ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ ỨNG DỤNG

XÁC SUẤT THỐNG KÊ (MT2013)

Bài tập lớn

Đề tài 1

Giảng viên hướng dẫn: Nguyễn Kiều Dung

STT	Họ và tên	MSSV	Lớp	Khoa/Ngành học	Công việc
1	Trần Nguyễn Thái Bình	2110051	L03	Khoa học và Kỹ thuật máy tính	Hoạt động 1
2	Phạm Hữu Huy	2111347	L08	Khoa học và Kỹ thuật máy tính	Cơ sở lý thuyết
3	Đỗ Nguyễn An Huy	2110193	L07	Khoa học và Kỹ thuật máy tính	Hoạt động 2

THÀNH PHỐ HỒ CHÍ MINH, 2022

Mục lục

1 Cơ sở lý thuyết

1.1 Định nghĩa thống kê và phân tích hồi quy

1.1.1 Khái niệm thống kê

Thống kê là bộ môn toán học nghiên cứu quy luật của các hiện tượng ngẫu nhiên có tính chất số lớn trên cơ sở thu thập và xử lý các số liệu thống kê (các kết quả quan sát). Nội dung chủ yếu của thống kê toán là xây dựng các phương pháp thu thập và xử lý các số liệu thống kê, nhằm rút ra các kết luận khoa học từ thực tiễn, dựa trên những thành tựu của lý thuyết xác suất.

Việc thu thập, sắp xếp, trình bày các số liệu của tổng thể hay của một mẫu được gọi là thống kê mô tả. Còn việc sử dụng các thông tin của mẫu để tiến hành các suy đoán, kết luận về tổng thể gọi là thống kê suy diễn.

1.1.2 Khái niệm phân tích hồi quy

Bài toán phân tích hồi quy là bài toán nghiên cứu mối liên hệ phụ thuộc của một biến (gọi là biến phụ thuộc) vào một hay nhiều biến khác (gọi là các biến độc lập), với ý tưởng ước lượng được giá trị trung bình (tổng thể) của biến phụ thuộc theo giá trị của các biến độc lập, dựa trên mẫu được biết trước.

1.2 Cơ sở lý thuyết mô hình hồi quy tuyến tính bội

Trong đời sống, kỹ thuật và đặc biệt là các ngành kinh tế, việc một yếu tố phụ thuộc vào nhiều yếu tố khác diễn ra khá thường xuyên. Để mô hình hóa các bài toán như thế, ta cần một mô hình có thể có nhiều biến độc lập, và hồi quy tuyến tính bội là một trong những mô hình đơn giản và nền tảng nhất có thể đáp ứng được yêu cầu đó.

Phương pháp hồi quy bội còn gọi là phương pháp hồi quy đa biến, dùng để phân tích mối quan hệ giữa nhiều biến số độc lập (tức là biến giải thích hay biến nguyên nhân) ảnh hưởng đến một biến phụ thuộc (tức là biến phân tích hay biến kết quả). Mô hình hồi quy bội dùng cho dự báo sử dụng nhiều hơn một biến độc lập.

1.2.1 Phương trình hồi quy tuyến tính bội

Tổng quan, với y là biến số phụ thuộc tuyến tính với k biến độc lập $x_1, x_2, x_3, \cdots, x_k$ (biến hồi quy). Khi đó mô hình hồi quy tuyến tính bội với k biến hồi quy có dạng như sau:

$$Y \quad = \quad \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + \epsilon$$

Trong đó:

- Tham số β_0 được gọi là hệ số tung độ gốc (hay còn gọi là hệ số chặn). Hệ số trên bằng giá trị trung bình của biến phụ thuộc Y khi các biến độc lập trong mô hình nhận giá trị bằng 0: $x_1 = x_2 = x_3 = \cdots = x_k = 0$. Trong thực tế, hệ số này ít được quan tâm.
- Các tham số $\beta_1, \beta_2, \cdots, \beta_k$ được gọi là các hệ số hồi quy riêng (hệ số độ dốc). β_k thể hiện sự thay đổi của Y theo mỗi đơn vị của x_k khi các biến còn lại được giữ nguyên.
- ϵ là thành phần ngẫu nhiên hay yếu tố nhiễu. Thực chất, mô hình này thường chỉ dự đoán tốt kỳ vọng của Y, chứ không phải giá trị thực tế của Y.

Với một tổng thể có k biến độc lập và n là số lần quan sát. Tổng thể trên được biểu diễn bằng hệ phương trình sau:

$$\begin{cases} Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13}. + \cdots + \beta_k X_{1k} + \epsilon_1 \\ Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \beta_3 X_{23} + \cdots + \beta_k X_{2k} + \epsilon_2 \\ \cdots \\ Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \beta_3 X_{n3} + \cdots + \beta_k X_{nk} + \epsilon_n \end{cases}$$

Với $\mathbf{Y}^T = \begin{pmatrix} Y_1 & Y_2 & \cdots & Y_n \end{pmatrix}$, $\boldsymbol{\epsilon}^T = \begin{pmatrix} \epsilon_1 & \epsilon_2 & \cdots & \epsilon_n \end{pmatrix}$ và $\boldsymbol{\beta}^T = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_n \end{pmatrix}$. Hệ phương trình có thể viết dưới dạng phương trình ma trận:

$$Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

1.2.2 Các giả thiết để xây dựng mô hình hồi quy tuyến tính bội

Ta đưa ra các giả thiết cơ bản cho mô hình hồi quy bội với n là số lần quan sát như sau:

(i) Việc ước lượng được dựa trên cơ sở mẫu ngẫu nhiên.

(ii) $\epsilon \sim N(0, \sigma^2)$ và độc lập với X_i, $i = \overline{1,k}$.

(iii) Giữa các biến độc lập X_j không có quan hệ đa cộng tuyến hoàn hảo, nghĩa là không tồn tại hằng số không đồng thời bằng 0 sao cho: $\lambda_1 X_1 + \lambda_2 X_2 + \cdots, +\lambda_2 X_k = 0$. Có thể nhận thấy nếu giữa các biến X_j với $j = \overline{1,n}$ có quan hệ cộng tuyến hoàn hảo thì có ít nhất một trong các biến này sẽ suy ra được từ các biến còn lại. Do đó, giả thiết (iii) được đưa ra để loại trừ tình huống này.

1.2.3 Ước lượng các tham số bằng phương pháp bình phương nhỏ nhất OLS

Có nhiều cách để xác định giá trị của các tham số, tuy nhiên trong đó phương pháp bình phương cực tiểu thường được sử dụng nhất. Phương pháp bình phương cực tiểu thường được sử dụng để ước tính hệ số hồi quy của mô hình hồi quy tuyến tính bội.

Giả sử, có n quan sát và k biến hồi quy thỏa $n > k$. Vậy tổng bình phương các sai số là:
$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1s}^{k} \beta_j X_{ij})^2$

Chúng ta đang muốn giảm giá trị L này theo các tham số β về giá trị nhỏ nhất.
Lấy đạo hàm riêng phần của biểu thức này theo mỗi biến trong k biến chưa biết. Để giá trị L đạt nhỏ nhất thì các tham số phải thỏa mãn hệ.

$$\begin{aligned}
\frac{\partial f(\hat{B}_0, \hat{B}_1, \cdots, \hat{B}_n)}{\partial(\hat{B}_0)} &= -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \beta_3 X_{i3} \cdots - \beta_k X_{ik}) &= 0 \\
\frac{\partial f(\hat{B}_0, \hat{B}_1, \cdots, \hat{B}_n)}{\partial(\hat{B}_1)} &= -2\sum_{i=1}^{n} X_{i1}(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \beta_3 X_{i3} \cdots - \beta_k X_{ik}) &= 0 \\
&\vdots \\
\frac{\partial f(\hat{B}_0, \hat{B}_1, \cdots, \hat{B}_n)}{\partial(\hat{B}_k)} &= -2\sum_{i=1}^{n} X_{ik}(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \beta_3 X_{i3} \cdots - \beta_k X_{ik}) &= 0
\end{aligned}$$

Đơn giản hệ phương trình trên:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + ... + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + ... + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i$$

.....

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + ... + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik}y_i$$

Với X^T là ma trận chuyển vị của X. Ta có

$$\boldsymbol{X^T X} = \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{ik}x_{i1} & \sum_{i=1}^n x_{ik}x_{i2} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}$$

Vậy hệ phương trình trên có thể biểu diễn dưới dạng ma trận: $X^T X\hat{\beta} = X^T Y$
$\implies \hat{\beta} = (X^T X)^{-1} X^T Y$
Biểu thức này được gọi là phương trình cơ bản của phương pháp OLS.

1.2.4 Ước lượng độ lệch chuẩn của sai số ngẫu nhiên

Sau khi đã có được các hàm ước lượng OLS của các hệ số hồi quy riêng phần, chúng ta có thể tính được độ lệch chuẩn của sai số ngẫu nhiên. Chúng ta cần có những sai số chuẩn vì hai mục đích chính: để thiết lập khoảng tin cậy và kiểm định các giả thiết thống kê. Với σ^2 là phương sai (phương sai có điều kiện không đổi) của yếu tố nhiễu tổng thể ϵ. Bằng phương pháp giải tích, có thể chứng minh rằng một hàm ước lượng không thiên lệch của σ^2.

$$\hat{\sigma}^2 = \frac{SSE}{n-k-1} = \frac{\sum \hat{\epsilon}_i^2}{n-k-1} => \hat{\sigma} = \sqrt{\frac{SSE}{n-k-1}}$$

Lưu ý rằng sự tương tự giữa hàm ước lượng σ^2 này và hàm ước lượng hai biến tương ứng với nó. Các bậc tự do bây giờ là $(n-k-1)$ bởi vì khi ước lượng trước hết chúng ta cần ước lượng $\beta_0, \beta_1, \cdots, \beta_k$ đã sử dụng $k+1$ bậc tự do.

1.2.5 Ước lượng phương sai của tham số bằng ma trận hiệp phương sai

Để kiểm định giả thuyết, tìm khoảng tin cậy, cũng như thực hiện các suy luận thống kê khác ta cần phải tìm $\text{Var}(\beta_j), j = \overline{1,k}$ và $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j)$. Phương pháp ma trận giúp ta có thể thực hiện điều này. Thuộc tính phương sai của $\hat{\beta}$ được biểu diễn trong ma trận hiệp phương sai sau:

$$\boldsymbol{Cov(\hat{\beta})} = \begin{pmatrix} Var(\hat{\beta}_1) & Cov(\hat{\beta}_i, \hat{\beta}_j) & \cdots & Cov(\hat{\beta}_i, \hat{\beta}_j) \\ Cov(\hat{\beta}_i, \hat{\beta}_j) & Var(\hat{\beta}_2) & \cdots & Cov(\hat{\beta}_i, \hat{\beta}_j) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\hat{\beta}_i, \hat{\beta}_j) & Cov(\hat{\beta}_i, \hat{\beta}_j) & \cdots & Var(\hat{\beta}_k) \end{pmatrix}$$

$$
\begin{aligned}
Cov(\hat{\beta}) &= E\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T\right] \\
&= E\left\{\left[(X^TX)^{-1}X^T\varepsilon\right]\left[(X^TX)^{-1}X^T\varepsilon^T\right]\right\} \\
&= E\left[(X^TX)^{-1}X^T\varepsilon\varepsilon^TX(XX^T)^{-1}\right] \\
&= (X^TX)^{-1}E(\varepsilon\varepsilon^T)X(XX^T)^{-1} \\
&= \sigma^2(X^TX)^{-1}
\end{aligned}
$$

Trong đó $Cov(\hat{\beta})$ được xác định bởi công thức dưới đây:

Từ $Cov(\hat{\beta})$ ta có thể xác định bất kỳ $Var(\hat{\beta}_1), Var(\hat{\beta}_2)$ hoặc hiệp phương sai cần thiết khác.

1.2.6 Kiểm định giả thuyết thống kê trong mô hình hồi quy tuyến tính bội

1.2.6.a Kiểm định ý nghĩa thống kê của các hệ số hồi qui

Mô hình được gọi là không có hiệu lực giải thích, hay nói cách khác không giải thích được sự thay đổi của biến Y, nếu toàn bộ các hệ số hồi quy riêng đều bằng 0. Vì vậy để kiểm định mức ý nghĩa của mô hình hay xác định xem có mối quan hệ tuyến tính tồn tại giữa Y và các biến hồi quy $X_i i = \overline{1,k}$ ta cần kiểm định bài toán sau: $\begin{cases} \mathrm{H}_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0 \\ \mathrm{H}_1 : \exists \beta_j \neq 0,\ j = \overline{1,n} \end{cases}$

Bác bỏ H_0 đồng nghĩa với việc ta chấp nhận có ít nhất một trong các biến hồi quy $X_1, X_2, \cdots, X_k$ có ảnh hưởng đến mô hình. Tổng bình phương SST được chia thành hai phần, gồm tổng bình phương do mô hình và tổng bình phương do chênh lệch.

$$SST = SSR + SSE$$

Nếu H_0 đúng, SSR/ϵ^2 là một biến ngẫu nhiên tuân theo phân phối chi bình phương (Chi-Square) với k bậc tự do, bằng với số lượng biến hồi quy trong mô hình. Chúng ta cũng có thể chỉ ra rằng SSE/ϵ^2 là một biến ngẫu nhiên tuân theo phân phối chi bình phương với n-p bậc tự do, và SSE và SSR là độc lập. Kiểm định thống kê cho H_0 là:

$$F_0 = \frac{SSR/k}{SSE/(n-p)} = \frac{MSR}{MSE} \quad (3)$$

Ta bác bỏ H_0 nếu giá trị kiểm định trong phương trình (3) lớn hơn $f_{(a,k,n-p)}$ có được từ việc tra bảng Fisher, ngược lại chấp nhận H_0.

1.2.6.b Kiểm định giả thuyết của từng biến độc lập

Nếu chúng ta kết luận được mô hình toàn diện có ý nghĩa. Điều này có nghĩa là có ít nhất một biến độc lập trong mô hình có thể giải thích được một cách có ý nghĩa cho biến thiên trong biến phụ thuộc. Tuy nhiên điều này không có nghĩa là tất cả các biến độc lập đưa vào mô hình đều có nghĩa. Chúng ta có thể kiểm định hệ số β_j, với $j = \overline{1,k}$ bằng phương pháp thông thường:
$\begin{cases} \mathrm{H}_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$

Tiêu chuẩn kiểm định $t_i = \dfrac{\hat{B}_i}{\hat{\sigma}_{\hat{B}_i}}$.
Nếu $|t_i| < t(\alpha/2, n-p)$ chấp nhận H_0 ngược lại bác bỏ H_0.

1.2.6.c Hệ số xác định hiệu chỉnh

Ta có thể đánh giá hàm hồi quy mẫu phù hợp với số liệu mẫu đến mức nào thông qua hệ số xác định bội R^2. Để tính toán:

$$R^2 = \frac{SSR}{SST}$$

Giá trị R^2 gần 1 cho thấy mô hình là tốt, có khả năng cao phù hợp với dữ liệu đã đưa vào, trong khi R^2 gần 0 chỉ ra rằng mô hình đang sử dụng không thật sự phù hợp để mô tả dữ liệu đầu vào.

Tuy nhiên một tính chất quan trọng của R^2 là nó sẽ tăng khi ta đưa thêm biến độc lập vào mô hình, việc đưa thêm một biến số bất kỳ vào mô hình nói chung sẽ làm gia tăng R^2, không kể nó có giúp giải thích thêm cho biến phụ thuộc hay không. Điều này ngụ ý rằng R^2 chưa phải là thước đo tốt khi muốn so sánh các mô hình với số biến khác nhau nên giá trị R^2 hiệu chỉnh thường được sử dụng trong thực tế hơn, do nó chỉ tăng thật sự khi số lượng dự đoán được cải thiện. Giá trị này được tính như sau:

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

1.2.6.d Kiểm định phân phối chuẩn

Kiểm định phân phối chuẩn là bước quan trọng trong thủ tục thống kê suy luận, giúp chúng ta xác định được cơ bản hình dạng chung của một phân phối, từ đó đánh giá kiểm định có bị lệch hay không, và có độ lệch dương hay âm. Các phương pháp thường sử dụng như sau:

1. Kiểm định phân phối chuẩn bằng biểu đồ.
2. Sử dụng kiểm định Shapiro-Wilk

   Phát biểu giả thuyết thống kê trong kiểm định Shapiro-Wilk:

   $$\begin{cases} \text{H}_0\text{: Biến cần kiểm định tuân theo phân phối chuẩn.} \\ \text{H}_1\text{: Biến cần kiểm định không tuân theo phân phối chuẩn} \end{cases}$$

   Giá trị của thống kê Shapiro-Wilk: $W = \dfrac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum (x_i - \overline{x})^2}$

   Với:

   - x_i: giá trị thứ i nhỏ nhất của x.
   - a_i: hằng số Shapiro-Wilk.

   Sử dụng thống kê Shapiro-Wilk để xác định giá trị p-value.

   Nếu giá trị p-value là nhỏ (so với một mốc đã quy ước), ta có thể bác bỏ giả thuyết H_0 hay nói cách khác biến cần kiểm định không tuân theo luật phân phối chuẩn.

2 Hoạt động 1

2.1 Mô tả tập dữ liệu

Tập tin **"gia_nha.csv"** chứa thông tin về giá bán ra thị trường (đơn vị đô la) của 21613 ngôi nhà ở quận King nước Mỹ trong khoảng thời gian từ tháng 5/2014 đến 5/2015. Bên cạnh giá nhà, dữ liệu còn bao gồm các thuộc tính mô tả chất lượng ngôi nhà. Dữ liệu gốc được cung cấp tại:. *https://www.kaggle.com/harlfoxem/housesalesprediction*.
Các biến chính trong bộ dữ liệu:

- **price:** Giá nhà được bán ra
- **floors:** Số tầng của ngôi nhà được phân loại từ 1-3.5.
- **condition:** Điều kiện kiến trúc của ngôi nhà từ 1 - 5, 1: rất tệ và 5: rất tốt
- **view:** Đánh giá quan cảnh xung quanh nhà theo mức độ từ thấp đến cao: 0 - 4.
- **sqft_above:** Diện tích ngôi nhà.
- **sqft_living:** Diện tích khuôn viên nhà.
- **sqft_basement:** Diện tích tầng hầm.

2.2 Đọc dữ liệu (Import Data)

Để đọc tập tin, dùng lệnh **read_csv** như sau:

```
library(readr)
setwd("E:/CTRR")
gia_nha <- read_csv("gia_nha.csv")
```

Dữ liệu nhận được như sau:

	X1	X.1	X	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
1	1	1	1	7129300520	2014-10-13	221900	3	1.00	1180	5650	1.0	0	0	3	7	1180	0	1955	0	98178	47.5112	-122.257	1340	5650
2	2	2	2	6414100192	2014-12-09	538000	3	2.25	2570	7242	2.0	0	0	3	7	2170	400	1951	1991	98125	47.7210	-122.319	1690	7639
3	3	3	3	5631500400	2015-02-25	180000	2	1.00	770	10000	1.0	0	0	3	6	770	0	1933	0	98028	47.7379	-122.233	2720	8062
4	4	4	4	2487200875	2014-12-09	604000	4	3.00	1960	5000	1.0	0	0	5	7	1050	910	1965	0	98136	47.5208	-122.393	1360	5000
5	5	5	5	1954400510	2015-02-18	510000	3	2.00	1680	8080	1.0	0	0	3	8	1680	0	1987	0	98074	47.6168	-122.045	1800	7503
6	6	6	6	7237550310	2014-05-12	1225000	4	4.50	5420	101930	1.0	0	0	3	11	3890	1530	2001	0	98053	47.6561	-122.005	4760	101930
7	7	7	7	1321400060	2014-06-27	257500	3	2.25	1715	6819	2.0	0	0	3	7	1715	0	1995	0	98003	47.3097	-122.327	2238	6819
8	8	8	8	2008000270	2015-01-15	291850	3	1.50	1060	9711	1.0	0	0	3	7	1060	0	1963	0	98198	47.4095	-122.315	1650	9711
9	9	9	9	2414600126	2015-04-15	229500	3	1.00	1780	7470	1.0	0	0	3	7	1050	730	1960	0	98146	47.5123	-122.337	1780	8113
10	10	10	10	3793500160	2015-03-12	323000	3	2.50	1890	6560	2.0	0	0	3	7	1890	0	2003	0	98038	47.3684	-122.031	2390	7570
11	11	11	11	1736800520	2015-04-03	662500	3	2.50	3560	9796	1.0	0	0	3	8	1860	1700	1965	0	98007	47.6007	-122.145	2210	8925
12	12	12	12	9212900260	2014-05-27	468000	2	1.00	1160	6000	1.0	0	0	4	7	860	300	1942	0	98115	47.6900	-122.292	1330	6000
13	13	13	13	114101516	2014-05-28	310000	3	1.00	1430	19901	1.5	0	0	4	7	1430	0	1927	0	98028	47.7558	-122.229	1780	12697
14	14	14	14	6054650070	2014-10-07	400000	3	1.75	1370	9680	1.0	0	0	4	7	1370	0	1977	0	98074	47.6127	-122.045	1370	10208
15	15	15	15	1175000570	2015-03-12	530000	5	2.00	1810	4850	1.5	0	0	3	7	1810	0	1900	0	98107	47.6700	-122.394	1360	4850
16	16	16	16	9297300055	2015-01-24	650000	4	3.00	2950	5000	2.0	0	3	3	9	1980	970	1979	0	98126	47.5714	-122.375	2140	4000
17	17	17	17	1875500060	2014-07-31	395000	3	2.00	1890	14040	2.0	0	0	3	7	1890	0	1994	0	98019	47.7277	-121.962	1890	14018
18	18	18	18	6865200140	2014-05-29	485000	4	1.00	1600	4300	1.5	0	0	4	7	1600	0	1916	0	98103	47.6648	-122.343	1610	4300

Hình 1: Dữ liệu nhập vào

2.3 Làm sạch dữ liệu (Data Cleaning): NA (dữ liệu khuyết)

Sử dụng hàm **subset** để loại các cột không phải là biến chính của bộ dữ liệu và lưu dataframe mới đó vào **new_df**. Sau đó, sử dụng hàm **is_na()** để kiểm tra tập dữ liệu có bị khuyết hay không:

```
new_df<-subset(gia_nha, select = c(price, floors, condition, view, sqft_above,
    sqft_living, sqft_basement))
colSums(is.na(new_df))
```

```
> #2/
> #Clean Data
> new_df<-subset(gia_nha, select = c(price, floors, condition, view, sqft_above, sqft_living, sqft_basement))
> colSums(is.na(new_df))
       price        floors     condition          view    sqft_above   sqft_living sqft_basement
          20             0             0             0             0             0             0
>
```

Hình 2: Kiểm tra dữ liệu khuyết

Sau khi chạy lệnh trên ta thấy cột **price** có 20 hàng có dữ liệu bị khuyết, vì vậy ta phải xử lý những dữ liệu khuyết đó. Chạy lệnh `apply(new_df, 2, function(x)sum(is.na(x))/length(x))` để kiểm tra tỷ lệ dữ liệu khuyết chiếm trên toàn tập dữ liệu thì ta được:

```
       price        floors     condition          view    sqft_above   sqft_living sqft_basement
0.000925369   0.000000000   0.000000000   0.000000000   0.000000000   0.000000000   0.000000000
```

Hình 3: Tỷ lệ dữ liệu khuyết từng cột

Nhận thấy tỷ lệ dữ liệu bị khuyết chiếm không đáng kể nên ta có thể loại bỏ các hàng có dữ liệu khuyết ra khỏi tập dữ liệu mà không sợ sai lệch quá lớn xảy ra thông qua hàm **na.omit()**:

```
new_df<-na.omit(new_df)
```

2.4 Làm rõ dữ liệu (Data Visualization)

2.4.1 Chuyển đổi biến

Lấy logarithm các cột **price, floors, sqft_above, sqft_living** để dễ dàng quan sát khi phân tích và trực quan hóa. Lưu ý, vì sqft_basement (diện tích tầng hầm) có những trường hợp bằng 0, nên ta cộng thêm 1 trước khi lấy logarithm nhằm tránh việc giá trị tiến về vô cùng:

```
new_df$sqft_basement<- new_df$sqft_basement + 1
new_df$price<-log(new_df$price)
new_df$sqft_above<-log(new_df$sqft_above)
new_df$sqft_living<-log(new_df$sqft_living)
new_df$sqft_basement<-log(new_df$sqft_basement)
```

2.4.2 Thống kê đơn biến

Trong các biến chính, **floors, condition, view** là các biến rời rạc phân loại. Ta sẽ lập bảng thống kê số lượng cho từng biến:

```
table_floors<-as.data.frame(table(new_df$floors))
table_condition<-as.data.frame(table(new_df$condition))
table_view<-as.data.frame(table(new_df$view))
```

	Var1	Freq
1	1	10672
2	1.5	1909
3	2	8230
4	2.5	161
5	3	613
6	3.5	8

	Var1	Freq
1	1	30
2	2	172
3	3	14016
4	4	5677
5	5	1698

	Var1	Freq
1	0	19472
2	1	331
3	2	962
4	3	509
5	4	319

Hình 4: Bảng thống kê theo thứ tự (từ trái sang phải) lần lượt của biến floors, condition, view

Với các biến còn lại là biến liên tục, ta sẽ tính giá trị thống kê mô tả bao gồm: trung bình, trung vị, độ lệch chuẩn, giá trị lớn nhất, giá trị nhỏ nhất.

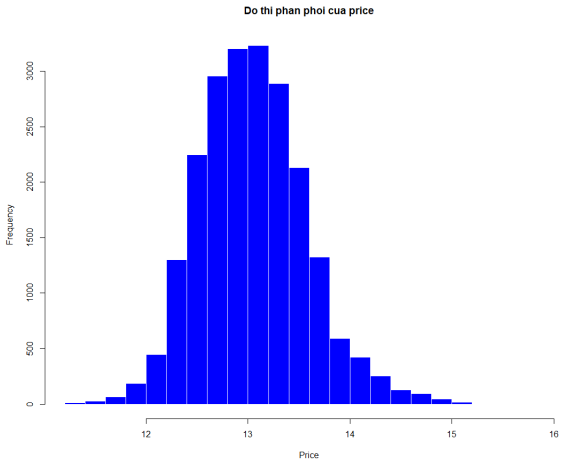
```
stat_table<-apply(new_df[,c("price", "sqft_above", "sqft_living", "sqft_basement")], 2,
    function(x){c(mean(x), median(x), sd(x), min(x), max(x))})
rownames(stat_table)<-c("mean", "median", "sd", "min", "max")
```

	price	sqft_above	sqft_living	sqft_basement
mean	13.047841	7.3948826	7.5503286	2.529186
median	13.017003	7.3524411	7.5548585	0.000000
sd	0.526574	0.4276433	0.4247722	3.170528
min	11.225243	5.6698809	5.6698809	0.000000
max	15.856731	9.1495282	9.5134035	8.480737

Hình 5: Bảng thống kê theo thứ tự (từ trái sang phải) lần lượt của biến floors, condition, view

Vì hoạt động sẽ đánh giá về giá nhà, nên ta sẽ dùng hàm **hist()** vẽ đồ thị phân phối của biến price để quan sát. Hàm **hist()** nhận giá trị đầu vào là cột *price* của bảng *new_ df*, main là tên của đồ thị, xlab là tên trục hoành và breaks là khoảng chia của mỗi khoảng:

```
hist(new_df$price, main="Do thi phan phoi cua price", col = "blue", border = "white",
    xlab = "Price", breaks = 20)
```



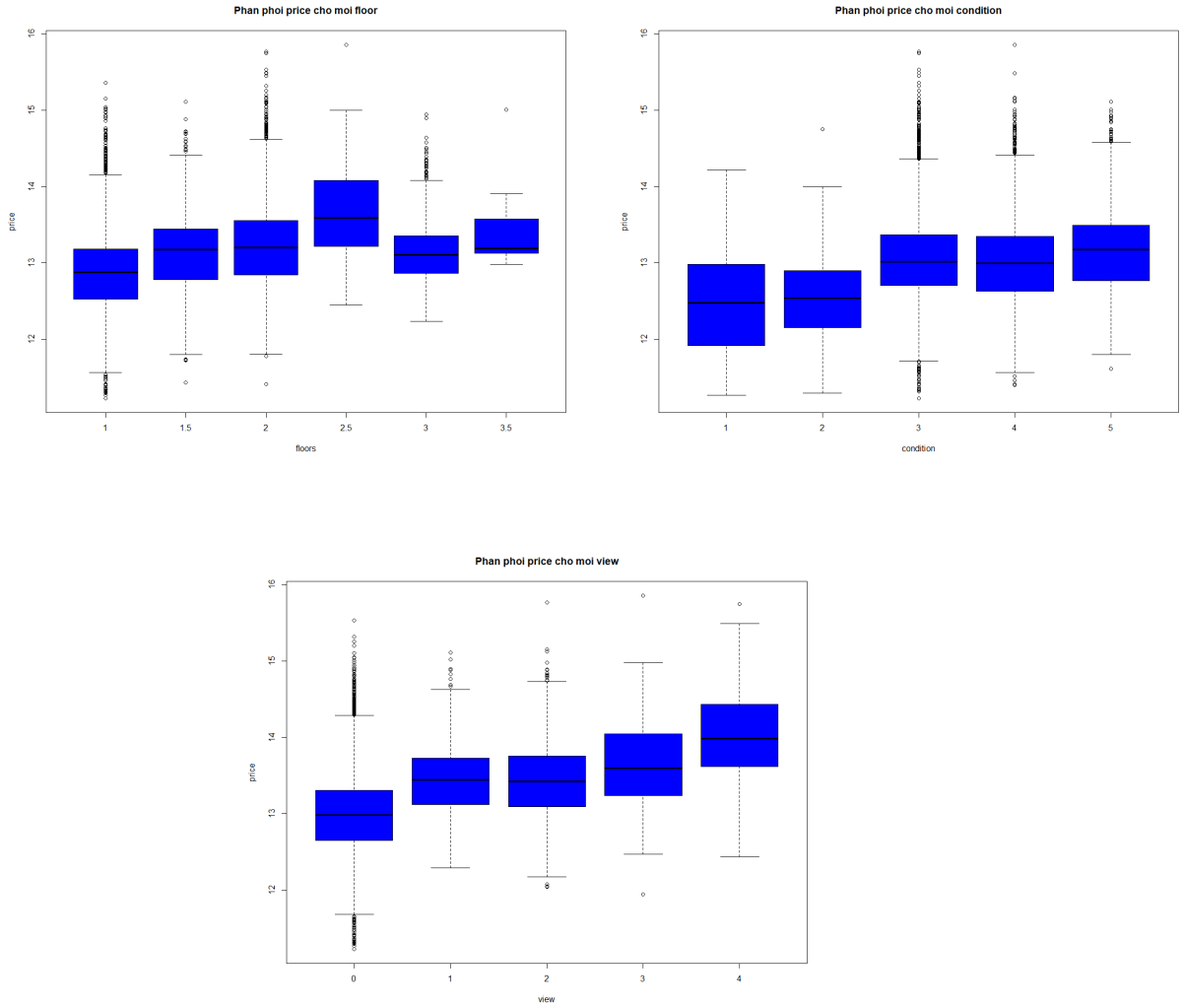
Hình 6: Đồ thị phân phối của price

Nhận xét: Qua đồ thị hình 6, ta có thể thấy rằng dữ liệu của *price* tuân theo phân phối chuẩn.

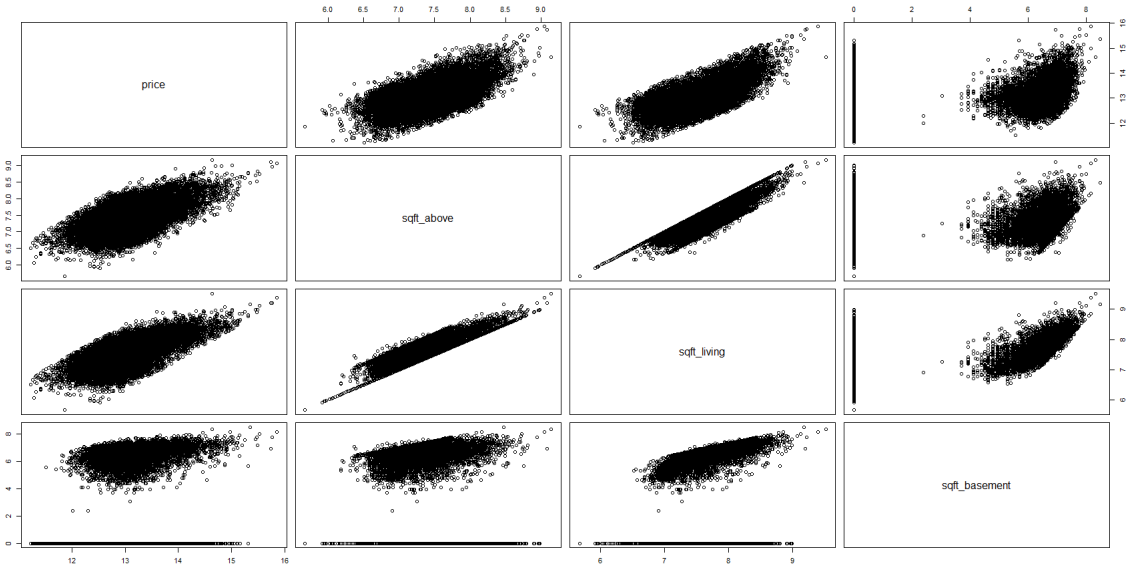
2.4.3 Thống kê đa biến

Ta sẽ tiếp tục trực quan hóa mối quan hệ giữa biến price và các biến còn lại qua đồ thị. Đối với các biến rời rạc, ta sẽ dùng đồ thị boxplot và hàm **boxplot()** để xem xét sự phân phối biến price theo các biến rời rạc. Đối với các biến liên tục còn lại, ta sẽ sử dụng lệnh **pairs()** để vẽ đồ thị phân tán giữa các đại lượng liên tục.

```
# Boxplot price for each floor.
boxplot(price~floors, new_df, main="Phan phoi price cho moi floor", col = "blue")

# Boxplot price for each condition
boxplot(price~condition, new_df, main="Phan phoi price cho moi condition", col = "blue")

# Boxplot price for each view
boxplot(price~view, new_df, main="Phan phoi price cho moi view", col = "blue")

#Pairs plot price for each sqft_above, sqft_living, sqft_basement
contTab = subset(new_df, select = c(price, sqft_above, sqft_living, sqft_basement))
pairs(contTab)
```



Hình 7: Phân phối của price lần lượt theo floors, condition, view.



Hình 8: Đồ thị phân tán giữa các đại lượng liên tục

2.5 Mô hình hồi quy tuyến tính

2.5.1 Xây dựng mô hình hồi quy tuyến tính bội

Ta sẽ xây dựng mô hình hồi quy tuyến tính bội với **price** là biến phụ thuộc và tất cả các biến còn lại là biến độc lập. Sử dụng lệnh **lm** để thực hiện. Sau đó dùng lệnh **summary** để xem kết quả:

```
linearModule <-lm(price~sqft_living+sqft_above+floors+condition+sqft_basement+view,
    data=new_df)
summary(linearModule)
```

Kết quả tổng quan về mô hình ta nhận được:

```
Call:
lm(formula = price ~ sqft_living + sqft_above + floors + condition +
    sqft_basement + view, data = new_df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.21674 -0.27523  0.01533  0.24741  1.45544

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.166053   0.051849 138.209  < 2e-16 ***
sqft_living   0.172981   0.029199   5.924 3.19e-09 ***
sqft_above    0.544631   0.029247  18.622  < 2e-16 ***
floors        0.102728   0.005834  17.609  < 2e-16 ***
condition     0.075292   0.004014  18.757  < 2e-16 ***
sqft_basement 0.042972   0.001975  21.761  < 2e-16 ***
view          0.125300   0.003404  36.804  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3671 on 21586 degrees of freedom
Multiple R-squared:  0.5141,    Adjusted R-squared:  0.514
F-statistic:  3807 on 6 and 21586 DF,  p-value: < 2.2e-16
```

Hình 9: Tổng quan về mô hình hồi quy

2.5.2 Ước lượng tham số và ảnh hưởng của các yếu tố tác động lên giá nhà

Hệ số của các biến là các giá trị trong cột **Estimate** trong bảng *Summary* hình 9 hoặc có thể xuất thông qua lệnh `linearModule$coefficients` :

```
> linearModule$coefficients
  (Intercept)   sqft_living    sqft_above        floors     condition sqft_basement          view
   7.16605337    0.17298105    0.54463145    0.10272805    0.07529167    0.04297234    0.12530011
```

Hình 10: Hệ số hồi quy

Ta thấy các hệ số (Estimate) của các biến **sqft_living, sqft_above, sqft_basement, floors, condition, view** đều dương nên kết luận các biến này tỷ lệ thuận lên **price** - giá nhà.
Ngoài ra ta còn tính được thêm các giá trị SSE, SSR, SST:

```
SSE <- (linearModule$residuals ^ 2) %>% sum()
SSR <- ((linearModule$fitted.values - mean(new_df$price)) ^ 2) %>% sum()
SST <- ((new_df$price - mean(new_df$price)) ^ 2) %>% sum()
```

Kết quả:

SSE	2908.98596191874
SSR	3078.04673940695
SST	5987.03270132569

2.5.3 Ước lượng độ lệch chuẩn của sai số

$$\hat{\sigma}^2 = \frac{SSE}{n-p} = 0.1347626$$

với n là số quan sát, và p là số lượng tham số hồi quy

```
#### Standard deviation of error
error.sd <- sqrt(SSE / linreg$df.residual)
error.sd
```

Như vậy độ lệch chuẩn ước lượng của sai số là:

$$\hat{\sigma} = \sqrt{0.1347626} = 0.3671003$$

2.5.4 Xác định hệ số R^2 hiệu chỉnh

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{2908.986/(21593-7)}{5987.033/(21593-1)} = 0.4860161$$

với n là số quan sát và p là số lượng tham số hồi quy.

2.5.5 Kiểm định đường hồi quy và các hệ số hồi quy

Trước tiên ta kiểm định đường hồi quy với mức ý nghĩa 0.01

- Giả thuyết H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$, với $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ lần lượt là hệ số hồi quy của **floors, condition, view, sqft_above, sqft_living, sqft_basement**.
- Giả thuyết H_1: $\exists i,\ \beta_i \neq 0$.

- Tiêu chuẩn kiểm định: $F_0 = \dfrac{SSR/6}{SSE/(21593-7)} \sim F(6, 21568)$
- Miền bác bỏ $D_{RR} = [f_{0.01,6,21568}, +\infty)$ với $f_{0.01,6,21568} = 2.802$.
- Giá trị của tiêu chuẩn kiểm định với mẫu hiện tại: $f_0 = 3806.751 \in D_{RR}$.

Như vậy, ta bác bỏ giải thuyết H_0. Đường hồi quy hiện tại có khả năng giải thích được biến **price**.
Tiếp theo ta lần lượt kiểm định từng hệ số hồi quy β_1, β_2, ..., β_6 với mức ý nghĩa 0.05.

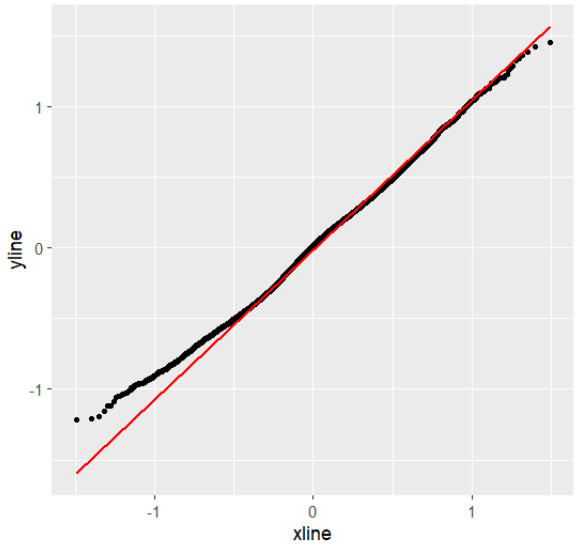
- Giả thuyết H_0: $\beta_i = 0$.
- Giả thuyết H_1: $\beta_i \neq 0$.

Quay trở lại ở bảng Summary hình 9, ta thấy các *p-value* đều nhỏ hơn $2e-16$, riêng biến **sqft_living** là $3.19e-09$ đều nhỏ hơn mức ý nghĩa là 0.05. Vậy ta có thể yêu tâm bác bỏ các giả thiết $\beta_i = 0$.

2.5.6 Kiểm định sự phù hợp của mô hình hồi quy tuyến tính

Ta phải kiểm tra rằng phần dư trong mô hình tuân theo phân phối chuẩn. Ta sử dụng lệnh **ggplot** để quan sát biểu đồ xác suất chuẩn của độ lệch dự đoán:

```
ggplot(mapping = aes(sample = linearModule$residuals)) +
      stat_qq_point(size = 1) + stat_qq_line(color = "red") +
labs(title = "Normal probability plot")
```



Hình 11: Biểu đồ xác suất chuẩn của độ lệch dự đoán

Trong biểu đồ, các điểm dữ liệu phần dư tập trung khá sát với đường chéo, như vậy phần dư có phân phối xấp xỉ chuẩn, giả định phân phối chuẩn của phần dư không bị vi phạm.

2.6 Dự đoán giá nhà ở quận King

Ta sử dụng lệnh **predict()** trong **R** để dự đoán giá nhà của một thuộc tính cho trước vào mô hình hồi quy ta đã xây dựng.
Ví dụ 1: Dự đoán giá nhà của thuộc tính x_1 : $floors = 1.5, condition = 5, view = 4, sqft\_above = mean(sqft\_above), sqft\_living = mean(sqft\_living), sqft\_basement = mean(sqft\_basement)$
Chạy đoạn chương trình sau:

```
x1<-data.frame(sqft_living = mean(new_df$sqft_living), sqft_above =
    mean(new_df$sqft_above), floors = max(new_df$floors), condition =
    max(new_df$condition),
               sqft_basement = mean(new_df$sqft_basement), view = min(new_df$view))

predict(linearModule, newdata = x1, interval = "confidence")
```

Kết quả ta nhận được:

	fit	lwr	upr
1	13.34429	13.31576	13.37283

Hình 12: Dự đoán giá nhà x_1

Giá nhà của x_1 được dự đoán nằm trong khoảng tin cậy $[13.31576; 13.37283]$ với giá trị trung bình là 13.34429.

3 Hoạt động 2

3.1 Yêu cầu

- Sinh viên tự tìm một bộ dữ liệu thuộc về chuyên ngành của mình. Khuyến khích sinh viên sử dụng dữ liệu thực tế có sẵn từ các thí nghiệm, khảo sát, dự án, ... trong chuyên ngành của mình. Ngoài ra sinh viên có thể tự tìm kiếm dữ liệu từ những nguồn khác hoặc tham khảo trong kho dữ liệu cung cấp trong tập tin "kho_du_lieu_BTL_xstk.xlsx".
- Sinh viên được tự do chọn phương pháp lý thuyết phù hợp để áp dụng phân tích dữ liệu của mình, nhưng phải đảm bảo 2 phần: Làm rõ dữ liệu (data visualization) và mô hình dữ liệu (model fitting).

3.2 Dữ liệu sử dụng

Trong hoạt động 2 này, nhóm chúng em lựa chọn bộ dữ liệu Computer Hardware dataset trong tập tin "kho_du_lieu_BTL_xstk.xlsx" [1].
Tập dữ liệu gồm có thông tin về 209 con chip CPU cùng với thông tin về hiệu năng tương đối của những con chip đó. Tập dữ liệu gồm có 9 thuộc tính:

1. **vendor name**: Tên nhà sản xuất con chip.
2. **model name**: Mã riêng biệt được gắn với mỗi con chip.
3. **MYCT**: Độ dài một chu kì clock của CPU, đơn vị nanosecond (ns).
4. **MMIN**: Kích thước RAM tối thiểu, đơn vị kilobyte (KB).
5. **MMAX**: Kích thước RAM tối đa, đơn vị kilobyte (KB).
6. **CACH**: Kích thước bộ nhớ đệm, đơn vị kilobyte (KB).
7. **CHMIN**: Số lượng kênh tối thiểu, tính theo đơn vị.
8. **CHMAX**: Số lượng kênh tối đa, tính theo đơn vị.
9. **PRP**: Hiệu năng tương đối được nhà sản xuất ghi nhận của CPU.
10. **ERP**: Hiệu năng tương đối mà các tác giả trong bài báo gốc ước tính được sử dụng mô hình hồi quy họ xây dựng.

3.3 Mục tiêu

Mục tiêu của nhóm là xây dựng một mô hình hồi quy tuyến tính bội có thể ước lượng được hiệu năng tương đối của CPU dựa trên các thông số của CPU đó.

Cụ thể hơn ở đây ta có,

- Biến dự đoán:
  - **MYCT**
  - **MMIN**
  - **MMAX**

  - **CACH**
  - **CHMIN**
  - **CHMAX**

- Biến được ước lượng: **PRP**

3.4 Đọc dữ liệu

Sau khi đọc dữ liệu từ file, ta thu được bảng dữ liệu như ở Hình 13.

```
original_data <- read.csv("Computer hardware.data", header = FALSE)
names(original_data) <- c("vendor name", "model name", "MYCT", "MMIN", "MMAX", "CACH",
    "CHMIN", "CHMAX", "PRP", "ERP")
original_data
```

vendor name <chr>	model name <chr>	MYCT <int>	MMIN <int>	MMAX <int>	CACH <int>	CHMIN <int>	CHMAX <int>	PRP <int>	ERP <int>
adviser	32/60	125	256	6000	256	16	128	198	199
amdahl	470v/7	29	8000	32000	32	8	32	269	253
amdahl	470v/7a	29	8000	32000	32	8	32	220	253
amdahl	470v/7b	29	8000	32000	32	8	32	172	253
amdahl	470v/7c	29	8000	16000	32	8	16	132	132
amdahl	470v/b	26	8000	32000	64	8	32	318	290
amdahl	580-5840	23	16000	32000	64	16	32	367	381
amdahl	580-5850	23	16000	32000	64	16	32	489	381
amdahl	580-5860	23	16000	64000	64	16	32	636	749
amdahl	580-5880	23	32000	64000	128	32	64	1144	1238

1-10 of 209 rows Previous 1 2 3 4 5 6 ... 21 Next

Hình 13: Dữ liệu gốc từ dataset

3.5 Làm sạch dữ liệu

3.5.1 Trích xuất các thuộc tính chính

```
new_data <- original_data %>% select(3:9)
new_data
```

Như đã đề cập ở trên, ở đây ta loại bỏ 3 thuộc tính không cần thiết cho việc xây dựng mô hình là **vendor name**, **model name** và **ERP**. Ta thu được bảng dữ liệu như ở Hình 15 với 7 thuộc tính.

3.5.2 Kiểm tra dữ liệu khuyết

Không có dữ liệu bị khuyết trong tập dữ liệu này.

```
new_data %>% sapply(function(col) sum(is.na(col)))
```

```
MYCT  MMIN  MMAX  CACH CHMIN CHMAX   PRP
   0     0     0     0     0     0     0
```

Hình 14: Số lượng dữ liệu khuyết ở mỗi biến

MYCT <int>	MMIN <int>	MMAX <int>	CACH <int>	CHMIN <int>	CHMAX <int>	PRP <int>
125	256	6000	256	16	128	198
29	8000	32000	32	8	32	269
29	8000	32000	32	8	32	220
29	8000	32000	32	8	32	172
29	8000	16000	32	8	16	132
26	8000	32000	64	8	32	318
23	16000	32000	64	16	32	367
23	16000	32000	64	16	32	489
23	16000	64000	64	16	32	636
23	32000	64000	128	32	64	1144

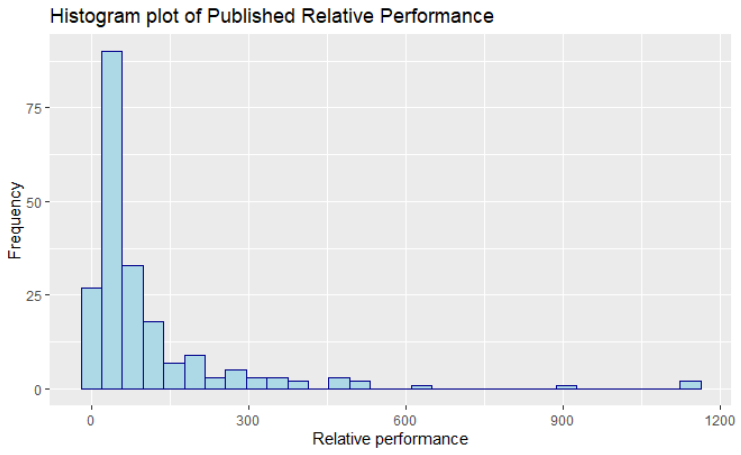
1-10 of 209 rows Previous 1 2 3 4 5 6 .. 21 Next

Hình 15: Dữ liệu gồm các thuộc tính chính

3.6 Làm rõ dữ liệu và Thống kê mô tả

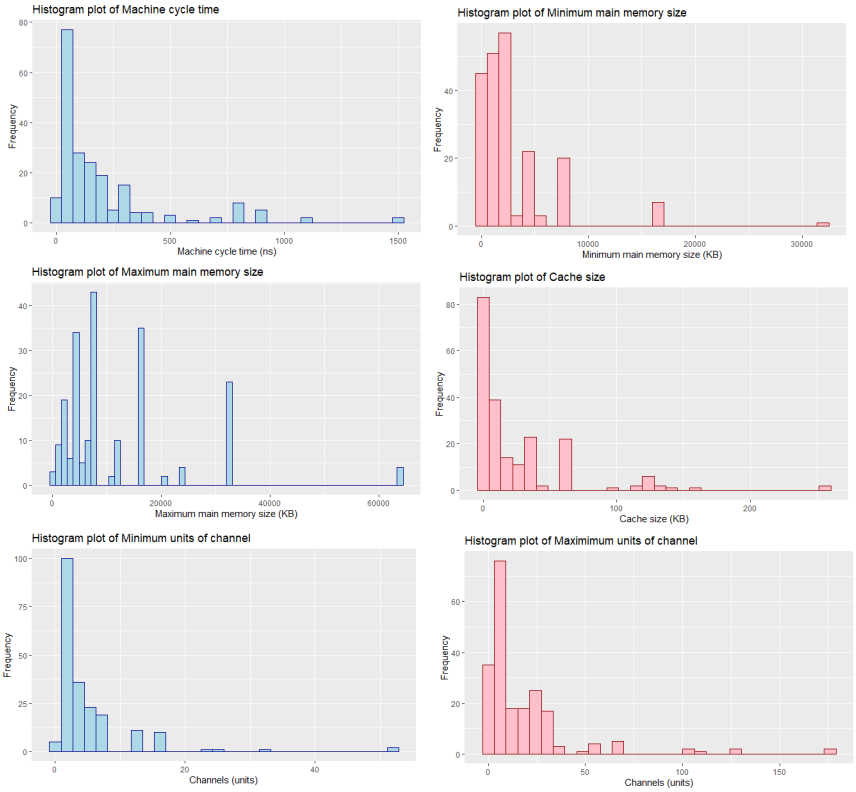
3.6.1 Chuyển đổi biến

Ta vẽ sơ đồ cột cho từng biến (Hình 17 và Hình 16) để có một cái nhìn tổng quan hơn về phân phối giá trị của các biến.



Hình 16: Biểu đồ cột thể hiện phân phối giá trị của biến được ước lượng **PRP**

Nhận thấy khoảng cách giữa 2 giá trị liên tiếp có vẻ tăng theo độ lớn của 2 giá trị đó. Ngoài ra, các giá trị tập trung nhiều ở gần 0 hơn.
Để cho các giá trị có phân phối đều hơn trên một khoảng giá trị, ta sẽ lấy logarithm của các biến. Lưu ý các biến **CACH**, **CHMIN**, **CHMAX** sẽ được cộng 1 trước khi lấy logarithm để



Hình 17: Biểu đồ cột thể hiện phân phối giá trị của các biến dự đoán

tránh trường hợp những biến này nhận giá trị bằng 0. Những biến còn lại không bao giờ nhận giá trị 0.

```
new_data$MYCT <- new_data$MYCT %>% log()
new_data$MMIN <- new_data$MMIN %>% log()
new_data$MMAX <- new_data$MMAX %>% log()
new_data$CACH <- (new_data$CACH + 1) %>% log()
new_data$CHMIN <- (new_data$CHMIN + 1) %>% log()
new_data$CHMAX <- (new_data$CHMAX + 1) %>% log()
new_data$PRP <- new_data$PRP %>% log()
```

3.6.2 Thống kê đơn biến

Ở đây, ta sẽ xem các biến như là các biến liên tục và xây dựng bảng tổng hợp một số thống kê quan trọng cho các biến sau khi được chuyển đổi.

```
means <- new_data %>% sapply(mean)
medians <- new_data %>% sapply(median)
sds <- new_data %>% sapply(sd)
mins <- new_data %>% sapply(min)
maxs <- new_data %>% sapply(max)
```

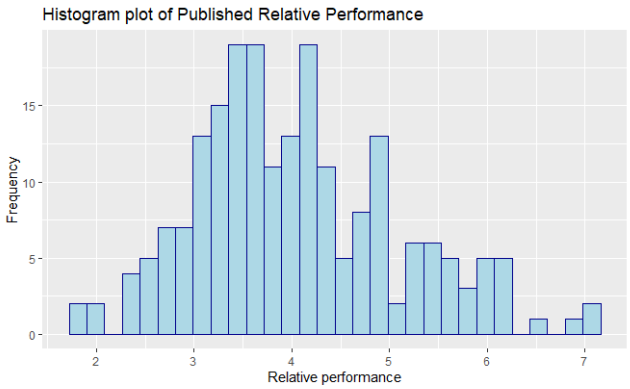
```
uniques <- new_data %>% sapply(function(col) length(unique(col)))
summary <- as.data.frame(cbind(means, medians, sds, mins, maxs, uniques))
names(summary) <- c("mean", "median", "sd", "min", "max", "unique")
summary
```

	mean <dbl>	median <dbl>	sd <dbl>	min <dbl>	max <dbl>	unique <dbl>
MYCT	4.746955	4.700480	1.0378866	2.833213	7.313220	60
MMIN	7.360234	7.600902	1.1038587	4.158883	10.373491	25
MMAX	8.922222	8.987197	1.0321500	4.158883	11.066638	23
CACH	2.075453	2.197225	1.7072286	0.000000	5.549076	22
CHMIN	1.355234	1.098612	0.8055484	0.000000	3.970292	15
CHMAX	2.407778	2.197225	1.0359947	0.000000	5.176150	31
PRP	4.037242	3.912023	1.0483648	1.791759	7.047517	116

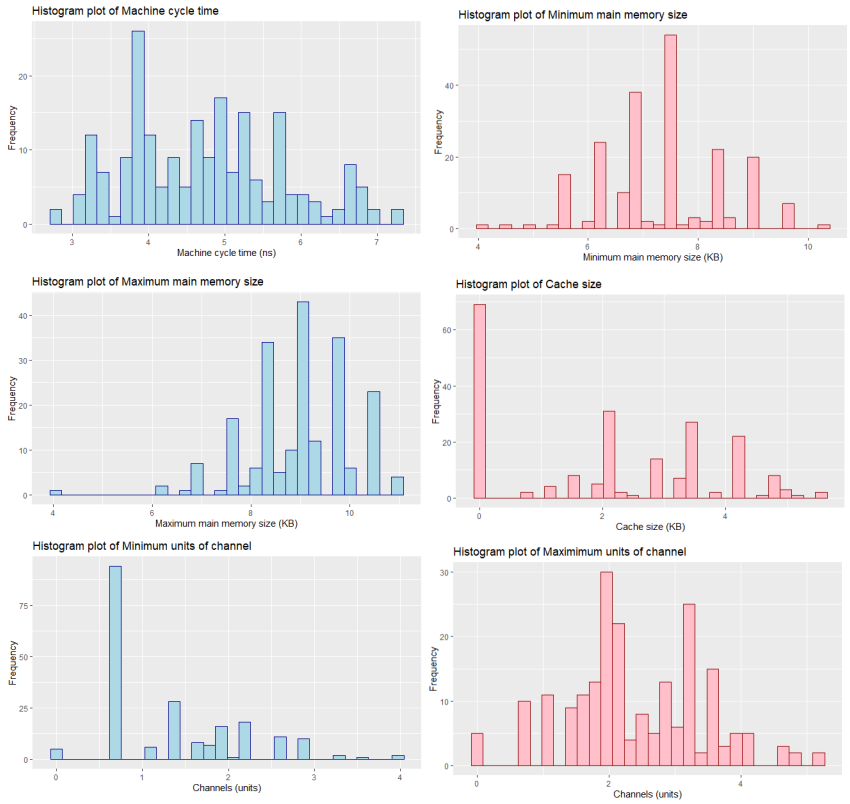
7 rows

Hình 18: Một số thống kê quan trọng

Ngoài các cột thể hiện các giá trị trung bình cộng, trung vị, độ lệch chuẩn, giá trị bé nhất và lớn nhất, còn có thêm một cột thể hiện số giá trị riêng biệt của mỗi biến (cột cuối của Hình 18). Ta có các biểu đồ cột cho từng biến ở Hình 20 và Hình 19.



Hình 19: Biểu đồ cột cho biến được ước lượng **PRP**



Hình 20: Biểu đồ cột cho các biến dự đoán

3.6.3 Thống kê đa biến

Ma trận hiệp phương sai giữa các biến được thể hiện ở Hình 21.

```
new_data %>% cov() %>% as.data.frame()
```

	MYCT <dbl>	MMIN <dbl>	MMAX <dbl>	CACH <dbl>	CHMIN <dbl>	CHMAX <dbl>	PRP <dbl>
MYCT	1.0772086	-0.8356683	-0.6782635	-1.0934254	-0.5166114	-0.5922359	-0.7636093
MMIN	-0.8356683	1.2185040	0.8354628	1.1320365	0.5076570	0.5054680	0.8845829
MMAX	-0.6782635	0.8354628	1.0653337	1.0859777	0.4037242	0.5483370	0.8622087
CACH	-1.0934254	1.1320365	1.0859777	2.9146294	0.7179031	0.8687128	1.3736763
CHMIN	-0.5166114	0.5076570	0.4037242	0.7179031	0.6489081	0.5923601	0.5680989
CHMAX	-0.5922359	0.5054680	0.5483370	0.8687128	0.5923601	1.0732850	0.6956917
PRP	-0.7636093	0.8845829	0.8622087	1.3736763	0.5680989	0.6956917	1.0990688

7 rows

Hình 21: Ma trận hiệp phương sai

Ma trận tương quan giữa các biến được thể hiện ở Hình 22.

```
new_data %>% cor() %>% as.data.frame()
```

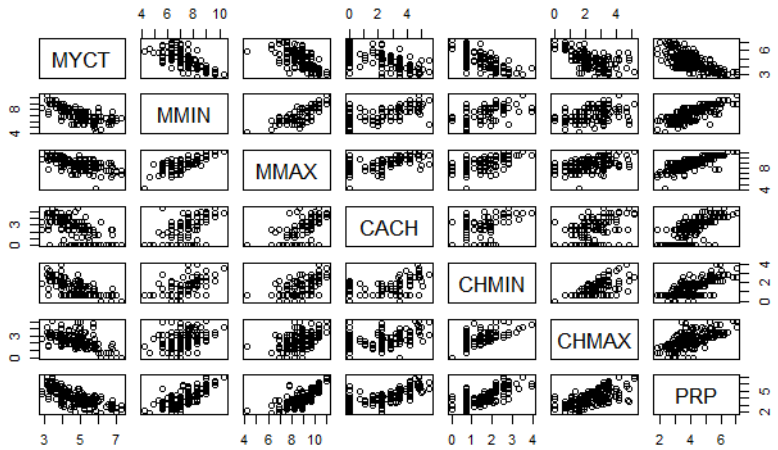
	MYCT <dbl>	MMIN <dbl>	MMAX <dbl>	CACH <dbl>	CHMIN <dbl>	CHMAX <dbl>	PRP <dbl>
MYCT	1.0000000	-0.7294080	-0.6331487	-0.6170887	-0.6179061	-0.5507916	-0.7017927
MMIN	-0.7294080	1.0000000	0.7332816	0.6006968	0.5709069	0.4420004	0.7643858
MMAX	-0.6331487	0.7332816	1.0000000	0.6162918	0.4855683	0.5127990	0.7968144
CACH	-0.6170887	0.6006968	0.6162918	1.0000000	0.5220145	0.4911646	0.7675034
CHMIN	-0.6179061	0.5709069	0.4855683	0.5220145	1.0000000	0.7098011	0.6726976
CHMAX	-0.5507916	0.4420004	0.5127990	0.4911646	0.7098011	1.0000000	0.6405409
PRP	-0.7017927	0.7643858	0.7968144	0.7675034	0.6726976	0.6405409	1.0000000

7 rows

Hình 22: Ma trận tương quan

Đồ thị mô tả sự tương quan giữa các biến được thể hiện ở Hình 23.

```
new_data %>% pairs()
```



Hình 23: Đồ thị phân tán giữa các cặp biến

3.7 Mô hình hồi quy tuyến tính

Ở đây ta sẽ ký hiệu $Y = PRP$, $X_1 = MYCT$, $X_2 = MMIN$, $X_3 = MMAX$, $X_4 = CACH$, $X_5 = CHMIN$ và $X_6 = CHMAX$.

3.7.1 Định nghĩa mô hình

Ta đặt ra giả thiết như sau:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon$$

Trong đó, $\beta_i \in \mathbb{R}$ và $\epsilon \sim \mathcal{N}(0, \sigma^2)$, ϵ độc lập với các biến ngẫu nhiên X_i, $i = \overline{1,6}$.

Nếu ta kí hiệu $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_6 \end{pmatrix}$, ta có thể viết:

$$Y = \begin{pmatrix} 1 & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 \end{pmatrix} \boldsymbol{\beta} + \epsilon$$

3.7.2 Một số thống kê mẫu

Ở đây, ta có một mẫu kích thước 209. Ta ký hiệu:

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_6^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_6^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(209)} & x_2^{(209)} & \cdots & x_6^{(209)} \end{pmatrix}$$

$$\boldsymbol{Y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(209)} \end{pmatrix}$$

gồm 209 hàng, mỗi hàng tương ứng với 1 phần tử của mẫu.
Dựa vào Hình 18, ta có:

- $\overline{x_1} = 4.75955$, $s_{x_1} = 1.0378866$, $S_{x_1x_1} = 224.05939$
- $\overline{x_2} = 7.360234$, $s_{x_2} = 1.1038587$, $S_{x_2x_2} = 253.44884$
- $\overline{x_3} = 8.922222$, $s_{x_3} = 1.0321500$, $S_{x_3x_3} = 221.58939$
- $\overline{x_4} = 2.075453$, $s_{x_4} = 1.7072286$, $S_{x_4x_4} = 606.24293$
- $\overline{x_5} = 1.355234$, $s_{x_5} = 0.8055484$, $S_{x_5x_5} = 134.97291$
- $\overline{x_6} = 2.407778$, $s_{x_6} = 1.0359947$, $S_{x_6x_6} = 223.24328$
- $\overline{y} = 4.037242$, $s_y = 1.0483648$, $S_{yy} = 228.60630$
- Ma trận hiệp phương sai được thể hiện ở Hình 21.

3.7.3 Ước lượng tham số

Ta đã biết $\boldsymbol{\beta}$ được ước lượng bằng công thức sau $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. Như vậy,

$$\begin{aligned} \hat{\beta_0} &= -1.297365 \\ \hat{\beta_1} &= 0.005072 \\ \hat{\beta_2} &= 0.194948 \\ \hat{\beta_3} &= 0.325042 \\ \hat{\beta_4} &= 0.188860 \\ \hat{\beta_5} &= 0.194507 \\ \hat{\beta_6} &= 0.132900 \end{aligned}$$

Sử dụng R ta thu được kết quả ước lượng như ở cột 1 Hình 24.

```
linreg <- lm(PRP ~ MYCT + MMIN + MMAX + CACH + CHMIN + CHMAX, new_data)
summary(linreg)$coefficients %>% as.data.frame() %>% select(1:2)
```

	Estimate <dbl>	Std. Error <dbl>
(Intercept)	-1.297364940	0.54678716
MYCT	0.005072295	0.04837169
MMIN	0.194948355	0.04895054
MMAX	0.325042010	0.04772017
CACH	0.188859695	0.02503585
CHMIN	0.194507256	0.06046611
CHMAX	0.132900186	0.04463259

7 rows

Hình 24: Ước lượng hệ số hồi quy và độ lệch chuẩn

Ngoài ra, ta tính toán thêm được:

```
SSE <- (linreg$residuals ^ 2) %>% sum()
SSR <- ((linreg$fitted.values - mean(new_data$PRP)) ^ 2) %>% sum()
SST <- ((new_data$PRP - mean(new_data$PRP)) ^ 2) %>% sum()
ss <- as.data.frame(c(SSE, SSR, SST), row.names = c("SSE", "SSR", "SST"))

names(ss) <- c("Value")
ss
```

$$\begin{array}{rcl} SSE & = & 39.07301 \\ SSR & = & 189.53330 \\ \hline SST & = & 228.60631 \end{array}$$

3.7.4 Ước lượng độ lệch chuẩn của sai số

$$\hat{\sigma}^2 = \frac{SSE}{n-p} = \frac{39.07301}{209-7} = 0.1934307$$

với n là số quan sát và p là số lượng tham số hồi quy.
Như vậy, độ lệch chuẩn ước lượng được của sai số là:

$$\hat{\sigma} = \sqrt{0.1934307} = 0.43980762$$

Lệnh R sau trả về kết quả xấp xỉ với kết quả ta thu được.

```
summary(linreg)$sigma
```

3.7.5 Xác định hệ số R^2 hiệu chỉnh

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{39.07301/(209-7)}{228.60631/(209-1)} = 0.8240048$$

với n là số quan sát và p là số lượng tham số hồi quy.

Lệnh R sau trả về kết quả xấp xỉ với kết quả ta thu được.

```
summary(linreg)$adj.r.squared
```

3.7.6 Xác định khoảng tin cậy của các hệ số hồi quy

Gọi $\hat{B}_0$, $\hat{B}_1$, ..., $\hat{B}_6$ là các hàm ước lượng không chệch các tham số β_0, β_1, ..., β_6 bằng phương pháp hồi quy tuyến tính.
$\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_6$ chính là giá trị của các hàm ước lượng tương ứng với mẫu hiện tại.
Xét ma trận $\boldsymbol{S} = \hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Khi đó $\hat{\sigma}^2_{\hat{B}_i} = \boldsymbol{S}_{(i+1)(i+1)}$, $i = \overline{0,6}$. Ta tính toán được

$$\begin{aligned}
\hat{\sigma}_{\hat{B}_0} &= 0.5468 \\
\hat{\sigma}_{\hat{B}_1} &= 0.0484 \\
\hat{\sigma}_{\hat{B}_2} &= 0.0490 \\
\hat{\sigma}_{\hat{B}_3} &= 0.0477 \\
\hat{\sigma}_{\hat{B}_4} &= 0.0250 \\
\hat{\sigma}_{\hat{B}_5} &= 0.0605 \\
\hat{\sigma}_{\hat{B}_6} &= 0.0446
\end{aligned}$$

Sử dụng R ta thu được kết quả như ở cột 2 Hình 24.
Chọn độ tin cậy $\gamma = 1 - \alpha = 0.95$.
Ta đã biết $\hat{B}_i$ xấp xỉ phân phối chuẩn, hay $\hat{B}_i \sim \mathcal{N}(\beta_i, \hat{\sigma}^2_{\hat{B}_i})$.
Xét hàm thống kê $G_i = \dfrac{\hat{B}_i - \beta_i}{\hat{\sigma}_{\hat{B}_i}}$. Bởi vì kích thước mẫu khá lớn nên $G_i \sim \mathcal{N}(0,1)$.
Như vậy, $P(|G_i| < z_{\frac{\alpha}{2}}) = 1 - \alpha \iff P(-z_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{B}_i} + \hat{B}_i < \beta_i < z_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{B}_i} + \hat{B}_i) = 1 - \alpha$. Khoảng ước lượng với độ tin cậy 95% cho các tham số là $(-z_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{B}_i} + \hat{\beta}_i, z_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{B}_i} + \hat{\beta}_i)$, cụ thể

$$\begin{aligned}
\beta_0 &\in (-2.369093, -0.225637) \\
\beta_1 &\in (-0.089792, 0.099936) \\
\beta_2 &\in (0.098908, 0.290988) \\
\beta_3 &\in (0.231550, 0.418534) \\
\beta_4 &\in (0.139860, 0.237860) \\
\beta_5 &\in (0.075927, 0.313087) \\
\beta_6 &\in (0.045484, 0.220316)
\end{aligned}$$

Sử dụng R ta có các khoảng tin cậy 95% cho các hệ số hồi quy ở Hình 25.

```
confint(linreg, level = 0.95) %>% as.data.frame()
```

	2.5 % <dbl>	97.5 % <dbl>
(Intercept)	-2.37550750	-0.2192224
MYCT	-0.09030591	0.1004505
MMIN	0.09842879	0.2914679
MMAX	0.23094846	0.4191356
CACH	0.13949457	0.2382248
CHMIN	0.07528155	0.3137330
CHMAX	0.04489466	0.2209057

7 rows

Hình 25: Khoảng tin cậy 95% cho các hệ số hồi quy

3.7.7 Kiểm định đường hồi quy và các hệ số hồi quy

Trước tiên ta kiểm định đường hồi quy với mức ý nghĩa 0.01.

- Giả thuyết H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$.
- Giả thuyết H_1: $\exists i,\ \beta_i \neq 0$.
- Tiêu chuẩn kiểm định: $F_0 = \dfrac{SSR/6}{SSE/(209-7)} \sim F(6, 202)$
- Miền bác bỏ $D_{RR} = [f_{0.01,6,202}, +\infty)$ với $f_{0.01,6,202} = 2.802$.
- Giá trị của tiêu chuẩn kiểm định với mẫu hiện tại: $f_0 = 163.3084944 \in D_{RR}$.

Như vậy, ta bác bỏ giả thuyết H_0. Đường hồi quy hiện tại có khả năng giải thích được biến **PRP**.
Tiếp theo ta lần lượt kiểm định từng hệ số hồi quy β_1, β_2, ..., β_6 với mức ý nghĩa 0.05.

- Giả thuyết H_0: $\beta_i = 0$.
- Giả thuyết H_1: $\beta_i \neq 0$.
- Tiêu chuẩn kiểm định: $T_i = \dfrac{\hat{B}_i}{\hat{\sigma}_{\hat{B}_i}} \sim \mathcal{N}(0,1)$ (do mẫu có kích thước khá lớn).
- Miền bác bỏ $D_{RR} = (-\infty, -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}, +\infty)$.

  Nhận thấy rằng $t_i = \dfrac{\hat{\beta}_i}{\hat{\sigma}_{\hat{B}_i}}$ thuộc miền bác bỏ $\iff$ khoảng ước lượng của β_i không chứa 0.

Như vậy, ta có thể bác bỏ được các giả thuyết $\beta_i = 0$ với $i = \overline{2,6}$. Riêng giả thuyết $\beta_1 = 0$ ta chưa bác bỏ được.
Sử dụng R, ta có các giá trị p-value như Hình 26.

```
summary(linreg)$coefficients %>% as.data.frame() %>% select(3:4)
```

Nhận thấy ngoại trừ giả thuyết $\beta_1 = 0$ ứng với biến **MYCT**, ta có thể yên tâm bác bỏ các giả thuyết $\beta_i = 0$ với $i = \overline{2,6}$.

	t value <dbl>	Pr(>|t|) <dbl>
(Intercept)	-2.3727056	1.859706e-02
MYCT	0.1048608	9.165903e-01
MMIN	3.9825579	9.511066e-05
MMAX	6.8114180	1.084486e-10
CACH	7.5435701	1.526483e-12
CHMIN	3.2167981	1.509796e-03
CHMAX	2.9776491	3.259959e-03

7 rows

Hình 26: Các giá trị p-value của các kiểm định hệ số hồi quy

3.7.8 Xác định khoảng tin cậy của giá trị trung bình của Y khi $\mathbf{x} = \mathbf{x_0}$

Cho trước giá trị của các biến dự đoán, ký hiệu $\mathbf{x_0^T} = \begin{pmatrix} 1 & x_1 & x_2 & \cdots & x_6 \end{pmatrix}$. Ta muốn ước lượng được khoảng tin cậy của $\mu_{Y|\mathbf{x_0}} \equiv E[Y|\mathbf{x_0}] = \mathbf{x_0^T}\boldsymbol{\beta}$.
Ta có hàm ước lượng $\hat{\mu}_{Y|\mathbf{x_0}} = \mathbf{x_0^T}\hat{\boldsymbol{B}}$.
Ta đã biết,

- $\hat{\mu}_{Y|\mathbf{x_0}}$ có phân phối chuẩn.
- $\hat{\mu}_{Y|\mathbf{x_0}}$ là hàm ước lượng không chệch, hay $E[\hat{\mu}_{Y|\mathbf{x_0}}] = \mu_{Y|\mathbf{x_0}}$.
- $D[\hat{\mu}_{Y|\mathbf{x_0}}] = \sigma^2\mathbf{x_0^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{x_0}$ được ước lượng bằng $\hat{\sigma}^2\mathbf{x_0^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{x_0}$.

Do kích thước mẫu lớn, $\dfrac{\hat{\mu}^2_{Y|\mathbf{x_0}} - \mu_{Y|\mathbf{x_0}}}{\sqrt{\hat{\sigma}^2\mathbf{x_0^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{x_0}}} \sim \mathcal{N}(0,1)$.
Chọn độ tin cậy $\gamma = 1 - \alpha = 0.95$.
Như vậy, $P(|\dfrac{\hat{\mu}_{Y|\mathbf{x_0}} - \mu_{Y|\mathbf{x_0}}}{\sqrt{\hat{\sigma}^2\mathbf{x_0^T}(X^TX)^{-1}\mathbf{x_0}}}| < z_{\frac{\alpha}{2}}) = 1 - \alpha$.
Hay $P(-z_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2\mathbf{x_0^T}(X^TX)^{-1}\mathbf{x_0}} + \hat{\mu}_{Y|\mathbf{x_0}} < \mu_{Y|\mathbf{x_0}} < z_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2\mathbf{x_0^T}(X^TX)^{-1}\mathbf{x_0}} + \hat{\mu}_{Y|\mathbf{x_0}}) = 1 - \alpha$.
Vậy khoảng ước lượng với độ tin cậy 95% cho $\mu_{Y|\mathbf{x_0}}$ là

$$(-1.96\sqrt{\hat{\sigma}^2\mathbf{x_0^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{x_0}} + \mathbf{x_0^T}\hat{\boldsymbol{\beta}}, 1.96\sqrt{\hat{\sigma}^2\mathbf{x_0^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{x_0}} + \mathbf{x_0^T}\hat{\boldsymbol{\beta}})$$

3.7.9 Xác định khoảng tin cậy của các giá trị dự đoán

Giả sử trong tương lai, ta thu được một quan sát đã biết giá trị của x_1, x_2, ..., x_6, ký hiệu $\mathbf{x_0^T} = \begin{pmatrix} 1 & x_1 & x_2 & \cdots & x_6 \end{pmatrix}$.
Ta muốn xác định được khoảng tin cậy của giá trị y_0 đối với quan sát này.
Ta có hàm ước lượng $\hat{Y}_0 = \mathbf{x_0^T}\hat{\boldsymbol{B}}$.
Ta đã biết:

- $\hat{Y}_0 - y_0$ có phân phối chuẩn.
- $E[\hat{Y}_0 - y_0] = 0$.

- $D[\hat{Y_0} - y_0] = D[\mathbf{x_0^T}\hat{\boldsymbol{B}} - \mathbf{x_0^T}\boldsymbol{\beta} - \epsilon_0] = \sigma^2(1 + \mathbf{x_0^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{x_0})$ được ước lượng bằng $\hat{\sigma}^2(1 + \mathbf{x_0^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{x_0})$.

Do kích thước mẫu lớn, $\dfrac{\hat{Y_0} - y_0}{\sqrt{\hat{\sigma}^2(1 + \mathbf{x_0^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{x_0})}} \sim \mathcal{N}(0, 1)$.

Chọn độ tin cậy $\gamma = 1 - \alpha = 0.95$.

Như vậy, $P(|\dfrac{\hat{Y_0} - y_0}{\sqrt{\hat{\sigma}^2(1 + \mathbf{x_0^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{x_0})}}| < z_{\frac{\alpha}{2}}) = 1 - \alpha$.

Hay

$$P(-z_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2(1 + \mathbf{x_0^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{x_0})} + \hat{Y_0} < y_0 < z_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2(1 + \mathbf{x_0^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{x_0})} + \hat{Y_0}) = 1 - \alpha$$

Vậy khoảng ước lượng với độ tin cậy 95% cho y_0 là

$$(-1.96\sqrt{\hat{\sigma}^2(1 + \mathbf{x_0^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{x_0})} + \mathbf{x_0^T}\hat{\boldsymbol{\beta}}, 1.96\sqrt{\hat{\sigma}^2(1 + \mathbf{x_0^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{x_0})} + \mathbf{x_0^T}\hat{\boldsymbol{\beta}})$$

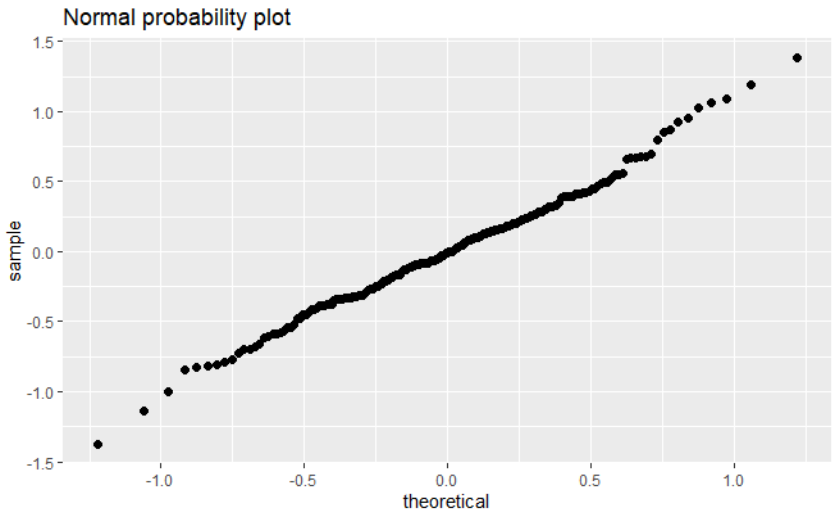
3.7.10 Kiểm định sự phù hợp của mô hình hồi quy tuyến tính

Khi định nghĩa mô hình, ta đã giả sử 2 điều sau:

- Sai số ngẫu nhiên ϵ có phân phối chuẩn $\mathcal{N}(0, \sigma^2)$.
- Sai số ngẫu nhiên ϵ phân phối độc lập với X_i, $i = \overline{1, 6}$.

Ta có thể quan sát biểu đồ xác suất chuẩn của độ lệch dự đoán ở Hình 27.

```
ggplot(mapping = aes(sample = linreg$residuals)) +
                        stat_qq_point(size = 2) +
                        labs(title = "Normal probability plot")
```



Hình 27: Biểu đồ xác suất chuẩn của độ lệch dự đoán

Vì biểu đồ gần với một đường thẳng, nên có thể cho rằng độ lệch chuẩn xấp xỉ phân phối chuẩn.

Ngoài ra, tiến hành kiểm định Shapiro-Wilk, với:

- Giả thuyết H0: Tổng thể độ lệch dự đoán có phân phối chuẩn.
- Giả thuyết H1: tổng thể độ lệch dự đoán không có phân phối chuẩn.

ta cũng thu được p-value bằng 0.3155.

```
shapiro.test(linreg$residuals)
```

Như vậy, ta chưa thể bác bỏ giả thuyết H0 và có thể chấp nhận nó.

Như vậy, mô hình hồi quy tuyến tính ở đây là phù hợp.

Tài liệu

[1] Phillip Ein-Dor and Jacob Feldmesser Ein-Dor, *Computer Hardware Dataset*, UCI Machine Learning Repository

[2] William N.Venables, David M.Smith and the R core team, *An Introduction to R: A Programming Environment for Data Analysis and Graphics*, Network Theory, [Bristol], 2009.

[3] Damodar N. Gujarati, Dawn C. Porter, *Basic Econometrics*, McGraw-Hill, Boston, 2009.

[4] Douglas C. Montgomery, George C. Runger, *Applied Statistics and Probability for Engineers*, Wiley, Hoboken, NJ, 2019.

[5] Hoàng Trọng, Chu Nguyễn Mộng Ngọc (2008), *Thống kê ứng dụng trong kinh tế - xã hội*, Nxb. Thống kê.

[6] Lê Thị Diệu Hiền (2010), *Hồi quy tuyến tính và ứng dụng*, Trường Đại học Cần Thơ