 http://dx.doi.org/10.29313/.v7i1.25541

Penerapan Metode Modifikasi *Hosmer-Lemeshow Test* pada Model Regresi Logistik Data Penderita Penyakit Hipertensi

Shavira Yuhadisi*, Suliadi

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Islam Bandung, Indonesia.

*shavirayuhadisi14@gmail.com, suliadi@gmail.com

Abstract. In this study, we will discuss the modification of the Hosmer-Lemeshow Test method. The Hosmer-Lemeshow test is a suitability test (Goodness of fit) based on the predicted probability values, the Hosmer-Lemeshow Test is widely used to test the suitability of the model using big data. The use of sufficiently large data in logistic regression analysis can create some test stability problems. Therefore, (Yu, Xu, & Zhu, 2017) propose to modify the Hosmer-Lemeshow Test method for big data which can minimize problems with the test power so that the test is more stable. The data used is data on people with hypertension in the city of Framingham, Massachusettes, which is obtained from the website kaggle.com. From this study it can be concluded that the logistic regression model used does not match the data on hypertension sufferers. So it is necessary to look for other models that are more suitable to the data.

Keywords: Goodness of fit, Logistic Regression, Modification of the Hosmer-Lemeshow Test Method.

Abstrak. Dalam penelitian ini akan dibahas mengenai modifikasi metode Hosmer-Lemeshow Test. Uji Hosmer-Lemeshow merupakan uji kesesuaian (Goodness of fit) berdasarkan nilai-nilai prediksi peluang, pengujian Hosmer-Lemeshow Test banyak digunakan untuk menguji kesesuaian model menggunakan data besar. Penggunaan data yang cukup besar dalam analisis regresi logistik dapat membuat beberapa masalah kestabilan uji. Oleh karena itu, (Yu, Xu, & Zhu, 2017) mengusulkan untuk memodifikasi metode Hosmer-Lemeshow Test untuk data besar yang dapat meminimalisir masalah pada kuasa uji sehingga ujinya lebih stabil. Data yang digunakan adalah data penderita penyakit hipertensi di kota Framingham, Massachusettes yang diperoleh dari situs website kaggle.com. Dari penelitian ini dapat disimpulkan bahwa model regresi logistik yang digunakan tidak cocok dengan data penderita penyakit hipertensi. Sehingga perlu dicari model lain yang lebih cocok dengan data.

Kata Kunci: *Goodness of fit* , Regresi Logistik, Modifikasi Metode *Hosmer-Lemeshow Test.*

1. Pendahuluan

Regresi logistik merupakan salah satu bagian dari analsis regresi yang sudah banyak dipakai orang untuk memodelkan hubungan antara peubah bebas dengan peubah tak bebas. Yang membedakan regresi logistik dengan regresi biasa adalah pada regresi biasa variabel responnya bertipe kontinyu sedangkan pada regresi logistik variabel responnya bersifat biner atau memiliki

2 kategori yaitu "0 dan 1" atau "sukses dan gagal".

Goodness of Fit Test yaitu suatu pengujian kesesuaian model dengan data. Apakah model yang digunakan sudah sesuai dengan data. Tes *goodness of fit* memberi tahu seberapa baik model yang digunakan sesuai dengan data (Lai & Liu, 2018). Ada beberapa metode yang bisa digunakan untuk menguji uji kesesuaian model regresi logistik, diantaranya *Pearson Chi-Square Statistic and Deviance* dan *Hosmer-Lemeshow Test*. Kelemahan dari *Hosmer-Lemeshow Test* adalah kuasa uji akan meningkat dengan ukuran sampel n, ini berarti jumlah ukuran sampel berpengaruh terhadap kuasa uji. Kuasa ujinya akan kurang jika ukuran sampelnya kecil. Dan jika ukuran sampel sangat besar, maka kuasa uji akan sangat tinggi sehingga cenderung menolak H_0 meskipun H_0 benar. Selain itu data dari populasi yang sama, uji ini akan cenderung menerima H_0 untuk ukuran sampel kecil dan cenderung menolak H_0 untuk sampel besar, dengan kata lain ada masalah kestabilan uji akibat perbedaan ukuran sampel.

Oleh karena itu, untuk mengatasi hal tersebut (Yu, Xu, & Zhu, 2017) mengusulkan untuk memodifikasi metode *Hosmer-Lemeshow Test* untuk data besar yang dapat meminimalisir masalah pada kuasa uji sehingga ujinya lebih stabil, artinya jika jika hipotesis nol benar maka akan cenderung menerima hipotesis nol dan jika hipotesis nol salah makan cenderung akan menolak hipotesis nol, baik ukuran sampel berukuran kecil atau besar.

2. Landasan Teori

Regresi Logistik

Misalkan Y1, Y2, ..., Yn adalah sampel acak yang berdistribusi Bernoulli dengan peubah bebas $X_1, X_2, \dots, X_k$ dan parameter $\pi(x)$, maka peluangnya adalah :

$$f(y_i) = \pi(x_i)^{y}(1-\pi(x_i))^{1-y_i} \quad \dots (2.1)$$

Dimana :

π_i = peluang kejadian ke-i.

y_i = peubah acak ke-i.

Model regresi logistik untuk peluang $P(Y=1)$ terhadap k variabel bebas $X_1, X_2, \dots, X_k$ adalah sebagai berikut. Untuk mempersingkat penulisan, maka kami tidak menuliskan indeks pengamatan. Maka model regresi logistik adalah sebagai berikut :

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}}{1+e^{(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}} \quad \dots (2.2)$$

Penaksiran Parameter Regresi Logistik

Untuk menduga koefisien regesi dilakukan dengan metode Maksimum Likelihood. Berdasarkan pada persamaan (2.1), maka untuk pengamatan $y_1, y_2, \dots, y_n$ fungsi likelihood-nya sebagai berikut :

$$L(\beta) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n}[\,\pi(x_i)^{y_i}\big(1-\pi(x_i)\big)^{1-y_i}] \quad \dots (2.3)$$

Lalu bentuk logaritma natural dari fungsi likelihood sebagai berikut :

$$\ln L(\beta) = \ln \prod_{i=1}^{n}[\,\pi(x_i)^{y_i}\big(1-\pi(x_i)\big)^{1-y_i}]$$

$$= \sum_{i=1}^{n}\{y_i \ln[\,\pi(x_i)] + (1-y_i)\ln(1-\pi(x_i)\} \quad \dots (2.4)$$

Setelah mendapatkan bentuk logaritmanya turunkan fungsi tersebut agar mendapatkan solusi atau hasil sebagai berikut :

Solusi untuk mendapatkan penaksiran parameter berdasarkan pada persamaan:

$$\hat{\beta}_{i+1} = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_k \end{pmatrix} - \begin{bmatrix} \frac{\partial^2 \ln L(\beta)}{\partial \beta_0^2} & \frac{\partial^2 \ln L(\beta)}{\partial \beta_0 \beta_1} & \cdots & \frac{\partial^2 \ln L(\beta)}{\partial \beta_0 \beta_k} \\ \frac{\partial^2 \ln L(\beta)}{\partial \beta_0 \beta_1} & \frac{\partial^2 \ln L(\beta)}{\partial \beta_1^2} & & \frac{\partial^2 \ln L(\beta)}{\partial \beta_1 \beta_k} \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 \ln L(\beta)}{\partial \beta_k \beta_1} & \frac{\partial^2 \ln L(\beta)}{\partial \beta_k \beta_2} & \cdots & \frac{\partial^2 \ln L(\beta)}{\partial \beta_k^2} \end{bmatrix}^{-1} \begin{pmatrix} \frac{\partial \ln L(\beta)}{\partial \beta_0} \\ \frac{\partial \ln L(\beta)}{\partial \beta_1} \\ \vdots \\ \frac{\partial \ln L(\beta)}{\partial \beta_k} \end{pmatrix} \quad \dots (2.5)$$

Proses iterasi ini akan berhenti ketika nilai $\hat{\beta}_{i+1} \cong \widehat{\beta}_i$ dimana dengan tingkat toleransi

kesalahan sebesar 10-6.

Untuk mengetahui faktor mana sajakah yang signifikan atau dapat mempengaruhi penyakit hipertensi maka dilakukan uji keberartian.

Hosmer-Lemeshow Test

Uji Hosmer-Lemeshow adalah uji kesesuaian (Goodness of fit) berdasarkan nilai-nilai prediksi peluang. Uji ini di populerkan oleh Hosmer dan Lemeshow. Prinsip dasar dari uji Hosmer-Lemeshow ini yaitu pertama dengan mengelompokkan $\hat{\pi}$ ke dalam g kelompok, lalu kemudian hitung $\hat{C}$ sebagai statistik ujinya (Hosmer & Lemeshow, 2000).

Hipotesis yang digunakan dalam pengujian Goodness of fit sebagai berikut:

H_0: Model yang digunakan sesuai dengan data.

H_1: Model yang digunakan tidak sesuai dengan data.

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1-\bar{\pi}_k)}$$

Dimana :

n'_k = total subjek kelompok ke-k.

$O_k = \sum_{j=1}^{c_k} y_i$; Nilai banyaknya sukses pada pengamatan kelompok ke-k dimana ck merupakan jumlah pola kovariat dalam desil.

$\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}}{n'_k}$; Nilai estimasi rata-rata peluang kelompok ke-k.

m_j = banyaknya nilai pengamatan pada baris $\bar{\pi}_k$.

$\hat{\pi}$ = banyaknya niai pengamatan pada kolom $\bar{\pi}_k$.

Jika Hipotesis nol benar, maka $\hat{C}$ akan berdistribusi chi-square dengan derajat bebas g-2 , sehingga nilai kritisnya untuk taraf uji α adalah $x^2_{(g-2;1-\alpha)}$ atau tolak H0 jika $\hat{C} > x^2_{(g-2;1-\alpha)}$

Modifikasi Hosmer-Lemeshow Test

Yu et al. (2017) menstabilkan kuasa uji melalui parameter non-sentral (λ), dengan mengalikan λ dengan suatu konstanta tertentu, di mana konstanta ini fungsi dari sampel size (n). Hal ini disebabkan karena kuasa uji dipengaruhi parameter non-sentral (λ). Seperti diketahui bahwa jika H0 benar maka $\hat{C}_g \sim x^2_{g-2}$. Sedangkan di bawah hipotesis alternatif, di mana model tidak cocok dengan data maka statistik Hosmer-Lemeshow yang digunakan mengikuti distribusi chi-square non-sentral, sebagai berikut :

$\hat{C}_g \sim x^2_{g-2}(\lambda)$

dimana :

λ = Parameter non-sentralis. Nilai λ dapat diestimasikan sebagai berikut :

$$\hat{\lambda} = (\hat{C}_g) - (g-2) = (\hat{C}_g) - v \quad \text{... (2.6)}$$

Dimana $\boldsymbol{v} = \boldsymbol{g} - \mathbf{2}$ dan $\hat{C}_{\boldsymbol{g}}$ adalah rata-rata yang dihasilkan dari sampel sebanyak K (Biasanya diperoleh dari simulasi).

Paul et al. (2012) juga merekomendasikan untuk ukuran sampel $1000 < n \leq 25000$ maka g yang digunakan adalah sebagai berikut :

$$g = \max\left(10, \min\left\{\frac{m}{2}, \frac{n-m}{2}, 2 + 8\left(\frac{n}{1000}\right)^2\right\}\right) \quad \text{... (2.7)}$$

Prosedur pengujian terbaru Hosmer-Lemeshow Test yang direkomendasikan oleh Yu et.al (2017) sebagai berikut :

Tahap 1 : Dari data yang berukuran n, kita menghitung statistik Hosmer-Lemeshow $\hat{C}$ dengan g = 10, lalu hitung λ.

Tahap 2 : Definisikan modifikasi λ_c

$$\lambda_c = c\lambda = c[T_g - (g-2)] \quad \text{... (2.8)}$$

dimana :

$$T_g = \sum_{j=1}^{g} \frac{(e_j - O_j)^2}{\overline{\text{ne}}_j(1 - \bar{\text{e}}_j)}$$

Tahap 3 : Jika $\lambda_c < 0$ maka terima H0 tetapi jika $\lambda_c \geq 0$, maka bangkitkan secara acak r dari distribusi $x^2_{g-2,\lambda}$. χ_c merupakan nilai kritis dari modifikasi Hosmer-Lemeshow, dimana χ_c merupakan solusi persamaan dari $\frac{1}{c}\int_0^{\infty}\left(1 - G_{m-2,x(\chi_c)}\right) f_{m-2}\left(\frac{x}{c} + (m-2)\right) dx = \alpha$.

Kita akan menolak hipotesis nol apabila nilai $\boldsymbol{r > \chi_c}$ dan sebaliknya kita akan menerima hiotesis nol apabila $\boldsymbol{r < \chi_c}$.

3. Hasil Penelitian dan Pembahasan

Data

Data yang digunakan merupakan mengenai data pasien penderita penyakit hipertensi di kota Farmingham, Massachusettes. Ukuran sampel yang digunakan sebanyak 4.190 data. Variabel tak bebas yang diamati memiliki 2 kategori yaitu memiliki penyakit hipertensi dan tidak memiliki penyakit hipertensi. Dan terdapat 6 variabel bebas yang diamati, sebagai berikut: X_1 = Jenis Kelamin (0 = Wanita, 1 = Pria), X_2 = Usia, X_3 = Jumlah Rokok yang Dihisap per hari, X_4 = Tekanan Darah Sistolik, X_5 = Tekanan Darah Diastolik, X_6 = BMI (*Body Mass Indeks*).

Tabel 1. Data Penyakit Hipertensi Penduduk di Kota Framingham, Massachusettes.

No	Memiliki Penyakit Hipertensi	Jenis Kelamin	Usia	Rokok yang Dihabiskan	Tekanan darah Sisitolik	Tekanan Darah Diastolik	BMI
1	0	1	39	0	106	70	26,97
2	0	0	46	0	121	81	28,73
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4190	0	0	52	0	133,5	83	21,47

Sumber : www.kaggle.com

Pemodelan Regresi Logistik

Pemodelan yang dilakukan menggunakan pemodelan dari analisis regresi logistik dengan hasil sebagai berikut :

$$\pi(\text{x}) = \frac{e^{(-25{,}844 + 0{,}294\,X_1 + 0{,}0295\,X_2 - 0{,}0003\,X_3 + 0{,}1082\,X_4 + 0{,}0812\,X_5 + 0{,}0674\,X_6)}}{1 + e^{(-25{,}844 + 0{,}294\,X_1 + 0{,}0295\,X_2 - 0{,}0003\,X_3 + 0{,}1082\,X_4 + 0{,}0812\,X_5 + 0{,}0674\,X_6)}}$$

Tabel 2. Nilai Odds Ratio Pada Data Penderita Penyakit Hipertensi di Kota Framingham, Massachusettes

Variabel	Odds Ratio
Jenis Kelamin	1.34
Usia	1.03
Rokok yang Dihabiskan	0,9997
Tekanan Darah Sistolik	1.11
Tekanan Darah Diastolik	1.08
BMI	1.07

Berdasarkan tabel 2 di atas nilai odds ratio dapat dijelaskan sebagai berikut:

1. Jenis Kelamin (X_1)

Seseorang dengan jenis kelamin pria memiliki resiko terkena penyakit hipertensi 1,34 kali dibandingkan dengan seseorang berjenis kelamin wanita.

2. Usia (X_2)

Seseorang yang usianya lebih tua 1 tahun mempunyai resiko terkena penyakit hipertensi sebesar 3% lebih tinggi dibandingkan dengan seseorang yang usianya lebih muda 1 tahun.

3. Rokok yang Dihabiskan (X_3)

Seseorang yang menghabiskan rokok lebih banyak 1 batang mempunyai resiko penyakit hipertensi sebesar 0,03% lebih rendah dibandingkan dengan orang yang tidak merokok.

4. Tekanan Darah Sistolik (X_4)

Seseorang yang tekanan darah sistoliknya meningkat 1 mmHg mempunyai resiko penyakit hipertensi sebesar 11% lebih tinggi dibandingkan dengan orang yang mempunyai tekanan darah sistoliknya normal yaitu sebesar 120 mmHg.

5. Tekanan Darah Diastolik (X_5)

Seseorang yang tekanan darah diastoliknya meningkat 1 mmHg mempunyai resiko penyakit hipertensi sebesar 8% lebih tinggi dibandingkan dengan orang yang mempunyai tekanan darah diastoliknya normal yaitu sebesar 80 mmHg.

6. BMI (X_6)

Seseorang yang BMI nya meningkat 1 satuan mempunyai resiko terkena penyakit hipertensi sebesar 7% lebih tinggi dibandingkan dengan orang yang mempunyai BMI nya normal.

Uji Keberartian

Uji keberatian ini akan melihat faktor mana sajakah yang dapat mempengaruhi penderita penyakit hipertensi.

Tabel 3. Tabel Hasil Uji Parsial Data Penderita Penyakit Hipertensi

Variabel	p-value
Jenis Kelamin	0,009
Usia	0,000
Rokok yang Dihabiskan per hari	0,949
Tekanan Darah Sistolik	0,000
Tekanan Darah Diastolik	0,000
BMI	0,000

Berdasarkan pada Tabel 3 diatas dapat dilihat bahwa *p-value* dari enam variabel yang mempengaruhi penyakit hipertensi hanya satu variabel yang nilainya tidak signifikan dengan $\alpha =$ 5% yaitu variabel rokok yang dihabiskan . Oleh karena itu dapat diartikan bahwa dengan tingkat kepercayaan sebesar 95% terdapat satu dari enam variabel yang tidak mempengaruhi penyakit hipertensi.

Uji Kesesuaian Dengan Modifikasi *Hosmer-Lemeshow Test* Pada Data Penderita Penyakit Hipertensi

Pada bagian ini akan dilakukan pengujian *Hosmer-Lemeshow Test* pada data penderita penyakit hipertensi di kota Framingham.

Hipotesis :

H_0 : Model yang digunakan sesuai dengan data.

H_1 : Model yang digunakan tidak sesuai dengan data.

Tabel 4. Hasil Perhitungan *Hosmer-Lemeshow Test* Pada Data Penderita Penyakit Hipertensi di Kota Framingham, Massachusettes

Pengujian	Derajat Bebas	Chi-Square
Hosmer-Lemeshow	8	38.63

Dari Tabel 4 dapat dilihat bahwa statistik uji *chi-square* diperoleh sebesar 38.63 dengan

derajat bebas sebesar 8. Dimana grup yang digunakan sebesar 10 dan dapat diperoleh nilai $x^2_{(g-2;1-\alpha)}$ adalah sebesar 2,733. Karena nilai statistik uji lebih besar daripada nilai tabel maka H_0 ditolak yang berarti bahwa model tidak sesuai dengan data.

Untuk memodifikasi *Hosmer-Lemeshow Test*, langkah selanjutnya yaitu menghitung dengan menggunakan hasil pada statistik uji *chi-square* di atas sebesar 38.63 , maka diperoleh nilai λ yang diestimasikan sebesar $\hat{\lambda} = 30.63$. Setelah mendapatkan nilai *chi-square* dan nilai parameter di atas, maka diperoleh nilai parameter yang telah di modifikasi sebesar $\lambda_c = 4.6097$. Karena nilai $\lambda_c \geq 0$ maka keputusan belum bisa diambil. Tahapan berikutnya adalah membangkitkan data (r) dari distribusi $x^2_{g-2,\lambda}$ dan diperoleh nilai sebesar 21.143. Dengan menggunakan taraf signifikansi sebesar 5% dengan derajat bebas 8, nilai kritis tabel untuk ukuran sampel 4.190 tidak ada, maka dilakukan interpolasi terhadap nilai z_0. Hasil interpolasi nilai kritis untuk ukuran sampel 4.190 yaitu 13.624. Karena nilai $r\ (21.143) > \chi_c(13.624)$ maka H_0 ditolak yang artinya model tidak cocok dengan data.

4. Kesimpulan

Berdasarkan penelitian yang telah dilakukan dan dibahas maka dapat disimpulkan bahwa model regresi logistik yang didapatkan yaitu : $\pi(\mathbf{x}) = \frac{e^{(-25{,}844+0{,}294\,X_1+0{,}0295\,X_2-0{,}0003\,X_3+\ 0{,}1082\,X_4+0{,}0812\,X_5+0{,}0674\,X_6)}}{1+e^{(-25{,}844+0{,}294\,X_1+0{,}0295\,X_2-0{,}0003\,X_3+\ 0{,}1082\,X_4+0{,}0812\,X_5+0{,}0674\,X_6)}}$ dengan hasil pengujian metode modifikasi *Hosmer-Lemeshow Test* menolak H_0 dimana $\boldsymbol{r\ (21.143) > \chi_c(13.624)}$ yang berarti bahwa model tidak cocok dengan data penderita penyakit hipertensi. Hasil dari uji *Hosmer-Lemeshow Test* juga menghasilkan kesimpulan yang sama, yaitu model yang digunakan tidak sesuai dengan data.

5. Saran

Kajian dalam penelitian ini adalah sampai pengujian kesesuain model dengan data, dan mencari model yang lebih cocok dengan data diluar cakupan dalam skripsi ini. Oleh karena itu disarankan untuk mencari model lain yang lebih cocok, dalam bentuk panambahan variabel bebas atau transformasi variabel bebas, seperti pengaruh kuadratik, logaritma, ataupun transformasi akar.

Daftar Pustaka

[1] Collett, D. (2003). *Modelling Binary Data.* Boca Raton London NewYork Washington, D.C: A CRC Press Company.

[2] Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression.* Canada: Jhon Wiley & Sons, Inc.

[3] Lai, X., & Liu, L. (2018). A simple test procedure in standardizing the power of Hosmer-Lemeshow test in large data sets. Journal of Statistical Computation and Simulation, 1.

[4] Li, P. (2015). *Logistic Regression.* New Brunswick United States: Rutgers University.

[5] Paul, P., Pennell, M. L., & Lemeshow, S. (2012). Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. Statistics in Medicine, 1-14.

[6] Yu, W., Xu, W., & Zhu, L. (2017). A Modified Hosmer-Lemeshow Test for Large Data Sets. *Communications in Statistics - Theory and Methods*, 15.