Mô Hình Hồi Quy Bội

Code

Nguyễn Thanh Nga - Học viện Ngân Hàng

9/22/2021

Tài liệu buổi thực hành seminar TKUD [TKUD19] 22/9/2021: Mô hình hồi quy bội

Speaker: NCS. Nguyễn Thanh Nga - Học viện Ngân Hàng

Chi tiết tại: https://sites.google.com/view/tkud/home?authuser=1 (https://sites.google.com/view/tkud/home?authuser=1)

Tài liệu thực hành có thể download tại đây (https://drive.google.com/drive/folders/1maNUAWyCcjXrU0m6hMgZNhjEI0jUI9Gu?usp=sharing).

(Chọn chuột phải tại chữ "tại đây", chọn open new tab)

Hoặc copy link này: https://drive.google.com/drive/folders/1maNUAWyCcjXrU0m6hMgZNhjEI0jUI9Gu?usp=sharing (https://drive.google.com/drive/folders/1maNUAWyCcjXrU0m6hMgZNhjEI0jUI9Gu?usp=sharing)

Code

1. Mô tả dữ liệu

Tập tin *marketing.csv* dùng để phân tích ảnh hưởng của các hình thức quảng cáo lên doanh thu. Dữ liệu bao gồm 4 biến:

- $youtube$, $facebook$ và $newspaper$ là số tiền chi cho quảng cáo (Đơn vị: triệu đồng)
- $sales$ là biến doanh thu. Bộ số liệu có 200 quan sát, thu thập tại 200 cửa hàng.

Code

```
##   X youtube facebook newspaper     sales
## 1 1  200.92   142.17    145.41  943.0419
## 2 2  156.26   129.85     62.70  856.2597
## 3 3  124.38   187.57    140.04  964.9689
## 4 4  157.69   187.48    143.94 1017.4412
## 5 5  158.23   222.41    116.04 1115.2990
## 6 6  132.48   181.55    119.84  932.3739
```

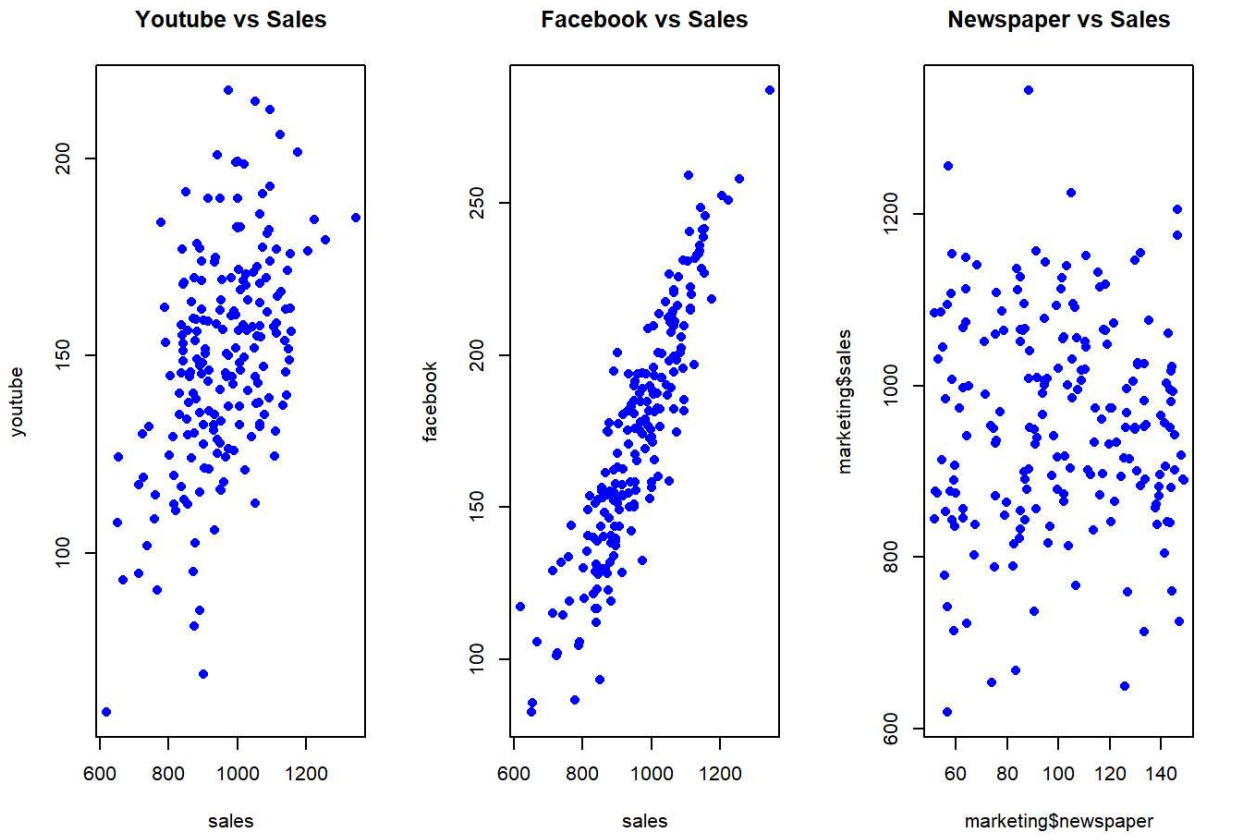
Code

```
## 'data.frame':    200 obs. of  5 variables:
##  $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ youtube  : num  201 156 124 158 158 ...
##  $ facebook : num  142 130 188 187 222 ...
##  $ newspaper: num  145.4 62.7 140 143.9 116 ...
##  $ sales    : num  943 856 965 1017 1115 ...
```

2. Phân tích tương quan

vẽ đồ thị phân tán giữa các cặp biến

Code



Code

Từ các đồ thị phân tán, ta có thể nhận xét rằng các biến $youtube$ và $facebook$ có mối quan hệ tuyến tính với biến $sales$ trong khi biến $newspaper$ không có. Mặt khác, mối quan hệ giữa $Newspaper$ và $sales$, nói một cách chính xác hơn là phi tuyến tính hơn là tuyến tính.

Hệ số tương quan tuyến tính (Correlation coefficient)

Hệ số tương quan đo lường mức độ quan hệ tuyến tính giữa hai biến, không phân biệt biến này phụ thuộc vào biến kia

Các phương pháp tính tương quan: * Hệ số tương quan Pearson * Hệ số tương quan spearman * Hệ số tương quan kendall

2.1. Hệ số tương quan Pearson

- Công thức tính hệ số tương quan Pearson trên R: cor(df, method = “pearson”)

Code

```
##                    X     youtube    facebook    newspaper        sales
## X          1.00000000 -0.04977015 -0.03918551 -0.177473371 -0.047532656
## youtube   -0.04977015  1.00000000  0.08401121  0.047806059  0.487083735
## facebook  -0.03918551  0.08401121  1.00000000 -0.039579633  0.903092760
## newspaper -0.17747337  0.04780606 -0.03957963  1.000000000 -0.002900308
## sales     -0.04753266  0.48708374  0.90309276 -0.002900308  1.000000000
```

Code

```
## [1] 0.9030928
```

2.2. Hệ số tương quan spearman

- Đánh giá mức độ tương quan của 2 hạng của 2 biến (rank-ordered variables), sử dụng khi phân phối của tổng thể được giả sử không phải là phân phối chuẩn hoặc trong trường hợp có các giá trị quan sát bất thường (lớn quá hoặc nhỏ quá)
- Công thức tính trên R: cor(df, method = “spearman”)

Code

```
##                     X      youtube    facebook     newspaper         sale
s
## X          1.000000000 -0.007446197 -0.06512643 -0.1753965822 -0.070609765
2
## youtube   -0.007446197  1.000000000  0.06036389  0.0620604644  0.442211718
6
## facebook  -0.065126427  0.060363895  1.00000000 -0.0332461435  0.899423160
1
## newspaper -0.175396582  0.062060464 -0.03324614  1.0000000000 -0.000452261
8
## sales     -0.070609765  0.442211719  0.89942316 -0.0004522618  1.000000000
0
```

Code

```
## [1] 0.4422117
```

2.3. Hệ số tương quan kendall

- Hệ số kendall ít dùng hơn so với 2 hệ số tương quan trên
- Công thức tính trên R: cor(df, method = “kendall”)

Code

```
##                     X      youtube    facebook   newspaper       sales
## X          1.000000000 -0.005025631 -0.04532891 -0.11734553 -0.04874372
## youtube   -0.005025631  1.000000000  0.03789516  0.04829995  0.30324656
## facebook  -0.045328911  0.037895160  1.00000000 -0.02326925  0.72656918
## newspaper -0.117345529  0.048299952 -0.02326925  1.00000000 -0.00587984
## sales     -0.048743719  0.303246559  0.72656918 -0.00587984  1.00000000
```

Code

```
## [1] 0.3032466
```

2.4. Kiểm định tương quan

- Cũng như phương pháp tính, kiểm định cũng có 3 method: Pearson, Spearman, Kendall
- Giả thiết kiểm định:
    - H0: Không có tương quan (hệ số tương quan = 0)
    - H1: Có tương quan
- Ví dụ: Sử dụng hàm cor.test

Code

```
##
##  Pearson's product-moment correlation
##
## data:  marketing$youtube and marketing$sales
## t = 7.8478, df = 198, p-value = 2.597e-13
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3735893 0.5862096
## sample estimates:
##       cor
## 0.4870837
```

Code

```
##
##  Pearson's product-moment correlation
##
## data:  marketing$facebook and marketing$sales
## t = 29.591, df = 198, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8738405 0.9258309
## sample estimates:
##       cor
## 0.9030928
```

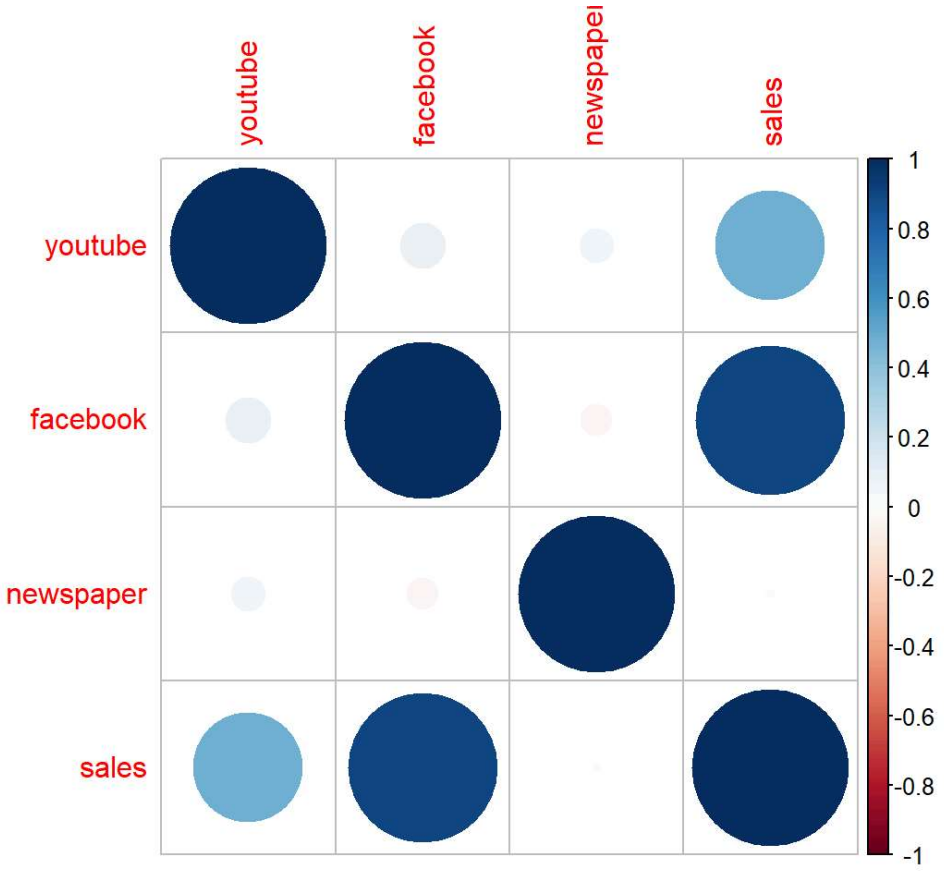
Code

```
##
##  Pearson's product-moment correlation
##
## data:  marketing$newspaper and marketing$sales
## t = -0.040811, df = 198, p-value = 0.9675
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1415844  0.1358954
## sample estimates:
##          cor
## -0.002900308
```

2.5.Một số đồ thị trong phân tích tương quan

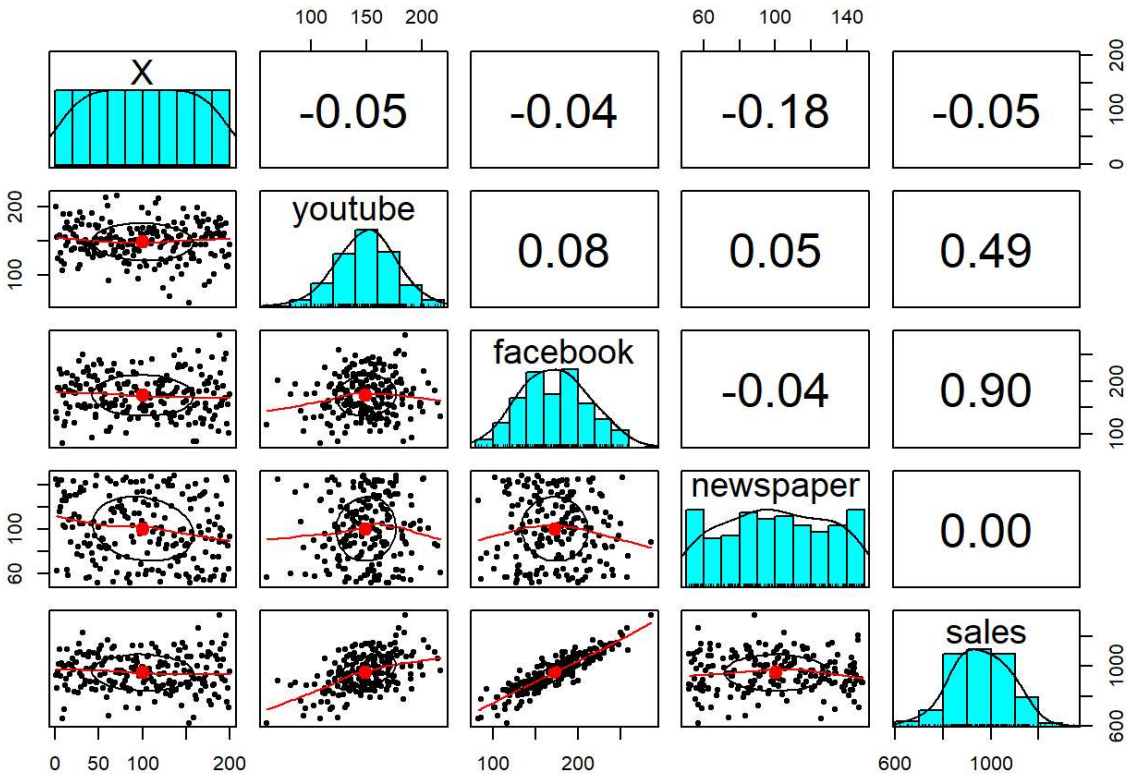
- corrplot (package: corrplot)

Code



- biểu đồ tương quan giữa các biến, hàm pairs.panels() trong gói *psych*

Code



3. Hồi quy tuyến tính đa biến

Mô hình hồi quy bội có dạng:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + . . + \beta_k X_k + u$$

Yêu cầu: xây dựng mô hình hồi quy bội (multiple regression) để phân tích tác động của các hình thức quảng cáo lên doanh thu và diễn giải kết quả. cụ thể:

- Biến phụ thuộc: doanh thu $sales$
- Biến độc lập: $youtube, facebook, newspaper$ lần lượt là số tiền chi cho quảng cáo trên Youtube, Facebook và trên báo chí

3.1.Ước lượng mô hình

Sử dụng lệnh **lm()**:

Code

```
##
## Call:
## lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.571  -9.826   0.581   9.575  40.175
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 202.19448    7.93632  25.477   <2e-16 ***
## youtube       1.86693    0.03832  48.714   <2e-16 ***
## facebook      2.73133    0.02668 102.382   <2e-16 ***
## newspaper     0.04969    0.03590   1.384    0.168
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.62 on 196 degrees of freedom
## Multiple R-squared:  0.986,  Adjusted R-squared:  0.9858
## F-statistic:  4605 on 3 and 196 DF,  p-value: < 2.2e-16
```

Vì biến $newspaper$ không có ý nghĩa trong mô hình hồi quy bội được xây dựng, ta có thể loại bỏ biến này và xây dựng lại mô hình hồi quy bội chỉ với hai biến $youtube$ và $facebook$:

Code

```
##
## Call:
## lm(formula = sales ~ youtube + facebook, data = marketing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.589  -9.631   0.468   9.658  38.351
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 207.06397    7.13046   29.04   <2e-16 ***
## youtube       1.86966    0.03836   48.74   <2e-16 ***
## facebook      2.72971    0.02671  102.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.65 on 197 degrees of freedom
## Multiple R-squared:  0.9859, Adjusted R-squared:  0.9857
## F-statistic:  6875 on 2 and 197 DF,  p-value: < 2.2e-16
```

Hệ số R^2 hiệu chỉnh bằng 0.9857 nghĩa là 98.57% sự biến thiên trong doanh thu $sales$ được giải thích bởi các biến $youtube$ và $facebook$ tức là chi phí chi cho việc quảng cáo trên Facebook và Youtube.

3.2. Tìm khoảng tin cậy cho các hệ số hồi quy

Sử dụng hàm **confint()**:

Code

```
##                 2.5 %     97.5 %
## (Intercept) 193.002136 221.125806
## youtube       1.794002   1.945312
## facebook      2.677027   2.782391
```

Code

```
##                 0.5 %     99.5 %
## (Intercept) 188.517527 225.610415
## youtube       1.769874   1.969440
## facebook      2.660225   2.799193
```

3.3. Kiểm định và lựa chọn mô hình

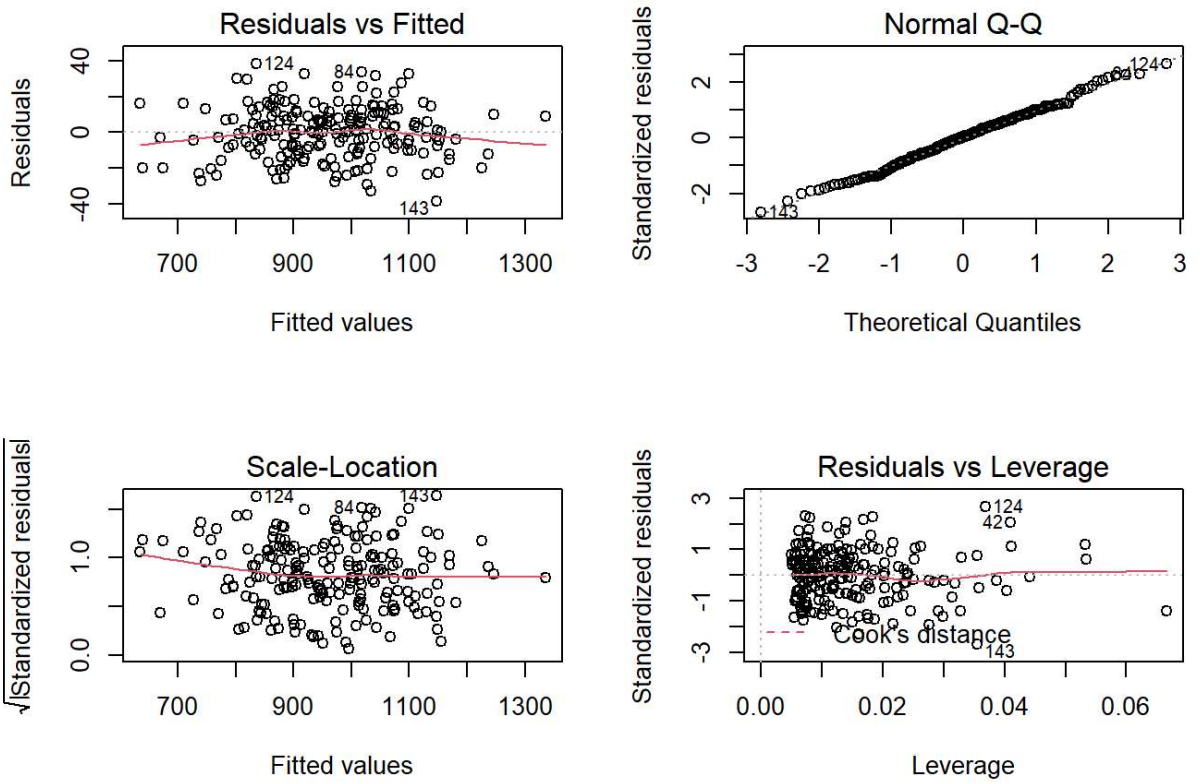
Nhắc lại một số giả thiết của mô hình hồi quy

1. **Tính tuyến tính của dữ liệu**
2. **Phần dư có phân phối chuẩn**
3. **Phần dư có trung bình bằng 0**

4. **Phương sai của phần dư là không đổi**

3.3.1 Dựa vào đồ thị phần dư

Code



Code

- Đồ thị thứ 1 (*Residuals vs Fitted*) Dùng để kiểm tra giả thiết tuyến tính của dữ liệu và giả thiết phần dư có trung bình bằng 0

Trục tung biểu thị giá trị của phần dư, trục hoành biểu thị giá trị tiên lượng ($\hat{y}_i$) của biến phụ thuộc. Nếu như giả thiết về tính tuyến tính của dữ liệu **KHÔNG** thỏa, ta sẽ quan sát thấy rằng đường màu đỏ trên đồ thị sẽ phân bố theo một hình mẫu (pattern) đặc trưng nào đó (ví dụ parabol). Nếu đường màu đỏ trên đồ thị phân tán là đường thẳng nằm ngang mà không phải là đường cong, thì giả thiết tính tuyến tính của dữ liệu được thỏa mãn. Giả thiết phần dư có trung bình bằng 0 thỏa mãn nếu đường màu đỏ gần với đường nằm ngang (ứng với phần dư =0).

- Đồ thị thứ 2 (*Normal Q-Q*) Dùng để kiểm tra giả thiết phần dư có phân phối chuẩn

Nếu các điểm thặng dư nằm trên cùng 1 đường thẳng thì điều kiện về phân phối chuẩn được thỏa.

- Đồ thị thứ 3 (*Scale - Location*) Dùng để kiểm định giả thiết phương sai của phần dư là không đổi

Trục tung là căn bậc hai của giá trị của phần dư (đã được chuẩn hóa), trục hoành là giá trị tiên lượng ($\hat{y}_i$) của biến phụ thuộc từ mô hình. Nếu như đường màu đỏ trên đồ thị là đường thẳng nằm ngang và các điểm thặng dư phân tán đều xung quanh đường thẳng này thì giả thiết thứ 4

được thỏa. Nếu như đường màu đỏ có độ dốc (hoặc cong) hoặc các điểm thặng dư phân tán không đều xung quanh đường thẳng này, thì giả thiết này bị vi phạm.

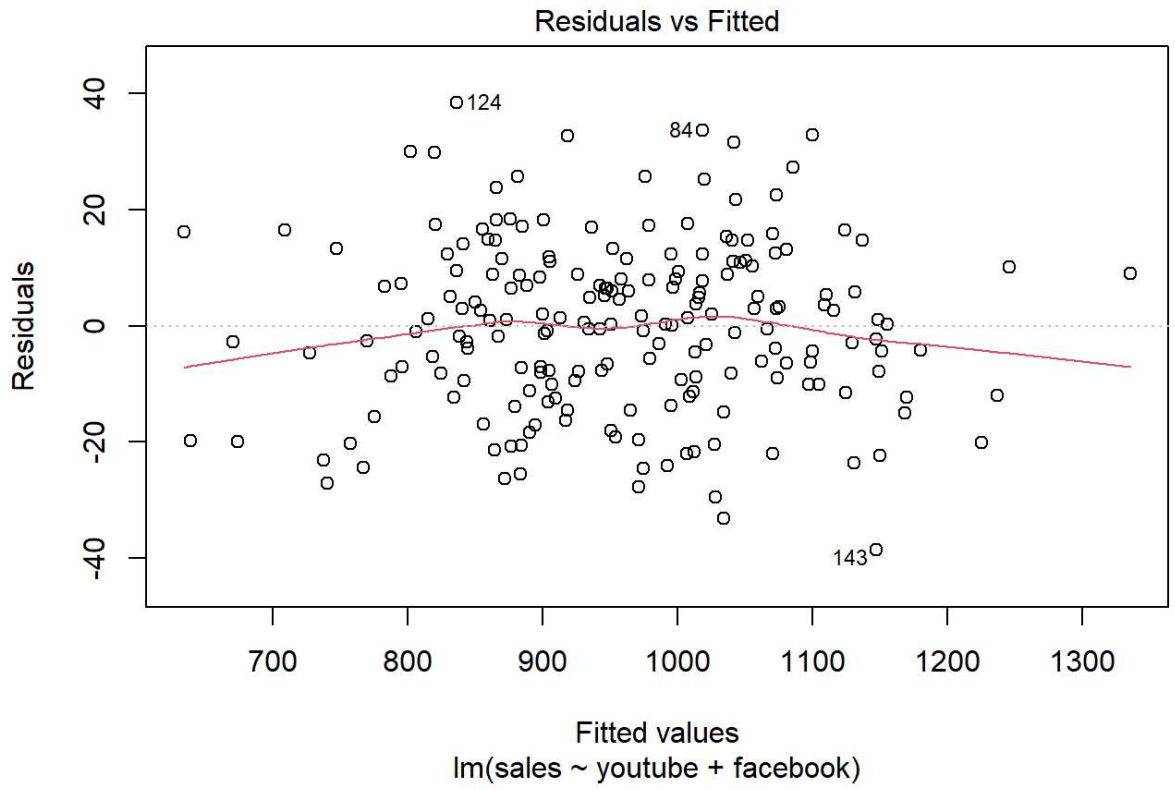
- Đồ thị thứ 4 (*Residuals vs Leverage*) cho phép xác định những điểm có ảnh hưởng cao (*influential observations*), nếu chúng có hiện diện trong bộ dữ liệu. Những điểm có ảnh hưởng cao này có thể là các điểm outliers, là những điểm có thể gây nhiều ảnh hưởng nhất khi phân tích dữ liệu. Nếu như ta quan sát thấy một đường thẳng màu đỏ đứt nét (Cook's distance (https://en.wikipedia.org/wiki/Cook%27s_distance)), và có một số điểm vượt qua đường thẳng khoảng cách này, nghĩa là các điểm đó là các điểm có ảnh hưởng cao. Nếu như ta chỉ quan sát thấy đường thẳng khoảng cách Cook ở góc của đồ thị và không có điểm nào vượt qua nó, nghĩa không có điểm nào thực sự có ảnh hưởng cao.

Nhận xét:

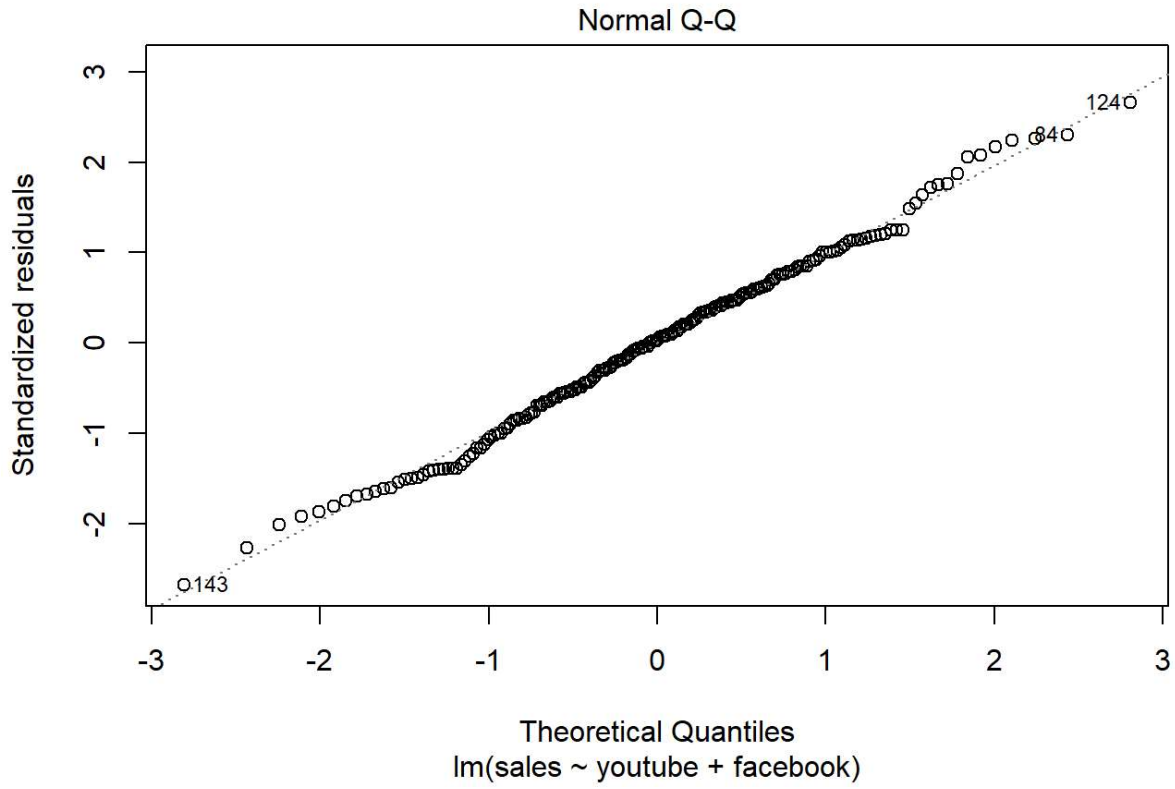
- Đồ thị thứ 1 (*Residuals vs Fitted*) cho thấy giả thiết về tính tuyến tính của dữ liệu hơi bị vi phạm. Tuy nhiên giả thiết trung bình của phần dư có thể coi là thỏa mãn
- Đồ thị *Normal Q-Q* cho thấy giả thiết phần dư có phân phối chuẩn được thỏa mãn.
- Đồ thị (*Scale - Location*) cho ta thấy rằng giả thiết về tính đồng nhất của phương sai cũng thỏa mãn.
- Đồ thị thứ tư chỉ ra có các quan trắc thứ 42, 124 và 143 có thể là các điểm có ảnh hưởng cao trong bộ dữ liệu.

Để vẽ từng đồ thị, ta dùng các lệnh sau:

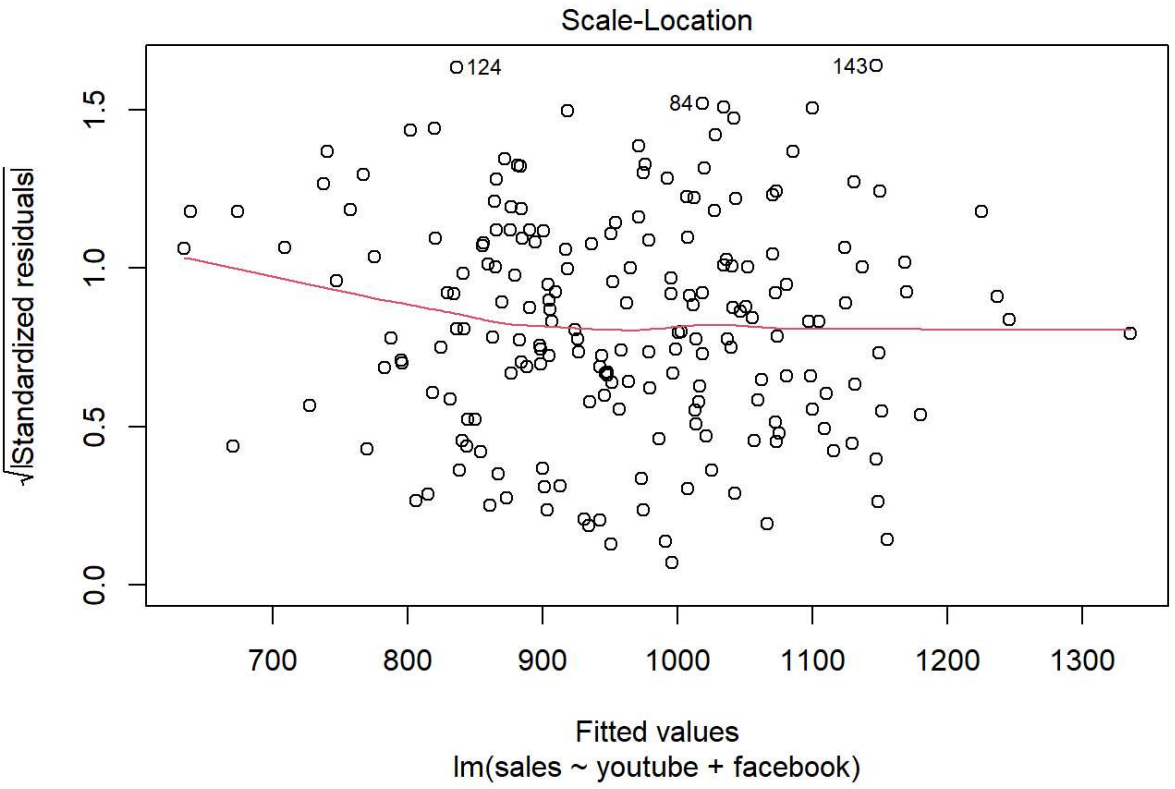
Code



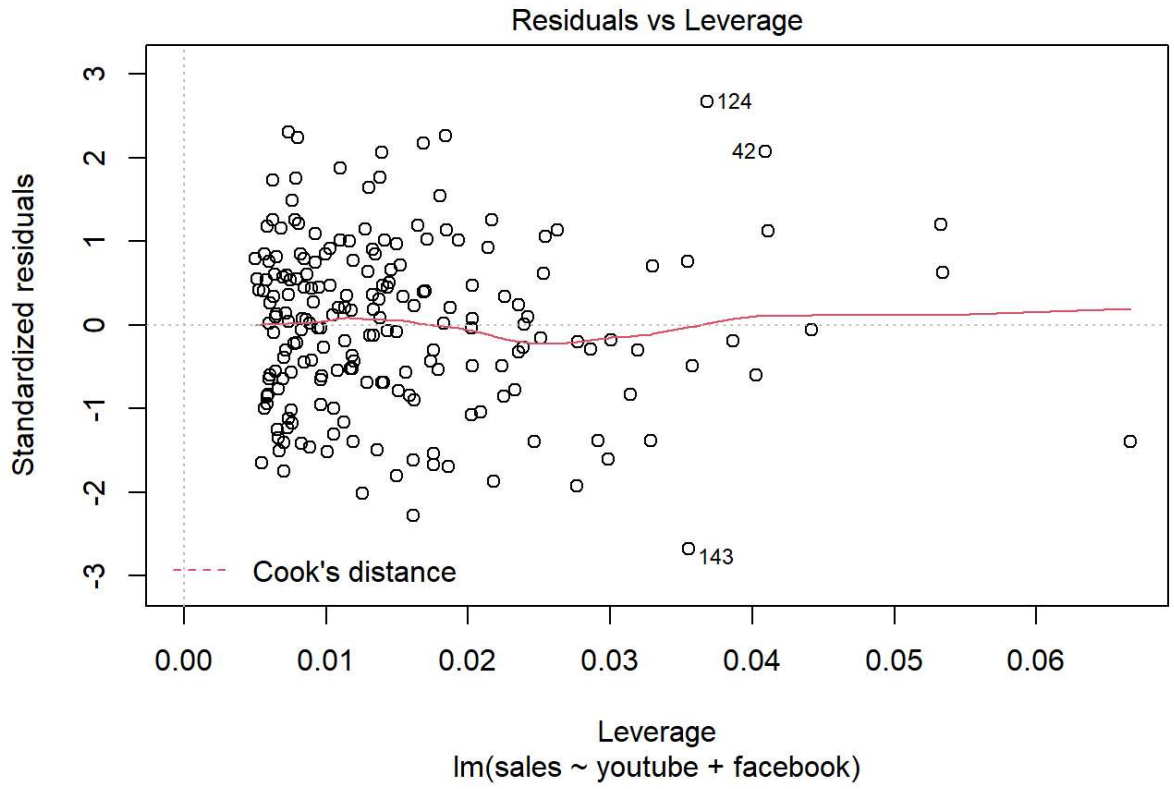
Code


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(sales ~ youtube + facebook)
124
84
143

Code


Scale-Location
√|Standardized residuals|
Fitted values
lm(sales ~ youtube + facebook)
124
143
84

Code



3.3.2 Dựa vào các kiểm định

Code

Giả thiết 1: Sai số ngẫu nhiên có phân phối chuẩn

Giả thiết kiểm định:

```
- H0: Dữ liệu có phân phối chuẩn
- H1: Dữ liệu không có phân phối chuẩn
```

Code

```
##
##  Shapiro-Wilk normality test
##
## data:  re
## W = 0.9945, p-value = 0.676
```

p-value =0.676 nên phần dư có phân phối chuẩn

Giả thiết 2: Kỳ vọng của sai số ngẫu nhiên tại mỗi giá trị bằng 0

Code

```
##
##  One Sample t-test
##
## data:  re
## t = -6.9193e-16, df = 199, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -2.032987  2.032987
## sample estimates:
##     mean of x
## -7.133487e-16
```

p-value = 1, do do mô hình thỏa mãn giả thiết 2

Giả thiết 3: Phương sai của sai số ngẫu nhiên không đổi

H0: phương sai sai số không đổi

H1: Phương sai sai số thay đổi

Code

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.2402166, Df = 1, p = 0.62405
```

p-value = 0.624, có thể kết luận phương sai sai số là không đổi.

Giả thiết 4: Giữa các biến độc lập không có mối quan hệ đa cộng tuyến hoàn hảo

$VIF_j = \frac{1}{1-R_j^2}$

R_j^2 là hệ số xác định của mô hình hồi qui tuyến tính phụ của biến độc lập X_j theo các biến độc lập còn lại của mô hình.

Code

```
##  youtube facebook
## 1.007108 1.007108
```

vif < 10 cho thấy các biến độc lập không có đa cộng tuyến

3.4. Dự báo

Code

Dự báo giá trị trung bình

Code

```
##          fit       lwr       upr
## 1   340.9315  330.1266  351.7365
## 2   362.0998  351.6488  372.5509
## 3   447.3154  437.8850  456.7458
## 4   633.2717  627.4215  639.1219
## 5   457.9871  449.3153  466.6590
## 6   882.0121  875.3312  888.6930
## 7   934.5121  929.9632  939.0609
## 8   896.9689  894.5891  899.3486
## 9   853.4203  847.4027  859.4380
## 10  649.5938  643.9473  655.2404
## 11  648.8345  641.5132  656.1557
## 12 1082.1027 1078.0787 1086.1267
## 13  925.6076  914.6529  936.5623
## 14 1123.1979 1118.8581 1127.5377
```

Dự báo giá trị cá biệt

Code

```
##          fit       lwr       upr
## 1   340.9315  310.0795  371.7836
## 2   362.0998  331.3699  392.8297
## 3   447.3154  416.9174  477.7133
## 4   633.2717  603.7874  662.7561
## 5   457.9871  427.8159  488.1584
## 6   882.0121  852.3517  911.6725
## 7   934.5121  905.2581  963.7660
## 8   896.9689  867.9729  925.9648
## 9   853.4203  823.9023  882.9384
## 10  649.5938  620.1492  679.0385
## 11  648.8345  619.0233  678.6456
## 12 1082.1027 1052.9257 1111.2797
## 13  925.6076  894.7028  956.5124
## 14 1123.1979 1093.9757 1152.4201
```

3.5. Xuất kết quả hệ số hồi quy

Xuất bảng hồi quy dạng tex, cho báo cáo file Word

Code

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                  sales
## ------------------------------------------------
## youtube                         1.870***
##                                  (0.038)
##
## facebook                        2.730***
##                                  (0.027)
##
## Constant                       207.064***
##                                  (7.130)
##
## ------------------------------------------------
## Observations                      200
## R2                               0.986
## Adjusted R2                      0.986
## Residual Std. Error       14.654 (df = 197)
## F Statistic          6,875.092*** (df = 2; 197)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Xuất bảng hồi quy dạng tex, cho báo cáo file latex.

Code

```
##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University.
E-mail: hlavac at fas.harvard.edu
## % Date and time: Tue, Sep 21, 2021 - 10:19:08 PM
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{\textit{Dependent variable:}} \\
## \cline{2-2}
## \\[-1.8ex] & sales \\
## \hline \\[-1.8ex]
##  youtube & 1.870$^{***}$ \\
##   & (0.038) \\
##   & \\
##  facebook & 2.730$^{***}$ \\
##   & (0.027) \\
##   & \\
##  Constant & 207.064$^{***}$ \\
##   & (7.130) \\
##   & \\
## \hline \\[-1.8ex]
## Observations & 200 \\
## R$^{2}$ & 0.986 \\
## Adjusted R$^{2}$ & 0.986 \\
## Residual Std. Error & 14.654 (df = 197) \\
## F Statistic & 6,875.092$^{***}$ (df = 2; 197) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{**
*}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

TRÂN TRỌNG MỜI ĐẠI BIỂU THAM DỰ VÀ TRÂN TRỌNG CẢM ƠN!