THỐNG KÊ ỨNG DỤNG

ĐỖ LÂN

dolan@tlu.edu.vn
Đại học Thủy Lợi

Ngày 15 tháng 10 năm 2018

Nội dung môn học

1. Tổng quan về Thống kê
2. Thu thập dữ liệu
3. Tóm tắt và trình bày dữ liệu bằng bảng và đồ thị
4. Tóm tắt dữ liệu bằng các đại lượng thống kê mô tả
5. Xác suất căn bản và biến ngẫu nhiên
6. Phân phối của tham số mẫu và ước lượng tham số tổng thể
7. Kiểm định giả thuyết về tham số một tổng thể
8. Kiểm định giả thuyết về tham số hai tổng thể
9. Phân tích phương sai
10. Kiểm định phi tham số
11. Kiểm định chi - bình phương
12. Hồi quy đơn biến
13. **Hồi quy đa biến**

Phần XI

Hồi quy đa biến

Mục lục

Nội dung chính được trình bày trong chương

- Trình bày định nghĩa mô hình hồi quy tuyến tính đa biến tổng thể;
- Giới thiệu phương pháp bình phương tối thiểu tìm phương trình hồi quy tuyến tính mẫu;
- Trình bày các kiến thức liên quan đến hồi quy: Hệ số xác định, hệ số xác định hiệu chỉnh, sai số chuẩn của ước lượng, kiểm định về hệ số độ dốc và khoảng tin cậy cho hệ số độ dốc;
- Trình bày bài toán dự báo giá trị của biến phụ thuộc theo các biến độc lập

Những kiến thức sinh viên phải nắm được trong chương

- Nắm được định nghĩa mô hình hồi quy tuyến tính đa biến tổng thể;
- Biết cách tìm phương trình hồi quy tuyến tính theo phương pháp bình phương tối thiểu;
- Nắm được các kiến thức liên quan đến hồi quy: Hệ số xác định, hệ số xác định hiệu chỉnh, sai số chuẩn của ước lượng, kỉnh định về hệ số độ dốc và khoảng tin cậy cho hệ số độ dốc;
- Biết cách dự báo giá trị của biến phụ thuộc theo các biến độc lập.

Mô hình hồi quy tuyến tính đa biến

Mô hình hồi quy tuyến tính liên hệ biến phụ thuộc Y theo các biến độc lập $X_1, X_2, ..., X_k$ (gọi là mô hình hồi quy tuyến tính đa biến) có dạng

$$\begin{aligned} Y &= E(Y|x_1, x_2, ...x_k) + \epsilon \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \end{aligned}$$

Mô hình hồi quy tuyến tính đa biến

Mô hình hồi quy tuyến tính liên hệ biến phụ thuộc Y theo các biến độc lập $X_1, X_2, ..., X_k$ (gọi là mô hình hồi quy tuyến tính đa biến) có dạng

$$\begin{aligned} Y &= E(Y|x_1, x_2, ...x_k) + \epsilon \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \end{aligned}$$

trong đó

Mô hình hồi quy tuyến tính đa biến

Mô hình hồi quy tuyến tính liên hệ biến phụ thuộc Y theo các biến độc lập $X_1, X_2, ..., X_k$ (gọi là mô hình hồi quy tuyến tính đa biến) có dạng

$$\begin{aligned} Y &= E(Y|x_1, x_2, ...x_k) + \epsilon \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \end{aligned}$$

trong đó

- $E(Y|x_1, x_2, ...x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ là trung bình của biến phụ thuộc Y khi các biến độc lập nhận giá trị $x_1, x_2, \cdots, x_k$.
- $\beta_0, \beta_1, \cdots, \beta_k$ (chưa biết) là các tham số của mô hình hồi quy:
  1. β_0 là trung bình của Y khi các biến $X_1, X_2, \cdots X_k$ nhận giá trị 0;
  2. β_i là độ thay đổi trung bình của Y khi biến X_i tăng lên một đơn vị và các biến khác giữ nguyên.
- ϵ là phần dư hay yếu tố nhiễu mô tả ảnh hưởng của các yếu tố khác với các biến độc lập đang xem xét tới biến phụ thuộc Y.

Các giả thiết của mô hình

1. Với mọi giá trị $x_1, x_2, \cdots, x_k$ của các biến độc lập, phần dư có phân phối chuẩn, trung bình bằng 0 và phương sai không phụ thuộc vào các giá trị $x_1, x_2, \cdots, x_k$.
2. Các phần dư ứng với các bộ giá trị khác nhau của các biến độc lập là độc lập.

Ví dụ

Trước đây một công ty kiểm toán sử dụng các nhân viên đến các cơ quan để tìm ra phần thuế mà cơ quan đó chưa trả hàng tháng. Gần đây họ sử dụng thêm hệ thống máy tính với hy vọng kết quả sẽ chính xác hơn. Bảng sau đây ghi lại số giờ lao động của các nhân viên (X_1), số giờ dùng máy tính (X_2) và số thuế chưa trả (Y) tìm thấy được trong vòng 10 tháng

Tháng	1	2	3	4	5	6	7	8	9	10
Số giờ LĐ của nhân viên	45	42	44	45	43	46	44	45	44	43
Số giờ dùng máy tính	16	14	15	13	13	14	16	16	15	15
Số thuế chưa trả	29	24	27	25	26	28	30	28	28	27

Ta muốn liên hệ số thuế chưa trả Y khi số giờ lao động của các nhân viên bằng x_1, số giờ dùng máy tính bằng x_2 bởi mô hình tuyến tính:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

1 Mô hình hồi quy tuyến tính đa biến

2 Ước lượng bằng phương pháp bình phương tối thiểu

Ước lượng bằng phương pháp bình phương tối thiểu

- Các tham số của mô hình $\beta_0, \beta_1, \ldots, \beta_k$ chưa biết và được ước lượng bởi các số $b_0, b_1, \ldots, b_k$ xác định từ dữ liệu mẫu.
- Phương trình

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

gọi là phương trình hồi qui tổng thể.

- Phương trình

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

gọi là phương trình hồi qui mẫu.

Ước lượng bằng phương pháp bình phương tối thiểu

- Với một mẫu được chọn ra, gọi y_i và $\hat{y}_i$ là giá trị thật và giá trị dự báo của Y ở quan sát thứ i. Đặt $e_i = y_i - \hat{y}_i$ là phần dư của quan sát thứ i.
- Các giá trị $b_0, b_1, \ldots, b_k$ được xác định sao cho tổng bình phương các phần dư

$$SSE = \sum_{i=1}^{k} e_i^2$$

nhỏ nhất. Phương pháp này được gọi là ước lượng theo bình phương bé nhất.

Ước lượng phương trình hồi quy tuyến tính mẫu trong R

- Để tìm phương trình hồi qui tuyến tính mẫu của số thuế chưa trả theo số giờ lao động của nhân viên và số giờ dùng máy tính ta thực hiện trong *R* như sau:

```
> SoGioNhanVien = c(45,42,44,45,43,46,44,45,44,43)
> SoGioMayTinh = c(16,14,15,13,13,14,16,16,15,15)
> SoThueChuaTra = c(29,24,27,25,26,28,30,28,28,27)
> summary(lm(SoThueChuaTra ~ SoGioNhanVien+SoGioMayTinh))
```

Ước lượng phương trình hồi quy tuyến tính mẫu trong R

Kết quả phân tích hồi quy trong R là:

```
Call:
lm(formula = SoThueChuaTra ~ SoGioNhanVien + SoGioMayTinh)
Residuals:
      Min        1Q    Median        3Q       Max
 -1.24668  -0.74702  -0.02321   0.51956   1.42706
Coefficients:
                Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)    -13.8196     13.3233   -1.037   0.33411
 SoGioNhanVien    0.5637      0.3033    1.859   0.10543
 SoGioMayTinh     1.0995      0.3131    3.511   0.00984  **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.071 on 7 degrees of freedom
Multiple R-Squared: 0.7289, Adjusted R-squared: 0.6515
F-statistic: 9.411 on 2 and 7 DF, p-value: 0.01037
```

Ước lượng phương trình hồi quy tuyến tính mẫu trong R

Kết luận:

- Ước lượng điểm cho $\beta_0, \beta_1, \beta_2$ là

$$b_0 = -13,8196, b_1 = 0.5637, b_2 = 1.0995.$$

Ước lượng phương trình hồi quy tuyến tính mẫu trong R

Kết luận:

- Ước lượng điểm cho $\beta_0, \beta_1, \beta_2$ là

$$b_0 = -13,8196, b_1 = 0.5637, b_2 = 1.0995.$$

- Vậy, phương trình đường hồi quy tuyến tính mẫu là

$$\begin{aligned} y &= b_0 + b_1 x_1 + b_2 x_2 \\ &= -13.8196 + 0.5637 x_1 + 1.0995 x_2. \end{aligned}$$

- Phương trình hồi quy cho ta thấy:
  - Trung bình số thuế chưa trả y biến thiên cùng chiều với số giờ làm việc của nhân viên (vì sao??) và cũng biến thiên cùng chiều với số giờ làm việc của máy tính (vì sao??);
  - Khi số giờ nhân viên tăng lên 1 và giữ nguyên số giờ máy tính, số thuế thu tăng thêm khoảng $b_1 = 0.5637$ đơn vị;
  - Khi số giờ máy tính tăng lên 1 và giữ nguyên số giờ nhân viên, số thuế thu tăng thêm khoảng $b_2 = 1.0995$ đơn vị.

Sai số bình phương trung bình và sai số chuẩn của ước lượng

Xét mô hình hồi qui tuyến tính tổng thể

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

với k biến độc lập và $k+1$ tham số $\beta_0, \beta_1, \ldots, \beta_k$. Nếu các giả thiết cho mô hình được thỏa mãn và SSE là tổng bình phương các phần dư thì:

1. Một ước lượng điểm cho phương sai chung của các nhiễu, σ^2, là sai số bình phương trung bình

$$s^2 = \frac{SSE}{n-(k+1)}$$

2. Một ước lượng điểm cho σ là sai số chuẩn

$$s = \sqrt{\frac{SSE}{n-(k+1)}}$$

Trong ví dụ trên sai số chuẩn của ước lượng là $s = 1.071$

Hệ số xác định bội, hệ số xác định hiệu chỉnh

Trong mô hình hồi qui tuyến tính:

1. Biến thiên toàn bộ là $SST = \sum(y_i - \overline{y})^2$
2. Biến thiên giải thích được là $SSR = \sum(\hat{y}_i - \overline{y})^2$
3. Biến thiên không giải thích được là $SSE = \sum(y_i - \hat{y}_i)^2$
4. $SST = SSR + SSE$
5. Hệ số xác định bội là

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

Tính chất của hệ số xác định bội

- $0 \leqslant R^2 \leqslant 1$.
- $R^2 = 1$ thì phương trình hồi qui giải thích 100% sự thay đổi của Y.
- $R^2 = 0$ thì phương trình hồi qui hoàn toàn không giải thích được sự thay đổi của Y.
- R^2 là hàm tăng theo số biến số độc lập của mô hình, tức là số biến giải thích càng lớn thì R^2 càng lớn. Kết quả này được lí giải như sau:
  - $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ không phụ thuộc vào số biến độc lập trong mô hình.
  - $$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(\hat{Y}_i - (b_1 + b_2X_{2i} + \ldots + b_kX_{ki}))^2$$
    là hàm giảm theo số biến độc lập trong mô hình.
  - Do đó $R^2 = 1 - \dfrac{SSE}{SST}$ là hàm tăng theo số biến độc lập của mô hình.

Hệ số xác định hiệu chỉnh

- Như vậy, tính phù hợp của mô hình tăng lên khi có nhiều biến giải thích mô hình hơn. Tuy nhiên, người ta muốn dụng một lượng biến giải thích vừa đủ sao cho vẫn có được mô hình phù hợp mà không cần quá tốn kèm khi phải thu thập quá nhiều thông tin về biến giải thích.
- Hơn nữa, nhiều khi đưa thêm một biến độc lập vào mô hình thì tác động riêng phần của các biến độc lập tới biến phụ thuộc lại không thực sự có ý nghĩa thống kê.
- Để đánh giá mức độ phù hợp của mô hình, trong đó cân nhắc tới số lượng biến giải thích, người ta sử dụng hệ số xác định hiệu chỉnh R^2_{adj}.

Hệ số xác định hiệu chỉnh

Định nghĩa

Hệ số xác định hiệu chỉnh $\overline{R}^2$ của phương trình hồi quy mẫu được xác định bởi công thức

$$R^2_{adj} = (R^2 - \frac{k}{n-1})(\frac{n-1}{n-(k-1)}.$$

Khoảng tin cậy cho β_i

- Khoảng tin cậy $100(1-\alpha)\%$ cho β_i là

$$[b_i - s_{b_i} t_{n-(k+1),\alpha/2}, b_i + s_{b_i} t_{n-(k+1),\alpha/2}]$$

- Để tìm khoảng tin cậy 95% cho các hệ số β_1, β_2 trong phương trình hồi qui về số thuế chưa trả theo số giờ làm việc của nhân viên và số giờ dùng máy tính, ta thực hiện lệnh sau:

```
> SoGioNhanVien = c(45,42,44,45,43,46,44,45,44,43)
> SoGioMayTinh = c(16,14,15,13,13,14,16,16,15,15)
> SoThueChuaTra = c(29,24,27,25,26,28,30,28,28,27)
> confint(lm(SoThueChuaTra ~ SoGioNhanVien+
SoGioMayTinh),
level = 0.95)
```

Khoảng tin cậy cho β_i

- Kết quả ước lượng độ tin cậy 95% cho hai hệ số β_1, β_2 của mô hình tổng thể trong R như sau

	2.5 %	97.5 %
(Intercept)	-45.3242267	17.684969
SoGioNhanVien	-0.1534683	1.280789
SoGioMayTinh	0.3590134	1.839926

Vậy khoảng tin cậy 95% cho β_1 là $[-0.1534683,\ 1.280789]$, cho β_2 là $[0.3590134,\ 1.839926]$

Ví dụ

Với số giờ lao động của nhân viên là $x_1 = 45$ và số giờ dùng máy tính là $x_2 = 15$:

a) Hãy tìm một ước lượng điểm và khoảng tin cậy 98% cho giá trị thật của số thuế chưa trả.

b) Hãy tìm một ước lượng điểm và khoảng tin cậy 98% cho giá trị trung bình của số thuế chưa trả.

Ví dụ

- Để tìm ước lượng điểm và khoảng tin cậy 98% cho giá trị thật của số thuế chưa trả, ta thực hiện trong *R* như sau:

```
> SoGioNhanVien = c(45,42,44,45,43,46,44,45,44,43)
> SoGioMayTinh = c(16,14,15,13,13,14,16,16,15,15)
> SoThueChuaTra = c(29,24,27,25,26,28,30,28,28,27)
> predict(lm(SoThueChuaTra ~
SoGioNhanVien+SoGioMayTinh),
data.frame(SoGioNhanVien = 45, SoGioMayTinh = 15),
interval = "prediction", level = 0.98)
```

- Kết quả trong *R* cho ta:

```
          fit       lwr       upr
[1,] 28.03714 24.57352 31.50075
```

- Kết quả này cho ta một ước lượng điểm cho giá trị thật của số thuế chưa trả là $\hat{y} = 28.03714$ và khoảng tin cậy 98% cho giá trị thật của số thuế chưa trả là $(24.57352, 31.50075)$.

Ví dụ

- Để tìm ước lượng điểm và khoảng tin cậy 98% cho giá trị trung bình của số thuế chưa trả, ta thực hiện trong *R* như sau:

```
> SoGioNhanVien = c(45,42,44,45,43,46,44,45,44,43)
> SoGioMayTinh = c(16,14,15,13,13,14,16,16,15,15)
> SoThueChuaTra = c(29,24,27,25,26,28,30,28,28,27)
> predict(lm(SoThueChuaTra ~ SoGioNhanVien+
SoGioMayTinh),
data.frame(SoGioNhanVien = 45, SoGioMayTinh = 15),
interval = "confidence", level = 0.98)
```

- Kết quả trong *R* cho ta:

```
          fit       lwr       upr
[1,] 28.03714 26.73549 29.33878
```

- Kết quả này cho ta một ước lượng điểm cho giá trị trung bình của số thuế chưa trả là $\hat{y} = 28.03714$ và khoảng tin cậy 98% cho giá trị trung bình của số thuế chưa trả là $(26.73549, 29.33878)$.