ĐẠI HỌC THĂNG LONG
Khoa Toán - Tin

ĐỀ CƯƠNG

Kinh tế lượng

Sinh viên thực hiện: Nguyễn Tú Anh
MSV: A29888

HÀ NỘI - 2019

Mục lục

I Mô hình hồi quy tuyến tính 3

1 Một số hàm cơ bản trong R 3
1.1 Hệ số xác định R^2 3
1.2 Khoảng tin cậy - confint 3
1.3 Dự báo - predict 3
1.4 Lấy phần dư trong hồi quy tuyến tính - resid 4
1.5 Tìm $\hat{Y}$ - fitted 4

2 Hệ thức F và kiểm định đồng thời 4

3 Kiểm định ý nghĩa kinh tế 5

4 Kiểm định giả thuyết về một ràng buộc giữa các hệ số hồi qui 5

II Một số dạng mô hình hồi quy thường gặp 7

1 Mô hình phi tuyến 7
1.1 log - log 7
1.2 log - lin 7
1.3 lin-log 7

2 Mô hình hồi quy với biến giả 7
2.1 Một biến lượng và một biến chất 7

III Vi phạm giả thuyết 8

1 Đa cộng tuyến 9
1.1 Khái niệm 9
1.2 Dấu hiệu 9
1.3 Khắc phục 9

2 Phương sai thay đổi 10
2.1 Khái niệm 10
2.2 Dấu hiệu 10
2.3 Phương pháp định lượng 10
2.4 Khắc phục 12

3 Tự tương quan 13
3.1 Khái niệm 13
3.2 Phát hiện 13
3.3 Khắc phục 14

IV Lựa chọn mô hình 16

1 Các bước lựa chọn mô hình 16

2 So sánh 2 mô hình - Kiểm định Wald 16

3 Kiểm định sự bỏ sót biến - Reset 17

Phần I
Mô hình hồi quy tuyến tính

1 Một số hàm cơ bản trong R

1.1 Hệ số xác định R^2

$R^2 = \alpha\%$ và $\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$

Ý nghĩa

$\alpha\%$ sự biến động của Y được giải thích qua sự biến động của X.

1.2 Khoảng tin cậy - confint

Trong R

Khoảng tin cậy 95%:
$> confint(lm(Y \sim X), level = 0.95)$
Hoặc $b_i - t_{n-k,\alpha/2}se(b_i) < B_i < b_i + t_{n-k,\alpha/2}se(b_i)$

Ý nghĩa

Khi X tăng 1 đơn vị thì Y trung tăng tối thiểu α đơn vị, tăng tối đa β đơn vị (Với α, β là cận dưới và cận trên của B_2 với mô hình tuyến tính $Y \sim X$)

1.3 Dự báo - predict

- Ước lượng điểm tại $X = 28$:
  $> predict(lm(Y \sim X), new = data.frame(X = 28))$
- Tìm khoảng dự báo 90% cho Y **cá biệt** tại $X = 28$:
  $> predict(lm(Y \sim X), newdata = data.frame(X = 28), \textbf{interval} = "\textbf{p}", level = 0.9)$
- Tìm khoảng dự báo 90% cho Y **trung bình** tại $X = 28$:
  $> predict(lm(Y \sim X), newdata = data.frame(X = 28), \textbf{interval} = "\textbf{c}", level = 0.9)$

1.4 Lấy phần dư trong hồi quy tuyến tính - resid

$> e = resid(lm(Y \sim X_2 + X_3 + .. + X_k))$

1.5 Tìm $\hat{Y}$ - fitted

$> fitted(lm(Y \sim X))$
Hoặc: $\hat{Y} = Y - resid(lm(Y \sim X))$

2 Hệ thức F và kiểm định đồng thời

Xét mô hình: $Y_i = B_1 + B_2X_{2i} + ... + B_kX_{ki} + U_i$ với $i = \overline{2, n}$

Ý nghĩa

Kiểm định đồng thời các hệ số bằng 0.

Bảng kiểm định đồng thời

H_0	H_1	F	Bác bỏ H_0	$p - value$
$B_2 = ... = B_k = 0$	$\exists B_i \neq 0$	$F = \frac{n-k}{k-1}\frac{R^2}{1-R^2}$	$F > F_{k-1,n-k,\alpha}$	$P(F_{k-1,n-k,\alpha} > F)$

Trong đó: $F_{k-1,n-k,\alpha} = qf(1-\alpha, k-1, n-k)$

Bậc

F tuân theo phân phối Fisher với $k - 1$ bậc tự do ở tử và $n - k$ bậc tự do ở mẫu.

Cách tính

Đã được tính sẵn trong hàm summary(lm())

3 Kiểm định ý nghĩa kinh tế

Kiểm định giả thuyết cho B_i

H_0	H_1	t	Bác bỏ H_0	$p-value$
$B_i \leq B_i^*$	$B_i > B_i^*$	$t = \frac{b_i - B_i^*}{se(B_i)}$	$t > t_{n-k,\alpha}$	$P(t_{n-k} > t)$
$B_i \geq B_i^*$	$B_i < B_i^*$	$t = \frac{b_i - B_i^*}{se(B_i)}$	$t < -t_{n-k,\alpha}$	$P(t_{n-k} < t)$
$B_i = B_i^*$	$B_i \neq B_i^*$	$t = \frac{b_i - B_i^*}{se(B_i)}$	$\lvert t\rvert > t_{n-k,\alpha/2}$	$2P(t_{n-k} > \lvert t\rvert)$

Với $se(B_i)$ xem ở hàm summary(), n là số quan sát, k là số biến.
$t_{n-k,\alpha} = qt(1-\alpha, n-k)$ (Trong R)

4 Kiểm định giả thuyết về một ràng buộc giữa các hệ số hồi qui

Cặp giả thuyết

H_0	H_1
$\alpha B_i + \beta B_j \leq \gamma$	$\alpha B_i + \beta B_j > \gamma$
$\alpha B_i + \beta B_j \geq \gamma$	$\alpha B_i + \beta B_j < \gamma$
$\alpha B_i + \beta B_j = \gamma$	$\alpha B_i + \beta B_j \neq \gamma$

Cách giải

Bài toán kiểm định về tổ hợp các hệ số hồi quy có dạng $AB = b$
Với bài toán kiểm định ở trên ta có:

B = $\begin{bmatrix} B_1 \\ ... \\ B_i \\ ... \\ B_j \\ ... \\ B_k \end{bmatrix}$, A = $\begin{bmatrix} 0 & ... & \alpha & ... & \beta & ... & 0 \end{bmatrix}$ và b = γ

Hàm glht()

- Gọi thư viện: **multcomp**
- **glht(model, linfct, rhs = 0, alt = c("t","l","g"))**

- model: là mô hình hồi quy
- linfct: là ma trận A
- rhs: là vector b
- alt: dạng **giả thuyết đối**

Ví dụ

Với mô hình $Y = B_1 + B_2X_2 + B_3X_3 + U$
Kiểm định $\alpha B_2 + \beta B_3 \leq \gamma$

Thực hiện lệnh:
$>$ library(multcomp)
$>$ A = matrix(c(0, α, β), nrow = 1)
$>$ summary(glht(lm(Y $\sim$ X1 + X2), linfct = A, rhs = γ,alt = "g"))

Phần II
Một số dạng mô hình hồi quy thường gặp

1 Mô hình phi tuyến

1.1 log - log

- Ví dụ: $\log(y) = B_1 + B_2 \log(X)$ (B_2 là Hệ số co giãn)
- Ý nghĩa B_2: Khi X tăng 1% thì Y tăng B_2%

1.2 log - lin

- Ví dụ: $\log(y) = B_1 + B_2 X$ (B_2 là Tốc độ tăng trưởng)
- Ý nghĩa B_2: Khi X tăng 1 đơn vị thì Y tăng B_2.100% đơn vị

1.3 lin-log

- Ví dụ: $Y = B_1 + B_2 \log(X)$
- Ý nghĩa B_2: Khi X tăng 1% thì Y tăng 0.01*B_2 đơn vị

2 Mô hình hồi quy với biến giả

2.1 Một biến lượng và một biến chất

Bài toán

Xây dựng mô hình hồi quy dự báo chi phí của khách du lịch theo thời gian du lịch và có xét đến quốc tịch - (Pháp, Mỹ, Trung Quốc, Việt Nam).
Giả sử mô hình có dạng: $Y = \alpha + \beta T + U$

Giải - Tổng quát

- Mã hóa biến quốc tịch theo bảng sau:

Quốc tịch	D_2	D_3	D_4
Pháp	1	0	0
Mỹ	0	1	0
Trung Quốc	0	0	1
Việt Nam	0	0	0

- Chi phí ban đầu khác nhau, tốc độ chi tiêu khác nhau (Tổng quát) nên ta đặt:
  $\alpha = B_1 + B_2D_2 + B_3D_3 + B_4D_4$
  $\beta = B_1^* + B_2^*D_2 + B_3^*D_3 + B_4^*D_4$

- Khi đó phương trình hồi quy có dạng:
  $Y = B_1 + B_2D_2 + B_3D_3 + B_4D_4 + B_1^*T + B_2^*D_2T + B_3^*D_3T + B_4^*D_4T + U$

- Tách riêng từng phương trình:
  - Việt Nam: $Y = B_1 + B_1^*T$
  - Pháp: $Y = (B_1 + B_2) + (B_1^* + B_2^*)T$
  - Mỹ: $Y = (B_1 + B_3) + (B_1^* + B_3^*)T$
  - TQ: $Y = (B_1 + B_4) + (B_1^* + B_4^*)T$

- Ý nghĩa các hệ số tiêu biểu:
  - B_1: Khi T = 0, Chi phí trung bình của một khách VN là B_1
  - B_1^*: Khi T tăng 1 đơn vị, Chi phí trung bình của một khách VN tăng B_1^* đơn vị
  - B_2: Khi T =0, Chi phí trung bình của một khách Pháp chênh lệch với 1 khách VN là B_2 đợn vị
  - B_2^*: Khi T tăng 1 đơn vị, Độ thay đổi của một khách Pháp lớn hơn khách VN là B_2^*

Phần III
Vi phạm giả thuyết

1 Đa cộng tuyến

1.1 Khái niệm

Các biến độc lập trong mô hình có tương quan.

1.2 Dấu hiệu

- **R^2 cao và p_{value} cao**
- **Hệ số tương quan cao giữa các biến độc lập**
- **Sử dụng mô hình hồi quy phụ** Hồi quy biến X_i theo các biến còn lại. Sử dụng thống kê F để kiểm định sự tương quan(Tính sẵn khi ta dùng hàm summary để hồi quy).

1.3 Khắc phục

Giả sử mô hình có dạng: $Y = B_1 + B_2X_2 + B_3X_3 + U$

- **Sử dụng thông tin tiên nghiệm**. Ví dụ: $B_2 = \alpha B_3$
- **Bỏ bớt biến độc lập** Ưu tiên giảm đa cộng tuyến đến R^2 cao.
- **Sai phân cấp 1** Dùng cho dữ liệu chuỗi thời gian. Lấy hiệu 2 phương trình tại thời điểm t và t-1.

  1 Lấy sai phân: $> Y^* = diff(Y); X_2^* = diff(X_2); X_3^* = diff(X_3)$

  2 Kiểm tra ĐCT: $> cor(data.frame(X_2^*, X_3^*))$

  3 Ước lượng MH mới: $lm(Y^* \sim \mathbf{0} + X_2^* + X_3^*)$

2 Phương sai thay đổi

2.1 Khái niệm

Phương sai của các sai số ngẫu nhiên U_i ứng với giá trị X_i là khác nhau, tức là:

$$var(U_i) = E(U_i)^2 = \sigma_i^2, \exists j : \sigma_i \neq \sigma_j$$

thì ta nói sai số ngẫu nhiên có phương sai thay đổi.

2.2 Dấu hiệu

- Dựa vào bản chất của vấn đề nghiên cức.
- Phương pháp đồ thị. Vẽ đồ thị ánh xạ của phần dư e_i^2 theo một biến độc lập nào đó hay theo Y
- Phương pháp định lượng.

2.3 Phương pháp định lượng

Kiểm định Park

Các bước tiến hành:

1 : Hồi qui gốc để tìm e: $e = resid(lm(Y \sim X_2 + ... + X_k))$

2 : Hồi qui: $ln(e^2) = A_1 + A_2 ln(X_i) + V$ với $i = \overline{2..k}$

3 : Mỗi hồi quy với **i** ở bước [2] ta tiến hành kiểm định cặp giả thuyết:

  - H_0: $A_2 = 0$ hay phương sai không đổi với biến X_i theo Park.
  - H_1: $A_2 \neq 0$ hay phương sai thay đổi.

Kiểm định Glejser

Giống kiểm định Park nhưng sử dụng 4 hàm sau (Ở bước 2 trong Park):

- $|e| = \alpha_1 + \alpha_2 X_i + V$
- $|e| = \alpha_1 + \alpha_2 \sqrt{X_i} + V$

- $|e| = \alpha_1 + \alpha_2 \frac{1}{X_i} + V$
- $|e| = \alpha_1 + \alpha_2 \frac{1}{\sqrt{X_i}} + V$

Kiểm định White

Giả sử xét mô hình: $Y = B_1 + B_2X_2 + B_3X_3 + U$ [*]

1 : Hồi quy mô hình gốc để lấy phần dư e;

2 : Hồi quy mô hình phụ:

$$e^2 = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_2^2 + \alpha_5 X_3^2 + \alpha_6 X_2 X_3 + V$$

3 : Kiểm định cặp giả thuyết:

- $H_0 : \alpha_2 = \alpha_3 = ... = \alpha_6 = 0$ Hay phương sai không đổi với White.
- $H_1 : \exists i : \alpha_i \neq 0$ hay phương sai thay đổi.

4 : Thực hiện trên R:
$> library(lmtest)$
$> bptest(lm(Y \sim X_2 + X_3), \sim X_2 +\ X_3 + I(X_2^2) + I(X_3^2) + I(X_2X_3))$

Kiểm định Goldfeld -Quandt

Cặp giả thuyết:

- H_0: Phương sai không đổi theo G-Q.
- H_1: Phương sai thay đổỉ.

Thực hiện trên R bằng hàm **gqtest**
Ví dụ với mô hình [*]: $> gqtest(Y \sim X_2 + X_3, fraction = 4, order.by = \sim X_2)$

2.4 Khắc phục

Xét mô hình: $Y = B_1 + B_2X_2 + U$ [**]

Đã biết phương sai tổng thể

- Giả sử $var(U) = \sigma^2$ đã biết.
- Chia cả 2 vế cho σ có:

$$\frac{Y}{\sigma} = B_1\frac{1}{\sigma} + B_2\frac{X_2}{\sigma} + \frac{U}{\sigma}$$

- Ta được phương trình viết gọn: $Y^* = B_1X_1^* + B_2X_2^* + U^*$
- Phương trình không còn phương sai thay đổi nữa vì $var(U^*) = 1$ (bất biến).

Chưa biết phương sai tổng thể

- **Giả thiết 1**: $var(U) = \alpha^2X_i^2$ (α đã biết)
  - Chia cả 2 vế cho X_i.
  - Khắc phục được vì mô hình mới có $var(U) = \alpha^2$ (bất biến)
- **Giả thiết 2**: $var(U) = \alpha^2X_i$ (α đã biết)
  - Chia cả 2 vế cho $\sqrt{X_i}$.
  - Khắc phục được vì mô hình mới có $var(U) = \alpha^2$ (bất biến)
- **Sử dụng phép biến đổi logarit**: để **giảm** mức độ phương sai thay đổi.
  - Sử dụng mô hình thay thế: $\ln(Y) = B_1 + B_2\ln(X) + V$.
  - ln làm giảm sự cách biệt giữa các giá trị.

3 Tự tương quan

3.1 Khái niệm

Tự tượng quan là: $\exists i, j, i \neq j : cov(U_i, U_j) \neq 0$

3.2 Phát hiện

Xét mô hình: $Y_t = B_1 + B_2X_{2t} + U_t$

Phương pháp đồ thị

Ta có thể vẽ 2 loại đồ thị sau:

- e_t (hoặc e_t^2) theo thời gian.
- e_t theo e_{t-1}

Nếu đồ thị ngẫu nhiên thì không có tư tương quan.

Kiểm định Durbin - Watson

Hồi quy phụ: $U_t = \rho U_{t-1} + V_t$ (Tự tương quan bậc nhất)

- Cặp giả thuyết:

  H_0 : Không có tự tương quan - $\rho = 0$

  H_1 : Có tự tương quan **dương** - $\rho > 0$

- Thực hiện trên R: $> dwtest(lm(Y \sim X_2), alt = "g")$
  ($alt = "g"$ vì kiểm định $\rho > 0$)

- Rút p_{value}, kết luận.

Kiểm định Breusch - Godfrey

Hồi quy phụ: $U_t = \rho_1 U_{t-1} + \rho_2 U_{t-2} + V_t$ (Tự tương quan bậc 2)

- Cặp giả thuyết:

$$H_0 \text{ : Không có tự tương quan - } \rho_1 = \rho_2 = 0$$
$$H_1 \text{ : Có tự tương quan - } \rho_1^2 + \rho_2^2 \neq 0 > 0$$

- Thực hiện trên R:
  $> library(lmtest)$
  $> bgtest(lm(Y \sim X_2), order = 2, type = "Chisq")$
  Trong đó:
  - **order** là bậc của tự tương quan.
  - **type = c("Chisq","F")** là kiểu thống kê kiểm định theo khi bình phương hay thống kê F.

- Rút p_{value}, kết luận.

3.3 Khắc phục

Xét mô hình: $Y_t = B_1 + B_2 X_t + U_t$
Giả sử $U_t = \rho U_{t-1} + V_t$

Biết cấu trúc tự tương quan

Nếu biết ρ ta biến đổi mô hình như sau:

- Ta có: $\rho Y_{t-1} = \rho B_1 + \rho B_2 X_{t-1} + \rho U_{t-1}$
- Suy ra: $Y_t - \rho Y_{t-1} = B_1(1-\rho) + B_2(X_t - \rho X_{t-1}) + V_t$
- Đặt:
  - $B_1^* = B_1(1-\rho)$
  - $B_2^* = B_2$
  - $Y_t^* = Y_t - \rho Y_{t-1}$
  - $X_t^* = X_t - \rho X_{t-1}$
- Ta được mô hình mới: $Y_t^* = B_1^* + B_2^* X_t^* + V_t$

Hồi quy theo mô hình mới ta sẽ thu được các tham số của mô hình gốc

Chưa biết cấu trúc tự tương quan

Nếu chưa biết ρ thì ta sẽ tìm ρ rồi làm tương tự như trường hợp đã biết bằng các cách sau:

- Công thức phần dư:
$$\rho = \frac{\sum e_t e_{t-1}}{\sum e_t^2}$$
- Sử dụng kiểm định Durbin Watson: $\rho = 1 - DW/2$
- Ước lượng qua mô hình hồi quy qua gốc e_t theo $e_{t-1} :> lm(e_t \sim e_{t-1}+0)$

Phần IV
Lựa chọn mô hình

1 Các bước lựa chọn mô hình

1 : Xác định số biến đọc lập trong mô hình.

- Bổ sung biến độc lập, xem có bỏ sót biến quan trọng không.
- Thanh lọc các biến không quan trọng.

2 : Chọn dạng hàm hồi quy dựa trên kiến thức chuyên ngành hoặc khảo sát mỗi quan hệ.

3 : Kiểm tra vi phạm giả thuyết và khắc phục chúng.

4 : Áp dụng tiêu chuẩn đánh giá mô hình như:

- Tiêu chuẩn R^2 hoặc $\overline{R}^2$
- Tiêu chuẩn log-likelihood - logLik(lm()): Càng lớn càng phù hợp.
- Tiêu chuẩn AIC - AIC(lm()): Càng nhỏ càng phù hợp.
- Tiêu chuẩn BIC - BIC(lm()): Càng nhỏ càng phù hợp.

2 So sánh 2 mô hình - Kiểm định Wald

Bài toán

So sánh 2 mô hình tuyến tính (m < k):

$$Y = B_1 + B_2X_2 + B_3X_3 + ... + B_kX_k + U.$$

$$Y = B_1 + B_2X_2 + B_3X_3 + ... + B_mX_m + U.$$

Cặp giả thuyết

- $H_0 : B_{m+1} = B_{m+2} = ... = B_k = 0$
- $H_1 : \exists B_i \neq 0, i \in m+1, m+2, ..., k$

Nếu chấp nhận H_0 thì 2 phương trình là như nhau về khả năng dự báo giá trị của biên phụ thuộc.
Nếu bấc bỏ H_0 thì mô hình nhiều biến độc lập hơn sẽ tốt hơn.

Lệnh trong R

$> waldtest(lm(Y \sim X_2 + X_3 + ... + X_k), lm(Y \sim X_2 + X_3 + ... + X_m))$

3 Kiểm định sự bỏ sót biến - Reset

Xét mô hình: $Y = B_1 + B_2X + U$

Cặp giả thuyết

- H_0: Không có hiện tượng bỏ sót biến độc lập.
- H_1: Có hiện tượng bỏ sót biến độc lập.

Lệnh trong R

$> resettest(lm(Y \sim X), power = 2:3, type = "fitted")$