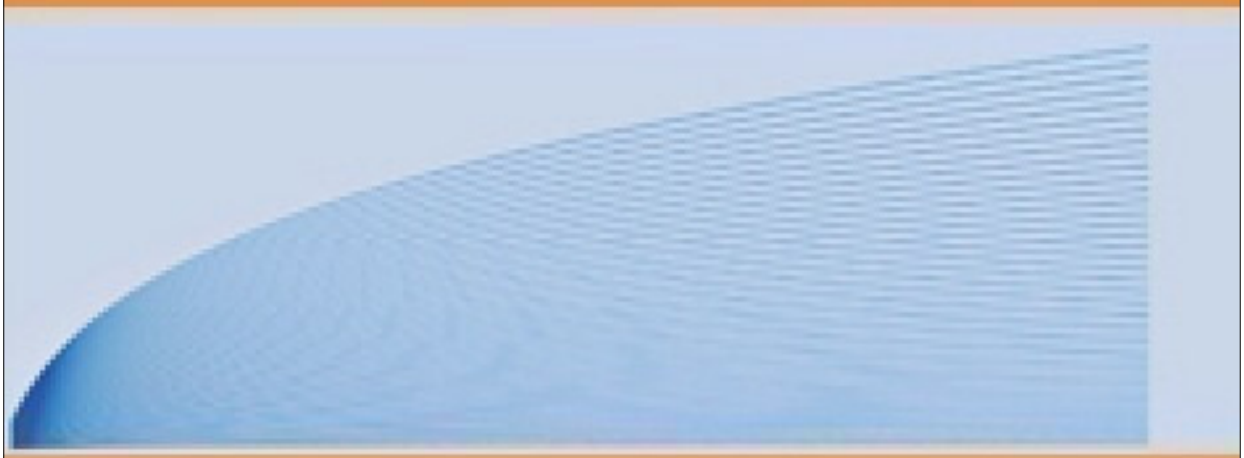
Johan Harlan

Analisis Regresi Logistik



Penerbit Gunadarma

ANALISIS REGRESI LOGISTIK

Johan Harlan

Penerbit Gunadarma

Analisis Regresi Logistik
Penulis : Johan Harlan
Cetakan Pertama, Agustus 2018

Disain cover : Joko Slameto

Diterbitkan pertama kali oleh Gunadarma
Jl. Margonda Raya No. 100, Pondokcina, Depok 16424
Telp. +62-21-78881112, 7863819 Faks. +62-21-7872829
e-mail : sektor@gunadarma.ac.id

KATA PENGANTAR

Analisis regresi logistik merupakan materi ajar Statistika yang sangat penting bagi mahasiswa maupun para peneliti, dapat disetarakan dengan keperluan mempelajari dan memahami analisis regresi linear. Walaupun demikian, di Indonesia umumnya regresi logistik tidak masuk dalam kurikulum Statistika bagi mahasiswa umum S1, karena pengajaran Statistika untuk mahasiswa S1 di Indonesia, terutama untuk mahasiswa jurusan non-eksakta, hampir seluruhnya masih berbasis manual, sedangkan regresi logistik dengan prosedur iteratif dan metode *maximum likelihood* praktis tidak mungkin dikerjakan tanpa bantuan komputer.

Dalam buku ini penulis berusaha membahas mengenai regresi logistik dari tingkat dasar sampai tingkat lanjut sederhana, dengan sedapat mungkin menghindari pembahasan dasar-dasar Statistika yang bersifat matematis. Pembaca diharapkan sudah terlebih dahulu memahami dasar-dasar dan aplikasi analisis regresi linear, karena dalam pembahasan regresi logistik acapkali digunakan penjelasan yang bersifat analogi dengan metode analisis regresi linear. Semua contoh-contoh yang dibahas dalam buku ini dibahas dengan menggunakan program komputer statistik Stata 15.

Penulis sangat mengharapkan saran dan kritik dari pembaca demi perbaikan kekurangan yang ada dalam isi buku ini.

Agustus 2018

Johan Harlan

DAFTAR ISI

BAB 1
BEBERAPA KONSEP DASAR

❖ Odds

Probabilitas (peluang) adalah pernyataan kuantitatif mengenai kemungkinan suatu kejadian akan terjadi. Ukuran probabilitas dikaitkan dengan suatu kejadian Y dan dinyatakan sebagai $P(Y)$ yang bernilai $0 \le P(Y) \le 1$. Odds suatu kejadian *Y*, dinyatakan sebagai $O(Y)$, adalah rasio probabilitas antara 2 *outcome* suatu variabel biner, yaitu rasio antara probabilitas terjadinya suatu kejadian *Y* dengan probabilitas tidak terjadinya kejadian *Y* tersebut:

$$\boxed{O(Y) = \frac{P(Y)}{1-P(Y)}} \qquad (1.1)$$

Jika peristiwa terjadinya suatu kejadian *Y* dinyatakan dengan nilai $Y = 1$ dan peristiwa tidak terjadinya kejadian *Y* dengan nilai $Y = 0$, maka odds kejadian *Y* adalah:

$$O(Y=1) = \frac{P(Y=1)}{1-P(Y=1)}$$

dan odds tidak terjadinya kejadian *Y* adalah:

$$O(Y=0) = \frac{P(Y=0)}{1-P(Y=0)} = \frac{1-P(Y=1)}{P(Y=1)} = \frac{1}{O(Y=1)}$$

Contoh 1.1:

Misalkan dimiliki data imaginer tentang 2 kesebelasan sepakbola ABC dan PQR. Data lampau menyatakan bahwa kedua kesebelasan pernah bertanding 10 kali dengan kemenangan 7 kali bagi kesebelasan ABC dan 3 kali bagi kesebelasan PQR. Untuk pertemuan kesebelas berikutnya prediksi probabilitas kemenangan ABC adalah:

$$P\,(\text{ABC} = 1) = \frac{7}{10}$$

Prediksi probabilitas kekalahan ABC adalah:

$$P\,(\text{ABC} = 0) = \frac{3}{10}$$

Sedangkan prediksi odds kemenangan ABC adalah:

$$O\big(\text{ABC} = 1\big) = \frac{7/10}{3/10} = \frac{7}{3}$$

Prediksi odds kekalahan ABC adalah:

$$O\big(\text{ABC} = 0\big) = \frac{3}{7} = \frac{1}{O\big(\text{ABC} = 1\big)}$$

❖ Log Odds

Log odds, dengan menggunakan konstante Euler ($e \approx 2.718$) sebagai bilangan pokok logaritma naturalis, lazimnya dituliskan sebagai ln odds. Log odds kejadian Y, disebut juga logit Y adalah:

$$\boxed{\text{logit } Y = \ln \text{ odds } Y = \ln \frac{P\big(Y = 1\big)}{1 - P\big(Y = 1\big)}} \qquad (1.2)$$

Pada tabel 1.1 berikut diperlihatkan beberapa nilai probabilitas Y $\left[P(Y)\right]$, odds Y, dan logit Y [ln odds Y].

Tabel 1.1 Beberapa nilai probabilitas *Y*, odds *Y*, dan logit *Y*

$P(Y=1)$	$O(Y=1)$	logit Y
0.01	0.01	–4.60
0.05	0.05	–2.94
0.10	0.11	–2.20
0.20	0.25	–1.39
0.50	1.00	0.00
0.80	4.00	1.39
0.90	9.00	2.20
0.95	19.0	2.94
0.99	99.0	4.60

Tampak bahwa rentang nilai probabilitas Y adalah $0 \leq P(Y=1) \leq 1$, rentang nilai odds $O(Y=1)$ adalah $0 \leq O(Y=1) \leq \infty$, sedangkan rentang nilai logit Y adalah $-\infty \leq$ logit $Y \leq \infty$. Tampak pula bahwa logit Y berdistribusi simetris dengan nilai nol (*null value*) sama dengan nol, sedangkan odds $O(Y=1)$ berdistribusi menceng ke kanan (*skewed to the right*).

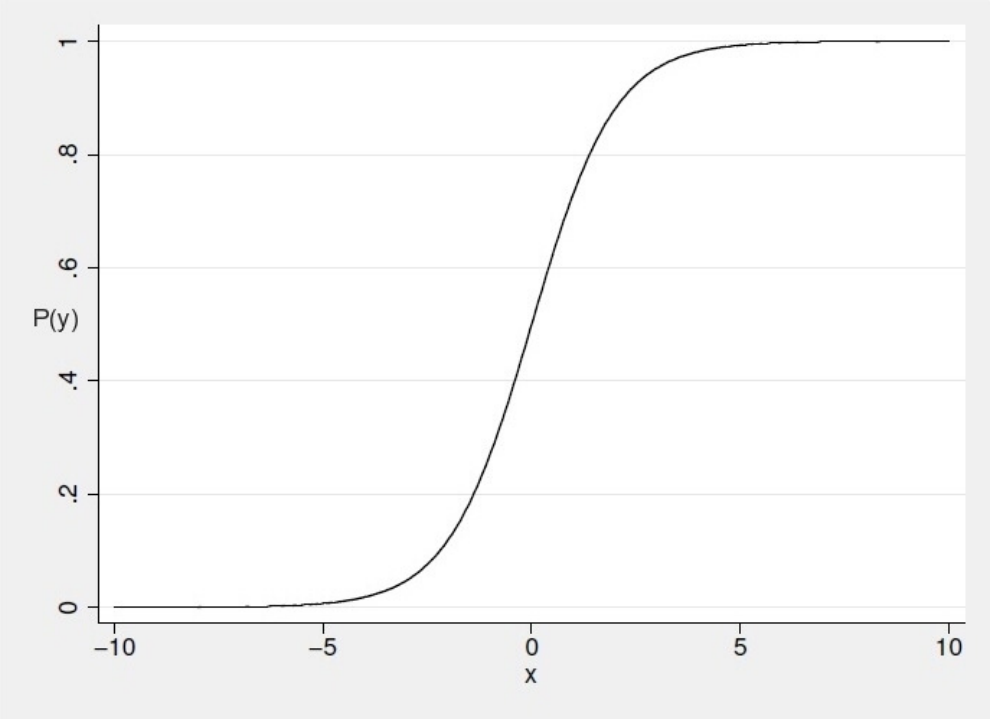
Dalam model regresi logistik, karena sifat-sifatnya tersebut logit Y dijadikan suku transformasi variabel dependen Y pada ruas kanan persamaan dengan ruas kiri berupa kombinasi linear variabel independen X:

$$\text{logit } Y = \ln \frac{P(Y=1)}{1-P(Y=1)} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \qquad (1.3)$$

Hasil yang diperoleh dari regresi logistik dalam logit Y dapat dikembalikan ke dalam bentuk probabilitas dengan persamaan:

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\ldots+\beta_p X_p)}} \qquad (1.4)$$

Grafik $P(Y=1)$ ini berupa kurve sigmoid (menyerupai huruf 'S') seperti terlihat pada gambar 1.1 berikut ini.



Gambar 1.1 Kurve sigmoid $P(Y=1)$

❖ Rasio Odds

Pada studi epidemiologi dengan **prediktor biner** sebagai variabel independen dan respons yang juga biner sebagai variabel dependen, ringkasan data dapat disajikan dalam bentuk tabel 2×2 berikut:

Tabel 1.2 Tabel 2×2 untuk prediktor biner

X = Prediktor	Y = Respons		Jumlah
	1 = Ada	0 = Tidak ada	
1 = Ada	a	b	n_1
0 = Tidak ada	c	d	n_2
Jumlah	m_1	m_2	n

Odds bersyarat Y, yaitu odds Y dengan syarat prediktor X ada ialah:

$$\hat{O}\left(Y \middle| X = 1\right) = a/b$$

Sedangkan odds Y dengan syarat prediktor tidak ada yaitu:

$$\hat{O}\left(Y \middle| X = 0\right) = c/d$$

Rasio antara keduanya dinamakan rasio odds (*odds ratio*), sebagai estimasi untuk nilai rasio odds dalam populasi, yaitu:

$$\boxed{\hat{OR} = \frac{a/b}{c/d} = \frac{ad}{bc}} \qquad (1.5.a)$$

Untuk **prediktor kontinu**, rasio odds dihitung sebagai rasio odds untuk dua keadaan dengan perubahan 1 satuan satuan variabel independen, dengan asumsi rasio ini konstan di sepanjang perubahan nilai variabel independen, yang ringkasan datanya disajikan pada tabel 1.3 berikut:

Tabel 1.3 Tabel 2×2 untuk prediktor kontinu

X = Prediktor	Y = Respons		Jumlah
	1 = Ada	0 = Tidak ada	
$X = x + 1$	a	b	n_1
$X = x$	c	d	n_2
Jumlah	m_1	m_2	n

Rasio odds untuk prediktor kontinu adalah:

$$\boxed{\hat{OR} = \frac{ad}{bc}} \qquad (1.5.b)$$

Pada model regresi logistik dengan 1 **prediktor biner** X, yaitu:

$$\text{logit } Y = \beta_0 + \beta_1 X$$

logit bersyarat Y untuk $X = 1$ dan $X = 0$ masing-masing adalah:

$$\text{logit } \left(Y \middle| X = 1\right) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$$

dan: $$\text{logit } \left(Y \middle| X = 0\right) = \beta_0 + \beta_1 \times 0 = \beta_0$$

Odds-nya masing-masing adalah:

$$O\left(Y \middle| X = 1\right) = \exp\left(\beta_0 + \beta_1\right)$$

dan: $$O\left(Y \middle| X = 0\right) = \exp\left(\beta_0\right)$$

Rasio odds-nya adalah:

$$\boxed{\hat{OR} = \frac{\exp\left(\beta_0 + \beta_1\right)}{\exp\left(\beta_0\right)} = e^{\beta_1}} \qquad (1.6.a)$$

Untuk model regresi logistik dengan 1 **prediktor kontinu** X, logit bersyarat Y untuk $X = x + 1$ dan $X = x$ masing-masing adalah:

$$\text{logit } \left(Y \middle| X = x+1\right) = \beta_0 + \beta_1 \times (x+1) = \beta_0 + \beta_1 + \beta_1 x$$

dan: $$\text{logit } \left(Y \middle| X = x\right) = \beta_0 + \beta_1 \times (x) = \beta_0 + \beta_1 xi$$

Odds-nya masing-masing adalah:

$$O\left(Y \middle| X = x+1\right) = \exp\left(\beta_0 + \beta_1 + \beta_1 x\right)$$

dan: $$O\left(Y \middle| X = x\right) = \exp\left(\beta_0 + \beta_1 x\right)$$

Rasio odds-nya adalah:

$$\boxed{\hat{OR} = \frac{\exp\left(\beta_0 + \beta_1 + \beta_1 x\right)}{\exp\left(\beta_0 + \beta_1 x\right)} = e^{\beta_1}} \qquad (1.6.b)$$

Untuk model regresi logistik dengan p variabel independen, yaitu:

$$\text{logit } Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

rasio odds dapat dihitung untuk masing-masing variabel independen, misalnya **untuk variabel independen ke-*j*** X_j adalah:

$$\boxed{\hat{OR}_j = e^{\beta_j}} \qquad (1.7)$$

BAB 2
REGRESI LOGISTIK DENGAN STATA

❖ Regresi Logistik Sederhana

Model regresi logistik sederhana adalah model regresi logistik dengan 1 prediktor variabel kontinu atau variabel indikator, yang dinyatakan sebagai:

$$\text{logit}\,(Y) = \beta_0 + \beta_1 X \tag{2.1}$$

Variabel indikator adalah variabel dengan nilai 0 atau 1. *Y* adalah respons biner yang juga bernilai 0 atau 1. Logit *Y* adalah:

$$\text{logit}\,(Y) = \ln \text{odds}\,(Y)$$

$$= \ln \frac{P(Y=1)}{1-P(Y=1)}$$

Berbeda dengan model regresi linear, pada ruas kanan persamaan model regresi logistik tidak ada suku galat. Selanjutnya diperoleh:

$$\ln \text{odds}\, Y = \beta_0 + \beta_1 X_1$$

$$\text{odds}\, Y = e^{\beta_0+\beta_1 X}$$

atau:
$$\frac{P(Y=1)}{1-P(Y=1)} = e^{\beta_0+\beta_1 X}$$

dan:
$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}} \tag{2.2}$$

Perintah Stata untuk melakukan analisis regresi logistik dan mengestimasi koefisien regresi logistik adalah:

logit *depvar indepvar* [*if*] [*in*] [, *options*]

depvar : Respons biner

indepvar : (Himpunan) prediktor

Perintah Stata untuk melakukan analisis regresi logistik sederhana dan mengestimasi rasio odds adalah:

logistic *depvar indepvar* [*if*] [*in*] [, *options*]

depvar : Respons biner

indepvar : Prediktor

Estimasi rasio odds juga dapat diperoleh dengan perintah:

logit *depvar indepvar* [*if*] [*in*], **or** [*options*]

Contoh 2.1 (prediktor biner):

. use "D:\Analisis Regresi Logistik\Data\apilog.dta", clear

File ini memuat data 1200 sekolah menengah di negara bagian California, Amerika Serikat. **api00** adalah variabel kontinu yang menyatakan pencapaian akademik tiap sekolah.

. sum api00

```
Variable |   Obs       Mean   Std. Dev.    Min     Max
---------+--------------------------------------------
   api00 | 1,200   674.2142   134.0295     346     967
```

Dengan *cut-off point* 745, variabel ini dikonversi menjadi variabel biner **hiqual** yang akan digunakan sebagai respons dalam contoh ini. Prediktornya adalah variabel biner **yr_rnd**, yang menyatakan apakah sekolah dibuka sepanjang tiap tahun kalender atau tidak.

. tab hiqual

```
 Hi Quality |
 School, Hi |
     vs Not |   Freq.   Percent     Cum.
------------+-----------------------------
   not high |      809     67.42    67.42
       high |      391     32.58   100.00
------------+-----------------------------
      Total |   1,200    100.00
```

. tab hiqual, nolabel

```
 Hi Quality |
 School, Hi |
     vs Not |   Freq.   Percent     Cum.
------------+-----------------------------
          0 |      809     67.42    67.42
          1 |      391     32.58   100.00
------------+-----------------------------
      Total |   1,200    100.00
```

. tab yr_rnd

```
 Year Round |
     School |   Freq.   Percent     Cum.
------------+-----------------------------
  not_yrrnd |      984     82.00    82.00
      yrrnd |      216     18.00   100.00
------------+-----------------------------
      Total |   1,200    100.00
```

. tab yr_rnd, nolabel

```
 Year Round |
     School |   Freq.   Percent     Cum.
------------+-----------------------------
          0 |      984     82.00    82.00
          1 |      216     18.00   100.00
------------+-----------------------------
      Total |   1,200    100.00
```

. tab yr_rnd hiqual, nolabel

```
           | Hi Quality School, Hi
Year Round |         vs Not
    School |      0       1 |  Total
-----------+----------------+-------
         0 |    613     371 |    984
         1 |    196      20 |    216
-----------+----------------+-------
     Total |    809     391 |  1,200
```

Untuk memperoleh estimasi koefisien regresi logistik:

. logit hiqual yr_rnd

```
Iteration 0: log likelihood = -757.42622
Iteration 1: log likelihood = -719.77388
Iteration 2: log likelihood = -718.62645
Iteration 3: log likelihood = -718.62623
Iteration 4: log likelihood = -718.62623
```

```
Logistic regression                     Number of obs =  1,200
                                        LR chi2(1)    =  77.60
                                        Prob > chi2   = 0.0000
Log likelihood = -718.62623             Pseudo R2     = 0.0512

------------------------------------------------------------------------------
      hiqual |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      yr_rnd |   -1.78022   .2437802    -7.30   0.000    -2.25802   -1.302419
       _cons |  -.5021629    .065778    -7.63   0.000    -.6310853   -.3732405
------------------------------------------------------------------------------
```

Tampak bahwa prediktor biner **yr_rnd** bermakna secara statistik. Estimasi model adalah:

logit hiqual = –0.502 – 1.708 **yr_rnd**

Untuk mengestimasi rasio odds perintahnya adalah:

. logistic hiqual yr_rnd

```
------------------------------------------------------------------------------
      hiqual | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      yr_rnd |   .1686011   .0411016    -7.30   0.000     .1045573    .2718733
       _cons |   .6052202   .0398101    -7.63   0.000     .5320141    .6884997
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Estimasi rasio odds juga dapat diperoleh pada perintah **logit** dengan opsi **or**.

. logit hiqual yr_rnd, or

```
------------------------------------------------------------------------------
hiqual | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------
yr_rnd |   .1686011   .0411016    -7.30   0.000     .1045573    .2718733
 _cons |   .6052202   .0398101    -7.63   0.000     .5320141    .6884997
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Contoh 2.2 (prediktor kontinu):

Pada contoh 2.2 ini akan digunakan file data yang sama **apilog.dta** dengan respons yang sama **hiqual**. Prediktornya adalah variabel kontinu **avg_ed** yang menyatakan rerata tingkat pendidikan (kisaran nilai 1 s.d. 5) orang tua siswa di tiap sekolah.

. sum avg_ed

```
Variable |       Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
  avg_ed |     1,158    2.753964    .7699517         1          5
```

Estimasi koefisien regresinya yaitu:

. logit hiqual avg_ed

```
Iteration 0: log likelihood = -730.68708
Iteration 1: log likelihood = -386.87925
Iteration 2: log likelihood = -355.07208
Iteration 3: log likelihood = -353.91734
Iteration 4: log likelihood = -353.91719
Iteration 5: log likelihood = -353.91719

Logistic regression                     Number of obs =  1,158
                                        LR chi2(1)    = 753.54
                                        Prob > chi2   = 0.0000
Log likelihood = -353.91719             Pseudo R2     = 0.5156

------------------------------------------------------------------------------
      hiqual |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avg_ed |   3.909635   .2383161    16.41   0.000     3.442544    4.376726
       _cons |  -12.30054    .731489   -16.82   0.000    -13.73423   -10.86684
------------------------------------------------------------------------------
```

Estimasi modelnya adalah:

logit hiqual = –12.301 + 3.910 **avg_ed**

Estimasi nilai rasio oddsnya adalah:

. logistic hiqual avg_ed

```
Logistic regression                     Number of obs =  1,158
                                        LR chi2(1)    = 753.54
                                        Prob > chi2   = 0.0000
Log likelihood = -353.91719             Pseudo R2     = 0.5156
```

```
------------------------------------------------------------------------------
hiqual | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------
avg_ed |   49.88075   11.88738    16.41   0.000     31.26641    79.57708
 _cons |   4.55e-06   3.33e-06   -16.82   0.000     1.08e-06    .0000191
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Hasil yang sama dapat diperoleh dengan perintah Stata berikut:

. logit hiqual avg_ed, or

```
------------------------------------------------------------------------------
hiqual | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------
avg_ed |   49.88075   11.88738    16.41   0.000     31.26641    79.57708
 _cons |   4.55e-06   3.33e-06   -16.82   0.000     1.08e-06    .0000191
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

❖ Regresi Logistik Ganda

Model regresi logistik ganda (*multiple logistic regression*) adalah model regresi logistik dengan lebih daripada 1 prediktor, yang dinyatakan sebagai:

$$\text{logit}\,(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \qquad (2.3)$$

$Y = \{0, 1\}$.

Perintah Stata untuk melakukan analisis regresi logistik sederhana dan mengestimasi koefisien regresi logistik adalah:

logit *depvar indepvars* [*if*] [*in*] [, *options*]

depvar : Respons biner

indepvars : Himpunan prediktor

Perintah Stata untuk melakukan analisis regresi logistik sederhana dan mengestimasi rasio odds adalah:

logistic *depvar indepvars* [*if*] [*in*] [, *options*]

depvar : Respons biner

indepvars : Himpunan prediktor

Estimasi rasio odds juga dapat diperoleh dengan perintah:

logit *depvar indepvars* [*if*] [*in*], **or** [*options*]

Contoh 2.3:

. use "D:\Analisis Regresi Logistik\Data\apilog.dta", clear

. list hiqual avg_ed yr_rnd in 1/10

```
     +---------------------------+
     |   hiqual   avg_ed   yr_rnd |
     |---------------------------|
  1. | not high     2.22     yrrn |
  2. | not high     3.02     not_ |
  3. | not high     1.76     yrrn |
  4. | not high     3.43     yrrn |
```

```
  5. | not high     2.68     not_ |
     |---------------------------|
  6. | not high     2.49     not_ |
  7. | not high     2.07     not_ |
  8. | not high     1.59     yrrn |
  9. |     high     2.71     not_ |
 10. | not high     2.75     not_ |
     +---------------------------+
```

. logit hiqual avg_ed yr_rnd

```
Iteration 0: log likelihood = -730.68708
Iteration 1: log likelihood = -384.29611
Iteration 2: log likelihood = -349.78416
Iteration 3: log likelihood = -348.21633
Iteration 4: log likelihood = -348.21614
Iteration 5: log likelihood = -348.21614

Logistic regression                 Number of obs   =       1158
                                    LR chi2(2)      =     764.94
                                    Prob > chi2     =     0.0000
Log likelihood = -348.21614         Pseudo R2       =     0.5234
```

```
------------------------------------------------------------------------------
hiqual |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------
avg_ed |    3.864344   .2410931    16.03   0.000     3.39181    4.336878
yr_rnd |   -1.091301   .3425414    -3.19   0.001    -1.762669   -.4199316
 _cons |   -12.05094   .7397089   -16.29   0.000    -13.50074   -10.60113
------------------------------------------------------------------------------
```

Diperoleh model empirik:

logit (highqual) = –12.0509 + 3.8643**(avg_ed)** – 1.0913**(yr_rnd)**

Estimasi nilai rasio odds dapat diperoleh langsung dengan perintah **logistic**.

. logistic hiqual avg_ed yr_rnd

```
Logistic regression                             Number of obs =    1158
                                                LR chi2(2)    =  764.94
                                                Prob > chi2   =  0.0000
Log likelihood = -348.21614                     Pseudo R2     =  0.5234
```

hiqual	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
avg_ed	47.67199	11.49339	16.03	0.000	29.7197	76.46842
yr_rnd	.3357795	.1150184	-3.19	0.001	.1715862	.6570917
_cons	5.84e-06	4.32e-06	-16.29	0.000	1.37e-06	.0000249

BAB 3
KELUARAN REGRESI LOGISTIK DENGAN STATA

❖ Blok Iterasi

Fungsi likelihood adalah suatu fungsi parameter model yang ada dalam populasi, yaitu:

$$L(\boldsymbol{\beta}) = L\left(\beta_0, \beta_1, \ldots, \beta_p\right) \tag{3.1}$$

Fungsi ini merepresentasikan probabilitas bersama (*joint probability*) atau likelihood untuk mengamati data yang dikumpulkan memiliki koefisien regresi logistik populasi $\{\beta_0, \beta_1, \ldots, \beta_p\}$. Fungsi likelihood sampel $\hat{L}$ memiliki sifat analogi dengan koefisien determinasi R^2 pada regresi linear, yaitu semakin banyak parameter dalam model semakin besar nilai R^2 pada regresi linear ataupun nilai $\hat{L}$ pada regresi logistik. Semakin besar nilai R^2 pada regresi linear ataupun nilai $\hat{L}$ pada regresi logistik, semakin baik kesesuaian model dengan data.

Metode estimasi *maximum likelihood* memaksimumkan nilai statistik log likelihood, yaitu $-2 \ln \hat{L}$, yang akan tercapai jika nilai estimasi parameternya $\{\beta_0, \beta_1, \ldots, \beta_p\}$ menghasilkan kesesuaian model yang terbaik (*the best fit*) dengan data. Blok iterasi pada keluaran Stata memuat gambaran pelaksanaan prosedur *maximum likehood* melalui sejumlah proses iterasi (pengulangan) untuk mencapai nilai maksimum tersebut, yang disebut juga sebagai pencapaian konvergensi statistik log likelihood. Jika

konvergensi tidak tercapai, adakalanya diperlukan penyederhanaan model berupa pengurangan jumlah parameter dalam model.

Contoh 3.1:

Lihat contoh 2.3 dengan file data **apilog.dta**, respons **hiqual** diregresikan terhadap **avg_ed** dan **yr_rnd** dengan model regresi logistik:

$$\text{logit } \textbf{hiqual} = \beta_0 + \beta_1 \textbf{ avg\_ed} + \beta_2 \textbf{ yr\_rnd}$$

Blok iterasinya adalah:

```
Iteration 0: log likelihood = -730.68708
Iteration 1: log likelihood = -384.29611
Iteration 2: log likelihood = -349.78416
Iteration 3: log likelihood = -348.21633
Iteration 4: log likelihood = -348.21614
Iteration 5: log likelihood = -348.21614
```

Tampak konvergensi tercapai setelah melalui 5 kali proses iterasi. Statistik log likehood maksimum, yaitu $-2 \ln \hat{L} = -348.21614$ menghasilkan estimasi parameter $\{\beta_0, \beta_1, \beta_2\}$ yang akan ditampilkan pada tabel koefisien regresi logistik.

❖ Blok Kesesuaian Model

Misalkan dimiliki 2 model regresi logistik M_1 dan M_2 untuk dataset yang sama dengan M_2 tersarang pada M_1. Misalkan pula model M_1

memiliki statistik log likelihood $-2 \ln L_1$ dengan jumlah parameter p_1, sedangkan model M_2 memiliki statistik log likelihood $-2 \ln L_2$ dengan jumlah parameter p_2. Maka uji statistik perbedaan antara kedua model dapat dilakukan dengan uji rasio likelihood, dengan statistik penguji *LR* (*likelihood ratio*):

$$\boxed{LR = -2 \ln L_1 - (-2 \ln L_2)} \qquad (3.2)$$

yang berdistribusi khi-kuadrat dengan derajat bebas $(p_1 - p_2)$.

Selanjutnya dimisalkan M_1 adalah model peneliti dengan likelihood L_K dengan jumlah parameter p_1 dan M_2 adalah model nol (*null model*), yaitu model regresi logistik dengan semua parameter (kecuali konstante β_0) bernilai nol, dengan likelihood L_0. Blok kesesuaian model pada keluaran Stata memuat hasil uji rasio likelihood antara model M_1 dan M_2, yang menguji hipotesis H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$. Statistik pengujinya adalah:

$$\boxed{LR = -2 \ln L_K - (-2 \ln L_0)} \qquad (3.2.a)$$

yang berdistribusi khi-kuadrat dengan derajat bebas $(p_1 - 1)$.

Pada blok kesesuaian model dilaporkan pula nilai `Pseudo R2`. Dalam regresi logistik dikenal berbagai definisi untuk `Pseudo R2`. Pada Stata definisi yang digunakan adalah p^2_{MF} (McFadden,1973), yaitu:

$$\boxed{p^2_{MF} = \frac{\ln L_0 - \ln L_K}{\ln L_0} = 1 - \frac{\ln L_K}{\ln L_0}} \qquad (3.3)$$

yang dapat dianggap sebagai analogi $R^2 = \frac{JKR}{JKT} = 1 - \frac{JKG}{JKT}$ pada regresi linear. Walaupun rentang nilai p^2_{MF} juga berkisar antara 0 dan 1, interpretasinya tak dapat dilakukan sebagaimana dengan koefisien determinasi R^2 pada regresi linear. Baik p^2_{MF} maupun berbagai definisi `Pseudo R2` lainnya tidak menyatakan proporsi variansi respons yang 'dijelaskan' oleh model seperti pada koefisien determinasi R^2 untuk regresi linear.

Contoh 3.2:

Lihat kembali data pada contoh 3.1. Blok kesesuaian model dapat dilihat pada keluaran Stata berikut (blok kesesuaian model adalah yang tertera di bagian kanan).

```
Logistic regression                 Number of obs   =      1158
                                    LR chi2(2)      =    764.94
                                    Prob > chi2     =    0.0000
Log likelihood = -348.21614         Pseudo R2       =    0.5234
```

Tampak jumlah pengamatan yaitu 1,158. Uji rasio likelihood terhadap model nol menghasilkan statistik penguji sebesar 764.94 yang berdistribusi khi-kuadrat dengan derajat bebas 2 dan nilai $p = 0.0000$.

Nilai `Pseudo R2` adalah `0.5234`, tetapi interpretasinya hanya dapat dilakukan dalam perbandingan dengan model lain dari dataset yang sama.

❖ Tabel Koefisien Regresi

Tabel koefisien regresi memuat nilai estimasi *maximum likelihood* koefisien regresi logistik setiap prediktor termasuk konstantenya, beserta *standard error*-nya, nilai statistik penguji uji Wald-nya yang berdistribusi *Z* dan nilai *p*-nya, serta estimasi interval untuk koefisien regresi.

Uji Wald adalah uji statistik untuk tiap koefisien regresi logistik β_j, menguji hipotesis $H_0: \beta_j = 0$.

Contoh 3.3:

Lihat kembali data pada contoh 3.1. Tabel koefisien regresinya adalah:

hiqual	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
avg_ed	3.864344	.2410931	16.03	0.000	3.39181	4.336878
yr_rnd	−1.091301	.3425414	−3.19	0.001	−1.762669	−.4199316
_cons	−12.05094	.7397089	−16.29	0.000	−13.50074	−10.60113

Estimasi model adalah:

logit **hiqual** = −12.05 + 3.86 **avg_ed** − 1.09 **yr_rnd**

Untuk prediktor **avg_ed** misalnya, uji Wald untuk hipotesis $H_0: \beta_1 = 0$ menghasilkan statistik penguji:

$$Z_{uji} = \frac{\hat{\beta}_1}{S\hat{E}\left(\hat{\beta}_1\right)} = \frac{3.864344}{0.2410931} = 16.03$$

yang berdistribusi Z dengan nilai p = 0.000. Interval konfidensi 95%-nya adalah:

$$3.392 \leq \beta_1 \leq 4.337$$

BAB 4
METODE ESTIMASI MAXIMUM LIKELIHOOD

❖ Fungsi Likelihood

Jika n menyatakan jumlah pengamatan, h menyatakan jumlah pengamatan dengan $Y = 1$, dan $(n - h)$ menyatakan jumlah pengamatan dengan $Y = 0$, serta π menyatakan probabilitas untuk memperoleh $Y = 1$, maka fungsi likelihood, yaitu probabilitas untuk memperoleh h pengamatan dengan $Y = 1$ di antara n pengamatan adalah:

$$L(\pi|h,n) = C_n^h \, \pi^h (1-\pi)^{n-h} \tag{4.1}$$

dengan $C_n^h = n!/h!(n-h)!$

Dari fungsi $F(z) = e^z/(1+e^z)$ untuk model regresi logistik, diperoleh probabilitas untuk memperoleh pengamatan dengan $Y = 1$, yaitu:

$$P(Y=1) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}} \tag{4.2}$$

Fungsi likelihood, yaitu probabilitas bersama untuk memperoleh h pengamatan dengan $Y = 1$ dan $(n - h)$ pengamatan dengan $Y = 0$ dari n pengamatan adalah:

$$L(\beta_k|f,n,h) = P(Y=1)^h \times \{1-P(Y=1)\}^{n-h}$$

$$\left(\frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}\right)^h \times \left(1-\frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}\right)^{n-h} \quad (4.3)$$

Dengan algoritma iteratif, himpunan nilai β_k diperoleh dengan memaksimalkan fungsi likelihood, yang dalam praktik lebih mudah dikerjakan terhadap $-2\ln L$, sehingga diperoleh estimator *maximum likelihood* untuk himpunan koefisien regresi logistik β_k.

❖ Uji Rasio Likelihood

Misalkan dimiliki 2 model regresi logistik untuk dataset yang sama dengan model regresi kedua tersarang dalam model pertama. Maka model pertama dinamakan model lengkap (*full model*), sedangkan model kedua dinamakan model tereduksi (*reduced model*).

Uji statistik untuk memperbandingkan kedua model tersebut dapat dilakukan dengan uji rasio likelihood. Jika model pertama memiliki fungsi likehood $-2\ln L_1$ dengan (p + k) parameter dan model kedua memiliki fungsi likehood $-2\ln L_2$ dengan p parameter, maka statistik pengujinya adalah:

$$LR = -2\ln L_1 - (-2\ln L_2) \quad (4.4)$$

yang berdistribusi khi-kuadrat dengan derajat bebas (p + k) – p = k.

Seandainya hasil uji statistik tidak menunjukkan perbedaan antara model lengkap dengan model tereduksi, maka berdasarkan prinsip parsimoni yang dipilih adalah model tereduksi.

Perintah Stata untuk uji rasio likelihood adalah:

. **logit** [*full_model*]

. **estimates store A**

. **logit** [*reduced_model*]

. **estimates store B**

. **lrtest A B**

Contoh 4.1:

Lihat file data **apilog.dta** pada contoh 2.3. Pada model lengkap, respons **hiqual** akan diregresikan terhadap **avg_ed** dan **yr_rnd** , sedangkan pada model tereduksi, respons **hiqual** hanya akan diregresikan terhadap **avg_ed**.

. **use "D:\Analisis Regresi Logistik\Data\apilog.dta"**

. **logit hiqual avg_ed yr_rnd**

```
Iteration 0: log likelihood = -730.68708
Iteration 1: log likelihood = -384.29611
Iteration 2: log likelihood = -349.78416
Iteration 3: log likelihood = -348.21633
Iteration 4: log likelihood = -348.21614
Iteration 5: log likelihood = -348.21614

Logistic regression                 Number of obs =  1,158
                                    LR chi2(2)    = 764.94
                                    Prob > chi2   = 0.0000
Log likelihood = -348.21614         Pseudo R2     = 0.5234
```

```
------------------------------------------------------------------------------
hiqual |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------------
avg_ed |   3.864344   .2410931    16.03   0.000      3.39181    4.336878
yr_rnd |  -1.091301   .3425414    -3.19   0.001    -1.762669   -.4199316
 _cons |  -12.05094   .7397089   -16.29   0.000    -13.50074   -10.60113
------------------------------------------------------------------------------
```

. estimate store A

. logit hiqual avg_ed

```
Iteration 0:  log likelihood = -730.68708
Iteration 1:  log likelihood = -386.87925
Iteration 2:  log likelihood = -355.07208
Iteration 3:  log likelihood = -353.91734
Iteration 4:  log likelihood = -353.91719
Iteration 5:  log likelihood = -353.91719

Logistic regression                             Number of obs =  1,158
                                                LR chi2(1)    = 753.54
                                                Prob > chi2   = 0.0000
Log likelihood = -353.91719                     Pseudo R2     = 0.5156

------------------------------------------------------------------------------
hiqual |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------------
avg_ed |   3.909635   .2383161    16.41   0.000     3.442544    4.376726
 _cons |  -12.30054    .731489   -16.82   0.000    -13.73423   -10.86684
------------------------------------------------------------------------------
```

```
. estimates store B

. lrtest A B

Likelihood-ratio test                 LR chi2(1)  =  11.40
(Assumption: B nested in A)           Prob > chi2 = 0.0007
```

Dengan p = 0.0007 tampak model **B** secara bermakna berbeda dengan model **A**, sehingga prediktor **yr_rnd** tetap dipertahankan dalam model.

❖ Uji Wald

Dengan uji rasio likelihood dapat diuji kemaknaan 1 ataupun beberapa prediktor sekaligus. Jika uji melibatkan dua atau lebih prediktor dan diperoleh hasil bermakna, tidak diketahui prediktor mana saja yang menyebabkan kemaknaan tersebut. Uji Wald menguji kemaknaan tiap prediktor satu demi satu, masing-masing terhadap hipotesis H_0: $\beta_j = 0$. Sebagian ahli Statistika menganggapnya sebagai pengujian ganda (*multiple testings*) yang memerlukan koreksi untuk kesalahan tipe I-nya, misalnya dengan metode Bonferroni.

Statistik penguji untuk uji Wald adalah:

$$Z_{Wald} = \frac{\beta_j}{SE(\beta_j)} \qquad (4.5)$$

yang berdistribusi normal standar.

Jika hendak digunakan koreksi Bonferroni, maka seandainya terdapat (p + 1) parameter dalam model ($\beta_0, \beta_1, \ldots, \beta_p$) dan akan digunakan tingkat signifikansi α, maka batas kemaknaan yang seharusnya digunakan adalah:

$$\alpha_{\text{nominal}} = \frac{\alpha}{p+1} \tag{4.6}$$

❖ Estimasi Interval

Estimator koefisien regresi diasumsikan berdistribusi normal, sehingga interval konfidensi (1 – α)-nya adalah:

$$\hat{\beta}_j \pm Z_{\alpha/2} \times S\hat{E}\left(\hat{\beta}_j\right) \tag{4.7}$$

Misalnya, interval konfidensi 95%-nya adalah:

$$\hat{\beta}_j \pm 1.96 \times S\hat{E}\left(\hat{\beta}_j\right)$$

Untuk rasio odds, keadaannya agak berbeda. Untuk perhitungan secara manual, rasio odds diasumsikan berdistribusi log-normal, yaitu logaritma rasio odds diasumsikan berdistribusi normal dengan variansi sampel (Woolf, 1955):

$$\hat{Var}\left(\ln \hat{OR}\right) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \tag{4.8}$$

dengan notasi *a*, *b*, *c*, dan *d* seperti pada tabel 1.2.

Interval konfidensi $(1 - \alpha)$ untuk $\ln OR$ adalah:

$$\ln \hat{OR} \pm Z_{\alpha/2} \times SE\left(\ln \hat{OR}\right)$$

atau:
$$\boxed{\ln \hat{OR} \pm Z_{\alpha/2} \times \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}} \quad (4.9)$$

Sehingga diperoleh interval konfidensi $(1 - \alpha)$ untuk $\ln OR$:

$$\boxed{\left[\left(\ln \hat{OR}\right)_B ; \left(\ln \hat{OR}\right)_A\right]} \quad (4.10)$$

Interval konfidensi $(1 - \alpha)$ untuk OR adalah:

$$\boxed{\left[\exp\left(\ln \hat{OR}\right)_B ; \exp\left(\ln \hat{OR}\right)_A\right]} \quad (4.10.a)$$

Contoh 4.2:

Lihat kembali file data **apilog.dta** pada pada contoh 2.1.

. use "D:\Analisis Regresi Logistik\Data\ apilog.dta", clear

. tab2 hiqual yr_rnd

```
-> tabulation of hiqual by yr_rnd

Hi Quality |
School, Hi |   Year Round School
    vs Not | not_yrrnd      yrrnd |     Total
-----------+----------------------+----------
  not high |       613        196 |       809
      high |       371         20 |       391
-----------+----------------------+----------
     Total |       984        216 |     1,200
```

Dengan cara manual, ln $\hat{OR}$ = ln (*ab*/*cd*) adalah:

```
. display ln((613*20)/(196*371))
  -1.7802195
```

$\hat{SE}$ ln $\hat{OR} = \sqrt{1/a + 1/b + 1/c + 1/d}$ adalah:

```
. display ((1/613 + 1/196 + 1/371 + 1/20))^.5
  .24378019
```

Batas bawah interval konfidensi 95% $\ln OR = \ln \hat{OR} - 1.96 \times \hat{SE}$ ln $\hat{OR}$ adalah:

```
. display (-1.7802195-1.96*.24378019)
  -2.2580287
```

Batas atas interval konfidensi 95% $\ln OR = \ln \hat{OR} + 1.96 \times \hat{SE}$ ln $\hat{OR}$ adalah:

```
. display (-1.7802195+1.96*.24378019)
-1.3024103
```

Batas bawah interval konfidensi 95% OR = exp −2.2580287 adalah:

```
. display exp(-2.2580287)
.10455639
```

Batas atas interval konfidensi 95% OR = exp −1.3024103 adalah:

```
. display exp(-1.3024103)
.2718757
```

Diperoleh interval konfidensi 95% untuk *OR*:

[0.1046 ; 0.2719]

Dengan perintah Stata **cc** diperoleh hasil yang agak berbeda:

```
. cc hiqual yr_rnd
```

```
                                                    Proportion
         |   Exposed   Unexposed  |    Total     Exposed
---------+------------------------+--------------------
   Cases |        20         371  |      391      0.0512
Controls |       196         613  |      809      0.2423
---------+------------------------+--------------------
   Total |       216         984  |     1200      0.1800
         |                        |
                | Point estimate  | [95% Conf. Interval]
                |-----------------+--------------------
     Odds ratio |    .1686011     |  .0991047   .2736994 (exact)
Prev. frac. ex. |    .8313989     |  .7263006   .9008953 (exact)
Prev. frac. pop |    .2014267     |
                +-----------------------------------------
                       chi2(1) = 65.24  Pr>chi2 = 0.0000
```

Interval konfidensi 95% untuk *OR* dengan perintah Stata **cc** adalah:

[0.0991 ; 0.2737]

Perbedaan ini disebabkan karena perintah Stata **cc** menggunakan metode *exact* untuk menghitung estimator $S\hat{E}\,O\hat{R}$. Dengan regresi logistik diperoleh:

. logistic hiqual yr_rnd

```
Logistic regression                     Number of obs =  1,200
                                        LR chi2(1)    =  77.60
                                        Prob > chi2   = 0.0000
Log likelihood = -718.62623             Pseudo R2     = 0.0512

------------------------------------------------------------------------------
      hiqual | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      yr_rnd |   .1686011   .0411016    -7.30   0.000     .1045573    .2718733
       _cons |   .6052202   .0398101    -7.63   0.000     .5320141    .6884997
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Tampak interval konfidensi 95% untuk *OR* adalah:

[0.1046 ; 0.2719]

dengan hasil yang lebih mendekati hasil perhitungan manual.

BAB 5
STRATEGI PEMODELAN

❖ Spesifikasi Variabel

Kriteria pemilihan dan seleksi variabel independen untuk model regresi logistik berbeda antar disiplin ilmu. Prinsip dasar utama yaitu mencari model yang paling parsimoni (*parcimonious*), yang masih mampu "menjelaskan" data. Model demikian secara numerik akan lebih stabil dalam arti kata memiliki estimasi interval koefisien regresi dan *standard error* yang relatif sempit, serta lebih mudah digeneralisasikan.

Dalam epidemiologi misalnya, dianjurkan untuk menginklusikan semua variabel yang relevan secara klinik ataupun intuitif dalam model, tanpa tergantung pada kemaknaan statistiknya. Praktik ini diharapkan akan mampu mengendalikan semua kemungkinan konfaunding pada *dataset*. Sebaliknya, inklusi terlalu banyak variabel dalam model, relatif terhadap jumlah anggota sampelnya dapat menyebabkan "*overfitting*", yaitu estimasi yang tak stabil dengan estimasi koefisien regresi dan *standard error* yang terlalu besar.

Secara umum, urutan langkah pada seleksi variabel yaitu:

1). Pilih calon variabel independen dengan teliti, lalu lakukan analisis univariabel tiap calon variabel independen ini dengan variabel dependen. Tiap calon variabel independen yang pada analisis univariabel ini menghasilkan nilai $p < 0.25$ dan secara substantif memiliki kemaknaan substantif menurut ranah bidang ilmu yang diteliti dapat dipertahankan menuju langkah berikutnya.

   Perhatikan bahwa calon prediktor yang pada analisis univariabel hanya menunjukkan asosiasi lemah dengan respons, mungkin menjadi prediktor penting pada analisis multivariabel.

2) Lakukan analisis multivariabel dengan variabel dependen, menggunakan himpunan variabel independen yang lolos dari seleksi tahap 1). Beberapa alternatif di sini:

- Inklusikan seluruh variabel independen yang dianggap "relevan secara ilmiah" dalam analisis multivariabel tanpa tergantung pada hasil analisis univariabel pada langkah 1). Seandainya jumlah variabel independen tersebut terlalu banyak (dalam perbandingan dengan ukuran sampel), dapat dilakukan seleksi ulang dengan mendefinisikan kembali kriteria "relevan secara ilmiah".

- Lakukan seleksi variabel dengan metode "*stepwise*". Inklusi dan ekslusi variabel dengan metode ini sepenuhnya didasarkan pada kriteria statistik, yaitu nilai *p*-nya. Dua versi metode *stepwise* ialah "seleksi ke depan" (*forward selection*) dan "eliminasi ke belakang" (*backward elimination*).

  Pada *forward selection*, analisis dimulai dengan 1 variabel independen yang nilai *p*-nya terkecil pada analisis univariabel, lalu setiap kali ditambah 1 variabel independen yang nilai *p*- nya terkecil berikutnya pada analisis univariabel. Lihat hasil uji Wald pada tiap penambahan variabel baru, tiap variabel yang tak bermakna secara statistik dikeluarkan kembali dari model. Penambahan ke depan berlanjut sampai semua calon variabel independen dicoba untuk dimasukkan dalam model.

  Pada *backward elimination*, analisis multivariabel dimulai dengan memasukkan semua variabel independen yang menunjukkan kemaknaan statistik pada analisis univariabel, lalu dicoba mengeliminasi 1 variabel independen yang pada uji Wald memiliki nilai *p* terbesar dan tak bermakna secara statistik. Lakukan kembali analisis multivariabel setelah mengeliminasi 1 variabel ini. Proses percobaan eliminasi ini berlanjut sampai yang tersisa seluruhnya adalah variabel independen yang bermakna secara statistik.

- Metode lain yang relatif panjang dan tak akan dibahas di sini yaitu "*best subsets selection*" yang memperbandingkan berbagai model

dengan jumlah maupun anggota himpunan variabel independen yang berbeda.

Prinsip dasar yang terpenting untuk diingat pada tahap ini, yaitu bahwa yang terutama seharusnya bertanggung jawab menelaah-ulang dan mengevaluasi model adalah peneliti, bukan komputer.

3) Verifikasikan kepentingan tiap variabel independen yang diinklusikan dalam model sebagai hasil langkah 2).

   a) Periksa kembali statistik uji Wald untuk tiap variabel.

   b) Bandingkan hasil estimasi tiap koefisien regresi dengan hasil estimasi pada analisis univariabel pada langkah 1). Perbedaan yang terlalu besar mungkin mengindikasikan bahwa 1 atau lebih variabel yang telah dieksklusikan sebenar masih perlu dipertahankan demi "penyesuaian" efek bersama seluruh variabel independen.

4) Periksa pemenuhan asumsi linearitas untuk logit variabel dependen.

5) Periksa semua kemungkinan interaksi antar variabel utama yang dihasilkan dari langkah-langkah di atas. Keputusan untuk menginklusikan suatu interaksi harus didasarkan atas pertimbangan statistik maupun praktik. Inklusi interaksi yang tak bermakna secara statistik hanya akan memperbesar rentang estimasi *standard error* tanpa mengubah nilai estimasi koefisiennya sendiri. Interaksi yang diinklusikan juga harus memiliki arti sesuai ranah bidang ilmu yang diteliti.

❖ Penilaian Interaksi

Interaksi adalah efek bersama 2 faktor atau lebih yang mempengaruhi terjadinya suatu peristiwa. “Faktor” yang dimaksud dalam regresi logistik adalah variabel independennya, sedangkan “peristiwa” adalah terjadinya respons atau variabel dependen.

Misalkan pada akhir langkah 4) seleksi variabel di atas, diperoleh 3 variabel independen, X_1, X_2, dan X_3. Selanjutnya dalam penilaian interaksi, ketiga variabel ini dinamakan efek utama (*main effects*). Dari 3 efek utama ini dapat dibentuk 3 suku interaksi 2-faktor, $X_1 * X_2$, $X_1 * X_3$, dan $X_2 * X_3$, serta 1 suku interaksi 3-faktor $X_1 * X_2 * X_3$.

Prinsip hirarki untuk interaksi menyatakan bahwa jika ada suatu suku interaksi dalam model, maka model harus menginklusikan semua efek utama pembentuk suku interaksi tersebut, beserta semua suku interaksi lebih rendah penyusun suku interaksi tersebut. Misalkan sebuah model memuat suku interaksi 3-faktor $X_1 * X_2 * X_3$, maka model harus pula menginklusikan efek utama X_1, X_2, dan X_3, serta suku interaksi 2-faktor $X_1 * X_2$, $X_1 * X_3$, dan $X_2 * X_3$, walaupun di antara efek utama dan suku interaksi 2-faktor ini mungkin ada yang tak bermakna secara statistik.

Jika dalam suatu model terdapat interaksi, maka keberadaan konfaunding menjadi tak dapat dinilai dan penilaiannya tidak akan dikerjakan. Dengan demikian, dalam suatu model penilaian interaksi akan dikerjakan terlebih dahulu sebelum penilaian konfaunding dan penilaian konfaunding hanya akan dilakukan jika tidak ditemukan interaksi.

Penilaian interaksi dimulai dari model lengkap yang memuat semua suku interaksi dan prosedur diawali dengan memeriksa suku interaksi tertinggi. Dalam praktik, pemeriksaan suku interaksi 3-faktor atau lebih

jarang dikerjakan, karena kesukaran untuk menjelaskan maknanya secara substantif dari segi ranah bidang ilmu yang diteliti.

Pada model dengan suku interaksi 3-faktor $X_1 * X_2 * X_3$ tadi, suku inilah yang harus diperiksa terlebih dahulu. Jika suku interaksi ini ditemukan bermakna, maka penilaian interaksi berhenti di sini dan suku interaksi 3-faktor tersebut dipertahankan dalam model. Sebaliknya jika suku ini ditemukan tak bermakna, suku interaksi 3-faktor ini dieksklusikan dari model dan penilaian interaksi dilanjutkan dengan pemeriksaan suku interaksi 2-faktor pada pemodelan baru, dimulai dari yang nilai *p*-nya terbesar. Seperti sebelumnya, tergantung hasil yang diperoleh, penilaian interaksi dapat berhenti di sini ataupun dilanjutkan dengan memeriksa suku interaksi 2-faktor berikutnya pada pemodelan baru, yang nilai *p*-nya kedua terbesar, dan seterusnya sampai semua suku dalam model bermakna secara statistik.

❖ Penilaian Konfaunding dan Pencapaian Presisi

Konfaunding (*confounding*) adalah peristiwa distorsi ukuran efek prediktor terhadap responsnya karena keberadaan faktor lain (konfaunder) yang mempengaruhi terjadinya respons. Ukuran efek yang dimaksud dalam regresi logistik adalah rasio odds.

Kedua prosedur seleksi variabel dan penilaian interaksi yang telah diuraikan di atas sedikit banyak melibatkan pengujian statistik dalam pelaksanaannya. Berbeda dengan kedua topik itu, penilaian konfaunding adalah isu validitas yang tidak menyangkut dan tidak akan menggunakan uji statistik dalam pelaksanaannya.

Tujuan terpenting dalam pengendalian konfaunding adalah mendapatkan estimasi yang valid (benar) dengan menghilangkan distorsi

efek oleh konfaunder, berbeda lagi dengan tujuan pencapaian presisi yang dikerjakan setelah pengendalian konfaunding, yaitu mencapai estimasi interval parameter yang sempit (presisi yang tinggi).

Setelah melalui langkah 4) seleksi variabel dan pada langkah 5) tidak ditemukan adanya interaksi, maka dimiliki himpunan lengkap variabel independen yang akan diinklusikan dalam model. Dari himpunan variabel independen ini dipilih satu prediktor terpenting yang akan diestimasi pengaruhnya terhadap respons. Selanjutnya berdasarkan pertimbangan substantif sesuai ranah bidang ilmu yang diteliti, dari himpunan variabel independen tersisa dipilih beberapa yang berpotensi menjadi kandidat konfaunder. Model dengan himpunan lengkap kandidat konfaunder ini dinamakan model "*gold standard*" (model baku emas).

Estimasi hubungan prediktor terpenting yang dipilih dengan respons dalam analisis univariabel pada langkah 1) seleksi variabel dinamakan "estimasi kasar" dan karena hubungan yang lazim digunakan dalam regresi logistik adalah rasio odds, estimasinya disebut rasio odds kasar (*crude odds ratio*). Estimasi yang diperoleh dari analisis multivariabel dengan menginklusikan kandidat konfaunder dan karena itu mengendalikan efek konfaunder tersebut, dinamakan "estimasi suaian", dan untuk rasio odds dinamakan ratio odds suaian (*adjusted odds ratio*) . Estimasi suaian yang diperoleh dari model dengan himpunan lengkap kandidat konfaunder dinamakan estimasi *gold standard*.

Selanjut untuk pencapaian presisi, dicoba berbagai model yang mengeksklusikan satu atau lebih kandidat konfaunder. Dari sekian banyak model yang diujicobakan ini, yang dipertahankan untuk seleksi lebih lanjut adalah model yang menghasilkan estimasi suaian yang sama dengan estimasi *gold standard.* Dalam tahap akhir seleksi ini, dari sejumlah model yang dipertahankan ini (termasuk model *gold standard*), dipilih satu model akhir yang memiliki estimasi dengan presisi tertinggi (estimasi intervalnya paling sempit).

BAB 6
PENILAIAN KESESUAIAN MODEL

❖ Deviansi

Deviansi (*deviance*) merupakan ukuran kebaikan-suai (*goodness-of-fit*; GOF) yang lazim digunakan untuk model regresi logistik. Deviansi adalah rasio antara fungsi likelihood model peneliti dengan fungsi likelihood model jenuh:

$$\text{Dev}(\hat{\boldsymbol{\beta}}) = -2 \ln \left(\hat{L}_C / \hat{L}_{\max} \right) \qquad (6.1)$$

$\hat{L}_C$: Likelihood model peneliti, yaitu model yang menggunakan estimasi koefisien regresi $\hat{\boldsymbol{\beta}}$ dan akan dihitung deviansinya.

$\hat{L}_{\max}$: Likelihood model jenuh

Model jenuh (*saturated model*) adalah model yang jumlah parameternya sama dengan ukuran sampel. Model jenuh akan menghasilkan prediksi nilai-nilai respons yang sempurna:

$$\left(\hat{Y}_i - Y_i \right) = 0 \text{ untuk } i = 1, 2, \ldots, n \qquad (6.1.a)$$

Deviansi memiliki rentang nilai yang berkisar dari nol sampai dengan positif tak berhingga. Jika model peneliti memiliki $(p + 1)$ parameter, maka deviansinya dianggap berdistribusi khi-kuadrat dengan derajat bebas $\{(n - (p + 1)\} = (n - p - 1)$. Uji hipotesis kebaikan-suai dengan statistik deviansi menguji hipotesis H_0 : Model sesuai data *vs* H_1 : Model tak-sesuai data.

Dalam kenyataannya, uji hipotesis kebaikan-suai dengan statistik deviansi ini dapat menggunakan individu anggota sampel ataupun kelompok pola kovariat sebagai unit analisis. Kelompok pola kovariat (*covariate pattern group*) adalah kelompok yang beranggotakan subjek yang memiliki himpunan nilai prediktor yang sama. Penggunaan unit analisis yang berbeda ini akan menghasilkan nilai statistik penguji yang berbeda (rumus perhitungannya memang berbeda) dengan derajat bebas yang berbeda pula.

Pada uji hipotesis yang menggunakan kelompok pola kovariat sebagai unit analisis, maka jika jumlah kelompok pola kovariat sama dengan k, statistik penguji dianggap berdistribusi khi-kuadrat dengan derajat bebas $\{k - (p + 1)\} = (k - p - 1)$.

Pada Stata, deviansi merupakan salah satu statistik post-estimasi, yang diperoleh langsung setelah perintah **logistic**:

```
. logistic [assessed_model]
. ldev
```

Perintah Stata **ldev** menggunakan kelompok pola kovariat sebagai unit analisis. Syarat penggunaan perintah perhitungan statistik deviansi ini yaitu jumlah parameter plus satu, yaitu $\{(p + 1) + 1\} = (p + 2)$ lebih kecil daripada jumlah pola kovariat. Perhatikan bahwa jika ada prediktor kontinu, jumlah kelompok pola kovariat akan relatif besar, tetapi jika seluruh prediktor adalah kategorik, jumlah kelompok pola kovariat ini umumnya sangat terbatas. Jika jumlah parameter plus satu sama dengan jumlah kelompok pola kovariat, maka statistik deviansi sama dengan nol.

Statistik deviansi juga dapat digunakan pada uji rasio likelihood yang memperbandingkan dua model hirarkis, yaitu model pertama tersarang dalam model kedua. Jika model pertama memiliki statistik deviansi $\text{Dev}_1\left(\hat{\boldsymbol{\beta}}\right)$ dengan jumlah parameter $\left(p_1 + 1\right)$ dan model kedua memiliki statistik

deviansi $\text{Dev}_2\left(\hat{\boldsymbol{\beta}}\right)$ dengan jumlah parameter $\left(p_2+1\right)$, maka statistik penguji rasio likelihood-nya adalah:

$$\boxed{\text{Dev}_1\left(\hat{\boldsymbol{\beta}}\right) - \text{Dev}_2\left(\hat{\boldsymbol{\beta}}\right)} \qquad (6.2)$$

yang berdistribusi khi-kuadrat dengan derajat bebas $\left(p_2+p_1\right)$.

Contoh 5.1:

. use "D:\Analisis Regresi Logistik\Data\apilog.dta", clear

. logistic hiqual avg_ed yr_rnd

```
Logistic regression                     Number of obs =  1,158
                                        LR chi2(2)    = 764.94
                                        Prob > chi2   = 0.0000
Log likelihood = -348.21614             Pseudo R2     = 0.5234

------------------------------------------------------------------------------
      hiqual | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avg_ed |   47.67199   11.49339    16.03   0.000     29.7197    76.46842
      yr_rnd |   .3357795   .1150184    -3.19   0.001    .1715862    .6570917
       _cons |   5.84e-06   4.32e-06   -16.29   0.000    1.37e-06    .0000249
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

. ldev

```
Logistic estimates for , goodness-of-fit test

          no. of observations =    1200
    no. of covariate patterns =     427
           Deviance chi2(424) = 2209.27
                       P>chi2 =  0.0000
```

Jumlah individu anggota sampel adalah n = 1200, tetapi unit analisis untuk perhitungan statistik deviansi adalah kelompok pola kovariat, yang jumlahnya yaitu k = 427. Statistik penguji deviansi berdistribusi khi-kuadrat dengan derajat bebas $(k - p - 1) = 424$.

Uji kebaikan-suai juga dapat dilakukan dengan statistik khi-kuadrat Pearson:

. estat gof

```
Logistic model for hiqual, goodness-of-fit test

       number of observations =    1158
 number of covariate patterns =     426
            Pearson chi2(423) = 2762.25
                  Prob > chi2 =  0.0000
```

❖ Uji Hosmer-Lemeshow

Sebagian ahli menganggap uji kebaikan-suai 1 model dengan statistik deviansi kurang valid karena pada uji untuk 1 model statistik deviansi kurang mendekati distribusi khi-kuadrat. Perbaikannya adalah dengan uji Hosmer-

Lemeshow, yang juga merupakan uji khi-kuadrat tetapi bukan terhadap kelompok-kelompok pola kovariat, melainkan kelompok kuantil. Kuantil yang lazim digunakan adalah desil, dengan membagi sampel menjadi 10 desil.

Perintah Stata untuk uji Hosmer-Lemeshow diberikan langsung setelah *fitting* model, yaitu:

estat gof, group(*k*) table

k : Jumlah kelompok, biasanya digunakan $k = 10$ untuk kelompok desil

Contoh 6.2:

. use "D:\Analisis Regresi Logistik\Data\apilog.dta", clear

. logit hiqual avg_ed yr_rnd

Uji kebaikan-suai Hosmer-Lemeshow:

. estat gof, group(10) table

```
Logistic model for hiqual, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0037 |     1 |   0.2 |   116 | 116.8 |   117 |
  |     2 | 0.0112 |     1 |   0.8 |   114 | 114.2 |   115 |
```

```
|     3 | 0.0269 |     2 |   2.2 |   114 | 113.8 |   116 |
|     4 | 0.0782 |     4 |   5.9 |   115 | 113.1 |   119 |
|     5 | 0.1502 |    13 |  12.8 |   102 | 102.2 |   115 |
|-------+--------+-------+-------+-------+-------+-------|
|     6 | 0.2926 |    22 |  24.5 |    91 |  88.5 |   113 |
|     7 | 0.4918 |    45 |  45.4 |    71 |  70.6 |   116 |
|     8 | 0.7552 |    78 |  76.4 |    42 |  43.6 |   120 |
|     9 | 0.9377 |   100 |  96.6 |    12 |  15.4 |   112 |
|    10 | 0.9993 |   111 | 112.3 |     4 |   2.7 |   115 |
+-------------------------------------------------------------+

  number of observations = 1158
        number of groups =   10
Hosmer-Lemeshow chi2(8) =    5.46
             Prob > chi2 =    0.7076
```

❖ Kriteria Informasi

Kriteria informasi (*informational criteria*) adalah statistik untuk model yang estimasi parameternya diperoleh dengan memaksimumkan fungsi likelihood-nya, digunakan untuk memperbandingkan kebaikan-suai 2 model hirarkis (salah satu model tersarang dalam model lainnya) ataupun 2 model non-hirarkis. Dua kriteria informasi yang dibahas di sini yaitu AIC (*Akaike's Informational Criteria*) dan BIC (*Bayesian Informational Criteria*).

Dikenal berbagai rumus perhitungan AIC dan BIC, yang menghasilkan nilai yang saling berbeda. Rumus AIC dan BIC yang digunakan pada Stata masing-masing adalah:

$$\boxed{\text{AIC} = -2 \ln (\textit{likelihood}) + 2\,(p+1)} \qquad (6.2)$$

$$\boxed{\text{BIC} = -2 \ln (\textit{likelihood}) + \ln (n) \times (p+1)} \qquad (6.3)$$

$(p + 1)$: Jumlah parameter yang diestimasi (termasuk konstante)

n : Jumlah pengamatan (ukuran sampel)

Untuk memperbandingkan 2 model dengan uji kriteria informasi harus digunakan rumus yang sama. Perintah Stata untuk memperoleh nilai AIC dan BIC yaitu:

estat ic

yang diberikan setelah perintah *fitting* model dengan **logit** atau **logistik**.

Pada perbandingan 2 model dengan statistik AIC dan BIC tidak dikenal distribusi statistik penguji, sehingga nilai *p*-nya tak dapat dihitung. Model yang dipilih adalah model dengan nilai AIC dan BIC yang lebih kecil. Kriteria penilaian selisih relatif nilai AIC antara 2 model A dan B dengan asumsi $\text{AIC}_A < \text{AIC}_B$ menurut Hilbe (2009) adalah:

Selisih AIC antara model A dan B *)	Interpretasi
$0 <$ Selisih ≤ 2.5	Kedua model tak berbeda
$2.5 <$ Selisih ≤ 6.0	Pilih model A jika $n > 256$
$6.0 <$ Selisih ≤ 9.9	Pilih model A jika $n > 64$
Selisih ≥ 10	Pilih model A

*) Persentasi selisih, yaitu $(1 - \text{AIC}_A / \text{AIC}_B) \times 100\%$

Kriteria penilaian selisih absolut nilai BIC antara 2 model A dan B dengan asumsi $\text{BIC}_A < \text{BIC}_B$ menurut Raftery (1986) adalah:

Selisih BIC antara model A dan B *)	Derajat perbedaan
0 – 2	Lemah
2 – 8	Positif
6 – 10	Kuat
> 10	Sangat kuat

Contoh 6.1:

. use "D:\Analisis Regresi Logistik\Data\apilog.dta", clear

. logit hiqual yr_rnd avg_ed

```
Iteration 0: log likelihood = -730.68708
Iteration 1: log likelihood = -384.29611
Iteration 2: log likelihood = -349.78416
Iteration 3: log likelihood = -348.21633
Iteration 4: log likelihood = -348.21614
Iteration 5: log likelihood = -348.21614

Logistic regression                     Number of obs =  1,158
                                        LR chi2(2)    = 764.94
                                        Prob > chi2   = 0.0000
Log likelihood = -348.21614             Pseudo R2     = 0.5234
```

```
------------------------------------------------------------------------------
hiqual |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------+----------------------------------------------------------------------
yr_rnd | -1.091301   .3425414    -3.19   0.001    -1.762669   -.4199316
avg_ed |  3.864344   .2410931    16.03   0.000     3.39181    4.336878
 _cons | -12.05094   .7397089    16.29   0.000    -13.50074   -10.60113
------------------------------------------------------------------------------
```

. estat ic

```
Akaike's information criterion and Bayesian information
criterion

------------------------------------------------------------------------------
Model |    Obs  ll(null)  ll(model)   df       AIC       BIC
------+-----------------------------------------------------------------------
      |  1,158 -730.6871  -348.2161    3  702.4323  717.5956
------------------------------------------------------------------------------
      Note: N=Obs used in calculating BIC; see [R] BIC note.
```

. logit hiqual avg_ed

```
Iteration 0: log likelihood = -730.68708
Iteration 1: log likelihood = -386.87925
Iteration 2: log likelihood = -355.07208
Iteration 3: log likelihood = -353.91734
Iteration 4: log likelihood = -353.91719
Iteration 5: log likelihood = -353.91719
```

```
Logistic regression                             Number of obs =  1,158
                                                LR chi2(1)    = 753.54
                                                Prob > chi2   = 0.0000
Log likelihood = -353.91719                     Pseudo R2     = 0.5156

------------------------------------------------------------------------------
      hiqual |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avg_ed |    3.909635   .2383161    16.41   0.000     3.442544    4.376726
       _cons |   -12.30054    .731489   -16.82   0.000    -13.73423   -10.86684
------------------------------------------------------------------------------
```

. estat ic

```
Akaike's information criterion and Bayesian information criterion

------------------------------------------------------------------------------
       Model |    Obs  ll(null)  ll(model)   df        AIC       BIC
-------------+----------------------------------------------------------------
             |  1,158 -730.6871  -353.9172    2   711.8344  721.9433
------------------------------------------------------------------------------
               Note: N=Obs used in calculating BIC; see [R] BIC note.
```

Selisih relatif nilai AIC kedua model adalah:

. display (1 - (702.4323/711.8344))*100

```
1.3208269
```

Selisih relatif nilai AIC kedua model adalah 1.32% dengan interpretasi menurut Hilbe bahwa kedua model tak berbeda.

```
. display (721.9433 - 717.5956)
4.3477
```

Selisih absolut nilai BIC kedua model adalah 4.35 dengan interpretasi adanya perbedaan positif antara kedua model menurut Ratcliffe.

Sebagai ringkasan akhir, penggunaan statistik deviansi, uji Hosmer-Lemeshow, dan kriteria informasi untuk penilaian kebaikan-suai model regresi logistik diperlihatkan pada tabel 6.1 berikut.

Tabel 6.1 Skema penilaian kebaikan-suai model regresi logistik

Uji GOF	Deviansi	Hosmer-Lemeshow	Kriteria informasi
1 model	√	√	
2 model hirarkis	√		√
2 model non-hirarkis			√

BAB 7
PENILAIAN TAMPILAN DISKRIMINATORIK

Tampilan diskriminatorik (*discriminatory performance*) adalah tampilan kemampuan suatu model statistik untuk mendiferensiasi subjek menurut responsnya, yang diprediksi memiliki respons positif dan yang diprediksi memiliki respons negatif, beserta ketepatan prediksinya. Ukuran kualitas diskriminatorik yang dibahas di sini adalah sensitivitas dan spesifisitas, serta kurve ROC.

❖ Sensitivitas dan Spesifisitas

Dari model regresi logistik dapat diprediksi ada tidaknya respons dalam nilai $\hat{Y} = 1$ dan $\hat{Y} = 0$. Prediksi $\hat{Y} = 1$ jika $P\left(Y \middle| \hat{\boldsymbol{\beta}}\right) \geq 0.5$ dan $\hat{Y} = 0$ jika $P\left(Y \middle| \hat{\boldsymbol{\beta}}\right) < 0.5$. Prediksi ini tidak selalu benar, dan kesesuaiannya dengan keberadaan respons menurut pengamatan (*observed data*) akan menentukan kualitas diskriminatorik suatu model regresi logistik, dengan model regresi logistik berperan sebagai uji diagnostik. Nilai *cutoff* tidak selalu harus 0.5, melainkan dapat diubah menurut keperluan peneliti. Tabel silang hubungan antara data pengamatan dengan prediksi respons diperlihatkan pada Tabel 7.1 dan 7.2.

Tabel 7.1. Tabel Klasifikasi: Hubungan antara prediksi dengan pengamatan respons

		Respons (pengamatan; *observed*)	
		Ada: $Y = 1$	**Tidak ada: $Y = 0$**
Prediksi Respons	**Ada: $\hat{Y} = 1$**	**Positif Benar:** n_{TP} ***a***	**Positif Palsu:** n_{FP} ***b***
	Tidak ada $\hat{Y} = 0$	***c*** **Negatif Palsu:** n_{FN}	***d*** **Negatif Benar:** n_{TN}

n_{TP} : Jumlah subjek yang terdeteksi positif benar (*true positive*)

n_{FP} : Jumlah subjek yang terdeteksi positif palsu (*false positive*)

n_{FN} : Jumlah subjek yang terdeteksi negatif palsu (*false negative*)

n_{TN} : Jumlah subjek yang terdeteksi negatif benar (*true negative*)

Tabel 7.2. Karakteristik dan definisi pada uji diagnostik

a. Sensitivitas dan Spesifisitas

		Pengamatan Respons (Data)		
		$Y = 1$	$Y = 0$	
Prediksi Respons	$\hat{Y} = 1$	a	b	$a + b$
	$\hat{Y} = 0$	c	d	$c + d$
		$a + c$	$b + d$	$a + b + c + d$

$$Se = \frac{a}{a+c} \qquad Sp = \frac{d}{b+d} \tag{7.1}$$

Se : Sensitivitas = $a/(a+c)$

Sp : Spesifisitas = $d/(b+d)$

b. Nilai Prediksi Positif dan Nilai Prediksi Negatif

		Pengamatan Respons (Data)			
		$Y = 1$	$Y = 0$		
Prediksi Respons	$\hat{Y} = 1$	a	b	$a + b$	$PPV = \frac{a}{a+b}$
	$\hat{Y} = 0$	c	d	$c + d$	$NPV = \frac{d}{c+d}$
		$a + c$	$b + d$	$a + b + c + d$	

PPV : Nilai prediksi positif (*positive predictive value*) = $a/(a+b)$

NPV : Nilai prediksi negatif (*negative predictive value*) = $d/(c+d)$

Kualitas tampilan diskriminatorik dinilai dengan dua parameter, yaitu **sensitivitas** dan **spesifisitas**-nya (tabel 7.2.a). Kedua parameter ini memiliki nilai yang konstan, yaitu bernilai sama dimanapun uji diskriminatorik dilakukan. Selain itu ada pula kuantitas yang dinamakan **nilai prediksi positif** dan **nilai prediksi negatif**. Kedua kuantitas terakhir dapat memiliki nilai yang berbeda jika uji dilakukan di tempat dengan kondisi yang berbeda.

- **Sensitivitas** (*Se*): Proporsi prediksi positif di antara yang memberi respons
- **Spesifisitas** (*Sp*): Proporsi prediksi negatif di antara yang tidak memberi respons
- **Nilai prediksi positif** (*PPV*) = Proporsi yang memberi respons di antara prediksi positif
- **Nilai prediksi negatif** (*NPV*) = Proporsi yang tidak memberi respons di antara prediksi negatif

Perintah Stata untuk memperoleh nilai-nilai sensitivitas, spesifisitas, nilai prediksi positif, dan nilai prediksi negatif adalah:

estat classification, cutoff(#)

yang langsung diberikan setelah *fitting* model regresi logistik. Nilai *default* untuk titik *cutoff* adalah 0.5.

Contoh 7.3:

```
. use "D:\Analisis Regresi Logistik\Data\apilog.dta", clear

. logit hiqual yr_rnd avg_ed
```

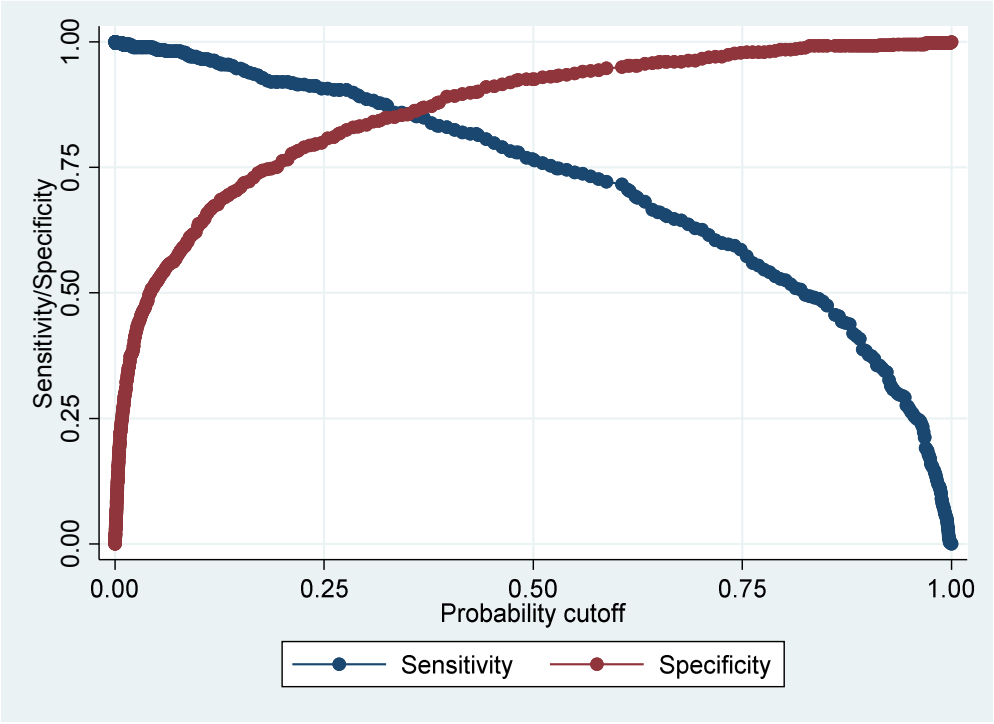
. estat classification

```
Logistic model for hiqual

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |       288            58  |        346
     -     |        89           723  |        812
-----------+--------------------------+-----------
   Total   |       377           781  |       1158

Classified + if predicted Pr(D) >= .5
True D defined as hiqual != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   76.39%
Specificity                     Pr( -|~D)   92.57%
Positive predictive value       Pr( D| +)   83.24%
Negative predictive value       Pr(~D| -)   89.04%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    7.43%
False - rate for true D         Pr( -| D)   23.61%
False + rate for classified +   Pr(~D| +)   16.76%
False - rate for classified -   Pr( D| -)   10.96%
--------------------------------------------------
Correctly classified                        87.31%
--------------------------------------------------
```

Tampak bahwa model peneliti memiliki kemampuan yang tinggi untuk mendeteksi subjek yang tidak memberi respons, yaitu dengan spesifisitas 92.6%, sebaliknya kemampuan untuk mendeteksi subjek yang memberi respons tidak terlalu tinggi, yaitu dengan sensitivitas yang hanya 76.4%. Secara keseluruhan, proporsi subjek yang diklasifikasikan dengan benar cukup baik, yaitu 87.3%.

Grafik sensitivitas dan spesifisitas untuk berbagai titik *cutoff* diperlihatkan sebagai berikut:

. lsens



Tampak bahwa semakin tinggi nilai titik *cutoff*, semakin rendah sensitivitas, sebaliknya spesifisitas menjadi semakin besar. Pada tabel 7.3 tampak bahwa dengan *cutoff* 0.00 (model ekstrim I), sensitivitas akan menjadi 1.00 karena seluruh respons akan terdeteksi, walaupun sebagian di antara prediksi respons adalah positif palsu (FP), sedangkan spesifisitas

menjadi 0.00 karena tidak ada respons negatif yang terdeteksi. Sebaliknya pada *cutoff* 1.00 (model ekstrim II), spesifisitas yang akan menjadi 1.00, sedangkan sensitivitas menjadi 0.00 karena seluruh prediksi respons adalah negatif, sehingga seluruh respons negatif terdeteksi walaupun sebagian di antaranya adalah negatif palsu (FN).

Tabel 7.3 Model diagnostika ekstrim

Model I:

***co* = 0.00: *Se* = 1.00, *Sp* = 0.00**

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	$\boldsymbol{n}_{\text{TP}}$	$\boldsymbol{n}_{\text{FP}}$
$\hat{Y} = 0$	0	0

Model II:

***co* = 1.00: *Se* = 0.00, *Sp* = 1.00**

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	0	0
$\hat{Y} = 0$	$\boldsymbol{n}_{\text{FN}}$	$\boldsymbol{n}_{\text{TN}}$

co : *cut-off point*

❖ Rasio Likelihood Positif dan Negatif

Rasio likelihood positif adalah:

$$\boxed{\text{LR+} = \frac{Se}{1-Sp}} \qquad (7.2.a)$$

Rasio likelihood negatif adalah:

$$\boxed{\text{LR}- = \frac{1-Se}{Sp}} \qquad (7.2.b)$$

Rasio likelihood positif dapat dinyatakan sebagai rasio antara proporsi subjek yang memberi respons dengan proporsi subjek yang tak memberi respons di antara subjek yang prediktornya positif. Dengan sudut pandang yang sama, rasio likelikood negatif adalah rasio antara proporsi

subjek memberi respons dengan proporsi subjek yang tak memberi respons di antara subjek yang prediktornya negatif. Karena nilai rasio likelihood positif dan negatif ini hanya dihitung berdasarkan sensitivitas dan spesifisitas, nilai-nilainya juga konstan, yaitu bernilai sama dimanapun uji diskriminatorik dilakukan.

Nilai rasio likelihood positif biasanya lebih besar daripada 1, sedangkan nilai rasio likelihood negatif umumnya lebih kecil daripada 1.

Contoh 7.4:

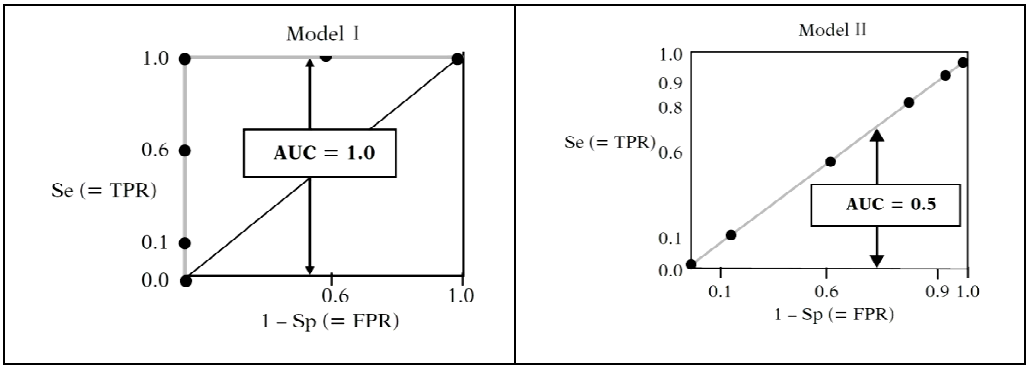
Lihat kembali hasil analisis data pada contoh 7.3. Sensitivitas adalah 76.39% dan spesifisitas adalah 92.57%. Maka rasio likelihood positif dan negatif masing-masing adalah:

```
. display 0.7639/(1 – 0.9257)
10.281292

. display (1 – 0.7639)/0.9257
.25505023
```

❖ Kurve ROC

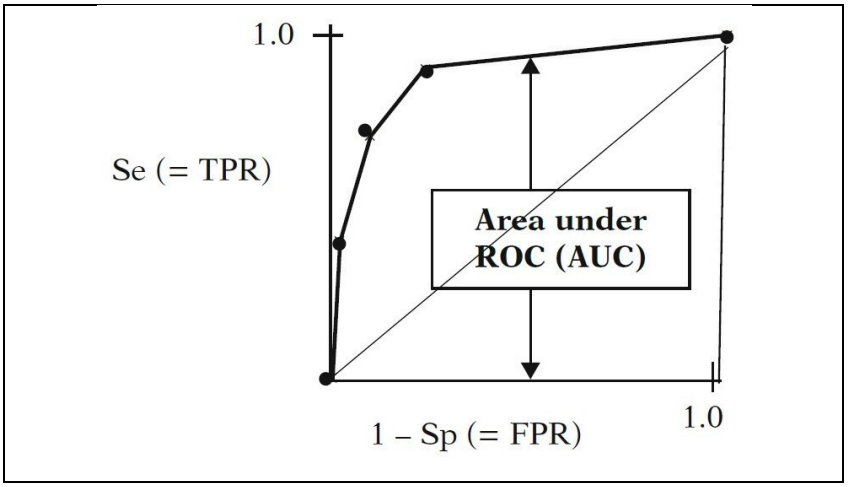
Kurve ROC (*Receiver Operating Characterictic*) pada mulanya dirancang untuk mendeteksi sinyal elektronik, antara lain sinyal radar di medan perang. Yang dibahas di sini adalah kurve ROC untuk model logistik, yaitu grafik sensitivitas (*Se*) pada sumbu *Y* dengan 1 – spesifisitas (1 – *Sp*) pada sumbu *X*.

Kurve ROC selain menampilkan kemampuan diskriminatorik model regresi logistik, juga menunjukkan seberapa baik *fitting* model dengan data yang ada. Beberapa contoh kurve ROC diperlihatkan pada gambar 7.1 dan 7.2 berikut. Kurve dibentuk oleh titik-titik potong nilai *Se* dengan (1 – *Sp*).



Gambar 7.1 Kurve ROC dengan AUC = 1.0 dan AUC = 0.5

Luas area di bawah kurve dinamakan AUC (*Area under ROC*), merupakan ukuran seberapa baiknya tampilan diskriminatorik model regresi maupun *fitting*-nya dengan data. Kurve pada gambar 7.1 kiri memperlihatkan AUC = 1.0 untuk model dengan tampilan diskriminatorik sempurna, sedangkan kurve pada gambar 7.1 kanan memperlihatkan AUC = 0.5 untuk model dengan tanpa tampilan diskriminatorik. Pada umumnya model regresi logistik memiliki nilai AUC di antara 0.5 dan 1.0 (gambar 7.2).



Gambar 7.2 Contoh kurve ROC

Semakin besar luas AUC, semakin baik tampilan diskriminatorik dan *fitting* suatu model regresi logistik. Pedoman kasar untuk penilaian AUC adalah sebagai berikut:

- 0.90-1.0 : Diskriminatorik sangat baik (*excellent*)
- 0.80-0.90 : Diskriminatorik baik (*good*)
- 0.70-0.80 : Diskriminatorik sedang (*fair*)
- 0.60-0.70 : Diskriminatorik buruk (*poor*)
- 0.50-0.60 : Diskriminatorik gagal (*failed*)

Kurve pada Stata dapat diperoleh dengan perintah **lroc** atau **roctab** yang diberikan langsung setelah *fitting* model. Untuk perintah **roctab** harus terlebih dahulu diprediksi probabilitas respons (misalnya disimpan dalam variabel **prob**), lalu diberikan perintah:

roctab *resp_var* **prob, graph**

resp_var : Variabel respons

Contoh 6.4:

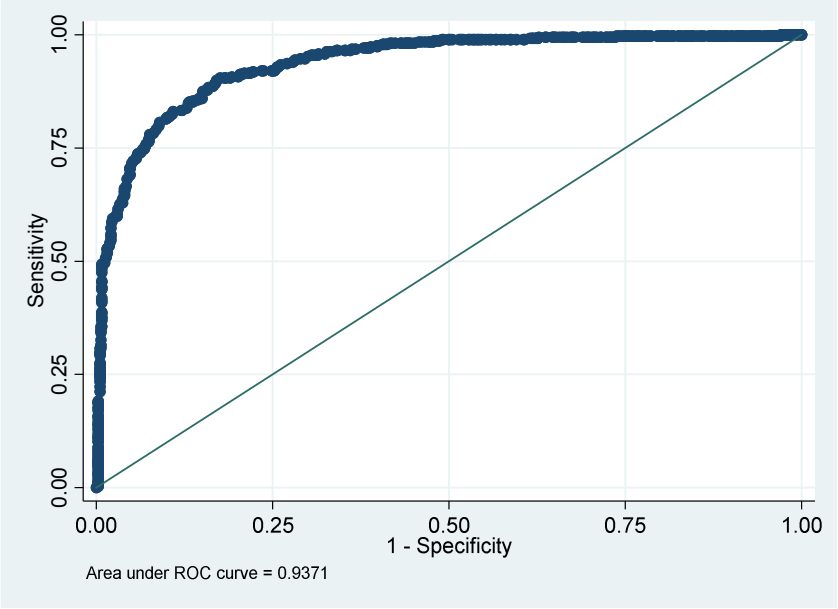
Lihat kembali file data pada contoh 7.3:

. use “D:\Analisis Regresi Logistik\Data\apilog.dta”, clear

. quietly logit hiqual yr_rnd avg_ed

. lroc

```
Logistic model for hiqual

number of observations =      1158
area under ROC curve   =    0.9371
```

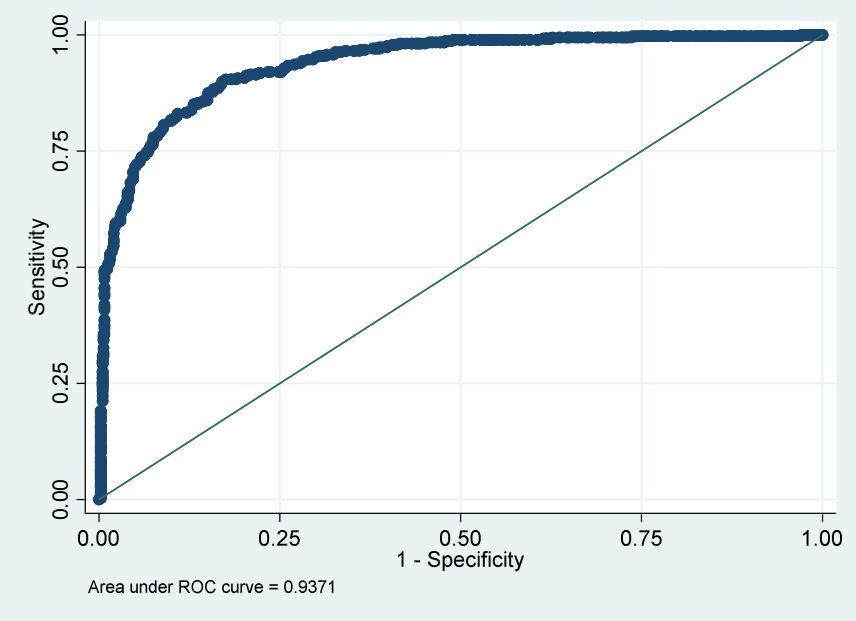


Luas AUC adalah 0.9371, yang menurut pedoman di atas tergolong diskriminatorik sangat baik (*excellent*). Selanjutnya diperlihatkan perolehan kurve ROC dengan perintah **roctab**.

. predict prob

```
(option pr assumed; Pr(hiqual))
(42 missing values generated)
```

. roctab hiqual prob, graph



Tampak kurve ROC yang diperoleh berikut luas area AUC sama dengan yang diperoleh dengan perintah **lroc**.

BAB 8
REGRESI LOGISTIK KONDISIONAL

Regresi logistik kondisional (*conditional logistic regression*) digunakan untuk data berpasangan (*paired subjects*), dengan tiap subjek memiliki respons positif ($Y = 1$) dipadankan dengan 1 atau lebih subjek memiliki respons negatif ($Y = 0$). Dalam bidang Epidemiologi rancangan demikian digunakan pada studi kasus-kontrol dengan *matching*. subjek dengan respons positif dinamakan sebagai kasus dan padanannya dengan respons negatif dinamakan kontrol. Tiap subjek dengan $Y = 1$ beserta 1 atau lebih padanannya dengan $Y = 0$ membentuk 1 grup.

Struktur data ini tidak sama dengan yang digunakan pada uji *t* berpasangan. Pada uji *t* berpasangan untuk studi eksperimental yang dipadankan adalah variabel independennya, yaitu satu anggota pasangan yang menerima perlakuan (*treatment*) yang diujicobakan dipadankan dengan satu anggota lain yang tidak menerima perlakuan sebagai kontrol.

❖ Tabel 2×2 dan Rasio Odds untuk Data Berpasangan

Penyajian ringkasan data biner berpasangan dalam bentuk tabel 2×2 tidak sama dengan penyajian untuk data independen (tak-berpasangan) seperti pada tabel 1.2. Untuk data biner berpasangan perlu diperhitungkan apakah suatu pasangan memiliki prediktor yang sama ($X_1 = 0$; $X_2 = 0$) atau ($X_1 = 1$; $X_2 = 1$) yang dinamakan pasangan konkordan (*concordant*); atau prediktor yang tidak sama ($X_1 = 1$; $X_2 = 0$) atau ($X_1 = 0$; $X_2 = 1$) yang dinamakan pasangan diskordan (*disconcordant*). Penyajian ringkasan data biner berpasangan demikian diperlihatkan pada tabel 8.1.

Tabel 8.1 Tabel 2×2 untuk data biner berpasangan

		Kontrol	
Kasus	$X_2 = 1$	$X_2 = 0$	
$X_1 = 1$	*e*	*f*	*a*
$X_1 = 0$	*g*	*h*	*b*
	c	*d*	*n'*

e dan *h* : Jumlah pasangan konkordan (1 ; 1) dan (0 ; 0)

f dan *g* : Jumlah pasangan diskordan (1 ; 0) dan (0 ; 1)

a : Jumlah kasus dengan prediktor positif

b : Jumlah kasus dengan prediktor negatif

c : Jumlah kontrol dengan prediktor positif

d : Jumlah kontrol dengan prediktor negatif

n' : Jumlah pasangan. Jumlah anggota sampel seluruhnya adalah $n = 2n'$.

Rasio odds adalah:

$$\boxed{\hat{OR} = \frac{f}{g}} \qquad (8.1)$$

Rasio odds f/g berdistribusi log-normal dengan transformasi logaritmanya yaitu ln $\hat{OR}$ berdistribusi normal dengan variansi $(1/f + 1/g)$. Rasio odds untuk data berpasangan ini dihitung sebagai rasio odds gabungan untuk *n'* strata, tiap grup merupakan 1 stratum. Rasio odds gabungannya dapat dihitung menggunakan rumus Mantel-Haenzel, yaitu:

$$\boxed{\hat{OR}_{\text{MH}} = \frac{\sum_{i=1}^{n'}(a_i d_i / n')}{\sum_{i=1}^{n'}(b_i c_i / n')}} \quad ; \quad i = 1, 2, \ldots, n' \qquad (8.2)$$

Tabel di atas memperlihatkan penyajian ringkasan data biner untuk 1 : 1 *matching*, yang disebut juga sebagai *pair-matching* dan paling sering digunakan. Dalam praktik dapat digunakan 1 : *m matching* dengan *m* berkisar antara 1 sampai dengan 5. Untuk 1 : 2 *matching*, yaitu tiap 1 kasus dipadankan dengan 2 kontrol, yang disebut juga sebagai *triplet-matching*, tabel penyajiannya adalah sebagai berikut:

Tabel 8.2 Tabel 2 × 3 untuk triplet matching

	Kontrol			
Kasus	2 prediktor positif	1 prediktor positif	0 prediktor positif	
$X_1 = 1$	f_2	f_1	f_0	a
$X_1 = 0$	g_2	g_1	g_0	b
	c_2	c_1	c_0	n'

Rasio odds untuk *triplet-matching* adalah:

$$\boxed{\hat{OR} = \frac{2f_0 + f_1}{2g_2 + g_1}} \qquad (8.3)$$

Perintah Stata untuk mengestimasi rasio odds untuk data *matched* adalah:

mhodds *resp_var pred_var* [*adjust_var(s)*] [*if*] [*in*] [, *mhodds_options*]

resp_var : Variabel respons

pred_var : Variabel prediktor

adjust_var(*s*) : (Himpunan) variabel kendali

Contoh 8.1:

. use "D:\Analisis Regresi Logistik\Data\lowbirth2.dta", clear

(Applied Logistic Regression, Hosmer & Lemeshow)

Menghitung rasio odds **ht** (prediktor) terhadap **low** (respons):

. mhodds low ht

```
Maximum likelihood estimate of the odds ratio
Comparing ht==1 vs. ht==0

------------------------------------------------------
Odds Ratio   chi2(1)     P>chi2     [95% Conf. Interval]
------------------------------------------------------
  2.523810      1.74     0.1870      0.607168  10.490689
------------------------------------------------------
```

Mengestimasi rasio odds **ht** (prediktor) terhadap **low** (respons) dengan mengendalikan faktor **smoke**:

. mhodds low ht smoke

```
Mantel-Haenszel estimate of the odds ratio
Comparing ht==1 vs. ht==0, controlling for smoke

------------------------------------------------------------
Odds Ratio   chi2(1)     P>chi2     [95% Conf. Interval]
------------------------------------------------------------
  2.617446      1.90     0.1686      0.630599  10.864307
------------------------------------------------------------
```

❖ Regresi Logistik Kondisional untuk 1 : 1 Matching

Regresi logistik kondisional digunakan untuk pemodelan regresi logistik data prediktor-respons dengan respons biner $Y = 1$ (kasus) dan $Y = 0$ (kontrol). Perintah Stata untuk regresi logistik kondisional adalah:

clogit *depvar indepvar(s)* [*if*] [*in*], **group**(*var_name*) [*options*]

depvar	:	Variabel dependen
indepvar(*s*)	:	(Himpunan) variabel independen
var_name	:	Nama variabel grup *matching*

Untuk mendapatkan estimasi nilai-nilai rasio odds, digunakan opsi [, **or**]. Seperti pada regresi logistik tak berpasangan, estimasi rasio odds adalah $\hat{OR}_j = e^{\hat{\beta}_j}$, walaupun hasil yang diperoleh tidak tepat sama dengan hasil perintah **mhodds**, karena metode estimasi yang digunakan berbeda.

Perhatikan bahwa variabel yang dipadankan tidak dapat dinilai hubungannya dengan respons, dan tidak boleh dimasukkan sebagai salah satu variabel independen dalam model.

Contoh 8.2:

. use "D:\Analisis Regresi Logistik\Data\lowbirth2.dta", clear

```
(Applied Logistic Regression, Hosmer & Lemeshow)
```

. sum

```
Variable |     Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
  pairid |     112        28.5    16.23587         1         56
     low |     112          .5    .5022472         0          1
     age |     112        22.5    4.316321        14         34
     lwt |     112    127.1696    30.46986        80        241
   smoke |     112    .4107143    .4941746         0          1
---------+--------------------------------------------------------
     ptd |     112    .2232143    .4182723         0          1
      ht |     112    .0892857    .2864373         0          1
      ui |     112    .1785714    .3847144         0          1
    race |     112    2.026786    .9050392         1          3
```

. list in 1/10

```
     +-------------------------------------------------------+
     | pairid  low   age   lwt  smoke   ptd   ht   ui    race |
     |-------------------------------------------------------|
  1. |      1    0    14   135      0     0    0    0   white |
  2. |      1    1    14   101      1     1    0    0   other |
  3. |      2    0    15    98      0     0    0    0   black |
  4. |      2    1    15   115      0     0    0    1   other |
  5. |      3    0    16    95      0     0    0    0   other |
     |-------------------------------------------------------|
  6. |      3    1    16   130      0     0    0    0   other |
  7. |      4    0    17   103      0     0    0    0   other |
  8. |      4    1    17   130      1     1    0    1   other |
  9. |      5    0    17   122      1     0    0    0   white |
 10. |      5    1    17   110      1     0    0    0   white |
     +-------------------------------------------------------+
```

Tampak bahwa **age** merupakan variabel yang dipadankan (di-*matched*), karena nilainya selalu sama untuk anggota **pairid** yang sama, karena itu tidak boleh dimasukkan sebagai salah satu prediktor dalam model regresi.

. tab race

```
    race of |
    mother: |
   1=white, |
   2=black, |
    3=other |      Freq.     Percent        Cum.
------------+-----------------------------------
      white |          44       39.29       39.29
      black |          21       18.75       58.04
      other |          47       41.96      100.00
------------+-----------------------------------
      Total |         112      100.00
```

. clogit low lwt smoke ptd ht ui i.race, group(pairid)

```
Iteration 0: log likelihood = -26.768693
Iteration 1: log likelihood = -25.810476
Iteration 2: log likelihood = -25.794296
Iteration 3: log likelihood = -25.794271
Iteration 4: log likelihood = -25.794271

Conditional (fixed-effects) logistic regression

                                  Number of obs =    112
                                  LR chi2(7)    =  26.04
                                  Prob > chi2   = 0.0005
Log likelihood = -25.794271       Pseudo R2     = 0.3355
```

```
------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lwt |  -.0183757   .0100806    -1.82   0.068    -.0381333    .0013819
       smoke |   1.400656   .6278396     2.23   0.026     .1701131    2.631199
         ptd |   1.808009   .7886502     2.29   0.022     .2622828    3.353735
          ht |   2.361152   1.086128     2.17   0.030     .2323796    4.489924
          ui |   1.401929   .6961585     2.01   0.044     .0374836    2.766375
             |
        race |
       black |   .5713643    .689645     0.83   0.407    -.7803149    1.923044
       other |  -.0253148   .6992044    -0.04   0.971     -1.39573    1.345101
------------------------------------------------------------------------------
```

. clogit, or

```
Conditional (fixed-effects) logistic regression

                                                Number of obs =    112
                                                LR chi2(7)    =  26.04
                                                Prob > chi2   = 0.0005
Log likelihood = -25.794271                     Pseudo R2     = 0.3355
```

```
------------------------------------------------------------------------------
         low | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lwt |   .9817921    .009897    -1.82   0.068     .9625847    1.001383
       smoke |   4.057862   2.547686     2.23   0.026     1.185439    13.89042
         ptd |   6.098293    4.80942     2.29   0.022     1.299894    28.60938
          ht |   10.60316   11.51639     2.17   0.030     1.261599    89.11467
          ui |    4.06303   2.828513     2.01   0.044     1.038195    15.90088
             |
        race |
       black |   1.770681   1.221141     0.83   0.407     .4582617     6.84175
       other |    .975003   .6817263    -0.04   0.971     .2476522    3.838573
------------------------------------------------------------------------------
```

Bandingkan hasil estimasi rasio odds di sini dengan contoh 8.1:

. clogit low ht, group(pairid) or

```
Iteration 0: log likelihood = -37.999811
Iteration 1: log likelihood = -37.993415
Iteration 2: log likelihood = -37.993413

Conditional (fixed-effects) logistic regression

                                                Number of obs =    112
                                                LR chi2(1)    =   1.65
                                                Prob > chi2   = 0.1996
Log likelihood = -37.993413                     Pseudo R2     = 0.0212
```

```
------------------------------------------------------------------------------
         low | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ht |   2.333333   1.610152    1.23   0.220     .6033812    9.023218
------------------------------------------------------------------------------
```

Tampak estimasi rasio odds $\hat{OR} = 2.33$, mendekati tetapi tidak sama dengan estimasi rasio odds $\hat{OR} = 2.52$ pada contoh 8.1.

. clogit low ht smoke, group(pairid) or

```
Iteration 0:  log likelihood = -34.269074
Iteration 1:  log likelihood = -34.264559
Iteration 2:  log likelihood = -34.264558

Conditional (fixed-effects) logistic regression

                                                Number of obs =    112
                                                LR chi2(2)    =   9.10
                                                Prob > chi2   = 0.0105
Log likelihood = -34.264558                     Pseudo R2     = 0.1173
```

```
------------------------------------------------------------------------------
         low | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ht |   2.910674   2.150782     1.45   0.148     .6839368    12.38715
       smoke |   2.964466   1.266002     2.54   0.011      1.28361    6.846361
------------------------------------------------------------------------------
```

Estimasi rasio odds **ht** terhadap **low** dengan mengendalikan **smoke** di sini adalah 2.91 (bandingkan dengan hasil estimasi 2.62 pada contoh 8.1).

❖ Regresi Logistik Kondisional untuk 1 : *m* Matching

Perintah Stata untuk regresi logistik kondisional 1 : *m matching* sama saja dengan perintah untuk *pair-matching* di atas, karena untuk opsi **group** pada perintah **clogit** tidak dispesifikasikan jumlah anggota grup.

Contoh 8.3:

Contoh di sini adalah untuk *triplet-matching*, yaitu untuk tiap 1 kasus dengan respins **mi** = 1 terdapat 2 kontrol dengan respons **mi** = 0.

```
. use "D:\Analisis Regresi Logistik\Data\mi.dta"

. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
match	117	20	11.30304	1	39
person	117	59	33.91902	1	117
mi	117	.3333333	.4734321	0	1
smk	117	.2820513	.4519337	0	1
sbp	117	136.4103	16.10641	120	160
ecg	117	.2051282	.405532	0	1
survtime	117	1.666667	.4734321	1	2

. list match mi smk sbp ecg in 1/9

	match	mi	smk	sbp	ecg
1.	1	1	0	160	1
2.	1	0	0	140	0
3.	1	0	0	120	0
4.	2	1	0	160	1
5.	2	0	0	140	0
6.	2	0	0	120	0
7.	3	1	0	160	0
8.	3	0	0	140	0
9.	3	0	0	120	0

. clogit mi smk sbp ecg, strata(match)

```
Iteration 0: log likelihood = -32.362396
Iteration 1: log likelihood = -31.750966
Iteration 2: log likelihood = -31.745464
Iteration 3: log likelihood = -31.745464

Conditional (fixed-effects) logistic regression

                                  Number of obs =    117
                                  LR chi2(3)    =  22.20
                                  Prob > chi2   = 0.0001
Log likelihood = -31.745464       Pseudo R2     = 0.2591

------------------------------------------------------------------------------
          mi |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         smk |   .7290581   .5612571     1.30   0.194    -.3709857    1.829102
         sbp |   .0456419   .0152469     2.99   0.003     .0157586    .0755252
         ecg |   1.599263   .8534146     1.87   0.061     -.073399    3.271925
------------------------------------------------------------------------------
```

. clogit mi smk sbp ecg, strata(match) or

```
Iteration 0: log likelihood = -32.362396
Iteration 1: log likelihood = -31.750966
Iteration 2: log likelihood = -31.745464
Iteration 3: log likelihood = -31.745464
```

```
Conditional (fixed-effects) logistic regression

                                                Number of obs =    117
                                                LR chi2(3)    =  22.20
                                                Prob > chi2   = 0.0001
Log likelihood = -31.745464                     Pseudo R2     = 0.2591

------------------------------------------------------------------------------
          mi | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         smk |   2.073127   1.163557     1.30   0.194      .6900538    6.22829
         sbp |   1.046699   .0159589     2.99   0.003      1.015883    1.07845
         ecg |   4.949382   4.223875     1.87   0.061       .92923    26.36203
------------------------------------------------------------------------------
```

BAB 9
REGRESI LOGISTIK ORDINAL

❖ Pengertian Regresi Logistik Ordinal

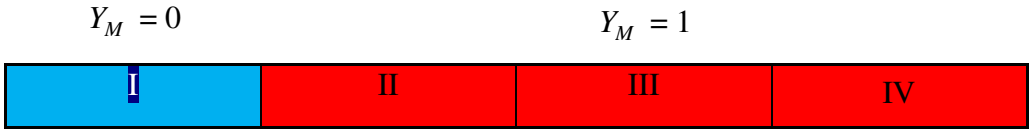
Regresi logistik ordinal adalah pemodelan regresi logistik untuk data prediktor-respons dengan respons kategorik ordinal non-biner (kategorik ordinal dengan jumlah kategori lebih daripada dua). Pengolahan data pada regresi logistik ordinal tetap dilakukan dengan menggunakan himpunan nilai prediktor yang sama, memisahkannya ke dalam dua bagian dengan respons modifikasi Y_M =1 dan Y_M = 0 seperti pada regresi logistik biasa, tetapi dilakukan secara berulang dengan memindah-mindahkan titik *cutoff* untuk respons-nya.

Misalkan dimiliki data prediktor-respons dengan respons kategorik ordinal yang memiliki 4 kategori, yaitu kategori I, II, III, dan IV. Maka regresi logistik biasa dilakukan 3 kali terhadap himpunan nilai prediktor yang sama, tetapi respons kategori I *vs* II-III-IV, respons kategori I-II *vs* III-IV, dan respons kategori I-II-III *vs* IV (gambar 9.1). Ketiga titik *cutoff* respons akan menjadi estimator konstante dalam tiap model.

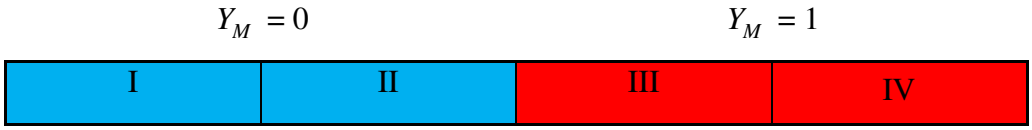
Sebagai hasil akan diperoleh 3 model regresi dengan estimasi koefisien regresi yang sama (karena menggunakan himpunan nilai prediktor yang sama), namun dengan konstante berbeda (karena menggunakan titik *cutoff* respons yang berbeda). Ketiga model ini biasanya disebut sebagai 1 model regresi saja, yaitu:

- Model pertama : $\text{logit}(Y_M) = \beta_{0-\text{I}} + \beta_1 X_1 + \ldots + \beta_p X_p$
- Model kedua : $\text{logit}(Y_M) = \beta_{0-\text{II}} + \beta_1 X_1 + \ldots + \beta_p X_p$
- Model ketiga : $\text{logit}(Y_M) = \beta_{0-\text{III}} + \beta_1 X_1 + \ldots + \beta_p X_p$

Regresi logistik pertama:



Regresi logistik kedua:



Regresi logistik ketiga

Gambar 9.1 Regresi logistik ordinal untuk respons dengan 4 kategori ordinal

❖ Regresi Logistik Ordinal dengan Stata

Berikut diperlihatkan *template* keluaran tabel koefisien regresi logistik ordinal dengan respons kategorik ordinal dengan 4 kategori pada Stata (tabel 9.1). Tampak tampilan 2 baris teratas pertama adalah sama seperti regresi logistik biasa, kecuali tidak ada suku konstante. Yang berbeda ialah adanya tambahan baris ketiga berupa nilai-nilai *cutoff* respons, yang merupakan estimator konstante pada ketiga model.

Perintah Stata untuk *fitting* regresi logistik ordinal adalah:

ologit *depvar indepvars* [*if*] [*in*] [, *options*]

depvar : Respons kategorik ordinal

indepvars: Himpunan prediktor

Untuk memperoleh estimasi nilai-nilai rasio odds, digunakan opsi **or**.

Tabel 9.1 Contoh tabel koefisien regresi logistik ordinal pada Stata

respons	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
prediktor 1	. . .	. . .	. .	. . .	. . .	. . .
prediktor 2	. . .	. . .	. .	. . .	. . .	. . .
prediktor 3	. . .	. . .	. .	. . .	. . .	. . .
/cut1	. . .	. . .			. . .	. . .
/cut2	. . .	. . .			. . .	. . .
/cut3	. . .	. . .			. . .	. . .

Contoh 9.1:

File data di sini memuat data tentang 400 orang lulusan sekolah menengah. Dengan prediktor pendidikan orang tua (**pared**), jenis sekolah menengah (swasta atau negeri; **public**), dan nilai IPK (**gpa**), respons-nya

adalah kecenderungan siswa untuk melamar ke jenjang perguruan tinggi (**apply**).

. use "D:\Analisis Regresi Logistik\ologit.dta", clear

. tab apply

```
          apply |      Freq.     Percent        Cum.
----------------+-----------------------------------
       unlikely |         220       55.00       55.00
somewhat likely |         140       35.00       90.00
    very likely |          40       10.00      100.00
----------------+-----------------------------------
          Total |         400      100.00
```

. tab apply, nolab

```
      apply |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         220       55.00       55.00
          1 |         140       35.00       90.00
          2 |          40       10.00      100.00
------------+-----------------------------------
      Total |         400      100.00
```

. tab apply pared

	At Least One Parent	
Applying to	has a Graduate Degree	
Graduate School	0	1	Total
unlikely	200	20	220
somewhat likely	110	30	140
very likely	27	13	40
Total	337	63	400

. tab apply public

	Undergraduate	
	Institution is Public	
Applying to	or Private	
Graduate School	0	1	Total
unlikely	189	31	220
somewhat likely	124	16	140
very likely	30	10	40
Total	343	57	400

. summarize gpa

Variable	Obs	Mean	Std. Dev.	Min	Max
gpa	400	2.998925	.3979409	1.9	4

. table apply, cont(mean gpa sd gpa)

```
-------------------------------------
Applying to     |
Graduate School |  mean(gpa)     sd(gpa)
----------------+--------------------
       unlikely |   2.952136     .403594
somewhat likely |   3.030071    .3893446
    very likely |    3.14725    .3560322
-------------------------------------
```

Pada perintah **ologit** berikut, **i.** di depan **pared** mengindikasikan bahwa **pared** adalah variabel kategorik dan akan diinklusikan dalam model sebagai serangkaian variabel indikator. Hal yang sama berlaku untuk **i.public**.

. ologit apply i.pared i.public gpa

```
Iteration 0: log likelihood = -370.60264
Iteration 1: log likelihood =   -358.605
Iteration 2: log likelihood = -358.51248
Iteration 3: log likelihood = -358.51244
Iteration 4: log likelihood = -358.51244

Ordered logistic regression           Number of obs =    400
                                      LR chi2(3)    =  24.18
                                      Prob > chi2   = 0.0000
Log likelihood = -358.51244           Pseudo R2     = 0.0326
```

```
------------------------------------------------------------------------------
       apply |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.pared |    1.047664   .2657891     3.94   0.000     .5267266    1.568601
   1.public |   -.0586828   .2978588    -0.20   0.844    -.6424754    .5251098
        gpa |    .6157458   .2606311     2.36   0.018     .1049183    1.126573
-------------+----------------------------------------------------------------
      /cut1 |    2.203323   .7795353                      .6754621    3.731184
      /cut2 |    4.298767   .8043147                      2.72234    5.875195
------------------------------------------------------------------------------
```

Statistik khi-kuadrat ratio *likelihood* sebesar 24.18 dengan nilai-*p* 0.0000 menyatakan model secara keseluruhan bermakna, dibandingkan dengan model nol tanpa prediktor.

Variabel **pared** dan **gpa** keduanya bermakna; variabel **public** tak bermakna. Untuk **pared**, peningkatan satu satuannya (dari 0 menjadi 1), akan menghasilkan peningkatan 1.05 kali lipat untuk log odds **apply**, dengan syarat semua variabel lain dalam model konstan. Untuk satu satuan peningkatan **gpa**, akan diperoleh 0.62 kali lipat log odds **apply**, dengan syarat semua variabel lain dalam model konstan. Titik potong (*cutpoints*) yang dilaporkan pada bagian akhir keluaran menyatakan lokasi variabel laten dipotong untuk menghasilkan tiga grup yang terbentuk pada data. Perhatikan bahwa variabel laten ini adalah kontinu.

Nilai rasio odds dapat diperoleh dengan mencantumkan opsi **or** setelah perintah **ologit**.

. ologit apply i.pared i.public gpa, or

```
Iteration 0: log likelihood = -370.60264
Iteration 1: log likelihood =   -358.605
Iteration 2: log likelihood = -358.51248
Iteration 3: log likelihood = -358.51244
Iteration 4: log likelihood = -358.51244

Ordered logistic regression                     Number of obs =    400
                                                LR chi2(3)    =  24.18
                                                Prob > chi2   = 0.0000
Log likelihood = -358.51244                     Pseudo R2     = 0.0326

------------------------------------------------------------------------------
       apply | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.pared |   2.850982   .7577601     3.94   0.000     1.69338    4.799927
    1.public |   .9430059   .2808826    -0.20   0.844    .5259888    1.690645
         gpa |   1.851037   .4824377     2.36   0.018     1.11062    3.085067
-------------+----------------------------------------------------------------
       /cut1 |   2.203323   .7795353                      .6754621    3.731184
       /cut2 |   4.298767   .8043147                      2.72234    5.875195
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation.
```

Pada keluaran di atas, hasil ditampilkan sebagai rasio odds proporsional. Untuk **pared**, satu satuan pertambahan nilainya, yaitu dari 0 menjadi 1, odds untuk **apply high** *vs* kategori kombinasi **middle-low** akan menjadi 2.85 kali lebih besar, dengan syarat semua variabel lain dalam model tetap konstan. Demikian pula, odds kombinasi kategori **apply**

middle-high *vs* **low** akan menjadi 2.85 kali lebih besar, dengan syarat semua variabel dalam model tetap konstan. Untuk satu unit peningkatan **gpa**, odds kategori **apply high** *vs* kategori **apply low-middle** menjadi 1.85 kali lebih besar, dengan syarat semua variabel lain dalam model tetap konstan. Berdasarkan asumsi proporsional odds, peningkatan yang sama sebesar 1.85 kali, juga ditemukan antara **apply low** *vs* kategori kombinasi **apply middle-high**.

Contoh 9.2:

File data di sini memuat data tentang 74 buah mobil di Amerika Serikat.

. use "D:\Analisis Regresi Logistik\Data\fullauto.dta", clear

```
(Automobile Models)
```

Variabel-variabel:

rep77 : Kondisi mobil pada saat reparasi pada tahun 1977

length : Panjang mobil dalam inci

mpg : Jarak tempuh mobil untuk 1 gallon BBM (*miles per gallon*)

. list rep77 mpg length in 1/10

```
     +------------------------+
     |   rep77   mpg   length |
     |------------------------|
  1. |    Fair    22      186 |
  2. |    Poor    17      173 |
  3. |       .    22      168 |
```

```
  4. | Average    23       174 |
  5. |    Fair    17       189 |
     |-----------------------|
  6. |    Good    25       177 |
  7. | Average    20       196 |
  8. |    Good    15       222 |
  9. |    Good    18       218 |
 10. |       .    26       170 |
     +-----------------------+
```

. tab rep77

```
     Repair |
Record 1977 |  Freq.   Percent    Cum.
------------+-----------------------
       Poor |      3      4.55    4.55
       Fair |     11     16.67   21.21
    Average |     27     40.91   62.12
       Good |     20     30.30   92.42
  Excellent |      5      7.58  100.00
------------+-----------------------
      Total |     66    100.00
```

rep77 adalah variabel kategorik ordinal, akan menjadi variabel dependen pada regresi logistik ordinal berikut:

. ologit rep77 mpg length

```
Iteration 0: log likelihood = -89.895098
Iteration 1: log likelihood = -86.062239
Iteration 2: log likelihood = -86.015487
Iteration 3: log likelihood = -86.015382
Iteration 4: log likelihood = -86.015382

Ordered logistic regression       Number of obs =     66
                                  LR chi2(2)    =   7.76
                                  Prob > chi2   = 0.0207
Log likelihood = -86.015382       Pseudo R2     = 0.0432

------------------------------------------------------------------------------
       rep77 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |   .1760509   .0675753     2.61   0.009     .0436057     .308496
      length |   .0322272   .0176631     1.82   0.068    -.0023918    .0668463
-------------+----------------------------------------------------------------
       /cut1 |   6.656575   4.570512                     -2.301464    15.61461
       /cut2 |   8.455557   4.573731                     -.5087906     17.4199
       /cut3 |   10.41195   4.633466                      1.33052    19.49337
       /cut4 |   12.59688   4.732698                      3.320966     21.8728
------------------------------------------------------------------------------
```

Pada regresi logistik ordinal, yang diperbandingkan selalu tetap 2 kelompok, sehingga analisis dilakukan dalam 4 tahap:

1. **rep77(2)**, **rep77(3)**, **rep77(4)**, dan **rep77(5)** seluruhnya diberi nilai 1; **rep77(1)** diberi nilai 0; lalu lakukan regresi logistik biasa terhadap 2 kelompok perbandingan.

2. **rep77(3)**, **rep77(4)**, dan **rep77(5)** seluruhnya diberi nilai 1; **rep77(1)** dan **rep77(2)** diberi nilai 0; lalu lakukan regresi logistik biasa terhadap 2 kelompok perbandingan.

3. **rep77(4)** dan **rep77(5)** diberi nilai 1; **rep77(1)**, **rep77(2)** dan **rep77(3)** diberi nilai 0; lalu lakukan regresi logistik biasa terhadap 2 kelompok perbandingan.

4. **rep77(5)** diberi nilai 1; **rep77(1)**, **rep77(2)**, **rep77(3)**, dan **rep77(4)** seluruhnya diberi nilai 0; lalu lakukan regresi logistik biasa terhadap 2 kelompok perbandingan.

Diperoleh model empirik:

rep77: **Poor** *vs* **Fair-Average-Good-Excellent**

logit (**rep77**) = 6.6566 + 0.1761(**mpg**) + 0.0322(**length**)

rep77: **Poor-Fair** *vs* **Average-Good-Excellent**

logit (**rep77**) = 8.4556 + 0.1761(**mpg**) + 0.0322(**length**)

rep77: **Poor-Fair-Average** *vs* **Good-Excellent**

logit (**rep77**) = 10.4119 + 0.1761(**mpg**) + 0.0322(**length**)

rep77: **Poor-Fair-Average-Good** *vs* **Excellent**

logit (**rep77**) = 12.5969 + 0.1761(**mpg**) + 0.0322(**length**)

Pada regresi logistik ordinal juga berlaku:

$$\hat{OR} = \exp\ b_1 \quad \text{atau} \qquad b_1 = \ln\ \hat{OR}$$

Estimasi nilai rasio odds diperoleh dengan menambahkan opsi ", **or**" pada perintah **ologit**.

. ologit rep77 mpg length, or

```
Iteration 0: log likelihood = -89.895098
Iteration 1: log likelihood = -86.062239
Iteration 2: log likelihood = -86.015487
Iteration 3: log likelihood = -86.015382
Iteration 4: log likelihood = -86.015382

Ordered logistic regression        Number of obs =     66
                                   LR chi2(2)    =   7.76
                                   Prob > chi2   = 0.0207
Log likelihood = -86.015382        Pseudo R2     = 0.0432

------------------------------------------------------------------------------
      rep77 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
         mpg |   1.192499   .0805834     2.61   0.009     1.04457    1.361376
      length |   1.032752   .0182416     1.82   0.068      .997611    1.069131
------------+-----------------------------------------------------------------
       /cut1 |   6.656575   4.570512                   -2.301464    15.61461
       /cut2 |   8.455557   4.573731                   -.5087906     17.4199
       /cut3 |   10.41195   4.633466                     1.33052    19.49337
```

```
 /cut4 |   12.59688  4.732698              3.320966    21.8728
------------------------------------------------------------------------------
```

Contoh 9.3:

Responden diberikan pernyataan berikut: “Seorang ibu yang bekerja dapat membentuk hubungan dengan anaknya sama hangatnya dan sama teguhnya seperti seorang ibu yang tak bekerja”.

Pilihan respons yaitu: 1=Sangat tidak setuju (*Strongly Disagree*; SD), 2=Tak setuju (*Disagree*; D), 3=Setuju (*Agree*; A), dan 4=Sangat setuju (*Strongly agree*; SA).

. use ordwarm2, clear

```
(77 & 89 General Social Survey)
```

. describe warm yr89 male white age ed prst

```
variable  storage   display  value     variable
  name      type    format   label     label
------------------------------------------------------------------------------
warm        byte    %10.0g   sd2sa     Mom can have warm
                                          relations with child
yr89        byte    %10.0g   yrlbl     Survey year: 1=1989 0=1977
male        byte    %10.0g   sexlbl    Gender: 1=male 0=female
white       byte    %10.0g   race2lbl  Race: 1=white 0=not white
age         byte    %10.0g             Age in years
ed          byte    %10.0g             Years of education
prst        byte    %10.0g             Occupational prestige
```

. sum warm yr89 male white age ed prst

Variable	Obs	Mean	Std. Dev.	Min	Max
warm	2,293	2.607501	.9282156	1	4
yr89	2,293	.3986044	.4897178	0	1
male	2,293	.4648932	.4988748	0	1
white	2,293	.8765809	.3289894	0	1
age	2,293	44.93546	16.77903	18	89
ed	2,293	12.21805	3.160827	0	20
prst	2,293	39.58526	14.49226	12	82

. tab warm

Mom can have warm relations with child	Freq.	Percent	Cum.
1SD	297	12.95	12.95
2D	723	31.53	44.48
3A	856	37.33	81.81
4SA	417	18.19	100.00
Total	2,293	100.00	

. ologit warm male yr89 white age ed prst, nolog

```
Ordered logistic regression                     Number of obs =  2,293
                                                LR chi2(6)    = 301.72
                                                Prob > chi2   = 0.0000
Log likelihood = -2844.9123                     Pseudo R2     = 0.0504

------------------------------------------------------------------------------
        warm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |  -.7332997   .0784827    -9.34   0.000    -.887123   -.5794765
        yr89 |   .5239025   .0798989     6.56   0.000     .3673036    .6805014
       white |  -.3911595   .1183808    -3.30   0.001    -.6231816   -.1591373
         age |  -.0216655   .0024683    -8.78   0.000    -.0265032   -.0168278
          ed |   .0671728    .015975     4.20   0.000     .0358624    .0984831
        prst |   .0060727   .0032929     1.84   0.065    -.0003813    .0125267
-------------+----------------------------------------------------------------
       /cut1 |  -2.465362   .2389128                     -2.933622   -1.997102
       /cut2 |   -.630904   .2333156                     -1.088194   -.1736138
       /cut3 |   1.261854    .234018                      .8031871    1.720521
------------------------------------------------------------------------------
```

. lrtest, saving(0)

. test male

```
 ( 1)  [warm]male = 0

           chi2(  1) = 87.30
         Prob > chi2 = 0.0000
```

. test age white male

```
( 1)  [warm]age = 0
( 2)  [warm]white = 0
( 3)  [warm]male = 0
```

. ologit warm yr89 white age ed prst, nolog

```
Ordered logistic regression                     Number of obs =  2,293
                                                LR chi2(5)    = 212.98
                                                Prob > chi2   = 0.0000
Log likelihood =  -2889.278                     Pseudo R2     = 0.0355

------------------------------------------------------------------------------
        warm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        yr89 |   .5486813   .0796222     6.89   0.000     .3926246    .704738
       white |  -.4248955   .1184223    -3.59   0.000    -.6569991   -.192792
         age |  -.0197197   .0024473    -8.06   0.000    -.0245164   -.014923
          ed |   .0659969   .0159375     4.14   0.000       .03476   .0972338
        prst |   .0046902    .003286     1.43   0.153    -.0017502   .0111306
-------------+----------------------------------------------------------------
```

```
/cut1 | -2.066166   .2337297                -2.524268   -1.608064
/cut2 | -.2697128   .2290783                -.7186981    .1792725
/cut3 |  1.566067   .2312319                 1.112861    2.019273
------------------------------------------------------------------
```

. lrtest

```
Likelihood-ratio test                     LR chi2(1)  =  88.73
(Assumption: . nested in LRTEST_0)        Prob > chi2 = 0.0000
```

BAB 10
REGRESI LOGISTIK MULTINOMIAL

❖ Risiko dan Rasio Risiko

Dalam bab ini dibahas mengenai risiko dan rasio risiko, karena pada keluaran regresi logistik multinomial yang akan diperoleh adalah rasio risiko, bukan rasio odds. Istilah risiko (*risk*) sebagai ukuran kuantitatif awalnya digunakan dalam Epidemiologi. Untuk itu akan disajikan kembali tabel 2×2 prediktor biner dan respons biner (tabel 10.1).

Prediktor	**Respons**		
	$Y = 1$	$Y = 0$	
$X = 1$	a	b	$n_1 = a + b$
$X = 0$	c	d	$n_2 = c + d$

Dalam Epidemiologi, dengan prediktor berupa pajanan (*exposure*) dan respons adalah penyakit (*disease*), maka risiko adalah proporsi subjek yang terkena penyakit selama suatu periode tertentu di antara sejumlah subjek yang pada awal periode seluruhnya sehat. Maka risiko penyakit dengan syarat prediktor ada ($X = 1$) adalah:

$$\hat{R}_1 = \frac{a}{a+b} = \frac{a}{n_1} \qquad (10.1.a)$$

Risiko penyakit dengan syarat risiko tidak ada ($X = 0$) adalah:

$$\hat{R}_2 = \frac{c}{c+d} = \frac{c}{n_2} \qquad (10.1.b)$$

Rasio antara keduanya, yaitu rasio risiko adalah:

$$\hat{R}R = \frac{\hat{R}_1}{\hat{R}_2} = \frac{a/n_1}{c/n_2} \qquad (10.2)$$

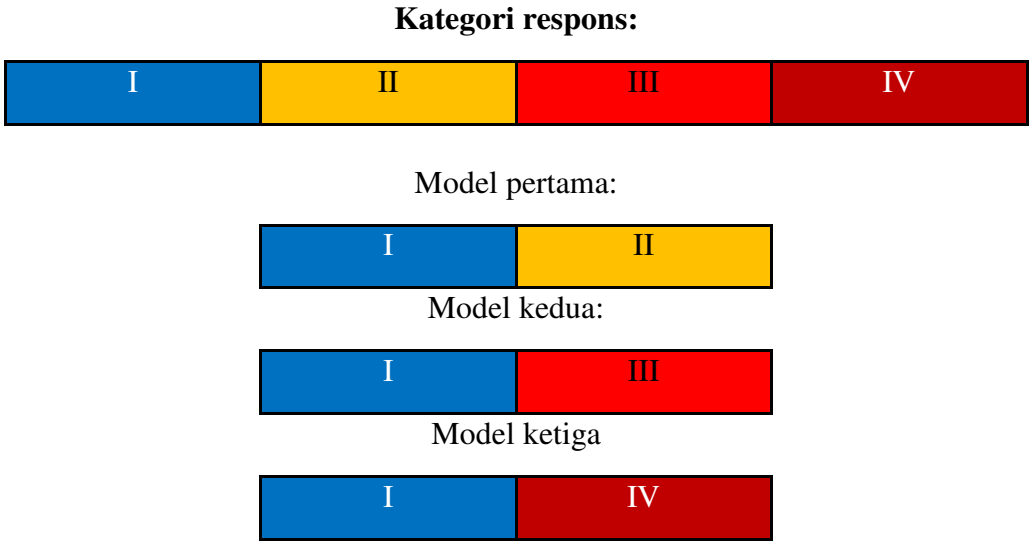
Dalam pengertian umum, risiko dapat diterjemahkan sebagai proporsi subjek yang mengalami suatu peristiwa dalam periode tertentu di antara sejumlah subjek yang pada awal periode belum pernah mengalami peristiwa tersebut. Sedangkan rasio risiko adalah rasio antara risiko dalam keadaan prediktor ada dengan risiko dalam keadaan prediktor tidak ada

❖ Pengertian Regresi Logistik Multinomial

Regresi logistik multinomial (regresi logistik politomi) adalah pemodelan regresi logistik untuk data prediktor-respons dengan respons kategorik nominal non-biner.

Misalkan dimiliki data prediktor-respons dengan respons berskala nominal non-biner yang memiliki *M* kategori, maka akan dipilih 1 kategori sebagai kategori dasar (*baseline*), dan tiap kategori lainnya masing-masing akan dibandingkan dengan kategori dasar ini, sehingga diperoleh (*M* – 1) model regresi logistik. Jika tidak dispesifikasikan, umumnya yang diambil untuk kategori dasar secara *default* adalah kategori dengan nilai respons terendah.

Jika respons memiliki 4 kategori, I, II, III, dan IV, maka secara *default* kategori I akan menjadi *baseline*, lalu dilakukan 3 kali pemodelan regresi logistik, II *vs* I, III *vs* I, dan IV *vs* I (gambar 10.1). Karena tiap pemodelan menggunakan himpunan nilai prediktor yang berbeda, akan diperoleh 3 model regresi logistik yang berbeda nilai-nilai estimasi koefisien regresinya maupun estimasi konstantenya.



Gambar 10.1 Kategori respons dan model regresi logistik multinomial

❖ Regresi Logistik Multinomial dengan Stata

Misalkan dimiliki data himpunan nilai prediktor-respons, respons adalah data kategorik multinomial non-biner dengan *M* kategori, maka pada *fitting* model regresi logistik multinomial akan diperoleh (*M* – 1) model regresi logistik estimasi.

Perintah Stata untuk *fitting* regresi logistik multinomial adalah:

mlogit *depvar indepvars* [*if*] [*in*] [, *options*]

depvar : Respons kategorik nominal

indepvars : Himpunan prediktor

Beberapa opsi:

basecategory(#)	:	Kategori dasar (*baseline*). *(#)* menyatakan nilai variabel dependen untuk kategori dasar. *Default*-nya adalah kategori dengan nilai terendah
rrr	:	Menampilkan nilai-nilai estimasi rasio risiko
nolog	:	Tak menampilkan nilai-nilai iterasi

Untuk memperoleh estimasi nilai-nilai rasio odds, digunakan opsi **or**.

Contoh 10.1:

File data **hsbdemo.dta** yang digunakan memuat data tentang 200 orang siswa sekolah lanjutan atas. Prediktornya adalah status sosial-ekonomi siswa (**ses**) dan kemampuan menulis siswa (**write**) dengan prediktor tipe program yang dipilih siswa di sekolah (**prog**).

```
. use "D:\Analisis Regresi Logistik\Data\hsbdemo", clear
(highschool and beyond (200 cases))

. tab prog ses, chi2
```

Program Type	Social Economic Status low	middle	high	Total
general	16	20	9	45
academic	19	44	42	105
vocation	12	31	7	50
Total	47	95	58	200

Pearson chi2(4) = 16.6044 Pr = 0.002

. table prog, con(mean write sd write)

Program Type	mean(write)	sd(write)
general	51.33333	9.397776
academic	56.25714	7.943343
vocation	46.76	9.318754

. mlogit prog i.ses write, base(2)

```
Iteration 0: log likelihood = -204.09667
Iteration 1: log likelihood = -180.80105
Iteration 2: log likelihood = -179.98724
Iteration 3: log likelihood = -179.98173
Iteration 4: log likelihood = -179.98173

Multinomial logistic regression       Number of obs =    200
                                      LR chi2(6)    =  48.23
                                      Prob > chi2   = 0.0000
Log likelihood = -179.98173           Pseudo R2     = 0.1182

------------------------------------------------------------------------------
        prog |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
general      |
         ses |
      middle |   -.533291   .4437321    -1.20   0.229    -1.40299      .336408
        high |  -1.162832   .5142195    -2.26   0.024    -2.170684    -.1549804
             |
       write |  -.0579284   .0214109    -2.71   0.007    -.0998931    -.0159637
       _cons |   2.852186   1.166439     2.45   0.014     .5660075     5.138365
-------------+----------------------------------------------------------------
academic     | (base outcome)
```

```
-----------+------------------------------------------------------------
vocation   |
      ses  |
   middle  |   .2913931   .4763737     0.61   0.541    -.6422822    1.225068
     high  |  -.9826703   .5955669    -1.65   0.099     -2.14996    .1846195
           |
     write |  -.1136026   .0222199    -5.11   0.000    -.1571528   -.0700524
     _cons |     5.2182   1.163549     4.48   0.000     2.937686    7.498714
-----------+------------------------------------------------------------
```

Dengan base(2) yaitu academic sebagai kategori dasar, diperoleh 2 model estimasi:

- general *vs* academic:

  logit **prog** = 2.85 – 0.53 **ses**(middle) – 1.16 **ses**(high) – 0.06 **write**

- vocation *vs* academic:

  logit **prog** = 5.22 + 0.29 **ses**(middle) – 0.98 **ses**(high) – 0.11 **write**

Tampak kedua model yang diperoleh berbeda, baik estimasi koefisien regresinya maupun estimasi konstantenya. Untuk general *vs* academic, prediktor yang bermakna adalah status sosial-ekonomi high (*vs* low) dan kemampuan menulis, sedangkan untuk vocation *vs* academic, prediktor yang bermakna hanya kemampuan menulis. Selanjutnya untuk memperoleh estimasi nilai-nilai rasio risiko, digunakan opsi "**rrr**".

. mlogit, rrr

```
Multinomial logistic regression       Number of obs =    200
                                      LR chi2(6)    =  48.23
                                      Prob > chi2   = 0.0000
Log likelihood = -179.98173           Pseudo R2     = 0.1182

------------------------------------------------------------------------------
        prog |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
general      |
         ses |
      middle |    .586671   .2603248    -1.20   0.229     .2458607    1.39991
        high |   .3125996   .1607448    -2.26   0.024     .1140996     .856432
             |
       write |   .9437175   .0202059    -2.71   0.007     .9049342     .984163
       _cons |   17.32562   20.20928     2.45   0.014     1.761221    170.4369
-------------+----------------------------------------------------------------
academic     |  (base outcome)
-------------+----------------------------------------------------------------
vocation     |
         ses |
      middle |   1.338291   .6375264     0.61   0.541     .5260904    3.404399
```

```
    high  | .3743103   .2229268  -1.65  0.099   .1164888   1.202761
          |
    write | .8926126   .0198338  -5.11  0.000   .8545734   .9323449
    _cons | 184.6016   214.793    4.48  0.000   18.87213   1805.719
------------------------------------------------------------------------------
Note: _cons estimates baseline relative risk for each outcome.
```

Perintah **test** berikut ini digunakan untuk menguji efek menyeluruh (*overall effect*) **2.ses** dan **3.ses** (**middle** *vs* **low** dan **high** *vs* **low**). Diperoleh hasil yang bermakna.

```
. test 2.ses 3.ses

 ( 1)  [general]2.ses = 0
 ( 2)  [academic]2o.ses = 0
 ( 3)  [vocation]2.ses = 0
 ( 4)  [general]3.ses = 0
 ( 5)  [academic]3o.ses = 0
 ( 6)  [vocation]3.ses = 0
       Constraint 2 dropped
       Constraint 5 dropped

           chi2(  4) = 10.82
         Prob > chi2 =  0.0287
```

Perintah berikut menguji kesamaan efek **3.ses** (**high** *vs* **low**) pada model regresi logistik **general** (*vs* **academic**) dan **vocation** (*vs* **academic**). Hasilnya tampak tak bermakna (tidak ada perbedaan antara kedua model).

. test [general]3.ses = [vocation]3.ses

```
 ( 1)  [general]3.ses - [vocation]3.ses = 0

           chi2(  1) = 0.08
         Prob > chi2 = 0.7811
```

Contoh 10.2:

File data berikut, **nomocc2.dta** memuat data tentang 337 individu. Anggota sampel dibedakan menurut 5 macam status pekerjaan (**occ**) yang akan menjadi variabel dependen yang berskala nominal, yaitu **Menial** (petugas rendahan), **BlueCol** (pekerja kerah biru, pekerja kasar di pabrik), **Craft** (pekerja keterampilan tangan), **WhiteCol** (pekerja kerah putih, pegawai kantor), dan **Prof** (tenaga profesional). Prediktor adalah **ed** (jumlah tahun pendidikan), **exper** (jumlah tahun pengalaman kerja), dan **white** (ras, kulit putih dan bukan kulit putih.

. use "D:\Analisis Regresi Logistik\Data\nomocc2.dta", clear

```
(1982 General Social Survey)
```

. tab occ

Occupation	Freq.	Percent	Cum.
Menial	31	9.20	9.20
BlueCol	69	20.47	29.67
Craft	84	24.93	54.60
WhiteCol	41	12.17	66.77
Prof	112	33.23	100.00
Total	337	100.00	

. list occ ed exper white in 1/10

	occ	ed	exper	white
1.	Menial	11	3	1
2.	Menial	12	14	1
3.	Menial	12	44	1
4.	Menial	12	18	1
5.	Menial	14	24	0
6.	Menial	13	38	1
7.	Menial	14	8	0
8.	Menial	14	19	1
9.	Menial	12	8	1
10.	Menial	12	3	1

. tab white

```
     Race: |
   1=white |
0=nonwhite |   Freq.   Percent     Cum.
-----------+-----------------------------
         0 |      28      8.31     8.31
         1 |     309     91.69   100.00
-----------+-----------------------------
     Total |     337    100.00
```

. sum ed exper

```
Variable |   Obs        Mean    Std. Dev.    Min     Max
---------+--------------------------------------------------
      ed |   337    13.09496    2.946427      3      20
   exper |   337    20.50148    13.95936      2      66
```

. mlogit occ ed exper white, base(5) nolog

```
Multinomial logistic regression                 Number of obs =    337
                                                LR chi2(12)   = 166.09
                                                Prob > chi2   = 0.0000
Log likelihood = -426.80048                     Pseudo R2     = 0.1629
```

occ	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Menial						
ed	-.7788519	.1146293	-6.79	0.000	-1.003521	-.5541826
exper	-.0356509	.018037	-1.98	0.048	-.0710028	-.000299
white	-1.774306	.7550543	-2.35	0.019	-3.254186	-.2944273
_cons	11.51833	1.849356	6.23	0.000	7.893659	15.143
BlueCol						
ed	-.8782767	.1005446	-8.74	0.000	-1.07534	-.6812128
exper	-.0309296	.0144086	-2.15	0.032	-.05917	-.0026893
white	-.5378027	.7996033	-0.67	0.501	-2.104996	1.029391
_cons	12.25956	1.668144	7.35	0.000	8.990061	15.52907
Craft						
ed	-.6850365	.0892996	-7.67	0.000	-.8600605	-.5100126
exper	-.0079671	.0127055	-0.63	0.531	-.0328693	.0169351
white	-1.301963	.647416	-2.01	0.044	-2.570875	-.0330509
_cons	10.42698	1.517943	6.87	0.000	7.451864	13.40209
WhiteCol						
ed	-.4256943	.0922192	-4.62	0.000	-.6064407	-.2449479
exper	-.001055	.0143582	-0.07	0.941	-.0291967	.0270866
white	-.2029212	.8693072	-0.23	0.815	-1.906732	1.50089
_cons	5.279722	1.684006	3.14	0.002	1.979132	8.580313

```
---------+-------------------------------------------------------
Prof     | (base outcome)
---------+-------------------------------------------------------
```

Tampak bahwa jumlah tahun pendidikan (**ed**) selalu berpengaruh bermakna terhadap jenis pekerjaan (**occ**) individu, sedangkan prediktor lain adakalanya tidak terlalu jelas ataupun tak berpengaruh terhadap jenis pekerjaan. Selanjutnya akan dicoba menilai pengaruh lama pendidikan (**ed**) dan lama pengalaman kerja (**exper**) secara terpisah pada ras **white** dan **nonwhite**.

. mlogit occ ed exper if white==1, base(5) nolog

```
Multinomial logistic regression                 Number of obs =    309
                                                LR chi2(8)    = 154.60
                                                Prob > chi2   = 0.0000
Log likelihood = -388.21313                     Pseudo R2     = 0.1660

------------------------------------------------------------------------------
         occ |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Menial       |
          ed |  -.8307509   .1297238    -6.40   0.000    -1.085005    -.576497
       exper |  -.0338038   .0192045    -1.76   0.078     -.071444     .0038364
       _cons |   10.34842   1.779603     5.82   0.000      6.86046     13.83637
-------------+----------------------------------------------------------------
BlueCol      |
          ed |  -.9225517   .1085452    -8.50   0.000    -1.135296    -.7098071
       exper |   -.031449   .0150766    -2.09   0.037    -.0609987    -.0018994
```

```
       _cons |  12.27337  1.507683    8.14  0.000    9.318363   15.22837
-------------+----------------------------------------------------------------
Craft        |
          ed |  -.687611  .0952882   -7.22  0.000   -.8743724  -.5008497
       exper | -.0002589  .0131021   -0.02  0.984   -.0259385   .0254207
       _cons |   9.01797   1.36333    6.61  0.000    6.345893   11.69005
-------------+----------------------------------------------------------------
WhiteCol     |
          ed | -.4196398  .0956209   -4.39  0.000   -.6070532  -.2322263
       exper |  .0008478  .0147558    0.06  0.954   -.0280731   .0297687
       _cons |  4.972966  1.421145    3.50  0.000    2.187572   7.758359
-------------+----------------------------------------------------------------
Prof         | (base outcome)
------------------------------------------------------------------------------
```

. mlogit occ ed exper if white==0, base(5) nolog

```
Multinomial logistic regression                 Number of obs =        28
                                                LR chi2(8)    =     17.79
                                                Prob > chi2   =    0.0228
Log likelihood = -32.779416                     Pseudo R2     =    0.2135
```

occ	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Menial						
ed	-.7012628	.3331149	-2.11	0.035	-1.354156	-.0483695
exper	-.1108415	.0741489	-1.49	0.135	-.2561706	.0344876
_cons	12.32779	6.053749	2.04	0.042	.4626601	24.19292
BlueCol						
ed	-.560695	.3283296	-1.71	0.088	-1.204209	.0828191
exper	-.0261099	.0682348	-0.38	0.702	-.1598477	.1076279
_cons	8.063397	6.008364	1.34	0.180	-3.712779	19.83957
Craft						
ed	-.882502	.3359808	-2.63	0.009	-1.541012	-.2239917
exper	-.1597929	.0744173	-2.15	0.032	-.305648	-.0139378
_cons	16.21925	6.059759	2.68	0.007	4.342344	28.09616
WhiteCol						
ed	-.5311514	.3698153	-1.44	0.151	-1.255976	.1936734
exper	-.0520881	.0838967	-0.62	0.535	-.2165227	.1123465
_cons	7.821371	6.805372	1.15	0.250	-5.516914	21.15966
Prof	(base outcome)					

Hasil yang agak berbeda ditemukan di sini, pada ras **white** faktor lama pendidikan (**ed**) tetap selalu berpengaruh bermakna terhadap jenis pendidikan, tetapi pada ras **nonwhite** tak selalu demikian, bahkan untuk jenis pekerjaan **WhiteCol** tampak tak bermakna ($p = 0.15$; lama pendidikan tidak menentukan apakah yang bersangkutan akan menjadi pekerja kerah putih atau tenaga profesional). Kemungkinan yaitu ada sebagian individu kulit berwarna yang walaupun memiliki pendidikan profesional, lapangan kerja yang tersedia hanya memungkinkan mereka menjadi pekerja kerah putih.

DAFTAR PUSTAKA

Agresti A. 2002. **Categorical Data Analysis**, 2nd Ed. Hoboken, New Jersey: John Wiley & Sons.

Agresti A. 2007. **An Introduction to Categorical Analysis**, 2nd Ed. Hoboken, New Jersey: John Wiley & Sons.

Agresti A. 2010. **Analysis of Ordinal Categorical Data**, 2nd Ed. Hoboken, New Jersey: John Wiley & Sons.

Christensen R. 1997. **Log-Linear Models and Logistic Regression**, 2nd Ed. New York: Springer-Verlag.

Harlan J. 2017. **Pemodelan Persamaan Struktural: III. Model Regresi Struktural dan Generalized SEM**. Jakarta: Penerbit Gunadarma.

Hilbe JM. 2009. **Logistic Regression Models**. Boca Raton, FL: Chapman & Hall.

Hosmer DW & Lemeshow. 2000. **Applied Logistic Regression**, 2nd Ed. New York: Wiley.

Kleinbaum DG & Klein M. 2010. **Logistic Regression: A Self-Learning Text**, 3rd Ed. New York: Springer.

Long JS & Freese J. 2014. **Regression Models for Categorical Dependent Variables Using Stata**, 3rd Ed. College Station, Texas: Stata Press.

Pearce N & Greenland S. 2005. "Confounding and Interaction". In: W Ahrens & I Pigeot (eds), **Handbook of Epidemiology**, 2nd Ed. Bremen: Springer, pp 659-684.

Rabe-Hesketh S & Everitt B. 2004. **A Handbook of Statistical Analyses Using Stata**, 3rd Ed. Boca Raton, FL: Chapman & Hall.

StataCorp. 2017. **Stata Base Reference Manual: Release 15**. College Station, Texas: Stata Press.

Szklo M & Nieto FJ. 2007. **Epidemiology: Beyond the Basics**, 2nd Ed. Sudbury, MA: Jones and Bartlett Publishers.

Upton GJG. 2017. **Categorical Data Analysis by Examples**. Hoboken, New Jersey: John Wiley & Sons.