BAB II

LANDASAN TEORI

Pada bab ini akan dibahas mengenai dasar-dasar teori yang akan digunakan dalam penulisan skripsi ini, yaitu mengenai data hirarki, model regresi 2-level, model logistik, estimasi parameter model logistik, uji signifikansi parameter dalam model regresi logistik, dan interpretasi parameter dalam model regresi logistik.

2.1 Data Hirarki

Pada berbagai disiplin ilmu, antara lain ilmu sosial dan biologi, sering dijumpai data populasi yang berstruktur hirarki. Data berstruktur hirarki yaitu data yang terdiri dari unit-unit yang diobservasi bersarang atau terkelompokkan dalam unit level yang lebih tinggi. Data hirarki disebut juga data multilevel atau data bersarang.

Sebagai contoh pada suatu penelitian mengenai pendidikan, data murid-murid yang diteliti bersarang pada sekolah-sekolah, selanjutnya sekolah-sekolah tersebut bersarang pada area tempat sekolah-sekolah tersebut berada. Data populasi yang demikian mempunyai struktur hirarki tiga tingkatan, atau disebut data tiga level, dalam hal ini unit level-1 adalah murid,

unit level-2 adalah sekolah, dan unit level-3 adalah area. Pada contoh di atas, jika hanya terdiri dari unit sekolah-sekolah dan unit murid-murid di dalam sekolah-sekolah tersebut, maka data populasi mempunyai struktur hirarki dua-level.

Pada data hirarki, unit-unit level-1 pada unit level-2 yang sama cenderung berkorelasi dibandingkan dengan unit-unit level-1 dari unit level-2 yang berbeda. Sehingga unit-unit pada level-2 yang sama cenderung mempunyai karakteristik yang hampir sama.

2.2 Model Regresi 2-Level

Data yang mempunyai struktur hirarki dapat dianalisis dengan beberapa pendekatan. Jika analisis regresi linier biasa dilakukan untuk menganalisis data hirarki, maka analisis dapat dilakukan pada unit-unit di level-1 saja atau di level-2 saja.

Jika analisis dilakukan pada level-1, struktur hirarki / pengelompokkan data diabaikan (*disaggregated*), artinya model regresi dibentuk dari seluruh data pengamatan level-1. Variasi antar unit-unit level-2 tidak dapat diketahui secara langsung, tapi masih bisa diukur dengan membuat model regresi untuk tiap unit level-2. Untuk jumlah unit level-2 yang sedikit mungkin prosedur penaksiran variasi antar unit-unit level-2 tersebut cukup efisien,

namun jika jumlah unit level-2 cukup banyak akan mengakibatkan banyaknya parameter-parameter yang harus diestimasi dalam model-model regresi yang terbentuk, sehingga prosedur tersebut menjadi tidak efisien.

Jika analisis dilakukan hanya pada unit-unit di level-2 saja (*aggregated*), maka data yang digunakan untuk membuat model regresi adalah rata-rata data respon dan rata-rata data variabel penjelas pada tiap-tiap unit level-2. Analisis dengan cara seperti itu akan mengakibatkan kesalahan interpretasi mengenai hubungan yang terbentuk.

Seperti yang telah dibahas pada subbab 2.1, pada data yang mempunyai struktur hirarki, unit-unit observasi pada level-1 dalam unit level-2 yang sama akan cenderung mempunyai sifat yang hampir sama, sehingga unit-unit observasi tersebut tidak sepenuhnya *independent*. Hal tersebut menjadi alasan mengapa analisis regresi linier biasa kurang tepat digunakan pada data yang mempunyai struktur hirarki.

Kekurangan-kekurangan yang terjadi jika data hirarki dianalisis menggunakan analisis regresi linier biasa dapat diatasi jika data menggunakan analisis multilevel. Untuk analisis data berstruktur dua-level dibutuhkan model regresi dua-level.

Model regresi dua-level dapat digolongkan dalam dua bentuk dasar, yaitu *random intercept model* dan *random slope model*.

2.2.1 *Random Intercept Model*

Random intercept model merupakan salah satu bentuk model regresi 2-level dimana perpotongan (*intercept*) pada model terhadap sumbu-*y* dinyatakan dalam bentuk *random*, tidak *fixed* seperti pada regresi linier biasa, *Intercept* yang berbeda-beda untuk tiap unit level-2 dapat digunakan untuk mengukur perbedaan antar unit level-2. *Random intercept model* dapat direpresentasikan dalam bentuk representasi multilevel sebagai berikut:

- Untuk model level-1, model *random-intercept* ditulis :

$$y_{ij} = \beta_{0j} + \sum_{p=1}^{P} \beta_p x_{pij} + \varepsilon_{ij} \qquad (2.1)$$

dengan

y_{ij} = variabel respon untuk unit ke-*i* pada level-1 dalam unit ke-*j* pada level-2

β_{0j} = *random intercept* untuk unit ke-*j* pada level-2

β_p = efek tetap (*fixed effects*) untuk variabel penjelas ke-*p*

x_{pij} = variabel penjelas ke-*p* di level-1 untuk unit ke-*i* pada level-1 dalam unit ke-*j* pada level-2

ε_{ij} = residual untuk unit ke-*i* pada level-1 dalam unit ke-*j* pada level-2 (residual level-1), diasumsikan berdistribusi $N(0,\sigma_\varepsilon^2)$

- Untuk model level-2 :

$$\beta_{0j} = \beta_0 + u_{0j} \tag{2.2}$$

dengan

β_0 = *fixed intercept*, merupakan rata-rata keseluruhan

u_{0j} = efek random (*error*) untuk unit ke-*j* pada level-2, diasumsikan berdistribusi $N(0,\sigma^2_{u_0})$

ε_{ij} dan u_{oj} diasumsikan saling bebas, $\text{cov}(\varepsilon_{ij}, u_{oj}) = 0$.

Pada model *random intercept*, notasi $j = 1, 2, \ldots, m$ menyatakan unit-unit level-2 dan $i = 1, 2, \ldots, n_j$ menyatakan unit-unit level-1 yang bersarang dalam unit ke-*j* pada level-2. Sehingga total observasi level-1 dalam seluruh unit level-2 adalah :

$$n = \sum_{j=1}^{m} n_j$$

Model (2.2) dapat disubstitusikan ke dalam model (2.1) sehingga model regresi 2-level dengan *random intercept* menjadi

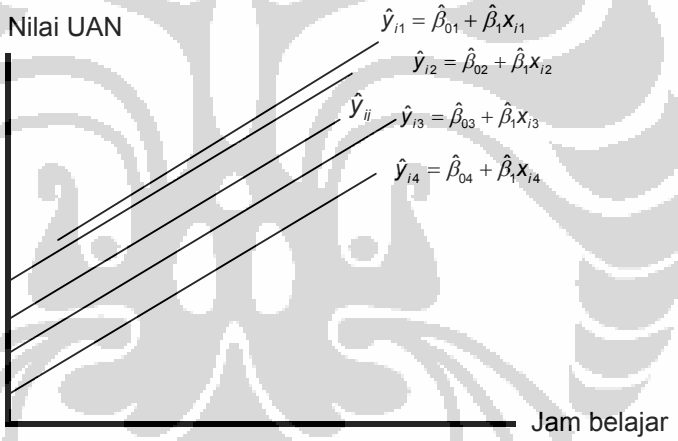
$$y_{ij} = \beta_0 + \sum_{p=1}^{P} \beta_p x_{pij} + u_{0j} + \varepsilon_{ij} \tag{2.3}$$

Model dalam persamaan (2.3) disebut juga sebagai *combine model*.

Parameter-parameter dalam model yang akan ditaksir adalah β_0 dan β_p sebagai *fixed parameter* serta σ^2_{u0} dan σ^2_ε sebagai *random parameter*. σ^2_ε

dan $\sigma^2_{u_0}$ masing-masing menyatakan variansi antar unit level-1 dan variansi antar unit level-2.

Pada contoh data hirarki dalam bidang pendidikan, misalnya ingin diketahui pengaruh jam belajar murid dengan nilai UAN. Jika murid-murid yang diteliti berasal dari empat sekolah, maka ilustrasi contoh data dua level yang dianalisis menggunakan model *random intercept* terlihat seperti pada gambar 1 :



Gambar 1. Model *random intercept*

Pada Gambar 1 masing-masing garis regresi menyatakan garis regresi untuk masing-masing sekolah, dimana terdapat empat sekolah. Sementara garis $\hat{y}_{ij}$ menyatakan model regresi untuk seluruh sekolah.Terlihat pada gambar, bahwa masing-masing sekolah mempunyai *intercept* yang berbeda. Perbedaan tersebut disebabkan oleh efek dari unit level-2 (dalam hal ini sekolah) yang terdapat pada intercept (u_{0j}).

Model (2.3) dapat juga dituliskan dalam bentuk vektor seperti berikut :

$$y_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + u_{0j} + \varepsilon_{ij} \quad (2.4)$$

dengan

y_{ij} = respon untuk unit ke-*i* pada level-1 dalam unit ke-*j* pada level-2

$\mathbf{x}_{ij}^{'}$ = vektor berisi kovariat untuk unit ke-*i* pada level-1 dalam unit ke-*j* pada level-2, berukuran *1x*(*P+1*), $\mathbf{x}_{ij}^{'} = [1 \quad x_{1ij} \quad x_{2ij} \quad \ldots \quad x_{Pij}]$

$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_P \end{bmatrix}$, $\boldsymbol{\beta}$ merupakan vektor berisi parameter-parameter *fixed* yang tidak diketahui, berukuran (*P+1*)*x1*,

ε_{ij} = residual untuk unit ke-*i* pada level-1 dalam unit ke-*j* pada level-2 (residual level-1), diasumsikan berdistribusi $N(0, \sigma_{\varepsilon}^2)$

u_{0j} = efek random (*error*) untuk unit ke-*j* pada level-2, diasumsikan berdistribusi $N(0, \sigma_{u_0}^2)$

Intra-class correlation

Dalam analisis multilevel untuk data 2-level dikenal istilah *intra-class correlation*, yaitu korelasi antar dua unit level-1 dalam unit level-2 yang sama. Seperti yang telah dijelaskan sebelumnya, dalam data yang mempunyai struktur hirarki, dua unit level-1 pada unit level-2 yang sama cenderung mempunyai karakteristik yang hampir sama dibandingkan dengan dua unit level-1 dari unit level-2 yang berbeda. Semakin tinggi nilai korelasi ini

menunjukkan semakin miripnya dua unit level-1 dari unit level-2 yang sama, dibandingkan dengan dua unit level-1 yang diambil dari dua unit level-2 yang berbeda. Hal tersebut mengindikasikan semakin besarnya pengaruh dari unit level-2 pada unit observasi level-1, sehingga penting dilakukan analisis yang memperhatikan struktur hirarki dari data (analisis multilevel).

Pada model regresi 2-level dengan *random intercept*, "*intra-class correlation*" merupakan rasio variansi antar unit level-2 terhadap total variansi. Dalam model regresi 2-level, total variansi adalah penjumlahan dari variansi level-1 dengan variansi level-2 atau dituliskan sebagai $\sigma_{u0}^2 + \sigma_{\varepsilon 0}^2$. Sehingga *intra-class correlation* dapat dinyatakan dalam bentuk :

$$\rho = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_{\varepsilon 0}^2} \quad , \qquad 0 \le \rho \le 1 \tag{2.5}$$

dimana ρ menyatakan *intra-class correlation*, atau disebut juga *intra-level-2-unit correlation*.

2.2.2 *Random Slope Model*

Berbeda dengan *random intercept model*, pada *random slope model* memungkinkan garis-garis regresi untuk tiap unit level-2 mempunyai kemiringan (*slope*) yang berbeda. Representasi multilevel dari *random slope model* dinyatakan dalam bentuk :

- Untuk model level-1 :

$$y_{ij} = \beta_{0j} + \sum_{p=1}^{P} \beta_p x_{pij} + \sum_{q=1}^{Q} \beta_{qj} z_{qj} + \varepsilon_{ij} \qquad (2.6)$$

y_{ij} = variabel respon untuk unit ke-*i* pada level-1 dalam unit ke-*j* pada level-2

β_{0j} = *random intercept* untuk unit ke-*j* pada level-2

β_p = efek tetap (*fixed effects*) untuk variabel penjelas ke-*p*, $p = 1,2,\ldots,P$

β_{qj} = *random slope* untuk variabel penjelas ke-*q* pada unit ke-*j* level-2, $q = 1,2,\ldots,P$

z_{qj} = variabel penjelas ke-*q* dengan $q = 1,2,\ldots,Q$ untuk unit ke-*j* pada level-2

x_{pij} = variabel penjelas ke-*p*, dengan $p = 1, 2, \ldots, P$ untuk unit level-1 ke-*i* dalam unit level-2 ke-*j*

ε_{ij} = residual untuk unit ke-*i* pada level-1 dalam unit ke-*j* pada level-2 (residual level-1), diasumsikan berdistribusi $N(0,\sigma_\varepsilon^2)$

- Untuk model level-2 :

$$\begin{aligned} \beta_{0j} &= \beta_0 + u_{0j} \\ \beta_{qj} &= u_{qj}, \qquad \text{untuk } q = 1,\ldots,Q \end{aligned} \qquad (2.7)$$

β_0 = *fixed intercept*, atau rata-rata keseluruhan

u_{0j} = efek random (*error*) untuk unit ke-*j* pada level-2, diasumsikan berdistribusi $N(0,\sigma_{u_0}^2)$

u_{qj} = efek random dari z_{qj} pada level-2

Pada *random slope model*, j = 1, 2, … , m menyatakan unit-unit level-2 dan i = 1, 2, … , n_j menyatakan unit-unit level-1 yang bersarang dalam tiap unit level-2. Total observasi level-1 dalam seluruh unit level-2 adalah :

$$n = \sum_{j=1}^{m} n_j$$

Secara umum model *random slope* dinyatakan dalam bentuk vektor adalah sebagai berikut :

$$y_{ij} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{j}^{'}\mathbf{u}_j + \varepsilon_{ij} \qquad (2.8)$$

dengan

y_{ij} = respon untuk unit ke-i pada level-1 dalam unit ke-j pada level-2

$\mathbf{x}_{ij}^{'}$ = vektor berisi variabel-variabel penjelas level-1, berukuran 1x(P+1)

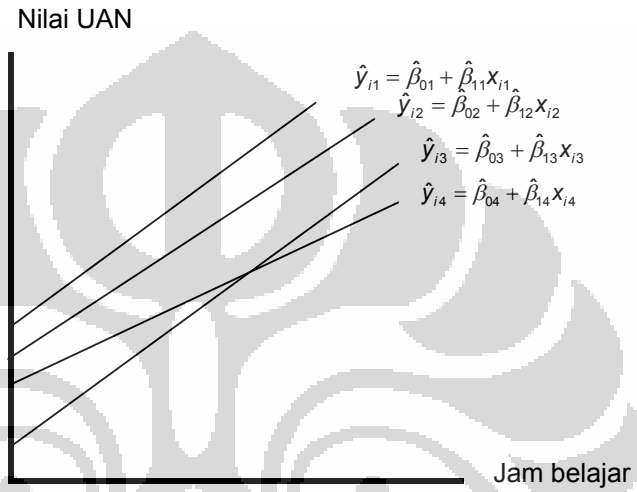
$\boldsymbol{\beta}$ = vektor berisi parameter-parameter *fixed* yang tidak diketahui yang bersesuaian dengan vektor $\mathbf{x}_{ij}^{'}$, berukuran (P+1)$x1$

$\mathbf{z}_{j}^{'}$ = vektor berisi variabel-veriabel penjelas level-2 untuk Q+1 efek random, $\mathbf{z}_{ij}^{'} = [1 \quad z_{1j} \quad z_{2j} \quad \ldots \quad z_{Qj}]$

$\mathbf{u}_j$ = vektor berisi efek random yang bersesuaian dengan vektor $\mathbf{z}_{j}^{'}$, berukuran (Q+1)$x1$, $\mathbf{u}_j = \begin{bmatrix} u_{0j} \\ u_{1j} \\ \vdots \\ u_{Qj} \end{bmatrix}$

ε_{ij} = residual untuk unit ke-i pada level-1 dalam unit ke-j pada level-2 (residual level-1), diasumsikan berdistribusi $N(0,\sigma_{\varepsilon}^{2})$

Menggunakan contoh pada ilmu pendidikan, dimana murid bersarang pada sekolah, maka ilustrasi contoh data dua level yang dianalisis menggunakan model *random slope* terlihat seperti pada gambar 2 :



Gambar 2. Model *random slope*

Pada Gambar 2, masing-masing garis regresi menyatakan garis regresi untuk masing-masing sekolah, dimana terdapat empat sekolah. Terlihat pada gambar, bahwa masing-masing sekolah mempunyai *intercept* dan *slope* yang berbeda. Perbedaan tersebut disebabkan oleh efek dari unit level-2 (dalam hal ini sekolah) yang terdapat pada intercept (u_{0j}) dan slope (u_{1j}) dalam masing-masing sekolah.

2.3 Generalized Least Square

Pada model regresi linier yang dinyatakan dalam persamaan

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \tag{2.9}$$

umumnya diasumsikan $E(\boldsymbol{\varepsilon}) = 0$ dan $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, sehingga parameter-parameter dalam model (2.9) tersebut dapat diestimasi menggunakan metode *Ordinary Least Square* (OLS). Taksiran parameter yang diperoleh dengan OLS adalah :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.10}$$

Namun pada kondisi-kondisi tertentu asumsi-asumsi tersebut tidak dapat terpenuhi, sehingga metode OLS tidak cocok untuk digunakan. Misalnya pada model regresi linier dalam persamaan (2.9) diperoleh variansi yang tidak sama, dinyatakan sebagai $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$, dengan **V** adalah matriks ukuran *nxn*. Interpretasi dari kondisi dimana diasumsikan $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$ adalah jika **V** adalah matriks diagonal namun elemen-elemen diagonalnya tidak sama, artinya observasi-observasi **y** tidak berkorelasi namun memiliki variansi yang tidak sama. Sementara jika terdapat entri-entri non-diagonal utama dari **V** yang tidak nol artinya observasi-observasi **y** dikatakan berkorelasi.

Pada model (2.9) yang mempunyai asumsi $E(\boldsymbol{\varepsilon}) = 0$ dan $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$ akan dilakukan pendekatan dengan mentransformasi model untuk kumpulan observasi supaya observasi-observasinya memenuhi asumsi-asumsi pada

metode *least square,* sehingga OLS dapat digunakan untuk data observasi yang telah ditransformasi.

$\sigma^2\mathbf{V}$ merupakan matriks kovarians dari *error*, maka **V** harus nonsingular dan definit positif sehingga terdapat matriks yang simetris dan nonsingular ukuran *nxn*, **K**, dimana **K'K = KK = V.** Matriks **K** disebut juga akar kuadrat dari matriks **V**.

Didefinisikan variabel-variabel yang baru :

$$\mathbf{z} = \mathbf{K}^{-1}\mathbf{y}, \quad \mathbf{B} = \mathbf{K}^{-1}\mathbf{X}, \quad \mathbf{g} = \mathbf{K}^{-1}\boldsymbol{\varepsilon}$$

Sedemikian sehingga model regresi linier dalam persamaan (2.9) menjadi

$$\mathbf{K}^{-1}\mathbf{y} = \mathbf{K}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{K}^{-1}\boldsymbol{\varepsilon} \tag{2.11}$$

Atau ditulis

$$\mathbf{z} = \mathbf{B}\boldsymbol{\beta} + \mathbf{g} \tag{2.12}$$

Error pada model yang ditransformasi pada persamaan (2.12) mempunyai mean nol,

$E(\mathbf{g}) = \mathbf{K}^{-1}E(\boldsymbol{\varepsilon}) = \mathbf{0}$, dan matriks kovariansi

$$\begin{aligned}
\text{var}(\mathbf{g}) &= E((\mathbf{g} - E(\mathbf{g}))(\mathbf{g} - E(\mathbf{g}))^T) \\
&= E(\mathbf{g}\mathbf{g}^T) \\
&= E(\mathbf{K}^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{K}^{-1}) \\
&= \mathbf{K}^{-1}E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)\mathbf{K}^{-1} \\
&= \sigma^2\mathbf{K}^{-1}\mathbf{V}\mathbf{K}^{-1} \\
&= \sigma^2\mathbf{K}^{-1}\mathbf{K}\mathbf{K}\mathbf{K}^{-1} \\
&= \sigma^2\mathbf{I}
\end{aligned} \tag{2.13}$$

Maka elemen-elemen dari **g** mempunyai mean nol dan variansi konstan dan tidak berkorelasi. Hal tersebut menunjukkan *error* dalam model (2.12)

memenuhi asumsi-asumsi pada metode OLS sehingga dapat diterapkan OLS sebagai estimasi parameter-parameter dalam model. Fungsi *least square*nya adalah

$$\begin{aligned} S(\boldsymbol{\beta}) &= \mathbf{g}^T\mathbf{g} = \boldsymbol{\varepsilon}^T\mathbf{V}^{-1}\boldsymbol{\varepsilon} \\ &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \end{aligned} \tag{2.14}$$

Dan diperoleh persamaan normal *least square*nya, yaitu

$$(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{2.15}$$

Sehingga diperoleh solusi untuk (2.15) :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{2.16}$$

Pada (2.16), $\hat{\boldsymbol{\beta}}$ disebut taksiran *generalized least square* untuk $\boldsymbol{\beta}$.

2.4 Model Regresi Logistik

Misal Y merupakan variabel respon biner yang bernilai 0 atau 1. Nilai Y = 1 menunjukkan suatu karakteristik terjadi dan nilai Y = 0 menunjukkan karakteristik tersebut tidak terjadi. Untuk menganalisis hubungan antara variabel respon Y dengan variabel-variabel penjelas dibentuk suatu model regresi. Misal model yang dibentuk untuk menganalisis hubungan antara variabel respon dengan variabel-variabel penjelasnya ditulis sebagai berikut :

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$
$$= [1 \quad x_{1i} \quad x_{2i} \quad \ldots \quad x_{Pi}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_P \end{bmatrix} + \varepsilon_i \qquad (2.17)$$

Dengan y_i merupakan variabel respon untuk observasi ke-*i,* dan y_i bernilai 0 atau 1. Diasumsikan y_i berdistribusi Bernoulli, dengan probabilitas y_i bernilai 1 adalah $\pi(\mathbf{x}_i)$ dan probabilitas y_i bernilai 0 adalah $1-\pi(\mathbf{x}_i)$, dituliskan $\Pr(y_i = 1) = \pi(\mathbf{x}_i)$ dan $\Pr(y_i = 0) = 1-\pi(\mathbf{x}_i)$. Ekspektasi dari *error* diasumsikan sama dengan nol, sehingga nilai ekspektasi dari y_i adalah :

$$\begin{aligned} E(y_i) &= y_i \Pr(y_i = 1) + y_i \Pr(y_i = 0) \\ &= 1(\pi(\mathbf{x}_i)) + 0(1-\pi(\mathbf{x}_i)) \\ &= \pi(\mathbf{x}_i) \end{aligned}$$

Hal tersebut mengakibatkan $E(y_i) = \mathbf{x}_i'\boldsymbol{\beta} = \pi(\mathbf{x}_i)$. Sehingga persamaan (2.17) dapat ditulis sebagai

$$y_i = \pi(\mathbf{x}_i) + \varepsilon_i \qquad (2.18)$$

Jika data respon biner dianalisis dengan menggunakan model linier seperti pada persamaan (2.18), maka *error* akan mempunyai dua nilai :

$\varepsilon_i = 1-\pi(\mathbf{x}_i)$ ketika y_i bernilai 1.

$\varepsilon_i = -\pi(\mathbf{x}_i)$ ketika y_i bernilai 0.

dan variansi dari observasi y_i adalah :

$$\begin{aligned}\sigma_{y_i}^2 &= E(y_i - E(y_i))^2 \\ &= (1-\pi(\mathbf{x}_i))^2\pi(\mathbf{x}_i) + (0-\pi(\mathbf{x}_i))^2(1-\pi(\mathbf{x}_i)) \\ &= \pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i)) \\ &= E(y_i)(1-E(y_i))\end{aligned}$$

Hal tersebut di atas menunjukkan ε_i tidak berdistribusi normal dan mempunyai variansi yang tidak tetap. Sehingga asumsi variansi sama pada model regresi linier tidak terpenuhi, artinya model regresi linier bukan model yang tepat untuk menganalisis data respon biner.

Pada data respon biner, nilai $E(y_i)$, bernilai antara 0 dan 1. Untuk mencari hubungan antara variabel respon yang biner dengan variabel-variabel penjelasnya sama dengan melakukan analisis hubungan antara $E(y_i)$ dengan variabel-variabel penjelas, atau ditulis $E(y_i)=\pi(\mathbf{x}_i)=\mathbf{x}_i'\boldsymbol{\beta}$. Tidak seperti nilai $E(y_i)$, $\mathbf{x}_i'\boldsymbol{\beta}$ tidak hanya bernilai antara 0 dan 1. Supaya nilai $\mathbf{x}_i'\boldsymbol{\beta}$ terletak antara 0 dan 1, diasumsikan hubungan antara $\pi(\mathbf{x}_i)$ dengan variabel-variabel penjelasnya mengikuti bentuk fungsi distribusi logistik, sehingga bentuk model regresi logistik adalah

$$\pi(\mathbf{x}_i) = \frac{1}{1+\exp(-(\beta_0+\beta_1 x_{1i}+...+\beta_P x_{Pi}))} = \frac{\exp(\beta_0+\beta_1 x_{1i}+...+\beta_P x_{Pi})}{1+\exp(\beta_0+\beta_1 x_{1i}+...+\beta_P x_{Pi})} \quad (2.19)$$

Dari persamaan (2.19) diperoleh

$$1-\pi(\mathbf{x}_i) = \frac{1}{1+\exp(\beta_0+\beta_1 x_{1i}+...+\beta_P x_{Pi})}$$

Rasio antara $\pi(\mathbf{x}_i)$ dengan $1-\pi(\mathbf{x}_i)$ dituliskan sebagai :

$$\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} = \frac{\left(\frac{\exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_P x_{Pi})}{1+\exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_P x_{Pi})}\right)}{\left(\frac{1}{1+\exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_P x_{Pi})}\right)} = \exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_P x_{Pi}) \quad (2.20)$$

Rasio $\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}$ pada persamaan (2.20) disebut sebagai *odds ratio*.

Bentuk $\pi(\mathbf{x}_i)$ pada persamaan (2.19) dapat dilakukan transformasi logit, yang didefinisikan sebagai :

$$g(\mathbf{x}_i) = \log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \beta_0 + \beta_1 x_{1i} + ... + \beta_P x_{Pi} \quad (2.21)$$

atau jika dinyatakan dalam bentuk vektor persamaan (2.21) menjadi

$$g(\mathbf{x}_i) = \log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \mathbf{x}_i'\boldsymbol{\beta} \quad (2.22)$$

dimana $g(\mathbf{x}_i)$ merupakan transformasi *logit*, dan dengan dilakukannya transformasi *logit*, $g(\mathbf{x}_i)$ mempunyai hubungan linier dengan parameter-parameternya.

2.3.1 Estimasi Parameter Model Logistik

Estimasi parameter-parameter dalam model logistik salah satunya dapat dilakukan dengan metode *maximum likelihood*. Penaksiran parameter dengan menggunakan metode *maximum likelihood* diperoleh dengan mencari taksiran parameter yang memaksimumkan fungsi *likelihood*.

Diketahui probabilitas bersyarat untuk respon dinyatakan $\Pr(Y=1|\mathbf{x})=\pi(\mathbf{x})$ dan $\Pr(Y=0|\mathbf{x})=1-\pi(\mathbf{x})$. Misal terdapat *n* buah observasi yang saling bebas. Y_i menyatakan variabel respon dari observasi ke-*i*, dimana *i* = 1,2,…,*n*. Diketahui probabilitas untuk $y_i=1$ (suatu karakteristik terjadi pada observasi ke-*i*) adalah $\pi(\mathbf{x}_i)$ dan probabilitas untuk $y_i=0$ adalah $1-\pi(\mathbf{x}_i)$. Maka pdf dari Y_i adalah

$$f(y_i)=[\pi(\mathbf{x}_i)]^{y_i}[1-\pi(\mathbf{x}_i)]^{1-y_i}$$

Karena observasi saling bebas, maka fungsi *likelihood* dapat diperoleh dengan mengalikan fungsi-fungsi kepadatan (pdf) dari Y_i

$$L(\boldsymbol{\beta})=f(y_1,...,y_n)=\prod_{i=1}^{n}[\pi(\mathbf{x}_i)]^{y_i}[1-\pi(\mathbf{x}_i)]^{1-y_i}$$

dimana $\boldsymbol{\beta}$ merupakan vektor berisi parameter-parameter tidak diketahui yang ingin ditaksir, $\boldsymbol{\beta}'=(\beta_0,\beta_1,...,\beta_P)$.

Untuk memudahkan mencari nilai $\beta_0,\beta_1,...,\beta_P$ yang memaksimumkan fungsi *likelihood* digunakan bentuk logaritma natural dari fungsi *likelihood*, yang kemudian disebut sebagai fungsi *log-likelihood*, yaitu :

$$l(\boldsymbol{\beta})=\ln\left(L(\boldsymbol{\beta})\right)$$

$$=\sum_{i=1}^{n}\left[y_i\ln(\pi(\mathbf{x}_i))+(1-y_i)\ln(1-\pi(\mathbf{x}_i))\right]$$

$$=\sum_{i=1}^{n}\left[y_i\ln\left(\frac{\exp(g(\mathbf{x}_i))}{1+\exp(g(\mathbf{x}_i))}\right)+(1-y_i)\ln\left(1-\left(\frac{\exp(g(\mathbf{x}_i))}{1+\exp(g(\mathbf{x}_i))}\right)\right)\right]$$

$$= \sum_{i=1}^{n}\left[y_i \left\{ \ln(\exp(g(\mathbf{x}_i))) - \ln\left(1+\exp(g(\mathbf{x}_i))\right)\right\} + (1-y_i)\left\{\ln(1) - \ln(1+\exp(g(\mathbf{x}_i)))\right\}\right]$$

$$= \sum_{i=1}^{n}\left[y_i \left\{ \ln(\exp(g(\mathbf{x}_i))) - \ln\left(1+\exp(g(\mathbf{x}_i))\right)\right\} - (1-y_i)\ln(1+\exp(g(\mathbf{x}_i)))\right]$$

$$= \sum_{i=1}^{n}\left[y_i \ln(\exp(g(\mathbf{x}_i))) - y_i \ln\left(1+\exp(g(\mathbf{x}_i))\right) - \ln(1+\exp(g(\mathbf{x}_i))) + y_i \ln(1+\exp(g(\mathbf{x}_i)))\right]$$

$$= \sum_{i=1}^{n}\left[y_i g(\mathbf{x}_i) - y_i \ln(1+\exp(g(\mathbf{x}_i))) - \ln(1+\exp(g(\mathbf{x}_i))) + y_i \ln(1+\exp(g(\mathbf{x}_i)))\right]$$

$$= \sum_{i=1}^{n}\left[y_i g(\mathbf{x}_i) - \ln(1+\exp(g(\mathbf{x}_i)))\right] \qquad (2.23)$$

dengan $g(\boldsymbol{x})$ seperti pada persamaan (2.22), sehingga (2.23) menjadi

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[y_i \mathbf{x}_i'\boldsymbol{\beta} - \ln(1+\exp(\mathbf{x}_i'\boldsymbol{\beta}))\right]$$

$$= \sum_{i=1}^{n}\left[y_i (\beta_0 + \beta_1 x_{1i} + \ldots + \beta_P x_{Pi}) - \ln(1+\exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_P x_{Pi}))\right]$$

Untuk mendapatkan nilai $\boldsymbol{\beta}$ yang memaksimalkan fungsi *log-likelihood*, diferensialkan fungsi *log-likelihood* terhadap β_p, p = 0,1, … ,P dan menyamakannya dengan nol, sehingga diperoleh persamaan *likelihood* sebagai berikut :

$$\frac{\partial \mathrm{l}(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n}\left[y_i - \frac{1}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}\exp(\mathbf{x}_i'\boldsymbol{\beta})\right]$$

$$= \sum_{i=1}^{n}\left[y_i - \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right]$$

$$= \sum_{i=1}^{n}\left[y_i - \pi(\mathbf{x}_i)\right] = 0$$

$$\frac{\partial \mathrm{l}(\boldsymbol{\beta})}{\partial \beta_1} = \sum_{i=1}^{n}\left[y_i x_{1i} - \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})} x_{1i} \right]$$

$$= \sum_{i=1}^{n}\left[y_i x_{1i} - \pi(\mathbf{x}_i) x_{1i} \right] = 0$$

$$\vdots$$

$$\frac{\partial \mathrm{l}(\boldsymbol{\beta})}{\partial \beta_P} = \sum_{i=1}^{n}\left[y_i x_{Pi} - \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})} x_{Pi} \right]$$

$$= \sum_{i=1}^{n}\left[y_i x_{Pi} - \pi(\mathbf{x}_i) x_{Pi} \right] = 0$$

Karena persamaan-persamaan *likelihood* yang diperoleh di atas tidak linier dalam β_p, maka perlu dilakukan perhitungan menggunakan metode numerik untuk mendapatkan taksiran dari β_p, yang dinyatakan dalam $\hat{\beta}_p$ dengan $p = 0,1, \ldots, P$.

Taksiran dari variansi dan kovariansi diperoleh dari turunan parsial kedua fungsi *likelihood*. Bentuk turunan parsial kedua dari fungsi *log-likelihood* adalah :

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_0^2} = -\sum_{i=1}^{n} \pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i))$$

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_0} = -\sum_{i=1}^{n} \pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i)) x_{1i}$$

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_P \partial \beta_0} = -\sum_{i=1}^{n} \pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i)) x_{Pi}$$

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_P} = -\sum_{i=1}^{n} \pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i)) x_{1i} x_{Pi}$$

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_P^2} = -\sum_{i=1}^{n} \pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i))x_{Pi}^2$$

Bentuk umum dari turunan parsial kedua fungsi *log-likelihood* adalah

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_p^2} = -\sum_{i=1}^{n} x_{pi}^2 \pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i)) \tag{2.24}$$

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_r \partial \beta_p} = -\sum_{i=1}^{n} x_{pi} x_{ri} \pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i)) \tag{2.25}$$

dimana $p, r = 0,1,2,...,P$.

Dari turunan parsial kedua fungsi *log-likelihood* dapat dibentuk sebuah matriks berukuran (P+1) x (P+1) yang isinya merupakan elemen-elemen negatif dari nilai-nilai dalam persamaan (2.24) dan (2.25). Matriks yang demikian disebut sebagai matriks informasi yang dinyatakan dengan $\mathbf{I}(\boldsymbol{\beta})$, yang bentuknya :

$$\mathbf{I}(\boldsymbol{\beta}) = \begin{bmatrix} -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_0^2} & -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_1 \beta_0} & \cdots & -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_P \beta_0} \\ -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_0 \beta_1} & -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_1^2} & \cdots & -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_P \beta_1} \\ \vdots & \vdots & & \vdots \\ -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_0 \beta_P} & -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_1 \beta_P} & \cdots & -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_P^2} \end{bmatrix}$$

Untuk mengetahui variansi dan kovariansi dari taksiran parameter dibentuk suatu matriks yang merupakan invers dari matriks informasi. Sehingga matriks taksiran variansi kovariansi dari $\hat{\boldsymbol{\beta}}$, yaitu $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})$, diperoleh dengan menginverskan taksiran matriks informasi, $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$.

Elemen diagonal utama ke-*p* dari matriks taksiran variansi kovariansi $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})$ menunjukkan taksiran variansi $\hat{\beta}_p$, yaitu $\hat{\sigma}^2_{\beta_p}$, dan elemen-elemen non-diagonalnya menunjukkan taksiran kovariansi dari $\hat{\beta}_p$ dan $\hat{\beta}_r$, yaitu $\hat{\sigma}_{(\beta_p,\beta_r)}$. Akar kuadrat dari $\hat{\sigma}^2_{\beta_p}$, yaitu $\hat{\sigma}_{\beta_p}$, merupakan taksiran standar *error* dari $\hat{\beta}_p$.

2.3.2 Uji Signifikansi Parameter Dalam Model Regresi Logistik

Pengujian signifikansi masing-masing parameter dalam model regresi logistik dapat dilakukan dengan menggunakan uji *Wald*.

Hipotesis pengujian parameter :

$H_0 : \beta_p = 0$, $p = 1,2,...,P$

$H_1 : \beta_p \neq 0$

Statistik uji Wald untuk menguji hipotesis di atas adalah :

$$W_p = \left[\frac{\hat{\beta}_p}{\hat{\sigma}_{\hat{\beta}_p}} \right]^2$$

dimana $W_p \sim \chi^2_{\alpha,1}$, dan α merupakan tingkat signifikansi.

Aturan keputusan : H_0 ditolak jika nilai W_p memenuhi $W_p \sim \chi^2_{\alpha,1}$. Jika H_0 ditolak artinya parameter yang diuji, β_p, signifikan pada tingkat signifikansi α. Hal tersebut menunjukkan bahwa variabel independen yang bersesuaian

dengan parameter β_p,yaitu x_p, berpengaruh secara signifikan terhadap variabel respon Y.

2.3.3 Interpretasi Parameter Dalam Model Regresi Logistik

Model regresi logistik berbentuk :

$$g(\mathbf{x}_i) = \log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_P x_{Pi}$$

Variabel-variabel independen dalam model regresi logistik seperti bentuk di atas dapat berupa variabel independen kontinu maupun kategorik. Terdapat beberapa cara dalam menginterpretasikan parameter-parameter dalam model regresi logistik dengan variabel independen berjenis kontinu ataupun kategorik.

Interpretasi Parameter Untuk Variabel Independen Kontinu

Untuk model regresi logistik dengan variabel independen kontinu, parameter β_p menunjukkan selisih logit untuk setiap perubahan 1 unit satuan nilai variabel independen kontinu x_{pi} (variabel independen kontinu ke-*p* untuk observasi ke-*i*, dengan *p* = 1,…,*P*) dan nilai-nilai variabel independen yang lainnya tetap.

Cara lain untuk menginterpretasikan parameter dalam model regresi logistik adalah dengan *odds ratio*. *Odds ratio* seperti ditunjukkan dalam persamaan (2.20) merupakan perbandingan dari dua nilai *odd*. Dari persamaan (2.20) diperoleh :

$$\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} = e^{\beta_0+\beta_1 x_{1i}+\ldots+\beta_P x_{Pi}} = e^{\beta_0} e^{\beta_1 x_{1i}} \ldots e^{\beta_P x_{Pi}}$$

Dari persamaan di atas, *odd* untuk variabel independen kontinu ke-*j* untuk observasi ke-*i* adalah :

$$\frac{\pi(x_{pi})}{1-\pi(x_{pi})} = e^{\beta_p x_{pi}} \tag{2.26}$$

dan *odd* untuk variabel independen kontinu ke-*p* yang nilainya bertambah satu unit satuan dibandingkan sebelumnya adalah

$$\frac{\pi(x_{pi}+1)}{1-\pi(x_{pi}+1)} = e^{\beta_p (x_{pi}+1)} \tag{2.27}$$

sehingga *odds ratio* dari *odd* pada persamaan (2.26) dengan *odd* pada persamaan (2.27) adalah :

$$\frac{\left(\frac{\pi(x_{pi})}{1-\pi(x_{pi})}\right)}{\left(\frac{\pi(x_{pi}+1)}{1-\pi(x_{pi}+1)}\right)} = e^{\beta_p} \tag{2.28}$$

Maka interpretasi *odd ratio* dari persamaan (2.28) atau interpretasi parameter β_p adalah untuk setiap kenaikan 1 unit satuan nilai variabel independen kontinu x_{pi} resiko terjadinya suatu karakteristik tertentu (Y=1)

akan naik sebesar e^{β_p} kali dibandingkan sebelumnya, dengan asumsi nilai-nilai variabel independen lainnya tetap.

Interpretasi Parameter Untuk Variabel Independen Kategorik

Untuk model regresi logistik dengan variabel independen kategorik, interpretasi parameter β_p dapat dilakukan dengan mencari *odds ratio*. *Odds ratio* didapat dari membandingkan nilai *odd* dari suatu kategori terhadap nilai *odd* dari kategori acuan suatu variabel independen kategorik.

Misal x_{pi} adalah variabel independen biner (mempunyai dua kategori) ke-*j* untuk observasi ke-*i*. *Odds ratio* antara *odd* untuk kategori 1 ($x_{pi}=1$) terhadap *odd* untuk kategori acuan ($x_{pi}=0$) adalah:

$$\frac{\left(\frac{\pi(x_{pi}=1)}{1-\pi(x_{pi}=1)}\right)}{\left(\frac{\pi(x_{pi}=0)}{1-\pi(x_{pi}=0)}\right)}=e^{\beta_p} \qquad (2.29)$$

Maka interpretasi *odd ratio* dari persamaan (2.29) atau interpretasi parameter β_p adalah resiko terjadinya suatu karakteristik tertentu (Y=1) untuk x_{pi} kategori 1 ($x_{pi}=1$) adalah sebesar e^{β_p} kali dibandingkan resiko terjadinya suatu karakteristik tertentu untuk x_{pi} kategori acuan ($x_{pi}=0$), dengan asumsi nilai-nilai variabel independen lainnya tetap.