Departemen Statistika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Institut Pertanian Bogor

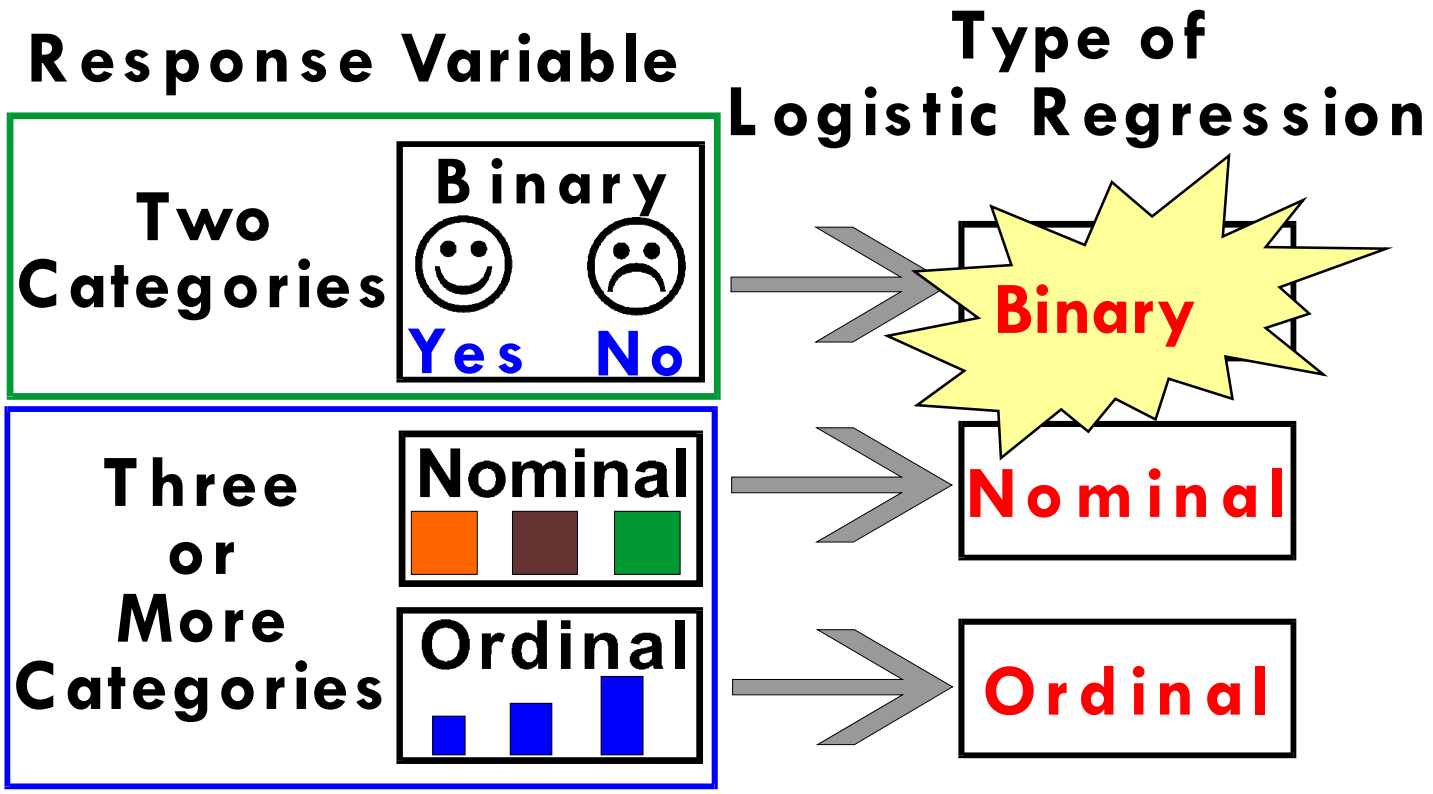
PEMODELAN KLASIFIKASI

PERTEMUAN #5/6

REGRESI LOGISTIK

Bagus Sartono
bagusco@ipb.ac.id bagusco@gmail.com

Types of Logistic Regression

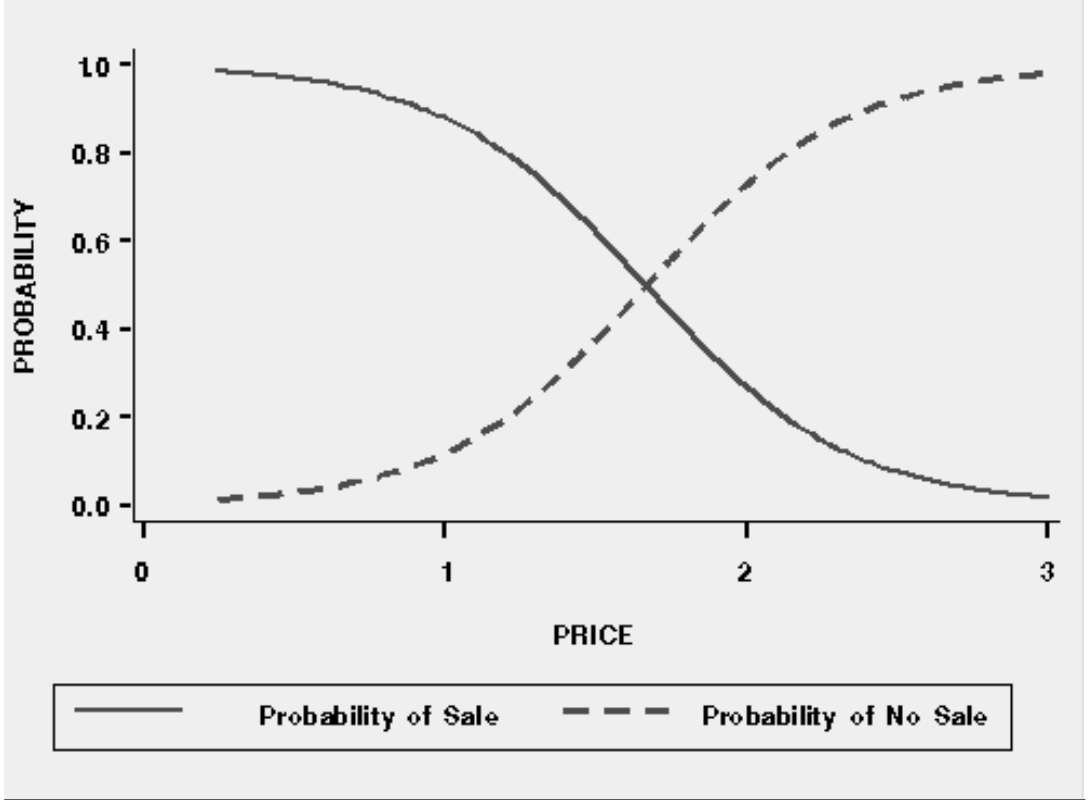


What Does Logistic Regression Do?

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

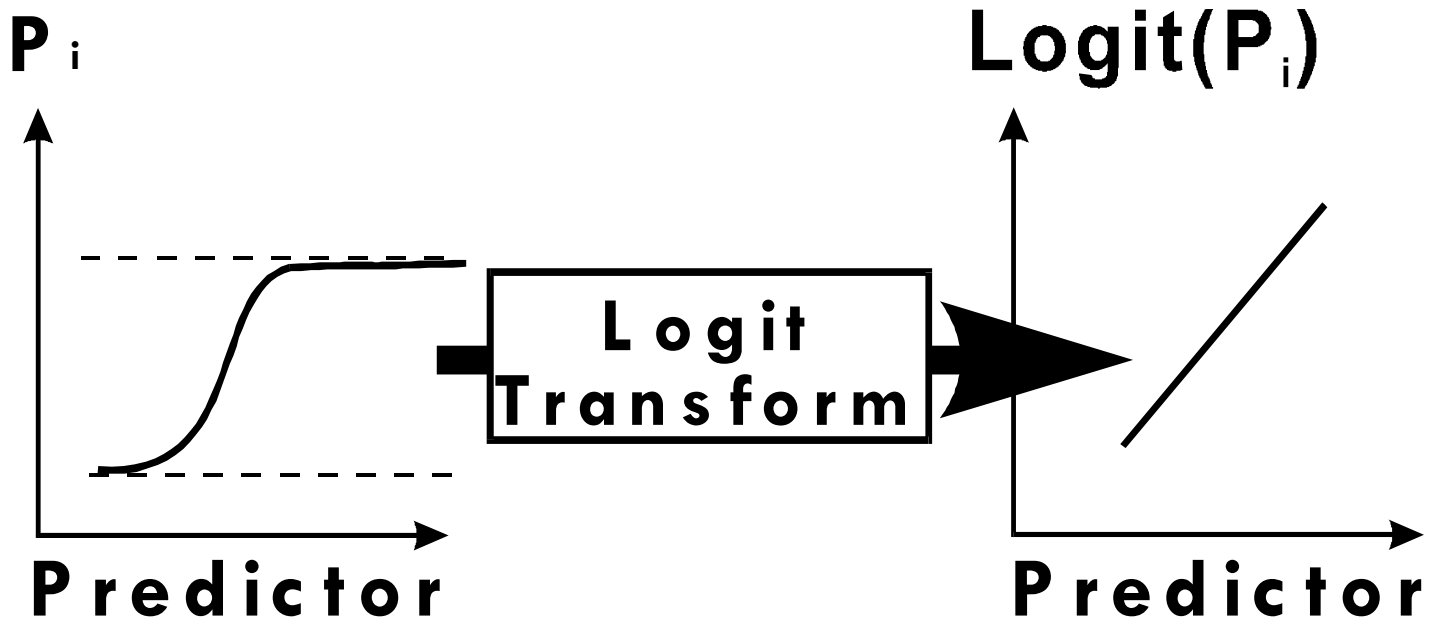
In other words, logistic regression is designed to describe probabilities associated with the values of the response variable.

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.

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

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).

Logistic Regression Model

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

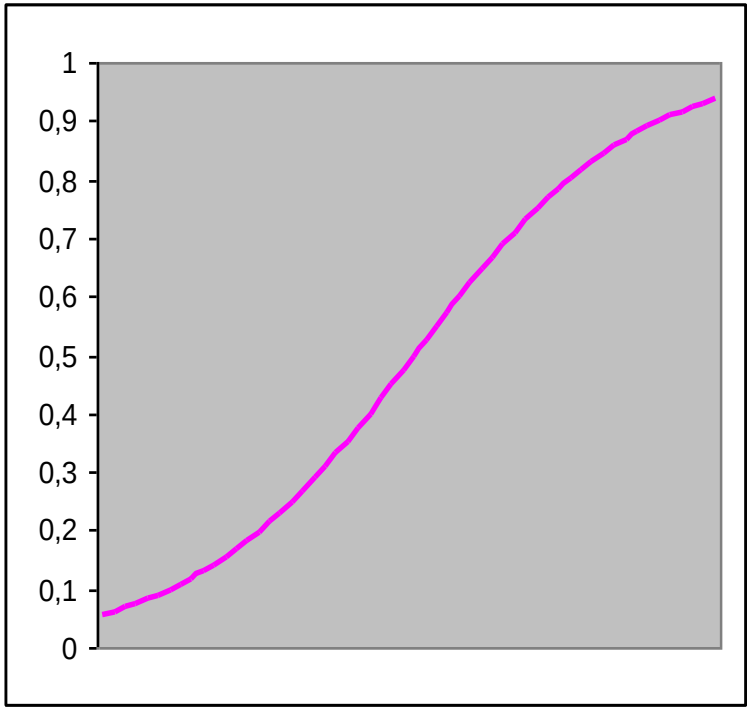
where

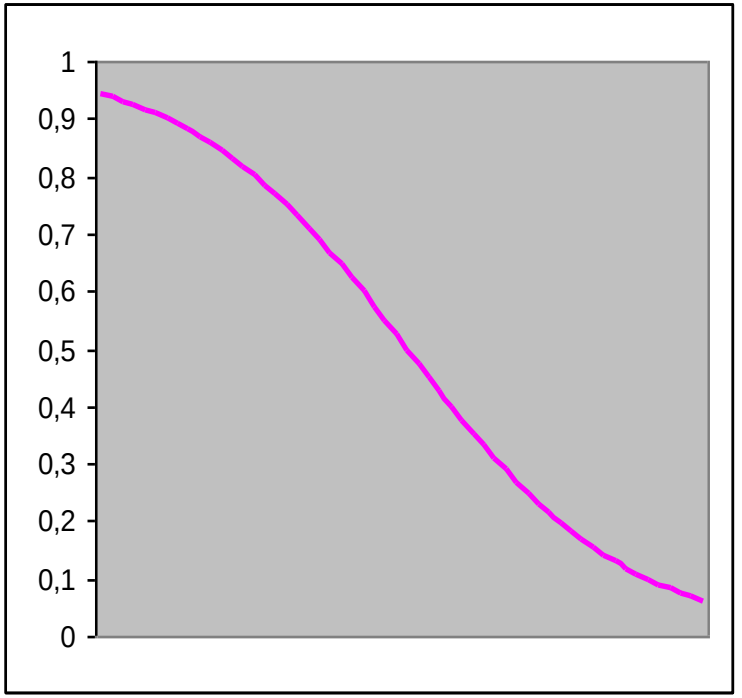
$\text{logit}(p_i)$	logit transformation of the probability of the event
β_0	intercept of the regression line
β_1	slope of the regression line.

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





Measures of Association : Odds Ratio

Measures of association are statistics that indicate the strength of the association between two variables.

One advantage of statistical modeling is that measures of association are often functions of the estimated parameters.

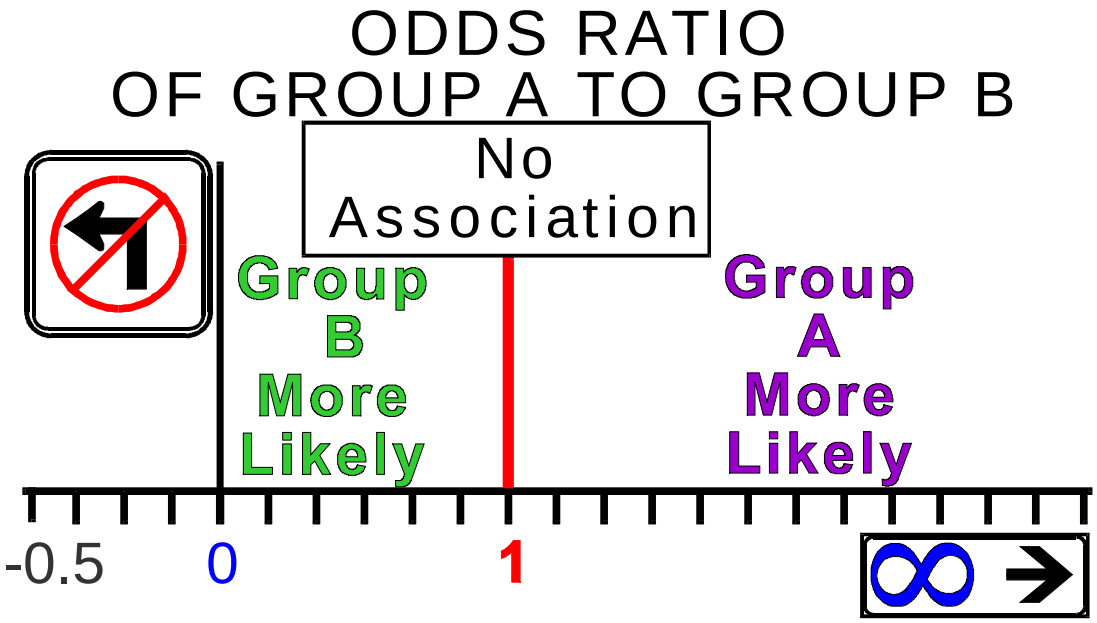
One measure of association that can be computed in logistic regression is the odds ratio.

WHAT IS AN ODDS RATIO?

An *odds ratio* indicates how much more likely, with respect to odds, a certain event occurs in one group relative to its occurrence in another group.

Example: How much more likely are females to purchase 100 dollars or more in products compared to males?

PROPERTIES OF THE ODDS RATIO



ODDS RATIO FROM A LOGISTIC REGRESSION MODEL

$$\psi_{ab} = e^{\beta(b-a)}$$

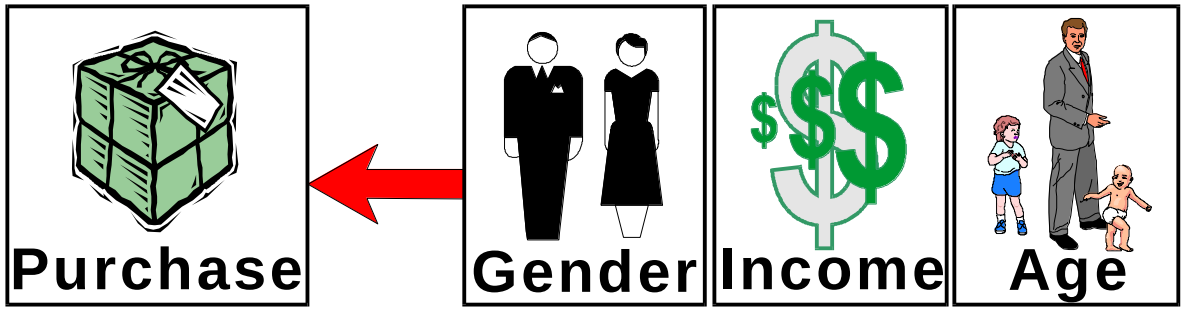
If $\beta > 0$ then $\psi > 1$

If $\beta = 0$ then $\psi = 1$

If $\beta < 0$ then $\psi < 1$

Note : ψ = odds ratio

Multiple Logistic Regression



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

Hypothesis Testing about Effect of X

Untuk memeriksa peranan variabel-variabel penjelas (X) dalam model, dilakukan pengujian terhadap parameter model (β). Pengujian secara simultan dilakukan menggunakan uji-G, sedangkan secara parsial menggunakan uji-Wald.

Simultaneous Test

Statistik uji-G adalah uji rasio kemungkinan (likelihood ratio test) yang digunakan untuk menguji peranan variabel penjelas di dalam model secara bersama-sama (Hosmer & Lemeshow, 1989). Rumus umum uji-G untuk menguji hipotesis :

H0 : $\beta 1 = \beta 2 = \ldots = \beta k = 0$

H1 : minimal ada satu β yang tidak sama dengan 0

adalah

$$G = 2\ln\left[\frac{likelihood \quad \tan pa \quad peubah \quad bebas}{likelihood \quad dengan \quad peubah \quad bebas}\right]$$

Statistik G ini, secara teoritis mengikuti sebaran $\chi 2$ dengan derajat bebas k.

Partial Test

Sementara itu, uji Wald digunakan untuk menguji parameter βi secara parsial. Hipotesis yang diuji adalah:

H0 : $\beta i = 0$

H1 : $\beta i \neq 0$

Formula statistik Wald adalah:

$$W = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Secara teori, statistik W ini mengikuti sebaran normal baku jika H0 benar.

```
wine <- read.csv("white_wine2.csv")
colnames(wine)

data <- wine[,c("alcohol", "density", "quality")]
head(data)
data$kelas.kualitas <- ifelse(data$quality > 6, 1, 0)


model.logistik <- glm(kelas.kualitas ~ alcohol+density,
                      data=data, family="binomial")

summary(model.logistik)

prob.prediksi <- predict(model.logistik, data, type="response")
prediksi <- ifelse(prob.prediksi>0.5, 1, 0)

library(caret)
confusionMatrix(prediksi, data$kelas.kualitas)
```