CHƯƠNG II

Mô hình hồi quy tuyến tính đơn biến

Dương Nam Hà (VNUA)

Nội dung chính

- I. Một số khái niệm TK cơ bản
- II. Giới thiệu mô hình hồi quy
- III. Hàm HQ tổng thể & HQ mẫu
- IV. Ước lượng các tham số - Phương pháp OLS
- V. Phân tích phương sai & Kiểm định giả thuyết thống kê

Dương Nam Hà (VNUA)

I. Một số khái niệm TK cơ bản

- **1.1 Biến ngẫu nhiên: Biến rời rạc** và **Biến liên tục**
- **1.2 Tổng thể** và **Mẫu**
- **1.3 Hàm phân phối xác suất (PDF, f(X))** và **Hàm mật độ phân phối xác suất tích lũy (CDF, F(X)):**
  - f(X) = lim ΔF(X)/ Δ x = F'(X)
- **1.4 Số trung bình:** $\bar{X} = \frac{\sum X}{n}$ hay $\bar{X} = \frac{\sum X f_i}{\sum f_i}$

Dương Nam Hà (VNUA)

I. Một số khái niệm… (tiếp)

- **1.5 Kì vọng** (E)
  - Biến rời rạc: $E(X) = \sum X.P(X) = \sum X.f(x)$
  - Biến liên tục: $E(X) = \int Xf(X)d(X)$
- **1.6 Phương sai** (Var, σ^2)
  - Biến rời rạc: $Var(X) = E[(X-\mu)^2] = \sum (X-\mu)^2.P(X)$
  - Biến liên tục: $Var(X) = E[(X-\mu)^2] = \int (X-\mu)^2 f(X)d(X)$

Dương Nam Hà (VNUA)

I. Một số khái niệm… (tiếp)

- **1.7 Độ lệch chuẩn** (σ)
  - Tổng thể: $\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$
  - Mẫu: $s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}$
- **1.8 Hiệp phương sai** (cov)
  - Tổng thể:

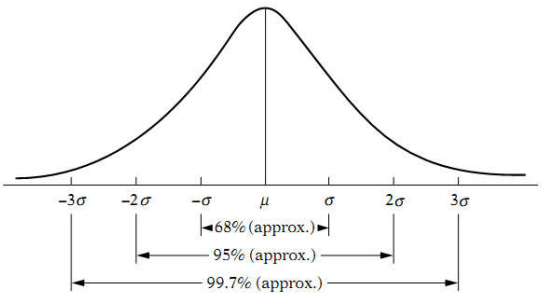
$$Cov(X,Y) = \sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_X\mu_Y$$

  - Mẫu:

$$Cov(X,Y) = \frac{\Sigma[(X-\bar{X})(Y-\bar{Y})]}{n-1}$$

I. Một số khái niệm… (tiếp)

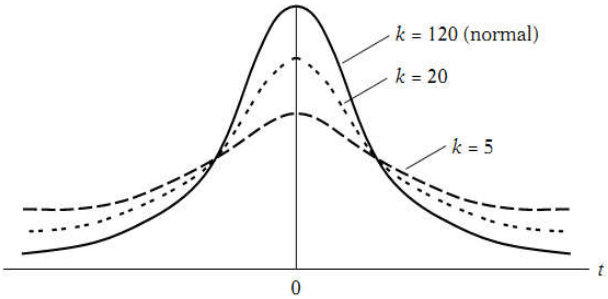
- **1.9 Phân phối chuẩn**



Hình 1: Diện tích dưới đường phân phối chuẩn
Nguồn: Gujarati, 2002

I. Một số khái niệm… (tiếp)

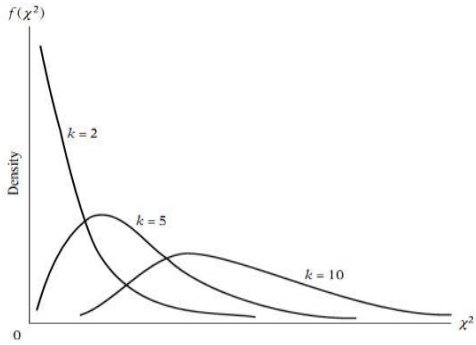
- **1.10 Phân phối Student (t)**



Hình 2: Phân phối t với một số bậc tự do cho trước
Nguồn: Gujarati, 2002

I. Một số khái niệm… (tiếp)

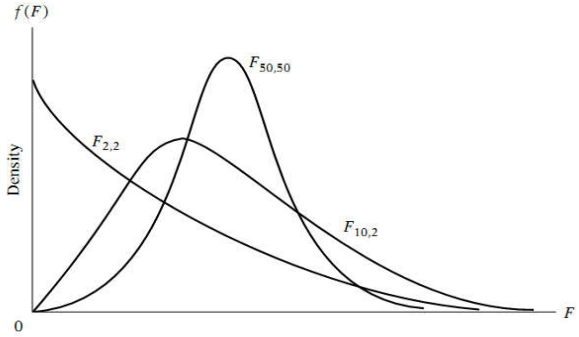
- **1.11 Phân phối Chi bình phương (χ^2)**



Hình 3: Hàm mật độ (density) của biến χ^2
Nguồn: Gujarati, 2002

I. Một số khái niệm… (tiếp)

- **1.12 Phân phối Fisher (F)**



Hình 4: Phân phối F với một số bậc tự do khác nhau
Nguồn: Gujarati, 2002

II. Giới thiệu mô hình hồi quy

- 2.1 Khái niệm
  - Mô tả và đánh giá mối quan hệ (Y và X)
  - Ví dụ?
- 2.2 Mục tiêu (riêng rẽ và phối hợp)
  - Ước lượng giá trị TB của Y
  - Phân tích ảnh hưởng của X → Y
  - Dự báo giá trị biến Y
  - Kiểm định/trắc nghiệm

II. Giới thiệu mô hình… (tiếp)

- 2.3 Phân biệt các mối quan hệ
  - A. QH chính xác vs. QH thống kê ngẫu nhiên
  - B. QH hồi quy vs. QH nhân quả
  - C. QH hồi quy vs. Tương quan
- 2.4 Số liệu trong phân tích hồi quy
  - A. Số liệu theo thời gian (Time Series Data)
  - B. Số liệu theo không gian (Cross-Section Data)
  - C. Số liệu kết hợp (Pooled/Panel Data)
  - D. Số liệu định tính (Qualitative Data)

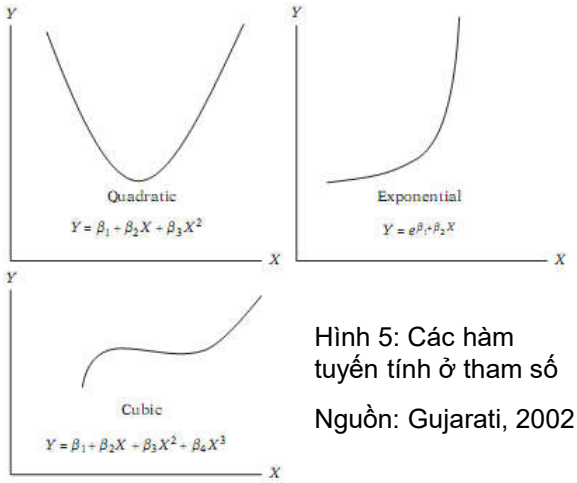
III. Hàm HQ tổng thể & HQ mẫu

- 3.1 Hàm HQ tổng thể
  - PRL: $E(Y|X_i) = B_1 + B_2X_i$
  - PRF: $Y_i = B_1 + B_2X_i + u_i$ hay $Y_i = E(Y|X_i) + u_i$
  - Trong đó: u_i là sai số ngẫu nhiên
  - Bản chất của u_i:
    - Phản ánh ảnh hưởng của biến khác;
    - Phản ánh sai sót trong điều tra/đvt;
    - Ảnh hưởng từ việc đơn giản hóa MH;
    - Sai số ngẫu nhiên không giải thích được (hành vi).

III. Hàm HQ… (tiếp)

- 3.2 Hàm HQ mẫu
  - SRL: $\hat{Y}_i = b_1 + b_2X_i$
  - SRF: $Y_i = b_1 + b_2X_i + e_i$
- 3.3 Hồi quy tuyến tính
  - Tuyến tính ở biến
  - Tuyến tính ở tham số

III. Hàm HQ… (tiếp)



Hình 5: Các hàm tuyến tính ở tham số

Nguồn: Gujarati, 2002

IV. Ước lượng các tham số - Phương pháp OLS

- 4.1 Giới thiệu
  - PRF: $Y_i = B_1 + B_2X_i + u_i$ với i = 1..n
  - → mục tiêu: ước lượng tham số tổng thể
  - → sử dụng số liệu mẫu, mô hình mẫu
  - SRF: $Y_i = b_1 + b_2X_i + e_i$
  - → PP OLS (Ordinary Least Squares) hay Σe_i^2 đạt cực tiểu

Dương Nam Hà (VNUA)

IV. Ước lượng… (tiếp)

- 4.2 Các giả thiết (Assumptions): Gaussian/CLRA/CLRM
  - (1) $Y_i = B_1 + B_2X_i + u_i$
  - (2) $E(u_i) = 0$
  - (3) $Var(u_i) = \sigma^2$ (constant)
  - (4) $Cov(u_i, u_j) = 0$
  - (5) $u_i \neq f(X_i)$ hay $Cov(X_i, u_i) = 0$
  - (6) $u_i \sim N(0, \sigma^2)$

Dương Nam Hà (VNUA)

IV. Ước lượng… (tiếp)

- Ngoài ra:
  - X_i được giả định là phi ngẫu nhiên (nonstochastic)
  - Số quan sát n > số tham số ƯL
  - Var (X) > 0
  - MHHQ là lựa chọn "chính xác"
  - Không có đa cộng tuyến hoàn hảo (HQ đa biến)

IV. Ước lượng… (tiếp)

- 4.3 PP bình phương nhỏ nhất (OLS)
  - PRF: $Y_i = B_1 + B_2X_i + u_i$ với i = 1..n
  - SRF: $Y_i = b_1 + b_2X_i + e_i$
  - ➔ $Z = \Sigma e_i^2 = \Sigma_i(Y_i - \hat{Y}_i)^2 = \Sigma_i(Y_i - b_1 - b_2X_i)^2 \rightarrow \min$
  - ➔ $\begin{cases} \partial Z/\partial b_1 = -2\Sigma_i(Y_i - b_1 - b_2X_i) = 0 & (4.1) \\ \partial Z/\partial b_2 = -2\Sigma_i(Y_i - b_1 - b_2X_i)X_i = 0 & (4.2) \end{cases}$
  - Giải hệ (4.1) và (4.2) → b_1, b_2?
  - Đặt: $\overline{X} = \frac{\sum_i X_i}{n}$ và $\overline{Y} = \frac{\sum_i Y_i}{n}$

IV. Ước lượng… (tiếp)

- (4.1) → $n\overline{Y} - nb_1 - nb_2\overline{X} = 0$
- $\Rightarrow \boxed{b_1 = \overline{Y} - b_2\overline{X}}$
- Thay b_1 vào (4.2):
- $\sum_i Y_i X_i - (\overline{Y} - b_2\overline{X})\sum_i X_i - b_2 \sum_i X_i^2 = 0$
- $\Rightarrow \sum_i Y_i X_i - \overline{Y}\sum_i X_i + b_2\overline{X}\sum_i X_i - b_2 \sum_i X_i^2 = 0$
- $\Rightarrow \sum_i Y_i X_i - \overline{Y}\sum_i X_i + b_2(\overline{X}\sum_i X_i - \sum_i X_i^2) = 0$
- $\Rightarrow \boxed{b_2 = \dfrac{\sum_i X_i Y_i - \overline{Y}\sum_i X_i}{\sum_i X_i^2 - \overline{X}\sum_i X_i} = \dfrac{\sum_i X_i Y_i - n\overline{Y}\,\overline{X}}{\sum_i X_i^2 - n\overline{X}^2} = \dfrac{\sum_i (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_i (X_i - \overline{X})^2}}$

Dương Nam Hà (VNUA)

IV. Ước lượng… (tiếp)

- Ví dụ:
  - Ước lượng b_1, b_2?
  - 3 cách?

Thu nhập (Xi)	Tiêu dùng (Yi)
18	16
21	17
24	23
27	24
30	26
33	27
36	32
39	31
42	36
45	36
48	43
51	45

Dương Nam Hà (VNUA)

IMPACT OF ADVERTISING EXPENDITURE

Firm	Impressions, millions	Expenditure, millions of 1983 dollars
1. Miller Lite	32.1	50.1
2. Pepsi	99.6	74.1
3. Stroh's	11.7	19.3
4. Fed'l Express	21.9	22.9
5. Burger King	60.8	82.4
6. Coca Cola	78.6	40.1
7. McDonald's	92.4	185.9
8. MCI	50.7	26.9
9. Diet Cola	21.4	20.4
10. Ford	40.1	166.2
11. Levi's	40.8	27.0
12. Bud Lite	10.4	45.6
13. ATT/Bell	88.9	154.9
14. Calvin Klein	12.0	5.0
15. Wendy's	29.2	49.7
16. Polaroid	38.0	26.9
17. Shasta	10.0	5.7
18. Meow Mix	12.3	7.6
19. Oscar Meyer	23.4	9.2
20. Crest	71.1	32.4
21. Kibbles 'N Bits	4.4	6.1

Source: http://lib.stat.cmu.edu/DASL/Datafiles/tvadadat.html

IV. Ước lượng… (tiếp)

- 4.4 Đặc tính của hàm HQ
  - Các ƯL OLS dễ tính toán vì dựa trên các lượng quan sát được
  - Là các ƯL điểm
  - Khi có các ƯL OLS thì SRL có các đặc điểm sau:
    - Đường SRL đi qua $(\overline{X}, \overline{Y})$: $\overline{Y} = b_1 + b_2\overline{X}$
    - $\overline{\hat{Y}} = \overline{Y}$
    - $\overline{e} = 0$
    - Cov $(X_i, e_i) = 0$
    - Cov$(\hat{Y}_i, e_i) = 0$

Dương Nam Hà (VNUA)

IV. Ước lượng… (tiếp)

- 4.5 Đặc điểm của tham số ƯL theo pp OLS:
  - Định lí Gauss-Markov (BLUE – Best Linear Unbiased Estimation) dựa trên CLRM
  - Tuyến tính (Linear):
    - $b_2 = \Sigma_i c_i y_i$; trong đó: $c_i = \frac{x_i}{\sum_i x_i^2}$; $y_i = (Y_i - \overline{Y})$
  - Không chệch (Unbiased): $E(b_j) = B_j$
    - Gợi ý: $y_i = B_2 x_i + u_i$
  - Tốt nhất (Best estimator): min *Var(b_j)*
    - *Tính var(b_2)?*

Dương Nam Hà (VNUA)

IV. Ước lượng… (tiếp)

- C/m: Giả sử $b^*_2 = \Sigma_i w_i y_i$
- → $E(b^*_2) = \Sigma_i w_i E(y_i) = \Sigma_i w_i E(B_2 x_i + u_i)$
- $= \Sigma_i w_i B_2 x_i = B_2 \Sigma_i w_i x_i$
- → $\Sigma_i w_i x_i = 1$
- Mặt khác: $var(b^*_2) = var(\Sigma_i w_i y_i) = \Sigma_i w_i^2 var(y_i)$
- Trong đó: $var(y_i) = var(u_i) = \sigma^2$
- → $var(b^*_2) = \sigma^2 \Sigma_i w_i^2 = \sigma^2 \Sigma_i (w_i - c_i + c_i)^2$
- $= \sigma^2 \Sigma_i (w_i - c_i)^2 + \sigma^2 \Sigma_i c_i^2 + 2\sigma^2 \Sigma_i (w_i - c_i) c_i$
- ➔ $var(b^*_2) = \sigma^2 \Sigma_i (w_i - c_i)^2 + var(b_2)$
- Hay $var(b^*_2) \geq var(b_2)$ hay $var(b_2)$ là min

Dương Nam Hà (VNUA)

V. Phân tích phương sai & Kiểm định giả thuyết thống kê

- 5.1 Phương sai & Sai số chuẩn của các b_j
  - $Var(b_1) = \frac{\sigma^2}{n}\frac{\sum_i X_i^2}{\sum_i (X_i-\bar{X})^2} = \frac{\sigma^2}{n^2}\frac{\sum_i X_i^2}{Var(X_i)}$
  - $Var(b_2) = \frac{\sigma^2}{\sum_i x_i^2} = \frac{1}{n}\frac{\sigma^2}{Var(X)}$
  - $se(b_2) = \sqrt{Var(b_2)}$; $se(b_1) = \sqrt{Var(b_1)}$
  - $\sigma^2_u \rightarrow \hat{\sigma}^2_u = S^2 = \frac{\sum_i e_i^2}{n-2}$

Dương Nam Hà (VNUA)

V. Phân tích phương sai… (tiếp)

- 5.2 ƯL khoảng tin cậy & KĐ giả thuyết TK
- 5.2.1 ƯL KTC cho tham số của PRF
  - Dựa trên tham số của SRF (b_j)
  - $P(b_2 - \delta \le B_2 \le b_2 + \delta) = p = (1 - \alpha)$
  - Trong đó: p là mức độ tin cậy; α là mức ý nghĩa thống kê (%)
  - Từ OLS: $b_2 \sim N(B_2, \sigma^2_{b2})$ với $\sigma^2_{b2} = Var(b_2)$
  - Và $Z = \frac{b_2 - B_2}{\sqrt{\frac{\sigma^2}{\sum_i x_i^2}}} \sim N(0,1)$

Dương Nam Hà (VNUA)

V. Phân tích phương sai… (tiếp)

- Tuy nhiên, chỉ biết $\hat{\sigma}_u^2$ → Z ~ t_{n-2}
- $t_{kđ(b2)} = \frac{b_2 - B_2}{Se(b_2)} \sim t_{(n-2)}$ (n-2 bậc tự do)
- Lựa chọn t_c tra bảng ($t_{n-2,\ \alpha/2}$) sao cho:
  - $P(-t_c \le t \le t_c) = 1 - \alpha$
  - → $P\left(-t_c \le \frac{b_2 - B_2}{Se(b_2)} \le t_c\right) = 1 - \alpha$
  - → $P(-t_c Se(b_2) + b_2 \le B_2 \le b_2 + t_c Se(b_2)) = 1 - \alpha$
  - ➔ KTC: $b_2 - t_c Se(b_2) \le B_2 \le b_2 + t_c Se(b_2)$
    - với mức ý nghĩa thống kê α
  - Tương tự với B_1

Dương Nam Hà (VNUA)

V. Phân tích phương sai… (tiếp)

- 5.2.2 Kiểm định GTTK
  - H_0: $B_2 = 0$
  - H_1: $B_2 \ne 0$ (hay H_A)
  - Sử dụng kiểm định t: $t = \frac{b_2 - B_2}{Se(b_2)}$
  - Nếu $B_2 = 0$ thì:
  - $-t_{(n-2,\ \alpha/2)} \le b_2/Se(b_2) \le +t_{(n-2,\ \alpha/2)}$ hay
  - $-t_{(n-2,\ \alpha/2)} \le t_{kđ} \le +t_{(n-2,\ \alpha/2)}$
  - Nguyên tắc:
    - $|t_{kđ}| < t_c$ → chấp nhận H_0, bác bỏ H_1
    - $|t_{kđ}| > t_c$ → bác bỏ H_0 hay chấp nhận H_1

Dương Nam Hà (VNUA)

V. Phân tích phương sai… (tiếp)

- Quan niệm trong kiểm định giả thuyết:
  - Chấp nhận vs. Bác bỏ
  - Sai lầm loại 1: Bác bỏ GT đúng
  - Sai lầm loại 2: Chấp nhận GT sai

V. Phân tích phương sai… (tiếp)

- 5.3 Phân tích phương sai
- 5.3.1 Độ chặt chẽ của MH
  - TSS = ESS + RSS
  - HSXĐ: $R^2 = ESS/TSS = 1 - RSS/TSS$
  - $0 \leq R^2 \leq 1$
  - Chú ý:
    - $R^2 < 0$: MH ƯL không có hằng số
    - R^2 thấp thì đường hồi quy phản ánh Y_i kém (có thể do nhiều nguyên nhân)

V. Phân tích phương sai… (tiếp)

- Mối quan hệ của TSS và Y:

$$TSS = \Sigma_i y_i^2 = \Sigma_i (Y_i - \bar{Y})^2 = \Sigma_i Y_i^2 - n\bar{Y}^2$$

V. Phân tích phương sai… (tiếp)

- 5.3.2 Kiểm định mô hình
  - H_0: $R^2 = 0$
  - H_1: $R^2 \neq 0$
  - Sử dụng kiểm định F: $F_{kđ} = \frac{ESS/1}{RSS/(n-2)} \sim F_{1}^{\lambda, n-2}$
  - Trong đó: λ là mức độ ý nghĩa thống kê
  - Nguyên tắc:
    - $\frac{ESS/1}{RSS/(n-2)} < F_{1}^{\lambda, n-2} \rightarrow$ Chấp nhận H_0
    - $\frac{ESS/1}{RSS/(n-2)} > F_{1}^{\lambda, n-2} \rightarrow$ Bác bỏ H_0, chấp nhận H_1

V. Phân tích phương sai… (tiếp)

- 5.4 Dự báo với MHHQ
- 5.4.1 Dự báo giá trị TB (E(Y) tại X_0)
  - $P(\hat{Y}_0 - Se(\hat{Y}_0)t_{(n-2,\ \alpha/2)} \le B_1 + B_2X_0 \le \hat{Y}_0 + Se(\hat{Y}_0)t_{(n-2,\ \alpha/2)}) = 1 - \alpha$
  - Trong đó: $Var(\hat{Y}_0) = var(b_1 + b_2X_0) =$
    $= var(b_1) + X_0^2var(b_2) + 2X_0cov(b_1,b_2)$
  - Tính $cov(b_1,b_2)$?
    - Gợi ý: $cov(b_1,b_2) = E[(b_1 - E(b_1))(b_2 - E(b_2))]$

$$b_1 = \bar{Y} - b_2\bar{X}$$

V. Phân tích phương sai… (tiếp)

- → kết quả:

$$cov(b_1,b_2) = -\bar{X}E(b_2 - B_2)^2 = -\bar{X}Var(b_2)$$

- Như vậy:

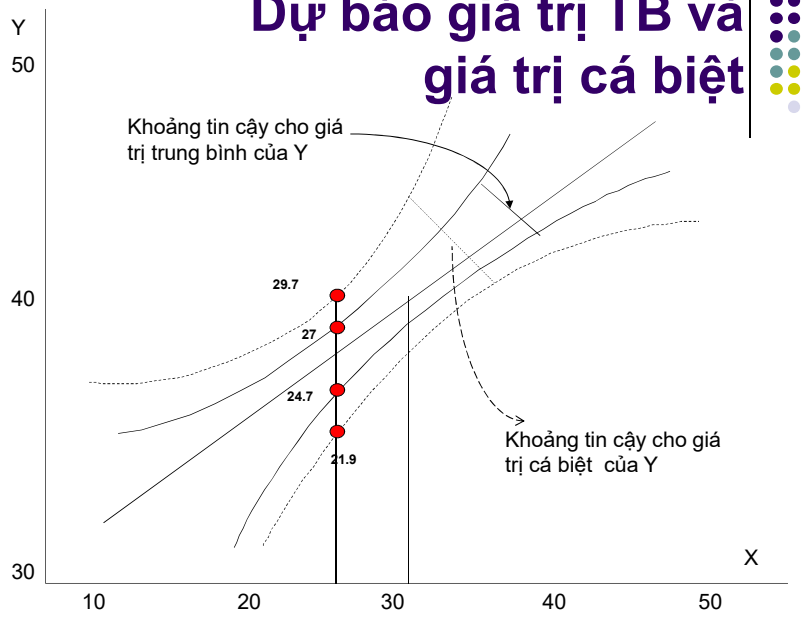
$$Var(\hat{Y}_0) = \sigma^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right)$$

- Đối với mẫu:

$$Se(\hat{Y}_0) = \hat{\sigma}\left(\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} x_i^2}}\right)$$

V. Phân tích phương sai… (tiếp)

- 5.4.2 Dự báo giá trị cá biệt (Y_0)
  - $\hat{Y}_0 - Se(Y_0 - \hat{Y}_0)t_{(n-2,\ \alpha/2)} \le Y_0 \le \hat{Y}_0 + Se(Y_0 - \hat{Y}_0)t_{(n-2,\ \alpha/2)}$
  - Trong đó:
    - $var(Y_0 - \hat{Y}_0) = var(b_1) + X_0^2 var(b_2) + 2X_0 cov(b_1,b_2) + var(u_0)$
  - Đối với mẫu:

$$Se(Y_0 - \hat{Y}_0) = \hat{\sigma}\left(\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} x_i^2}}\right)$$

Dự báo giá trị TB và giá trị cá biệt