Chapter 11: Simple Linear Regression & Correlation

MAS291 - STATISTICS & PROBABILITY

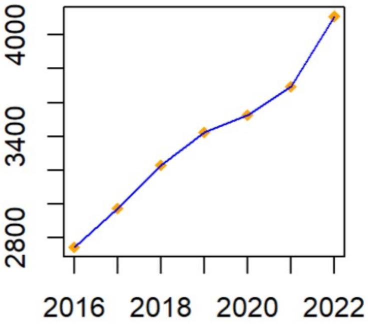
Ly Anh Duong

duongla3@fe.edu.vn

Table of Contents

Introduction – GDP per capita (2016-2022) Vietnam

Year	GDP per capita ($)
2016	2746
2017	2974
2018	3231
2019	3425
2020	3526
2021	3694
2022	4110



Our goal: to build a model to predict GDP per capita in 2025

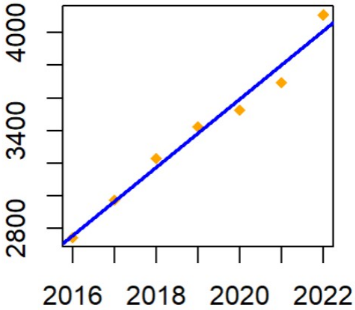
Introduction – GDP per capita (2016-2022) Vietnam

Fitted Regression Line:

GDP per capita = 208*year – 416781

In 2025:

GDP per capita = 208*2025 – 416781 = $4635



Regression (an empirical model)

We have two variables x (nonrandom), Y (random): numerical data.
We believe that Y depends in some way on x:

$$E(Y|x)=f(x)$$

Dependent variable
Response variable
Target variable

Independent variable
Predictor
Explanatory variable
Regressor

Example: (x, Y) pairs:
$x=$ study time and $Y=$ score on a test.
$x=$ smoking frequency and $Y=$ age of first heart attack.
Given information about x and Y, we would like to predict future values of Y for particular values of x. $\Longrightarrow$ Estimate $\mathrm{E}(Y|x)$.

Simple linear regression

1 Simple Linear Regression

Assume that each observation x, Y can be describe by the model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

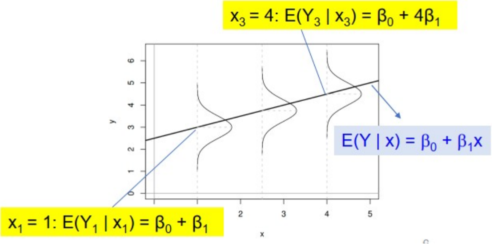
where ϵ is a random error with mean $E(\epsilon) = 0$ and unknown variance σ^2.

$$E(Y|x) = \beta_0 + \beta_1 x$$

β_0 and β_1 are unknown regression coefficients $\Longrightarrow$ to be estimated.

$\Longrightarrow$ Estimate, perform hypothesis tests on these parameters

$\Longrightarrow$ prediction



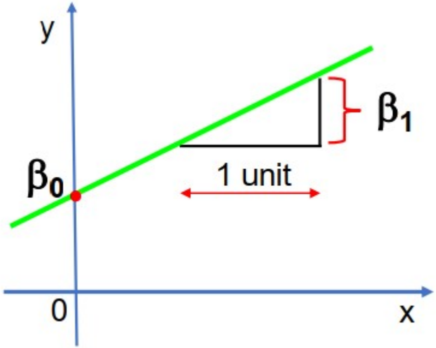
Intercept and Slope

- **Regression model:**

$$Y = \beta_0 + \beta_1 x + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\Longrightarrow E(Y|x) = \beta_0 + \beta_1 x$$

- **Intercept:** $\beta_0 = E(Y|x = 0)$
- **Slope β_1:** β_1 is how much Y changes (on average) when x increases by 1 unit.



- **Regression model:**

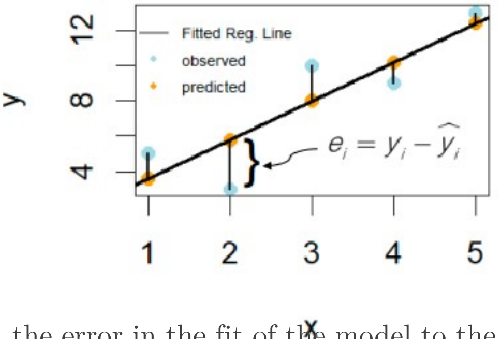
$$Y = \beta_0 + \beta_1 x + \epsilon$$

- **Estimated regression line:**

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1} x$$

- **Residual of i^{th} datapoint:**

$$e_i = y_i - \hat{y}_i$$

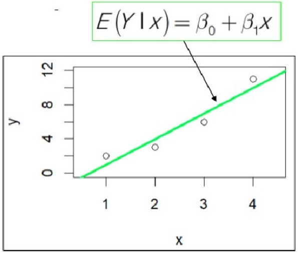


Error term $e_i = y_i - \hat{y}_i$, is called i^{th} residual, the error in the fit of the model to the i^{th} observation y_i.

Method of OLS (Ordinary least Squares) - Example

- Given observations

x	1	2	3	4
y	2	3	6	11



- Model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

$Y_1 = \beta_0 + \beta_1 x_1 + \epsilon_1,\ Y_2 = \beta_0 + \beta_1 x_2 + \epsilon_2$

$Y_3 = \beta_0 + \beta_1 x_3 + \epsilon_3,\ Y_4 = \beta_0 + \beta_1 x_4 + \epsilon_4$

$$L(\beta_0, \beta_1) = (2 - \beta_0 - \beta_1)^2 + (3 - \beta_0 - 2\beta_1)^2 + (6 - \beta_0 - 3\beta_1)^2 + (11 - \beta_0 - 4\beta_1)^2$$

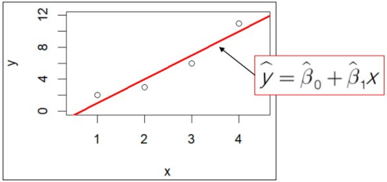
Method of OLS (Ordinary least Squares) - Example

1 Simple Linear Regression

$$L\left(\beta_0, \beta_1\right) = \left(2 - \beta_0 - \beta_1\right)^2 + \left(3 - \beta_0 - 2\beta_1\right)^2 + \left(6 - \beta_0 - 3\beta_1\right)^2 + \left(11 - \beta_0 - 4\beta_1\right)^2$$

- Given observations

x	1	2	3	4
y	2	3	6	11



$$\frac{\partial L}{\partial \beta_0} = -2\left(22 - 4\beta_0 - 10\beta_1\right),$$

$$\frac{\partial L}{\partial \beta_1} = -2\left(70 - 10\beta_0 - 30\beta_1\right)$$

OLS estimates of β_0, β_1

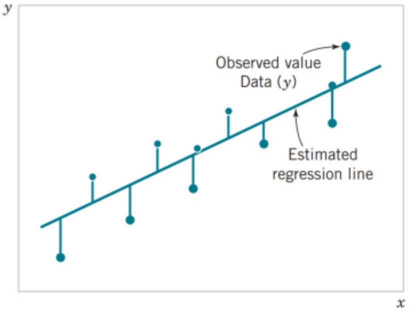
$$\frac{\partial L(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_0} = 0, \quad \frac{\partial L(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_1} = 0$$

$$\Longleftrightarrow \hat{\beta}_0 = -2, \ \hat{\beta}_1 = 3$$

Method of least squares (OLS)

The method of least squares is used to estimate the parameters β_0 and β_1 by minimizing L, the sum of the squares of the vertical deviations.

$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$



$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0$$

Simplifying these two equations yields:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

The least squares estimates of the intercept and slope in the simple linear regression model are:

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}}$$

where $\bar{y} = \dfrac{1}{n}\sum_{i=1}^n y_i$ and $\bar{x} = \dfrac{1}{n}\sum_{i=1}^n x_i$.

Notation:

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}$$

Estimating σ^2 Error sum of squares SS_E

1 Simple Linear Regression

The error sum of squares is:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SS_T - \beta_1 S_{xy}$$

where $SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i)^2 - n\bar{y}^2$ (Total Sum of Squares)
It can be shown that the expected value of the error sum of squares is:

$$E(SSE) = (n-2)\sigma^2$$

An **unbiased estimator** of σ^2 is:

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

Table of Contents

Properties of the least squares Estimators

- **Slope Properties:**

$$E(\hat{\beta}_1) = \beta_1 \quad \text{and} \quad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

- **Intercept Properties:**

$$E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad V(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$$

In simple linear regression, the estimated standard error of the slope and intercept are:

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad \text{and} \quad se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

Table of Contents

t-test on β_1

Suppose we wish to test

$$H_0 : \beta_1 = \beta_{1,0} \quad \text{vs} \quad H_1 : \beta_1 \neq \beta_{1,0}$$

An appropriate test statistic would be

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

We would reject the null hypothesis if

$$|t_0| > t_{\alpha/2, n-2}$$

t-test on β_0

Suppose we wish to test

$$H_0 : \beta_0 = \beta_{0,0} \quad \text{vs} \quad H_1 : \beta_0 \neq \beta_{0,0}$$

An appropriate test statistic would be

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}} = \frac{\hat{\beta}_0 - \beta_{0,0}}{se(\hat{\beta}_0)}$$

We would reject the null hypothesis if

$$|t_0| > t_{\alpha/2,n-2}$$

ANOVA Identity:

$$\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$(SS_T = SS_R + SS_E)$$

Total variation = Explained variation + Unexplained variation.

SS_R: Regression Sum of Squares
$\Longrightarrow$ (variation explained by linear model)

SS_E: Error Sum of Squares
$\Longrightarrow$ (unexplained variation)

ANOVA F-test

F-test ($F_{1,n-2}$ distribution):

$$F_0 = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E}, \text{ Note that } MS_E = \hat{\sigma}^2$$

Reject $H_0: \beta_1 = 0$ if $f_0 > f_{\alpha,1,n-2}$

TABLE 11.3 Analysis of Variance for Testing Significance of Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R/MS_E
Error	$SS_E = SS_T - \hat{\beta}_1 S_{xy}$	$n-2$	MS_E	
Total	SS_T	$n-1$		

We reject $H_0: \beta_1 = 0$ when F is large – that is, when the explained variation is large relative to the unexplained variation.

Table of Contents

We assume that the joint distribution of X_i and Y_i is the bivariate normal distribution presented in Chapter 5, and μ_Y and σ_Y^2 are the mean and variance of Y, μ_X and σ_X^2 are the mean and variance of X, and ρ is the **correlation coefficient** between Y and X.

Correlation coefficient:

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \; -1 \le \rho \le 1$$

where σ_{XY}: covariance

Sample correlation coefficient:

$$R = \frac{\sum_{i=1}^n Y_i(X_i - \bar{X})}{\left[\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2\right]^{1/2}} = \frac{S_{XY}}{(S_{XX} S_{YY})^{1/2}}$$

Test Statistic for Zero Correlation

Hypotheses: $H_0 : \rho = 0 \quad \text{vs} \quad H_1 : \rho \neq 0$

The appropriate test statistic for these hypotheses is: $T_0 = \dfrac{R\sqrt{n-2}}{\sqrt{1-R^2}}$

(t distribution with $n-2$ degrees of freedom if $H_0 : \rho = 0$ is true)

Reject H_0 if $|T_0| > t_{\alpha/2, n-2}$

Note that:

$$R = \frac{\sum_{i=1}^{n} Y_i (X_i - \bar{X})}{\left[\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2\right]^{1/2}} = \frac{S_{XY}}{(S_{XX} S_{YY})^{1/2}}$$

$$R^2 = \hat{\beta}_1^2 \frac{S_{XX}}{S_{YY}} = \frac{\hat{\beta}_1 S_{XY}}{SS_T} = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}, \quad 0 \leq R^2 \leq 1$$

$$\hat{\beta}_1 = \left(\frac{SS_T}{S_{XX}}\right)^{1/2} R$$

Example

4 Correlation

Chapter 1 (Section 1.3) describes an application of regression analysis in which an engineer at a semiconductor assembly plant is investigating the relationship between pull strength of a wire bond and two factors: wire length and die height. In this example, we consider only one of the factors, the wire length. A random sample of 25 units is selected and tested, and the wire bond pull strength and wire length are observed for each unit. We assume that pull strength and wire length are jointly normally distributed.

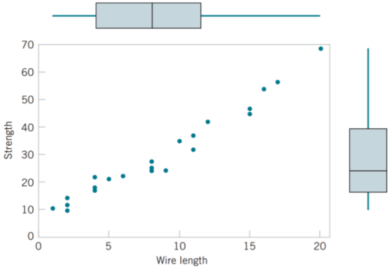
Table 1-2 Wire Bond Pull Strength Data

Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2
1	9.95	2	50
2	24.45	8	110
3	31.75	11	120
4	35.00	10	550
5	25.02	8	295
6	16.86	4	200
7	14.38	2	375
8	9.60	2	52
9	24.35	9	100
10	27.50	8	300
11	17.08	4	412
12	37.00	11	400
13	41.95	12	500
14	11.66	2	360
15	21.65	4	205
16	17.89	4	400
17	69.00	20	600
18	10.30	1	585
19	34.93	10	540
20	46.59	15	250
21	44.88	15	290
22	54.12	16	510
23	56.63	17	590
24	22.13	6	100
25	21.15	5	400

Example

4 Correlation

Figure 11.13 shows a scatter diagram of wire bond strength versus wire length. We have displayed box plots of each individual variable on the scatter diagram




```
Strength = 5.11 + 2.90 Length
Predictor    Coef    SE Coef     T      P
Constant     5.115   1.146      4.46   0.000
Length       2.9027  0.1170    24.80   0.000
S = 3.093   R-sq = 96.4%  R-sq(adj) = 96.2%
PRESS = 272.144   R-sq(pred) = 95.54%
Analysis of Variance
Source       DF     SS       MS       F       P
Regression    1   5885.9   5885.9   615.08  0.000
Residual     23    220.1      9.6
  Error
Total        24   6105.9
```

$S_{xx} = 698.56$ and $S_{xy} = 2027.7132$, and the sample correlation coefficient is:

$$r = \frac{S_{xy}}{[S_{xx} SS_T]^{1/2}} = \frac{2027.7132}{[(698.56)(6105.9)]^{1/2}} = 0.9818$$

Note that:

$$r^2 = (0.9818)^2 = 0.9640$$

Suppose that we wish to test the hypothesis:

$$H_0 : \rho = 0 \quad \text{vs} \quad H_1 : \rho \neq 0$$

With $\alpha = 0.05$, we can compute the t-statistic of Equation 11-46 as:

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.9818\sqrt{23}}{\sqrt{1-0.9640}} = 24.8$$

This statistic is also reported in the Minitab output as a test of $H_0 : \beta_1 = 0$.
Because $t_{0.025,23} = 2.069$, we reject H_0 and conclude that the correlation coefficient $\rho \neq 0$.
Finally, we may construct an approximate 95% confidence interval on ρ from Equation 10-57. Since arctanh $r =$ arctanh $0.9818 = 2.3452$, Equation 11-50 becomes:

$$\tanh\left(2.3452 - \frac{1.96}{\sqrt{22}}\right) \leq \rho \leq \tanh\left(2.3452 + \frac{1.96}{\sqrt{22}}\right)$$

which reduces to:

$$0.9585 \leq \rho \leq 0.9921$$

Q&A

Thank you for listening!