Penggunaan SAS untuk Analisis Data

PERKENALAN

Dr. Bagus Sartono

Dewi Kiswani Bodro, M.Si

Mulianto Raharjo, S.Stat

Departemen Statistika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Institut Pertanian Bogor

IPB University
Bogor Indonesia

Outline

Hari #1:

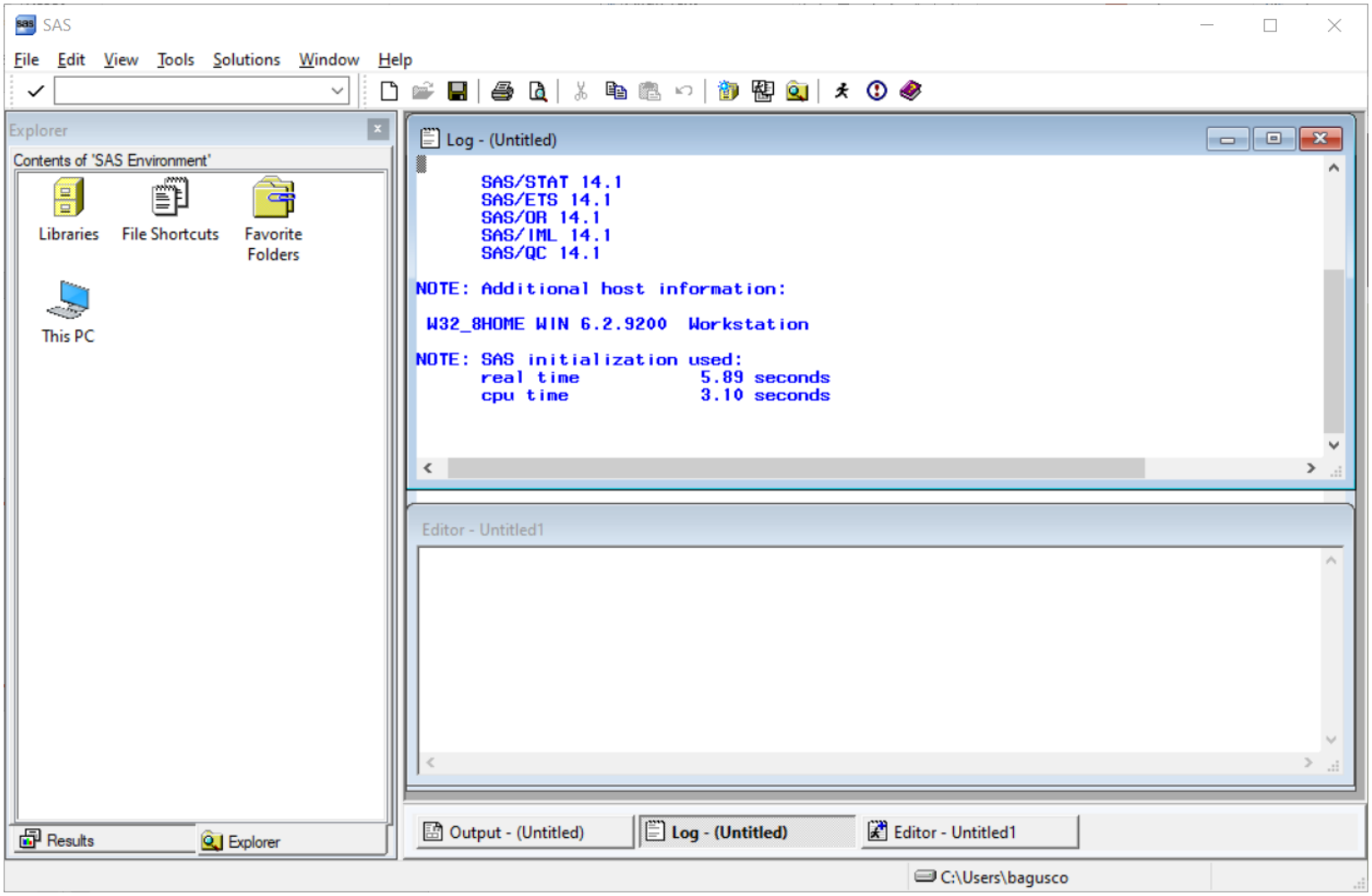
- Bekerja dengan SAS Dataset
- Analisis Statistika Deskriptif
- Peringkasan Data

Hari #2:

- Pengantar Pemodelan Regresi Linear
- Pengantar Pemodelan Regresi Logistik
- Penyusunan Credit Scoring

Bekerja dengan SAS dataset

SAS Working Environment



SAS Working Environment

- Program Editor: jendela untuk membuat program yang akan dijalankan untuk melakukan analisis tertentu
- Log: jendela yang menampilkan pesan setelah program dijalankan, termasuk pesan kesalahan pada program atau peringatan lainnya.
- Output: jendela yang menampilkan output hasil kerja dari prosedur analisis setelah program dijalankan
- Explorer: jendela yang menampilkan direktori/library yang ada dan objek SAS (terutama data) yang ada di dalamnya
- Graph: jendela untuk menampilkan grafik resolusi tinggi yang dihasilkan setelah program dijalankan

Membaca data secara in-stream

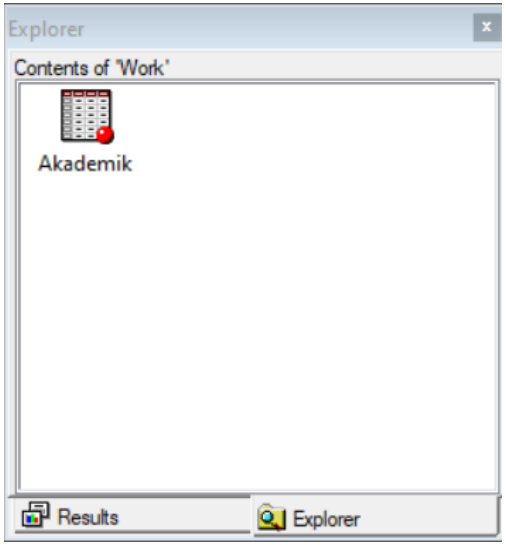
```
data akademik;
input nama$  matematika fisika ekonomi;
datalines;
Andini        80     85     77
Budiman       67     62     89
Rida          54     88     65
Sinta         76     81     69
Rama          88     61     57
;
run;
```

- Membuat SAS dataset dengan nama “akademik” yang berisi empat kolom/variabel: “nama”, “matematika”, “fisika”, “ekonomi”
- Kolom “nama” bertipe string, yang lainnya bertipe numerik
- Data antar kolom/variabel dipisahkan dengan tab atau spasi
- Perhatikan aturan pemberian nama data dan nama kolom!

Membaca data secara in-stream

Catatan pada window LOG yang mengindikasikan program berhasil dijalankan

```
NOTE: The data set WORK.AKADEMIK has 5 observations and 4 variables.
NOTE: DATA statement used (Total process time):
      real time           0.17 seconds
      cpu time            0.07 seconds
```



Terdapat icon data dengan nama akademik pada library WORK yang dapat terlihat pada jendela EXPLORER

PROC PRINT

PROC PRINT berguna untuk mencetak isi dari suatu dataset

```
proc print data=akademik;
run;
```

The SAS System

Obs	nama	matematika	fisika	ekonomi
1	Andini	80	85	77
2	Budiman	67	62	89
3	Rida	54	88	65
4	Sinta	76	81	69
5	Rama	88	61	57

PROC PRINT

Mencetak hanya tiga baris/observasi pertama dari dataset “akademik”

```
proc print data=akademik (obs=3);
run;
```

The SAS System

Obs	nama	matematika	fisika	ekonomi
1	Andini	80	85	77
2	Budiman	67	62	89
3	Rida	54	88	65

PROC PRINT

Mencetak observasi yang nilai matematika-nya lebih dari 75

```
proc print data=akademik;
where matematika > 75;
run;

proc print data=akademik;
where matematika gt 75;
run;
```

The SAS System

Obs	nama	matematika	fisika	ekonomi
1	Andini	80	85	77
4	Sinta	76	81	69
5	Rama	88	61	57

TITLE

Memberikan judul pada output

```
title1 "Daftar Jagoan Matematika";
proc print data=akademik;
where matematika > 75;
run;
```

Daftar Jagoan Matematika

Obs	nama	matematika	fisika	ekonomi
1	Andini	80	85	77
4	Sinta	76	81	69
5	Rama	88	61	57

Latihan

Buatlah sebuah dataset dengan nama “biodata” yang berisi variabel dengan data sebagai berikut

Obs	nama	usia	tinggibadan	gender
1	Budiman	19	162	1
2	Rida	26	154	2
3	Sinta	22	158	2
4	Ratna	19	167	2
5	Herman	23	178	1
6	Andini	16	169	2
7	Rama	17	166	1

Gunakan PROC PRINT untuk mencetak data observasi yang berjenis kelamin laki-laki (gender bernilai 1)

PROC CONTENTS

```
proc contents data=akademik;
run;
```

Data Set Name	WORK.AKADEMIK	Observations	5
Member Type	DATA	Variables	4
Engine	V9	Indexes	0
Created	10/30/2019 09:01:04	Observation Length	32
Last Modified	10/30/2019 09:01:04	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_32		
Encoding	wlatin1 Western (Windows)		

Alphabetic List of Variables and Attributes

#	Variable	Type	Len
4	ekonomi	Num	8
3	fisika	Num	8
2	matematika	Num	8
1	nama	Char	8

Copy, Subsetting, Filtering

```
*mengcopy utuh dataset "akademik"
 menjadi dataset baru bernama "akademik1";
data akademik1;
set akademik;
run;
```

The SAS System

Obs	nama	matematika	fisika	ekonomi
1	Andini	80	85	77
2	Budiman	67	62	89
3	Rida	54	88	65
4	Sinta	76	81	69
5	Rama	88	61	57

Copy, Subsetting, Filtering

```
*mengcopy kolom nama dan matematika dari akademik
 menjadi dataset bernama "matematika";
data matematika (keep = nama matematika);
set akademik;
run;
```

The SAS System

Obs	nama	matematika
1	Andini	80
2	Budiman	67
3	Rida	54
4	Sinta	76
5	Rama	88

Copy, Subsetting, Filtering

```
*mengcopy dataset akademik kecuali kolom ekonomi
 menjadi dataset bernama "ipa";
data ipa (drop = ekonomi);
set akademik;
run;
```

The SAS System

Obs	nama	matematika	fisika
1	Andini	80	85
2	Budiman	67	62
3	Rida	54	88
4	Sinta	76	81
5	Rama	88	61

WHERE STATEMENT

```
*mengcopy hanya observasi yang
nilai matematika lebih dari 75;
data jagomatematika;
set akademik;
where matematika > 75;
run;
```

The SAS System

Obs	nama	matematika	fisika	ekonomi
1	Andini	80	85	77
2	Sinta	76	81	69
3	Rama	88	61	57

WHERE STATEMENT

```
*mengcopy hanya observasi yang
nilai matematika lebih dari 75
dan nilai fisika > 75;
data jagoipa;
set akademik;
where matematika > 75 and fisika > 75;
run;
```

The SAS System

Obs	nama	matematika	fisika	ekonomi
1	Andini	80	85	77
2	Sinta	76	81	69

IF STATEMENT

```
*atau;
*mengcopy hanya observasi yang
nilai matematika lebih dari 75;
data jagomatematika1;
set akademik;
if matematika > 75;
run;
```

Latihan

- Gunakan dataset BIODATA
  - Buat dataset baru yang berisi amatan dengan kriteria gender berkode 2, beri nama “perempuan”
  - Buat dataset baru dengan nama “tinggi” yang hanya berisi kolom/variabel nama dan tinggibadan
  - Buat dataset baru dengan nama remaja yang berisi individu berusia kurang dari 20 tahun

Membuat variabel baru

```
*menambahkan variabel rata-rata nilai;
data akademik;
set akademik;
ratarata = (matematika + fisika + ekonomi)/3;
run;
```

The SAS System

Obs	nama	matematika	fisika	ekonomi	ratarata
1	Andini	80	85	77	80.6667
2	Budiman	67	62	89	72.6667
3	Rida	54	88	65	69.0000
4	Sinta	76	81	69	75.3333
5	Rama	88	61	57	68.6667

If-else

```
*membuat variabel baru menggunakan if-else;
data akademik;
set akademik;
if ratarata > 80 then nilaiakhir = 'lulus';
else nilaiakhir  = 'gagal';
run;
```

The SAS System

Obs	nama	matematika	fisika	ekonomi	ratarata	nilaiakhir
1	Andini	80	85	77	80.6667	lulus
2	Budiman	67	62	89	72.6667	gagal
3	Rida	54	88	65	69.0000	gagal
4	Sinta	76	81	69	75.3333	gagal
5	Rama	88	61	57	68.6667	gagal

If-else

```
*membuat variabel baru menggunakan if-else;
data akademik;
set akademik;
if ratarata > 80 then nilaiakhir = 'lulus';
else if ratarata > 75 then nilaiakhir  = 'cadangan';
else nilaiakhir = 'gagal';
run;
```

The SAS System

Obs	nama	matematika	fisika	ekonomi	ratarata	nilaiakhir
1	Andini	80	85	77	80.6667	lulus
2	Budiman	67	62	89	72.6667	gagal
3	Rida	54	88	65	69.0000	gagal
4	Sinta	76	81	69	75.3333	cadangan
5	Rama	88	61	57	68.6667	gagal

Merging

Perhatikan data “akademik” dan “biodata”

“akademik”

Obs	nama	matematika	fisika	ekonomi	ratarata	nilaiakhir
1	Andini	80	85	77	80.6667	lulus
2	Budiman	67	62	89	72.6667	gagal
3	Rida	54	88	65	69.0000	gagal
4	Sinta	76	81	69	75.3333	cadangan
5	Rama	88	61	57	68.6667	gagal

“biodata”

Obs	nama	usia	tinggibadan	gender
1	Budiman	19	162	1
2	Rida	26	154	2
3	Sinta	22	158	2
4	Ratna	19	167	2
5	Herman	23	178	1
6	Andini	16	169	2
7	Rama	17	166	1

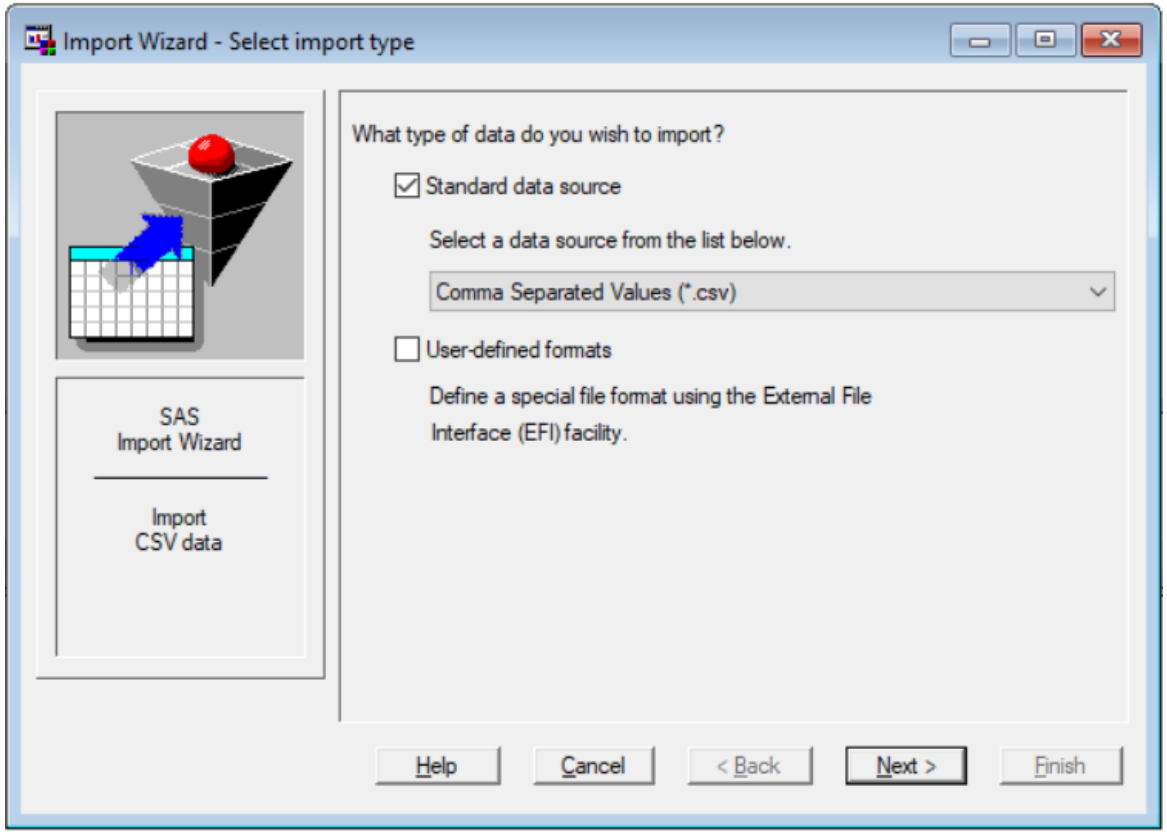
ingin ditambahkan variabel-variabel nilai dari data dataset “akademik” ke dataset “biodata” menggunakan variabel/kolom “nama”

Merging

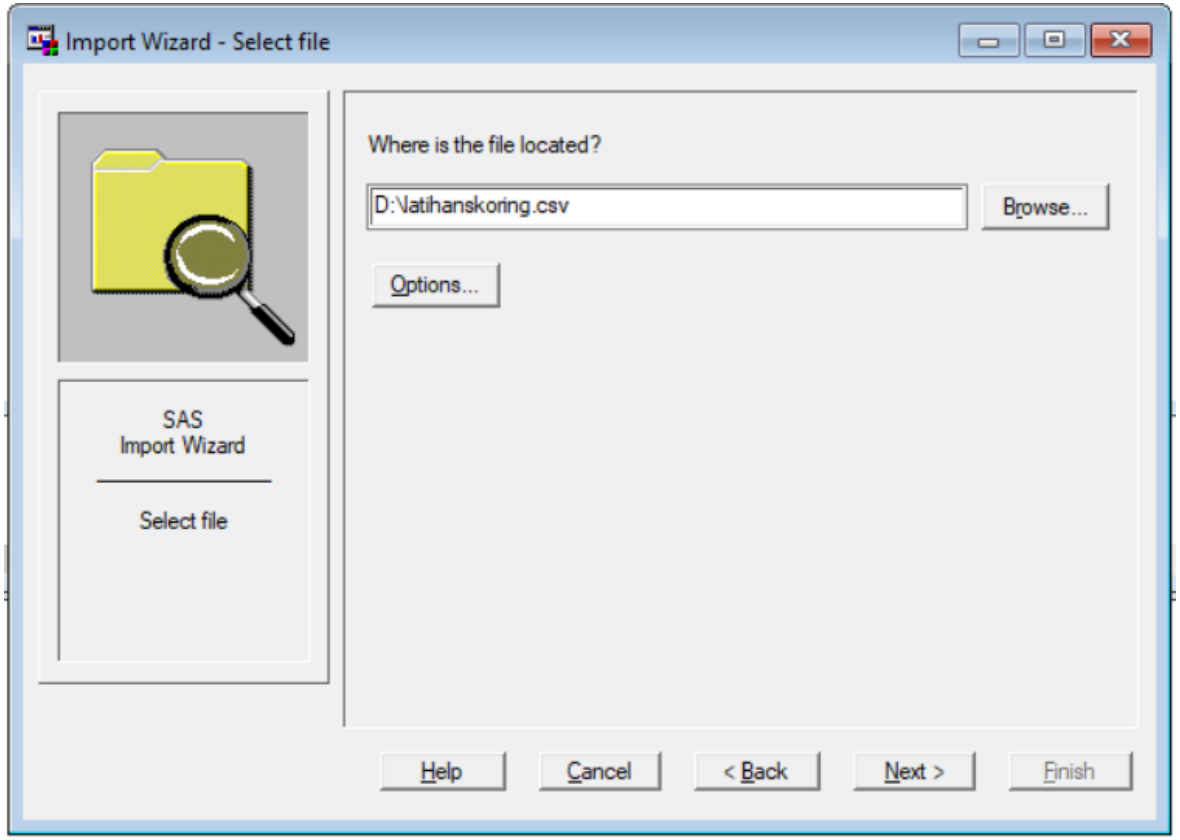
```
*menambahkan variabel-variabel nilai dari
data dataset "akademik" ke dataset "biodata"
menggunakan variabel/kolom "nama";
proc sort data=biodata;
by nama;
proc sort data=akademik;
by nama;
data gabungan;
merge biodata akademik;
by nama;
run;
```

Obs	nama	usia	tinggibadan	gender	matematika	fisika	ekonomi	ratarata	nilaiakhir
1	Andini	16	169	2	80	85	77	80.6667	lulus
2	Budiman	19	162	1	67	62	89	72.6667	gagal
3	Herman	23	178	1	.	.	.	.	
4	Rama	17	166	1	88	61	57	68.6667	gagal
5	Ratna	19	167	2	.	.	.	.	
6	Rida	26	154	2	54	88	65	69.0000	gagal
7	Sinta	22	158	2	76	81	69	75.3333	cadangan

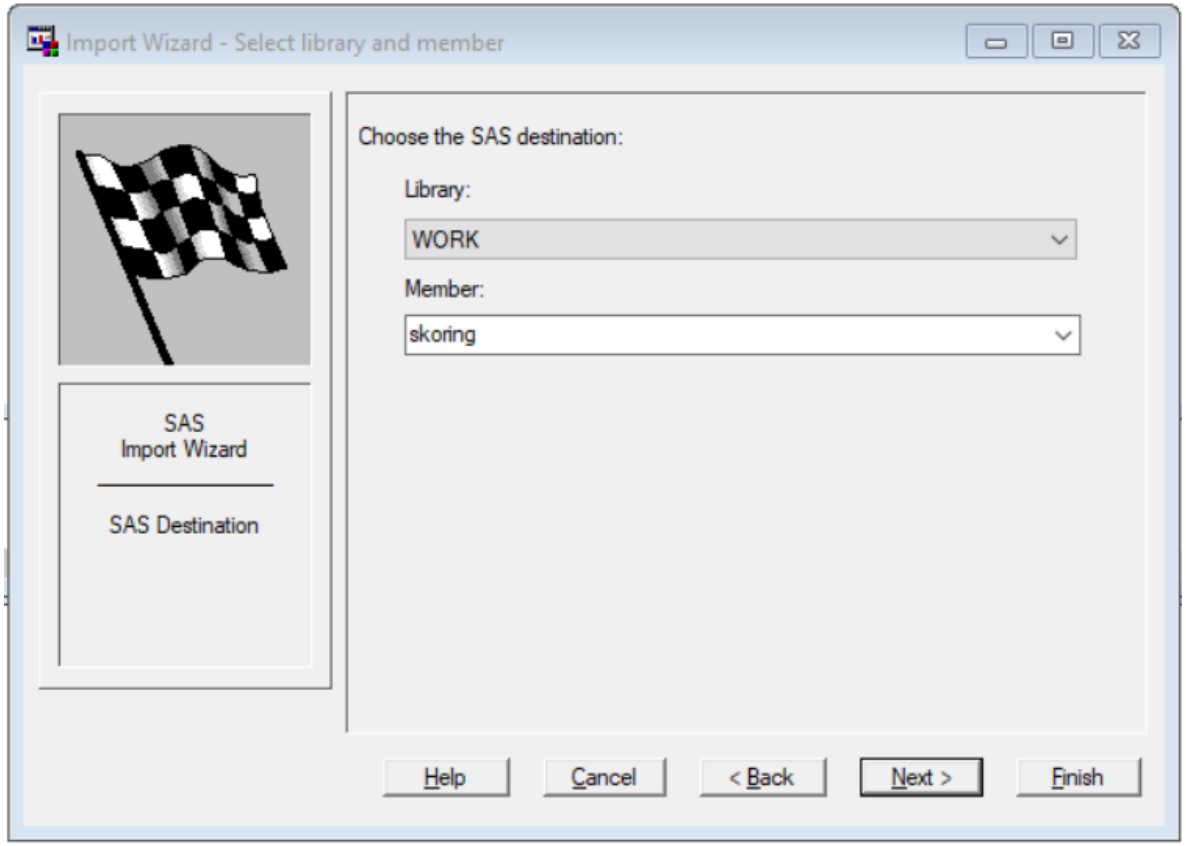
Meng-import file eksternal



Meng-import file eksternal



Meng-import file eksternal



Meng-import file eksternal

```
5000 rows created in WORK.skoring from D:\latihanskoring.csv.


NOTE: WORK.SKORING data set was successfully created.
NOTE: The data set WORK.SKORING has 5000 observations and 8 variables.
```

Obs	DBR	usia	jeniskelamin	marital_status	duration_job	pekerjaan	pendidikan	status
1	0.217740432	31	pria	menikah	15	Karyawan Swasta	Sarjana	1
2	0.196980012	25	pria	single	13	Karyawan Swasta	Sarjana	1
3	0.23158596	39	pria	menikah	11	PNS	Sarjana	1

Peringkasan data

PROC TABULATE

PROC TABULATE

Fitur eksplorasi data dari PROC TABULATE meliputi:

→Mengkontruksi tabel

→Membedakan antara classification variables dan analysis variables

→Menampilkan statistik yang spesifik

→Memformat nilai

→Memberikan label untuk variable dan statistik

PROC TABULATE Syntax

General form untuk PROC TABULATE:

```
PROC TABULATE DATA=dataset <options>;
        CLASS classvariables;
        VAR analysis-variables;
        TABLE pageexpression,
              rowexpression,
              columnexpression </ option(s)>;
RUN;
```

Penentuan Classification Variables

Statement CLASS menentukan variable untuk digunakan sebagai klasifikasi, pengelompokan, atau variabel.

```
PROC TABULATE DATA=dataset <options>;
        CLASS classvariables;
        VAR analysis-variables;
        TABLE pageexpression,
              rowexpression,
              columnexpression </ option(s)>;
RUN;
```

Contoh yang termasuk class variables: **`jeniskelamin`**, **`Pekerjaan`**, and **`Pendidikan`**.

Penentuan Analysis Variables

Pernyataan VAR mengidentifikasi variabel yang akan digunakan sebagai variabel analisis.

PROC TABULATE DATA=*dataset* <*options*>;
    CLASS *classvariables*;
    VAR *analysis-variables*;
    TABLE *pageexpression*,
        rowexpression,
        columnexpression </ *option*(*s*)>;
RUN;

Contoh yang termasuk class variables: **`DBR`** and **`duration_job`**

Penentuan Table Structure

Pernyataan TABEL mengidentifikasi struktur dan format tabel.

```
PROC TABULATE DATA=dataset <options>;
        CLASS classvariables;
        VAR analysis-variables;
        TABLE pageexpression,
              rowexpression,
              columnexpression </ option(s)>;
RUN;
```

Statistic keyword

<*variable name*>***statistic-keyword**;

Descriptive Statistics	Quantile Statistics
COLPCTN	MEDIAN | P50
PCTSUM	P1
COLPCTSUM	Q3 | P75
MAX	P90
ROWPCTN	P95
MEAN	P5
ROWPCTSUM	P10
MIN	P99
STDDEV / STD	Q1 | P25
N	QRANGE
STDERR	
NMISS	
SUM	
PAGEPCTSUM	
PCTN	
VAR	

Penggunaan Class Variables (variabel kategorik)

```
/*Menghasilkan frekuensi masing-masing jenis kelamin*/
```

```
proc tabulate data=a.datalatihan;
        class jeniskelamin;
        table jeniskelamin;
run;
```

jeniskelamin	
pria	wanita
N	N
2994	2006

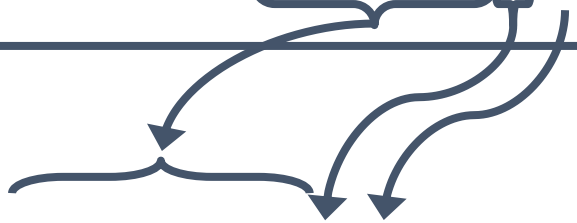
```
proc tabulate data=a.datalatihan;
        class jeniskelamin;
        table jeniskelamin, n;
run;
```

	N
jeniskelamin	
pria	2994
wanita	2006

Option Table “ALL”

Keyword ALL dapat digunakan untuk menghasilkan total.

```
proc tabulate data=a.datalatihan;
        class jeniskelamin;
        table jeniskelamin all;
run;
```



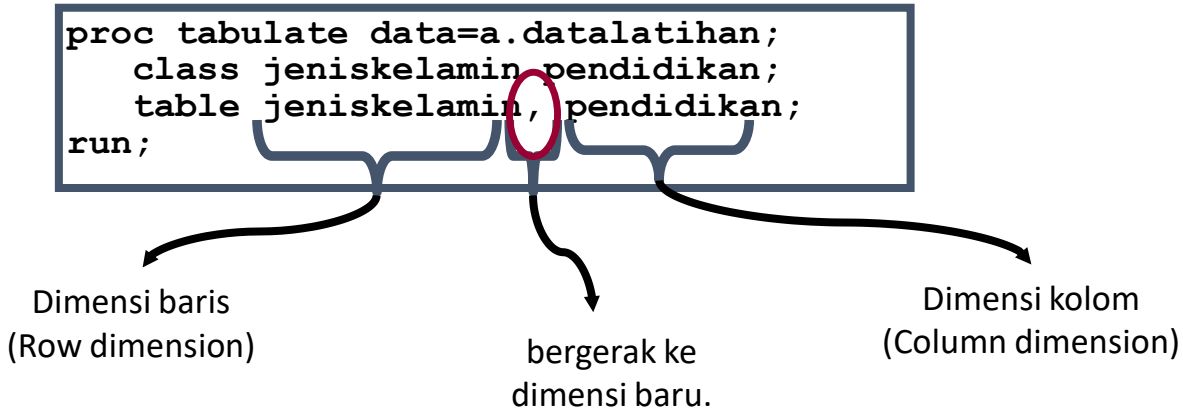
jeniskelamin		All
pria	wanita	
N	N	N
2994	2006	5000

```
proc tabulate data=a.datalatihan;
        class jeniskelamin;
        table jeniskelamin all, n;
run;
```

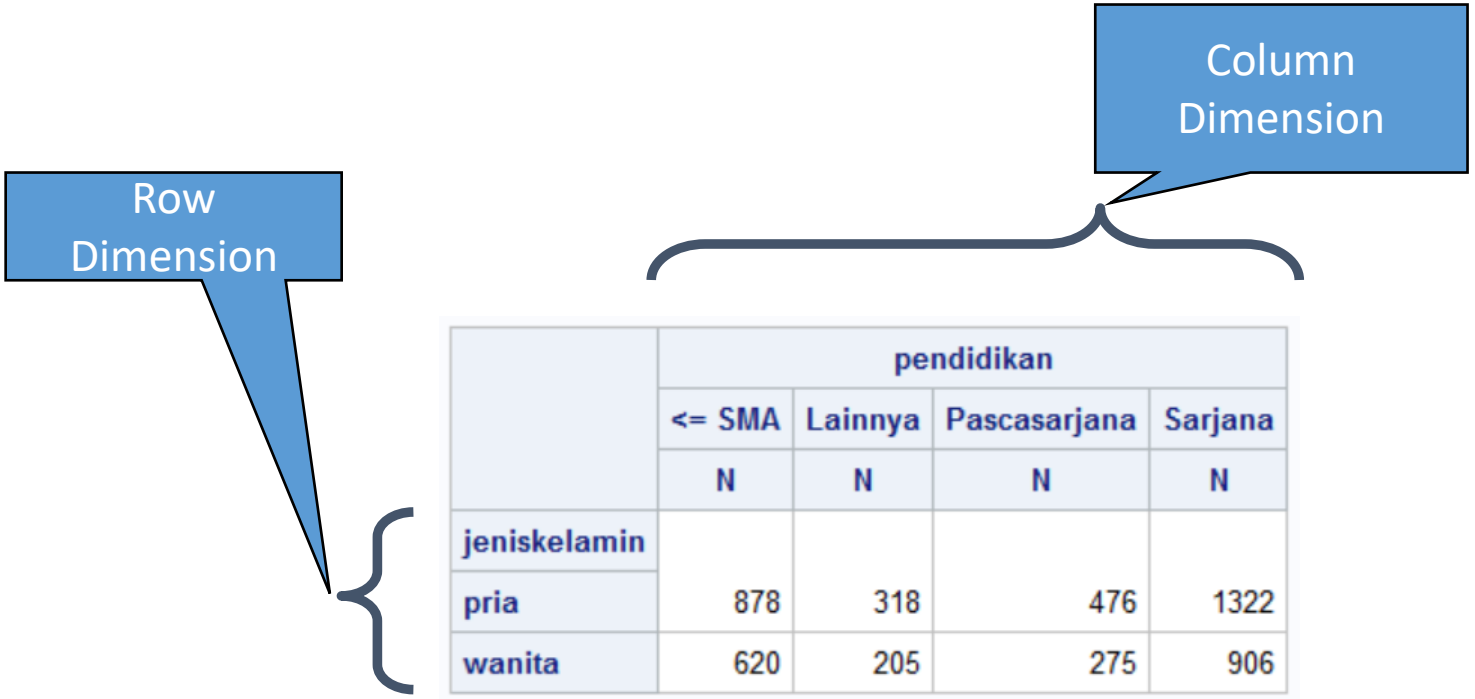
	N
jeniskelamin	
pria	2994
wanita	2006
All	5000

Two-Dimensional Tables

Koma (,) dalam pernyataan TABLE mengarahkan tabel untuk pindah ke dimensi yang berbeda.



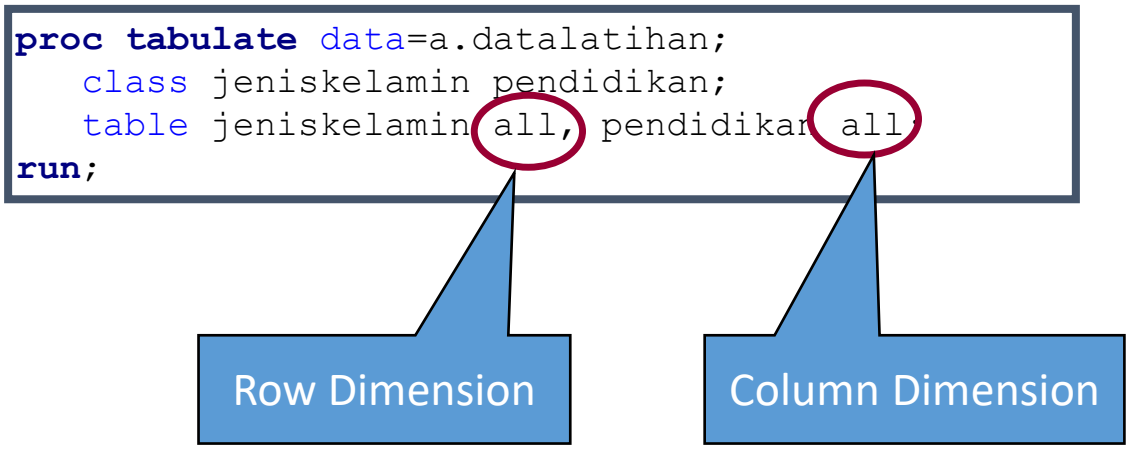
Two-Dimensional Tables



	pendidikan			
	<= SMA	Lainnya	Pascasarjana	Sarjana
	N	N	N	N
jeniskelamin				
pria	878	318	476	1322
wanita	620	205	275	906

Option Table “ALL”

ALL keyword menampilkan total untuk dimensi yang dipilih.



Option Table “ALL”

“All” in the Column Dimension

“All” in the Row Dimension

	pendidikan				All
	<= SMA	Lainnya	Pascasarjana	Sarjana	
	N	N	N	N	N
jeniskelamin					
pria	878	318	476	1322	2994
wanita	620	205	275	906	2006
All	1498	523	751	2228	5000

Option Table “PCTN”

```
/*Menampilkan persentase untuk masing-masing gender*/
```

```
proc tabulate data=a.datalatihan;
        class jeniskelamin;
        table jeniskelamin all,colpctn;
run;
```

	ColPctN
jeniskelamin	
pria	59.88
wanita	40.12
All	100.00

```
proc tabulate data=a.datalatihan;
        class jeniskelamin;
        table jeniskelamin all,n colpctn;
run;
```

	N	ColPctN
jeniskelamin		
pria	2994	59.88
wanita	2006	40.12
All	5000	100.00

Penggunaan Analysis Variables (variabel numerik)

```
/*Menampilkan summary variabel (mean) numerik: DBR*/
```

```
proc tabulate data=a.datalatihan;
        var dbr;
        table dbr*mean;
run;
```

DBR
Mean
0.23

Penggunaan Analysis Variables (variabel numerik)

```
/*Menampilkan rata-rata DBR untuk setiap jenis pekerjaan*/
```

```
proc tabulate data=a.datalatihan format=10.7;
        var dbr;
        class pekerjaan;
        table pekerjaan,dbr*mean;
run;
```

	DBR
	Mean
pekerjaan	
Karyawan Swasta	0.2315873
Lainnya	0.2304095
PNS	0.2300787
Wirausaha	0.2311403

Penggunaan ALL pada variabel numerik

```
/*Menampilkan rata-rata DBR untuk setiap jenis pekerjaan*/
```

```
proc tabulate data=a.datalatihan format=10.7;
          var dbr;
          class pekerjaan;
          table pekerjaan all,dbr*mean;
run;
```

	DBR
	Mean
pekerjaan	
Karyawan Swasta	0.2315873
Lainnya	0.2304095
PNS	0.2300787
Wirausaha	0.2311403
All	0.2309036

Menampilkan beberapa nilai statistik

```
/*Menampilkan beberapa nilai statistik untuk setiap jenis pekerjaan*/

proc tabulate data=a.datalatihan format=10.7;
        var dbr;
        class pekerjaan;
        table pekerjaan all,dbr*(mean std min max);
run;
```

	DBR			
	Mean	Std	Min	Max
pekerjaan				
Karyawan Swasta	0.2315873	0.0284562	0.1437537	0.3000000
Lainnya	0.2304095	0.0283421	0.1487555	0.3000000
PNS	0.2300787	0.0273979	0.1518443	0.3000000
Wirausaha	0.2311403	0.0281851	0.1478455	0.3000000
All	0.2309036	0.0281345	0.1437537	0.3000000

Penggunaan beberapa variabel numerik

```
/*Menampilkan rata-rata beberapa variabel numerik untuk setiap jenis
pekerjaan*/
proc tabulate data=a.datalatihan format=10.7;
        var dbr duration_job;
        class pekerjaan;
        table pekerjaan all,(dbr duration_job)*mean;
run;
```

	DBR	duration_job
	Mean	Mean
pekerjaan		
Karyawan Swasta	0.2315873	15.5577299
Lainnya	0.2304095	15.6457926
PNS	0.2300787	15.7492904
Wirausaha	0.2311403	15.4041787
All	0.2309036	15.5736000

Crosstab

```
/*Crosstab antara jenis kelamin dan jenis pekerjaan dengan variabel
numerik dbr dan durjob*/
```

```
proc tabulate data=a.datalatihan format=10.7;
        var dbr duration_job;
        class pekerjaan jeniskelamin;
        table pekerjaan all,jeniskelamin*(dbr duration_job)*(mean std);
run;
```

	jeniskelamin							
	pria				wanita			
	DBR		duration_job		DBR		duration_job	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
pekerjaan								
Karyawan Swasta	0.2311105	0.0284076	15.3040254	7.0525782	0.2323516	0.0285415	15.9643463	7.5876120
Lainnya	0.2298312	0.0290261	15.5321252	7.6526407	0.2312552	0.0273235	15.8120482	7.8116938
PNS	0.2303779	0.0280590	16.0358891	7.6689266	0.2296658	0.0264839	15.3536036	7.4838403
Wirausaha	0.2299381	0.0280109	15.5385542	7.3882724	0.2329284	0.0283733	15.2043011	7.4265656
All	0.2303761	0.0283449	15.5651303	7.3987158	0.2316908	0.0278060	15.5862413	7.5687565

Crosstab - transpose

```
/*Crosstab antara jenis kelamin dan jenis pekerjaan dengan variabel
numerik dbr dan durjob*/
proc tabulate data=a.datalatihan format=10.7;
        var dbr;
        class pekerjaan jeniskelamin;
        table jeniskelamin*dbr*(mean std),pekerjaan;
run;
```

			pekerjaan			
			Karyawan Swasta	Lainnya	PNS	Wirausaha
jeniskelamin						
pria	DBR	Mean	0.2311105	0.2298312	0.2303779	0.2299381
		Std	0.0284076	0.0290261	0.0280590	0.0280109
wanita	DBR	Mean	0.2323516	0.2312552	0.2296658	0.2329284
		Std	0.0285415	0.0273235	0.0264839	0.0283733

Option Table “COLPCTN”

```
proc tabulate data=a.datalatihan format=10.7;
        var dbr;
        class pekerjaan jeniskelamin;
        table pekerjaan all,jeniskelamin*dbr*(mean n colpctn);
run;
```

	jeniskelamin					
	pria			wanita		
	DBR			DBR		
	Mean	N	ColPctN	Mean	N	ColPctN
pekerjaan						
Karyawan Swasta	0.2311105	944	31.5297261	0.2323516	589	29.3619143
Lainnya	0.2298312	607	20.2738811	0.2312552	415	20.6879362
PNS	0.2303779	613	20.4742819	0.2296658	444	22.1335992
Wirausaha	0.2299381	830	27.7221109	0.2329284	558	27.8165503
All	0.2303761	2994	100.000000	0.2316908	2006	100.000000

WHERE Statement

Pernyataan WHERE dapat digunakan dalam PROC TABULATE untuk mengelompokkan data.

```
proc tabulate data=a.datalatihan;
   where pekerjaan in ("PNS");
   class pekerjaan jeniskelamin;
   table pekerjaan, jeniskelamin;
run;
```

	jeniskelamin	
	pria	wanita
	N	N
pekerjaan		
PNS	613	444

WHERE Statement

```
proc tabulate data=a.datalatihan format = 10.7;
   where pekerjaan in ("PNS","Wirausaha");
   class pekerjaan jeniskelamin;
   var dbr;
   table pekerjaan, jeniskelamin*dbr*mean;
run;
```

DBR nested dalam jeniskelamin

	jeniskelamin	
	pria	wanita
	DBR	DBR
	Mean	Mean
pekerjaan		
PNS	0.2303779	0.2296658
Wirausaha	0.2299381	0.2329284

LATIHAN

- Hasilkan output seperti tampilan di bawah ini:

1.

	N
pendidikan	
<= SMA	1498
Lainnya	523
Pascasarjana	751
Sarjana	2228
All	5000

2.

usia			
Mean	Std	Min	Max
34.37	6.25	18.00	69.00

3.

	pendidikan								RowPctN	N
	<= SMA		Lainnya		Pascasarjana		Sarjana			
	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN	N		
jeniskelamin										
pria	29.33	878	10.62	318	15.90	476	44.15	1322	100.00	2994
wanita	30.91	620	10.22	205	13.71	275	45.16	906	100.00	2006

LATIHAN

- Hasilkan output seperti tampilan di bawah ini:

4.

	usia		
	Mean	Max	NMiss
pekerjaan			
Karyawan Swasta	34.51	60.00	0
Lainnya	34.45	65.00	0
PNS	34.32	69.00	0
Wirausaha	34.19	58.00	0
All	34.37	69.00	0

5.

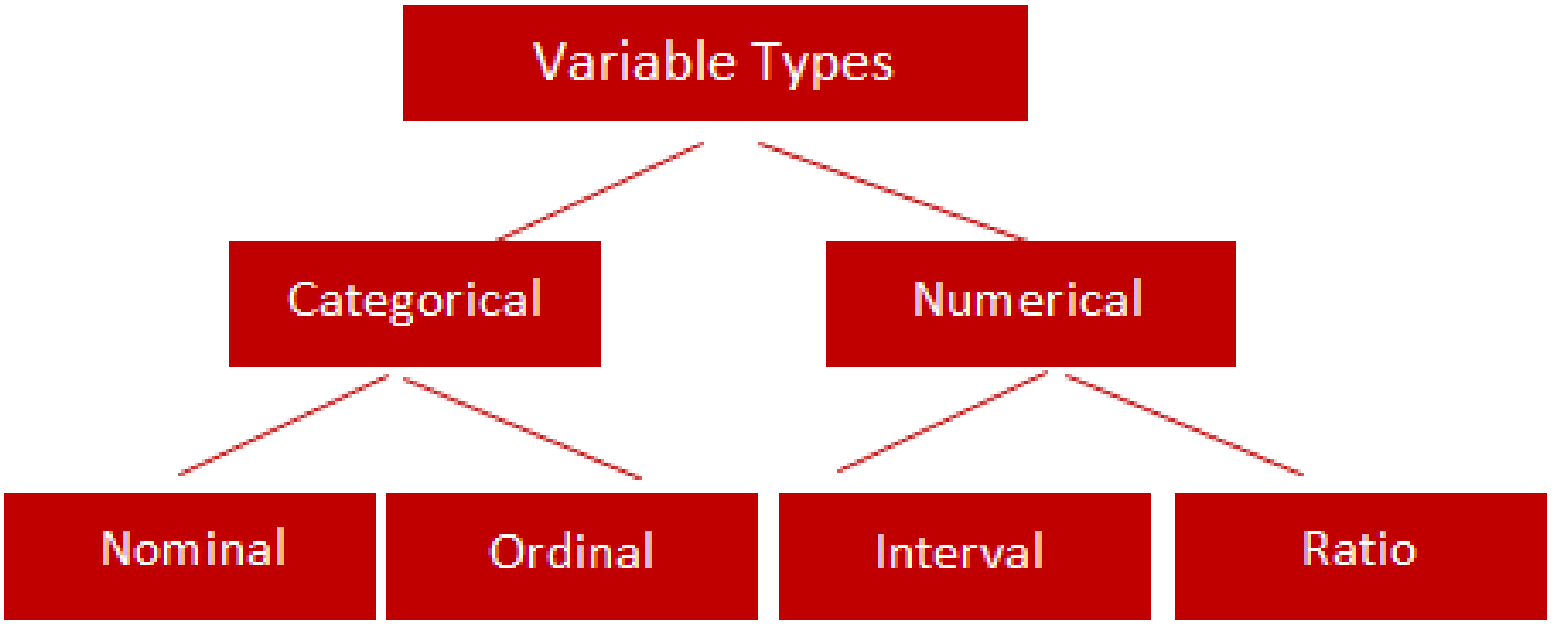
	pendidikan			
	Pascasarjana		Sarjana	
	usia		usia	
	Mean	Std	Mean	Std
jeniskelamin				
wanita	34.8800	7.0034	34.2671	6.3088
All	34.8800	7.0034	34.2671	6.3088

6.

	jeniskelamin		All
	pria	wanita	
	DBR	DBR	DBR
	Mean	Mean	Mean
pekerjaan			
PNS	0.2303779	0.2296658	0.2300787
Wirausaha	0.2299381	0.2329284	0.2311403
All	0.2301249	0.2314827	0.2306813

STATISTIK deskriptif untuk Data NUMERIK

Jenis Variabel



Jenis Variabel

- Numerik
  - Misal: income, age, number of dependants
  - Terhadapnya dapat dilakukan operasi-operasi aritmatika

- Kategorik
  - Misal: gender, occupation, residential ownership
  - Ada yang bersifat ordinal, ada yang bersifat nominal

STATISTIK DESKRIPSI DATA NUMERIK

Ukuran Pemusatan	• Rata-Rata (rataan) • Median • Modus
Ukuran Penyebaran	• Ragam (variance) • Simpangan baku (standard deviation) • Inter Quartile Range (Q3 – Q1)
Ukuran Bentuk Sebaran	• Skewness • Kurtosis

PROC MEANS

```
*menampilkan beberapa statistik deskriptif;
*untuk variabel DBR;
proc means data=skoring;
var dbr;
run;
```

The MEANS Procedure

Analysis Variable : DBR				
N	Mean	Std Dev	Minimum	Maximum
5000	0.2309036	0.0281345	0.1437537	0.3000000

PROC MEANS

```
*menampilkan beberapa statistik deskriptif pilihan;
*untuk variabel DBR;
proc means data=skoring min max mean std var median;
var dbr;
run;
```

The MEANS Procedure

Analysis Variable : DBR					
Minimum	Maximum	Mean	Std Dev	Variance	Median
0.1437537	0.3000000	0.2309036	0.0281345	0.000791552	0.2297512

PROC MEANS

```
*menampilkan beberapa statistik deskriptif;
*untuk variabel DBR dan USIA sekaligus;
proc means data=skoring;
var dbr usia;
run;
```

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
DBR	5000	0.2309036	0.0281345	0.1437537	0.3000000
usia	5000	34.3710000	6.2489654	18.0000000	69.0000000

PROC MEANS

```
*menampilkan beberapa statistik deskriptif;
*untuk variabel DBR berdasarkan STATUS KREDIT;
proc means data=skoring;
var dbr;
class status;
run;
```

The MEANS Procedure

Analysis Variable : DBR						
status	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	1139	1139	0.2384886	0.0270932	0.1596618	0.3000000
1	3861	3861	0.2286660	0.0280487	0.1437537	0.3000000

PROC UNIVARIATE

```
*menampilkan statistik deskriptif
*variabel DBR menggunakan PROC UNIVARIATE;
proc univariate data=skoring;
var dbr;
run;
```

The UNIVARIATE Procedure
Variable: DBR

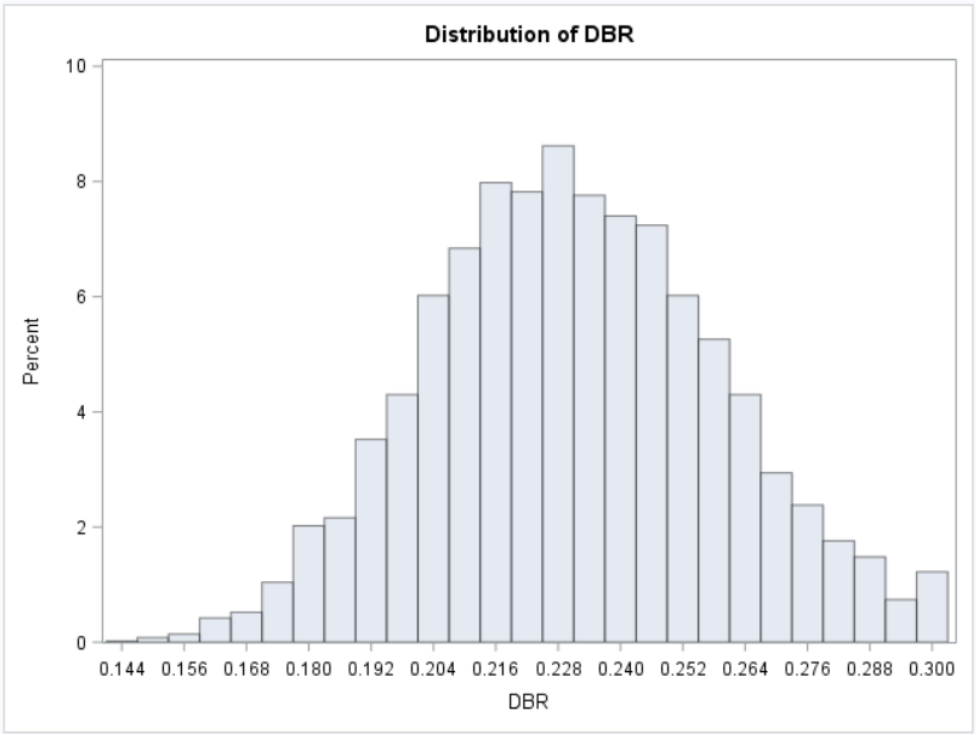
Moments			
N	5000	Sum Weights	5000
Mean	0.23090355	Sum Observations	1154.51776
Std Deviation	0.02813453	Variance	0.00079155
Skewness	0.10105182	Kurtosis	-0.2857775
Uncorrected SS	270.539222	Corrected SS	3.95696811
Coeff Variation	12.1845387	Std Error Mean	0.00039788

Basic Statistical Measures			
Location		Variability	
Mean	0.230904	Std Deviation	0.02813
Median	0.229751	Variance	0.0007916
Mode	0.300000	Range	0.15625
		Interquartile Range	0.03905

Quantiles (Definition 5)	
Level	Quantile
100% Max	0.300000
99%	0.299795
95%	0.279621
90%	0.267686
75% Q3	0.250148
50% Median	0.229751
25% Q1	0.211100
10%	0.195060
5%	0.185223
1%	0.169064
0% Min	0.143754

PROC UNIVARIATE

```
*menampilkan histogram variabel DBR
 menggunakan PROC UNIVARIATE;
proc univariate data=skoring;
var dbr;
histogram dbr;
run;
```



PENGENALAN REGRESI LINIER

Outline

- Pengenalan Pemodelan Prediktif
- Analisis Korelasi
- Analisis Regresi Linear
- Analisis Regresi Logistik

bagian 1

Pengenalan Pemodelan Prediktif

Predictive Model

- model statistik yang digunakan untuk menduga kejadian yang belum diketahui
- disusun berdasarkan data historis (baik data transaksional maupun operasional)
- dalam dunia bisnis sering digunakan untuk menduga resiko dan peluang bisnis (opportunity), berdasarkan karakteristik atau atribut tertentu
- yang ingin diduga (target variable) dapat berupa variabel numerik (misal: harga, profit, exchange rate) atau berupa variabel kategorik/kelas (misal: tingkat resiko, kemauan membeli, keberhasilan)

Contoh Penggunaan Model Prediktif

Credit Scoring – Approval

- Digunakan untuk menduga apakah seorang applicant (pemohon) kredit termasuk pada pemohon yang akan lancar bayar (good) atau tidak (bad).
- Didasarkan pada berbagai variabel demografi, sosial ekonomi, atau history kredit yang dimiliki.
- Variabel targetnya bersifat kategorik: Good / Bad

Contoh Penggunaan Model Prediktif

Residence Pricing Strategy

- Digunakan untuk menduga nilai/harga rumah yang pantas untuk ditawarkan kepada calon pembeli

- Didasarkan pada berbagai variabel karakteristik bangunan (umur, bahan, luas bangunan, luas lahan, dsb), fasilitas (tempat tidur, kamar mandi, garasi, kolam renang), lokasi, infrastruktur pendukung di sekitar rumah (tol, mal/pasar, rumah sakit, olahraga, taman, dsb).

- Variabel targetnya bersifat numerik: Rp atau $

Contoh Penggunaan Model Prediktif

Disease Detection

- Digunakan untuk mendeteksi tingkat keterjangkitan suatu penyakit pada individu

- Didasarkan pada berbagai variabel gaya hidup (makanan, merokok, konsumsi alkohol, dsb), riwayat penyakit orang tua, kondisi lingkungan tempat tinggal.

- Variabel targetnya bersifat kategorik: Ya / Tidak

Contoh Penggunaan Model Prediktif

Propensity Model

- Digunakan untuk mengidentifikasi kecenderungan atau peluang seseorang untuk tertarik pada program penawaran/kampanye produk tertentu.
  - Customer Acquisition
  - Up-Sell & Cross-Sell
  - Credit/Debt Card Activation
- Didasarkan pada informasi demografi, transaksi, dll
- Variabel Target bersifat kategorik: Tertarik / Tidak Tertarik

Komponen dalam Model Prediktif

- **Variabel Target**

Variabel yang akan diprediksi, sering juga disebut variabel respon (response variable) atau variabel tak-bebas (dependent variable). Notasi yang sering digunakan: Y.

- **Variabel Input**

Variabel yang akan digunakan untuk memprediksi, sering juga disebut variabel penjelas (predictor variable) atau variabel bebas (independent variable). Notasi yang sering digunakan: X

- **Model Fungsional**

Sebuah fungsi matematis yang menghubungkan nilai X dengan nilai Y. Model statistika yang banyak digunakan antara lain:
-Regresi Linear
-Regresi Logistik
-Decision Tree

Pemodelan

$$y = f(x_1, x_2, \cdots, x_p) + \varepsilon$$

- Membangun miniatur dari dunia nyata
  - dinyatakan dalam satu atau beberapa fungsi matematis

- Menyederhanakan fenomena nyata sehingga mudah memahami pola umum yang ada
  - memberikan penjelasan terhadap perubahan
  - memberikan penjelasan tentang perbedaan yang terjadi
  - menemukan faktor yang menyebabkan perubahan dan perbedaan

Pemodelan

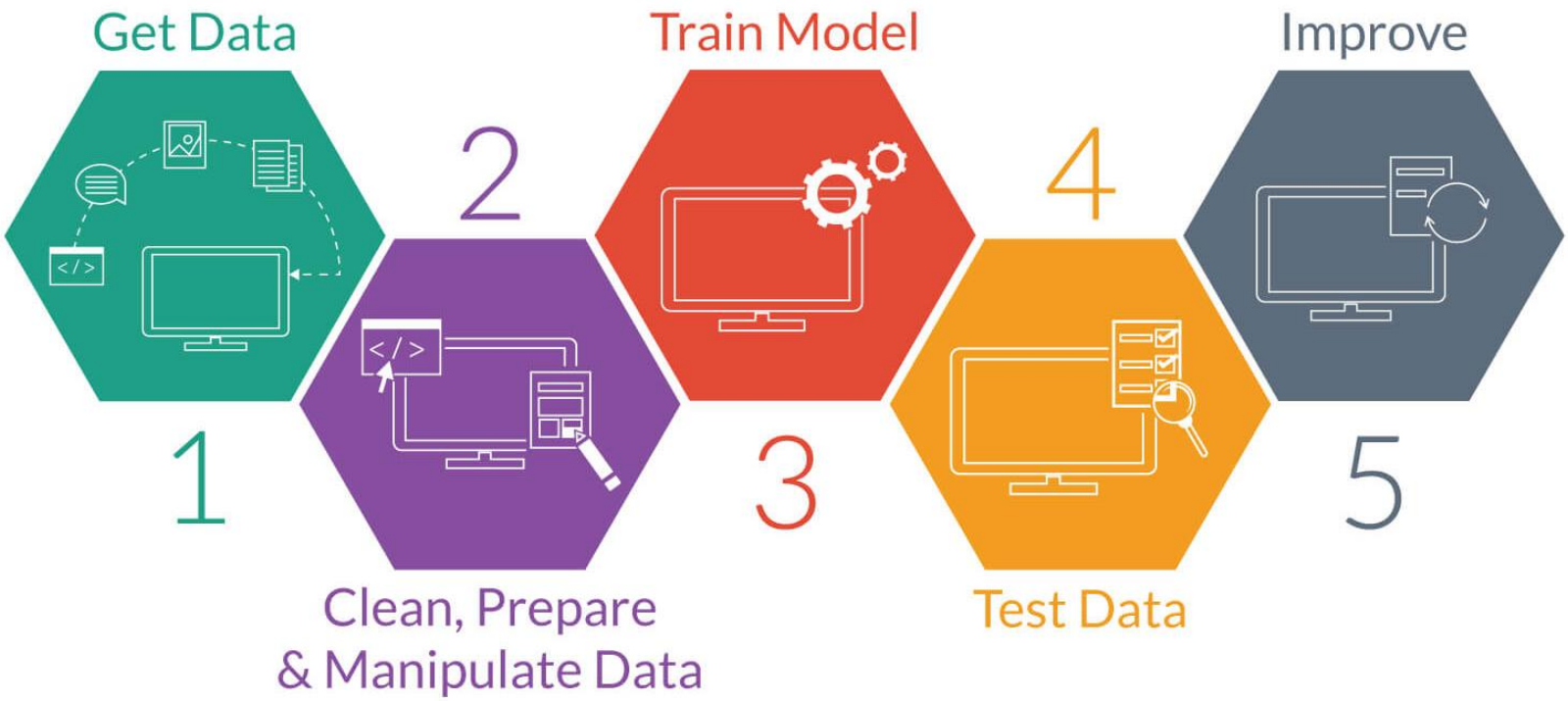
- Tujuan/Manfaat:
  - Sering digunakan untuk meng-explore dataset yang dimiliki
  - Digunakan untuk melakukan prediksi berdasarkan informasi dari variabel prediktor
  - Digunakan untuk mengkaji dan memahami bagaimana suatu variabel berhubungan dengan variabel yang lain

- Are not perfect
  - All models are wrong, but some are useful" (alm. GEP Box)

Beberapa Model Statistika yang Populer

Jenis Variabel Target	Model Statistika
Numerik	Regresi Linear
Kategorik	Regresi Logistik Pohon Klasifikasi (Classification Tree)

Tahapan Pemodelan



bagian 2

Analisis Korelasi

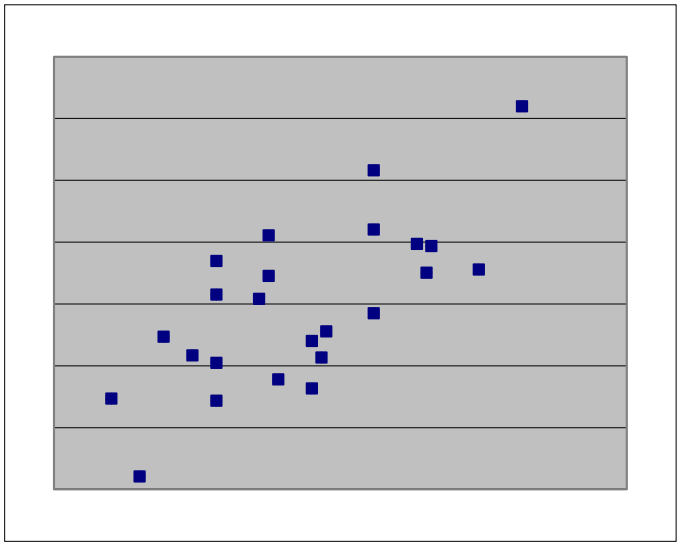
Hubungan Antar Peubah

- Dari setiap objek/individu/tempat/dll dapat diukur/dicatat/diamati lebih dari satu buah peubah.

- Nilai dari suatu peubah bersifat:
  - saling bebas dengan peubah lain
  - saling terkait dengan peubah lain

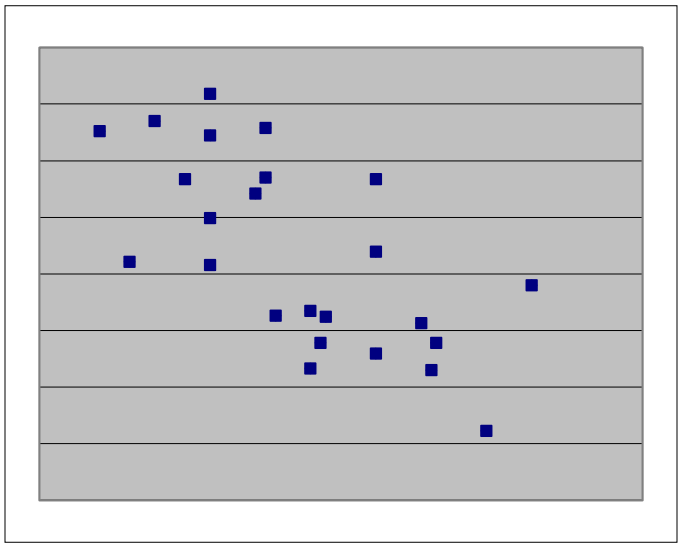
Hubungan antar Peubah

berat badan



tinggi badan

suhu rata-rata



ketinggian tempat

Koefisien Korelasi

- Diperlukan sebuah ukuran yang dapat mencirikan keeratan hubungan antar dua peubah.

- Koefisien Korelasi (ρ ; baca: rho)
  - nilainya: $-1 \leq \rho \leq 1$
  - Tanda plus/minus menunjukkan arah hubungan
  - besar/magnitude menunjukkan kekuatan hubungan
  - koefisien korelasi data contoh dinotasikan r

Koefisien Korelasi (Pearson)

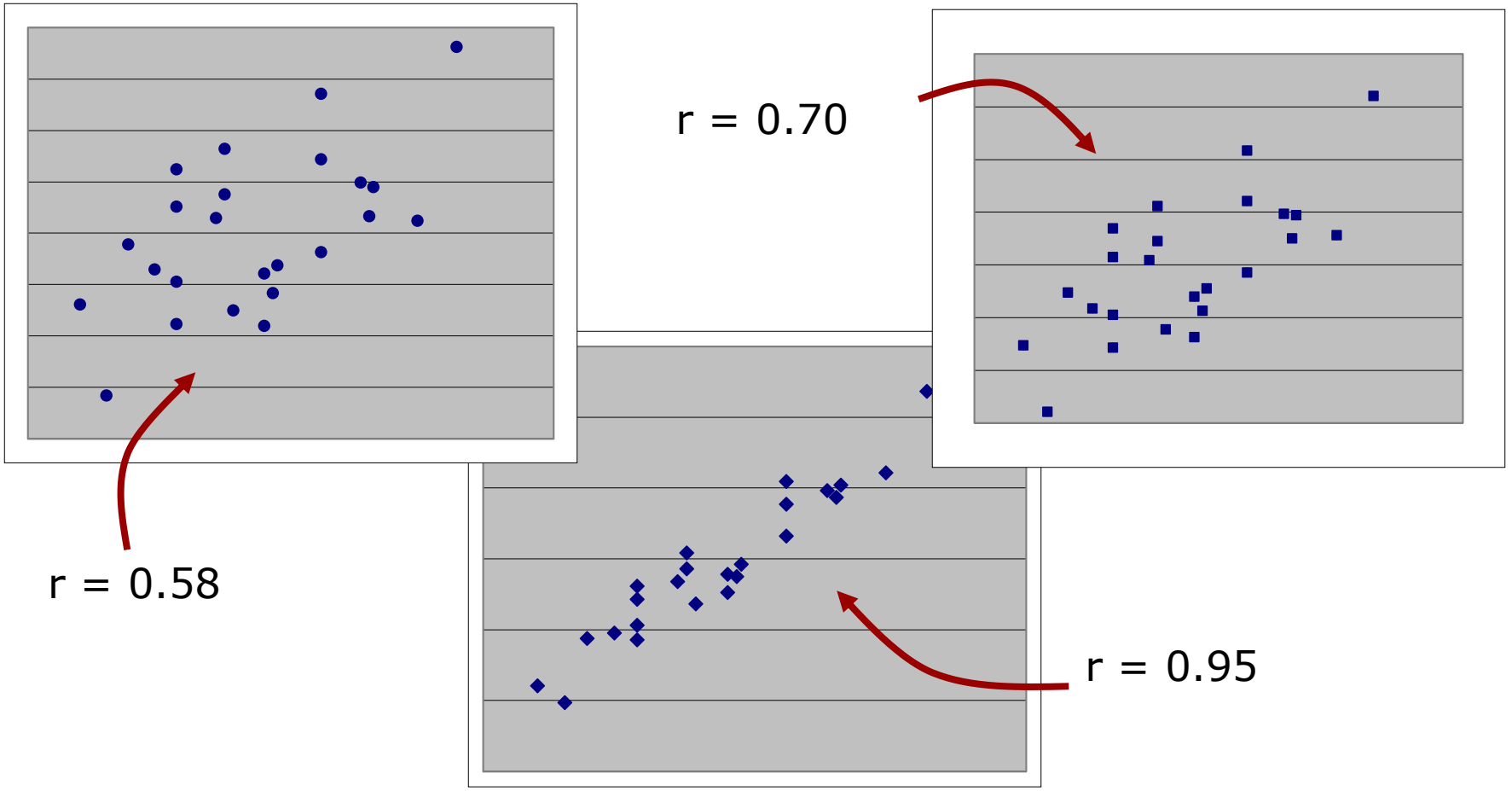
Jika ada dua peubah X dan Y, korelasi antara keduanya adalah

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

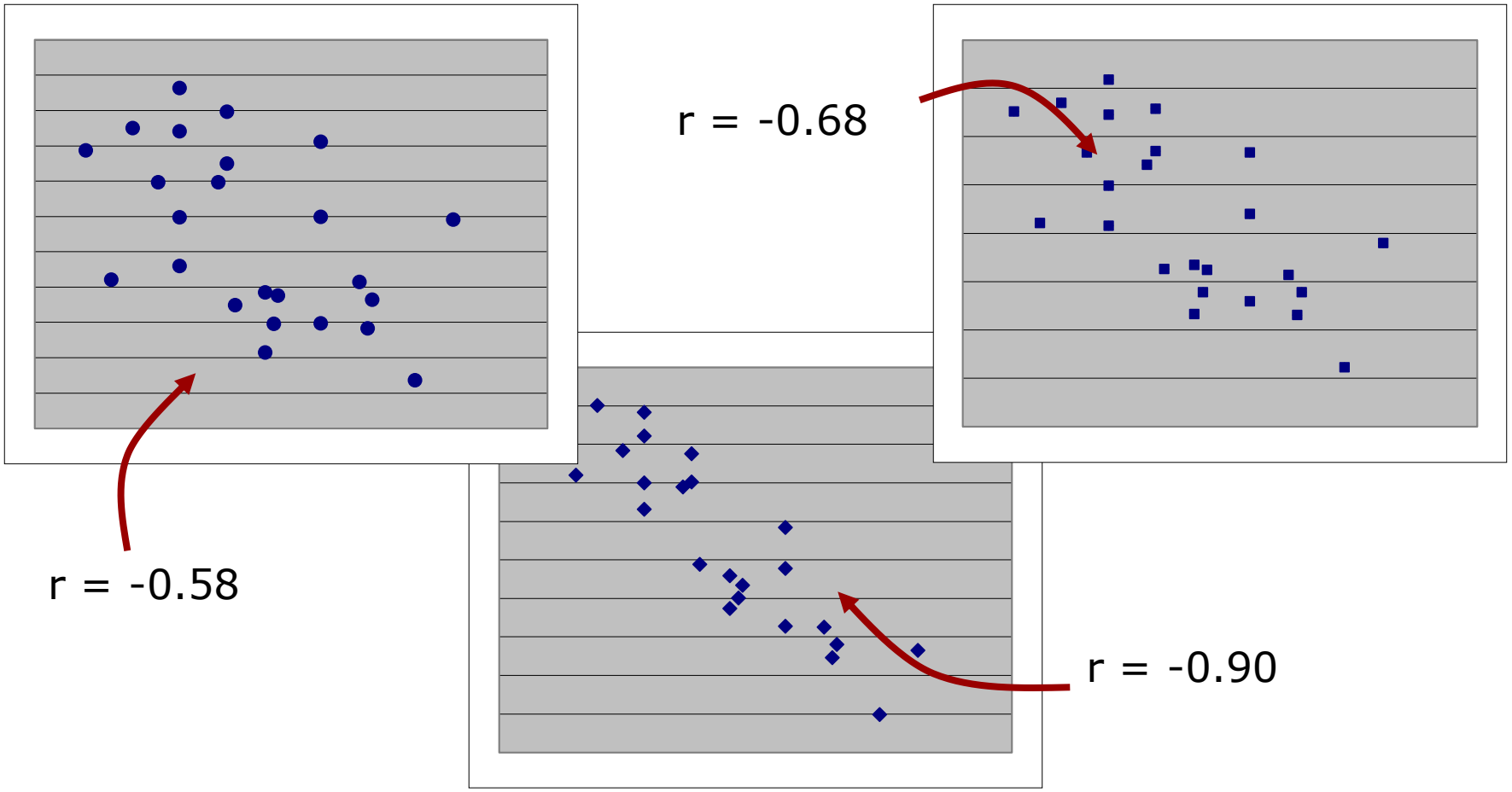
Correlation Coefficient Scale

+ r values	Positive	– r values	Negative
 1.0	Perfect +	 –1.0	Perfect –
 .8 to .99	Very strong+	 –.8 to –.99	Very strong–
 .6 to .8	Strong +	 –.6 to –.8	Strong –
 .4 to .6	Moderate +	 –.4 to –.6	Moderate –
 .2 to .4	Weak +	 –.2 to –.4	Weak –
 0 to .2	Very weak +	 0 to –.2	Very weak –

Koefisien Korelasi (+)



Koefisien Korelasi (-)



bagian 3

Pemodelan Regresi Linear

Pengantar: Regresi Linear

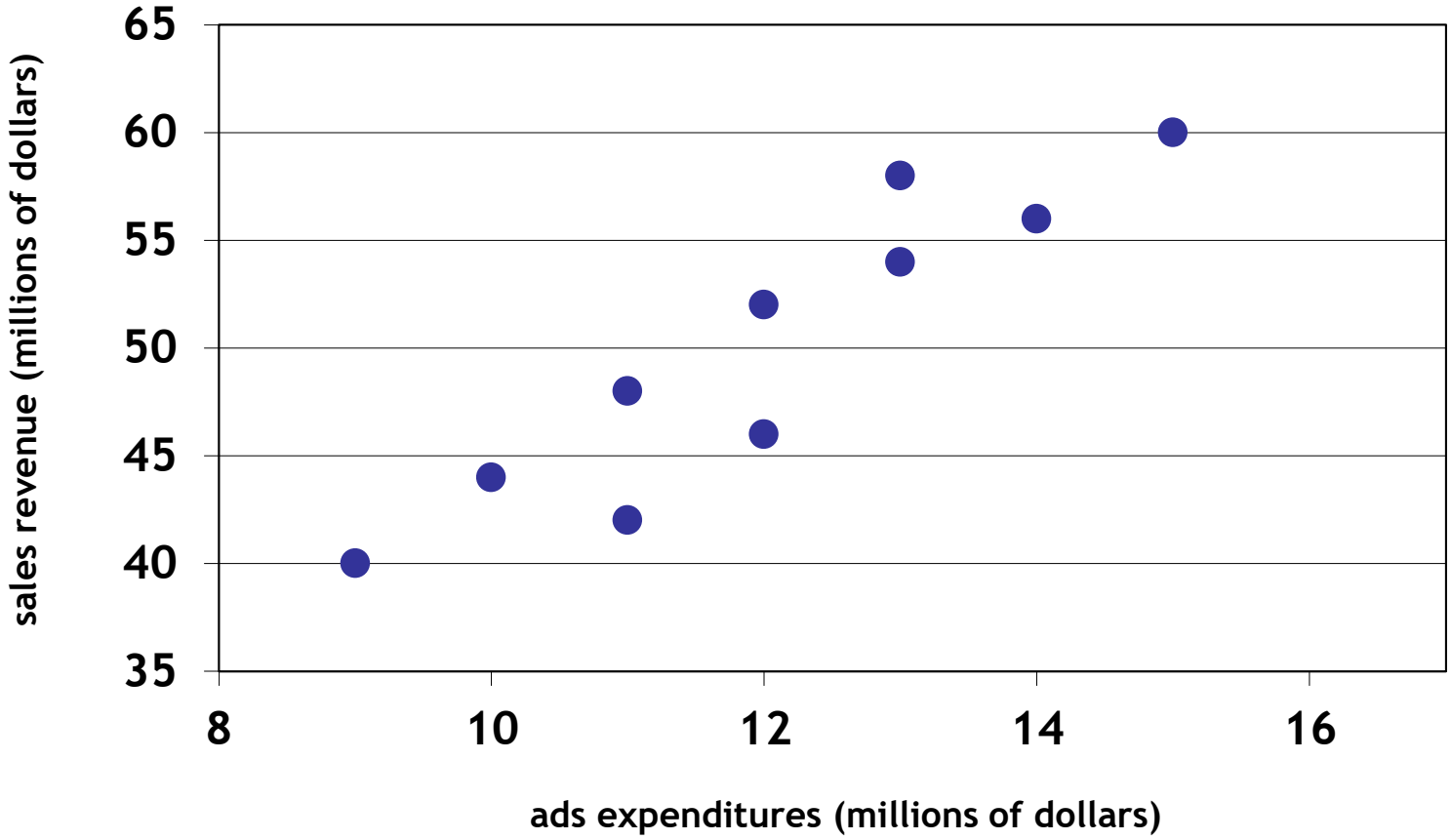
- Terdapat 2 peubah numerik : peubah yang satu mempengaruhi peubah yang lain
- Peubah yang mempengaruhi → X, peubah bebas (independent), peubah penjelas (explanatory)
- Peubah yang dipengaruhi → Y, peubah tak bebas (dependent), peubah respon (response)

Pengantar: Regresi Linear

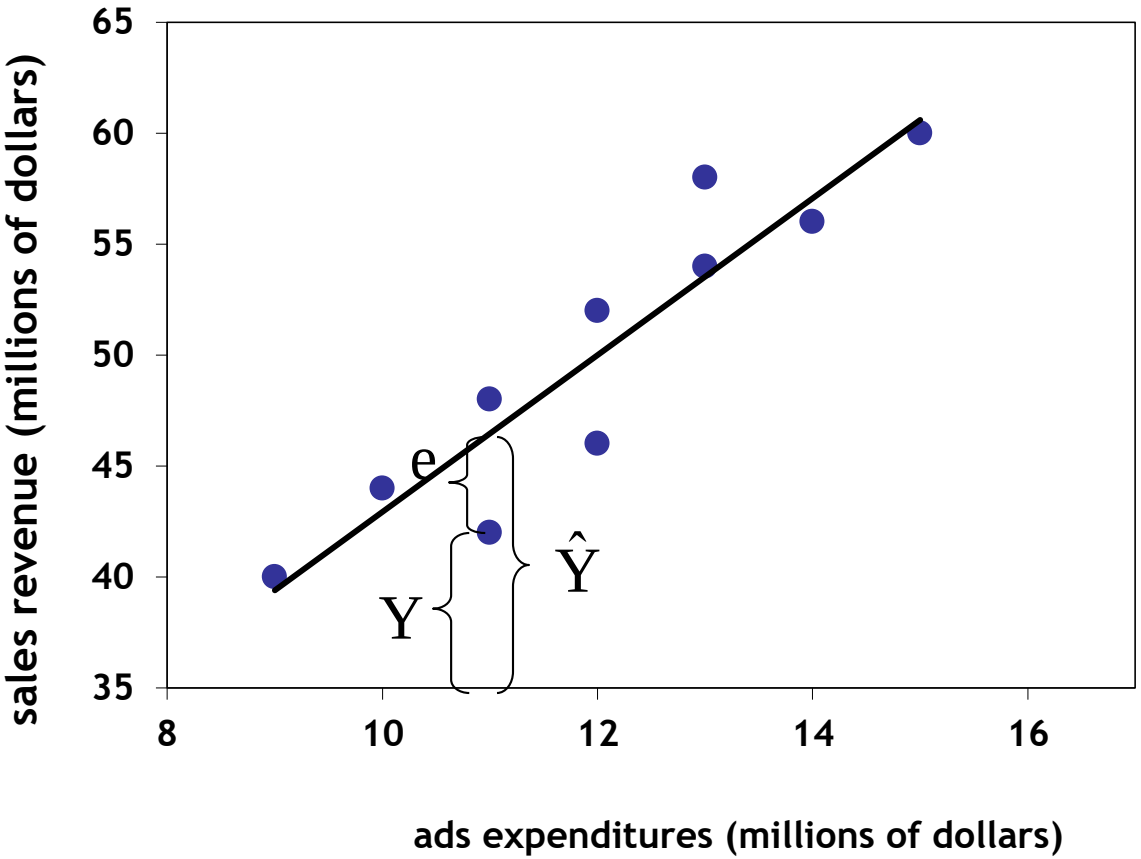
Misalnya ingin melihat hubungan antara pengeluaran untuk iklan (ads expenditures, X) dengan penerimaan melalui penjualan (sales revenue, Y)

Bulan	1	2	3	4	5	6	7	8	9	10
X	10	9	11	12	11	12	13	13	14	15
Y	44	40	42	46	48	52	54	58	56	60

Pengantar: Regresi Linear



Pengantar: Regresi Linear



Ingin dibuat model

$$Y = a + bX$$

Model memuat error, selisih nilai sebenarnya dengan dugaan berdasar model

$$e = Y - \hat{Y}$$

Bagaimana mendapatkan a dan b?

Metode yang digunakan : OLS (ordinary least squares/kuadrat terkecil), mencari a dan b sehingga jumlah kuadrat error paling kecil

Cari penduga a dan b sehingga

minimum $$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[Y_i - \hat{a} - \hat{b} X_i \right]^2$$

Bagaimana mendapatkan a dan b?

$$\hat{b} = \frac{\sum_{i=1}^{n} [X_i - \overline{X}][Y_i - \overline{Y}]}{\sum_{i=1}^{n} [X_i - \overline{X}]^2}$$

$$\hat{a} = \overline{Y} - \hat{b}\overline{X}$$

Rata-rata Y

Rata-rata X

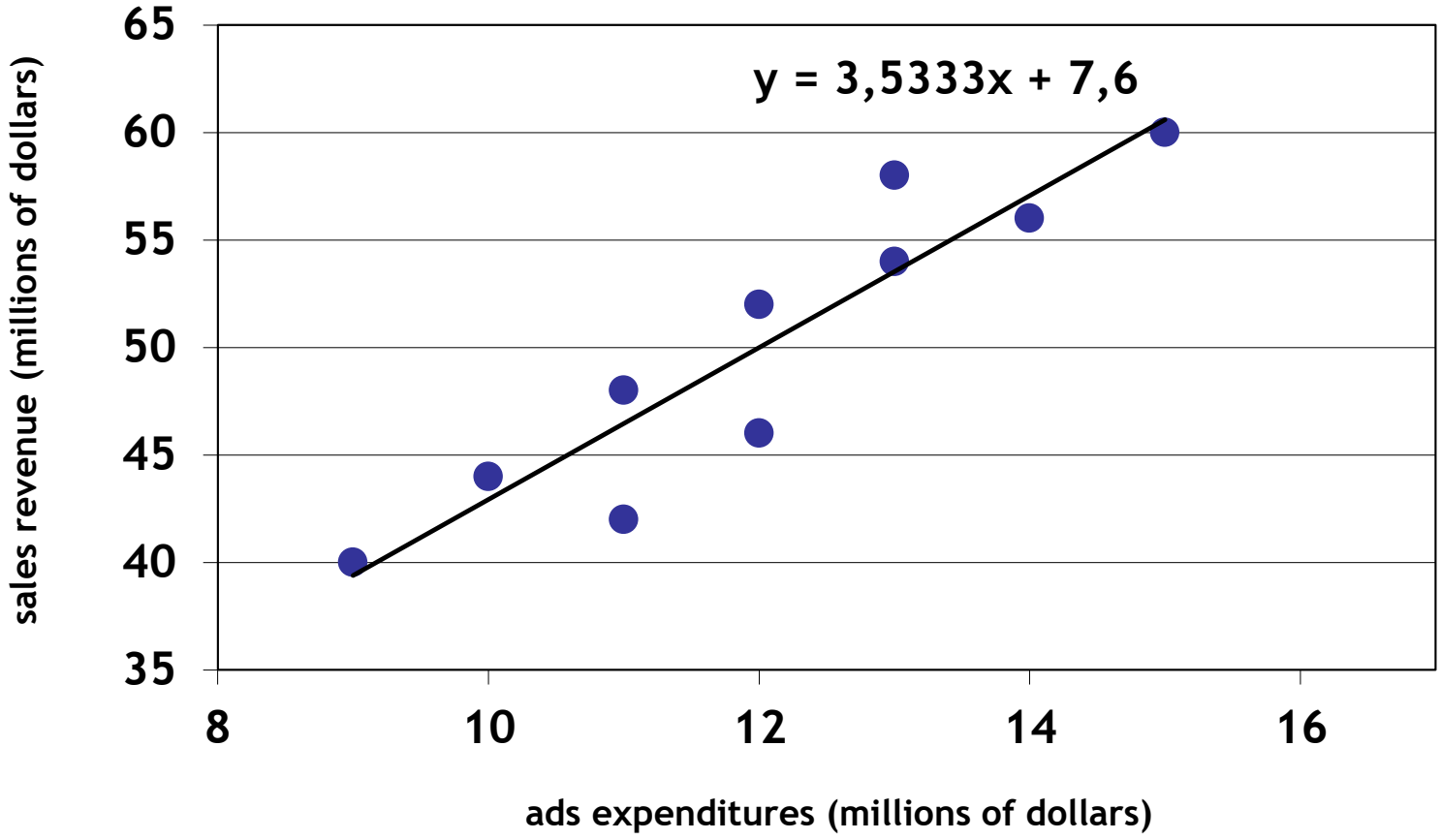
Ilustrasi Perhitungan

X	Y	(X - Xbar)	(Y - Ybar)	(X - Xbar)(Y - Ybar)	$(X - Xbar)^2$
10	44	-2	-6	12	4
9	40	-3	-10	30	9
11	42	-1	-8	8	1
12	46	0	-4	0	0
11	48	-1	-2	2	1
12	52	0	2	0	0
13	54	1	4	4	1
13	58	1	8	8	1
14	56	2	6	12	4
15	60	3	10	30	9

$\overline{X} = 12$ $\quad \sum [X - \overline{X}][Y - \overline{Y}] = 106$ **b = 106 / 30 = 3.533**

$\overline{Y} = 50$ $\quad \sum [X - \overline{X}]^2 = 30$ **a = 50 – 3.533 (12) = 7.60**

Plot X-Y


y = 3,5333x + 7,6
65
60
55
50
45
40
35
8
10
12
14
16
sales revenue (millions of dollars)
ads expenditures (millions of dollars)

Interpretasi a dan b

$Y = 7.6 + 3.53 X$

Penerimaan = 7.6 + 3.53 Belanja Iklan

- a = intersep/intercept = besarnya nilai Y ketika X sebesar 0
- b = gradient/slope = besarnya perubahan nilai Y ketika X berubah satu satuan. Tanda koefisien b menunjukkan arah hubungan X dan Y

Pada kasus ilustrasi

- a = 7.6 → besarnya sales revenue jika tidak ada belanja iklan adalah 7.6 juta dolar
- b = 3.533 → jika belanja iklan dinaikkan 1 juta dolar maka sales revenue naik 3.533 juta dolar

Uji Signifikasi Koefisien b

H_0 : b = 0 (artinya X tidak mempengaruhi Y)

H_1 : b ≠ 0 (artinya X mempengaruhi Y)

statistik uji

$$t = \frac{\hat{b}}{s_{\hat{b}}}$$

$$s_{\hat{b}} = \sqrt{\frac{\sum_{i=1}^{n}\left[Y_i - \hat{Y}_i\right]^2}{(n-2)\sum_{i=1}^{n}\left[X_i - \overline{X}\right]^2}}$$

Tolak H_0 jika nilai |t| melebihi nilai t pada tabel dengan derajat bebas (n-2) dengan tingkat kesalahan $\alpha/2$

Uji Signifikasi Koefisien b

- Nilai $s_b = \sqrt{}(65.47 / (8)(30)) = 0.52$
- Nilai t = 3.53 / 0.52 = 6.79
- Nilai t pada tabel (db = 8, α = 5%) = 2.306
- Kesimpulan : Tolak H_0, data mendukung kesimpulan adanya pengaruh ads expenditure terhadap sales revenue.

Ukuran Kebaikan Model

- Menggunakan koefisien determinasi (R^2, R-squared)
- R-squared bernilai antara 0 s/d 1
- R-squared adalah persentase keragaman data yang mampu diterangkan oleh model
- R-squared tinggi adalah indikasi model yang baik

Ukuran Kebaikan Model

$$R^2 = \frac{\sum \left[\hat{Y}_i - \bar{Y}\right]^2}{\sum \left[Y_i - \bar{Y}\right]^2} = 1 - \frac{\sum e^2}{\sum \left[Y_i - \bar{Y}\right]^2}$$

- Model dalam ilustrasi bisa ditunjukkan memiliki R-squared 0.85 atau 85%

Regresi Linear Berganda (multiple linear regression)

Pengantar: Regresi Linier Berganda

- Simple Linear Regression (regresi linear sederhana): hanya ada 1 peubah penjelas. Modelnya:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Multiple Linear Regression (regresi linear berganda): melibatkan lebih dari satu peubah penjelas. Modelnya:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Pengantar: Regresi Linier Berganda

- Harga jual rumah dipengaruhi oleh beberapa peubah penjelas, misalnya: (1) luas bangunan, (2) umur bangunan, (3) jarak lokasi rumah ke jalan raya.

- Tinggi pohon tanaman tertentu dipengaruhi oleh (1) umur tanaman, (2) dosis pupuk yang diberikan, (3) kandungan hara tanah, (4) curah hujan di lokasi penanaman.

Notasi Matriks Model Regresi

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_2 \end{bmatrix}$$

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

Penduga Kuadrat Terkecil (least square estimator)

Penduga OLS (ordinary least squares) bagi β adalah

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

bersifat BLUE (best linear unbiased estimator) jika

$E(\varepsilon) = 0$

$var(\varepsilon) = \sigma^2$

$cov(\varepsilon_i, \varepsilon_j) = 0$

ε menyebar normal

Pengujian Pengaruh Variabel Penjelas

- Uji Simultan
  - ANOVA digunakan untuk menguji secara simultan pengaruh seluruh X terhadap Y
    - H_0: semua $b_i = 0$ (tidak ada X yang berpengaruh terhadap Y)
    - H_1: ada $b_i \neq 0$ (ada X yang berpengaruh terhadap Y)

- Jika nilai p-value kecil, disimpulkan tolak H_0. dengan kata lain, jika nilai p-value kecil berarti ada X yang berpengaruh terhadap Y.

- Sebaliknya, jika p-value besar, maka tidak ada X yang pengaruhnya signifikan.

Pengujian Pengaruh Variabel Penjelas

- Uji Parsial → uji t

$$\text{statistik uji}$$

$$t = \frac{\hat{b}}{s_{\hat{b}}}$$

- Menguji pengaruh setiap variabel penjelas satu persatu.

- Dilakukan jika uji simultan menyatakan tolak H_0 (signifikan)

Kebaikan model regresi

- Dilihat dari nilai koefisien determinasi (R^2) → merupakan ukuran seberapa besar keragaman dari peubah respon (y) dapat dijelaskan oleh model (peubah penjelas (x))

- Nilainya antara 0-100%, semakin mendekati 100% maka semakin bagus

$$R^2 = \frac{SSR}{SST} \qquad R^2_{adj} = 1 - \frac{SSE/dfe}{SST/dft} = 1 - \frac{MSE}{MST}$$

Ilustrasi

- Model
  - Variabel Respon: Harga (harga jual rumah dalam satuan juta rupiah)
  - Variabel Penjelas:
    - Luasbangunan (luas bangunan rumah dalam satuan meter persegi)
    - Umur (umur bangunan dalam satuan tahun)
    - Kamarmandi (banyaknya kamar mandi di dalam rumah)

BACA DATA

- Lakukan dulu import file RUMAH.CSV, menjadi SAS dataset dengan nama “rumah”

```
*isi dataset rumah;
proc contents data=rumah;
run;
```

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
1	ID	Num	8	BEST12.	BEST32.
7	dekattol	Num	8	BEST12.	BEST32.
8	harga	Num	8	BEST12.	BEST32.
6	kamarmandi	Num	8	BEST12.	BEST32.
5	kamartidur	Num	8	BEST12.	BEST32.
2	luasbangunan	Num	8	BEST12.	BEST32.
3	luastanah	Num	8	BEST12.	BEST32.
4	umur	Num	8	BEST12.	BEST32.

PROC PRINT

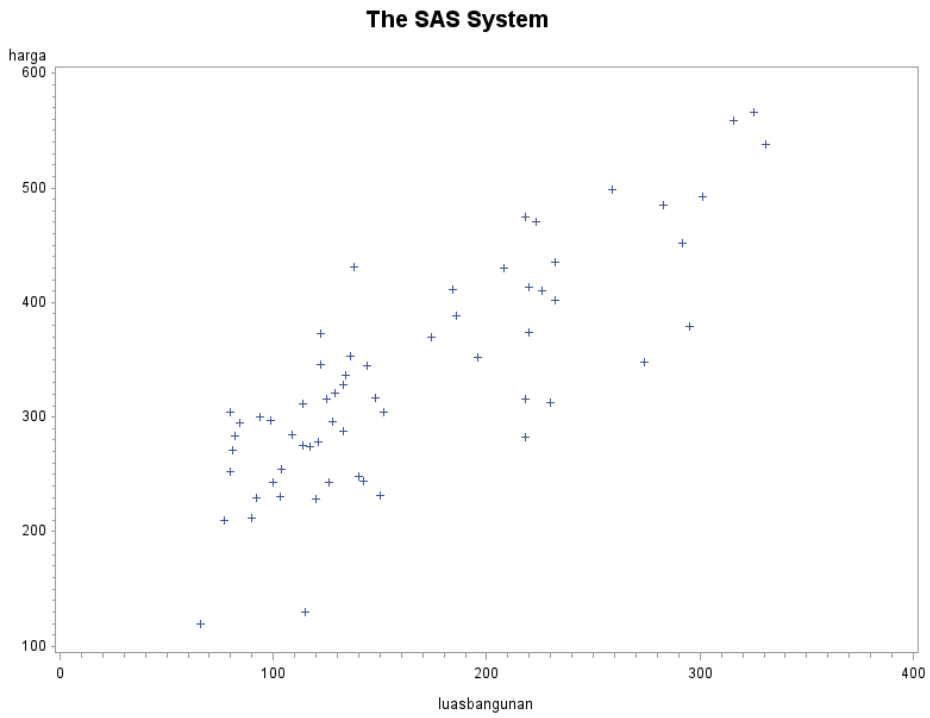
```
*mencetak 3 baris pertama dataset rumah;
proc print data=rumah (obs=3);
run;
```

The SAS System

Obs	ID	luasbangunan	luastanah	umur	kamartidur	kamarmandi	dekattol	harga
1	1	109	121	13	3	2	0	285
2	2	90	108	20	3	2	0	212
3	3	325	362	26	5	4	0	566

PROC GPLOT

```
*scatter plot antara luas bangunan dengan harga rumah;
proc gplot data=rumah;
plot harga*luasbangunan;
run;
quit;
```



PROC CORR

```
*korelasi antara luas
bangunan dengan harga rumah;
proc corr data=rumah;
var harga;
with luasbangunan;
run;
```

The CORR Procedure

1 With Variables:	luasbangunan
1 Variables:	harga

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
luasbangunan	62	164.59677	72.27644	10205	66.00000	331.00000
harga	62	334.91935	97.07396	20765	119.00000	566.00000

Pearson Correlation Coefficients, N = 62
Prob > |r| under H0: Rho=0

	harga
luasbangunan	0.82104 <.0001

PROC REG

```
*model regresi dengan Y = harga
 prediktor: (1) luas bangunan,
            (2) umur bangunan,
            (3) banyaknya kamar mandi;

proc reg data=rumah;
model harga = luasbangunan umur kamarmandi;
run;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	472121	157374	88.87	<.0001
Error	58	102703	1770.74523		
Corrected Total	61	574825			

Root MSE	42.08022	R-Square	0.8213
Dependent Mean	334.91935	Adj R-Sq	0.8121
Coeff Var	12.56429		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	239.24918	19.12089	12.51	<.0001
luasbangunan	1	1.14007	0.10456	10.90	<.0001
umur	1	-3.85913	0.55870	-6.91	<.0001
kamarmandi	1	1.13656	6.77455	0.17	0.8673

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	239.24918	19.12089	12.51	<.0001
luasbangunan	1	1.14007	0.10456	10.90	<.0001
umur	1	-3.85913	0.55870	-6.91	<.0001
kamarmandi	1	1.13656	6.77455	0.17	0.8673

Dugaan Koefisien Model Regresi

Dugaan modelnya adalah
Harga = 239 + 1.14 luasbangn – 3.86 umur + 1.14 Kamarmandi

Koefisien luasbangunan: positif
Koefisien Umur: negatif
Koefisien Kamarmandi: positif

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	472121	157374	88.87	<.0001
Error	58	102703	1770.74523		
Corrected Total	61	574825			

Root MSE	42.08022	R-Sq
Dependent Mean	334.91935	Adj R
Coeff Var	12.56429	

Parameter Estimates					
Variable	DF	Parameter Estimate	Standar Erro		
Intercept	1	239.24918	19.1208		
luasbangunan	1	1.14007	0.1045		
umur	1	-3.85913	0.5587		
kamarmandi	1	1.13656	6.77455	0.17	0.8673

Uji Simultan

H_0: semua X tidak berpengaruh
H_1: ada X yang berpengaruh

p-value (Pr > F) bernilai sangat kecil, sehingga disimpulkan Tolak H_0

Lanjutkan ke pengujian parsial untuk melihat variabel penjelas mana saja yang pengaruhnya signifikan.

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	472121	157374	88.87
Error	58	102703	1770.74523	
Corrected Total	61	574825		

Root MSE	42.08022	R-Square	0.8213
Dependent Mean	334.91935	Adj R-Sq	0.8121
Coeff Var	12.56429		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	239.24918	19.12089	12.51	<.0001
luasbangunan	1	1.14007	0.10456	10.90	<.0001
umur	1	-3.85913	0.55870	-6.91	<.0001
kamarmandi	1	1.13656	6.77455	0.17	0.8673

<u>Uji Partial</u>

Untuk masing-masing X dilakukan pengujian
H_0: X tidak berpengaruh
H_1: X berpengaruh

p-value (Pr > |t|) bernilai kecil, → pengaruh X signifikan (Tolak H_0)

Luasbangunan: signifikan
Umur : signifikan
Kamarmandi: tidak signifikan

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	472121	157374	88.87	<.0001
Error	58	102703	1770.74523		
Corrected Total	61	574825			

Root MSE	42.08022	R-Square	0.8213
Dependent Mean	334.91935	Adj R-Sq	0.8121
Coeff Var	12.56429		

Kebaikan Model

$R^2 = 0.8213$

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	239.24918	19.12089	12.51	<.0001
luasbangunan	1	1.14007	0.10456	10.90	<.0001
umur	1	-3.85913	0.55870	-6.91	<.0001
kamarmandi	1	1.13656	6.77455	0.17	0.8673

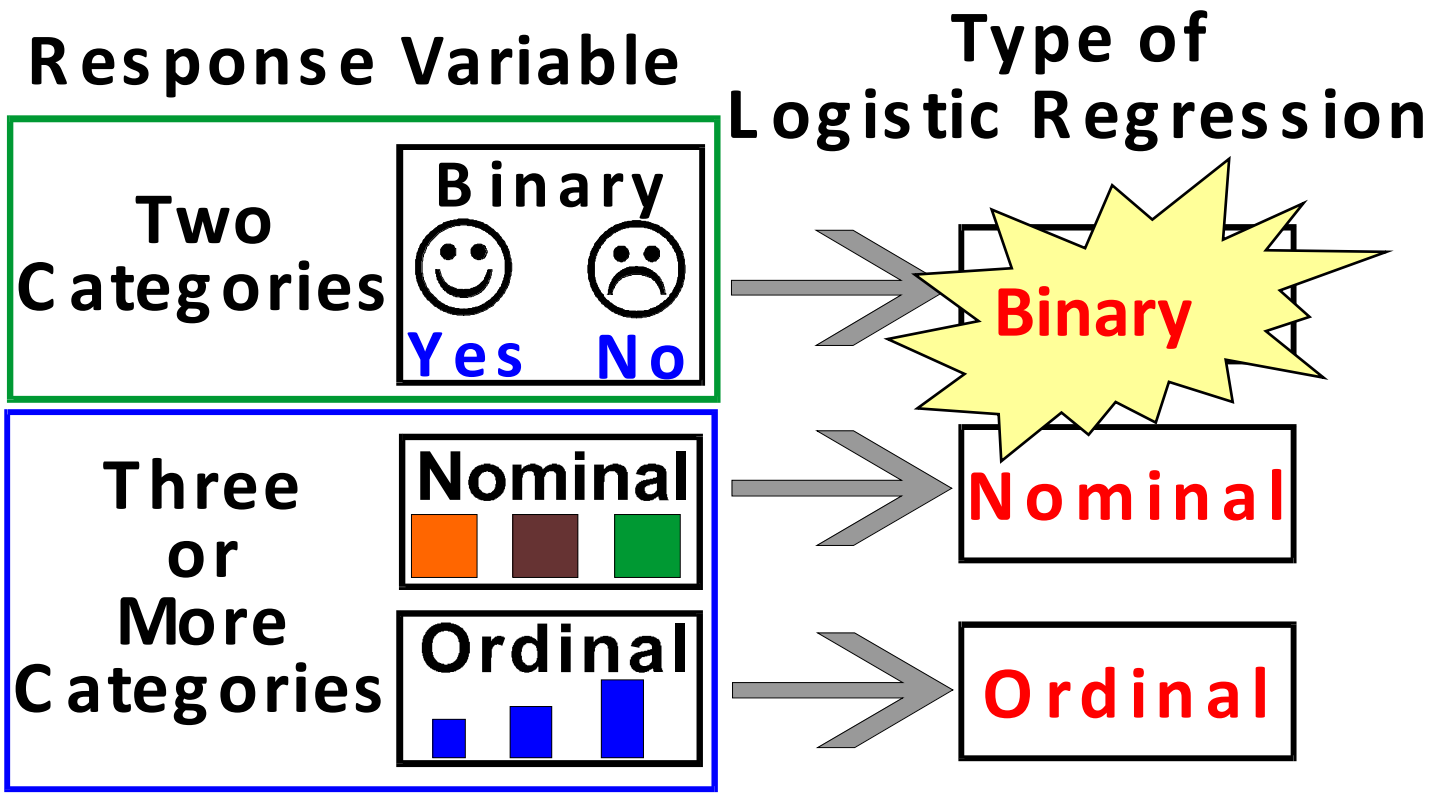
bagian 4

Pemodelan Regresi Logistik

Objectives

- Explain the concepts of logistic regression.
- Fit a binary logistic regression model.
- Fit a binary logistic regression model with interactions.

Tipe dalam Regresi Logistik



Apa yang dikerjakan Regresi Logistik

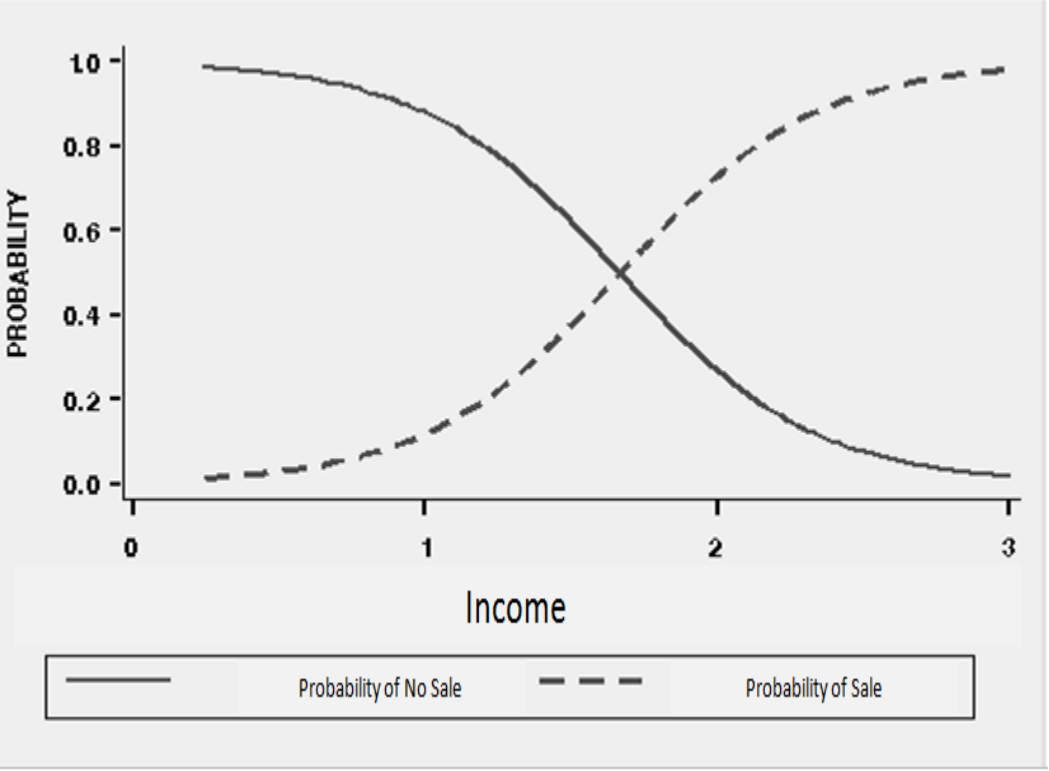
•Regresi Logistik menggunakan variabel prediktor (yang bersifat numerik ataupun kategorik) untuk memprediksi kejadian tertentu.

•Dengan kata lain, regresi logistik didesain untuk memberikan informasi mengenai peluang (probability) terjadinya suatu nilai dari variabel target.

•Tidak seperti dalam model regresi linear, yang langsung bisa diperoleh nilai dugaan Y karena bentuk modelnya adalah Y fungsi dari variabel-variabel penjelas, pada model regresi logistik yang dimodelkan adalah nilai peluang terjadinya kategori tertentu.

•Dalam propensity model, model ini digunakan untuk menduga peluang seseorang tertarik terhadap suatu penawaran tertentu.

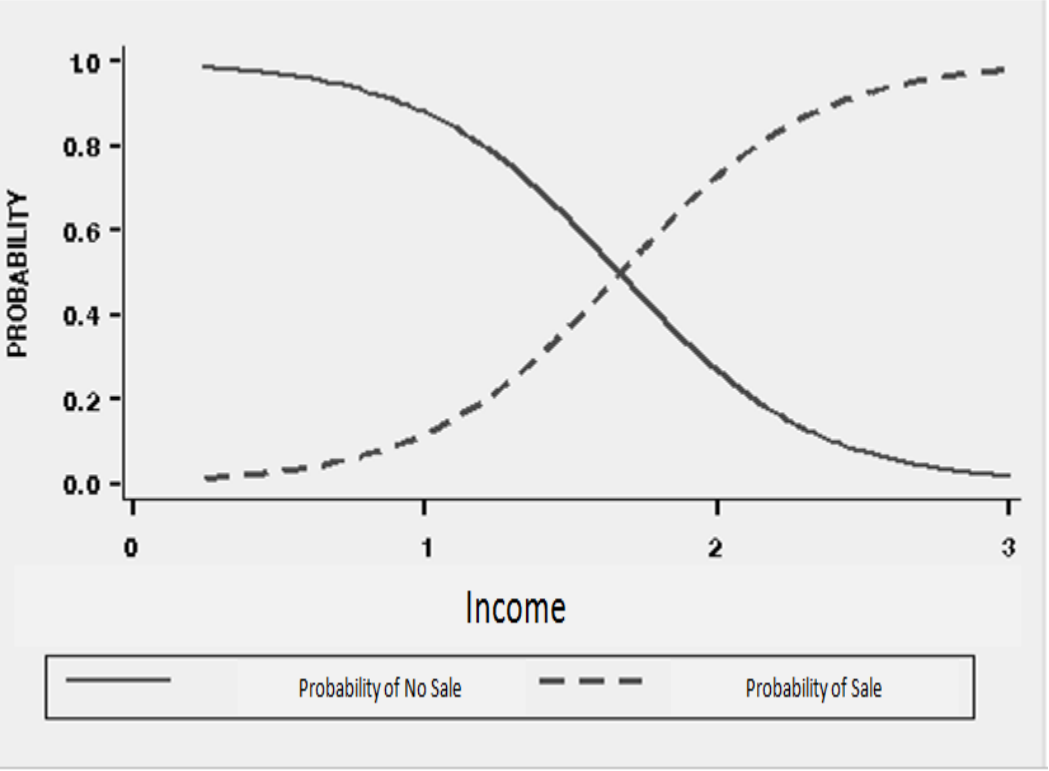
Kurva Regresi Logistik



This graph shows the relationship between the probability of SALE to INCOME.

Umumnya bentuk hubungan antara besarnya variabel prediktor X dengan besarnya peluang suatu kejadian merupakan kurva yang berbentu S (S-shaped curve) seperti yang digambarkan di atas.

Kurva Regresi Logistik

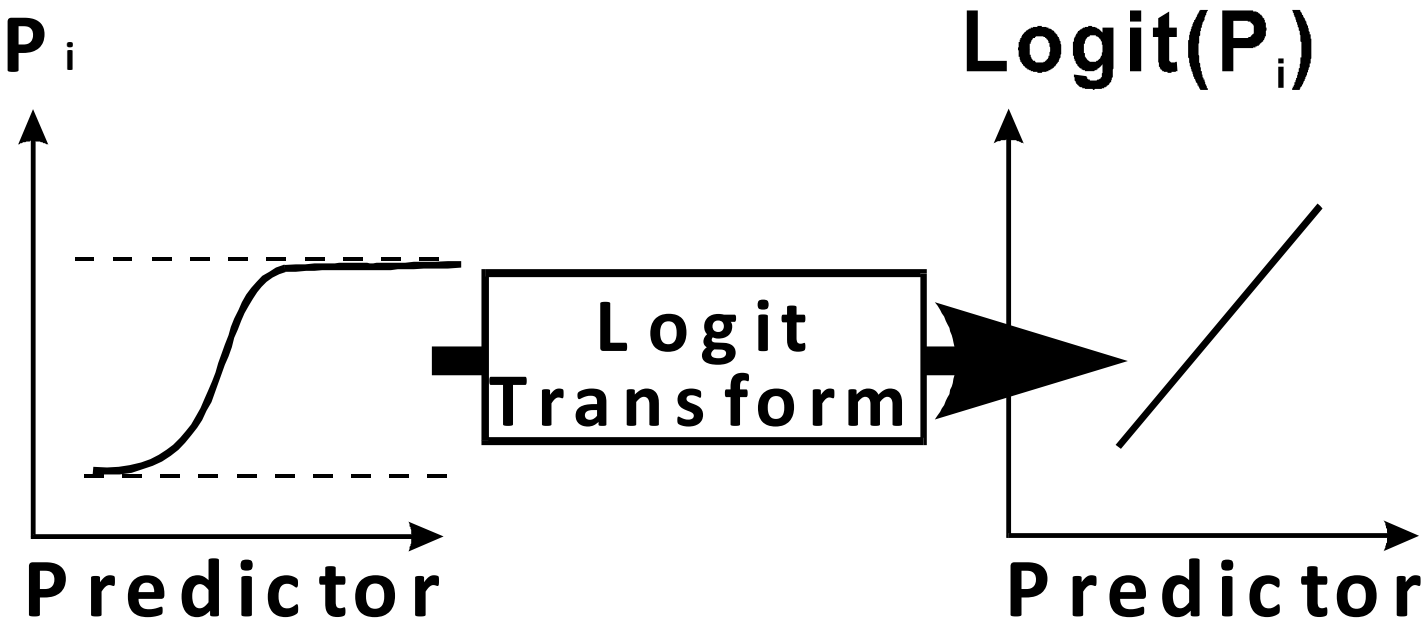


- Jika hubungannya **positif** (garis putus-putus pada gambar) akan ditunjukkan dengan meningkatnya nilai peluang jika nilai X semakin tinggi.
- Pada gambar di samping diilustrasikan untuk individu dengan income yang semakin tinggi, maka peluang terjadinya penjualan semakin besar.

Karena hanya ada dua kategori: SALE dan NO-SALE, maka untuk P(Y=SALE) menurun diikuti secara otomatis dengan P(Y=NOSALE) yang meningkat. Dan total keduanya adalah 1 (satu).

.

Asumsi



- Secara matematis, pemodelan hubungan dalam bentuk S-curve lebih sulit dibandingkan dengan model yang berbentuk linear.
- Untuk mengatasi hal tersebut, dalam teknik komputasinya dilakukan transformasi agar diperoleh bentuk linear. Selain mudah dalam hal penghitungan, bentuk linear juga umumnya lebih mudah dalam hal interpretasi model yang diperoleh.
- Salah satu bentuk transformasi yang dapat digunakan untuk hal yang dijelaskan di atas adalah transformasi logit.

Logit Transformation

Logistic regression models transformed probabilities called logits.

$$\mathbf{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right)$$

where

i indexes all cases (observations).

p_i is the probability the event (a sale, for example) occurs in the i^{th} case.

log is the natural log (to the base e).

Andaikan p adalah nilai peluang kejadian kategory tertentu. Dalam model, nilai p adalah P(nasabah tertarik pada penawaran), P(nasabah default), P(terkena penyakit), dan sebagainya.

Nilai p / (1 – p) yang merupakan rasio antara peluang sesuatu terjadi dengan peluang tidak terjadi disebut sebagai ODD.

Transformasi logit, adalah logaritma natural dari nilai odd.

Model Regresi Logistik

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_1$$

where

$\text{logit}(p_i)$	logit transformation of the probability of the event
β_0	intercept of the regression line
β_1	slope of the regression line.

- Model regresi logistik adalah model linear antara logit(p) dengan variabel penjelas X. Seperti halnya dalam regresi linear, kita bisa mendapatkan nilai-nilai intersep dan slope dari model tersebut.

- Berbeda dengan regresi linear yang menggunakan metode kuadrat terkecil (least squares method) dalam menentukan b_0 dan b_1, model regresi logistik menggunakan metode iteratively reweighted least squares (IRLS) dan metode maximum likelihood (ML).

Model Regresi Logistik

$$P(Y=1) = p = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}$$

Perhatikan bahwa

Log(p/(1-p)) = $b_0 + b_1X$

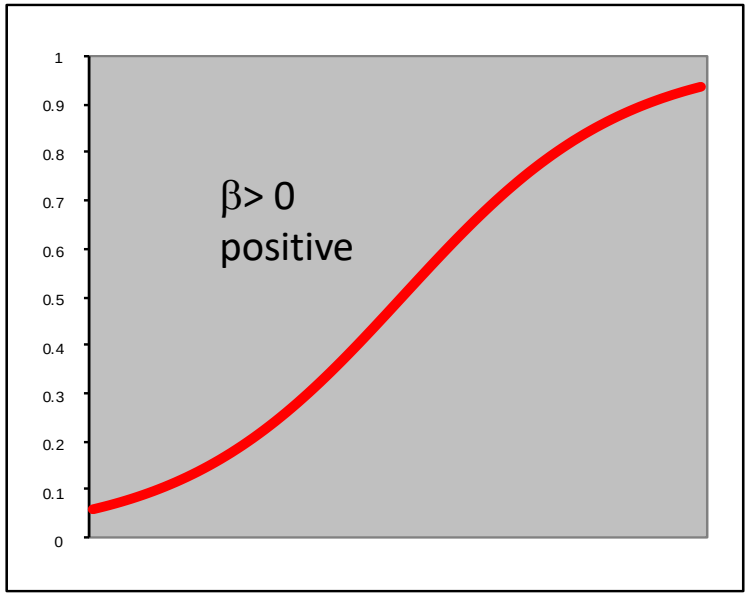
p/(1-p) = exp($b_0 + b_1X$)

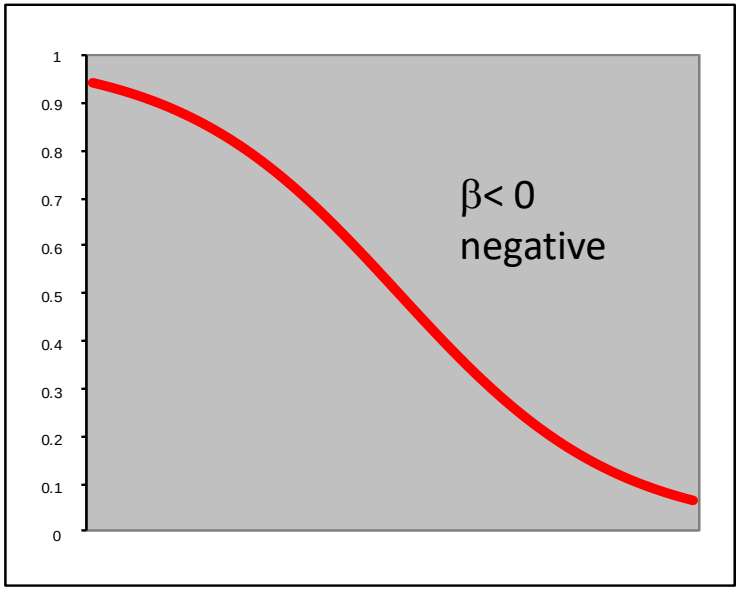
p = exp($b_0 + b_1X$) – p exp($b_0 + b_1X$)

p (1 + exp($b_0 + b_1X$)) = exp($b_0 + b_1X$)

Sehingga p = exp($b_0 + b_1X$) / (1 + exp($b_0 + b_1X$))

Model Regresi Logistik



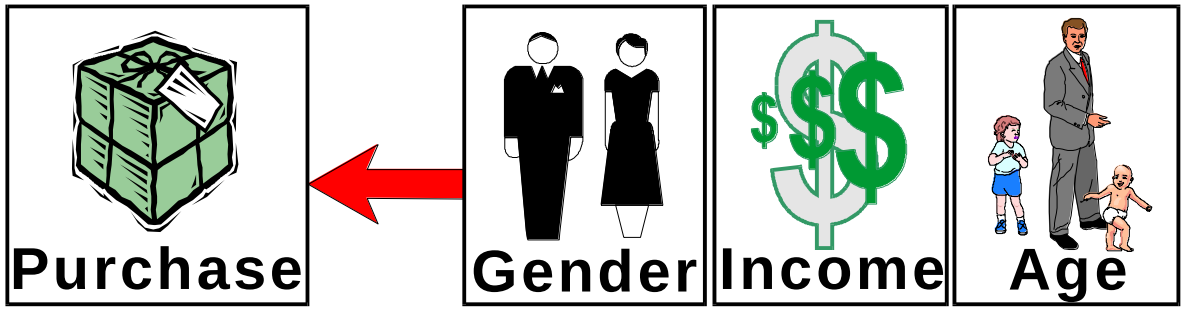


Berdasarkan model persamaan regresi logistik, bentuk kurva yang dihasilkan kemungkinan salah satu dari gambar di atas.

Jika model memiliki koefisien slope yang positif maka peluang suatu kejadian akan meningkat seiring dengan peningkatan nilai variabel penjelas.

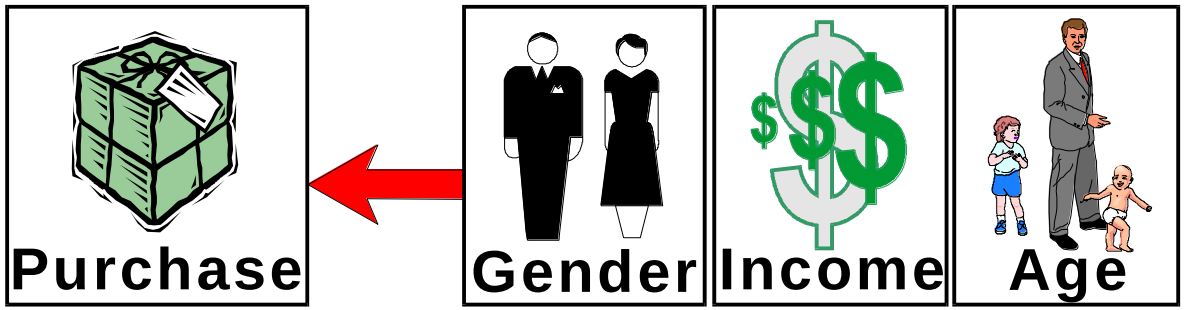
Sebaliknya jika koefisiennya negatif, peluang kejadiannya akan menurun untuk nilai variabel penjelas yang semakin tinggi.

Model Regresi Logistik



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

Model Regresi Logistik dengan Beberapa Prediktor



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

$$P(Y=1) = p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}}$$

penggunaan model regresi logistik untuk prediksi

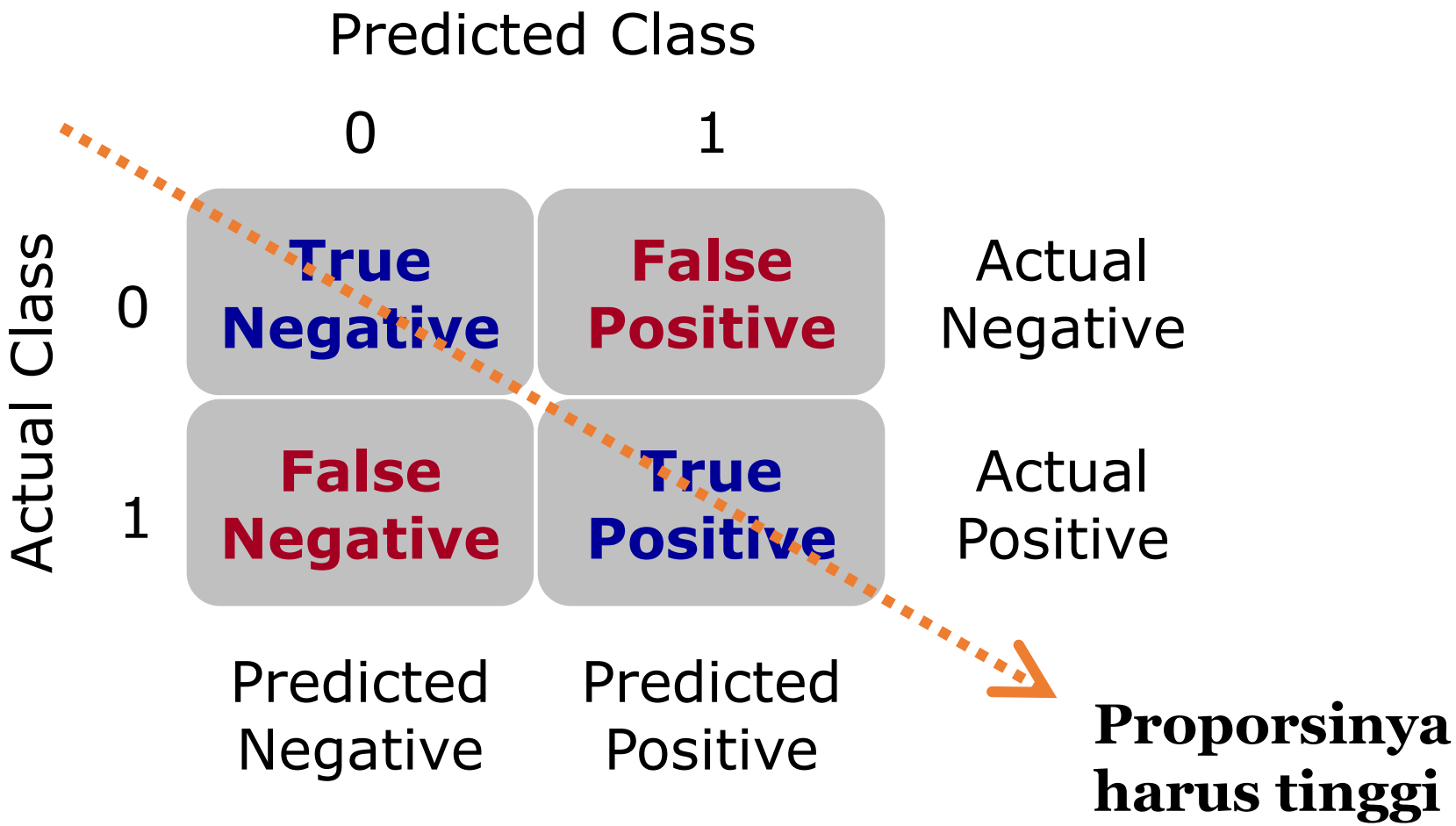
$$P(Y=1) = p = \frac{e^{-2+0.4*(gender=Lk)+0.7*pendpt-0.08*usia}}{1+e^{-2+0.4*(gender=Lk)+0.7*pendpt-0.08*usia}}$$

Gender	Laki-Laki	Laki-Laki	Perempuan
Pendapatan	6	8	7.5
Usia	26	28	50
P(membeli)	0.627	0.853	0.321

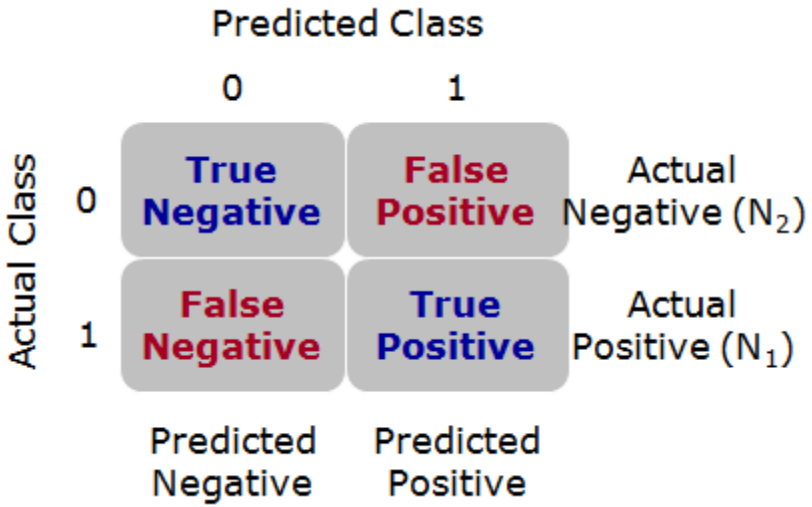
Penentuan Prediksi Kelas menggunakan Cut-Off Tertentu

		ACTUAL CLASS		PREDICTED CLASS	
Case	x	response	*P*	cut-off=.5	cut-off=.25
1		75.8	0		.32
	0			1	
2		68.3	1		.40
	0			1	
3		14.1	1		.92
	1			1	
4		99.2	0		.06
	0			0	
5		65.4	1		.52
	1			1	
6		68.7	1		.39
	0			1	
7		76.7	1		.22
	0			0	

Ukuran Kebaikan/Keseuaian Prediksi



Ukuran Kebaikan/Keseuaian Prediksi

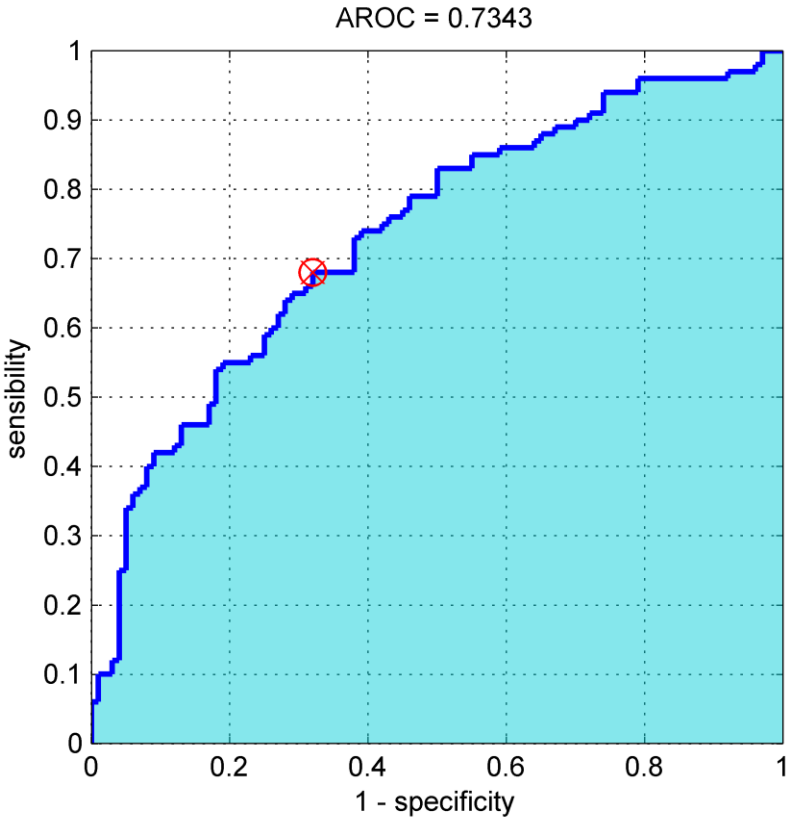


Accuracy = (TP + TN) / #N

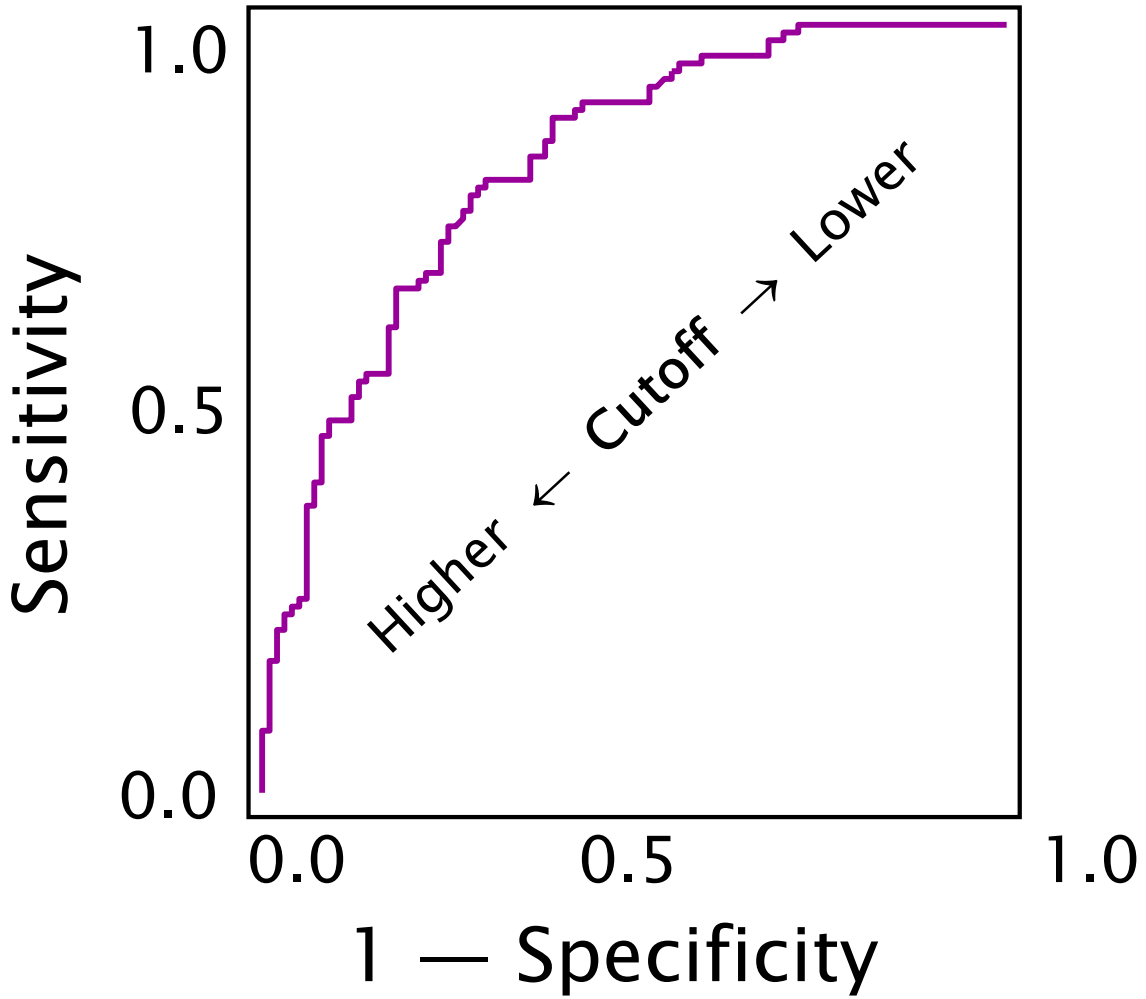
Sensitivity = TP/N1

Specificity = TN/N2

ROC (receiver operating characteristic) Curve



ROC (receiver operating characterintic) Curve



ROC Curve

- An ideal ROC curve is a vertical line from the origin to a sensitivity of 1, and then a horizontal line along sensitivity of 1 for all level of (1 – specificity)
- ROC curve is usually used to compare competing model
- A numerical measure of how close the ROC curve match the ideal curve is computed by comparing the area under the curve to 1. It turns out the area is equal to the c statistic.

Ilustrasi

- Respon: external rating (OK, NO)
- Prediktor:
  - return on equity
  - return on asset
  - cost to income ratio

PROC LOGISTIC

```
********* regresi logistik *****************;
libname a 'D:/';
proc logistic data = a.rating outmodel=modelrating;
model eksternal_rating (event = "OK") =
      Return_on_equity
      Return_on_Asset
      Cost_to_Income_Ratio;
run;
```

OUTPUT : PROC LOGISTIC

Response Profile		
Ordered Value	Eksternal_rating	Total Frequency
1	NO	81
2	OK	29

Probability modeled is Eksternal_rating='OK'.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.9957	1.8408	2.6484	0.1037
Return_on_equity	1	10.4903	4.6121	5.1735	0.0229
Return_on_Asset	1	199.7	71.9641	7.7005	0.0055
Cost_to_Income_Ratio	1	-4.9575	2.9008	2.9208	0.0874

OUTPUT : PROC LOGISTIC

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.9957	1.8408	2.6484	0.1037
Return_on_equity	1	10.4903	4.6121	5.1735	0.0229
Return_on_Asset	1	199.7	71.9641	7.7005	0.0055
Cost_to_Income_Ratio	1	-4.9575	2.9008	2.9208	0.0874

Koefisien Return on Equaity: POSITIF → semakin tinggi RoE, perusahaan cenderung memiliki rating OK

Koefisien Return on Asset: POSITIF → semakin tinggi RoA, perusahaan cenderung memiliki rating OK

Koefisien Cost to Income Ratio: NEGATIF → semakin tinggi cost-to-income ratio, perusahaan cenderung memiliki rating NO

OUTPUT : PROC LOGISTIC

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.9957	1.8408	2.6484	0.1037
Return_on_equity	1	10.4903	4.6121	5.1735	0.0229
Return_on_Asset	1	199.7	71.9641	7.7005	0.0055
Cost_to_Income_Ratio	1	-4.9575	2.9008	2.9208	0.0874

Bentuk Model:

$$P(rating = OK) = \frac{e^{-2.99+10.49ROE+199.7ROA-4.96CIR}}{1 + e^{-2.99+10.49ROE+199.7ROA-4.96CIR}}$$

PROC LOGISTIC

```
*ingin dilakukan prediksi terhadap dua perusahaan;
*Perusahaan #1: RoE = 0.1, RoA=0.02, CIR=0.8;
*Perusahaan #2: RoE = 0.2, RoA=0.02, CIR=0.6;
data maudiprediksi;
input Return_on_equity Return_on_Asset Cost_to_Income_Ratio;
cards;
0.1             0.02   0.8
0.2             0.02   0.6
;
proc logistic inmodel=modelrating;
score data=maudiprediksi out=prediksi;
run;
proc print data=prediksi;
run;
```

The SAS System

Obs	Return_on_equity	Return_on_Asset	Cost_to_Income_Ratio	I_Eksternal_rating	P_NO	P_OK
1	0.1	0.02	0.8	NO	0.87199	0.12801
2	0.2	0.02	0.6	OK	0.46957	0.53043

PROC LOGISTIC

```
*menentukan tingkat ketepatan/akurasi model;
*tahap 1: lakukan pendugaan terhadap keseluruhan data;
proc logistic inmodel=modelrating;
score data=a.rating out=prediksirating;
run;

*tahap 2: bandingkan rating asli dengan prediksi;
proc tabulate data=prediksirating;
class eksternal_rating i_eksternal_rating;
table eksternal_rating, i_eksternal_rating*n;
table eksternal_rating, i_eksternal_rating*pctn;
run;
```

	Into: Eksternal_rating	
	NO	OK
	N	N
Eksternal_rating		
NO	78	3
OK	10	19

	Into: Eksternal_rating	
	NO	OK
	PctN	PctN
Eksternal_rating		
NO	70.91	2.73
OK	9.09	17.27

AKURASI = 70.91% + 17.27% = 88.18%

LATIHAN

- Gunakan data LATIHANSKORING
- Buat model regresi logistik dengan
  - Y = status
  - X = DBR, gender, pekerjaan
- Prediksikan status kredit seseorang dengan karakteristik:
  - DBR = 25%
  - Gender = wanita
  - Pekerjaan = PNS

The Essential of Credit Scoring Model

Yang Harus Dikuasai dalam Pembuatan Model Skoring

Bagus Sartono

2019

Pengantar Mengenai Model Credit Scoring dan Menilai Kebaikan Dari Suatu Model

Apa itu model skoring kredit?

- Model statistika yang berguna dalam menghasilkan skor sebagai bahan untuk mengambil keputusan mengenai kategori resiko kredit (calon) nasabah, baik perorangan maupun perusahaan.

- Kategori resiko ini hanya merupakan dugaan, sehingga ada kemungkinan bahwa keputusan yang diambil adalah salah.

- Perlu proses yang baik untuk menyusun model skoring agar tingkat kesalahan itu minimum.

Ilustrasi: Scorecard dan Threshold

Attribute	Category	Score
Gender	FEMALE	165
	MALE	107
Age Group	<= 25	108
	25 – 30	123
	31 – 35	121
	36 – 40	133
	41 – 45	128
	> 45	186
Residence Ownership	OTHERS	98
	OWNED	172
	PARENT	115
	RENT	83
Number of Dependants	0	169
	1	140
	2	136
	3	122
	4	89
	> 4	92

Score	Odds (good)
640	200.0
620	100.0
600	50.0
580	25.0
560	12.5
540	6.3
520	3.1
500	1.6
480	0.8
460	0.4
440	0.2

“Accepted (Good) jika score lebih dari 540”

Bagaimana menggunakan scorecard?

Attribute	Category	Score
Gender	FEMALE	165
	MALE	107
Age Group	<= 25	108
	25 – 30	123
	31 – 35	121
	36 – 40	133
	41 – 45	128
	> 45	186
Residence Ownership	OTHERS	98
	OWNED	172
	PARENT	115
	RENT	83
Number of Dependants	0	169
	1	140
	2	136
	3	122
	4	89
	> 4	92

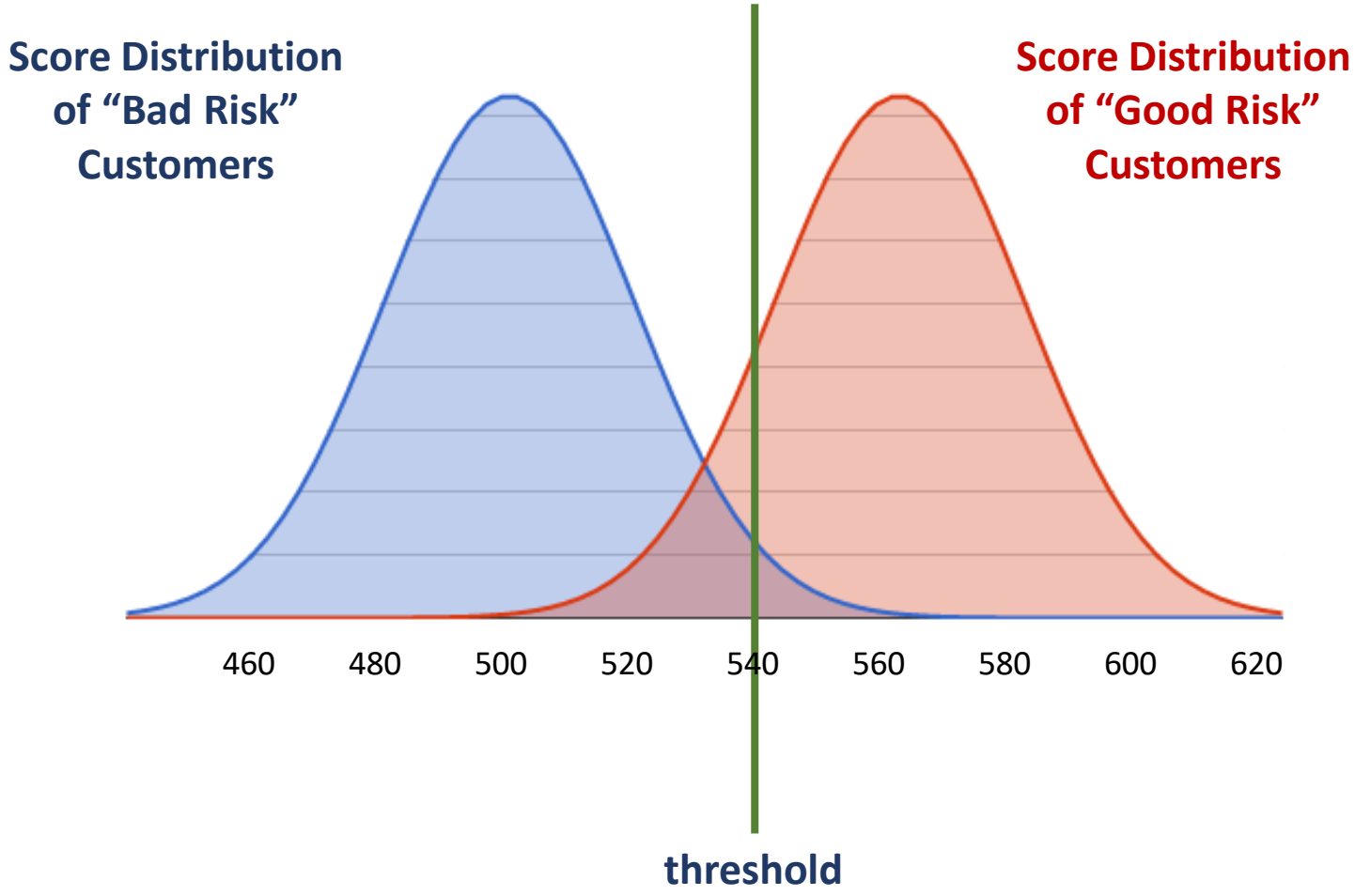
Female	165
37 yr old	133
rent house	83
1 dependant	140
Total Score	521
Decision	REJECTED

Apakah itu scorecard yang bagus?

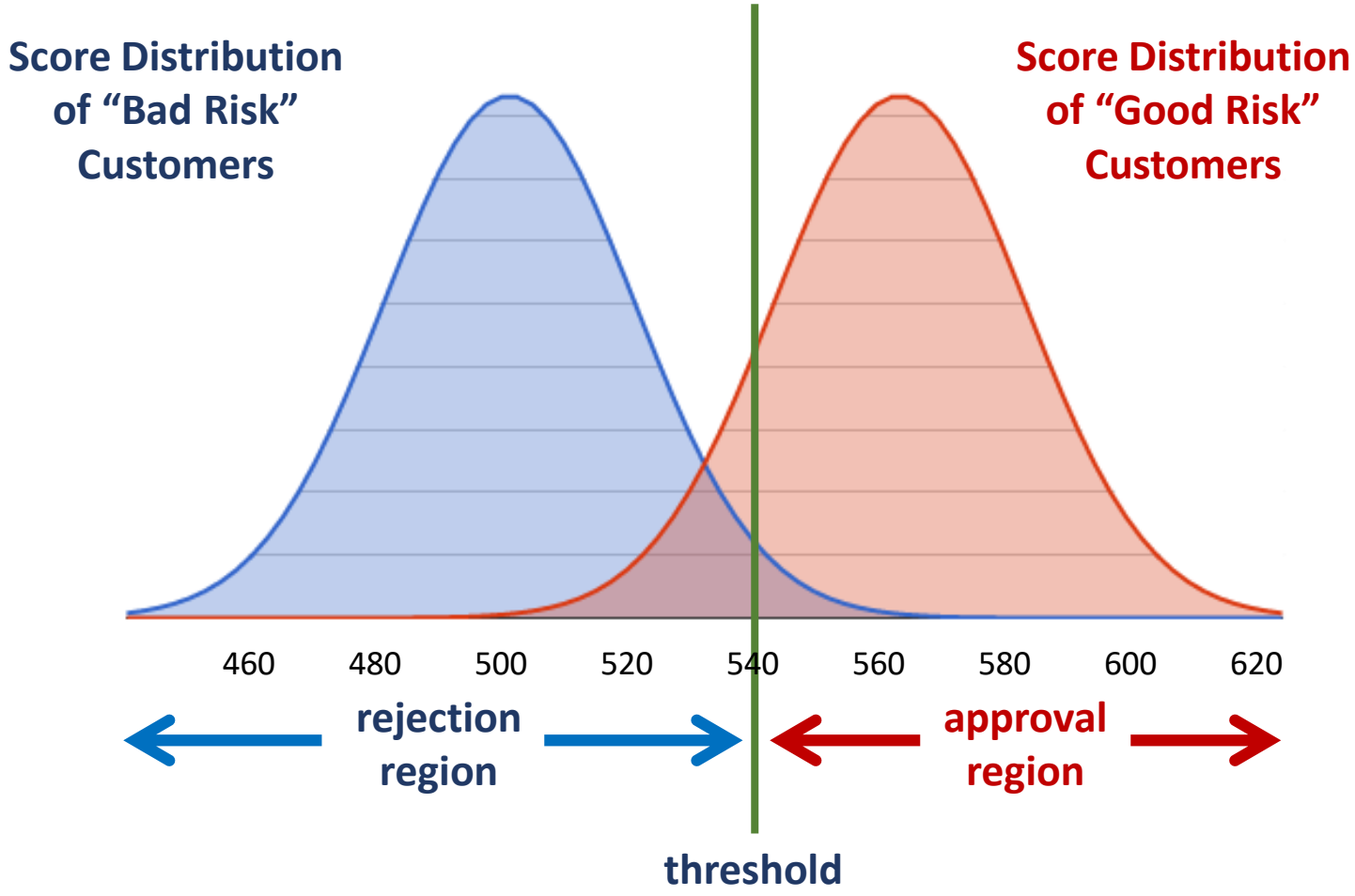
Memberikan skor berbeda antara nasabah GOOD dan nasabah BAD

Memberikan padanan skor dan resiko sesuai dengan rancangan pembuatannya

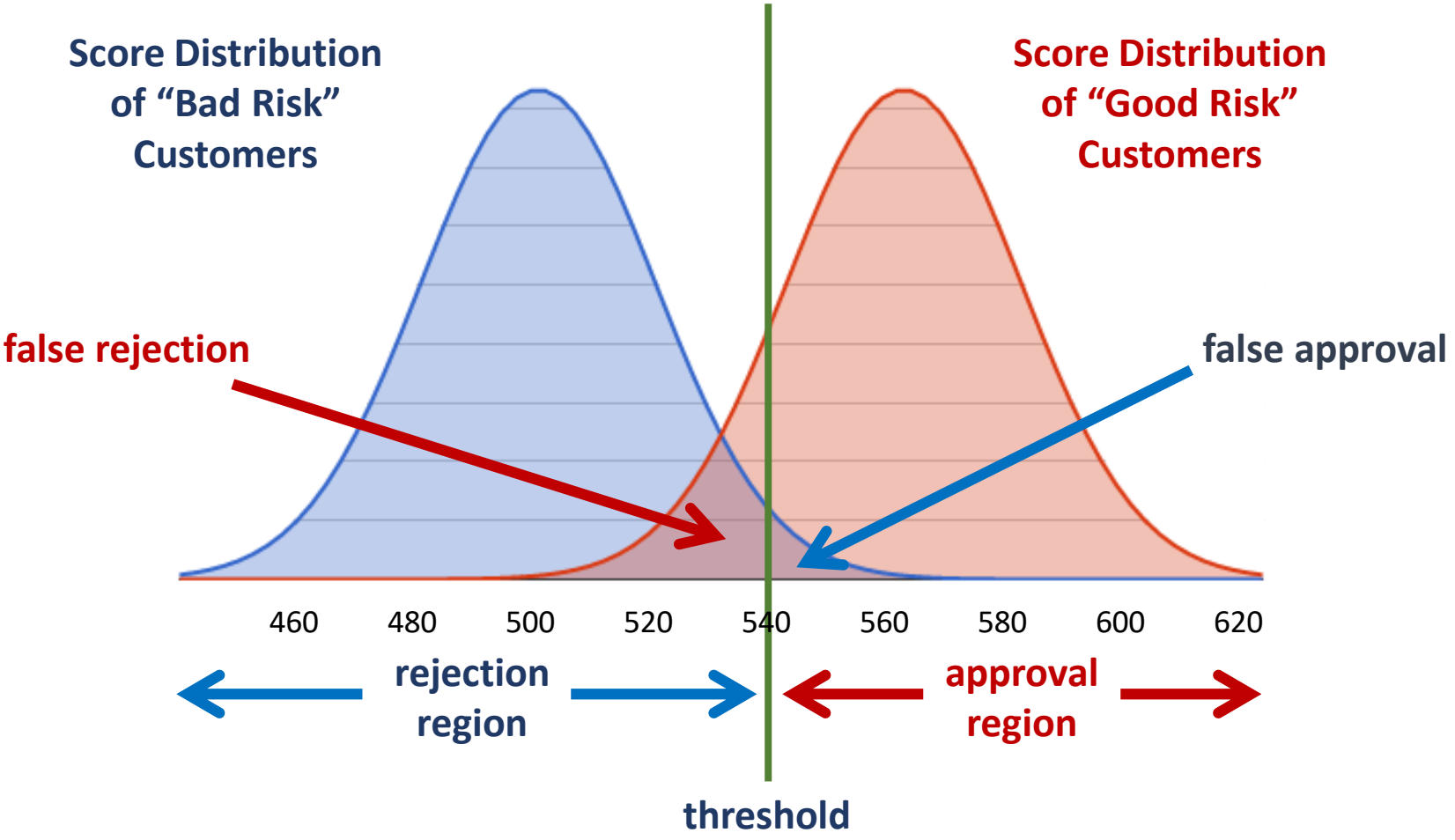
Apakah itu scorecard yang bagus?



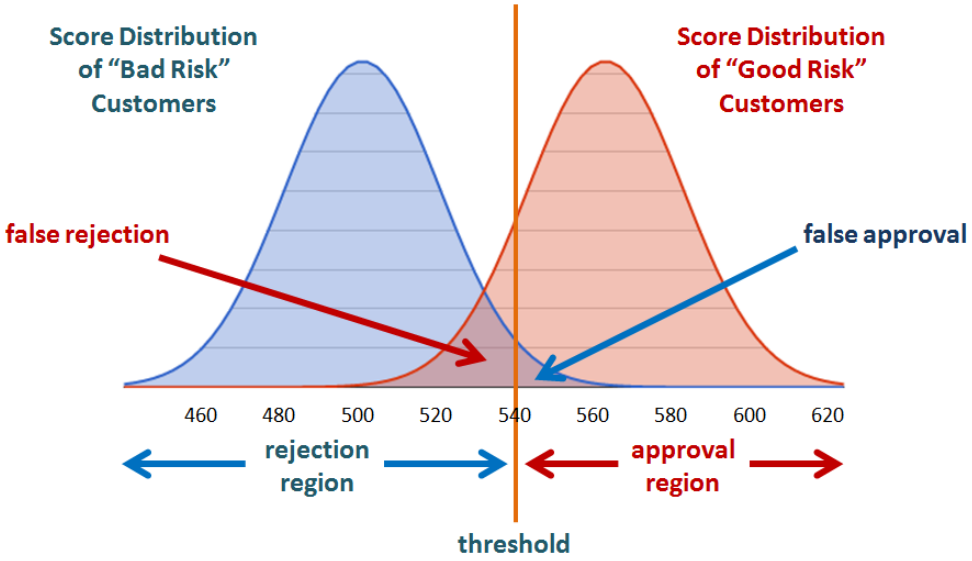
Apakah itu scorecard yang bagus?



Apakah itu scorecard yang bagus?



Apakah itu scorecard yang bagus?



Scorecard yang bagus memiliki false rate yang rendah:

-False approval rate

-False rejection rate

-Total false rate

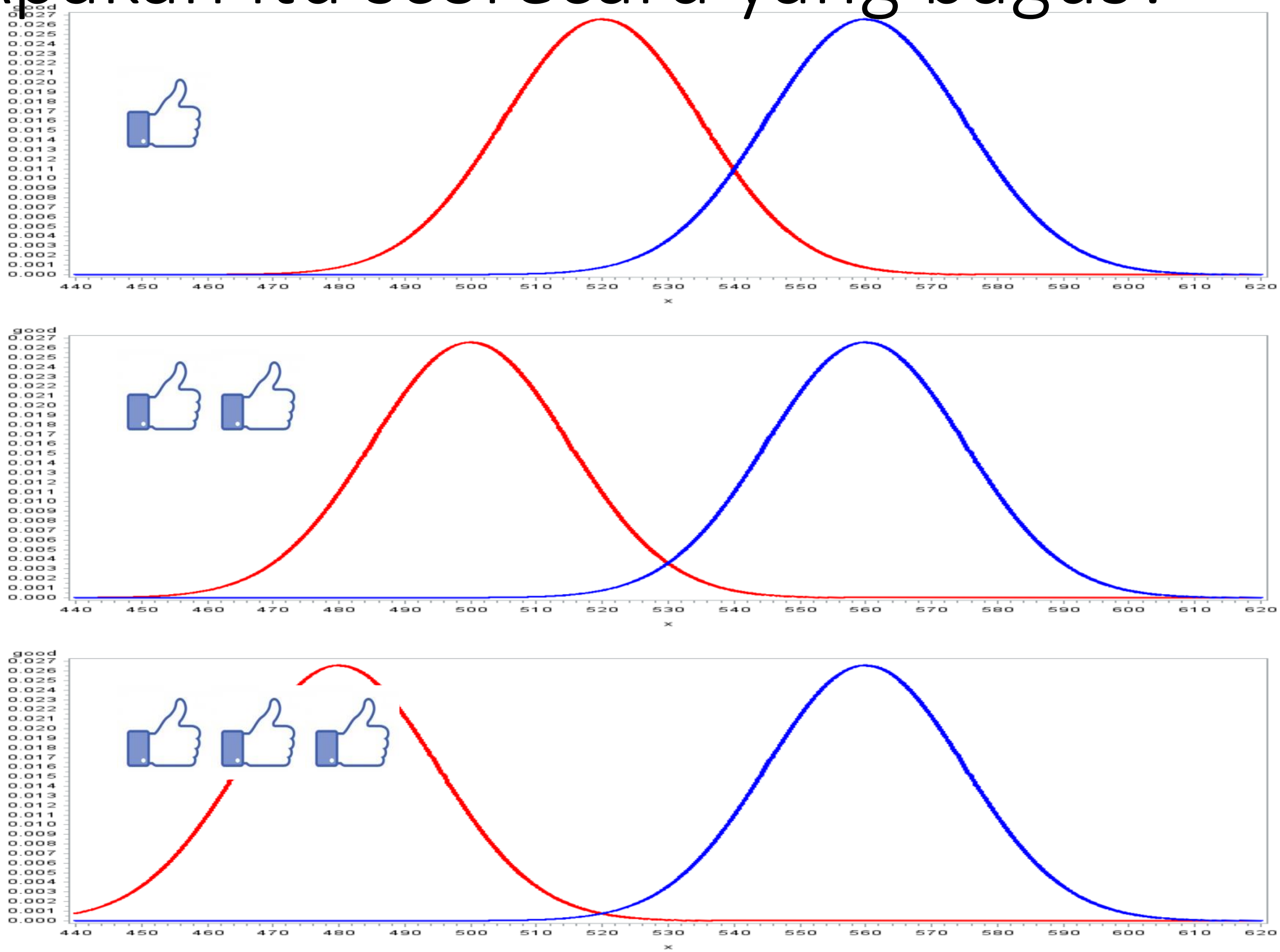
catatan:

1 – False approval rate = sensitivity

1 – False rejection rate = specificity

1 – Total false rate = accuracy

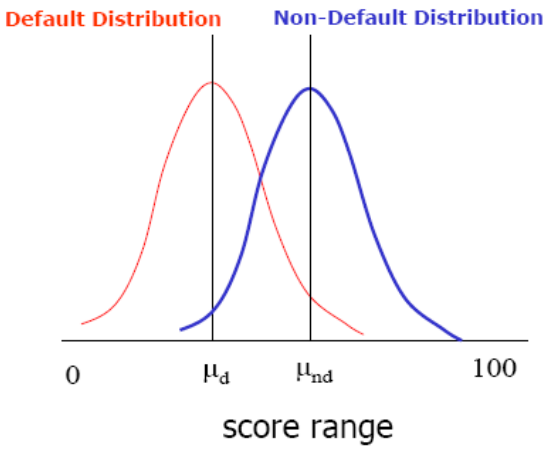
Apakah itu scorecard yang bagus?



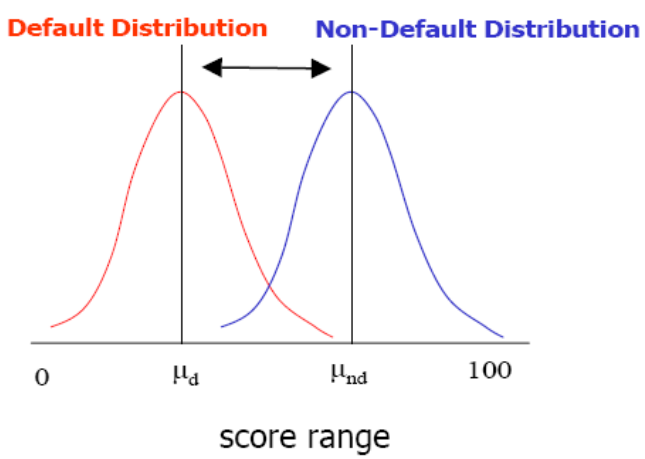
Mengevaluasi Model Skoring

- Kolmogorov-Smirnov (KS)
- Melihat apakah model mampu menghasilkan skor yang dapat membedakan Bad-Good

Model 1



Model 2



- Empirical Distribution Function based test

K-S Statistic

The *empirical distribution function* (EDF) of a sample is defined as the following function:

$$F(x) \quad = \quad \frac{1}{n} \left(\textbf{number of } x_j \leq x \right)$$

If there are two class levels, two-sample Kolmogorov-Smirnov test statistic D as

$$D = \max_j \left| F_1(x_j) - F_2(x_j) \right| \quad \textbf{where } j = 1, 2, \ldots, n$$

Model Assessment using K-S Statistic

Band	#Bad	#Good	%Bad	%Good	Cum% Bad	Cum% Good	Diff
1							
2							
3							
…							
…							
k							

1 2 3 4

5

KS statistic = maximum diff

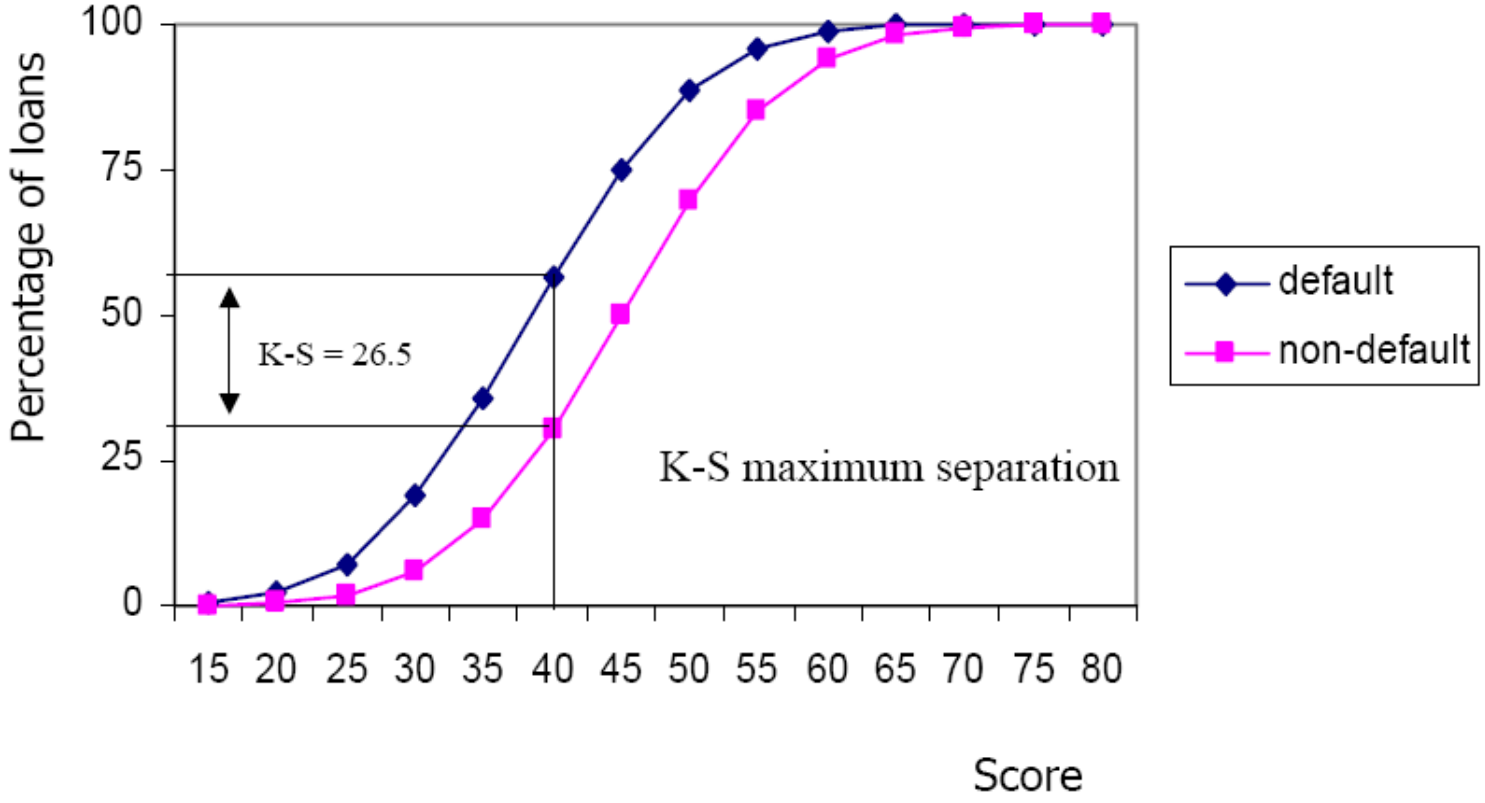
Model Assessment using K-S test

Obs (i)	Score Range lower	upper	Distributions Default (#)	Non-Def (#)	Cumulative Distributions Default (%)	Non-Def (%)	K-S
1	0	15	82	458	0.34	0.05	0.29
2	15	20	428	3205	2.12	0.37	1.75
3	20	25	1235	13886	7.24	1.75	5.49
4	25	30	2778	41657	18.77	5.92	12.85
5	30	35	4074	91645	35.69	15.09	20.60
6	35	40	5092	152741	56.82	30.36	**26.46**
7	40	45	4365	196381	74.94	50.00	24.94
8	45	50	3274	196381	88.53	69.64	18.89
9	50	55	1698	152741	95.58	84.91	10.67
10	55	60	764	91645	98.75	94.08	4.67
11	60	65	232	41657	99.71	98.24	1.47
12	65	70	58	13886	99.95	99.63	0.32
13	70	75	9	3205	99.99	99.95	0.04
14	75	80	1	458	100	100	0.00
15	80	100	1	31	100	100	0

(82+428)/24092

Total Bad = 24092

Kolmogorov-Smirnov test visualization



Population Stability Index

- Compares closeness of the predicted defaults to actual defaults by score
  - Exp % corresponds to the predicted default rate in a score band
  - Act % is the actual default rate in a score band
- Stability index =

$$100 * \sum_{i=1}^{NbBands} \left((Exp\%(i) - Act\%(i)) * \ln\left(\frac{Exp\%(i)}{Act\%(i)} \right) \right)$$

- As the index lower, the two population's characteristics become similar

Normal	Caution	Danger
< 10	10 – 25	> 25

Hands-On

- Ada data…. berisi nilai-nilai variabel prediktor dan status good/bad
- Berikan skor sesuai scorecard di atas
- Lihat perbedaan sebarannya antara nasabah good dan nasabah bad… hitung KS-nya
- Lihat peluang setiap band… bandingkan dengan rancangan yang ada pada scorecard

Pengenalan Pemodelan Regresi Logistik biner

Pemodelan

- Membangun miniatur dari dunia nyata
  - dinyatakan dalam satu atau beberapa fungsi matematis
- Menyederhanakan fenomenya nyata sehingga mudah memahami pola umum yang ada
  - memberikan penjelasan terhadap perubahan
  - memberikan penjelasan tentang perbedaan yang terjadi
  - menemukan faktor yang menyebabkan perubahan dan perbedaan

Komponen Model

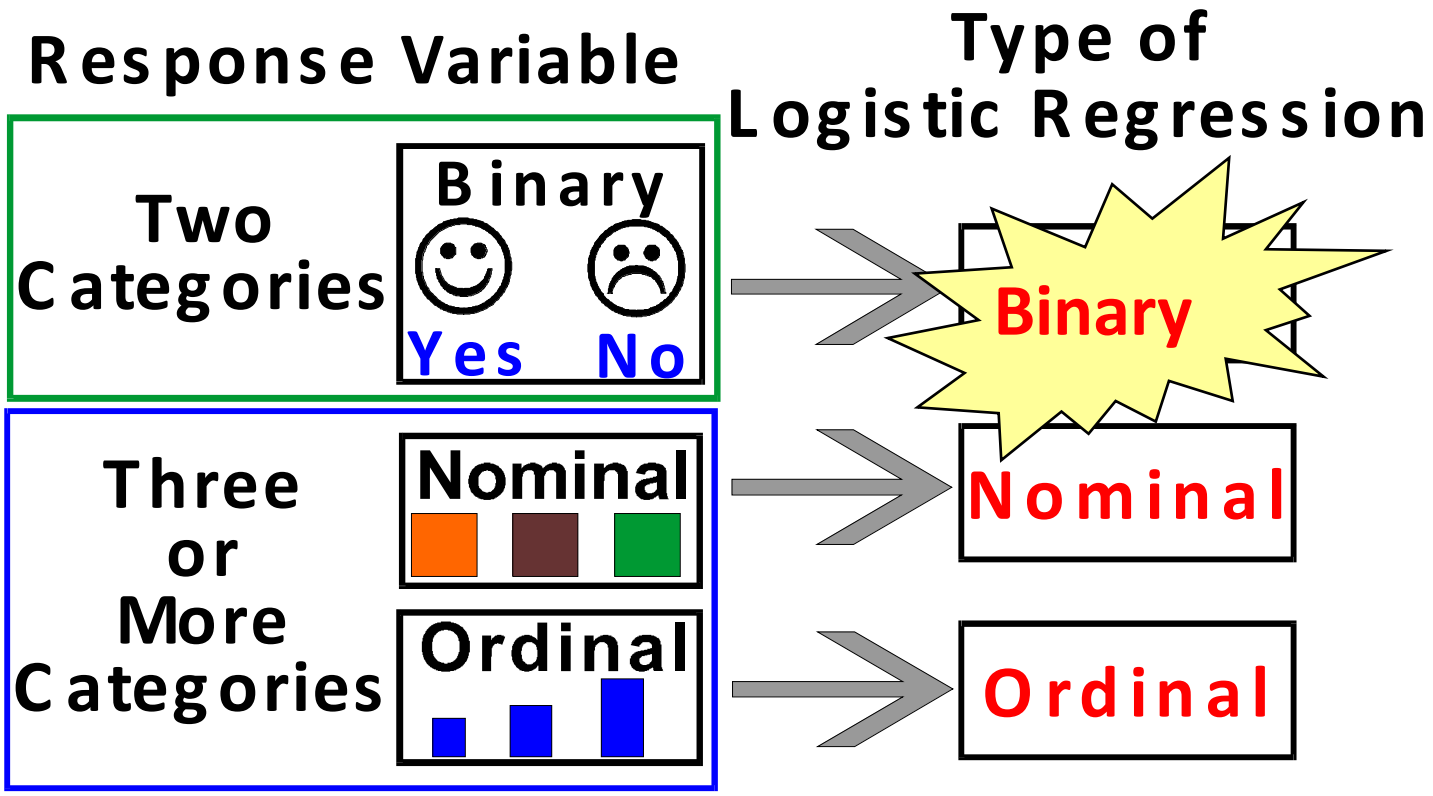
$$y = f(x_1, x_2, \cdots, x_p) + \varepsilon$$

Y	X
Output Target	Input
Respon	Penjelas (explanatory) Prediktor (predictor) Faktor (factors)
Dependent	Independent

Models

- Powerful predictors for optimizing performance
- Powerful summaries for understanding
- Used to explore data set
- Are not perfect
  - "All models are wrong, but some are useful"
  - "Statisticians, like artists, have the bad habit of falli in love with their models"

Types of Logistic Regression

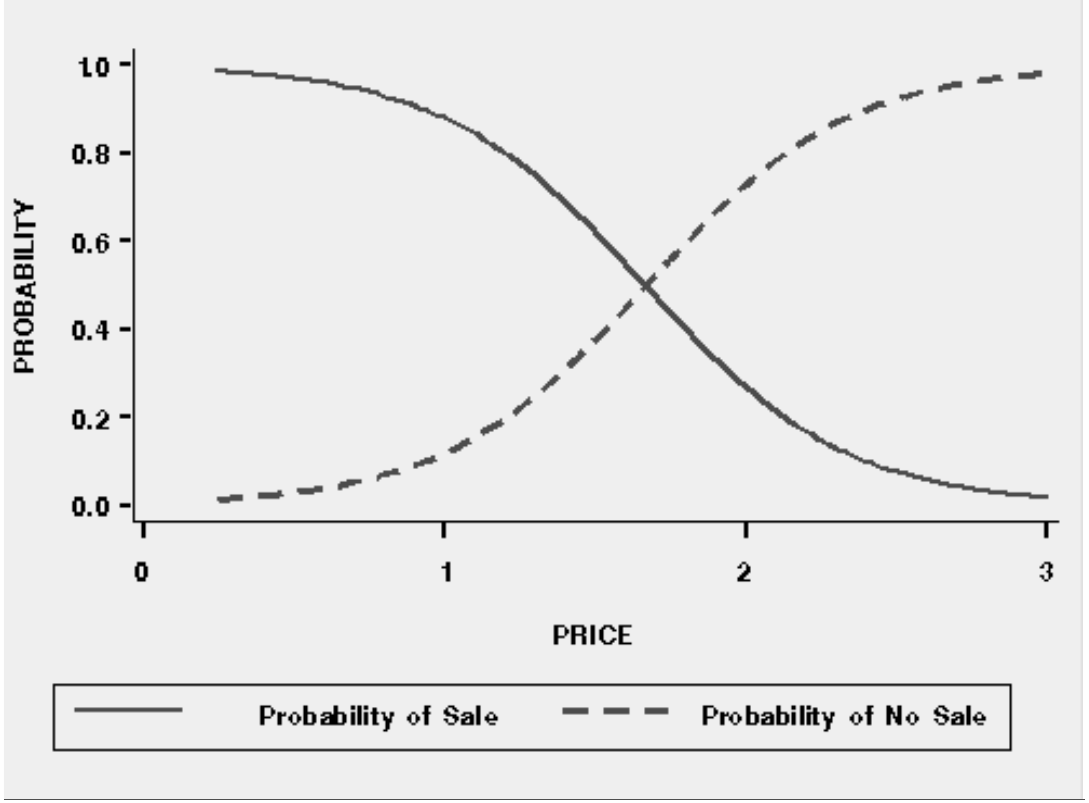


What Does Logistic Regression Do?

The logistic regression model uses the predictor variables, which can be **categorical or continuous**, to predict the probability of specific outcomes.

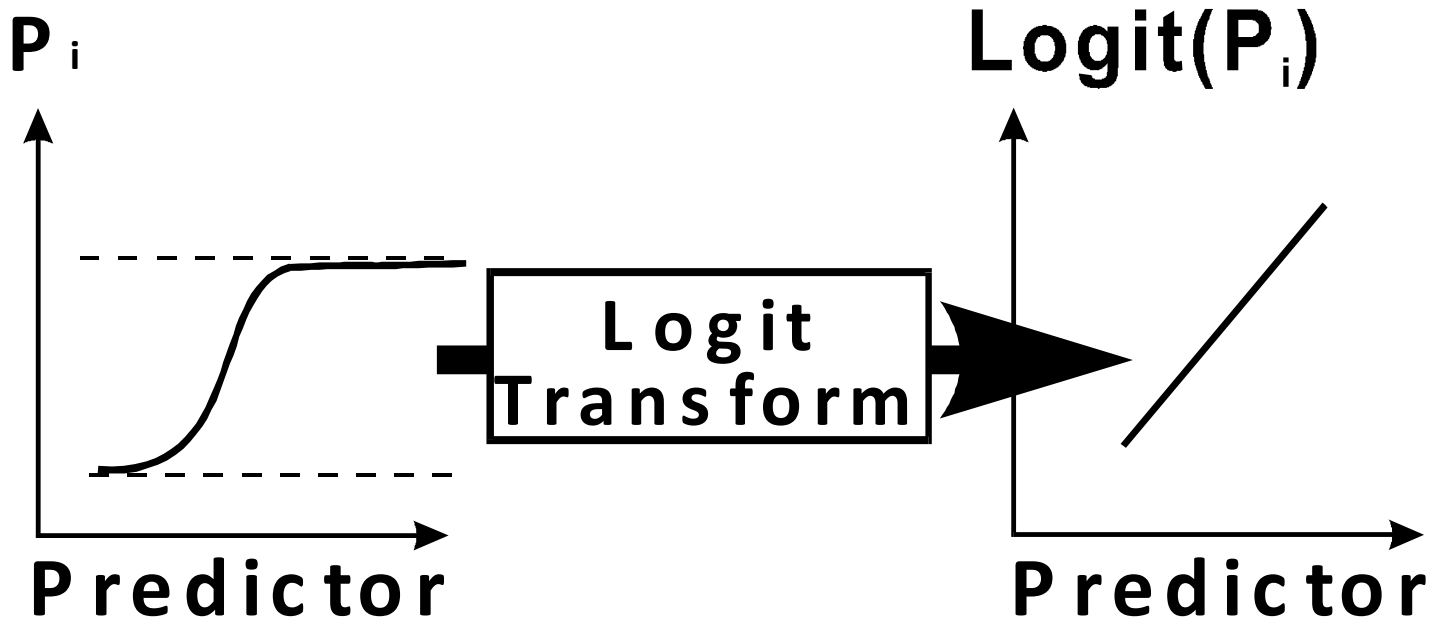
In other words, logistic regression is designed to describe probabilities associated with the values of the response variable.

Logistic Regression Curves



This graph shows the relationship between the probability of SALE to PRICE.

Assumption



Logit Transformation

Logistic regression models transformed probabilities called logits.

where $\mathbf{logit}(p_i) = \mathbf{log}\left(\frac{p_i}{1-p_i}\right)$

i indexes all cases (observations).

p_i is the probability the event (a sale, for example) occurs in the i^{th} case.

log is the natural log (to the base e).

Logistic Regression Model

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_1$$

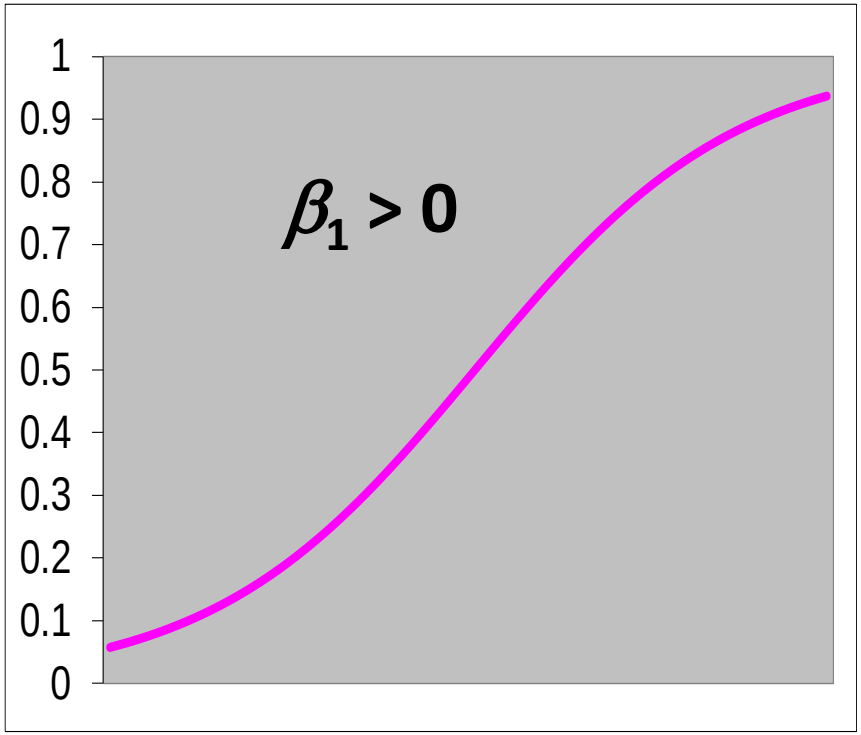
where

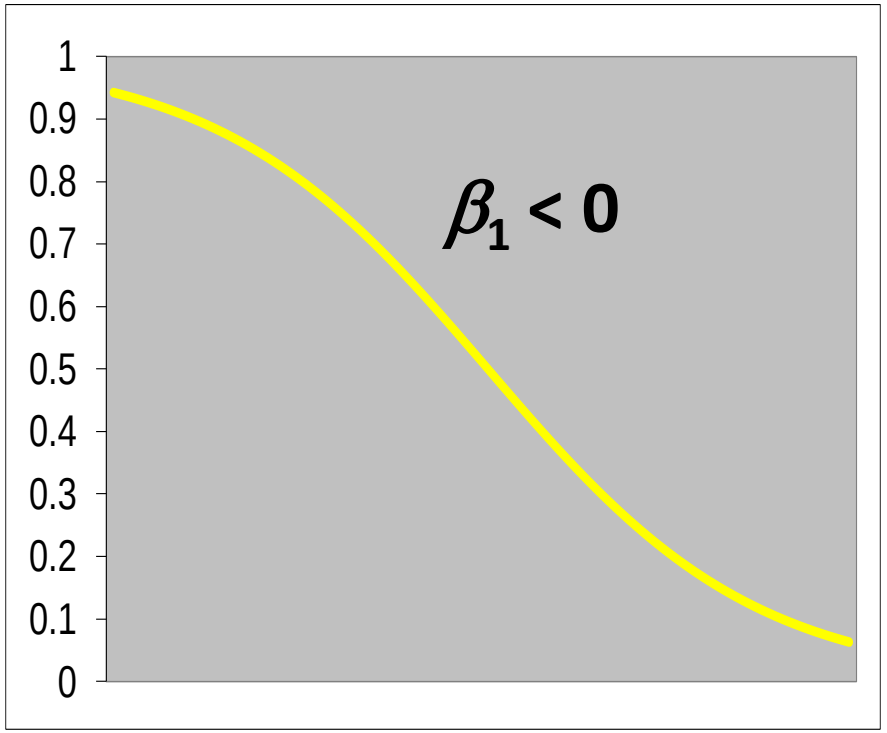
$\text{logit}(p_i)$	logit transformation of the probability of the event
β_0	intercept of the regression line
β_1	slope of the regression line.

Logistic Regression Model

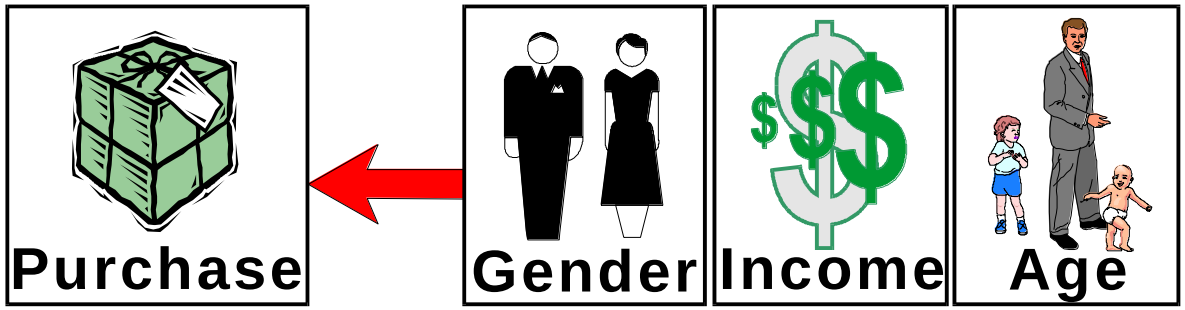
$$P(Y=1)=\pi=\frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}$$

Logistic Regression Model





Multiple Logistic Regression



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

Ilustrasi

- Respon: external rating (OK, NO)
- Prediktor:
  - return on equity
  - return on asset
  - cost to income ratio

```
proc logistic data = a.rating outmodel=modelrating;
model eksternal_rating (event = "OK") =
        Return_on_equity
        Return_on_Asset
        Cost_to_Income_Ratio;
run;
```

Response Profile		
Ordered Value	Eksternal_rating	Total Frequency
1	NO	81
2	OK	29

Probability modeled is Eksternal_rating='OK'.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.9957	1.8408	2.6484	0.1037
Return_on_equity	1	10.4903	4.6121	5.1735	0.0229
Return_on_Asset	1	199.7	71.9641	7.7005	0.0055
Cost_to_Income_Ratio	1	-4.9575	2.9008	2.9208	0.0874

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.9957	1.8408	2.6484	0.1037
Return_on_equity	1	10.4903	4.6121	5.1735	0.0229
Return_on_Asset	1	199.7	71.9641	7.7005	0.0055
Cost_to_Income_Ratio	1	-4.9575	2.9008	2.9208	0.0874

$$P(rating = OK) = \frac{e^{-2.99+10.49ROE+199.7ROA-4.96CIR}}{1 + e^{-2.99+10.49ROE+199.7ROA-4.96CIR}}$$

```
data maudiprediksi;
input Return_on_equity Return_on_Asset Cost_to_Income_Ratio;
cards;
0.1             0.02    0.8
0.2             0.02    0.6
;

proc logistic inmodel=modelrating;
score data=maudiprediksi out=prediksi;
run;
```

Obs	Return on equity	Return on Asset	Cost to Income Ratio	Eksternal rating	prediksi
1	0.2493	0.0301	0.48030	OK	0.96265
2	0.1764	0.0162	0.61955	OK	0.27260
3	0.1094	0.0055	0.65080	NO	0.01841
4	0.0626	0.0073	0.60580	NO	0.02015
5	0.1800	0.0094	0.56000	NO	0.11853
6	0.0982	0.0084	0.58180	NO	0.04022
7	0.2427	0.0094	0.52160	NO	0.23897
8	0.1493	0.0048	0.78570	NO	0.01254
9	0.1701	0.0056	0.57700	NO	0.04957
10	0.0947	0.0094	0.48440	NO	0.07402
11	0.1748	0.0116	0.52420	OK	0.19090
12	0.2429	0.0063	0.73110	NO	0.05658

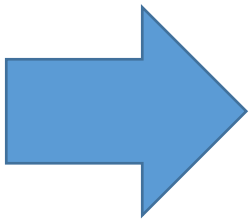
Diskretisasi

Discretization/Binning/Bucketing

Andaikan dataset berisi *N* observasi, proses diskretiasi terhadap variabel numerik *A* adalah mengubah nilai variabel tersebut menjadi *m* interval $D = \{[d_0, d_1], (d_1, d_2], \ldots, (d_{m-1}, d_m]\}$, dengan d_0 adalah nilai terkecil, d_m adalah nilai terbesar, dan $d_i < d_{i+i}$, untuk $i = 0, 1, \ldots, m-1$.

Diskretisasi

6.58
15.35
14.24
6.22
1.82
2.11
13.77
5.65
15.58
12.46
13.05
11.64
10.91
14.31
7.42

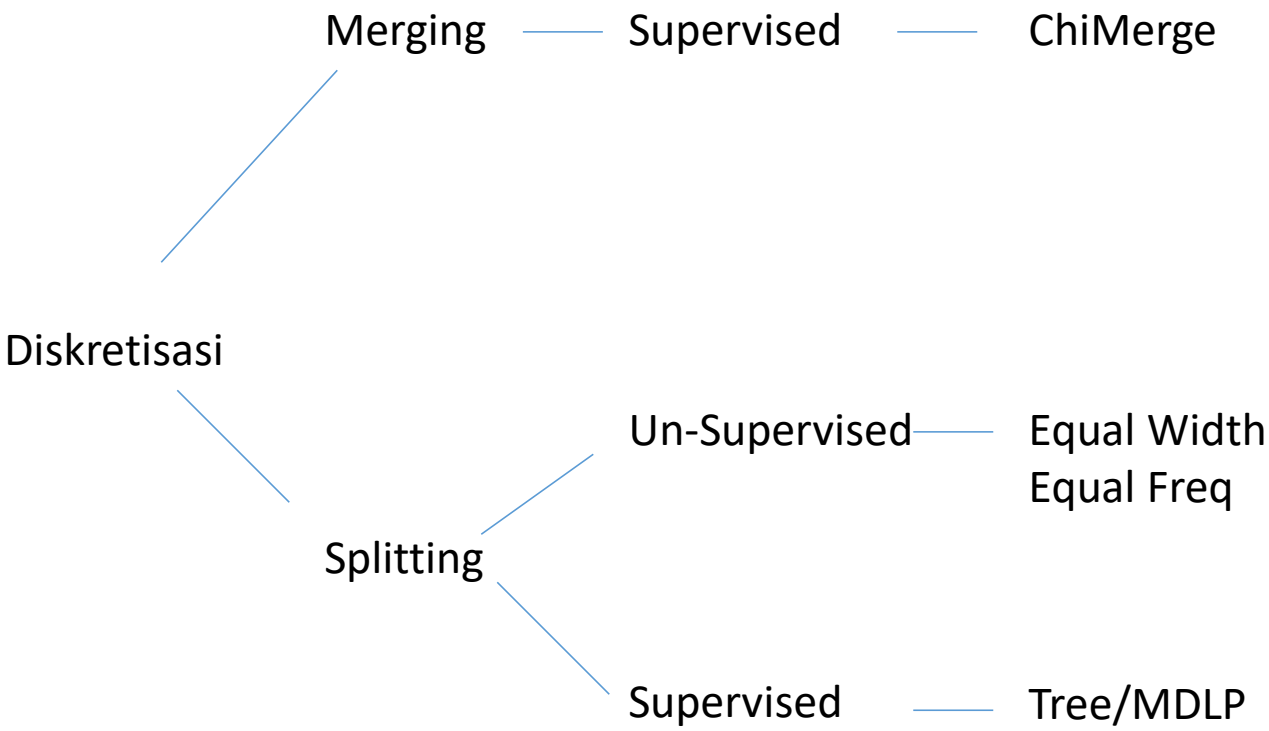


$X \leq 5$

$5 < X \leq 10$

$10 < X \leq 15$

$X > 15$

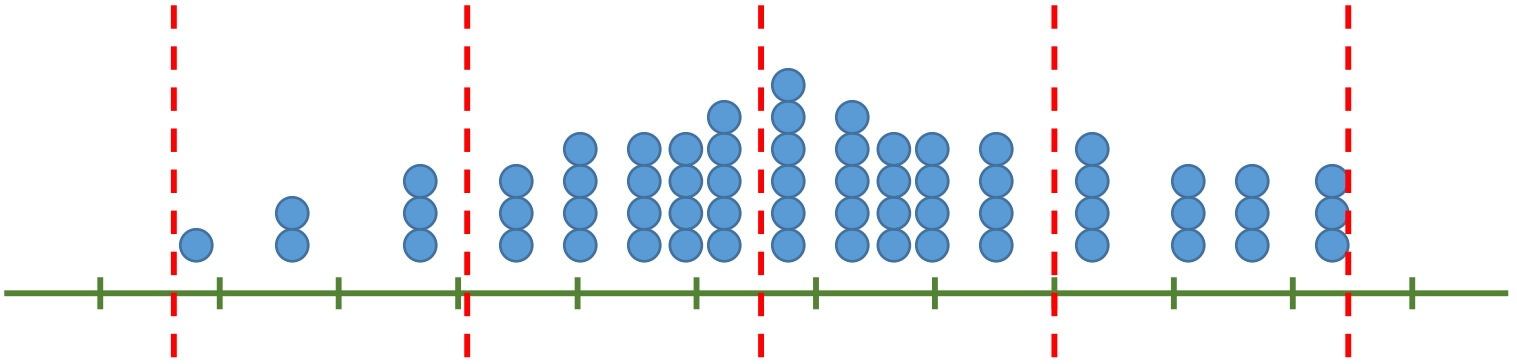

Diskretisasi
Merging
Supervised
ChiMerge
Splitting
Un-Supervised
Equal Width
Equal Freq
Supervised
Tree/MDLP

Equal Width dan Equal Frequency

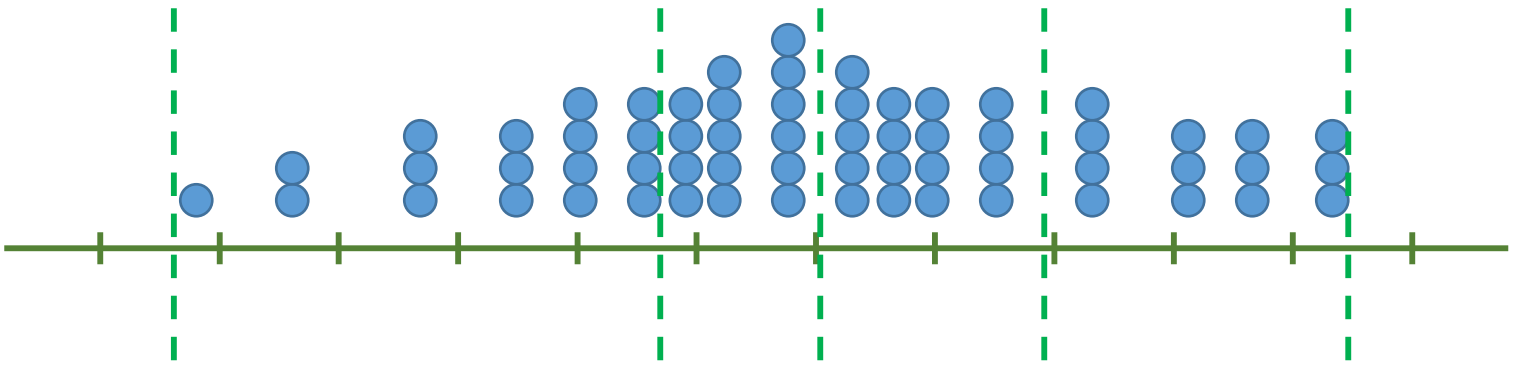
- In equal width, the continuous range of a feature is divided into intervals that have an equal width and each interval represents a bin. The arity can be calculated by the relationship between the chosen width for each interval and the total length of the attribute range.

- In equal frequency, an equal number of continuous values are placed in each bin. Thus, the width of each interval is computed by dividing the length of the attribute range by the desired arity.

Unsupervised Discretization: Equal Width Discretization



Unsupervised Discretization: Equal Freq Discretization



Hands-On

- Diskretisasi
- Pake hasil diskretisasi untuk regresi logistik

Tahapan Pembuatan Model Skoring

Tahapan

- Eksplorasi data dan Data cleansing
- Pemilihan variabel prediktor
  - WoE (weight of evidence)
  - IV (information value)
- Penyusunan model skoring awal
- Reject Inferece (khusus untuk approval scoring)
- Penyusunan model skoring dan scorecard
- Validasi model skoring

Eksplorasi Data

- Mengetahui gambaran umum mengenai karakteristik data yang akan digunakan
  - Jenis variabel: kategorik vs numerik
  - Distribusi nilai variabel
    - Kode dan kesalahan pengkodean
    - Outliers
  - Perlunya menyusun variabel baru (derivative variabel)
    - misal, variabel usia dari variabel tanggal lahir

Jenis Variabel

- Numerik
  - Misal: income, age, number of dependants
  - Terhadapnya dapat dilakukan operasi-operasi aritmatika
- Kategorik
  - Misal: gender, occupation, residential ownership
  - Ada yang bersifat ordinal, ada yang bersifat nominal

Jenis Variabel

- Variabel numerik bisa dijadikan variabel kategorik melalui proses discretization/binning/bucketing

- Misal, nilai income dikelompok-kelompokkan menjadi
  - Kurang dari 5 juta per bulan
  - 5 – 10 juta per bulan
  - 10 – 20 juta per bulan
  - Lebih dari 20 juta per bulan

Pemilihan Variabel

- Tidak semua variabel yang ada pada data layak untuk jadi prediktor dalam model skoring
- Hanya variabel yang memiliki pengaruh terhadap status good/bad saja yang pantas untuk dijadikan prediktor
- Variabel prediktor memiliki pengaruh jika untuk nilai variabel yang berbeda maka proporsi good/bad-nya berbeda
- Perbedaan tersebut dapat dilihat menggunakan nilai WoE (weight of evidence)

Weight of Evidence

$$WoE(X = k) = \log\left(\frac{P(X = k \mid Good)}{P(X = k \mid Bad)}\right)$$

Age	*Count*	*P(X=k)*	*Good*	*P(X=k|Good)*	*Bad*	*P(X=k|Bad)*	*Bad Rate*	*WoE*
Missing	1,000	2.50%	860	2.38%	140	3.65%	14.00%	-0.428
18–22	4,000	10.00%	3,040	8.41%	960	25.00%	24.00%	-1.089
23–26	6,000	15.00%	4,920	13.61%	1,080	28.13%	18.00%	-0.726
27–29	9,000	22.50%	8,100	22.40%	900	23.44%	10.00%	-0.045
30–35	10,000	25.00%	9,500	26.27%	500	13.02%	5.00%	0.702
35–44	7,000	17.50%	6,800	18.81%	200	5.21%	2.86%	1.284
44+	3,000	7.50%	2,940	8.13%	60	1.56%	2.00%	1.651
Total	40,000	100%	36,160	100%	3,840	100%	9.60%	

Cara membuat kelompok

- Discretization/Binning/Bucketing

- Binning variabel numerik
  - Awali dengan banyak bin/bucket, kemudian gabung bin dengan bad-rate atau WoE yang sama sehingga jumlah bin/bucket menjadi lebih sedikit
  - “Missing” is grouped separately
  - Rule of thumb: “minimum 5% in each bucket”
  - There is no bucket with 0 counts for good or bad.
  - The bad rate and WOE are sufficiently different from one bucket to the next
  - The WOE for nonmissing values also follows a logical distribution, for example: going from negative to positive without any reversals

Information Value

$$IV = \sum_{k=1}^{b} (P(X = k \mid Good) - P(X = k \mid Bad)) * WoE$$

- Mengukur kekuatan pengaruh suatu prediktor terhadap status good/bad
  - Less than 0.02: unpredictive
  - 0.02 to 0.1:weak
  - 0.1 to 0.3: medium
  - 0.3 +: strong

- Digunakan sebagai salah satu ukuran dalam pemilihan variabel untuk model skoring

Information Value

Age	*Count*	*P(X=k)*	*Good*	*P(X=k|Good)*	*Bad*	*P(X=k|Bad)*	*Bad Rate*	*WoE*
Missing	1,000	2.50%	860	2.38%	140	3.65%	14.00%	-0.428
18–22	4,000	10.00%	3,040	8.41%	960	25.00%	24.00%	-1.089
23–26	6,000	15.00%	4,920	13.61%	1,080	28.13%	18.00%	-0.726
27–29	9,000	22.50%	8,100	22.40%	900	23.44%	10.00%	-0.045
30–35	10,000	25.00%	9,500	26.27%	500	13.02%	5.00%	0.702
35–44	7,000	17.50%	6,800	18.81%	200	5.21%	2.86%	1.284
44+	3,000	7.50%	2,940	8.13%	60	1.56%	2.00%	1.651
Total	40,000	100%	36,160	100%	3,840	100%	9.60%	

IV = (2.38% - 3.655)*(-0.428) + (8.41% - 25.00*(-1.089) + ….
= 0.668

Penyusunan Model Skoring

- Gunakan nilai WoE sebagai pengganti nilai asli dari setiap variabel sesuai dengan bucket-nya masing-masing

- Susun model regresi logistik dengan variabel baru berisi WoE menjadi predictor variable dan status good/bad sebagai response variable

- Jika diperlukan, gunakan teknik-teknik penyeleksian variabel (forward, backward, stepwise) untuk memastikan kelayakan variabel yang disertakan dalam model.

Hands-On

Data

#	Variable	Type	Label
1	ID	Num	ID
2	Age	Num	Age
3	Gender	Char	Gender
4	Residence_Ownership	Char	Residence Ownership
5	number_of_dependants	Num	number of dependants
6	status	Char	status

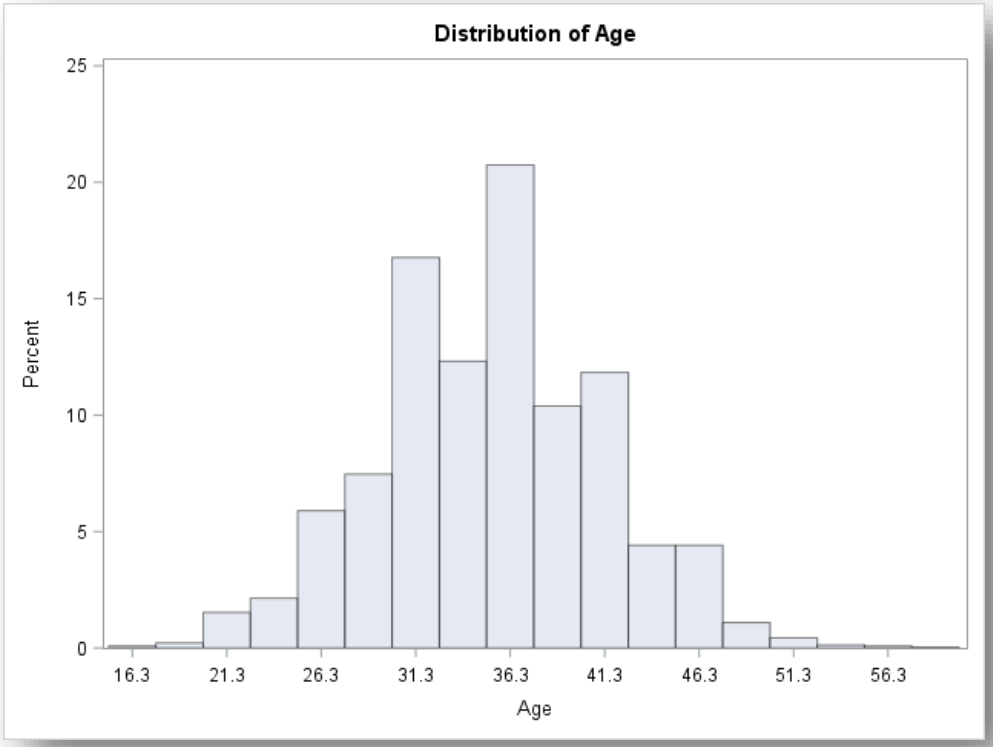
Peran dalam Pemodelan

Age	→	Input / Independent
Gender	→	Input / Independent
Residence_Ownership	→	Input / Independent
number_of_dependants	→	Input / Independent
status	→	Target / Dependent

Jenis/Tipe Peubah Input

Age	→	Numerik → perlu binning
Gender	→	Kategorik
Residence_Ownership	→	Kategorik
number_of_dependants	→	Numerik → perlu binning

Melihat Sebaran Nilai Variabel AGE



Quantiles (Definition 5)	
Level	Quantile
100% Max	59
99%	49
95%	45
90%	43
75% Q3	39
50% Median	35
25% Q1	31
10%	28
5%	25
1%	21
0% Min	16

```
proc univariate data=data.datascoring;
var age;
histogram age;
run;
```

Melakukan Binning Variabel AGE

```
data data.datascoring;
set data.datascoring;
if age <= 25 then agegroup = 1;
else if age <= 30 then agegroup = 2;
else if age <= 35 then agegroup = 3;
else if age <= 40 then agegroup = 4;
else if age <= 45 then agegroup = 5;
else agegroup = 6;
run;

proc tabulate data=data.datascoring;
class agegroup;
table agegroup all, n colpctn;
run;
```

	N	ColPctN
agegroup		
1	134	5.85
2	394	17.21
3	680	29.69
4	671	29.30
5	311	13.58
6	100	4.37
All	2290	100.00

Melihat Sebaran Nilai number_of_dependants

```
proc tabulate data=data.datascoring;
class number_of_dependants;
table number_of_dependants all, n colpctn;
run;
```

	N	ColPctN
number of dependants		
0	533	23.28
1	491	21.44
2	305	13.32
3	331	14.45
4	326	14.24
5	304	13.28
All	2290	100.00

Karena setiap nilai sudah cukup banyak frekuensinya → tidak diperlukan binning/diskretisasi

Setiap nilai menjadi bin

Menghitung WOE Gender

**** menghitung WOE dari variabel GENDER ***;
* tahapan: 1. Menghitung P(Gender | Good) dan P(Gender|Bad);

proc tabulate data=data.datascoring out=WOEgender;
class gender status; tables gender, status*colpctn; **run;**

proc transpose data=woegender out=woegender;
var pctn_01; by gender; id status; **run;**

* tahapan: 2. hitung WoE dengan formula WoE = log(P(-|GOOD)/P(-|BAD));
data WOEgender;
set WOEgender; WOEgender = log(GOOD / BAD); **run;**

* tahapan: 3. Berikan nilai WoE Gender pada data lengkap
    (datascoring);
data woegender (keep = gender woegender);
set woegender; **run;**
proc sort data=data.datascoring;
by gender; **run;**

data data.datascoring;
merge data.datascoring woegender; by gender; **run;**

Obs	Gender	WOEgender
1	FEMALE	0.88171
2	MALE	-0.46165

Menghitung WOE Residence

```
**** menghitung WOE dari variabel RESIDENCE ***;
proc tabulate data=data.datascoring out=WOEresidence;
class residence_ownership status; tables residence_ownership, status*colpctn; run;

proc transpose data=woeresidence out=woeresidence;
var pctn_01;  by residence_ownership;  id status;  run;

data WOEresidence;
set WOEresidence;  WOEresidence = log(GOOD / BAD);  run;

data woeresidence   keep = residence_ownership woeresidence);
set woeresidence;
run;

proc sort data=data.datascoring;
by residence_ownership ; run;

data data.datascoring;
merge data.datascoring woeresidence;
by residence_ownership ;
run;
```

Obs	Residence_Ownership	WOEresidence
1	OTHERS	-0.77202
2	OWNED	1.21297
3	PARENTS	-0.32148
4	RENT	-1.18076

Menghitung WOE Age Group

**** menghitung WOE dari variabel agegroup ***;
proc tabulate data=data.datascoring out=WOEagegroup;
class agegroup status; tables agegroup, status*colpctn; **run;**

proc transpose data=woeagegroup out=woeagegroup;
var pctn_01; by agegroup; id status; **run;**

data WOEagegroup;
set WOEagegroup; WOEagegroup = log(GOOD / BAD); **run;**

data woeagegroup (keep = agegroup woeagegroup);
set woeagegroup; **run;**

proc sort data=data.datascoring;
by agegroup; **run;**

data data.datascoring;
merge data.datascoring woeagegroup; by agegroup; **run;**

Obs	agegroup	WOEagegroup
1	1	-0.41277
2	2	-0.08443
3	3	-0.13305
4	4	0.12881
5	5	0.01846
6	6	1.27891

Menghitung WOE Number of Dependants

```
* Menghitung WoE untuk variabel NUMBER OF DEPENDANTS;
proc tabulate data=data.datascoring;
class number_of_dependants;  tables number_of_dependants, n colpctn;  run;

proc tabulate data=data.datascoring out=WOEdependants;
class number_of_dependants status;  tables number_of_dependants, status*colpctn;  run;

proc transpose data=woedependants out=woedependants;
var pctn_01;  by number_of_dependants;   id status;  run;

data WOEdependants;  set WOEdependants;
WOEdependants = log(GOOD / BAD);  run;

data woedependants
(keep = number_of_dependants woedependants);
set woedependants;  run;

proc sort data=data.datascoring;  by number_of_dependants;  run;

data data.datascoring;
merge data.datascoring woedependants;  by number_of_dependants;  run;
```

Obs	number_of_dependants	WOEdependants
1	0	0.90721
2	1	0.27666
3	2	0.18904
4	3	-0.10972
5	4	-0.82406
6	5	-0.75300

Menghitung Information Value dari Gender

```
**** menghitung INFORMATION VALUE dari GENDER;
proc tabulate data=data.datascoring out=WOEgender;
class gender status;
tables gender, status*colpctn;
run;
proc transpose data=woegender out=woegender;
var pctn_01;
by gender;
id status;
run;
data WOEgender;
set WOEgender;
WOEgender = log(GOOD / BAD);
IVgender = (GOOD - BAD) * WOEgender /100;
run;
proc tabulate data=WOEgender;
var IVgender;
tables sum, IVgender;
run;
```

	IVgender
Sum	0.39

Menghitung Information Value dari Age

```
proc tabulate data=data.datascoring out=WOEagegroup;
class agegroup status;
tables agegroup, status*colpctn;
run;
proc transpose data=woeagegroup out=woeagegroup;
var pctn_01;
by agegroup;
id status;
run;
data WOEagegroup;
set WOEagegroup;
WOEagegroup = log(GOOD / BAD);
IVagegroup = (GOOD - BAD) * WOEagegroup / 100;
run;
proc tabulate data=WOEagegroup;
var IVagegroup;
tables sum, IVagegroup;
run;
```

	IVagegroup
Sum	0.07

Menghitung Information Value dari Residence

```
proc tabulate data=data.datascoring out=WOEresidence;
class residence_ownership status;
tables residence_ownership, status*colpctn;
run;
proc transpose data=woeresidence out=woeresidence;
var pctn_01;
by residence_ownership;
id status;
run;
data WOEresidence;
set WOEresidence;
WOEresidence = log(GOOD / BAD);
IVresidence = (GOOD - BAD) * WOEresidence / 100;
run;
proc tabulate data=WOEresidence;
var IVresidence;
tables sum, IVresidence;
run;
```

	IVresidence
Sum	1.12

Menghitung Information Value dari Number of Dependants

```
proc tabulate data=data.datascoring out=WOEdependants;
class number_of_dependants status;
tables number_of_dependants, status*colpctn;
run;
proc transpose data=woedependants out=woedependants;
var pctn_01;
by number_of_dependants;
id status;
run;
data WOEdependants;
set WOEdependants;
WOEdependants = log(GOOD / BAD);
IVdependants = (GOOD - BAD) * WOEdependants / 100;
run;
proc tabulate data=WOEdependants;
var IVdependants;
tables sum, IVdependants;
run;
```

	IVdependants
Sum	0.37

Menentukan Bobot Setiap Variabel

********* menentukan bobot masing-masing variabel;
proc logistic data=data.datascoring outest=bobot;
model status (event = 'GOOD') = WOEgender WOEagegroup WOEresidence WOEdependants;
run;

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.7177	0.0601	142.4263	<.0001
WOEgender	1	1.4986	0.1017	217.0604	<.0001
WOEagegroup	1	1.6055	0.2668	36.2019	<.0001
WOEresidence	1	1.2996	0.0611	452.1798	<.0001
WOEdependants	1	1.6036	0.1067	226.0576	<.0001

```
data WOEgender (keep = category WOE input);
set WOEgender; length input $ 20;
input = 'WOEgender'; category = gender; WOE = WOEgender;   run;

data WOEagegroup (keep = category WOE input);
set WOEagegroup; length input $ 20;
input = 'WOEagegroup'; category = compress(agegroup); WOE = WOEagegroup;  run;

data WOEresidence (keep = category WOE input);
set WOEresidence; length input $ 20;
input = 'WOEresidence'; category = residence_ownership; WOE = WOEresidence; run;

data WOEdependants (keep = category WOE input);
set WOEdependants; length input $ 20;
input = 'WOEdependants'; category = compress(number_of_dependants); WOE = WOEdependants;
run;

data WOEall;
set WOEgender WOEagegroup WOEresidence WOEdependants;
run;
```

Obs	input	category	WOE
1	WOEgender	FEMALE	0.88171
2	WOEgender	MALE	-0.46165
3	WOEagegroup	1	-0.41277
4	WOEagegroup	2	-0.08443
5	WOEagegroup	3	-0.13305
6	WOEagegroup	4	0.12881
7	WOEagegroup	5	0.01846
8	WOEagegroup	6	1.27891
9	WOEresidence	OTHERS	-0.77202
10	WOEresidence	OWNED	1.21297
11	WOEresidence	PARENT	-0.32148
12	WOEresidence	RENT	-1.18076
13	WOEdependants	0	0.90721
14	WOEdependants	1	0.27666
15	WOEdependants	2	0.18904
16	WOEdependants	3	-0.10972
17	WOEdependants	4	-0.82406
18	WOEdependants	5	-0.75300

Parameter	DF	Estimate
Intercept	1	0.7177
WOEgender	1	1.4986
WOEagegroup	1	1.6055
WOEresidence	1	1.2996
WOEdependants	1	1.6036

Pembuatan Scorecard

- Model regresi logistik yang diperoleh sebenarnya sudah dapat dipergunakan untuk menghasilkan skor

- Skor yang dihasilkan berupa nilai peluang seorang customer untuk menjadi ‘bad’-customer (atau sebaliknya menjadi good-customer). Nilainya antara 0 dan 1.

- Skor tersebut diperoleh dengan memasukkan nilai-nilai variabel prediktor ke dalam model regresi logistik.

Penskalaan

- Proses scaling (penskalaan) seringkali diperlukan terhadap hasil regresi logistik

- Penskalaan tidak mempengaruhi power dari model skoring

- Alasan penggunaan penskalaan antara lain:
  - Implementability of the scorecard into application processing software.
  - Ease of understanding by staff (e.g., discrete numbers are easier to work with).
  - Continuity with existing scorecards or other scorecards in thecompany. This avoids retraining on scorecard usage and interpretation of scores.

Penskalaan

$$\text{Score} = \text{Offset} + \text{Factor } \textit{ln (odds)}$$

- nilai OFFSET dan FACTOR dapat diperoleh jika telah didefinisikan
  - nilai skor yang diinginkan untuk odds tertentu
  - nilai pdo (points to double the odds), yaitu besarnya kenaikan skor yang menyebabkan odds-nya menjadi dua kali lipat

Penskalaan

$$\text{Score} = \text{Offset} + \text{Factor}\ ln\ (odds)$$

$$\text{Score} + pdo = \text{Offset} + \text{Factor}\ ln\ (2*odds)$$

- Misal, scorecard yang diinginkan memiliki odds of 50:1 pada nilai 600 dan odds-nya akan dua kali lipat kalau skornya bertambah 20 points (*pdo = 20)*

- Maka akan diperoleh

  Factor = 20 / *ln (2) = 28.8539*

  Offset = 600 – {28.8539 *ln (50)} = 487.123*

- Sehingga

  Score = 487.123 + 28.8539 *ln (odds)*

Penskalaan

- Ingat bahwa, dalam model regresi logistik

$$\ln(\text{odds}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

atau

$$\ln(\text{odds}) = \beta_0 + \beta_1 \text{WoE}_1 + \beta_2 \text{WoE}_2 + \ldots + \beta_p \text{WoE}_p$$

$$\ln(\text{odds}) = \beta_0 + \sum \beta_i \text{WoE}_i$$

Sehingga

Score = Offset + Factor *ln (odds)*

Score = Offset + Factor * ($\beta_0 + \sum \beta_i \text{WoE}_i$)

Score = $\sum$ (($\beta_i \text{WoE}_i + \beta_0/p$) * factor + Offset / p)

Scorecard

- Dengan demikian, skor dari suatu individu merupakan penjumlahan dari skor untuk setiap variabel prediktor

- Skor dari setiap variabel prediktor diperoleh menggunakan formula

$$(\beta_i WoE_i + \beta_0/p) * \text{factor} + \text{Offset} / p$$

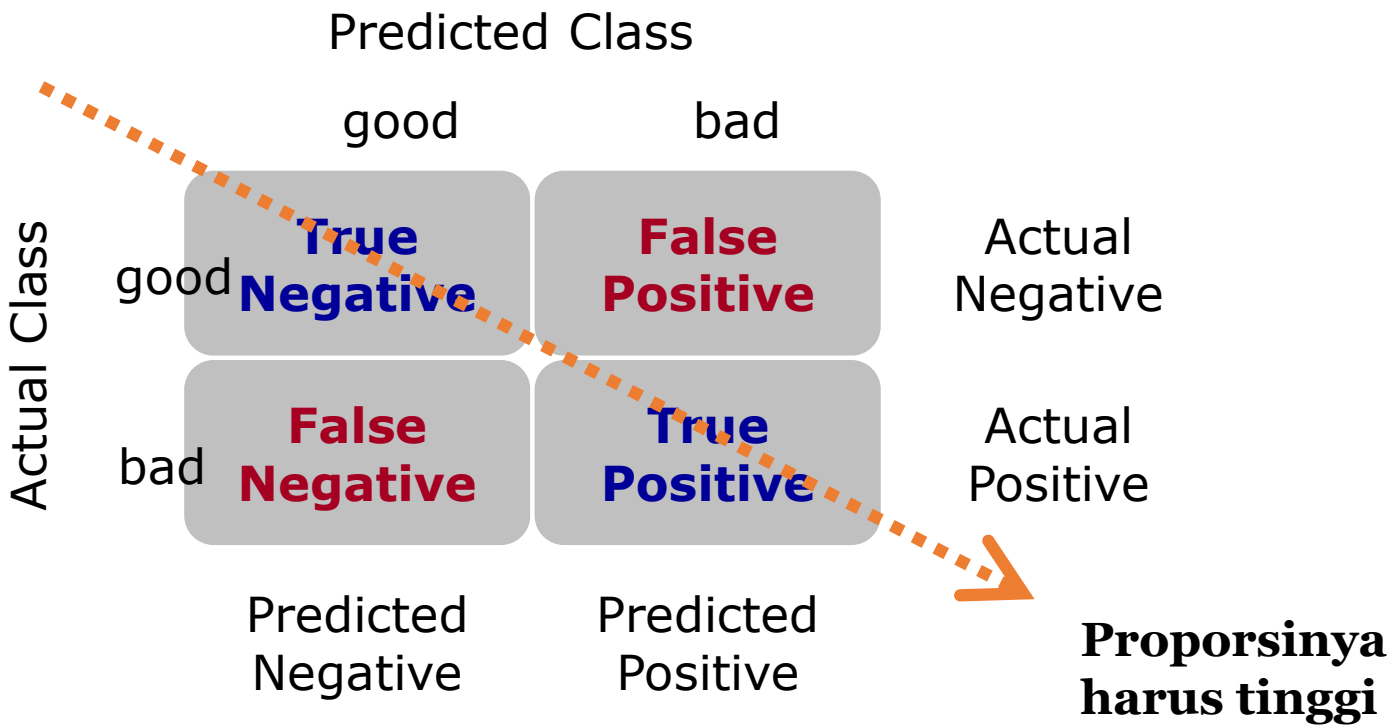
```
data _null_;
set bobot;
if _n_=1 then call symput("b0", intercept);
if _n_=1 then call symput("bgender", WOEgender);
if _n_=1 then call symput("bagegroup", WOEagegroup);
if _n_=1 then call symput("bresidence", WOEresidence);
if _n_=1 then call symput("bdependants", WOEdependants);
run;

data WOEall (drop = factor offset);
set WOEall;
Factor = 20 / log (2);
Offset = 600 - factor * log (50);
if input = 'WOEgender' then score = (&bgender * WOE + &b0 / 4) * factor + offset / 4;
if input = 'WOEagegroup' then score = (&bagegroup * WOE + &b0 / 4) * factor + offset / 4;
if input = 'WOEresidence' then score = (&bresidence * WOE + &b0 / 4) * factor + offset / 4;
if input = 'WOEdependants' then score = (&bdependants * WOE + &b0 / 4) * factor + offset / 4;
score = round(score);
run;

proc print data=WOEall;
run;
```

Scorecard yang Dihasilkan

Obs	input	category	WOE	score
1	WOEgender	FEMALE	0.88171	165
2	WOEgender	MALE	-0.46165	107
3	WOEagegroup	1	-0.41277	108
4	WOEagegroup	2	-0.08443	123
5	WOEagegroup	3	-0.13305	121
6	WOEagegroup	4	0.12881	133
7	WOEagegroup	5	0.01846	128
8	WOEagegroup	6	1.27891	186
9	WOEresidence	OTHERS	-0.77202	98
10	WOEresidence	OWNED	1.21297	172
11	WOEresidence	PARENT	-0.32148	115
12	WOEresidence	RENT	-1.18076	83
13	WOEdependants	0	0.90721	169
14	WOEdependants	1	0.27666	140
15	WOEdependants	2	0.18904	136
16	WOEdependants	3	-0.10972	122
17	WOEdependants	4	-0.82406	89
18	WOEdependants	5	-0.75300	92

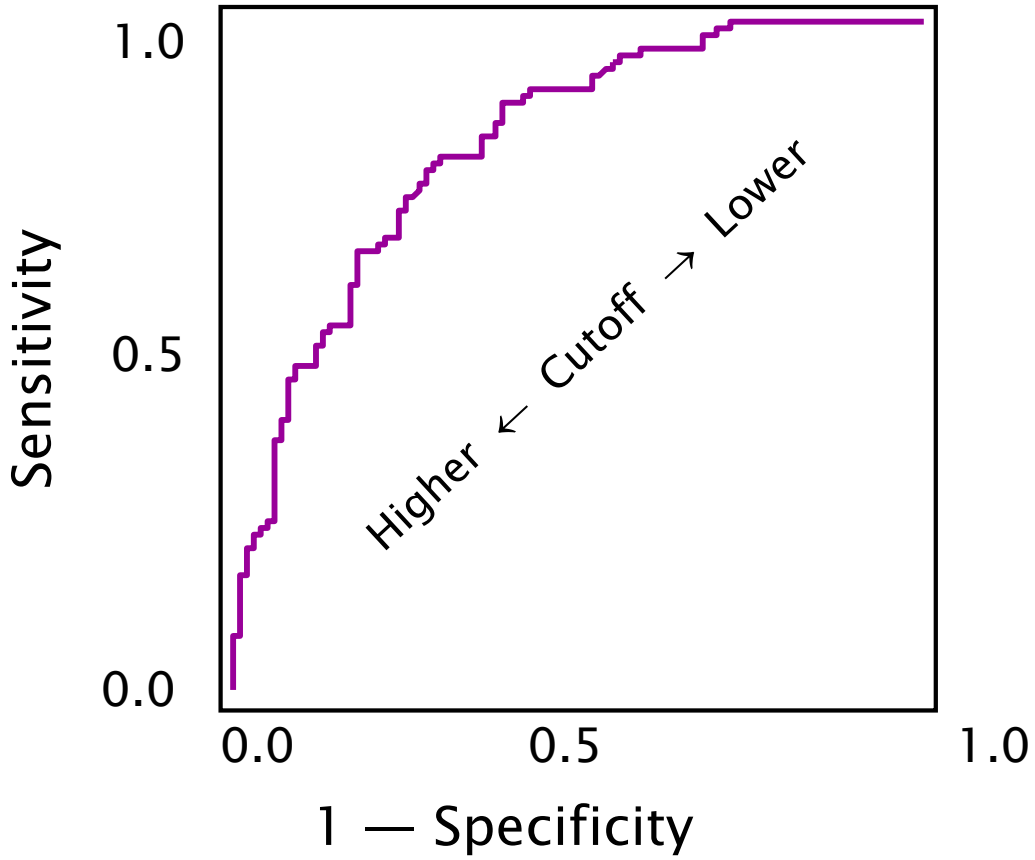
Mengevaluasi Model Skoring

- Misclassification
  - Classification table / Confusion table



Mengevaluasi Model Skoring

- ROC-curve



Validating the Model

- Model yang dihasilkan harus dapat memberikan performan yang baik tidak hanya pada data yang digunakan dalam menyusun model (learning set, in-sample), tetapi juga mempunyai kemampuan prediktif yang baik pada gugus data lain (testing set, out-sample)
  - Scoring an alternate data set
  - Population Stability Index

Scoring an Alternate Data Set

- Models often deteriorate when scored on a data set that was not used in model development
- A similar campaign from a different time period or geographic area is good for validation
- Methods for scoring
  - Within the model development process
  - modeling new data

Menilai seberapa bagus model yang diperoleh

Tahapan:
- Menghitung score dari setiap individu
- Menentukan prediksi kelas status
- Membandingkan prediksi dengan status yang sebenarnya

```
data data.datascoring;
set data.datascoring;
Factor = 20 / log (2);
Offset = 600 - factor * log (50);
SCOREgender = round((&bgender * WOEgender + &b0 / 4) * factor + offset / 4);
SCOREagegroup = round((&bagegroup * WOEagegroup + &b0 / 4) * factor + offset / 4);
SCOREresidence = round((&bresidence * WOEresidence + &b0 / 4) * factor + offset / 4);
SCOREdependants = round((&bdependants * WOEdependants + &b0 / 4) * factor + offset / 4);
SCOREtotal = sum(SCOREgender, SCOREagegroup, SCOREresidence, SCOREdependants);
run;
```

Menilai seberapa bagus model yang diperoleh

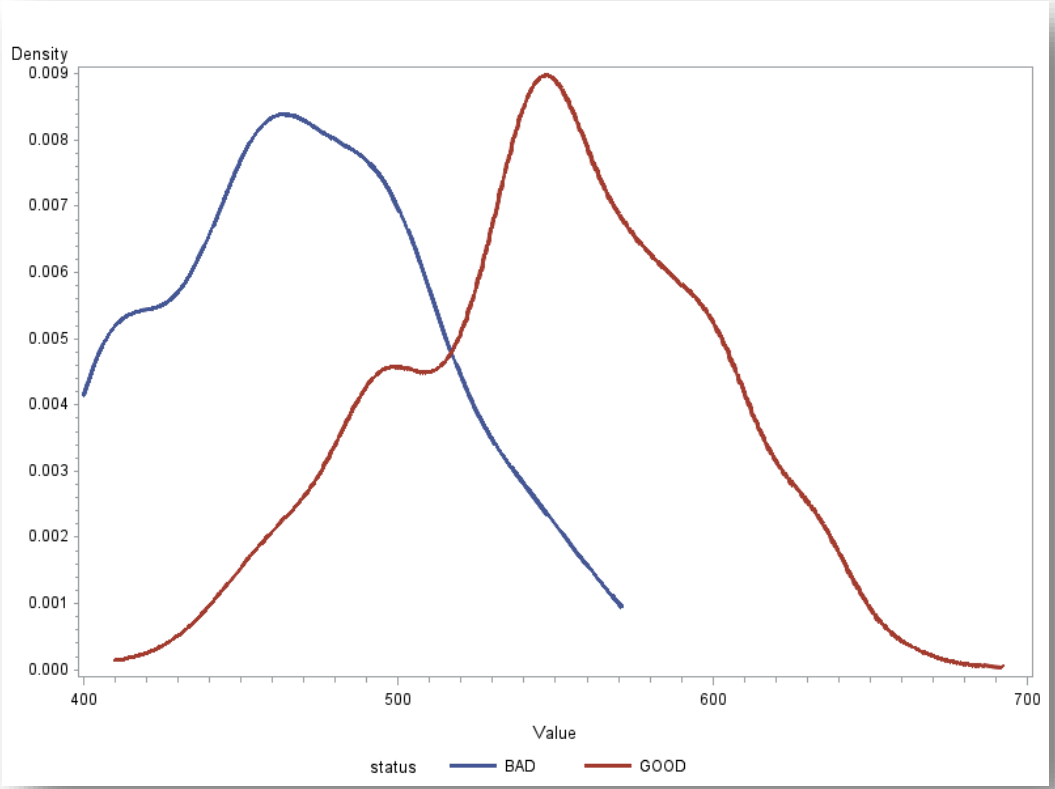
```
data data.datascoring;
set data.datascoring;
if SCOREtotal > 500 then predict = "GOOD";
else predict = "BAD ";
run;

proc tabulate data=data.datascoring;
class status predict;
table status, predict*(n pctn rowpctn);
run;
```

	predict					
	BAD			GOOD		
	N	PctN	RowPctN	N	PctN	RowPctN
status						
BAD	570	24.89	75.70	183	7.99	24.30
GOOD	282	12.31	18.35	1255	54.80	81.65

Accuracy = ?
Sensitivity = ?
Specificity = ?

Menilai seberapa bagus model yang diperoleh

```
proc sort data=data.datascoring;
by status;

proc kde data=data.datascoring;
univar SCOREtotal / out=density bwm=3;
by status;
run;

symbol1 i=join w=2;
symbol2 i=join w=2;
proc gplot data=density;
plot density*value=status;
run;
quit;
```



Reject Inference

- Pada kasus pembuatan approval scoring, penggunaan data customer penerima kredit dapat menyebabkan bias.

- Hal ini dikarenakan data yang digunakan hanya melibatkan individu yang terpilih (tidak secara acak) oleh proses seleksi approval sebelumnya.

- Dengan demikian, data yang digunakan memiliki sifat keterwakilan (representativeness) yang rendah.

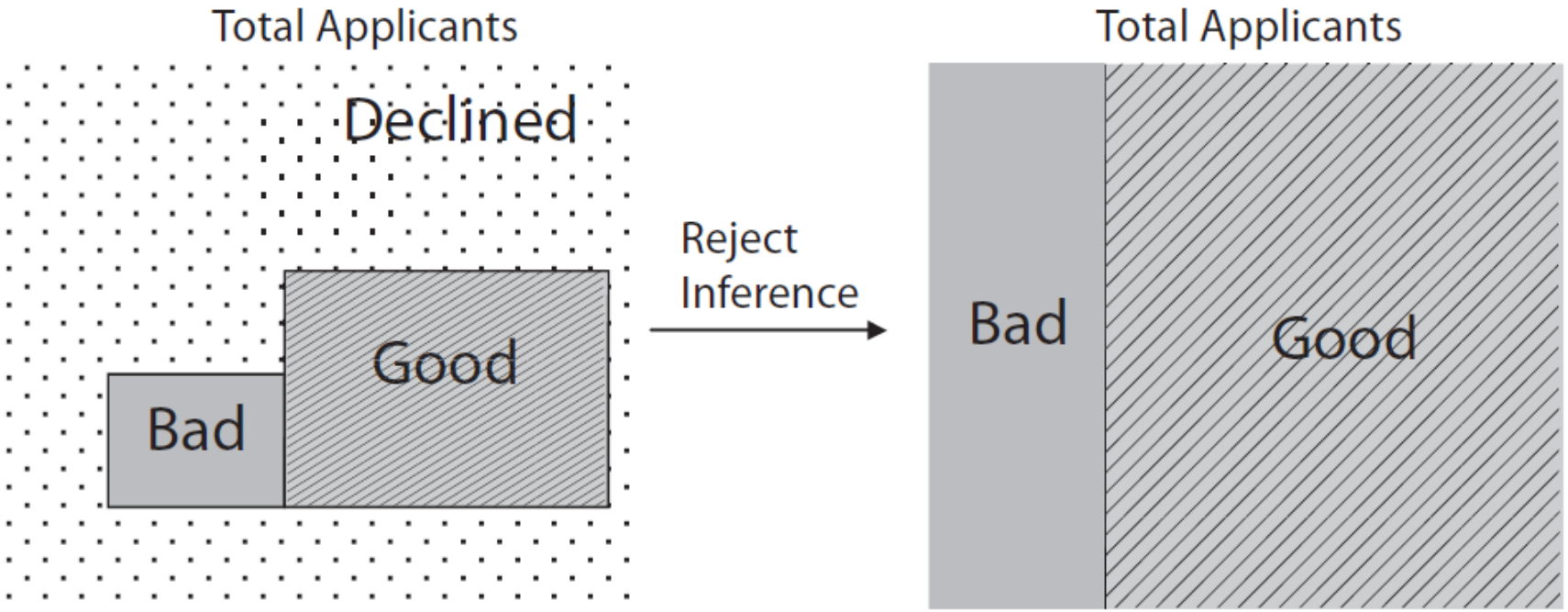
Reject Inference

- Karena itu, ada baiknya melibatkan juga data-data individu yang ditolak pada seleksi proses approval sebelumnya.

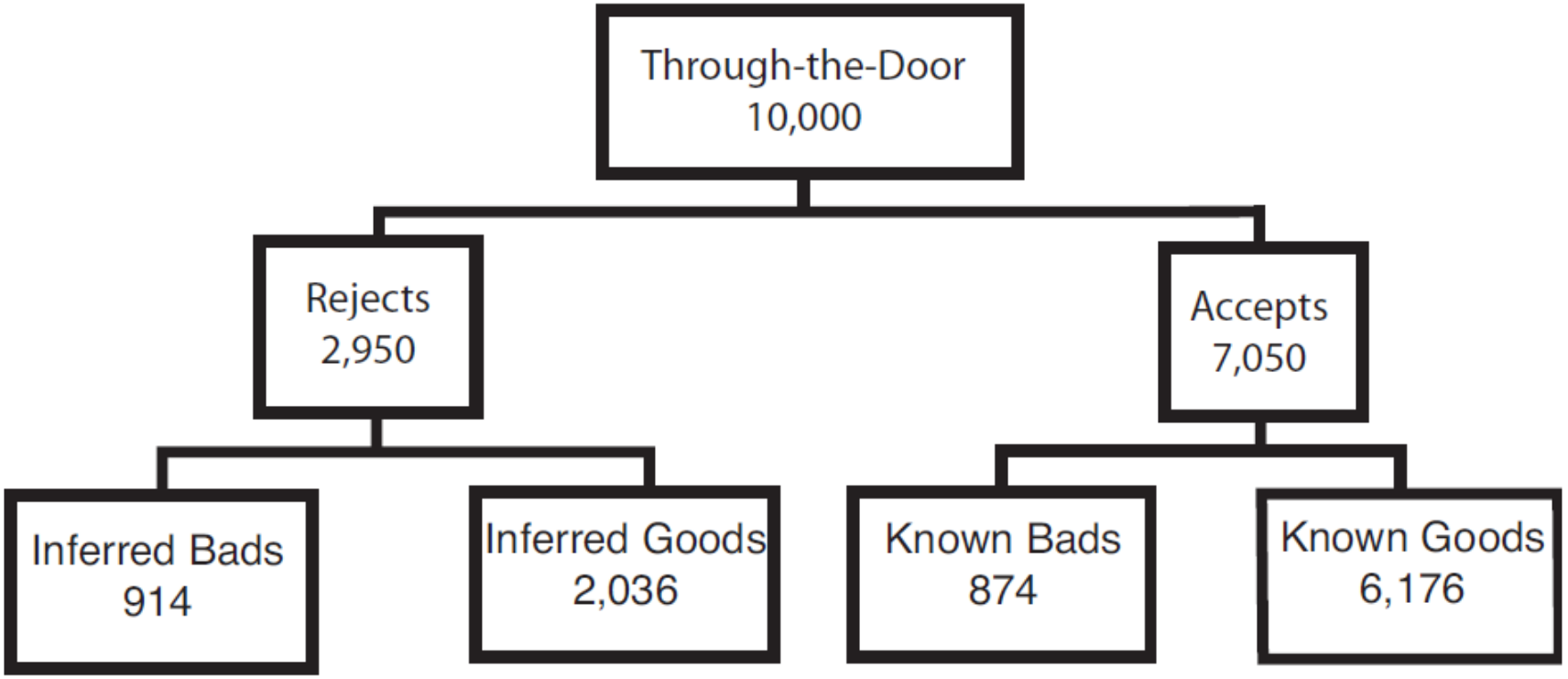
- Yang menjadi persoalan, individu yang ditolak tersebut tidak diketahui status good/bad-nya karena memang tidak menjadi customer dari produk kredit.

- Perlu upaya untuk memberikan status good/bad pada data individu yang ditolak agar bisa digunakan.

Reject Inference



Reject Inference



Teknik melakukan Reject Inference

- **Simple Augmentation (hard cutoff)**
  - Step 1 Build a model using known goods and bads
  - Step 2 Score rejects using this model and establish their expected bad rates, or *p(bad).*
  - Step 3 Set an expected bad rate level above which an account isdeemed “bad”; all applicants below this level are conversely classified as “good.”
  - Step 4 Add the inferred goods and bads to the known goods/bads and remodel.

Teknik melakukan Reject Inference

- **Nearest Neighbor (Clustering)**
  - Step 1 Create two sets of clusters—one each for known goods and bads.
  - Step 2 Run rejects through both clusters.
  - Step 3 Compare Euclidean distances to assign most likely performance (i.e., if a reject is closer to a “good” cluster than a “bad” one, then it is likely a good).
  - Step 4 Combine accepts and rejects to create inferred dataset, and remodel.