Phenomenon , 2017, Vol. 07 (No. 2), pp. 137-163
JURNAL PHENOMENON
http://phenomenon@walisongo.ac.id

REGRESI LOGISTIK: MENAKSIR PROBABILITAS PERISTIWA VARIABEL BINARI

Ibnu Hadjar[1]
[1]*Fakultas Ilmu Tarbiyah dan Keguruan, UIN Walisongo, Semarang*

Abstract

Prediction through a statistical modeling is a major strength of linear regression analysis. However, this technique requires that the dependent variable or criterion must have continuum data and a normal score distribution so that it can not be used if the criterion variable is a binary variable (nominal in two categories, having a score of 1 or 0). Logistic regression is an appropriate technique to replace it. Instead of generating estimates of criterion variable scores, modeling with logistic regression yields probability estimates of event occurrence (score 1) on the criterion variable. The estimate of probability is obtained through a regression model that yields odds and log odds, which is basically a different way of disclosure of the same fact with respect to probability estimates of the event occurrence of all events in the criterion variable.

Keywords: logistic regression, binary variables, odds, log odds, logit.

Abstrak

Prediksi melalui suatu pemodelan statistic merupakan kekuatan utama dari analisis regresi linier. Akan tetapi, teknik ini mensyaratkan variabel dependen atau kriteria harus memiliki data kontinum dan sebaran skor yang normal sehingga ia tidak dapat digunakan jika variabel kriterianya merupakan variabel binary (nominal dengan dua kategori, memiliki skor 1 atau 0). Regresi logistik merupakan teknik yang tepat untuk menggantikannya. Alih-alih menghasilkan taksiran skor variabel kriteria, pemodelan dengan regresi logistik menghasilkan taksiran probabilitas munculnya peristiwa (skor 1) pada variabel kriteria. Taksiran probabilitas tersebut diperoleh melalui model regresi yang menghasilkan nilai odds dan log odds, yang pada dasarnya merupakan cara pengungkapan yang berbeda tentang kenyataan yang sama terkait taksiran probabilitas munculnya peristiwa dari seluruh kejadian dalam variabel kriteria.

Kata kunci: regresi logistik, variabel binary, odds, log odds, logit.

PENDAHULUAN

Sebagai salah satu pemodelan statistik, analisis regresi merupakan teknik statistik yang banyak digunakan dalam penelitian untuk menaksir hubungan antar variabel. Fokus analisis ini adalah hubungan antar satu variabel dependen (juga disebut variabel

Email: ibnu_hadjar@walisongo.ac.id

ISSN: 2088-7868, e-ISSN 2502–5708

luaran [*output*], respon, atau kriteria) dengan satu atau lebih variabel independen (juga disebut variabel eksplanatoris atau prediktor). Teknik analisis ini membantu peneliti memahami bagaimana skor variabel dependen berubah sejalan dengan perubahan skor variabel independennya (Pedhazur, 1982).

Berdasarkan pengamatan penulis, hampir semua penelitian yang menggunakan regresi yang dilaporkan di jurnal ilmiah di Indonesia memakai teknik regresi linier, yang melibatkan variabel dengan skor kontinum untuk variabel dependen/output/kriterianya. Teknik regresi ini mengasumsikan bahwa skor variabel kriteria atau dependen tersebar secara normal, memiliki hubugan linier, dan nilai varian sama lintas kelompok variabel independen atau prediktornya (Glass & Hopkins, 1984). Asumsi ini tidak dapat dipenuhi jika variabel kriterianya berupa *variabel binari*, yakni variabel yang hanya memiliki dua kemungkinan nilai, biasanya menggunakan nilai 0 dan 1 (Upto & Cook, 2011), misalnya: kelulusan (lulus = 1; tidak lulus = 0), kehadiran (hadir = 1; absen = 0), berdzikir (melakukan = 1; tidak melakukan = 0).

Dalam rangka mengatasi hal ini, para pakar statistik mengembangkan alternatif model analisis yang disebut *regresi logistik* atau *analisis logit* (Hair dkk., 1995; Reed & Wu, 2013). Regresi logistik merupakan teknik untuk pemodelan probabilitas terjadinya peristiwa dari sisi kesesuaiannya (Loh, 2006). Model ini menggunakan probabilitas linier sehingga memberikan rentangan teknik diagnostik dan penjelasan untuk variabel dependen non-parametrik, khususnya pengukuran binari (Askar, Usluel, & Mumcu, 2006). Model logit ini banyak digunakan karena adanya kesamaan dengan teknik analisis regresi linier, terutama terkait dengan uji statistiknya, kemampuannya memadukan efek non linier dan diagnostik. Model ini memiliki banyak hasil analisis yang serupa dengan regresi, tetapi menggunakan metode koefisien dan penaksiran yang berbeda. Alih-alih meminimalkan penyimpangan kuadrat (kuadrat terkecil atau *least squares*), model logit memaksimalkan kemungkinan bahwa suatu peristiwa akan terjadi.

Karena karakteristiknya tersebut, model regresi logistik ini mulai banyak digunakan dalam penelitian yang diterbitkan dalam jurnal ilmiah internasional, khususnya bila *outcome*-nya berupa variabel binari (Reed & Wu, 2013). Namun demikian, penggunaan model dalam kepustakaan berbahasa Indonesia masih sangat terbatas. Hal ini di antaranya karena masih sangat terbatasnya sumber pustaka yang membahas model regresi logistik ini. Artikel ini bertujuan untuk menyajikan

pembahasan secara ringkas model logistik ini. Kajian ini akan difokuskan pada konsep dasarnya disertai dengan contoh penggunaannya. Karena keterbatasan, pembahasan dibatasi pada model yang melibatkan satu variabel independen saja, baik yang binari maupun kontinum.

METODE PENELITIAN

Penelitian ini termasuk penelitian kepustakaan terhadap buku-buku referensi yang bersifat pengembangan atau implementasi teori yang telah ada, dan relevansinya dengan perkembangan zaman sekarang. Penelitian kepustakaan juga sering disebut dengan istilah penelitian Kepustakaan (Library Research). menurut Noeng Muhadjir, penelitian kepustakaan itu lebih memerlukan olahan filosofis dan teoritis daripada uji empiris dilapangan (Noeng Muhadjir, 1996:169). Karena sifatnya teoritis dan filosofis , penelitian kepustakaan ini sering menggunakan pendekatan filosofis (philosophical approach) daripada pendekatan yang lain. Metode penelitiannya mencakup sumber data, pengumpulan data, dan analisis data.

PEMBAHASAN

BEBERAPA KONSEP TERKAIT

Meskipun ada kesamaan dengan regresi linier dalam hal prediksi, regresi logistik menggunakan konsep dan istilah yang berbeda dan sekaligus penafsiran yang berbeda pula (Garson, 200). Untuk memudahkan pemahaman, terlebih dahulu akan dibahas konsep-konsep terkait, di antaranya adalah *probabilitas, odds, log odds,* dan *rasio odds.* Untuk memahami konsep-konsep tersebut, contoh-contoh penghitungan akan mengacu pada data (fiktif) pendaftar seleksi calon mahasiswa baru Universitas Tugu Muda (UTM), yang ringkasannya disajikan dalam suatu tabel kontingensi (Press, & Wilson, 1978), sebagaimana dalam tabel 1. Dalam contoh ini, sebagai variabel luaran/*output* (dependen/kriteria) adalah hasil seleksi calon mahasiswa baru (variabel binari, dengan skor kategori: Gagal = 0 dan Sukses/lulus diterima = 1). Sedangkan sebagai variabel masukan/input (independen/prediktor) adalah Reputasi Asal Sekolah (variabel binari, dengan skor kategori: Reguler = 0, Unggulan = 1) dan Nilai Ujian Nasional/NUN (kontinum, dengan skor merentang dari 0 – 60). Tabel 1 menyajikan kontingensi dua jalur yang menyajikan rangkuman data kelompok utama (Hasil Seleksi dan Reputasi Asal Sekolah) dan interaksi (antar keduanya) berikut ini.

Tabel 1. Data pendaftar calon mahasiswa baru UTM berdasarkan hasil seleksi dan Reputasi asal sekolah

Hasil Seleksi/Y	Reputasi Asal Sekolah/X		Total
	Reguler (X=0)	Unggulan (X=1)	
Gagal (Y=0)	$n_{(X=0;Y=0)} = 29$	$n_{(X=1;Y=0)} = 11$	$n_{(Y=0)} = 40$
Sukses (Y=1)	$n_{(X=0;Y=1)} = 31$	$n_{(X=1;Y=1)} = 29$	$n_{(Y=1)} = 60$
Total	$n_{(X=0)} = 60$	$n_{(X=1)} = 40$	n = 100

Probabilitas

Regresi logistik digunakan untuk menaksir pengaruh relatif variabel independen (eksplanatoris) pada variabel dependen (luaran/*output*). Oleh karena itu pembahasannya diawali dari konsep probabilitas (*p*), yakni *proporsi munculnya suatu peristiwa dari seluruh kejadian* (Glass & Hopkins, 1984). Nilai probabilitas munculnya peristiwa, $p_{(Y=1)}$, diperoleh dari jumlah munculnya peristiwa, $n_{(Y=1)}$, dibagi jumlah seluruh kejadian, n, sehingga dapat dirumuskan sebagai berikut:

$$Probabilitas\ peristiwa = p_{(Y=1)} = \frac{jumlah\ peristiwa}{jumlah\ seluruh\ kejadian} = \frac{n_{(Y=1)}}{n}$$

Dari tabel 1, sebagai contoh, diketahui bahwa dari 100 pendaftar/n (jumlah seluruh kejadian), 60 di antaranya diterima atau sukses/$n_{(Y=1)}$ (jumlah munculnya peristiwa sukses). Karena itu, probabilitas sukses adalah:

$$p_{(Y=1)} = \frac{n_{(Y=1)}}{n} = \frac{60}{100} = 0{,}6$$

Sedangkan probabilitas tidak munculnya peristiwa tidak diterima atau gagal, $p_{(Y=0)}$, adalah 1 dikurangi probabilitas munculnya peristiwa, atau:

$$p_{(Y=0)} = 1 - p_{(Y=1)} = 1 - 0{,}6 = 0{,}4$$

Lebih lanjut, berdasarkan reputasi asal sekolah, X, probabilitas sukses peserta yang berasal dari sekolah regular, $p_{(X=0;Y=1)}$, adalah jumlah peserta sukses dari sekolah tersebut ($n_{[X=0;Y=1]}$ = 31) dibagi jumlah peserta dari sekolah tersebut ($n_{[X=0]}$ = 60), sehingga:

$$p_{(X=0;Y=1)} = \frac{n_{(X=0;Y=1)}}{n_{(X=0)}} = \frac{31}{60} = 0{,}517.$$

Sedangkan probabilitas sukses peserta yang berasal dari sekolah unggulan/$p_{(X=1;Y=1)}$ adalah jumlah peserta sukses dari sekolah unggulan ($n_{[X=1;Y=1]}$ = 29) dibagi jumlah peserta dari sekolah tersebut ($n_{[X=1]}$ = 40), sehingga:

$$p_{(X=1;Y=1)} = \frac{n_{(X=1;Y=1)}}{p_{(X=1)}} = \frac{29}{40} = 0{,}725.$$

Dengan demikian, probabilitas sukses (munculnya peristiwa) berbanding terbalik

dengan probabilitas gagal (tidak munculnya peristiwa). Semakin tinggi probabilitas sukses, semakin rendah probabilitas gagal, dan sebaliknya. Nilai probabilitas tersebut dapat merentang dari 0 (diperoleh ketika *tidak ada* peristiwa yang muncul dalam seluruh kejadian) sampai 1 (diperoleh ketika *jumlah* peristiwa yang muncul sama dengan seluruh kejadian atau *tidak ada peristiwa yang tidak muncul* dari seluruh kejadian).

Odds

Jika probabilitas berkenaan dengan peristiwa yang terjadi, *odds* merupakan rerata jumlah peristiwa yang diharapkan terjadi untuk setiap tidak munculnya peristiwa. Dalam contoh pendaftaran masuk calon mahasiswa UTM, odds merupakan rerata jumlah pendaftar yang diharapkan sukses untuk setiap pendaftar yang gagal diterima. Odds (*O*), merupakan rasio probabilitas munculnya peristiwa ($p_{[Y=1]}$) dan probabilitas tidak munculnya peristiwa dari seluruh kejadian ($p_{[Y=0]}$) (Upton & Cook, 2011) atau:

$$Odds = O = \frac{Proporsi\ Munculnya\ Peristiwa}{Proporsi\ Tidak\ Munculnya\ Peristiwa} = \frac{p_{(Y=1)}}{p_{(Y=0)}} = \frac{p_{(Y=1)}}{1 - p_{(Y=1)}}$$

Dalam kasus pendaftar calon mahasiswa baru tersebut, odds sukses (pendaftar yang diterima sebagai calon mahasiswa baru) adalah:

$$O = \frac{p_{(Y=1)}}{1 - p_{(Y=1)}} = \frac{0{,}6}{1 - 0{,}6} = \frac{0{,}6}{0{,}4} = 1{,}5$$

Berdasarkan hasil tersebut, odds sukses adalah 1,5; yang berarti bahwa probabilitas atau kemungkinan pendaftar ujian masuk calon mahasiswa baru UTM yang sukes (pendaftar diterima) adalah 1,5 kali atau 50 persen lebih besar dari kemungkinan gagal (tidak diterima). Dengan kata lain, setiap 1 pendaftar yang gagal, ada rata-rata ada 1,5 pendaftar yang sukses.

Contoh lebih lanjut, berdasarkan reputasi asal sekolah, odds sukses peserta seleksi yang berasal dari sekolah reguler ($n_{[X=0;Y=1]}$ = 60; $p_{[X=0;Y=1]}$ = 0,517) adalah:

$$O = \frac{p_{(X=0;Y=1)}}{1-p_{(X=0;Y=1)}} = \frac{0{,}517}{1-0{,}517} = \frac{0{,}517}{0{,}483} = 1{,}07.$$

Sedangkan rerata peluang sukses yang diharapkan untuk peserta yang berasal dari sekolah dengan reputasi unggulan ($n_{[X=1;Y=1]}$= 40; $p_{[X=1;Y=1]}$ = 0,725) adalah:

$$O = \frac{p_{(X=1;Y=1)}}{1-p_{(X=1;Y=1)}} = \frac{0{,}725}{1-0{,}725} = \frac{0{,}725}{0{,}275} = 2{,}636.$$

Hasil ini menunjukkan bahwa rerata peluang sukses/diterima yang diharapkan untuk pendaftar dari sekolah reguler adalah 1,07 atau 7 persen lebih tinggi dari pada rerata

peluang gagal peserta dari sekolah tersebut. Dengan kata lain, untuk setiap pendaftar yang gagal diharapkan terdapat 1,07 pendaftar yang sukses diterima sebagai calon mahasiswa di UTM. Sedangkan probabilitas sukses/diterima pendaftar calon mahasiswa baru yang berasal dari sekolah unggulan adalah 2,636 kali atau 163,6 persen lebih tinggi dari pada rerata peluang gagal peserta yang berasal dari sekolah dengan reputasi yang sama (unggulan). Dengan kata lain, setiap 1 pendaftar yang gagal diharapkan terdapat 2,636 pendaftar yang sukses.

Rasio Odds

Untuk membandingkan rerata peluang munculnya peristiwa yang diharapkan dari dua peristiwa yang berbeda kondisi digunakan *Rasio odds* atau *odds ratio* (Minitab, 2016), disingkat *RO*. RO merupakan pengukuran kekuatan hubungan antar dua variabel binary (Hailpern, & Visintainer, 2003). Rasio tersebut merupakan odds dari peristiwa tertentu dalam kondisi tertentu dibagi odds dari peristiwa yang sama dalam kondisi yang berbeda. Rasio ini menentukan apakah suatu kelompok memiliki peluang yang lebih atau kurang dibandingkan kelompok lain. *RO* antara kelompok 1 dan kelompok 2 dapat dihitung dengan rumus berikut:

$$Rasio\ Odds = RO = \frac{Odds\ 1}{Odds\ 2} = \frac{O_1}{O_2}$$

Dari contoh penghitungan Odss sebelumnya diketahui bahwa Odss sukses pendaftar dari sekolah regular ($X = 0; Y = 1$), sebagai kelompok 1, adalah O_1 =1,08 dan Odss sukses pendaftar dari sekolah unggulan (($X = 0; Y = 1$), sebagai kelompok 2, adalah O_2= 2,70. Karena itu nilai rasio odds sukses antar kedua kelompok reputasi asal sekolah (reguler dibanding unggulan) adalah:

$$RO = \frac{O_1}{O_2} = \frac{O_{(X=0;Y=1)}}{O_{(X=1;Y=1)}} = \frac{1{,}08}{2{,}70} = 0{,}4.$$

Dengan nilai *RO* ini dapat disimpulkan bahwa rerata peluang sukses pedaftar calon mahasiswa baru yang berasal dari sekolah dengan reputasi reguler 0,4 kali atau 40 persen dari pada rerata peluang pedaftar yang berasal dari sekolah dengan reputasi unggulan. Hal ini berarti bahwa setiap 1 pendaftar sekolah unggulan yang sukses, rata-rata terdapat 0,4 pendaftar dari sekolah reguler yang sukses. Jika penghitungan dibalik (unggulan dibanding reguler), maka:

$$RO = \frac{O_2}{O_1} = \frac{O_{(X=1;Y=1)}}{O_{(X=0;Y=1)}} = \frac{2{,}70}{1{,}08} = 2{,}497$$

Dengan nilai *RO* ini dapat disimpulkan bahwa rerata peluang sukses pedaftar calon mahasiswa baru yang berasal dari sekolah dengan reputasi unggulan 2,497 ~ 2,5 kali atau 149,7 persen lebih besar dari pada rerata peluang sukses pedaftar yang berasal dari sekolah dengan reputasi reguler. Hal ini berarti bahwa setiap 1 pendaftar dari sekolah reguler yang sukses, rata-rata terdapat 2,5 pedaftar dari sekolah dengan reputasi unggulan yang sukses. Hasil ini sama dengan sebelumnya, hanya cara pengungkapannya yang berbeda. Dengan demikian, semakin besar nilai *RO*, semakin besar hubungan antara kedua variabel, dan sebaliknya.

Log Odds

Log odds, juga disebut *logit* **(Gullickson, 2005),** merupakan cara alternatif untuk mengungkapkan probabilitas. Log Odds (disingkat LO) merupakan logaritma natural dari Odds, sehingga dapat dirumuskan:

$$LO = \ln(O)$$

Di mana *LO* adalah nilai log odds, *ln* adalah logaritma natural, yaitu logaritma dengan basis nilai *e* (eksponen, yakni nilai konstan yang irasional dan transendental, dengan nilai sekitar 2,718281828459~2,71828 [Oty & Elliott, 2015]), karena itu juga sering ditulis sebagai $\log_e$ (Togneti, 1998), dan *O* adalah nilai odds.

Dari contoh penghitungan sebelumnya diketahui bahwa nilai odds sukses pendaftar seleksi calon mahasiswa baru di UTM adalah $O_{(Y=1)}$ = 1,5, odds sukses pendaftar dari sekolah reguler adalah $O_{(X=0;Y=1)}$ = 1,083 dan odds sukses pendaftar dari sekolah unggulan adalah $O_{(X=1;Y=1)}$ = 2,704. Karena itu, nilai log odds masing-masing secara berturut-turut adalah:

$$LO_{(Y=1)} = \ln(O_{[Y=1]}) = log_{2{,}71828}(1{,}5) \ = 0{,}405.$$
$$LO_{(X=0;Y=1)} = \ln(O_{[X=0Y=1]}) = log_{2{,}71828}(1{,}083) = 0{,}08.$$
$$LO_{(X=1;Y=1)} = \ln(O_{[X=1Y=1]} = log_{2{,}71828}(2{,}704) = 0{,}995.$$

Contoh lain, jika tiga nilai odds adalah *O* = 0,25; *O* = 1; dan *O* = 4 maka nilai *LO* masing-masing adalah:

$$LO = \ln(O) = log_{2{,}71828}(0{,}25) = -1{,}3683.$$
$$LO = \ln(O) = log_{2{,}71828}(1) \quad = 0.$$
$$LO = \ln(O) = log_{2{,}71828}(4) \quad = 1{,}386.$$

Dari contoh tersebut, tranformasi odds yang bernilai lebih kecil dari 1 (*O*<1) akan

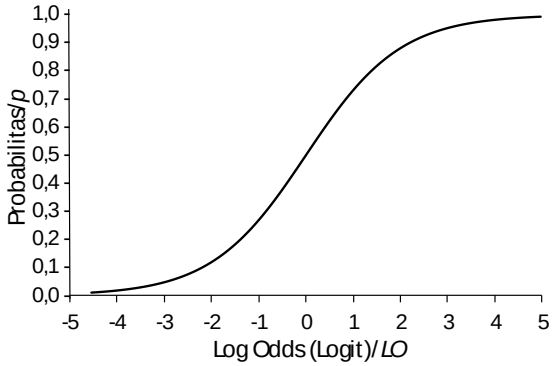
memperoleh nilai log odds negatif (lebih kecil dari 0, $LO<0$), yang bernilai lebih besar dari 1 ($O>1$) akan memperoleh nilai log odds positif (lebih besar dari 0, $LO>0$). Sedangkan jika nilai odds sama dengan 1 ($O = 1$), maka nilai log odds sama dengan 0 ($LO = 0$). Transformasi dari odds ke log odds merupakan transformasi yang bersifat monotonik, di mana semakin besar odds, semakin besar log odds, dan sebaliknya.

Karena nilai odds merupakan transformasi dari nilai probabilitas, maka nilai ketiga konsep tersebut juga berhubungan. Tabel berikut ini memperlihatkan beberapa probabilitas (p) dengan hasil transformasinya ke nilai odds (O) dan, selanjutnya, ke log odds (LO) .

Tabel 15.6. Transformasi dari nilai probabilitas (p) ke Odds (O) dan Log Odds (LO)

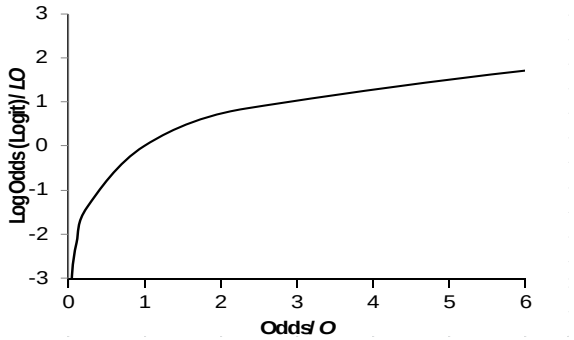
p	*O*	*LO*
0,01	0,01	-4,60
0,05	0,05	-2,94
0,1	0,11	-2,20
0,2	0,25	-1,39
0,5	1,00	0,00
0,7	2,33	0,85
0,9	9,00	2,20
0,95	19,00	2,94
0,99	99,00	4,60

Tabel tersebut memperlihatkan bahwa ketiga konsep memiliki hubungan bivariat (berpasangan) yang positif (semakin meningkat nilai pada satu konsep, semakin meningkat pula nilai pada satu konsep yang lain. Hubungan antar nilai probabilitas dengan nilai log odds hasil transformasinya tidak linier (membentuk garis lurus). Hal ini karena nilai probabilitas bersifat monotonik dan positif, nilai log odds bersifat monotonik dan simetris, negatif dan positif, dengan titik tengahnya LO = 0. Karena itu, hubungan keduanya digambarkan dalam plot yang membentuk huruf S, sebagaimana gambar berikut ini.



Gambar 1. Bentuk hubungan transformatif logistik antara Log Odds dan Probabilitas

Sebagaimana hubungan di atas, hubungan antara nilai odds dan log odds hasil transformasinya juga bersifat monotonik, tetapi tidak linier. Akan tetapi berbeda dari sebelumnya, karena nilai odds bergerak landai ketika nilai di bawah 1 dan meningkat secara tajam setelah nilai 1, sementar nilai log odds bergerak simetris dengan titik tengahnya nilai 0 yang merupakan transformasi nilai odds sama dengan 1. Gambar berikut memperlihatkan plot hubungan antara odds dan log odds.



Gambar 2. Bentuk hubungan transformasi logistik antara Odds dan Log Odds

Transformasi dari probabilitas ke log odds, disebut *tranformasi logit* (Peng, Lee & Ingersoll, 2002), diperlukan untuk mengatasi kesulitan dalam membuat model probabilitas yang memiliki rentangan yang terbatas. Transformasi ini dilakukan untuk memetakan probabilitas yang memiliki rentangan antar 0 sampai1 ke log odds yang memiliki rentangan – ∞ sampai ∞, negatif tak terbatas sampai positif tak terbatas. Dengan demikian, rentangan nilai hasil transformasi ini memungkinkan untuk dilakukan penaksiran nilai variabel binari secara linier, sebagaimana yang menjadi tujuan analisis regresi.

MODEL REGRESI LOGISTIK

Sebagaimana regresi liner, regresi logistik bertujuan untuk menaksir nilai variabel luaran/*output* (kriteria atau dependen, *Y*) berdasarkan skor variabel eksplanatoris (prediktor atau independen, *X*). Tidak seperti regresi biasa, variabel kriteria dalam regresi logistik adalah variabel binary (Strömbergsson, 2009), yang memiliki skor dikotomi, 1 (untuk munculnya peristiwa) dan 0 (untuk tidak munculnya peristiwa). Misalnya, variabel kelulusan dalam suatu ujian: lulus/sukses = 1, tidak lulus/gagal = 0; keanggotaan dalam organisasi keagamaan: anggota = 1, bukan anggota = 0. Dengan adanya keterbatasan skor ini, model regresi linier tidak dapat digunakan karena persyaratan linieritas tidak dapat dipenuhi.

Konsep matematis utama yang mendasari regresi logistik adalah logit—*logarithm natural* dari rasio odds (Peng, Lee & Ingersoll, 2002). Model regresi logistik memungkinkan membentuk hubungan antara variabel kriteria/dependen binari dan satu atau lebih variabel prediktor/independen melalui proses transformasi probabilitas perolehan skor binari ke nilai logit atau log odds. Karena itu, regresi logistik ini memodelkan probabilitas logit yang tertransformasikan sebagai suatu hubungan linier dengan variabel prediktor. Dalam model ini nilai variabel kriteria ditransformasikan ke dalam bentuk skala logit sehingga dapat memiliki rentangan nilai yang tak terbatas, dari -∞ sampai ∞. Regresi logistik *Y* pada X_1, X_2, … X_k menaksir nilai parameter β_0, β_1, β_1,. . . , β_k melalui metode *maximum likelihood* (Strömbergsson, 2009) dengan persamaan berikut:

$$logit(p_{(Y=1)}) = log\frac{p_{(Y=1)}}{1-p_{(Y=1)}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_k X_k$$

Di mana $logit(p_{[Y=1]})$ atau $log\frac{p_{(Y=1)}}{1-p_{(Y=1)}}$ adalah taksiran nilai probabilitas munculnya peristiwa pada variabel dependen (*Y*), β_0 adalah intersep atau koefisien regresi ketika skor (semua) variabel prediktor sama dengan 0 (nol), $\beta_1 + \beta_2 \ldots + \beta_k$ adalah slop/koefisien regresi atau besarnya pengaruh masing-masing variabel prediktor ketika pengaruh variabel prediktor lain dikontrol, dan X_1, X_2, $\ldots X_k$ adalah skor masing-masing prediktor X_1, X_2, $\ldots X_k$. Model tersebut digunakan bila jumlah variabel prediktornya lebih dari satu. Jika hanya satu prediktor, maka komponen koefisien dalam model persamaan tersebut hanya ada dua, yaitu $\beta_0 + \beta X$. Sedangkan jika tanpa prediktor, maka hanya ada satu komponen koefisien dalam model persamaan, yaitu β_0.

Untuk lebih jelasnya, masing-masing model akan dibahas dalam bagian-bagian berikut.

Sebagai ilustrasi aplikasi dari rumus tersebut, pembahasan akan menggunakan contoh data yang telah disajikan dalam tabel 1. Dalam tabel tersebut telah disajikan serangkaian data yang melibatkan amatan sebanyak 100 dengan variabel kriteria adalah kelulusan dalam seleksi tes masuk UTM (sebagai variabel *Y*), yang menunjukkan apakah pendaftar gagal atau sukses diterima sebagai calon mahasiswa baru. Nilai $p_{(Y=1)}$ = prob(*Y*=1). Sedangkan variabel prediktornya adalah reputasi asal sekolah (dengan skor kategori: 0 = reguler dan 1 = unggulan) dan NUN (Nilai ujian nasional, dengan rentangan skor teoritis 0 – 60). Untuk memudahkan pemahaman tentang aplikasi persamaan regresi logistik ini, pembahasan akan difokuskan pada pemahaman koefisien regresi. Sedangkan komponen lainya akan dibahas kemudian. Dalam pembahasan berikut, penghitungan koefisien dalam model yang sederhana dilakukan secara manual. Sebagai pembanding, pembahasan berikut juga akan menyajikan hasil penghitungan dengan menggunakan bantuan program statistik SPSS. Sedangkan penghitungan yang memiliki tingkat kerumitan yang tinggi akan dilakukan dengan menggunakan bantuan program statistik SPSS sehingga pembahasan akan difokuskan pada hasil analisis, tanpa disertai pembahasan langkah-langkah teknis penghitungannya.

Regresi logistik tanpa prediktor

Untuk memudahkan pemahaman, pembahasan akan dimulai dari model regresi logistik yang paling sederhana, yakni ***model tanpa prediktor***, juga disebut *regresi nihil*, yang menjadi dasar untuk model selanjutnya (Reed & Wu, 2013). Persamaan regresi tanpa prediktor ini hanya melibatkan variabel dependen (Y) saja, dengan kategori "sukses" atau terjadinya peristiwa, sebagai acuan. Karena tidak melibatkan prediktor, maka model persamaannya untuk menaksir probabilitas sukses hanya melibatkan nilai konstan (intersep), sehingga rumus persamaannya adalah:

$$logit(p_{[Y=1]}) = \beta_0$$

Di mana $logit(p_{[Y=1]})$ adalah taksiran nilai logit atau log odds dan β_0 adalah nilai konstan/intersep. Koefisien regresi dalam persamaan ini adalah nilai konstan, yang merupakan nilai log odds sukses (*Y*=1). Karena nilai log odds merupakan transformasi dari nilai odds, maka terlebih dahulu perlu dihitung nilai odds sukses (log odds = *LO* = ln[*O*]). Lebih lanjut, karena nilai odds diperoleh dari transformasi nilai probabilitas/*p* (odds = *O* = *p*/[*1*-*p*]), maka dapat dirumuskan sebagai berikut:

$$logit(p_{[Y=1]}) = \beta_0 = LO = \ln(O) = ln\frac{p_{(Y=1)}}{1 - p_{(Y=1)}}$$

Dalam kasus penerimaan mahasiswa baru UTM, sebagaimana contoh tersebut di atas, jumlah pendaftar yang sukses (*Y*=1, yakni pendaftar yang diterima sebagai calon mahasiswa baru) adalah 60 dari total 100 pendaftar. Karena itu, probabilitas sukses ($p_{[Y=1]}$) adalah:

$$p_{(Y=1)} = \frac{n_{(Y=1)}}{n} = \frac{Jumlah\ sukses}{Jumlah\ total} = \frac{60}{100} = 0{,}6$$

dengan nilai probabilitas sukses tersebut (p = 0,6), maka nilai odds sukses adalah:

$$O = \frac{p_{(Y=1)}}{1 - p_{(Y=1)}} = \frac{0{,}6}{1 - 0{,}6} = \frac{0{,}6}{0{,}4} = 1{,}5$$

Berdasarkan nilai odds tersebut, maka nilai log odds sukses (*LO*) adalah:

$$LO = \ln(O) = log_{2{,}71828}(1{,}5) \ = 0{,}405$$

Berdasarkan hasil tersebut, taksiran nilai odds sukses tanpa prediktor sama dengan nilai konstan (intersep/β_0) atau sama dengan log odds (logit) sukses, yakni 0,405. Dengan kata lain, intersep dari model regresi logistik tanpa variabel prediktor adalah taksiran log odds sukses untuk seluruh populasi.

Karena pemaknaan probabilitas dalam bentuk log odds sulit dipahami oleh mereka yang tidak atau kurang memiliki pemahaman matematika dengan baik, maka nilai log odds tersebut dapat ditransformasikan ke nilai odds atau nilai probabilitas. Untuk melakukan transformasi ke odds dapat dilakukan dengan rumus dan contoh penghitungan sebagai berikut:

$O = e^{LO}$
$= 2{,}71828^{0{,}405}$
$= 1{,}4993 \sim 1{,}5.$

Dengan hasil tersebut dapat disimpulkan bahwa taksiran odds sukses adalah 1,5 kali atau 50 persen lebih besar dari pada taksiran probabilitas gagal. Dengan kata lain, setiap 2 pendaftar yang gagal ditaksir terdapat 3 (= 2 × 1,5) pendaftar yang sukses diterima sebagai calon mahasiswa baru di UTM. Dalam penghitungan dengan menggunakan bantuan program statistik SPSS, hasilnya disajikan dalam tabel hasil *print out* berikut ini.

Variables in the Equation[a]

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.405	.204	3.946	1	.047	1.500

[a]Tabel dikutip dari *printout* hasil analisis program SPSS.

Dalam tabel di atas, koefisien regresi atau log odds untuk konstan atau *constant* disajikan dalam kolom B (sebagai lambang koefisien regresi, dengan nilai 0,405. Sedangkan odds disajikan dalam kolom Exp(B), dengan nilai 1,500. Nilai ini diperoleh dari eksponen log odds atau $e^{0,405} = 2,71828^{0,405}$.

Selanjutnya, karena tujuan regresi logistik adalah untuk memperoleh nilai taksiran probabilitas sukses, maka hasil penghitungan log odds tersebut perlu ditransformsikan ke nilai probabilitas. Transfomasi nilai log odds ke nilai probabilitas dapat dilakukan melalui nilai odds atau secara langsung (tanpa melalui odds). Untuk transformasi melalui nilai odds dapat dilakukan dengan rumus dan contoh penghitungan sebagai berikut:

$$p = O/(1+O) = 1,5/(1+1,5) = 1,5/2,5 = 0,6.$$

Sedangkan untuk transformasi langsung dari log odds dapat dilakukan dengan rumus dan contoh penghitungan sebagai berikut:

$p = e^{LO}/(1+e^{LO})$
$= 2,71828^{0,405}/(1+2,71828^{0,405}) = 1,4993/(1+1,4993) = 1,4993/2,4993 = 0,599988 \sim 0,6.$

Nilai tersebut sama dengan nilai probabilitas hasil penghitungan sebelumnya. Dengan demikian, taksiran nilai probabilitas sukses tanpa menggunakan variabel prediktor adalah sama dengan probabilitas sukses dari seluruh subjek. Dalam contoh seleksi calon mahasiswa baru UTM, taksiran nilai probabilitas sukses/diterima sama dengan nilai probabilitas sukses dari jumlah seluruh pendaftar calon mahasiswa baru UTM.

Regresi logistik dengan 1 prediktor variabel binari

Sebagai kelanjutan dari persamaan sebelumnya, dalam pembahasan berikut ini model akan ditambahkan satu variabel prediktor, eksplanatoris atau independen (*X*) yang berupa variabel binari (dikotomi, memiliki dua kategori masing-masing dengan skor 0 dan 1). Hal ini berarti bahwa taksiran nilai log odds variabel kriteria, luaran/*output*, atau dependen *Y* (yang juga berupa variabel binari, memiliki dua kategori

masing-masing dengan skor 0 dan 1) didasarkan pada skor variabel independen binari (*X*) (Peng, Lee, Ingersoll, 2002). Model atau persamaan regresi logistik dengan satu prediktor variabel binari ini dapat dirumuskan dalam model hubungan linier berikut ini:

$$logit(p_{[Y=1]}) = \beta_0 + \beta X$$

Di mana $logit(p_{[Y=1]})$ adalah taksiran nilai log odds variabel kriteria atau dependen dengan kategori sukses (*Y*=1), β_0 adalah nilai konstan/intersep (yakni nilai log odds ketika skor *X* = 0), β adalah nilai slop atau log odds ketika variabel *X* ditambahkan sebagai prediktor), dan *X* adalah skor variabel independen/ prediktor untuk kategori yang menjadi konsen dalam analisis (Wuensch, 2014).

Dalam kasus seleksi calon mahasiswa baru UTM, sebagai contoh, hal ini berarti taksiran probabilitas kelulusan/sukses (*Y*=1) pendaftar yang diterima berdasarkan reputasi asal sekolah unggulan (*X*=1). Untuk contoh, penghitungan merujuk pada ringkasan data dalam tabel 2. Fokus dalam pembahasan ini adalah nilai odds sukses pendaftar dari sekolah reguler dan odds sukses pendaftar dari sekolah unggulan ($O_{[X=0;Y=1]}$ dan $O_{[X=1;Y=1]}$). Selanjutnya, penghitungan dilakukan dengan menggunakan program SPSS, dengan fokus pada nilai odds (kolom B). Ringkasan hasil analisis disajikan sebagai berikut.

*Variables in the Equation**

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Reputasi	.903	.438	4.241	1	.039	2.466
	Constant	.067	.258	.067	1	.796	1.069

a. Variable(s) entered on step 1: Reputasi.

* Tabel dikutip dari *printout* hasil analisis program SPSS.

Dengan demikian, persamaan regresi logistik untuk memprediksi atau menaksir probabilitas sukses pendaftar dari sekolah unggulan ($X_1 = 1$) adalah:

$$logit(p_y) = \beta_0 + \beta_1 X_1 = 0{,}067+0{,}903(1) = 0{,}97.$$

atau

$$logit(p_{sukses}) = \beta_0 + \beta_1(\text{Unggulan}) = 0{,}067+0{,}903(1) = 0{,}97$$

Selanjutnya, jika logit/log odds tersebut ditransformasikan ke probabilitas, maka nilainya adalah:

$$p = e^{LO}/(1+e^{LO})$$
$$= 2{,}71828^{0{,}97}/(1+2{,}71828^{0{,}97}) = 2{,}637943/(1+2{,}637943)= 2{,}637943/3{,}637943$$
$$= 0{,}725119 \sim 0{,}725.$$

Dengan hasil tersebut, taksiran probabilitas sukses pendaftar seleksi masuk calon

mahasiswa baru UTM yang berasal dari sekolah unggulan (X_1 = 1;Y=1) adalah p = 0,725 atau 72,5 persen dari total pendaftar yang berasal dari sekolah tersebut.

Regresi logistik dengan 1 prediktor variabel kontinum

Regresi logistik dapat menggunaka satu prediktor atau variabel independen (*X*) yang berupa variabel kontinum (Garson, 2014; Reed & Wu, 2013). Pada dasarnya regresi ini sama dengan sebelumnya, kecuali dalam nilai variabel independen. Dalam regresi ini, taksiran nilai log odds variabel dependen *Y* (yang berupa variabel binary) didasarkan pada skor kontinum pada variabel independennya (*X*), dengan model yang sama dengan sebelumnya, yakni:

$$logit(p_y) = \beta_0 + \beta X$$

Di mana $logit(p_y)$ adalah taksiran nilai log odds variabel dependen atau kriteria (Y), β_0 adalah nilai konstan/intersep (yakni nilai log odds ketika $X = 0$), β adalah nilai slop atau log odds ketika skor *X* ditambahkan sebagai prediktor), dan *X* adalah skor variabel independen/prediktor yang menjadi konsen).

Dalam kasus seleksi calon mahasiswa baru UTM, sebagai contoh, hal ini berarti taksiran probabilitas kelulusan/sukses (*Y*=1) pendaftar untuk diterima berdasarkan NUN atau Nilai Ujian Nasional (*X*), yang berupa data kontinum. Walaupun persamaan regresi yang digunakan sama dengan persamaan untuk regresi logistik dengan prediktor variabel binari, langkah-langkah penghitungan dan penafsiran hasilnya berbeda karena variasi dan karakteristik skor variabel kontinum berbeda (lebih banyak variasi). Skor variabel binari merupakan skala nominal yang hanya berfungsi untuk membedakan variasi karakteristik satu dari lainnya. Sedangkan skor variabel kontinum merupakan skala interval atau rasio yang di samping berfungsi membedakan juga menunjukkan tingkatan dengan ukuran jarak yang sistematis, di samping banyak variasinya.

Karena variasi karakteristik variabel prediktor yang kontinum yang lebih banyak tersebut, penghitungan logit menjadi lebih rumit dan memerlukan langkah-langkah yang lebih panjang sehingga akan sulit dilakukan secara manual. Karena itu, untuk memudahkan memberikan pemahaman tentang konsep dan penafsiran hasilnya, dalam bahasan berikut ini akan langsung digunakan hasil penghitungan yang dilakukan dengan menggunakan program SPSS. Hasil analisis tersebut disajikan dalam *print out* tabel *Variables in the Equation*. Dalam tabel berikut ini akan disajikan hasil analisis regresi logistik yang melibatkan variabel NUN (Nilai Ujian Nasional), sebagai variabel prediktor, eksplanatoris atau independen (*X*=NUN), dan variabel kelulusan (sukses atau gagal) pendaftar calon mahasiswa baru UTM, sebagai variabel kriteria atau dependen (*Y*=kelulusan). Data selengkapnya untuk kedua variabel disajikan dalam tabel 15.1. Pembahasan hasil analisis akan dibatasi pada logit(*p*) atau koefisien regresi logistik.

*Variables in the Equation**

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	NUN	.119	.056	4.547	1	.033	1.126
	Constant	-4.547	2.321	3.839	1	.050	.011

a. *Variable(s) entered on step* 1: NUN.
*Tabel dikutip dari *printout* hasil analisis program SPSS.

Dalam tabel tersebut, koefisien regresi untuk konstan (intersep) dan NUN, yakni koefisien regresi *Y* pada *X*, atau (slop) disajikan dalam kolom B. Nilai intersep atau konstan (*constant*) merupakan log odds pendaftar yang memiliki skor nol untuk NUN ($X = 0$), yakni LO = -4,547. Dengan hasil LO ini, nilai odds sukses pendaftar calon mahasiswa baru UTM jika nilai NUN-nya sama dengan nol (0) adalah:

$$O = e^{LO}$$
$$= 2{,}71828^{-4{,}547}$$
$$= 0{,}010599 \sim 0{,}011.$$

Nilai odds atau intersep dalam model ini sama dengan nilai log odds untuk dapat diterima sebagai calon mahasiswa baru jika nilai NUN berada pada nilai nol hipotetis. Untuk menafsirkan taksiran intersep dan koefisien NUN dilakukan melalui persamaan berikut:

$$logit(p_y) = \beta_0 + \beta X$$

Di mana $logit(p_y)$ adalah taksiran nilai log odds variabel dependen atau kriteria (kelulusan), β_0 adalah nilai konstan/intersep atau nilai log odds ketika $X = 0$, [NUN = 0]), β adalah nilai koefisien regresi atau slop atau log odds ketika skor *X* ditambahkan sebagai prediktor 1, dan *X* adalah skor variabel independen/prediktor untuk kategori yang menjadi konsen (X = nilai NUN). Berdasarkan nilai koefisien dalam tabel tersebut di atas, maka nilai logit(p_y)/log odds adalah:

$$logit(p_y) = -4{,}547 + 0{,}119X$$

atau

$$logit(p_{sukses}) = -4{,}547 + 0{,}119(\text{NUN})$$

Untuk menguji taksiran kenaikan atau perubahan 1 unit *X* (NUN) pada perubahan logit, berikut ini akan dihitung nilai logit sukses jika nilai NUN = 37 dan 38, sebagai contoh, yaitu:

$$logit(p_y|X=37) = -4{,}547 + 0{,}119(37) = -0{,}144$$

dan

$$logit(p_y|X=38) = -4{,}547 + 0{,}119(38) = -0{,}025$$

Sedangkan nilai logit sukses jika nilai NUN = 59 dan 60, sebagai contoh yang lain, yaitu:

$$logit(p_y|X=59) = -4{,}547 + 0{,}119(59) = 2{,}474$$

dan

$$logit(p_y|X=60) = -4{,}547 + 0{,}119(60) = 2{,}953$$

Perbedaan antar log odds/logit(p_y) dalam masing-masing pasangan yang memiliki perbedaan satu unit skor X_1 (mis. 38 – 37 = 1 dan 60 – 59 = 1) tersebut adalah:

$$logit(p_y|X=38) - logit(p_y|X=37) = (-0{,}025) - (-0{,}144) = 0{,}119$$

dan

$$logit(p_y|X=60) - logit(p_y|X=59) = 2{,}953 - 2{,}474 = 0{,}119$$

Dengan hasil persamaan tersebut dapat disimpulkan bahwa koefisien regresi/slop (β) merupakan perbedaan log odds antara dua skor NUN yang berbeda 1 unit. Dengan kata lain, setiap kenaikan satu unit skor NUN (X) secara linier akan terjadi perubahan nilai logit atau log odds sebesar 0,119.

Selanjutnya, sesuai dengan tujuan regresi logistik, yakni untuk menaksir probabilitas terjadinya peristiwa (sukses) berdasarkan skor variabel prediktor/independen, nilai logit tersebut ditransformasikan ke nilai p, dengan menggunakan rumus sebagaimana dinyatakan sebelumnya, yakni:

$$p = e^{LO}/(1+e^{LO})$$

Dengan demikian, taksiran nilai logit probabilitas sukses pendaftar dengan nilai NUN = X_1=37 dan X_1=38 (yang masing-masing memiliki nilai logit(p_y) atau LO = –0,144 dan –0,025) adalah:

$$\begin{aligned} P(p_y|X=37) &= e^{LO}/(1+e^{LO}) \\ &= 2{,}71828^{-0{,}144}/(1+2{,}71828^{-0{,}144}) \\ &= 0{,}866888/1{,}866888 \\ &= 0{,}464349 \sim 0{,}464. \end{aligned}$$

dan

$$\begin{aligned} P(p_y|X=38) &= e^{LO}/(1+e^{LO}) \\ &= 2{,}71828^{-0{,}025}/(1+2{,}71828^{-0{,}025}) \\ &= 0{,}97531/1{,}97531 \\ &= 0{,}49375 \sim 0{,}494. \end{aligned}$$

Hasil tersebut menunjukkan bahwa taksiran probabilitas sukses pendaftar yang memiliki nilia NUN = 37 dan 38 adalah 0,464 dan 0,494; dengan perbedaan 0,03.

Sedangkan taksiran probabilitas sukses pendaftar dengan nilai NUN = X=59 dan X=60 (yang masing-masing memiliki nilai logit(py) atau LO = 2,474 dan 2,953) adalah:

$$\begin{aligned} P(p_y|X=59) &= e^{LO}/(1+e^{LO}) \\ &= 2{,}71828^{2{,}474}/(1+2{,}71828^{2{,}474}) \\ &= 11{,}86981/12{,}86981 \\ &= 0{,}9222987 \sim 0{,}922 \end{aligned}$$

dan

$$\begin{aligned} P(p_y|X=60) &= e^{LO}/(1+e^{LO}) \\ &= 2{,}71828^{2{,}953}/(1+2{,}71828^{2{,}953}) \\ &= 13{,}3697976/14{,}3697976 \\ &= 0{,}930409597 \sim 0{,}93 \end{aligned}$$

Dari hasil tersebut dapat diketahui bahwa taksiran probabilitas sukses pendaftar yang memiliki nilia NUN = 59 dan 60 adalah 0,922 dan 0,930; dengan perbedaan 0,008.

Dengan demikian, walaupun perbedaan skor antara pasangan dan nilai perbedaan logit sama, besaran taksiran probabilitas sukses berbeda sejalan dengan perbedaan skor. Perbedaan taksiran probabilitas antara skor NUN 37 dan 38 (dengan perbedaan skor 1 dan logit 0,119) adalah 0,03. Sedangkan perbedaan taksiran besaran probabilitas antara skor NUN 59 dan 60

(dengan perbedaan skor 1 dan logit 0,119) adalah 0,008. Hal ini menunjukkan bahwa perbedaan probabilitas semakin besar ketika semakin mendekati nilai log odds sama dengan 0 (atau p = 0,5) dan, sebaliknya, perbedaan probabilitas semakin kecil ketika semakin jauh dari nilai log odds sama dengan 0 (baik positif maupun negatif).

LOGIKA REGRESI LOGISTIK

Dalam regresi linier (di mana variabel dependennya, *Y*, merupakan variabel kontinum [memiliki skor yang merentang]), untuk mendapatkan taksiran nilai *Y* berdasarkan nilai *X* yang telah diketahui dapat diperoleh melalui rumusan persamaan regresi sebagai berikut:

$$Y' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_k X_k$$

Di mana *Y'* adalah taksiran nilai *Y*, β_0 adalah nilai intersep (nilai *Y* jika skor *X* = 0), β_1, β_2 … β_k adalah slop/koefisien regresi atau besarnya pengaruh masing-masing variabel prediktor ketika pengaruh variabel prediktor lain dikontrol, dan X_1, X_2, $\ldots X_k$ adalah skor masing-masing prediktor X_1, X_2, $\ldots X_k$. Dengan menggunakan rumus tersebut, nilai *Y'* dapat memiliki rentang yang tak terbatas serta dapat bernilai positif maupun negatif.

Dalam regresi logistik, variabel luaran (dependen) merupakan variabel binary (dalam contoh di atas adalah hasil seleksi calon mahasiswa baru), dengan skor 1 (*Y*=1; munculnya peristiwa, lulus) atau 0 (*Y*=0; tidak munculnya peristiwa, gagal). Karena nilai *Y* hanya ada dua kemungkinan (1 atau 0), maka memprediksinya berdasarkan skor *X* menjadi tidak rasional karena nilai taksiran dari hasil persamaan tersebut dapat tak terbatas (di atas 1 atau di bawah 0), sesuai dengan variasi nilai *X* (misalnya nilai ujian nasional/NUN). Alih-alih menggunakan nilai taksiran skor asli (1 atau 0), nilai taksiran yang diprediksikan dalam regresi logistik adalah nilai proporsi atau probabilitas (*p*) munculnya peristiwa pada variabel luaran/dependen/kriteria, sehingga nilai taksiran yang diprediksikan dapat merentang dari 0 (terendah) sampai 1 (tertinggi).

Persamaan regresi linier tersebut tidak dapat digunakan secara langsung, karena adanya dua masalah utama, yakni *statistik* dan *konseptual*. Secara statistik, luaran binari mencederai asumsi normalitas dan homosidastisitas yang dipersyaratkan dalam regresi linier. Dalam regresi linier diasumsikan bahwa skor variabel luaran atau kriteria sebagian besarnya akan berada di sekitar nilai yang diprediksikan melaui persamaan regresi dan terdistribusikan secara normal. Dalam regresi logistik asumsi ini tidak dapat

dipenuhi karena nilai variabel luaran yang diprediksikan hanya memiliki nilai 0 (jika *Y*=0) atau 1 (jika *Y*=1). Karena hanya ada dua alternatif, maka sebaran residu (selisih antara *Y* dan *Y'*) tidak bisa normal. Di samping itu, asumsi homosidestisitas juga tercederai karena varian galat menjadi konstan untuk semua jenjang *X*.

Secara konseptual, proporsi dan probabilitas terbatasi, terendah nilai 0 dan tertinggi nilai 1, sehingga tidak bisa melampaui batas tersebut. Sedangkan nilai taksiran dalam regresi linier dapat melampaui batas atas (*Y'*>1) atau batas bawah (*Y'*<0) dari rentang tersebut. Probabilitas lulus dalam seleksi penerimaan mahasiswa baru, misalnya, tidak mungkin bisa lebih besar dari 1 atau lebih kecil dari 0. Karena itu, taksiran dengan nilai tersebut (0 – 1) tidak dapat digunakan, semata-mata karena tidak masuk akal.

Masalah keterbatasan nilai probabilitas tersebut sebenarnya dapat diatasi dengan memotong nilai *Y*>1 sehingga menjadi maksimum 1. Garis regresi akan langsung ke nilai 1, sehingga setiap kenaikan nilai *X* di atas titik ini tidak akan memiliki pengaruh pada taksiran *Y*. Demikian juga, nila *Y*<0 dipotong menjadi minimal 0 sehingga pengurangan nilai *X* di bawah titik menjadi 0. Dengan demikian, pemotongan linieritas ini secara teoritis tidak dapat diterima karena tidak masuk akal.

Hubungan antara *X* dan *Y* dapat digambarkan dalam kurva berbentuk *S*. Dalam regresi linier, garis regresi diasumsikan memiliki pengaruh pada *Y* yang sama di titik manapun dalam distribusi nilai *X*. Kurva berbentuk *S* menggambarkan hubungan non linier antara *X* dan *Y*, di mana hubungannya mendekati linier antara probabilitas 0,2 sampai 0,8 dan menjadi datar saat probabilitas mendekati 1 (di bagian atas) dan 0 (di bagian bawah). Perubahan 1 unit nilai *X* pada bagian tengah distribusi memiliki pengaruh yang cukup berarti pada perubahan taksiran *Y*. Akan tetapi, untuk mendapatkan tingkat perubahan nilai taksiran *Y* yang sama diperlukan perubahan nilai X yang lebih besar pada titik distribusi di ujung garis *S*, mendekati nilai 1(di bagian atas) dan 0 (di bagian bawah). Hal ini memberi petunjuk bahwa secara konseptual kurva berbentuk *S* lebih masuk akal dari pada garis linier untuk mengatasi taksiran probabilitas tersebut.

Secara matematis, kurva berbentuk *S* dapat dibuat berdasarkan fungsi logistik, yakni proses mengubah data secara sistematis dari log odds ke proporsi. Ada banyak

cara matematis untuk menghasilkan kurva berbentuk *S*, tetapi fungsi logistik adalah yang paling populer dan paling mudah untuk menafsirkan. Fungsi hanyalah sebuah proses yang mentransfer data dengan cara yang sistematis - dalam contoh ini mengubah peluang logit ke proporsi.

Sebagaimana dibahas sebelumnya, variabel eksplanatoris atau independen /prediktor memiliki hubungan linier dan aditif dengan log odds terjadinya peristiwa. Karena itu, nilai log odds atau logit(p_y) dapat di taksir dari variabel eksplanatoris dengan menggunakan model persamaan (Askar, Usluel, & Mumcu, 2006) sebagai berikut:

$$logit(p_y) = \beta_0 + \beta X$$

Di mana $logit(p_y)$ adalah taksiran nilai log odds variabel luaran /dependen/kriteria (*Y*), β_0 adalah nilai konstan/intersep (yakni nilai log odds ketika $X = 0$), β adalah nilai slop atau log odds ketika skor *X* ditambahkan sebagai prediktor), dan *X* adalah skor variabel eksplanatoris/independen/prediktor yang menjadi konsen.

Untuk memperjelas hal ini, marilah kita kembali ke contoh penggunaan analisis regresi logistik dengan satu prediktor variabel kontinum, sebagai contoh, sebagaimana disajikan dalam bagian sebelumnya. Dalam contoh tersebut, sebagai variabel luaran/ kriteria (*Y*) adalah hasil seleksi calon mahasiswa baru UTM (*Y*=1, lulus; *Y*=0, gagal) dan sebagai variabel eksplanatoris/prediktornya adalah nilai Ujian Nasional atau NUN (*X*). Hasil luaran analisis dengan menggunakan program SPSS adalah sebagai berikut:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	NUN	.119	.056	4.547	1	.033	1.126
	Constant	-4.547	2.321	3.839	1	.050	.011

a. *Variable(s) entered on step* 1: NUN.

Dari tabel tersebut dapat diketahui bahwa nilai konstan adalah koefisien Constant (dalam kolom B), sehingga β_0 = -4,546; nilai slop adalah koefisien NUN (dalam kolom B), sehingga β = 0,119. Dengan demikian, taksiran nilai log odds variabel luaran (*Y*) berdasarkan nilai variabel eksplanatorisnya (*X*) adalah:

$$logit(p_y) = \beta_0 + \beta X$$
$$= -4{,}546 + 0{,}119X$$

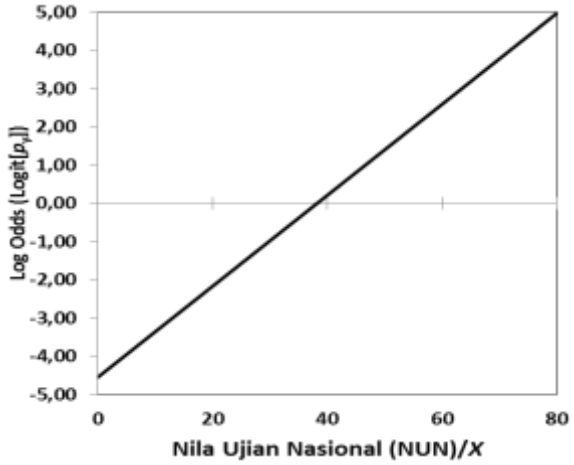
Berdasarkan persamaan tersebut, jika 5 orang pedaftar yang masing-masing memiliki

NUN, X = 20, 30, 40, 50, 60, maka nilai logit masing-masing adalah:

$X = 20 \rightarrow logit(p_y) = -4{,}546 + 0{,}119(20) = -2{,}166$
$X = 30 \rightarrow logit(p_y) = -4{,}546 + 0{,}119(30) = -0{,}976$
$X = 40 \rightarrow logit(p_y) = -4{,}546 + 0{,}119(40) = 0{,}214$
$X = 50 \rightarrow logit(p_y) = -4{,}546 + 0{,}119(50) = 1{,}404$
$X = 60 \rightarrow logit(p_y) = -4{,}546 + 0{,}119(60) = 2{,}594$

Dengan menggunakan logit atau log odds sebagai taksiran probabilitas terjadinya peristiwa, masalah batas atas dan bawah dari nilai probabilitas (yakni maksimal bernilai 1 dan minimal bernilai 0), yang menjadi kendala dalam hubungan linier, dapat terpecahkan. Hal ini dimungkinkan karena nilai logit tidak memiliki batas atas (nilai maksimal) maupun batas bawah (nilai minimal) sehingga nilai taksiran Y berdasarkan nilai X (berapapun nilai) dapat diperoleh dengan menggunakan persamaan regresi linier tersebut. Pengaruh perubahan setiap unit nilai X, secara konsisten dan sistematis diikuti perubahan pada nilai logit.

Jika penghitungan logit tersebut diteruskan untuk nilai X yang lain, maka hubungan antara NUN dan Logit(p_y) dapat divisualisasikan dalam grafik sebagai berikut:



Gambar 3. Bentuk hubungan antara Nilai Ujian Nasional, sebagai prediktor/eksplanatoris dan nilai Log Odds, sebagai kriteria/luaran

Grafik tersebut memperlihatkan hubungan linier antara Nilai Ujian Nasional (NUN) sebagai variabel eksplanatoris/prediktor atau independen (X), dengan nilai Log Odds/logit, sebagai variabel luaran/ kriteria atau dependen (Y). Dengan demikian, penggunaan log odds dalam menaksir probabilitas tersebut dapat memecahkan masalah batas atas dan bawah dari nilai probabilitas, yakni 1 dan 0.

Meskipun permasalahan linieritas terpecahkan, penalaran nilai log odds sebagai

representasi tingkat probabilitas terjadinya peristiwa sulit dipahami, terutama bagi orang awam. Untuk memudahkan pemahaman penalaran tersebut, nilai logit dapat ditransformasikan atau dirubah menjadi nilai probabilitas atau proporsi. Sebagaimana dibahas sebelumnya, transformasi dari nilai logit menjadi nilai proporsi/probabilitas tersebut dapat dilakukan dengan rumus berikut:

$$p = e^{LO}/(1+e^{LO})$$

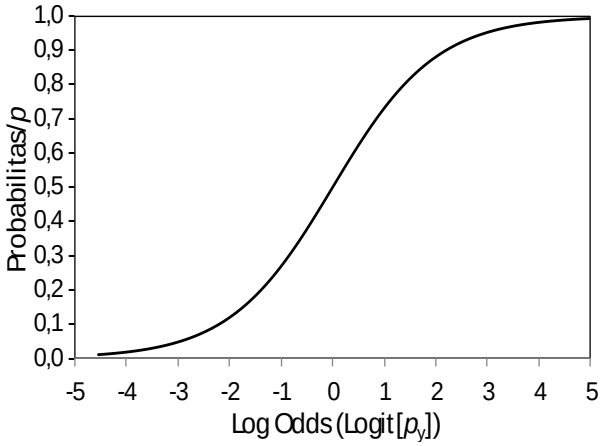
di mana *p* adalah nilai probabilitas, *e* adalah nilai konstan eksponen (yakni 2,71828), LO adalah nilai log odds atau logit yang ditaksir dari nilai variabel eksplanatori/prediktor/ independen. Berdasarkan rumus tersebut, probabilitas diterima sebagai mahasiswa baru UTM untuk 5 orang pedaftar yang masing-masing memiliki NUN, *X* = 20, 30, 40, 50, dan 60 serta *logit*(p_y) =LO = -2,166; -0,976; 0,124; 1,404; 2,594; sebagaimana telah dihitung sebelumnya tersebut adalah sebagai berikut:

X = 20; LO = -2,166→p = $2{,}71828^{-2{,}166}/(1+2{,}71828^{-2{,}166})= 0{,}103$
X = 30; LO = -0,976→p = $2{,}71828^{-0{,}976}/(1+2{,}71828^{-0{,}976})= 0{,}274$
X = 40; LO = 0,214→p = $2{,}71828^{0{,}214}/(1+2{,}71828^{0{,}214})= 0{,}553$
X = 50; LO = 1,404→p = $2{,}71828^{1{,}404}/(1+2{,}71828^{1{,}404})= 0{,}803$
X = 60; LO = 2,594→p = $2{,}71828^{2{,}594}/(1+2{,}71828^{2{,}594})= 0{,}930$

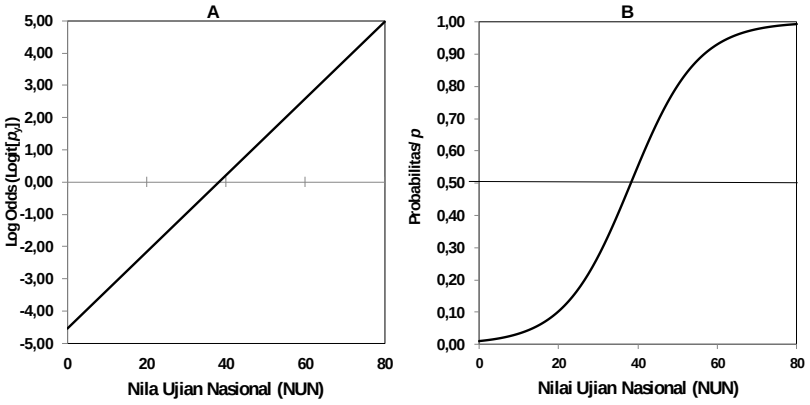
Jika penghitungan diteruskan untuk logit dari transformasi nilai NUN yang lain, maka akan diperoleh hubungan non linier yang tergambarkan dalam kurva bentuk S, sebagaimana telah dibahas sebelumnya. Perubahan nilai di sekitar (dekat) titik tengah distribusi; logit(p_y) = 0,0; memiliki pengaruh perubahan probabilitas yang lebih substantif (besar) dibandingkan perubahan pada titik yang lebih jauh dari titik tengah distribusi tersebut. Perubahan 0,5 dari logit(p_y) = 0,0 ke 0,5 menyebabkan perubahan taksiran proporsi 0,122 (dari *p* = 0,50 ke *p* = 0,62). Sedangkan perubahan 0,5 dari logit(p_y) = 3,0 ke 3,5 menyebabkan perubahan proporsi 0,018 (dari *p* = 0,953 ke *p* = 0,971). Hubungan antara logit atau log odds/LO (sebagai variabel eksplanatoris/prediktor) dengan nilai probabilitas/proporsi/p hasil transformasinya tersebut (sebagai variabel luaran/kriteria) dapat divisualisasikan dalam kurva berbentuk *S* berikut ini.

Selanjutnya, karena nilai variabel eksplanatoris/prediktor (*X*) memiliki hubungan linier dengan nilai log odds/logit, maka hubungan antara variabel eksplanatoris/prediktor (*X*) dengan variabel luaran/kriteria, nilai probabilitas/proporsi (*Y*), juga memiliki bentuk yang sama, yakni non linier bentuk *S*. Berdasarkan data nilai

ujian nasional (NUN) yang dimiliki pendaftar calon mahasiswa baru UTM, maka hubungan antara NUN (*X*) dengan taksiran peluang atau probabilitas lulus untuk diterima (*Y*) direpresentasikan secara visual dalam kurva berbentuk S. Dengan demikian, pengaruh NUN (*X*) terhadap logit(p_y) direpresentasikan dalam kurva berbentuk linier, sedangkan pengaruh NUN (*X*) terhadap taksiran probabiltas kelulusan dalam seleksi calon mahasiswa baru, p_y, direpresentasikan dalam kurva non linier berbentuk *S*. Kedua bentuk hubungan tersebut dapat divisualisasikan dalam dua grafik berikut ini.



Gambar 4. Bentuk hubungan antara nilai Log Odds dan Probabilitas lulus peserta seleksi calon mahasiswa baru UTM.



Gambar 5. Bentuk hubungan transformasi logistik antara NUN dengan Logit(p_y), gambar A, dan antara NUN dengan Probabilitas/p, gambar B.

Dalam gambar A tersebut di atas, pengaruh NUN pada hasil seleksi (direpresentasikan dengan nilai logit) digambarkan dengan garis linier/lurus. Sedangkan dalam gambar B, pengaruh NUN pada hasil seleksi (direpresentasikan dengan nilai probabilitas) digambarkan dengan garis kurva berbentuk *S*. Walaupun bentuk dan nilainya berbeda,

keduanya mempresentasikan peluang sukses yang sama bagi pendaftar seleksi calon mahasiswa baru UTM berdasarkan nilai ujian nasional (NUN) yang mereka miliki.

UJI SIGNIFIKANSI KOEFISIEN: UJI WALD

Untuk menguji hipotesis nol, yakni apakah koefisien regresi logistik (logit) pada suatu variabel independen secara signifikan berbeda dari 0, digunakan uji Wald (Lee & Su, 2015). Pada dasarnya uji ini merupakan yakni uji Kai Kuadrat (*Wald chi-square test*). Uji Wald digunakan untuk menentukan signifikansi statistik untuk koefisien logit setiap variabel independen. Nilai Wald dapat diperoleh dari rasio antara kuadrat log odds (b^2) dengan nilai kuadrat galat baku (gb^2) atau *standar error* (*SE*), yang dapat dirumuskan sebagai berikut:

$$Wald = b^2/(gb)^2$$

Karena Wald merupakan nilai Kai Kuadrat (χ^2), maka selanjutnya, nilai χ^2 tersebut dibandingkan dengan nilai χ^2 kriteria untuk taraf signifikansi (*p*) dan derajat kebebasan (dk) tertentu (nilai tersebut dapat diperoleh dari tabel χ^2kritis). Dalam *printout* program statistik, taraf signifikansi untuk nilai Wald hasil penghitungan disajikan dalam kolom Sign. (*significance*) sehingga tidak perlu lagi membandingkan dengan nilai yang ada dalam tabel χ^2kritis. Alih-alih, untuk pengujian signifikansi tinggal membandingkan nilai taraf signifikansi tersebut dengan nilai taraf signfikansi kritis yang sebelumnya telah tetapkan menjadi kriteria penerimaan atau penolakan hipotesis nol. Untuk memudahkan pemahaman, *print out* hasil analisis dengan menggunakan program SPSS disajikan kembali sebagaimana tabel berikut. Dalam tabel tersebut, nilai koefisien regresi logistik atau log odds untuk masing-masing variabel independen disajikan dalam kolom B dan galat bakunya (standard error) disajikan dalam Kolom S.E. Nilai Kai Kuadrat Wald disajikan dalam kolom Wald, diikuti nilai derajat kebebasan, kolom df (*degrees of freedom*) dan taraf signifikansinya, kolom Sig.

*Variables in the Equation**

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	NUN	.119	.056	4.547	1	.033	1.126
	Constant	-4.547	2.321	3.839	1	.050	.011

a. *Variable(s) entered on step* 1: NUN.
*Tabel dikutip dari *printout* hasil analisis program SPSS.

Berdasarkan tabel tersebut di atas, sebagai contoh, nilai Wald untuk variabel NUN (Nilai ujian nasional) adalah $\chi^2 = 4{,}547$ (diperoleh dari $B^2/S.E.^2 = 0{,}119^2/0{,}056^2$).

Dengan derajat kebebasan dk = 1, nilai Wald tersebut memiliki taraf signfikansi 3,3 persen (p =0,033), lebih kecil dari taraf signfikansi yang dijadikan kriteria penerimaan hipotesis nol, yakni 5 persen (nilai kritis $\alpha \leq 0{,}05$). Dengan hasil yang demikian ini, maka hipotesis nol yang menyatakan bahwa nilai konstan/koefisien/log odds sama dengan 0 ditolak, sehingga dapat disimpulkan bahwa konstan tidak sama dengan 0. Meskipun demikian, taraf signifikansi tersebut lebih besar dari 1 persen (nilai kritis $\alpha \leq 0{,}01$). Dengan kata lain, nilai log odds tersebut signifikan pada taraf 5 persen, tetapi tidak signifikan pada taraf 1 persen. Jika nilai taraf signifikansi kritis yang ditentukan adalah 5 persen, dengan demikian, nilai koefisien regresi logistik atau log odds nilai NUN (b = 0,119) dapat digunakan dalam model taksiran probabilitas sukses pendaftar calon mahasiswa baru UTM. Jika yang digunakan sebagai kriteria taraf signifikansi kritis adalah 1 persen, NUN tidak dapat dimasukkan dalam model taksiran tersebut.

Tabel tersebut juga memperlihatkan nilai Wald untuk konstan atau intersep, yakni $\chi^2 = 3{,}839$ (diperoleh dari $B^2/S.E.^2 = -4{,}547^2/2{,}321^2$). Dengan derajat kebebasan sama dengan 1, nilai Wald tersebut memiliki taraf signfikansi 5 persen (p =0,050), sama dengan taraf signfikansi yang dijadikan kriteria penerimaan hipotesis nol, yakni 5 persen (nilai kritis $\alpha \leq 0{,}05$). Dengan hasil ini, maka hipotesis nol yang menyatakan bahwa nilai konstan/koefisien/log odds sama dengan 0 ditolak, sehingga dapat disimpulkan bahwa konstan tidak sama dengan 0. Dengan demikian, nilai koefisien regresi logistik atau log odds NUN dapat digunakan dalam model taksiran probabilitas sukses pendaftar calon mahasiswa baru UTM.

SIMPULAN

Regresi logistik digunakan jika variabel output/kriteria/dependenya merupakan variabel binari, memiliki dua kategori, yakni apakah suatu peristiwa terjadi atau tidak terjadi (Y=1 atau Y=0). Sebagaimana regresi linier, regresi logistik digunakan untuk memprediksi probabilitas bahwa Y sama dengan 1 (bukan 0) untuk nilai X tertentu. Yakni, jika X dan Y memiliki korelasi positif, probabilitas bahwa seseorang akan memperoleh skor Y=1 akan meningkat sejalan dengan meningkatnya nilai X. Hubungan antara variabel independen dan dependen diasumsikan menyerupai kurva berbentuk-S, yakni ketika variabel independen berada dalam tingkat terendah, probabilitas mendekati nol, sejalan dengan peningkatan nilai variabel independen, nilai probabilitas meningkat,

tetapi kemudian melandai sehingga mendekati nilai satu, walaupun tak pernah melampauinya. Regresi logistik menggunakan istilah yang berbeda dari regresi linier, di antaranya adalah probabilitas, odds, log odds, dan rasio odds.

Regresi logistik memiliki model serupa dengan model regresi linier, yakni $p_{(Y=1)}$ = tetapi dengan konsep dan pemaknaan yang berbeda, yakni $logit(p_y) = \beta_0 + \beta X$. Untuk mengevaluasi apakah model regresi logistik cocok dengan data dapat dilakukan dengan Uji Wald, yang pada dasarnya merupakan χ^2 yang menguji apakah pengaruh masing-masing prediktor signifikan dalam memprediksi probabilitas terjadinya peristiwa pada variabel output.

DAFTAR RUJUKAN

Askar, P., Usluel, Y. K. & Mumcu, F. K. (2006). Logistic Regression Modeling for Predicting Task-Related ICT Use in Teaching. *Educational Technology & Society, 9* (2), 141-151.

Garson, G. David, 2014, *Logistik Regression: Binomial and Multinomial.* Asheboro, NC, USA: Statistical Publishing Associates.

Garson, G. D. (2009). *"Logistic Regression" from Statnotes: Topics in Multivariate Analysis.* Diunduh 21 Juni 2016 from ttp://faculty.chass.ncsu.edu/garson/pa765/statnote.htm.

Glass, G. V., & Hopkins, K. D. (1984). Statistical Methods in Education and Psychology (2nd ed.). Englewood Cliffs, N. J.: Prentice-Hall.

Gullickson, A. (2005). *Odds and Probabilities*. Diunduh pada 24 Januari 2017, dari: http://pages.uoregon.edu/aarong/teaching/G4075_Outline/node15.html.

Hailpern, Susan M. & Paul F. Visintainer, 2003, Odds ratios and logistic regression: Further examples of their use and interpretation, *The Stata Journal*, *3*(3), hh. 213–225.

Hair, J.F., R.E. Anderson, R.L. Tatham, & W.C. Black, 1995, *Multivariate Data Anaysis with Readings*, Englewood Cliffs, Amerika Serikat: Prentice Hall International.

Lee, M.C. & Su, L.E. (2015). Comparison Of Wavelet Network And Logistic Regression In Predicting Enterprise Financial Distress. *International Journal of Computer Science & Information Technology*, 7(3), 83-96. DOI:10.5121/ijcsit.2015.7307 83.

Loh, Wei-Yin, 2006, Logistic Regression Tree Analysis, dalam H. Pham, ed. *Handbook of Engineering Statistics*, London: Springer, hh. 537–549.

Minitab Inc. 2015, *Minitab Support*, *diunduh 20 September 2016, dari* http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regression-and-correlation/logistic-regression/what-is-the-odds-ratio/.

Oty, K. & Elliott, B. (2015). *Algebra for the Sciences*. *Diunduh 20 September 2016* http://www.

se.edu/ dept/math/ files/2015/06/Book-as-of-June-1-2015.pdf

Pedhazur, E.J., *Multiple regression in behavioral research: Explanation and prediction*, New York: Holt, Rinehart, & Winston, 1982.

Peng, C.J., Lee, K.L. & Ingersoll, G.M. (2002). An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research, 96*(1), 3-14.

Press, S.J. & Wilson, S. (1978). Choosing Between Logistic Regression and Discriminant Analysis. *Journal of the American Statistical Association, 73*(364), 699-705.

Reed, P. & Wu, Y. (2013). Logistic regression for risk factor modelling in stuttering research. *Journal of Fluency Disorders*, 38, 88–101.

Strömbergsson, S. (2009). *Binary Logistic Regression and its application to data from a study of children's recognition of their own recorded voices.* Diunduh 23 Pebruari 2017 dari: http://stp.lingfil.uu.se/~nivre/statmet/strombergsson.pdf.

Tognetti, K. (1998). *e-the exponential-the magic number of growth*. Diunduh 23 Pebruari 2017 dari: https://www.austms.org.au/Modules/Exp/exp.pdf.

Upton, Graham & Ian Cook, 2011, *A Dictionary of Statistics*, New York: Oxford University Press.

Wuensch, K.L. (2014). *Binary Logistic Regression with SPSS*. Diunduh 24 Maret 2016, dari: http://core.ecu.edu/psyc/wuenschk/MV/multReg/Logistic-SPSS.pdf.

Yoo, W., Ference, B. A., Cote, M. L., & Schwartz, A. (2012). A Comparison of Logistic Regression, Logic Regression, Classification Tree, and Random Forests to Identify Effective Gene-Gene and Gene-Environmental Interactions. *International Journal of Applied Science and Technology*, *2*(7), 268.