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

PHÂN TÍCH VÀ DỰ ĐOÁN HIỆU NĂNG TƯƠNG ĐỐI CỦA CPU

Giảng viên hướng dẫn: Nguyễn Tiến Dũng
Lớp L05 - Nhóm 9

STT	Họ và tên	MSSV	Khoa/Ngành học	Công việc	Tỉ lệ
1	Nguyễn Tuấn Minh	2110359	Khoa học và Kỹ thuật máy tính	Mô hình hồi quy tuyến tính Code R	**100%**
2	Phạm Anh Kiệt	2110304	Khoa học và Kỹ thuật máy tính	Cơ sở lý thuyết Thống kê mô tả	**100%**
3	Trần Minh Khoa	2110278	Khoa học và Kỹ thuật máy tính	Soạn báo cáo Code R	**100%**

THÀNH PHỐ HỒ CHÍ MINH, 04/2023

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 dữ liệu thống kê (dạng số hoặc không phải dạng số), nhằm rút ra các kết luận khoa học từ thực tiễn, dựa trên những thành tựu của lý thuyết xác suất.

1.1.2 Phân loại thống kê

Thống kê được chia làm hai loại chính là Descriptive Statistics (Thống kê mô tả) và Inferential Statistics (Thống kê suy diễn). 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.

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

1.2.3.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:

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

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.

1.2.3.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.

1.2.3.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:

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

2 Công việc

2.1 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.
- Sinh viên tự chọn phương pháp lý thuyết phù hợp để áp dụng phân tích dữ liệu của mình.

2.2 Dữ liệu sử dụng

Trong hoạt động 2 này, nhóm chúng em lựa chọn bộ dữ liệu Computer Hardware dataset.
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. **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.

2.3 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**

2.4 Đọc dữ liệu

Sau khi đọc dữ liệu từ file, ta thu được bảng dữ liệu như ở Hình 1.

```
#Nhap du lieu
data = read.csv("machine.data", header = FALSE)
colnames(data) = c("vendor name", "model name", "MYCT", "MMIN", "MMAX", "CACH",
    "CHMIN", "CHMAX", "PRP", "ERP")
paste("Du lieu co", ncol(data), "cot,", nrow(data), "hang")
```

```
[1] "Du lieu co 10 cot, 209 hang"
```

	vendor name	model name	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP	ERP
1	adviser	32/60	125	256	6000	256	16	128	198	199
2	amdahl	470v/7	29	8000	32000	32	8	32	269	253
3	amdahl	470v/7a	29	8000	32000	32	8	32	220	253
4	amdahl	470v/7b	29	8000	32000	32	8	32	172	253
5	amdahl	470v/7c	29	8000	16000	32	8	16	132	132
6	amdahl	470v/b	26	8000	32000	64	8	32	318	290
7	amdahl	580-5840	23	16000	32000	64	16	32	367	381
8	amdahl	580-5850	23	16000	32000	64	16	32	489	381
9	amdahl	580-5860	23	16000	64000	64	16	32	636	749
10	amdahl	580-5880	23	32000	64000	128	32	64	1144	1238
11	apollo	dn320	400	1000	3000	0	1	2	38	23
12	apollo	dn420	400	512	3500	4	1	6	40	24
13	basf	7/65	60	2000	8000	65	1	8	92	70
14	basf	7/68	50	4000	16000	65	1	8	138	117
15	bti	5000	350	64	64	0	1	4	10	15
16	bti	8000	200	512	16000	0	4	32	35	64
17	burroughs	b1955	167	524	2000	8	4	15	19	23
18	burroughs	b2900	143	512	5000	0	7	32	28	29

Hình 1: Dữ liệu gốc từ dataset

2.5 Làm sạch dữ liệu

2.5.1 Trích xuất các thuộc tính chính

```
#Lam sach du lieu
data = data %>% select(3:9)
```

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 3 với 7 thuộc tính.

2.5.2 Kiểm tra dữ liệu khuyết

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

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

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

Hình 2: 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 3: Dữ liệu gồm các thuộc tính chính

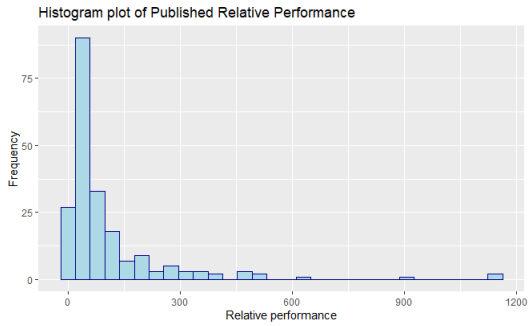
2.6 Làm rõ dữ liệu và Thống kê mô tả

2.6.1 Chuyển đổi biến

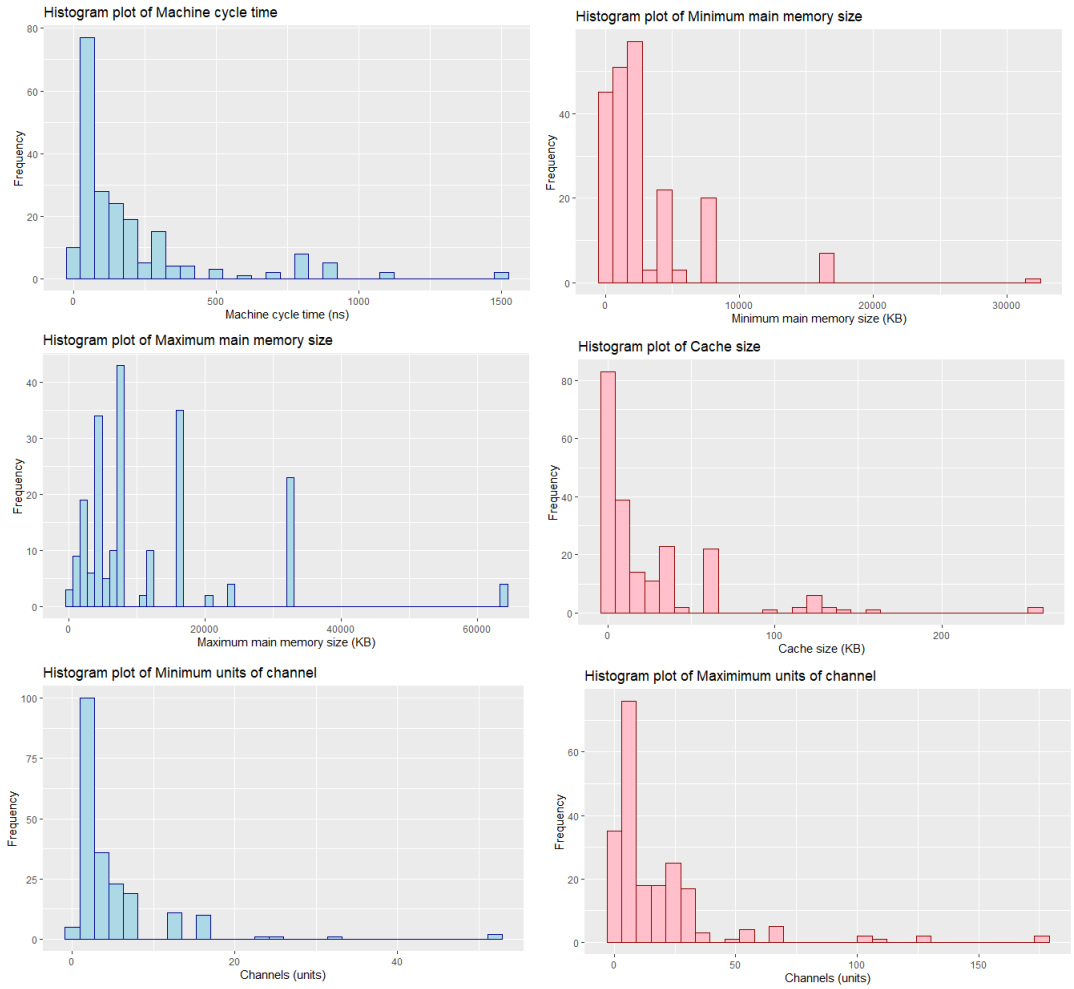
Ta vẽ sơ đồ cột cho từng biến (Hình 5 và Hình 4) để có một cái nhìn tổng quan hơn về phân phối giá trị của các biến.

```
#Do thi histogram
#PRP
data %>% ggplot(aes(x=PRP)) +
geom_histogram(color="darkblue", fill="lightblue") +
labs(title="Histogram plot of Published Relative Performance", x="Relative
    performance", y = "Frequency")
```

Làm tương tự với các biến còn lại (MYCT, MMIN, MMAX, CACH, CHMIN, CHMAX) ta được các biểu đồ cột thể hiện phân phối giá trị các biến như sau.



Hình 4: Biểu đồ cột thể hiện phân phối giá trị của biến được ước lượng **PRP**

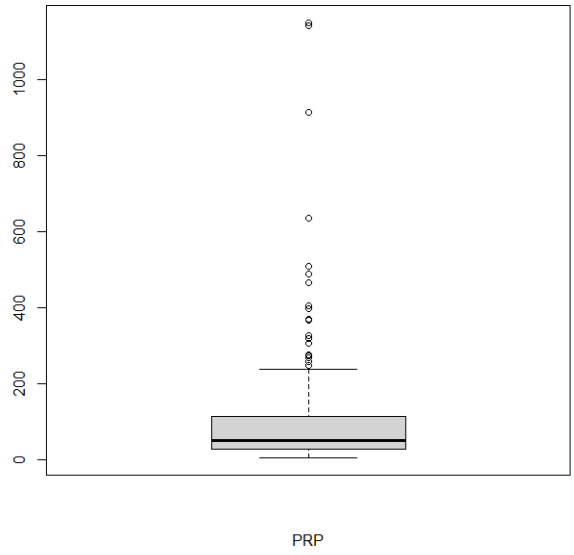


Hình 5: Biểu đồ cột thể hiện phân phối giá trị của các biến dự đoán

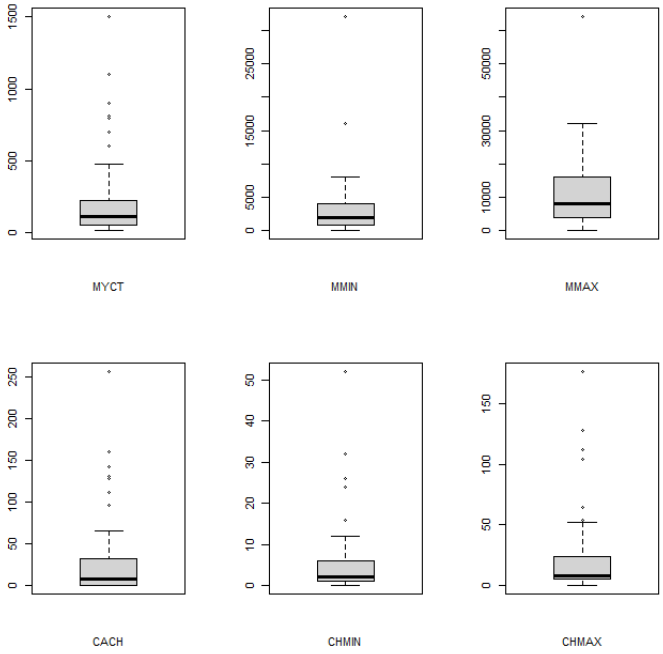
Ta nhận thấy có một vài điểm đòn bẩy cao (high leverage point), tiếp tục sử dụng biểu đồ hình hộp (boxplot) để xem xét các điểm này ở (Hình 7 và Hình 6)

```
#Boxplot
boxplot(data$MYCT)
boxplot(data$MMIN)
boxplot(data$MMAX)
boxplot(data$CACH)
boxplot(data$CHMIN)
boxplot(data$CHMAX)
boxplot(data$PRP)
```

Ở các biến được ước lượng và cả biến dự đoán, ta nhận thấy có vài giá trị có khả năng là outliers, tuy nhiên cũng có khả năng chỉ là một điểm đòn bẩy cao (high leverage point) và không ảnh hưởng đến độ chính xác của mô hình hồi quy tuyến tính. Tuy nhiên việc xác định xem các giá



Hình 6: Biểu đồ hình hộp thể hiện phân phối giá trị của biến được ước lượng **PRP**



Hình 7: Biểu đồ hình hộp thể hiện phân phối giá trị của các biến dự đoán

trị đó là outliers hay không là tương đối phức tạp. Bên cạnh đó, nhận thấy các giá trị có tần suất xuất hiện cao tập trung rất gần nhau trong những giá trị có tần suất thấp lại nằm càng ngày càng xa, hay nói cách khác đồ thị bị lệch phải quá nhiều.
Do đó, để cho các giá trị có phân phối chuẩn hơn và giảm sự ảnh hưởng của các điểm đòn bẩy cao để có được mô hình hồi quy tuyến tính tốt hơn, ta sẽ lấy logarithm của các biến. Lưu ý các biến **CACH**, **CHMIN**, **CHMAX** sẽ được cộng 1 trước khi lấy logarithm để tránh trường hợp những biến này nhận giá trị bằng 0, các biến còn lại không bao giờ nhận giá trị 0.

```
#Xu ly du lieu
data$MYCT <- data$MYCT %>% log()
data$MMIN <- data$MMIN %>% log()
data$MMAX <- data$MMAX %>% log()
data$CACH <- (data$CACH + 1) %>% log()
data$CHMIN <- (data$CHMIN + 1) %>% log()
data$CHMAX <- (data$CHMAX + 1) %>% log()
data$PRP <- data$PRP %>% log()
```

2.6.2 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.

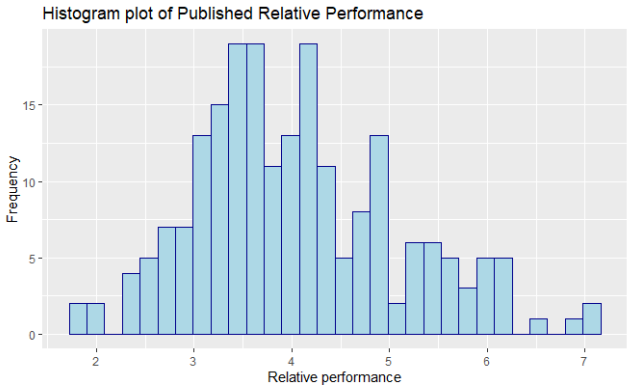
```
#Thong ke don bien
means <- data %>% sapply(mean)
medians <- data %>% sapply(median)
sds <- data %>% sapply(sd)
mins <- data %>% sapply(min)
maxs <- data %>% sapply(max)
uniques <- data %>% sapply(function(col) length(unique(col)))
summary <- as.data.frame(cbind(means, medians, sds, mins, maxs))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

	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

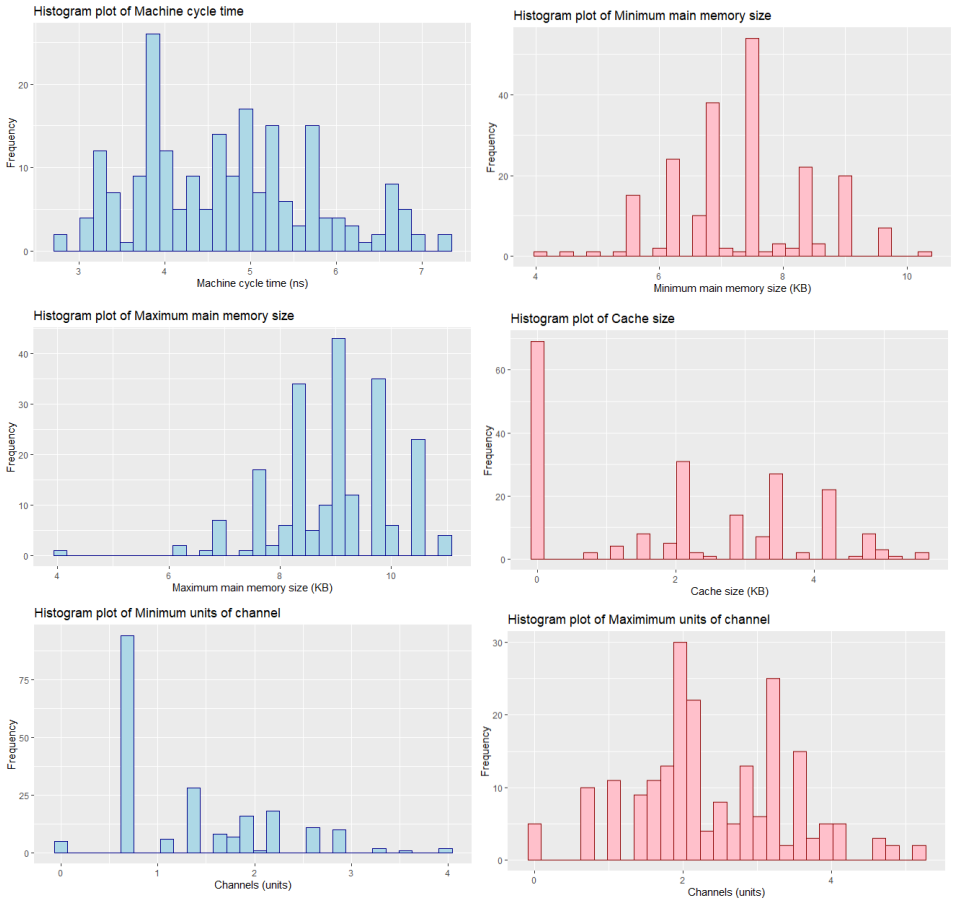
7 rows

Hình 8: Một số thống kê quan trọng

Ngoài các cột thể hiện các giá trị trung bình cộng, trung vị, độ lệch chuẩn, giá trị nhỏ nhất và lớn nhất, còn có thêm một cột thể hiện số giá trị riêng biệt của mỗi biến (cột cuối của Hình 8). Ta có các biểu đồ cột cho từng biến ở Hình 10 và Hình 9 và biểu đồ hình hộp cho từng biến ở Hình 11

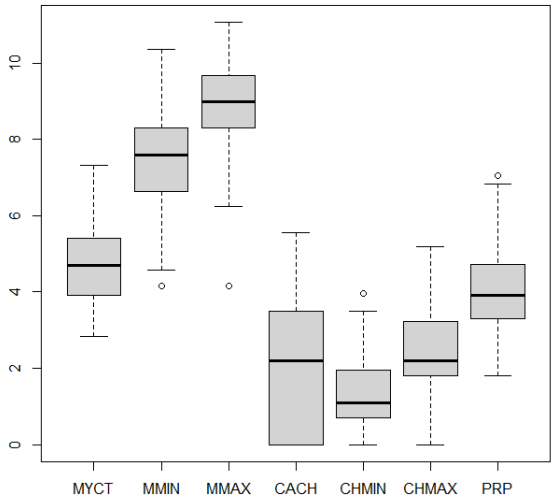


Hình 9: Biểu đồ cột cho biến được ước lượng **PRP**



Hình 10: Biểu đồ cột cho các biến dự đoán

Có thể thấy sau khi được biến đổi, sự phân phối giá trị của các biến dự đoán và biến được ước lượng đã trở nên gần với phân phối chuẩn hơn, đồng thời các điểm đòn bẩy cao ở dạng logarithm



Hình 11: Biểu đồ hình hộp cho các biến sau khi biến đổi dữ liệu

của các biến thể hiện sự chêch lệch ít hơn so với các điểm còn lại do đó giảm thiểu các điểm đòn bẩy cao.

2.6.3 Thống kê đa biến

Ma trận hiệp phương sai giữa các biến được thể hiện ở Hình 12.

```
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 12: Ma trận hiệp phương sai

Ma trận tương quan giữa các biến được thể hiện ở Hình 13.

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

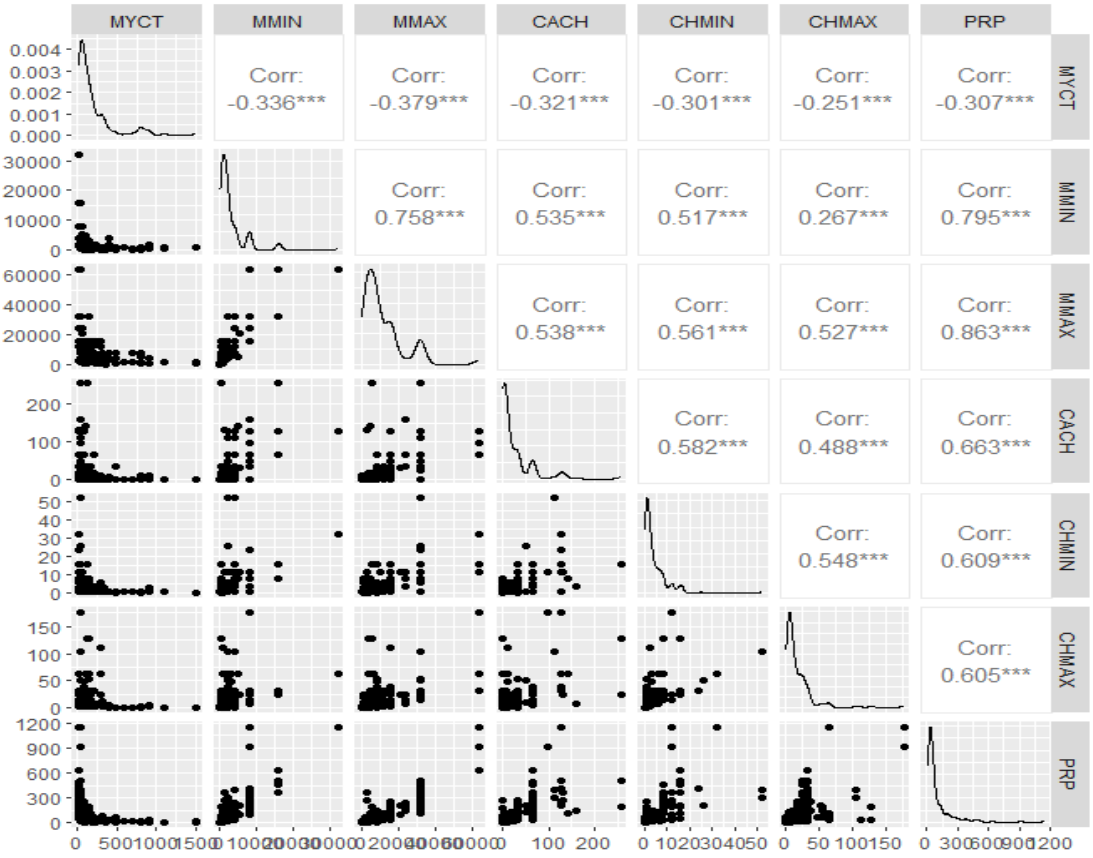
	MYCT <dbl>	MMIN <dbl>	MMAX <dbl>	CACH <dbl>	CHMIN <dbl>	CHMAX <dbl>	PRP <dbl>
MYCT	1.0000000	-0.7294080	-0.6331487	-0.6170887	-0.6179061	-0.5507916	-0.7017927
MMIN	-0.7294080	1.0000000	0.7332816	0.6006968	0.5709069	0.4420004	0.7643858
MMAX	-0.6331487	0.7332816	1.0000000	0.6162918	0.4855683	0.5127990	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 13: Ma trận tương quan

Đồ thị mô tả sự tương quan giữa các cặp biến được vẽ trước và sau khi biến đổi bằng hàm ggpairs() với sự kết hợp của thư viện ggplot2 và pakage mở rộng GGally thể hiện ở Hình 15 và Hình 14

```
library(ggplot2)
library(GGally)
ggpairs(data)
```



Hình 14: Đồ thị phân tán giữa các cặp biến trước khi biến đổi

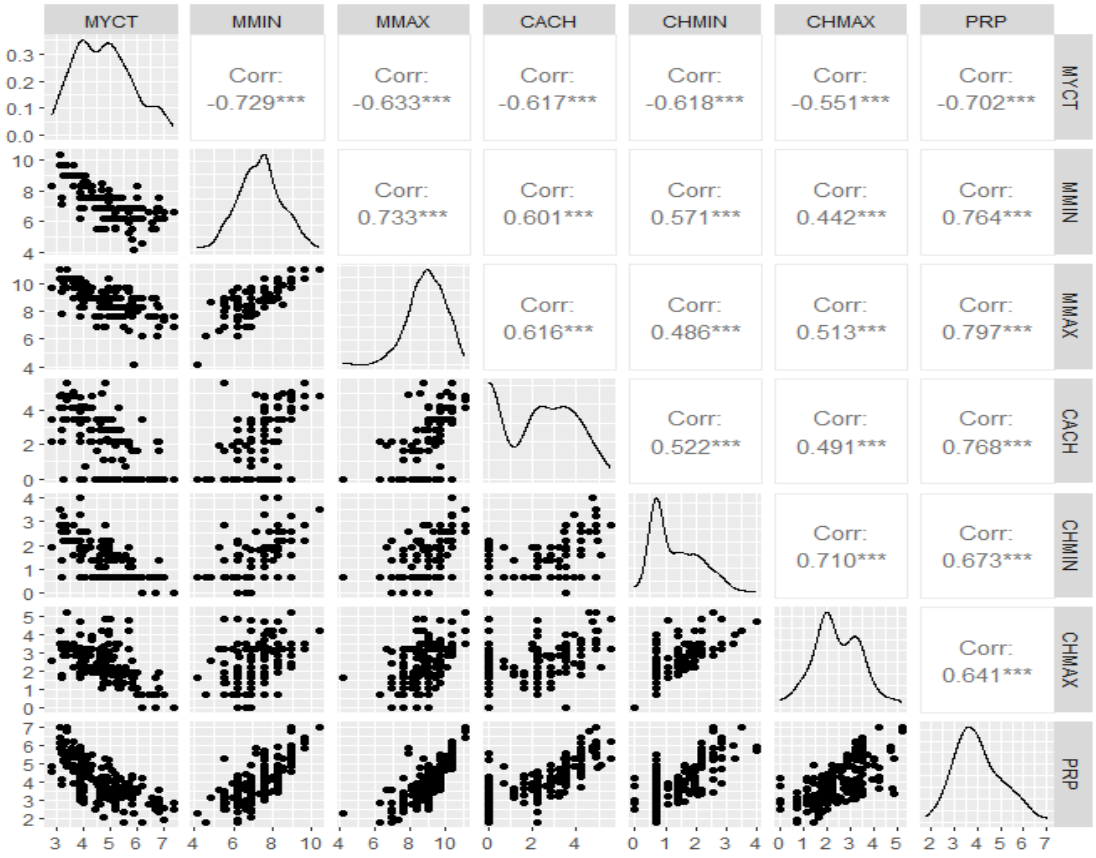
2.7 Mô hình hồi quy tuyến tính

2.7.1 Phân chia tập dữ liệu

Tập dữ liệu hiện tại sẽ được chia thành hai tập dữ liệu khác nhau, một tập huấn luyện (training set) và một tập kiểm tra (testing set) với tỉ lệ được chọn lần lượt là 90% và 10% của tập dữ liệu ban đầu.

```
#Spliting dataset
set.seed(5)
sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE, prob=c(0.9,0.1))
train <- data[sample , ]
test <- data[!sample ,]
```

Khoảng 10% dữ liệu sẽ được lấy ngẫu nhiên vào tập testing data, 90% còn lại sẽ được dùng để xây dựng mô hình hồi quy tuyến tính. do ta có kích thước mẫu là 209, tỉ lệ kích thước giữa tập huấn luyện và tập kiểm tra có được là 186:23 với seed(5). Tập huấn luyện có kích thước 186 sẽ được dùng để xây dựng mô hình ở những bước tiếp theo



Hình 15: Đồ thị phân tán giữa các cặp biến sau khi biến đổi

2.7.2 Xây dựng mô hình

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

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

Ta xây dựng mô hình hồi quy tuyến tính như sau:

```
model = lm(PRP ~ MYCT + MMIN + MMAX + CACH + CHMIN + CHMAX, data)
```

2.7.3 Một số thống kê mẫu

Ở đây, ta có một mẫu kích thước 186. Ta ký hiệu:

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_6^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_6^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(186)} & x_2^{(186)} & \cdots & x_6^{(186)} \end{pmatrix}$$

$$\boldsymbol{Y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(186)} \end{pmatrix}$$

gồm 186 hàng, mỗi hàng tương ứng với 1 phần tử của mẫu.
Tính toán lại các đặc trưng của mẫu ở hình Hình 16 và hiệp phương sai ở Hình 17

	mean	median	sd	min	max	unique
MYCT	4.738022	4.653960	1.0478164	2.833213	7.313220	58
MMIN	7.370196	7.600902	1.1198040	4.158883	10.373491	25
MMAX	8.924448	8.987197	1.0481548	4.158883	11.066638	23
CACH	2.083282	2.197225	1.7048947	0.000000	5.549076	22
CHMIN	1.368344	1.386294	0.8173163	0.000000	3.970292	15
CHMAX	2.416511	2.197225	1.0485477	0.000000	5.176150	30
PRP	4.037431	3.912023	1.0603703	1.791759	7.047517	109

Hình 16: Các đặc trưng quan trọng của mẫu

	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
MYCT	1.0979193	-0.8464740	-0.6838538	-1.1099295	-0.5303318	-0.5939801	-0.7668762
MMIN	-0.8464740	1.2539610	0.8580170	1.1710597	0.5176327	0.5050522	0.9084216
MMAX	-0.6838538	0.8580170	1.0986285	1.0915465	0.4089855	0.5539467	0.8744880
CACH	-1.1099295	1.1710597	1.0915465	2.9066660	0.7420274	0.8601628	1.3872439
CHMIN	-0.5303318	0.5176327	0.4089855	0.7420274	0.6680060	0.6137350	0.5860381
CHMAX	-0.5939801	0.5050522	0.5539467	0.8601628	0.6137350	1.0994522	0.7036458
PRP	-0.7668762	0.9084216	0.8744880	1.3872439	0.5860381	0.7036458	1.1243852

Hình 17: Ma trận hiệp phương sai

- $\overline{x_1} = 4.738022$, $s_{x_1} = 1.0478164$, $S_{x_1x_1} = 203.1150705$
- $\overline{x_2} = 7.370196$, $s_{x_2} = 1.119804$, $S_{x_2x_2} = 231.982785$
- $\overline{x_3} = 8.924448$, $s_{x_3} = 1.0481548$, $S_{x_3x_3} = 203.2462725$

- $\overline{x_4} = 2.083282$, $s_{x_4} = 1.7048947$, $S_{x_4x_4} = 537.73321$
- $\overline{x_5} = 1.368344$, $s_{x_5} = 0.8173163$, $S_{x_5x_5} = 123.58111$
- $\overline{x_6} = 2.416511$ $s_{x_6} = 1.0485477$, $S_{x_6x_6} = 203.398657$
- $\overline{y} = 4.037431$, $s_y = 1.0603703$, $S_{yy} = 208.011262$

2.7.4 Ước lượng tham số

Sử dụng R ta thu được kết quả ước lượng như ở cột 1 Hình 18.

```
summary(model)$coefficients %>% as.data.frame() %>% select(1:2)
```

Như vậy,

$$\begin{aligned}
\hat{\beta}_0 &= -1.390614 \\
\hat{\beta}_1 &= 0.022670 \\
\hat{\beta}_2 &= 0.206128 \\
\hat{\beta}_3 &= 0.313080 \\
\hat{\beta}_4 &= 0.193014 \\
\hat{\beta}_5 &= 0.207103 \\
\hat{\beta}_6 &= 0.133200
\end{aligned}$$

	Estimate	Std. Error
(Intercept)	-1.39061427	0.57929395
MYCT	0.02266975	0.05164862
MMIN	0.20612791	0.05247698
MMAX	0.31307964	0.05086216
CACH	0.19301398	0.02752655
CHMIN	0.20710314	0.06605578
CHMAX	0.13319953	0.04835813

Hình 18: Ước lượng hệ số hồi quy và độ lệch chuẩn

Ngoài ra, ta tính toán thêm được:

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

$$\begin{array}{rcl} SSE & = & 36.608094 \\ SSR & = & 171.403174 \\ \hline SST & = & 208.011268 \end{array}$$

2.7.5 Ước lượng độ lệch chuẩn của sai số

$$\hat{\sigma}^2 = \frac{SSE}{n-p} = \frac{36.608094}{186-7} = 0.204514$$

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:

$$\hat{\sigma} = \sqrt{0.204514} = 0.452233$$

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

```
summary(model)$sigma
```

2.7.6 Xác định hệ số R^2 hiệu chỉnh

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{36.608094/(186-7)}{208.011268/(186-1)} = 0.8181099$$

với n là số quan sát và p là số lượng tham số hồi quy. Thông thường, nếu giải thích độ chính xác của mô hình bằng hệ số R^2 có thể dẫn đến mô hình được đánh giá quá cao (overestimation) do hệ số R^2 sẽ luôn tăng khi ta thêm các biến dự đoán ngay cả khi biến đó không ảnh hưởng đến giá trị dự đoán. Chính vì vậy hệ số R^2 hiệu chỉnh được sử dụng nhằm tránh xảy ra overestimation, khi tăng số lượng biến dự đoán lên nhưng không tạo ra sự khác biệt của mô hình, hệ số R^2 sẽ có xu hướng giảm xuống. Lệnh R sau trả về kết quả xấp xỉ với kết quả ta thu được.

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

2.7.7 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: $F_0 = \dfrac{SSR/6}{SSE/(186-7)} \sim F(6, 179)$
- Miền bác bỏ $D_{RR} = [f_{0.01,6,179}, +\infty)$ với $f_{0.01,6,179} = 2.904$.
- Giá trị của tiêu chuẩn kiểm định với mẫu hiện tại: $f_0 = 139,682990 \in D_{RR}$.

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 $D_{RR} = (-\infty, -t_{\frac{\alpha}{2}}(n-7)) \cup (t_{\frac{\alpha}{2}}(n-7), +\infty)$.

  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 $i = \overline{2,6}$. Đối với giả thuyết $\beta_1 = 0$ ta chưa bác bỏ được. Ở đây, có khả năng biến PRP không phụ thuộc vào biến MYCT, nghĩa là sự có mặt của biến MYCT không những không làm tăng mà có thể còn làm giảm khả năng giải thích của mô hình hồi quy tuyến tính. Trong khi xây dựng mô hình, ta vẫn sẽ coi MYCT là một biến dự đoán, sau đó sẽ loại bỏ biến MYCT ra khỏi mô hình và tính toán lại các giá trị.
Sử dụng R, ta có các giá trị p-value như Hình 19.

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

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

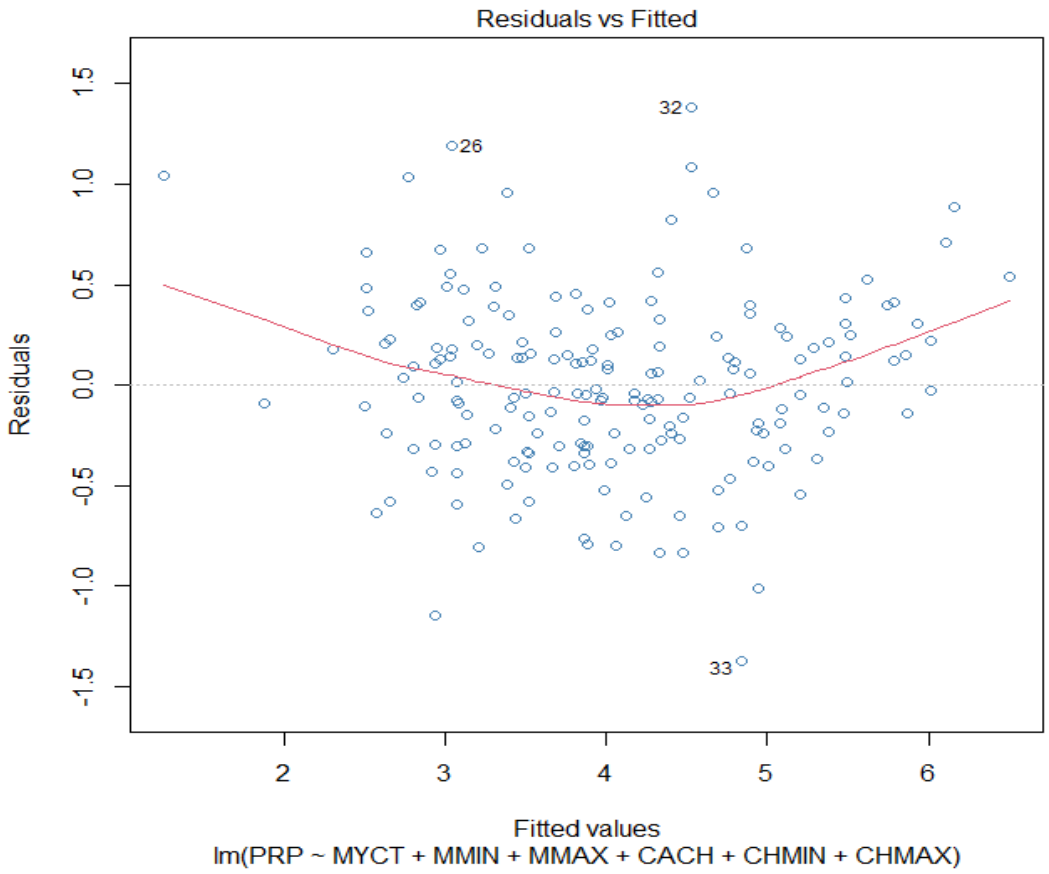
Nhắc lại các giả định của mô hình hồi quy đã được đề cập ở phần cơ sở lý thuyết:

- Tính tuyến tính của dữ liệu
- Sai số có phân phối chuẩn.
- Phương sai của các sai số không đổi, $\epsilon_i \in \mathcal{N}(0, \sigma^2)$.
- Sai số ngẫu nhiên ϵ phân phối độc lập với X_i, $i = \overline{1,6}$.

Tính tuyến tính của dữ liệu:

	t value	Pr(>|t|)
(Intercept)	-2.4005330	1.739608e-02
MYCT	0.4389227	6.612469e-01
MMIN	3.9279685	1.221703e-04
MMAX	6.1554533	4.786983e-09
CACH	7.0119196	4.640582e-11
CHMIN	3.1352765	2.007172e-03
CHMAX	2.7544392	6.486602e-03

Hình 19: Các giá trị p-value của các kiểm định hệ số hồi quy



Hình 20: Đồ thị Residuals vs Fitted

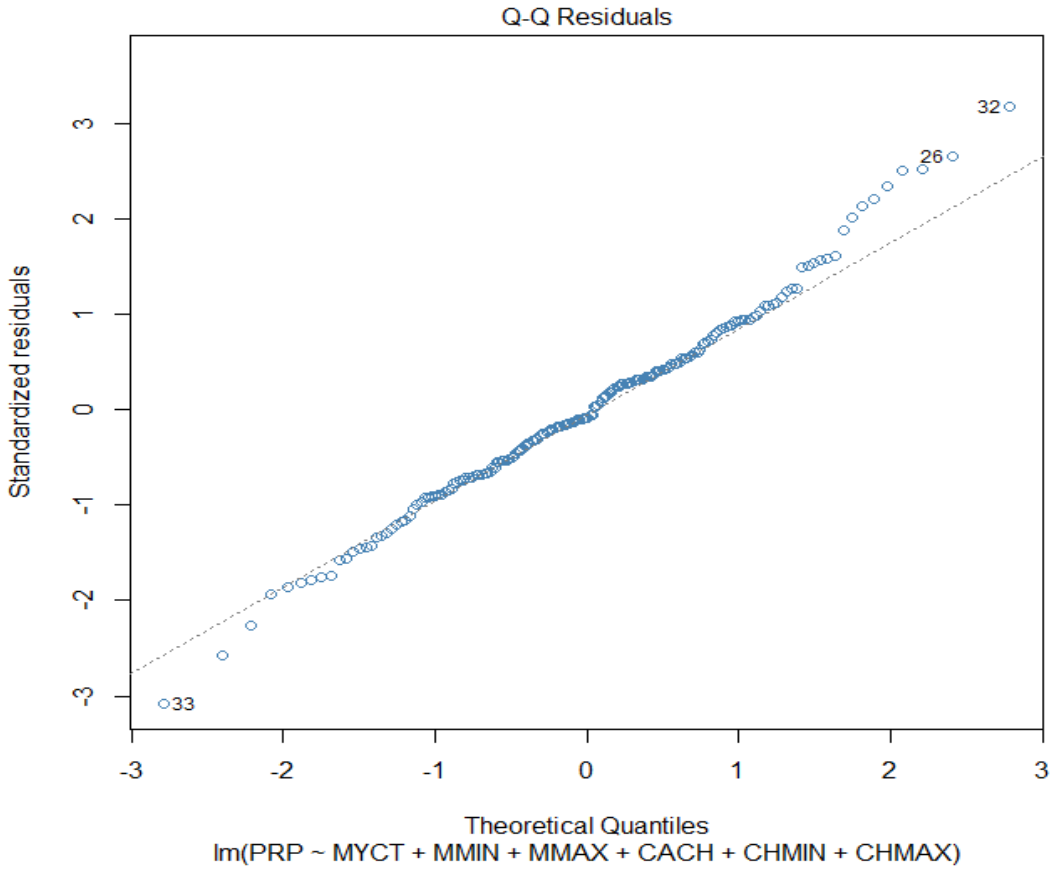
```
plot(model, col = "steel blue", which = 1)
```

Đồ thị Residuals vs Fitted vẽ các giá trị dự báo với các giá trị phần dư (sai số) tương ứng, dùng

để kiểm tra tính tuyến tính của dữ liệu và tính đồng nhất của các phương sai sai số. Nếu như đường thẳng màu đỏ trên đồ thị phân bố theo một hình mẫu đặc trưng nào đó (ví dụ parabol) thì có thể nói là do tính tuyến tính của dữ liệu không thỏa. Ngược lại giả thuyết tính tuyến tính của dữ liệu sẽ thỏa mãn khi đường màu đó nằm ngang. Đối với mô hình hồi quy tuyến tính đang xây dựng, có thể thấy đường màu đó có xu hướng cong, do đó tính tuyến tính của dữ liệu đã bị vi phạm, giả thiết trung bình ở đây được thỏa mãn do đường màu đó nằm khá gần với đường nằm ngang.

Phân phối chuẩn của phần dư

```
plot(model, col = "steel blue", which = 2)
```

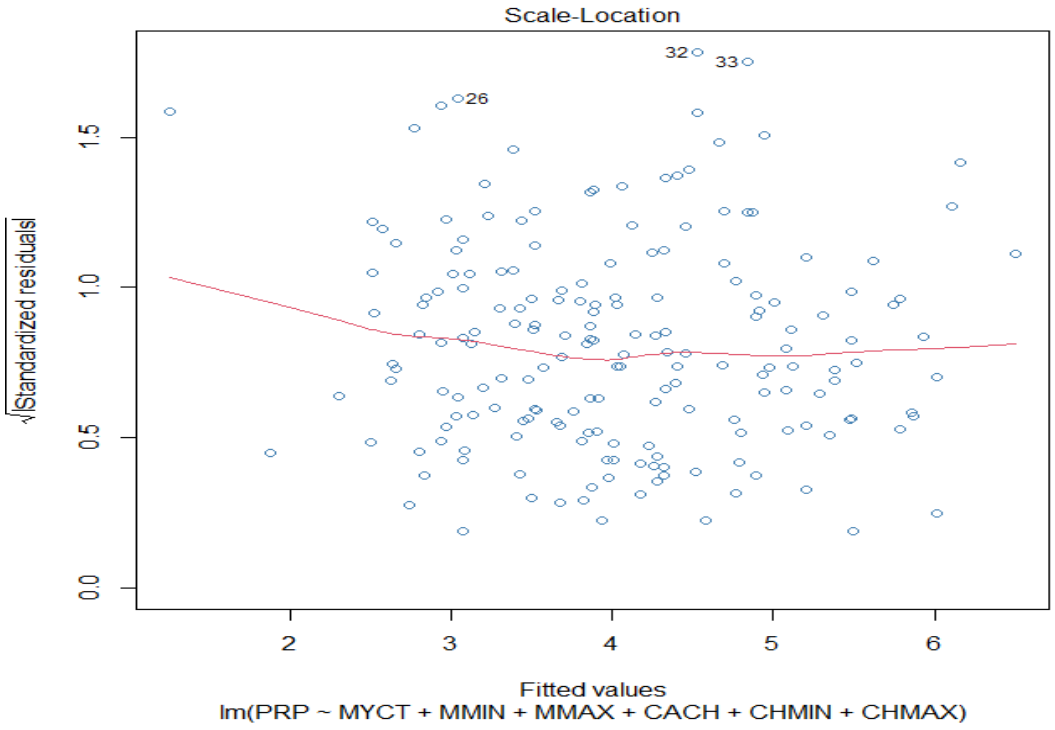


Hình 21: Đồ thị Q-Q Residuals

Đồ thị Q-Q Residuals (Normal Q-Q) dùng để kiểm tra giả thiết phần dư có phân phối chuẩn hay không, nếu các điểm phần dư (residuals) cùng nằm trên một đường thẳng thì điều kiện phân phối chuẩn được thỏa. Có thể thấy ở mô hình hồi quy tuyến tính đã được xây dựng đồ thị Q-Q Residuals như trên có các điểm đều nằm xấp xỉ theo một đường thẳng, do đó giả định này là chấp nhận được.

Phương sai của các sai số là không đổi:

```
plot(model, col = "steel blue", which = 3)
```



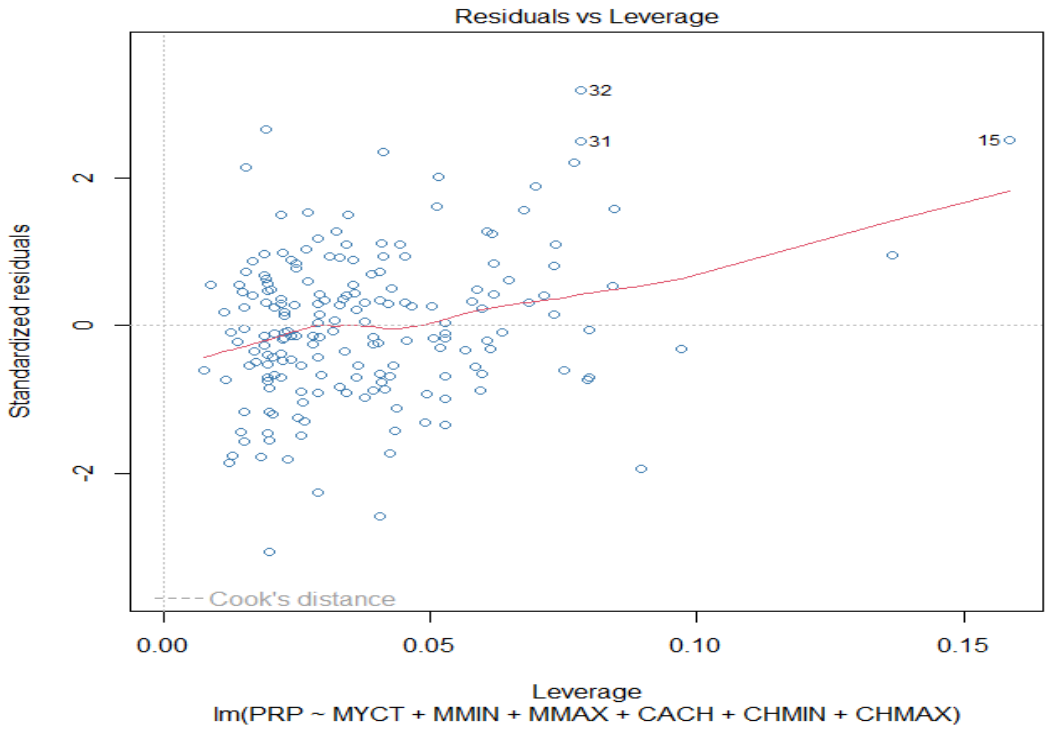
Hình 22: Đồ thị Scale Location

Đồ thị Scale - Location vẽ căn bậc hai của các giá trị phần dư được chuẩn hóa với các giá trị dự báo, được dùng để kiểm tra giả định phương sai của các sai số là không đổi. Nếu như đường màu đỏ trên đồ thị là đường thẳng nằm ngang và các điểm phần dư phân tán đều xung quanh đường thẳng này thì giả định được thỏa. Nếu như đường màu đỏ có độ dốc (hoặc cong) hoặc các điểm phân dư phân tán không đều xung quanh đường thẳng này, thì giả định bị vi phạm. Ở đồ thị Scale-Location của mô hình hồi quy tuyến tính đang xây dựng, đường màu đỏ có bị ảnh hưởng bởi các giá trị dự báo, tuy nhiên ảnh hưởng này là tương đối nhỏ, bên cạnh đó các điểm phần dư phân tán khá đều xung quanh đường màu đỏ này. Do đó, giả thuyết về phương sai của các sai số không đổi là chấp nhận được.

Outliers và các điểm đòn bẩy cao (high leverage points)

```
plot(model, col = "steel blue", which = 5)
```

Cho phép xác định những điểm có ảnh hưởng cao (influential observations), nếu chúng có hiện diện trong bộ dữ liệu. Những điểm có ảnh hưởng cao này có thể là các điểm outliers, là những điểm có thể gây nhiều ảnh hưởng nhất khi phân tích dữ liệu. Nếu như ta quan sát thấy một đường thẳng màu đỏ đứt nét (Cook's distance), và có một số điểm vượt qua đường thẳng khoảng



Hình 23: Đồ thị Residuals Leverage

cách này, nghĩa là các điểm đó là các điểm có ảnh hưởng cao. Nếu như ta chỉ quan sát thấy đường thẳng khoảng cách Cook ở góc của đồ thị và không có điểm nào vượt qua nó, nghĩa không có điểm nào thực sự có ảnh hưởng cao. Ở mô hình hồi quy tuyến tính đang xây dựng có 3 điểm có thể là các điểm có ảnh hưởng lớn là các quan trắc thứ 15, 31 và 32.

Kiểm định giả thuyết giữa các biến độc lập không có mối quan hệ đa cộng tuyến hoàn hảo

Một trong những giả định của mô hình hồi quy tuyến tính cổ điển (CLRM) là không có mối quan hệ tuyến tính chính xác (exact linear relationship) giữa các biến giải thích. Nếu có một hoặc nhiều mối quan hệ như vậy giữa các biến giải thích thì chúng ta gọi ngắn gọn là đa cộng tuyến hoặc cộng tuyến (multicollinearity hoặc collinearity).

Khi các biến giải thích cộng tuyến, suy diễn thống kê trở nên không vững, đặc biệt là khi có cộng tuyến gần hoàn hảo. Bởi vì nếu hai biến có cộng tuyến cao thì rất khó tách biệt tác động riêng của từng biến lên biến phụ thuộc. VIF là **hệ số phóng đại phương sai (variance-inflating factor)**: một thước đo mức độ trong đó phương sai của ước lượng OLS bị phóng đại do cộng tuyến.

$$VIF_j = \frac{1}{1-R_j^2}$$

R_j^2 là hệ số xác định của mô hình hồi qui tuyến tính phụ của biến độc lập X_j theo các biến độc lập còn lại của mô hình. Nếu 2 biến độc lập, thì R = 0, nên VIF = 1, nghĩa là phương sai trong

hồi quy bội sẽ đúng bằng phương sai trong hồi quy đơn. Còn nếu 2 biến cộng tuyến, R gần bằng 1, khi đó VIF sẽ rất lớn, và phương sai của từng hệ số hồi quy sẽ rất lớn. Ta kiểm tra trong R bằng code:

```
vif (model) %>% as.data .frame() %>% select(1:1)
```

MYCT	2.649327
MMIN	3.123700
MMAX	2.570917
CACH	1.992262
CHMIN	2.636633
CHMAX	2.325750

Hình 24: Kiểm định tính đa cộng tuyến

Nhận thấy các giá trị cộng tuyến là không quá lớn (<10) nên hiện tượng cộng tuyến không xảy ra trong mô hình

2.8 Dự đoán hiệu năng tương đối của một CPU

Ở đây chung ta sẽ sử dụng hàm predict() cho tập kiểm tra (testing dataset) mà chúng ta đã phân chia từ trước

```
#Predict
predicts <- predict(model, newdata = test)
actuals <- test$PRP
evaluate <- data.frame(actuals ,predicts)
summary(evaluate)
```

actuals		predicts	
Min.	:2.565	Min.	:2.753
1st Qu.	:3.349	1st Qu.	:3.219
Median	:3.714	Median	:3.787
Mean	:4.036	Mean	:3.965
3rd Qu.	:4.909	3rd Qu.	:4.559
Max.	:6.455	Max.	:5.999

Hình 25: Dự đoán các đặc trưng của tập kiểm tra

Hình 26 và Hình 25 cho thấy kết quả dự đoán của mô hình được xây dựng. Có thể thấy mô hình dự đoán cho ra kết quả xấp xỉ tốt so với dữ liệu chính xác, các đặc trưng cơ bản của kết quả dự đoán và thực tế là tương đối xấp xỉ nhau, đặc biệt là ở các giá trị trung vị, kỳ vọng và các tứ phân vị
Tính Mean Squared Error để đánh giá kết quả dự đoán, ta có giá trị Mean Squared Error (MSE) là:

actuals	predicts
5.393628	5.381623
6.455199	5.998811
3.295837	3.028649
3.295837	3.263119
4.276666	4.157968
4.927254	4.051068
3.583519	3.256974
3.258097	3.573261
4.969813	4.524981
4.969813	4.741992
3.465736	3.451328
3.912023	4.256156
2.772589	3.113264
3.688879	3.510857
3.178054	2.915618
2.564949	2.753182
4.890349	4.833497
3.737670	3.787148
5.356586	5.422715
3.401197	3.108006
3.713572	4.297893
3.465736	3.181842
4.248495	4.592091

Hình 26: Kết quả dự đoán so với thực tế

```
#MSE
MSE <- ((evaluate$actuals evaluate$predicts)^2) %>% sum()
MSE <- MSE/23
```

$$\text{MSE} = 0.109792$$

Giá trị MSE thông thường cho biết mức độ "gần" của đường hồi quy với một tập các điểm. Giá trị này tính bằng cách lấy bình phương khoảng cách từ các điểm đến đường hồi quy. Với giá trị MSE trên thì có thể thấy kết quả dự đoán của mô hình là tương đối xấp xỉ được với giá trị thực tế.

Dựa trên kết quả của tổng hợp ở Hình 27, mô hình hồi quy tuyến tính được xây dựng có giá trị R^2 là 88.31% và giá trị R^2 hiệu chỉnh là 87.75%. Có thể thấy mô hình được xây dựng có thể giải thích được khoảng 88% các giá trị. Độ chính xác của mô hình là tương đối ổn tuy nhiên ta xét lại giả thiết

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.65887 -0.19912 -0.02437  0.21584  0.80471

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02104    0.32654   0.064    0.949
predicts     1.01245    0.08040  12.593 2.98e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3387 on 21 degrees of freedom
Multiple R-squared:  0.8831,    Adjusted R-squared:  0.8775
F-statistic: 158.6 on 1 and 21 DF,  p-value: 2.978e-11
```

Hình 27: Độ chính xác của mô hình

2.9 Loại bỏ MYCT ra khỏi mô hình

Tương tự như những bước để xây dựng và kiểm tra sự phù hợp của mô hình hồi quy tuyến tính, chỉ khác ở chỗ giờ đây mô hình chỉ còn được xây dựng bởi các biến MMIN, MMAX, CACH, CHMIN, CHMAX để dự đoán giá trị của PRP.
Xây dựng mô hình và ước lượng tham số:

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

	Estimate	Std. Error
(Intercept)	-1.1857722	0.34241948
MMIN	0.1973764	0.04843260
MMAX	0.3114599	0.05061423
CACH	0.1905747	0.02689921
CHMIN	0.2020548	0.06490070
CHMAX	0.1305148	0.04786204

Hình 28: Ước lượng hệ số hồi quy và độ lệch chuẩn

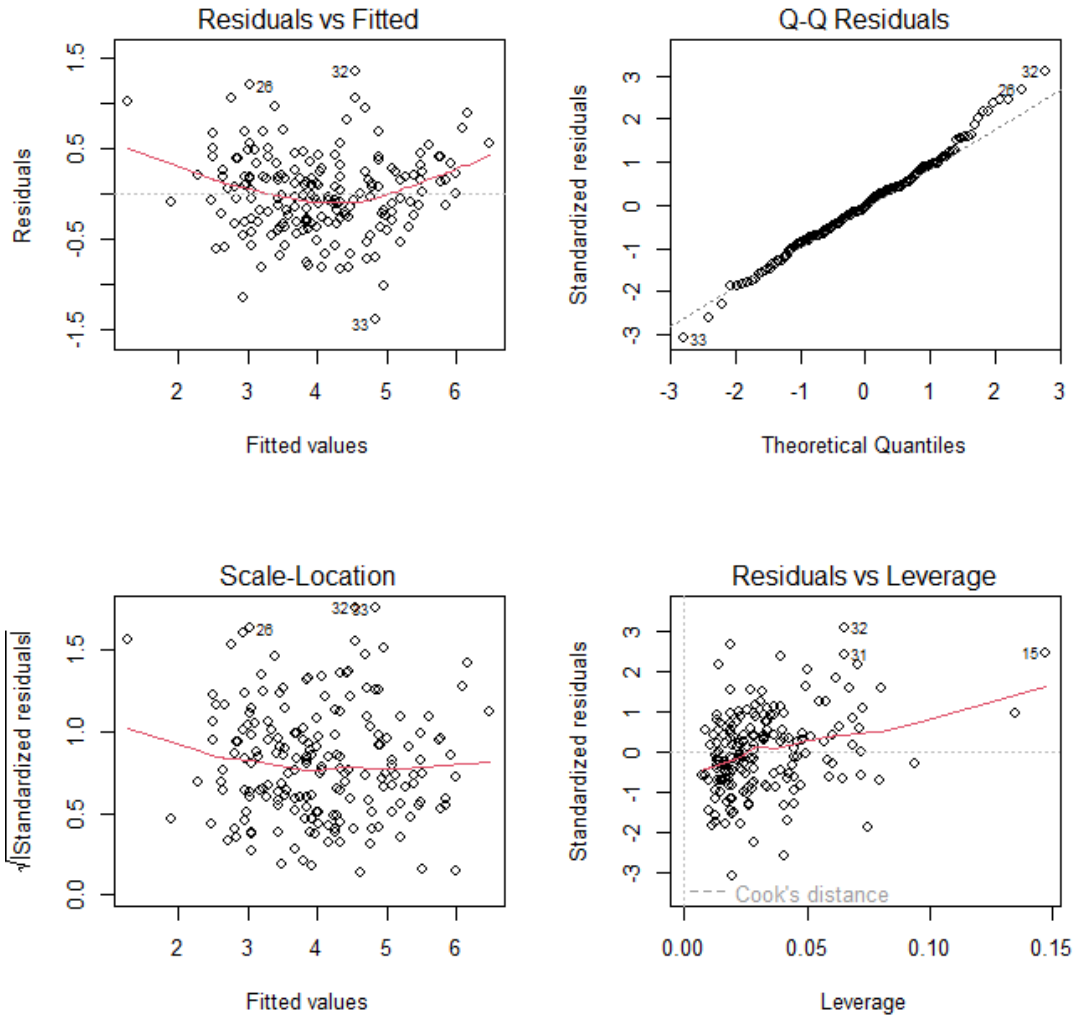
Bên cạnh đó tính toán lại các giá trị:

$$\begin{aligned} SSE &= 36.647495 \\ SSR &= 171.363773 \\ \hline SST &= 208.011268 \end{aligned}$$

Hệ số R^2 hiệu chỉnh:

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{36.647495/(186-6)}{208.011268/(186-1)} = 0.818926$$

Có thể thấy, các giá trị ước lượng hệ số hồi quy, độ lệch chuẩn SSE, SSR, SST gần như không có sự thay đổi khi ta loại bỏ biến MYCT, trong khi đó hệ số R^2 hiệu chỉnh lại tăng từ 0.8181099 lên 0.8189258 khi loại bỏ MYCT. Do đó tới đây ta có thể chấp nhận biến MYCT không ảnh hưởng đến biến được dự đoán PRP.
Tiếp tục đến với các đồ thị kiểm định mô hình hồi quy tuyến tính và kiểm định đa cộng tuyến ở Hình 29 và Hình 30.



Hình 29: Kiểm định mô hình hồi quy tuyến tính

MMIN	2.672758
MMAX	2.557384
CACH	1.911059
CHMIN	2.556696
CHMAX	2.288541

Hình 30: Kiểm định tính đa cộng tuyến

Hầu như không có sự khác biệt ở các yếu tố kiểm định khi ta loại bỏ biến MYCT ra khỏi mô hình. Cuối cùng là kết quả dự đoán ở Hình 31 và các đặc trưng của tập kiểm tra Hình 32.

actuals	predicts
5.393628	5.385667
6.455199	5.996057
3.295837	3.030082
3.295837	3.262222
4.276666	4.156780
4.927254	4.053620
3.583519	3.260246
3.258097	3.584298
4.969813	4.527354
4.969813	4.743242
3.465736	3.461138
3.912023	4.235126
2.772589	3.103354
3.688879	3.509696
3.178054	2.918937
2.564949	2.723452
4.890349	4.860607
3.737670	3.803540
5.356586	5.421285
3.401197	3.111029
3.713572	4.283056
3.465736	3.189148
4.248495	4.581480

Hình 31: Kết quả dự đoán so với thực tế

```
    actuals         predicts
Min.   :2.565   Min.   :2.723
1st Qu.:3.349   1st Qu.:3.225
Median :3.714   Median :3.804
Mean   :4.036   Mean   :3.965
3rd Qu.:4.909   3rd Qu.:4.554
Max.   :6.455   Max.   :5.996
```

Hình 32: Các đặc trưng của kết quả dự đoán so với thực tế

Đồng thời ta cũng tính toán được giá trị MSE = 0.107033, giá trị này cho thấy kết quả dự đoán có sự sai số ít hơn so với khi xây dựng mô hình hồi quy tuyến tính có sự xuất hiện của biến MYCT. Độ chính xác cũng được cải thiện với hệ số R^2 và R^2 hiệu chỉnh đều tăng.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.64398 -0.18341 -0.02785  0.21165  0.80207

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01984    0.32175   0.062    0.951
predicts     1.01276    0.07922  12.785 2.24e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3341 on 21 degrees of freedom
Multiple R-squared:  0.8862,    Adjusted R-squared:  0.8807
F-statistic: 163.5 on 1 and 21 DF,  p-value: 2.244e-11
```

Hình 33: Độ chính xác của mô hình

Kết luận: việc có hay không sự xuất hiện của biến MYCT không làm thay đổi quá nhiều kết quả dự đoán của mô hình hồi quy tuyến tính, do đó có thể coi như biến PRP không phụ thuộc vào biến MYCT. Hơn nữa, việc thêm biến MYCT vào mô hình thậm chí còn có thể làm giảm độ chính xác của mô hình dự đoán.

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