# Yêu cầu:

1. Viết mã giả của thuật toán tìm phương trình hồi quy phi tuyến.

2. Đánh giá độ phức theo thời gian của thuật toán tìm phương trình hồi quy phi tuyến ở trên.

3. Với tập dữ liệu về Việt Nam tùy chọn, dùng MS Excel, ngôn ngữ R và Python, hãy tìm phương trình hồi quy phi tuyến.

# Thực hiện:

## Bảng phân công công việc:

|  |  |  |
| --- | --- | --- |
| **Thành viên** | | **Công việc phân công** |
| **Mã học viên** | **Họ và Tên** |
| 240104026 | Huỳnh Quốc Bữu | * Viết mã giả của thuật toán tìm phương trình hồi quy phi tuyến * Tìm phương trình hồi quy phi tuyến bằng R * Tìm phương trình hồi quy phi tuyến bằng Python |
| 240104039 | Nguyễn Đức Linh | * Đánh giá độ phức theo thời gian của thuật toán tìm phương trình hồi quy phi tuyến * Chọn tập dữ liệu & xử lý dữ liệu * Tìm phương trình hồi quy phi tuyến bằng Excel |

## 1. Viết mã giả của thuật toán tìm phương trình hồi quy phi tuyến:

### 1.1 Mô tả bài toán:

Cho một tập dữ liệu gồm n quan sát:

Với:

* : là vector biến độc lập đầu vào có k phần tử
* : là biến phụ thuộc đầu ra

Bảng dữ liệu:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Mục tiêu: tìm hàm hồi qui phi tuyến sao cho:

Có nhiều mô hình hồi qui phi tuyến, trong bài tập nhóm chọn **mô hình hồi quy phi tuyến đa thức bậc d** (Polynomial d degree) [1]

Chọn là một hàm đa thức bậc d loại **polynomial d degree**  có dạng:

Với:

* : chỉ số mũ từng biến
* : tổng các bậc
* : là dạng của một đơn thức (monomial) tương tương ứng
* : hệ số tương ứng
* là tổng các monomial trong đa thức

Bảng mô tả các phần tử đa thức (nomial) ví dụ với k =3 và d = 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Tổng các nomial trong đa thức

Khi đó hàm hôì quy phi tuyến là:

Bảng mô tả các phần tử đa thức (nomial) tổng quát với k biến độc lập và bậc d

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Tổng các nomial trong đa thức:

Hàm hôì quy phi tuyến là:

Gọi ánh xạ

Phương trình hồi qui phi tuyến trên có thể viết lại như sau:

Do đó, bài toán:

với và tập dữ liệu

có thể biến đổi thành bài toán:

với và tập dữ liệu

Z được biến biến đổi qua ánh xạ

Dễ dàng nhận ra là hàm tuyến tính

Ta có thể giải bằng phương pháp bình phương tối thiểu (Ordinary Least Squares - OLS) [2]

Ta tính được:

### 1.2 Mã giả của thuật toán tìm phương trình hồi quy phi tuyến ở trên:

**Bước 1:** Nhập tập dữ liệu

với : là vector biến độc lập đầu vào có k phần tử

**Bước 2:** Xác định ma trận , vector

**Bước 3:** Chọn mô hình hồi quy phi tuyến là mô hình hồi qui đa thức bậc d

**Bước 4:** Tìm ma trận ma trận từ ma trận , trong đó

**Bước 5:** Tính ma trân chuyển vị

**Bước 6:** Tính ma trận

**Bước 7:** Tính ma nghịch đảo

**Bước 8:** Tính

**Bước 9:** Xuất kết quả tính được

## 2. Đánh giá độ phức theo thời gian của thuật toán tìm phương trình hồi quy phi tuyến ở trên

**Bước 1:** Nhập tập dữ liệu

với : là vector biến độc lập đầu vào có k phần tử

* **BigO1 = O(1)** vì đây chỉ là thao tác nhập liệu nên coi độ phức tạp là hằng số

**Bước 2:** Xác định ma trận , vector

**For i=1 to n**:

**For j=1 to k:**

X[i][j] = dữ liệu input

y[i] = dữ liệu đầu ra

* **BigO2 = O()**

**Bước 3:** Chọn mô hình hồi quy phi tuyến là mô hình hồi qui đa thức bậc d

* **BigO3 = O(1)**

**Bước 4:** Tìm ma trận ma trận từ ma trận , trong đó

**For i = 1 to n:**

m = 0

**For from 0 to d:**

**For from 0 to d:**

**...**

**For from 0 to d:**

**if ( ) <= d:**

don\_thuc = 1

**For t = 1 to k:**

don\_thuc = don\_thuc \*

Z[i][c] = don\_thuc

m = m + 1

* **BigO4 = O() = O()**

**Bước 5:** Tính ma trân chuyển vị

**For i = 1 to n:**

**For j = 1 to m:**

= Z[i][j]

* **BigO5 = O(n.m)**

**Bước 6:** Tính ma trận

A <- ma trận m x m

**For p = 1 to m:**

**For q = 1 to m:**

A[p][q] = 0

**For r = 1 to n:**

A[p][q] += ] \* Z[r][q]

* **BigO6 = O()**

**Bước 7:** Tính ma nghịch đảo

* BigO7 = O()

Đây là 1 phép toán phức tạp, thường dùng thư viện. Một số phương pháp như phương pháp Gauss-Jordan – phổ biến nhất trong tính toán thực tế – độ phức tạp trung bình là O(), ngoài ra có phương pháp của VICTOR PAN đạt BigO ~[3].

**Bước 8:** Tính

**For p = 1 to m:**

w[p] = 0

**For i = 1 to n:**

w[p] += [p][i] \* y[i]

**For p = 1 to m:**

β[p] = 0

**For q = 1 to m:**

β[p] += [p][q] \* w[q]

* **BigO8 = O( )**

**Bước 9:** Xuất kết quả tính được

* **BigO9 = O(m)**

**Độ phức tạp của tổng hợp của thuật toán tìm phương trình hồi qui phi tuyến là:**

## 3. Với tập dữ liệu về Việt Nam tùy chọn, dùng MS Excel, ngôn ngữ R và Python, hãy tìm phương trình hồi quy phi tuyến

### 3.1 Chọn tập dữ liệu về Việt Nam

1. Chọn tập dữ liệu về giá nhà tại Việt Nam để phân tích
2. Nguồn dữ liệu: <https://www.kaggle.com/datasets/nguyentiennhan/vietnam-housing-dataset-2024/data> [4]
3. File dữ liệu: Github-Code-Team11/datasets/vietnam\_housing\_dataset.csv
4. Giới thiệu tập dữ liệu:

Bộ dữ liệu này chứa thông tin về các bất động sản nhà ở tại Việt Nam. Nó bao gồm các đặc điểm chi tiết của từng bất động sản như vị trí, đặc điểm vật lý, tình trạng pháp lý và nội thất, cùng với giá bán. Bộ dữ liệu này rất hữu ích cho việc phân tích bất động sản, xây dựng mô hình dự đoán giá và phân tích xu hướng thị trường.

Nguồn thu thập dữ liệu: Dữ liệu được thu thập (crawled) từ trang web batdongsan.vn.

1. Mô tả các cột dữ liệu:

* Address (Địa chỉ): Địa chỉ đầy đủ của bất động sản, bao gồm tên dự án (nếu có), đường, phường, quận và thành phố.
* Area (Diện tích): Tổng diện tích của bất động sản, đơn vị tính là mét vuông (m²).
* Frontage (Mặt tiền): Chiều rộng mặt trước của bất động sản, đơn vị tính là mét (m).
* Access Road (Đường vào): Chiều rộng của con đường dẫn vào bất động sản, đơn vị tính là mét (m).
* House Direction (Hướng nhà): Hướng chính mà mặt trước của căn nhà quay về (ví dụ: Đông, Tây, Nam, Bắc).
* Balcony Direction (Hướng ban công): Hướng mà ban công quay về.
* Floors (Số tầng): Tổng số tầng của căn nhà.
* Bedrooms (Số phòng ngủ): Số lượng phòng ngủ trong bất động sản.
* Bathrooms (Số phòng tắm): Số lượng phòng tắm trong bất động sản.
* Legal Status (Tình trạng pháp lý): Trạng thái pháp lý của bất động sản, ví dụ như có sổ hồng, đang trong hợp đồng mua bán, v.v.
* Furniture State (Tình trạng nội thất): Tình trạng trang bị nội thất, ví dụ như đầy đủ nội thất, nội thất cơ bản hoặc chưa có nội thất.
* Price (Giá bán): Giá của bất động sản, đơn vị tính là **tỷ đồng Việt Nam (VND)**.

### 3.2 Tiền xử lý dữ liệu:

* Lọc bỏ dữ liệu rỗng hoặc NA
* Trích xuất thông tin Provine từ trường Address
* Chuẩn hóa định dạng dữ liệu
* Encode các cột text (chưa làm)

**Code python tiền xử lý dữ liệu** [5]

A screen shot of a computer program

AI-generated content may be incorrect. A screen shot of a computer code

AI-generated content may be incorrect.

Figure 1. Hình ảnh code python tiền xử lý dữ liệu

* input file: vietnam\_housing\_dataset.csv [4]
* output file: hw2-nonlinear-regression-excel.xlsx [5]

**file hw2-nonlinear-regression-excel.xlsx sẽ dùng làm data đầu vào cho các phân tích bên dưới**

### 3.3 Tìm phương trình hồi quy phi tuyến bằng Excel: [5]

**Bước 1:** Chọn X, y và bậc của mô hình Polynomial

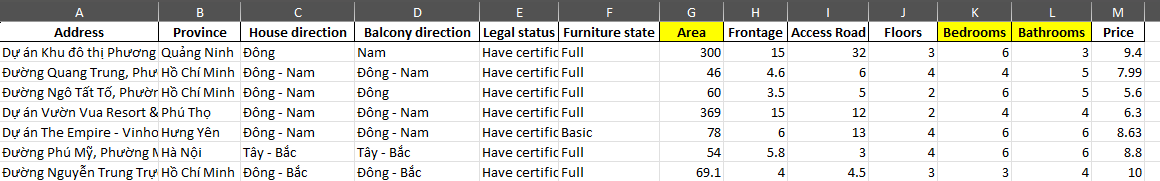
Với dữ liệu sau khi được tiền xử lý, ta chọn:

Biến độc lập đầu vào là: **X = [“Area”, “Bedrooms”, “Bathrooms”]**

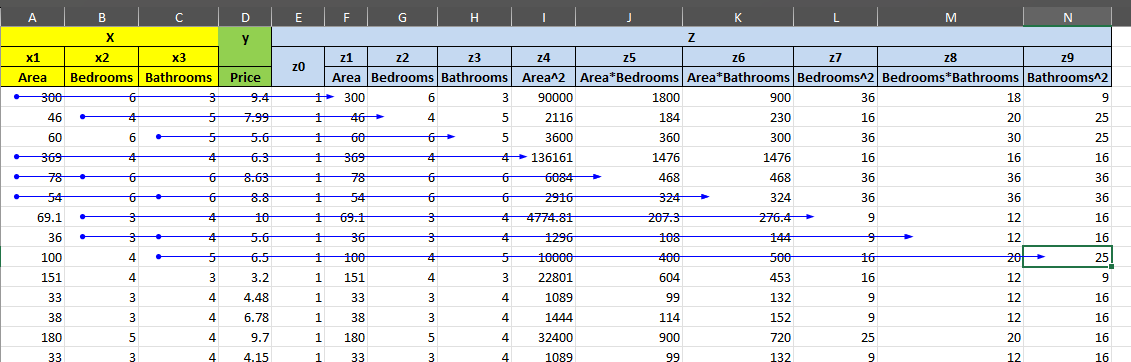
Biến phụ thuộc đầu ra là **y = [“Price”]**

Chọn bậc của mô hình Polynomial là **d = 2**

Ta tính được tổng các đơn thức (mononomial) là:

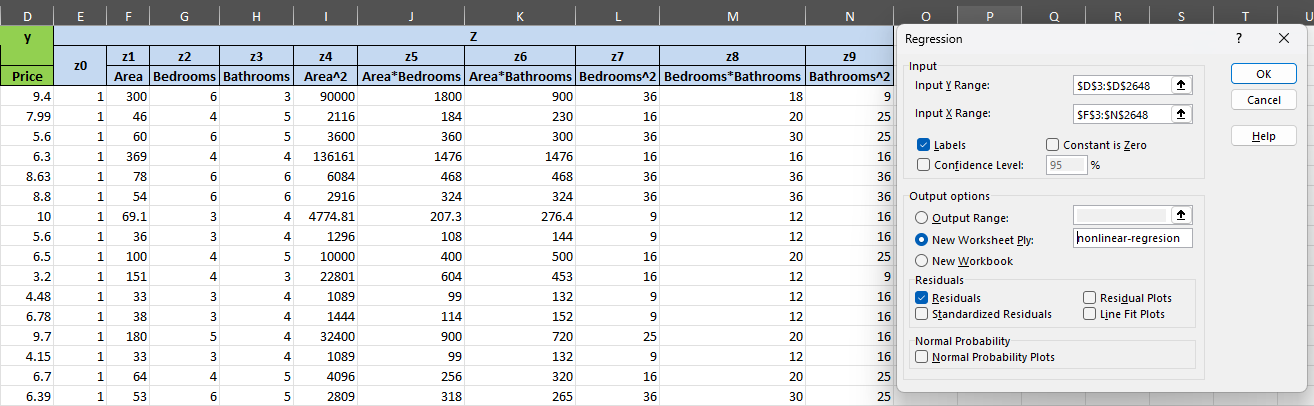


**Bước 2: Ánh xạ**



**Bước 3: chạy mô hình hồi qui tuyến tính với Z và y**

Ghi chú: không chọn z0 trong Input X Range



**Bước 4: Quan sát kết quả sau khi chạy xong** [5]

A screenshot of a spreadsheet

AI-generated content may be incorrect.

Kết quả cần tìm là cột Coefficients (thông tin highlight xanh lá cây)

**Bước 5: Nhận xét kết quả:**

**Summary**

* Cái **R Square ≈** **0.277**, **có nghĩa** là từ **những biến số đầu vào lựa chọn (x: Area, Bedrooms, Bathrooms và các đơn thức bậc 2 của chúng) chỉ giải thích được ≈** **27.7% biến thiên của giá (Price)**. Cái này cho thấy, dù mô hình có ý nghĩa thống kê (xem mục F và Significance F ở dưới), nhưng vẫn còn rất nhiều yếu tố khác ảnh hưởng đến giá mà chưa được đưa vào (Như địa chỉ là 1 biến rất quan trọng)
* Adjusted R-Square (0.275): Chỉ số này điều chỉnh R-Square theo số lượng biến trong mô hình. Sự thay đổi không đáng kể so với R-Square, cho thấy **mô hình vẫn chưa thực sự tốt**.
* Standard Error (1.9): Đây là sai số chuẩn của ước lượng, cho thấy mức độ sai lệch của các dự báo so với giá trị thực tế.

A table with black text and black letters

AI-generated content may be incorrect.

Figure 2. Hình ảnh kết quả tổng quan trên Excel

**ANOVA (Analysis of Variance):**

* Giá trị F khoảng 112.5 với Significance F siêu nhỏ (6.11E-179), cho thấy tổng thể mấy biến giải thích có ý nghĩa thống kê rất mạnh. Nói một cách đơn giản, từ những cái này, nếu gộp lại thì mô hình có ảnh hưởng rõ rệt lên giá, mặc dù từng biến riêng lẻ có thể không hết đều ý nghĩa.

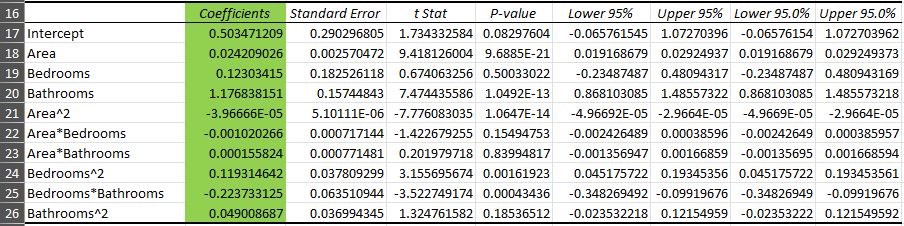
A screenshot of a table

AI-generated content may be incorrect.

Figure 3. Hình kết quả ANOVA trên Excel

**Các hệ số (Coefficients):**

Các hệ số thể hiện tác động của từng biến vào giá trị dự đoán:



* Bathrooms:
  + Hệ số ≈ 1.17
  + t Stat & P-value: t ≈ 7.47, p ≈ 1.049E-13

Đây là một biến có ảnh hưởng tích cực và có ý nghĩa thống kê. Có thể hiểu, khi số phòng tắm tăng lên 1 đơn vị thì giá có xu hướng tăng trung bình khoảng 1.17 đơn vị (các đơn vị tùy thuộc vào thang đo giá trị).→ Cái này cho thấy một mối quan hệ trực quan: nhiều phòng tắm thường đi kèm với giá cao hơn. (Vì thường nhiều phòng tắm sẽ kéo theo những chỉ số khác nữa)

* Bedrooms:
  + Hệ số: 0.123
  + t Stat & P-value: t ≈ 0.67, p ≈ 0.5

Biến này có nghĩa là Bedroom của bất động sản cũng có tác động tích cực nhưng không quá mạnh.

**Hiệu ứng phi tuyến:**

Các biến có dạng bậc hai như với hệ số âm cho thấy có mối quan hệ cong, tức là sau một mức độ nhất định thì hiệu ứng của biến đó lên giá sẽ giảm đi (hiệu ứng biên giảm dần). Cái này khá hay, vì ở những trường hợp thực tế, tăng diện tích hay cải thiện đường vào ban đầu có ảnh hưởng mạnh, nhưng sau đó tác động này bão hòa lại.

### 3.4 Tìm phương trình hồi quy phi tuyến bằng Python: [5]

**Bước 1: Chuẩn bị data file**

Là file excel của bước tiền xử lý dữ liệu

**Bước 2: Chuẩn bị code python** [5]

A screen shot of a computer

AI-generated content may be incorrect.

A black screen with orange and purple text

AI-generated content may be incorrect.

A computer screen with white text

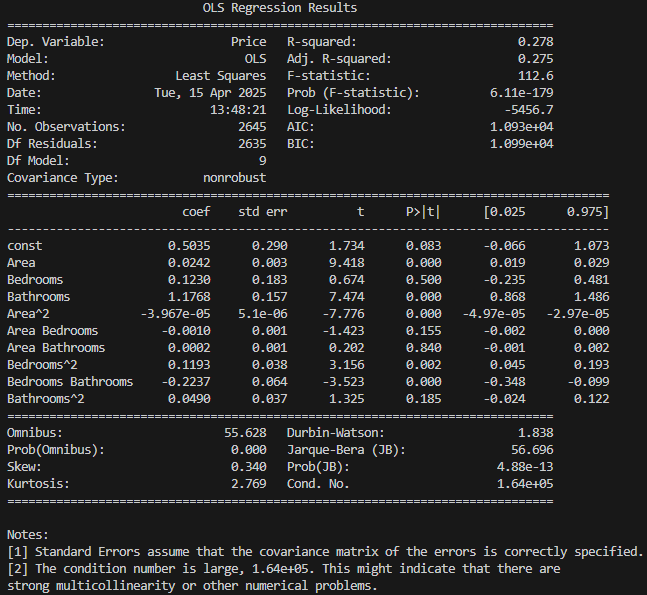
AI-generated content may be incorrect.

**Bước 3: chạy code python**

A white text on a black background

AI-generated content may be incorrect.

**Bước 4: Quan sát kết quả:**

****

Kết quả cần tìm là cột coef

**Bước 5: So sánh kết quả chạy bằng Excel**

Kết quả chạy mô hình với python giống với chạy bằng Excel

### 3.5 Tìm phương trình hồi quy phi tuyến bằng R: [5]

**Bước 1: Chuẩn bị data file**

Là file excel của bước tiền xử lý dữ liệu

**Bước 2: Chuẩn bị code R** [5]

A screen shot of a computer program

AI-generated content may be incorrect.

A computer screen shot of a program code

AI-generated content may be incorrect.

A black screen with white text

AI-generated content may be incorrect.

**Bước 3: chạy code R**

A black and orange text

AI-generated content may be incorrect.

**Bước 4: Quan sát kết quả:**

A screenshot of a computer

AI-generated content may be incorrect.

Kết quả cần tìm là cột Estimate

**Bước 5: So sánh kết quả chạy bằng Excel và Python**

Kết quả chạy mô hình với R giống với chạy bằng Excel và Python

## Kiểm tra kết quả với ChatGPT

**Nhận xét:** Câu trả lời của chatGPT khá là chung chung, mã giả chưa đúng, cách tính độ phức tạp cũng không hoàn thiện như bài làm của nhóm.

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.A black screen with white text

AI-generated content may be incorrect.

Figure 5. Hình ảnh so sánh kết quả với chatGPT

# References

|  |  |
| --- | --- |
| [1] | G. James, D. Witten, T. Hastie và R. Tibshirani, An Introduction to Statistical Learning with Applications in R, New York: Springer, 2021. |
| [2] | D. C. P. E. A. &. V. G. G. Montgomery, Introduction to Linear Regression Analysis, 5th Edition, Wiley, 2012. |
| [3] | V. PANe, “HOW CAN WE SPEED UP MATRIX MULTIPLICATION?,” *Society for Industrial and Applied Mathematics,* tập 26, p. 393, 1984. |
| [4] | N. T. Nhan, “Kaggle,” [Trực tuyến]. Available: https://www.kaggle.com/datasets/nguyentiennhan/vietnam-housing-dataset-2024/data. |
| [5] | T.-L. P. t. d. l. k. d. n. cao, “sourcode-hw2-team11,” [Trực tuyến]. Available: https://github.com/buuhq-it/uit-msc-data-analyst-assignments/tree/main/hw2. |