**Regresi Logistik**

Regresi logistik adalah sebuah pendekatan untuk membuat model prediksi seperti halnya regresi linear atau yang biasa disebut dengan istilah *Ordinary Least Squares (OLS) regression*. Perbedaannya adalah pada regresi logistik, peneliti memprediksi variabel terikat yang berskala dikotomi. Skala dikotomi yang dimaksud adalah skala data nominal dengan dua kategori, misalnya: Ya dan Tidak, Baik dan Buruk atau Tinggi dan Rendah.

Apabila pada OLS mewajibkan syarat atau asumsi bahwa error varians (residual) terdistribusi secara normal. Sebaliknya, pada regresi ini tidak dibutuhkan asumsi tersebut sebab pada regresi jenis logistik ini mengikuti distribusi logistik.

Asumsi Regresi Logistik

Asumsi Regresi Logistik antara lain:

1. Regresi logistik tidak membutuhkan hubungan linier antara variabel independen dengan variabel dependen.
2. Variabel independen tidak memerlukan asumsi *multivariate normality.*
3. Asumsi homokedastisitas tidak diperlukan
4. Variabel bebas tidak perlu diubah ke dalam bentuk metrik (interval atau skala ratio).
5. Variabel dependen harus bersifat dikotomi (2 kategori, misal: tinggi dan rendah atau baik dan buruk)
6. Variabel independen tidak harus memiliki keragaman yang sama antar kelompok variabel
7. Kategori dalam variabel independen harus terpisah satu sama lain atau bersifat eksklusif
8. Sampel yang diperlukan dalam jumlah relatif besar, minimum dibutuhkan hingga 50 sampel data untuk sebuah variabel prediktor (independen).
9. Dapat menyeleksi hubungan karena menggunakan pendekatan non linier log transformasi untuk memprediksi odds ratio. Odd dalam regresi logistik sering dinyatakan sebagai probabilitas.

Model Persamaan Regresi Logistik

Model persamaan aljabar layaknya OLS yang biasa kita gunakan adalah berikut: Y = B0 + B1X + e. Dimana e adalah error varians atau residual. Dengan model regresi ini, tidak menggunakan interpretasi yang sama seperti halnya persamaan regresi OLS. Model Persamaan yang terbentuk berbeda dengan persamaan OLS.

Berikut persamaannya:



Ln: Logaritma Natural.Di mana:

B0 + B1X: Persamaan yang biasa dikenal dalam OLS.

Sedangkan P Aksen adalah probabilitas logistik yang didapat rumus sebagai berikut:



Di mana:

exp atau ditulis “e” adalah fungsi exponen.

(Perlu diingat bahwa exponen merupakan kebalikan dari logaritma natural. Sedangkan logaritma natural adalah bentuk logaritma namun dengan nilai konstanta 2,71828182845904 atau biasa dibulatkan menjadi 2,72).

Dengan model persamaan di atas, tentunya akan sangat sulit untuk menginterprestasikan koefisien regresinya. Oleh karena itu maka diperkenalkanlah istilah Odds Ratio atau yang biasa disingkat Exp(B) atau OR. Exp(B) merupakan exponen dari koefisien regresi. Jadi misalkan nilai slope dari regresi adalah sebesar 0,80, maka Exp(B) dapat diperkirakan sebagai berikut:



Nilai Odds Ratio

Besarnya nilai Exp(B) dapat diartikan sebagai berikut:

Misalnya nilai Exp (B) pengaruh rokok terhadap terhadap kanker paru adalah sebesar 2,23, maka disimpulkan bahwa orang yang merokok lebih beresiko untuk mengalami kanker paru dibadningkan dengan orang yang tidak merokok. Interprestasi ini diartikan apabila pengkodean kategori pada tiap variabel sebagai berikut:

1. Variabel bebas adalah Rokok: Kode 0 untuk tidak merokok, kode 1 untuk merokok.
2. Variabel terikat adalah kanker Paru: Kode 0 untuk tidak mengalami kanker paru, kode 1 untuk mengalami kanker paru.

Pseudo R Square

Perbedaan lainnya yaitu pada regresi ini tidak ada nilai “R Square” untuk mengukur besarnya pengaruh simultan beberapa variabel bebas terhadap variabel terikat. Dalam regresi logistik dikenal istilah *Pseudo R Square*, yaitu nilai **R Square Semu** yang maksudnya sama atau identik dengan R Square pada OLS.

Jika pada OLS menggunakan uji F Anova untuk mengukur tingkat signifikansi dan seberapa baik model persamaan yang terbentuk, maka pada regresi ini menggunakan Nilai Chi-Square. Perhitungan nilai Chi-Square ini berdasarkan perhitungan Maximum Likelihood.

Tahap Analisis Regresi Logistik

Kemudian pada menu, klik **Analyze -> Regression -> Binary Logistic**. Kemudian masukkan variabel terikat ke kotak **dependent** dan masukkan semua variabel bebas ke kotak **Covariates**.



Save Regresi Logistik

Tekan tombol **Save** lalu centang **Probabilities, Group membership, Unstandardized dan Studentized**kemudian klik **Continue**.



Tekan tombol **Options** lalu centang **Classification plots, Hosmer-lemeshow goodness-of-fit, Casewise listing residuals**dan pilih **Outliers outside**dan isi dengan angka **2, Correlation of estimates, Iteration history, CI for exp(B)**dan isi dengan **95**.

Opsi Regresi Logistik

Sedangkan nilai *Maximum iteration* biarkan tetap sebesar 20 dan nilai *Classification Cutoff*tetap 0.5. Nilai ini disebut dengan *the cut value* atau *prior probability*, yaitu peluang suatu observasi untuk masuk ke dalam salah satu kelompok sebelum karakteristik variabel penjelasnya diketahui. Apabila kita tidak mempunyai informasi tambahan tentang data kita, maka bisa langsung menggunakan nilai default yaitu 0,5. Jika tidak ada penelitian sebelumnya, dapat digunakan *classification cutoff* sebesar 0,5. Namun, jika ada penelitian lain yang telah meneliti maka bisa dinaikkan/diturunkan *classification cutoff* sesuai hasil penelitian.



Kemudian pada jendela utama, klik **OK** dan segera lihat Output anda

Interprestasi Regresi Logistik dengan SPSS

Setelah anda mempelajari tutorial [Regresi Logistik dengan SPSS](https://www.statistikian.com/2015/02/regresi-logistik-dengan-spss.html), maka saatnya kita belajar interprestasi regresi logistik dengan SPSS. Langsung saja anda buka output yang dihasilkan!



Di atas pada tabel *Case Processing Summary*adalah ringkasan jumlah sampel, yaitu sebanyak 200 sampel.



Di atas adalah kode variabel dependen. Yaitu kategori “Tidak Mengalami Kanker” dengan kode 0 dan “Mengalami Kanker” dengan kode 1. Oleh karena yang diberi kode 1 adalah “Mengalami Kanker”, maka “Mengalami Kanker” menjadi referensi atau efek dari sebab. Sebab yang dimaksud adalah kejadian yang dihipotesiskan sebagai penyebab munculnya efek atau masalah. Dalam hal ini, merokok (kode 1) dan ada riwayat keluarga (kode 1) menjadi sebab yang dapat meningkatkan resiko terjadinya kanker paru (kode 1).

Iteration History: Interprestasi Regresi Logistik dengan SPSS

Iteration History



Di atas: Tabel Iteration History pada block 0 atau saat variabel independen tidak dimasukkan dalam model: N=200 mendapatkan Nilai -2 Log Likelihood: 276,939.

Degree of Freedom (DF) = N – 1 = 200-1=199. Chi-Square (X2) Tabel Pada DF 199 dan Probabilitas 0.05 = 232,912.

Nilai -2 Log Likelihood (276,939) > X2 tabel (232,912) sehingga menolak H0, maka menunjukkan bahwa model sebelum memasukkan variabel independen adalah TIDAK FIT dengan data.

Classification Table: Interprestasi Regresi Logistik dengan SPSS

Classification Table



Di atas pada tabel Classifacation Table: Merupakan tabel kontingensi 2 x 2 yang seharusnya terjadi atau disebut juga frekuensi harapan berdasarkan data empiris variabel dependen, di mana jumlah sampel yang memiliki kategori variabel dependen referensi atau akibat buruk (kode 1) yaitu “Mengalami Kanker” sebanyak 104. Sedangkan yang “Tidak mengalami Kanker” sebanyak 96 orang.  Jumlah sampel sebanyak 200 orang. Sehingga nilai overall percentage sebelum variabel independen dimasukkan ke dalam model sebesar: 104/200 = 52,0%.

Variabel Dalam Persamaan Tahap Beginning: Interprestasi Regresi Logistik dengan SPSS

Variabel Dalam Persamaan Tahap Beginning



Di atas pada tabel *Variables in The Equation*: Saat sebelum var independen di masukkan ke dalam model, maka belum ada variabel independen di dalam model. Nilai Slope atau Koefisien Beta (B) dari Konstanta adalah sebesar 0,080 dengan Odds Ratio atau Exp(B) sebesar 1,083. Nilai Signifikansi atau p value dari uji Wald sebesar 0,572.

Perlu diingat bahwa nilai B identik dengan koefisien beta pada *Ordinary Least Square (OLS)*atau regresi linear. Sedangkan Uji Wald identik dengan t parsial pada OLS. Sedangkan Exp(B) adalah nilai eksponen dari B, maka Exp(0,080) = 1,083.

Variabel Tidak Dalam Persamaan Tahap Beginning: Interprestasi Regresi Logistik dengan SPSS

Variabel Tidak Dalam Persamaan Tahap Beginning



Di atas pada tabel *Variables not in the Equation*: Menunjukkan variabel yang belum dimasukkan ke dalam model regresi, yaitu variabel X1 dan X2. Di mana X1 adalah variabel merokok dan X2 adalah variabel riwayat keluarga.

Tahap Entry Variabel: Interprestasi Regresi Logistik dengan SPSS

Tahap Entry Variabel



Di atas pada tabel Iteration history Block 1 atau saat variabel independen dimasukkan dalam model: N=200. Degree of Freedom (DF) = N – jumlah variabel independen – 1 = 200-2-1=197. Chi-Square (X2) Tabel Pada DF 197 dan Prob 0.05 = 230,746.

Nilai -2 Log Likelihood (207,575) < X2 tabel (230,746) sehingga menerima H0, maka menunjukkan bahwa model dengan memasukkan variabel independen adalah FIT dengan data. Hal ini berbeda dengan *Block Beginning*di atas, di mana saat sebelum variabel independen dimasukkan ke dalam model, model TIDAK FIT dengan data.

Hasil Omnibus Test



Nilai X2 69,394 > X2 tabel pada DF 2 (jumlah variabel independen 2) yaitu 5,991 atau dengan signifikansi sebesar 0,000 (< 0,05) sehingga menolak H0, yang menunjukkan bahwa penambahan variabel independen DAPAT memberikan pengaruh nyata terhadap model, atau dengan kata lain model dinyatakan FIT.

Interprestasi Regresi Logistik dengan SPSS: Jawaban Hipotesis

Perlu diingat jika pada *OLS* untuk menguji signifikansi simultan menggunakan uji F, sedangkan pada regresi logistik menggunakan nilai *Chi-Square* dari selisih antara *-2 Log likelihood*sebelum variabel independen masuk model dan *-2 Log likelihood* setelah variabel independen masuk model. Pengujian ini disebut juga dengan pengujian *Maximum likelihood*.

Sehingga jawaban terhadap hipotesis pengaruh simultan variabel independen terhadap variabel dependen adalah menerima H1 dan menolak H0 atau yang berarti ada pengaruh signifikan secara simultan merokok dan riwayat keluarga terhadap kejadian kanker paru oleh karena nilai p value Chi-Square sebesar 0,000 di mana < Alpha 0,05 atau nilai Chi-Square Hitung 69,364 > Chi-Square tabel 5,991.

Pseudo R Square



Di atas pada tabel *Model Summary*: Untuk melihat kemampuan variabel independen dalam menjelaskan variabel dependen, digunakan nilai *Cox & Snell R Square* dan *Nagelkerke R Square.* Nilai-nilai tersebut disebut juga dengan *Pseudo R-Square* atau jika pada regresi linear (OLS) lebih dikenal dengan istilah *R-Square*.

Nilai *Nagelkerke R Square* sebesar 0,391 dan *Cox & Snell R Square* 0,293, yang menunjukkan bahwa kemampuan variabel independen dalam menjelaskan variabel dependen adalah sebesar 0,391 atau 39,1% dan terdapat 100% – 39,1% = 60,9% faktor lain di luar model yang menjelaskan variabel dependen.

Hosmer and Lemeshow Test

Hosmer and Lemeshow Test adalah uji *Goodness of fit test (GoF)*, yaitu uji untuk menentukan apakah model yang dibentuk sudah tepat atau tidak. Dikatakan tepat apabila tidak ada perbedaan signifikan antara model dengan nilai observasinya.



Nilai *Chi Square* tabel untuk DF 1 (Jumlah variabel independen – 1) pada taraf signifikansi 0,05 adalah sebesar 3,841. Karena nilai *Chi Square Hosmer and Lemeshow* hitung 13,671 > Chi Square table 3,841 atau nilai signifikansi sebesar 0,000 (< 0,05) sehingga menolak H0, yang menunjukkan bahwa model TIDAK dapat diterima dan pengujian hipotesis TIDAK dapat dilakukan sebab ada perbedaan signifikan antara model dengan nilai observasinya.

Oleh karena dalam tutorial Interprestasi regresi logistik dengan SPSS ini, nilai Hosmer and Lemeshow Test menolak H0, sebaiknya anda mencoba untuk membuat agar nilai Hosmer and Lemeshow Test menerima H0.

Cara Mengatasi Hosmer Lemeshow

Caranya adalah dengan mengubah model persamaan regresi logistik dengan menambahkan variabel interaksi antar variabel independen. Misal pada kasus di sini, dengan menambahkan variabel interaksi antara X1 dan X2.

Lebih jelasnya akan saya bahas pada bagian akhir dalam artikel ini yaitu pada bagian VARIABEL INTERAKSI. Sehingga anda untuk sementara bisa melanjutkan pembelajaran ini, seolah-olah hasil uji Hosmer and Lemeshow Test menerima H0.

Classification Result



Berdasarkan tabel *Classification Table* di atas, jumlah sampel yang tidak mengalami kanker 78 + 18 = 96 orang. Yang benar-benar tidak mengalami kanker sebanyak 78 orang dan yang seharusnya tidak mengalami kanker namun mengalami, sebanyak 18 orang. Jumlah sampel yang mengalami kanker 27 + 77 = 104 orang. Yang benar-benar mengalami kanker sebanyak 27 orang dan yang seharusnya mengalami kanker namun tidak mengalami, sebanyak 77 orang.

Dalam Interprestasi regresi logistik dengan SPSS: Tabel di atas memberikan nilai *overall percentage* sebesar (78+77)/200 = 77,5% yang berarti ketepatan model penelitian ini adalah sebesar 77,5%.

Pendugaan Parameter

Pendugaan Parameter



Lihat tabel *Variabel in the equation* di atas: semua variabel independen nilai P value uji wald (Sig) < 0,05, artinya masing-masing variabel mempunyai pengaruh parsial yang signifikan terhadap Y di dalam model. X1 atau merokok mempunyai nilai Sig Wald 0,000 < 0,05 sehingga menolak H0 atau yang berarti merokok memberikan pengaruh parsial yang signifikan terhadap kejadian kanker paru. X2 atau riwayat keluarga mempunyai nilai Sig Wald 0,031 < 0,05 sehingga menolak H0 atau yang berarti riwayat keluarga memberikan pengaruh parsial yang signifikan terhadap kejadian kanker paru.

Odds Ratio

Besarnya pengaruh ditunjukkan dengan nilai EXP (B) atau disebut juga ODDS RATIO (OR). Variabel Merokok dengan OR 6,277 maka orang yang merokok (kode 1 variabel independen), lebih beresiko mengalami kanker paru (kode 1 variabel dependen) sebanyak 6,277 kali lipat di bandingkan orang yang tidak merokok (kode 0 variabel independen). Nilai B = Logaritma Natural dari 6,277 = 1,837. Oleh karena nilai B bernilai positif, maka merokok mempunyai hubungan positif dengan kejadian kanker.

Variabel Riwayat Keluarga dengan OR 2,645 maka orang yang ada riwayat keluarga (kode 1 variabel independen), lebih beresiko mengalami kanker paru (kode 1 variabel dependen) sebanyak 2,645 kali lipat di bandingkan orang yang tidak ada riwayat keluarga (kode 0 variabel independen). Nilai B = Logaritma Natural dari 2,645 = 0,973. Oleh karena nilai B bernilai positif, maka riwayat keluarga mempunyai hubungan positif dengan kejadian kanker.

Persamaan Regresi Logistik

Berdasarkan nilai-nilai B pada perhitungan di atas, maka model persamaan yang dibentuk adalah sebagai berikut

Ln P/1-P = -1,214 + 1,837 Rokok + 0,973 Riwayat Keluarga.

Atau bisa menggunakan rumus turunan dari persamaan di atas, yaitu:

Probabilitas = exp(-1,214 + 1,837 Rokok + 0,973 Riwayat Keluarga) / 1 + exp(-1,214 + 1,837 Rokok + 0,973 Riwayat Keluarga).

Contoh Penggunaan Persamaan Regresi Logistik

Misalkan sampel yang merokok dan ada riwayat keluarga, maka merokok=1 dan riwayat keluarga=1. Jika dimasukkan ke dalam model persamaan di atas, maka sebagai berikut:

Probabilitas atau *Predicted* = (exp(-1,214 + (1,837 x 1) + (0,973 x 1))) / (1 + exp(-1,214 + (1,837 x 1) + (0,973 x 1))).

Probabilitas atau *Predicted* = (exp(-1,214 + (1,837) + (0,973))) / (1 + exp(-1,214 + (1,837) + (0,973))).

Probabilitas atau *Predicted* = 0.8315.

Nilai Predicted Regresi Logistik

Oleh karena *Predicted*: 0,8315 > 0,5 maka nilai *Predicted Group Membership*dari sampel di atas adalah 1. Di mana 1 adalah kode mengalami kanker. Jadi jika sampel merokok (kode 1) dan ada riwayat keluarga (kode 1) maka prediksinya adalah mengalami kanker (kode 1). Jika seandainya sampel yang bersangkutan ternyata faktanya tidak mengalami kanker (kode 0) maka sampel tersebut keluar dari nilai prediksi.

Besarnya perbedaan atau yang disebut dengan *Residual* = *Predicted Group Membership – Predicted.*Pada kasus di atas di mana orang yang merokok dan ada riwayat keluarga namun faktanya tidak mengalami kanker, maka *Residual* = 0 – 0,8315 = -0,8315.



Di atas adalah grafik klasifikasi dari *observed group*dan *predicted group membership*.

Outlier

Nilai Outlier



Agar anda paham Interprestasi Regresi Logistik dengan SPSS, maka perhatikan output di atas, yaitu output hasil deteksi outlier atau data pencilan. Pada kasus dalam tutorial ini tidak ada outlier dengan notifikasi seperti di atas, yaitu: *The casewise plot is not produced because no outliers were found*. Seandainya ada outlier, maka tampilan akan berubah dalam bentuk tabel yang berisi daftar sampel yang menjadi outlier beserta nilai *Studentized Residual*.Dinyatakan outlier apabila nilai Absolut dari *Studentized Residual*dari sampel tersebut lebih dari 2 (> 2). Di mana *Studentized Residual*adalah standarisasi berdasarkan*Mean*dan *standart deviasi*dari nilai *Residual* yang telah dibahas di atas.

Variabel Interaksi

Variabel interaksi adalah variabel yang merupakan hasil interaksi antar variabel independen. Variabel ini ada kalanya diperlukan untuk dimasukkan ke dalam model regresi logistik dengan alasan karena secara substantif memang ada hubungan antar variabel independen atau untuk mengatasi masalah *Goodness of Fit Test* di mana menolak H0.

Caranya adalah dengan menambahkan variabel interaksi ke dalam variabel independen, yaitu pada saat memasukkan variabel independen, seleksi variabel-variabel independen yang akan diinteraksikan, kemudian klik tombol “‘>a\*b>'”.



Selanjutnya lakukan proses seperti yang sudah dijelaskan di atas