Lecture 11

SIMPLE LINEAR REGRESSION & CORRELATION

FPT Education

FPT UNIVERSITY

Department of Mathematics

Võ Văn Nam

Contents

1. Simple Linear Regression

Regression analysis

- **Regression analysis** is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable.
  - Explain the impact of changes in an independent variable on the dependent variable.
- Dependent variable Y: the variable we wish to predict or explain.
- Independent variable X: the variable used to predict or explain the dependent variable.
- A scatter plot can be used to:
  - Visualize the relationship between X and Y variables.
  - Help suggest a starting point for regression analysis.

FPT Education
FPT UNIVERSITY

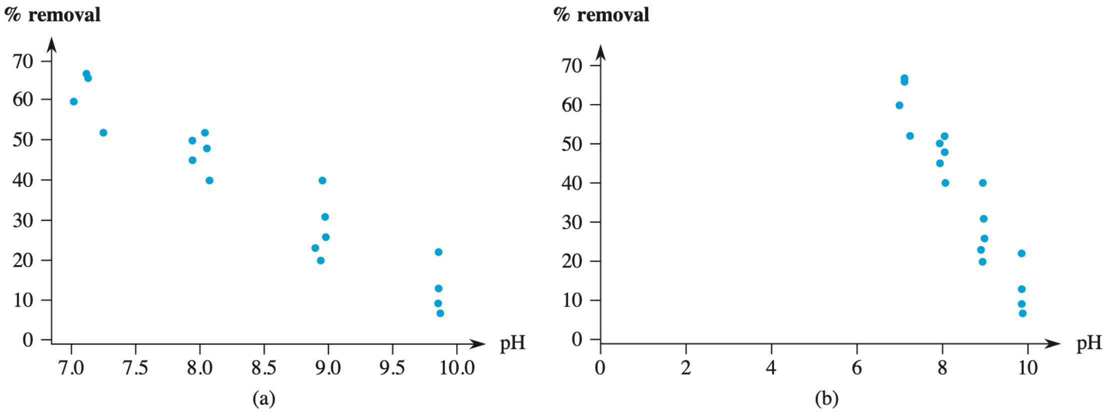
Example

Arsenic is found in many ground-waters and some surface waters. Recent health effects research has prompted the Environmental Protection Agency to reduce allow- able arsenic levels in drinking water so that many water systems are no longer com- pliant with standards. This has spurred interest in the development of methods to remove arsenic. The accompanying data on $x =$ pH and $y =$ arsenic removed (%) by a particular process was read from a scatter plot in the article Optimizing Arsenic Removal During Iron Removal: Theoretical and Practical Considerations (*J. of Water Supply Res. and Tech.*, 2005: 545560).

x	7.01	7.11	7.12	7.24	7.94	7.94	8.04	8.05	8.07
y	60	67	66	52	50	45	52	48	40

x	8.90	8.94	8.95	8.97	8.98	9.85	9.86	9.86	9.87
y	23	20	40	31	26	9	22	13	7

Example (cont')



Minitab scatter plots of data

Simple Linear Regression Model

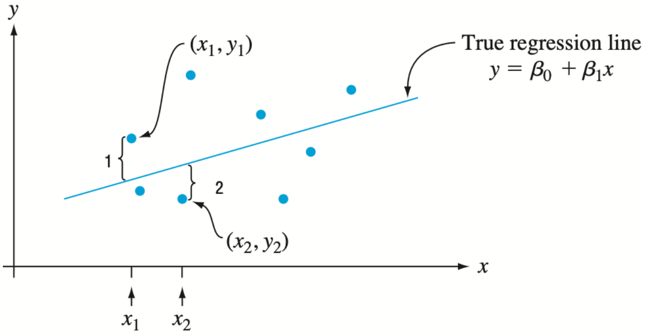
There are parameters β_0, β_1, and σ^2, such that for any fixed value of the independent variable x, the dependent variable is a random variable related to x through the model equation

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

in which $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is the random error of the model.

Simple Linear Regression Model (cont')

Without ε, any observed pair (x, y) would correspond to a point falling exactly on the line $y = \beta_0 + \beta_1 x$, called the true (or population) regression line. The inclusion of the random error term allows (x, y) to fall either above the true regression line (when $\varepsilon > 0$) or below the line (when $\varepsilon < 0$). The points (x_1, y_1), ..., (x_n, y_n) resulting from n independent observations will then be scattered about the true regression line.



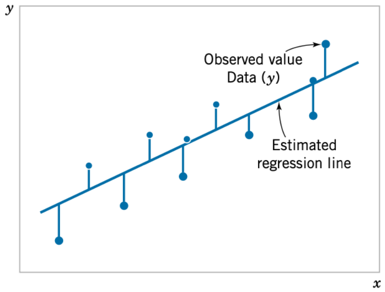
Simple Linear Regression

- Sample contains n data points $(x_i, y_i), i = 1, 2..., n$.
- The point estimates for $\beta_0, \beta_1, \sigma^2$ are denoted by $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$.
- Estimated regression equation (best-fit line) is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Question. How to find point estimates for $\beta_0, \beta_1, \sigma^2$ from samples?
To estimate the regression coefficients, we use Least Squares method, it means

$$\min SS_E = \sum_{i=1}^{n} \varepsilon_i^2$$

where residual $\varepsilon_i = y_i - \hat{y}_i$.



Estimated regression line

Theorem (Best-fit line)

The point estimates of β_0, β_1, say $\hat{\beta}_0, \hat{\beta}_1$ are:

$$\textit{Intercept: } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \textit{Slope: } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$

in which

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

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

Example

A mail-order firm is interested in estimating the number of order that need to be processed on a given day from the weight of the mail received. A close monitoring of mail on 4 randomly selected business days produced the results below.
Find the equation of the least squares regression line relating the number of orders to the weight of the mail and use this equation to predict the number of orders when $x = 25$.

Mails (x)	10	12	13	17
Orders (y)	8	11	12	15

FPT Education
FPT UNIVERSITY

Example (cont')

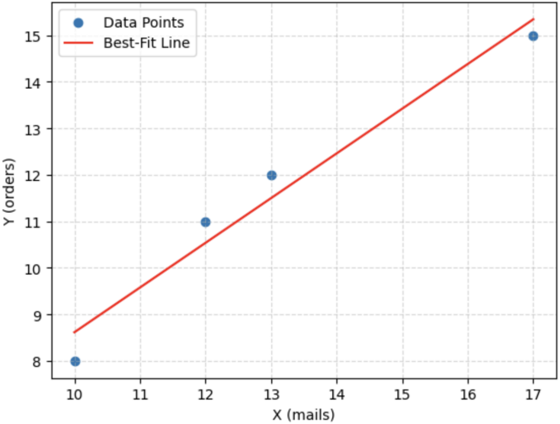


Figure: Best-fit line $\hat{y} = -1 + 0.9615x$

Standard error of estimate

- Total sum of squares

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

- Regression sum of squares

$$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{xy}$$

- Error sum of squares

$$SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SS_T - SS_R$$

- An unbiased estimator of σ^2

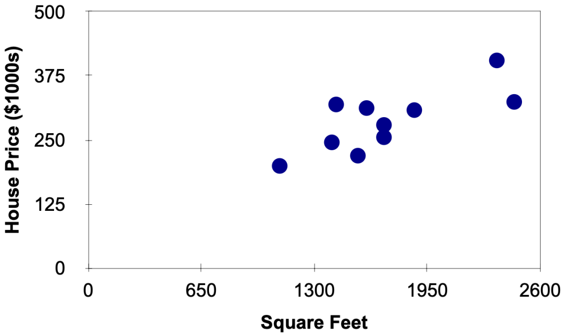
$$\hat{\sigma}^2 = \frac{SS_E}{n-2}$$

Example. Find error sum of squares and the estimate of the variance of the random error in the previous example.

FPT Education
FPT UNIVERSITY

Use Regression in Excel

The following data was determined for 10 randomly selected houses.
Find the estimated regression line and error sum of squares.



House Price in $1000s	Square Feet (X)
245	1,400
312	1,600
279	1,700
308	1,875
199	1,100
219	1,550
405	2,350
324	2,450
319	1,425
255	1,700

Use Regression in Excel (cont')

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is

House Price = 98.24833 + 0.10977 * Square Feet

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

2. Hypothesis Test in Simple Linear Regression

Estimated standard error of the Slope and Intercept

- Estimated of regression slope β_1 is $\hat{\beta}_1$
- Estimated of regression intercept β_0 is $\hat{\beta}_0$
- Estimated standard error of the slope is $se(\hat{\beta}_1) = \sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}$
- Estimated standard error of the intercept is $se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right)}$
- We use t-test with degree of freedom $df = n - 2$ to test for

$$H_0 : \beta_i = \beta_{i,0}, \ i = 1, 2$$

FPT Education
FPT UNIVERSITY

Test hypothesis about the Slope and Intercept

	Test on slope	Test on y-intercept
Null Hypothesis	$H_0 : \beta_1 = \beta_{1,0}$	$H_0 : \beta_0 = \beta_{0,0}$
Alternative Hypothesis	$H_1 : \beta_1 \neq \beta_{1,0}$	$H_1 : \beta_0 \neq \beta_{0,0}$
Test statistic	$t_0 = \dfrac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$	$t_0 = \dfrac{\hat{\beta}_0 - \beta_{0,0}}{se(\hat{\beta}_0)}$
Reject H_0	$\lvert t_0 \rvert > t_{\alpha/2,n-2}$	$\lvert t_0 \rvert > t_{\alpha/2,n-2}$

Example. (continue the previous example) At significance level $\alpha = 0.05$,

- Test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$,
- Test $H_0 : \beta_0 = 100$ vs $H_1 : \beta_0 \neq 100$.

FPT Education
FPT UNIVERSITY

Test for significance of regression

- If $\beta_1 = 0$ then X is NOT significant in explaining the values of Y. We say that the (linear) regression is not significant.
- To formally test the significance of the regression, we can use t-test for

$$H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0.$$

  - If we reject $H_0 : \beta_1 = 0$, we support $H_1 : \beta_1 \neq 0$, and then the regression is **significant**.
  - If we fail to reject $H_0 : \beta_1 = 0$, the regression is **not significant**.

3. Correlation

Correlation coefficient

The **correlation coefficient** is a statistical measure that quantifies the strength and direction of a linear relationship between two variables X and Y. It is denoted by the symbol ρ and takes values between -1 and 1.

Properties of ρ

- $\rho \sim 1$ then there is a **strong positive** linear regression.
- $\rho \sim -1$ then there is a **strong negative**linear regression.
- $\rho \sim 0$ then linear relation between X and Y is weak.

Sample correlation coefficient R

- The sample correlation coefficient, denoted as r, is a statistic that measures the strength and direction of a linear relationship between two variables in a sample. It is an estimate of the population correlation coefficient (ρ).
- The sample correlation coefficient is calculated using the following formula:

$$R = \frac{\sum (x_i - \bar{x}) y_i}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} SS_T}}.$$

- It is used to assess the strength and direction of the linear relationship within the observed data.
- R and β_1 have same sign.

Coefficient of determination R^2

The **coefficient of determination**, denoted as R^2, is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model.
The formula for R^2 is:

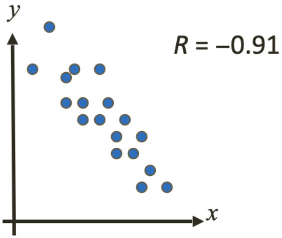
$$R^2 = 1 - \frac{SS_E}{SS_T} = \frac{SS_R}{SS_T}.$$

The interpretation of R^2 is as follows:

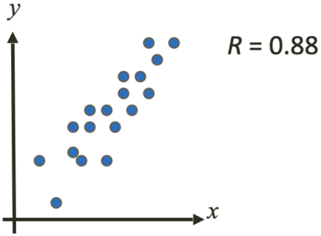
- $R^2 = 0$ indicates that the model does not explain any variability in the dependent variable.
- $R^2 = 1$ indicates that the model perfectly explains the variability in the dependent variable.
- $0 < R^2 < 1$ indicates the proportion of variability explained by the model.

FPT Education
FPT UNIVERSITY

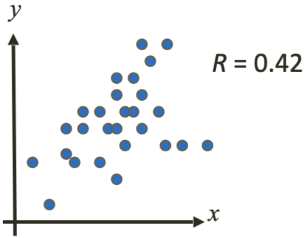
Correlation and R



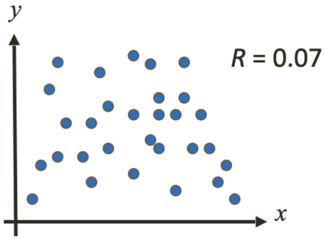
Strong negative correlation



Strong positive correlation



Weak positive correlation



Nonlinear Correlation

Examples

Example 1. In a regression problem the following pairs of (x, y) are given $(-4; 8); (-1; 3); (0; 0); (1; -3)$.
What does this indicate about the value of coefficient of correlation and coefficient of determination?

Example 2. The least squares regression line is $\hat{y} = -2.87 - 1.6x$ and a coefficient of determination of 0.36.
What is the coefficient of correlation?

FPT Education
FPT UNIVERSITY

Hypothesis Test for Zero Correlation

- Test hypothesis

$$H_0 : \rho = 0$$

- Test statistic

$$T_0 = \frac{R\sqrt{n-2}}{1-R^2}$$

has a t-distribution with $n-2$ degrees of freedom if H_0 is True.

Alternative Hypothesis	Critical Values	Reject H_0
$H_1 : \rho \neq 0$	$t_{\alpha/2,n-2}, \ -t_{\alpha/2,n-2}$	$\|T_0\| > t_{\alpha/2,n-2}$
$H_1 : \rho > 0$	$t_{\alpha/2,n-2}$	$T_0 > t_{\alpha/2,n-2}$
$H_1 : \rho < 0$	$-t_{\alpha/2,n-2}$	$T_0 < -t_{\alpha/2,n-2}$

Example

You want to explore the relationship between the grades students receive on their first two exams. For a sample of 25 students, you find a correlation of 0.45.
What is your conclusion in testing $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$ at significant level $\alpha = 0.05$?

Python Codes: Calculate R

In Python, you can calculate the sample correlation coefficient using libraries such as NumPy or pandas. Here's an example using NumPy:

```python
import numpy as np

# Example data
X = np.array([1, 2, 3, 4, 5])
Y = np.array([2, 3, 5, 4, 6])

# Calculate sample correlation coefficient
r = np.corrcoef(X, Y)[0, 1]

print(f"The sample correlation coefficient (r) is: {r:.4f}")
```

Python Codes: Calculate R^2

In Python, you can calculate the coefficient of determination using libraries such as scikit-learn or statsmodels. Here's an example using scikit-learn:

```python
from sklearn.metrics import r2_score

# Example data
observed_values = [2, 4, 5, 4, 5]
predicted_values = [1.8, 4.2, 4.7, 3.9, 5.2]

# Calculate R^2
r_squared = r2_score(observed_values, predicted_values)

print(f"The coefficient of determination (R^2) is: {r_squared:.4f}")
```

Thank you!

namvv14@fe.edu.vn