Analisis Regresi Beserta Metode Evaluasinya

Kuliah 2 - STA1382 Teknik Pembelajaran Mesin

Septian Rahardiantoro

Outline

- Pengantar Pemodelan Statistika
- Regresi linier beserta metode evaluasinya
- Regresi logistik beserta metode evaluasinya

Pengantar Pemodelan Statistika

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

- 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" (GEP Box)

Beberapa Model Statistika yang Populer

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

Regresi Linier

- Syarat Utama: Variabel output (Y) bersifat numerik

- Variabel prediktor (X)
    - numerik OK, kategorik OK
    - satu OK, lebih dari satu OK

- Bentuk model

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

- Analisis Regresi digunakan untuk:
  - Menjelaskan dampak perubahan peubah prediktor terhadap peubah respon
  - Memprediksi nilai dari peubah respon berdasarkan nilai dari setidaknya sebuah peubah prediktor

Peubah Respon (peubah tak bebas, peubah terikat, dependent variable): peubah yang ingin kita jelaskan

Peubah Prediktor (peubah bebas, independent variable): peubah yang digunakan untuk menjelaskan peubah respon

Regresi Linier Sederhana

- Suatu pendekatan untuk memprediksi peubah respon kuantitatif Y berdasarkan sebuah peubah prediktor X

- Pendekatan ini mengasumsikan bahwa ada hubungan linier antara X dan Y

The population regression model:

Dependent Variable

Population y intercept

Population Slope Coefficient

Independent Variable

Random Error term, or residual

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{Linear component}} + \underbrace{\varepsilon}_{\text{Random Error component}}$$

- Pendugaan koefisien
  - Misalkan $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ adalah prediksi untuk Y berdasarkan nilai ke-i peubah X (dengan $i = 1,2,3 \dots, n$)
  - Maka residual ke-i didefinisikan oleh:
$$e_i = y_i - \hat{y}_i \rightarrow e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$
  - JKG (Jumlah Kuadrat Galat) didefinisikan oleh:
$$JKG = e_1^2 + e_2^2 + \cdots + e_n^2$$
$$JKG = \left(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)^2 + \cdots + \left(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right)^2$$
  - Penduga MKT (Metode Kuadrat Terkecil), memilih $\hat{\beta}_0$ dan $\hat{\beta}_1$ yang meminimumkan JKG. Dengan perhitungan kalkulus diperoleh:
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}; \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Ilustrasi kontur dan plot 3D pada JKG (RSS) untuk model dengan Y = sales dan X = TV

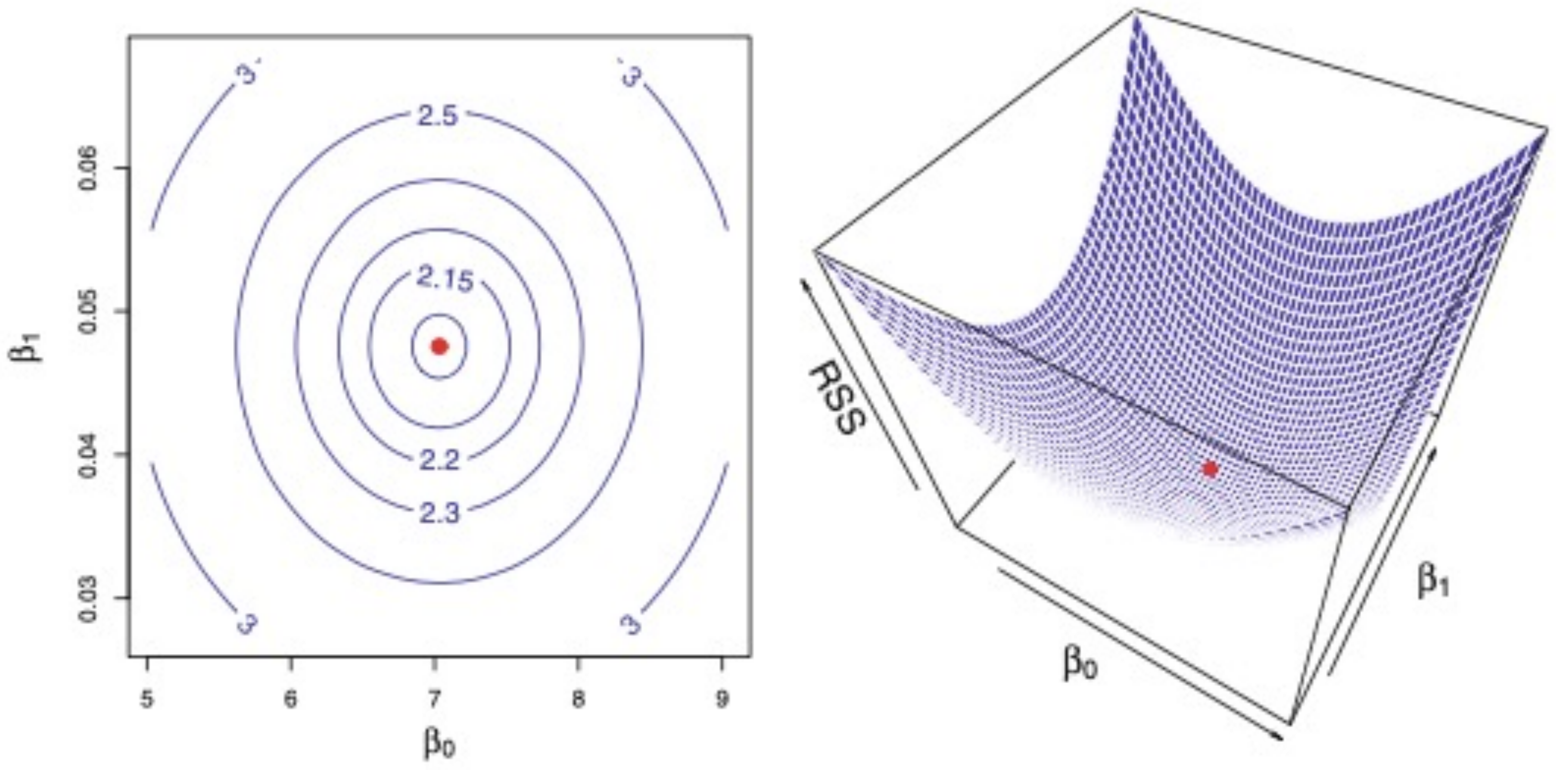
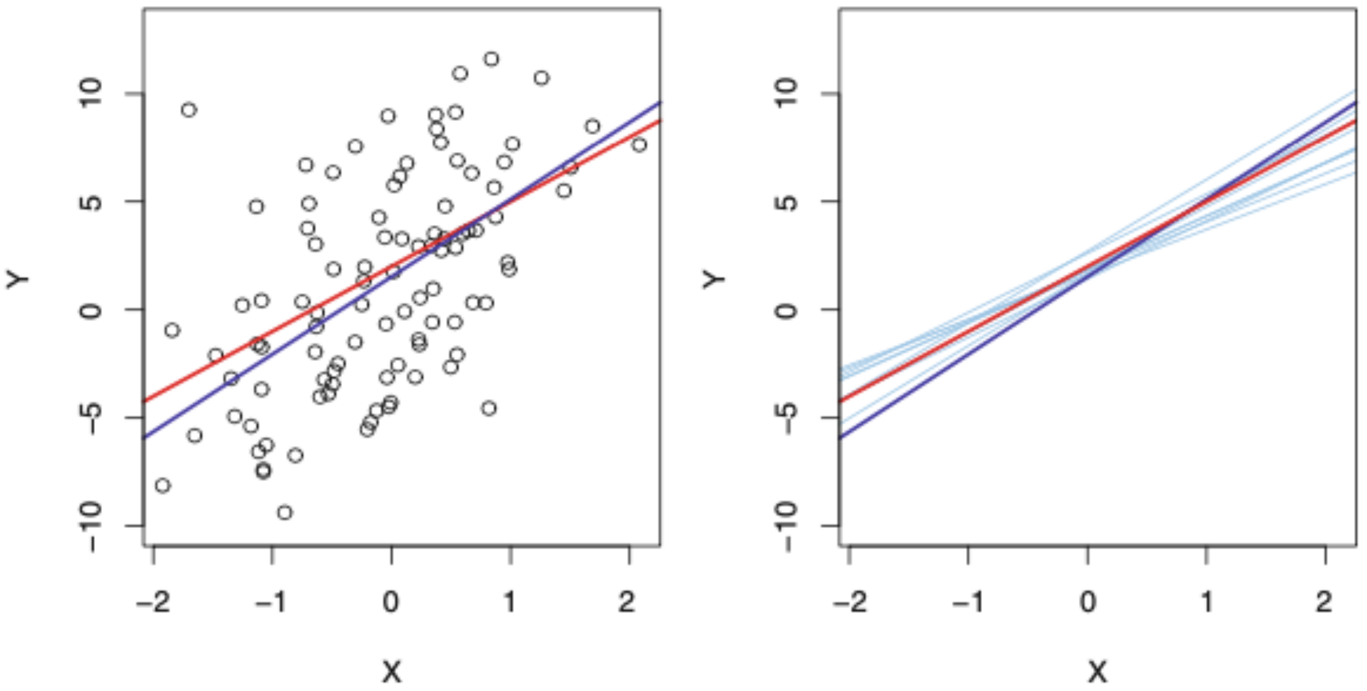


FIGURE 3.2. *Contour and three-dimensional plots of the RSS on the* Advertising *data, using* sales *as the response and* TV *as the predictor. The red dots correspond to the least squares estimates* $\hat{\beta}_0$ *and* $\hat{\beta}_1$*, given by (3.4).*

Menilai Akurasi Penduga Koefisien



Simulasi.

Kiri: Garis merah mewakili hubungan sebenarnya, $f(X) = 2 + 3X$, yang dikenal sebagai garis regresi populasi. Garis biru adalah garis kuadrat terkecil (MKT); yang merupakan dugaan kuadrat terkecil untuk $f(X)$ berdasarkan data yang diamati, ditampilkan dalam warna hitam.

Kanan: Garis regresi populasi ditampilkan lagi dengan warna merah, dan garis kuadrat terkecil berwarna biru tua. Dengan warna biru muda, sepuluh garis kuadrat terkecil ditampilkan, masing-masing dihitung berdasarkan kumpulan pengamatan acak yang terpisah. Setiap garis kuadrat terkecil berbeda, tetapi rata-rata garis kuadrat terkecil cukup dekat dengan garis regresi populasi.

- Sekilas, perbedaan antara garis regresi populasi dan garis kuadrat terkecil mungkin tampak halus dan membingungkan.

- Dalam kasus ini, diketahui satu kumpulan data, namun terdapat banyak garis berbeda menggambarkan hubungan antara prediktor dan respons?

- Sehingga, muncul pertanyaan seberapa dekat penduga $\hat{\beta}_0$ dan $\hat{\beta}_1$ terhadap β_0 dan β_1
  - Hal ini dapat diselidiki dengan standar error (galat baku) $\hat{\beta}_0$ dan $\hat{\beta}_1$
  $$SE\left(\hat{\beta}_0\right)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]; SE\left(\hat{\beta}_1\right)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
  - Secara kasar, galat baku memberi tahu kita jumlah rata-rata perkiraan pendugaan berbeda dari nilai parameter aktualnya
  - Selang kepercayaan bagi β_0 dan β_1 (taraf nyata 95%)
  $$\hat{\beta}_0 \pm 2 \times SE\left(\hat{\beta}_0\right);\ \ \hat{\beta}_1 \pm 2 \times SE\left(\hat{\beta}_1\right)$$
  - Uji hipotesis $\beta_1 \rightarrow H_0\colon \beta_1 = 0; H_1\colon \beta_1 \neq 0$ dengan statistik uji $t = \frac{\hat{\beta}_1 - 0}{SE\left(\hat{\beta}_1\right)}$

Menilai Akurasi Model

- Kualitas kecocokan regresi linier biasanya dinilai menggunakan dua besaran terkait: Galat Baku Residual (residual standard error) dan statistik R^2

- **Galat Baku Residual**
  - Galat Baku Residual merupakan dugaan simpangan baku dari residual, yakni jumlah rata-rata respon yang akan menyimpang dari garis regresi yang sebenarnya.

$$\text{Galat Baku Residual} = \sqrt{\frac{1}{n-2}JKG} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

  - Galat Baku Residual dianggap sebagai ukuran kecocokan model dengan data.
    - Jika prediksi yang diperoleh dengan menggunakan model sangat dekat dengan nilai hasil sebenarnya—yaitu, jika $\hat{y}_i \approx y_i$ untuk $i = 1, \ldots, n$—maka Galat Baku Residual akan menjadi kecil, dan kita dapat menyimpulkan bahwa model tersebut sangat cocok dengan data.
    - Di sisi lain, jika $\hat{y}_i$ sangat jauh dari y_i untuk satu atau lebih pengamatan, maka Galat Baku Residual mungkin cukup besar, menunjukkan bahwa model tidak sesuai dengan data dengan baik.

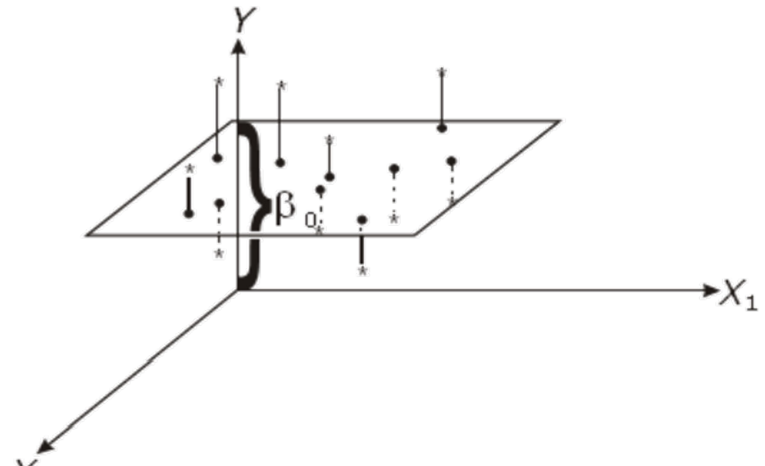
- **Statistik $\boldsymbol{R^2}$**
  - Galat Baku Residual memberikan ukuran mutlak ketidaksesuaian model dengan data.
  - Tetapi karena diukur dalam satuan Y, tidak selalu jelas apa yang dimaksud dengan Galat Baku Residual yang baik.
  - Statistik R^2 memberikan alternatif ukuran kecocokan model.
  - Bentuknya berupa proporsi (proporsi ragam yang dijelaskan) sehingga selalu mengambil nilai antara 0 dan 1, dan tidak bergantung pada skala Y.

$$R^2 = \frac{JKT - JKG}{JKT} = 1 - \frac{JKG}{JKT} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

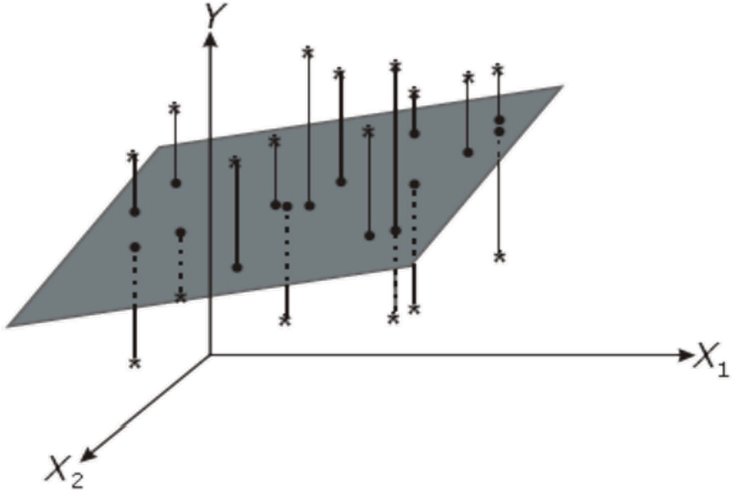
  - R^2 mengukur proporsi keragaman dalam Y yang dapat dijelaskan dengan menggunakan X.

Regresi Linier Berganda

- Analisis regresi linear berganda:
  - Secara umum, kita memodelkan peubah respon Y sebagai fungsi linier dari k peubah prediktor (X) sebagai:
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$
  - Atau dalam notasi matriks $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$
  - Jika kita memiliki dua variabel X, model dapat diilustrasikan sebagai berikut



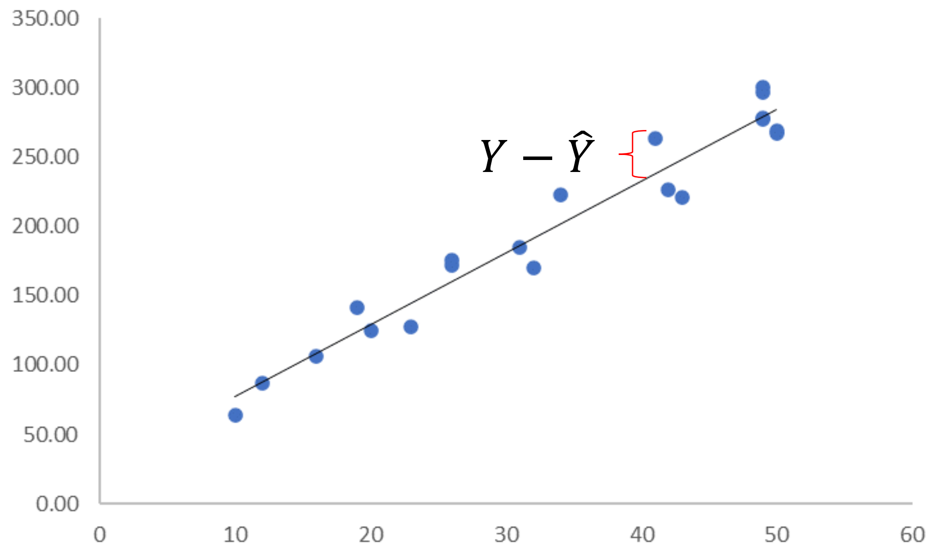
Tidak ada hubungan



Ada hubungan

- Pendugaan koefisien regresi:
  - Pendugaan koefisien regresi diperoleh dengan meminimumkan jumlah kuadrat galat (residual) → OLS (Ordinary Least Square) atau MKT (Metode Kuadrat Terkecil)
  - Dalam hal ini dicari dugaan dari $\beta_j, j = 0,1,2 \ldots, k$ yang meminimumkan $\sum_i \varepsilon^2$, dengan $\varepsilon = Y - \hat{Y}$, yang dalam notasi matriks diperoleh

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$



Asumsi model regresi linear

Nilai mean dari peubah Y dimodelkan secara akurat oleh fungsi linier dari peubah-peubah X

$$Y = \beta_0 + \underbrace{\beta_1 X_1 + \cdots + \beta_k X_k}_{} + \varepsilon$$

Galat bersifat independen/ saling bebas (tidak ada autokorelasi)

Antar peubah X tidak ada multikolinearitas

Galat acak diasumsikan menyebar normal dengan nilai tengah nol dan memiliki ragam yang konstan σ^2 (ragam homogen)

Menilai Akurasi Model

- Dua ukuran numerik yang paling umum untuk mengidentifikasi kecocokan model adalah Galat Baku Residual dan R^2.

- Nilai-nilai ini dihitung dan ditafsirkan dengan cara yang sama seperti untuk regresi linier sederhana.

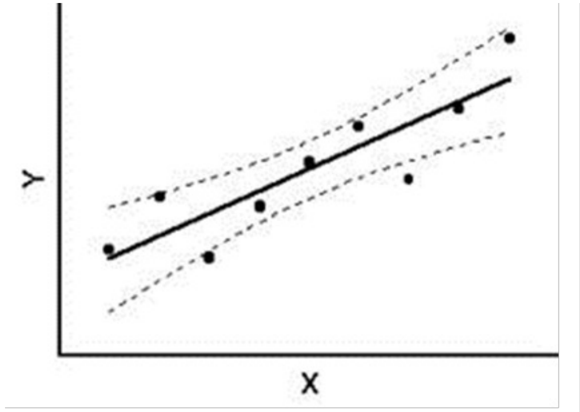
$$\text{Galat Baku Residual} = \sqrt{\frac{1}{n-p-1} JKG} = \sqrt{\frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Menilai Akurasi Prediksi

- Pemodelan prediktif merupakan masalah pengembangan model menggunakan data historis untuk membuat prediksi pada data baru yang belum dimiliki jawabannya.

- Berikut ini beberapa ukuran evaluasi dalam konteks prediksi untuk model regresi
  1. Selang Kepercayaan Prediksi (Confidence Interval)

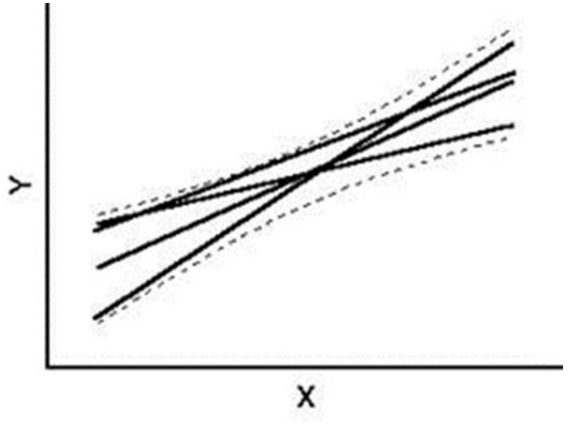
$$\hat{y}_h \pm t_{\left(1-\frac{\alpha}{2};n-2\right)} \times \sqrt{KTG\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$$



Selang Kepercayaan

  2. Selang Prediksi (Prediction Interval)

$$\hat{y}_h \pm t_{\left(1-\frac{\alpha}{2};n-2\right)} \times \sqrt{KTG\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$$



Selang Prediksi

3. MSE (Mean Squared Error) atau KTG

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

4. RMSE (Root Mean Squared Error)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

5. MAE (Mean Absolute Error)

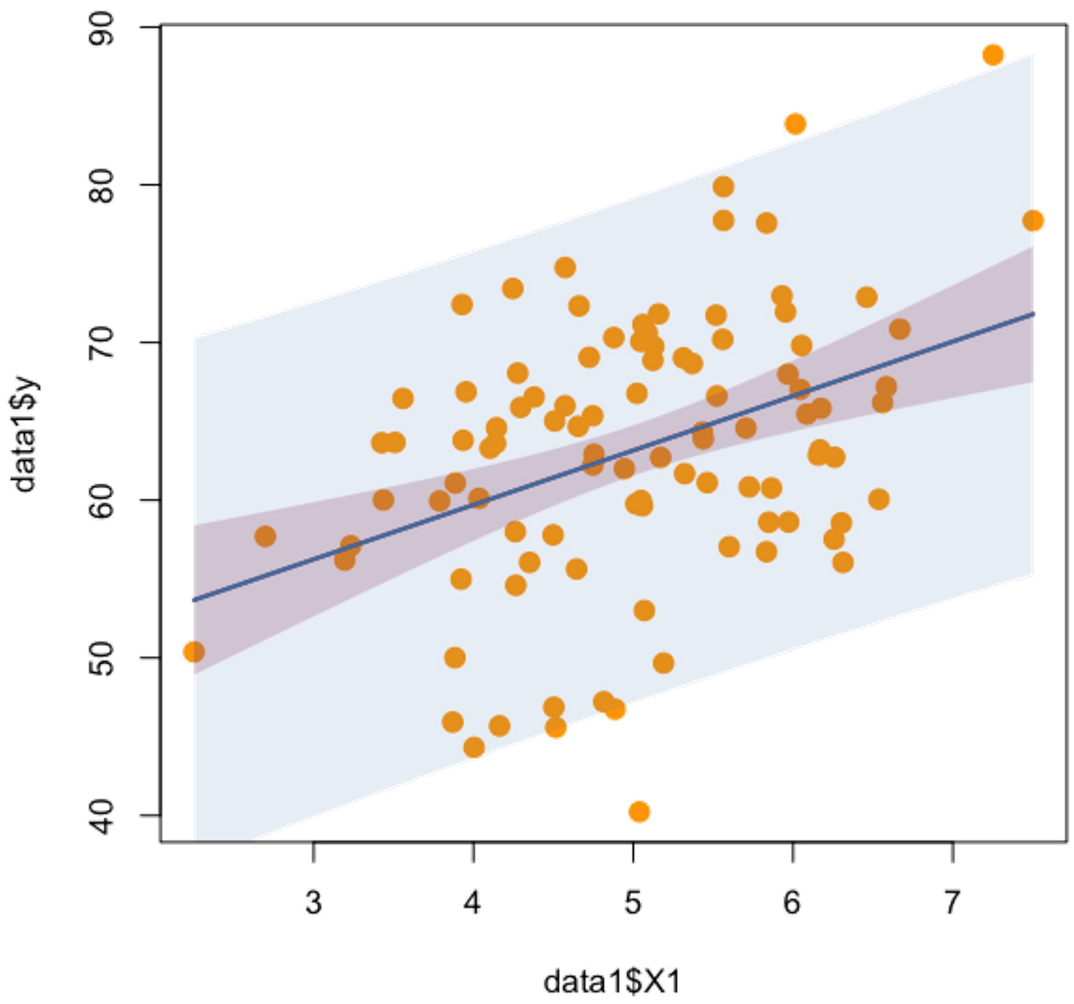
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Contoh 1

```
#simulasi analisis regresi
beta <- c(3,5,7)
set.seed(123456)
Xa <- matrix(rnorm(200,5,1),100,2)
X <- cbind(1,Xa)
e <- rnorm(100,0,4)
y <- X%*%beta+e
data1 <-
data.frame(y=y,X1=Xa[,1],X2=Xa[,2])

##model regresi
mod1 <- lm(y~X1+X2,data=data1)
summary(mod1)
```

```
> summary(mod1)

Call:
lm(formula = y ~ X1 + X2, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.8883 -2.4990  0.0499  2.3688 10.2160 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.2994     3.2282   0.712    0.478    
X1            4.9915     0.4102  12.168   <2e-16 ***
X2            7.2377     0.4142  17.473   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.961 on 97 degrees of freedom
Multiple R-squared:  0.7965,	Adjusted R-squared:  0.7923 
F-statistic: 189.8 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
plot(data1$X1,data1$y,pch=20,col="orange", cex=2)
library(DescTools)
mod2 <- lm(y~X1,data=data1)
lines(mod2,col="red") #conf interval
lines(mod2,col="steelblue", pred.level=0.95) #pred interval
```



```r
##evaluasi
yduga <- fitted.values(mod1)
MSE <- mean((data1$y-yduga)^2)
RMSE <- sqrt(MSE)
MAE <- mean(abs(data1$y-yduga))
```

```
> MSE
[1] 15.21544
> RMSE
[1] 3.900697
> MAE
[1] 3.029892
```

Regresi Logistik

- Model regresi yang diterapkan untuk peubah respon Y dengan skala kategorik

- Peubah respon Y dapat terdiri dari 2 kategori (biner), maupun lebih dari 2 kategori (multinomial) yang dapat urutan (ordinal) maupun tidak (nominal)

- Daripada memodelkan respon Y ini secara langsung, regresi logistik memodelkan peluang bahwa Y termasuk dalam kategori tertentu

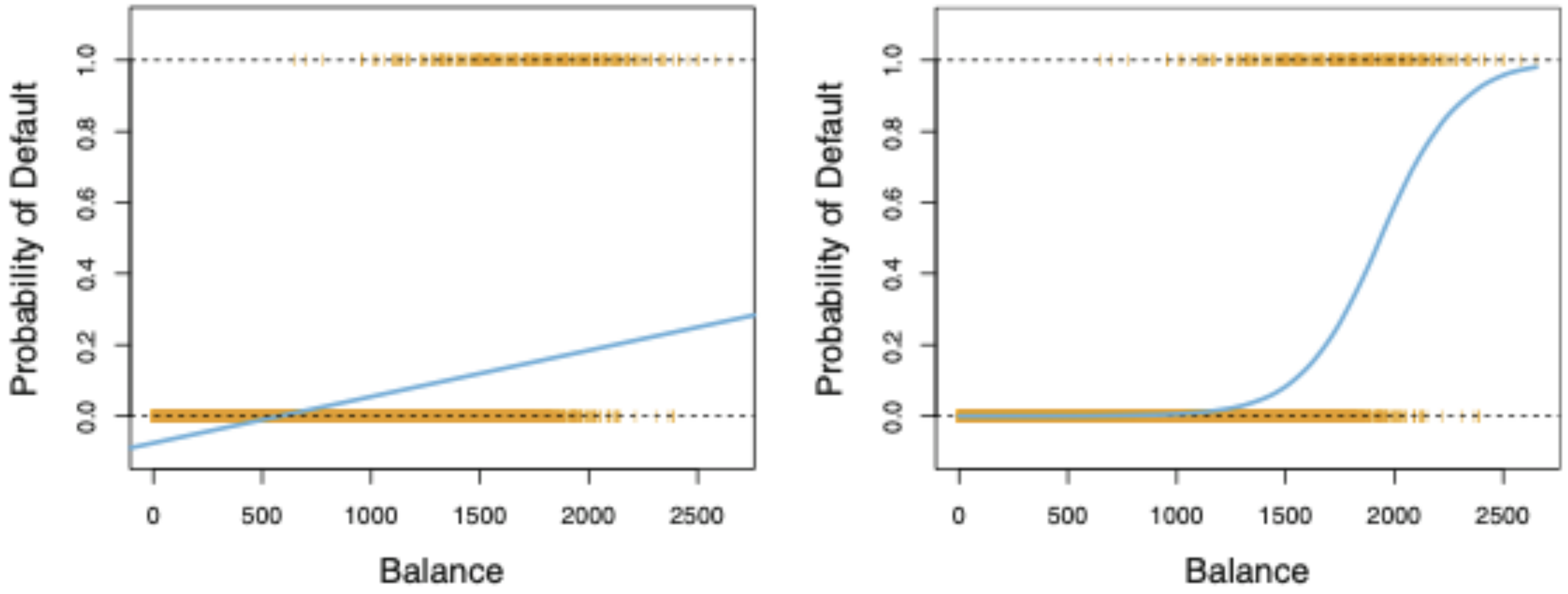


FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default`*(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

Model Logistik

- Model hubungan antara $p(X) = \mathrm{P}(Y = 1|X)$ dan X, dalam hal ini digunakan kode 0 atau 1 untuk kategori peubah respon

$$p(X) = \beta_0 + \beta_1 X$$

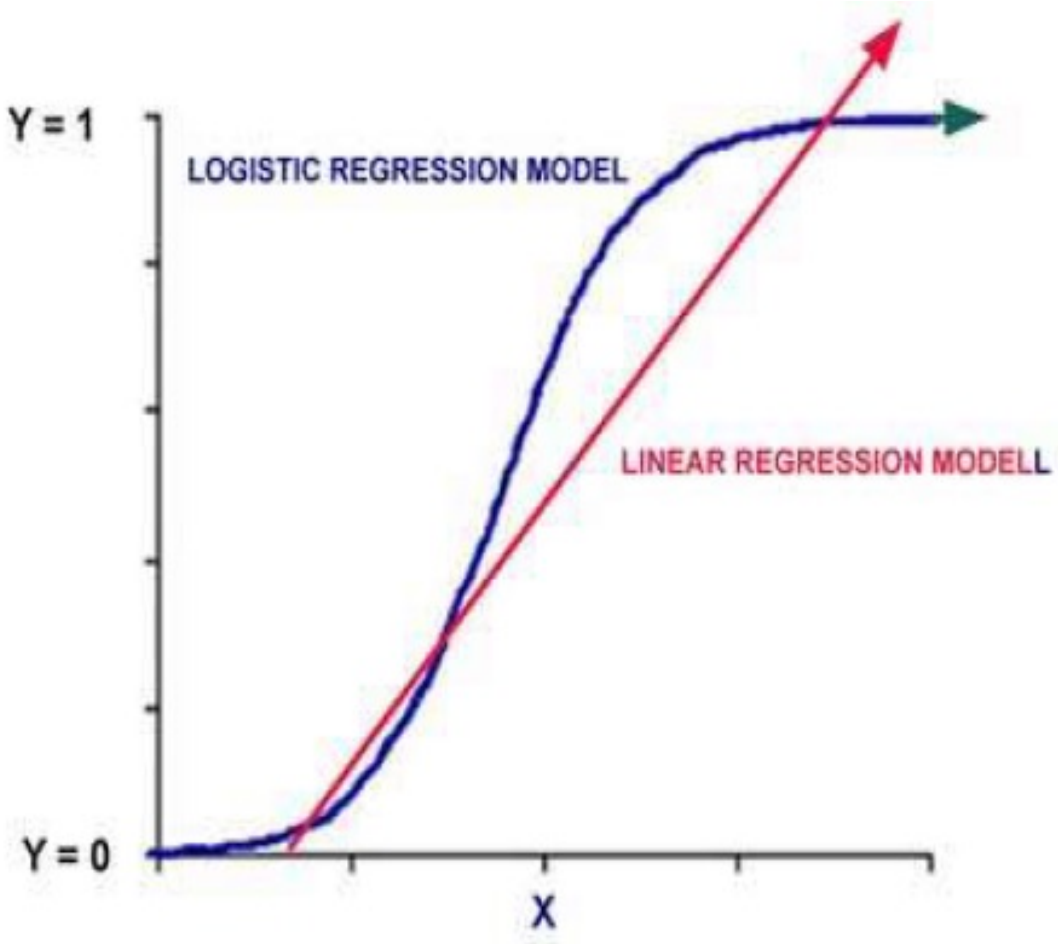
- Dengan fungsi logistik:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \quad \rightarrow \quad \frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X}$$

- Nilai $\frac{p(X)}{1-p(X)}$ disebut dengan odds, berkisar dari 0 s.d ∞
- Dengan menerapkan logaritma natural, maka diperoleh persamaan log-odds atau logit

$$\ln\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

- Dalam model regresi logistik, meningkatkan X sebesar satu unit mengubah log-odds sebesar β_1, atau setara dengan mengalikan odds dengan e^{β_1}



- $\beta > 0$ maka kurva akan naik
- $\beta < 0$ maka kurva akan turun
- Jika $\beta = 0$ maka nilai berapapun nilai $p(X)$ konstan, berapapun nilai X → kurva akan menjadi garis horizontal

$$\ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

Interpretasi nilai β_1

- 1 kenaikan X akan meningkatkan $\ln\left(\frac{p(X)}{1-p(X)}\right)$ sebesar β_1 satuan
- Dengan kata lain, 1 kenaikan X akan meningkatkan $\left(\frac{p(X)}{1-p(X)}\right)$ sebesar e^{β_1} satuan
- $\left(\frac{p(X)}{1-p(X)}\right)$ disebut dengan odds → peluang dari kejadian terjadi dibagi dengan peluang dari kejadian tidak terjadi
- Artinya odds akan meningkat secara sebesar e^{β_1} untuk setiap kenaikan 1 unit X
- e^{β_1}: odds ratio (OR)

$$OR = e^{\beta_1} = \frac{odds(X = x + 1)}{odds(X = x)}$$

- Odds Ratio menunjukkan seberapa besar kemungkinan, sehubungan dengan peluang, suatu peristiwa tertentu terjadi dalam satu kelompok dibandingkan dengan kejadiannya di kelompok lain.

- Pendugaan koefisien regresi logistik
  - Menggunakan metode maksimum likelihood, dengan fungsi likelihood:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=1} \left(1 - p(x_{i'})\right)$$

  - Sehingga, dugaan $\hat{\beta}_0$ dan $\hat{\beta}_1$ dipilih yang memaksimumkan nilai fungsi likelihood

- Prediksi
  - Peluang $p(X)$ dapat diprediksi dengan:

$$\hat{p}(X) = \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 X}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 X}}$$

  - Pada umumnya, jika $\hat{p}(X) \geq 0.5 \ \rightarrow \ \hat{y} = 1$ dan sebaliknya

- Model regresi logistik untuk lebih dari satu prediktor

$$\ln\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \ \ \rightarrow p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

Evaluasi Prediksi

Confusion Matrix

		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

TN	True Negative
FP	False Positive
FN	False Negative
TP	True Positive

Model Performance

Accuracy $= \dfrac{TN+TP}{TN+FP+FN+TP}$

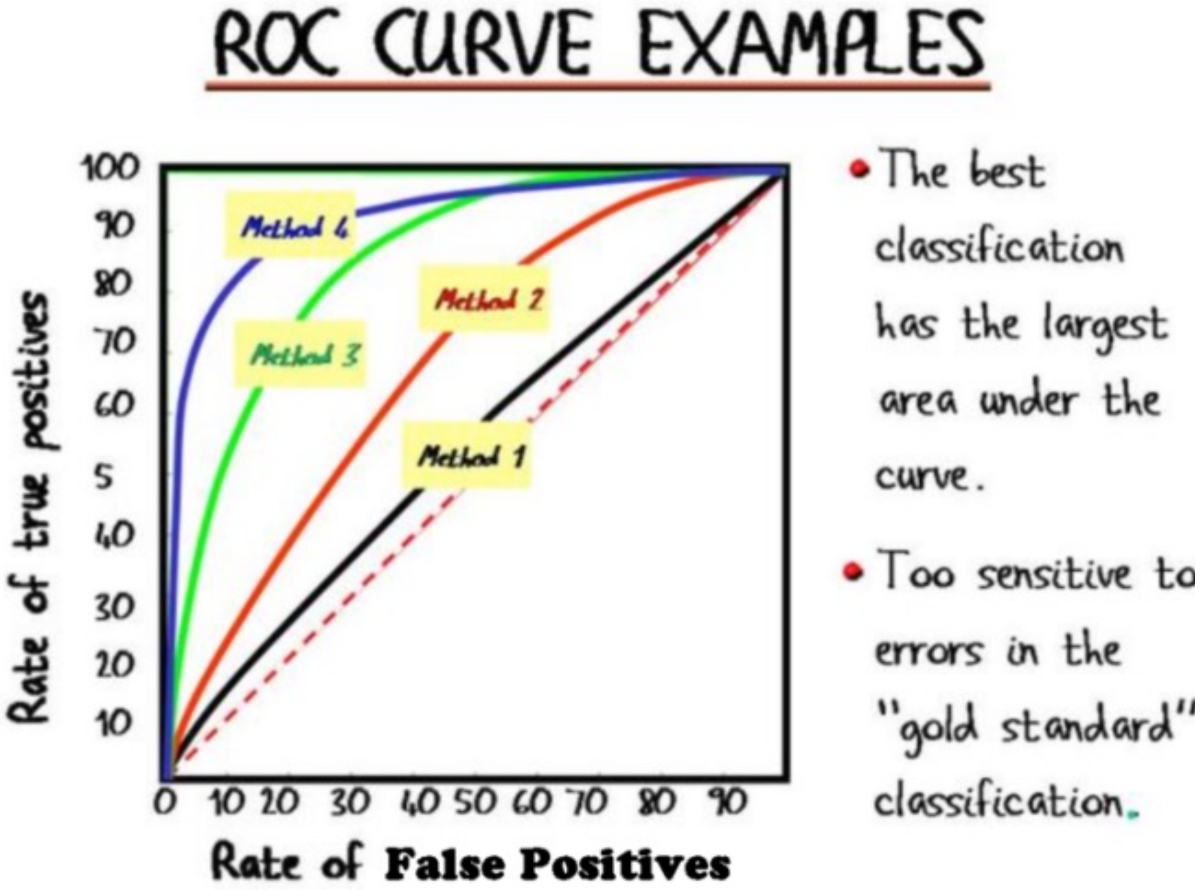
Precision $= \dfrac{TP}{FP+TP}$

Sensitivity $= \dfrac{TP}{TP+FN}$

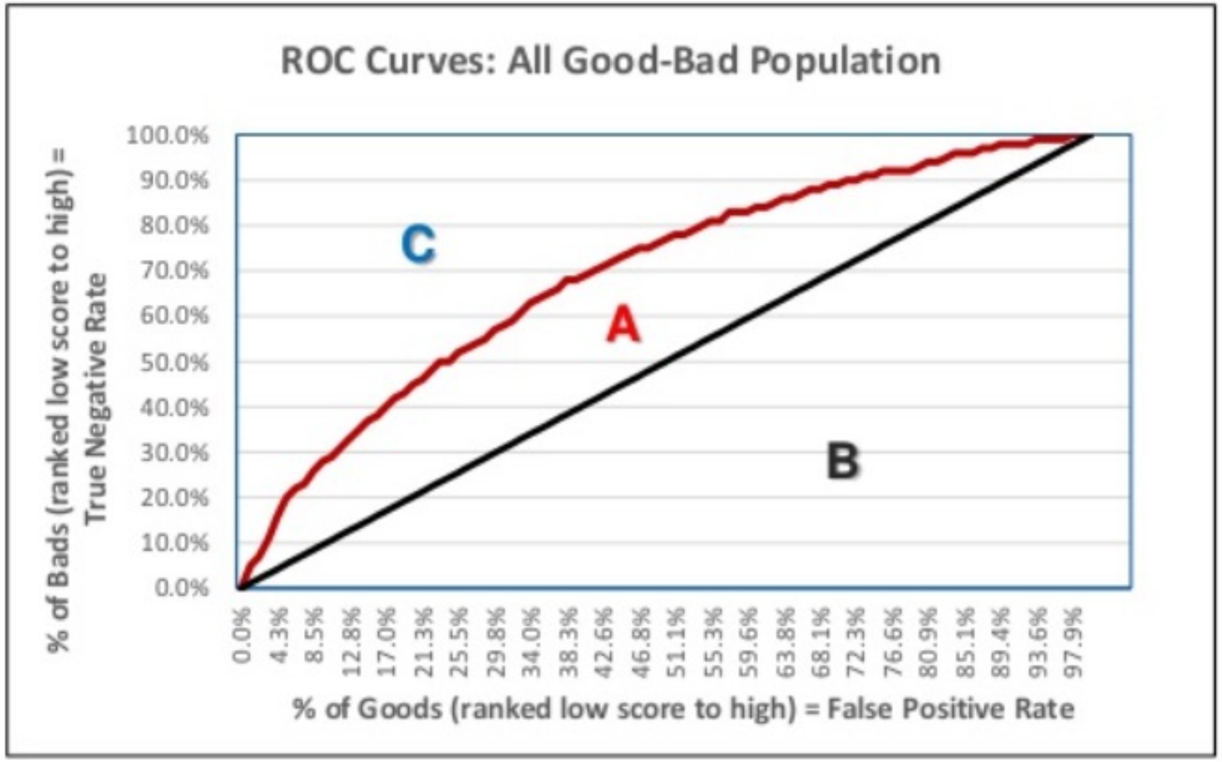
Specificity $= \dfrac{TN}{TN+FP}$

Receiver operating characteristic (ROC)

Sensitivity



1 - Specificity



AUC = Area A + Area B

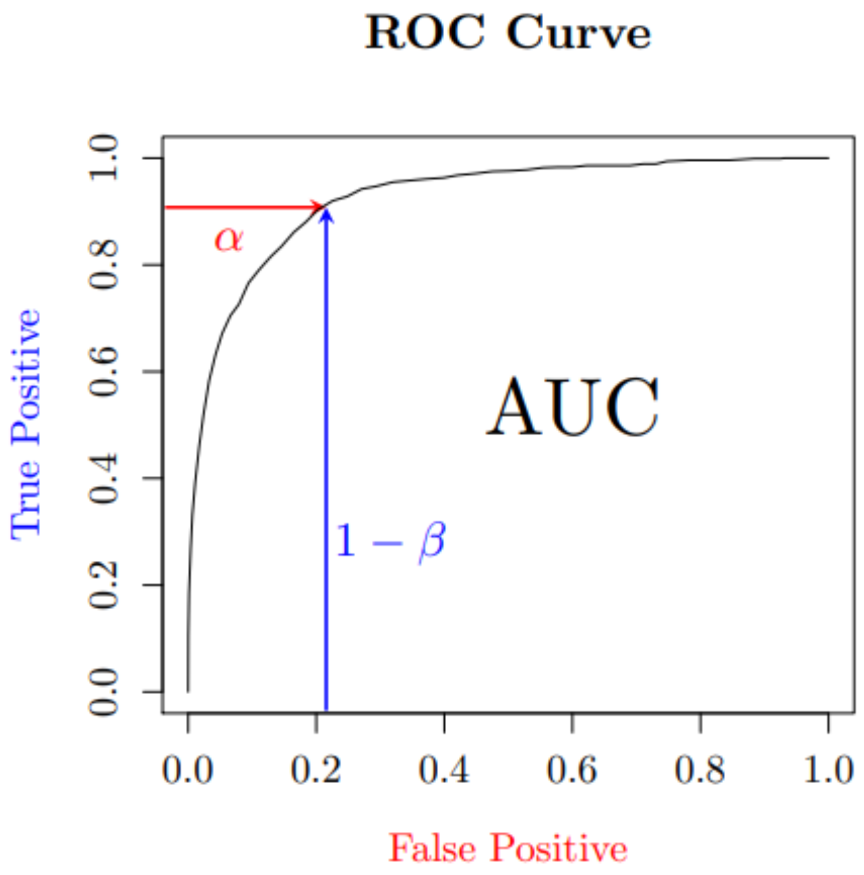
Gini Coefficient = 2 x AUC - 1

Receiver operating characteristic (ROC)

	Sick	Healthy
Treating as sick	True Positive	False Positive
Treating as healthy	False Negative	True Negative

$$\text{Power} = \frac{TP}{TP+FN} = 1-\beta$$

$$\text{False Positive Rate} = \frac{FP}{FP+TN} = \alpha$$



Contoh 2: Latihan dengan R

German Credit Data Set

The German Credit Data contains data on 20 variables and the classification whether an applicant is considered a Good or a Bad credit risk for 1000 loan applicants.

The data contains 1000 observations (700 good loans, 300 bad loans)

A predictive model developed on this data is expected to provide a bank manager guidance for making a decision whether to approve a loan to a prospective applicant based on his/her profiles.

```
#Membaca data
#Membaca data
install.packages("fairml")
library(fairml)

data(german.credit)
str(german.credit)

str(germancredit)
```

```
> str(german.credit)
'data.frame':	1000 obs. of  21 variables:
 $ Account_status          : Factor w/ 4 levels "< 0 DM",">= 200 DM",..: 1 3 4 $
 $ Duration                : num  6 48 12 42 24 36 24 36 12 30 ...
 $ Credit_history          : Factor w/ 5 levels "all credits at this bank paid $
 $ Purpose                 : Factor w/ 10 levels "business","car (new)",..: 8 8$
 $ Credit_amount           : num  1169 5951 2096 7882 4870 ...
 $ Savings_bonds           : Factor w/ 5 levels "< 100 DM",">= 1000 DM",..: 5 1$
 $ Present_employment_since: Factor w/ 5 levels "< 1 year",">= 7 years",..: 2 3$
 $ Installment_rate        : num  4 2 2 2 3 2 3 2 2 4 ...
 $ Other_debtors_guarantors: Factor w/ 3 levels "co-applicant",..: 3 3 3 2 3 3 $
 $ Resident_since          : num  4 2 3 4 4 4 4 2 4 2 ...
 $ Property                : Factor w/ 4 levels "building society savings agree$
 $ Age                     : num  67 22 49 45 53 35 53 35 61 28 ...
 $ Other_installment_plans : Factor w/ 3 levels "bank","none",..: 2 2 2 2 2 2 2$
 $ Housing                 : Factor w/ 3 levels "rent","own","for free": 2 2 2 $
 $ Existing_credits        : num  2 1 1 1 2 1 1 1 1 2 ...
 $ Job                     : Factor w/ 4 levels "management / self-employed / h$
 $ People_maintenance_for  : num  1 1 2 2 2 2 1 1 1 1 ...
 $ Telephone               : Factor w/ 2 levels "none","yes": 2 1 1 1 1 2 1 2 1$
 $ Foreign_worker          : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2$
 $ Credit_risk             : Factor w/ 2 levels "BAD","GOOD": 2 1 2 2 1 2 2 2 2$
 $ Gender                  : Factor w/ 2 levels "Female","Male": 1 2 1 1 1 1 1 $
```

Peubah:

1. Account_status: a factor with four levels representing the amount of money in the account or "no chcking account".
2. Duration: a continuous variable, the duration in months.
3. Credit_history: a factor with five levels representing possible credit history backgrounds.
4. Purpose: a factor with ten levels representing possible reasons for taking out a loan.
5. Credit_amount: a continuous variable.
6. Savings_bonds: a factor with five levels representing amount of money available in savings and bonds or "unknown / no savings account".
7. Present_employment_since: a factor with five levels representing the length of tenure in the current employment or "unemployed".
8. Installment_rate: a continuous variable, the installment rate in percentage of disposable income.
9. Other_debtors_guarantors: a factor with levels "none", "co-applicant" and "guarantor".
10. Resident_since: a continuous variable, number of years in the current residence.
11. Property: a factor with four levels describing the type of property to be bought or "unknown / no property".
12. Age: a continuous variable, the age in years.
13. Other_installment_plans: a factor with levels "bank", "none" and "stores".
14. Housing: a factor with levels "rent", "own" and "for free".
15. Existing_credits: a continuous variable, the number of existing credit lines at this bank.
16. Job: a factor with four levels for different job descriptions.
17. People_maintenance_for: a continuous variable, the number of people being liable to provide maintenance for.
18. Telephone: a factor with levels "none" and "yes".
19. Foreign_worker: a factor with levels "no" and "yes".
20. Credit_risk: a factor with levels "BAD" and "GOOD". → Response variable
21. Gender: a factor with levels "Male" and "Female".

```
##Konstruksi model dengan data training 80%, dan data testing 20%
library(caret)
set.seed(12420246)
in.train <- createDataPartition(as.factor(german.credit$Credit_risk), p=0.8, list=FALSE)
german.credit.train <- german.credit[in.train,]
german.credit.test <- german.credit[-in.train,]

credit.glm0 <- glm(Credit_risk ~ ., family = binomial, german.credit.train)
credit.glm.step <- step(credit.glm0)
credit.glm.step$anova

summary(credit.glm.step)
```

```
Step:  AIC=776.26
Credit_risk ~ Account_status + Duration + Credit_history + Purpose +
    Credit_amount + Savings_bonds + Present_employment_since +
    Installment_rate + Other_debtors_guarantors + Other_installment_plans +
    Housing + Foreign_worker + Gender

                           Df Deviance    AIC
<none>                          704.26 776.26
- Gender                    1   706.37 776.37
- Other_debtors_guarantors  2   709.13 777.13
- Housing                   2   709.65 777.65
- Credit_amount             1   708.79 778.79
- Other_installment_plans   2   712.36 780.36
- Foreign_worker            1   711.15 781.15
- Present_employment_since  4   717.36 781.36
- Installment_rate          1   711.82 781.82
- Duration                  1   714.48 784.48
- Credit_history            4   726.07 790.07
- Savings_bonds             4   726.68 790.68
- Purpose                   9   738.98 792.98
- Account_status            3   758.74 824.74
```

```
> credit.glm.step$anova
                      Step Df  Deviance Resid. Df Resid. Dev      AIC
1                          NA        NA       753   697.4557 791.4557
2               - Property  3 1.2504839       756   698.7062 786.7062
3                    - Job  3 2.2388200       759   700.9450 782.9450
4              - Telephone  1 0.4198912       760   701.3649 781.3649
5 - People_maintenance_for  1 0.5168207       761   701.8817 779.8817
6         - Resident_since  1 0.5915425       762   702.4732 778.4732
7                    - Age  1 0.8074193       763   703.2807 777.2807
8      - Existing_credits  1 0.9775937       764   704.2583 776.2583
```

```r
###Final model
credit.glm.final <- glm(Credit_risk ~ Account_status + Duration + Credit_history + Purpose +
    Credit_amount + Savings_bonds + Present_employment_since +
    Installment_rate + Other_debtors_guarantors + Other_installment_plans +
    Housing + Foreign_worker + Gender, family = binomial, german.credit.train)

summary(credit.glm.final)
```

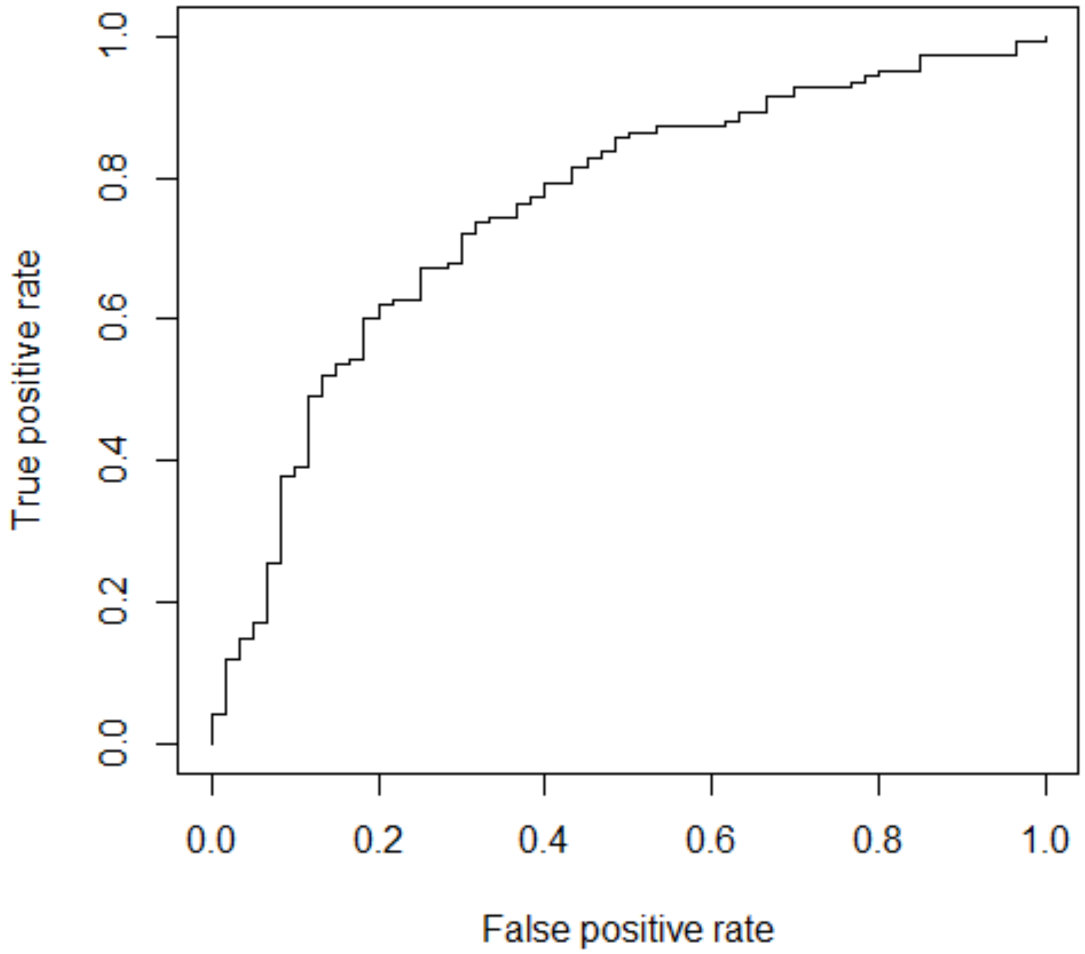
```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 977.38  on 799  degrees of freedom
Residual deviance: 704.26  on 764  degrees of freedom
AIC: 776.26

Number of Fisher Scoring iterations: 5
```

```
###Evaluation
#ConfusionMatrix
fit.final <- fitted.values(credit.glm.final)
pred.final <- ifelse(fit.final>=0.5,"GOOD","BAD")

tab <- table(german.credit.train$Credit_risk,pred.final, dnn = c("Truth", "Predicted"))
tab
acc <- sum(diag(tab))/sum(tab)
acc
```

```
> tab
      Predicted
Truth  BAD GOOD
  BAD  132  108
  GOOD  56  504
```

```
> acc
[1] 0.795
```

```
install.packages("ROCR")
library(ROCR)
pred<-
prediction(predict.glm(credit.glm.final,german.credit.test),german.credit.test$Credit_risk)
perf <- performance(pred,"tpr","fpr")
plot(perf)
```




```
AUC.final<-performance(pred, measure = "auc")@y.values[[1]]
AUC.final
[1] 0.7552381
```

Latihan 1 (B/S)

1. Jika $\hat{y}_i$ sangat jauh dari y_i untuk satu atau lebih pengamatan, maka Galat Baku Residual akan bernilai kecil yang menunjukkan bahwa model sesuai dengan data dengan baik.
2. Jika selisih $\hat{y}_i$ terhadap y_i besar untuk banyak pengamatan, maka R^2 akan bernilai besar.
3. Sensitivity merupakan perbandingan dari True Positive dengan True Negative yang dijumlahkan dengan False Positive.
4. Semakin besar nilai AUC maka semakin kecil nilai Gini.
5. Presisi merupakan rasio dari True Positive dengan total dari pengamatan.

Latihan 2 (Isian Singkat)

Diketahui tabel kontingensi:

Status Kelulusan	Laki-laki	Perempuan	Total
Tidak Lulus	35	40	75
Lulus	15	10	25
Total	50	50	**100**

1. Tentukan nilai odds laki-laki yang lulus
2. Tentukan nilai odds perempuan yang lulus
3. Tentukan nilai odds ratio yang lulus untuk perempuan terhadap laki-laki

Form Penyetaraan Kegiatan ke MK TPM

https://ipb.link/penyetaraan-tpm

Terima kasih ☺