BAB II
TINJAUAN PUSTAKA

2.1. Tabel Kontingensi

Tabel kontingensi merupakan suatu metode statistik yang menggambarkan dua atau lebih variabel secara simultan dan hasilnya ditampilkan dalam bentuk tabel yang merefleksikan distribusi bersama dua atau lebih variabel dengan jumlah kategori yang terbatas. Tabel kontingensi sering disebut juga dengan *cross tabulation* yang digunakan untuk mengetahui hubungan antara dua atau lebih variabel penelitian yang bukan hubungan sebab akibat [11].

Tabel 2.1 Tabel Kontingensi

A		B				Total
		B_1	B_2	$\cdots$	B_j	
	A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	n_1
	A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	n_2
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
	A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$n_{i..}$
	Total	n_1	n_2	$\cdots$	n_j	$n_{..}$

2.2. Uji Independensi

Uji independensi dilakukan untuk mengetahui ada atau tidaknya hubungan antara variabel respon dengan variabel prediktor. Pengujian tersebut dilakukan dengan melihat nilai *Chi-Square* [12].

Hipotesis yang digunakan adalah hipotesis nol (H_0) dan hipotesis alternatif (H_1). Hipotesis nol yaitu tidak ada hubungan antara variabel respon dengan variabel prediktor sedangkan hipotesis alternatifnya adalah terdapat hubungan antara variabel respon dengan variabel prediktor [12].

Statistik Uji :

$$X^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\left(n_{ij} - \hat{m}_{ij}\right)^2}{\hat{m}_{ij}} \quad (2.1)$$

Bersasarkan statistik uji tersebut, ditolak H_0 jika nilai X^2 yang diperoleh lebih besar dari nilai tabel $X^2_{a,(i-1)(j-1)}$.

2.3. Uji Korelasi *Spearman*

Uji *Spearman* merupakan metode korelasi yang dikemukakan oleh *Carl Spearman* pada Tahun 1904. Metode ini diperlukan untuk mengukur keeratan hubungan antara dua variabel. Kedua variabel itu tidak harus mengikuti distribusi normal dan kondisi variabel tidak diketahui sama. Korelasi rank dipergunakan apabila pengukuran kuantitatif secara eksak tidak mungkin dilakukan [13].

Perhitungan koefisien korelasi rank dinotasikan dengan ρ. Langkah-langkah perhitungan tersebut sebagai berikut.

a) Nilai pengamatan dari dua variabel yang akan diukur hubungannya diberi jenjang. Apabila ada nilai pengamatan yang sama dihitung jenjang rata-ratanya.
b) Setiap pasang jenjang dihitung perbedaannya.
c) Perbedaan setiap pasang jenjang tersebut dikuadratkan dan dihitung jumlahnya.
d) Nilai ρ (koefisien korelasi *Spearman*) dihitung dengan rumus:

$$\rho = 1 - \frac{6\sum bi^2}{n(n^2-1)}$$

Berdasarkan rumus tersebut diperoleh keterangan ρ adalah koefisien korelasi *spearman*, bi menunjukkan perbedaan setiap pasang rank dan n merupakan jumlah pasangan rank. Hipotesis H_0 yang akan diuji menyatakan bahwa dua variabel yang diteliti dengan nilai jenjang itu independen artinya tidak ada hubungan antara variabel yang satu dengan yang lainnya.

Kriteria pengambilan keputusan dari korelasi *spearman* yaitu H_0 diterima apabila ρ hitung lebih kecil dari ρ tabel dan H_0 ditolak apabila ρ hitung lebih besar dari ρ tabel. Nilai ρ tabel dapat dilihat pada tabel *spearman.* Kriteria tingkat kekuatan korelasi yang dihasilkan dalam uji *spearman* yaitu jika nilai koefisien korelasi yang dihasilkan sebesar 0.00 sampai 0.25 maka hubungannya sangat lemah, 0.26 sampai 0.5 maka hubungannya cukup, 0.51 hingga 0.75 sama dengan hubungannya kuat, 0.76 hingga 0.99 hubungannya sangat kuat dan jika dihasilkan sebesar 1 maka hubungannya sempurna.

Arah korelasi dilihat pada angka koefisien korelasi sebagaimana tingkat korelasi. Besarnya nilai koefisien tersebut terletak antara +1 sampai -1. Jika koefisien korelasi bernilai positif, maka hubungan kedua variabel dikatakan searah. Sebaliknya, jika koefisien korelasi bernilai negatif maka hubungan kedua variabel tidak searah. Tidak searah artinya jika variabel X meningkat maka variabel Y akan menurun [14].

Kekuatan dan arah korelasi akan mempunyai arti jika hubungan antar variabel tersebut signifikan. Dikatakan ada hubungan yang signifikan, jika nilai Sig. (2-tailed) hasil perhitungan lebih kecil dari nilai 0.05 atau 0.01. Sementara itu, jika nilai Sig. (2-tailed) lebih besar dari 0.05 atau 0.01 maka hubungan antar variabel tersebut dapat dikatakan tidak signifikan [14].

2.4. Regresi Logistik Multinominal

Regresi Logistik merupakan sebuah metode analisis statistik untuk menggambarkan hubungan antara variabel terikat dengan variabel bebas yang mempunyai dua atau lebih kategori maupun interval [15]. Regresi logistik terbagi menjadi dua yaitu regresi logistik biner dan regresi logistik multinomial. Regresi logistik biner adalah suatu regresi yang digunakan untuk menggambarkan hubungan variabel babas dengan sekumpulan variabel terikat, dimana variabel terikat bersifat biner atau dikotomus. Variabel kategori yang tidak memiliki urutan disebut sebagai variabel nominal sedangkan yang memiliki urutan disebut variabel oridinal. Kedua jenis variabel ini , baik nominal maupun ordinal sering disebut juga variabel multinomial.

Regresi logistik multinomial merupakan regresi logistik yang digunakan saat variabel dependen mempunyai skala yang bersifat multinominal. Skala multinomial adalah suatu pengukuran yang dikategorikan menjadi lebih dari dua kategori [15]. Model regresi logistik adalah sebagai berikut.

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

dengan $g(x) = \beta_0 + \beta_1 x_1 + \ ... + \beta_p x_p$

Secara umum, bentuk dari fungsi logit dengan variabel respon yang terdiri dari tiga kategori adalah sebagai berikut.

$$g_j(x) = \beta_{j0} + \beta_{j1} x_1 + \beta_{j2} x_2 + \ ... + \beta_{jp} x_p$$

Cumulative Logit Models didapatkan dengan membandingkan peluang kumulatif yaitu peluang kurang dari atau sama dengan kategori respon ke-j pada p variabel prediktor yang di nyatakan dalam vektor x_i $P(Y \leq j|x_i)$, dengan peluang lebih besar dari kategori respon ke-j, $P(Y > j|x_i)$ [15]. Berikut rumus *cumulative logit models.*

$$Logit\ P(Y \leq j|x_i) = \log\left(\frac{P(Y \leq j|x_i)}{P(Y > j|x_i)}\right)$$

Suatu variabel respon dengan tiga kategori akan membentuk dua persamaan logit, dimana masing-masing persamaan ini membentuk regresi logistik multinomial yang membandingkan suatu kelompok kategori terhadap pembanding, yaitu sebagai berikut.

$$g_1(x) = log\frac{P(Y = 2|x)}{P(Y = 1|x)}$$
$$= log\frac{\pi_2(x)}{\pi_1(x)} = \beta_{10} + \beta_{11} x_1 + \beta_{12} x_2 + \ \cdots + \beta_{1p} x_p$$

$$g_2(x) = log\frac{P(Y = 3|x)}{P(Y = 1|x)}$$
$$= log\frac{\pi_3(x)}{\pi_1(x)} = \beta_{20} + \beta_{21} x_1 + \beta_{22} x_2 + \ \cdots + \beta_{2p} x_p$$

Berdasarkan kedua peluang kumulatif pada persamaan diatas, didapatkan peluang untuk masing-masing kategori respon sebagai berikut.

$$P(Y = 1|x) = \pi_1(x) = \frac{\exp g_1(x)}{1 + \exp g_1(x) + \exp g_2(x)}$$

$$P(Y = 2|x) = \pi_2(x) = \frac{\exp g_2(x)}{1 + \exp g_1(x) + \exp g_2(x)}$$

$$P(Y = 3|x) = \pi_3(x) = \frac{1}{1 + \exp g_1(x) + \exp g_2(x)}$$

Peluang untuk kategori respon tersebut digunakan untuk melihat ketepatan model yang dihasilkan. Peluang nilai sukses yang dihasilkan berkisaran 0 sampai 1.

a. Estimasi Parameter *Maximum Likelihood*

Estimasi parameter adalah pendugaan karakteristik populasi (parameter) dengan menggunakan karakteristik sampel (statistik). Estimasi parameter regresi logistik multinomial menggunakan metode *Maximum Likelihood Estimation* (MLE). Metode *Maximum Likelihood Estimation* merupakan metode yang digunakan untuk menaksir parameter-parameter model regresi logistik dengan memberikan nilai estimasi β dengan memaksimumkan fungsi *Likelihood* [11]. Berikut fungsi *Likelihood* untuk sampel dengan n sampel random.

$$L(\beta) = \prod_{i=1}^{n} \pi_0(x_i)^{y_{0i}} \pi_1(x_i)^{y_{1i}} \pi_2(x_i)^{y_{2i}}$$

dengan i = 1,2,….,J

Dari persamaan diatas didapatkan fungsi ln- *Likelihood* sebagai berikut.

$$L(\beta) = \ln\{1(\beta)\} = \left\{\prod_{i=1}^{n} \pi_0(x_i)^{y_{0i}} \pi_1(x_i)^{y_{1i}} \pi_2(x_i)^{y_{2i}}\right\}$$

$$= \sum_{i=1}^{n} [y_{0i}\, ln\{\pi_0(x_i)\} + y_{1i}\, ln\{\pi_1(x_i)\} + y_{2i}\, ln\{\pi_2(x_i)\}]$$

$$= \sum_{i=1}^{n}\left[y_{0i}\ln\left\{\frac{1}{1+e^{g_1(x_i)}+e^{g_2(x_i)}}\right\}+y_{1i}\ln\left\{\frac{e^{g_1(x_i)}}{1+e^{g_1(x_i)}+e^{g_2(x_i)}}\right\}\right.$$
$$\left.+\ y_{2i}\ln\left\{\frac{e^{g_2(x_i)}}{1+e^{g_1(x_i)}+e^{g_2(x_i)}}\right\}\right]$$

$$= \sum_{i=1}^{n}\left[\ln\left\{\frac{y_{0i}e^{0}+y_{1i}e^{g_1(x_i)}+y_{2i}e^{g_2(x_i)}}{1+e^{g_1(x_i)}+e^{g_2(x_i)}}\right\}\right]$$

$$L(\beta)\ = \sum_{i=1}^{n}[y_{0i}\ ln\{\pi_0(x_i)\}+\ y_{1i}\ ln\{\pi_1(x_i)\}+\ y_{2i}\ ln\{\pi_2(x_i)\}]$$

Maksimum ln-*Likelihood* diperoleh dengan mendiferensikan $L(\beta)$ terhadap β dan menyamakan dengan nol. *Maximum Likelihood Estimator* (MLE) merupakan metode untuk mengestimasi varians dan kovarians dari taksiran β yang diperoleh dari turunan kedua fungsi ln- *Likelihood.* Untuk mendapatkan nilai tersebut digunakan metode iterasi *newton raphson* [11].

Metode *newton raphson* digunakan apabila langkah mengestimasi parameter menggunakan maksimum *likelihood* menghasilkan persamaan yang tidak *closed form* [16]. Metode *newton raphson* adalah salah satu metode untuk mencari akar penyelesaian dari f(x) = 0 melalui perhitungan yang iteratif, sehingga lebih mudah jika dikerjakan dengan bantuan program pada komputer.

Persamaan *likelihood* dengan parameter β dapat diselesaikan sehingga memperoleh nilai estimator $\hat{\beta}$ dengan menggunakan metode *newton raphson.* Rumus estimasi untuk paramter $\hat{\beta}$ pada iterasi ke- (t+1) dalam proses iterasi $t = 0, 1, 2, \ldots$) dituliskan dalam teorema sebagai berikut.

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{\,(t)} - \left(H^{(t)}\right)^{-1} q^{(t)}$$

Berdasarkan rumus tersebut diketahui bahwa $\hat{\beta}^{(t+1)}$ merupakan estimasi parameter β pada iterasi ke (t+1), $\hat{\beta}^{\,(t)}$ estimasi parameter β pada iterasi ke t, $q^{(t)}$ adalah matriks turunan pertama dari fungsi *likelihood* dan $H^{(t)}$ sama dengan matriks turunan kedua fungsi *likelihood*.

dengan H merupakan matriks *Hessian*

$$H = \left(\frac{\partial^2 L(\beta)}{\partial \beta_a \partial \beta_b}\right)$$

$$q^t = \left(\frac{\partial L(\beta)}{\partial \beta_0}, \frac{\partial L(\beta)}{\partial \beta_1}, \cdots, \frac{\partial L(\beta)}{\partial \beta_p}\right)$$

Langkah-langkah metode iterasi *newton raphson* adalah sebagai berikut.

1. Menentukan nilai awal estimasi parameter yaitu $\beta^{(0)}$
2. Mencari matriks *Hessian* $H^{(0)}$ dan matriks $q^{(0)}$
3. Iterasi berlanjut untuk t kurang dari 0
4. Langkah tersebut dilakukan terus menerus sehingga didapatkan estimasi parameter $\hat{\beta}$ yang mencapai kondisi:

$$\left|\frac{\hat{\beta}^{(t+1)} - \hat{\beta}^{(t)}}{\hat{\beta}^{(t)}}\right| \leq d; d > 0$$

Proses iterasi dengan menggunakan metode *newton raphson* hingga didapatkan nilai $\hat{\beta}$ yang konvergen yaitu sampai $\left|\frac{\hat{\beta}^{(t+1)} - \hat{\beta}^{(t)}}{\hat{\beta}^{(t)}}\right|$ kurang dari sama dengan d, dengan d bilangan yang sangat kecil tetapi lebih besar dari 0 [17]. Kendala dalam pemakaian metode *newton raphson* adalah keharusan menghitung nilai turunan fungsi. Tidak mudah dilakukan manual terutama oleh fungsi-fungsi tertentu, sekalipun perhitungan dilakukan dengan kalkulator atau komputer. Oleh karena itu perlu *software* untuk perhitungan.

Parameter yang telah diperoleh perlu diuji signifikansinya, dengan melakukan pengujian statistik. Dalam model regresi logistik terdapat dua jenis pengujian yaitu pengujian secara serentak dan pengujian secara parsial [15].

b. Pengujian Parameter Secara Serentak

Pengujian secara serentak digunakan untuk mengetahui pengaruh variabel prediktor dalam model secara bersama-sama. Hipotesis yang digunakan adalah hipotesis nol (H_0) yaitu $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ (tidak ada

pengaruh variabel prediktor terhadap model) dan hipotesis alternatifnya (H_1) adalah minimal ada satu pengaruh variabel prediktor terhadap model ($\beta_k \neq 0$, $k = 1,2,\dots,p$).

Statistik uji :

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_2}{n}\right)^{n_2}\left(\frac{n_3}{n}\right)^{n_3}}{\prod_{i=1}^{n}[\pi_1(x_i)^{y_{1i}}\pi_2(x_i)^{y_{2i}}\pi_3(x_i)^{y_{3i}}]}\right]$$

dengan

$$n_1 = \sum_{i=1}^{n} y_{1i}\,, n_2 = \sum_{i=1}^{n} y_{2i}\,, n_3 = \sum_{i=1}^{n} y_{3i}\,, dan\ n = n_1 + n_2 + n_3$$

Keterangan dari rumus yang diperoleh yaitu n_1 adalah banyaknya nilai observasi Y sama dengan 1, n_2 banyaknya nilai observasi Y sama dengan 2 dan n_3 merupakan banyaknya nilai observasi Y sama dengan 3.

Statistik uji G^2 mengikuti distribusi *Chi-Square*, sehingga untuk memperoleh keputusan dilakukan perbandingan dengan $X^2_{\propto,db}$ [15]. Kriteria penolakan (Tolak H_0) jika nilai G lebih besar dari $X^2_{\propto,db}$ dimana derajat bebas sama dengan k (banyaknya variabel prediktor).

c. Pengujian Parameter Secara Parsial

Pengujian parsial dilakukan untuk mengetahui apakah variabel prediktor berpengaruh signifikan atau tidak terhadap variabel respon. Uji ini digunakan untuk melihat apakah suatu variabel prediktor layak masuk dalam model [11].

Hipotesi yang digunakan H_0 adalah β_k sama dengan 0 (variabel prediktor tidak berpengaruh signifikan terhadap model) sedangkan H_1 merupakan β_k tidak sama dengan 0, k sama dengan $1,2,\cdots,p$ (variabel prediktor berpengaruh signifikan terhadap model).

Statistik uji :

$$W_k = \frac{\hat{\beta}_k}{S\hat{\mathrm{E}}(\hat{\beta}_k)}$$

Rasio yang dihasilkan dari statistik uji dibawah hipotesis H_0, akan mengikuti distribusi normal baku [15]. Sehingga untuk memperoleh

keputusan dilakukan perbandingan dengan distribusi normal baku (Z). Kriteria penolakan (Tolak H_0) jika nilai $|W_k|$ lebih besar dari $Z_{\propto/2}$.

2.5. Pengujian Kesesuaian Model

Dari estimasi model regresi logistik yang telah diperoleh, selanjutnya dilakukan pengujian kesesuaian model. Statistik uji yang digunakan adalah *Goodness of Fit* [15]. Hipotesis yang digunakan yaitu H_0 merupakan model sesuai (tidak ada perbedaan antara hasil observasi dengan hasil prediksi) sedangkan H_1 adalah model tidak sesuai (terdapat perbedaan antara hasil observasi dengan hasil prediksi).

Statistik Uji yang digunakan :

$$\hat{C} = \sum_{i=1}^{k} \frac{(O_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)} \tag{2.2}$$

Kriteria penolakan adalah tolak H_0 jika $\hat{C}$ lebih besar dari $X^2_{\propto,db}$ dengan derajat bebasnya $db = p - (k+1)$ dimana k adalah jumlah variabel prediktor. Interpetasi model dalam regresi logistik menggunakan nilai *odds ratio* yang menunjukkan perbandingan berapa kali lipat kenaikan atau penurunan angka kejadian Ysama dengan j terhadap Y sama dengan 1 sebagai kategori pembanding jika nilai variabel prediktor (x) berubah sebesar nilai tertentu [15]. Sebagaimana persamaan berikut :

$$\Psi_{ab} = OR_j(a,b) = \frac{P(Y=j|x=a)/P(Y=1|x=a)}{P(Y=j|x=b)/P(Y=1|x=b)}$$

Hubungan antara *odds ratio* terhadap parameter model (β) adalah :

$$\Psi_{ab} = \exp(\hat{\beta})$$

Jika Ψ kurang dari 1 menunjukkan bahwa antar kedua variabel terdapat hubungan negatif dan jika Ψ lebih dari 1 menunjukkan bahwa antar kedua variabel terdapat hubungan positif.

2.6. Evaluasi Ketepatan Klasifikasi

Evaluasi ketepatan klasifikasi merupakan langkah untuk melihat suatu peluang kesalahan yang dilakukan oleh suatu fungsi klasifikasi. Nilai APER

(Apparent Error Rate) menyatakan nilai proporsi sampel yang salah diklasifikasikan oleh fungsi klasifikasi [18]. Penentuan ketepatan pengklasifikasian dapat dilihat dari tabel klasifikasi sebagai berikut.

Tabel 2.2. Perhitungan Ketepatan Pengklasifikasian

Actual membership	(Wichern, 1992)*Predicted Membership*		
	$y = 1$	$y = 2$	$y = 3$
$y = 1$	n_{11}	n_{12}	n_{13}
$y = 2$	n_{21}	n_{22}	n_{23}
$y = 3$	n_{31}	n_{32}	n_{33}

Berdasarkan tabel diatas menjelaskan bahwa n_{11} adalah jumlah Y_i dari y sama dengan 1 tepat diklasifikasikan sebagai y sama dengan 2, n_{12} adalah jumlah Y_i dari y sama dengan 1 tepat diklasifikasikan sebagai y sama dengan 3, n_{13} merupakan jumlah Y_i dari y sama dengan 1 tepat diklasifikasikan sebagai y sama dengan 3, n_{21} merupakan jumlah Y_i dari y sama dengan 2 tepat diklasifikasikan sebagai y sama dengan 1, n_{22} yaitu merupakan jumlah Y_i dari y sama dengan 2 tepat diklasifikasikan sebagai y sama dengan 2, n_{23} adalah jumlah Y_i dari y sama dengan 2 tepat diklasifikasikan sebagai y sama dengan 3, n_{31} adalah jumlah Y_i dari y sama dengan 3 tepat diklasifikasikan sebagai y sama dengan 1, n_{32} merupakan jumlah Y_i dari y sama dengan 3 tepat diklasifikasikan sebagai y sama dengan 2, dan n_{33} adalah jumlah Y_i dari y sama dengan 3 tepat diklasifikasikan sebagai y sama dengan 3.

Tabel 2.2 juga menjelaskan tentang hasil prediksi dari tiap kategori respon dengan y sama dengan 1 sampai y sama dengan 3 di posisi baris merupakan data kategori secara observasi sedangkan y sama dengan 1 sampai y sama dengan 3 di posisi kolom merupakan hasil prediksi data dari kategori variabel respon. Persentase ketepatan model dapat dihitung dengan selisih antara 1 dengan hasil perhitungan APER.

$$APER\ (\%) = \frac{n_{12} + n_{13} + n_{21} + n_{23} + n_{31} + n_{32}}{n_{11} + n_{12} + n_{13} + \ \ldots + n_{33}} \times 100\%$$

2.7. *Odds Ratio*

Odds (ukuran asosiasi pada regresi logistik) adalah rasio peluang antara kejadian sukses dan kejadian tidak sukses dari peubah respon. *Odds Ratio* (OR) merupakan ukuran risiko atau kecenderungan untuk mengalami kejadian 'sukses' antara satu kategori dengan kategori lainnya. Didefinisikan sebagai rasio dari *odds* untuk X sama dengan 1 terhadap X sama dengan 0. *Odds Ratio* ini menyatakan risiko atau kecenderungan pengaruh observasi dengan X sama dengan 1 adalah berapa kali lipat jika dibandingkan dengan observasi dengan X sama dengan 0.

Interpetasi dalam regresi logistik menggunakan nilai *odds ratio* yang menunjukkan perbandingan berapa kali lipat kenaikan atau penurunan angka kejadian Y sama dengan i terhadap Y sama dengan 0 sebagai kategori pembanding jika nilai variabel prediktor (x) berubah sebesar nilai tertentu [19].

$$\Psi_1 = \frac{\pi_1(1)/\pi_0(1)}{\pi_1(0)/\pi_0(0)}$$

$$\Psi_2 = \frac{\pi_2(1)/\pi_0(1)}{\pi_2(0)/\pi_0(0)}$$

Jika 1 kurang dari Ψ menunjukkan bahwa antar kedua variabel terdapat hubungan negatif dan jika 1 lebih dari Ψ menunjukkan bahwa antar kedua variabel terdapat hubungan positif.