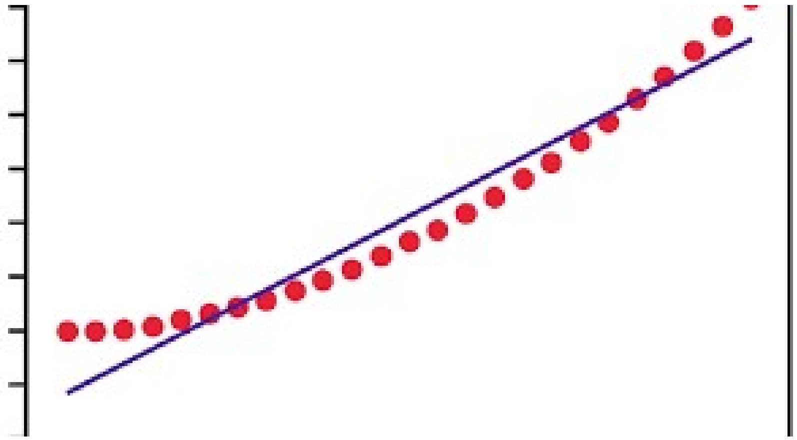
PHÂN TÍCH DỮ LIỆU KINH DOANH

LAB03. PHÂN TÍCH HỒI QUY

(Regression Analysis)



CÔNG CỤ: R, PYTHON, EXCEL

Trình bày: Nguyễn Minh Nhựt

SĐT: 0939013911 - 09851734105

3.1. Tổng quan về phân tích hồi quy

- ***Đặt vấn đề***

Trong nghiên cứu ta thường kiểm định các ***giả thuyết về mối quan hệ*** của *hai* hay ***nhiều biến***.

- ***Trong hồi quy tuyến tính có hai dạng hồi quy:***

SLR (Simple Linear Regression): Hồi quy tuyến tính đơn biến.

$$Y = \beta_0 + \beta_1 X + e$$

MLR (Multiple Linear Regression): Hồi quy tuyến tính đa biến. (Xem slide trang 53)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + e$$

3.2. Ví dụ về hồi quy tuyến tính đơn biến

- ***Thời gian học tập của sinh viên và điểm số***

Time (minutes)	60	120	200	90	10	20	30	50	80
Score	7	8	9	8.5	4.5	5	6.5	7	8

$$Score = 5,436 + 0,022 * Time$$

X là thời gian
y là số điểm

- ***Cách tính như thế nào?***

3.3. Hệ số tương quan r một biến

- Cho X và Y là hai biến chúng ta khảo sát. Hệ số tương quan r của X và Y được tính như sau:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- **Hệ số tương quan nằm trong khoảng $-1 \leq r \leq 1$**
- Nếu **r = 0** ta nói X và Y không tương quan
- Nếu **r > 0** ta nói X và Y tương quan thuận (X tăng Y tăng)
- Nếu **r < 0** ta nói X và Y tương quan nghịch (X tăng Y giảm hoặc ngược lại)
- Càng về gần 1 hoặc -1 độ **tương quan càng mạnh**

3.4. Hệ số tương quan r 2 biến phụ thuộc

- Cho x1, x2 và Y là ba biến chúng ta khảo sát. Hệ số tương quan r của x1, x2 và Y được tính như sau:

$$R = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}}$$

- Tìm hiểu cách tính **hệ số tương quan nhiều biến** (từ 3 biến trở lên)?
- Hệ số tương quan bao nhiêu là phù hợp tại sao?
- Link video tham khảo: https://www.youtube.com/watch?v=Wzj4l2GbTZI

3.5. Chỉ số R Square

- Là **thước đo mô hình nghiên cứu** phù hợp hay không?
- Đồng thời R square còn cho biết các nhân tố trong mô hình phụ thuộc **bao nhiêu phần trăm** trong quá trình nghiên cứu.

$$\boxed{R^2 = \frac{SSR}{SST}}$$

- SSR: Tổng bình phương biến thiên độ lệch ***tiên lượng*** và giá trị trung bình
- SST: Tổng bình phương biến thiên độ lệch ***quan sát*** và giá trị trung bình.
- SSE: Tổng bình phương biến thiên độ lệch ***tiên lượng*** và ***quan sát***

3.5. Chỉ số R Square

- SSR: Tổng bình phương biến thiên độ lệch ***tiên lượng*** và giá trị trung bình

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

3.5. Chỉ số R Square

- SST: Tổng bình phương biến thiên độ lệch ***quan sát*** và giá trị trung bình.

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

3.5. Chỉ số R Square

- SSE: Tổng bình phương biến thiên độ lệch ***tiên lượng*** và ***quan sát***

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

3.6. Chỉ số R Square hiệu chỉnh

- Tại sao cần R square hiệu chỉnh?

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

- Công thức R square hiệu chỉnh

$$R^2 adjusted = 1 - \frac{SSE}{SST}\frac{(n-1)}{(n-k)}$$

- Tại sao thêm (n-1)/(n-k)?
- R square bao nhiêu % là đủ để nghiên cứu?

3.7 Thực hành hồi quy tuyến tính

- ***Ngôn ngữ R***

Bước 1: Import dữ liệu Market Value vào trong ngôn ngữ R.

Bước 2: Dùng hàm lm để biểu diễn mô hình hồi quy tuyến tính linear model.

Bước 3: Chọn giá trị phù hợp cho mô hình hồi quy

```
> reg1 = lm(Market.Value~Square.Feet+House.Age)
> summary(reg1)

Call:
lm(formula = Market.Value ~ Square.Feet + House.Age)

Residuals:
   Min     1Q Median     3Q    Max
 -9164  -4220  -2175   2487  30968

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 47331.382  13884.347   3.409  0.00153 **
Square.Feet    40.911      6.697   6.109 3.65e-07 ***
House.Age    -825.161    607.313  -1.359  0.18205
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7212 on 39 degrees of freedom
Multiple R-squared:  0.5558,	Adjusted R-squared:  0.533
F-statistic:  24.4 on 2 and 39 DF,  p-value: 1.344e-07
```

Nhìn vào Coefficients ta thấy Pr(>|t|) của House.Age > 0.05. Nên ta có thể thực hiện việc chọn lại mô hình loại đi House.Age xem kết quả có khả quan hơn hay không?

3.7 Thực hành hồi quy tuyến tính

- ***Ngôn ngữ R***

Bước 4: Thực hiện mô hình hồi quy tuyến tính với giá trị Square.Feet

```
> req2 = lm(Market.Value~Square.Feet)
> summary(req2)

Call:
lm(formula = Market.Value ~ Square.Feet)

Residuals:
   Min     1Q Median     3Q    Max
 -8067  -4327  -1923   3097  32634

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 32673.220   8831.951   3.699  0.00065 ***
Square.Feet    35.036      5.167   6.780  3.8e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7288 on 40 degrees of freedom
Multiple R-squared:  0.5347,    Adjusted R-squared:  0.5231
F-statistic: 45.97 on 1 and 40 DF,  p-value: 3.798e-08
```

Nhận sét về độ tương quan R-squared, Adjusted R-squared ta thấy mô hình sau là phù hợp.

Nên ta có

Market.Value = 32673.220+35.036*Square.Feet

3.7 Thực hành hồi quy tuyến tính

- ***Ngôn ngữ R***

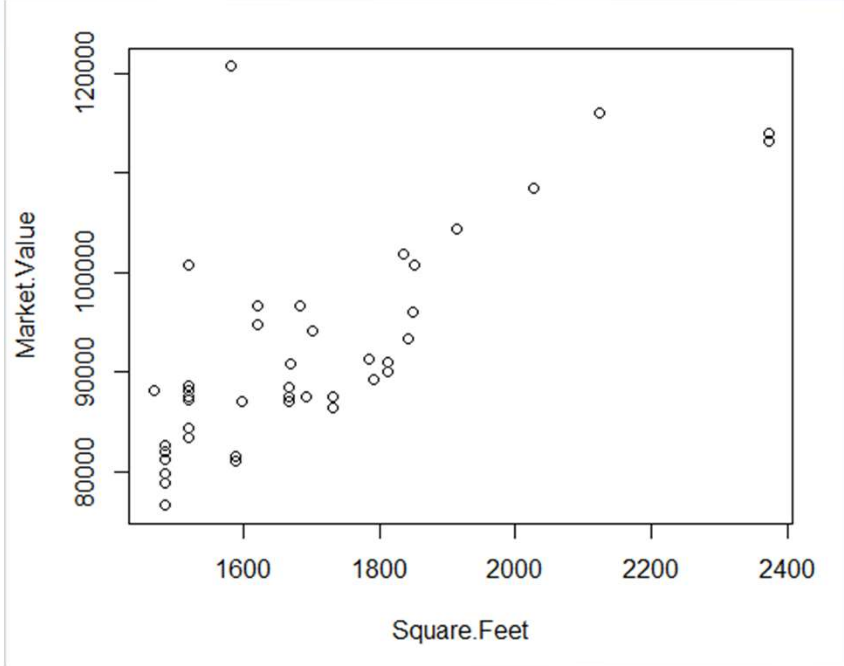
Bước 4: Thực hiện mô hình hồi quy tuyến tính với giá trị Square.Feet

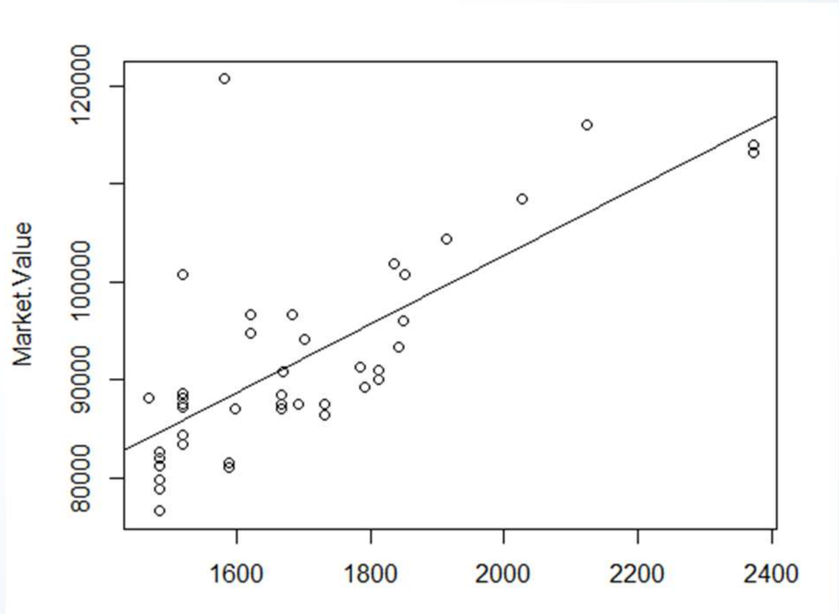
Bước 5: Vẽ hình biểu diễn mối liên hệ Market.Value và Square.Feet

```
>plot(y~x)
```

Bước 6: Biểu diễn phương trình hồi qui Market.Value và Square.Feet

```
>abline(lm(y~x))
```





3.7 Thực hành hồi quy tuyến tính

- *Ngôn ngữ Python*

Các thư viện cần thiết

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
```

Bước 1: Import dữ liệu Market Value vào trong ngôn ngữ Python.

Bước 2: Lấy ra biến phụ thuộc Y và biến độc lập X

```
x, y =
np.array(data["Square Feet"]).reshape((-1, 1)),
np.array( data["Market Value"]).reshape((-1, 1))
```

Bước 3: Dùng hàm LinearRegression() trong thư viện sklearn() để đưa ra mô hình theo biến x và y.

```
model = LinearRegression()
model.fit(x,y)
```

3.7 Thực hành hồi quy tuyến tính

- *Ngôn ngữ Python*

Tham khảo thư viện

https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html

https://www.statology.org/sklearn-linear-regression-summary/

Lấy các giá trị thông dụng của mô hình hồi quy tuyến tính

Giá trị thông dụng	Mô tả giá trị
`model.intercept_`	Hệ số chắn
`model.coef_`	Hệ số thành phần
`model.score(x, y)`	Giá trị R square

Bước 4. Từ các giá trị thông dụng cho mô hình hồi quy tuyến tính hãy lập ra một bảng giống như ngôn ngữ R

3.7 Thực hành hồi quy tuyến tính

- *Ngôn ngữ Python*

Trường hợp hồi quy tuyến tính đa biến

```
x, y =
np.array(data[["Square Feet","House Age"]]).reshape((-1, 2)),
np.array( data["Market Value"]).reshape((-1, 1))
```

Các bước còn lại làm tương tự hồi quy tuyến tính đơn biến

Thực hiện với tập dữ liệu Colleges and Universities

3.8 Mô hình hồi quy phi tuyến tính

- ***Giới thiệu về mô hình phi tuyến tính***

- ***Mô hình hồi quy phi tuyến*** khi ít nhất một tham số là hàm số phi tuyến. Qua **một số ví dụ** như sau:

$$\log(Y) = \beta_0 + \beta_1 X + e \qquad Y = \exp(\beta_0 + \beta_1 X + e)$$

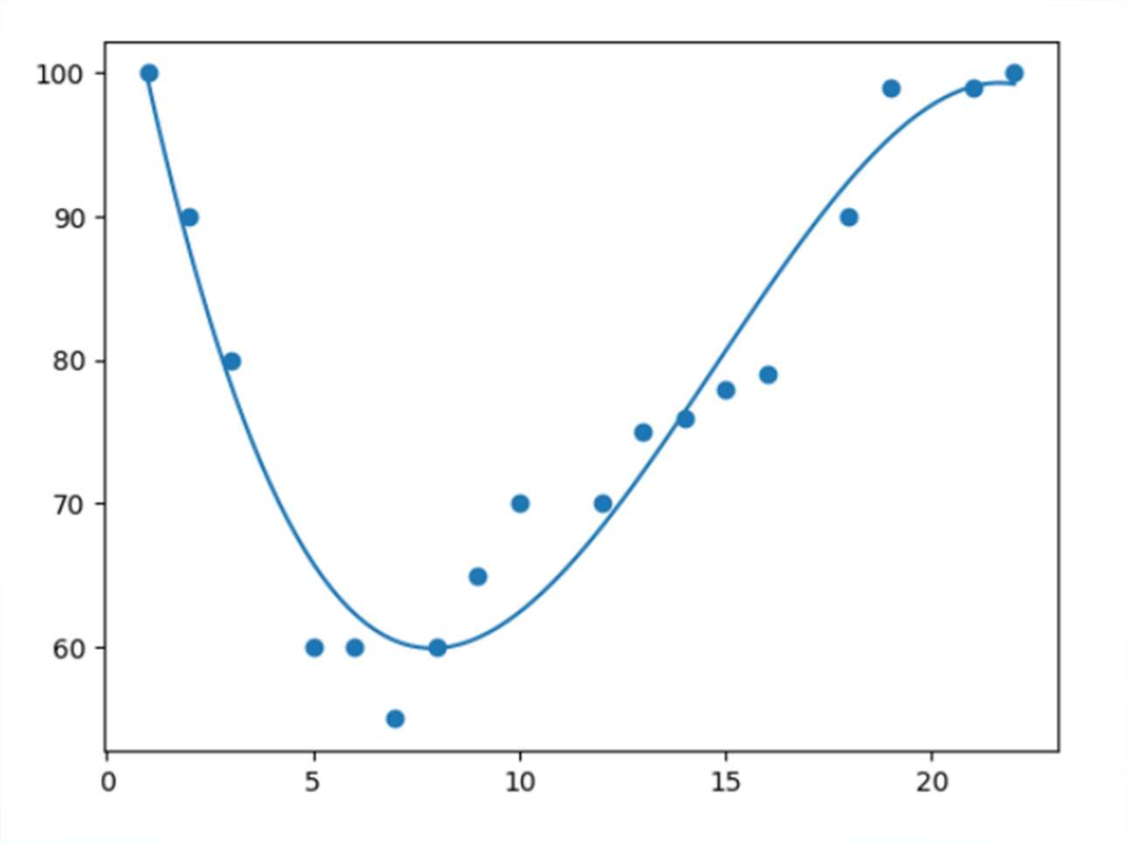
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2{}^2 + e$$

- ***Xác định hàm phi tuyến nào là hợp lý?***

3.8 Mô hình hồi quy phi tuyến tính

- ***Giới thiệu về mô hình phi tuyến tính***

Minh họa hàm phi tuyến



3.9 Một số hàm phi tuyến trong ngôn ngữ R, Python

- ***Một số hàm phi tuyến trong ngôn ngữ R***

Tên hàm	Formula	Giải thích
poly	poly(x,degree=z, raw=TRUE)	Hàm đa thức
logarithms	log(x)	Hàm ln()
sin	sin(x)	Hàm sin
cos	cos(x)	Hàm cos
	log(x,base=y)	logy(x)

Tham khảo: https://econometricsr.hieunguyenphi.com/ham-hi-quy-phi-tuyn.html

3.10 Mô hình hồi quy Logistic

- ***Phân tích hồi quy Logistic*** là một kỹ thuật thống kê xem mối liên hệ giữa **một biến độc lập** (là biến số hoặc biến phân loại) với **một biến phụ thuộc** là dạng nhị phân (0 hoặc 1).
- ***Dạng tuyến tính của phương trình hồi quy logistic:***

$$Y = \beta_0 + \beta_1 X + e$$

- **Trong đó Y là biến phụ thuộc nhị phân, X là biến độc lập.**
- Vì Y là một **biến nhị phân (0 và 1)** tuân theo luật phân phối nhị thức. Do đó mô hình hồi quy tuyến tính thông thường không thể áp dụng được.

3.10 Mô hình hồi quy Logistic

- ***Theo luật phân phối nhị thức***

Gọi P là ***xác suất xảy ra sự kiện A*** và 1-P là ***biến cố đối của sự kiện A.***

- Chỉ số **ODDs = P/(1-P)**
  - Nếu **ODDs > 1** xác suất biến cố A xảy ra **khả năng cao** hơn biến cố đối của nó.
  - Nếu **ODDs < 1** xác suất biến cố A xảy ra **khả năng thấp** hơn biến cố đối của nó.
  - Nếu **ODDs = 1** xác suất biến cố A xảy ra **khả năng bằng** biến cố đối của nó.
- Từ chỉ số ODDs ta chuyển Y trong **phương trình hồi quy tuyến tính→ log(ODDs).** Phương trình hồi quy logistic có dạng khác:

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X + e$$

3.11 Thực hiện mô hình hồi quy logistic trên ngôn ngữ R

- Tập dữ liệu **Graduate School Survey.**
- Trong ngôn ngữ R dùng hàm `glm()` để phân tích hồi quy logistic. Với tham số `family = binomial.`

```
> logistic <- glm(Plan.to.attend.graduate.school~Undergraduate.GPA,family = binomial)
> summary(logistic)

Call:
glm(formula = Plan.to.attend.graduate.school ~ Undergraduate.GPA,
    family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5976  -0.8444   0.2483   0.7797   1.8741

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -10.909      4.585  -2.379   0.0174 *
Undergraduate.GPA    3.593      1.463   2.456   0.0140 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 39.429  on 29  degrees of freedom
Residual deviance: 29.494  on 28  degrees of freedom
AIC: 33.494

Number of Fisher Scoring iterations: 5
```

Từ **kết quả ta được phương trình** hồi quy như sau:

$$\log\left(\frac{P}{1-P}\right) = -10.909 + 3.593 * UndergraduateGPA + \varepsilon$$

=> Ta suy ra được kết quả như sau:

$$\left(\frac{p}{1-p}\right) = e^{-10,909+3,593*(undergraduateGPA)}$$

3.11 Thực hiện mô hình hồi quy logistic trên ngôn ngữ R

- Với ví dụ 3.11 làm ở phía trên hãy thử với trường hợp Odd0 và Odd1 tức UndergraduateGPA = 0 và 1 rồi lập tỉ lệ giữa hai phần này để xem xét khi tăng 1 điểm tỉ lệ tham dự lễ tốt nghiệp là bao nhiêu lần?

- Ta có $\left(\frac{p}{1-p}\right) = e^{-10,909+3,593*(undergraduateGPA)}$. Ta đặt hệ số p/(1-p) là odd
- Đặt Odd_0 và undergraduateGPA = 0 thì $Odd_0 = e^{-10,909}$
- Đặt Odd_1 và undergraduateGPA = 1 thì $Odd_1 = e^{-10,909+3,593}$
- Tỉ số $\frac{Odd_1}{Odd_0} = \frac{e^{-10,909+3,593}}{e^{-10,909}} \approx 36,359$
- Lúc này ta có thể diễn dịch, cứ điểm undergraduateGPA lên 1 đơn vị thì khả năng đi dự tốt nghiệp tăng lên tỉ lệ có kế hoạch tham dự lễ tốt nghiệp tăng lên 36,359 lần, nếu tăng 0,1 điểm GPA thì tỉ lệ tham dự lễ tốt nghiệp tăng lên 3,6359 lần

$$\log\left(\frac{P}{1-P}\right) = -10.909 + 3.593 * UndergraduateGPA + \varepsilon$$

3.12 Thực hiện mô hình hồi quy logistic trên Python

- Tập dữ liệu **Graduate School Survey.**
- Để thực hiện hồi quy logistic trên **Python** dùng thông qua hàm **`LogisticRegression().`**
- Cũng giống như **`LinearRegresion()`** cách import cũng tương tự.

`from sklearn.linear_model import LogisticRegression`

- Chuẩn bị dữ liệu Y là giá trị nhị phân, X là giá trị số hoặc phân loại.
- Thực hiện các bước như **hồi quy tuyến tính đơn biến, đa biến.**
- **Cách gọi tham khảo ví dụ bên dưới:**

```python
model = LogisticRegression(solver='liblinear', C=10.0, random_state=0)
model.fit(x, y)
```

- Về tìm các giá trị kiểm dịnh làm tương tự như **LinearRegression.**
- So sánh kết quả Odd1 và Odd0 cũng làm tương tự như trong ngôn ngữ R.