_j|p_0)}\right)\right] \leq 0$$

Hal ini berakibat bahwa :

$$E[\ln(L_n(p))] \leq E[\ln(L_n(p_0))]$$

Jadi $E[\ln(L_n(p))]$ adalah maksimal untuk $p = p_0$ dan dalam hal ini dapat ditunjukkan bahwa nilai maksimum ini adalah tunggal.

3. PENAKSIRAN MAKSIMUM LIKELIHOOD BAGI MODEL LOGIT

Misalkan $(Y_1, X_1), \dots, (Y_n, X_n)$ adalah sample acak dari distribusi logit bersyarat :

$$P[Y_j = 1 | X_j] = \frac{1}{1 + \exp(-\alpha_0 - \beta_0 X_j)}$$

$$P[Y_j = 0 | X_j] = \frac{\exp(-\alpha_0 - \beta_0 X_j)}{1 + \exp(-\alpha_0 - \beta_0 X_j)}$$

dimana α_0 dan β_0 adalah nilai parameter yang ditaksir. Model ini disebut model logit karena

$$P[Y_j = 1 | X_j] = F(\alpha_0 + \beta_0 X_j)$$

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

adalah fungsi distribusi dari distribusi logistik (logit)

fungsi probabilitas bersyarat yang terkandung adalah :

$$f(y | X_j, \alpha_0, \beta_0) = F(\alpha_0 + \beta_0 X_j)^y (1 - F(\alpha_0 + \beta_0 X_j))^{1-y}$$

$$= F(\alpha_0 + \beta_0 X_j) \quad \text{jika } y = 1$$

$$= 1 - F(\alpha_0 + \beta_0 X_j) \quad \text{jika } y = 0$$

sekarang fungsi log likelihood bersyarat adalah :

$$\ln(L_n(\alpha, \beta)) = \sum_{j=1}^n \ln(f(Y_j | X_j, \alpha, \beta))$$

$$= \sum_{j=1}^n Y_j \ln(F(\alpha + \beta X_j)) + \sum_{j=1}^n (1 - Y_j) \ln(1 - F(\alpha + \beta X_j))$$

$$= - \sum_{j=1}^n Y_j \ln(1 + \exp(-\alpha - \beta X_j)) - \sum_{j=1}^n (1 - Y_j) \ln(1 + \exp(-\alpha - \beta X_j))$$

$$+ \sum_{j=1}^n (1 - Y_j) \ln(\exp(-\alpha - \beta X_j))$$

$$= - \sum_{j=1}^n (1 - Y_j)(\alpha + \beta X_j) - \sum_{j=1}^n \ln(1 + \exp(-\alpha - \beta X_j))$$

sehingga diperoleh :

$$\begin{aligned} E \left[\ln \left(\frac{f(Y_j | X_j, \alpha, \beta)}{f(Y_j | X_j, \alpha_0, \beta_0)} \right) \middle| X_j \right] &\leq E \left[\frac{f(Y_j | X_j, \alpha, \beta)}{f(Y_j | X_j, \alpha_0, \beta_0)} \middle| X_j \right] - 1 \\ &= \frac{f(1 | X_j, \alpha, \beta)}{f(1 | X_j, \alpha_0, \beta_0)} P[Y_j = 1 | X_j] + \frac{f(0 | X_j, \alpha, \beta)}{f(0 | X_j, \alpha_0, \beta_0)} P[Y_j = 0 | X_j] - 1 \\ &= \frac{f(1 | X_j, \alpha, \beta)}{f(1 | X_j, \alpha_0, \beta_0)} f(1 | X_j, \alpha_0, \beta_0) + \frac{f(0 | X_j, \alpha, \beta)}{f(0 | X_j, \alpha_0, \beta_0)} f(0 | X_j, \alpha_0, \beta_0) - 1 \\ &= f(1 | X_j, \alpha, \beta) + f(0 | X_j, \alpha, \beta) - 1 \\ &= \frac{1}{1 + \exp(-\alpha - \beta X_j)} + \frac{\exp(-\alpha - \beta X_j)}{1 + \exp(-\alpha - \beta X_j)} - 1 = 0 \end{aligned}$$

Jadi ;

$$E[\ln(Ln(\alpha, \beta)) | X_1, \dots, X_n] \leq E[\ln(Ln(\alpha_0, \beta_0)) | X_1, \dots, X_n]$$

selanjutnya hasil ini memberikan arah dalam menaksir α_0 dan β_0 dengan memaksimalkan $\ln(Ln(\alpha, \beta))$ untuk nilai α dan β :

$$\ln(Ln(\hat{\alpha}, \hat{\beta})) = \max_{\alpha, \beta} \ln(Ln(\alpha, \beta))$$

syarat pertama bagi nilai maksimum adalah :

$$0 = \frac{\partial \ln(Ln(\hat{\alpha}, \hat{\beta}))}{\partial \hat{\alpha}} = - \sum_{j=1}^n (1 - Y_j) + \sum_{j=1}^n \frac{\exp(-\hat{\alpha} - \hat{\beta} X_j)}{1 + \exp(-\hat{\alpha} - \hat{\beta} X_j)}$$

$$0 = \frac{\partial \ln(Ln(\hat{\alpha}, \hat{\beta}))}{\partial \hat{\beta}} = - \sum_{j=1}^n (1 - Y_j) X_j + \sum_{j=1}^n \frac{\exp(-\hat{\alpha} - \hat{\beta} X_j) X_j}{1 + \exp(-\hat{\alpha} - \hat{\beta} X_j)}$$

penaksir maksimum likelihood diperoleh dari pemecahan secara numeric yang dapat dikerjakan dengan cepat melalui software ekonometri.

4. ASIMTOTIS NORMAL DAN PEMBANGKITAN NILAI STATISTIK UJI t

Telah ditunjukkan bahwa jika ukuran sampel n cukup besar ;

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \cong N(0, \sigma_\alpha^2) \text{ dan } \sqrt{n}(\hat{\beta} - \beta_0) \cong N(0, \sigma_\beta^2)$$

jika $\hat{\sigma}_\alpha^2$ dan $\hat{\sigma}_\beta^2$ sebagai penaksir konsisten dari masing-masing σ_α^2 dan σ_β^2 telah dihitung secara numeris, akan didapat untuk berikutnya :

$$\frac{\sqrt{n}(\hat{\alpha} - \alpha_0)}{\hat{\sigma}_\alpha} \cong N(0,1) \text{ dan } \frac{\sqrt{n}(\hat{\beta} - \beta_0)}{\hat{\sigma}_\beta} \cong N(0,1)$$

Hasil ini dapat digunakan untuk menguji apakah α_0 dan β_0 berbeda secara statistic dengan 0 apa tidak.

Misalkan ingin diuji hipotesis null $\beta_0 = 0$.

Hipotesis ini tidak mengakibatkan probabilitas bersyarat $P[Y_j = 1 | X_j]$ tidak bergantung pada X_j maka dibawah penerimaan hipotesis null $\beta_0 = 0$ akan diperoleh :

$$\hat{t}_\beta = \frac{\sqrt{n}\hat{\beta}}{\hat{\sigma}_\beta} \cong N(0,1)$$

Statistik \hat{t}_β dinamakan nilai t yang dibangkitkan karena ini digunakan pada jalan yang sama sebagaimana nilai t dalam regresi linier.

Misal pada daerah kritis 5% pada uji dua sisi berdasarkan distribusi normal diperoleh 1,96, maka dalam hal ini criteria penolakan H null : $\beta_0 = 0$ pada tingkat signifikansi 5% adalah $|\hat{t}_\beta| > 1,96$

5. KESIMPULAN

1. Misalkan $(Y_1, X_1), \dots, (Y_n, X_n)$ adalah sample acak dari distribusi logit bersyarat :

$$P[Y_j = 1 | X_j] = \frac{1}{1 + \exp(-\alpha_0 - \beta_0 X_j)}$$

$$P[Y_j = 0 | X_j] = \frac{\exp(-\alpha_0 - \beta_0 X_j)}{1 + \exp(-\alpha_0 - \beta_0 X_j)}$$

dimana α_0 dan β_0 adalah nilai parameter yang ditaksir maka dalam menaksir α_0 dan β_0 dilakukan dengan memaksimalkan $\ln(Ln(\alpha, \beta))$ untuk nilai α dan β dimana

$$\ln(Ln(\hat{\alpha}, \hat{\beta})) = \max_{\alpha, \beta} \ln(Ln(\alpha, \beta)) \quad \text{dan}$$

$$\ln(Ln(\alpha, \beta)) = \sum_{j=1}^n \ln(f(Y_j | X_j, \alpha, \beta))$$

2. Jika ukuran sample n cukup besar ;

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \cong N(0, \sigma_\alpha^2) \text{ dan } \sqrt{n}(\hat{\beta} - \beta_0) \cong N(0, \sigma_\beta^2)$$

jika $\hat{\sigma}_\alpha^2$ dan $\hat{\sigma}_\beta^2$ sebagai penaksir konsisten dari masing-masing σ_α^2 dan σ_β^2 telah dihitung secara numeris, akan didapat untuk berikutnya :

$$\frac{\sqrt{n}(\hat{\alpha} - \alpha_0)}{\hat{\sigma}_\alpha} \cong N(0,1) \text{ dan } \frac{\sqrt{n}(\hat{\beta} - \beta_0)}{\hat{\sigma}_\beta} \cong N(0,1)$$

Hasil ini dapat digunakan untuk menguji apakah α_0 dan β_0 berbeda secara statistik dengan 0 apa tidak.

DAFTAR PUSTAKA:

1. Mood, Alexander; Introduction to The Theory of Statistics ; McGraw-Hill; 1974
2. Bierens, Herman J; The Logit Model : Estimation, Testing and Interpretation, <http://econ.La.psu.edu/~hbierens/ML-logit.pdf>
3. Rusgiyono, A., *Reduksi Bias melalui Nonparametrik Bootstrap*, Thesis Magister Matematika ITB, 1998