Phân tích hồi quy một yếu tố

caihuuthuc

April 2020

1 Lý thuyết

1.1 Phương trình tổng quát

Mô hình hồi quy tuyến tính đơn giản giữa một biến phụ thuộc y và một biến độc lập x được phát biểu rằng:

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{1}$$

Phương trình trên giả sử biến y bằng một hằng số α cộng với một hệ số β liên quan tới biến x. Trong phương trình trên, α là *chặn* (intercept, tức lúc $x_i = 0$), β là độ dốc (slope hay gradient). Và ϵ_i gọi là phần dư (residual), tuân theo phân phối chuẩn với trung bình 0 và phương sai σ^2.

Các thông số α, β, σ^2 phải được tính từ dữ liệu. Phương pháp để ước tính các thông số này được gọi là *phương pháp bình phương nhỏ nhất* (least squares method). Phương pháp này tìm các giá trị α, β sao cho

$$\sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2 \tag{2}$$

nhỏ nhất.

Sau một vài công thức biến đổi toán học, ta có được công thức để tính ước số $\hat{\alpha}$ và $\hat{\beta}$ là:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3}$$

và

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{4}$$

Sau khi đã có hai ước số $\hat{\alpha}$ và $\hat{\beta}$, ta có thể tính ước số theo từng biến độc lập:

$$\hat{y_i} = \hat{\alpha} + \hat{\beta} x_i \tag{5}$$

Từ đó, ta có thể tính ước số của phương sai phần dư:

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y_i})^2}{n - 2} \tag{6}$$

1.2 Kiểm định cho ước số β

Trong phân tích hồi quy tuyến tính, ta muốn biết hệ số β bằng không hay khác không. Nếu hệ số β bằng 0, tức là biến y chỉ xoay quanh giá trị trung bình và sai số ngẫu nhiên ϵ chứ không hề phụ thuộc vào giá trị của x. Nếu hệ số β khác không, chúng ta có bằng chứng để phát biểu rằng x và y có liên quan đến nhau. Để kiểm định giả thuyết $\beta = 0$, chúng ta dùng kiểm định t:

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} \tag{7}$$

trong đó $SE(\hat{\beta})$ là sai số chuẩn (standard error) của ước số $\hat{\beta}$.

Trong biểu thức trên, t tuân theo luật phân phối t với n - 2 bậc tự do.

1.3 Đánh giá độ thích hợp của mô hình

Một thước đo thường được sử dụng để đánh giá sự phù hợp của mô hình hồi quy tuyến tính là hệ số xác định R bình phương (R-squared). Xuất phát từ ý tưởng xem toàn bộ biến thiên quan sát của biến phụ thuộc chia thành hai phần: phần biến thiên do hồi quy và phần biến thiên không do hồi quy (còn gọi là phần dư):

$$\sum_{i=1}^{n} (y_i - \bar{Y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{Y})^2 \tag{8}$$

Ta ký hiệu:

- SST $= \sum_{i=1}^{n} (y_i - \bar{Y})^2$: Sum Square Total, tức toàn bộ biến thiên của dữ liệu quan sát được
- SSR $= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$: Sum Square Regression, tức biến thiên mà mô hình hồi quy có thể giải thích được.
- SSE $= \sum_{i=1}^{n} (\hat{y}_i - \bar{Y})^2$: Sum Square Error (Residuals), tức biến thiên mà mô hình hồi quy không giải thích được.

Từ đó ta có thể viết gọn:

$$SST = SSR + SSE \tag{9}$$

Từ đó,ta có thể tính hệ số xác định R^2:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{10}$$

Giá trị của R bình phương dao động từ 0 đến 1. Nếu phần biến thiên do phần dư càng nhỏ, nghĩa là khoảng cách giữa các điểm quan sát đến đường thẳng hồi quy càng nhỏ, thì phần biến thiên do hồi quy sẽ càng cao, do đó hệ số xác định càng cao. Với một mô hình cụ thể nào đó, nếu hệ số $R^2 = 0.6$, ta có thể phát biểu rằng mô hình giải thích 60% dữ liệu.

2 Thực hành với ngôn ngữ R

2.1 Load dữ liệu

Ta sử dụng bộ dữ liệu cars trong thư viện car để làm ví dụ cho hồi quy tuyến tính. Dữ liệu này gồm 50 quan sát (observations), trong đó mỗi quan sát bao gồm hai giá trị:: speed - tốc độ của xe lúc bắt đầu ngừng, và dist - khoảng cách đi thêm để xe ngừng hẳn.

```
require(car)
head(cars)
```

```
  speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10
```

2.2 Phân tích ban đầu bằng các biểu đồ

Mục tiêu là ta cần phải xây dựng một mô hình dự đoán khoảng cách cần đi thêm (dist, đơn vị ft) dựa vào tốc độ của xe (speed, đơn vị mph). Nhưng trước khi bắt đầu xây dựng mô hình, ta cần vẽ các biểu đồ để kiểm tra các tính chất của dữ liệu:

- Biểu đồ tán xạ (scatter plot): Để xem liệu có mối quan hệ tuyến tính tăng/giảm giữa biến độc lập và biến phụ thuộc hay không
- Biểu đồ hộp (box plot): Để xem liệu có điểm ngoại vi (outliers) nào trong dữ liệu hay không (những điểm nằm ngoài 1.5 * IQR tính từ bách phân vị 75% có thể là các điểm ngoại vi). Dữ liệu có điểm ngoại vi có thể ảnh hưởng đến độ dốc (slope) của đường thẳng hồi quy.
- Biểu đồ mật độ (desity plot): Để xem mật độ phân phối của dữ liệu.

Biểu đồ tán xạ (scatter plot) Ta vẽ biểu đồ tán xạ bằng lệnh sau.

```
scatter.smooth(cars$speed, cars$dist)
```

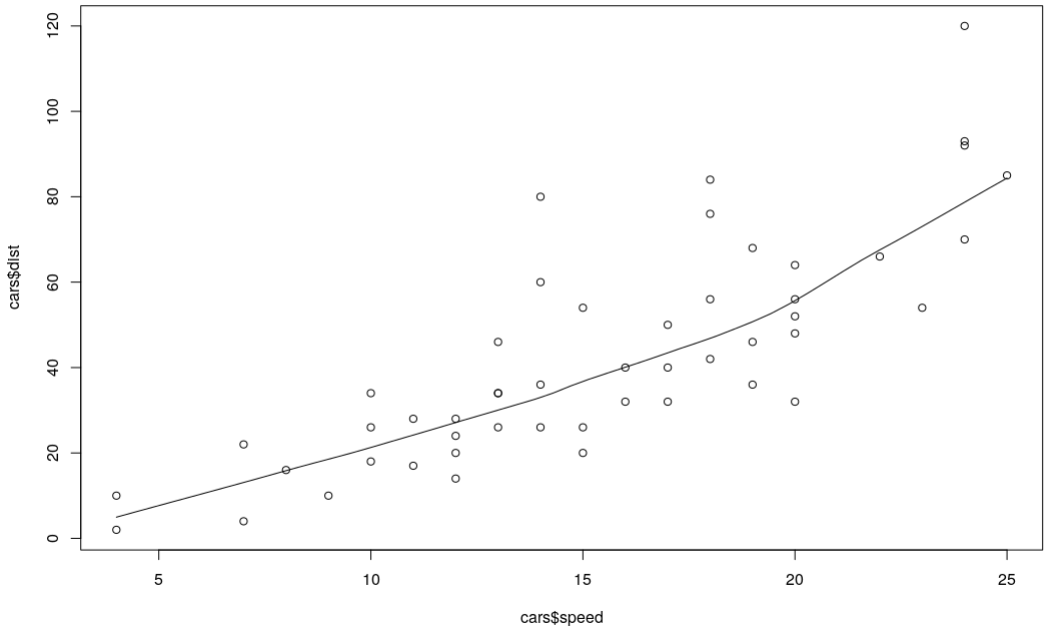

Figure 1: Biểu đồ tán xạ của biến *dist* và *speed*

Biểu đồ tán xạ cùng với đường làm mịn có hướng tăng dần cho chúng ta thấy mối quan hệ tăng tuyến tính giữa speed và dist.

```
par(mfrow=c(1, 2))

boxplot(
    cars$speed,
    main="Speed",
    sub=paste("Outlier rows: ", boxplot.stats(cars$speed)$out))

boxplot(
    cars$dist,
    main="Distance",
    sub=paste("Outlier rows: ", boxplot.stats(cars$dist)$out))

par(mfrow=c(1,1))
```

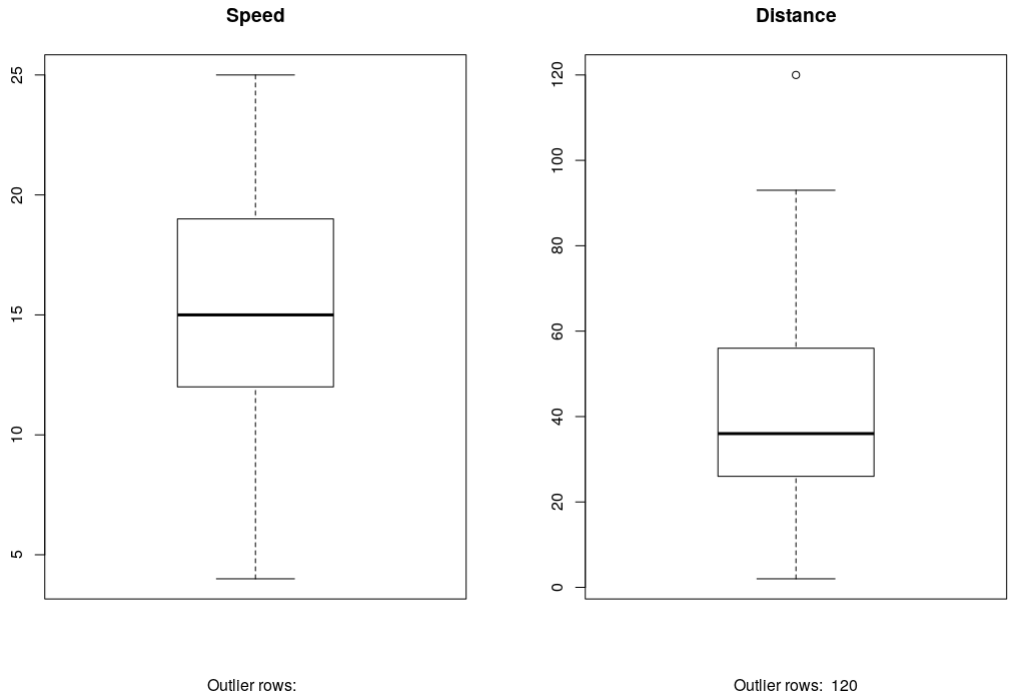

Figure 2: Biểu đồ hộp của biến *dist* và *speed*

Qua biểu đồ 2 ta thấy chỉ có một điểm của biến dist có thể là ngoại vi. (Vì ta chưa kiểm định nên chưa kết luận được là ngoại vi).

Biểu đồ mật độ (density plot) Ta vẽ biểu đồ mật độ của hai biến 'dist' và 'speed' bằng cách lệnh dưới đây.

```
library(e1071)

par(mfrow=c(1, 2))  # divide graph area in 2 columns

plot(
    density(cars$speed),
    main="Density Plot: Speed",
    ylab="Frequency",
    sub=paste("Skewness:", round(e1071::skewness(cars$speed), 2)))

polygon(density(cars$speed), col="red")

plot(
```

```
    density(cars$dist),
    main="Density Plot: Distance",
    ylab="Frequency",
    sub=paste("Skewness:", round(e1071::skewness(cars$dist), 2)))

polygon(density(cars$dist), col="red")

par(mfrow=c(1, 1))
```

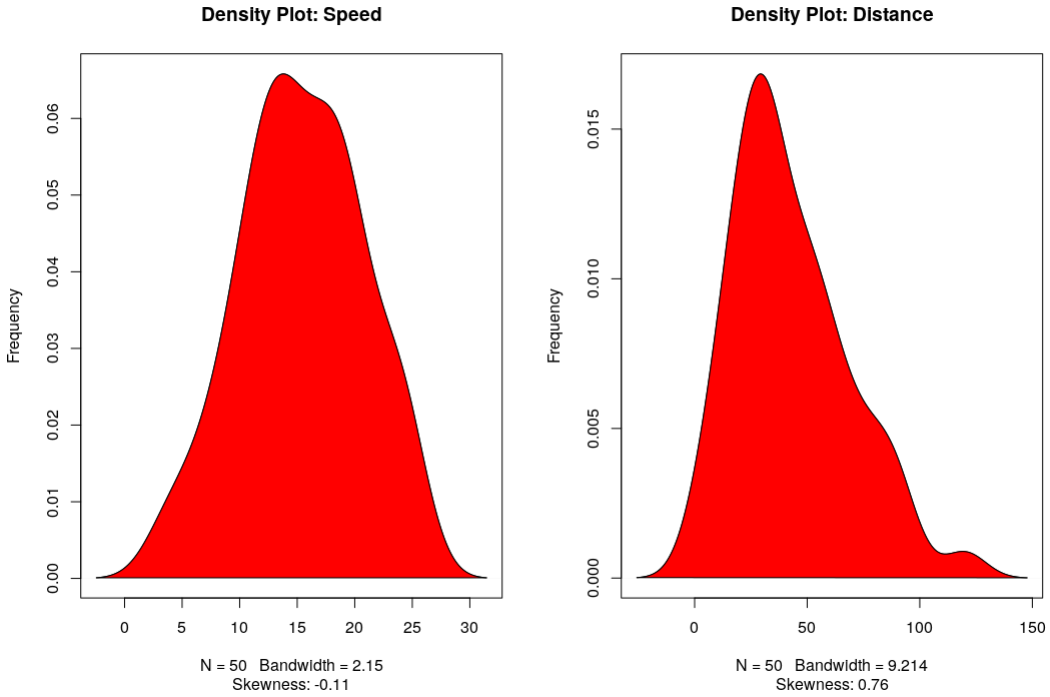

Figure 3: Biểu đồ mật độ của biến *dist* và *speed*

Ta có trực giác rằng biến speed tuân theo luật phân phối chuẩn còn biến dist thì không. Để kiểm tra, ta sử dụng kiểm định shapiro:

```
shapiro.test(cars$speed)

        Shapiro-Wilk normality test

data:  cars$speed
W = 0.97765, p-value = 0.4576

shapiro.test(cars$dist)
```

```
        Shapiro-Wilk normality test

data:  cars$dist
W = 0.95144, p-value = 0.0391
```

Trị số $p > 0.05$ nên biến speed tuân theo luật phân phối chuẩn. Còn biến dist không tuân theo luật phân phối chuẩn do có trị số $p < 0.05$

2.3 Phân tích hệ số tương quan

Hệ số tương quan là một chỉ số thống kê đo lường môi liên hệ tương quan giữa hai biến số, x và y. Hệ số tương quan nhận giá trị liên tục từ -1 đến 1. Hệ số tương quan bằng 0, hoặc gần 0, có nghĩa là hai biến số không có liên hệ gì với nhau. Nêu hệ số tương quan bằng 1 hoặc -1 có nghĩa là hai biến số có tương quan tuyệt đối. Nêu hệ số tương quan âm, thì hai biến số có mối quan hệ nghịch biến (khi x tăng thì y giảm). Nêu hệ số tương quan dương thì hai biến số tương quan đồng biến, khi x tăng thì y tăng.

Ta ước tính hệ số tương quan của speed và dist bằng lệnh:

```
cor(cars$dist, cars$speed)
```

```
[1] 0.8068949
```

Và kiểm định giả thiết hệ số tương quan bằng 0 bằng lệnh:

```
cor.test(cars$dist, cars$speed)
```

```
        Pearson's product-moment correlation

data:  cars$dist and cars$speed
t = 9.464, df = 48, p-value = 1.49e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6816422 0.8862036
sample estimates:
      cor
0.8068949
```

Kết quả phân tích cho thấy kiểm định $t = 9.464$ với trị số $p = 1.49e - 12 < 0.05$. Do đó, ta có bằng chứng kết luận rằng mối liên hệ giữa speed và dist có ý nghĩa thống kê.

2.4 Phân tích hồi quy tuyến tính đơn giản bằng R

Hàm 'lm' (viết tắt của **l**inear **m**odel) trong R có thể tính toán giá trị của các ước số $\hat{\alpha}$, $\hat{\beta}$ và s^2 một cách nhanh chóng.

```
lm(cars$dist ~ cars$speed)

Call:
lm(formula = cars$dist ~ cars$speed)

Coefficients:
(Intercept)   cars$speed
    -17.579        3.932
```

Trong lệnh trên, "cars$dist ~cars$speed" có nghĩa là mô tả cars$dist như là một hàm số của cars$speed. Kết quả tính toán cho thấy $\hat{\alpha} = -17.579$ và $\hat{\beta} = 3.932$. Nói cách khác, với hai thông số này, ta có thể ước tính số khoảng cách cần để xe ngừng hẳn dist từ bất kì tốc độ nào bằng phương trình:

$$\hat{y_i} = -17.579 + 3.932x_i \tag{11}$$

Phương trình này có nghĩa là khi tốc độ lúc xe bắt đầu ngưng tăng thêm 1 mph thì xe phải chạy thêm 3.932 ft để ngừng hẳn.

Ta có thể xem thêm các thông tin khác về mô hình tuyến tính bằng lệnh 'summary'

```
summary(lm(cars$dist ~ cars$speed))


Call:
lm(formula = cars$dist ~ cars$speed)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
cars$speed    3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Ta có thể chia phần kết quả này thành 3 phần:

- Phần 1, Residuals, mô tả phần dư (residuals) của mô hình hồi quy.
- Phần 2, Coefficients, trình bày ước số $\hat{\alpha}$ của $\hat{\beta}$ cùng với sai số chuẩn (standard error) và giá trị của kiểm định t. Giá trị kiểm định t cho $\hat{\beta}$

bằng 9.464 với trị số p = 1.49e-12. Nói cách khác, chúng ta có bằng chứng để cho rằng mối liên hệ giữa dist và speed, và mối liên hệ này có ý nghĩa thống kê.

- Phần 3 cho chúng ta thông tin về phương sai của phần dư. Ở mô hình này là $s^2 = 15.38$. Hệ số $R^2 = 0.6511$ có nghĩa những khác biệt về speed‘ giải thích 65.11% những khác biệt về dist‘ giữa các xe.

2.5 Mô hình tiên lượng

Sau khi đã có được các ước số của α và β, ta có thể vẽ đường biểu diễn mối liên hệ giữa hai biến số.

```
plot(cars, pch=16)
abline(lm(cars$dist ~ cars$speed))
```

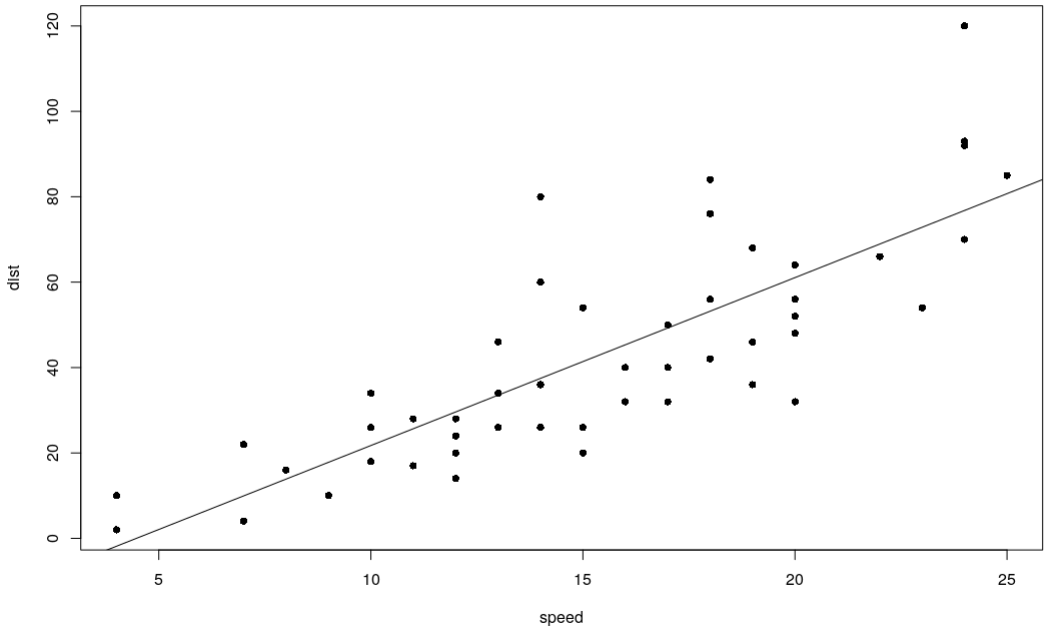

Figure 4: Đường thẳng hồi quy thể hiện mối quan hệ giữa hai biến số