BÀI TẬP TUẦN 6: HỒI QUY TUYẾN TÍNH

1. Dữ liệu tuyensinhdaihoc

Nạp thư viện

In [85]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [86]:
```
# Đọc file
df = pd.read_csv('../data/dulieuxettuyendaihoc.csv', header=0, delimiter=',', encoding='utf-8')
```

In [87]:
```
'''
Phân tích hồi quy tuyến tính
Mục đích: Phân tích tác động hay ảnh hưởng giữa các yếu tố đến mục tiêu (thường
đùng cho các biến (yếu tố) định lượng)
Thường vẽ biểu đồ Scatter để khám phá mối tương quan tuyến tính trước khi khám
phá quan hệ hồi quy tuyến tính
PHƯƠNG PHÁP
1. Xác định biến độc lập (yếu tố) và biến phụ thuộc (mục tiêu)
2. Ghi ra phương trình hồi quy tuyến tính tổng quát y = f(x)
3. Chạy dữ liệu mô hình
4. Đọc các giá trị quan trọng và kết luận
5. Dự báo giá trị biến phụ thuộc khi biết trước giá trị biến độc lập
'''
```

Out[87]:
```
'\nPhân tích hồi quy tuyến tính\nMục đích: Phân tích tác động hay ảnh hưởng giữa các yếu tố đến mục tiêu
(thường\nđùng cho các biến (yếu tố) định lượng)\nThường vẽ biểu đồ Scatter để khám phá mối tương qua
n tuyến tính trước khi khám\nphá quan hệ hồi quy tuyến tính\nPHƯƠNG PHÁP\n1. Xác định biến độc lập
(yếu tố) và biến phụ thuộc (mục tiêu)\n2. Ghi ra phương trình hồi quy tuyến tính tổng quát y = f(x)\n3. Ch
ạy dữ liệu mô hình\n4. Đọc các giá trị quan trọng và kết luận\n5. Dự báo giá trị biến phụ thuộc khi biết trư
ớc giá trị biến độc lập\n'
```

In [88]:
```
# Hãy cho biết sự ảnh hưởng của điểm học kì 1 năm lớp 12 đến điểm học kì 2 năm lớp 12
#import statsmodels.api as sm
#pip install statsmodels –cài đặt thư viện này để sử dụng thư viện
#statsmadels.api
import statsmodels.api as sm
#linear regression
```

In [89]:
```
'''
1. Biến độc lập: học kì 1 (T5)
Biến phụ thuộc : học kì 2 (T6)
2. T6 = f(T5) = Ao + A1*T5 + epsilon
3. Chạy mô hình
4. Đọc và hiểu kết quả
'''
```

Out[89]: '\n1. Biến độc lập: học kì 1 (T5)\nBiến phụ thuộc : học kì 2 (T6)\n2. T6 = f(T5) = Ao + A1*T5 + epsilon\n3. Chạy mô hình\n4. Đọc và hiểu kết quả\n'

In [90]:
```
# adding a constant
X_with_constant = sm.add_constant(df[["T5"]].values)
y = df[['T6']].values
# performing the regression
result = sm.OLS(y,X_with_constant).fit()
# Result of statsmodels
print(result.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.606
Model:                            OLS   Adj. R-squared:                  0.602
Method:                 Least Squares   F-statistic:                     151.0
Date:                Fri, 06 Oct 2023   Prob (F-statistic):           1.48e-21
Time:                        07:49:43   Log-Likelihood:                -125.76
No. Observations:                 100   AIC:                             255.5
Df Residuals:                      98   BIC:                             260.7
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.1130      0.402      5.257      0.000       1.315       2.911
x1             0.7182      0.058     12.286      0.000       0.602       0.834
==============================================================================
Omnibus:                        1.387   Durbin-Watson:                   1.738
Prob(Omnibus):                  0.500   Jarque-Bera (JB):                0.860
Skew:                          -0.104   Prob(JB):                        0.650
Kurtosis:                       3.404   Cond. No.                         32.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [91]:
```
'''
Căn cứ vào điểm số T5 sẽ giải thích được 60% sự thay đổi của T6 (Adj. R-squared)
Prob (F-statistic) < 0.05: cho biết mô hình có khả năng phù hợp cho tổng thể [nó
là p-value]
const (2,11) chính là A0
x1 là A1
=> T6 = 2.113 + 0.7182 * T5
P>|t| = 0.000 rất nhỏ < 5% => x1 tương ứng với T5 (x` = T5) => biến T5 có ý nghĩa
thống kê trong phương trình này hay nói T5 có
ý nghĩa tham gia đánh giá tác động tới biến T6
'''
'''
Bước 5: Giả sử T5 = 7.5, dự báo T6 = 7.499
'''
```

Out[91]: '\nBước 5: Giả sử T5 = 7.5, dự báo T6 = 7.499\n'

In [92]:
```
df = df[['T5', 'T6', 'GT', 'DT', 'KV', 'KT', 'NGONNGU', 'TOANLOGICPHANTICH', 'GIAIQUYETVANDE', '
df.rename(columns={
    'TOANLOGICPHANTICH': 'LOGIC',
    'GIAIQUYETVANDE': 'UNGXU',
    'DINHHUONGNGHENGHIEP': 'HUONGNGHIEP'
}, inplace=True)
```

In [93]:
```
# Khám phá sự ảnh hưởng của T6 đến điểm thi LOGIC
# adding a constant
X_with_constant = sm.add_constant(df[["T6"]].values)
y = df[['LOGIC']].values
# performing the regression
result = sm.OLS(y,X_with_constant).fit()
```

In [94]:
```
# Result of statsmodels
print(result.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.091
Model:                            OLS   Adj. R-squared:                  0.082
Method:                 Least Squares   F-statistic:                     9.798
Date:                Fri, 06 Oct 2023   Prob (F-statistic):            0.00230
Time:                        07:49:43   Log-Likelihood:                -142.46
No. Observations:                 100   AIC:                             288.9
Df Residuals:                      98   BIC:                             294.1
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.6287      0.529      4.965      0.000       1.578       3.679
x1             0.2344      0.075      3.130      0.002       0.086       0.383
==============================================================================
Omnibus:                       12.364   Durbin-Watson:                   1.974
Prob(Omnibus):                  0.002   Jarque-Bera (JB):               14.780
Skew:                           0.671   Prob(JB):                     0.000617
Kurtosis:                       4.322   Cond. No.                         37.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [95]:
```
'''
Adj. R-squared = 8%: quá ít, không thể giải thích cho điểm LOGIC, nói cách khác k
dựa T6 để gthich dc
Prob (F-statistic) = 0.00230 => phù hợp, bé hơn 0.05, có ý nghĩa thống kê
LOGIC = 2.6287 + 0.2344 * T6
P>|t| = 0.000 < 0.05: T6 có ý nghĩa thống kê
Giả sử T6 = 7.0
=> LOGIC = 4.2695
'''
```

Out[95]:
```
'\nAdj. R-squared = 8%: quá ít, không thể giải thích cho điểm LOGIC, nói cách khác k\ndựa T6 để gthich dc\nProb (F-statistic) = 0.00230 => phù hợp, bé hơn 0.05, có ý nghĩa thống kê\nLOGIC = 2.6287 + 0.2344 * T6\nP>|t| = 0.000 < 0.05: T6 có ý nghĩa thống kê\nGiả sử T6 = 7.0\n=> LOGIC = 4.2695\n'
```

In [96]:
```
# Khám phá sự ảnh hưởng của T6 đến điểm thi LOGIC
# adding a constant
X_with_constant = sm.add_constant(df[["T6"]].values)
y = df[['LOGIC']].values
# performing the regression
result = sm.OLS(y,X_with_constant).fit()
# Result of statsmodels
print(result.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.091
Model:                            OLS   Adj. R-squared:                  0.082
Method:                 Least Squares   F-statistic:                     9.798
Date:                Fri, 06 Oct 2023   Prob (F-statistic):            0.00230
Time:                        07:49:43   Log-Likelihood:                -142.46
No. Observations:                 100   AIC:                             288.9
Df Residuals:                      98   BIC:                             294.1
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.6287      0.529      4.965      0.000       1.578       3.679
x1             0.2344      0.075      3.130      0.002       0.086       0.383
==============================================================================
Omnibus:                       12.364   Durbin-Watson:                   1.974
Prob(Omnibus):                  0.002   Jarque-Bera (JB):               14.780
Skew:                           0.671   Prob(JB):                     0.000617
Kurtosis:                       4.322   Cond. No.                         37.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [97]:
```
'''
Adj. R-squared = 8%: quá ít, không thể giải thích cho điểm LOGIC, nói cách khác k
dựa T6 để gthich dc
Prob (F-statistic) = 0.00230 => phù hợp, bé hơn 0.05, có ý nghĩa thống kê
LOGIC = 2.6287 + 0.2344 * T6
P>|t| = 0.000 < 0.05: T6 có ý nghĩa thống kê
Giả sử T6 = 7.0
=> LOGIC = 4.2695
'''
```

Out[97]:
```
'\nAdj. R-squared = 8%: quá ít, không thể giải thích cho điểm LOGIC, nói cách khác k\ndựa T6 để gthich dc\nProb (F-statistic) = 0.00230 => phù hợp, bé hơn 0.05, có ý nghĩa thống kê\nLOGIC = 2.6287 + 0.2344 * T6\nP>|t| = 0.000 < 0.05: T6 có ý nghĩa thống kê\nGiả sử T6 = 7.0\n=> LOGIC = 4.2695\n'
```

In [98]:
```
X_with_constant = sm.add_constant(df[["T5","T6"]].values)
y = df[['LOGIC']].values
# performing the regression
result = sm.OLS(y,X_with_constant).fit()
# Result of statsmodels
print(result.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.097
Model:                            OLS   Adj. R-squared:                  0.079
Method:                 Least Squares   F-statistic:                     5.226
Date:                Fri, 06 Oct 2023   Prob (F-statistic):            0.00699
Time:                        07:49:43   Log-Likelihood:                -142.11
No. Observations:                 100   AIC:                             290.2
Df Residuals:                      97   BIC:                             298.0
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.7072      0.539      5.026      0.000       1.638       3.776
x1            -0.0913      0.110     -0.828      0.410      -0.310       0.128
x2             0.3115      0.120      2.606      0.011       0.074       0.549
==============================================================================
Omnibus:                       14.098   Durbin-Watson:                   2.005
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               17.272
Skew:                           0.745   Prob(JB):                     0.000178
Kurtosis:                       4.387   Cond. No.                         52.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [99]:
```
'''
1. Độc lập T5,T6 phụ thuộc Logic
2. Logic f(T5,T6)
Logic = A0 + A1*T5 + A2*T6 * epsilon = 2.7072 - 0.0913*T5 + 0.3115 * T6 + epsilon
'''
```

Out[99]:
```
'\n1. Độc lập T5,T6 phụ thuộc Logic\n2. Logic f(T5,T6)\nLogic = A0 + A1*T5 + A2*T6 * epsilon = 2.707
2 - 0.0913*T5 + 0.3115 * T6 + epsilon\n'
```

In [100]:
```
# Hãy phân tích sự ảnh hưởng của điểm toán học kì 1,2 năm lớp 12 đến điểm NGONNGU
X_with_constant = sm.add_constant(df[["T5","T6"]].values)
y = df[['NGONNGU']].values
# performing the regression
result = sm.OLS(y,X_with_constant).fit()
# Result of statsmodels
print(result.summary())
'''
Adj. R-squared = 1% : quá ít, k gthich dc gì
Prob (F-statistic): 7%
'''
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.006
Model:                            OLS   Adj. R-squared:                 -0.014
Method:                 Least Squares   F-statistic:                    0.3109
Date:                Fri, 06 Oct 2023   Prob (F-statistic):              0.734
Time:                        07:49:43   Log-Likelihood:                -176.45
No. Observations:                 100   AIC:                             358.9
Df Residuals:                      97   BIC:                             366.7
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.8860      0.759      5.117      0.000       2.379       5.393
x1             0.1114      0.155      0.716      0.475      -0.197       0.420
x2            -0.1289      0.169     -0.765      0.446      -0.463       0.206
==============================================================================
Omnibus:                        3.571   Durbin-Watson:                   1.774
Prob(Omnibus):                  0.168   Jarque-Bera (JB):                2.978
Skew:                           0.314   Prob(JB):                        0.226
Kurtosis:                       2.433   Cond. No.                         52.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Out[100]: '\nAdj. R-squared = 1% : quá ít, k gthich dc gì\nProb (F-statistic): 7%\n'

In [101]:
```
# Đánh giá mức độ tác động giữa các yếu tố đến 1 đối tượng bằng phân tích hồi quy tuyến tính
# Hãy cho biết mức độ tác động của T5, T6 (độc lập) đến điểm LOGIC (phụ thuộc)
# adding a constant
X = df[["T5","T6"]].values
y = df[['LOGIC']].values
# performing the regresssion
result = sm.OLS(y,X).fit()
# result of statsmodels
print(result.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared (uncentered):                   0.934
Model:                            OLS   Adj. R-squared (uncentered):              0.933
Method:                 Least Squares   F-statistic:                              694.9
Date:                Fri, 06 Oct 2023   Prob (F-statistic):                    1.30e-58
Time:                        07:49:43   Log-Likelihood:                         -153.68
No. Observations:                 100   AIC:                                      311.4
Df Residuals:                      98   BIC:                                      316.6
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.0063      0.121      0.052      0.959      -0.234       0.247
x2             0.5934      0.118      5.031      0.000       0.359       0.827
==============================================================================
Omnibus:                        9.328   Durbin-Watson:                   1.966
Prob(Omnibus):                  0.009   Jarque-Bera (JB):                9.293
Skew:                           0.636   Prob(JB):                      0.00960
Kurtosis:                       3.783   Cond. No.                         14.6
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [102]:
```
'''
x2 = |0.5934| => T6 tác động mạnh hơn so với T5 (x1 = |0.0063|)
vì x2 dương nên tác động tích cực, còn âm mới tác động tiêu cực (nghịch biến)
'''
# Đánh giá mức độ tác động giữa các yếu tố đến 1 đối tượng bằng phân tích hồi quy tuyến tính
# Hãy cho biết mức độ tác động của T5, T6 đến điểm UNGXU
# adding a constant
X = df[["T5","T6"]].values
y = df[['UNGXU']].values
# performing the regresssion
result = sm.OLS(y,X).fit()
# result of statsmodels
print(result.summary())
'''
x2 = |0.5318| => T6 tác động mạnh hơn so với T5 (x1 = |0.1519|)
vì x2 dương nên tác động tích cực (đồng biến), còn âm mới tác động tiêu cực
(nghịch biến)
'''
```

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                      y   R-squared (uncentered):                   0.926
Model:                            OLS   Adj. R-squared (uncentered):              0.924
Method:                 Least Squares   F-statistic:                              612.1
Date:                Fri, 06 Oct 2023   Prob (F-statistic):                    4.24e-56
Time:                        07:49:43   Log-Likelihood:                         -172.41
No. Observations:                 100   AIC:                                      348.8
Df Residuals:                      98   BIC:                                      354.0
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.1519      0.146      1.039      0.301      -0.138       0.442
x2             0.5318      0.142      3.738      0.000       0.249       0.814
==============================================================================
Omnibus:                        0.142   Durbin-Watson:                   1.874
Prob(Omnibus):                  0.931   Jarque-Bera (JB):                0.323
Skew:                          -0.025   Prob(JB):                        0.851
Kurtosis:                       2.726   Cond. No.                         14.6
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Out[102]: '\nx2 = |0.5318| => T6 tác động mạnh hơn so với T5 (x1 = |0.1519|)\nvì x2 dương nên tác động tích cực (đồng biến), còn âm mới tác động tiêu cực\n(nghịch biến)\n'