Analisis Regresi Linier

Kuliah 4 STA1381 – Pengantar Sains Data

Septian Rahardiantoro

Outline

- Pengantar Pemodelan Statistika
- Analisis Regresi
  - Regresi Linier Sederhana
  - Regresi Linier Berganda
- Dummy Variable

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)

Analisis Regresi

→Analisis statistika yang memanfaatkan hubungan sebab akibat antara dua atau lebih peubah kuantitatif sehingga salah satu peubah dapat diramalkan dari peubah lainnya.

- Hubungan Antar Peubah:
  - Fungsional (deterministik) → Y=f(X) ; misalnya: Y=10X
  - Statistik (stokastik) → amatan tidak jatuh pas pada kurva (terdapat galat)
    - Mis: IQ vs Prestasi, Berat vs Tinggi, Dosis Pupuk vs Produksi

Analisis Regresi

- 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

- Syarat Utama: Variabel output (Y) bersifat numerik

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

- Bentuk model

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

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

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

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:

Population y intercept

Population Slope Coefficient

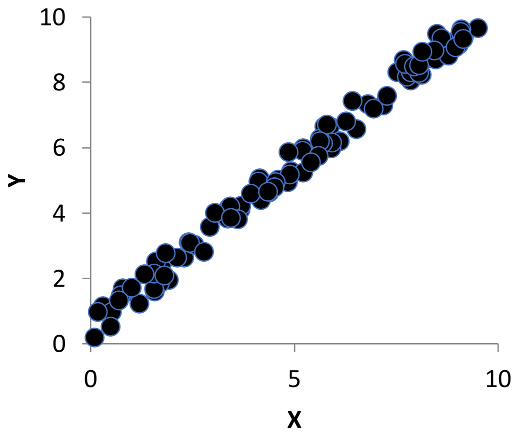
Independent Variable

Random Error term, or residual

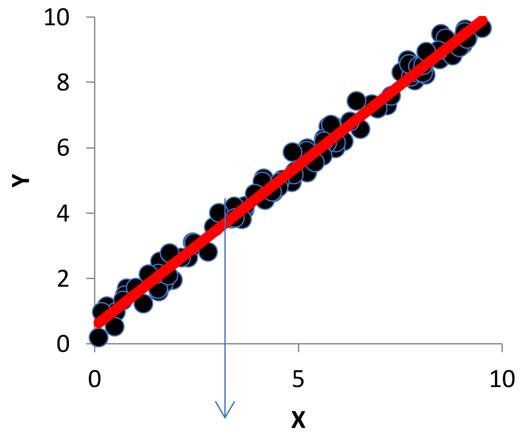
Dependent Variable

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

Regresi Linier Sederhana



Plot data Y dengan X



diperoleh persamaan garis regresi

$$\hat{Y} = b_0 + b_1 X$$

Sehingga:

Model regresi

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

b_0 b_1

diduga oleh

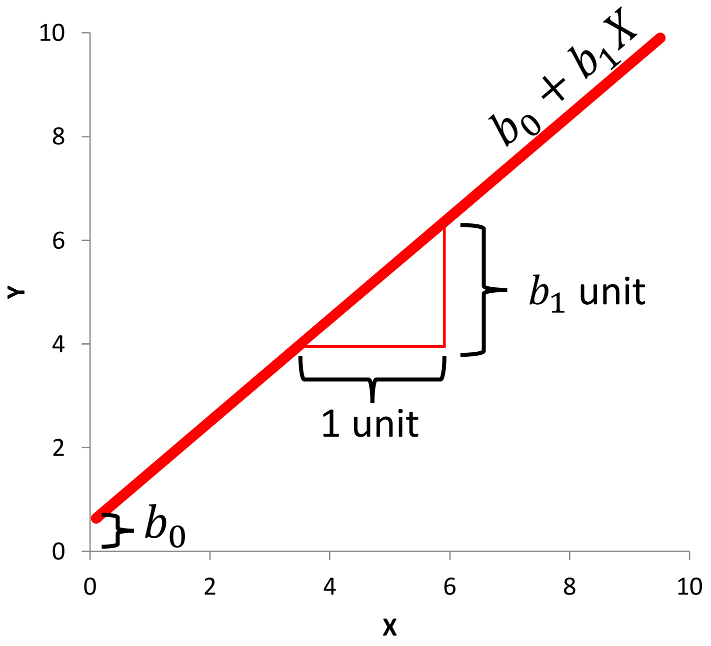
Persamaan regresi

$$\hat{Y} = b_0 + b_1 X$$

Regresi Linier Sederhana

- Maka
  - β_0 diduga oleh b_0
  - β_1 diduga oleh b_1

Interpretasi?



b_0 adalah nilai rataan Y ketika X= 0
(tidak dapat diinterpretasikan oleh X)

numerik
numerik

b_1 adalah perubahan nilai rataan Y untuk setiap perubahan 1 satuan X.

Pendugaan Parameter

- Misalkan $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ adalah prediksi untuk Y berdasarkan nilai ke-i peubah X (dengan $i = 1,2,3 \ldots, n$)

$y = \beta_0 + \beta_1 x + e$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

- Maka residual ke-i didefinisikan oleh:

$$e_i = y_i - \hat{y}_i \rightarrow e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

$e = y - \hat{y}$

$= y - (\hat{\beta}_0 + \hat{\beta}_1 x)$

$e = y - \hat{\beta}_0 - \hat{\beta}_1 x$

- JKG (Jumlah Kuadrat Galat) didefinisikan oleh:

$$JKG = e_1^2 + e_2^2 + \cdots + e_n^2 \rightarrow \Sigma e_i^2$$

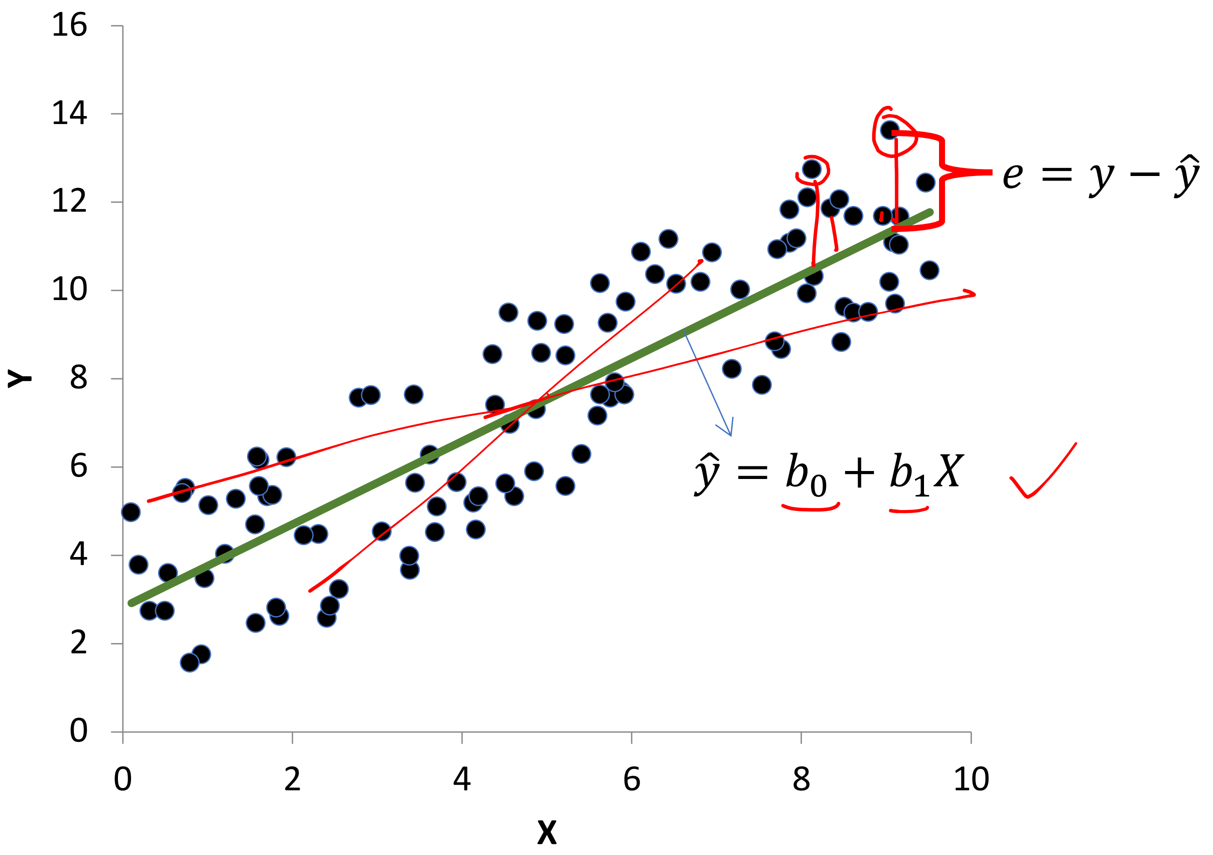
$$JKG = \left(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)^2 + \cdots + \left(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right)^2$$

- Penduga MKT (Metode Kuadrat Terkecil), memilih $\hat{\beta}_0$ dan $\hat{\beta}_1$ yang meminimumkan JKG. Dengan perhitungan kalkulus diperoleh:

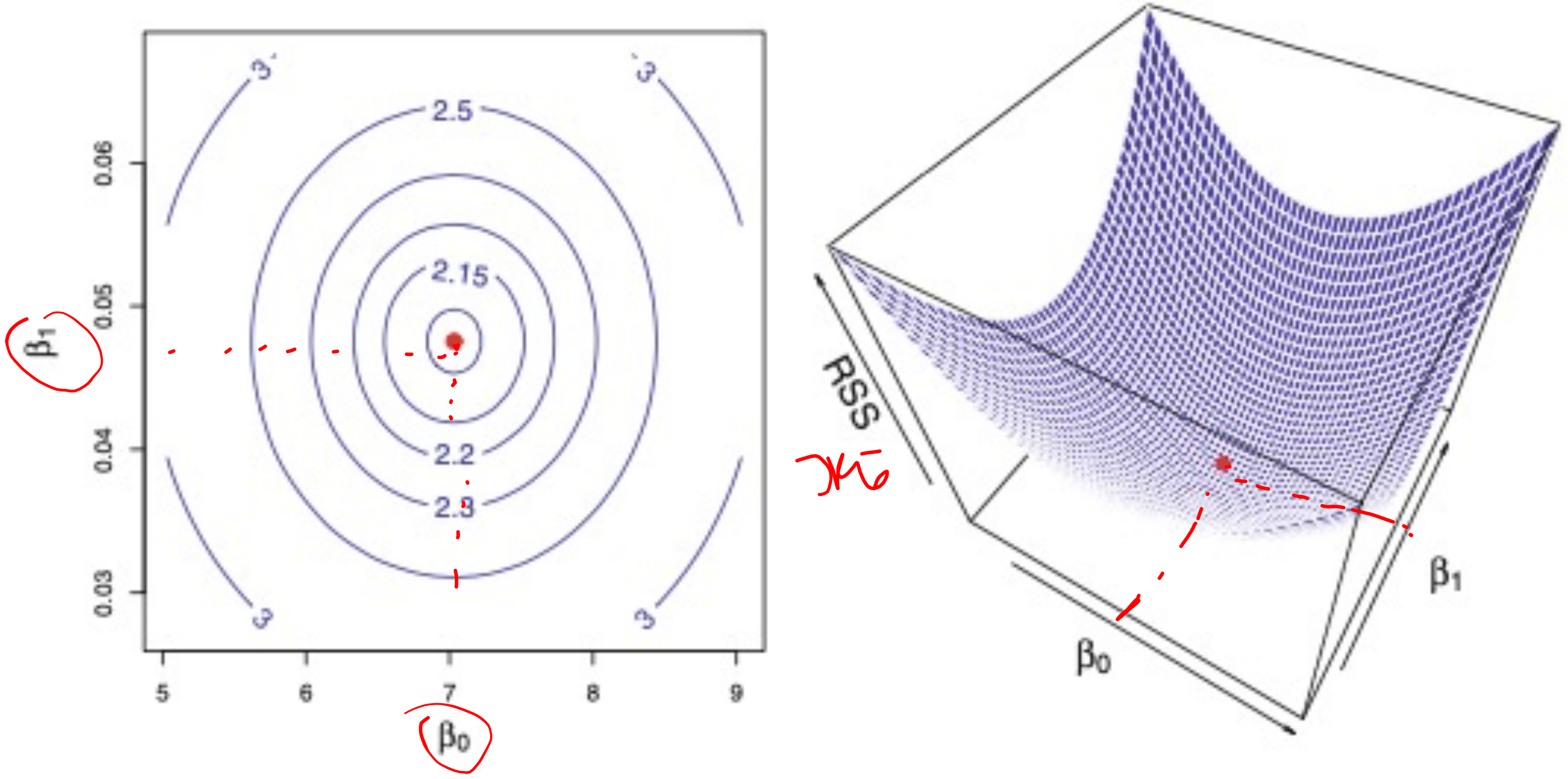
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}; \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Pendugaan Parameter

Metode Kuadrat Terkecil → Meminimumkan jumlah kuadrat galat



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



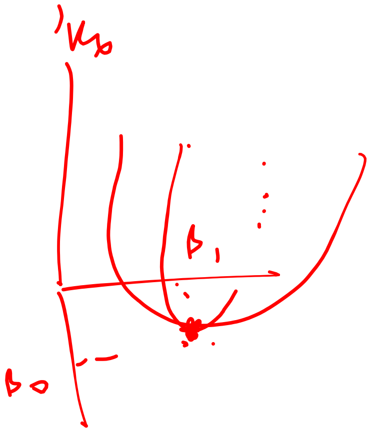
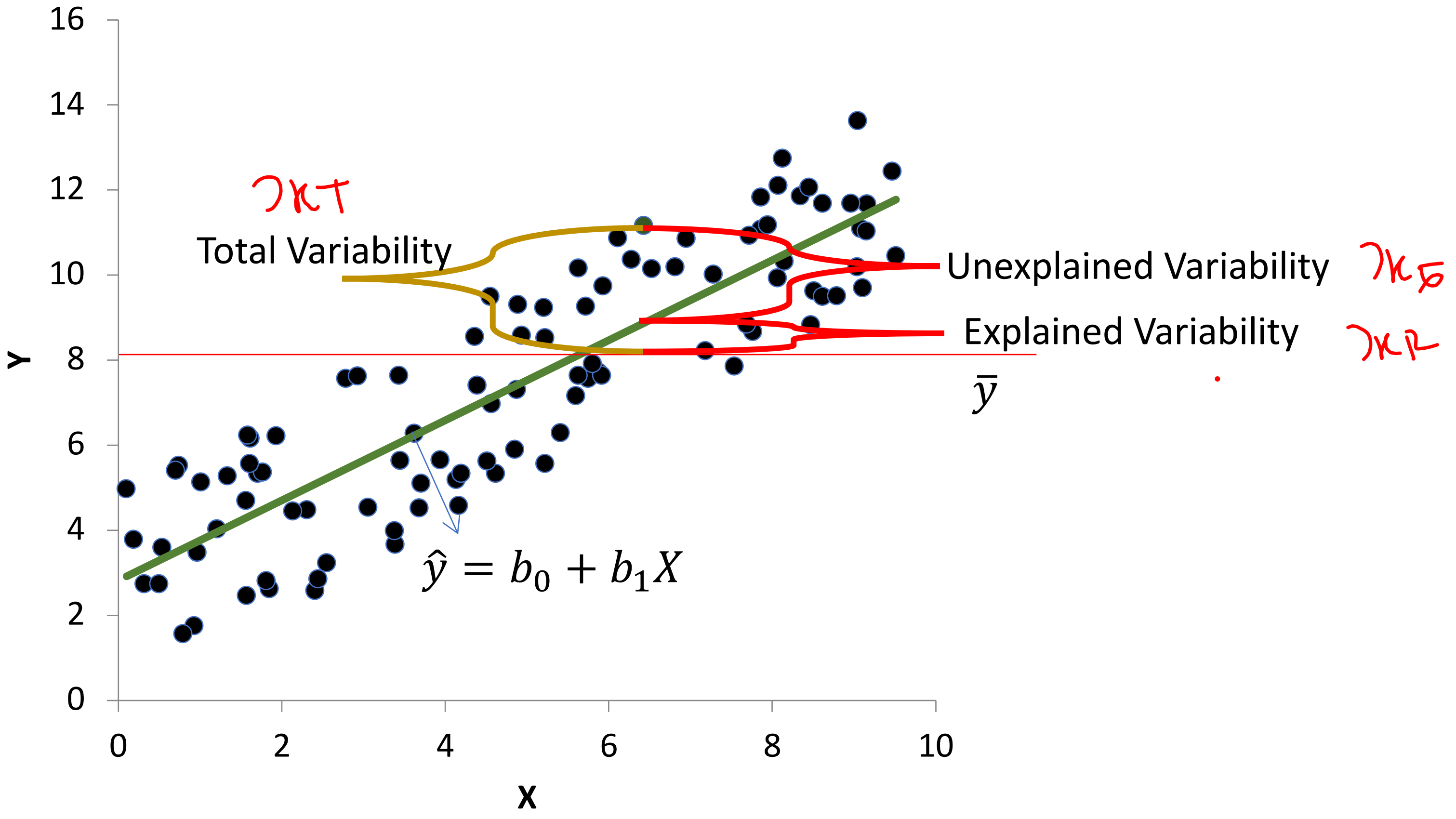


FIGURE 3.2. *Contour and three-dimensional plots of the RSS on the* `Advertising` *data, using* `sales` *as the response and* `TV` *as the predictor. The red dots correspond to the least squares estimates* $\hat{\beta}_0$ *and* $\hat{\beta}_1$*, given by (3.4).*

Pendugaan Parameter

Keragaman yang dapat dijelaskan dan tidak dapat dijelaskan



Regresi Linier Berganda

- Analisis regresi linear berganda:
  - Secara umum, kita memodelkan peubah respon Y sebagai fungsi linier dari k peubah prediktor (X) sebagai:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

n = byk persamaan

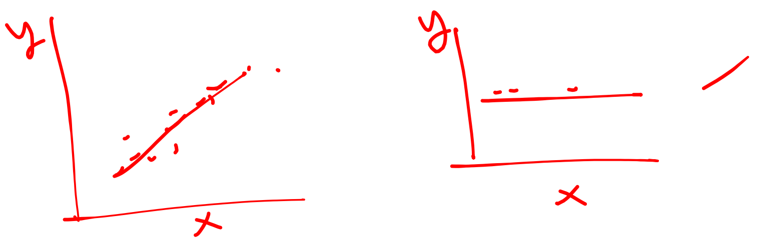
  - Atau dalam notasi matriks $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ → $n \times 1$

$n \times 1$ $\quad$ $n \times (k+1)$ $\quad$ $(k+1) \times 1$

$$\underset{n \times 1}{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \underset{n \times (k+1)}{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{k1} \\ 1 & X_{12} & & X_{k2} \\ 1 & X_{13} & & X_{k3} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1n} & & X_{kn} \end{pmatrix} \qquad \underset{(k+1) \times 1}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

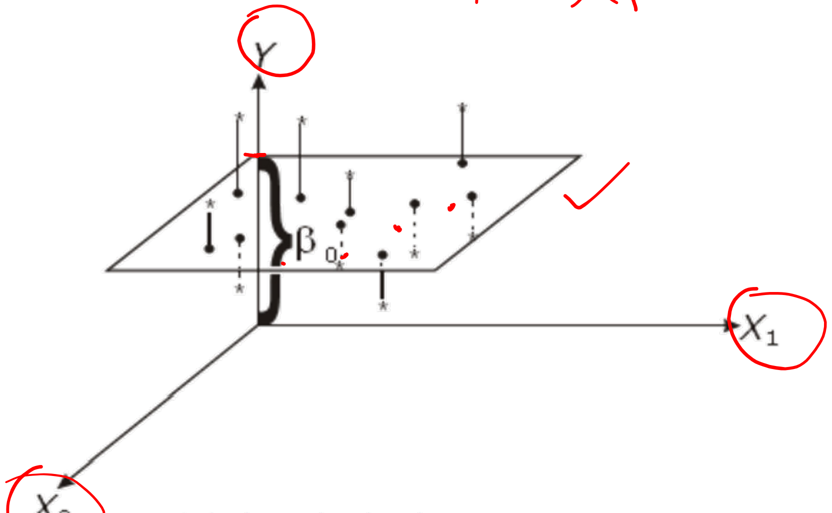
reg linier sederhana

$Y \sim X$ →

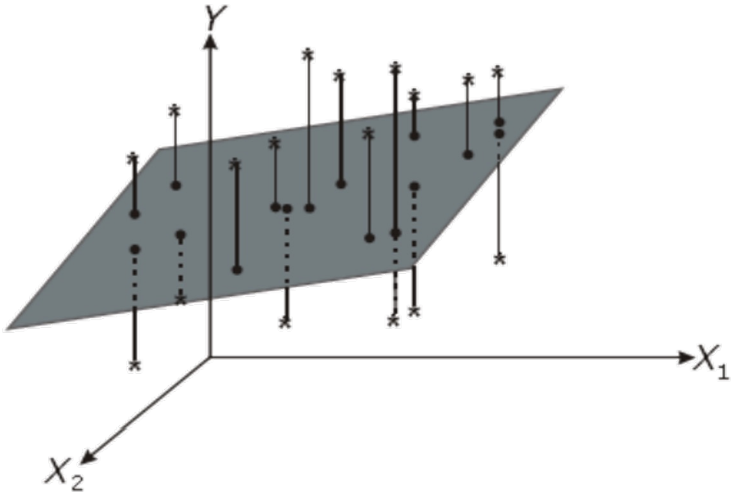


- Jika kita memiliki dua variabel X, model dapat diilustrasikan sebagai berikut

reg linier berganda $Y \sim X_1 + X_2$



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 $\beta_j, j = 0,1,2 \dots, k$ yang meminimumkan $\sum_i \varepsilon^2$, dengan $\varepsilon = Y - \hat{Y}$, yang dalam notasi matriks diperoleh

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

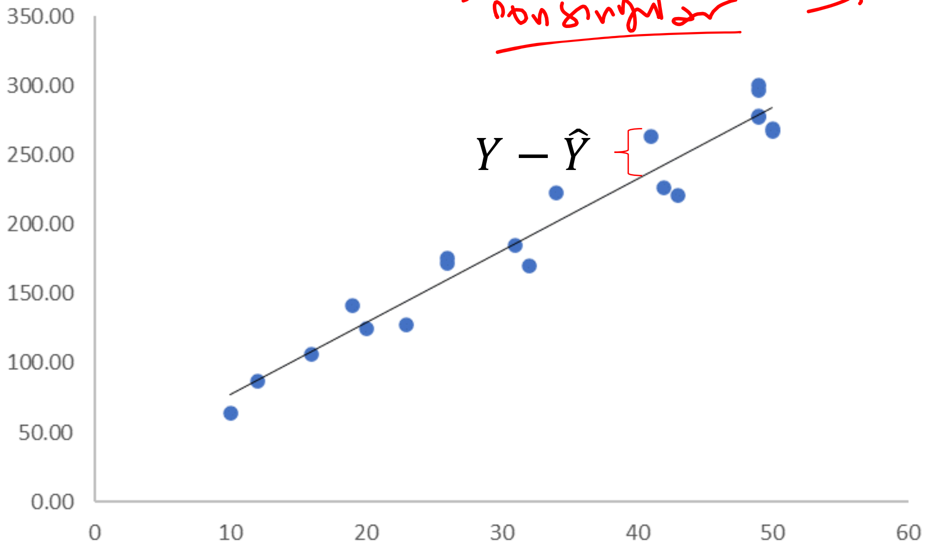
nonsingular → tdk ada multikol

korelasi tinggi

$$\begin{pmatrix} X_1 & X_2 \\ 1 & 5 \\ 2 & 6 \\ 3 & 7 \\ 4 & 8 \end{pmatrix}$$

$X_1 = X_2 - 4$

X_1 bisa dinyatakan dgn X_2



Asumsi model regresi linear

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

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

$\varepsilon \sim N(0, \sigma^2)$

$E(\varepsilon) = 0$

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

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

Antar peubah X tidak ada multikolinearitas

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

Galat bersifat independen/ saling bebas (tidak ada autokorelasi)

Kesesuaian 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.

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

  - 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 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.

Dummy Variable (Peubah Boneka)

- Dummy variable diterapkan pada peubah prediktor dengan skala kategorik
- Banyaknya dummy variable yang dibentuk dari satu peubah kategorik adalah *k*-1, dengan *k* adalah banyaknya kategori dalam peubah tersebut.
- Misalkan pada peubah pekerjaan yang terdiri dari 3 kategori: Pegawai BUMN, Pegawai swasta, dan PNS
- Pilih satu kategori sebagai reference, misalkan Pegawai BUMN

Kategori	D1	D2
Pegawai BUMN	0	0
Pegawai swasta	1	0
PNS	0	1

X (1, 2, 3) → $k=3$

→ dX_2 dX_3

X. → A B C D

Org	X	D_1	D_2	D_3
1	A	0	0	0
2	B	1	0	0
3	C	0	1	0
4	D	0	0	1
5	B	1	0	0
6	C	0	1	0
7.	A	0	0	0

Ilustrasi Kasus

Misalkan telah diperoleh data scraping harga rumah beserta karakteristiknya

Ingin diketahui pengaruh setiap karakteristiknya ke harga rumah tersebut.

Data: RUMAH.TXT

- ID, Identification number
- sales_price, Sales price of residence (dollars) Y
- X1, Finished area of residence (square feet)
- X2, Total number of bedrooms in residence
- X3, Total number of bathrooms in residence
- X4, Presence or absence of air conditioning: 1 if yes; 0 otherwise
- X5, Number of cars that garage will hold
- X6, Presence or absence of swimming pool: 1 if yes; 0 otherwise
- X7, Year property was originally constructed → usia = 2023 − X7
- X8, Index for quality of construction: 1.Indicates high quality; 2.Indicates medium quality; 3.Indicates low quality → siapkan dummy variable
- X9, Qualitative indicator of architectural style
- X10, Lot size (square feet)
- X11, Presence or absence of adjacency to highway: 1 if yes; 0 otherwise

```
rumah <- read.table("D:/rumah.txt", header=TRUE)
head(rumah)
```

Penjelasan:

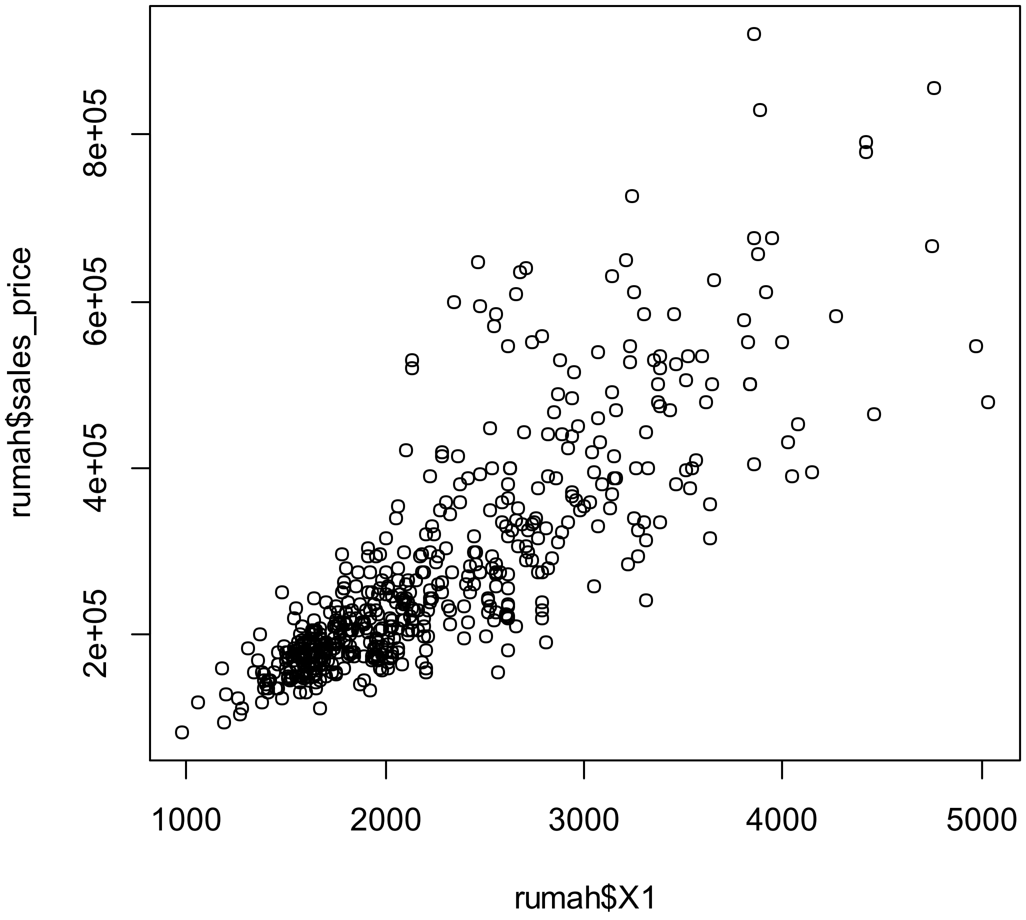
Baris #1: membaca file TXT dengan nama “rumah.txt” yang tersimpan pada folder D. Opsi “header=TRUE” mengatakan bahwa baris pertama pada file yang dibaca merupakan nama-nama kolom. Hasil pembacaan file tersebut disimpan dalam bentuk data frame di R dengan nama “rumah”

Baris #2: mencetak 6 (enam) baris pertama dari data frame “rumah”

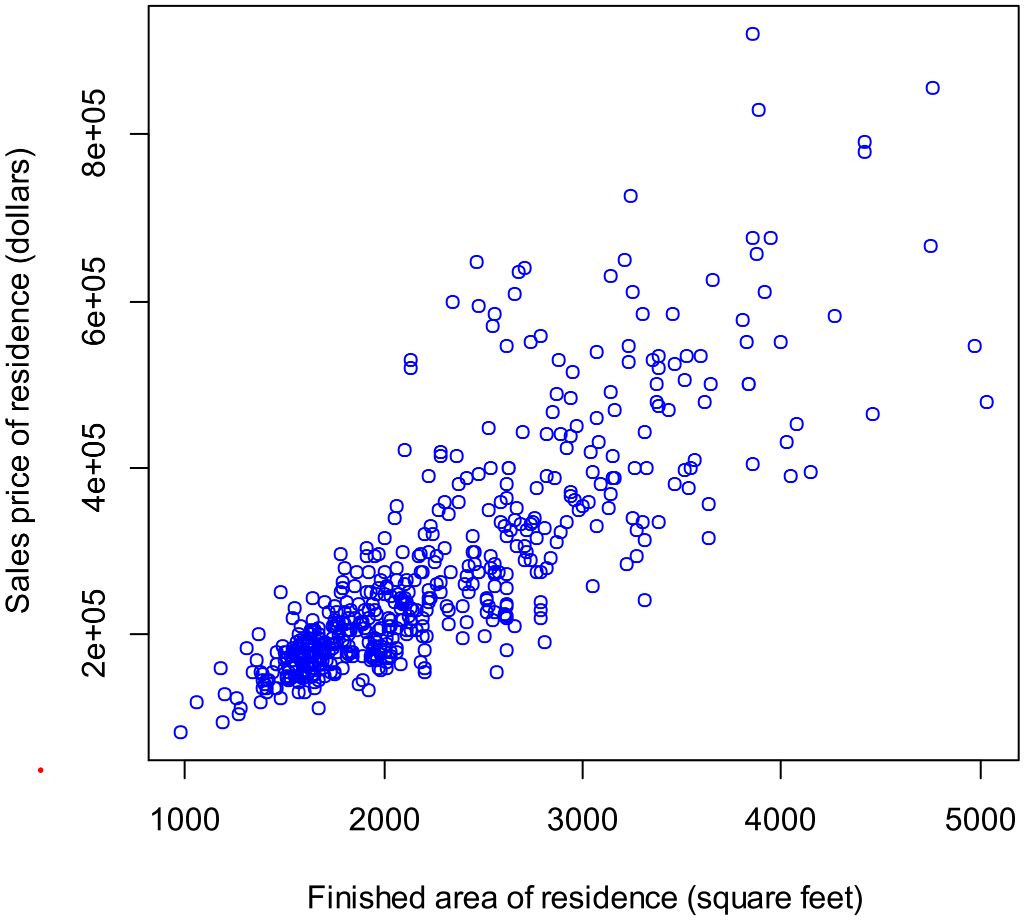
R Console

```
> rumah <- read.table("D:/rumah.txt", header=TRUE)
> head(rumah)
  ID sales_price   X1 X2 X3 X4 X5 X6   X7 X8 X9   X10 X11
1  1      360000 3032  4  4  1  2  0 1972  2  1 22221   0
2  2      340000 2058  4  2  1  2  0 1976  2  1 22912   0
3  3      250000 1780  4  3  1  2  0 1980  2  1 21345   0
4  4      205500 1638  4  2  1  2  0 1963  2  1 17342   0
5  5      275500 2196  4  3  1  2  0 1968  2  7 21786   0
6  6      248000 1966  4  3  1  5  1 1972  2  1 18902   0
>
```

```
plot(rumah$X1, rumah$sales_price)
```




```
plot(rumah$X1, rumah$sales_price,
xlab="Finished area of residence (square feet)",
ylab="Sales price of residence (dollars)",
col="blue")
```



Model Regresi Linier Sederhana: Y ~ X1

```
lm(sales_price ~ 1 + X1, data=rumah)
```

R Console

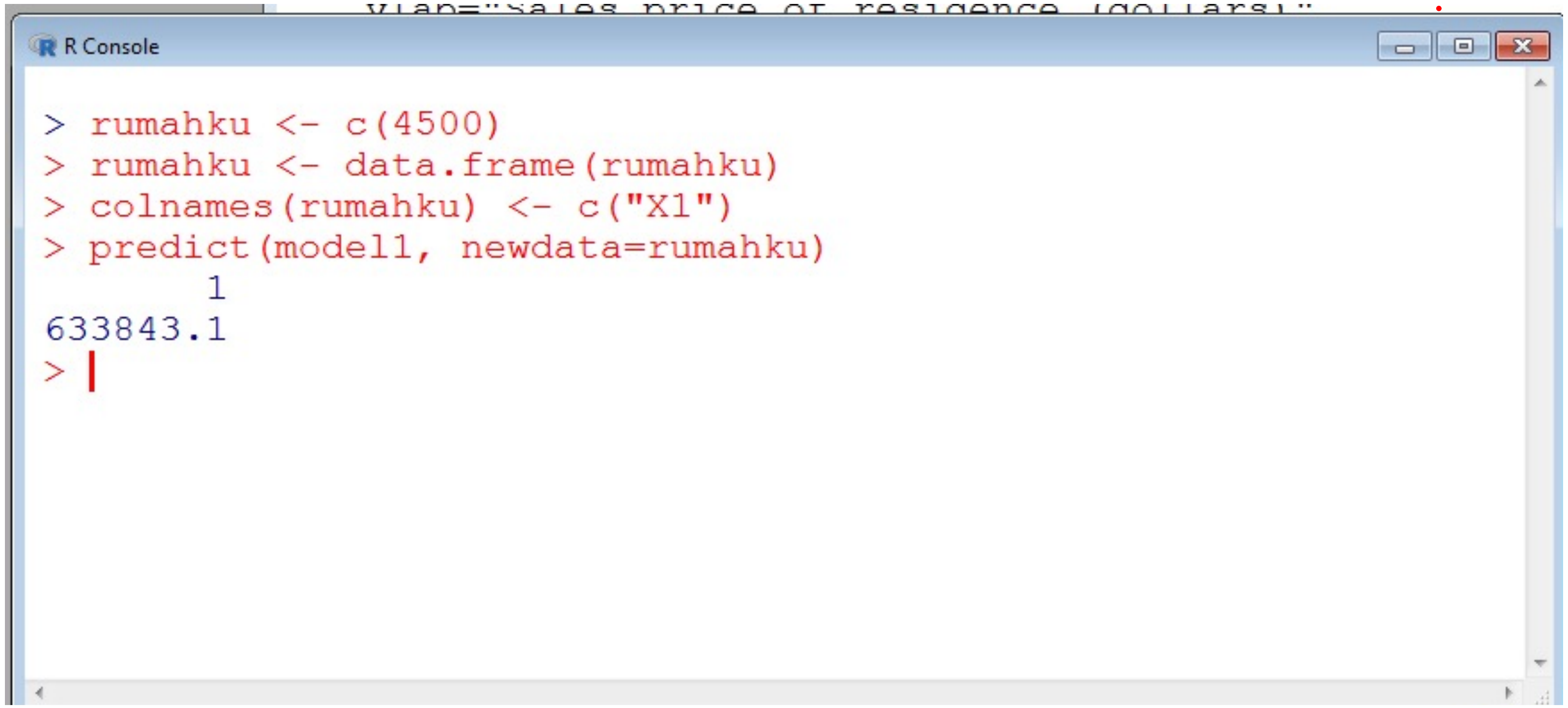
```
> lm(sales_price ~ 1 + X1, data=rumah)

Call:
lm(formula = sales_price ~ 1 + X1, data = rumah)

Coefficients:
(Intercept)           X1
     -81433          159

> 
```

```
model1 <- lm(sales_price ~ 1 + X1, data=rumah)
summary(model1)
```

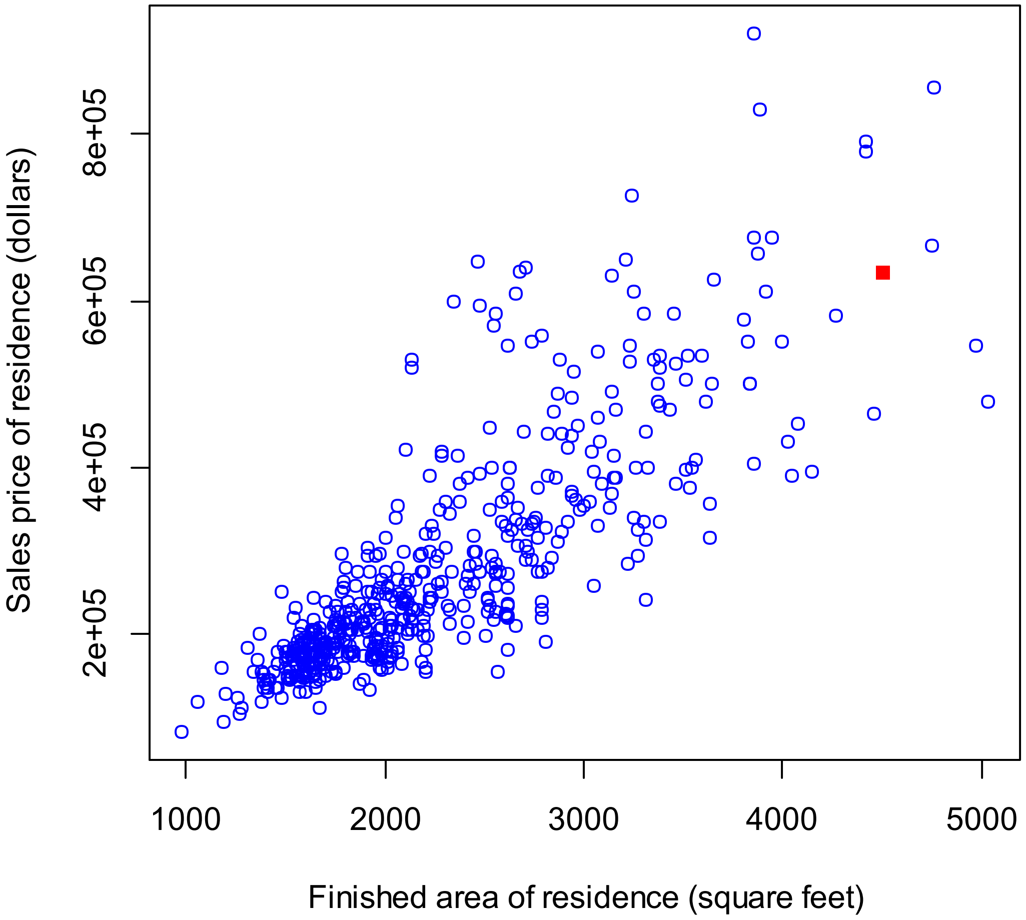
```
R Console

> model1 <- lm(sales_price ~ 1 + X1, data=rumah)
> summary(model1)

Call:
lm(formula = sales_price ~ 1 + X1, data = rumah)

Residuals:
    Min      1Q  Median      3Q     Max 
-239405  -39840   -7641   23515  388362 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -81432.946  11551.846  -7.049 5.74e-12 ***
X1             158.950      4.875  32.605  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 79120 on 520 degrees of freedom
Multiple R-squared:  0.6715,    Adjusted R-squared:  0.6709 
F-statistic:  1063 on 1 and 520 DF,  p-value: < 2.2e-16
```

Memprediksi Harga Rumah Seluas 4500 ft^2

```
rumahku <- c(4500)
rumahku <- data.frame(rumahku)
colnames(rumahku) <- c("X1")
predict(model1, newdata=rumahku)
```



```
plot(rumah$X1, rumah$sales_price,
       xlab="Finished area of residence (square feet)",
       ylab="Sales price of residence (dollars)",
       col="blue")
points(rumahku$X1, predict(model1, newdata=rumahku),
col="red", pch=15)
```



Model Regresi Linier Berganda

```
rumah$umur = 2023 - rumah$X7
rumah$ind.med <- ifelse(rumah$X8==2,1,0)
rumah$ind.low <- ifelse(rumah$X8==3,1,0)
model2 <- lm(sales_price ~ 1 + X1 + umur + ind.med + ind.low, data=rumah)
summary(model2)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.585e+05  2.102e+04   12.30  < 2e-16 ***
X1           9.756e+01  5.396e+00   18.08  < 2e-16 ***
umur        -1.119e+03  1.953e+02   -5.73  1.7e-08 ***
ind.med     -1.524e+05  1.040e+04  -14.65  < 2e-16 ***
ind.low     -1.709e+05  1.404e+04  -12.17  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 61770 on 517 degrees of freedom
Multiple R-squared:  0.801,   Adjusted R-squared:  0.7994
F-statistic: 520.1 on 4 and 517 DF,  p-value: < 2.2e-16
```

```
coef(model2)
> coef(model2)
  (Intercept)           X1         umur      ind.med      ind.low
 258477.75900     97.55652  -1119.15626 -152382.80986 -170876.66457
```

rumah indeks high

3 persamaan

① I. High → Int + X1 + umur

② I. med → Int + X1 + umur + ind.med

③ I. low → Int + X1 + umur + low

- Interpretasi
  - dugaan model yang diperoleh

    sales_price = 258478 + 97.56 X1 – 1119.16 Umur – 152382.80 Indeks Medium – 170876.66 Indeks Low
  - UMUR
    - Koef = -1119.16, artinya jika rumah bertambah tua satu tahun, harganya turun dengan rata-rata sebesar 1119.16 dollar (signifikan)
  - X1 (luas bangunan)
    - Koef = 97.56, artinya jika luas rumah bertambah satu feet^2, harganya naik dengan rata-rata sebesar 97.56 dollar (signifikan)
  - Indeks Medium
    - Rumah dengan indeks medium memiliki rata-rata harga rumah 152382.80 dollar lebih rendah dari rumah dengan indeks tinggi (high) (signifikan)

      referensi
  - Indeks Low
    - Rumah dengan indeks medium (low) memiliki rata-rata harga rumah 170876.66 dollar lebih rendah dari rumah dengan indeks tinggi (high) (signifikan)

      referensi

- atau

Langsung gunakan factor pada peubah kategorik

Kategori yang menjadi reference selalu kategori yang pertama

```
rumah$ind <- as.factor(rumah$X8)
str(rumah)
model3 <- lm(sales_price ~ 1 + X1 + umur + ind, data=rumah)
summary(model3)
coef(model3)
```

```
> coef(model3)
  (Intercept)            X1          umur          ind2          ind3
 258477.75900      97.55652   -1119.15626 -152382.80986 -170876.66457
```

Lalu bagaimana jika reference kategori peubah X8 diganti menjadi indeks “Low”?

```
##referensi X8 diganti menjadi "Low"
rumah$ind.high <- ifelse(rumah$X8==1,1,0)
str(rumah)
model4 <- lm(sales_price ~ 1 + X1 + umur + ind.high + ind.med, data=rumah)
summary(model4)
coef(model4)

> coef(model4)
 (Intercept)           X1         umur     ind.high      ind.med
 87601.09443     97.55652  -1119.15626 170876.66457  18493.85470
```

```
##dengan factor dari X8 (ind)
model5 <- lm(sales_price ~ 1 + X1 + umur + relevel(ind,ref=3), data=rumah)
summary(model5)
coef(model5)

> coef(model5)
          (Intercept)                    X1                  umur
          87601.09443              97.55652           -1119.15626
relevel(ind, ref = 3)1 relevel(ind, ref = 3)2
         170876.66457           18493.85470
```

Terima kasih ☺