
SUY LUẬN THỐNG KÊ

BÁO CÁO CUỐI KỲ

SINH VIÊN THỰC HIỆN: NHÓM 26

NGUYỄN ĐÌNH TUẤN LONG - 20216941

NGUYỄN VĂN TUẤN - 20216965

VŨ VĂN HUY - 20216931

GIẢNG VIÊN:

PGS.TS. NGUYỄN THỊ THU THỦY

BỘ MÔN TOÁN ỨNG DỤNG

SAMI – HUST

MỤC LỤC

Chương 1. THỰC HÀNH LÀM VIỆC VỚI DỮ LIỆU TRÊN PHẦN MỀM R	4
1.1 Nhập dữ liệu (trực tiếp và ghi nhập dữ liệu với file)	4
1.1.1 Thực hiện nhập dữ liệu từ file	4
1.1.2 Thực hiện thêm dữ liệu	5
1.2 Thao tác với dữ liệu (chiết, xuất dữ liệu, ghép nối dữ liệu, chia nhóm dữ liệu).	5
1.2.1 Chiết dữ liệu	5
1.2.2 Ghép nối dữ liệu	6
1.2.3 Chia nhóm dữ liệu	6
1.2.4 Xuất dữ liệu	7
1.3 Lập bảng tần số, bảng chia khoảng trong R.	8
1.3.1 Lập bảng tần số	8
1.3.2 Lập bảng chia khoảng trong R.	8
1.4 Vẽ các loại biểu đồ trong R.	10
1.4.1 Histogram cho dữ liệu ban đầu với dữ liệu đã chia khoảng	10
1.4.2 Biểu đồ cột	11
1.4.3 Biểu đồ scatter plot cho mối quan hệ giữa hai biến:	12
1.5 Tính các đặc trưng mẫu trong R.	13
1.5.1 Tính trung bình mẫu	13
1.5.2 Tính phương sai mẫu	14
1.5.3 Tính độ lệch chuẩn mẫu	14
1.5.4 Tính các đặc trưng mẫu trong R bằng 1 hàm	15
Chương 2. Kiểm định giả thuyết thống kê	16
2.1 Bài toán	16
2.2 Nội dung	16
Chương 3. XÂY DỰNG MÔ HÌNH HỒI QUY TUYẾN TÍNH BỘI	22
3.1 Mô tả bài toán và bộ dữ liệu	22
3.2 Phương pháp thực hiện	24
3.3 Đánh giá mô hình	28

Chương 4. Kết luận cuối cùng	29
-------------------------------------	-----------

Tài liệu tham khảo	30
---------------------------	-----------

Lời mở đầu

Suy luận thống kê là một lĩnh vực hấp dẫn và quan trọng của thống kê, trong đó ta sử dụng các dữ liệu thu thập được từ một mẫu nhỏ để suy ra các đặc điểm của một tổng thể lớn. Suy luận thống kê có thể giúp ta đưa ra các kết luận về một hiện tượng, dự đoán một xu hướng, kiểm tra một giả thuyết và ước lượng một khoảng tin cậy cho các tham số quan tâm. Suy luận thống kê có vai trò quan trọng trong nhiều lĩnh vực như khoa học, kỹ thuật, kinh tế, y tế, giáo dục và xã hội, bởi vì nó cho phép ta khai thác các thông tin có giá trị từ các dữ liệu có sẵn. Suy luận thống kê cần phải tuân theo các nguyên tắc và phương pháp khoa học để đảm bảo tính chính xác và tin cậy của các kết quả.

Qua học phần Suy luận thống kê MI3031, chúng em đã được trang bị những kiến thức lý thuyết và kỹ năng tính toán về mẫu thống kê nhằm phân tích các số liệu về các lĩnh vực. Không chỉ được cung cấp về lý thuyết về thống kê, chúng em còn được trang bị thêm những kỹ năng cơ bản sử dụng phần mềm R. Đã biết cách vận dụng phần mềm R để giải quyết các vấn đề liên quan đến môn học.

Cuối cùng, bọn em muốn gửi lời cảm ơn đến cô Nguyễn Thị Thu Thủy, người đã truyền tải tất cả những kiến thức Suy luận thống kê một cách rất dễ hiểu. Nhờ cô mà bọn em đã hiểu thống kê quan trọng như thế nào trong các lĩnh vực. Ngoài kỹ năng tính toán đã được cô cung cấp, cô còn tạo điều kiện để bọn em làm việc nhóm, thuyết trình, làm bọn em có kinh nghiệm hơn trong làm việc nhóm và báo cáo khoa học sau này ạ.

Chương 1

THỰC HÀNH LÀM VIỆC VỚI DỮ LIỆU TRÊN PHẦN MỀM R

1.1 Nhập dữ liệu (trực tiếp và ghi nhập dữ liệu với file)

1.1.1 Thực hiện nhập dữ liệu từ file

- Mã nguồn thực hiện nhập file

```
1 data <- read.csv("D:\\Downloads\\Suy_luan_thong_ke\\penguins.csv")
2 data
```

- Kết quả thực hiện:

```
> data <- read.csv("D:\\Downloads\\Suy_luan_thong_ke\\penguins.csv")
> data
```

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
1	Adelie	Torgersen	39.1	18.7	181	3750	male
2	Adelie	Torgersen	39.5	17.4	186	3800	female
3	Adelie	Torgersen	40.3	18.0	195	3250	female
4	Adelie	Torgersen	NA	NA	NA	NA	<NA>
5	Adelie	Torgersen	36.7	19.3	193	3450	female
6	Adelie	Torgersen	39.3	20.6	190	3650	male
7	Adelie	Torgersen	38.9	17.8	181	3625	female
8	Adelie	Torgersen	39.2	19.6	195	4675	male
9	Adelie	Torgersen	34.1	18.1	193	3475	<NA>
10	Adelie	Torgersen	42.0	20.2	190	4250	<NA>
11	Adelie	Torgersen	37.8	17.1	186	3300	<NA>
12	Adelie	Torgersen	37.8	17.3	180	3700	<NA>
13	Adelie	Torgersen	41.1	17.6	182	3200	female

1.1.2 Thực hiện thêm dữ liệu

- Mã nguồn thực hiện việc thêm dữ liệu:

```
1 newly_added_row <- subset(data, species == "Adelie" & island == "
  Torgersen" & bill_length_mm == 40)
2
3 # Thêm dòng mới
4 data <- rbind(data, new_row)
5 newly_added_row <- subset(data, species == "Adelie", island == "Torgersen",
  bill_length_mm == 40)
6 # In ra dòng vừa thêm
7 print(newly_added_row)
```

- Kết quả thực hiện:

```
> newly_added_row <- subset(data, species == "Adelie" & island == "Torgersen" & bill_length_mm == 40)
> print(newly_added_row)
  species island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex
345 Adelie Torgersen          40          18.7             181        3750 male
```

Hình 1.1: In ra dòng vừa được thêm vào

1.2 Thao tác với dữ liệu (chiết, xuất dữ liệu, ghép nối dữ liệu, chia nhóm dữ liệu).

1.2.1 Chiết dữ liệu

- Mã nguồn thực hiện chiết dữ liệu

```
1 adelie_penguins <- subset(data, species == "Adelie")
2 adelie_penguins
```

- Kết quả

```
> adelic_penguins <- subset(data, species == "Adelie")
> adelic_penguins
```

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
1	Adelie	Torgersen	39.1	18.7	181	3750	male
2	Adelie	Torgersen	39.5	17.4	186	3800	female
3	Adelie	Torgersen	40.3	18.0	195	3250	female
4	Adelie	Torgersen	NA	NA	NA	NA	<NA>
5	Adelie	Torgersen	36.7	19.3	193	3450	female
6	Adelie	Torgersen	39.3	20.6	190	3650	male
7	Adelie	Torgersen	38.9	17.8	181	3625	female
8	Adelie	Torgersen	39.2	19.6	195	4675	male
9	Adelie	Torgersen	34.1	18.1	193	3475	<NA>
10	Adelie	Torgersen	42.0	20.2	190	4250	<NA>
11	Adelie	Torgersen	37.8	17.1	186	3300	<NA>
12	Adelie	Torgersen	37.8	17.3	180	3700	<NA>
13	Adelie	Torgersen	41.1	17.6	182	3200	female

Hình 1.2: Chiết những dữ liệu có tên loài là Adelie

1.2.2 Ghép nối dữ liệu

- Mã nguồn ghép nối dữ liệu

```
1 merged_data <- merge(adelic_penguins, data, by = "species")
2 merged_data
3 n <- nrow(data)
4 m <- nrow(merged_data)
5 print(n) # In số hàng của dataframe cũ
6 print(m) # In số hàng
```

- Kết quả:

```
> print(n)
[1] 344
> print(m)
[1] 23104
```

Hình 1.3: Ta thấy số lượng hàng ở dataframe mới nhiều hơn số lượng hàng ở dataframe cũ sau khi thực hiện ghép nối

1.2.3 Chia nhóm dữ liệu

- Mã nguồn

```
1 library(dplyr)
2 # Chia nhóm dữ liệu theo biến "species"
3 penguins_grouped <- data %>% group_by(species)
```

```
4 penguins_grouped
```

```
5
```

- Kết quả thực hiện:

```
R 4.3.1 · D:/Downloads/Suy_luan_thong_ke/ ↗
# Groups:   species [3]
#   species island  bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex
#   <chr>   <chr>      <dbl>         <dbl>         <dbl>         <dbl> <chr>
1 Adelie  Torgersen    39.1           18.7           181           3750 male
2 Adelie  Torgersen    39.5           17.4           186           3800 female
3 Adelie  Torgersen    40.3            18           195           3250 female
4 Adelie  Torgersen    NA              NA              NA              NA NA
5 Adelie  Torgersen    36.7           19.3           193           3450 female
6 Adelie  Torgersen    39.3           20.6           190           3650 male
7 Adelie  Torgersen    38.9           17.8           181           3625 female
8 Adelie  Torgersen    39.2           19.6           195           4675 male
9 Adelie  Torgersen    34.1           18.1           193           3475 NA
10 Adelie Torgersen    42             20.2           190           4250 NA
# i 336 more rows
# i Use `print(n = ...)` to see more rows
```

Hình 1.4: Thực hiện nhóm những dữ liệu có chung loài

1.2.4 Xuất dữ liệu

```
1 # Đường dẫn đến nơi bạn muốn lưu tệp tin
2 duong_dan_cu_the <- "D:\\Downloads\\Suy_luan_thong_ke\\penguins.csv"
3
4 # Xuất dữ liệu
5 write.csv(adelie_penguins, file = duong_dan_cu_the, row.names = FALSE)
```


1.3 Lập bảng tần số, bảng chia khoảng trong R.

1.3.1 Lập bảng tần số

- Mã nguồn

```
1 # Tạo bảng tần số cho biến "species"
2 freq_table <- table(penguins$species)
3
4 # In bảng tần số
5 print(freq_table)
```

- Kết quả:

```
> print(freq_table)
      Adelie Chinstrap      Gentoo
      152         68       124
> |
```

1.3.2 Lập bảng chia khoảng trong R.

- Mã nguồn

```
1 # Xác định các khoảng giá trị cho biến "bill_length_mm"
2 breaks <- c(30, 40, 50, 60, 70, 80) # Đây là các khoảng bạn muốn xác đị
  nh
3
4 # Lập bảng chia khoảng và đếm tần suất
5 data %>%
6   mutate(bill_length_group = cut(bill_length_mm, breaks)) %>%
7   group_by(bill_length_group) %>%
8   summarise(count = n())
```

- Kết quả:

	bill_length_group	count
	<i><fct></i>	<i><int></i>
1	(30,40]	102
2	(40,50]	190
3	(50,60]	52
4	NA	2

Hình 1.5: Bảng chia khoảng cho kích thước của mỏ chim cánh cụt

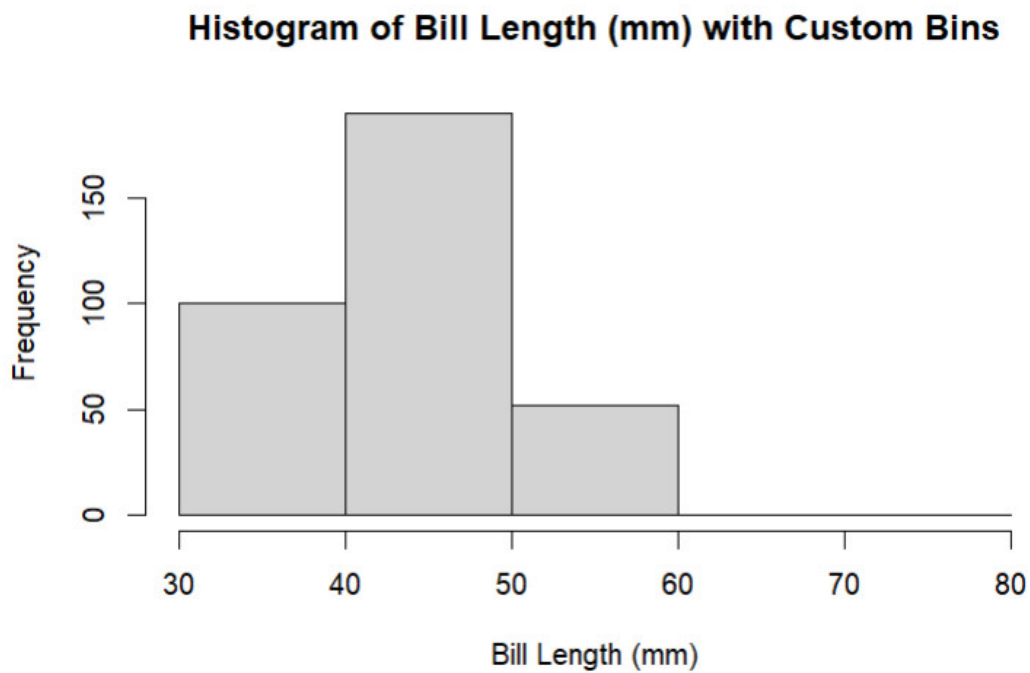
1.4 Vẽ các loại biểu đồ trong R.

1.4.1 Histogram cho dữ liệu ban đầu với dữ liệu đã chia khoảng

- Mã nguồn:

```
1 hist(data\bill_length_mm, breaks = c(30, 40, 50, 60, 70, 80),  
2     main = "Histogram of Bill Length (mm) with Custom Bin",  
3     xlab = "Bill Length (mm)")
```

- Kết quả:



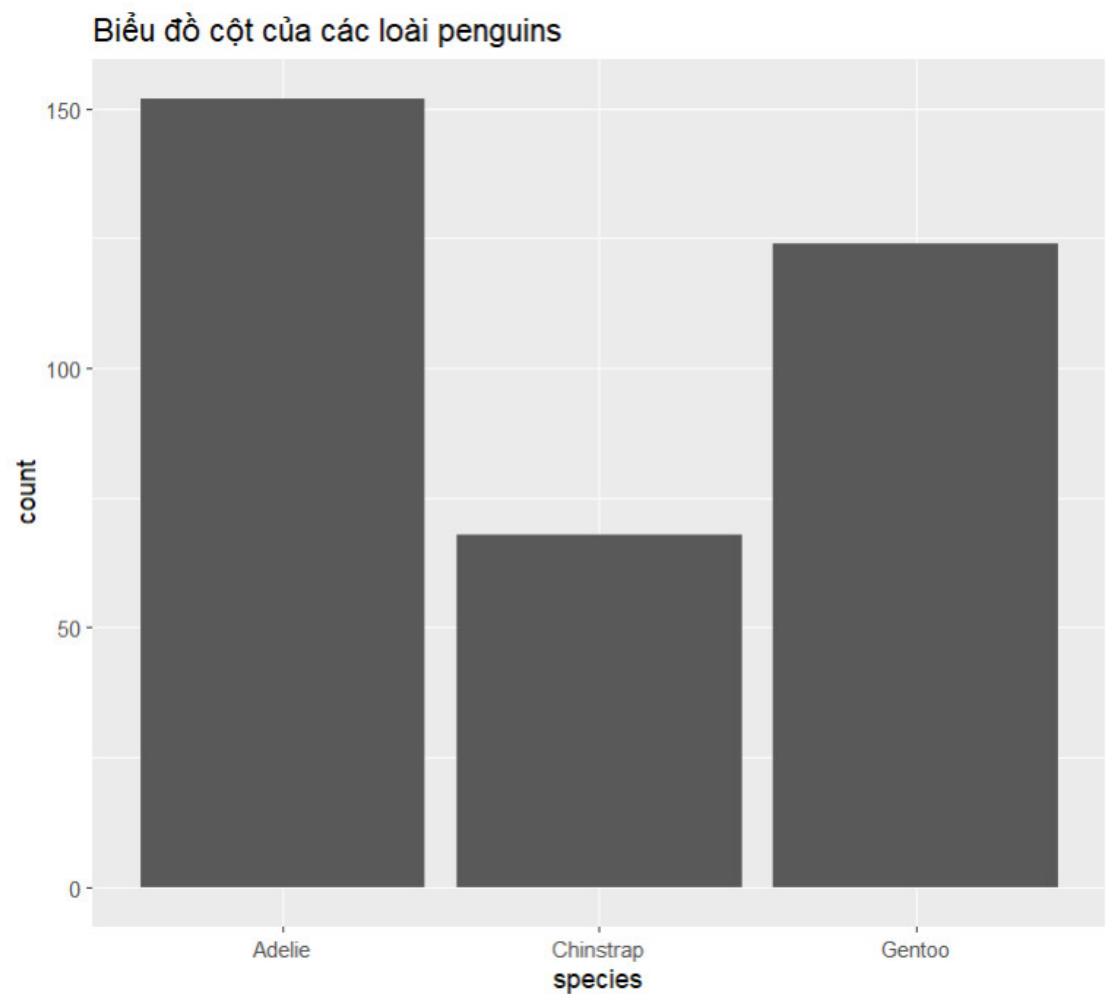
Hình 1.6: Biểu đồ thể hiện sự phân bố của dữ liệu dựa vào kích thước mỏ chim cánh cụt

1.4.2 Biểu đồ cột

- Mã nguồn:

```
1 # Tạo biểu đồ cột
2 ggplot(penguins, aes(x = species)) +
3   geom_bar() +
4   ggtitle("Biểu đồ cột của các loài penguins")
```

- Kết quả:

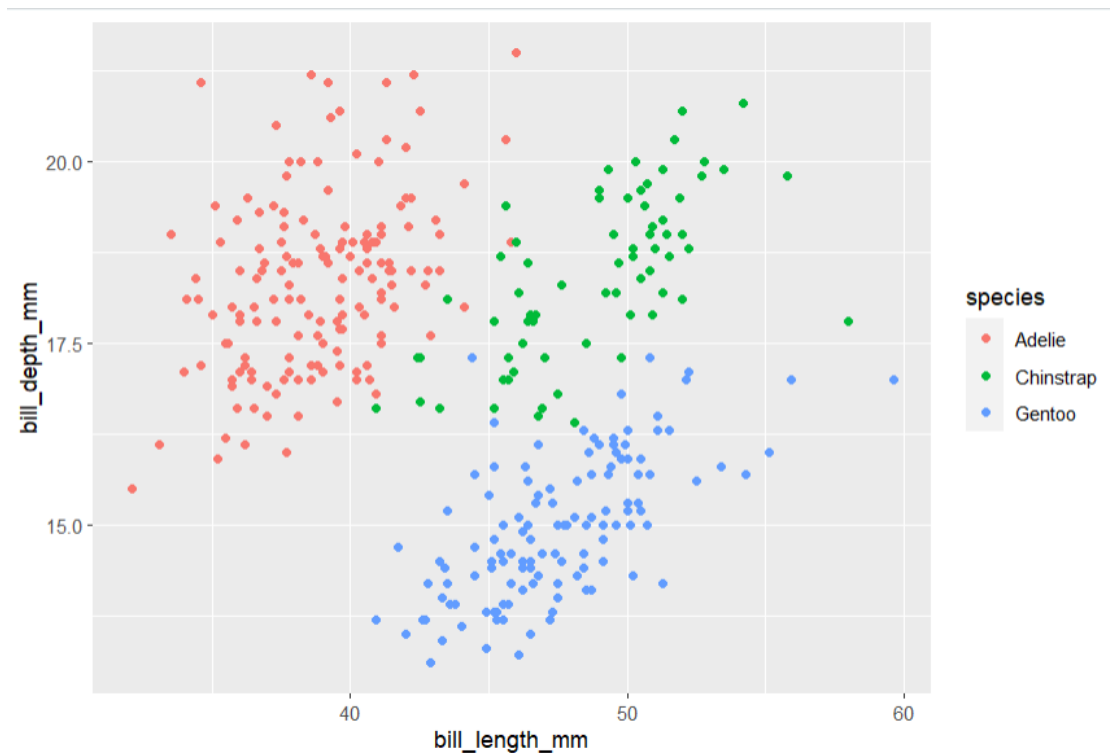


1.4.3 Biểu đồ scatter plot cho mối quan hệ giữa hai biến:

- Mã nguồn:

```
1 # Biểu đồ scatter plot
2 ggplot(data, aes(x = bill\_length\_mm, y = bill\_depth\_mm, color =
  species)) + geom\_point()
```

- Kết quả:



Hình 1.7: Trực quan dữ liệu của các loài

1.5 Tính các đặc trưng mẫu trong R.

1.5.1 Tính trung bình mẫu

- Mã nguồn:

```
1 # Tính trung bình mẫu của biến "bill_length_mm" trong bộ dữ liệu penguins  
  sử dụng công thức  
2 data1 <- data$bill_length_mm  
3  
4 # Tính tổng của các giá trị hợp lệ (loại bỏ NA)  
5 total <- sum(data1, na.rm = TRUE)  
6  
7 # Tính số lượng giá trị hợp lệ  
8 count <- sum(!is.na(data1))  
9  
10 # Tính trung bình mẫu  
11 mean_value <- total / count  
12  
13 # In trung bình mẫu  
14 print(mean_value)
```

- Kết quả:

```
> # In trung bình mẫu  
> print(mean_value)  
[1] 43.92193  
> |
```

Hình 1.8: Tính trung bình mẫu của biến "bill_length_mm" trong bộ dữ liệu

1.5.2 Tính phương sai mẫu

- Mã nguồn:

```
1 data1_cleaned <- na.omit(data1)
2 # Tính phương sai
3 variance_value <- sum((data1_cleaned - mean_value)^2) / (length(data1) -
  1)
4
5 # In phương sai
6 print(variance_value)
```

- Kết quả:

```
> data1_cleaned <- na.omit(data1)
> # Tính phương sai
> variance_value <- sum((data1_cleaned - mean_value)^2) / (length(data1) - 1)
> # In phương sai
> print(variance_value)
[1] 29.63325
```

1.5.3 Tính độ lệch chuẩn mẫu

- Mã nguồn:

```
1 # Tính độ lệch chuẩn
2 std_deviation_value <- sqrt(variance_value)
3
4 # In độ lệch chuẩn
5 print(std_deviation_value)
```

- Kết quả:

```
> # Tính độ lệch chuẩn
> std_deviation_value <- sqrt(variance_value)
> # In độ lệch chuẩn
> print(std_deviation_value)
[1] 5.443643
```

1.5.4 Tính các đặc trưng mẫu trong R bằng 1 hàm

```
1 summary(data)
```

- Kết quả

```
> summary(dabien1)
```

Call:

```
lm(formula = data$Close ~ data$High + data$Low)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-4.0724	-0.8740	-0.0407	0.8956	6.1237

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.42201	0.27602	1.529	0.127
data\$High	0.45617	0.03151	14.478	<2e-16 ***
data\$Low	0.54186	0.03153	17.186	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.35 on 750 degrees of freedom

Multiple R-squared: 0.9976, Adjusted R-squared: 0.9976

F-statistic: 1.554e+05 on 2 and 750 DF, p-value: < 2.2e-16

```
> |
```

Hình 1.9: Tính toán các đặc trưng mẫu của tất cả các biến trong dữ liệu

Chương 2

Kiểm định giả thuyết thống kê

2.1 Bài toán

Lấy ví dụ về một bài toán kiểm định giả thuyết về tham số của một tổng thể, xây dựng công thức và tính xác suất mắc sai lầm loại I, sai lầm loại II trong các trường hợp:

- Kích thước mẫu khác nhau.
- Điểm tới hạn khác nhau
- Giá thực của tham số khác nhau... Từ đó, rút ra nhận xét về ảnh hưởng của các yếu tố tới xác suất mắc sai lầm loại I, sai lầm loại II.

2.2 Nội dung

Giả sử ta muốn kiểm tra xem trung bình chiều cao của nam sinh viên tại trường Đại học Quốc gia Hà Nội có bằng 170 cm hay không. Ta lấy một mẫu ngẫu nhiên gồm 100 nam sinh viên và tính được trung bình mẫu là 168 cm, độ lệch chuẩn mẫu là 10 cm. Ta giả sử rằng chiều cao của nam sinh viên có phân phối chuẩn.

Ta đặt giả thuyết không $H_0 : \mu = 170$ và giả thuyết đối $H_1 : \mu \neq 170$. Ta chọn mức ý nghĩa là 5%. Ta sử dụng công thức sau để tính giá trị kiểm định:

$$z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Trong đó, \bar{x} là trung bình mẫu, μ_0 là giá trị kiểm định của trung bình, s là độ lệch chuẩn mẫu, n là kích thước mẫu.

Thay các giá trị vào công thức, ta được:

$$z = \frac{168 - 170}{\frac{10}{\sqrt{100}}} = -2$$

Ta so sánh giá trị này với điểm tới hạn của phân phối chuẩn tiêu chuẩn hai phía với mức ý nghĩa 5%, là -1.96 và 1.96. Vì $|z| > 1.96$, ta bác bỏ giả thuyết không và kết luận rằng trung bình chiều cao của nam sinh viên tại trường Đại học Quốc gia Hà Nội khác 170 cm.

Xác suất mắc sai lầm loại I trong bài toán này là $\alpha = 0.05$, tức là xác suất ta bác bỏ H_0 khi nó đúng. Xác suất mắc sai lầm loại II trong bài toán này là β , tức là xác suất ta chấp nhận H_0 khi nó sai. Để tính β , ta cần biết giá trị thực của tham số μ , tức là trung bình chiều cao thực sự của nam sinh viên tại trường Đại học Quốc gia Hà Nội. Giả sử μ có giá trị là 165 cm, ta có thể tính β như sau:

$$\beta = P(Z < z_1 | \mu = 165) + P(Z > z_2 | \mu = 165)$$

Trong đó, z_1 và z_2 là các điểm tới hạn của vùng chấp nhận H_0 , được tính như sau:

$$z_1 = \frac{\mu_0 - \mu}{\frac{s}{\sqrt{n}}} - z_{\alpha/2} = \frac{170 - 165}{\frac{10}{\sqrt{100}}} - 1.96 = -0.04$$

$$z_2 = \frac{\mu_0 - \mu}{\frac{s}{\sqrt{n}}} + z_{\alpha/2} = \frac{170 - 165}{\frac{10}{\sqrt{100}}} + 1.96 = 3.96$$

Sử dụng bảng phân phối chuẩn tiêu chuẩn, ta có:

$$\beta = P(Z < -0.04) + P(Z > 3.96) \approx 0.0158 + 0.00004 \approx 0.01584$$

Vậy xác suất mắc sai lầm loại II trong bài toán này khi $\mu = 165$ là khoảng 1.58%.

- **Trường hợp 1:** $n = 50$, $z_{\alpha/2} = 1.96$, $\mu = 165$. Đây là trường hợp ta giảm kích thước mẫu, giữ nguyên điểm tới hạn và giá trị thực của tham số. Ta có:

$$z = \frac{168 - 170}{\frac{10}{\sqrt{50}}} \approx -1.41$$

Vì $|z| \leq z_{\alpha/2}$, ta chấp nhận H_0 . Ta có:

$$\alpha = P(Z < -1.96) + P(Z > 1.96) \approx 0.025 + 0.025 = 0.05$$

Vậy xác suất mắc sai lầm loại I trong trường hợp này là 5%. So với bài toán gốc, ta thấy rằng khi giảm kích thước mẫu, ta không thay đổi xác suất mắc sai lầm loại I. Điều này là do ta đã đặt ra một ngưỡng cho xác suất mắc sai lầm loại I, và ta sẽ bác bỏ H_0 nếu giá trị kiểm định vượt quá ngưỡng này.

$$\beta = P(Z < z_1 | \mu = 165) + P(Z > z_2 | \mu = 165)$$

Trong đó:

$$z_1 = \frac{170 - 165}{\frac{10}{\sqrt{50}}} - 1.96 \approx -0.59$$

$$z_2 = \frac{170 - 165}{\frac{10}{\sqrt{50}}} + 1.96 \approx 3.41$$

Sử dụng bảng phân phối chuẩn tiêu chuẩn, ta có:

$$\beta \approx 0.2776 + 0.00032 \approx 0.27792$$

Vậy xác suất mắc sai lầm loại II trong trường hợp này là khoảng 27.79% . So với bài toán gốc, ta thấy rằng khi giảm kích thước mẫu, ta đã làm tăng β , tức là làm tăng xác suất mắc sai lầm loại II. Điều này là do khi kích thước mẫu nhỏ, độ chính xác của ước lượng mẫu cũng thấp hơn, dẫn tới việc khó phân biệt được giữa các giả thuyết khác nhau.

- **Trường hợp 2:** $n = 100$, $z_{\alpha/2} = 2.58$, $\mu = 165$. Đây là trường hợp ta tăng điểm tới hạn, giữ nguyên kích thước mẫu và giá trị thực của tham số. Ta có:

$$z = \frac{168 - 170}{\frac{10}{\sqrt{100}}} = -2$$

Vì $|z| > z_{\alpha/2}$, ta bác bỏ H_0 . Ta có:

$$\alpha = P(Z < -2.58) + P(Z > 2.58) \approx 0.005 + 0.005 = 0.01$$

Vậy xác suất mắc sai lầm loại I trong trường hợp này là 1%. So với bài toán gốc, ta thấy rằng khi tăng điểm tới hạn, ta đã làm giảm xác suất mắc sai lầm loại I. Điều này là do khi ta thu hẹp vùng bác bỏ H_0 , ta cũng làm cho việc bác bỏ H_0 khó khăn hơn, dẫn tới việc giảm xác suất mắc sai lầm loại I.

$$\beta = P(Z < z_1 | \mu = 165) + P(Z > z_2 | \mu = 165)$$

Trong đó:

$$z_1 = \frac{170 - 165}{\frac{10}{\sqrt{100}}} - 2.58 \approx -0.42$$

$$z_2 = \frac{170 - 165}{\frac{10}{\sqrt{100}}} + 2.58 \approx 5.42$$

Sử dụng bảng phân phối chuẩn tiêu chuẩn, ta có:

$$\beta \approx 0.3372 + 0 \approx 0.3372$$

Vậy xác suất mắc sai lầm loại II trong trường hợp này là khoảng 33.72% . So với bài toán gốc, ta thấy rằng khi tăng điểm tới hạn, ta đã làm tăng β , tức là làm tăng xác suất mắc sai lầm loại II. Điều này là do khi ta mở rộng vùng bác bỏ H_0 , ta cũng làm cho việc phân biệt được giữa các giả thuyết khó khăn hơn, dẫn tới việc có thể bỏ sót những trường hợp H_0 sai mà ta lại chấp nhận.

- **Trường hợp 3:** $n = 100, z_{\alpha/2} = 1.96, \mu = 167$. Đây là trường hợp ta tăng giá trị thực của tham số, giữ nguyên kích thước mẫu và điểm tới hạn. Ta có:

$$z = \frac{168 - 170}{\frac{10}{\sqrt{100}}} = -2$$

Vì $|z| < z_{\alpha/2}$, ta chấp nhận H_0 . Ta có:

$$\alpha = P(Z < -2.58) + P(Z > 2.58) \approx 0.005 + 0.005 = 0.01$$

Vậy xác suất mắc sai lầm loại I trong trường hợp này là 1%. So với bài toán gốc, ta thấy rằng khi tăng điểm tới hạn, ta đã làm giảm xác suất mắc sai lầm loại I. Điều này là do khi ta thu hẹp vùng bác bỏ H_0 , ta cũng làm cho việc bác bỏ H_0 khó khăn hơn, dẫn tới việc giảm xác suất mắc sai lầm loại I.

Trong đó:

$$z_1 = \frac{170 - 167}{\frac{10}{\sqrt{100}}} - 1.96 \approx -1.13$$

$$z_2 = \frac{170 - 167}{\frac{10}{\sqrt{100}}} + 1.96 \approx 2.87$$

Sử dụng bảng phân phối chuẩn tiêu chuẩn, ta có:

$$\beta \approx 0.1292 + 0.002 \approx 0.1312$$

Vậy xác suất mắc sai lầm loại II trong trường hợp này là khoảng 13.12%. So với bài toán gốc, ta thấy rằng khi tăng giá trị thực của tham số, ta đã làm tăng β , tức là làm tăng xác suất mắc sai lầm loại II. Điều này là do khi giá trị thực của tham số gần với giá trị kiểm định hơn, ta khó phân biệt được giữa H_0 và H_1 , dẫn tới việc có nhiều khả năng chấp nhận H_0 khi nó sai.

Từ các ví dụ trên, ta có thể thấy rằng các yếu tố như kích thước mẫu, điểm tới hạn, giá trị thực của tham số đều có ảnh hưởng tới xác suất mắc sai lầm loại I và loại II trong bài toán kiểm định giả thuyết về tham số của một tổng thể. Ta cần cân nhắc kỹ khi lựa chọn các yếu tố này để đảm bảo kết quả kiểm định có độ tin cậy cao và ít bị sai lệch.

Kết luận

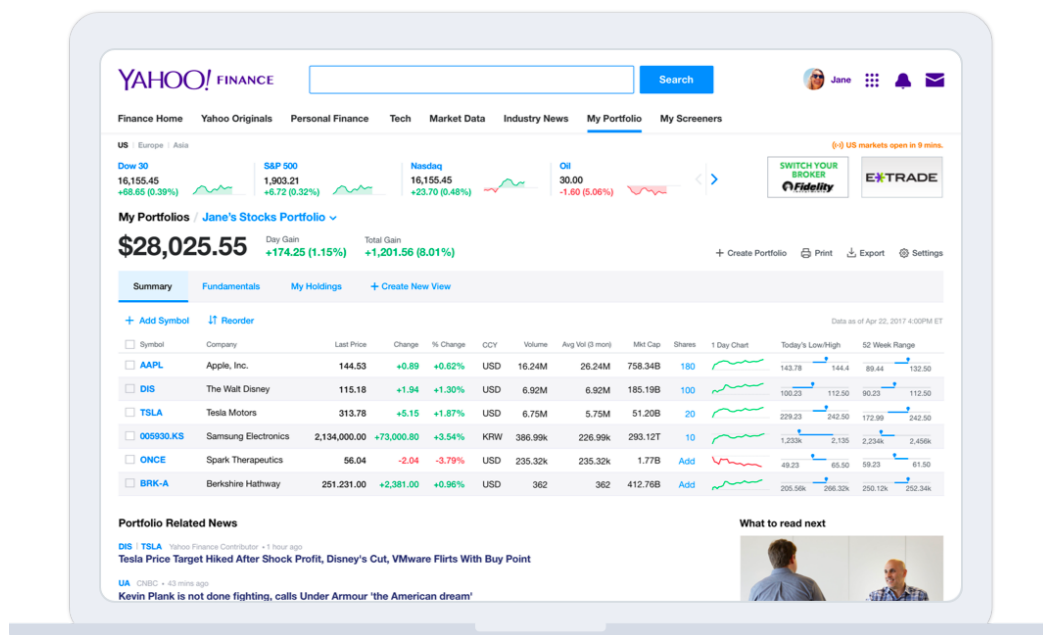
Nếu ta thay đổi các yếu tố như kích thước mẫu, điểm tới hạn, giá trị thực của tham số, ta sẽ thấy ảnh hưởng của chúng tới xác suất mắc sai lầm loại I và loại II như sau:

- Nếu ta tăng kích thước mẫu, ta sẽ giảm được cả α và β , tức là giảm được cả xác suất mắc sai lầm loại I và loại II. Điều này là do khi kích thước mẫu lớn, độ chính xác của ước lượng mẫu cũng cao hơn, dẫn tới việc phân biệt được giữa các giả thuyết khác nhau dễ dàng hơn.
- Nếu ta giảm điểm tới hạn, tức là giảm khoảng cách từ giá trị kiểm định đến vùng bác bỏ H_0 , ta sẽ giảm được α , tức là giảm được xác suất mắc sai lầm loại I. Tuy nhiên, điều này cũng làm tăng β , tức là tăng xác suất mắc sai lầm loại II. Điều này là do khi ta thu hẹp vùng bác bỏ H_0 , ta cũng làm cho việc phân biệt được giữa các giả thuyết khó khăn hơn, dẫn tới việc có thể bỏ sót những trường hợp H_0 sai mà ta lại chấp nhận.
- Nếu ta thay đổi giá trị thực của tham số, tức là thay đổi khoảng cách từ giá trị kiểm định đến giá trị thực, ta sẽ ảnh hưởng tới β , tức là ảnh hưởng tới xác suất mắc sai lầm loại II. Nếu khoảng cách này càng lớn, β càng nhỏ, tức là xác suất mắc sai lầm loại II càng nhỏ. Điều này là do khi khoảng cách này lớn, ta có thể phân biệt được giữa H_0 và H_1 dễ dàng hơn, dẫn tới việc ít có khả năng chấp nhận H_0 khi nó sai. Ngược lại, nếu khoảng cách này càng nhỏ, β càng lớn, tức là xác suất mắc sai lầm loại II càng lớn. Điều này là do khi khoảng cách này nhỏ, ta khó phân biệt được giữa H_0 và H_1 , dẫn tới việc có nhiều khả năng chấp nhận H_0 khi nó sai.

Chương 3

XÂY DỰNG MÔ HÌNH HỒI QUY TUYẾN TÍNH BỘI

3.1 Mô tả bài toán và bộ dữ liệu

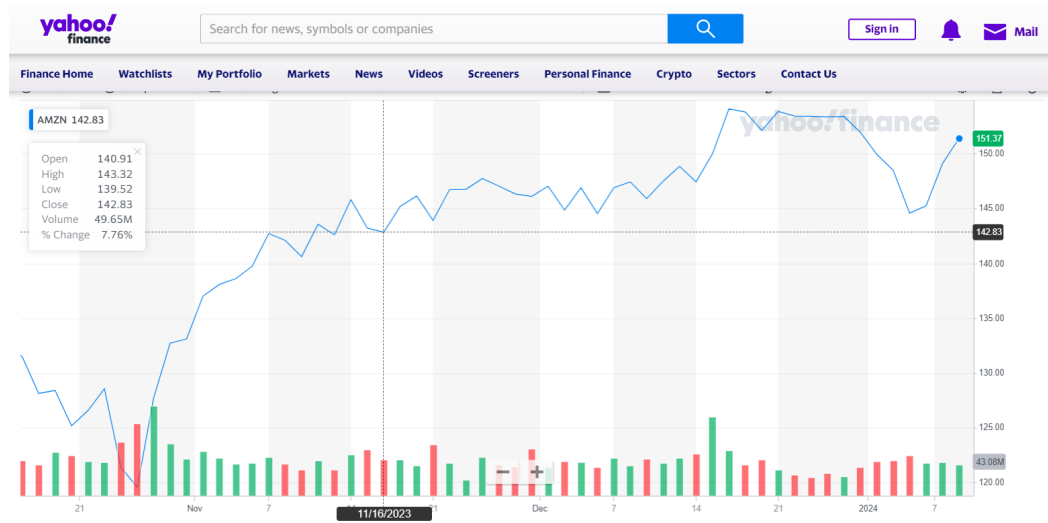


Hình 3.1: Yahoo Finance

Yahoo Finance là một dịch vụ tài chính trực tuyến cung cấp thông tin về thị trường tài chính, cổ phiếu, tỷ giá ngoại tệ, và tin tức kinh tế. Người dùng có thể theo dõi giá cổ phiếu, xem biểu đồ giá thời gian thực, đọc tin tức tài chính, và sử dụng các công cụ phân tích để hỗ trợ quyết định đầu tư.

Yahoo Finance cũng cung cấp dữ liệu lịch sử cổ phiếu, thông tin chi tiết về công ty, và các công cụ khác nhau để theo dõi và quản lý danh mục đầu tư. Đây là một nguồn thông tin quan

trọng cho những người quan tâm đến thị trường tài chính và đầu tư.



Hình 3.2: Cổ phiếu Amazon

Dữ liệu thu thập trực tiếp từ Yahoo Finance trong 2 năm gần nhất 2022 - 2023 cung cấp các chỉ số quan trọng liên quan đến cổ phiếu, tạo ra một bộ dữ liệu lý tưởng để áp dụng mô hình hồi quy tuyến tính.

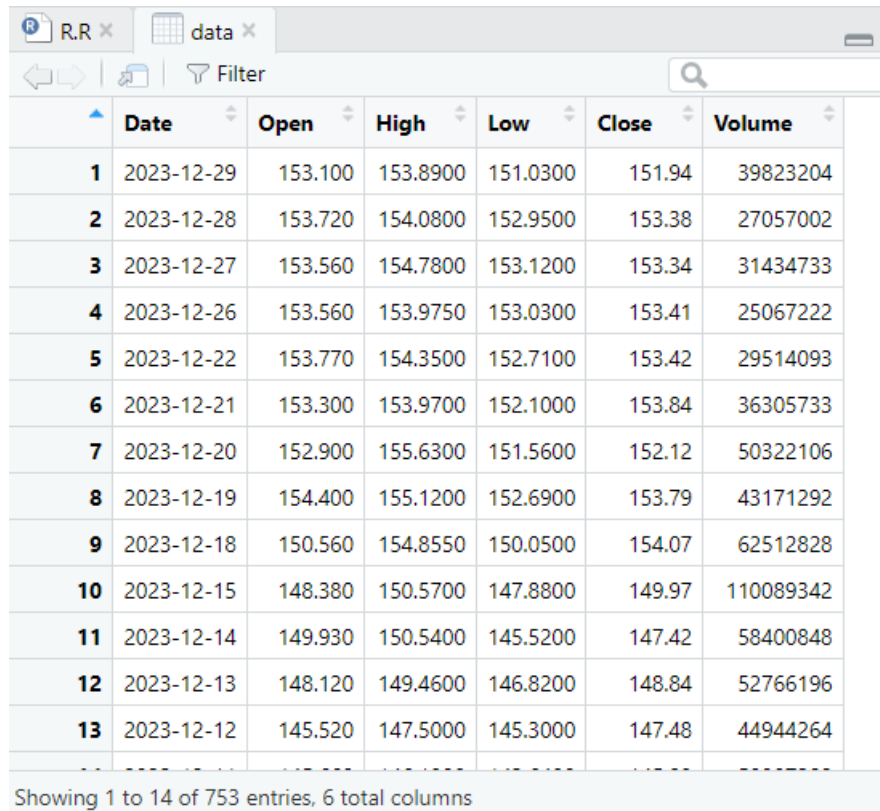
Trong những biến như:

- Giá mở cửa (Open)
- Giá cao nhất (High)
- Giá thấp nhất (Low)
- Khối lượng giao dịch (Volume)

⇒ Chúng ta có thể tận dụng mối quan hệ số liên quan để dự đoán Giá đóng cửa (Close)

Các thông số này có thể giúp xây dựng một mô hình hồi quy tuyến tính, trong đó các biến độc lập như Giá mở cửa, Giá cao nhất, Giá thấp nhất, và Khối lượng giao dịch sẽ được sử dụng để dự đoán biến phụ thuộc, tức là Giá đóng cửa. Mục tiêu là tìm ra một mô hình có thể mô tả mối quan hệ tuyến tính giữa các yếu tố này để dự đoán giá cổ phiếu trong các phiên giao dịch tiếp theo.

Sự linh hoạt và khả năng dự đoán của mô hình hồi quy tuyến tính có thể giúp đưa ra quyết định đầu tư thông minh, dựa trên hiểu biết chặt chẽ về diễn biến giá cổ phiếu từ các thông số quan trọng.



	Date	Open	High	Low	Close	Volume
1	2023-12-29	153.100	153.8900	151.0300	151.94	39823204
2	2023-12-28	153.720	154.0800	152.9500	153.38	27057002
3	2023-12-27	153.560	154.7800	153.1200	153.34	31434733
4	2023-12-26	153.560	153.9750	153.0300	153.41	25067222
5	2023-12-22	153.770	154.3500	152.7100	153.42	29514093
6	2023-12-21	153.300	153.9700	152.1000	153.84	36305733
7	2023-12-20	152.900	155.6300	151.5600	152.12	50322106
8	2023-12-19	154.400	155.1200	152.6900	153.79	43171292
9	2023-12-18	150.560	154.8550	150.0500	154.07	62512828
10	2023-12-15	148.380	150.5700	147.8800	149.97	110089342
11	2023-12-14	149.930	150.5400	145.5200	147.42	58400848
12	2023-12-13	148.120	149.4600	146.8200	148.84	52766196
13	2023-12-12	145.520	147.5000	145.3000	147.48	44944264

Showing 1 to 14 of 753 entries, 6 total columns

Hình 3.3: View dữ liệu

Dữ liệu gồm 756 hàng, 6 cột.

3.2 Phương pháp thực hiện

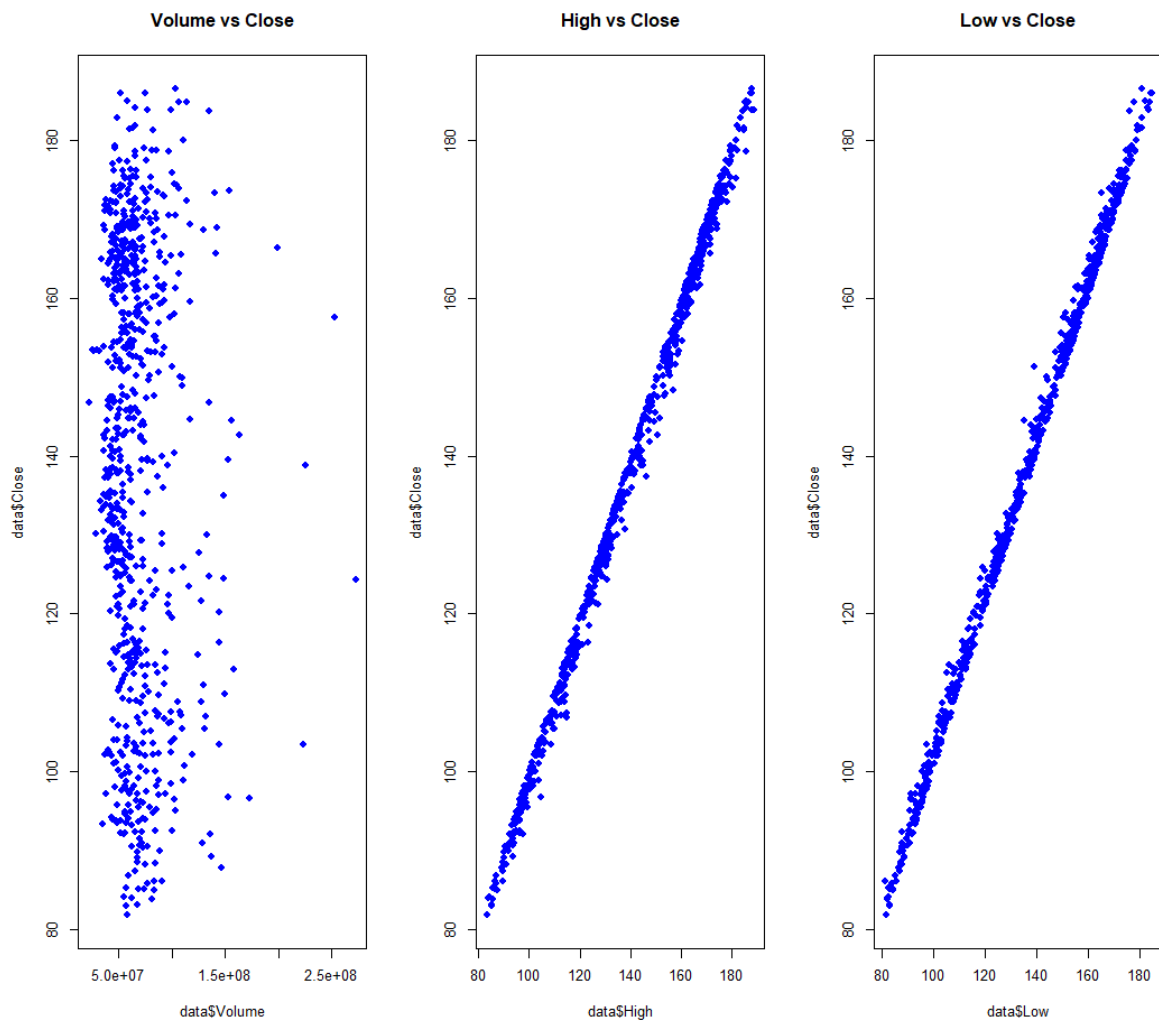
Sử dụng biến Open, High, Low, Volume để dự đoán giá Close.

```
1 dabien <- lm(data$Close ~ data$Open + data$High + data$Low + data$Volume)
```

predicted_data biểu thị cho giá trị dự đoán dựa trên mô hình.

```
1 predicted_data <- data.frame(Predicted = predict(dabien), Observed =  
  data$Close)
```

Vẽ đồ thị phân tán giữa các cặp biến:

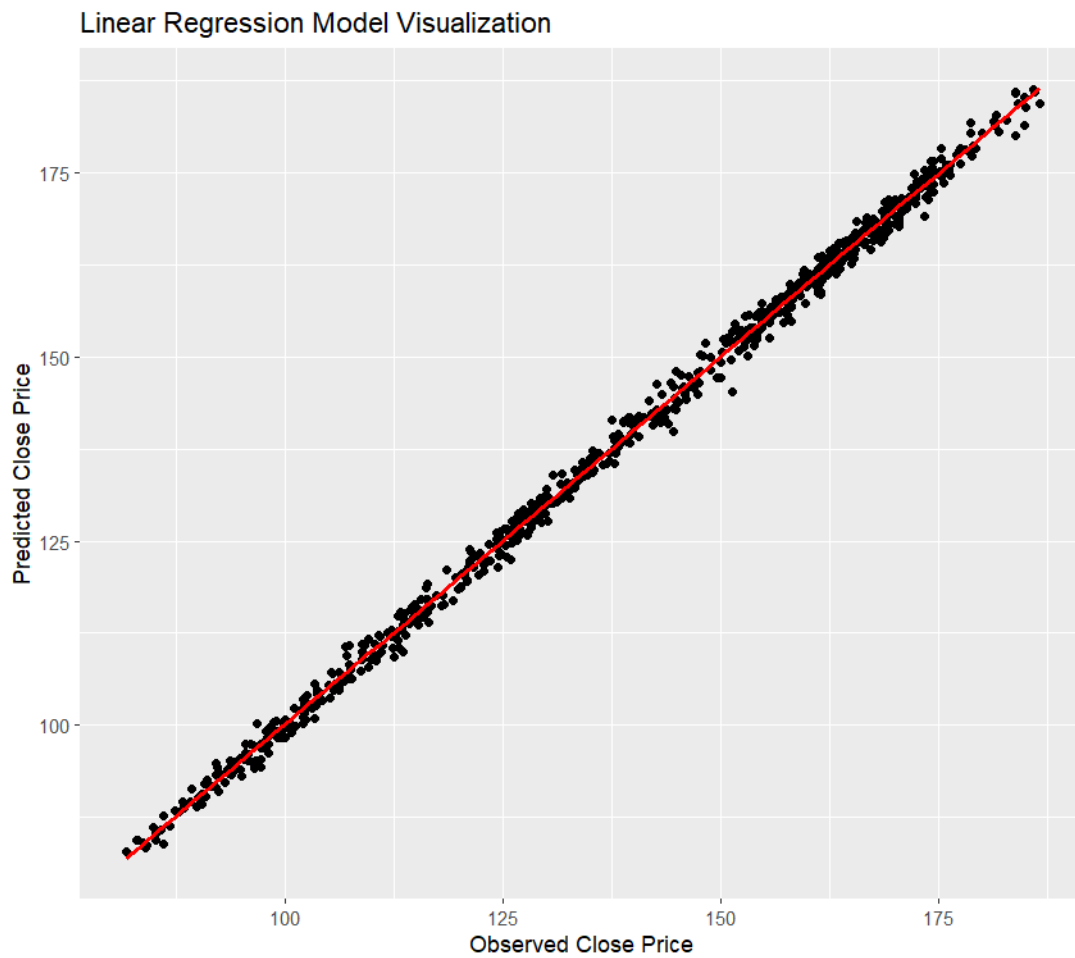


Hình 3.4: Kiểm tra tương quan giữa các biến

Nhận xét:

- Đồ thị đám mây điểm giữa Volume và Close cho thấy chúng tương quan khá kém.
- Đối với đồ thị giữa High, Low và Close các điểm nằm trên cùng 1 đường tuyến tính
→ Chúng tương quan tốt.

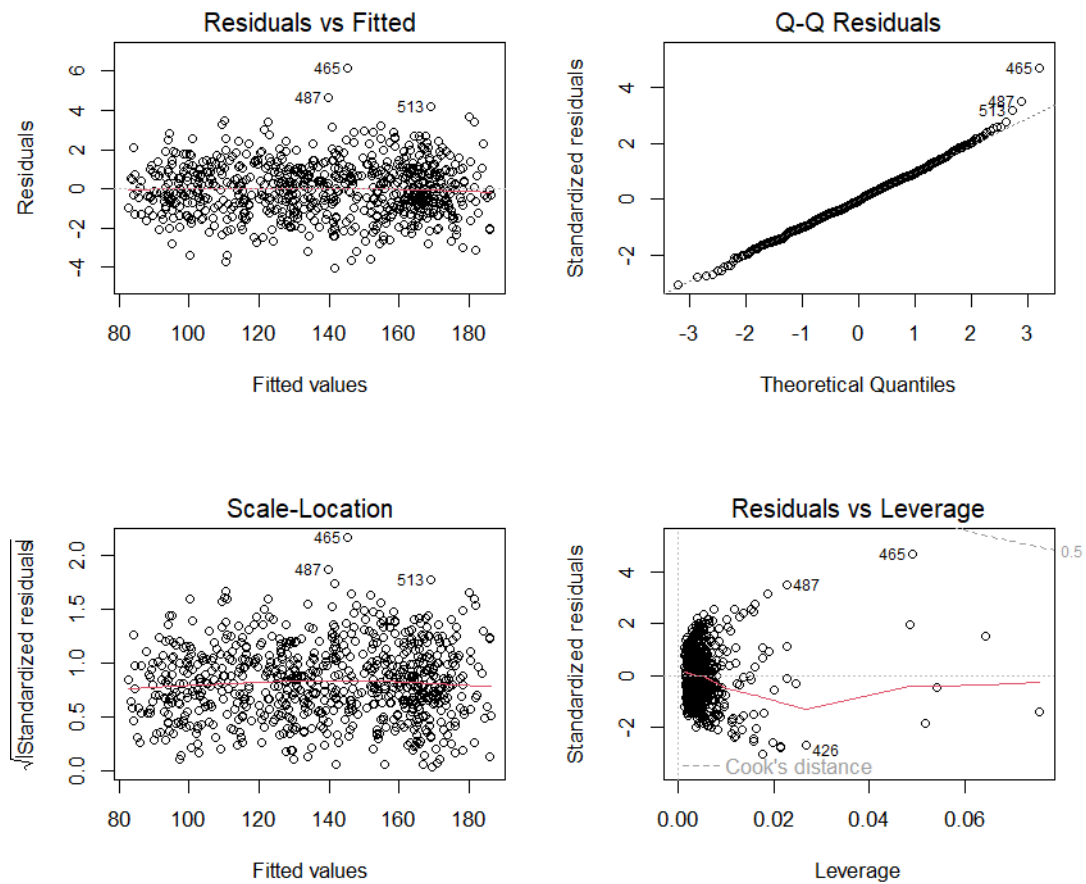
Thực hiện vẽ đồ thị để xem xét sự tương quan:



Hình 3.5: Xây dựng đường trực quan

Nhận xét:

Nhìn chung, đồ thị này cho thấy mối quan hệ tuyến tính chặt chẽ giữa giá đóng cửa thực tế và giá đóng cửa dự đoán. Các điểm dữ liệu nằm gần đường hồi quy tuyến tính, cho thấy rằng mô hình hồi quy phù hợp với dữ liệu khá tốt.



Hình 3.6: Đồ thị phần dư

Nhận xét:

- Đồ thị thứ 1 (Residuals vs Fitted) cho thấy giả thiết về tính tuyến tính của dữ liệu hơi bị vi phạm. Tuy nhiên giả thiết trung bình của phần dư có thể coi là thỏa mãn
- Đồ thị Normal Q-Q cho thấy giả thiết phần dư có phân phối chuẩn được thỏa mãn.
- Đồ thị (Scale - Location) cho ta thấy rằng giả thiết về tính đồng nhất của phương sai cũng thỏa mãn.
- Đồ thị thứ tư chỉ ra có các quan trắc thứ 487, 426 và 465 có thể là các điểm có ảnh hưởng cao trong bộ dữ liệu.

3.3 Đánh giá mô hình

Tính hệ số $R_squared$: 0.9976

Một số các thông số khác:

```
> summary(dabien1)

Call:
lm(formula = data$Close ~ data$High + data$Low)

Residuals:
    Min       1Q   Median       3Q      Max
-4.0724 -0.8740 -0.0407  0.8956  6.1237

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.42201    0.27602   1.529   0.127
data$High    0.45617    0.03151  14.478 <2e-16 ***
data$Low     0.54186    0.03153  17.186 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.35 on 750 degrees of freedom
Multiple R-squared:  0.9976,    Adjusted R-squared:  0.9976
F-statistic: 1.554e+05 on 2 and 750 DF,  p-value: < 2.2e-16

> |
```

Hình 3.7: Thông tin khác

Nhận xét: Mô hình có độ chính xác cao (Multiple R-squared gần 1), giả sử rằng mô hình phản ánh tốt sự biến động của dữ liệu. F-statistic và p-value của nó cũng thể hiện sự ý nghĩa toàn cục của mô hình.

Một phần cho ra kết quả trên do dữ liệu trên đã quá sạch, đã thông qua các bước chuẩn hóa khi đưa lên mạng.

Chương 4

Kết luận cuối cùng

Trong bài báo cáo này, chúng em đã được làm việc với dữ liệu bằng phần mềm R, kiểm định giả thuyết thống kê và xây dựng mô hình hồi quy tuyến tính bội. Trong mục làm việc với R, bọn em đã thực hiện những thao tác cơ bản như: Nhập dữ liệu, thao tác chiết, ghép nối, xuất dữ liệu, lập bảng tần số, thực hiện vẽ các biểu đồ để thấy sự phân bố của dữ liệu và cuối cùng là tính những đặc trưng mẫu bằng hàm có sẵn trong R và tự xây dựng công thức để tính.

Trong phần kiểm định giả thuyết thống kê, nhóm em đã giải quyết bài toán kiểm định về trung bình tổng thể, xây dựng công thức để tính toán các xác suất mắc sai lầm loại I, loại II trong các trường hợp kích thước mẫu khác nhau, điểm tới hạn khác nhau và giá trị thực của các tham số khác nhau, cuối cùng là đưa ra nhận xét về sự ảnh hưởng của các yếu tố đến xác suất mắc sai lầm loại I, loại II.

Trong phần cuối cùng, nhóm đã thực hiện xây dựng mô hình tuyến tính bội để dự đoán giá đóng cửa cổ phiếu dựa trên bộ dữ liệu về giá cổ phiếu của Amazon. Nhóm đã biết cách đánh giá mô hình hồi quy với hệ số $R_squared$, F-statistic và p_value.

Tài liệu tham khảo

- [1] Nguyễn Thị Thu Thủy, Bài giảng “Suy luận thống kê”, Đại học Bách khoa Hà Nội, 2023.
- [2] Nguyễn Thị Thanh Hương, Suy luận thống kê và ứng dụng, Nhà xuất bản Đại học Quốc gia Hà Nội, 2021.
- [3] Phạm Văn Thành, Nguyễn Thị Thanh Hương, Suy luận thống kê với R, Nhà xuất bản Đại học Quốc gia Hà Nội, 2020.
- [4] Nguyễn Thị Hồng Hạnh, Nguyễn Thị Thanh Hương, Suy luận thống kê với SPSS, Nhà xuất bản Đại học Quốc gia Hà Nội, 2019.
- [5] Nguyễn Thị Thanh Hương, Nguyễn Thị Hồng Hạnh, Suy luận thống kê với Excel, Nhà xuất bản Đại học Quốc gia Hà Nội, 2018.