TEST 3: Chapter 8, 9, 10, 11

Chapter 8: Statistical Intervals for a Single Sample

8.1 Introduction

Confidence interval (khoảng tin cậy) là khoảng mà ta tính ra từ các dữ liệu của sample để dự đoán thông số của Population. Ví dụ 1-α level of confidence của μ tức là:
Xác suất để μ nằm trong khoảng tin cậy đó là 1-α

$$P(L \leq \theta \leq U) = 1 - \alpha$$

L: Lower-confidence limit
U: Upper-confidence limit

> Ví dụ: Ta không biết tuổi thọ của tất cả các con chó là bao nhiêu, ta sẽ thi thu thập dữ liệu của 100 con. Vậy ta sẽ có sample mean, ví dụ sample mean = 10 năm. Sau đó ta tính khoảng tin cậy 95%, tức α = 5%, ra [7.5, 12,5] chẳng hạn, thì tức là mean μ của Population sẽ có 95% khả năng nằm trong khoảng đó.

8.2 Confidence Interval for a Population Mean (Khoảng tin cậy cho mean)

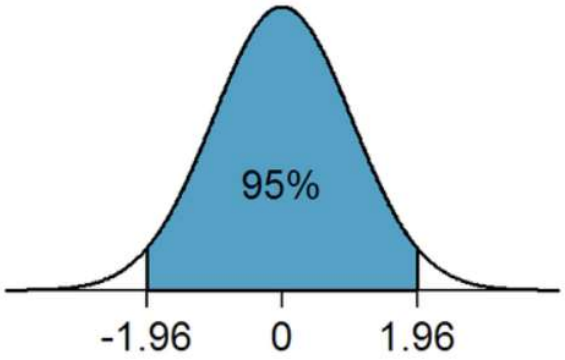
Ta sẽ chia ra 3 trường hợp

- Trường hợp 1: variance σ^2 đã biết.
- Trường hợp 2: Đây là 1 phân bố bất kỳ, không nhất thiết phải là normal, nhưng phải có size lớn (≥ 40).
- Trường hợp 3: variance σ^2 chưa biết.

Trường hợp 1: variance σ^2 đã biết
1–α confidence interval cho mean μ là:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Zα/2 là giá trị của Z để xác suất P(Z > zα/2) = α/2.
Để dễ hình dung, ta có α=5%, $Z_{\alpha/2} = Z_{0.025}$ = 1.96, hay P(Z>1.96) = α/2 = 0.025



VD: Tuổi thọ của bóng đèn được biết là có phân bố chuẩn và có σ = 25 giờ. Ta có một sample gồm 20 bóng và chúng có tuổi thọ trung bình là 1014 giờ. Tìm 95% confidence interval cho mean μ của loại bóng đèn này .
Phân tích: Population Loại bóng đèn này phân bố chuẩn nên sample cũng có phân bố chuẩn, chúng ta đã biết standard deviation σ = 25 giờ. Giải: Ta có thêm một cái sample gồm 20 bóng có tuổi thọ trung bình là 1014 => $\overline{x}$ = 1014, n=20. Ta phải tính độ tin cậy 95%, tức α = 5%, Zα/2 = 1.96. Thay vào tính đc 1003.04 ≤ μ ≤1024.96 .

Còn 1 dạng trong cái này, đó là bắt tính số lượng (n) để có thể đạt được một khoảng confidence interval nhất định.Ta thấy rằng để dự đoán mean μ của tổng thể, thì càng nhiều dữ liệu càng tốt, điều đố khá dễ hiểu và bạn cũng có thể nhìn vào phương trình trên kia để thấy điều đó, khi n càng lớn thì confidence interval càng thu hẹp lại. Vì vậy để tự tin rằng ta có một khoảng confidence interval bằng bao nhiêu đó thì ta cũng phải có số lượng dữ liệu nhất định. Công thức:

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$$

E ở đây là $|\overline{x} - \mu|$
VD: Lấy lại VD ở trên, thử tính lượng bóng đèn cần để ta có 95% confident để sai số khi ta ước tính mean μ của tổng thể < 5 giờ.
Giải: Ta có : α = 5%, => α/2 = 0.025. σ = 25 giờ, E = 5 giờ, thay vào công thức n = 96.05, tuy nhiên ta luôn phải làm tròn lên nên n=97.

Trường hợp 2: Đây là 1 phân bố bất kỳ, không nhất thiết phải là normal, nhưng phải có size lớn (≥ 40).
1–α confidence interval cho mean μ:

$$\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}$$

Bởi vì đây không phải phân bố chuẩn nên không có σ nhé. Chỉ có S cho cái sample đấy thôi, nhưng cũng chả khác nhau đâu.

Trường hợp 3: variance σ² chưa biết biết.
Vì σ chưa biết nên chỉ có S, hãy nhớ trường hợp σ chưa biết, đồng thời là phân bố chuẩn thì mới dùng t, còn không phải dùng z. Công thức:

$$\overline{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \le \mu \le \overline{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

8.3 Large-Sample Confidence Interval for a Population Proportion (Confidence Interval cho xác suất)

Nó cũng giống như mean thôi, mỗi tội là tìm Confidence Interval cho xác suất. Công thức:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Trong đó $\hat{p} = \frac{x}{n}$ là xác suất của sample, x là cái lượng thỏa mãn, còn n là kích thước sample. Nếu để ý, bạn sẽ thấy pˆ(1-pˆ) trong công thức chính là σ² giống như phần binorminal, nên suy cho cùng, công thức này y hệt phần trên.

VD: 1000 ca ung thư được lấy ngẫu nhiên, và 823 ca chết sớm. Tính 95% confidence interval cho tỉ lệ chết sớm.
Giải: Ta có x = 823, n = 1000 => pˆ = 0.823. α=5% = 0.05 => α/2 = 0.025 => Zα/2 = 1.96. Vậy ta tính được 0.799 ≤ p ≤ 0.847.

Trong chap này ta mới chỉ đọc thấy confidence interval được giới hạn 2 đầu, nhưng trong bất kỳ dạng nào đề có các trường hợp confidence interval 1 phía, Ví dụ như Phần 8.1 trường hợp 1, ta sẽ có các bài toán bắt tính confidence interval cho 1 phía,

A $(1-\alpha)$ upper-confidence bound for μ is

$$\mu \le \overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}}$$

A $(1-\alpha)$ lower-confidence bound for μ is

$$\mu \ge \overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}}$$

Mọi trường hợp 1 phía thì Zα/2 sẽ được thay bằng Zα.

Chapter 9: Test of Hypotheses for a Single Sample (Kiểm chứng giả thuyết cho 1 cái sample)

Chapter 8 ta đã biết cách tính confidence interval, thì chapter này ta sẽ dùng nó để xem 1 cái kết luận nào đó có sai hay không. Ví dụ có 1 người nói số giờ làm trung bình trong 1 ngày của người Việt là 4.5 tiếng , thì để kiểm chứng câu nói này, ta sẽ đi thu thập dữ liệu 1000 người, sau đó ta sẽ tính confidence interval như trên chapter 8, rồi sau đó so sánh 2 dữ liệu để đưa ra kết luận.

9.1 Hypothesis Testing (kiểm chứng giả thuyết)

Statistical hypothesis (giả thuyết) là 1 cái phát biểu về Parameter của Population. Nhớ lại rằng Parameter là các thông số của Population, còn Statistic là của Sample, rõ ràng là các Parameter thì chúng ta chưa thể biết, ta chỉ có thể ước chừng chúng

bằng cách lấy 1 sample để tính confidence interval như ở chapter 8.
Vì vậy ta sẽ lấy 1 cái sample để tính confidence interval, sau đó so sánh với cái giả thuyết để đưa ra kết luận (VD như cái đóng khung ở trên). Và với ví dự đó, ta có :

$H_0 : \mu = 4.5 \longleftarrow$ null hypothesis

$H_1 : \mu \neq 4.5 \longleftarrow$ alternative hypothesis

Trong 1 vài trường hợp, có thể giả thuyết sẽ là 1 phía:

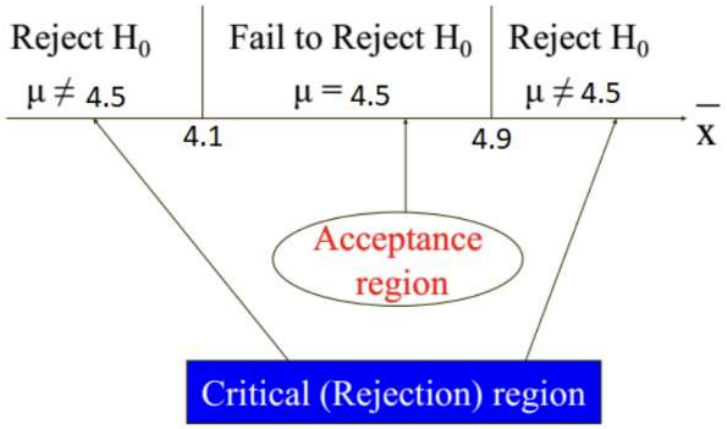
$$H_0 : \quad \mu = 4.5$$
$$H_1 : \quad \mu > 4.5$$

or

$$H_0 : \quad \mu = 4.5$$
$$H_1 : \quad \mu < 4.5$$

VD: 1 phát biểu rằng chiều cao trung bình người Việt lớn hơn 1m65. khi đó H1 : μ>1m65.

Như đã thấy ở trên, H0 là null hypothesis, H1 là alternative hypothesis.
- Nếu reject H0 thì có nghĩa là ta có bằng chứng đủ mạnh để kết luận rằng H1 đã đúng.
- Còn nếu không reject H0, thì ta không có đủ bằng chứng là H0 đã đúng.



Note: Nhìn biết đồ trên, nó có ý nghĩa rằng ta có 1 cái sample và ta tính được rằng khoảng tin cậy confidence interval là [4.1, 4.9], tức là H0 cho μ bằng bất kỳ số nào từ 4.1 đến 4.9, ta đều phải fail to reject, còn H0 mà nằm ngoài khoảng đấy, thì ta reject . Và nên nhớ 1 điều, H0 thì luôn luôn là dấu "=", và H1 thì ghi là khác một giá trị cố định, nhưng thực chất ta chỉ reject H0 và chấp nhận H1 khi giá trị của H0 nằm ngoài khoảng tin cậy (confidence interval).

VD: Có 1 người nghĩ rằng chiều cao trung bình của người Việt là 1m65, để ông đấy xem suy đoán của mình có đúng hay không, ông ta đi khảo sát 1000 người ở Hà Nội. Ông ta tính được khoảng tin cậy là [1m55, 1m66], Vì vậy ông ta phải fail to reject H0, bởi vì H0: μ = 1m65, nằm trong khoảng tin cậy.

Tuy nhiên, các bạn phải hiểu rằng ta chỉ tính trên sample để kết luận, nên đôi khi kết luận của ta có thể sai do dữ liệu ta thu thập là toàn những trường hợp đặc biệt.

Như khi ta tính khoảng tin cậy 95% thì ta có 95% khả năng là đúng, rõ ràng ta vẫn có 5% khả năng kết luận sai. Và sự sai sót ấy chia thành 2 trường hợp:
- Type 1 error : Reject H0 khi mà nó đúng. Để dễ hình dung hãy nhìn lại hình bên trên, nó có nghĩa rằng tiên đoán trước đó của mình về mean là đúng, nhưng do thu thập được toàn dữ liệu cùi mà ta tính được khoảng tin cậy bị lệch đi, dẫn đến mean μ của ta nằm ngoài khoảng tin cậy.
- Type 2 error : Fail to reject H0, tức là mình không bác bỏ H0 trong khi nó sai, ngược lại với Type 1.

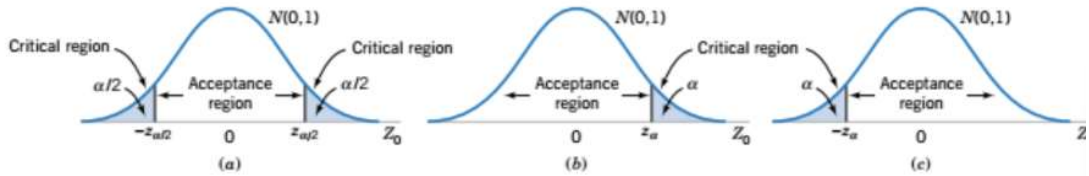
9.2 Tests on the Mean of a Normal Distribution N(μ, σ²)

Trong phần 9.1 mình đã giải thích khá rõ về reject và fail to reject, các bạn cũng đã hiểu về confidence interval. Và trong các bài toán cụ thể, các bạn sẽ gặp 2 trường hợp để tính toán rồi đưa ra kết luận về H0:

- Case 1: Variance σ² của Population đã biết.
Test Statistic:

$$Z_0 = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$$

Trong bài toán cụ thể, ta sẽ tính Test Statistic, sau đó so sánh nó với các Z của giới hạn của khoảng tin cậy(confidence interval). Nếu nằm ngoài thì reject H0 thôi.

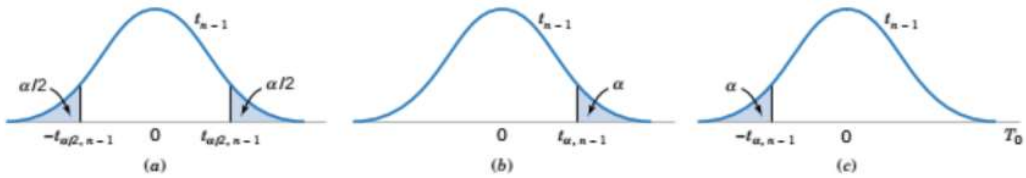


Như hình trên,
(a) là khoảng tin cậy có 2 phía với H1: μ ≠ μ0
(b) là khoảng tin cậy 1 phía với H1: μ > μ0
(c) là khoảng tin cậy 1 phía với H1: μ < μ0.

- Case 2: Variance σ² của Population chưa biết.
Test Statistic:

$$T_0 = \frac{\overline{X} - \mu_0}{S}\sqrt{n}$$

Khi không biết σ thì dùng t, khi biết σ thì dùng z, rất dễ, không khác gì nhau.



9.3 Tests on a population proportion (Kiểm định xác suất)

Phần trên thì kiểm định trên mean, bây giờ thì kiểm định xác suất, cách tính thì không khác gì nhau.

$$H_0: \quad p = p_0$$
$$H_1: \quad p \neq p_0 \ (\text{ or } H_1 : p > p_0, \text{ or } H_1 : p < p_0)$$

Thay vì tiên đoán trước giá trị trung bình của cái gì đấy, thì H0 của phần này sẽ là tiên đoán trước xác suất. Và ta cũng loại chúng nếu chúng nằm ngoài khoảng tin cậy thôi.

Ta sẽ có xác suất trong sample là p mũ, sau đó ta sẽ tính Test Statistic:

$$Z_0 = \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$$

Rồi so sánh với Z của các giới hạn của khoảng tin cậy confidence interval:

> Ví dụ: Một tạp chí nói rằng 1 nửa số tiến sĩ sẽ học tiếp sau khi tốt nghiệp. Dữ liệu từ một khảo sát cho thấy 117 người trong số 484 người ở trường X học tiếp sau khi tốt nghiệp. Câu hỏi: với α = 0.05, đưa ra kết luận về phát biểu trước đó.
> Giải:
> H0: P0 = 0.5 (1 nửa)
> Ta có P mũ = 117/484 = 0.24. Test Statistic Z0 = -11.44 .
> α = 0.05 => Zα/2 = $Z_{0.025}$ = 1.96. Mà |Z0| = 11.44 > 1.96, nằm ngoài khoảng tin cậy nên ta reject H0.

Chapter 10: Statistical Inference for Two Samples (Suy luận thống kê cho samples)

> Chapter 9 chúng ta đã làm quen với suy luận thống kê của 1 sample, về mean μ và xác suất p. Còn chapter 10 thì cũng sẽ làm về mean μ và xác suất p, chỉ khác là sẽ thực hiện trên 2 sample, 2 cái trừ cho nhau. Sẽ có 2 cái chính là tính confidence interval cho hiệu của 2 mean μ của 2 tổng thể khác nhau, và kiểm định H0 của nó.

10.1 Inference on the Difference in Means of Two Normal Distributions (Hiệu 2 mean)

Cũng như khi tính confidence interval cho 1 sample, thì phần này cũng chia thành 2 trường hợp: σ đã biết và chưa biết.

- Case 1: σ đã biết

Khi đó, hiệu 2 mean sẽ có :

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Khi này bạn sẽ dễ dàng đoán được công thức của 1-α confidence interval của hiệu 2 mean:

$$\overline{x}_1 - \overline{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \overline{x}_1 - \overline{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- Case 2: σ chưa biết

Khi mà σ chưa biết thì ta chỉ tính được các variance s của sample thôi. Và khi này, sẽ có 1 cái variance chung được gọi là pooled estimator của σ^2

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

where

$S_1^2 =$ the variance of the sample taken from population 1

$S_2^2 =$ the variance of the sample taken from population 2

Công thức 1-α confidence interval của hiệu 2 mean, ghi nhớ là t sẽ có n1+n2-2 bậc tự do:

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,n_1+n_2-2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,n_1+n_2-2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Note: Các trường hợp đều có những bài toán tính confidence interval 1 phía. Và ta chỉ cần thay α/2 bằng α với những bài toán ấy.

Hypothesis Tests on the Difference in Means

Cho hypothesis:

$$\begin{aligned} &H_0: \quad \mu_1 - \mu_2 = \Delta_0 \\ &H_1: \quad \mu_1 - \mu_2 \neq \Delta_0 \\ &(\text{ or} \quad H_1: \mu_1 - \mu_2 > \Delta_0, \text{ or } H_1: \mu_1 - \mu_2 < \Delta_0) \end{aligned}$$

Khi này H0 sẽ là một phát biểu về độ chênh lệch giữa mean của 2 tổng thể: ∆0.

VD: Một người cho rằng độ chênh lệch tuổi thọ trung bình giữa nam hơn nữ là 2 tuổi. Khi đó ta có H0: μ1 − μ2 = 2 tuổi.

Trong các bài toán cụ thể, ta lại gặp 2 trường hợp : σ đã biết và chưa biết. Và chúng ta làm y hệt các bài toán về H0 trước đó: tính test Statistic và so sánh với các giới hạn của confidence interval, nếu nằm ngoài thì reject H0

- Case 1: σ đã biết

Test Statistic:

$$Z_0 = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Case 2: σ chưa biết

Test Statistic:

$$T_0 = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

(Sp là pooled estimator)

10.6 Inference on the Two Proportions (Hiệu xác suất của 2 tổng thể)

1-α confidence interval của hiệu 2 xác suất:

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \le p_1 - p_2 \le \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

p mũ là xác suất trong sample.

Tests on the Difference in Population Proportions

$$\begin{aligned} H_0 &: \quad p_1 = p_2 \\ H_1 &: \quad p_1 \neq p_2 \ (\text{ or } H_1 : p_1 > p_2, \text{ or } H_1 : p_1 < p_2) \end{aligned}$$

Tương tự như các bài toán trước đó, ta có test statistic:

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Chapter 11: Simple Linear Regression and Correlation (Hồi quy tuyến tính)

Nếu chỉ học qua môn thì gần như không cần học sâu chương này (Với điều kiện những chương trước đã khá vững :D).

11.1 Empirical Models (Mô hình thực nghiệm)

Chúng ta sẽ có 2 cái dữ liệu khác nhau, và ta xem chúng ta xem chúng có mối liên hệ chặt chẽ với nhau không. Vậy ta sẽ lập 1 phương trình tuyến tính giữa 2 cái dữ liệu này, xem có thể dự báo trước giá trị của biến này theo biến kia hay không. Biến cần dự đoán là dependent variable và biến mà mình dùng nó để suy ra biến kia là independent variables.

Ví dụ: cho 1 tập các dữ liệu về nhiệt độ ban ngày ở HN, và 1 tập các dữ liệu về các mặt đường bị rạn nứt và tất nhiên 2 tập này phải liên kết từng cặp với nhau, rồi lập ra một phương trình tuyến tính giữa 2 thứ đó. Khi này nếu như ta thấy nó có mối tương quan lớn thì sau đó ta có thể dự đoán về số vết nứt nhờ vào nhiệt độ. Và nhiệt độ ở

HN là independent variable và số vết nứt là dependent variable

Linear Regression function (hồi quy tuyến tính) là hàm khi xây dựng mối tương quan giữa 2 dữ liệu kia :

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

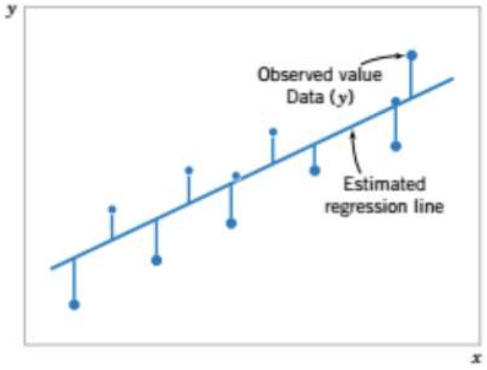
Thực ra công thức trên chả khác gì Y = a.X + b, nhưng viết khác bởi nó có các mục đích khác nhau.

11.2 Simple Linear Regression

Có n cặp dữ liệu:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

Từ các cặp dữ liệu trên, ta vẽ ra hàm tuyến tính như ở phần 11.1 sao cho nó là "best fit" với các dữ liệu. Và chúng ta dùng phương pháp đó là method of least squares, tức là tối thiểu hóa tổng bình phương các sai số(ε). Để dễ hiểu hơn, bạn thấy ta có phương trình hồi quy như ở 11.1, nhưng rõ ràng ta không thể tính giá trị của y_i bằng cách gán x_i như các bài toán bình thường bởi vì luôn có sai số:



Vì vậy, với sai số là ε, thì mỗi y_i ta sẽ có :

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

Do đó, tổng bình phương sai số được nhắc ở trên là :

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

Note: Trên thực tế, ta không có hàm hồi quy cho toàn bộ dữ liệu được nên ta chỉ ước tính nó bằng những dữ liệu mà ta có sẵn, Và từ trước đến giờ mình chỉ đang giới thiệu về lí thuyết, các bạn không cần nhớ các công thức ở trên nhưng phải hiểu. Và về sau các công thức cũng giống nhưng kí hiệu khác, bởi vì trên là các công thức cho tổng thể, mà ta chỉ tính trên các dữ liệu có sẵn (sample) thôi.
Các tham số trong phương trình hồi quy tuyến tính đó là:

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

với :

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

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

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Và từ đó, ta có estimated linear regression line:

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

Sai số residual (giống ε của tổng thể):

$$e_i = y_i - \hat{y}_i$$

Tổng bình phương các residual e_i là error sum of squares:

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

where

$$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

Từ đó, ta có công thức tính ước tính của σ^2 :

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

Sợ nhiều người lú, thì thực ra phần này chỉ gói gọn rằng:

- Hồi quy tuyến tính (Regression line):

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

- Residual (kiểu sai số, độ lệch):

$$e_i = y_i - \hat{y}_i$$

- Phương sai, tổng bình phương các độ lệch:

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

Trên đó là các tham số mấu chốt, còn tính thế nào thì các bạn phải hiểu nó đặc trưng cho cái gì.

11.3 Properties of the Least Squares Estimators (Các ước tính)

estimated standard error of the slope (ước tính sai số của slope) và estimated standard error of the intercept(ước tính sai số của intercept):

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

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

11.4 Hypothesis Tests in Simple Linear Regression (Kiểm định H0 trên hồi quy tuyến tính)

Test trên Slope

Hypotheses:

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

Nếu đã hiểu về phần kiểm định H0 rồi thì phần này cũng sẽ không xa lạ gì, cũng chỉ là kiểm định về một cái phát biểu bằng cách tính test statistic rồi so sánh với biên rồi kết luận.

Test statistic:

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$$

Reject H0 nếu :

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

(Nhớ là n − 2 degrees of freedom.)

Trường hợp đặc biệt:

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

Nếu Failure to reject H0, tức $\beta_1 = 0$ thì sẽ không có mối quan hệ giữa X và Y.

Test trên Intercept

Hypotheses:

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

Test statistic:

Reject H0 nếu :

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

(Nhớ là n − 2 degrees of freedom.)

11.8 Correlation (Hệ số tương quan)

Hệ số tương quan của X và Y là ρ, nhưng với sample, nó sẽ là:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

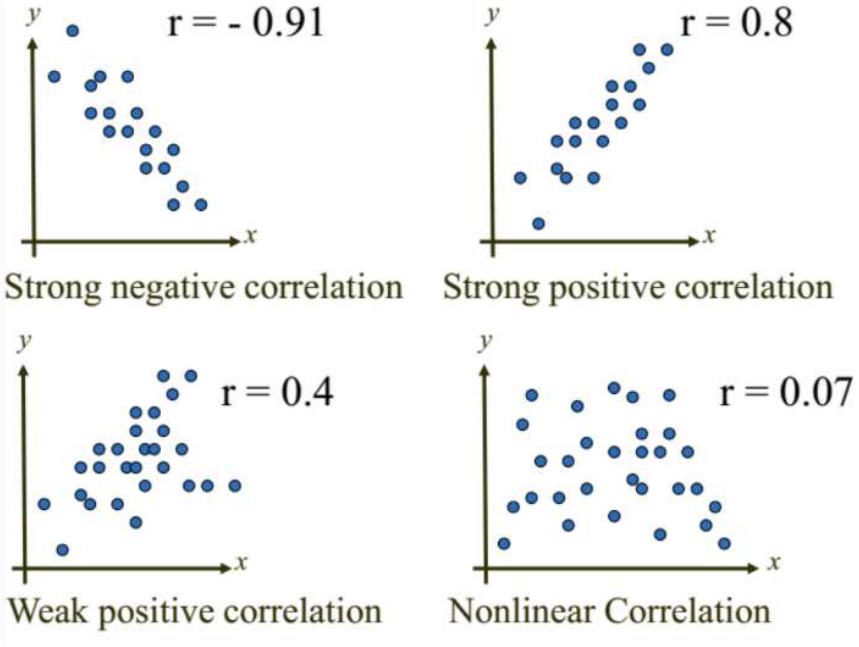
Note:

.) $-1 \le r \le 1$.

.) Nó đặc trưng cho mối liên hệ giữa 2 dữ liệu đó

.) r và β_1 có cùng dấu

.) r^2 gọi là coefficient of determination (hệ số xác định)



Như vậy, r càng tiến về 0 thì mối liên hệ giữa X và Y càng thấp, và ngược lại.

Test Hypotheses about the Correlation Coefficient

Hypotheses:

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

Test statistic:

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Reject H0 nếu :

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

Messi is GOAT, nếu bạn thấy đúng thì chứng tỏ bạn đã tiếp thu được nhiều kiến thức từ môn học này, good luck!