

VIF: trùng câu (Cột)

Biến 1: tuổi

Biến 2: năm sinh ← gần như giống

ISLP = viết công thức, xem thống kê, học bản chất

1. ISLP dùng khi học & phân tích thống kê
2. Sklearn dùng khi train & dự đoán

LINEAR - CHẠY RA SỐ

OLS - Tại sao ra được ?

GIẢI THÍCH MỐI QUAN HỆ TUYẾN TÍNH

OLS (statsmodels) : Dùng khi muốn hiểu rõ bài toán (VÌ SAO?)

- Cho biết:
 - số a, số b
 - có đáng tin không
 - đúng hay sai nhiều

Boston → y → MS → fit → transform → X → OLS(y, X) → fit → summary

1. nhập (thư viện) [ISLP]
2. chọn y => y = Boston['medv']
3. khai báo MS
4. Fit transform
5. transform (X)
6. OLS (.fit)
7. results.summary

1. nhập (thư viện) [ISLP]

bảng **insurance** có các cột:

- **age** (tuổi)
- **bmi** (chỉ số cơ thể)
- **charges** (tiền bảo hiểm) $\leftarrow y$

2. chọn y

3. khai báo MS (CHỌN CỘT X) - Dùng tuổi và bmi để giải thích tiền bảo hiểm

```
from ISLP.models import ModelSpec as MS
```

```
design = MS(['age', 'bmi'])
```

4. Fit transform - Máy học cách tạo bảng X

```
design = design.fit(insurance)
```

5. transform (X) - Có bảng X để đưa vào mô hình

```
X = design.transform(insurance)
```

6. OLS (.fit)

7. results.summary

=====

- **ModelSpec** = chọn nguyên liệu
- **fit()** = học cách nấu
- **transform()** = nấu ra món ăn (X)

=====



VÍ DỤ CỤ THỂ: TIỀN BẢO HIỂM

Giả sử dữ liệu có 3 cột:

age	bmi	charges
25	22	3000
40	30	12000
60	35	20000

👉 **charges** = tiền bảo hiểm (y)

🧱 BƯỚC 1 — ModelSpec = CHỌN NGUYÊN LIỆU

```
design = MS(['age', 'bmi'])
```

👉 Nghĩa là:

“Tôi chỉ dùng **age** và **bmi** để giải thích **charges**”

📌 Chưa có toán học gì cả, chỉ là chọn cột.

🧠 BƯỚC 2 — fit() = HỌC CÁCH NẤU

```
design = design.fit(data)
```

👉 Lúc này máy làm gì?

- Nhìn bảng dữ liệu gốc
- Ghi nhớ:
 - có bao nhiêu dòng
 - thứ tự cột
 - kiểu dữ liệu
- Nếu có:
 - `poly(age, 2)` → nhớ phải tạo age^2
 - `interaction` → nhớ phải nhân cột

📌 Chưa tạo **X**, chỉ nhớ cách tạo **X**

🍲 BƯỚC 3 — transform() = NẤU RA MÓN ĂN (X)

```
X = design.transform(data)
```

👉 KẾT QUẢ THỰC SỰ

X trông như thế này:

	intercept	age	bmi
1		25	22
1		40	30
1		60	35

🔴 Đây là **bảng số** mà OLS hiểu được.

BƯỚC 4 — OLS HỌC CÔNG THỨC

```
model = sm.OLS(y, X)
results = model.fit()
```

Máy tìm ra:

$\text{charges} = 1000 + 200 \times \text{age} + 300 \times \text{bmi}$

VẬY ModelSpec GIÚP GIẢI THÍCH MODEL NHƯ THẾ NÀO?

❓ “Tiền bảo hiểm tăng vì cái gì?”

👉 Nhìn hệ số:

- age: +200 → lớn tuổi hơn → tiền tăng
- bmi: +300 → bmi cao → tiền tăng mạnh hơn

👉 Bạn phải:

- tự tạo intercept
 - tự tạo age^2
 - tự nhân cột
 - rất dễ sai
-

VÍ DỤ NÂNG CAO: THÊM POLY

```
design = MS(['age', poly('age', 2), 'bmi'])
```

Sau `transform()`:

intercept age age² bmi

👉 Giải thích:

- tuổi càng lớn → tiền tăng **nhANH dãn**

CÂU CHỐT ĐỂ NHỚ (QUAN TRỌNG)

**ModelSpec quyết định bạn đang giải thích điều gì
fit & transform biến ý tưởng đó thành toán học**