Tóm tắt nội dung Chapter 11

Hồi quy và tương quan

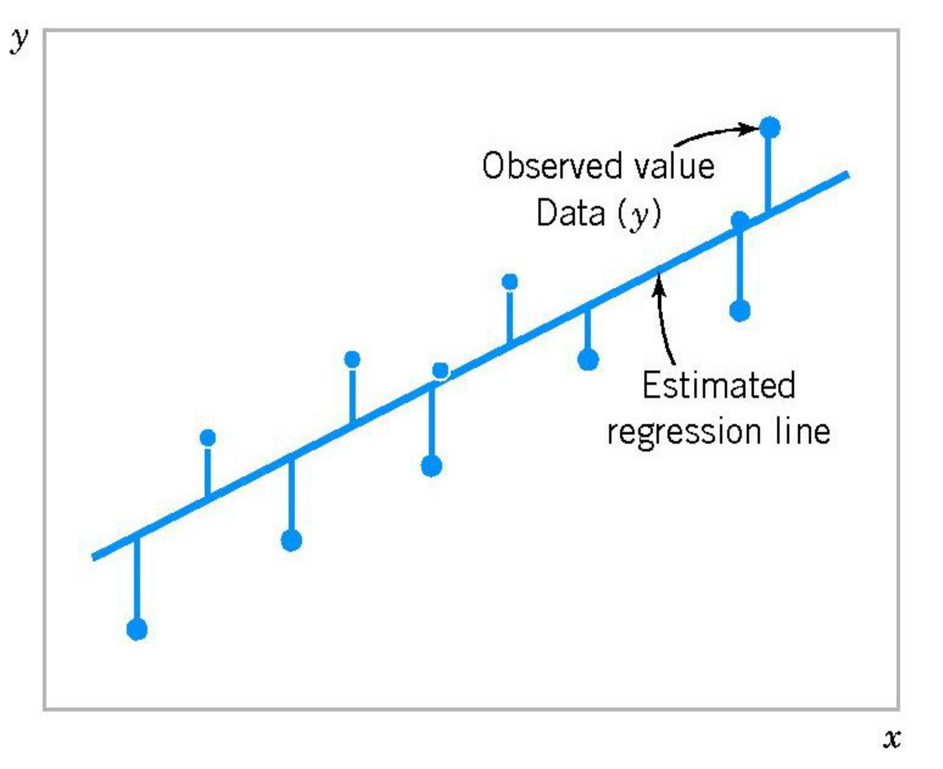
Hồi quy: Regression

Xét n cặp quan sát (x_1, y_1), (x_2, y_2), …, (x_n, y_n):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ , \ i = 1, \ldots, n$$

intercept

slope



Hồi quy: Regression

Theorem

Phương trình hồi quy tuyến tính đơn (**Estimated or fitted regression line**)

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

Trong đó

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

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Hồi quy: Regression

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

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

Các sai số (errors) ε_i

Chúng ta luôn giả sử rằng các sai số là độc lập với nhau và có cùng phân phối chuẩn $N(0, \sigma^2)$

Hồi quy: Regression

Ước lượng điểm cho σ^2

Theorem

An **unbiased estimator** of σ^2 is

$\hat{\sigma}$ Standard error

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

error sum of squares

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

where $SS_E = SS_T - \hat{\beta}_1 S_{xy}$

$$SS_E = SS_T - \hat{\beta}_1 S_{xy} \qquad SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2$$

Sử dụng Excel

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.310670688							
5	R Square	0.096516276							
6	Adjusted R Square	-0.084180468							
7	Standard Error	4.612004796							
8	Observations	7							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	11.36134454	11.36134454	0.534134008	0.497668094			
13	Residual	5	106.3529412	21.27058824					
14	Total	6	117.7142857						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	13.64705882	3.33234126	4.09533651	0.009397596	5.081002915	22.21311473	5.081002915	22.21311473
18	X Variable 1	-0.764705882	1.046331539	-0.730844722	0.497668094	-3.454386728	1.924974964	-3.454386728	1.924974964

$\hat{\sigma}$

$SS_R = SS_T - SS_E$

SS_T

$\hat{\beta}_1$

$\hat{\beta}_0$

Kiểm định giả thiết trong mô hình hồi quy

Test on the β_1

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

Test statistic

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

has the t distribution with n - 2 degrees of freedom.

If $|t_0| > t_{\alpha/2,\, n-2}$: reject H_0

If $|t_0| < t_{\alpha/2,\, n-2}$: fail to reject H_0

Kiểm định giả thiết trong mô hình hồi quy

Test on the β_0

If $|t_0| > t_{\alpha/2,\ n-2}$: reject H_0

If $|t_0| < t_{\alpha/2,\ n-2}$: fail to reject H_0

$$H_0\colon \beta_0 = \beta_{0,0}$$
$$H_1\colon \beta_0 \neq \beta_{0,0}$$

Test statistic

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

Confidence Intervals on the Slope and Intercept

Đoạn tin cậy 100(1-α)% cho β_1 trong mô hình hồi quy là

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

Đoạn tin cậy 100(1-α)% cho β_0 là

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

Hồi quy: Regression

Confidence Interval on the Mean Response

$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

A 100(1-α)% confidence interval about the mean response at the value of $x=x_0$ is given by

$$\hat{\mu}_{Y|x_0} - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$

$$\le \mu_{Y|x_0} \le \hat{\mu}_{Y|x_0} + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$

Hồi quy: Regression

Prediction of New Observations

A 100(1-α)% prediction interval on a future observation Y_0 at the value x_0 is given by

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$

$$\leq Y_0 \leq \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$

Hệ số tương quan: Correlation

Definition

The **sample correlation coefficient**

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX} SS_T}}$$

Note that

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

We may also write:

$$R^2 = \hat{\beta}_1^2 \frac{S_{XX}}{S_{YY}} = \frac{\hat{\beta}_1 S_{XY}}{SS_T} = \frac{SS_R}{SS_T}$$

Hệ số tương quan: Correlation

Test on the ρ

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

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

has the *t* distribution with *n* - 2 degrees of freedom.

If $|t_0| > t_{\alpha/2,\, n-2}$: reject H_0

If $|t_0| < t_{\alpha/2,\, n-2}$: fail to reject H_0