Lecture 10: Regression & Correlation

Applied Statistics - STAN - 5.37 & 5.38
8 & 9 January 2021
Lecturer: Erika Siregar, SST, MS

Today’s Agenda

- Correlation
- Regression
- Multiple Regression

Overview

- Basically we'll learn about relationships between variables in dataset.
- What are variables? Take a look at these datasets.
- Datasets normally have > 1 variables.
- People are curious about whether there's relationship between pairs of variables.
- if relationship exists, what & how strong? → **correlation**.
- From the related variables, could we create a model so that we can predict y based on x? → **regression**
- Remember: only make a prediction using regression if correlation exists between the x (independent) and y (dependent).

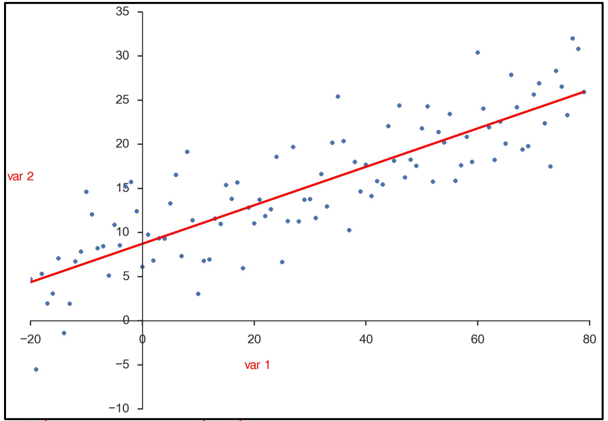
```
> datasets::airquality
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
7    23     299  8.6   65     5   7
```

```
> datasets::mtcars
                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
```

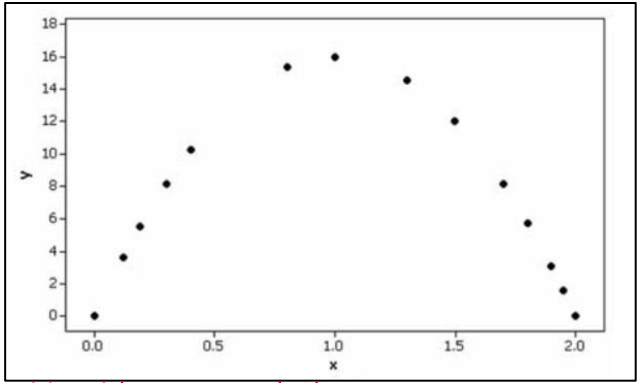
```
> head(datasets::iris, 7)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
7          4.6         3.4          1.4         0.3  setosa
```

Correlation

- Correlation:
  - relationship/association between 2 variables.
  - Values of one variable are somehow associated with the values of other variable.
- 2 types of correlation:
  - Linear → when the scatter plot of 2 correlated variables form an 'approximately' straight line.
  - Non linear → **not covered in this course.**



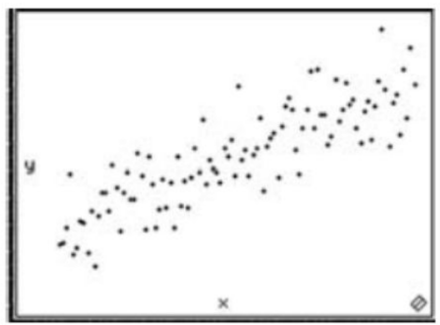
Linear correlation



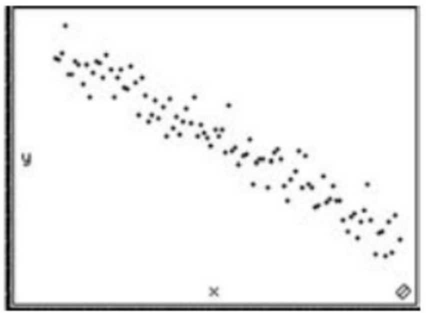
Non Linear correlation

Correlation (2)

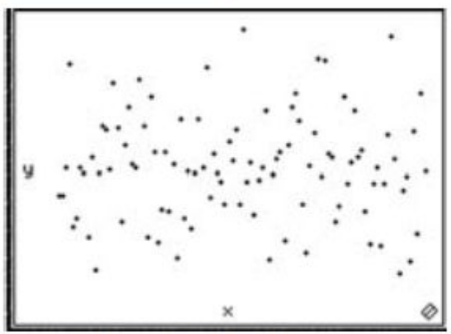
- Symbol:
  - ϱ (for population data)
  - r (for sample data)
  - Values are between -1 and 1 ->
    $-1 <= r <= 1$
- The relationship could be:
  - Linear negative correlation (-) → -1 = perfect negative
  - Linear positive correlation (+) → 1 = perfect positive correlation
  - No correlation (0)
  - ~~Non linear~~
- straight-line pattern → strong correlation



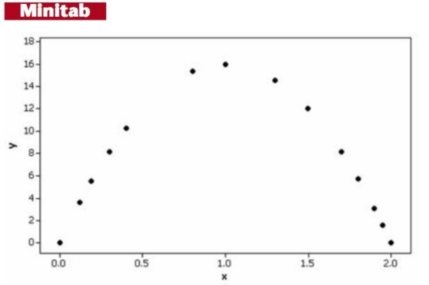
(a) Positive correlation: $r = 0.851$



(b) Negative correlation: $r = -0.965$



(c) No correlation: $r = 0$



(d) Nonlinear relationship: $r = -0.087$

Working with Correlation

1. **Detection** → does linear correlation exist?
   a. Scatter plot → subjective, eye measurement
   b. Checking on the requirements:
      i. Simple random sample
      ii. Scatter plot shows an ‘approximately’ straight pattern
      iii. Outliers are removed (if any)
   c. Calculation:
      i. Formula → cumbersome
      ii. Technology
   d. Hypothesis Test.
2. **Measurement**: how strong is the linear correlation?
   a. Represented with symbol ‘*r*’.
   b. Positif?
   c. Negative?
   d. No correlation.

How to compute correlation?

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

OR

$$r = \frac{\Sigma(z_x z_y)}{n - 1}$$

What a complicated formula. But technology could save you.

With R: cor(x, y)

Example of using correlation formula (1)

Table 10-2 Calculating r with Formula 10-1

x (Shoe Print)	y (Height)	x^2	y^2	xy
29.7	175.3	882.09	30730.09	5206.41
29.7	177.8	882.09	31612.84	5280.66
31.4	185.4	985.96	34373.16	5821.56
31.8	175.3	1011.24	30730.09	5574.54
27.6	172.7	761.76	29825.29	4766.52
$\Sigma x = 150.2$	$\Sigma y = 886.5$	$\Sigma x^2 = 4523.14$	$\Sigma y^2 = 157271.47$	$\Sigma xy = 26649.69$

Using Formula 10-1 with the results from Table 10-2, r is calculated as follows:

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

$$= \frac{5(26649.69) - (150.2)(886.5)}{\sqrt{5(4523.14) - (150.2)^2}\sqrt{5(157271.47) - (886.5)^2}}$$

$$= \frac{96.15}{\sqrt{55.66}\sqrt{475.10}} = 0.591$$

Example of using correlation formula (2)

Table 10-3 Calculating *r* with Formula 10-2

x (Shoe Print)	y (Height)	z_x	z_y	$z_x \cdot z_y$
29.7	175.3	−0.20381	−0.41035	0.08363
29.7	177.8	−0.20381	0.10259	−0.02091
31.4	185.4	0.81524	1.66191	1.35485
31.8	175.3	1.05501	−0.41035	−0.43292
27.6	172.7	−1.46263	−0.94380	1.38043
				$\Sigma(z_x \cdot z_y) = 2.36508$

$$z_x = \frac{x - \bar{x}}{s_x} = \frac{29.7 - 30.04}{1.66823} = -0.20381$$

$$r = \frac{\Sigma(z_x \cdot z_y)}{n - 1} = \frac{2.36508}{4} = 0.591$$

Using R:

```
> cor(foot_height$foot, foot_height$height)
[1] 0.5912691
```

Properties of Linear Correlation (r)

1. if all values of either variable are **converted to a different scale**, the value of **r does not change**.
2. The value of r is not affected by the choice of x and y. **Interchange all x- and y-values** and the value of r **will not change**.
3. R is very **sensitive to outliers**, they can dramatically affect its value.
4. correlation does not imply causality

Coefficient of Determination (r^2)

1. r=correlation
2. r^2 = proportion of the variation in **y** that is explained by the linear relationship between **x and y**.

$$r^2 = \frac{\text{explained variation.}}{\text{total variation}}$$

Example:
Using the pizza data fare costs, we found that r = 0.988. What proportion of the variation in the **subway fare (y)** can be explained by the variation in the **costs of a slice of pizza (x)**?

- **With r = 0.988, we get r2 = 0.976.**
- We conclude that 0.976 (or about 98%) of the variation in the cost of a subway fares can be explained by the linear relationship between the costs of pizza and subway fares.
- This implies that about 2% of the variation in costs of subway fares cannot be explained by the costs of pizza.

Hypothesis Test for Correlation (1)

1. Hypothesis:
   H0: $\rho = 0$
   H1: $\rho \neq 0$ or $\rho < 0$ or $\rho > 0$
2. 2 Choices of Test statistics:
   a. **r:**
      Decision:
      i. **Reject H0**: $|r| >$ critical value → there is sufficient evidence to support the claim of a linear correlation
      ii. **Fail to reject H0**: $|r| <=$ critical value → there is not sufficient evidence to support the claim of a linear correlation.
      iii. Critical value → Refer to Table A-5 in Triola's book (p.588)
      iv. Critical value using R:

```
# Pearson's critical value
critical.r <- function( n, alpha = .05 ) {
  df <- n - 2
  critical.t <- qt(alpha/2, df, lower.tail = F)
  critical.r <- sqrt( (critical.t^2) / ( (critical.t^2) + df ) )
  return(critical.r)
}
# Example usage: Critical correlation coefficient at sample size of n = 100
critical.r(23, 0.05)
```

Hypothesis Test for Correlation (2)

1. Hypothesis:
   H0: $\rho = 0$
   H1: $\rho \neq 0$ or $\rho < 0$ or $\rho > 0$
2. 2 Choices of Test statistics:
   b. **t**:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

df = n-2

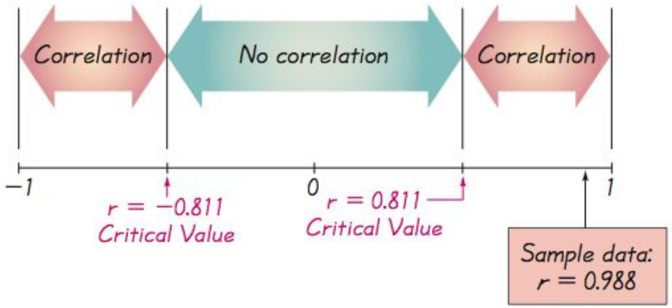
Decision:

i. **Reject H0**: p-value <= α → there is sufficient evidence to support the claim of a linear correlation

ii. **Fail to reject H0**: p-value > α. → there is **not sufficient evidence to support** the claim of a linear correlation.

Example 1

1. Use the paired pizza subway fare data, test the claim that there is a linear correlation between the costs of a slice of pizza and the subway fares. Use a 0.05 significance level.



Answer:

- H0: $\rho = 0$
- H1: $\rho \neq 0$
- $r = 0.988$.
- The critical value of $r = 0.811$ (Table A-5 with $n = 6$ and $\alpha = 0.05$)
- Because $|0.988| > 0.811$, we reject H0. (Rejecting "no linear correlation" indicates that there is a linear correlation.)

```
> cor(pizza$cost_of_pizza, pizza$subway_fare)
[1] 0.9878109
```

```
> critical.r(6, 0.05)
[1] 0.8114014
```

We conclude that there is sufficient evidence to support the claim of a linear correlation between costs of a slice of pizza and subway fares.

Example 1 (2)

Using test statistics t:

The linear correlation coefficient is r = 0.988 and n = 6 (six pairs of data), so the test statistic is $t = \dfrac{r}{\sqrt{\dfrac{1-r^2}{n-2}}} = \dfrac{0.988}{\sqrt{\dfrac{1-0.988^2}{6-2}}} = 12.793$

With df = 4, P-value < 0.01.

```
> cor.test(pizza$cost_of_pizza, pizza$subway_fare)

        Pearson's product-moment correlation

data:  pizza$cost_of_pizza and pizza$subway_fare
t = 12.692, df = 4, p-value = 0.000222
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8886647 0.9987251
sample estimates:
      cor
0.9878109
```

Example 2

2. With the choco data and α = 5%, compute:
   a. Test statistics r → try yourself based on slide 8
   b. Test statistics t

```
> cor.test(choco$chocolate, choco$nobel)

	Pearson's product-moment correlation

data:  choco$chocolate and choco$nobel
t = 6.123, df = 21, p-value = 4.477e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5797205 0.9118788
sample estimates:
      cor
0.8006078
```

- Decision?
- Interpretation?

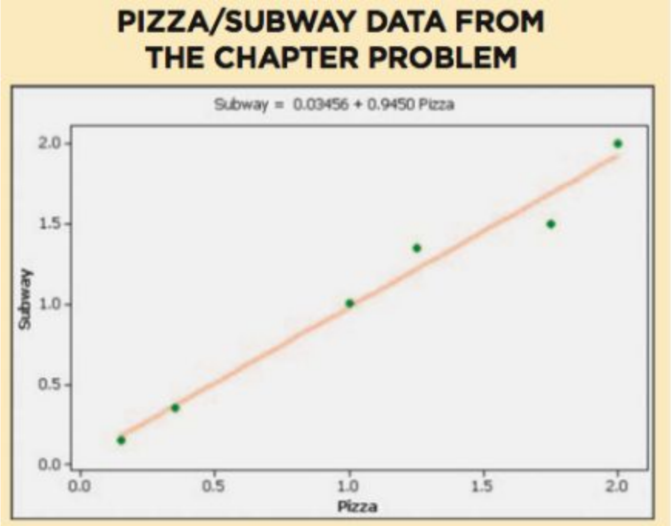
Part II: Regression

Regression

- The **regression equation** expresses a **relationship between x** (called the **explanatory** variable, **predictor** variable or **independent** variable), and **y** (called the **response** variable or **dependent** variable).
  - Equation: $\hat{y} = b_1 x + b_0$ → remember in math: $y = mx + b$?
  - b_1 = slope or gradient
  - b_0 = intercept
- It represents the straight line that best fits the paired sample data
- The **best-fitting straight line** is called a **regression line** a.k.a **line of best fit** a.k.a **least squares line**

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = r\frac{s_y}{s_x}$$



The slope **b1** represents the **marginal change** in y that occurs when x changes by one unit (besarnya perubahan y, ketika x berubah sebesar 1 unit).

Example

Table 10-1 Cost of a Slice of Pizza, Subway Fare, and the CPI

Year	1960	1973	1986	1995	2002	2003
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00
CPI	30.2	48.3	112.3	162.2	191.9	197.8

Based on the data above, create the regression equation, in which the explanatory variable (or **x** variable) is the **cost** of a slice of pizza and the response variable (or **y** variable) is the corresponding cost of a **subway fare**.

Tips: use R function lm().
Usage:
lm(y ~ x, data)

Answer

R script:

```
library(readr)
pizza <- read_csv('pizza.csv')

lm(subway_fare ~ cost_of_pizza, data = pizza)
```

Output:

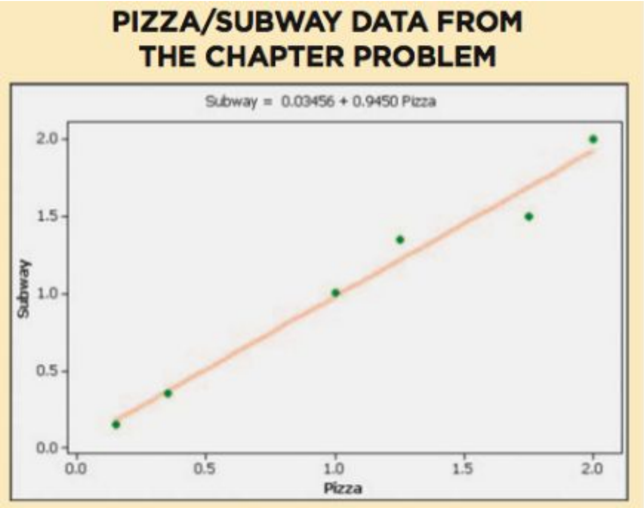
```
Coefficients:
  (Intercept)  cost_of_pizza
      0.03456        0.94502
```

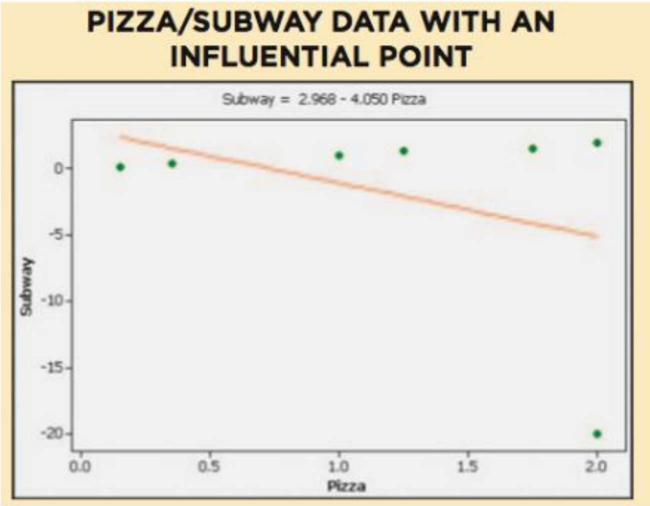
Regression Equation:

$\hat{y} = 0.03456 + 0.94502x$

year	cost_of_pizza	subway_fare	cpi
1960	0.15	0.15	30.2
1973	0.35	0.35	48.3
1986	1	1	112.3
1995	1.25	1.35	162.2
2002	1.75	1.5	191.9
2003	2	2	197.8

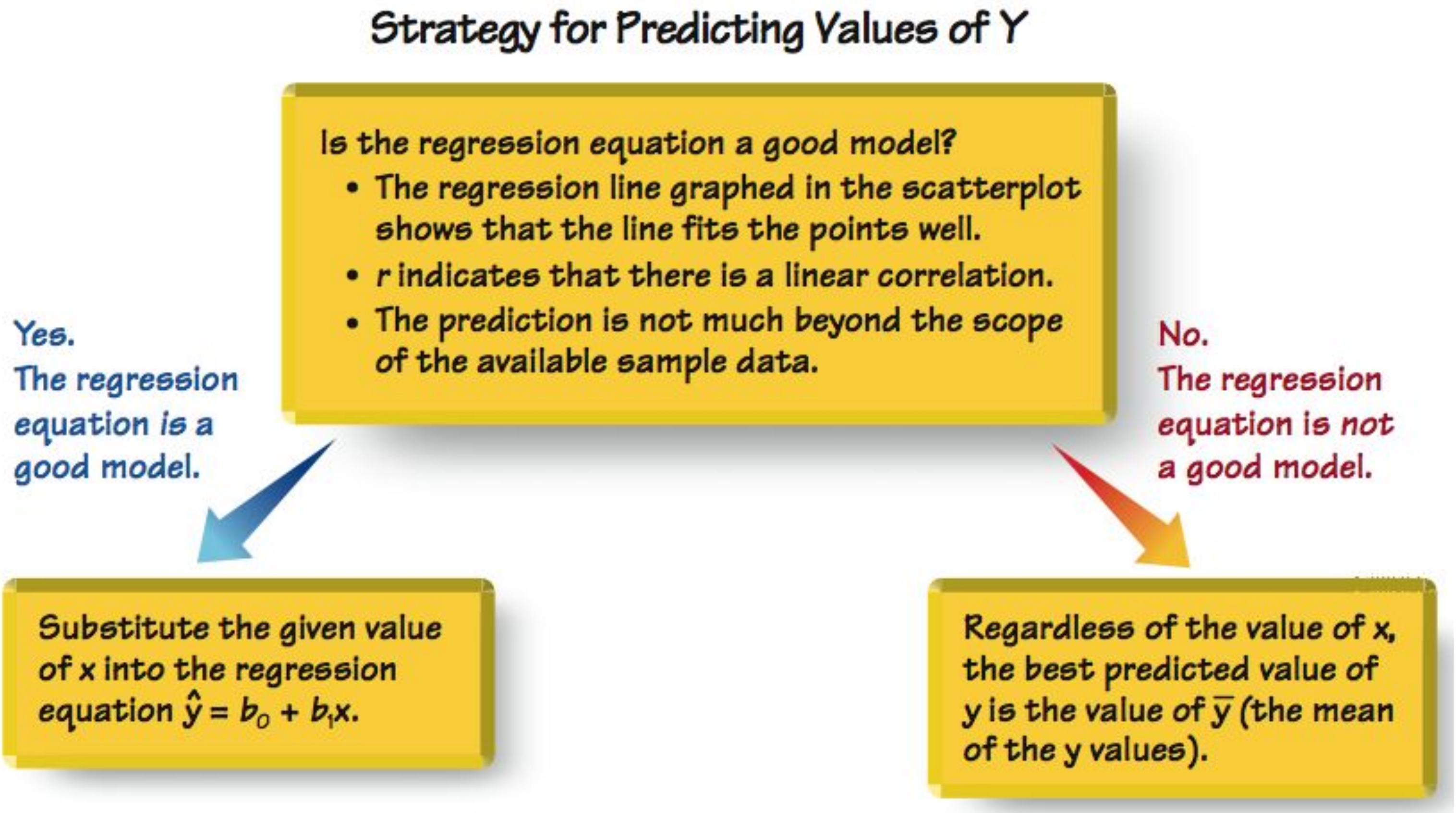
Influential Points





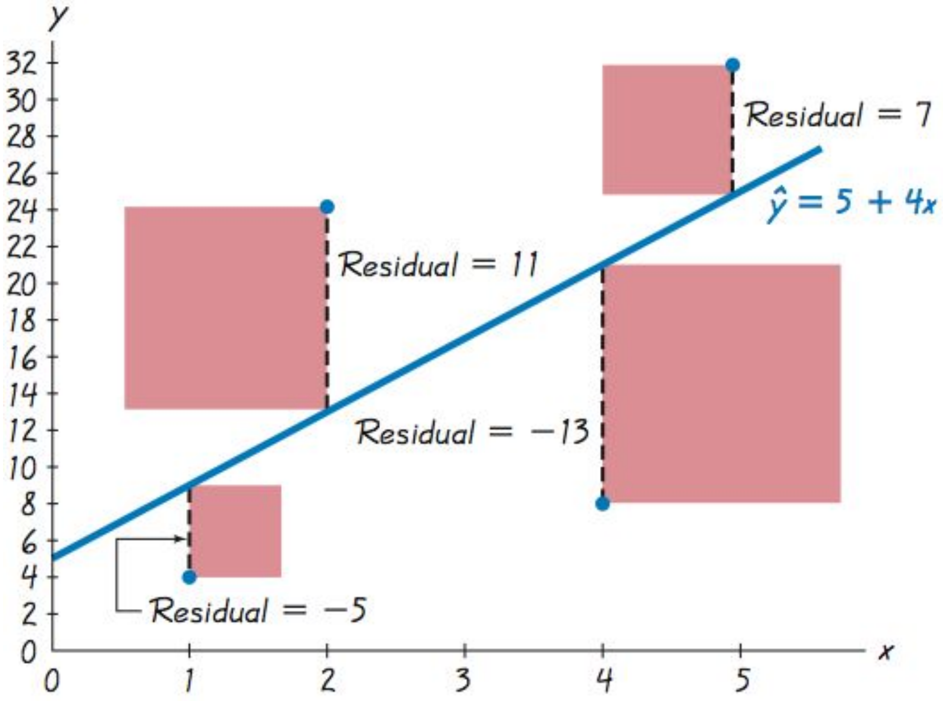
year	cost_of_pizza	subway_fare	cpi
1960	0.15	0.15	30.2
1973	0.35	0.35	48.3
1986	1	1	112.3
1995	1.25	1.35	162.2
2002	1.75	1.5	191.9
2003	2	2	197.8
2015	2	-20	

Strategy for predicting values of Y



Evaluating the Model

1. residual/error: observed y – predicted y = **y – ŷ**.



the **least-squares property** → if the sum of the squares of the residuals is the smallest sum possible

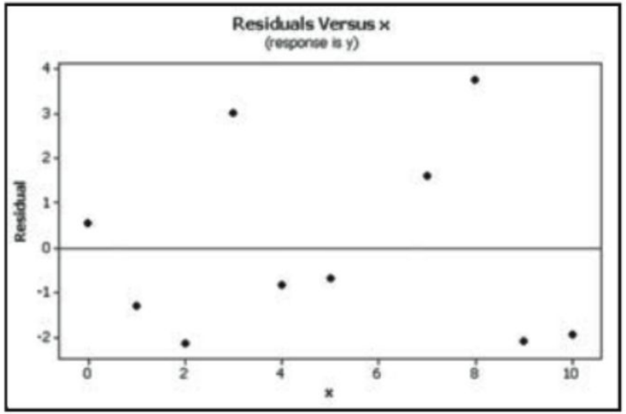
Residual Plot: a scatter plot of (x, y – ŷ).

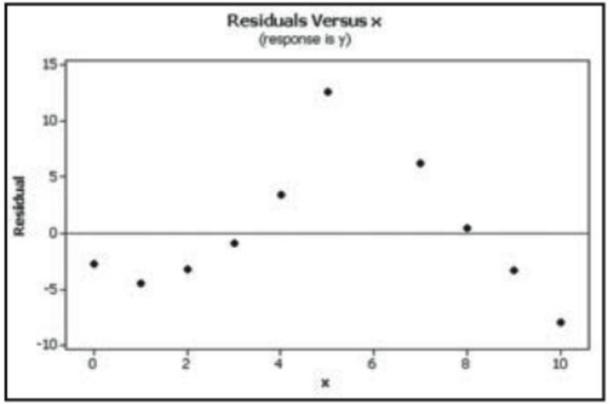
Residual plot suggests whether a regression equation is a good model or not.

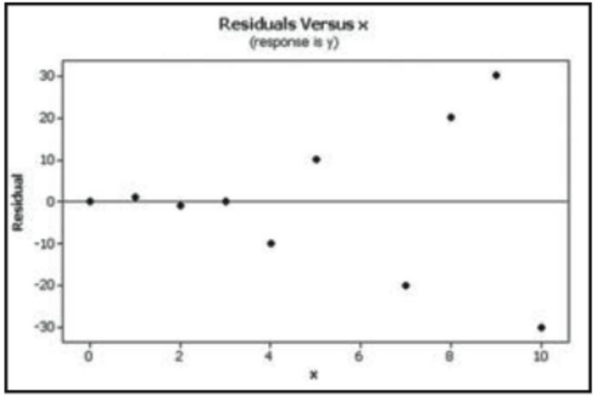
Residual Plots

Residual plots pattern:

1. should have no pattern (other than a straight-line pattern)
2. should not become thicker (or thinner) when viewed from left to right.

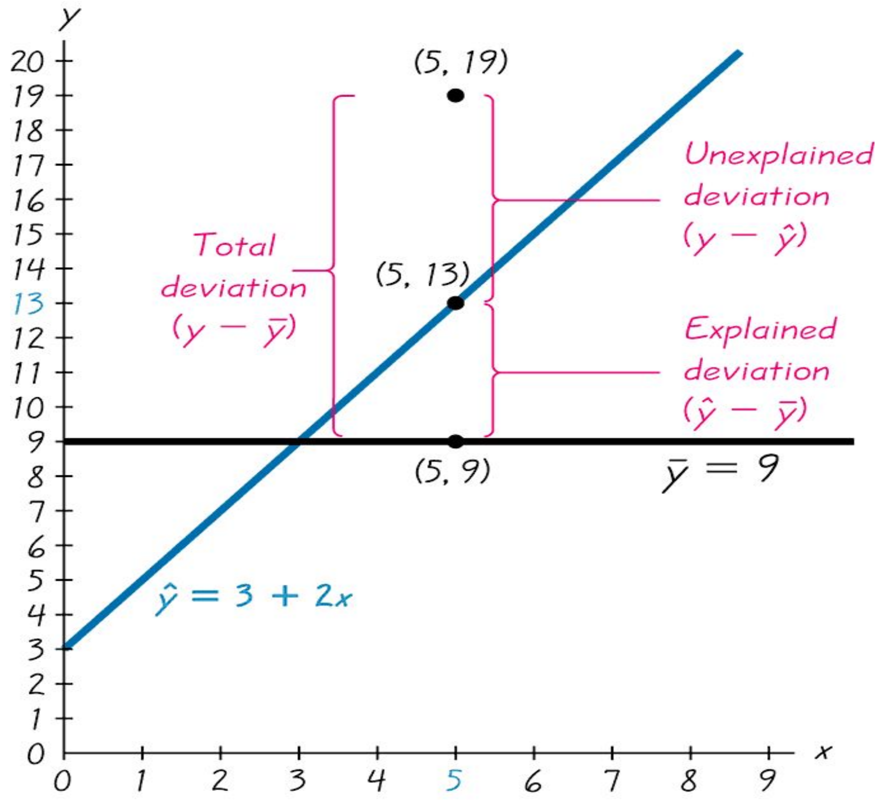






Which residual plot suggests that the regression equation is a good model.

Deviation pictured in residual plot



- Total deviation
- Explained deviation
- Unexplained deviation

Standard error (se)

(total deviation) = (explained deviation) + (unexplained deviation)

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

(total variation) = (explained variation) + (unexplained variation)

$$\Sigma (y - \bar{y})^2 = \Sigma (\hat{y} - \bar{y})^2 + \Sigma (y - \hat{y})^2$$

Standard Error

$$s_e = \sqrt{\frac{\Sigma (y - \hat{y})^2}{n - 2}}$$

or

$$s_e = \sqrt{\frac{\Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy}{n - 2}}$$

- The residual standard error (RSE) is a way to measure the standard deviation of the residuals in a regression model.
- <<< RSE, the better the model.
- RSE can be a useful metric to use when **comparing two or more models** to determine which model best fits the data.

R Script:

summary(regression_model)

Example

Use Formula 10-6 to find the standard error of estimate s_e for the paired pizza/subway fare data listed in Table 10-1in the Chapter Problem.

$n = 6$

$\Sigma y^2 = 9.2175$

$\Sigma y = 6.35$

$\Sigma xy = 9.4575$

$b_0 = 0.034560171$

$b_1 = 0.94502138$

$$s_e = \sqrt{\frac{\Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy}{n - 2}}$$

	year	cost_of_pizza	subway_fare	cpi
1	1960	0.15	0.15	30.2
2	1973	0.35	0.35	48.3
3	1986	1.00	1.00	112.3
4	1995	1.25	1.35	162.2
5	2002	1.75	1.50	191.9
6	2003	2.00	2.00	197.8

$$s_e = \sqrt{\frac{9.2175 - (0.034560171)(6.35) - (0.94502138)(9.4575)}{6 - 2}}$$

$s_e = 0.12298700 = 0.123$

Try it with R

Prediction Interval

- an interval estimate of a predicted value of y

$$\hat{y} - E < y < \hat{y} + E$$

Where:

  - $$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}}$$
  - x_0 represents the given value of x
  - $t_{\alpha/2}$ has n – 2 degrees of freedom

Example

For the paired pizza/subway fare costs from the Chapter Problem, we have found that for a pizza cost of \$2.25, the best predicted cost of a subway fare is \$2.16. Construct a 95% prediction interval for the cost of a subway fare, given that a slice of pizza costs \$2.25 (so that x = 2.25).

Answer:

$$E = t_{\alpha/2}\, s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}}$$

$$E = (2.776)(0.12298700)\sqrt{1 + \frac{1}{6} \quad \frac{6(2.25 - 1.0833333)^2}{6(9.77) - (6.50)^2}}$$

$$E = (2.776)(0.12298700)(1.2905606) = 0.441$$

Construct the CI

$$\hat{y} - E < y < \hat{y} + E$$

$$2.16 - 0.441 < y < 2.16 + 0.441$$

$$1.72 < y < 2.60$$

Try it with R

Multiple Regression

- 1 y, many x-es

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k.$$

- Adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{(n-1)}{[n-(k+1)]}(1 - R^2)$$

where n = sample size

k = number of predictor (x) variables

[redacted]

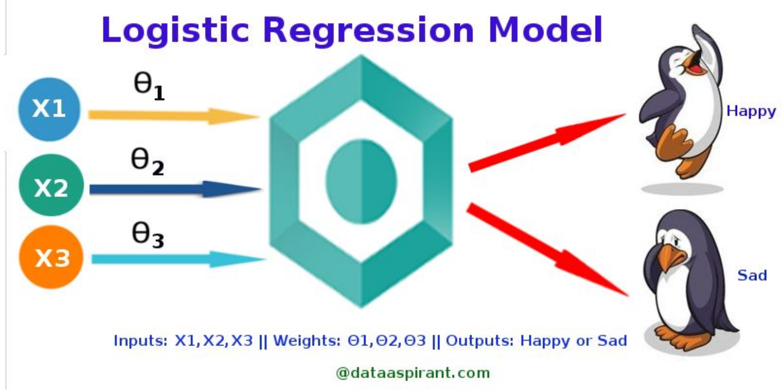
Height of Mother	Height of Father	Height of Daughter
63	64	58.6
67	65	64.7
64	67	65.3
60	72	61.0
65	72	65.4
67	72	67.4
59	67	60.9
60	71	63.1
58	66	60.0
72	75	71.1
63	69	62.2
67	70	67.2
62	69	63.4
69	62	68.4
63	66	62.2
64	76	64.7
63	69	59.6
64	68	61.0
60	66	64.0
65	68	65.4

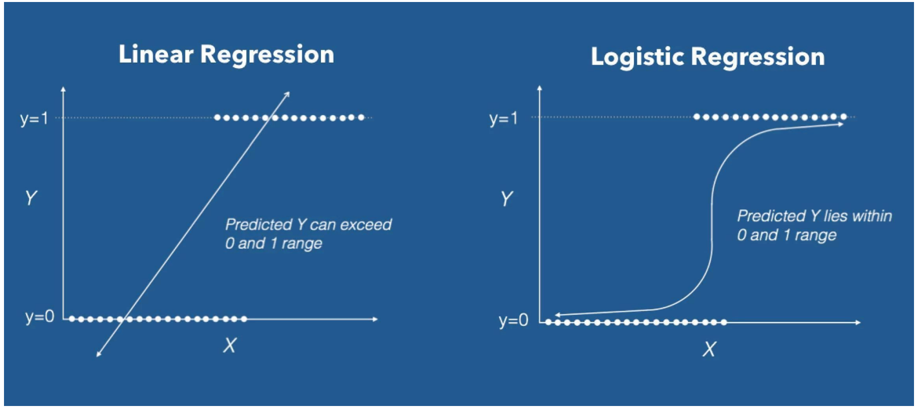
Find the multiple regression equation in which the response (y) variable is the height of a daughter and the predictor (x) variables are the height of the mother and height of the father.

```
> daughter_lm <- lm(daughter ~ mother + father, data = height)
> print(daughter_lm)

Call:
lm(formula = daughter ~ mother + father, data = height)

Coefficients:
(Intercept)       mother       father
     7.4543       0.7072       0.1636

> summary(daughter_lm)

Call:
lm(formula = daughter ~ mother + father, data = height)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8805 -0.6942  0.5915  0.8651  3.3138

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.4543    10.8804   0.685    0.503
mother        0.7072     0.1289   5.488    4e-05 ***
father        0.1636     0.1266   1.293    0.213
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.94 on 17 degrees of freedom
Multiple R-squared:  0.6752,    Adjusted R-squared:  0.637
F-statistic: 17.67 on 2 and 17 DF,  p-value: 7.057e-05
```

Logistic Variable

- Dichotomous/binary dependent variable →0/1, male/female, yes/no, etc





In R: glm(y ~ x1 + x2 + … + xk)

Exercise

Using data ‘salary.csv’

1. Uji hipotesis korelasi.
2. R dan R2 (manual or R) dan interpretasinya + Buat scatter plotnya
3. Buat regression equation + residual plot
4. Hitung standard error + make a prediction + interval estimate

Thanks!

ANY QUESTIONS?