8. REGRESSION ANALYSIS

Correlation, simple linear regression, multiple linear regression

Universiti Kebangsaan Malaysia
The National University of Malaysia

Correlation

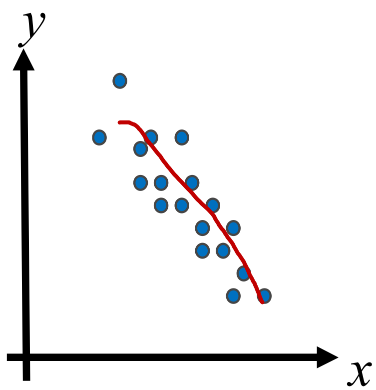
Correlation

- We are interested in looking at the relationship between two variables.

- The data can be represented by ordered pairs (x, y)
  - x is the independent (or explanatory) variable
  - y is the dependent (or response) variable

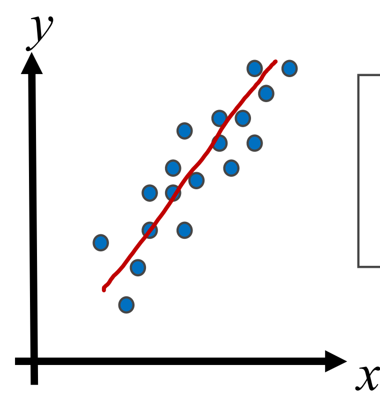
- We can draw a scatter plot to visually inspect the relationship

Types of correlation



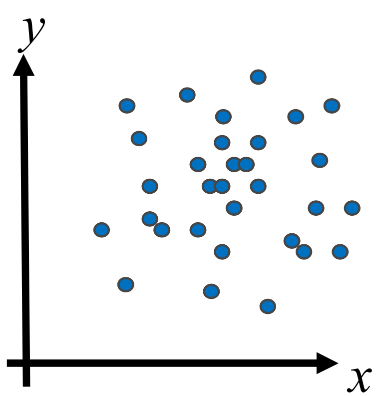
As x increases, y tends to decrease.

Negative Linear Correlation

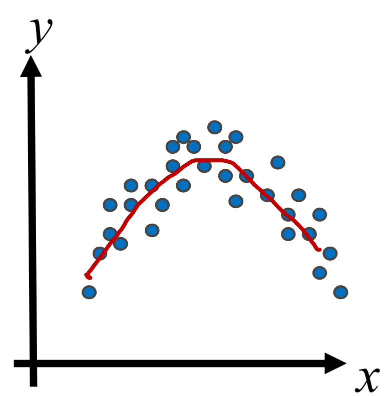


As x increases, y tends to increase.

Positive Linear Correlation



No Correlation



Nonlinear Correlation

Correlation coefficient

- A measure of the strength and the direction of a linear relationship between two variables.
- The symbol r represents the sample correlation coefficient.
- A formula for r is

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

- This is also called the Pearson correlation coefficient.
- The population correlation coefficient is represented by ρ (rho).

Correlation coefficient

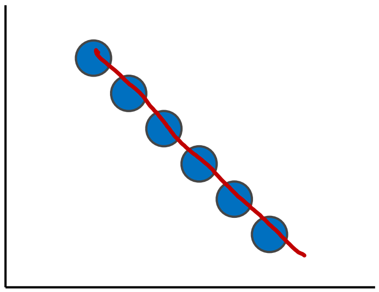
- The range of the correlation coefficient is -1 to 1.

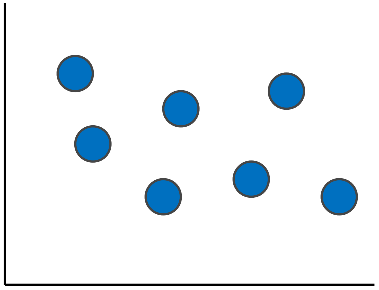
-1 | 0 | 1

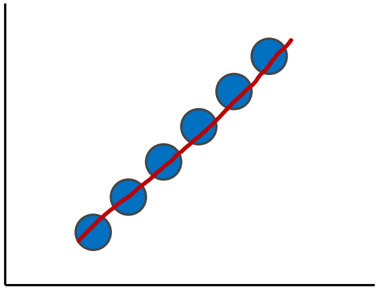
If $r = -1$ there is a perfect negative correlation

If r is close to 0 there is no linear correlation

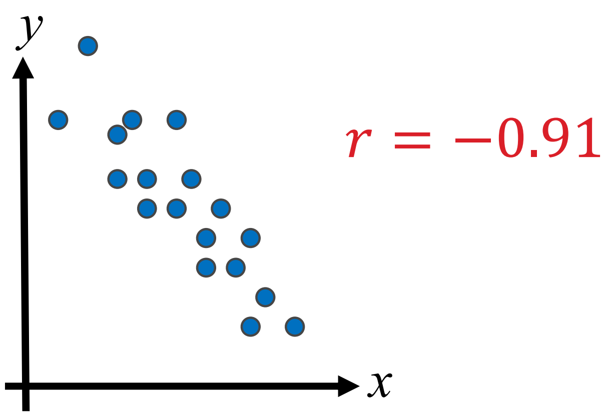
If $r = 1$ there is a perfect positive correlation



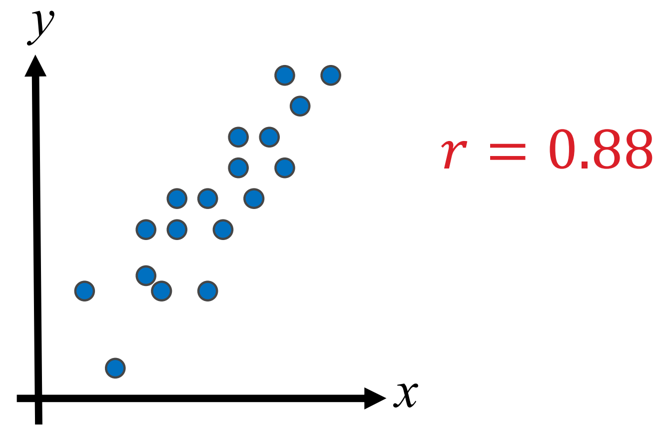




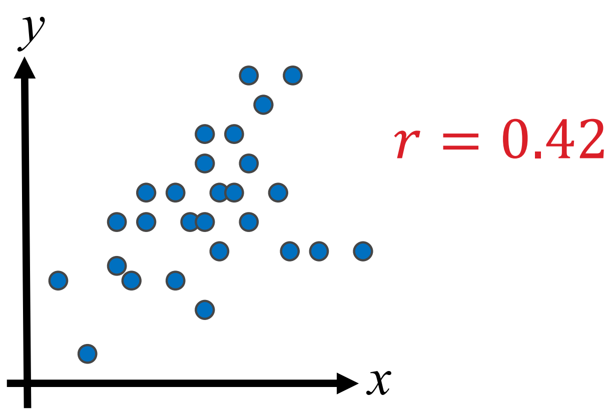
Correlation coefficient



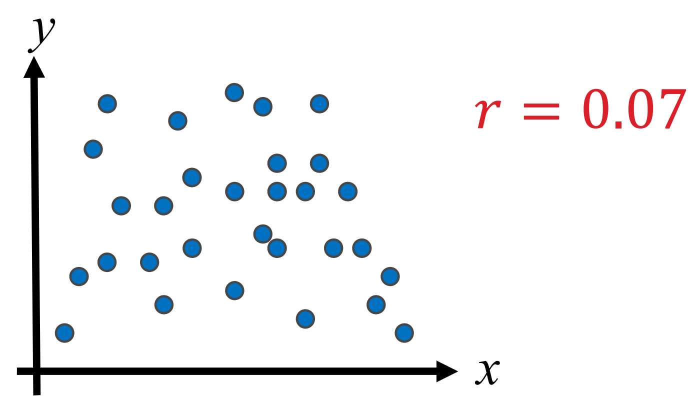
Strong negative correlation



Strong positive correlation



Weak positive correlation



Weak Correlation

UNIVERSITI KEBANGSAAN MALAYSIA
The National University of Malaysia

Introduction to regression analysis

Introduction to regression analysis

- Regression analysis is a statistical technique that is useful for studying relationship between variables.
- For examples:
  - Relationship between expenses and monthly income of one family.
  - Relationship between air pollution rates and the number of vehicles on the road.
  - Relationship between age and salary.
- In regression analysis, we study the relationship between 2 or more variables and predict the value of the variable of interests.

What about linear regression?

- The most basic type of regression, is the linear regression.

- For linear regression, we assume that there is an underlying linear relationship between the dependent variable y and the independent variable x.

Basic steps in regression analysis

1. Plot the variables and look at the relationship between them.
2. Is it linear/non-linear relationship?
3. Predict the parameters in the model.
4. Check for the suitability of the model build based on the data collected. Should the model be modified or accepted?
5. Prediction from the model

- Note: In Simple Linear Regression, only two variables are considered, y and x. In Multiple Linear Regression, there are more than one explanatory variables.

Simple linear regression

Empirical model

- The equation of straight line relating two variables is

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

- This equation represents the exact relation between x and y; where each point of (x, y) should lie on the straight line.
- Since data points do not fall exactly on a straight line, due to the error in the data we collected, so we need to modify the equation by adding the error term in the model.

Empirical model (cont)

- The simple linear regression model is given by

epsilon

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

for an observation i with response variable y_i and explanatory variable x_i, where ε_i is the random error term.

- Only error on y is considered. The error on x can be ignored since x is treated as fixed/control/known.

- ε_i is a random variable.

Assumptions

1. ε_i has zero mean.
2. ε_i and ε_j are not correlated or independent with each other.
3. Variance of ε_i is constant and the same for all observations.
4. ε_i is normally distributed.

- In short,

$$\varepsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2); \qquad i = 1,2, \dots, n$$

Example

Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

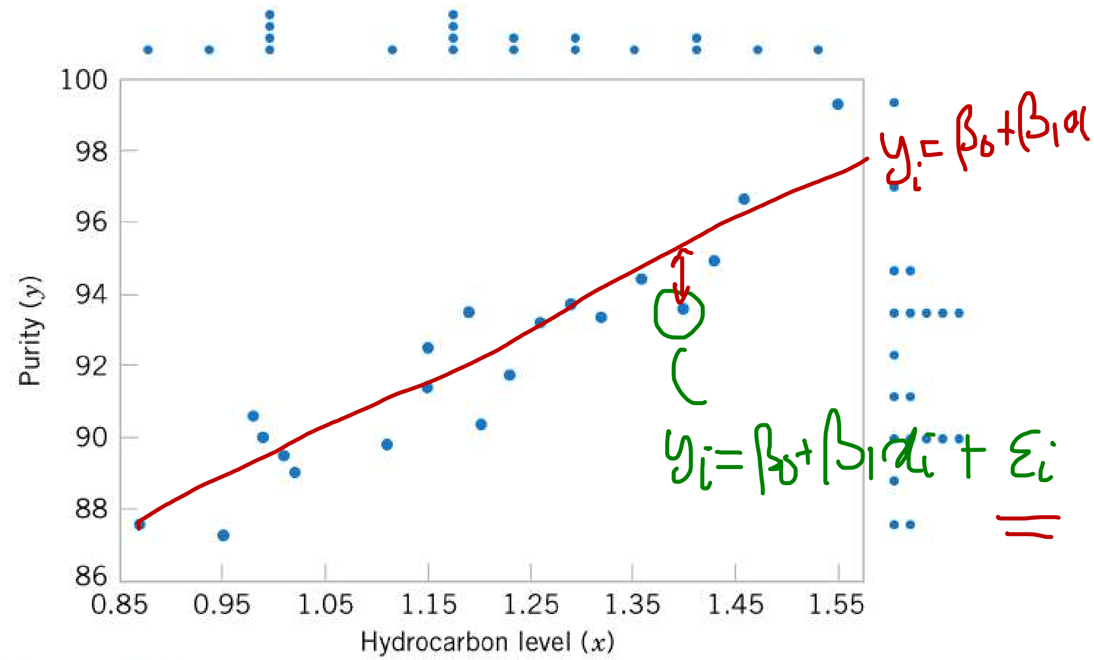


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

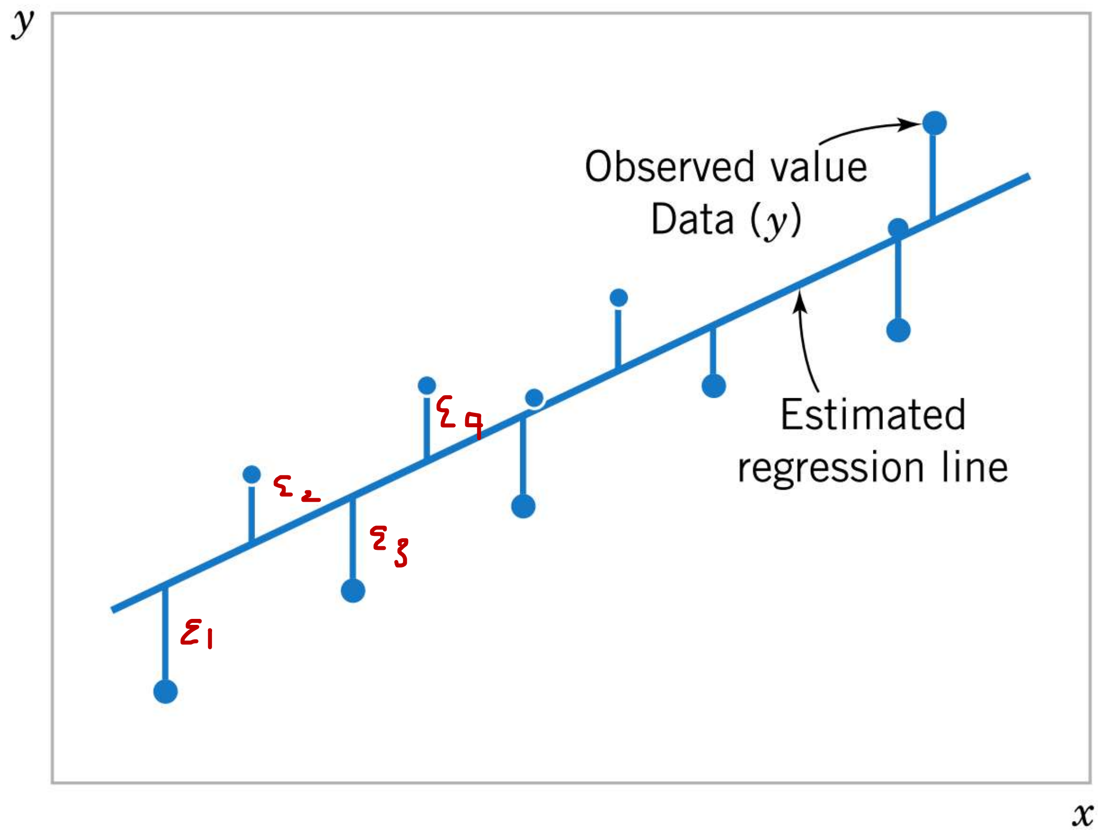
- The plot suggests a linear relationship between hydrocarbon level (x) and purity (y).

Ordinary least squares

- The next step in regression analysis is to estimate the parameters based on the observed data.

- In SLR, we want to get the best straight line which all/most points lie on this line.

- The method of ordinary least squares is used to estimate the parameters, β_0 and β_1 by minimizing the sum of the squares of the vertical deviations between the points y_i and the straight line.



- The sum of the squares of the deviations of the observations from the true regression line is

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

Ordinary least squares (cont)

- The least squares point estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$ can be estimated by

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$S_{xx} = \sum(x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

Ordinary least squares (cont)

- The fitted or estimated regression line is therefore

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

- Each pair of observations satisfies the relationship

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i; \quad i = 1, \dots, n$$

where $e_i = y_i - \hat{y}_i$ is called the residual.

Example (oxygen purity)

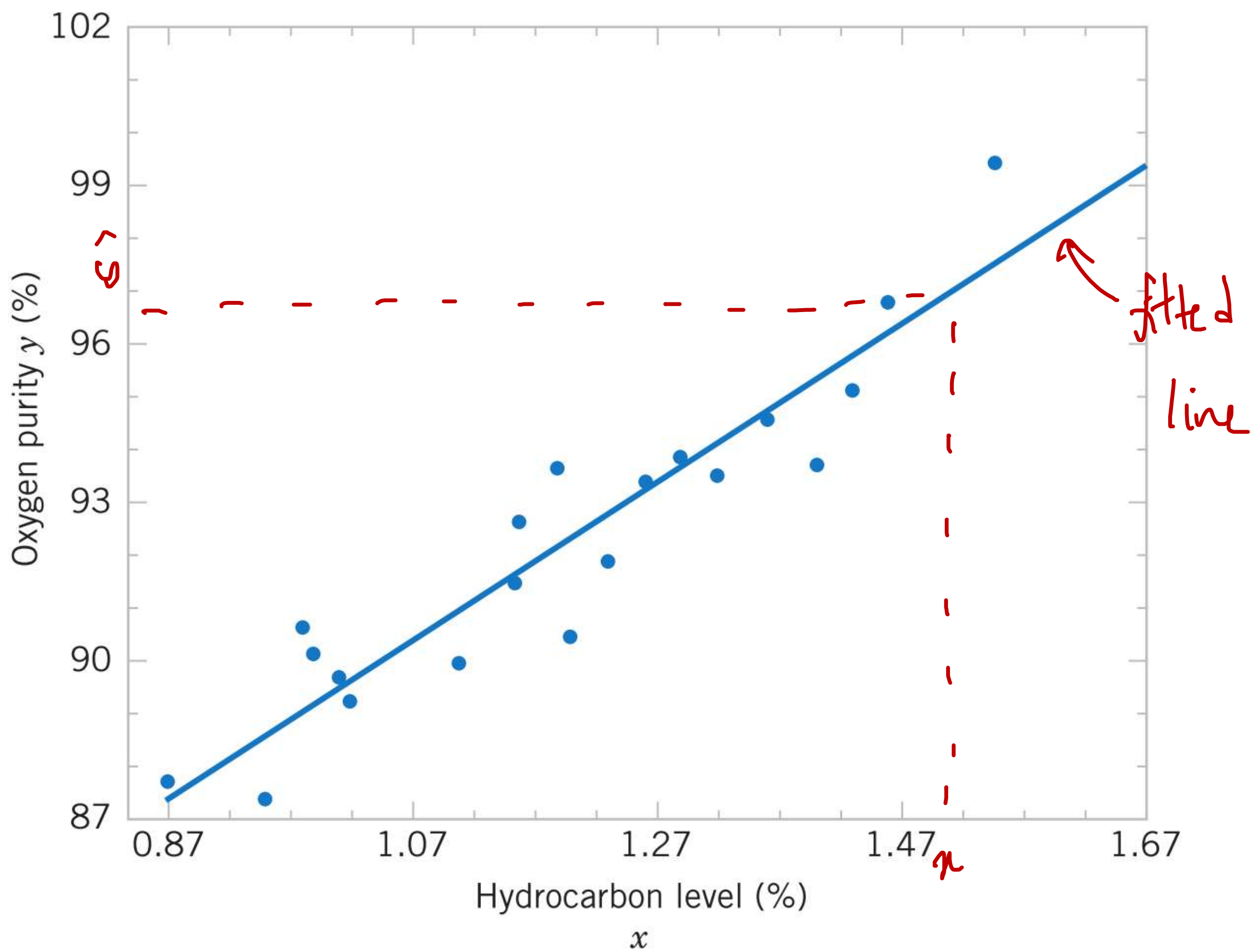


Figure 11-4 Scatter plot of oxygen purity y versus hydrocarbon level x and regression model $\hat{y} = 74.283 + 14.947x$.

$\hat{\beta}_1 = 14.947 > 0$

positive linear relationship.

Multiple linear regression

Multiple linear regression

- The dependent variable or response Y may be related to k independent or regressor variables.
- The model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \\ &= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \end{aligned}$$

  for $i = 1, \dots, n$ and $j = 1, \dots, k$ is called a multiple linear regression with k regressor or independent variables.
- The parameters β_j are called regression coefficients.
- The parameter β_j represents the expected change in response Y per unit change in x_j when all remaining regressor variables are held constant.
- Assumptions used are similar to SLR ($\varepsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$)

Matrix approach

- Suppose the model relating the regressors to the response is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \qquad i = 1, \dots, n$$

- In matrix notation, it can be written as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

Least squares estimates

- We wish to find the vector of least squares estimators that minimizes

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = (\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta})$$

- The resulting least squares estimator is the solution for $\boldsymbol{\beta}$ in the equation

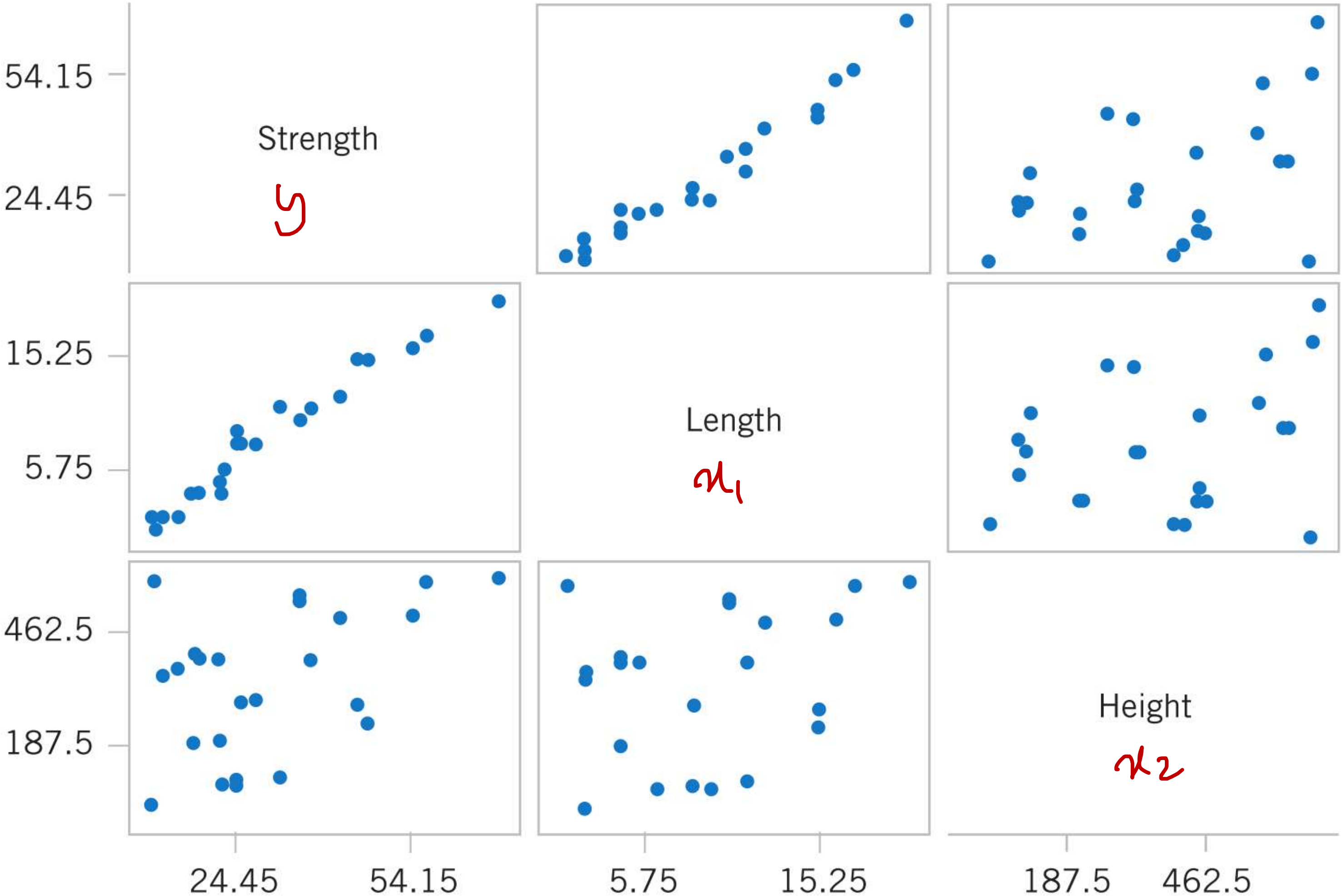
$$\frac{\partial L}{\partial \boldsymbol{\beta}} = 0 \Rightarrow \boxed{\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}}$$

Example (wire pull strength)

- 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.

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

Example (wire pull strength)



Plot indicates strong linear relationship between strength and wire length.

Figure 12-4 Matrix of scatter plots (from Minitab) for the wire bond pull strength data in Table 12-2.

Example (wire pull strength)

- The model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

y: wire pull strength

x_1: wire length

x_2: die height

ε: random error term

- The estimated parameters:

$$\hat{\beta}_0 = 2.26379, \qquad \hat{\beta}_1 = 2.74427, \qquad \hat{\beta}_2 = 0.01253$$

- Therefore, the fitted regression is

$$\hat{y} = 2.26379 + 2.74427x_1 + 0.01253x_2$$

Summary

- Correlation

- Simple linear regression
  - Model and assumption
  - Ordinary least squares method

- Multiple linear regression
  - Model and assumption
  - Parameter estimation using matrix approach