Chương 2
MH hồi quy hai biến
Ước lượng và kiểm định giả thuyết

2.1. Phương pháp bình phương bé nhất

Hàm hồi quy mẫu?

- Trong thực tế, ta chỉ có mẫu, ko có tổng thể
- V/đ: đoán tham số tổng thể dựa vào một mẫu của tổng thể (hai tham số tổng thể β_1 và β_2)
- Khái niệm hàm hồi quy mẫu:

2.1 Phương pháp bình phương bé nhất

(Carl Friedrich Gauss- nhà toán học Đức đưa ra)

a. Nội dung

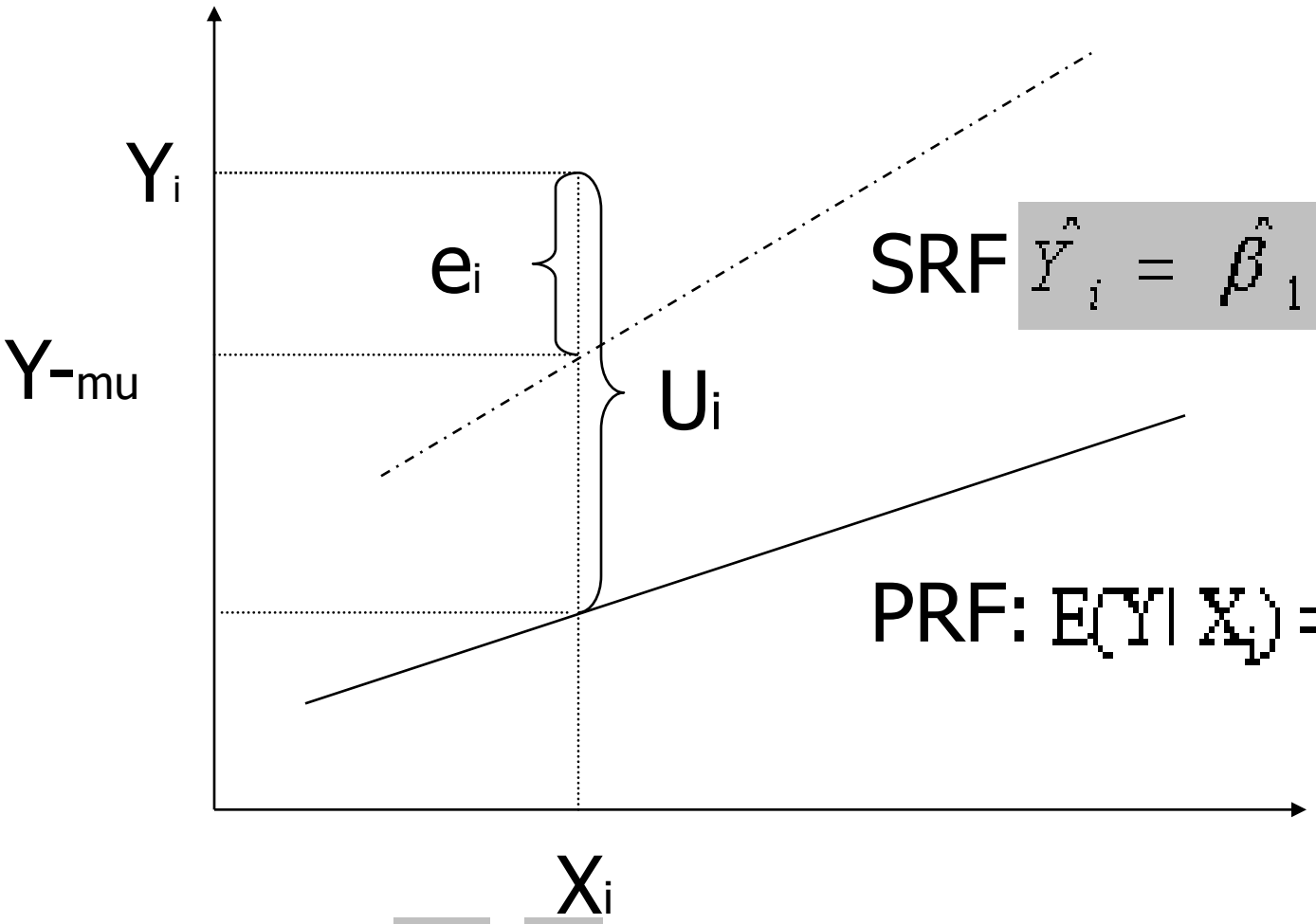
PRF: $E(Y|X_i) = \beta_1 + \beta_2 X_i$

Giá trị quan sát Y_i: $Y_i = E(Y|X_i) + U_i$

$= \beta_1 + \beta_2 X_i + U_i$

SRF: $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

Giá trị quan sát Y_i: $Y_i = \hat{Y}_i + e_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$



SRF $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

PRF: $E(Y|X_i) = \beta_1 + \beta_2 X_i$

V/đ: Tìm $\hat{\beta}_1, \hat{\beta}_2$ gần nhất với β_1, β_2 ?

Ước lượng bình phương bé nhất
(Least Squares Estimation)

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \Rightarrow \min$$

Đã biết

- Cần tìm?

$$\overline{X} = \sum_{i=1}^{n} X_i / n, \quad \overline{Y} = \sum_{i=1}^{n} Y_i / n, \quad x_i = X_i - \overline{X}, \quad y_i = Y_i - \overline{Y}$$

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}, \quad \hat{\beta}_2 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

Kết quả tính bằng phương pháp bình phương bé nhất.

$$\overline{X} = \sum_{i=1}^{n} X_i \Big/ n, \qquad \overline{Y} = \sum_{i=1}^{n} Y_i \Big/ n,$$

$$x_i = X_i - \overline{X}, \qquad y_i = Y_i - \overline{Y}$$

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}, \qquad \hat{\beta}_2 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

b. Tính chất của các ước lượng bình phương nhỏ nhất

1. $\hat{\beta}_1, \hat{\beta}_2$ xác định duy nhất ứng với n cặp quan sát (X_i, Y_i).
2. Ước lượng điểm $\hat{\beta}_1, \hat{\beta}_2$ là ngẫu nhiên.
3.1. SRF đi qua trung bình mẫu $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$.
3.2. Giá trị trung bình của $\hat{Y}_i$ bằng giá trị trung bình của các quan sát.
3.3. Giá trị trung bình của các phần dư bằng 0: $\sum_{i=1}^{n} e_i = 0$

3.4. Các phần dư e_i không tương quan $\hat{Y}_i$: $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$.
3.5. Các phần dư e_i không tương quan $\hat{X}_i$: $\sum_{i=1}^{n} \hat{X}_i e_i = 0$.

Xem trang 63, 64 và Appendix 3A sách Gujarati

2.2. Các giả thiết cơ bản của phương pháp bình phương nhỏ nhất.

- Giả thiết 1: Biến giải thích là phi ngẫu nhiên.
- Giả thiết 2: $E(U_i|X_i) = 0$
- Giả thiết 3: $Var(U_i|X_i) = Var(U_i|X_i) = \sigma^2$
- Giả thiết 4: Không có sự tương quan giữa các U_i $Cov(U_i, U_j) = 0, \quad \forall i \neq j$
- Giả thiết 5: U_i, X_i không tương quan nhau.

$$Cov(U_i, X_i) = 0$$

Chú ý quan trọng từ phần xác suất

Nếu mẫu ngẫu nhiên cỡ n rút ra từ tổng thể vô hạn với trung bình β và phương sai σ^2

Thì

$$E(\bar{Y}) = \beta$$

and variance

$$\text{var}(\bar{Y}) = \frac{\sigma^2}{n}.$$

2.3. Độ chính xác của các ước lượng bình phương nhỏ nhất.

$$Var(\hat{\beta}_2)=\frac{\sigma^2}{\sum_{i=1}^{n}x_i^2}\quad;\qquad Se(\hat{\beta}_2)=\frac{\sigma}{\sqrt{\sum_{i=1}^{n}x_i^2}}$$

$$Var(\hat{\beta}_1)=\frac{\sum_{i=1}^{n}X_i^2}{n\sum_{i=1}^{n}x_i^2}\sigma^2\;;\qquad Se(\hat{\beta}_1)=\sqrt{\frac{\sum_{i=1}^{n}X_i^2}{n\sum_{i=1}^{n}x_i^2}}\;\sigma$$

σ^2 Được ước lượng bằng ước lượng không chệch của nó:

$$\hat{\sigma}^2=\frac{\sum_{i=1}^{n}e_i^2}{n-2}\;;\qquad \hat{\sigma}=\sqrt{\frac{\sum_{i=1}^{n}e_i^2}{n-2}}$$

Định lý Gauss - Markov:

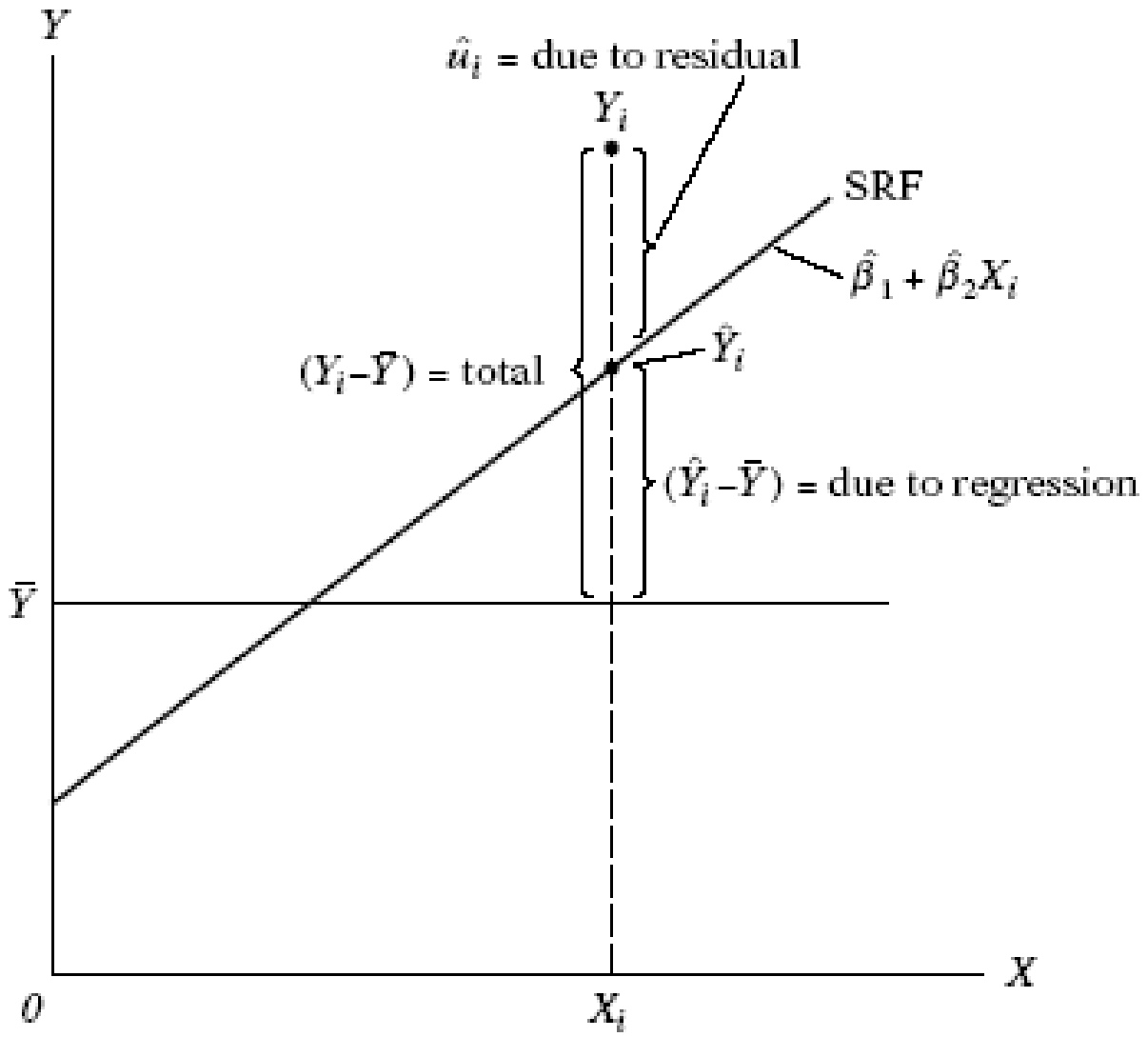
- Với các giả thiết 1-5 của phương pháp bình phương bé nhất, các ước lượng bình phương nhỏ nhất là các ước lượng tuyến tính, không chệch và có phương sai nhỏ nhất trong lớp các ước lượng tuyến tính không chệch.

  (C/m: xem trang 101-106 Gujarati)

(*Phương pháp ước lượng hợp lý tối đa đ/v hàm tuyến tính cũng cho ta kết quả như vậy, nhưng về mặt trực quan và mặt toán học phức tạp hơn OLS*)

2.4. Hệ số r^2 đo độ phù hợp của hàm hồi quy mẫu SRF

- Sơ đồ ven.



- Một số KN

TSS=ESS+RSS

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

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

ESS= $\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{\hat{Y}})^2 = \sum (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_2^2 \sum x_i^2$

$$1 = \frac{\text{ESS}}{\text{TSS}} + \frac{\text{RSS}}{\text{TSS}}$$

$$= \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} + \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2}$$

$$r^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\text{ESS}}{\text{TSS}}$$

$$r^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2}$$

$$= 1 - \frac{\text{RSS}}{\text{TSS}}$$

- Ý nghĩa r^2

Tính chất r^2

- r^2 không âm (mô hình 2 biến có hệ số chặn).
- $0 \leq r^2 \leq 1$

+ Nếu $r^2 = 1$ thì MH hoàn hảo $\hat{Y}_i = Y_i$

+ Nếu $r^2 = 0$ thì không có tương quan giữa biến phụ thuộc và biến giải thích ($\hat{\beta}_2 = 0$).

- Các tính chất của hệ số tương quan r

(tr38 KTL, page 86 Guarati)

2.5. Phân bố xác suất của U_i

- Giả thiết 6: U_i có phân bố $N(0, \sigma^2)$

Các ước lượng OLS $\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2$ có các tính chất:

1. Không chệch.
2. Phương sai cực tiểu.
3. Khi số quan sát đủ lớn, các ƯL đó xấp xỉ với giá trị thực của phân bố.
4. $\hat{\beta}_1$ có phân bố chuẩn:

$$\text{Mean:} \qquad E(\hat{\beta}_1) = \beta_1$$

$$\text{var}(\hat{\beta}_1): \qquad \sigma^2_{\hat{\beta}_1} = \frac{\sum X_i^2}{n \sum x_i^2} \sigma^2$$

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0, 1)$$

5. $\hat{\beta}_2$ có phân bố chuẩn:

$$\text{Mean:} \qquad E(\hat{\beta}_2) = \beta_2$$

$$\text{var}\,(\hat{\beta}_2): \qquad \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum x_i^2}$$

$$\hat{\beta}_2 \sim N(\beta_2, \sigma^2_{\hat{\beta}_2})$$

$$Z = \frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}} \sim N(0, 1)$$

6. $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$

7. Trong các ước lượng không chệch của β_1, β_2 (có thể tuyến tính hoặc không), $\hat{\beta}_1, \hat{\beta}_2$ co phương sai bé nhất.

8. Y_i có phân bố chuẩn:

$$Y_i \sim N(\beta_1 + \beta_2 X_i, \sigma^2)$$

2.6. Khoảng tin cậy và kiểm định giả thiết về các hệ số hồi quy.

- 1. Khoảng tin cậy β_2 , β_1

$$t = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} \sim T(n-2)$$

- df=n-2

$$\Pr\left(-t_{\alpha/2} \leq t \leq t_{\alpha/2}\right) = 1 - \alpha$$

$$\Pr\left[-t_{\alpha/2} \leq \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} \leq t_{\alpha/2}\right] = 1 - \alpha$$

$$\Pr\left[\hat{\beta}_2 - t_{\alpha/2}\, se(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2}\, se(\hat{\beta}_2)\right] = 1 - \alpha$$

Vậy $$\hat{\beta}_2 - t_{\alpha/2}\, se(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2}\, se(\hat{\beta}_2)$$

Tương tự:

- df=n-2

$$\Pr\left[\hat{\beta}_1 - t_{\alpha/2}\,\mathrm{se}\,(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2}\,\mathrm{se}\,(\hat{\beta}_1)\right] = 1-\alpha$$

Vậy:

$$\hat{\beta}_1 - t_{\alpha/2}\,\mathrm{se}\,(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2}\,\mathrm{se}\,(\hat{\beta}_1)$$

- 2. Khoảng tin cậy σ^2
- df=n-2

$$\chi^2 = (n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

$$\Pr\left(\chi^2_{1-\alpha/2} \le \chi^2 \le \chi^2_{\alpha/2}\right) = 1-\alpha$$

$$\Pr\left[(n-2)\frac{\hat{\sigma}^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le (n-2)\frac{\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}\right] = 1-\alpha$$

- Vậy $$(n-2)\frac{\hat{\sigma}^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le (n-2)\frac{\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}$$

- 3. Kiểm định giả thiết:
  - df=n-2, t= $\dfrac{\hat{\beta}_2 - \beta_2^*}{se\,(\hat{\beta}_2)}$

THE *t* TEST OF SIGNIFICANCE: DECISION RULES

Type of hypothesis	H_0: the null hypothesis	H_1: the alternative hypothesis	Decision rule: reject H_0 if
Two-tail	$\beta_2 = \beta_2^*$	$\beta_2 \neq \beta_2^*$	$\lvert t\rvert > t_{\alpha/2,df}$
Right-tail	$\beta_2 \le \beta_2^*$	$\beta_2 > \beta_2^*$	$t > t_{\alpha,df}$
Left-tail	$\beta_2 \ge \beta_2^*$	$\beta_2 < \beta_2^*$	$t < -t_{\alpha,df}$

Notes: β_2^* is the hypothesized numerical value of β_2.
$|t|$ means the absolute value of t.
t_α or $t_{\alpha/2}$ means the critical t value at the α or $\alpha/2$ level of significance.
df: degrees of freedom, $(n-2)$ for the two-variable model, $(n-3)$ for the three-variable model, and so on.

(Kđ giả thiết về β_1 tương tự)

A SUMMARY OF THE χ^2 TEST

- df=n-2 , $\chi^2 = (n-2)\dfrac{\hat{\sigma}^2}{\sigma^2}$

H_0: the null hypothesis	H_1: the alternative hypothesis	Critical region: reject H_0 if
$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\dfrac{\text{df}(\hat{\sigma}^2)}{\sigma_0^2} > \chi^2_{\alpha,\text{df}}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\dfrac{\text{df}(\hat{\sigma}^2)}{\sigma_0^2} < \chi^2_{(1-\alpha),\text{df}}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\dfrac{\text{df}(\hat{\sigma}^2)}{\sigma_0^2} > \chi^2_{\alpha/2,\text{df}}$ or $< \chi^2_{(1-\alpha/2),\text{df}}$

Note: σ_0^2 is the value of σ^2 under the null hypothesis. The first subscript on χ^2 in the last column is the level of significance, and the second subscript is the degrees of freedom. These are critical chi-square values. Note that df is $(n-2)$ for the two-variable regression model, $(n-3)$ for the three-variable regression model, and so on.

2.7. Kđ sự phù hợp của hàm hồi quy, phân tích hồi quy và phân tích phương sai

$$F = \left(\frac{(\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n} x_i^2}{\hat{\sigma}^2} / 1 \right) : \left(\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} / (n-2) \right) = \frac{(\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n} x_i^2}{\hat{\sigma}^2}$$

$\sim F(1, n-2)$

Chúng ta Kđ cặp giả thiết:

- $H_0: \beta_2 = 0$
- $H_1: \beta_2 \neq 0$

$$F = \frac{\hat{\beta}_2^2 \sum_{i=1}^{n} x_i^2}{\hat{\sigma}^2} > F_\alpha(1, n-2)$$

Bác bỏ giả thiết H_0

- Mặt khác

$$F = \frac{\hat{\beta}_2^2 \sum_{i=1}^{n} x_i^2}{\hat{\sigma}^2} = \frac{ESS/1}{RSS/(n-2)} = \frac{TSS \cdot r^2/1}{(1-r^2)\,TSS/(n-2)} = \frac{r^2}{(1-r^2)} \times \frac{n-2}{1}$$

Do đó quá trình phân tích phương sai cho phép ta phán đoán thống kê về độ thích hợp hàm hồi quy.

- bác bỏ giả thiết: H_0: $\beta_2 = 0$

- tương đương bác bỏ giả thiết H_0: $r^2 = 0$

Chú ý: ANOVA xét hàm 2 biến có hệ số chặn.

2.8. Phân tích hồi quy và dự báo

Hai loại dự báo

a- Dự báo giá trị trung bình $E(Y| X_0)$.

b- Dự báo giá trị cá biệt của Y với $X = X_0$.

$$\text{var}(\hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right]$$

$$t = \frac{\hat{Y}_0 - (\beta_1 + \beta_2 X_0)}{\text{se}(\hat{Y}_0)} \sim T(n-2)$$

$$\Pr[\hat{\beta}_1 + \hat{\beta}_2 X_0 - t_{\alpha/2}\,\text{se}(\hat{Y}_0) \le \beta_1 + \beta_2 X_0 \le \hat{\beta}_1 + \hat{\beta}_2 X_0 + t_{\alpha/2}\,\text{se}(\hat{Y}_0)] = 1 - \alpha$$

$$\hat{Y}_0 - t_{\alpha/2} Se(\hat{Y}_0) \le E(Y \mid X_0) \le \hat{Y}_0 + t_{\alpha/2} Se(\hat{Y}_0)$$

Var(Y_0)= $$\text{var}(Y_0 - \hat{Y}_0) = E[Y_0 - \hat{Y}_0]^2 = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right]$$

$$t = \frac{Y_0 - \hat{Y}_0}{Se(Y_0)} = \frac{Y_0 - \hat{Y}_0}{se(Y_0 - \hat{Y}_0)} \sim T(n-2)$$

- Khoảng tin cậy của Y_0 được xác định bởi:

$$P[\hat{Y}_0 - t_{\alpha/2} Se(Y_0) \le Y_0 \le \hat{Y}_0 + t_{\alpha/2} Se(Y_0)] = 1 - \alpha$$

- Bài tập: