

BÁO CÁO ĐỒ ÁN 01: REGRESSION

TÊN MÔN HỌC: NHẬP MÔN HỌC MÁY

ĐỀ TÀI: CHI PHÍ SỬ DỤNG DỊCH VỤ Y TẾ

GIẢNG VIÊN: NGUYỄN TIẾN HUY

THỨ TỰ NHÓM: 07

THÀNH VIÊN:

- 18120184 Nguyễn Nguyên Khang
- 18120189 Trần Đăng Khoa
- 18120264 Nguyễn Duy Vũ
- 18120283 Nguyễn Chiêu Bản
- 18120286 Nguyễn Quốc Bảo

PHÂN CÔNG:

Công việc	Thực hiện	Mức độ hoàn thành
Khám phá dữ liệu cơ bản	Vũ	100%
Tiền xử lý dữ liệu	Vũ	100%
Mô hình hóa dữ liệu	Bản, Bảo	100%
Phân tích dữ liệu tìm Insight	Khang, Khoa	100%

Mục Lục

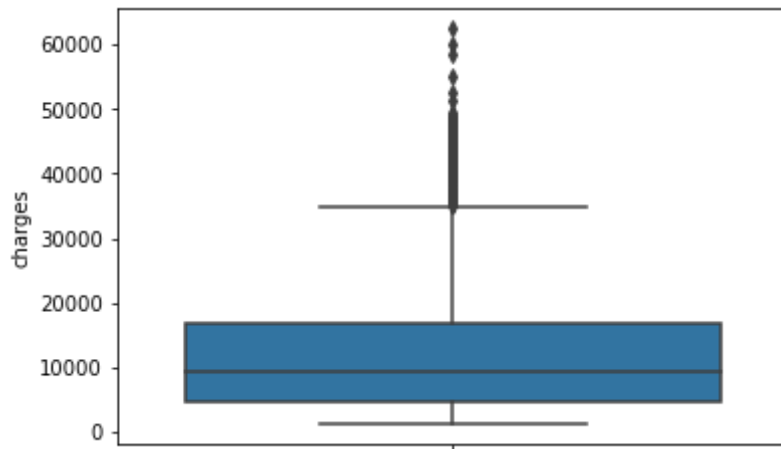
- I. Phân tích dữ liệu
 - 1. Vẽ biểu đồ một biến và nhận xét
 - 2. Vẽ biểu đồ các biến tương quan và nhận xét
 - 3. VIF
 - 4. Insight: Sex có ảnh hưởng đến Smoker?
 - 5. Insight: Trung bình của 'age', 'bmi', 'children' có bằng nhau đôi một
 - 6. Insight: Sự phụ thuộc của charges vào sex, smoker, age, bmi, children
- II. Thuật toán sử dụng
 - 1. Cách thức đánh giá mô hình
 - 2. Thuật toán SVR
 - a. Giới thiệu SVR
 - b. Sử dụng SVR từ thư viện Scikit-learn
- III. Tham khảo

I. Phân tích dữ liệu

1. Vẽ biểu đồ một biến và nhận xét

In [13]: `Biến charges:`

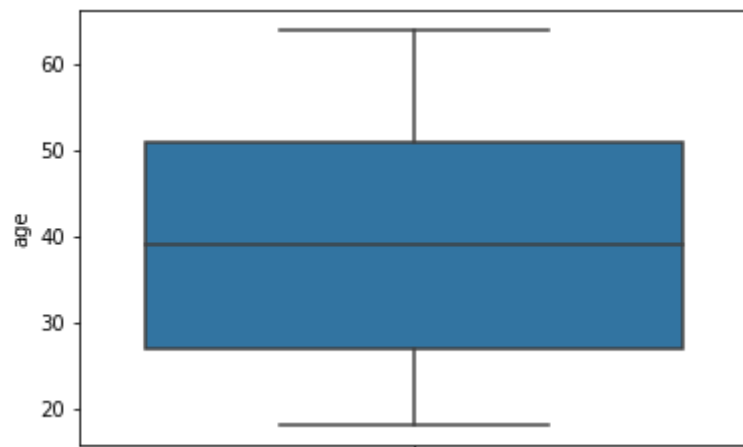
Out[13]: `<AxesSubplot:ylabel='charges'>`



Nhận xét: Biến charges có phân bố bị lệch trái, nhiều outlier

In [14]: `Biến age:`

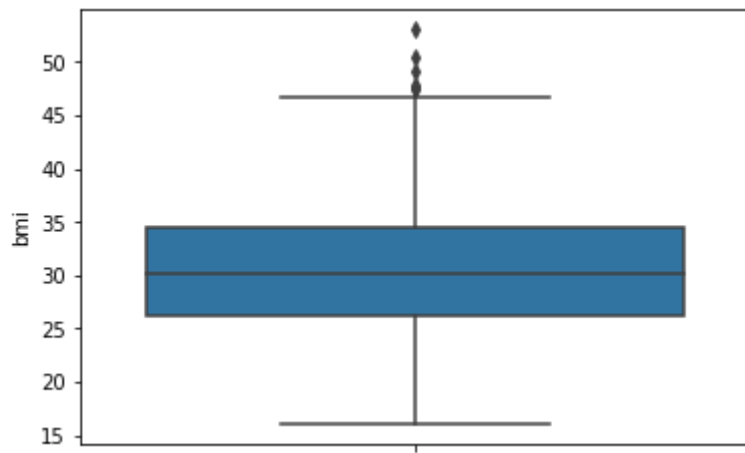
Out[14]: `<AxesSubplot:ylabel='age'>`



Nhận xét: Biến age có phân bố chuẩn

In [15]: `Biến bmi:`

Out[15]: `<AxesSubplot:ylabel='bmi'>`

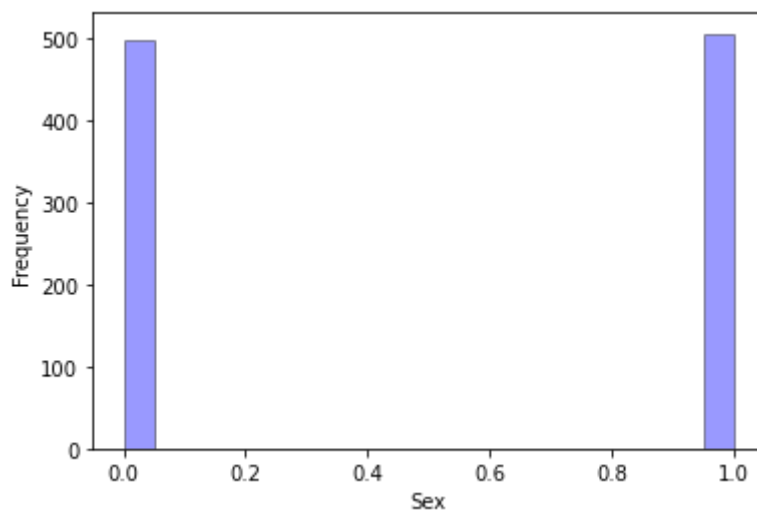


Nhận xét: Biến bmi có phân bố chuẩn, tồn tại outlier

In [18]: **Biến sex:**

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2551: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

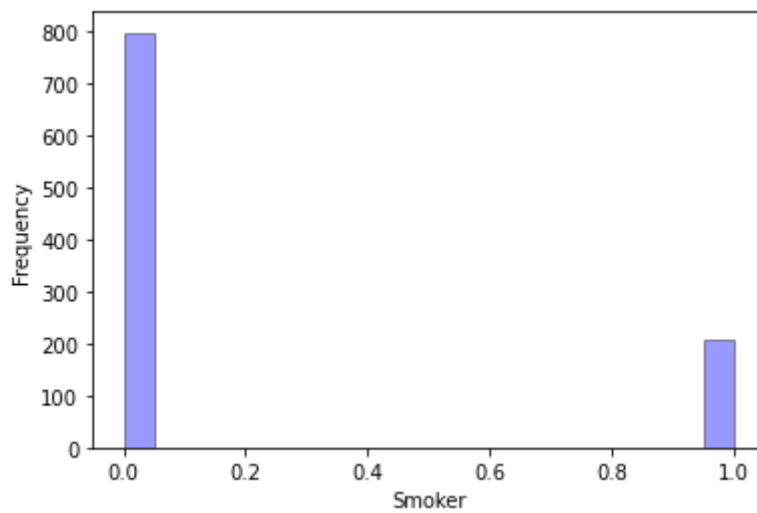
Out[18]: Text(0, 0.5, 'Frequency')



Nhận xét: Tỷ lệ nam nữ bằng nhau

In [19]: **Biến smoker:**

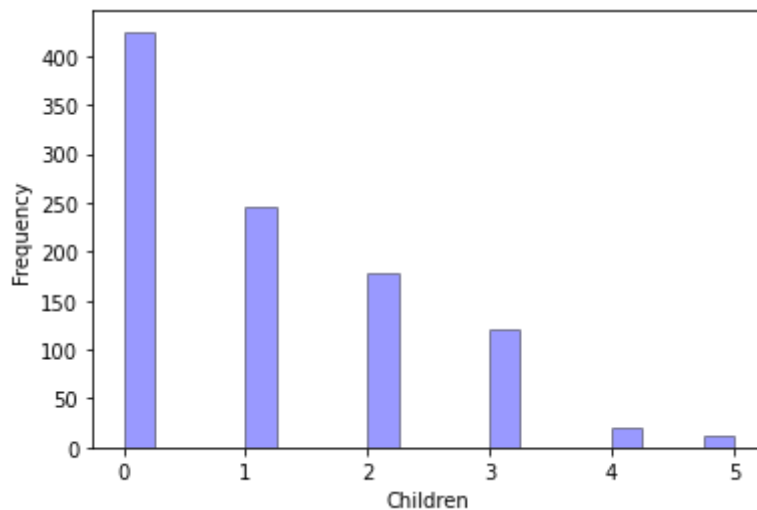
Out[19]: Text(0, 0.5, 'Frequency')



Nhận xét: Tỷ lệ người không hút thuốc gấp 4 lần người hút thuốc

In [20]: `Biến children:`

Out[20]: `Text(0, 0.5, 'Frequency')`



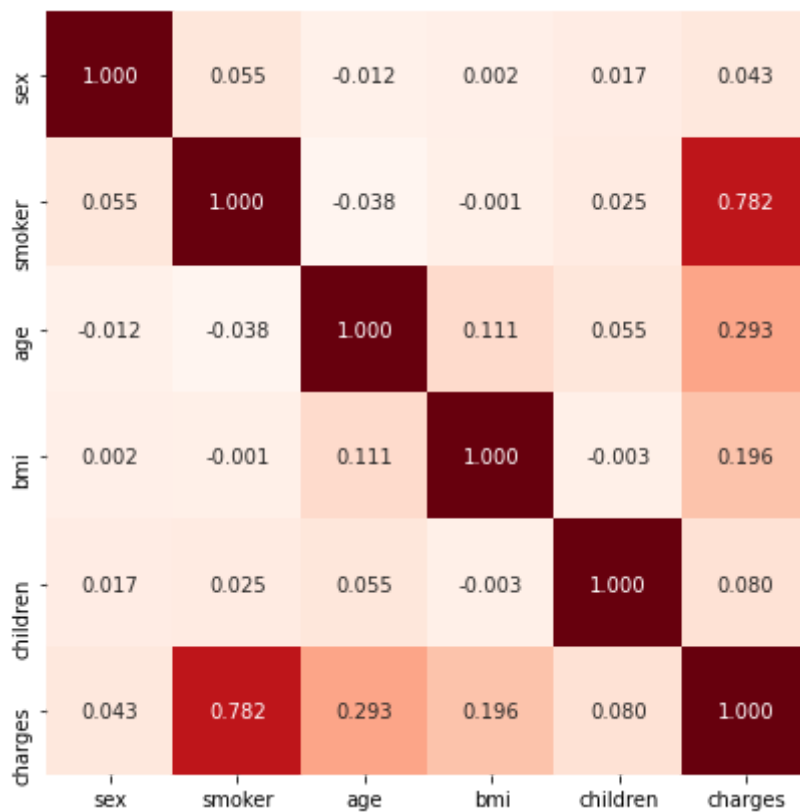
Nhận xét: Tỷ lệ người có càng nhiều con giảm dần

2. Vẽ biểu đồ các biến tương quan và nhận xét

Trước tiên, ta tính ma trận tương quan

In [21]:

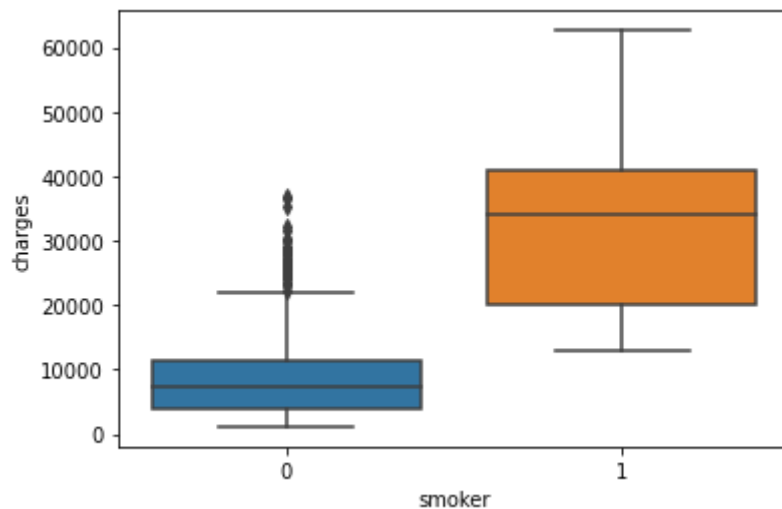
Out[21]: `<AxesSubplot:>`



Có thể thấy những thuộc tính như age (yếu), bmi (yếu), smoker (mạnh) có tương quan với thuộc tính charges

In [22]: Biểu đồ thể hiện sự mất tiền vào chi phí y tế của người có hút thuốc

Out[22]: <AxesSubplot:xlabel='smoker', ylabel='charges'>

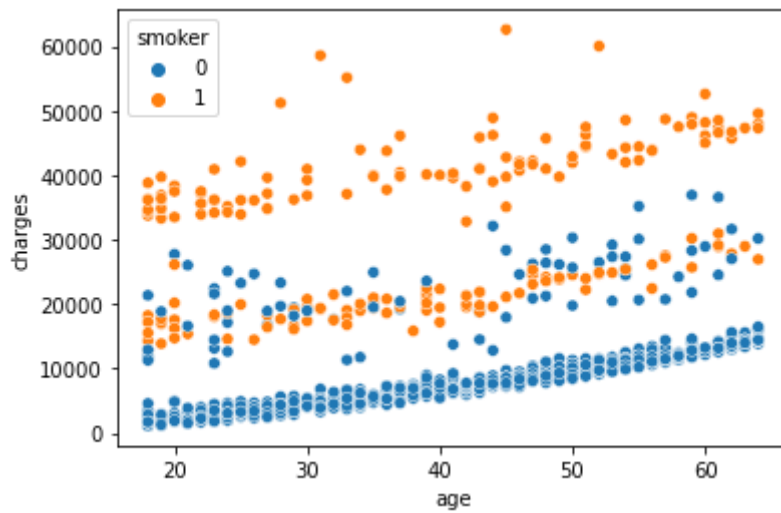


Biểu đồ trên cho ta thấy người hút thuốc thì có chi phí y tế cao hơn, cụ thể :

- hơn 75% người hút thuốc trả chi phí cao hơn hầu hết tất cả người không hút thuốc
- chi phí thấp nhất của người hút thuốc chỉ nhỉnh hơn một chút so với chi phí của 75% người không hút thuốc.
- nếu chi phí dưới 10k, xác suất cao là người đó không hút thuốc
- nếu chi phí trên 20k, xác suất cao là người đó hút thuốc

In [23]: Phân bố của chi phí y tế theo độ tuổi

Out[23]: <AxesSubplot:xlabel='age', ylabel='charges'>

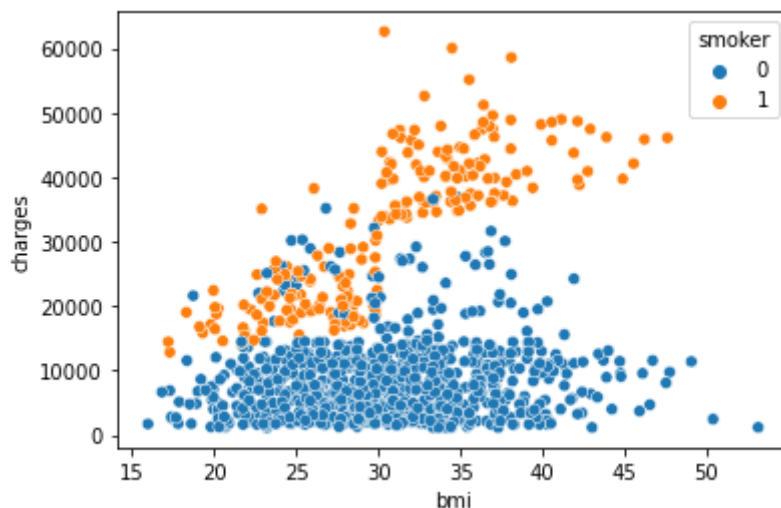


Nhìn vào biểu đồ trên, ta thấy

- người càng cao tuổi thì số tiền chi cho y tế càng nhiều
- Nếu dưới 35 tuổi và không hút thuốc thì khả năng cao chi phí dưới 6k

In [25]: Phân bố của chi phí y tế theo bmi

Out[25]: <AxesSubplot:xlabel='bmi', ylabel='charges'>



- Người hút thuốc và có chỉ số BMI lớn hơn 30 thì chi phí tối thiểu là khoảng 30k

3. VIF

In [26]: Bảng VIF

	feature	VIF
0	sex	1.966855
1	smoker	1.254563
2	age	7.658193
3	bmi	8.638958
4	children	1.816248

1 = Không tương quan [1]

Giữa 1 và 5 = Tương quan vừa [1]

Lớn hơn 5 = Tương quan mạnh [1]

Ta thấy các biến `sex`, `smoker`, `children` tương quan vừa với các biến còn lại.

`age` và `bmi` có sự tương quan mạnh với các biến còn lại

Nên thu thập thêm data để giảm sự phụ thuộc giữa các biến

4. Insight: Sex có ảnh hưởng đến Smoker?

H_0 : sex và smoker độc lập nhau

H_A : sex và smoker phụ thuộc nhau

Đặt:

$A = \text{sex}, A_1 = \text{male}, A_2 = \text{female}$

$B = \text{smoker}, B_1 = \text{yes}, B_2 = \text{no}$

Ta có:

$H_0: P(A_i \cap B_j) = P(A_i)P(B_j)$

$H_A: P(A_i \cap B_j) \neq P(A_i)P(B_j)$

Phần dưới sẽ trình bày về mặt toán học lần sử dụng thư viện `scipy.stats` để tính toán

```
In [32]: contingency
```

```
Out[32]:
```

	no	yes	Pr(Ai)
sex			
female	406	91	0.495513
male	391	115	0.504487

Ta đã tính được $Pr(A_i)$ như bảng trên và

$Pr(B_1) = 0.2053838484546361$

$Pr(B_2) = 0.7946161515453639$

Đến đây ta có thể tính:

Giá trị mong đợi E :

Do kỳ vọng A và B độc lập: (1)

$$E_{ij} = Pr(A_i) \times Pr(B_j) \times N[2]$$

hay

$$E_{ij} = \frac{(\text{Tổng dòng} \times \text{Tổng cột})}{\text{Tổng bảng}} [3]$$

với bảng là bảng contingency

Giá trị χ^2 :

$$\chi^2 = \sum \frac{(O - E)^2}{E} [2][3] \quad (2)$$

với O là giá trị thực sự và E là giá trị mong đợi

Giá trị dof: Degree of freedom

dof cho χ^2 độc lập:

$$dof = v = rc - 1 - (r - 1) - (c - 1) = (r - 1)(c - 1) [2] \\ = 1 \quad (3)$$

Chọn mức ý nghĩa:

$$\alpha = 0.05 \quad (4)$$

Tra bảng Chi Squared với $\alpha = 0.05$, $dof = 1$ ta được critical value = 3.841459

Chấp nhận H_0 nếu

$$\chi_v^2 \leq 3.841459 \quad (5)$$

Ta có thể sử dụng `chi2_contingency` của thư viện `spicy` để tính toán, các giá trị tính được từ thư viện và kết luận là:

p-value là: 0.09827321674727184

chi = 2.733346, critical value = 3.841459

Với mức ý nghĩa 0.05, ta bác bỏ H_A và chấp nhận H_0 .

Kết luận: sex và smoker độc lập.

Ta kiểm tra, không dùng thư viện, được kết quả như sau:

In [36]:

```
chi_square = 2.997908815661011
```

Out[36]:

	smoker	sex	count	Expected value	(O_ij - E_ij)^2/E_ij
0	no	male	391	402.075773	0.305099

	smoker	sex	count	Expected value	(O_ij - E_ij)^2/E_ij
1	yes	male	115	103.924227	1.180406
2	no	female	406	394.924227	0.310623
3	yes	female	91	102.075773	1.201781

Ta thấy:

$$\chi_v^2 = 2.997908815661011 < 3.841459 \quad (6)$$

Vậy bác bỏ H_A với mức ý nghĩa 0.05, chấp nhận H_0

Kết luận: sex và smoker độc lập

5. Insight: Trung bình của 'age', 'bmi', 'children' có bằng nhau đôi một

Sử dụng z-test ta tính được như sau:

Ho: trung bình age = trung bình bmi

Ha: trung bình age khác trung bình bmi

stat=18.131, p=0.000

KẾT LUẬN: Với mức ý nghĩa 0.05, ta chấp nhận Ha, bác bỏ Ho.

Trung bình age khác trung bình bmi

Ho: trung bình bmi = trung bình children

Ha: trung bình bmi khác trung bình children

stat=151.866, p=0.000

KẾT LUẬN: Với mức ý nghĩa 0.05, ta chấp nhận Ha, bác bỏ Ho.

Trung bình bmi khác trung bình children

Ho: trung bình children = trung bình age

Ha: trung bình children khác trung bình age

stat=-85.747, p=0.000

KẾT LUẬN: Với mức ý nghĩa 0.05, ta chấp nhận Ha, bác bỏ Ho.

6. Insight: Sự phụ thuộc của charges vào sex, smoker, age, bmi, children

Ta huấn luyện bằng mô hình OLS Regression:

In [38]:

```
=====
                        OLS Regression Results
=====
Dep. Variable:          charges    R-squared:                0.744
Model:                  OLS        Adj. R-squared:           0.743
Method:                 Least Squares    F-statistic:             580.0
Date:                  Fri, 14 May 2021    Prob (F-statistic):      3.98e-292
Time:                  20:59:36          Log-Likelihood:          -10164.
No. Observations:      1003           AIC:                    2.034e+04
Df Residuals:          997            BIC:                    2.037e+04
Df Model:               5
Covariance Type:       nonrobust
=====
                        coef      std err          t      P>|t|      [0.025      0.975]
-----
const          -1.23e+04    1121.474    -10.968    0.000    -1.45e+04    -1.01e+04
sex              66.0697     386.570      0.171    0.864    -692.515     824.655
smoker          2.363e+04    478.849     49.344    0.000     2.27e+04     2.46e+04
age             260.0358      13.870     18.748    0.000      232.818     287.254
bmi             327.5600      32.307     10.139    0.000      264.162     390.958
children        434.8043     160.590      2.708    0.007     119.671     749.938
=====
Omnibus:            241.068    Durbin-Watson:           2.068
Prob(Omnibus):      0.000    Jarque-Bera (JB):        592.918
Skew:               1.268    Prob(JB):                1.78e-129
Kurtosis:           5.785    Cond. No.                 299.
=====
```

Notes:

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

Kết luận:

- Biến sex không có ý nghĩa (có thể loại bỏ)
- Biến smoker có ý nghĩa đối với mô hình về mặt thống kê (với mức ý nghĩa (***) hay p-value = 0.000)
- Biến age có ý nghĩa đối với mô hình về mặt thống kê (với mức ý nghĩa (***) hay p-value = 0.000)
- Biến bmi có ý nghĩa đối với mô hình về mặt thống kê (với mức ý nghĩa (***) hay p-value = 0.000)
- Biến children không có ý nghĩa (có thể loại bỏ)
- Mô hình có thể giải thích được 74.3% sự thay đổi của biến charges
- Mô hình tương đối tốt (p-value = 1.78e-129)

Ta huấn luyện lại mô hình dựa theo kết luận trên

In [39]:

```
=====
                        OLS Regression Results
=====
```

```

Dep. Variable:          charges    R-squared:                0.742
Model:                  OLS        Adj. R-squared:           0.741
Method:                 Least Squares    F-statistic:             959.1
Date:                   Fri, 14 May 2021    Prob (F-statistic):      1.63e-293
Time:                   21:00:36    Log-Likelihood:          -10168.
No. Observations:       1003    AIC:                    2.034e+04
Df Residuals:           999    BIC:                    2.036e+04
Df Model:                3
Covariance Type:        nonrobust

```

```

=====
              coef      std err          t      P>|t|      [0.025      0.975]
-----
const      -1.185e+04    1097.537     -10.799     0.000     -1.4e+04     -9698.979
smoker       2.367e+04     479.257      49.386     0.000     2.27e+04     2.46e+04
age         262.1450      13.884      18.881     0.000      234.900      289.390
bmi         326.7252       32.392      10.086     0.000      263.160      390.290
=====
Omnibus:                238.957    Durbin-Watson:           2.060
Prob(Omnibus):           0.000    Jarque-Bera (JB):        580.364
Skew:                    1.263    Prob(JB):                9.45e-127
Kurtosis:                 5.741    Cond. No.                 291.
=====

```

Notes:

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

In [40]:

```

Parameters:  const      -11852.720452
smoker       23668.497446
age          262.144961
bmi          326.725200
dtype: float64

```

Ta thấy:

- Cứ tăng 1 tuổi thì chi phí y tế cá nhân tăng 262.144961, tăng 1 chỉ số bmi thì tăng 326.725200 chi phí y tế cá nhân
- Riêng với smoker, người có hút thuốc thì có chi phí y tế cá nhân cao hơn người không hút thuốc đến 23668.497446

II. Thuật toán sử dụng

Trong lab này, nhóm em sử dụng 2 thuật toán chính là Simple linear regression và Support Vector Regression (SVR) (là một dạng mở rộng của Support Vector MachineMachine).

1. Cách thức đánh giá mô hình

- Nhóm em sử dụng độ đo R-squared để đánh giá mô hình
- Phương pháp lấy mẫu để đánh giá mô hình là K-Fold Cross-Validation với k = 10

2. Thuật toán SVR

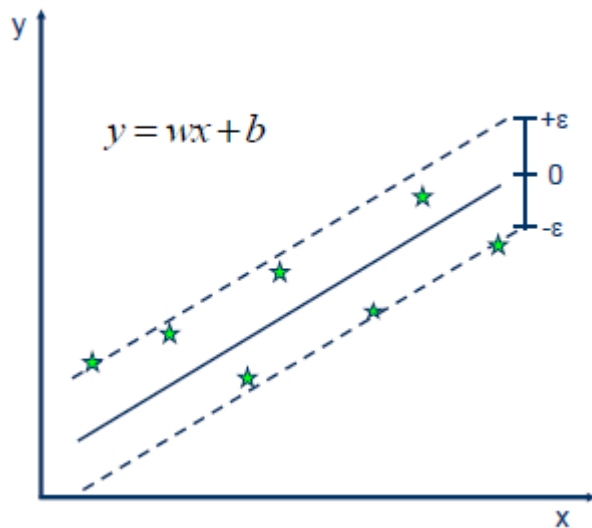
a. Giới thiệu SVR

Support Vector Machine (SVM)

SVM là bài toán phân lớp, và đi tìm mặt phân cách sao cho margin tìm được là lớn nhất, đồng nghĩa với việc các điểm dữ liệu an toàn nhất so với mặt phân cách.

Support Vector Regression (SVR)

SVR là một biến thể của SVM để dùng cho bài toán hồi quy. SVR tìm cách cực tiểu margin sao cho có thể chứa nhiều điểm dữ liệu nhất có thể.



• Solution:

$$\min \frac{1}{2} \|w\|^2$$

• Constraints:

$$y_i - wx_i - b \leq \epsilon$$

$$wx_i + b - y_i \leq \epsilon$$

b. Sử dụng SVR từ thư viện Scikit-learn

Nhóm chọn ra mô hình tốt nhất...

III. Tham khảo

[1]. [Stephanie - Variance Inflation Factor - Statisticshowto.com](#)

[2]. <https://www3.nd.edu/~rwilliam/stats1/x51.pdf>

[3]. <https://towardsdatascience.com/gentle-introduction-to-chi-square-test-for-independence-7182a7414a95>