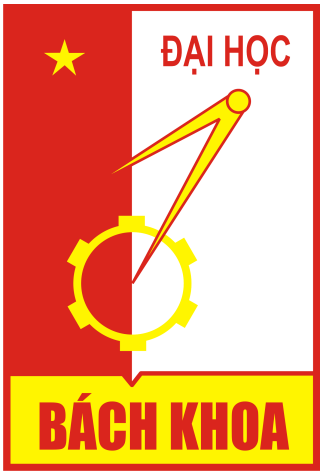
ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC



ĐỒ ÁN I

MÔ HÌNH HỒI QUY TUYẾN TÍNH VÀ ỨNG DỤNG TRONG DỰ BÁO GIÁ XE CŨ

Chuyên ngành: Toán ứng dụng

Giảng viên hướng dẫn: TS. Nguyễn Cảnh Nam
Sinh viên thực hiện: Vũ Thành Đạt
MSSV: 20206275
Lớp: Hệ thống thông tin 01-K65

HÀ NỘI, 07/2023

NHẬN XÉT CỦA GIẢNG VIÊN

1. Mục tiêu

(a)

(b)

(c)

2. Nội dung

(a)

(b)

(c)

3. Đánh giá kết quả đạt được

(a)

(b)

(c)

Hà Nội, ngày ... tháng ... năm 2023

Giảng viên hướng dẫn

TS. Nguyễn Cảnh Nam

Lời cảm ơn

Trong kỳ học vừa qua, tuy được học tập và làm việc với giảng viên hướng dẫn trong một thời gian ngắn nhưng bản thân em tự nhận thấy đã học hỏi được rất nhiều kiến thức và kỹ năng cần thiết và bổ ích. Báo cáo là tổng hợp những nội dung cơ bản nhất của chủ đề nghiên cứu dưới sự hướng dẫn tận tình của TS. Nguyễn Cảnh Nam cùng những kết quả có được từ quá trình làm việc nghiêm túc. Tuy nhiên, do hạn chế về thời gian và kiến thức bản thân nên không thể tránh khỏi những thiếu sót. Em rất mong nhận được những đánh giá và góp ý từ thầy cô. Em xin chân thành cảm ơn.

Hà Nội, 25 tháng 07 năm 2023

Tác giả đồ án

Vũ Thành Đạt

Lời mở đầu

Hồi quy tuyến tính là phương pháp đơn giản nhưng hiệu quả để dự đoán và giải thích sự biến động của biến đầu ra dựa trên các biến đầu vào. Với tính linh hoạt và khả năng mô hình hóa các mối quan hệ tuyến tính giữa các biến, phương pháp này đã được sử dụng rộng rãi trong nhiều lĩnh vực, từ kinh tế, tài chính đến y tế, khoa học xã hội và kỹ thuật. Việc áp dụng hồi quy tuyến tính cần phải được thực hiện một cách cẩn thận và chính xác để tránh các sai sót và kết quả không chính xác. Điều này đặc biệt quan trọng trong các lĩnh vực như y tế và kỹ thuật, khi sự chính xác và độ tin cậy của mô hình có thể ảnh hưởng đến việc ra quyết định và kết quả của nghiên cứu.

Ở trong nội dung bài báo cáo này chúng ta sẽ đi tìm hiểu các nội dung như tổng quan về mô hình hồi quy tuyến tính, mô hình hồi quy tuyến tính đơn, đa biến, kiểm tra sự phù hợp của mô hình hồi quy tuyến tính đa biến và áp dụng mô hình vào việc ứng dụng, giải quyết bài toán dự báo thực tế.

Tóm tắt nội dung báo cáo

Báo cáo đồ án I sẽ gồm các nội dung như sau:

- Giới thiệu sơ lược về phân tích hồi quy tuyến tính. Trình bày về mô hình hồi quy tuyến tính và xây dựng mô hình bằng phương pháp bình phương cực tiểu.
- Thực hiện bài toán kiểm định và ước lượng khoảng cho hệ số hồi quy, kiểm định mức ý nghĩa của mô hình hồi quy tuyến tính đa biến.
- Kiểm tra một số vấn đề liên quan đến giả thiết ước lượng của mô hình hồi quy tuyến tính đa biến.
- Thử nghiệm số với bài toán xây dựng mô hình dự báo giá xe ô tô cũ.

Quy tắc viết báo cáo và ký hiệu

Trong báo cáo này có sử dụng một số tên viết tắt cho các thuật ngữ như sau:

OLS	Ordinary Least Square (Phương pháp bình phương cực tiểu)
SRM	Sample regression mode (Mô hình hồi quy mẫu)
SRF	Sample regression function (Hàm hồi quy mẫu)

Đồng thời, một số ký hiệu trong báo cáo như sau:

β	Hệ số hồi quy
$\hat{\beta}$	Giá trị hệ số hồi quy ước lượng.
ε	Sai số
e	Phần dư
$E(X)$	Kỳ vọng
$V(X)$	Phương sai
$cov(X, Y)$	Hiệp phương sai
R^2	Hệ số xác định
r_{ij}	Hệ số tương quan giữa X_i và X_j

Mục lục

Chương 1

Giới thiệu về mô hình hồi quy tuyến tính

1.1 Giới thiệu

Trong nhiều lĩnh vực, việc mô hình hóa mối quan hệ giữa các biến là một nhu cầu thiết yếu để đưa ra những quyết định đúng đắn và hiệu quả. Hồi quy tuyến tính là một trong những phương pháp thống kê phổ biến nhất được sử dụng để mô hình hóa mối quan hệ giữa một biến phụ thuộc và một hay nhiều biến độc lập. Phương pháp này là một công cụ mạnh mẽ để dự đoán giá trị của một biến phụ thuộc dựa trên các biến độc lập, hoặc để tìm ra mối quan hệ giữa các biến trong một tập dữ liệu. Nó cũng cung cấp cho chúng ta khả năng kiểm tra giả thuyết về mối quan hệ giữa các biến và tìm ra biến quan trọng nhất.

Ta xem xét ví dụ nghiên cứu sự phụ thuộc của giá xe cũ vào các yếu tố như: năm sản xuất xe, số Km đã đi, loại động cơ của xe, số chỗ của xe. Trong trường hợp này, giá xe là biến phụ thuộc (cần dự đoán) và năm sản xuất xe, số Km đã đi, loại động cơ của xe, số chỗ của xe là các biến độc lập.

Mục tiêu của hồi quy tuyến tính là sử dụng một hàm số tuyến tính để dự đoán giá trị của biến phụ thuộc. Hàm số này được xây dựng dựa trên một tập hợp các giá trị dữ liệu đã được thu thập và xử lý, dự đoán giá trị của biến phụ thuộc trên tập giá trị của biến độc lập. Chúng ta có thể sử dụng các phương pháp thống kê để xác định mức độ ảnh hưởng của các biến độc lập lên biến phụ thuộc và đánh giá độ chính xác của mô hình.

1.2 Dữ liệu cho phân tích

Trong phân tích hồi quy chúng ta thường dùng 3 loại dữ liệu phổ biến sau đây:

- **Dữ liệu chéo:** là các dữ liệu được thu thập tại cùng một thời điểm(thời kỳ) nhưng ở các không gian (địa phương, đơn vị,...) khác nhau và được lấy một cách ngẫu nhiên từ tổng thể nghiên cứu.
  Ví dụ: Dữ liệu số lượng khách du lịch theo độ tuổi và nơi ở của khách du lịch tại các tỉnh miền Bắc trong đợt nghỉ lễ 30 tháng 4 năm 2023.

- **Dữ liệu chuỗi thời gian:** là loại dữ liệu được thu thập trong quá trình theo dõi các quan sát của một biến trong khoảng thời gian liên tiếp, theo định kỳ hoặc không định kỳ. Dữ liệu chuỗi thời gian thường được sử dụng để mô hình hóa và dự báo các xu hướng, mô hình tương quan giữa các biến và đưa ra các dự đoán về tương lai.
  Ví dụ: Dữ liệu về số lượng khách du lịch đến thăm một điểm du lịch trong suốt một năm.

- **Dữ liệu hỗn hợp:** là loại dữ liệu được thu thập theo không gian và thời gian.
  Ví dụ: Dữ liệu về số lượng khách du lịch đến một điểm du lịch trong suốt một năm (dữ liệu chuỗi thời gian), được phân tích theo độ tuổi và nơi ở của các khách du lịch (dữ liệu chéo).

Chương 2

Mô hình hồi quy tuyến tính

2.1 Mô hình hồi quy tuyến tính

Mô hình hồi quy tuyến tính là một mô hình trong đó giả định mối quan hệ tuyến tính giữa các biến độc lập X và biến phụ thuộc đầu ra duy nhất Y. Nói cách khác, Y có thể được tính toán từ sự kết hợp tuyến tính của các biến đầu vào X. Biểu diễn của mô hình hồi quy tuyến tính là một phương trình tuyến tính kết hợp tập giá trị đầu vào cụ thể X và nghiệm là đầu ra dự đoán cho tập giá trị đầu vào đó Y.

Phương trình đưa ra một hệ số tỷ lệ cho mỗi giá trị biến độc lập gọi là hệ số hồi quy. Ngoài hệ số của biến độc lập trong phương trình còn có thêm một hệ số gọi là hệ số chặn. Tính tuyến tính trong mô hình hồi quy là tuyến tính ở các hệ số hồi quy và có thể tuyến tính hoặc phi tuyến ở các biến X và Y.
Ví dụ:

$$Y = \beta_0 + \beta_1 X^3$$

mô hình trên tuyến tính ở các hệ số hồi quy nên được coi là mô hình hồi quy tuyến tính.

Khi có một biến độc lập X phương pháp này được gọi là hồi quy tuyến tính đơn giản *(Simple Linear Regression)*.

$$Y = \beta_0 + \beta_1 X \tag{2.1}$$

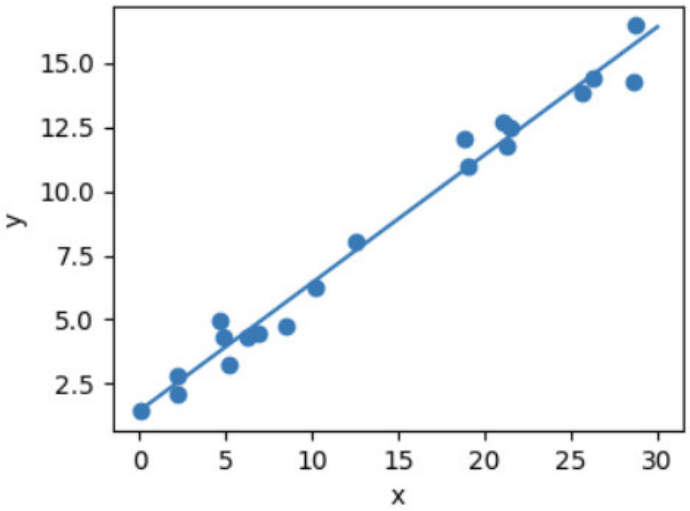
Trong đó:

- Y là biến phụ thuộc cần dự báo.
- β_0 là hệ số chặn, cho biết giá trị trung bình của biến phụ thuộc Y khi biến X = 0.
- β_1 là hệ số góc, cho biết khi X tăng lên 1 đơn vị thì giá trị của Y thay đổi như thế nào.

Tuy nhiên trong thực tế, từ dữ liệu của X ta không thể có dự báo giá trị của Y mà không có sai số. Vì Y còn phụ thuộc các yếu tố khác mà ta không đưa được vào mô hình vì một số lý do như không có sẵn dữ liệu, khó đo lường,... và gọi là nhiễu - sai số(ε) và các biến này ảnh hưởng lên Y không đáng kể. Ta viết lại phương trình (2.1) như sau:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{2.2}$$

Khi có dữ liệu của X và Y ta sẽ biểu diễn được các điểm trên đồ thị thành một đám mây điểm. Nếu các điểm tập trung quanh một đường thẳng thì có thể cho rằng đám mây điểm là phù hợp.



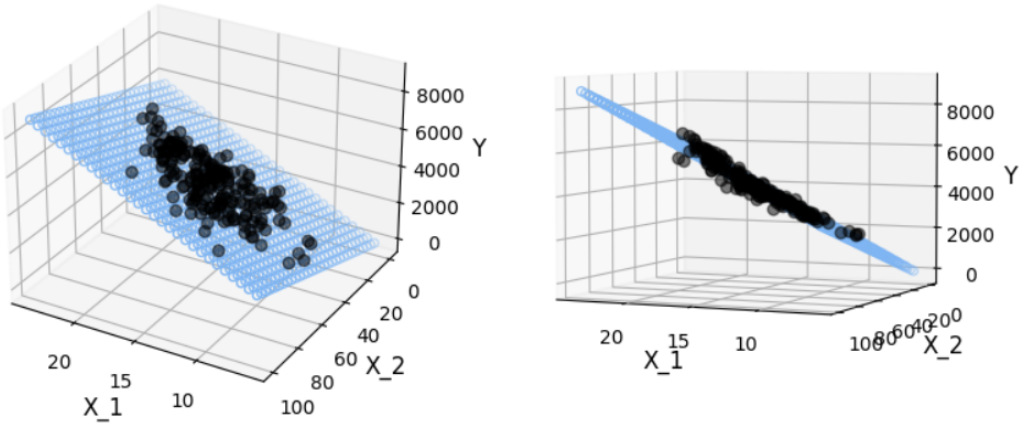
Hình 2.1: Đường hồi quy tuyến tính đơn biến

Tương tự, khi có nhiều biến đầu vào ta có phương pháp là hồi quy tuyến tính đa biến *(Multiple Linear Regression)*.

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon \tag{2.3}$$

Trong trường hợp này, với k = 2 biểu diễn của Y được gọi là mặt phẳng:

Với k > 2, biểu diễn của Y được gọi là siêu phẳng.



Hình 2.2: Mặt phẳng hồi quy tuyến tính 2 biến

2.2 Phương pháp ước lượng bình phương cục tiểu

Phương pháp ước lượng thông dụng nhất trong mô hình hồi quy tuyến tính là phương pháp bình phương cực tiểu *(Ordinary Least Square - OLS)*. Phương pháp này được giới thiệu bởi Gauss vào những năm cuối thế kỷ XVIII. Đây là một phương pháp tối ưu hóa để lựa chọn một đường phù hợp nhất cho một dải dữ liệu ứng với cực trị của tổng các bình phương sai số thống kê giữa đường phù hợp nhất và dữ liệu[1].

Bài toán đặt ra dựa trên mẫu quan sát được, tìm giá trị của các tham số β và σ^2 sao cho tổng bình phương sai số giữa các giá trị quan sát và các giá trị dự đoán của mô hình là nhỏ nhất.

2.2.1 Mô hình hồi quy tuyến tính đơn biến

Xét mô hình hồi quy tuyến tính đơn biến:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

Giả thiết phương pháp bình phương cực tiểu[2]:

i. $\mathrm{E}(\varepsilon) = 0$

ii. $\mathrm{V}(\varepsilon) = \sigma^2$, phương sai sai số không đổi.

iii. $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$ (các sai số ε là không tương quan với nhau)

Ta có $\widehat{\beta_0}$, $\widehat{\beta_1}$ là các ước lượng cần tìm của β_0, β_1 với dữ liệu từ mẫu kích thước n $W = \Big((x_1, y_1), ..., (x_i, y_i)\Big)$, $\widehat{y_i}$ là các giá trị ước lượng được, khi đó ta có hàm hồi quy mẫu

(Sample regression function - SRF):

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i, i = 1, 2, ..., n \tag{2.4}$$

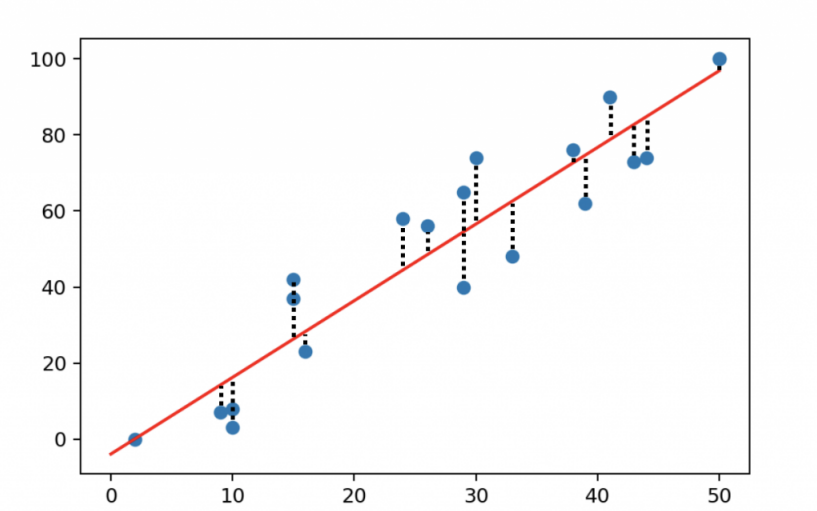
Ta có mô hình hồi quy mẫu *(Sample regression model - SRM)*:

$$y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + e_i, i = 1, 2, ..., n \tag{2.5}$$

Sự sai lệch giữa giá trị thực tế y_i và giá trị ước lượng tương ứng từ hàm hồi quy mẫu $\widehat{y}_i$ được gọi là phần dư *(residuals)*, ký hiệu e_i

$$e_i = y_i - \widehat{y}_i$$

Bản chất của các phần dư e_i giống như các sai số ngẫu nhiên ε_i.



Hình 2.3: Phần dư hàm hồi quy tuyến tính đơn

Ta có tổng bình phương phần dư:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

Phương pháp bình phương cực tiểu đưa về giải bài toán cực trị tìm $\widehat{\beta_0}$, $\widehat{\beta_1}$ sao cho cực tiểu hàm số:

$$S(\widehat{\beta}_0, \widehat{\beta}_1) = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 \to Min. \tag{2.6}$$

Đạo hàm $S(\widehat{\beta}_0, \widehat{\beta}_1)$ theo $\widehat{\beta}_0$ và $\widehat{\beta}_1$:

$$\begin{cases} \dfrac{\partial S}{\partial \widehat{\beta}_0} = 0 \\ \\ \dfrac{\partial S}{\partial \widehat{\beta}_1} = 0 \end{cases} \Leftrightarrow \begin{cases} -2\sum\limits_{i=1}^{n}(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i) = 0 \\ \\ -2\sum\limits_{i=1}^{n}(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i)(x_i) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \sum\limits_{i=1}^{n} y_i = n\widehat{\beta}_0 + \widehat{\beta}_1 \sum\limits_{i=1}^{n} x_i \\ \\ \sum\limits_{i=1}^{n} x_i y_i = \widehat{\beta}_0 \sum\limits_{i=1}^{n} x_i + \widehat{\beta}_1 \sum\limits_{i=1}^{n} x_i^2 \end{cases}$$

Giải hệ này ta nhận được:

$$\begin{cases} \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x} \\ \sum\limits_{i=1}^{n} x_i y_i = (\bar{y} - \hat{\beta}_1\bar{x})n\bar{x} + \hat{\beta}_1 \sum\limits_{i=1}^{n} {x_i}^2 \end{cases} \Leftrightarrow \begin{cases} \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x} \\ \sum\limits_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \widehat{\beta}_1(\sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2) \end{cases}$$

Với $S_{xy} = \sum\limits_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$ và $S_{xx} = \sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2$. Ta có công thức ước lượng hệ số hồi quy:

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x} \quad \text{và} \quad \widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{2.7}$$

Nhận xét:
Các công thức ước lượng (2.7) cho thấy rằng $\widehat{\beta}_0$ và $\widehat{\beta}_1$ sẽ nhận các giá trị khác nhau với các mẫu khác nhau, hay $\widehat{\beta}_0$ và $\widehat{\beta}_1$ là các biến ngẫu nhiên.

2.2.2 Mô hình hồi quy tuyến tính đa biến

Xét mô hình hồi quy tuyến tính đa biến, biến Y phụ thuộc tuyến tính vào các yếu tố chính X_i theo như phương trình:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon$$

Trong đó β_i với $i = 0, 1, 2, ..., k$ là các hệ số hồi quy chưa biết, ε là sai số ngẫu nhiên.

Tiến hành n quan sát độc lập đồng thời về k + 1 biến $X_1, ..., X_k, Y$. Mô hình hoàn chỉnh trở thành:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{k1} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{12} + \cdots + \beta_k x_{k2} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{1n} + \cdots + \beta_k x_{kn} + \varepsilon_n \end{aligned} \tag{2.8}$$

Trong đó:

- β_i là các hệ số hồi quy, Y là biến phụ thuộc, X_i là các biến độc lập (các yếu tố chính).
- ε là phần nhiễu - sai số (đại diện cho tất cả các biến không được đưa vào mô hình do các lý do như không có sẵn dữ liệu, khó đo lường và các biến này ảnh hưởng lên Y là không đáng kể)

Mô hình trên có thể được viết dưới dạng ma trận như sau:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Hoặc viết dưới dạng tổng quát:

$$\underbrace{Y}_{n\text{x}1} = \underbrace{X}_{n\text{x}(k+1)} \cdot \underbrace{\beta}_{(k+1)\text{x}1} + \underbrace{\varepsilon}_{n\text{x}1} \tag{2.9}$$

Ước lượng hệ số hồi quy bằng phương pháp bình phương cực tiểu

Các giả thiết phương pháp bình phương cực tiểu[3]:

Giả thiết 1:

i. Việc ước lượng dựa trên cở sở mẫu ngẫu nhiên(X,Y)
Trong thực tế, không thể điều tra được tổng thể, thay vào đó dựa vào số liệu mẫu ngẫu nhiên để đảm bảo tính đại diện cho quần thể ước lượng mô hình.

ii. $\mathrm{E}(\varepsilon|X) = 0$
Với giá trị X = X_i có nhiều giá trị trong ε nhưng trung bình bằng 0, các yếu tố không ảnh hưởng lên Y

iii. $V(\varepsilon|X) = \sigma^2$

Giả thiết này đảm bảo rằng các sai số không phụ thuộc vào giá trị của các biến độc lập và chúng có phân phối đồng nhất trong toàn bộ quần thể.

iv. Tồn tại ma trận nghịch đảo $(X^TX)^{-1}$.

Giả thiết có nghĩa là không có sự tương quan hoặc phụ thuộc tuyến tính giữa các biến độc lập trong mô hình.

v. Sai số ngẫu nhiên tuân theo phân phối chuẩn:

$$\varepsilon_i \sim N(0, \sigma^2)$$

Làm cơ sở để ước lượng, kiểm định và dự báo về các tham số trong mô hình.

vi. $\text{cov}(\varepsilon_i, \varepsilon_j) \neq 0$, $\forall i \neq j$. Các sai số không có sự tương quan với nhau.

Trong thực tế, thường không thể điều tra toàn bộ tổng thể. Khi đó thay vì điều tra tổng thể, ta chỉ có thể dựa vào mẫu và hàm hồi quy xây dựng trên mẫu được gọi là hàm hồi quy mẫu (*SRF*). Với mẫu kích thước n: $W = (y_i, x_{1i}, \cdots, x_{ki})$, *SRF* được định dạng như sau:

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + ... + \widehat{\beta}_k x_{ki} \tag{2.10}$$

với đại lượng $\hat{\beta}$ là ước lượng cần tìm của β với thông tin từ mẫu trên và mô hình hồi quy mẫu (*SRM*):

$$y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + ... + \widehat{\beta}_k x_{ki} + e_i \tag{2.11}$$

Tương tự như với mô hình hồi quy đơn biến, phương pháp bình phương cực tiểu nhằm xác định các giá trị β_j, j = 0, 1, . . . , k sao cho tổng bình phương các phần dư là nhỏ nhất.

$$\begin{aligned} S(\widehat{\beta}) &= \sum_{i=1}^{n} e^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1i} - \cdots - \widehat{\beta}_k x_{ki})^2 \\ &= \sum_{i=1}^{n} ee^T = (Y - X\widehat{\beta})^T (Y - X\widehat{\beta}) \end{aligned}$$

với đại lượng $\hat{\beta}$ là ước lượng cần tìm của β với thông tin từ mẫu trên.

Phương pháp bình phương cực tiểu đưa về việc giải bài toán cực trị tìm vectơ $\widehat{\beta}$ sao cho cực tiểu hàm số:

$$S(\widehat{\beta}) = (Y - X\widehat{\beta})^T (Y - X\widehat{\beta}) = Y^TY - 2\widehat{\beta}^T X^T Y + \widehat{\beta}^T X^T X\widehat{\beta} \to Min$$

Đạo hàm $S(\widehat{\beta})$ theo $\widehat{\beta}$ ta có:

$$\frac{\partial S}{\partial \widehat{\beta}} = -2X^TY + 2X^TX\widehat{\beta} = 0$$

$$\Leftrightarrow X^TX\widehat{\beta} = X^TY$$

Theo giả thuyết 1(iv), tồn tại ma trận nghịch đảo $(X^TX)^{-1}$ nên ước lượng bình phương cực tiểu có dạng:

$$\widehat{\beta} = (X^TX)^{-1}X^TY \tag{2.12}$$

Vector các giá trị dự báo $\widehat{Y}$ tương ứng với các giá trị quan sát Y là:

$$\widehat{Y} = X\widehat{\beta} = X(X^TX)^{-1}X^TY = HY. \tag{2.13}$$

với $H = X(X^TX)^{-1}X^T$ cấp $(n \times n)$

Các phần dư còn có thể được viết dưới dạng ma trận như sau:

$$e = Y - \widehat{Y} = (I_n - H)Y$$

Tính chất ước lượng bằng phương pháp bình phương cực tiểu

- Ước lượng $\widehat{\beta}$ là ước lượng không chệch của β với:

$$E(\widehat{\beta}) = \beta; \tag{2.14}$$

- Ma trận hiệp phương sai của các hệ số ước lượng:

$$cov(\widehat{\beta}) = \sigma^2(X^TX)^{-1} = \sigma^2C \tag{2.15}$$

  ở đây C := $(X^TX)^{-1}$ là ma trận đối xứng, $C = C_{ij}(i, j = 0, 1...k)$

- Phần dư e có tính chất:

$$E(e) = 0; cov(e) = \sigma^2(I - H) \tag{2.16}$$

- $\widehat{\sigma}^2 = \dfrac{e^Te}{n-k-1} = \sum\limits_{j=1}^{n} \dfrac{e_j^2}{n-k-1}$ là ước lượng không chệch của σ^2.

Chứng minh:

- Ta xét các phép biến đổi sau đây:

$$E(\widehat{\beta}) = E\left((X^TX)\right)^{-1}X^TY = E(\left(X^TX\right)^{-1}X^T(X\beta+\varepsilon))$$
$$= E(\left(X^TX\right)^{-1}X^TX\beta + \left(X^TX\right)^{-1}X^T\varepsilon) = \beta$$

vì E(ε) = 0 và $\left(X^TX\right)^{-1}X^TX\beta$ = I là ma trận đơn vị.

- Ta sử dụng công thức $cov(y,y) = V(y)$ và $V(Ay) = A.V(y).A^T$, suy ra:

$$cov(\widehat{\beta}) = \left(X^TX\right)^{-1}X^Tcov(Y)X\left(X^TX\right)^{-1}$$
$$= \sigma^2\left(X^TX\right)^{-1}X^TIX\left(X^TX\right)^{-1}$$
$$= \sigma^2\left(X^TX\right)^{-1}$$

- Do $e = (I-H)Y$ nên:

$$E(e) = (I-H)E(Y) = (I-H)X\beta = X\beta - HX\beta$$
$$= X\beta - X.(X^TX)^{-1}X^TX.\beta = X\beta - X\beta = 0$$
$$\text{cov}(e) = (I-H)I(I-H)\sigma^2 = \sigma^2(I-H)$$

- Có *trace* là tổng các phần tử trên đường chéo chính của ma trận vuông và $tr(A-B) = tr(A) - tr(B)$. Từ đó suy ra:

$$E\left(e^Te\right) = \sum_{j=1}^{n}E(e_j^2) = \text{tr}(\text{cov}(e)) = \sigma^2\,\text{tr}\,(I_n - H)$$
$$= \sigma^2(n - \text{tr}(H))$$

Bên cạnh đó:

$$\text{tr}(H) = \text{tr}\left(X\left(X^TX\right)^{-1}X^T\right) = \text{tr}\left(\left(X^TX\right)^{-1}X^TX\right) = \text{tr}\,(I_{k+1}) = k+1$$

$\Rightarrow$ Ước lượng không chệch của σ^2 là: $\widehat{\sigma}^2 = \dfrac{e^Te}{n-k-1}$

2.3 Hệ số xác định R^2

Ta có:

$$Y_i = \widehat{Y}_i + e_i$$

$$\Leftrightarrow Y_i - \bar{Y} = \widehat{Y}_i - \bar{Y} + e_i$$

Tổng bình phương 2 vế đẳng thức trên:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2$$

$$\Leftrightarrow \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Đặt:

- $SS_E = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, thể hiện sự thay đổi của Y do các yếu tố khác không nghiên cứu ngoài các biến độc lập X và gọi là tổng bình phương sai số *(sum of squares for error)*
- $SS_R = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$, thể hiện sự biến đổi của giá trị ước lượng $\widehat{Y}$ quanh giá trị trung bình mẫu của nó và gọi là tổng bình phương hồi quy *(sum of squares for regression).*
- $SS_T = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ là độ giao động trong mẫu của biến phụ thuộc, thể hiện sự biến đổi của biến Y quanh giá trị trung bình mẫu của nó *(sum of squares for total).*

Khi đó ta có:

$$SS_T = SS_R + SS_E. \tag{2.17}$$

2.3.1 Hệ số xác định R^2:

Ký hiệu R^2 được sử dụng cho hệ số xác định của hàm hồi quy bội.

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \tag{2.18}$$

gọi là bình phương của hệ số xác định, đó là tỷ lệ biến thiên của các biến Y được giải thích bởi các biến $X_1, \cdots, X_k$.
Do SS_T, SS_R, SS_E đều không âm nên R^2 nhận giá trị trong [0; 1].

- $R^2 = 1$, nghĩa là đường hồi quy giải thích hoàn toàn sự thay đổi của Y.
- $R^2 = 0$, mô hình không giải thích sự thay đổi nào của Y.
- R^2 bằng phần trăm sự thay đổi của biến phụ thuộc được giải thích bởi các biến độc lập trong mô hình.

R^2 là giá trị gắn liền với mẫu, đo mức độ phù hợp của mô hình với số liệu mẫu.

2.3.2 Hệ số R^2 hiệu chỉnh

Khi tiến hành thêm bất kỳ một biến nào vào mô hình (dù biến này có ý nghĩa hay không) thì R^2 sẽ không bao giờ giảm. Không thể dùng R^2 làm tiêu chuẩn để xem xét việc đưa thêm một biến giải thích vào mô hình. R^2 còn phụ thuộc vào số bậc tự do của SS_E, SS_T tương ứng là (n-k-1) và (n-1). Để ngăn chặn tình trạng trên, người ta đưa ra khái niệm R^2 hiệu chỉnh - ký hiệu $\overline{R}^2$ và được định nghĩa:

$$\overline{R}^2 = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)} \tag{2.19}$$

Thuật ngữ "điều chỉnh" nghĩa là điều chỉnh số bậc tự do, mà số bậc tự do này phụ thuộc vào số biến giải thích(k) trong mô hình. Việc thêm một biến dẫn đến tăng R^2 nhưng cũng làm giảm đi một bậc tự do. $\overline{R}^2$sẽ chỉ tăng khi một biến được thêm vào mô hình nếu biến mới làm giảm bình phương trung bình sai số. R^2 và $\overline{R}^2$ càng gần 1 thì mô hình càng có ý nghĩa.

2.4 Kiểm định giả thuyết về các hệ số hồi quy

Theo giả thiết 1: ε_j có cùng phân phối chuẩn $N(0, \sigma^2)$ và độc lập, tức là $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ có phân phối chuẩn $N_n(0, \sigma^2 I_n)$, từ đó ta có:

Giả thiết 2:

i. $\widehat{\beta}$ có phân phối chuẩn $N(\beta, \sigma^2(X^TX)^{-1})$

ii. $\dfrac{(n-k-1)\widehat{\sigma}^2}{\sigma^2} = \dfrac{\Sigma_{j=1}^n e_j^2}{\sigma^2}$ có phân phối χ^2 với (n-k-1) bậc tự do.

iii. $\widehat{\beta}, \widehat{\sigma}^2$ là độc lập.

2.4.1 Kiểm định về một hệ số hồi quy

Khi xây dựng mô hình hồi quy, ta giả sử rằng tất cả các biến độc lập tham gia vào mô hình hồi quy. Tuy nhiên trong thực tế sẽ có một vài biến sẽ không tham gia vào hàm hồi quy, hoặc có ít ảnh hưởng đến biến phụ thuộc và tùy theo mục tiêu nghiên cứu ta có thể lược bỏ biến đó nếu biến không quan trọng và ảnh hưởng đáng kể đến mô hình.

Nếu X_j không tác động đến biến Y thì $\beta_j = 0$ và ngược lại, nếu X_j tác động đến biến Y thì $\beta_j \neq 0$. Từ đó để kiểm tra vấn đề này, ta xét cặp giả thuyết:

$$H_0 : \beta_j = 0, H_1 : \beta_j \neq 0$$

Với giả thiết $\widehat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$ thỏa mãn. Xét thống kê:

$$T_0 = \frac{\widehat{\beta_j} - \beta_j}{\sqrt{\sigma^2(X^TX)^{-1}}} \tag{2.20}$$

do không xác định được σ^2 nên ta sẽ sử dụng $\widehat{\sigma}^2$ là một ước lượng điểm của σ^2 để thực hiện tính toán.

Nếu giả thuyết H_0 đúng và các giả thuyết 1,2 thỏa mãn suy ra T tuân theo quy luật Student với số bậc tự do là (n-k-1)
Ta có miền bác bỏ giả thuyết H_0:

$$W_\alpha = \left\{T_{qs} = \frac{\hat{\beta}_j}{\sqrt{\widehat{\sigma}^2(X^TX)^{-1}}}; |T_{qs}| > t_{\alpha/2,n-k-1}\right\} \tag{2.21}$$

2.4.2 Kiểm định mức ý nghĩa hồi quy

Xét mô hình hồi quy tuyến tính:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

Kiểm định tính có ý nghĩa của hồi quy là một kiểm định để xác định xem có mối quan hệ tuyến tính giữa phản hồi Y và bất kỳ biến độc lập nào X_1, X_2, ..., X_k hay không. Quy trình này thường được coi là một kiểm định về tính phù hợp của mô hình. Cặp giả thuyết thích hợp là:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k.$$
$$H_1 : \beta_j \neq 0 \text{ với ít nhất một j, j = 1,2,..,k.}$$

Giả thuyết H_1 có nghĩa là các biến độc lập không tham gia vào biểu thức tuyến tính, ngược lại đối thuyết H_1 nói rằng có ít nhất một trong các biến này có liên quan đến mô hình. Quy trình kiểm định là một sự tổng quát hóa của phân tích phương sai được sử dụng trong hồi quy tuyến tính[4].

$$SS_T = SS_R + SS_E$$

Với $\frac{SS_R}{\sigma^2} \sim \chi^2_k$. Số bậc tự do k trong phân phối này bằng với số biến hồi quy trong mô hình hồi quy. Theo giả thiết 2(ii), $\frac{SS_E}{\sigma^2} \sim \chi^2_{n-k-1}$. SS_E và SS_R độc lập.

Từ đó ta chọn tiêu chuẩn kiểm định:

$$F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MS_R}{MS_E}$$

trong đó, MS_R và MS_E là bình phương trung bình, tính bằng cách chia tổng bình phương cho số bậc tự do tương ứng.

Nếu giả thuyết $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$ là đúng thì F_0 có phân phối Fisher với k và n-k-1 bậc tự do.

Và ta bác bỏ giả thuyết H_0 nếu:

$$F_{qs} > f_{\alpha,k,n-k-1}$$

Khi đó ta kết luận rằng mô hình không giải thích sự thay đổi nào của Y. Thủ tục kiểm định thường được tóm tắt trong bảng phân tích phương sai (analysis-of-variance table) như sau:

Nguồn	**Tổng bình phương**	**Bậc tự do**	**Bình phương trung bình**	F_0
Hàm hồi quy	SS_R	k	$MS_R = \frac{SS_R}{k}$	$\frac{MS_R}{MS_E}$
Phần dư	SS_E	$n-k-1$	$MS_E = \frac{SS_E}{n-k-1}$	
Tổng	SS_T	$n-1$		

Bảng 2.1: Bảng phân tích phương sai

- Mặt khác với công thức: $R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_R}$. Tiêu chuẩn kiểm định F có thể viết lại như sau:

$$F_0 = \frac{R^2/k}{(1-R^2)/(n-k-1)} \tag{2.22}$$

cho nên quá trình phân tích phương sai cho phép đưa ra phán đoán thống kê về độ thích hợp của hàm hồi quy. Cặp giả thuyết kiểm định có thể được viết lại như sau:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$ hay $R^2 = 0$.

$H_1 : \beta_j \neq 0$ với ít nhất một j, j = 1,2,..,k hay $R^2 > 0$

2.5 Ước lượng khoảng của mô hình hồi quy

2.5.1 Khoảng tin cậy của các hệ số hồi quy β_j

Xét mô hình hồi quy tuyến tính $Y = X\beta + \varepsilon$. Khi đó khoảng tin cậy hai phía với độ tin cậy $(1-\alpha)$ của β xác định bởi:

$$\widehat{\beta_j} \pm t_{\alpha/2,n-k-1}\sqrt{\widehat{\sigma}^2 C_{jj}}$$

trong đó $t_{\alpha/2,n-k-1}\sqrt{\widehat{\sigma}^2 C_{jj}}$ là phân vị trên mức α của phân phối Student với bậc tự do là $n-k-1$.

Chứng minh:

Vì $\widehat{\beta}-\beta$ có phân phối chuẩn $N(0, \sigma^2(X^TX)^{-1})$ nên $\widehat{\beta}_j - \beta_j$ có phân phối chuẩn $N(0, \sigma^2 C_{jj})$, với C_{jj} là phần tử trên đường chéo chính ma trận $(X^TX)^{-1}$ và $\dfrac{(n-k-1)\widehat{\sigma}^2}{\sigma^2}$ có phân phối χ^2 với (n-k-1) bậc tự do. Do đó:

$$\Rightarrow \frac{(\widehat{\beta}-\beta)/\sqrt{\sigma^2(X^TX)^{-1}}}{\sqrt{\widehat{\sigma}^2/\sigma^2}} = \frac{\widehat{\beta_j}-\beta_j}{\sqrt{\widehat{\sigma^2}C_{jj}}} \sim t_{n-k-1}$$

$$\Rightarrow P\left(\frac{|\widehat{\beta_j}-\beta_j|}{\sqrt{\widehat{\sigma^2}C_{jj}}} < t_{\alpha/2,n-k-1}\right) = 1-\alpha$$

$\Rightarrow$ khoảng tin cậy của β_j với mức tin cậy $(1-\alpha)$ là :

$$\widehat{\beta_j} \pm t_{\alpha/2,n-k-1}\sqrt{\widehat{\sigma}^2 C_{jj}} \tag{2.23}$$

Nhận xét:

Với độ tin vậy là $(1-\alpha)$ khi biến X_j tăng 1 đơn vị và các yếu tố khác không đổi thì trung bình của biến Y tăng trong khoảng này.

Tương tự:

- Với khoảng tin cậy ước lượng giá trị lớn nhất cho hệ số hồi quy:

$$\beta_j < \widehat{\beta_j} + t_{\alpha,n-k-1}\sqrt{\widehat{\sigma}^2 C_{jj}}$$

- Khoảng tin cậy ước lượng giá trị bé nhất cho hệ số hồi quy:

$$\beta_j > \widehat{\beta_j} - t_{\alpha,n-k-1}\sqrt{\widehat{\sigma}^2 C_{jj}}$$

2.5.2 Các yếu tố ảnh hưởng đến khoảng tin cậy

Với độ tin cậy cố định, chúng ta sẽ quan tâm đến độ dài khoảng tin cậy đối xứng, khoảng tin cậy hẹp sẽ cho biết chính xác hơn về giá trị của hệ số cần ước lượng và ngược lại khoảng tin cậy quá rộng thì thông tin cho giá trị cần ước lượng kém chính xác hơn. Công thức (2.23) cho thấy độ dài khoảng tin cậy cho hệ số β_j bằng: $2t_{\alpha/2,n-k-1}\sqrt{\widehat{\sigma}^2 C_{jj}}$. Giá trị này phụ thuộc vào các yếu tố:

- Số bậc tự do (n-k-1): số bậc tự do càng bé thì $t_{\alpha/2,n-k-1}$ càng lớn và như vậy khoảng tin cậy càng rộng
- Mối tương quan tuyến tính giữa X_j và các biến độc lập còn lại trong mô hình, mối tương quan này được đo bởi R_j^2. Mối tương quan tuyến tính càng chặt thì R_j^2 này càng cao dẫn đến $\sqrt{\widehat{\sigma}^2 C_{jj}}$ lớn và khoảng sẽ rộng. Khi R_j^2 gần đến 1, khoảng tin cậy sẽ rộng ra vô cùng và ước lượng không còn ý nghĩa.

2.6 Một số vấn đề liên quan trong mô hình hồi quy tuyến tính

2.6.1 Hiện tượng đa cộng tuyến

Đa cộng tuyến là hiện tượng trong mô hình hồi quy tuyến tính, khi hai hoặc nhiều biến độc lập trong mô hình có mức độ tương quan cao với nhau. Khi có đa cộng tuyến, sự ảnh hưởng của các biến độc lập đến biến phụ thuộc không thể được ước tính một cách chính xác, và do đó làm giảm tính tin cậy của mô hình.

Ở trong chương trước theo giải thiết 1, các biến độc lập không tương quan tức tồn tại ma trận $(XX^T)^{-1}$ nên các công thức ước lượng ở trên đã được suy ra dễ dàng. Tuy nhiên trong thực tế việc xuất hiện mối liên quan giữa các biến độc lập làm cho các suy diễn từ mô hình không còn được chính xác.

Ta xét mô hình hồi quy gồm 2 biến độc lập X_1, X_2 với số liệu mẫu:

$$y_i = \beta_i x_{1i} + \beta_2 x_{2i} + \varepsilon_i, i = 1, \cdots, n \tag{2.24}$$

Ta có $X^T X$ là ma trận tương quan:

$$X^T X = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$$

trong đó: $r_{12} = \dfrac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2}}$

Từ đó ta có:

$$Var(\widehat{\beta}) = \sigma^2(X^TX)^{-1} = \sigma^2.\frac{1}{1-r_{12}^2}.\begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}.$$

$$\Rightarrow V(\widehat{\beta_1}) = \frac{1}{1-r_{12}^2}.\sigma^2 = V(\widehat{\beta_2}) \tag{2.25}$$

$$cov(\widehat{\beta_1}, \widehat{\beta_2}) = -\frac{r_{12}}{1-r_{12}^2}.\sigma^2 \tag{2.26}$$

Nếu biến X_1, X_2 tương quan với nhau $r_{12} \approx 1$(hoặc -1), suy ra $V(\widehat{\beta_1}) = V(\widehat{\beta_2}) \approx \infty$
Phương sai vô hạn nên không thể ước lượng hệ số hồi quy được.

Cách nhận biết đa cộng tuyến:

Một biện pháp đơn giản để phát hiện đa tuyến tính là kiểm tra ma trận tương quan. Nếu các biến hồi quy X_i và X_j gần như tuyến tính phụ thuộc vào nhau, thì $|r_{ij}|$ sẽ có giá trị gần bằng 1. Thông thường $|r_{ij}| > 0.75$ thì xảy ra hiện tượng đa cộng tuyến.

$$r_{ij} = s_{ij}/\sqrt{s_{jj}s_{ii}} \tag{2.27}$$

trong đó:

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ik} - \overline{x_i}) \times (x_{jk} - \overline{x_j})$$

Khắc phục

Thực hiện bỏ bớt biến độc lập ra khỏi mô hình nếu có thể.

- Tính các hệ số tương quan mẫu r_{ij}. Xem cặp biến giải thích nào có quan hệ chặt chẽ với nhau. Biến $X_1.X_2, ..., X_k$ là các biến độc độc lập, Y là biến phụ thuộc và X_i, X_j có tương quan chặt chẽ với nhau.

- Tính mối tương quan của biến X_i, X_j với biến Y theo công thức:

$$r_{yi} = s_{yi}/\sqrt{s_{ii}s_{yy}}.$$

trong đó $s_{yy} = \dfrac{1}{n-1}\sum_{k=1}^{n}(y_k - \overline{y}) \times (y_k - \overline{y})$; $s_{yi} = \dfrac{1}{n-1}\sum_{k=1}^{n}(y_k - \overline{y}) \times (x_{ik} - \overline{x_i})$.

- Khi đó:

– Ta loại biến X_i ra khỏi mô hình nếu Biến X_i tương quan với Y nhỏ hơn biến X_j:$|r_{yi}| < |r_{yj}|$.

– Ta loại biến X_i ra khỏi mô hình nếu Biến X_j tương quan với Y nhỏ hơn biến X_i: $|r_{yi}| > |r_{yj}|$.

Bản chất của bước này chính là ta giữ lại biến có độ tương quan với biến phụ thuộc Y cao hơn và loại bỏ biến có độ tương quan với Y thấp hơn.

- Thực hiện hồi quy sau khi ma trận X đã loại bỏ biến.

2.6.2 Phương sai sai số thay đổi

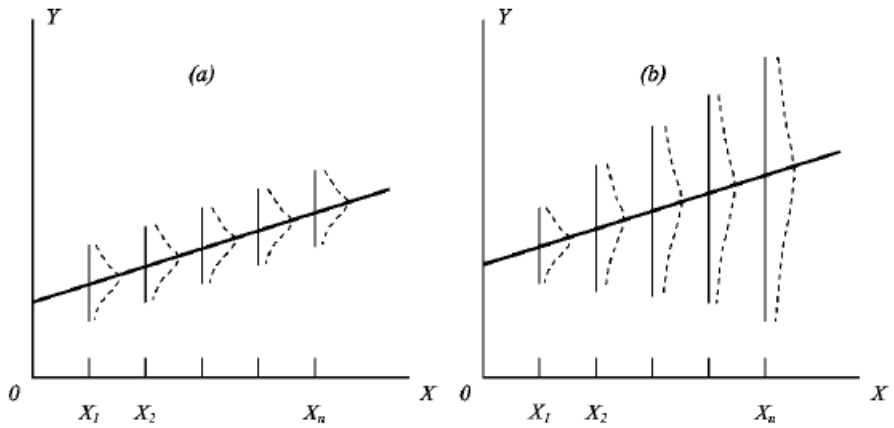
Một vấn đề khác mà mô hình cũng có thể gặp phải, đó là phương sai của sai số thay đổi.

$$Var(\varepsilon_i) = \sigma^2$$

Trong thực tế, sai số có thể tăng hoặc giảm khi giá trị của biến độc lập thay đổi:

$$Var(\varepsilon_i) = \sigma_i^2$$

Khi đó xảy ra hiện tượng phương sai sai số thay đổi



Hình 2.4: (a)Phương sai sai số không đổi (b) Phương sai sai số thay đổi

Hậu quả: Các ước lượng hệ số hồi quy vẫn là các ước lượng không chệch, tuy nhiên không còn là các ước lượng hiệu quả nhất.
Theo công thức tính chất ước lượng bình phương cực tiểu:

$$E(\widehat{\beta}) = \beta$$

$$V(\widehat{\beta}) = \sigma^2.C_{jj}$$

trong đó, C_{jj} là các phần tử trên đường chéo ma trận $(X^TX)^{-1}$.

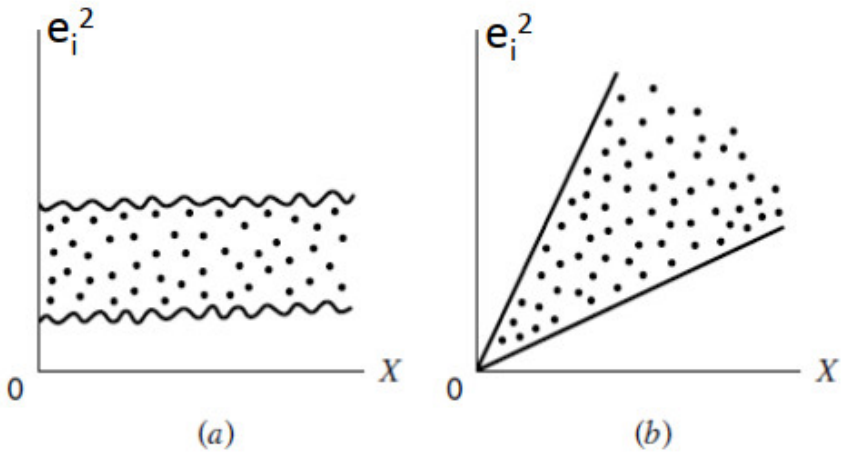
Nếu phương sai sai số thay đổi: $V(\varepsilon_i) = \sigma_i^2$, dẫn đến không xác định được phương sai nhỏ nhất nên $\widehat{\beta}$ không phải là ước lượng có phương sai tốt nhất.

Do không xác định được tổng thể, ta sẽ sử dụng phần dư(e) là ước lượng điểm của sai số(ε) với số liệu mẫu cho trước.

Khảo sát đồ thị

- Ước lượng mô hình bằng OLS, tìm được các phần dư e_i
- Vẽ đồ thị của các phần dư e_i hoặc e_i^2 theo X.
- Căn cứ vào các đồ thị để chuẩn đoán về hiện tượng phương sai sai số thay đổi. Nếu độ rộng của e_i hoặc e_i^2 tăng hoặc giảm khi X tăng thì xảy ra hiện tượng phương sai sai số thay đổi.

Ví dụ:



Hình 2.5: Khảo sát đồ thị phần dư

Đồ thị (a) ta thấy không xảy ra hiện tượng phương sai sai số thay đổi. Đồ thị (b) phương sai sai số thay đổi tuyến tính theo X.

Kiểm định White

Bản chất của kiểm định White[5] là kiểm tra tính đồng nhất về phương sai của mô hình hồi quy tuyến tính bằng cách đo lường sự phụ thuộc của phương sai của sai số dự báo vào giá trị của các biến độc lập. Xét mô hình hồi quy tuyến tính 2 biến:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{2.28}$$

Các bước tiến hành:

- Ước lượng mô hình hồi quy gốc (2.28) tìm được các phần dư e_i.
- Ước lượng mô hình hồi quy phụ bằng phương pháp OLS:

$$e_i^2 = \lambda_0 + \lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_1^2 + \lambda_4 X_2^2 + \lambda_5 X_1 X_2 + v_i \tag{2.29}$$

  thu được hệ số xác định R_e^2.
  Trong đó: v_i là sai số ngẫu nhiên thoả mãn mọi giả thiết của OLS.

  Mô hình hồi quy phụ gồm $m = 6 = 2k + 1 + C_k^2$ hệ số λ, với k là số biến độc lập trong mô hình gốc.

- Cặp giả thuyết kiểm định:

$$H_0 : \lambda_1 = \lambda_2 = \cdots = \lambda_5 = 0$$
$$H_1 : \lambda_1^2 + \lambda_2^2 + \cdots + \lambda_5^2 \neq 0$$

- Tiêu chuẩn kiểm định:

$$\chi^2 = n \times R_e^2 \sim \chi^2(5)$$

  với 5 là số biến độc lập trong mô hình không kể hệ số chặn.
  Trường hợp với mô hình tổng quát thì tiêu chuẩn kiểm định: $\chi^2 = n \times R_e^2 \sim \chi^2(m-1)$, với m là số hệ số của mô hình hồi quy phụ.

- Với mẫu cụ thể và với mức ý nghĩa α cho trước, ta bác bỏ giả thuyết H_0 nếu:

$$\chi_{qs}^2 > \chi_\alpha^2(5)$$

Mô hình hồi quy phụ bắt buộc phải có hệ số chặn, có thể không có các số hạng chéo nhưng cũng có thể có bậc cao hơn của các biến độc lập.

Khắc phục

a, Trường hợp đã biết σ_i^2

Xét mô hình: $Y_i = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon_i$
Giả sử mô hình có phương sai sai số thay đổi, khi đó ta khắc phục bằng cách chia cả 2 vế của mô hình trên cho σ_i(với $\sigma_i \neq 0$) ta thu được mô hình sau:

$$\frac{Y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{X_1}{\sigma_i} + ... + \beta_k \frac{X_k}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i}$$

$$Y_i^* = \beta_0^* + \beta_1^* X_1 + ... + \beta_k^* X_k + u_i \tag{2.30}$$

Có:

$$Var(u_i) = Var(\frac{\varepsilon_i}{\sigma_i}) = \frac{1}{\sigma_i} Var(\varepsilon_i) = \frac{1}{\sigma_i} \sigma_i = 1$$

Từ đó mô hình (2.30) có phương sai sai số không đổi. Như vậy ta sẽ ước lượng mô hình trên thay cho mô hình gốc. Sau khi được kết quả ta nhân cả 2 vế với ε_i để quay lại mô hình ban đầu.

b, Trường hợp chưa biết σ_i^2

Trường hợp này ta dựa trên giả thiết phương sai sai số thay đổi như thế nào theo các yếu tố thì có thể chia 2 vế của phương trình hồi quy cho căn bậc 2 của yếu tố đó. Từ đó phương trình hồi quy có thể biến đổi như sau:

- $\sigma^2 = \lambda . X_i^2$, đồ thị giữa e^2 và X_i có dạng Parabol.
  Chia mô hình cho biến độc lập (X_i):

$$\frac{Y_i}{X_i} = \beta_0 \frac{1}{X_i} + \beta_1 \frac{X_1}{X_i} + ... + \beta_k \frac{X_k}{X_i} + \frac{\varepsilon_i}{X_i}$$

$$Y_i^* = \beta_0^* + \beta_1^* X_1 + ... + \beta_k^* X_k + u_i \tag{2.31}$$

$$Var(u_i) = Var(\frac{\varepsilon_i}{X_i}) = \frac{1}{X_i} Var(\varepsilon_i) = \frac{1}{X_i} X_i = 1$$

  Mô hình hồi quy trên thỏa mãn phương sai sai số không đổi.

- Tương tự chia cho căn bậc 2 biến độc lập ($\sqrt{X_i}$), nếu giả thuyết $\sigma^2 = \lambda . X_i$.

2.6.3 Hiện tượng tự tương quan

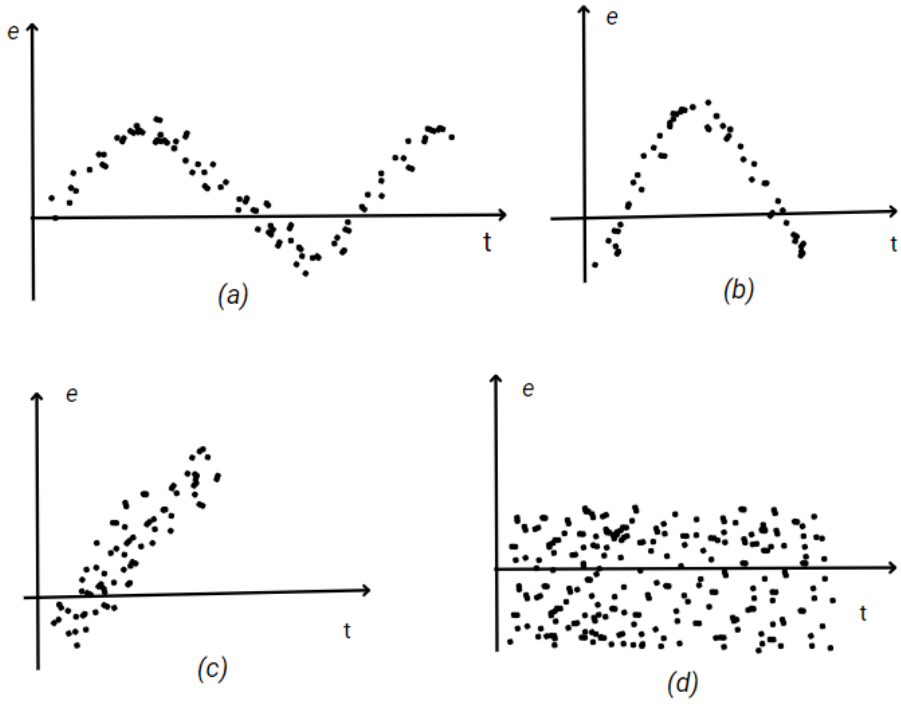
Trong mô hình hồi quy ta luôn giả thiết không có sự tương quan giữa các sai số ngẫu nhiên. Tức là yếu tố ngẫu nhiên gắn với một quan sát nào đó không bị ảnh hưởng bởi yếu tố ngẫu nhiên gắn với một quan sát khác. Tuy nhiên trong thực tế giả thiết này có thể bị vi phạm.

- Xét mô hình với số liệu chéo hay số liệu không gian: $Y_i = \beta X_i + \varepsilon_i$
  Giả thiết có thể bị vi phạm: $cov(\varepsilon_i, \varepsilon_j) \neq 0; (\forall i \neq j)$

- Xét mô hình sau với số liệu theo thời gian:$Y_t = \beta X_t + \varepsilon_t$
  Giả thiết có thể bị vi phạm: $cov(\varepsilon_t, \varepsilon_{t+k}) \neq 0;$(tồn tại k $\neq 0$)

Hiện tượng tự tương quan thường xảy ra với các số liệu theo thời gian.

Kiểm tra bằng đồ thị



Hình 2.6: Đồ thị kiểm tra tính tương quan theo thời gian

Ta thấy đồ thị a,b,c các giá trị e_i đều có phân theo một trình tự nào đó theo thời gian t, nên xảy ra hiện tượng tự tương quan. Còn ở đồ thị d các giá trị e_i xáo trộn một cách ngẫu nhiên không theo trình tự nên không có hiện tượng tự tương quan.

Kiểm tra bằng phương pháp Durbin-Watson

Tiêu chuẩn Durbin - Watson[6] là tiêu chuẩn thống kê được đưa ra để đánh giá, kiểm định sự tự tương quan trong các phần dư của mô hình hồi quy. Trị số thống kê Durbin - Watson (kí hiệu là DW) nhận giá trị từ 0 đến 4. Giá trị $DW = 2$ cho ta biết trong mô hình không có sự tự tương quan giữa các phần dư.

Công thức thống kê Durbin - Watson:

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \approx 2 - 2\frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2} = 2 - 2r \tag{2.32}$$

Do $-1 \leq r \leq 1$ nên $0 \leq DW \leq 4$.

Các bước tiến hành:

- Hồi quy mô hình gốc tìm được e_t, e_{t-1}

- Tính thống kê DW theo công thức(2.32)
- Với mức ý nghĩa α cho trước tra bảng Durbin Watson ta tìm được hai số $d_1(k, n, \alpha) < d_2(k, n, \alpha)$, khi đó dựa vào các kết quả sau để đưa ra kết luận:
  - Nếu $0 \leq DW < d_1$: có tự tương quan dương
  - Nếu $d_1 \leq DW \leq d_2$: không có kết luận.
  - Nếu $d_2 < DW < 4 - d_2$: không có tự tương quan.
  - Nếu $4 - d_2 \leq DW \leq 4 - d_1$: không có kết luận.
  - Nếu $4 - d_1 < DW \leq 4$: có tự tương quan âm.

Chương 3

Thử nghiệm số

Từ cở sở lý thuyết ở trên, trong báo cáo này em áp dụng vào việc xây dựng mô hình hồi quy tuyến tính dự báo giá xe ô tô cũ. Bộ dữ liệu em sử dụng gồm có các đặc trưng sau:

- **Brand**: Tên thương hiệu của xe.
- **Year**: Năm sản xuất xe.
- **Kilometers**: Tổng số Km mà xe đã đi.
- **Fuel**: Loại nhiên liệu xe sử dụng.
- **Transmission**: Loại hộp số xe.
- **Ower**: Thứ tự chủ sở hữu cũ của xe.
- **Mileage**: Mức tiêu thụ nhiên liệu.
- **Engine**: Thể tích xylanh của động cơ.
- **Power**: Công suất tối đa.
- **Seat**: Số ghế ngồi của xe.
- **Price**: Giá bán của xe.

Thực hiện đổi biến năm sản xuất **Year** thành biến tuổi của xe **Ageofcar**. Bộ dữ liệu được sử dụng về giá của các loại xe cũ phổ thông, không dự báo về giá của các loại xe cổ.

Sau đây là quá trình xử lý dữ liệu và xây dựng mô hình dự báo:

Bước 1: Tiền xử lý dữ liệu

- Tổng quan về bộ dữ liệu:

	Brand	Kilometers	Fuel	Transmission	Owner	Mileage	Engine	Power	Seats	Price	Ageofcar
count	5086	5.086000e+03	5086	5086	5086	5084.000000	5063.000000	4990.000000	5058.000000	4959.000000	5086.000000
unique	10	NaN	5	2	4	NaN	NaN	NaN	NaN	NaN	NaN
top	Maruti	NaN	Diesel	Manual	First	NaN	NaN	NaN	NaN	NaN	NaN
freq	1257	NaN	2680	3690	4208	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	5.836126e+04	NaN	NaN	NaN	18.525744	1605.632431	111.837162	5.310399	9.270617	8.533032
std	NaN	9.759020e+04	NaN	NaN	NaN	4.224189	584.300268	50.975248	0.813991	10.227709	3.228193
min	NaN	1.710000e+02	NaN	NaN	NaN	7.810000	72.000000	34.200000	0.000000	0.450000	3.000000
25%	NaN	3.390025e+04	NaN	NaN	NaN	15.600000	1197.000000	75.000000	5.000000	3.650000	6.000000
50%	NaN	5.222200e+04	NaN	NaN	NaN	18.500000	1493.000000	91.100000	5.000000	5.750000	8.000000
75%	NaN	7.200000e+04	NaN	NaN	NaN	21.430000	1968.000000	136.000000	5.000000	9.950000	10.000000
max	NaN	6.500000e+06	NaN	NaN	NaN	28.400000	5461.000000	450.000000	10.000000	93.670000	24.000000

Hình 3.1: Tổng quan về bộ dữ liệu

Từ bảng mô tả dữ liệu ở trên ta có thể thấy được:

  - Số lượng hàng có giá trị của các cột khác nhau do xuất hiện các giá trị NULL nên ta cần xử lý mất mát của bộ dữ liệu này.
  - Các cột **Brand**, **Fuel**, **Owner** là các cột có dự liệu không phải dạng số nên ta cần đặt biến giả cho các giá trị trong các cột này.
  - Các cột dữ liệu còn lại là các dữ liệu dạng số.

- **Xử lý giá trị NULL:**
  - Từ thống kê về các giá trị NULL ta thấy các giá trị mất mát không vượt quá 5% của các cột dữ liệu nên ta có thể loại bỏ các giá trị này.

```
data.isnull().sum()
```

```
Name               0
Year               0
Kilometers         0
Fuel               0
Transmission       0
Owner_Type         0
Mileage            2
Engine            23
Power             96
Seats             28
Price            127
dtype: int64
```

Hình 3.2: Thống kê số lượng các giá trị NULL

– Bộ dữ liệu sau khi loại bỏ các giá trị NULL:

	Brand	Kilometers	Fuel	Transmission	Owner	Mileage	Engine	Power	Seats	Price	Ageofcar
0	Mercedes-Benz	35277	Petrol	Automatic	First	7.81	5461.0	362.90	5.0	30.00	12
1	BMW	65329	Petrol	Automatic	First	7.94	4395.0	450.00	4.0	20.72	12
2	BMW	5900	Petrol	Automatic	First	7.94	4395.0	450.00	4.0	47.50	11
3	Mercedes-Benz	35000	Petrol	Automatic	First	8.10	5461.0	387.30	2.0	29.50	12
4	BMW	49000	Diesel	Automatic	First	8.20	2993.0	245.00	4.0	29.00	9
...	...	...	...	...	...	...	...	...	...	...	...
4861	Maruti	39437	Diesel	Manual	First	28.40	1248.0	74.00	5.0	5.15	7
4862	Maruti	103000	Diesel	Manual	Second	28.40	1248.0	74.00	5.0	5.25	8
4863	Maruti	48000	Diesel	Manual	First	28.40	1248.0	73.75	5.0	6.00	5
4864	Maruti	27000	Diesel	Manual	First	28.40	1248.0	73.75	5.0	5.90	4
4865	Maruti	27365	Diesel	Manual	First	28.40	1248.0	74.00	5.0	4.75	8

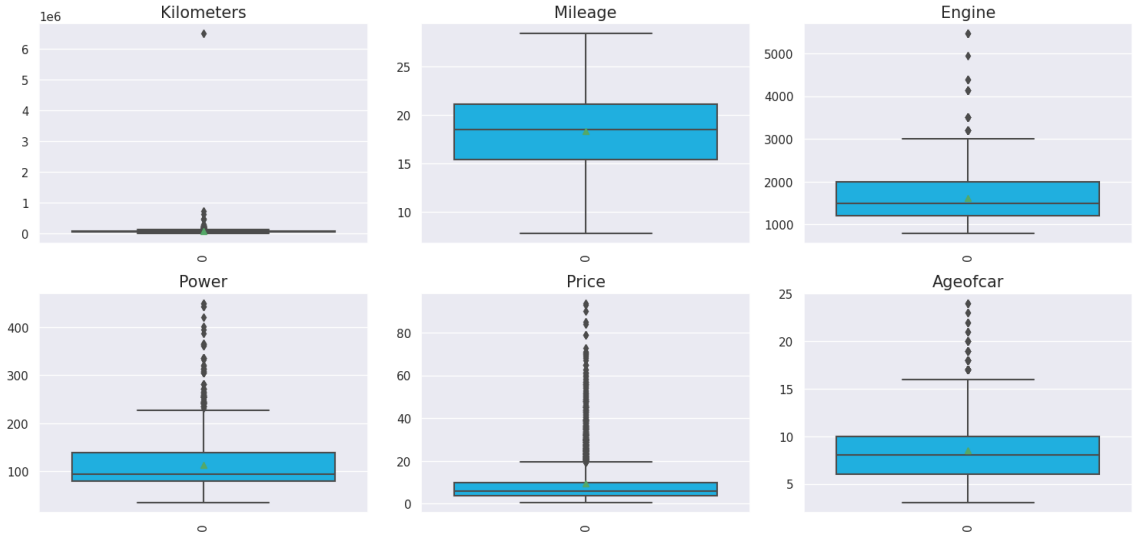
4866 rows × 11 columns

Hình 3.3: Bộ dữ liệu sau khi loại bỏ các giá trị NULL

- **Các điểm ngoại lệ:**

  Ta sử dụng trực quan các điểm ngoại lệ thông qua biểu đồ boxlpot để xem xét các điểm ngoại lệ ảnh hưởng đến hiệu suất của mô hình.

  Dựa trên đồ thị boxplot, ta xác định các giá trị nằm ngoài hai đường thẳng từ hộp đến giá trị cao nhất và thấp nhất trong khoảng biến thiên. Các giá trị này được coi là outlier.



Hình 3.4: Boxplot mô tả phân phối dữ liệu của các biến

Từ đồ thị trên ta thấy các biến **Kilometers**, **Engine**, **Power**, **Price**, **Ageofcar** xuất hiện nhiều các điểm ngoại lại, để tránh giảm hiệu suất mô hình dự đoán ta cần loại bỏ các điểm ngoại lệ.

Sau khi loại bỏ outlier bộ dữ liệu còn 3918 hàng.

	Brand	Kilometers	Fuel	Transmission	Owner	Mileage	Engine	Power	Seats	Price	Ageofcar
count	3918	3918.000000	3918	3918	3918	3918.000000	3918.000000	3918.000000	3918.000000	3918.000000	3918.000000
unique	10	NaN	4	2	4	NaN	NaN	NaN	NaN	NaN	NaN
top	Maruti	NaN	Petrol	Manual	First	NaN	NaN	NaN	NaN	NaN	NaN
freq	1086	NaN	2128	3223	3286	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	52934.822614	NaN	NaN	NaN	19.125651	1453.740939	97.310240	5.266207	6.014191	8.392292
std	NaN	25913.918507	NaN	NaN	NaN	3.953435	433.823043	32.333991	0.730346	3.393573	2.863861
min	NaN	171.000000	NaN	NaN	NaN	9.000000	793.000000	34.200000	4.000000	0.690000	3.000000
25%	NaN	33431.250000	NaN	NaN	NaN	16.470000	1197.000000	74.000000	5.000000	3.500000	6.000000
50%	NaN	52000.000000	NaN	NaN	NaN	18.900000	1298.000000	88.500000	5.000000	5.250000	8.000000
75%	NaN	69918.000000	NaN	NaN	NaN	22.070000	1582.000000	117.300000	5.000000	7.540000	10.000000
max	NaN	128000.000000	NaN	NaN	NaN	28.400000	2997.000000	218.000000	9.000000	17.110000	16.000000

Hình 3.5: Dữ liệu sau khi loại bỏ các outlier

- **Tạo biến giả:**
  - Với mỗi cột có m giá trị, ta cần tạo m-1 biến giả. Giá trị không được tạo biến giả sẽ làm cơ sở.
    - Cột **Brand**: giá trị 'Audi' làm cơ sở, còn lại là biến giả.
    - Cột **Fuel**: giá trị 'Diesel' làm cơ cở.
    - Cột **Transmission**: giá trị 'Automatic' làm cơ sở.
    - Cột **Owner**: giá trị 'First' làm cơ sở.
  - Kết quả:

	Kilometers	Mileage	Engine	Power	Seats	Price	Ageofcar	Brand_BMW	Brand_Ford	Brand_Honda	...	Brand_Mercedes-Benz	Brand_Toyota	Brand_Volkswagen	Fuel_Diesel	Fuel_LPG
12	120000	9.00	2997.0	218.00	5.0	2.65	16	0	0	1	...	0	0	0	0	0
17	77000	9.70	1995.0	163.50	5.0	5.50	14	0	0	0	...	0	0	0	0	0
18	99100	9.74	1984.0	208.00	5.0	10.00	13	0	0	0	...	0	0	0	0	0
20	42000	9.80	2354.0	180.00	5.0	6.25	10	0	0	1	...	0	0	0	0	0
21	60000	9.80	2354.0	180.00	5.0	3.50	13	0	0	1	...	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4861	39437	28.40	1248.0	74.00	5.0	5.15	7	0	0	0	...	0	0	0	1	0
4862	103000	28.40	1248.0	74.00	5.0	5.25	8	0	0	0	...	0	0	0	1	0
4863	48000	28.40	1248.0	73.75	5.0	6.00	5	0	0	0	...	0	0	0	1	0
4864	27000	28.40	1248.0	73.75	5.0	5.90	4	0	0	0	...	0	0	0	1	0
4865	27365	28.40	1248.0	74.00	5.0	4.75	8	0	0	0	...	0	0	0	1	0

3918 rows × 23 columns

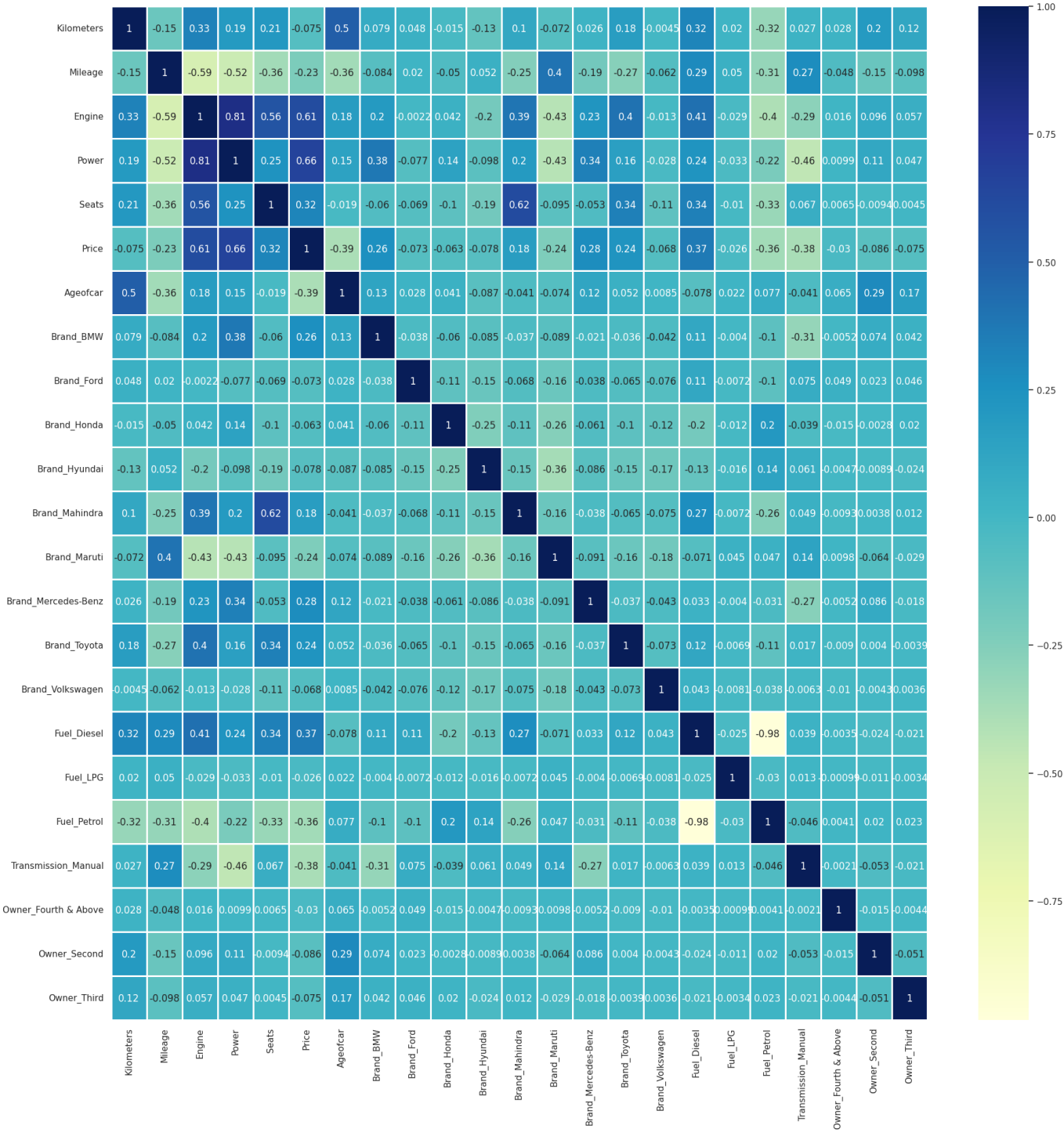
Hình 3.6: Dữ liệu sau khi đặt biến giả

- **Kiểm tra hiện tượng đa cộng tuyến:**

– Sử dụng ma trận tương quan kiểm tra hiện tượng đa cộng tuyến.

– Giữa 2 biến độc lập có hệ số tương quan $|r| > 0.75$ ta sẽ loại bỏ 1 trong 2 biến để tránh hiện tượng đa cộng tuyến. Biến chúng ta đang cần dự đoán là biến **Price** nên ta sẽ loại bỏ biến độc lập có hệ số tương quan với **Price** thấp hơn.

– Ma trận tương quan:



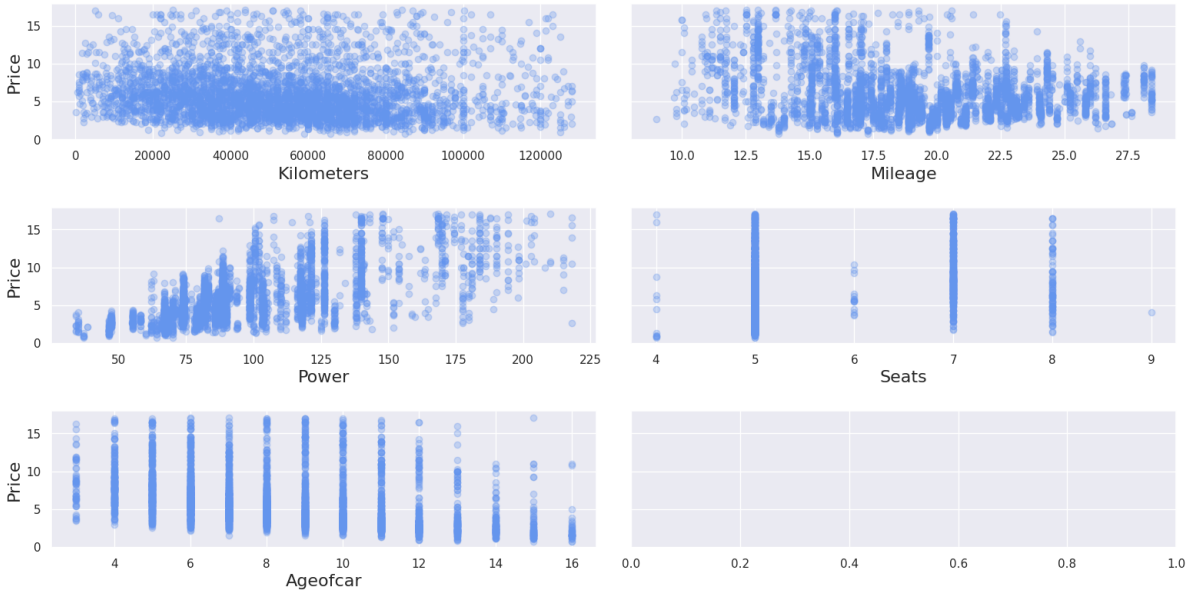
Hình 3.7: Ma trận tương quan

– Ta thấy có 2 cặp biến **Power** - **Engine** và **Fuel_Diesel** - **Fuel_Petrol** có hệ số tương quan $|r| > 0.75$. Ta cần loại bỏ 1 trong 2 biến để tránh hiện tượng đa cộng tuyến:

  * Loại bỏ biến **Engine** vì biến **Engine** có hệ số tương quan với biến dự đoán **Price** thấp hơn so với biến **Power**.
  * Tương tự loại biến **Fuel_Diesel** ra khỏi mô hình.

Bước 2: Xây dựng mô hình

- Xem xét sự phụ thuộc của biến dự báo **Price** với các biến qua đồ thị phân tán:

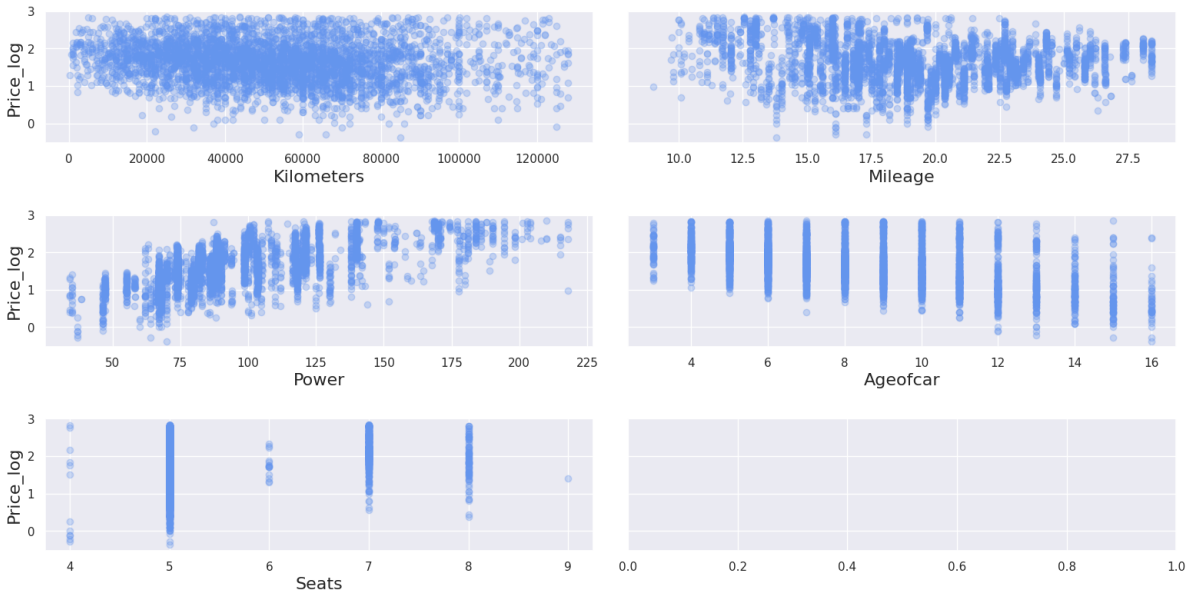


Hình 3.8: Sự phụ thuộc của biến **Price** với các biến.

=> Ta thấy dữ liệu trong các đồ thị có xu hướng tạo thành một đường cong nên để đảm bảo tính tuyến tính ta sẽ tính giá trị log(**Price**) để đưa các đồ thị về dạng đường thẳng.

– Kết quả:

- Xây dựng mô hình hồi quy tuyến tính bằng phương pháp bình phương cực tiểu.
  – Ta chia bộ dữ liệu thành 2 phần tập *train* và *test* một cách ngẫu nhiên, với tỷ lệ 80% cho tập *train* và 20% cho tập *test*.
  – Xây dựng mô hình bằng phương pháp bình phương cực tiểu-*OLS* có sẵn trong thư viện **statsmodels.api** với:
    * X: 'Kilometers', 'Mileage', 'Power', 'Seats','Ageofcar', 'Brand_BMW', 'Brand_Ford', 'Brand_Honda', 'Brand_Hyundai', 'Brand_Mahindra', 'Brand_Maruti', 'Brand_Mercedes-Benz', 'Brand_Toyota', 'Brand_Volkswagen',



Hình 3.9: Sự phụ thuộc biến log(**Price**) với các biến.

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state = 42)
x_train.reset_index()
print("x_train:",x_train.shape)
print("x_test:",x_test.shape)
print("y_train:",y_train.shape)
print("y_test:",y_test.shape)
```

```
x_train: (3134, 20)
x_test: (784, 20)
y_train: (3134,)
y_test: (784,)
```

'Fuel_LPG', 'Fuel_Petrol', 'Transmission_Manual', 'Owner_Fourth&Above', 'Owner_Second', 'Owner_Third'.

* Y: 'Price_log'.

```
import statsmodels.api as sm
result = sm.OLS(y_train, x_train_ac).fit()
print(result.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Price_log   R-squared:                       0.834
Model:                            OLS   Adj. R-squared:                  0.833
Method:                 Least Squares   F-statistic:                     782.0
Date:                Fri, 14 Jul 2023   Prob (F-statistic):              0.00
Time:                        04:49:59   Log-Likelihood:                209.18
No. Observations:                3134   AIC:                            -376.4
Df Residuals:                    3113   BIC:                            -249.3
Df Model:                          20
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                            2.3452      0.099     23.660      0.000       2.151       2.540
Kilometers                   -6.22e-07   2.05e-07     -3.035      0.002   -1.02e-06   -2.2e-07
Mileage                        -0.0154      0.002     -7.517      0.000      -0.019      -0.011
Power                           0.0086      0.000     39.411      0.000       0.008       0.009
Seats                           0.0768      0.010      8.035      0.000       0.058       0.096
Ageofcar                       -0.1190      0.002    -61.952      0.000      -0.123      -0.115
Brand_BMW                      -0.1816      0.050     -3.656      0.000      -0.279      -0.084
Brand_Ford                     -0.4719      0.046    -10.318      0.000      -0.562      -0.382
Brand_Honda                    -0.4306      0.044     -9.821      0.000      -0.517      -0.345
Brand_Hyundai                  -0.4065      0.044     -9.325      0.000      -0.492      -0.321
Brand_Mahindra                 -0.6029      0.048    -12.547      0.000      -0.697      -0.509
Brand_Maruti                   -0.4013      0.045     -8.961      0.000      -0.489      -0.314
Brand_Mercedes-Benz            -0.0365      0.049     -0.752      0.452      -0.132       0.059
Brand_Toyota                   -0.2434      0.047     -5.226      0.000      -0.335      -0.152
Brand_Volkswagen               -0.4302      0.045     -9.639      0.000      -0.518      -0.343
Fuel_LPG                        0.0036      0.161      0.022      0.982      -0.312       0.319
Fuel_Petrol                    -0.2523      0.014    -18.314      0.000      -0.279      -0.225
Transmission_Manual            -0.0897      0.013     -6.910      0.000      -0.115      -0.064
Owner_Fourth & Above           -0.0129      0.115     -0.113      0.910      -0.237       0.212
Owner_Second                   -0.0466      0.012     -3.793      0.000      -0.071      -0.023
Owner_Third                    -0.1170      0.034     -3.411      0.001      -0.184      -0.050
==============================================================================
Omnibus:                      209.356   Durbin-Watson:                   2.008
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              568.436
Skew:                          -0.362   Prob(JB):                    3.68e-124
Kurtosis:                       4.957   Cond. No.                     2.33e+06
==============================================================================
```

Hình 3.10: Kết quả chạy mô hình

=> Ta thấy giá trị $R^2 = 0.834$ và R^2 - hiệu chỉnh $= 0.833$ nên mô hình của chúng ta khá phù hợp với bộ dữ liệu.

* Các hệ số β tương ứng với các biến ước lượng được. Biến phụ thuộc là $log(\textbf{Price})$.

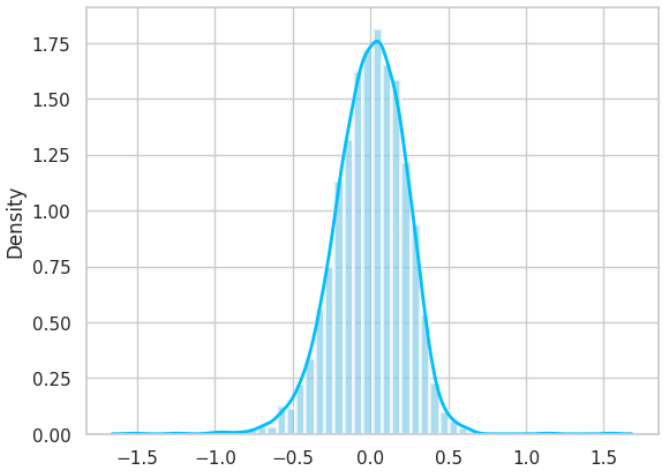
	Name	B
0	const	2.345217e+00
1	Kilometers	-6.220392e-07
2	Mileage	-1.539118e-02
3	Power	8.615465e-03
4	Seats	7.681353e-02
5	Ageofcar	-1.189824e-01
6	Brand_BMW	-1.815620e-01
7	Brand_Ford	-4.718577e-01
8	Brand_Honda	-4.305697e-01
9	Brand_Hyundai	-4.065416e-01
10	Brand_Mahindra	-6.028948e-01
11	Brand_Maruti	-4.013494e-01
12	Brand_Mercedes-Benz	-3.654838e-02
13	Brand_Toyota	-2.434237e-01
14	Brand_Volkswagen	-4.302008e-01
15	Fuel_LPG	3.586503e-03
16	Fuel_Petrol	-2.523302e-01
17	Transmission_Manual	-8.972660e-02
18	Owner_Fourth & Above	-1.293006e-02
19	Owner_Second	-4.663245e-02
20	Owner_Third	-1.170110e-01

Hình 3.11: Hệ số hồi quy ước lượng OLS

- Kiểm tra sai số của mô hình:
  - Từ mô hình với hệ số β vừa tính được và sử dụng bộ dữ liệu *train*, ta sẽ đi dự đoán giá của xe. Từ đó ta tính các phần dư theo công thức:

$$\text{e} = \text{y\_train} - \text{y\_train\_predict}$$

  - Trong đó:
    - y_train: Giá thực tế của bộ dữ liệu train.
    - y_train_predict: Giá dự đoán của bộ dữ liệu train.
  - Đồ thị phân phối của phần dư:



Hình 3.12: Phân phối của e

=> Ta thấy đồ thị phân phối của e có xấp xỉ với dạng đồ thị phân phối chuẩn.

– Giá trị kỳ vọng của phần dư:

```
residual= result.resid
np.mean(residual)
```

```
1.4706933820104392e-13
```

=> Thỏa mãn giả thiết kỳ vọng sai số = 0

– Kiểm định White kiểm tra phương sai sai số thay đổi:

```
white_test = het_white(result.resid, exog)
print('p-value:', white_test[1])
```

```
p-value: 0.8622196152463149
```

=> Ta có: p-value > 0.05 nên thỏa mãn giả thiết phương sai sai số thay đổi.

– Phương pháp Durbin-Watson kiểm tra hiện tượng tự tương quan:

```
dw_test_results = sm.stats.stattools.durbin_watson(result.resid)
print("DW = ", dw_test_results)
```

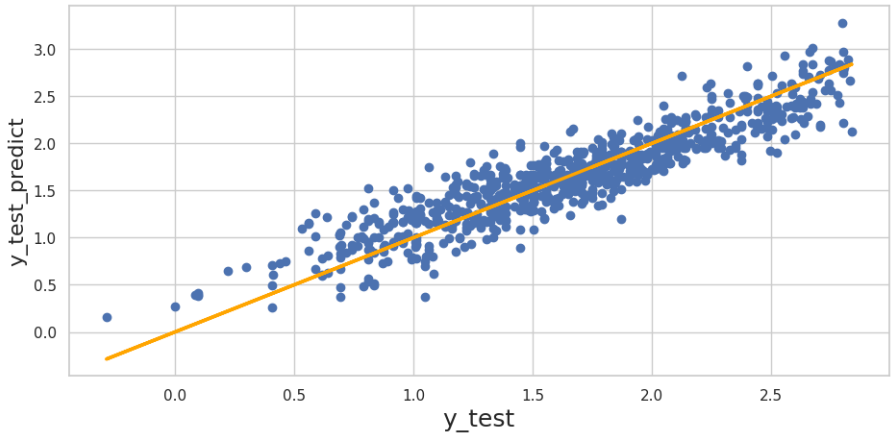
```
DW =  2.008106891649495
```

=> Giá trị DW ≈ 2, suy ra mô hình không có sự tự tương quan giữa các phần dư.

Như vậy, mô hình thỏa mãn các giả định cần thiết của phương pháp ước lượng bình phương cực tiểu.

Bước 3: Đánh giá mô hình

- Ta sử dụng bộ dữ liệu *test* để đánh giá mô hình.
- Từ mô hình với các giá trị β vừa tính được ta sẽ tính giá trị dự báo của bộ dữ liệu *test* và biểu diễn trên đồ thị:



- Trong đó:
  - Trục tung là giá trị dự đoán sử dụng mô hình đã ước lượng trên bộ dữ liệu *test*.
  - Trục hoành là giá trị thực tế.
  - Đường màu cam là đường $y\_test = y\_test\_predict$.
- Nhận xét:
  - Ta thấy mô hình của chúng ta dự đoán khá đúng với các xe có giá ở mức trung bình do các điểm tập trung sát và dày đặc quanh đường thẳng $y\_test = y\_test\_predict$.
  - Đối với các xe có giá thấp và cao, mô hình có vẻ dự đoán không tốt do các điểm có xu hướng phân tán qua đường $y\_test = y\_test\_predict$..
- Tiến hành xem xét cụ thể qua sai số và tỷ lệ phần trăm sai lệch:

```
df_predict['Thực tế'] = np.exp(y_test)
df_predict = df_predict.reset_index(drop= True)
df_predict['Phần dư'] = df_predict['Thực tế'] - df_predict['Dự báo']
df_predict['Chênh lệch %'] = (df_predict['Thực tế'] / df_predict['Dự báo']-1) * 100
df_predict.sort_values(by = ['Chênh lệch %'])
```

	Dự báo	Thực tế	Phần dư	Chênh lệch %
349	4.595196	2.25	-2.345196	-51.035825
577	5.731007	2.90	-2.831007	-49.398071
350	3.538009	1.80	-1.738009	-49.123934
529	3.182075	1.75	-1.432075	-45.004439
677	15.209426	8.38	-6.829426	-44.902589
...	...	...	...	...
19	9.198018	16.50	7.301982	79.386471
47	6.712419	12.50	5.787581	86.221980
184	1.457210	2.85	1.392790	95.579165
330	3.307130	6.50	3.192870	96.545014
467	8.354215	17.09	8.735785	104.567397

784 rows × 4 columns

- Nhận xét:
  - Sự chênh lệch lớn nhất là 104,57%, mô hình dự báo giá thấp hơn hơn khoảng 2 lần so với thực tế.
  - Sự chênh lệch nhỏ nhất là -51,035%, mô hình dự báo giá cao hơn khoảng 1,5 lần so với thực tế.

  Sự sai lệch này xuất hiện do một số yếu tố thực tế mà chưa đưa vào mô hình như: thị trường xe cũ, yếu tố cung cầu khi nghiên cứu, giá xe mới, chi phí thuế và một số chi phí khác,...

Kết luận:

- Mô hình dự báo giá xe cũ bằng phương pháp bình phương cực tiểu với hệ số xác định $R^2 = 0.834$, khá cao và thỏa mãn các giả thiết mà phương pháp đặt ra .Từ đó cho thấy đây là một mô hình khá phù hợp để dự báo giá xe ô tô cũ.
- Mô hình có thể đưa thêm một số yếu tố để dự đoán hoặc tăng kích thước mẫu điều tra để cải thiện độ chính xác cả mô hình.

Kết luận

Với những nội dung trên, từ đó thấy được phân tích hồi quy tuyến tính là một phương pháp cơ bản nhưng đã hiệu quả để dự đoán và giải thích sự biến động của biến phụ thuộc dựa trên các biến độc lập. Báo cáo đã trình bày cách xây dựng và ước lượng mô hình hồi quy tuyến tính bằng phương pháp bình phương cực tiểu, cũng như cách kiểm tra sự phù hợp của mô hình. Thực hiện kiểm tra các giả thiết liên quan như kiểm tra hiện tượng đa cộng tuyến, phương sai sai số thay đổi và hiện tượng tự tương quan. Từ đó áp dụng vào việc xây dựng mô hình hồi quy tuyến tính dự báo giá xe cũ, vận dụng được kiến thức cũng như kỹ năng sử dụng công cụ lập trình trong việc phân tích, xây dựng mô hình dự báo.

Kết quả của đồ án cho thấy rằng mô hình hồi quy tuyến tính có thể được sử dụng để dự đoán giá trị của một biến phụ thuộc dựa trên một hoặc nhiều biến độc lập. Tuy nhiên, cần chú ý đến các giả định của mô hình và thực hiện các kỹ thuật kiểm tra để đảm bảo tính phù hợp của mô hình.

Hướng phát triển đề tài:

- Kiểm định thêm về các giả thuyết hệ số hồi quy có ràng buộc giữa các hệ số. Suy diễn thống kế từ mô hình đưa ra các dự báo trong các lĩnh vực cụ thể.
- Tìm hiểu thêm nhiều kỹ thuật để kiểm định, lựa chọn, xây dựng mô hình phù hợp với mức ý nghĩa cao.
- Vận dụng để xây dựng mô hình dự báo với số liệu chuỗi thời gian.

Tài liệu tham khảo

[1] Linear Regression Analysis: Theory and Computing, Yan, Xin; Su, Xiao Gang. -(10).

[2] Applied multivariate statistical analysis. Johnson R. A., & Wichern D. W. (1992), -(361).

[3] Applied multivariate statistical analysis. Johnson R. A., & Wichern D. W, (1992). -(370 - 371).

[4] Introduction to Linear Regression Analysis, 5th ed - Walter A. Shewhart and Samuel S. Wilkswilks-(85 - 86).

[5] Applied Econometrics. Asteriou, Dimitrios, Hall, Stephen G - (123).

[6] Applied Econometrics. Asteriou, Dimitrios, Hall, Stephen G - (140 -142).