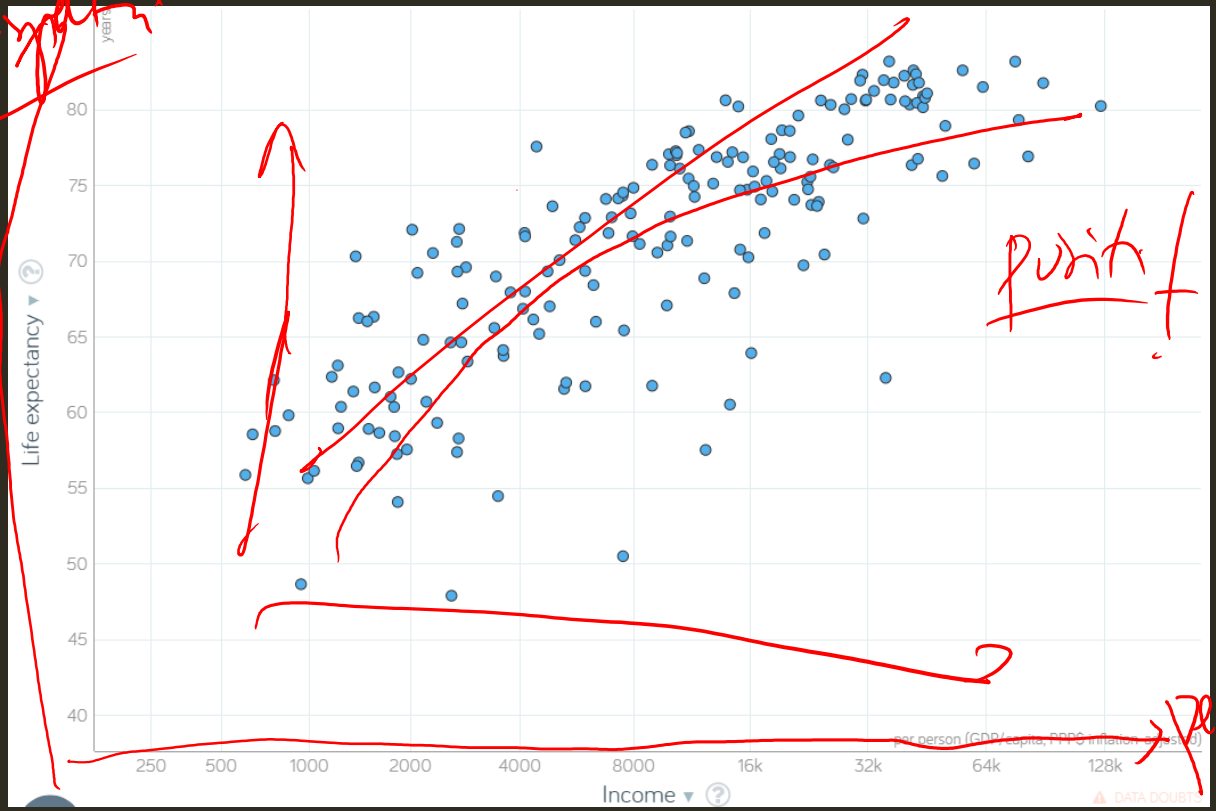
# FUNDAMENTALS OF REGRESSION ANALYSIS

Ventje Jeremias Lewi Engel, M.T., CEH
Prodi Informatika
Institut Teknologi Harapan Bangsa

2020

# REFRESHMENT

1. Jelaskan hubungan antara variabel ekspetasi hidup vs penghasilan yang ada di scatter plot berikut!



# REFRESHMENT

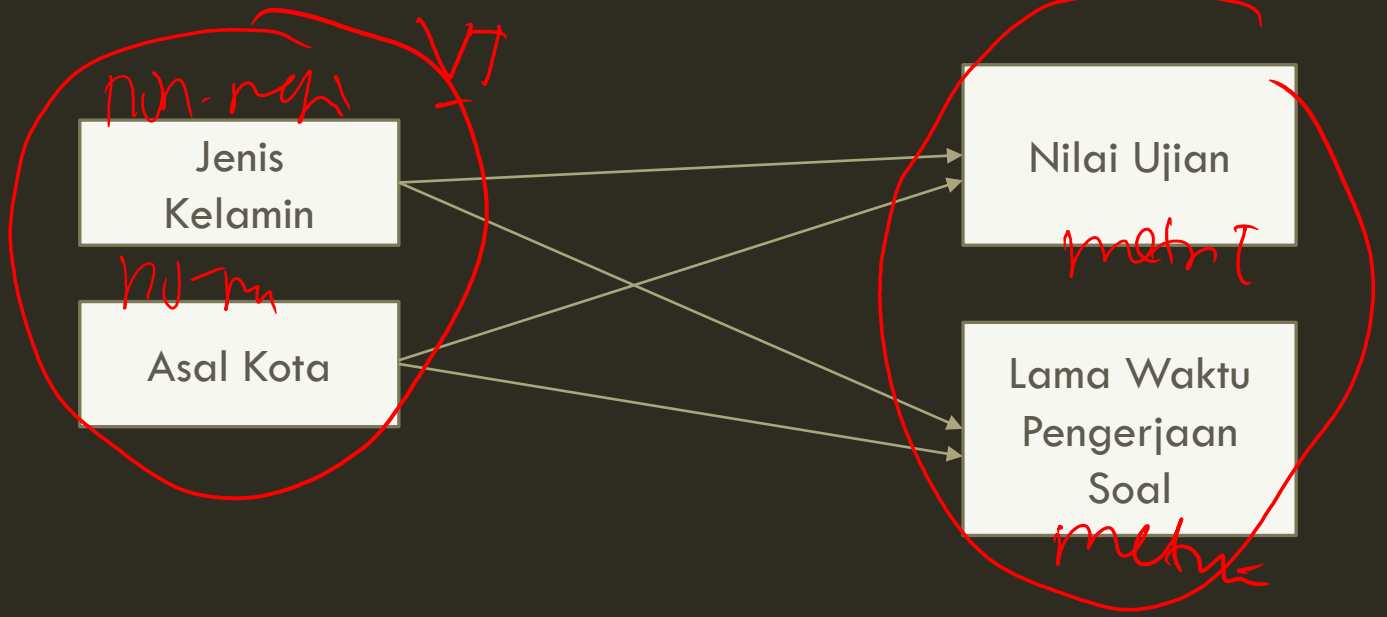
2. Sebutkan contoh hubungan ~~regresi~~ dari tabel 1!
3. Sebutkan contoh hubungan Canonical Correlation Analysis dari tabel 1!

Tabel 1

| Country | Income per person ($, 2012) | Life expectancy (years, 2012) |
|---|---|---|
| Afghanistan | 1359.7 | 60.254 |
| Albania | 6969.3 | 77.185 |
| Algeria | 6419.1 | 70.874 |
| … | … | … |
| Zimbabwe | 545.3 | 58.142 |

# REFRESHMENT

4. Jelaskan hubungan yang terdapat pada bagan berikut:



# TUJUAN PEMBELAJARAN

Setelah mengikuti perkuliahan ini, diharapkan mahasiswa mampu:

1. Membuat model regresi linier sederhana dari dataset
2. Membedakan regresi dengan korelasi
3. Mengevaluasi model regresi linier yang dihasilkan

**IMPORTANT!**

# OUTLINE

Hubungan & Model Regresi Linier

Regression vs Correlation

Langkah-langkah dan Asumsi Regresi Linier

Evaluasi Regresi

# REGRESI ADALAH CARA MENCARI PARAMETER MODEL UNTUK HUBUNGAN INPUT & OUTPUT

**OUTPUT**
**(Dependent Variables)**

**INPUT**
**(Independent Variables, Predictors)**

| | Y | $X_1$ | $X_2$ | $X_3$ | … | $X_n$ |
|---|---|---|---|---|---|---|
| Case 1 | | | | | | |
| Case 2 | | | | | | |
| Case 3 | | | | | | |
| ⋮ | | | | | | |
| Case m | | | | | | |

**Relationship ?**

**LINEAR MODEL**

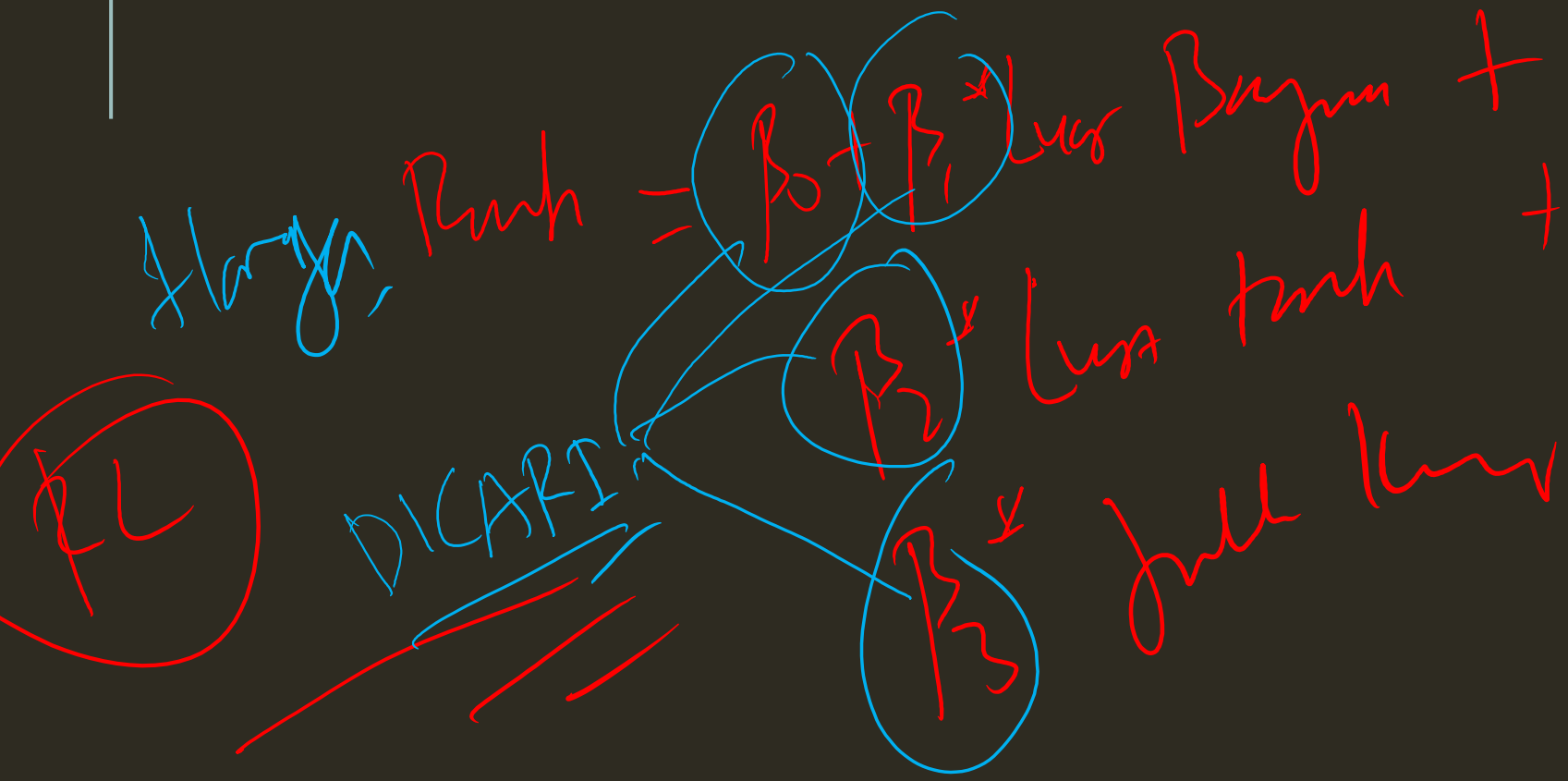
$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + e$$

$e$ (error)

**PARAMETER**

**Intercept/bias**

**coefficients**


Harga Rumah = β0 + β1 Luas Bangunan +
β2 Luas tanah +
β3 Jarak lokasi
DICARI

# SETELAH PARAMETER MODEL DIESTIMASI, KITA BISA MELAKUKAN PREDIKSI

Y-hat

*LINEAR MODEL*

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

*PREDICT !*

*Prediction* $\hat{Y}$

| $X_1$ | $X_2$ | $X_3$ | … | $X_n$ |
|---|---|---|---|---|
| | | | | |

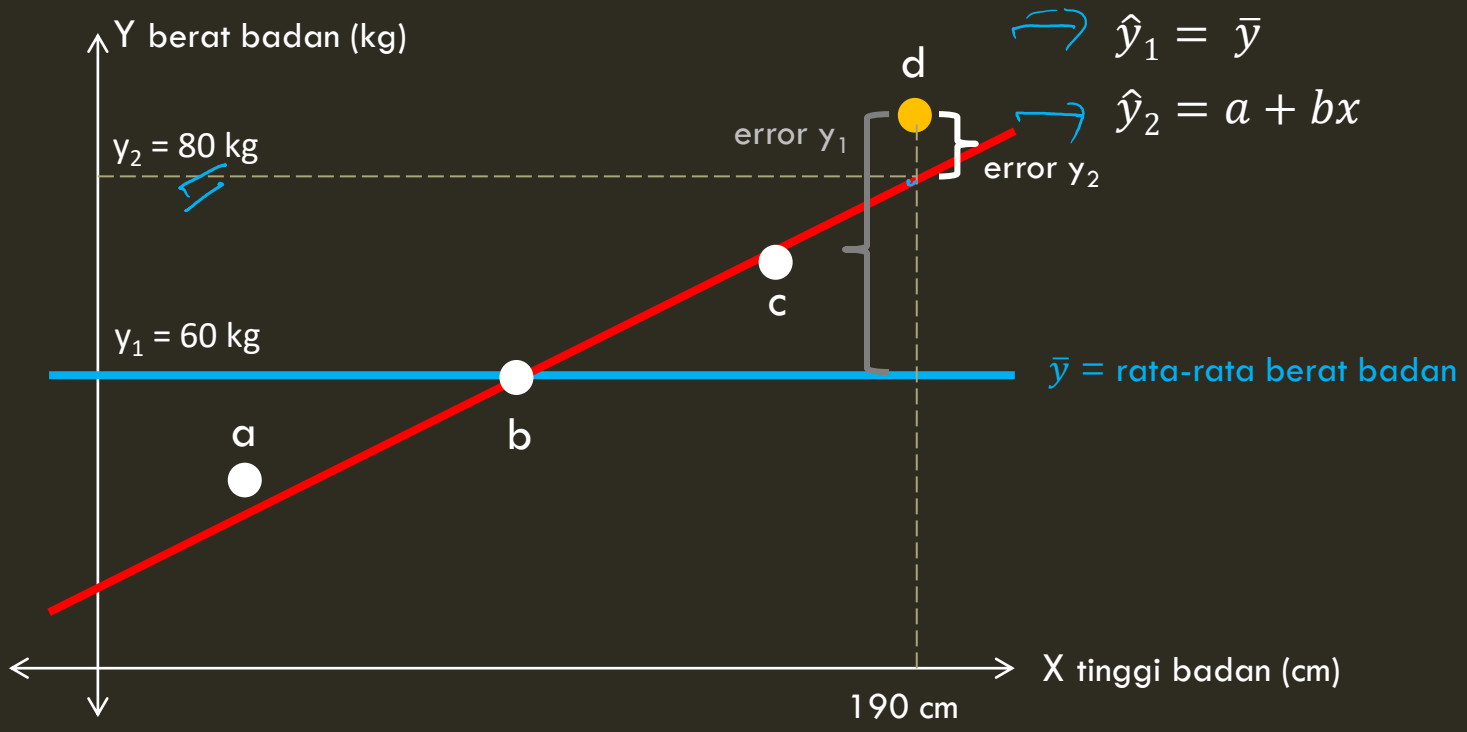
*Actual*

| Y |
|---|
| |

| $X_1$ | $X_2$ | $X_3$ | … | $X_n$ |
|---|---|---|---|---|
| | | | | |

*Prediction Error* $\hat{Y} - Y$

# ILUSTRASI PREDIKSI BERAT BADAN MENGGUNAKAN TINGGI BADAN




Prediksi menggunakan regresi linier (error $y_2$) menghasilkan total error lebih kecil dibandingkan total error dari dugaan rata-rata.

# TUJUAN ANALISIS REGRESI

Regresi linier adalah teknik statistik untuk:

- **Menganalisis efek** (*explanation*) dari variabel-variabel independen (*predictor*) terhadap variable dependen (*criterion*).
- **Memprediksi** (*prediction*) nilai sebuah variable dependen dengan menggunakan satu *(simple regression)* atau lebih variable independen *(multiple regression)*.

# JENIS ANALISIS REGRESI


- Berapa jumlah variabel independen?
  - 1 : Simple regression
  - >1 : Multiple regression
- Bagaimana bentuk garis regresi?
  - Linear : Linear regression
  - Nonlinear : Nonlinear regression
- Apa jenis data variable dependen?
  - Kontinyu/Metris : Simple & Multiple regression
  - Binomial/binary : Logistic regression

# MODEL REGRESI LINEAR SEDERHANA

$$Y = b_0 + bX_1 + e$$

$$\hat{Y} = b_0 + bX_1$$

Y = Berat badan actual (kg)

Y hat = berat badan yang diprediksi (kg)

$X_1$ = Tinggi badan (cm)

$b_0$ = konstanta

e = error

# MODEL MULTIPLE REGRESI LINEAR

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + e$$

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

Y = Konsumsi rumah tangga (rupiah per bulan)

Y hat = Konsumsi rumah tangga yang diprediksi (rupiah per bulan)

$X_1$ = Pendapatan rumah tangga (rupiah per bulan)

$X_2$ = Jumlah anggota keluarga (orang)

$X_3$ = Lokasi tempat tinggal (kota atau desa)

$b_n$ = konstanta

# MODEL MULTIPLE REGRESI LINEAR

$$\text{Variate (Y hat)} = X_1b_1 + X_2b_2 + \ldots + X_nb_n$$

- Nilai variat (Y hat) akan dihitung untuk setiap respon.
- Nilai Y hat adalah kombinasi linear dari seluruh gabungan variable yang menghasilkan prediksi paling baik.

# REGRESI NONLINIER JUGA BISA DILAKUKAN

$$Y = b_0 + b_1 X$$

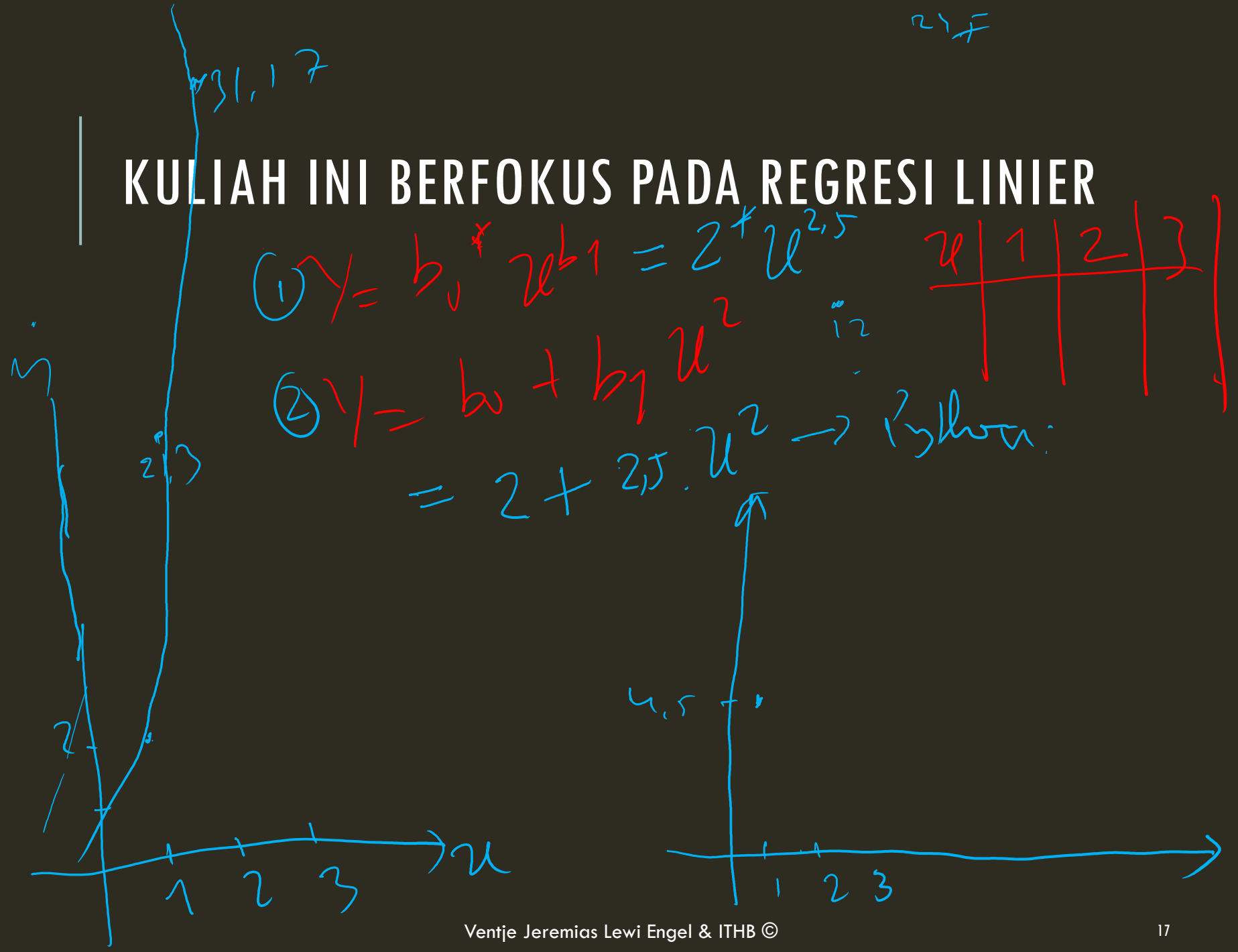
$$Y = b_0 * X^{b_1}$$

$$Y = b_0 + b_1 X^2$$

$$Y = b_0 + \exp(b_1 X)$$

$$Y = b_0 + \cos(b_1 X) + \sin(b_1 X)$$

# KULIAH INI BERFOKUS PADA REGRESI LINIER



# OUTLINE

Hubungan & Model Regresi Linier

Regression vs Correlation

Langkah-langkah dan Asumsi Regresi Linier

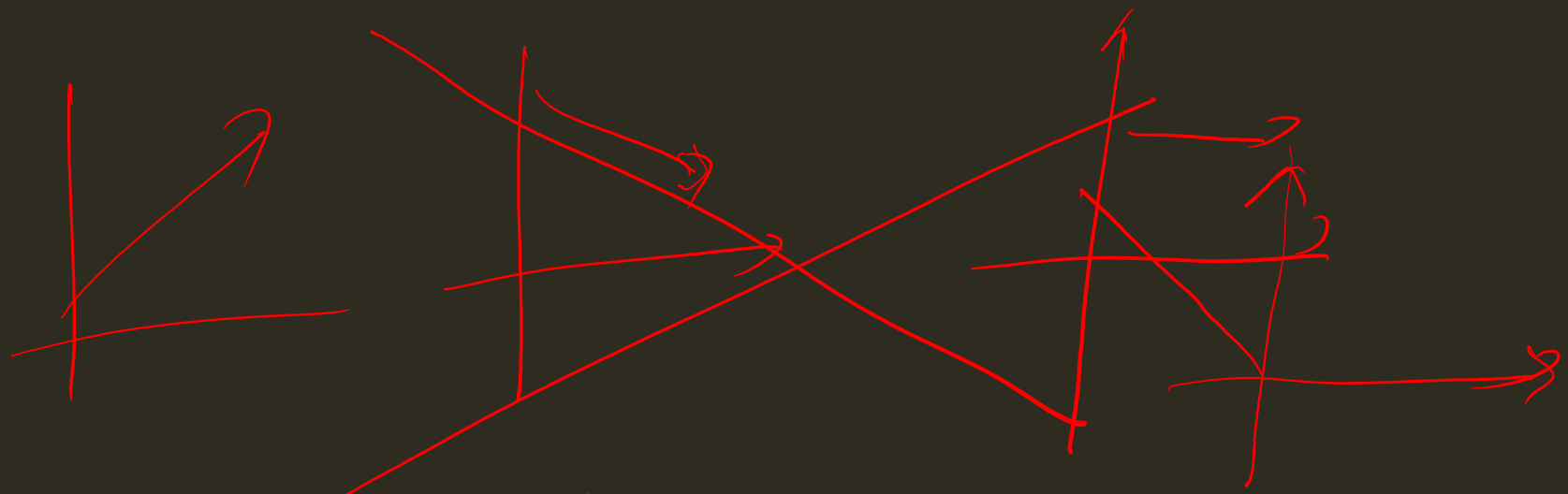
Evaluasi Regresi

# KORELASI BUKAN REGRESI

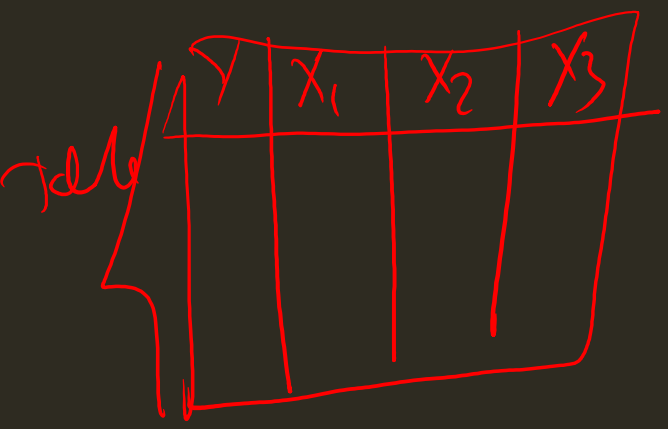
- Analisis regresi menganalisis efek antara variable independen dengan variable dependen.
- Efek ini bisa dilihat sebagai korelasi, tetapi keduanya berbeda. Korelasi untuk melihat apakah ada hubungan dan bagaimana bentuk hubungannya.
- Sedangkan, regresi menjelaskan hubungan tersebut secara satu arah dari variabel independen ke variabel dependen

# CORRELATION VS REGRESSION

| Aspek | Correlation | Regression |
|---|---|---|
| Hubungan | Hanya melihat hubungan (relationship) yang terjadi | Melihat efek dari variabel independen ke variabel dependen |
| Arah Hubungan | Korelasi antara A dan B adalah sama dengan korelasi antara B dan A | Arah hubungan adalah dari variabel independen ke variabel dependen |
| Representasi | Sebuah nilai | Sebuah formula statistik |
| Kausalitas | Bukan kausalitas | Bisa menjadi dasar untuk peneltian kausalitas menggunakan SEM |



# Regresi Linier




Formula :

$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$

$Y = 3 + (-2) X_1 + 1{,}5 X_2 + (-3) X_3$

# OUTLINE

Hubungan & Model Regresi Linier

Regression vs Correlation

Langkah-langkah dan Asumsi Regresi Linier

Evaluasi Regresi

# PERTIMBANGAN SEBELUM MELAKUKAN ANALISIS REGRESI

Tiga pertimbangan utama dalam penggunaan analisis regresi:

1. Kecocokan dengan masalah yang diteliti (prediction atau explanation)
2. Penentuan *statistical relationship*
3. Pemilihan variable dependen dan independen
   - Pastikan ada teori yang mendukung pemilihan variable.
   - Adanya *measurement error* pada variable, terutama pada variable dependen. Bisa diatasi dengan *summated scales* atau SEM
   - *Specification error: inclusion of irrelevant variables or exclusion of relevant variables.*

Specification error → pemilihan VI.

Y → harga rumah

X → luas bangunan, luas tanah, kualitas

variabel baru : lokasi, jumlah kamar, desain arsitek.

exclude relevant variabel lagi

# RULE OF THUMB

- Dengan analisis regresi, error yang dihasilkan tidak dapat dipisahkan antara error karena kesalahan prediksi atau error karena pengukuran (*measurement*).
- Lebih baik menggunakan variable independen yang banyak walaupun tampak tidak relevan (karena dampaknya hanya kesulitan interpretasi), daripada mengambil resiko mengabaikan sebuah variable yang relevan (yang dapat menghasilkan bias terhadap hasil regresi).

# DUMMY VARIABLES

- Dalam analisis regresi dan model machine learning, lebih aman untuk membuat variabel dummy menggunakan one-hot encoding

Label Encoding

| Food Name | Categorical # | Calories |
|---|---|---|
| Apple | 1 | 95 |
| Chicken | 2 | 231 |
| Broccoli | 3 | 50 |

→

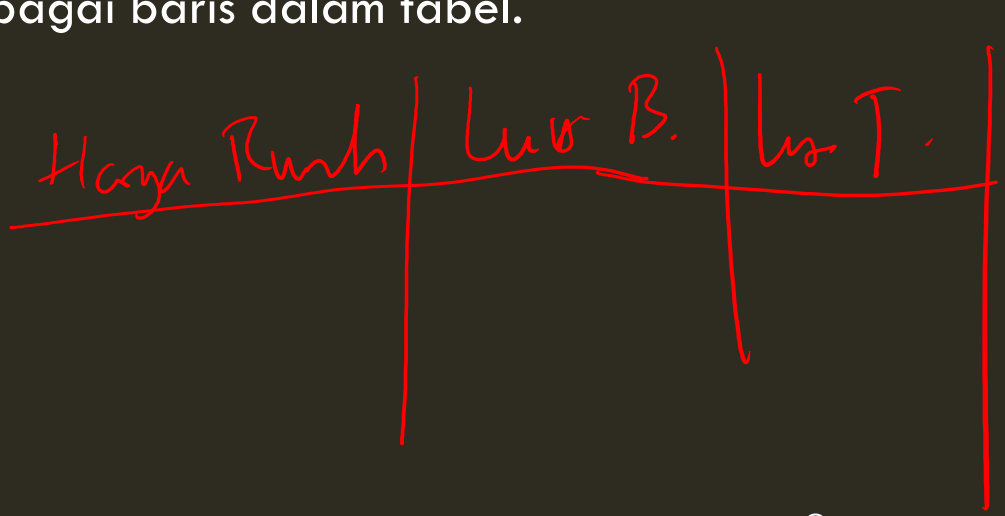
One Hot Encoding

| Apple | Chicken | Broccoli | Calories |
|---|---|---|---|
| 1 | 0 | 0 | 95 |
| 0 | 1 | 0 | 231 |
| 0 | 0 | 1 | 50 |

- **Jumlah variabel dummy = Jumlah Variasi Data - 1**

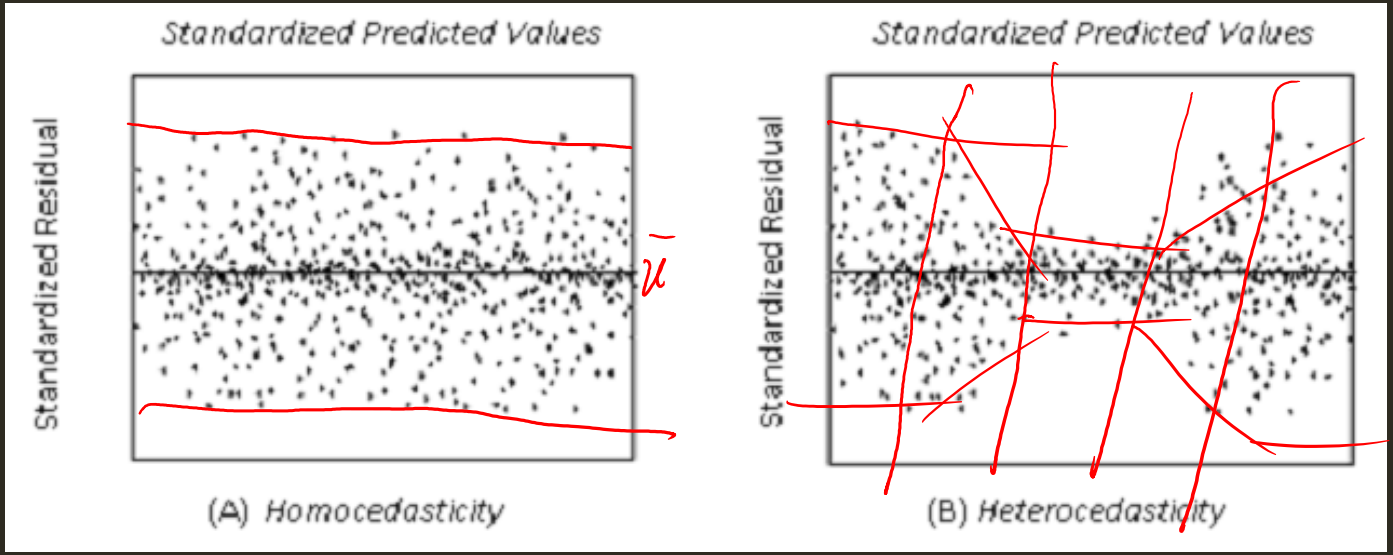
# ASUMSI REGRESI LINIER

- Normality
  - Menurut Central Limit Theorem (CLT) bila data diambil dari sampling random dan berjumlah besar, maka distribusi data akan semakin mendekati normal. Asal data tidak bersifat binomial.
- No Autocorrelation
  - Autocorrelation = Hubungan antar observasi data harus tidak ada kaitan atau saling mempengaruhi. Observasi data direpresentasikan sebagai baris dalam tabel.



# ASUMSI REGRESI LINIER

- Homoscedasticity
  - Keadaan ketika error variance adalah sama pada setiap level variabel independen (prediktor)



# ASUMSI REGRESI LINIER

- No Multicollinearity
  - Multicollinearity = Situasi yang menunjukkan adanya korelasi atau hubungan kuat antara dua variabel independen atau lebih dalam sebuah model regresi.
  - Biasa diuji menggunakan nilai VIF (Variation Inflation Factor) dari setiap variabel independen.

    VIF = 1 : no multicollinearity
    1 < VIF < 6 : ada sedikit multicolliearity, masih bisa diterima
    6 < VIF : ada multicollinearity
  - Untuk mendapatkan multicollinearity di bawah 6 bisa dibilang adalah jarang. Kadang kita bisa menerima nilai VIF > 6 dengan catatan.

# OUTLINE

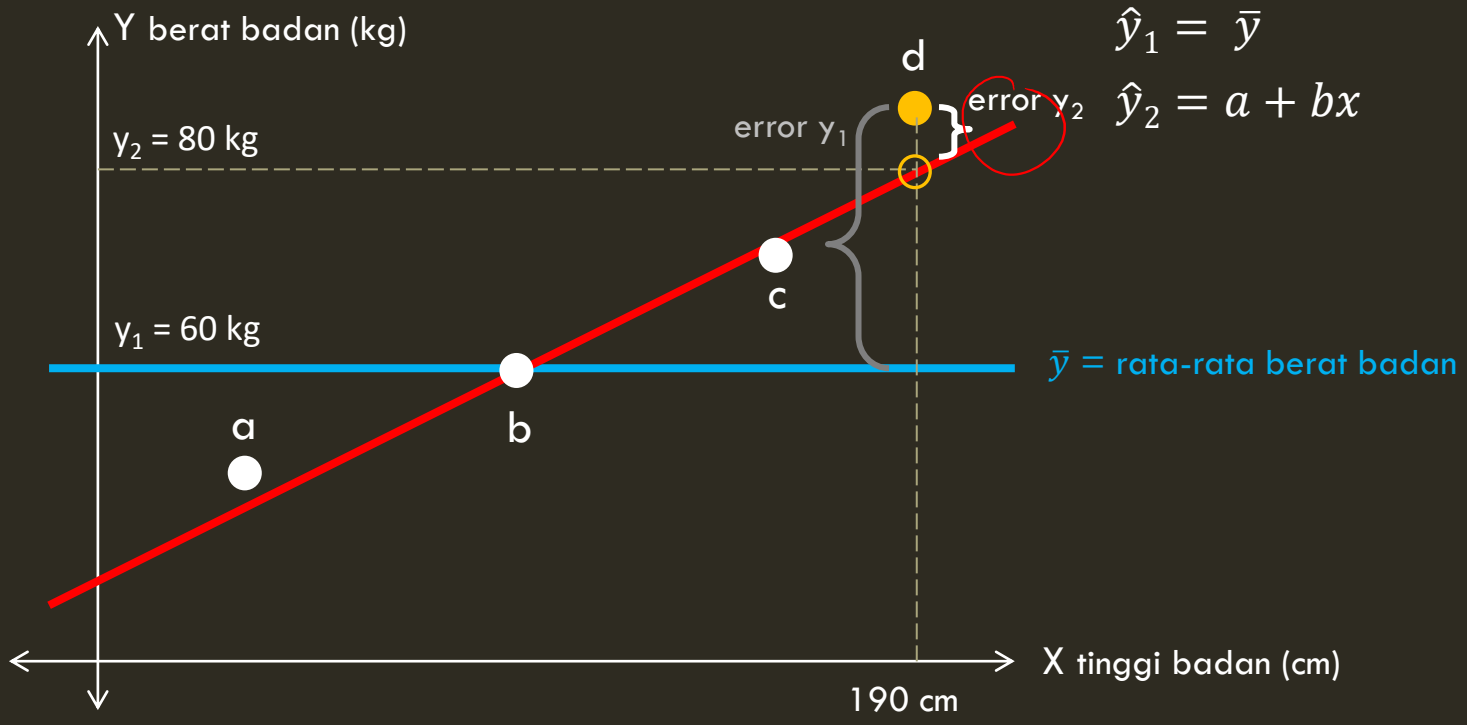
Hubungan & Model Regresi Linier

Regression vs Correlation

Langkah-langkah dan Asumsi Regresi Linier

Evaluasi Regresi

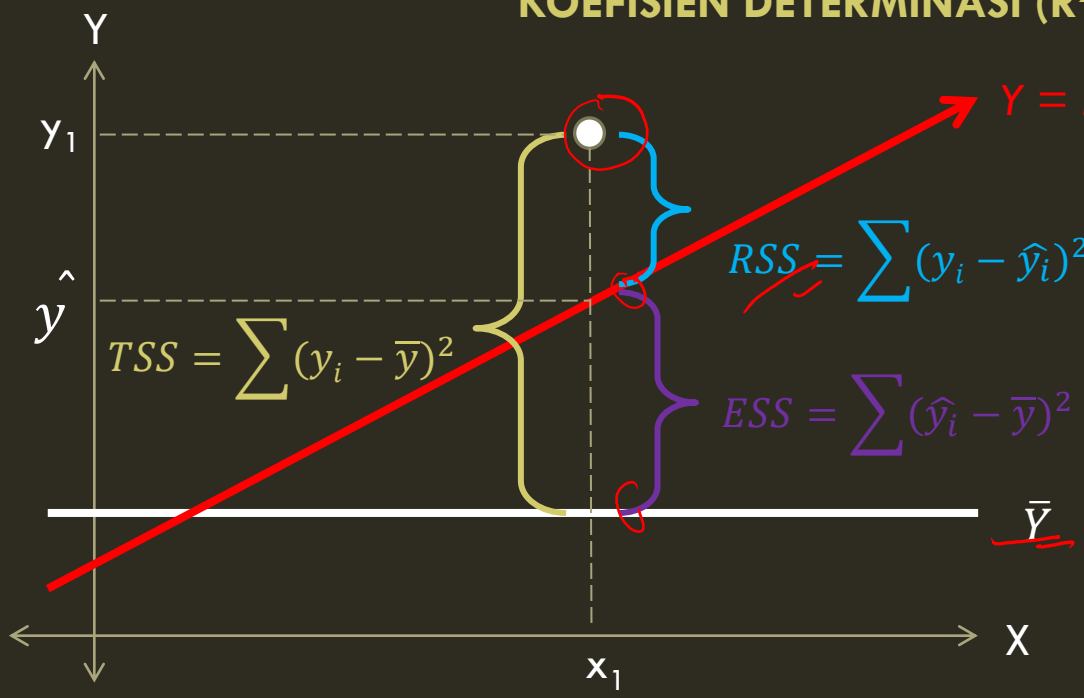
# REGRESI LINIER AKAN MENGHASILKAN PREDIKSI DENGAN ERROR YANG LEBIH KECIL




Prediksi menggunakan regresi linear (error $y_2$) menghasilkan total error lebih kecil dibandingkan total error dari dugaan rata-rata.

# MENILAI OVERALL FIT (KECOCOKAN MODEL): GUNAKAN $R^2$

**KOEFISIEN DETERMINASI ($R^2$)**




$$R^2 = \frac{ESS}{TSS}$$

*$R^2$ adalah persentase explained dibandingkan total variance*

$$F = \frac{ESS/dfreg}{RSS/dfres}$$

TSS = Total Sum of Square (Total Variance) = RSS + ESS
RSS = Residual Sum of Square (Unexplained Variance)
ESS = Explained Sum of Square (Explained Variance)
Dfreg = jumlah koefisien dalam model – 1
Dfres = jumlah sampel – jumlah koefisien dalam model

# MENILAI SIGNIFIKANSI HUBUNGAN IV DAN DV SECARA KESELURUHAN: F-TEST

- Dalam statistik, signifikan artinya nilai sebuah estimasi memang nyata, bukan nol, dan hal ini terjadi bukan karena kebetulan (sampel acak). Signifikan artinya peluang bahwa nilai sesungguhnya dari sebuah estimasi = nol sangatlah kecil (<5%)
- Bila uji F test menunjukkan konstan dan koefisien signifikan, artinya keterkaitan (korelasi) independent variable terhadap dependent variabel memang nyata, dan bukan terjadi karena kebetulan.

- Nilai F dihitung untuk melakukan F test (ANOVA)
- F test dilakukan untuk menguji apakah konstanta dan koefisien masing-masing independen variable tidak sama dengan 0.
- Model regresi $Y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$
- $H_0 : b_0 = b_1 = b_2 = \ldots = b_n = 0$
- $H_1$ : ada koefisien yang tidak nol

# MENILAI SIGNIFIKANSI MASING-MASING KOEFISIEN: T-TEST

- Uji t (t-test) akan menguji signifikansi masing-masing konstant dan koefisien secara individual (tepisah) dan menentukan apakah setiap estimasi tersebut berbeda dari nol bukan karena kebetulan (peluang bahwa angka sesungguhnya adalah nol < 5%) 0,05
- Bila koefisien dari sebuah variabel independent signifikan, artinya variabel tersebut memang memiliki pengaruh (sekecil apapun itu) terhadap variabel dependent, dan hal ini bukan terdeteksi secara kebetulan akibat sampel acak.

# CONTOH OUTPUT MULTIPLE LINEAR REGRESSION

**TABLE 15** Multiple Regression Results Adding $X_3$ (Firm Size) as an Independent Variable by Using a Dummy Variable

**Stepwise Regression with Transformed Variables**

| | |
|---|---|
| Multiple *R* | .895 |
| Coefficient of Determination ($R^2$) | .801 |
| Adjusted $R^2$ | .788 |
| Standard error of the estimate | .548 |

**Analysis of Variance**

| | Sum of Squares | *df* | Mean Square | *F* | Sig. |
|---|---|---|---|---|---|
| Regression | 112.669 | 6 | 18.778 | 62.464 | .000 |
| Residual | 27.958 | 93 | .301 | | |
| Total | 140.628 | 99 | | | |

**Variables Entered into the Regression Model**

| | *Regression Coefficients* | | | *Statistical Significance* | | *Correlations* | | | *Collinearity Statistics* | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Variables Entered** | **B** | **Std. Error** | **Beta** | ***t*** | **Sig.** | **Zero-order** | **Partial** | **Part** | **Tolerance** | **VIF** |
| (Constant) | −1.250 | .492 | | −2.542 | .013 | | | | | |
| $X_9$ Complaint Resolution | .300 | .060 | .304 | 4.994 | .000 | .603 | .460 | .231 | .576 | 1.736 |
| $X_6$ Product Quality | .365 | .046 | .427 | 7.881 | .000 | .486 | .633 | .364 | .727 | 1.375 |
| $X_{12}$ Salesforce Image | .701 | .093 | .631 | 7.507 | .000 | .500 | .614 | .347 | .303 | 3.304 |
| $X_7$ E-Commerce | −.333 | .135 | −.196 | −2.473 | .015 | .283 | −.248 | −.114 | .341 | 2.935 |
| $X_{11}$ Product Line | .203 | .061 | .224 | 3.323 | .001 | .551 | .326 | .154 | .469 | 2.130 |
| $X_3$ Firm Size | .271 | .123 | .114 | 2.207 | .030 | .229 | .223 | .102 | .798 | 1.253 |

# APA YANG SUDAH DIPELAJARI?

# THANK YOU

Ventje Jeremias Lewi Engel, M.T.
Prodi Informatika
Institut Teknologi Harapan Bangsa