Phân tích hồi quy

11-1 Giới thiệu

11-2 Covariance và Correlation Coefficient

11-3 Ví dụ: dữ liệu sửa chữa máy tính

11-4 Mô hình hồi quy tuyến tính đơn biến

11-5 Việc ước lượng tham số

11-6 Kiểm định giả thuyết

11-7 Khoảng tin cậy

11-8 Dự đoán

11-9 Measuring the quality of fit

Giới thiệu

- Phân tích hồi quy là một phương pháp đơn giản về mặt khái niệm để nghiên cứu các mối quan hệ giữa các biến.
- Mối quan hệ được thể hiện dưới dạng một phương trình hoặc mô hình kết nối biến phản hồi (*response*) hoặc biến phụ thuộc (*dependent*) và một hoặc nhiều biến giải thích (*explanatory*) hoặc dự đoán (*predictor*)
- Mô hình hồi quy

$$Y = f(X_1, X_2, \cdots, X_p) + \varepsilon,$$

Giới thiệu

- Ví dụ về mô hình hồi quy tuyến tính:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon,$$

- Trong đó $\beta_0, \beta_1, \ldots, \beta_p$ là *hệ số tương quan hồi quy của mô hình hay còn gọi là các tham số hồi quy*, các hệ số này là các hằng số, chưa biết, được xác định dựa vào tập dữ liệu.

Các bước trong phân tích hồi quy

- Phân tích hồi quy bao gồm các bước sau:
  - Phát biểu bài toán
  - Xác định các biến phụ thuộc
  - Thu thập dữ liệu
  - Xây dựng và lựa chọn mô hình
  - Sử dụng mô hình để giải quyết bài toán đưa ra.

Covariance và Correlation Coefficient

- Giả sử chúng ta có các quan sát trên **n đối tượng** bao gồm biến phản hồi (**response) *Y*** và biến dự đoán (**predictor) X**. Kết quả ghi lại như sau:

Observation Number	Response Variable Y	Predictor X
1	y_1	x_1
2	y_2	x_2
$\vdots$	$\vdots$	$\vdots$
n	y_n	x_n

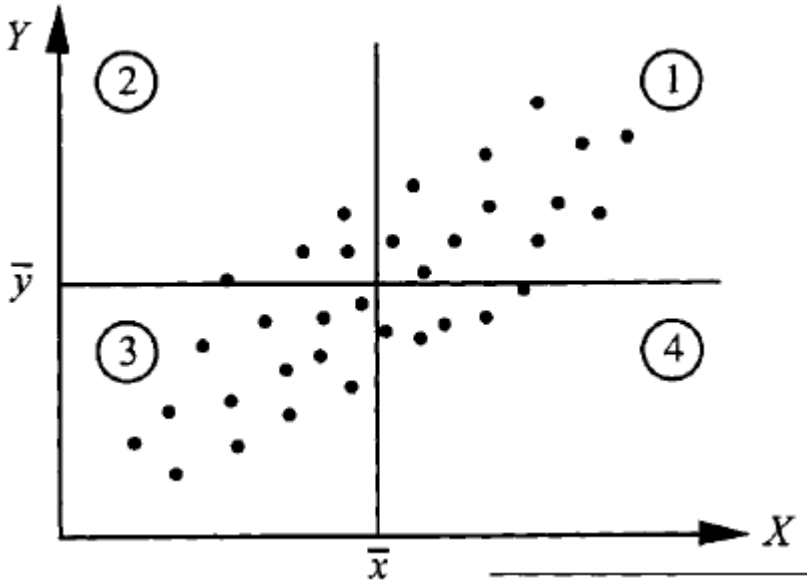
$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad \text{and} \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n},$$

Covariance và Correlation Coefficient

- Chúng ta mong muốn đo **cả hướng** và **độ mạnh** về mối quan hệ giữa Y và X.
- Ta sử dụng 2 đại lượng để đo là ***covariance*** và ***correlation coefficient***.

Covariance và Correlation Coefficient

- X tăng *Y* cũng tăng: các điểm nằm ở vùng thứ 1 và thứ 3.
- X tăng *Y* giảm: các điểm nằm ở vùng thứ 2 và thứ 4.



Quadrant	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})(x_i - \bar{x})$
1	+	+	+
2	+	−	−
3	−	−	+
4	−	+	−

Covariance và Correlation Coefficient

- ***Covariance** giữa Y* và X thể hiện hướng **của mối quan hệ tuyến tính giữa Y và X**.
- Cov(Y, X) **không cho ta biết độ mạnh** của mối quan hệ giữa Y và X
- Cor(Y,X): cho ta biết hướng
- và độ mạnh mối quan hệ giữa
- X và Y

$$\mathrm{Cov}(Y,X) = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{n-1}$$

$$\mathrm{Cor}(Y,X) = \frac{\mathrm{Cov}(Y,X)}{s_y s_x}$$

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

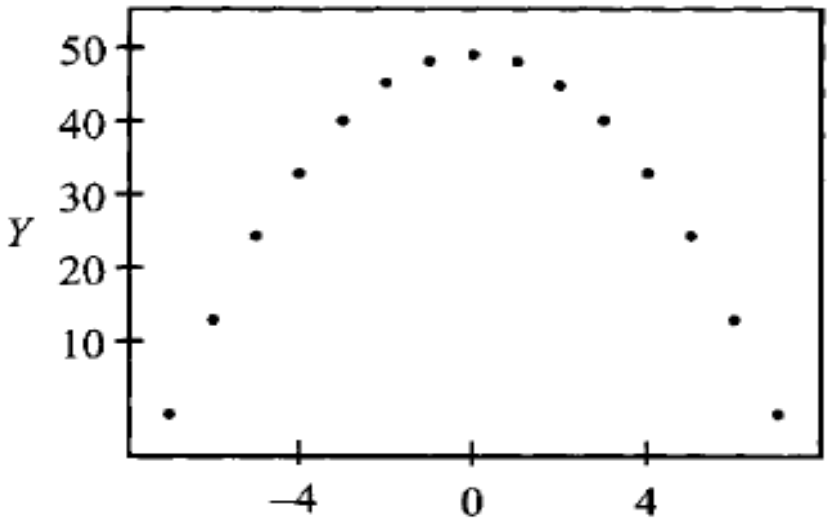
$$-1 \le \mathrm{Cor}(Y,X) \le 1$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$$

Covariance và Correlation Coefficient

Data Set with a Perfect Nonlinear Relationship Between Y and X, Yet $\text{Cor}(X, Y) = 0$

Y	X	Y	X	Y	X
1	−7	46	−2	41	3
14	−6	49	−1	34	4
25	−5	50	0	25	5
34	−4	49	1	14	6
41	−3	46	2	1	7



Ví dụ: dữ liệu sửa chữa máy tính

- ***Ví dụ:*** xem xét trường hợp của một công ty tiếp thị và sửa chữa máy tính nhỏ.
- Để nghiên cứu mối quan hệ giữa thời gian của một cuộc gọi đến dịch vụ và số lượng linh kiện điện tử trong máy tính phải được sửa chữa hoặc thay thế, dữ liệu mẫu của các cuộc gọi đến được lưu lại như sau:

Ví dụ: dữ liệu sửa chữa máy tính

- Dữ liệu bao gồm thời gian của cuộc gọi tính bằng phút (biến response) và số lượng linh kiện sửa chữa (biến predictor)

Length of Service Calls (in Minutes) and Number of Units Repaired

Row	Minutes	Units	Row	Minutes	Units
1	23	1	8	97	6
2	29	2	9	109	7
3	49	3	10	119	8
4	64	4	11	149	9
5	74	4	12	145	9
6	87	5	13	154	10
7	96	6	14	166	10

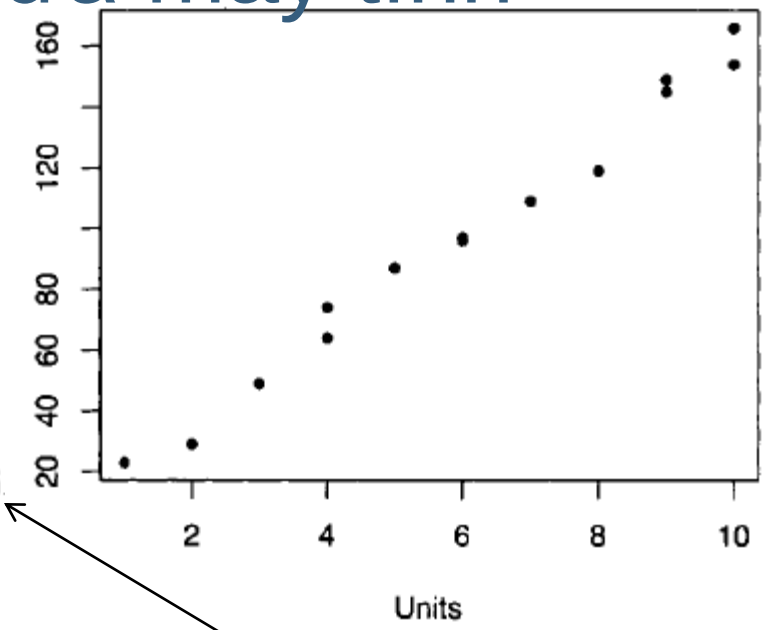
Ví dụ: dữ liệu sửa chữa máy tính

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{1361}{14} = 97.21 \quad \text{and} \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{84}{14} = 6,$$

$$\text{Cov}(Y, X) = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{n-1} = \frac{1768}{13} = 136,$$

and

$$\text{Cor}(Y, X) = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (x_i - \bar{x})^2}} = \frac{1768}{\sqrt{27768.36 \times 114}} = 0.996$$



Length of Service Calls (in Minutes) and Number of Units Repaired

Row	Minutes	Units	Row	Minutes	Units
1	23	1	8	97	6
2	29	2	9	109	7
3	49	3	10	119	8
4	64	4	11	149	9
5	74	4	12	145	9
6	87	5	13	154	10
7	96	6	14	166	10

There is a strong positive relationship between repair time and units repaired

Ví dụ: dữ liệu sửa chữa máy tính

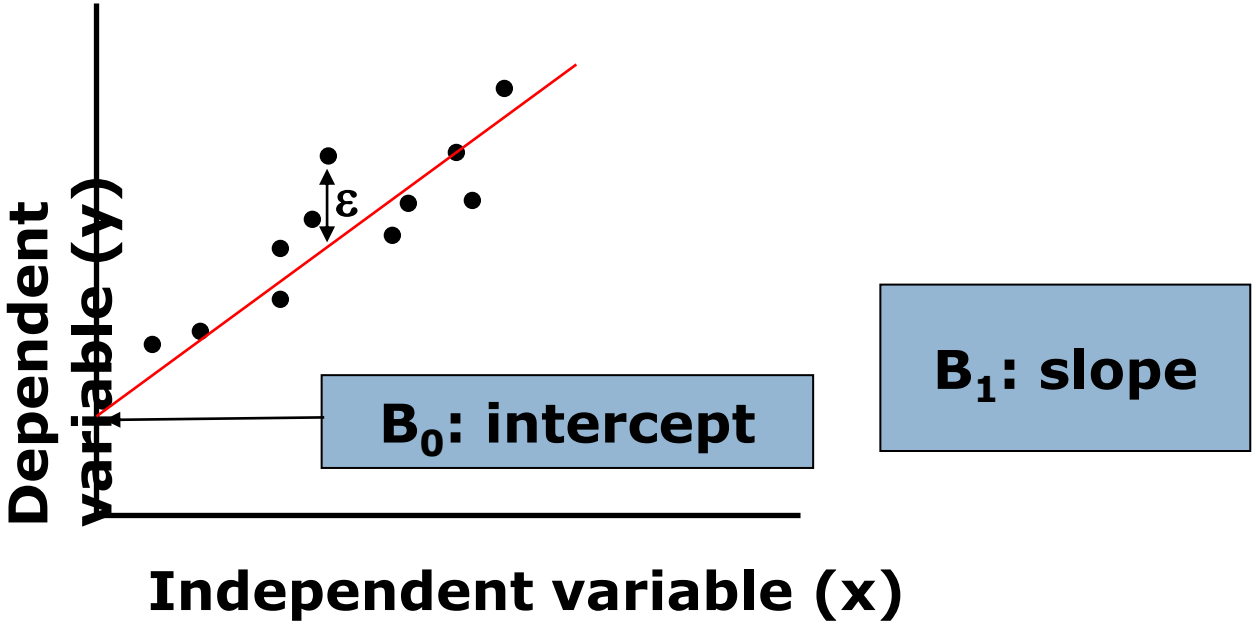
- Mặc dù Cor(Y, *X)* là đại lượng hữu ích để đo hướng và độ mạnh của mối quan hệ tuyến tính, nó không thể được sử dụng cho mục đích dự đoán.
- → **Phân tích hồi quy:** có thể được sử dụng không chỉ để đo hướng và độ mạnh của mối quan hệ giữa các biến phản hồi và biến dự đoán mà còn để mô tả mối quan hệ đó về mặt dự đoán dạng số.

Hồi quy tuyến tính đơn biến

- Mối quan hệ giữa biến phản hồi Y và biến dự đoán X được quy định bởi mô hình tuyến tính:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Trong đó β_0 và β_1*: hệ số tương quan hồi quy của mô hình hay còn gọi là các tham số.*
- ε: random disturbance or error



Hồi quy tuyến tính đơn biến

- Dữ liệu quan sát được lưu ở bảng dưới đây:

Observation Number	Response Variable Y	Predictor X
1	y_1	x_1
2	y_2	x_2
$\vdots$	$\vdots$	$\vdots$
n	y_n	x_n

Có thể được viết như sau:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \cdots, n,$$

ε_i represents the error in the approximation of y_i.

Hồi quy tuyến tính đơn biến

- Quay trở lại dữ liệu sửa chữa máy tính, giả sử công ty muốn dự đoán số phút gọi vào tổng đài dịch vụ của khách dựa vào số linh kiện cần sửa chữa, mô hình hồi quy như sau:

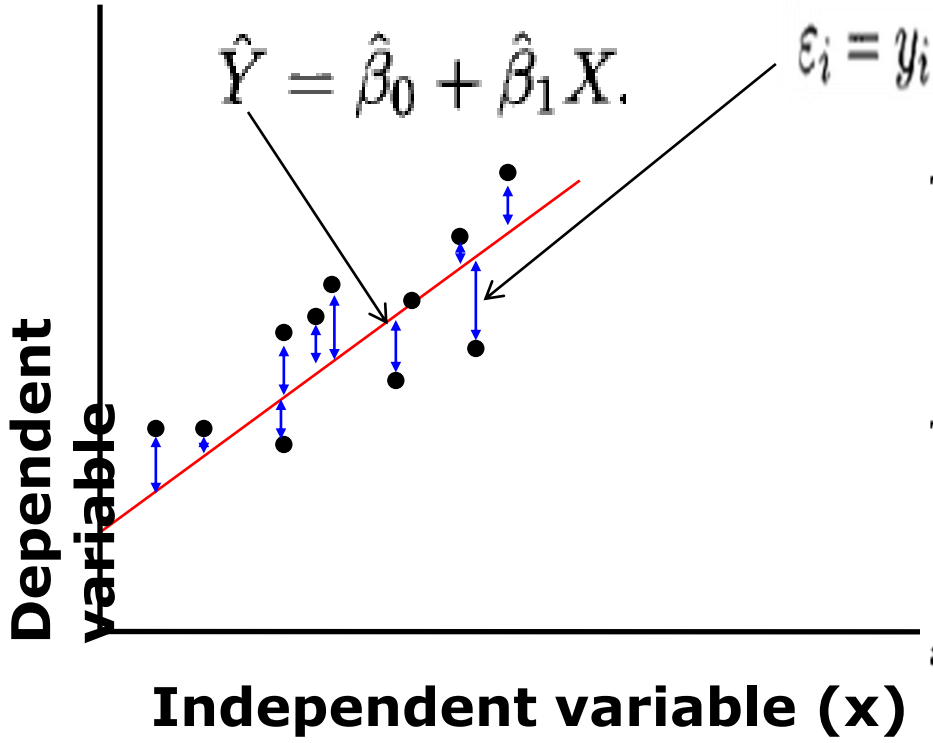
$$\text{Minutes} = \beta_0 + \beta_1 \text{Units} + \varepsilon,$$

Length of Service Calls (in Minutes) and Number of Units Repaired

Row	Minutes	Units	Row	Minutes	Units
1	23	1	8	97	6
2	29	2	9	109	7
3	49	3	10	119	8
4	64	4	11	149	9
5	74	4	12	145	9
6	87	5	13	154	10
7	96	6	14	166	10

Việc ước lượng tham số

- Chúng ta ước lượng các tham số bằng phương pháp phổ biến bình phương tối thiểu.
- Chúng ta giảm thiểu tổng bình phương của khoảng cách dọc từ mỗi điểm đến đường thẳng



$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i, \quad i = 1, 2, \cdots, n$$

The sum of squares of these distances can then be written as

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

The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize $S(\beta_0, \beta_1)$ are given by

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

and

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

Việc ước lượng tham số

- Đối với mỗi điểm quan sát trong tập dữ liệu, chúng ta có thể tính toán

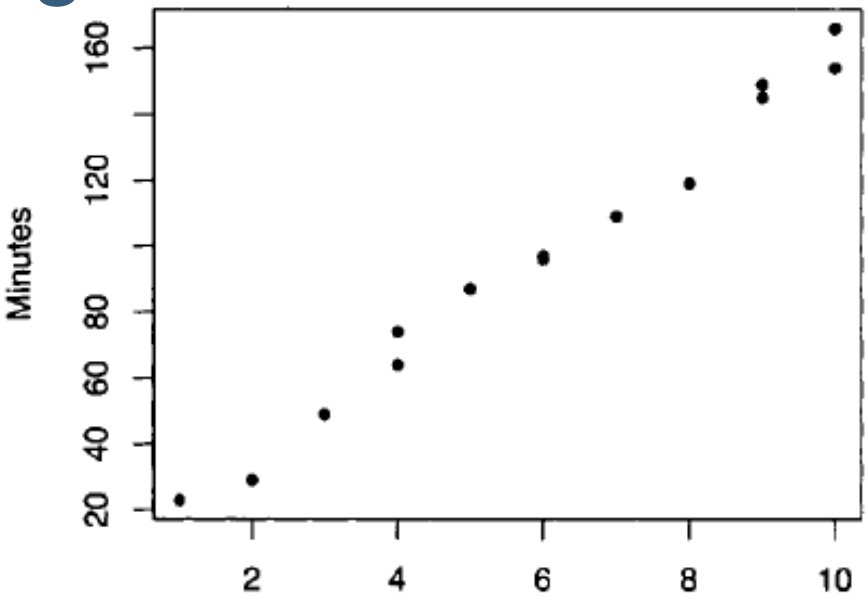
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, 2, \cdots, n.$$

- Khoảng cách dọc tương ứng với giá trị quan sát thứ i là:

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \cdots, n. \qquad \textit{least squares residuals.}$$

Việc ước lượng tham số

- Dựa vào dữ liệu có sẵn, chúng ta ước lượng các tham số β_0 và β_1
- Tìm 1 đường thẳng phù hợp nhất đi qua các điểm trong đồ thị scatter plot của biến phản hồi và biến dự đoán sao cho tổng bình phương khoảng cách từ điểm đó đến đoạn thẳng là nhỏ nhất



Length of Service Calls (in Minutes) and Number of Units Repaired

Row	Minutes	Units	Row	Minutes	Units
1	23	1	8	97	6
2	29	2	9	109	7
3	49	3	10	119	8
4	64	4	11	149	9
5	74	4	12	145	9
6	87	5	13	154	10
7	96	6	14	166	10

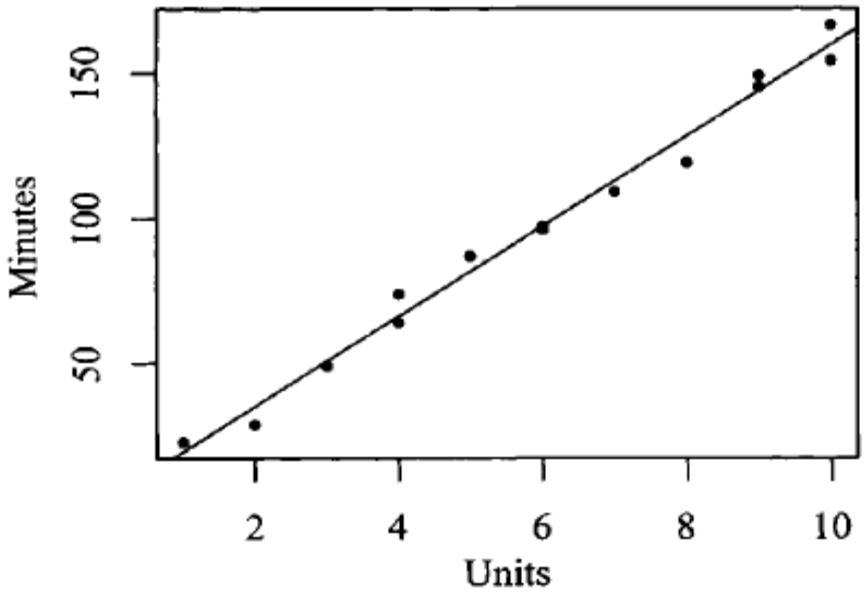
Việc ước lượng tham số

Length of Service Calls (in Minutes) and Number of Units Repaired

Row	Minutes	Units	Row	Minutes	Units
1	23	1	8	97	6
2	29	2	9	109	7
3	49	3	10	119	8
4	64	4	11	149	9
5	74	4	12	145	9
6	87	5	13	154	10
7	96	6	14	166	10

$$\hat{\beta}_1 = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} = \frac{1768}{114} = 15.509,$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 97.21 - 15.509 \times 6 = 4.162.$$

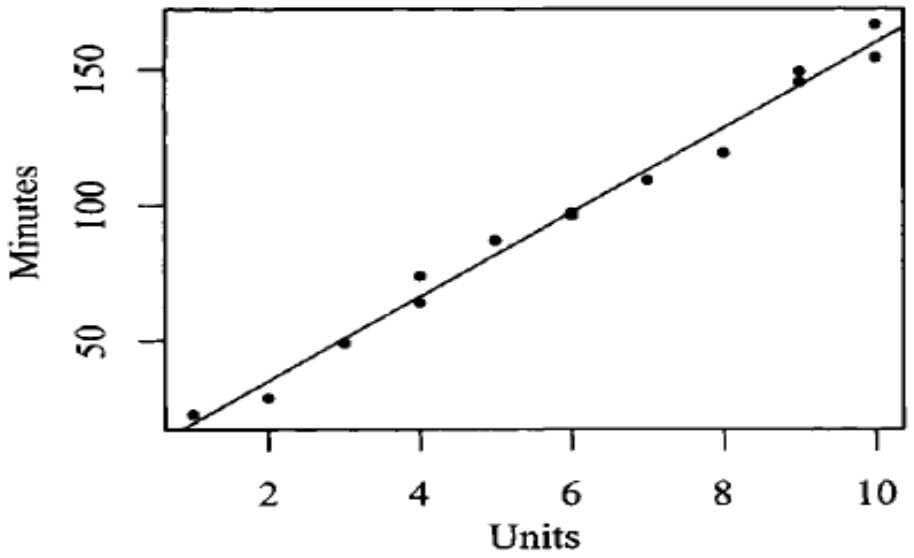


Việc ước lượng tham số

- Phương trình đường hồi quy như sau:

$$\text{Minutes} = 4.162 + 15.509\text{Units}$$

- Hằng số của phương trình cho ta biết thời gian bắt đầu cho mỗi lần gọi sửa chữa xấp xỉ là 4 phút
- Hệ số hồi quy là 15.5 có nghĩa là mất 15.509 để gọi sửa chữa 1 đơn vị linh kiện.



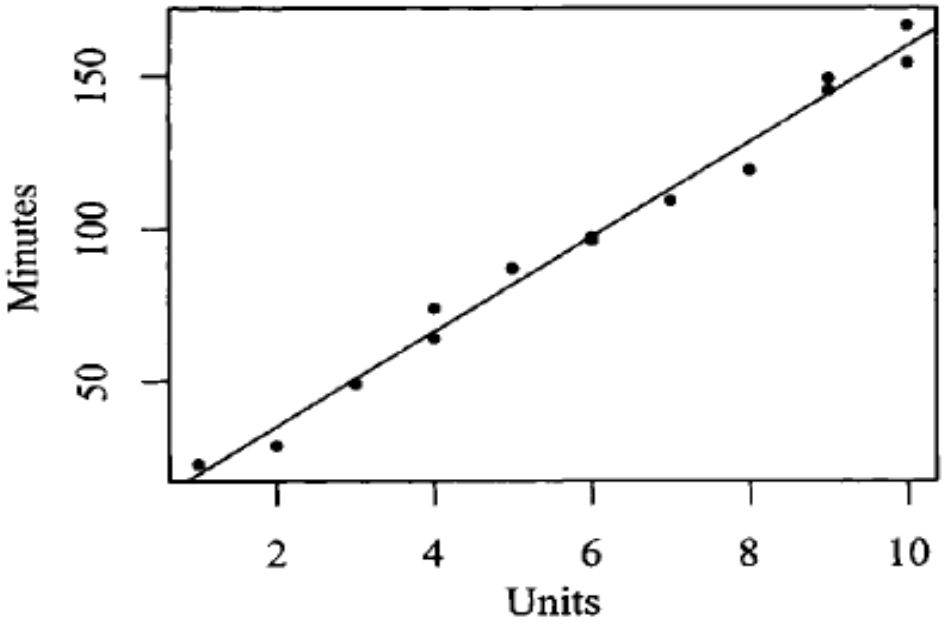
Từ phương trình đường hồi quy, ta dự đoán:

Thời gian cho cuộc gọi đến dịch vụ để sửa 4 linh kiện là:

$$\hat{y} = 4.162 + 15.509 \times 4 = 66.20$$

Việc ước lượng tham số

- Trong phân tích của chúng ta, chúng ta chỉ đưa ra một giả định, đó là Y và X có quan hệ tuyến tính.
- Cách để kiểm tra giả định trên là vẽ scatter plot của **biến phản hồi và biến dự đoán**, sau đó vẽ đường hồi quy đi qua các điểm đó



- Đường hồi quy là được xác nhận nếu các điều kiện giả định cho bài toán hồi quy được thỏa mãn

Các giả định cho bài toán hồi quy

- **1.** Assumptions about the form of the model (examining the **scatter plot**)
- **2.** Assumptions about the errors
  - The error ε_i, i = 1,2, ... , *n* has a **normal distribution**.
  - The errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ have **mean zero**
  - The errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ have the same (but unknown) variance σ^2 (as the **homogeneity** or the **homoscedasticity** assumption)
  - The errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, are independent of each other (their pairwise covariances are zero).
- **3. Assumptions about the observations**
  - All observations are equally **reliable** and have an approximately equal role in determining the regression results and in influencing conclusion

Kiểm định giả thuyết

- Kiểm định giả thuyết về tham số hồi quy β_0, β_1.
- Kiểm định giả thuyết $\beta_0 = 0$ (không có mối quan hệ tuyến tính nào giữa Y và X): dùng scatter plot.

Kiểm định giả thuyết

- **H_0: $\beta_1 = \beta_1^0$ (β_1^0 : constant chosen by investigator)**
- **H_1: $\beta_1 \# \beta_1^0$**

t-Test

$$\text{s.e.}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}},$$

$$t_1 = \frac{\hat{\beta}_1 - \beta_1^0}{\text{s.e.}(\hat{\beta}_1)} \qquad \hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{\text{SSE}}{n-2}$$

- **H_0** bị bác bỏ tại mức có ý nghĩa α nếu:

$$|t_1| \geq t_{(n-2,\alpha/2)}$$

hoặc

$$p(|t_1|) \leq \alpha$$

Kiểm định giả thuyết

- **Ví dụ:** giả sử rằng ban quản lý dự kiến việc tăng thời gian phục vụ cho mỗi đơn vị bổ sung sẽ được sửa chữa là 12 phút.
- Thực hiện kiểm định sau:
- **H_0: β_1= 12**
- **H_1: β_1# 12**

$$t_1 = \frac{\hat{\beta}_1 - 12}{\text{s.e.}(\hat{\beta}_1)} = \frac{15.509 - 12}{0.505} = 6.948$$

$$t_1 = \frac{\hat{\beta}_1 - \beta_1^0}{\text{s.e.}(\hat{\beta}_1)}$$

$$|t_1| \geq t_{(n-2,\alpha/2)}$$

$t_{(n-2,\ \alpha/2)}$= $t_{(12,\ 0.025)}$=2.18 (xem bảng sau)

- **t_1= 6.948>2.18: → bác bỏ H_0**
- Management's estimate of their increase in time for each additional component to be repaired is not supported by the data.

Kiểm định giả thuyết

t-Distribution with n Degrees of Freedom

	α				
n	0.10	0.05	0.025	0.010	0.005
1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.97	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.42	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
12	1.36	1.78	2.18	2.68	3.06
14	1.34	1.76	2.14	2.62	2.98
16	1.34	1.75	2.12	2.58	2.92
18	1.33	1.73	2.10	2.55	2.88
20	1.32	1.72	2.09	2.53	2.84
30	1.31	1.70	2.04	2.46	2.75
40	1.30	1.68	2.02	2.42	2.70
60	1.30	1.67	2.00	2.39	2.66
120	1.29	1.66	1.98	2.36	2.62
∞	1.28	1.64	1.96	2.33	2.58

Khoảng tin cậy

- Để xây dựng các khoảng tin cậy cho các tham số hồi quy, chúng ta cũng cần giả sử rằng ε có phân phối chuẩn
- (1- α) x 100% confidence interval for β_0 is given by
- (1- α) x 100% confidence interval for β_1 is given by

$$\hat{\beta}_0 \pm t_{(n-2,\alpha/2)} \times \text{s.e.}(\hat{\beta}_0)$$

$$\hat{\beta}_1 \pm t_{(n-2,\alpha/2)} \times \text{s.e.}(\hat{\beta}_1)$$

$$\text{s.e.}(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}} \qquad \hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = \frac{\text{SSE}}{n-2}$$

$$\text{s.e.}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}},$$

Khoảng tin cậy

- Từ bảng dưới, ta thấy rằng 95% khoảng tin cậy cho β_1:

$$15.509 \pm 2.18 \times 0.505 = (14.408, 16.610)$$

Variable	Coefficient	s.e.	t-Test	p-value
Constant	4.162	3.355	1.24	0.2385
Units	15.509	0.505	30.71	< 0.0001

- **Thời gian gia tăng cần thiết cho mỗi đơn vị linh kiện bị hỏng là từ 14 đến 17 phút.**

Dự đoán

- Phương trình hồi quy đơn giản có thể được sử dụng để dự đoán giá trị của biến phản hồi (response) bằng các giá trị cụ thể của biến dự đoán (predictor)
- Giá trị dự đoán là $\hat{y}_0$ tương ứng với x_0 theo công thức sau:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

- Khoảng ước lượng cho $\hat{y}_0$ với hệ số tin cậy 1-α là:

$$\hat{y}_0 \pm t_{(n-2,\alpha/2)}\, \text{s.e.}(\hat{y}_0)$$

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{\text{SSE}}{n-2}$$

$$\text{s.e.}(\hat{y}_0) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Dự đoán

- **Ví dụ:** chúng ta muốn dự đoán thời gian của một cuộc gọi đến dịch vụ trong đó 4 linh kiện phải được sửa chữa.

$$\hat{y}_4 = 4.162 + 15.509 \times 4 = 66.20$$

$$\text{s.e.}(\hat{y}_4) = 5.392\sqrt{1 + \frac{1}{14} + \frac{(4-6)^2}{114}} = 5.67$$

- Khoảng ước lượng cho $\hat{y}_4$ *với hệ số tin cậy là 1-α là:*

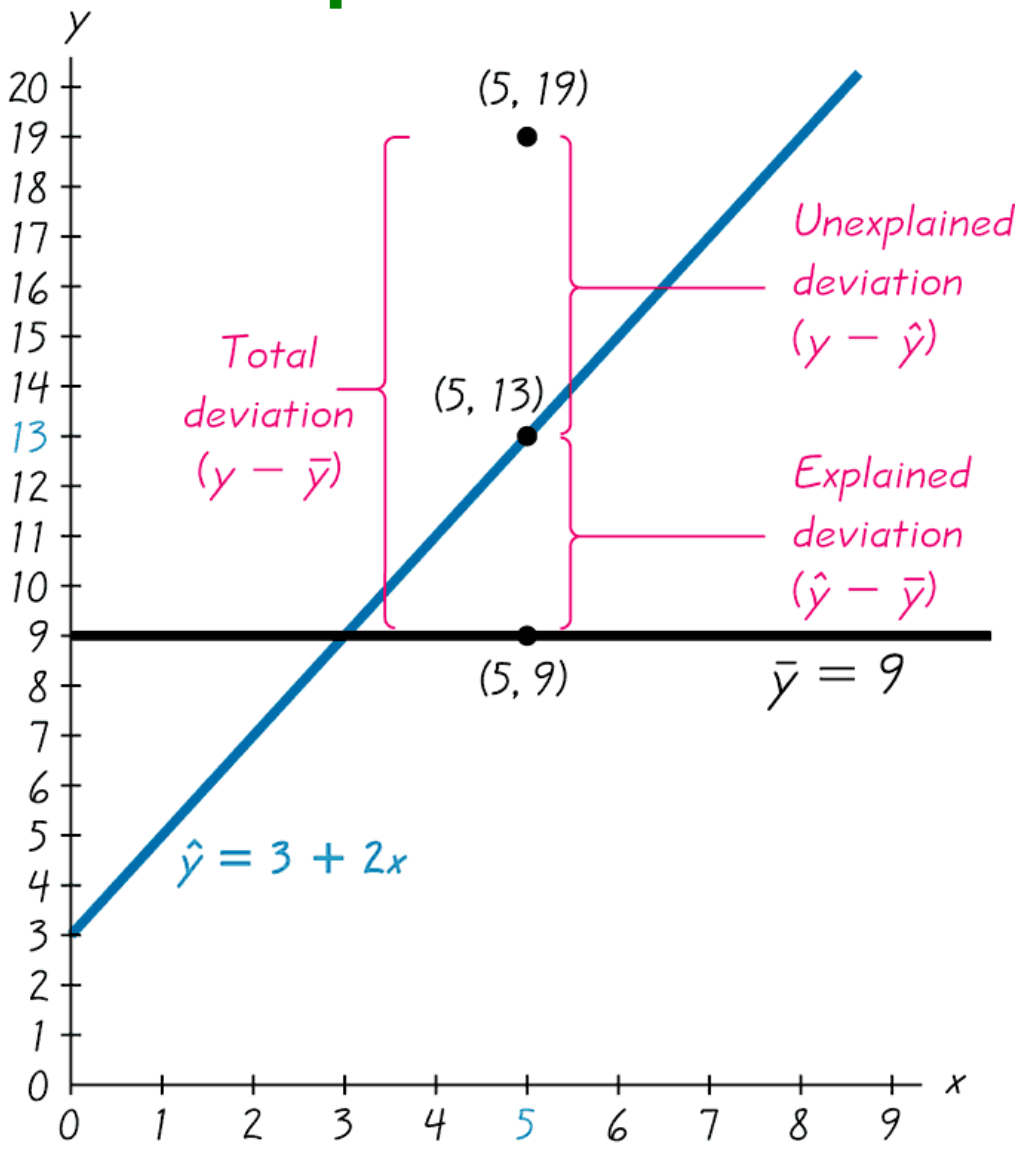
$$\hat{y}_0 \pm t_{(n-2,\alpha/2)}\,\text{s.e.}(\hat{y}_0)$$

$66.2 \pm 2.18*5.67$

Measuring the quality of fit

- Sau khi điều chỉnh mô hình tuyến tính liên quan đến Y với X, chúng ta không chỉ quan tâm đến việc liệu mối quan hệ tuyến tính có tồn tại không mà còn đo lường chất lượng của sự phù hợp của mô hình với dữ liệu.
- Chất lượng của sự phù hợp có thể được đánh giá bằng một trong những cách liên quan sau đây:
  - **Kiểm định giả thuyết H_0**
  - Kiểm tra Cor(Y, X)
  - Xem xét biểu đồ scatter plot của *Y* so với $\hat{Y}$
  - Measuring the quality of fit

Explained and Unexplained Variation



Assume the following:

- There is sufficient evidence of a linear correlation.
- The equation of the line is

$$\hat{y} = 3 + 2x$$

- The mean of the *y*-values is 9.
- One of the pairs of sample data is $x = 5$ and $y = 19$.
- The point (5,13) is on the regression line.

Explained and Unexplained Variation

The figure shows above lies on the regression line, but (5,19) does not.

We arrive at:

$$\text{Total Deviation (from } \bar{y} = 9) \text{ of the point } (5, 19) = y - \bar{y} = 19 - 9 = 10.$$

$$\text{Explained Deviation (from } \bar{y} = 9) \text{ of the point } (5, 19) = \hat{y} - \bar{y} = 13 - 9 = 4.$$

$$\text{Unexplained Deviation (from } \bar{y} = 9) \text{ of the point } (5, 19) = y - \hat{y} = 19 - 13 = 6.$$

Relationships

(total deviation) = (explained deviation) + (unexplained deviation)

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

(total variation) = (explained variation) + (unexplained variation)

$$\Sigma(y - \bar{y})^2 = \Sigma(\hat{y} - \bar{y})^2 + \Sigma(y - \hat{y})^2$$

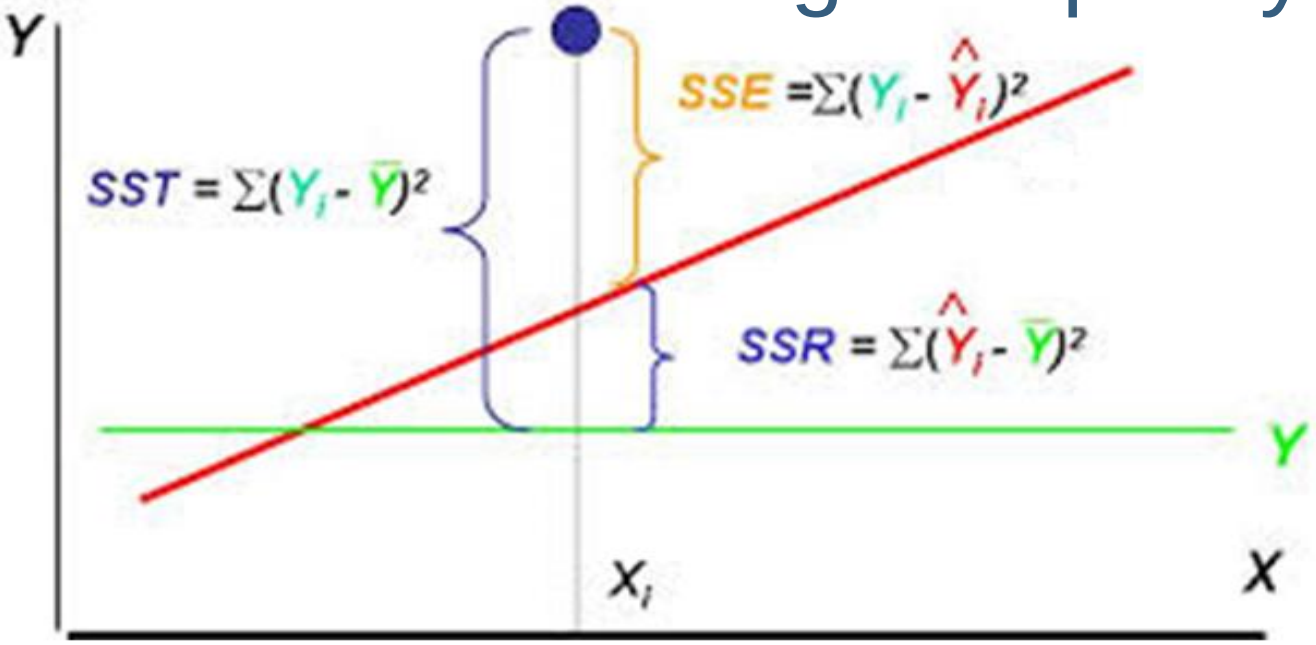
Definition

The coefficient of determination is the amount of the variation in *y* that is explained by the regression line.

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

The value of r^2 is the proportion of the variation in *y* that is explained by the linear relationship between *x* and *y*.

Measuring the quality of fit



$$SST = SSR + SSE.$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

SSE: đo lường lỗi trong dự đoán này

Measuring the quality of fit

Computer Repair Data

i	x_i	y_i	$\hat{y}_i$	e_i	i	x_i	y_i	$\hat{y}_i$	e_i
1	1	23	19.67	3.33	8	6	97	97.21	−0.21
2	2	29	35.18	−6.18	9	7	109	112.72	−3.72
3	3	49	50.69	−1.69	10	8	119	128.23	−9.23
4	4	64	66.20	−2.20	11	9	149	143.74	5.26
5	4	74	66.20	7.80	12	9	145	143.74	1.26
6	5	87	81.71	5.29	13	10	154	159.25	−5.25
7	6	96	97.21	−1.21	14	10	166	159.25	6.75

$$\text{SST} = \text{SSR} + \text{SSE}.$$

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

- Sử dụng dữ liệu Sửa chữa máy tính, các fitted values và residuals trong bảng trên, chúng ta tính được:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{348.848}{27768.348} = 0.987.$$

- Giá trị $R^2 = 0.987$ cho thấy rằng gần **99% tổng biến thiên của biến phản hồi (phút) được tính bằng biến dự đoán (đơn vị).**
- Giá trị cao của R^2 biểu thị mối quan hệ tuyến tính mạnh mẽ giữa thời gian phục vụ và số lượng đơn vị được sửa chữa trong một cuộc gọi dịch vụ.