Statistics & Probability

Chapter 11: SIMPLE LINEAR REGRESSION AND CORRELATION

FPT University

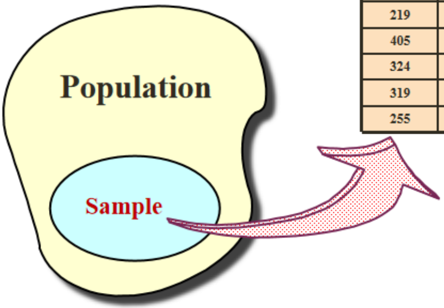
Department of Mathematics

Quy Nhon, 2023

Table of Contents

Random Sample

House Price in $1000s (Y)	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



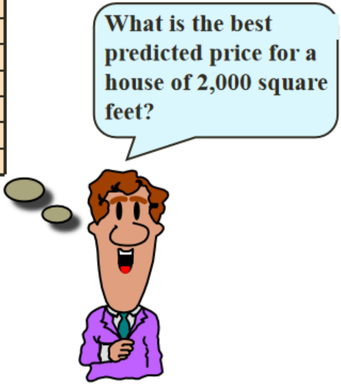


Table of Contents

Introduction

1. **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.

2. **Dependent variable** Y: the variable we wish to predict or explain.

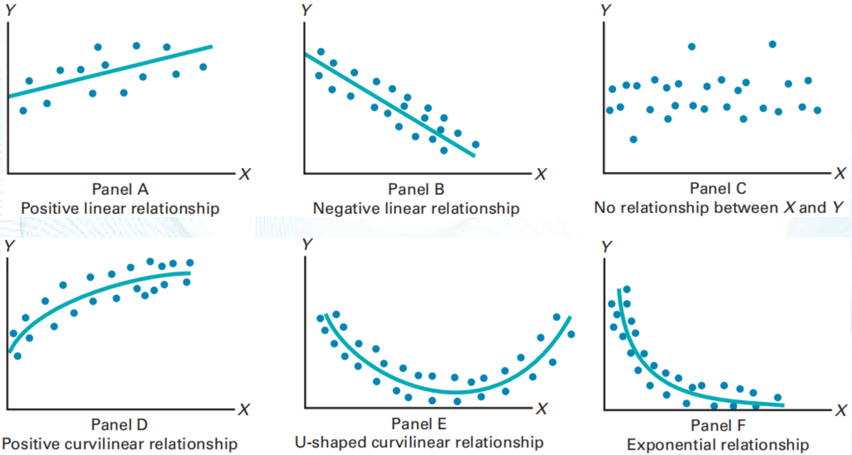
3. **Independent variable** X: the variable used to predict or explain the dependent variable.

4. A **scatter plot** can be used to:

   + Visualize the relationship between X and Y variables.

   + Help suggest a starting point for regression analysis.

Six types of relationships found in scatter plots:



Panel A
Positive linear relationship

Panel B
Negative linear relationship

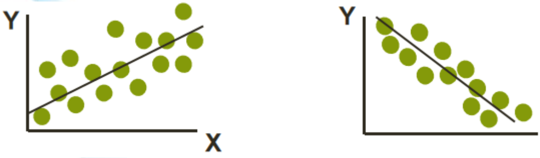
Panel C
No relationship between X and Y

Panel D
Positive curvilinear relationship

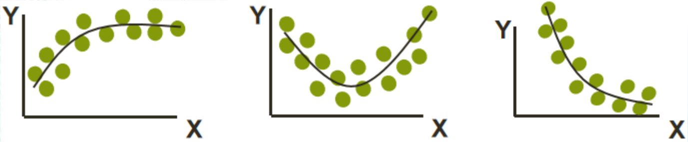
Panel E
U-shaped curvilinear relationship

Panel F
Exponential relationship

1. **Linear relationships**:



2. **Curvilinear relationships**:



3. **No relationships**:

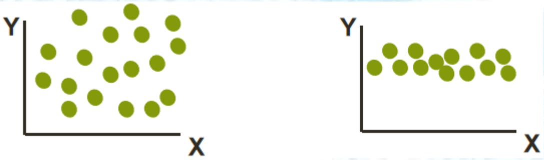
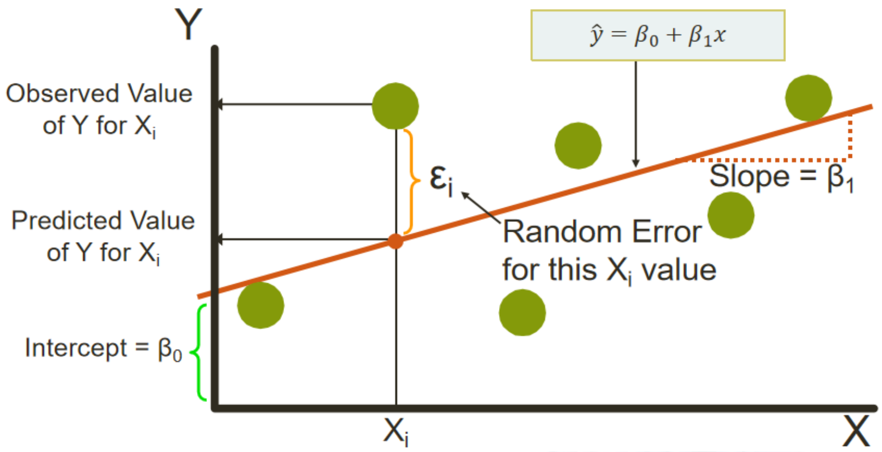


Table of Contents

Simple linear regression model

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

where $\varepsilon \sim N(0, \sigma^2)$ is the random error of the model.



Sample contain n data points (x_i, y_i) where $i = 1, 2, \ldots, n$.

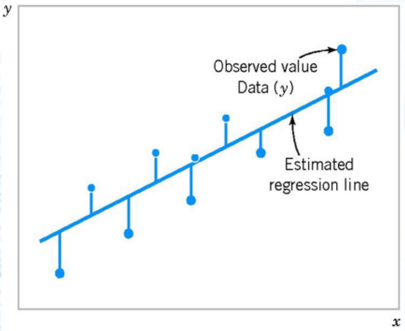
- The **point estimates** for $\beta_0, \beta_1, \sigma^2$ are denoted by $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$.
- **Estimated regression equation** (best-fit line) is

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

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

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

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



Estimated regression equation (best-fit line) is

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

Best-Fit Line

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

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

where

- $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$
- $S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}.$

Recall

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

where

- $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$
- $S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$.

Example

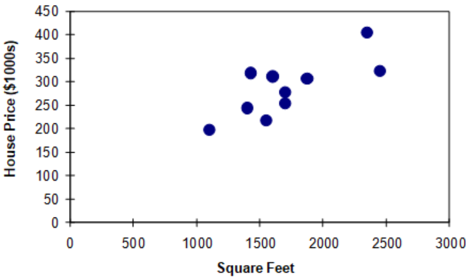
A mail-order firm is interested in estimating the number of order that need to be processed on a given day from the weight of the mail received. A close monitoring of mail on 4 randomly selected business days produced the results below. Find the equation of the least squares regression line relating the number of orders to the weight of the mail and use this equation to predict the number of orders when $x = 15$.

Mail (x)	10	12	13	17
Orders (y)	8	10	6	10

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 (Y)	**Square feet** (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700



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

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

The regression equation is

$$\widehat{(\text{house price})} = 98.24833 + 0.10977 \cdot (\text{square feet}).$$

Standard Error of Estimate

Question: How well the model describes the data?

- **Total sum of squares** is

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

- **Regression sum of squares** is

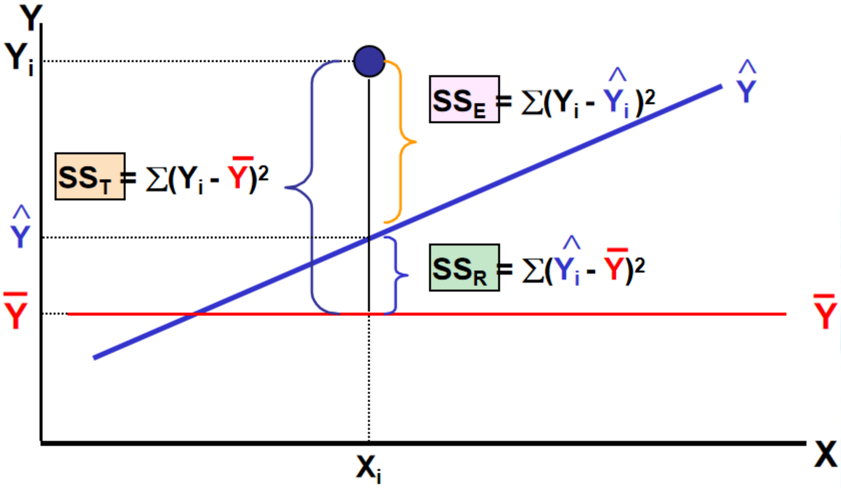
$$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = \hat{\beta}_1 S_{xy}.$$

- **Error sum of squares** is

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

- An **unbiased estimator** of σ^2 is

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



Y
Y_i
$\hat{Y}$
$\bar{Y}$
$SS_E = \sum(Y_i - \hat{Y}_i)^2$
$\hat{Y}$
$SS_T = \sum(Y_i - \bar{Y})^2$
$SS_R = \sum(\hat{Y}_i - \bar{Y})^2$
$\bar{Y}$
X_i
X

Recall

- **Total sum of squares** is $SS_T = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}.$
- **Regression sum of squares** is $SS_R = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = \hat{\beta}_1 S_{xy}.$
- **Error sum of squares** is $SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SS_T - SS_R.$
- An **unbiased estimator** of σ^2 is $\hat{\sigma}^2 = \dfrac{SS_E}{n-2}.$

Quiz

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 error sum of squares and the estimate of the variance of the random error.

Mail (x)	10	12	13	17
Orders (y)	8	10	6	10

Table of Contents

Estimated Standard Errors

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

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

respectively, where $\hat{\sigma}^2 = \dfrac{SS_E}{n-2}$.

Table of Contents

FUQN

Remark

- 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

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}.$$

- Estimated standard error of the intercept is

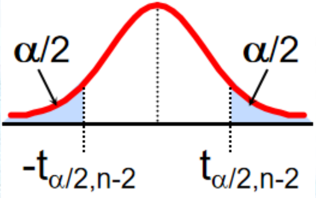
$$se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}.$$

- We use t-test with degree of freedom $df = n - 2$ to test for

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

where $i = 0, 1$.

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



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.

Mail (x)	10	12	13	17
Orders (y)	8	10	6	10

At level of significance $\alpha = 0.05$.

1. Test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.
2. Test $H_0 : \beta_0 = 100$ versus $H_1 : \beta_0 \neq 100$.

Remarks

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

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0.$$

3. If we reject $H_0 : \beta_1 = 0$, we support $H_1 : \ \beta_1 \neq 0$; then the regression is significant.
4. If we fail to reject $H_0 : \beta_1 = 0$, the regression is not significant.

Table of Contents

To measure the strength of the linear relationship between X and Y we can use the **correlation coefficient** ρ.

Remarks

1. $-1 \leq \rho \leq 1$.
2. If $\rho \sim 1$, there is a strong positive linear regression.
3. If $\rho \sim -1$, there is a strong negative linear regression.
4. If $\rho \sim 0$, linear relation between X and Y is weak.

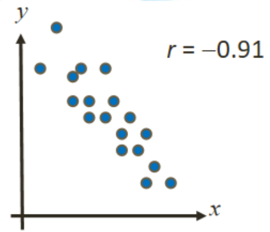
Sample Correlation Coefficient

$$R = \frac{S_{xy}}{\sqrt{S_{xx} SS_T}}.$$

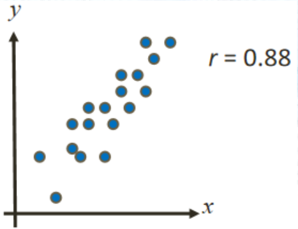
Remarks

1. $-1 \leq R \leq 1$.
2. The **coefficient of determination** $R^2 = \dfrac{SS_R}{SS_T}$ is often used to judge the adequacy of a regression model.
3. R and $\hat{\beta}_1$ have same sign.
4. Both R and R^2 measure the strength of a linear relationship.

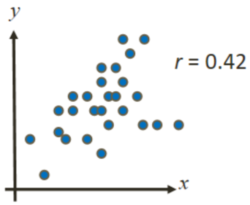
1. Strong negative correlation



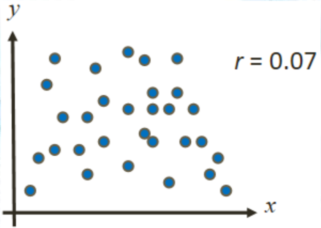
2. Strong positive correlation



3. Weak positive correlation



4. Nonlinear Correlation



FUQN

Sample Correlation Coefficient

$$R = \frac{S_{xy}}{\sqrt{S_{xx} SS_T}}.$$

Quiz 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?

Quiz 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?

Test For Zero Correlation

1. Test hypotheses

$$H_0 : \rho = 0.$$

2. Test statistic

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

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

3. Rejection region

Alternative hypothesis	**Critical value**	**Reject H_0**
$H_1 : \rho \neq 0$	$\pm t_{\alpha/2,n-2}$	$\lvert t_0 \rvert > t_{\alpha/2,n-2}$
$H_1 : \rho > 0$	$t_{\alpha,n-2}$	$t_0 > t_{\alpha,n-2}$
$H_1 : \rho < 0$	$-t_{\alpha,n-2}$	$t_0 < -t_{\alpha,n-2}$

Test statistic

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

Quiz

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

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

at significant level $\alpha = 0.05$.

Quiz*

Suppose you are interested in determining the relationship between the number of absences x and the final grades y of students from a statistic class. For a sample of 10 observations, you have the following information

$$\sum_{i=1}^{10} x_i = 304, \quad \sum_{i=1}^{10} y_i = 345, \quad \sum_{i=1}^{10} x_i y_i = 11312, \quad \sum_{i=1}^{10} x_i^2 = 11030, \quad \sum_{i=1}^{10} y_i^2 = 13547.$$

Find the sample regression line. Compute SS_T, SS_R, SS_E.

Thank you!