Correlation and Regression

Jan Michael C. Yap
Core Facility for Bioinformatics
jcyap@up.edu.ph

Outline

- Correlation
- Regression

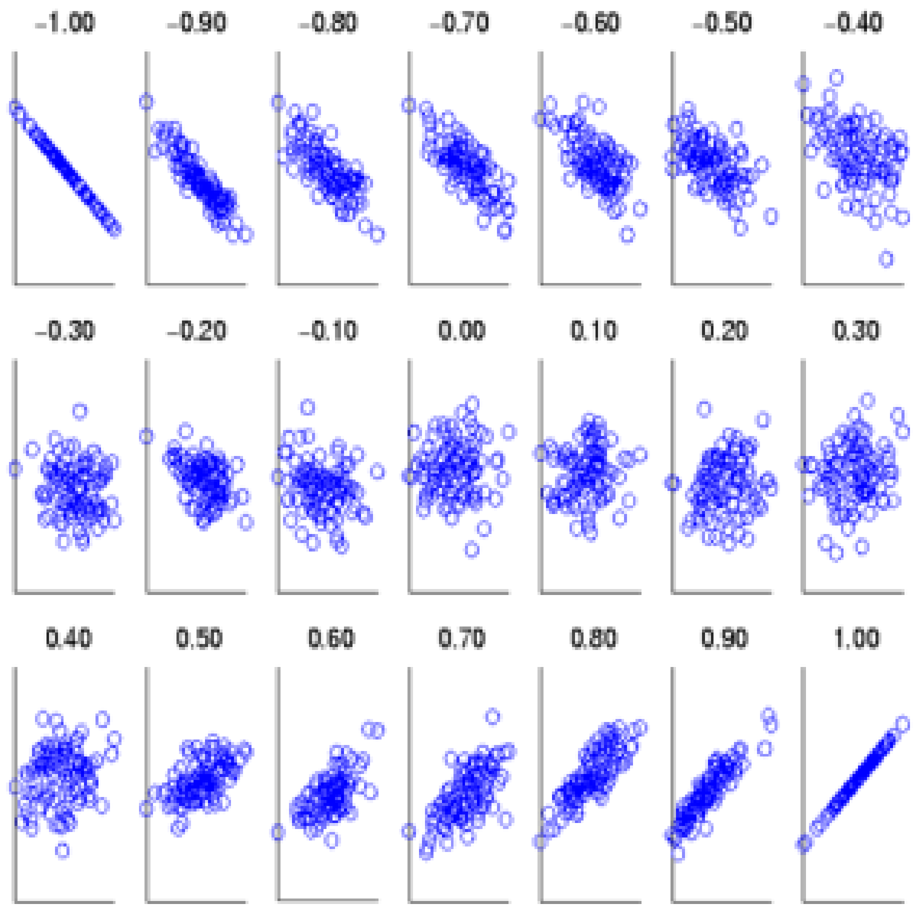
Outline

- **Correlation**
- Regression

(Statistical) Correlation

- Correlation describes the **degree of association or relationship between two datasets**
  - Can be extended to more than two datasets by performing **pairwise tests for all combinations of datasets**.
  - Assumption: values between two datasets are **paired**.
- Idea: do **values** between two datasets more or less **increase and decrease jointly?**
- Range of values: **[-1,1]**
  - **1: positively correlated; -1: negatively correlated; 0: no correlation / independent**
- Caveat: **correlation does not necessarily imply causation**
  - But can be treated as evidence towards it

Visualizing Correlation



Han J et al. (2011). Data Mining: Concepts and Techniques (3rd ed)

Pearson's Product-Moment Correlation Coefficient

- Given two datasets, X and Y, the Pearson's product-moment correlation coefficient is computed as:

$$r_{x,y} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

X	10	13	14	16	18	19	20	7	11	3
Y	13	14	8	5	9	11	17	16	5	1

$r_{x,y} = 0.341916$

PGC
PHILIPPINE GENOME CENTER

Rank Correlation Coefficients

- Makes use of the **order of the values (i.e. rank) when sorted according to magnitude**
- Preprocessing step: **arrange the values in each dataset and assign ranks**
- **Spearman's ρ**
  - Apply **Pearson's correlation on the ranks**

PGC
PHILIPPINE GENOME CENTER

Rank Correlation Coefficients

- Kendall's τ

$$\tau = \frac{n_c - n_d}{0.5n(n-1)}$$

  - n_c refers to **concordant pairs,** i.e. number of samples where if rank(x_i) > rank(x_j) then rank(y_i) > rank(y_j) OR if where if rank(x_i) < rank(x_j) then rank(y_i) < rank(y_j), for $1 \leq i < j \leq n$
  - n_d refers to **discordant pairs,** i.e. number of samples where if rank(x_i) > rank(x_j) then rank(y_i) < rank(y_j) OR if where if rank(x_i) < rank(x_j) then rank(y_i) > rank(y_j), for $1 \leq i < j \leq n$

PHILIPPINE GENOME CENTER

Rank Correlation Coefficients

X	10	13	14	16	18	19	20	7	11	3
Y	13	14	8	5	9	11	17	16	5	1

Rank X	8	6	5	4	3	2	1	9	7	10
Rank Y	4	3	7	8	6	5	1	2	8	10

$\rho = 0.289704$

C	3	3	5	4	4	3	3	1	1	0
D	6	5	2	2	1	1	0	1	0	0

$\tau = 0.2$

PHILIPPINE GENOME CENTER

Significance Testing for Correlation Coefficients

- Pearson (and Spearman)
  - Pearson's correlation coefficient value is first subjected to **Fisher transformation, F(r) = arctanh(r)**
  - F(r) is **normally distributed with mean = F(r_0) and standard deviation = 1 / $\sqrt{n-3}$**
- Kendall
  - Testing the significance requires computing a **different standard normally distributed test statistic**:

$$z_A = \frac{3(nc - nd)}{\sqrt{n(n-1)(2n+5)/2}}$$

PHILIPPINE GENOME CENTER

Outline

- Correlation
- **Regression**

Regression

- Regression is a modeling technique used to show the **relationship of one dependent variable with one or more independent variables**.
- The dependent variable is also termed as **response or output variable.**
- The independent variable is also termed as **predictor, input, or explanatory variable.**

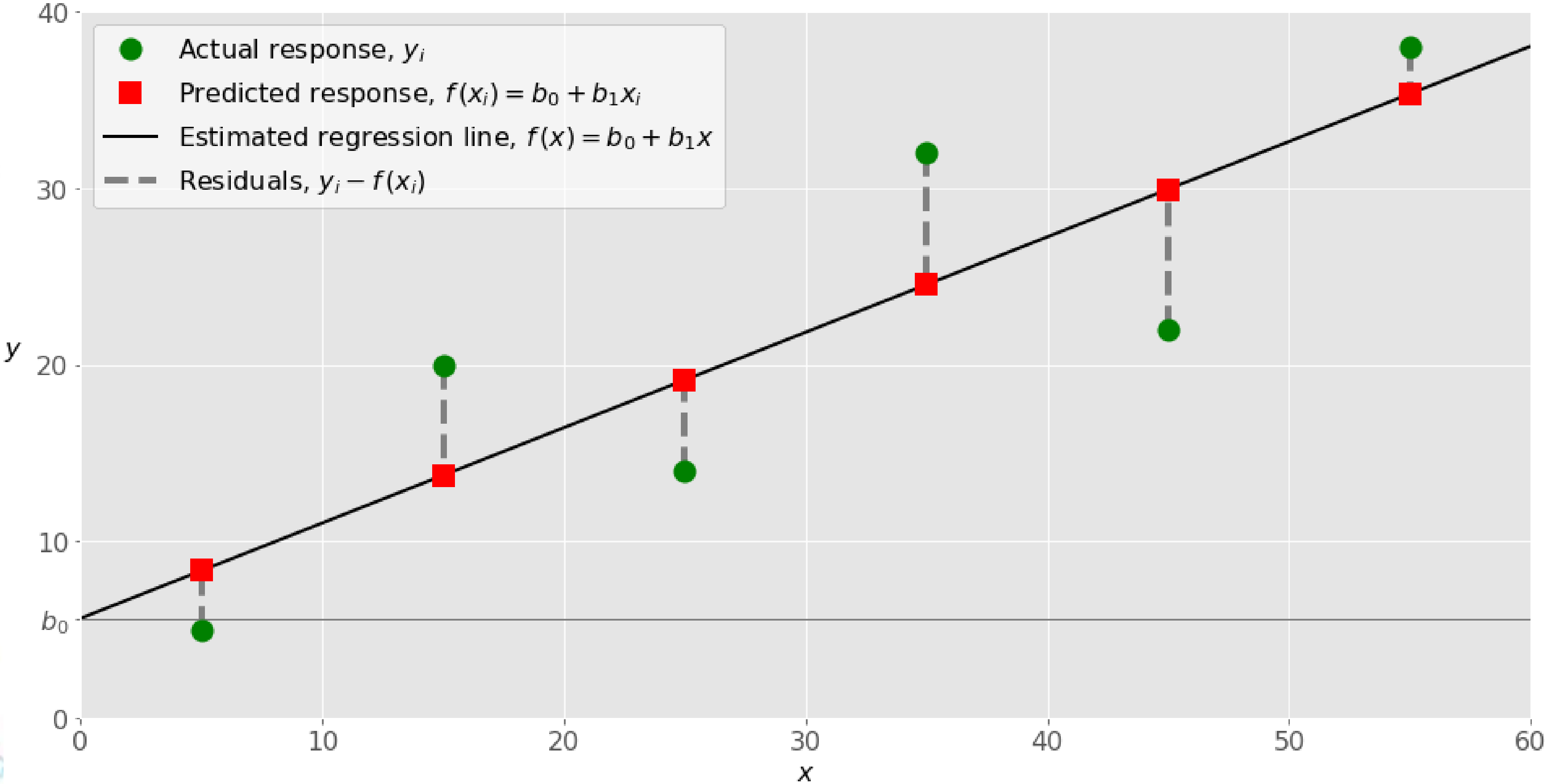
PHILIPPINE GENOME CENTER

Regression

- Regression is done by **fitting a function** wherein the dependent variable is determined by the independent variable(s)
- Let Y be the dependent variable, and X be the independent variable, we try to **regress Y using X via a function f** such that **Y = f(X)**
- For multiple independent variables, $X_1, X_2, \ldots, X_n$, we have $Y = f(X_1, X_2, \ldots, X_n)$.
- This has implicit assumption that the **dependent variable(s) are "causal" to the independent variable**

PHILIPPINE GENOME CENTER

(Least Squares) Linear Regression

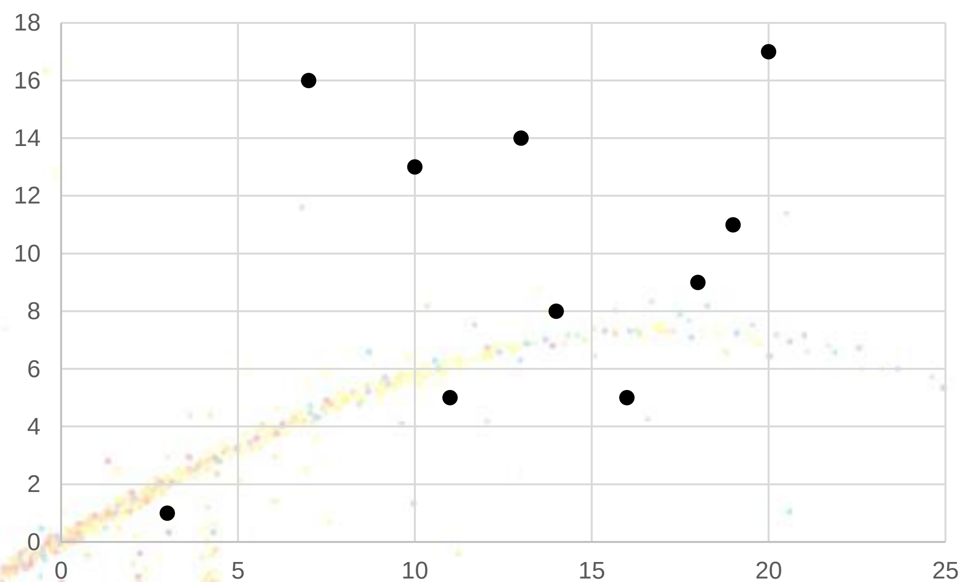


https://files.realpython.com/media/fig-lin-reg.a506035b654a.png

PHILIPPINE GENOME CENTER

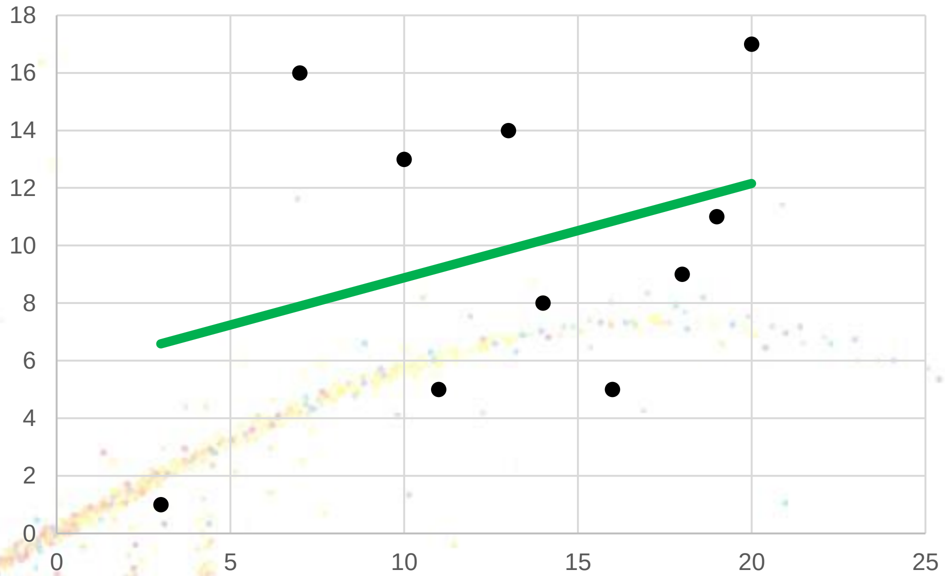
Least Squares Linear Regression

X	10	13	14	16	18	19	20	7	11	3
Y	13	14	8	5	9	11	17	16	5	1



Least Squares Linear Regression

X	10	13	14	16	18	19	20	7	11	3
Y	13	14	8	5	9	11	17	16	5	1



Least Squares Linear Regression

$$y = mx + b$$

$$b = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$m = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

X	10	13	14	16	18	19	20	7	11	3
Y	13	14	8	5	9	11	17	16	5	1

m = 0.32763 b = 5.60803

https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/

Significance Testing for Linear Regression

$$\boldsymbol{s_{est}} = \sqrt{\frac{\sum(\boldsymbol{y} - \widehat{\boldsymbol{y}})^2}{\boldsymbol{N} - 2}}$$

- **Standard error** of the regression model
- Perform **t-test** where the t-statistic is computed as

$$t = \frac{\boldsymbol{s_{est}}}{\sqrt{\sum(x - \bar{x})^2}}$$

onlinestatbook.com/2/regression/accuracy.html

Relationship between Correlation and Regression

$$m = r_{x,y} \frac{stdev(y)}{stdev(x)}$$

X	10	13	14	16	18	19	20	7	11	3
Y	13	14	8	5	9	11	17	16	5	1

$r_{x,y}$ = 0.341916

m = 0.32763   b = 5.60803

stdev(x) = 5.1856
stdev(y) = 4.9689

Multiple (Linear) Regression

- Multiple regression is done when the **dependent variable is regressed using more than one independent variables** (at a time)
- It is done to **assess effect of one independent variable to joint effects of multiple independent variables.**
- In practice, multiple regression is done only on **linear models**.

$$y = m_1 x_1 + m_2 x_2 + \cdots + b$$

THANK YOU VERY MUCH! ☺

PHILIPPINE GENOME CENTER