Báo cáo Nhóm 12

Bài toán: Dùng phần mềm R xây dựng mô hình hồi quy tuyến tính bội số >= 2 để xác định các hệ số B0, B1,….Bk

Bước 1: Xây dựng biến ngẫu nhiên X1, X2, X3:

```
  X1 = rnorm(100),     # Biến ngẫu nhiên cho X1 (từ 0 đến 1)
  X2 = rnorm(100),     # Biến ngẫu nhiên cho X2 (từ 0 đến 1)
  X3 = rnorm(100),     # Biến ngẫu nhiên cho X3 (từ 0 đến 1)
```

- Xây dựng biến phụ thuộc Y và data frame:

```
set.seed(123)
my_data <- data.frame(
  X1 = rnorm(100),     # Biến ngẫu nhiên cho X1 (từ 0 đến 1)
  X2 = rnorm(100),     # Biến ngẫu nhiên cho X2 (từ 0 đến 1)
  X3 = rnorm(100),     # Biến ngẫu nhiên cho X3 (từ 0 đến 1)
  Y = 2*X1 + 3*X2 + 4*X3 + rnorm(100)  # BIến phụ thuộc Y
)
```

Lệnh set.seed trong R được sử dụng để đặt giá trị hạt giống (seed) cho quá trình sinh dữ liệu ngẫu nhiên.

Bước 2: Kiểm tra dữ liệu: head(my_data)

```
> head(my_data)
           X1         X2         X3          Y
1  1.07401226 -0.7282191  0.3562833 -4.6578505
2 -0.02734697 -1.5404424 -0.6580102 -9.7113837
3 -0.03333034 -0.6930946  0.8552022 -3.0014454
4 -1.51606762  0.1188494  1.1529362 -0.1317315
5  0.79038534 -1.3647095  0.2762746 -0.5624108
6 -0.21073418  0.5899827  0.1441047 -6.4235110
```

- Xây dựng model:

```
model <- lm(Y ~ X1 + X2 + X3, data = my_data)

summary(model)


# Y =. 0.1829 - 0.4335X1 + 0.1661X2 -0.5818X3
```

- Dùng hàm summary để tóm tắt cho đối tượng R, như một mô hình hồi quy tuyến tính (linear regression model), một ma trận, hoặc một vector.

```
> summary(model)

Call:
lm(formula = Y ~ X1 + X2 + X3, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max
-15.0549  -3.7001   0.0544   4.1315  16.3359

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1829     0.5877   0.311    0.756
X1           -0.4335     0.5687  -0.762    0.448
X2            0.1661     0.5742   0.289    0.773
X3           -0.5818     0.5560  -1.046    0.298

Residual standard error: 5.77 on 96 degrees of freedom
Multiple R-squared:  0.01619,   Adjusted R-squared:  -0.01456
F-statistic: 0.5265 on 3 and 96 DF,  p-value: 0.6652
```

1. Giải thích: **Intercept (Chệch):**
    - Estimate: 0.1829
    - Std. Error (Độ lệch chuẩn): 0.5877
    - t value (Giá trị t): 0.311
    - Pr(>|t|) (Giá trị p): 0.756
    - **Estimate:** Đây là ước lượng cho hệ số chệch (intercept), tức là giá trị dự đoán cho biến phụ thuộc (Y) khi tất cả các biến độc lập đều bằng 0.
    - **Std. Error:** Đây là độ lệch chuẩn của ước lượng. Nó đo lường sự biến động hoặc không chắc chắn trong ước lượng.
    - **t value:** Giá trị t được tính bằng cách chia Estimate cho Std. Error. Nếu giá trị t lớn, có thể có bằng chứng cho sự ảnh hưởng đáng kể của biến độc lập đối với biến phụ thuộc.
    - **Pr(>|t|):** Giá trị p liên quan đến giá trị t và cho biết xác suất của việc quan sát giá trị t lớn như vậy (hoặc lớn hơn) nếu giả sử giá trị thực sự của hệ số là 0. Trong trường hợp này, giá trị p là 0.756, nên chúng ta không có đủ bằng chứng để bác bỏ giả thuyết rằng Intercept bằng 0.
2. **X1, X2, X3:**
    - Tương tự như Intercept, mỗi biến độc lập (X1, X2, X3) có các giá trị Estimate, Std. Error, t value và Pr(>|t|).

- **Estimate:** Đây là ước lượng cho hệ số tương ứng với mỗi biến độc lập. Nếu Estimate không bằng 0 (với độ lệch chuẩn nhỏ và giá trị t lớn), có thể có bằng chứng cho tác động của biến độc lập đó đối với biến phụ thuộc.
- **Std. Error, t value, Pr(>|t|):** Các giá trị này tương tự như đã giải thích ở trên.

Từ đó xây dựng mô hình:

Y =. 0.1829 - 0.4335X1 + 0.1661X2 -0.5818X3

Bước 3: Tìm khoảng tin cậy cho các hệ số hồi quy, dùng hàm **confint():**

```
> confint(model)
                2.5 %    97.5 %
(Intercept) -0.9837260 1.3495948
X1          -1.5624665 0.6953686
X2          -0.9736403 1.3057617
X3          -1.6854315 0.5218084
```

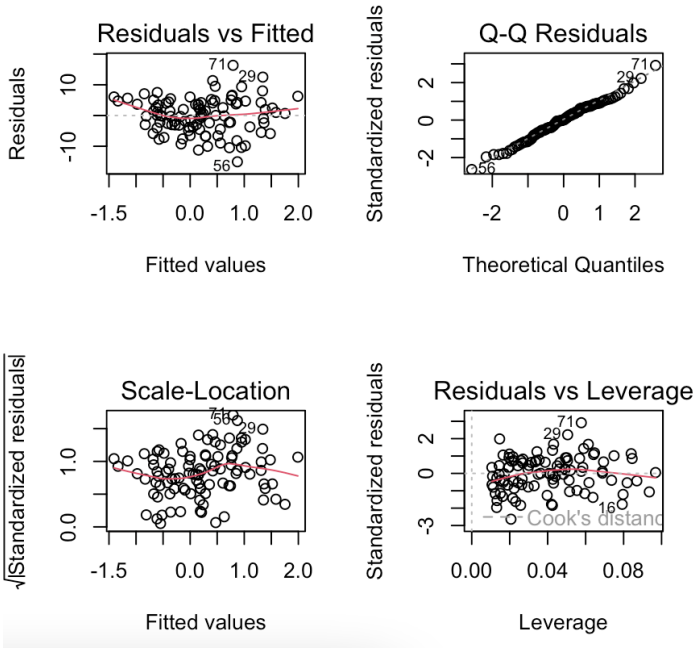
hoảng tin cậy 97,5% cho các hệ số hồi quy cho bởi:

$$-0.9837260 \leq \beta_0 \leq 1.3495948$$
$$-1.5624665 \leq \beta_1 \leq 0.6953686$$
$$-0.9736403 \leq \beta_2 \leq 1.3057617$$
$$-1.6854315 \leq \beta_3 \leq 0.5218084$$

Bước 4: Kiểm tra các giả định của mô hình:

Ta thực hiện phân tích thặng dư để kiểm tra các giả định của mô hình:

```
par(mfrow = c(2, 2))
plot(model)
```



- Đồ thị thứ 1 (*Residuals vs Fitted*) vẽ các giá trị dự báo ($\hat{y}_i$) với các giá trị thặng dư (sai số) tương ứng, dùng để kiểm tra tính tuyến tính của dữ liệu (giả định 1) và tính đồng nhất của các phương sai sai số (giả định 3). Nếu như giả định về tính tuyến tính của dữ liệu **KHÔNG** thỏa, ta sẽ quan sát thấy rằng các điểm thặng dư (residuals) 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 định tính tuyến tính của dữ liệu được thỏa mãn. Để kiểm tra giả định thứ 3 (phương sai đồng nhất) thì các điểm thặng dư phải phân tán đều nhau xung quanh đường thẳng $y=0$.
- Đồ thị thứ 2 (*Normal Q-Q*) cho phép kiểm tra giả định về phân phối chuẩn của các sai số. 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*) vẽ căn bậc hai của các giá trị thặng dư được chuẩn hóa với các giá trị dự báo, được dùng để kiểm tra giả định thứ 3 (phương sai của các sai số là hằng số). 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 định thứ 3 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 định thứ 3 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), 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.

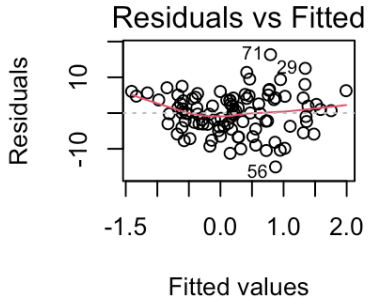
Nhận xét:

- Đồ thị *Normal Q-Q* cho thấy giả định sai số có phân phối chuẩn được thỏa mãn.

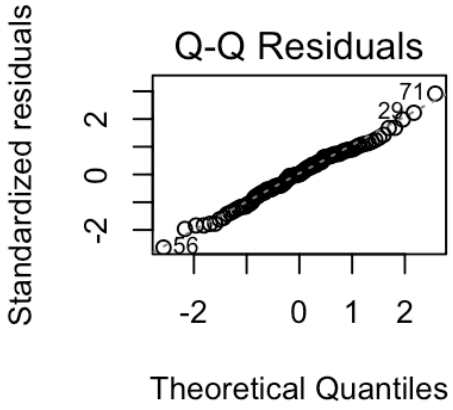
- Đồ thị thứ 1 (*Residuals vs Fitted*) cho thấy giả định về tính tuyến tính của dữ liệu hơi bị vi phạm, ta có thể thấy rằng sự vi phạm này bởi vì mối quan hệ giữa sales□□□□□ và youtube□□□□□□□ là phi tuyến tính.
- Đồ thị thứ 1 và thứ 3 (*Scale - Location*) cho ta thấy rằng giả định về tính đồng nhất của phương sai cũng hơi bị vi phạm. Tuy nhiên, ta cũng thấy này sự vi phạm này tương đối nhỏ và có thể chấp nhận được.

Bước 5: Kiểm tra từng đồ thị:

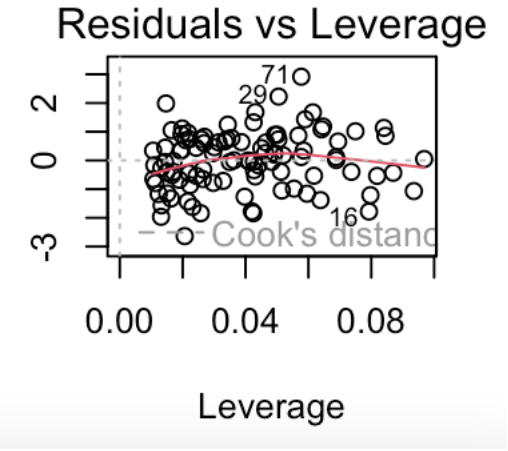
- plot(model, 1)



- plot(model, 2)



- plot(model, 5)


Residuals vs Leverage
71
29
16
Cook's distanc
-3
0
2
0.00
0.04
0.08
Leverage