

MODUL 10

ANALISIS REGRESI LOGISTIK



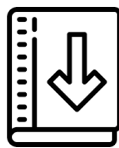
CAPAIAN PEMBELAJARAN

1. Praktikan mampu mengimplementasikan penggunaan software R untuk analisis regresi logistik



KEBUTUHAN ALAT/BAHAN/SOFTWARE

1. Komputer
2. Software R



DASAR TEORI

Dalam regresi linear, baik sederhana maupun berganda, variabel tak bebas bersifat metrik (interval atau rasio), sedangkan dalam regresi logistik, **variabel tak bebas bersifat nonmetrik** (memiliki kategori). Pada regresi linear, variabel bebas bersifat metrik (interval atau rasio), sedangkan dalam regresi logistik, **variabel bebas dapat bersifat metrik atau nonmetrik atau kombinasi dari keduanya**.

$$\hat{Y}_{\text{(binary nonmetrik)}} = \frac{X_1 + X_2 + X_3 + \dots + X_n}{\text{(non metrik dan metrik)}}$$

Pada regresi logistik, jika variabel tak bebas memiliki dua kategori, maka disebut regresi logistik biner (*binary regression logistic*). Namun, jika variabel tak bebas memiliki lebih dari dua kategori, maka disebut regresi logistik multinomial (*multinomial/polychotomous logistic regression*).

Secara umum, persamaan regresi logistik sederhana (melibatkan satu variabel bebas) memiliki bentuk sebagai berikut $\ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = \alpha + \beta x$.

Perhatikan bahwa $P(y=1)$ menyatakan probabilitas terjadinya kejadian sukses (*success*), sedangkan $1-P(y=1)$ menyatakan probabilitas terjadinya kejadian gagal (*failure*). Rasio dari $\frac{P(y=1)}{1-P(y=1)}$ disebut dengan odds. Sebagai contoh

misalkan $P(y=1)=0.8$ maka $\frac{P(y=1)}{1-P(y=1)} = \frac{0.8}{1-0.8} = 4$. Nilai 4 tersebut dapat

diartikan kejadian untuk terjadinya sukses 4 kali lebih mungkin dibandingkan untuk terjadinya gagal.

Persamaan regresi logistik sederhana untuk probabilitas terjadinya sukses memiliki bentuk sebagai berikut.

$$P(y=1) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

Persamaan regresi logistik untuk probabilitas dapat digunakan untuk mengestimasi probabilitas atau kemungkinan terjadinya suatu variabel tak bebas.

Persamaan regresi logistic biner berganda memiliki bentuk umum

$\ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$, dan persamaan regresi logistik

biner berganda untuk probabilitas terjadinya sukses memiliki

bentuk umum $P(y=1) = \frac{e^{\alpha+\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}}{1+e^{\alpha+\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}}$.



PRAKTIK

Praktik 1 (Input dan Import Data)

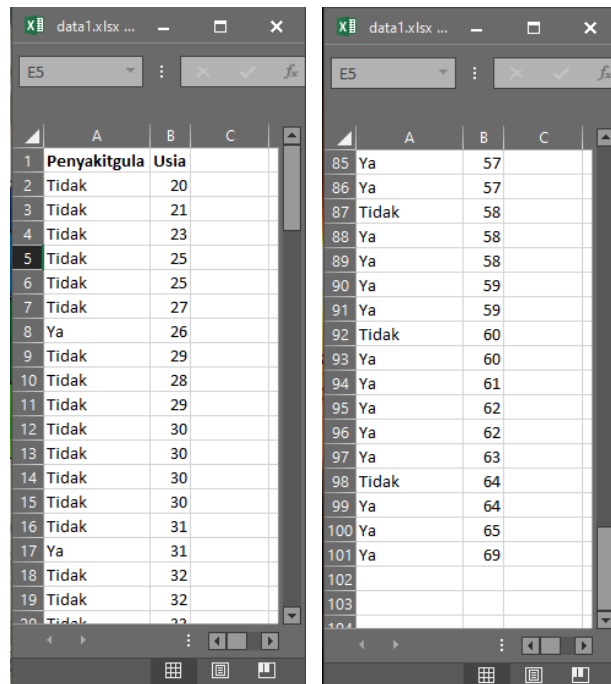
Andaikan diberikan data dari 100 responden mengenai usia, serta ada tidaknya penyakit gula pada Tabel 1 berikut:

Tabel 1

Responden	Penyakit Gula	Usia
1	Tidak	20
2	Tidak	21
3	Tidak	23
4	Tidak	25
5	Tidak	25
6	Tidak	27
7	Ya	26
8	Tidak	29
9	Tidak	28
10	Tidak	29
11	Tidak	30
12	Tidak	30
13	Tidak	30
14	Tidak	30
15	Tidak	31
16	Ya	31
17	Tidak	32
18	Tidak	32
19	Tidak	33
20	Tidak	34
21	Tidak	34
22	Tidak	34
23	Ya	34
24	Tidak	34
25	Tidak	34
26	Tidak	35
27	Tidak	35
28	Tidak	36
29	Ya	36
30	Tidak	36
31	Tidak	37
32	Ya	37
33	Tidak	37
34	Tidak	38
35	Tidak	38
36	Tidak	39
37	Ya	39
38	Tidak	40
39	Ya	40
40	Tidak	41
41	Tidak	41
42	Tidak	41
43	Tidak	42
44	Tidak	42
45	Ya	43
46	Tidak	43
47	Tidak	43
48	Ya	43
49	Tidak	44
50	Tidak	44

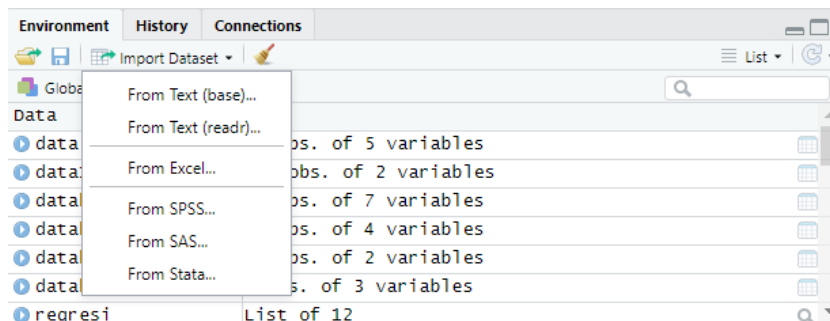
Responden	Penyakit Gula	Usia
51	Ya	44
52	Ya	44
53	Tidak	45
54	Ya	45
55	Tidak	46
56	Ya	46
57	Tidak	47
58	Tidak	47
59	Ya	47
60	Tidak	48
61	Ya	48
62	Ya	48
63	Tidak	49
64	Tidak	49
65	Ya	49
66	Tidak	50
67	Ya	50
68	Tidak	51
69	Tidak	52
70	Ya	52
71	Ya	53
72	Ya	53
73	Ya	54
74	Tidak	55
75	Ya	55
76	Ya	55
77	Ya	56
78	Ya	56
79	Ya	56
80	Tidak	57
81	Tidak	57
82	Ya	57
83	Ya	57
84	Ya	57
85	Ya	57
86	Tidak	58
87	Ya	58
88	Ya	58
89	Ya	59
90	Ya	59
91	Tidak	60
92	Ya	60
93	Ya	61
94	Ya	62
95	Ya	62
96	Ya	63
97	Tidak	64
98	Ya	64
99	Ya	65
100	Ya	69

Inputkan data Tabel 1 pada Ms. Excel, kemudian simpan dengan nama **data1.xlsx** (Gambar 1).



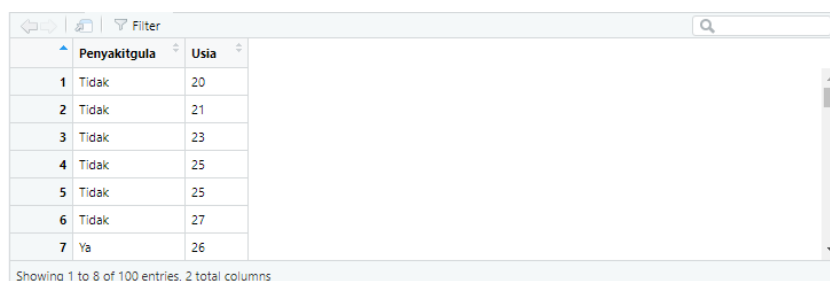
Gambar 1

Selanjutnya, pada Gambar 2, lakukan import file **data1.xlsx** pada R dengan cara lihat jendela **Environment**, pilih **Import Dataset**, pilih **From Excel**.



Gambar 2

Hasil dari langkah ini terlihat pada Gambar 3.

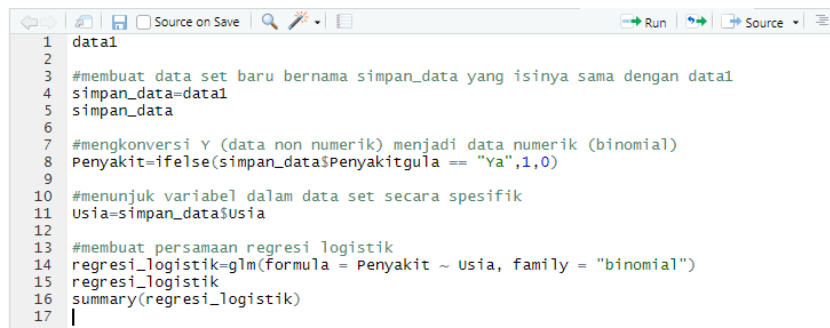


Gambar 3

Praktik 2 (Mengestimasi Persamaan Regresi Logistik)

Berdasarkan data pada Tabel 1 diketahui variabel tak bebas (*dependen*) **penyakit gula** bersifat non-metrik, yakni berupa kategori. Kategori “Ya” diberi kode angka 1, sementara kategori “Tidak” diberi kode angka 0. Pada variabel bebas (*independen*) **usia** bersifat metric. **Salah satu syarat penggunaan metode regresi logistik ialah data pada variabel tak bebas bersifat non-metrik (kategori).**

Input

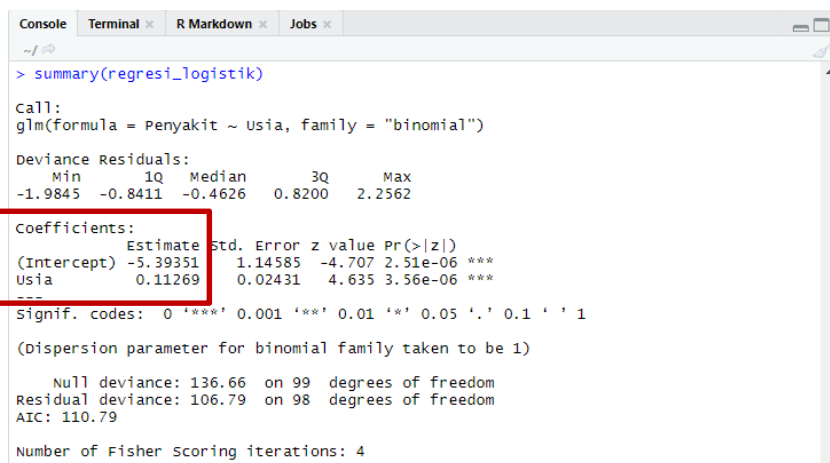


```
1 data1
2
3 #membuat data set baru bernama simpan_data yang isinya sama dengan data1
4 simpan_data=data1
5 simpan_data
6
7 #mengkonversi Y (data non numerik) menjadi data numerik (binomial)
8 Penyakit=ifelse(simpan_data$Penyakitgula == "Ya",1,0)
9
10 #menunjuk variabel dalam data set secara spesifik
11 Usia=simpan_data$Usia
12
13 #membuat persamaan regresi logistik
14 regresi_logistik=glm(formula = Penyakit ~ Usia, family = "binomial")
15 regresi_logistik
16 summary(regresi_logistik)
17 |
```

Gambar 4

Ketika menginputkan, harus dilakukan konversi pada variabel tak bebas (*dependen*) **penyakit gula** yang bersifat non-metrik menjadi metric (lihat Gambar 4, baris 8). Untuk mencari persamaan regresi logistic dapat menggunakan fungsi `glm` (lihat Gambar 4, baris 14).

Output



```
Console Terminal R Markdown Jobs
~/
> summary(regresi_logistik)

Call:
glm(formula = Penyakit ~ Usia, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9845  -0.8411  -0.4626   0.8200   2.2562

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.39351    1.14585  -4.707 2.51e-06 ***
Usia         0.11269    0.02431   4.635 3.56e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 106.79  on 98  degrees of freedom
AIC: 110.79

Number of Fisher scoring iterations: 4
```

Gambar 5

Analisis

Dari Gambar 5, diperoleh persamaan regresi logistic untuk memprediksi

probabilitas terjadinya penyakit gula, yaitu: $\hat{P}(y=1) = \frac{e^{-5.3935+0.11269(\text{Usia})}}{1+e^{-5.3935+0.11269(\text{Usia})}}$.

Praktik 3 (Memprediksi Nilai Peluang Responden)

Persamaan regresi logistik $\hat{P}(y=1) = \frac{e^{-5.3935+0.11269(\text{Usia})}}{1+e^{-5.3935+0.11269(\text{Usia})}}$, dapat digunakan untuk memprediksi atau mengestimasi peluang terjadinya penyakit gula, berdasarkan usia responden, yaitu dengan cara mensubstitusikan nilai setiap usia ke dalam persamaan tersebut. Contoh:

$$\text{Usia} = 20 \rightarrow \hat{P}(y=1) = \frac{e^{-5.3935+0.11269(20)}}{1+e^{-5.3935+0.11269(20)}} = 0.41498653$$

$$\text{Usia} = 60 \rightarrow \hat{P}(y=1) = \frac{e^{-5.3935+0.11269(60)}}{1+e^{-5.3935+0.11269(60)}} = 0.797039037$$

Selanjutnya, misalkan akan dicari prediksi peluang seseorang tidak terkena penyakit gula adalah:

$$\text{Usia} = 60 \rightarrow 1 - \hat{P}(y=1) = 1 - \frac{e^{-5.3935+0.11269(60)}}{1+e^{-5.3935+0.11269(60)}} = 1 - 0.797039037 = 0.202960963$$

Perhatikan bahwa

$$\text{Usia} = 60 \rightarrow \frac{\hat{P}(y=1)}{1 - \hat{P}(y=1)} = \frac{0.797039037}{0.202960963} = 3.92 \cong 4$$

Nilai tersebut dapat diartikan, ketika seseorang berusia 60 tahun, diprediksi terjadinya penyakit gula 4 kali lebih mungkin, dibandingkan tidak terkena penyakit gula.

Input dan Output

```

18 #Mengestimasi atau Memprediksi Nilai Peluang atau Probabilitas Responden
19 probabilitas=predict(regresi_logistik,type = "response")
20 probabilitas
21
21:1 (Top Level)
R Script

```

Console

```

1      2      3      4      5      6      7
0.04150107 0.04622284 0.05723942 0.07068705 0.07068705 0.08700246 0.07845763
8      9     10     11     12     13     14
0.10665149 0.09638058 0.10665149 0.11787416 0.11787416 0.11787416 0.11787416
15     16     17     18     19     20     21
0.13010577 0.13010577 0.14340030 0.14340030 0.15780686 0.17336784 0.17336784
22     23     24     25     26     27     28
0.17336784 0.17336784 0.17336784 0.17336784 0.19011686 0.19011686 0.20807669
29     30     31     32     33     34     35
0.20807669 0.20807669 0.22725707 0.22725707 0.22725707 0.24765257 0.24765257
36     37     38     39     40     41     42
0.26924073 0.26924073 0.29198043 0.29198043 0.31581068 0.31581068 0.31581068
43     44     45     46     47     48     49
0.34065009 0.34065009 0.36639696 0.36639696 0.36639696 0.36639696 0.39293012
50     51     52     53     54     55     56
0.39293012 0.39293012 0.39293012 0.42011072 0.42011072 0.44778465 0.44778465
57     58     59     60     61     62     63
0.47578584 0.47578584 0.47578584 0.50394011 0.50394011 0.50394011 0.53206941
64     65     66     67     68     69     70
0.53206941 0.53206941 0.55999634 0.55999634 0.58754856 0.61456298 0.61456298
71     72     73     74     75     76     77
0.64088948 0.64088948 0.66639391 0.69096044 0.69096044 0.69096044 0.71449296
78     79     80     81     82     83     84
0.71449296 0.71449296 0.73691584 0.73691584 0.73691584 0.73691584 0.73691584
85     86     87     88     89     90     91
0.73691584 0.75817372 0.75817372 0.75817372 0.77823083 0.77823083 0.79706966
92     93     94     95     96     97     98
0.79706966 0.81468924 0.83110315 0.83110315 0.84633742 0.86042834 0.86042834
99     100
0.87342038 0.91546901
>

```

Gambar 6

Untuk memprediksi nilai peluang atau probabilitas responden dapat menggunakan fungsi `predict` (lihat Gambar 6, baris 19).

Analisis

Tabel 2

Responden	Usia	Penyakit Gula	Probabilitas
1	20	Tidak	0.04150107
2	21	Tidak	0.04622284
...
7	26	Ya	0.07845763
...
100	69	Ya	0.91546901

Praktik 4 (Memprediksi Keanggotaan Responden dalam Kelompok)

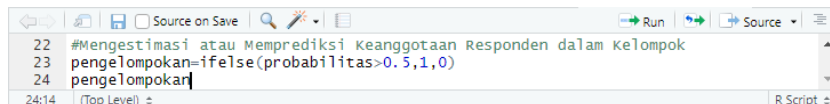
Pada pembahasan sebelumnya, telah dihitung nilai prediksi probabilitas terjadinya penyakit gula untuk tiap-tiap responden.

Input

Berdasarkan nilai prediksi peluang, dapat diprediksi apakah responden tersebut masuk ke dalam kelompok terkena penyakit gula “Ya” atau tidak terkena penyakit

gula “Tidak”. Pemberian kategori pada variabel Probabilitas menggunakan perintah `ifelse` (lihat Gambar 7, baris 24), dengan kriteria:

- Apabila nilai prediksi peluang responden $> 0,5$, maka responden tersebut diprediksi masuk ke dalam kelompok terkena penyakit gula berlabel **1**.
- Apabila nilai prediksi peluang responden $< 0,5$, maka responden tersebut diprediksi masuk ke dalam kelompok tidak terkena penyakit gula berlabel **0**.



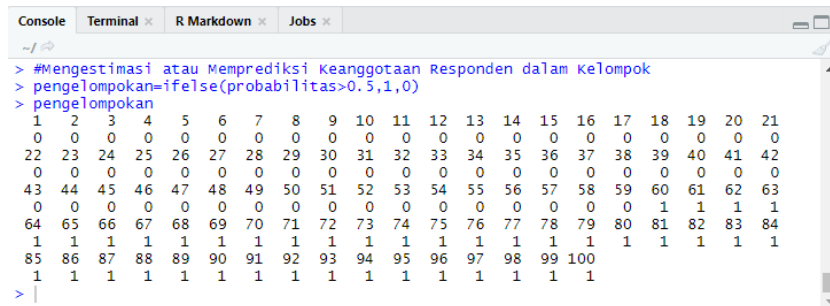
```

22 #Mengestimasi atau Memprediksi Keanggotaan Responden dalam Kelompok
23 pengelompokan=ifelse(probabilitas>0.5,1,0)
24 pengelompokan

```

Gambar 7

Output



```

> #Mengestimasi atau Memprediksi Keanggotaan Responden dalam Kelompok
> pengelompokan=ifelse(probabilitas>0.5,1,0)
> pengelompokan
  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  1  1
64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84
1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
> |

```

Gambar 8

Analisis

Tabel 3

Resp.	Usia	Penyakit Gula	Probabilitas	Pengelompokan
1	20	Tidak	0.04150107	0
2	21	Tidak	0.04622284	0
...
7	26	Ya	0.07845763	1
...
100	69	Ya	0.91546901	1

Berdasarkan Tabel 3, dapat dianalisis bahwa:

- Diketahui prediksi peluang responden ke-1 terkena penyakit gula sebesar 0,04150107, yakni $< 0,5$, maka responden ke-1 diprediksi masuk ke dalam kelompok tidak terkena penyakit gula **0**. Diketahui **pada keadaan sebenarnya, responden ke-1 memang tidak terkena penyakit gula (tidak terjadi kesalahan klasifikasi)**.
- Diketahui prediksi peluang responden ke-2 terkena penyakit gula sebesar 0,04622284, yakni $< 0,5$, maka responden ke-2 diprediksi masuk ke dalam

kelompok tidak terkena penyakit gula **0**. Diketahui **pada keadaan sebenarnya, responden ke-2 memang tidak terkena penyakit gula (tidak terjadi kesalahan klasifikasi)**.

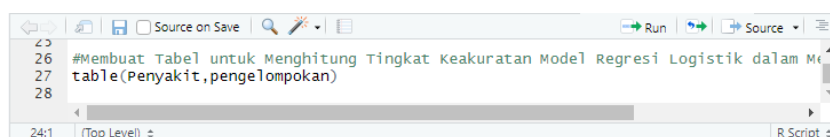
- Diketahui prediksi peluang responden ke-7 terkena penyakit gula sebesar 0,07845763, yakni $< 0,5$, maka responden ke-7 diprediksi masuk ke dalam kelompok tidak terkena penyakit gula **0**. Diketahui **pada keadaan sebenarnya, responden ke-7 terkena penyakit gula (terjadi kesalahan klasifikasi)**.

Praktik 5 (Menghitung Keakuratan Model Regresi Logistik dalam Memprediksi Pengelompokan)

Pada pembahasan sebelumnya, berdasarkan nilai prediksi peluang dari responden, dapat diprediksi responden tersebut masuk ke dalam kelompok tidak terkena penyakit gula **0** atau terkena penyakit gula **1**. Dalam proses pengelompokkan tersebut, bisa saja terjadi kesalahan pengelompokkan. Sebagai contoh, responden ke-7 diprediksi masuk ke dalam kelompok tidak terkena penyakit gula **0**. Diketahui pada keadaan sebenarnya, responden ke-7 terkena penyakit gula (terjadi kesalahan klasifikasi).

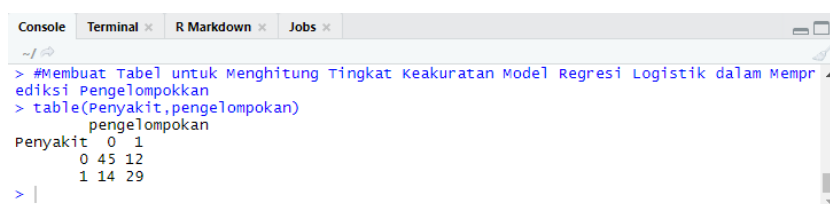
Input

Untuk melihat keakuratan model regresi dalam memprediksi pengelompokan dengan kondisi sebenarnya, dapat dibuat tabel silang dari variabel Pengelompokan dan Penyakit. Tabel tersebut dapat dibuat dengan menggunakan perintah `table` (lihat Gambar 9, baris 27).



Gambar 9

Output



Gambar 10

Analisis

Dari Gambar 10, jumlah keakuratan model regresi dengan kondisi sebenarnya dapat diperoleh jika Penyakit dan Pengelompokan sama (Penyakit = 0 dan Pengelompokan = 0, atau Penyakit = 1 dan Pengelompokan = 1). Jika Penyakit berbeda dengan Pengelompokan artinya model regresi tidak akurat, seperti pada responden ke-7. Jadi, tingkat akurasi model regresi terhadap kondisi sebenarnya

$$\text{adalah } \frac{\sum \text{responden akurat}}{\sum \text{responden}} = \frac{45 + 29}{45 + 12 + 14 + 29} = \frac{74}{100} = 0.74 \text{ atau } 74\%.$$

Praktik 6 (Grafik Usia v/s Nilai Prediksi Probabilitas)

Berikut disajikan grafik antara usia (sumbu horizontal) dan nilai prediksi probabilitas (sumbu vertikal) (nonlinear).

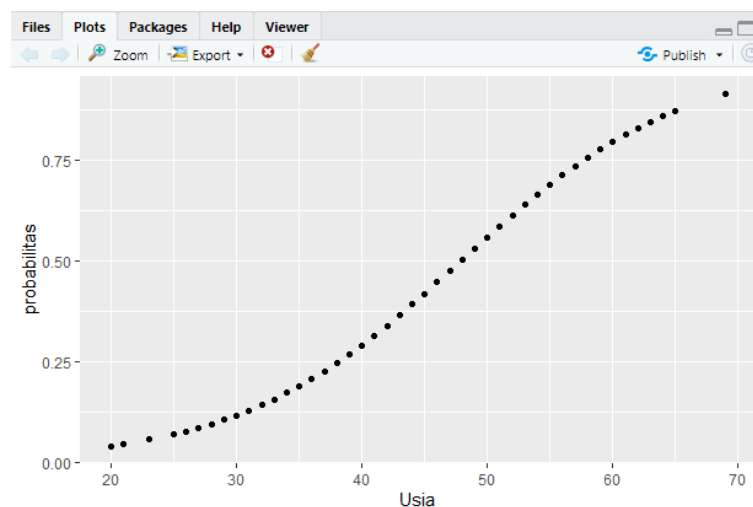
Input

```
28  
29 #Grafik Usia v/s Nilai Prediksi Probabilitas  
30 library(ggplot2)  
31 ggplot(simpan_data,aes(Usia, probabilitas)) + geom_point()  
32
```

Gambar 11

Untuk membuat grafik tersebut dapat menggunakan perintah `ggplot` (lihat Gambar 11, baris 31), dengan sebelumnya mengaktifkan library `ggplot2`.

Output



Gambar 12

Analisis

Dari Gambar 12, terlihat bahwa semakin tinggi usia maka semakin tinggi probabilitas terkena penyakit gula, walaupun grafik berbentuk nonlinear.



LATIHAN

Diketahui data status (Y) dan kadar hemoglobin (X) dari sejumlah pasien di suatu rumah sakit.

Resp.	Status	Kadar Hemoglobin
1	Tidak Sembuh Total	15.2
2	Tidak Sembuh Total	17.5
3	Tidak Sembuh Total	10.7
4	Tidak Sembuh Total	16.9
5	Sembuh Total	15.4
6	Tidak Sembuh Total	17.3
7	Tidak Sembuh Total	11.7
8	Sembuh Total	14.6
9	Sembuh Total	15.3
10	Tidak Sembuh Total	9.3
11	Tidak Sembuh Total	13.7
12	Sembuh Total	16.8
13	Tidak Sembuh Total	9.7
14	Tidak Sembuh Total	14.3
15	Sembuh Total	18.9

1. Inputkan data di atas!
2. Estimasi persamaan regresi logistiknya!
3. Prediksi nilai peluang tiap responden!
4. Prediksi keanggotaan responden dalam kelompok!
5. Hitung keakuratan model regresinya!
6. Buatlah grafiknya!



TUGAS

Diketahui data status (Y) dan kadar hemoglobin (X) dari sejumlah pasien di suatu rumah sakit.

Resp.	Status	Jumlah pelatihan yang diikuti
1	Diterima	6
2	Diterima	8
3	Diterima	4
4	Diterima	6
5	Diterima	6
6	Tidak diterima	4
7	Tidak diterima	4
8	Tidak diterima	4
9	Tidak diterima	4
10	Tidak diterima	6

1. Inputkan data di atas!
2. Estimasi persamaan regresi logistiknya!
3. Prediksi nilai peluang tiap responden!
4. Prediksi keanggotaan responden dalam kelompok!
5. Hitung keakuratan model regresinya!
6. Buatlah grafiknya!



REFERENSI

- [1] Gio, P.U., Effendie, A.R. 2017. Belajar Bahasa Pemrograman R (Dilengkapi Cara Membuat Aplikasi Olah Data Sederhana dengan R Shiny). Medan: USU Press.