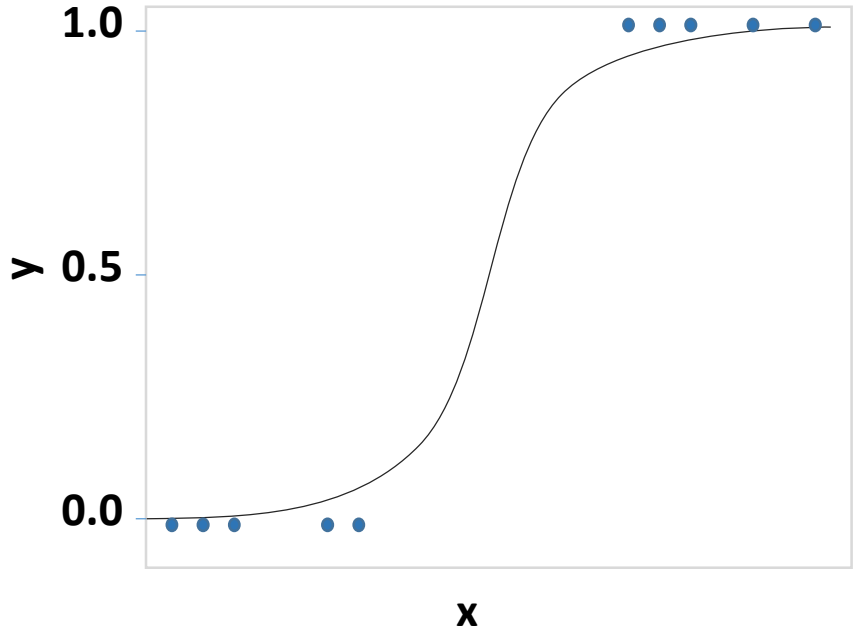
SUPERVISED LEARNING

~Logistic Regression~

Regresi Logistik

- Regresi logistik adalah suatu metode statistika untuk menjelaskan hubungan antara variabel respon yang bersifat kategori (kualitatif) dan variabel prediktor yang bersifat kualitatif maupun kuantitatif.
- Regresi logistik termasuk *generalized linear model.*
- Dalam *machine learning*, regresi logistik termasuk *supervised learning*.

Kurva regresi logistik dengan variabel respon bersifat biner



Macam-Macam Regresi Logistik

Berdasarkan Skala Data dari Variabel Respon

- **Regresi Logistik Biner**
  - seringkali hanya disebut dengan regresi logistik
  - variabel respon terdiri dari 2 kategori (biner)
- **Regresi Logistik Multinomial**
  - variabel respon memiliki skala data nominal dan terdiri lebih dari 2 kategori
- **Regresi Logistik Ordinal**
  - variabel respon memiliki skala data ordinal dan terdiri lebih dari 2 kategori

Kegunaan Regresi Logistik

- Regresi Logistik biasanya digunakan untuk klasifikasi, karena dapat menghasilkan nilai yang sesuai dengan probabilitas dari kategori tertentu.
- Regresi Logistik (juga disebut Regresi Logit) biasanya digunakan untuk memperkirakan probabilitas bahwa suatu observasi termasuk dalam kategori tertentu (misalnya, berapa probabilitas bahwa email ini adalah spam?).
  - Jika perkiraan probabilitas lebih besar dari 0,5 atau 50%, maka model memprediksi bahwa observasi termasuk kategori berlabel "1",
  - Jika probabilitas kurang dari 0,5, maka model memprediksi bahwa observasi termasuk kategori berlabel "0".

Geron, 2017

Analisis Regresi Logistik Biner

- Analisis regresi logistik biner atau selanjutnya disebut regresi logistik.
- Membuat model dengan p variabel prediktor

Transformasi logit

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)} \Rightarrow logit[\pi(x)] = log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- Menguji signifikansi parameter (statistik uji: Wald, *Likelihood Ratio Test*)
- Menilai kebaikan model (*goodness of fit*)
- Menginterpretasikan odds ratio (untuk mengetahui efek / hubungan)
- Memilih model (*backward elimination, forward selection,* dll)
- Mengevaluasi kinerja model (confusion matrix)

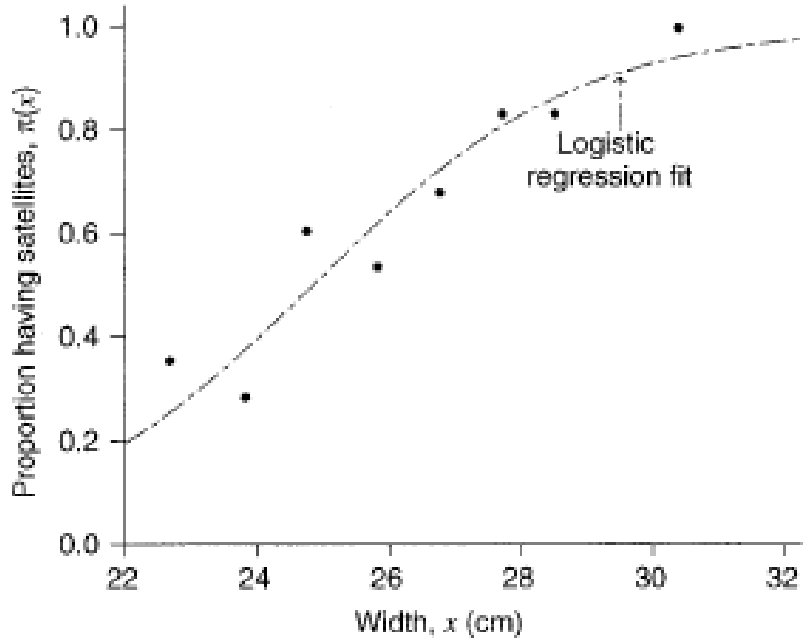
Model Regresi Logistik Biner

- Model regresi logistik dengan *p* variabel prediktor

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}$$

- Model regresi logistik ditransformasi dengan transformasi logit

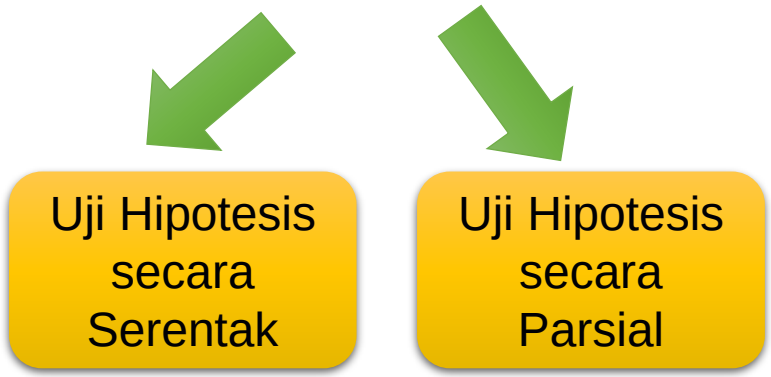
$$logit[\pi(x)] = log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$



Agresti (2002)

Uji Signifikansi Parameter

Ada atau tidak adanya hubungan yang nyata (signifikan) antara variabel prediktor dengan variabel respon dalam model dilakukan dengan pengujian signifikansi parameter.



Uji hipotesis secara serentak

- **Hipotesis**

$$H_0 = \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

- **Statistik Uji**

$$G = -2\ ln\left[\frac{\left[\frac{n_1}{n}\right]^{n_1}\left[\frac{n_0}{n}\right]^{n_0}}{\prod_{i=1}^{n}[\pi_i]^{y_i}[1-\pi_i]^{(1-y_i)}}\right]$$

dengan

$$n_0 = \sum_{i=1}^{n}(1-y_i)$$

dan

$$n_1 = \sum_{i=1}^{n} y_i$$

Daerah Kritis

H_0 ditolak jika

$$G > \chi^2_{(\alpha,p)}$$

atau

P-*value* < α

dengan tingkat signifikansi α dan derajat kebebasan *p*

Hosmer & Lemeshow (2000)

Uji hipotesis secara parsial

- **Hipotesis**

$$H_0: \beta_j = 0, j = 1,2,\dots,p$$
$$H_1: \beta_j \neq 0, j = 1,2,\dots,p$$

- **Statistik Uji**

$$W_j^2 = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}\right)^2, j = 1,2,\dots,p$$

- **Daerah Kritis**

H_0 ditolak jika

$$W^2 > \chi^2_{(\alpha,v)}$$

atau

P-*value* < α

dengan tingkat signifikansi α dan derajat kebebasan *v*

Hosmer & Lemeshow (2000)

Uji Kesesuaian Model (Goodness of Fit)

Hipotesis

H_0 : Tidak ada perbedaan antara hasil pengamatan dan hasil prediksi model

H_1 :Ada perbedaan antara hasil pengamatan dan hasil prediksi model

Statistik Uji

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)}$$

Daerah Kritis

H_0 ditolak jika $\hat{C} > \chi^2_{(\alpha, g-2)}$

atau P-*value* < α

Hosmer & Lemeshow (2000)

Interpretasi Model Regresi Logistik Biner

- **Variabel Prediktor (Kategorik)**

Interpretasi menggunakan nilai
Odds Ratio

$$OR = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}$$

sehingga

$$OR = e^{\beta_1}$$

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)}$$

Jika variabel prediktor bersifat biner / dikotomi:

	Variabel prediktor (X)	
Variabel respon (Y)	x = 1	x = 0
y = 1	$\pi(1) = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$	$\pi(0) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$
y = 0	$1 - \pi(1) = \frac{1}{1 + \exp(\beta_0 + \beta_1)}$	$1 - \pi(0) = \frac{1}{1 + \exp(\beta_0)}$

Hosmer & Lemeshow (2000)

Interpretasi

Variabel Prediktor (Kontinu)

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}$$

Koefisien β_1 mengacu pada pengaruh X_1 pada log odds untuk Y = 1, (variabel prediktor lainnya dianggap konstan).
Misalnya, **exp(β_1)** adalah efek multiplikatif pada peluang kenaikan 1 satuan di X_1, pada tingkat X lainnya tetap.

Agresti (2002)

Pemilihan Model

- Banyak prosedur pemilihan model yang ada, tidak ada yang selalu terbaik. Beberapa metode pemilihan model, yaitu *backward elimination*, *forward selection*, dll.
- Perhatian yang berlaku untuk regresi biasa berlaku untuk semua *generalized linear model*. Misalnya, model dengan beberapa prediktor mungkin menderita multikolinearitas, yaitu adanya korelasi yang erat antara variabel prediktor. Diantara dampak dari adanya multikolinieritas adalah membuat tidak ada satu variabel prediktor yang signifikan ketika semua variabel prediktor lain berada dalam model.

Agresti (2002)

Klasifikasi

Confusion Matrix

	Prediksi		Total
Aktual	y = 1	y = 0	
y = 1	*TP*	*FN*	*P*
y = 0	*FP*	*TN*	*N*

Implementasi

- Beberapa aplikasi penilaian kredit menggunakan regresi logistik untuk memodelkan probabilitas bahwa suatu subjek layak mendapatkan kredit. Misalnya, probabilitas bahwa subjek membayar tagihan tepat waktu dapat menggunakan prediktor seperti ukuran tagihan, pendapatan tahunan, pekerjaan, dan aspek lainnya.
- Seorang peneliti menggunakan regresi logistik untuk memprediksi kemungkinan seorang pasien menderita kanker payudara atau tidak berdasarkan beberapa prediktor.
- Menggunakan regresi logistik untuk mengklasifikasikan email sebagai spam atau tidak.
- dll

Data : Breast Cancer Wisconsin, dengan Python

```
# Making Confusion Matrix and calculating accuracy score
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score

model = LogisticRegression()

#Fit the model
model.fit(x_train, y_train)
y_pred = model.predict(x_test)

mylist = []
# Confusion Matrix
cm = confusion_matrix(y_test, y_pred)
# accuracy score
acc_logreg = accuracy_score(y_test, y_pred)

mylist.append(acc_logreg)
print(cm)
print(acc_logreg,'%')
```

```
[[104   0]
 [  4  48]]
0.9743589743589743 %
```

Data: Katarak, dengan R

```
> step1=glm(katarak~usia+jk+asamurat,family=binomial(logit),data=katarak)
> summary(step1)

Call:
glm(formula = katarak ~ usia + jk + asamurat, family = binomial(logit),
    data = katarak)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3916   0.3435   0.5920   0.6544   1.3749

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.6527     0.3695   4.473 7.70e-06 ***
usia         -1.8855     0.4142  -4.552 5.33e-06 ***
jk           -0.2204     0.4087  -0.539   0.5897
asamurat      1.1482     0.5916   1.941   0.0523 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 185.02  on 164  degrees of freedom
Residual deviance: 152.93  on 161  degrees of freedom
AIC: 160.93

Number of Fisher Scoring iterations: 5
```

```
> step2=glm(katarak~usia+asamurat,family=binomial(logit),data=katarak)
> summary(step2)

Call:
glm(formula = katarak ~ usia + asamurat, family = binomial(logit),
    data = katarak)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3346   0.3682   0.6263   0.6263   1.3234

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.5293     0.2850   5.366 8.04e-08 ***
usia         -1.8660     0.4123  -4.526 6.02e-06 ***
asamurat      1.1281     0.5876   1.920   0.0549 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 185.02  on 164  degrees of freedom
Residual deviance: 153.22  on 162  degrees of freedom
AIC: 159.22

Number of Fisher Scoring iterations: 5
```

Data: Katarak, dengan R (contd.)

Daftar Pustaka

- Agresti, A. 2002. *Categorical Data Analysis, 2nd ed. New Jersey: John Wiley & Sons, Inc.*
- Geron, A. 2017. *Hands-On Machine Learning with Scikit-Learn & Tensorflow. California*: *O'Reilly Media, Inc*.
- Hosmer, D.W. and Lemeshow, S. 2000. *Applied Logistic Regression, 2nd ed. New York: John Wiley & Sons, Inc.*

Terima Kasih