MODEL PILIHAN KUALITATIF

Oleh
Bambang Juanda
https://bambangjuanda.com/

Seringkali dalam suatu survei kita berhadapan dengan peubah kualitatif yang mempunyai skala pengukuran **nominal** atau **ordinal**. Nilai-nilai peubah respons kualitatif ini terbatas (*limited dependent variable*), bahkan sering hanya bernilai dua kemungkinan saja. Misalnya, **apakah seseorang membeli mobil atau tidak**; memilih atau tidak dalam Pilkada (pemilihan kepala daerah); punya penyakit jantung koroner atau tidak; dan masih banyak contoh lainnya. Peubah kualitatif yang hanya mempunyai dua kemungkinan nilai ini disebut **peubah biner**.

Meskipun logis kita memperkirakan suatu hubungan langsung antara pendapatan dan perilaku pembelian, namun kita tidak dapat yakin apakah masing-masing konsumen dengan pendapatan tertentu pasti akan membeli produk. Oleh karena itu, **tujuan model pilihan kualitatif** adalah **menentukan peluang** bahwa individu dengan karakteristik-karakteristik tertentu akan memilih suatu pilihan tertentu dari beberapa alternatif yang tersedia. Jika pilihannya hanya ada dua alternatif disebut *model pilihan biner*.

Overview

Response		Analysis	
Continuous	→	Linear Regression Analysis	
Categorical	→	Logistic Regression Analysis	-Model Peluang Linear -Model Probit

Ilustrasi

Studi mengenai pengaruh tingkat pendapatan, jenis kelamin dan umur terhadap membeli tidaknya seseorang pada suatu produk yang dijual dengan harga tertentu.

Peubah Penjelas (bebas): umur, jenis kelamin dan tingkat pendapatan

Peubah Respons(Y): membeli (=1) atau tidak (=0)

Ilustrasi utk 1 Peubah Bebas

Studi mengenai pengaruh tk pendapatan atau jenis kelamin (X) terhadap membeli tidaknya seseorang (Y) pada suatu produk yang dijual dengan harga tertentu.

Peubah Penjelas (bebas):

Tk Pendapatan: X = Rp …… juta

atau Jenis Kelamin: $X = \begin{cases} 1, \text{ jika Wanita} \\ 0, \text{ jika Pria} \end{cases}$

Peubah Respons: $Y = \begin{cases} 1, \text{ jika membeli} \\ 0, \text{ jika tidak membeli} \end{cases}$

1. Model Peluang Linear

$Y_i = \alpha + \beta X_i + \varepsilon_i$ (10.1)

Dimana X_i = nilai karakteristik (misalnya pendapatan) individu ke-i,

Y_i = 1 , jika pilihan kesatu dipilih (misalnya membeli mobil)

0 , jika pilihan kedua dipilih (tidak membeli mobil).

ε_i = peubah acak yang menyebar bebas dengan nilai tengah 0.

Untuk menginterpretasikan persamaan (10.1) kita tentukan nilai harapan dari masing-masing pengamatan peubah respons Y_i :

$E(Y_i) = \alpha + \beta X_i$ (10.2)

Karena Y_i hanya mempunyai kemungkinan dua macam nilai (1 dan 0), kita dapat menggambarkan sebaran peluang Y dengan memisalkan:

$P_i = P(Y_i=1)$ dan $1-P_i = P(Y_i=0)$,

sehingga $E(Y_i) = 1\,(P_i) + 0\,(1-P_i) = P_i$. (10.3)

model (10.1) → peluang bahwa individu konsumen ke-i dengan pendapatan tertentu (X_i) akan membeli mobil.

Slope garis mengukur pengaruh perubahan 1 unit pendapatan terhadap perubahan peluang membeli mobil

Dugaan Model Peluang Linear

$$P\mathbf{i} = \begin{cases} \alpha + \beta X\mathbf{i} & , \text{jika } 0<(\alpha+\beta X\mathbf{i})<1 \\ 1 & , \text{jika } (\alpha+\beta X\mathbf{i}) \geq 1 \\ 0 & , \text{jika } (\alpha+\beta X\mathbf{i}) \leq 0 \end{cases} \qquad (10.4)$$

Sebaran Peluang bagi ε_i

Y_i	ε_i	Peluang
1	$1 - \alpha - \beta X_i$	P_i
0	$- \alpha - \beta X_i$	$1 - P_i$

$E(\varepsilon_i) = (1 - \alpha - \beta X_i)\, P_i + (-\alpha - \beta X_i)\,(1 - P_i) = 0$

sehingga $P_i = \alpha + \beta X_i$

$(1 - P_i) = 1 - \alpha - \beta X_i$

Ragam komponen sisaan

$$E(\varepsilon_i^2) = (1-\alpha-\beta X_i)^2 P_i + (-\alpha-\beta X_i)^2 (1-P_i) = P_i(1-P_i)$$

$$Var(Y_i) = E(Y_i^2) - \left[E(Y_i)\right]^2 = E(Y_i)[1-E(Y_i)] = P_i(1-P_i) = \sigma_i^2 = E(\varepsilon_i^2)$$

Jadi, **peubah Y menyebar menurut sebaran (distribusi) peluang Bernouli**.

→ Masalah heteroskedastisitas

Kendala dalam model peluang linear →perlu transformasi model (linear) awal sedemikian rupa sehingga prediksi nilai Y berada dalam selang (0;1) untuk semua nilai peubah bebas X. Salah satu bentuk transformasi yang mempunyai karakteristik seperti ini adalah fungsi peluang kumulatif (*cumulative probability function*), F.[1] Sebaran peluangnya dapat direpresentasikan dalam bentuk:

$$P_{\mathbf{i}} = F(\alpha + \beta X_{\mathbf{i}}) = F(Z_{\mathbf{i}})$$

Sebenarnya banyak fungsi peluang kumulatif yang mungkin dapat digunakan, namun disini hanya dua macam yang dipertimbangkan, yaitu **fungsi peluang normal** dan **logistik kumulatif**.

[1] Fungsi peluang kumulatif adalah $F(x_{\mathbf{i}})$=Peluang $(X \leq x_{\mathbf{i}})$

Model Probit

$P_i = F(\alpha + \beta X_i) = F(Z_i)$

asumsikan ada suatu **indeks** Z_i yg bernilai kontinu secara teoritis, yg ditentukan oleh nilai peubah penjelas X shg dapat ditulis:

$Z_i = \alpha + \beta X_i$

asumsikan bahwa Z merupakan peubah acak yang menyebar normal sehingga peluang bahwa Z lebih kecil (atau sama dengan) Z_i dapat dihitung dari fungsi peluang normal kumulatif. Untuk fungsi peluang normal baku kumulatif dapat dituliskan dalam rumus:

$$P_i = F(Z_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_i} e^{-s^2/2} ds$$

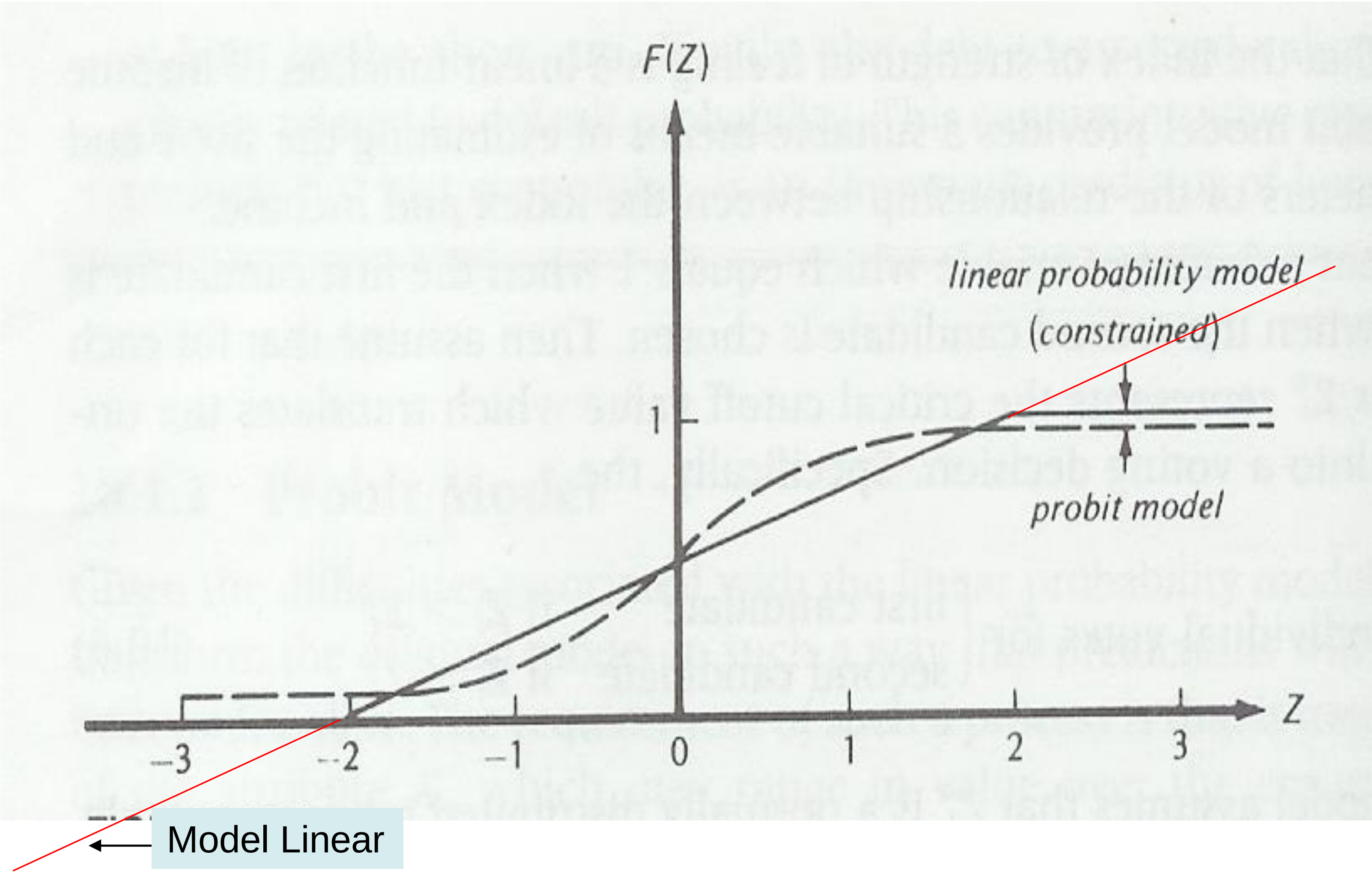
dimana s: peubah acak menyebar normal dgn nilai tengah 0 dan ragam1. Dgn rumus transformasi diatas, peubah P_i akan bernilai dlm selang (0;1). P_i menggambarkan peluang individu berkarakteristik (berpendapatan) X_i memilih pilihan-1 (beli mobil). Karena nilai peluang ini diukur berdasarkan luas daerah dibawah kurva normal baku dari -~ sampai Z_i, maka peluang pilihan-1 (beli mobil) makin tinggi jika nilai indeks Z_i makin tinggi. Untuk menduga indeks Z_i, kita menggunakan kebalikan (*inverse*) dari fungsi normal baku kumulatif (10.9) dengan:

$Z_i = F^{-1}(P_i) = \alpha + \beta X_i$

Hubungan Nilai Indeks Z dan Sebaran Peluang Normal Kumulatifnya

Z	F(Z)	Z	F(Z)
-3.0	.001	0.5	.691
-2.5	.006	1.0	.841
-2.0	.023	1.5	.933
-1.5	.067	2.0	.977
-1.0	.159	2.5	.994
-0.5	.309	3.0	.999
0.0	.500	3.5	.999

Model (Peluang) Linear vs Model Probit



Meskipun model probit lebih menarik dari model peluang linear, namun untuk menduga parameter koefisiennya menggunakan pendugaan kemungkinan maksimum (*maximum likelihood*, ML) non linear. Selain itu, justifikasi atau interpretasi koefisiennya agak terbatas. Oleh karena itu sebaiknya menggunakan model logit yang dibahas dalam subbab berikut

Model Regresi Logistik (Model logit)

menggunakan peubah penjelasnya (dpt peubah kategorik atau peubah numerik) untuk menduga peluang kejadian tertentu dari peubah respons kategori.

Model Logit Sederhana : $E(Y_i = 1 / X_i) = \dfrac{e^{(\beta_0 + \beta_1 X_i)}}{1 + e^{(\beta_0 + \beta_1 X_i)}}$

$$P(X_i) = P_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_i)}} = \frac{1}{1 + e^{-g(X)}}$$

Interpretasi: Peluang kejadian tertentu dari peubah respons kategori (misalnya membeli) jika pendapatannya X_i

Sebaran Logistik menyerupai kurva berbentuk S, sehingga interpretasinya logis. $0 \leq E(Y/X) \leq 1$

Transformasi Logit

Peluang kejadian tertentu dari peubah respons kategori (p_i), ditransformasi shg

$$\mathbf{logit}(p_i) = \mathbf{log}\left(\frac{p_i}{1 - p_i}\right) = g(x_i) = \beta_0 + \beta_1 X_i$$

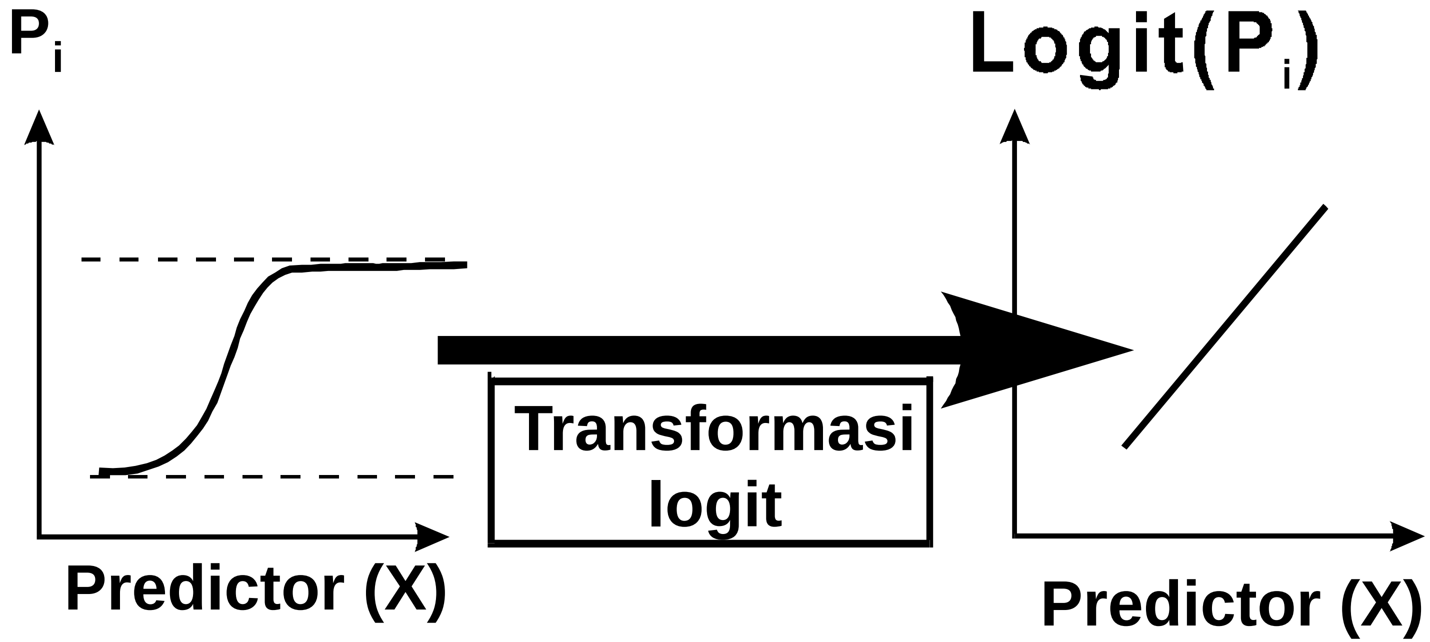
i indeks semua kasus (observasi 1,2,..,n).

p_i peluang kejadian (misalnya, membeli) terjadi untuk kasus ke-*i*.

log adalah natural log (bilangan dasar e).

Fungsi g(x) sudah Linear dalam Parameter,
dan $-\sim \leq g(x) \leq \sim$, shg dpt diduga dgn OLS

Assumption
(peubah X berskala Interval)



Interpretasi Koefisien Model Logit

Utk Peubah Bebas biner, mis Jenis Kelamin (X=1, X=0)

$$P(X_i) = \frac{e^{(\beta_0 + \beta_1 X_i)}}{1 + e^{(\beta_0 + \beta_1 X_i)}}$$

	X=1	X=0
Y=1	$P(1) = \frac{e^{(\beta_0+\beta_1)}}{1+e^{(\beta_0+\beta_1)}}$	$P(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
Y=0	$1-P(1) = \frac{1}{1+e^{(\beta_0+\beta_1)}}$	$1-P(0) = \frac{1}{1+e^{\beta_0}}$
Jumlah	1	1

P(1) : Peluang membeli produk utk konsumen Wanita

P(0) : Peluang membeli produk utk konsumen Pria

$$Odd_{wanita} = \frac{P(1)}{1-P(1)}$$

$$Odd_{pria} = \frac{P(0)}{1-P(0)}$$

$$OddsRatio = \frac{P(1)}{1-P(1)} \Big/ \frac{P(0)}{1-P(0)} = e^{\beta_1}$$

Interpretasi Koefisien

$$\log \frac{P(X_i)}{1-P(X_i)} = g(X_i) = \beta_0 + \beta_1 X_i$$

$\beta_1 = g(X+1) – g(X)$

utk X biner: $\beta_1 = g(1) – g(0)$

$$= \log \frac{P(1)}{1-P(1)} - \log \frac{P(0)}{1-P(01)} = \log \frac{P(1)/(1-P(1))}{P(0)/(1-P(0))}$$

Ukuran Asosiasi

Odds Ratio: $\psi = \frac{P(1)/[1-P(1)]}{P(0)/[1-P(0)]} = e^{\beta_1}$

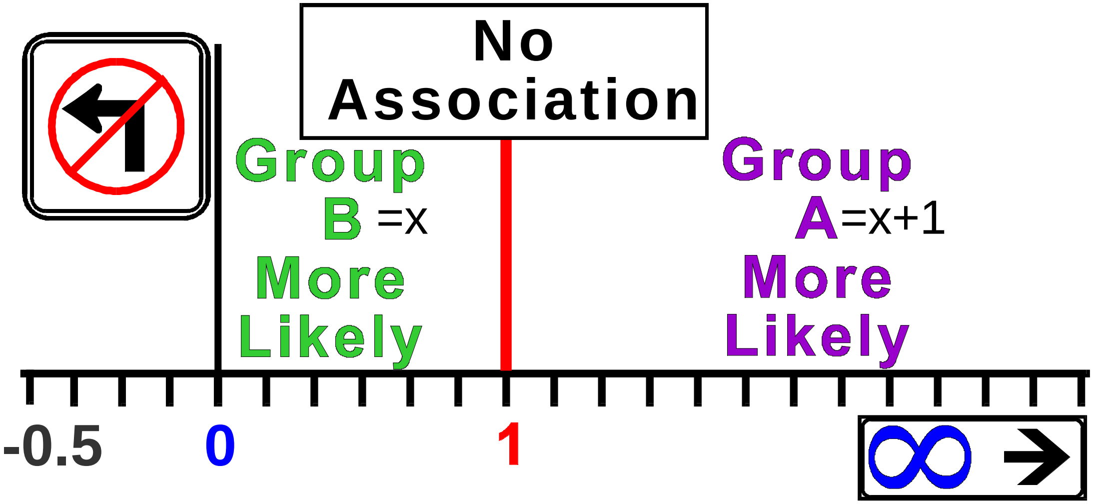
"Berapa kali Kemungkinan membeli utk konsumen Wanita dibandingkan Konsumen Pria"

Interpretasi Pendekatan Peluang Relatif P(1)/P(0) ini berlaku bila P(x) kecil

Utk X kontinu, exp(β1) *: Berapa kali Kemungkinan membelinya jika X naik 1 unit*

Properties of the Odds Ratio

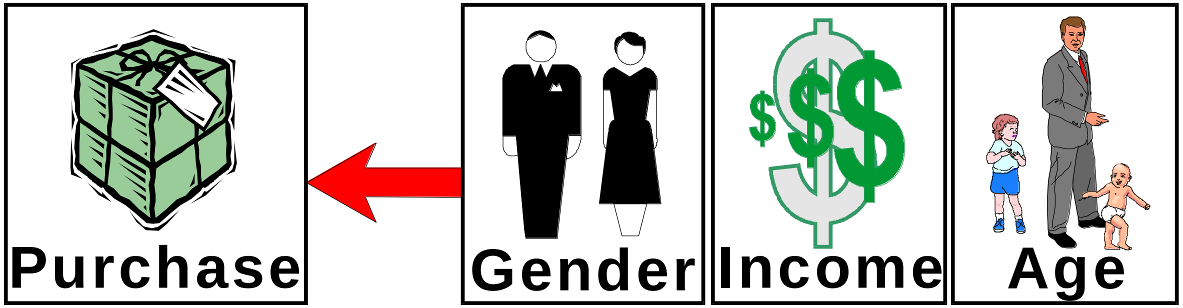
ODDS RATIO
OF GROUP A TO GROUP B



Note:

- SK (1-α) 100% bagi *Odds Ratio:* exp($c\ \hat{\beta} \pm z_{\alpha/2}\ c\ s_{\hat{\beta}}$)
- Dlm realitas ΔP(x) jika x berbeda 1 unit (1→2 dgn 10→11) dapat cukup berbeda. →Dilema utk peubah kontinu dimodelkan linear dlm model logit. Jika yakin bahwa logit tdk linear dgn *covariate* → *grouping* (*Dummy*)

Multiple Logistic Regression



$$\text{logit}\,(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Ilustrasi model utk mengkaji pengaruh jenis kelamin (X_1), umur (X_2), dan tingkat pendapatan (X_2) terhadap membeli tidaknya seseorang pada suatu produk yang dijual dengan harga tertentu.

$$P(X_i) = \frac{e^{(\beta_0 + \beta_1 X_i + \beta_2 X_{2i} + \beta_3 X_{3i})}}{1 + e^{(\beta_0 + \beta_1 X_i + \beta_2 X_{2i} + \beta_3 X_{3i})}}$$

logit (p_i) = $\log \frac{P(X_i)}{1 - P(X_i)} = g(\underline{X}_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$

Utk Peubah Bebas X kontinu, seringkali 1 unit terlalu kecil atau besar utk dipertimbangkan → Pendugaan utk perubahan "c" unit

$$g(x+c) - g(x) = c\,\beta_1$$

Odds Ratio-nya: $\psi(c) = \psi(x + c, x) = e^{c\beta_1}$

Pengujian Model dgn p Peubah Bebas

Uji Model secara keseluruhan:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_1: ada $\beta_j \neq 0$

→ *Likelihood Ratio Test Statistics* (G) ~ $\chi^2_{(p)}$

Uji **parsial** koefisien:

H_0: $\beta_j = 0$

H_1: $\beta_j \neq 0$

→ *WaldTest Statistics* (W) ~ Z

Statistics> Binary Outcome> Logistic Regression

```
. logit y x2_umur x3_jk d1_s d2_t
Logistic regression                               Number of obs   =        150
                                                  LR chi2(4)      =      11.24
                                                  Prob > chi2     =     0.0240
Log likelihood = -96.087961                       Pseudo R2       =     0.0553
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     x2_umur |  -.0251297   .0438814    -0.57   0.567    -.1111356    .0608763
       x3_jk |   .6852562   .3617553     1.89   0.058    -.0237711    1.394283
        d1_s |  -.0336209   .5029032    -0.07   0.947    -1.019293    .9520512
        d2_t |   .9570934   .4103268     2.33   0.020     .1528675    1.761319
       _cons |  -.2124465   1.727446    -0.12   0.902    -3.598178    3.173285
------------------------------------------------------------------------------
```

```
. logistic y x2_umur x3_jk d1_s d2_t
Logistic regression                               Number of obs   =        150
                                                  LR chi2(4)      =      11.24
                                                  Prob > chi2     =     0.0240
Log likelihood = -96.087961                       Pseudo R2       =     0.0553
------------------------------------------------------------------------------
           y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     x2_umur |   .9751834   .0427924    -0.57   0.567     .8948174    1.062767
       x3_jk |    1.98428   .7178238     1.89   0.058     .9765092    4.032084
        d1_s |    .966938   .4862762    -0.07   0.947     .3608499    2.591019
        d2_t |   2.604116   1.068539     2.33   0.020     1.165171     5.82011
       _cons |   .8086036   1.396819    -0.12   0.902     .0273735    23.88583
------------------------------------------------------------------------------
```

Statistics> Binary Outcomes> Postestimation> Classification Statistics after….

```
Logistic model for y

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        17            17  |         34
     -     |        45            71  |        116
-----------+--------------------------+-----------
   Total   |        62            88  |        150

Classified + if predicted Pr(D) >= .5
True D defined as y != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   27.42%
Specificity                     Pr( -|~D)   80.68%
Positive predictive value       Pr( D| +)   50.00%
Negative predictive value       Pr(~D| -)   61.21%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   19.32%
False - rate for true D         Pr( -| D)   72.58%
False + rate for classified +   Pr(~D| +)   50.00%
False - rate for classified -   Pr( D| -)   38.79%
--------------------------------------------------
Correctly classified                        58.67%
--------------------------------------------------
```

P(+/D) = 17/62 =27.42%
Untuk responden pembeli (D) maka Peluang diklasifikasikan membeli=27%

P(-/~D) =71/88 =80.68%
Untuk responden yg tidak beli (~D) maka Peluang diklasifikasikan tidak membeli=81%

Jika dugaan P(Y=1)≥0.5 → + (beli)

Jika Model ini mengklasifikasikan "+", Peluang benarnya = 17/34 = 50.00%.
Jika mengklasifikasikan "-", Peluang benarnya = 71/116 = 61.21%.

Model ini mengklasifikasikan dengan benar 58.67% =(17+71)/150

Categorical Variables Codings

		Frequency	Parameter coding (1)	(2)
INCOME	Low	132	1,000	,000
	Medium	144	,000	1,000
	High	155	,000	,000
GENDER	Female	240	1,000	
	Male	191	,000	

Classification Table[a]

	Observed		Predicted PURCHASE 0	1	Percentage Correct
Step 1	PURCHASE	0	236	33	87,7
		1	131	31	19,1
	Overall Percentage				61,9

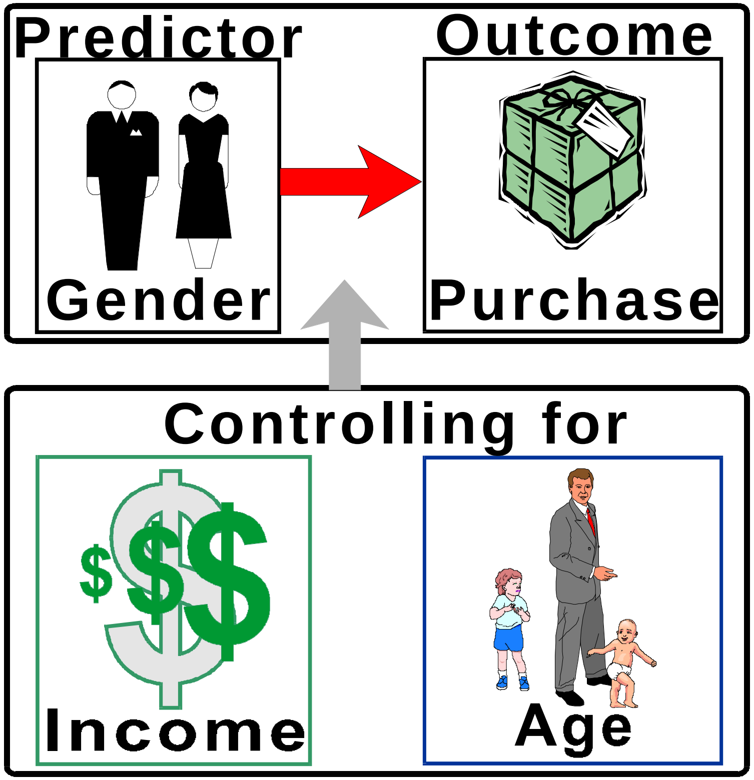
a. The cut value is ,500

Variables in the Equation

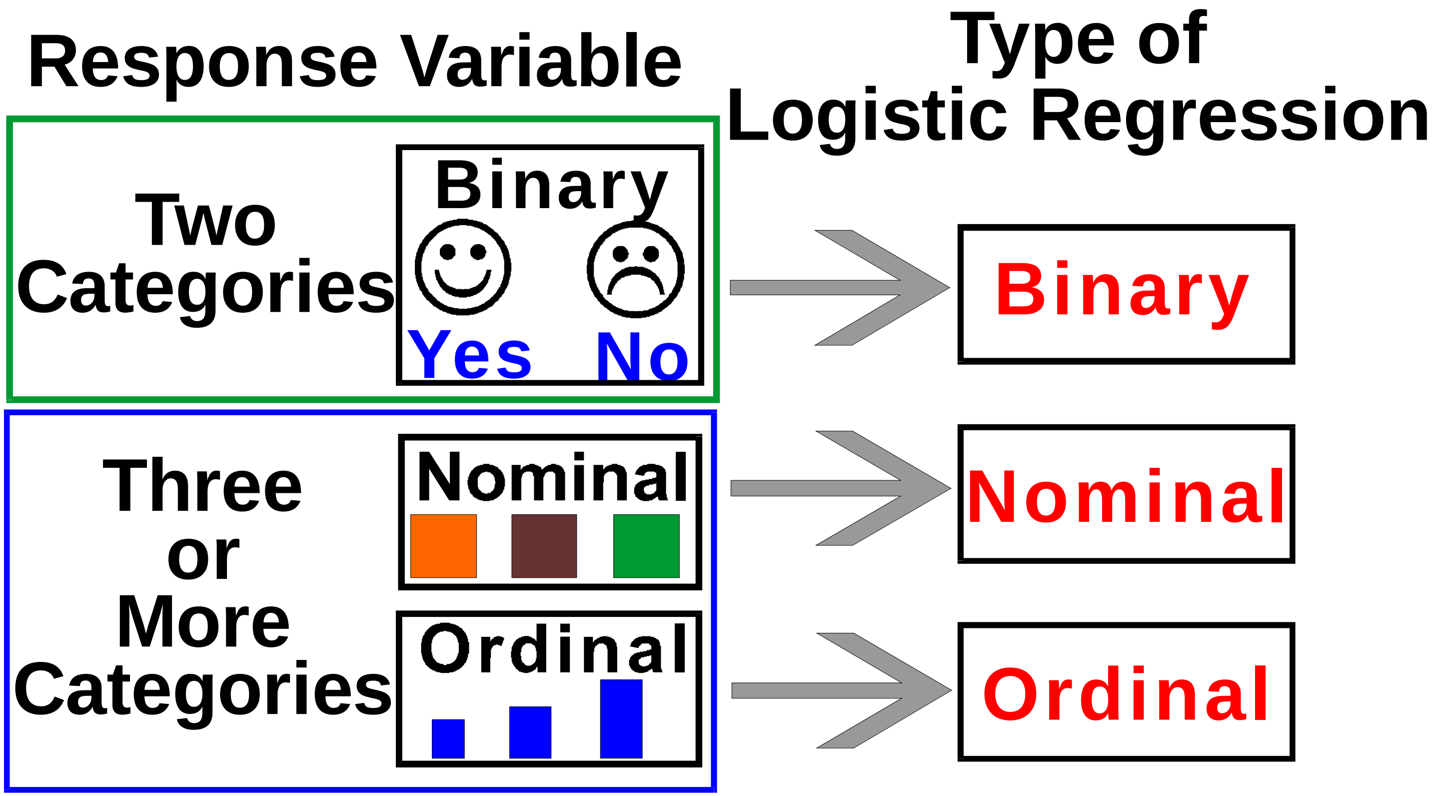
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.025	.018	1.974	1	.160	1.026
	GENDER(1)	.511	.209	5.954	1	.015	1.667
	INCOME			12.305	2	.002	
	INCOME(1)	-.787	.253	9.676	1	.002	.455
	INCOME(2)	-.686	.243	7.945	1	.005	.503
	Constant	-1.325	.720	3.382	1	.066	.266

a. Variable(s) entered on step 1: AGE, GENDER, INCOME.

Adjusted Odds Ratio



Types of Logistic Regression



Model Logistik Multinomial

Ilustrasi data tentang jenis asuransi kesehatan yang tersedia untuk 616 orang yang mengalami depresi psikologis di Amerika Serikat (Tarlov *et al*. 1989; Wells *et al*. 1989). Asuransi dikategorikan sebagai program ganti rugi (*indemnity* yaitu asuransi biaya-untuk-layanan reguler) atau paket prabayar (*prepaid*). Kemungkinan ketiga adalah bahwa orang tsb tidak memiliki asuransi apa pun (*insure*). Kita ingin mengkaji faktor-faktor demografis yang terkait dengan pilihan asuransi. Salah satu faktor demografis dalam data adalah ras peserta, yang dikodekan sebagai putih (*white*) atau bukan putih (*nonwhite*):

Model Logistik Multinomial

https://www.stata.com/manuals13/rmlogit.pdf

```
webuse sysdsn1
tabulate insure nonwhite, chi2 col      /*Ras putih lebih banyak
                                  paket prabayar(prepaid)*/

mlogit insure nonwhite
mlogit insure nonwhite, base(2)
mlogit, rrr                   /*relative-risk ratios=odds ratio*/

mlogit insure age male nonwhite i.site
        /* defaultnya 1=indemnity(ganti rugi) sbg base outcome */
mlogit insure age male nonwhite i.site, baseoutcome(3)
                                /* 3=uninsure sbg base outcome */
mlogit insure age male nonwhite i.site, rrr
    /*reports the estimated coefficients transformed to relative
    -riskratios, that is, exp(b) rather than b */
mlogit, rrr
```

Hipotesis: **Sabuk Pengaman akan membuat Pengendara Lebih aman jika terjadi KECELAKAAN**. Pengendara yang menggunakan Sabuk Pengaman lebih besar Peluangnya mengalami cidera lebih ringan dibandingkan yg tdk menggunakan

Gender	Location	SeatBelt	Respon				
			1	2	3	4	5
Female	Urban	No	7287	175	720	91	10
		Yes	11587	126	577	48	8
	Rural	No	3246	73	710	159	31
		Yes	6134	94	564	82	17
Male	Urban	No	10381	136	566	96	14
		Yes	10969	83	259	37	1
	Rural	No	6123	141	710	188	45
		Yes	6693	74	353	74	12

1. Tidak ada yang luka
2. Terjadi luka ringan
3. Terjadi luka dan memerlukan rawat jalan
4. Terjadi luka dan memerlukan rawat inap
5. Meninggal

Regresi Logistik Ordinal

- Y skala ordinal yg punya c kategori
- Peluang Kumulatif $P(Y \leq j)$: peluang respons Y pd kategori 1,2,...,j
- $P(Y \leq j) = P(Y=1) + P(Y=2) + ... + P(Y=j)$
- $P(Y \leq 1) \leq P(Y \leq 2) \leq ..... \leq P(Y \leq c) = 1$
- *Odd ratio* = $P(Y \leq j) / P(Y > j) = \exp(\beta)$
- Logit $P(Y \leq j) = \text{Log} \{ P(Y \leq j) / P(Y > j) \}$
- Logit $P(Y \leq j) = \alpha j + \beta X$; $j = 1,2,..,c-1$
- Asumsi: pengaruh X sama utk tiap peluang kumulatif. Jika tdk, gunakan regresi logistik nominal.
- $\beta > 0$: Peluang utk <u>nilai order lebih kecil</u>, lebih besar jika X naik satu unit