Proses inferensi pada model logit

Agus Rusgiyono

Abstracts

Let $\{ Y_1, Y_2, Y_3, \cdots, Y_n \}$ 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 linier 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 estimstors 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 hipothesis that $p = p_0$, with a large sample size used fact that $\frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1-p_0)}} \cong N[0,1]$

Keyword : Estimator, unbiased estimator, test statistic

1. Pendahuluan

Diketahui $\{ Y_1, Y_2, Y_3, \cdots, Y_n \}$ 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 $X_i , i = 1,2,3,\ldots,k$ 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 varibel respon dengan variabel penjelas yakni $\log \frac{p(X)}{1-p(X)} = a + b_i X_i \quad i = 1,2,3,\ldots k$

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 maximum likelihhod

$$L_n(p) = f(Y_1|p).f(Y_2|p).f(Y_3|p)\ldots f(Y_n|p) = \prod_{j=1}^{n} f(Y_j|p)$$

$$= \prod_{j=1}^{n} p^{Y_j}(1-p)^{1-Y_j}$$

$$p^{\sum_{j=1}^{n} Y_j}(1-p)^{n-\sum_{j=1}^{n} Y_j}$$

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

3. Penaksir maksimum Likelihood untuk proporsi p

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

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

Fungsi likelihood dalam kasus ini didefinisikan sebagai ;

$$L_n(p) = f(Y_1|p).f(Y_2|p).f(Y_3|p)\ldots f(Y_n|p) = \prod_{j=1}^{n} f(Y_j|p)$$

$$= \prod_{j=1}^{n} p^{Y_j}(1-p)^{1-Y_j}$$

$$= p^{\sum_{j=1}^{n} Y_j}(1-p)^{n-\sum_{j=1}^{n} Y_j}$$

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

Dasar pemikiran taksiran maximum likelihood sekarang adalah memilih p sedemikian hingga probabilitas dari sampel yang diambil maksimal.
Karena memaximalkan $L_n(p)$ equivalent dengan memaximalkan

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

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

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

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

Perhatikan bahwa ini adalah estimator tak bias (Agus Rusgiyono, 1998) karena :

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

lebih dari itu dari teorema limit pusat didapat :

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

dimana :

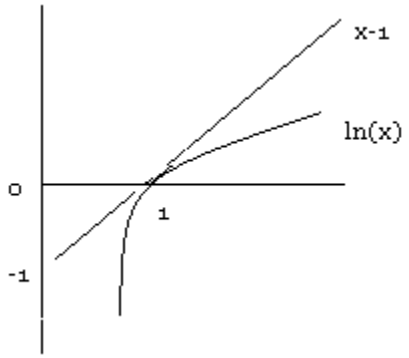
$$\sigma_0^2 = Var(Y_j) = E\left[(Y_j - p_0)^2\right]$$
$$= (1-p_0)^2 p_0 + (-p_0)^2 (1-p_0)$$
$$= p_0(1-p_0)$$

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 : $2\sqrt{n}(\hat{p} - 0{,}5) = \frac{\sqrt{n}(\hat{p} - 0{,}5)}{\sqrt{0{,}5x0{,}5}} \cong N[0,1]$, 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. Hall ini diilustrasikan dalam gambar sbb :



$$\ln(x) \le x - 1$$

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

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

Dengan mengambil nilai harapan matematisnya diperoleh :

$$E\left[\ \ln\left(\frac{f(Y_j|p)}{f(Y_j|p_0)}\right)\ \right] \le E\left[\ \frac{f(Y_j|p)}{f(Y_j|p_0)}\ \right] - 1$$

$$= \frac{f(1|p)}{f(1|p_0)} P\left[Y_j = 1\right] + \frac{f(0|p)}{f(0|p_0)} P\left[Y_j = 0\right] - 1$$

$$= \frac{p}{p_0} p_0 + \frac{1-p}{1-p_0}(1-p_0) - 1$$

$$= p + 1 - p - 1 = 0$$

selanjutnya didapat :

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

Hal ini berakibat bahwa :

$$E\left[\ln(L_n(p))\right] \le E\left[\ln(L_n(p_0))\right]$$

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

4. Penaksiran maximum likelihood bagi model logit

Misalkan $(Y_1, X_1),\ldots,(Y_{n1}, X_n)$ adalah sample acak dari distribusi logit bersyarat :

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

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

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

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

adalah fungsi distribusi dari distribusi logistik (logit)

fungsi probabilitas bersyarat yang terkandung adalah :

$$f(y \middle| X_j, \alpha_0, \beta_0) = F(\alpha_0 + \beta_0 X_j)^y (1 - F(\alpha_0 + \beta_0 X_j))^{1-y}$$
$$= F(\alpha_0 + \beta_0 X_j) \quad jika \quad y = 1$$
$$= 1 - F(\alpha_0 + \beta_0 X_j) \quad jika \quad y = 0$$

sekarang fungsi log likelihood bersyarat adalah :

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

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

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

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

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

sehingga diperoleh :

$$E\left[\ln\left(\frac{f(Y_j|X_j,\alpha,\beta)}{f(Y_j|X_j,\alpha_0,\beta_0)}\right)\Big|X_j\right] \le E\left[\frac{f(Y_j|X_j,\alpha,\beta)}{f(Y_j|X_j,\alpha_0,\beta_0)}\Big|X_j\right]-1$$

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

$$= \frac{f(1|X_j,\alpha,\beta)}{f(1|X_j,\alpha_0,\beta_0)} f(1|X_j,\alpha_0,\beta_0) + \frac{f(0|X_j,\alpha,\beta)}{f(0|X_j,\alpha_0,\beta_0)} f(0|X_j,\alpha_0,\beta_0)-1$$

$$= f(1|X_j,\alpha,\beta) + f(0|X_j,\alpha,\beta)-1$$

$$= \frac{1}{1+\exp(-\alpha-\beta X_j)} + \frac{\exp(-\alpha-\beta X_j)}{1+\exp(-\alpha-\beta X_j)} - 1 = 0$$

Jadi ;

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

selanjutnya hasil ini memberikan arah dalam menaksir α_0 dan β_0 dengan memaximalkan ln(Ln(α,β)) untuk nilai α dan β :

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

syarat pertama bagi nilai maximum 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 maximum likelihood diperoleh dari pemecahan secara numeric yang dapat dikerjakan dengan cepat melalui software ekonometri.

5. Asimtotis normal dan pembangkitan nilai statistic uji t

Telah ditunjukkan bahwa jika ukuran sample n cukup besar ;

$\sqrt{n}(\hat{\alpha}-\alpha_0) \cong N(0,\sigma_\alpha^2)$ 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 masinng-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 β_0 =0 .
Hipotesis ini tidak mengakibatkan probabilitas bersyarat $P[Y_j = 1 | X_j]$ tidak bergantung pada X_j maka dibawah penerimaan hipotesis null β_0 =0 akan diperoleh :

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

Statistik $\hat{\sigma}_\beta$ 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 : β_0 =0 pada tingkat signifikansi 5% adalah $|\hat{\sigma}_\beta| > 1{,}96$

6. Kesimpulan

1. misalkan $(Y_1, X_1),\ldots,(Y_{n1}, X_n)$ 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 memaximalkan ln(Ln(α, β)) 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 ;

$\sqrt{n}(\hat{\alpha} - \alpha_0) \cong N(0, \sigma_\alpha^2)$ 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.

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,Agus : Reduksi bias melalui non parametrik boostrap,Thesis magister matematika ITB, 1998