**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC BÁCH KHOA**


# BÁO CÁO BÀI TẬP LỚN

**Bộ môn**: Xác suất thống kê

**Giáo viên:** *Nguyễn Thị Kiều Dung*

**Nhóm GT52– Lớp L04 – Học kỳ 231**

| STT | Họ và tên | MSSV | Lớp | Khoa |
|---|---|---|---|---|
| 1 | Hồ Trần An Phong | 2212551 | L04 | Kỹ thuật Giao thông |
| 2 | Trần Ngô Nhật Luân | 2211956 | L04 | Kỹ thuật Giao thông |
| 3 | Trần Phát Lộc | 2211935 | L04 | Kỹ thuật Giao thông |
| 4 | Hoàng Quốc Huy | 2211172 | L02 | Kỹ thuật Giao thông |
| 5 | Nguyễn Hoàng Huy | 2211215 | L02 | Kỹ thuật Giao thông |
| **6** | **Châu Trần Phúc Thiện** | **2213244** | **L04** | Kỹ thuật Giao thông |

# THÔNG TIN NHÓM THỰC HIỆN

| Họ và tên | MSSV | Mô tả đóng góp | Điểm chia |
|---|---|---|---|
| Hồ Trần An Phong (L04) | 2212551 | Phương pháp ANOVA | |
| Trần Ngô Nhật Luân (L04) | 2211956 | Tổng hợp dư liệu – tổng quan dữ liệu | |
| Trần Phát Lộc (L04) | 2211935 | Kiểm định 1 mẫu, 2 mẫu | |
| Hoàng Quốc Huy (L02) | 2211172 | Thảo luận và mở rộng | |
| Nguyễn Hoàng Huy (L02) | 2211215 | Phương pháp hồi quy | |
| Châu Trần Phúc Thiện (L04) | 2213244 | Code | |

# MỤC LỤC

# 1. TỔNG QUAN VỀ DỮ LIỆU

## 1.1. Ngữ cảnh dữ liệu

Bộ dữ liệu "Airlines Traffic Passenger Statistics" chứa thông tin về số liệu thống kê hành khách của các hãng hàng không. Nó bao gồm thông tin về các hãng hàng không, sân bay và khu vực mà các chuyến bay khởi hành và đến. Nó cũng bao gồm thông tin về loại hoạt động, loại giá, nhà ga, khu vực lên máy bay và số lượng hành khách.

Tập dữ liệu được lấy từ Open Flight chứa thông tin về số liệu thống kê hành khách không lưu theo hãng hàng không trong năm 2017. Dữ liệu bao gồm số lượng hành khách, hãng hàng không khai thác, hãng hàng không được công bố, khu vực địa lý, mã loại hoạt động, mã danh mục giá, nhà ga, khu vực lên máy bay, năm và tháng của chuyến bay.

## 1.2. Mục đích của việc thống kê dữ liệu

Số liệu thống kê hành khách không lưu có thể là một công cụ hữu ích để hiểu ngành hàng không và lập kế hoạch du lịch.

Số liệu thống kê về hành khách trong ngành hàng không có thể được sử dụng để dự đoán xu hướng du lịch hàng không trong tương lai.

Dữ liệu cũng có thể được sử dụng để tạo ra bản đồ nhiệt về mô hình giao thông hàng không.

Dữ liệu có thể được sử dụng để nghiên cứu tác động của các yếu tố khác nhau đến số lượng hành khách giao thông hàng không, chẳng hạn như thời gian trong năm hoặc ngày, giá vé máy bay hoặc số lượng chuyến bay do một hãng hàng không cung cấp.

## 1.3. Các loại biến có trong dữ liệu

+ Biến phụ thuộc:

- Biến Activity period, phụ thuộc vào biến Year và Month thể hiện thời gian thực hiện hoạt động.
- Biến Operating Airline IATA Code sẽ là biến phụ thuộc, phụ thuộc vào biến Operating Airline và biến Published Airline cũng có thể là biến phụ thuộc vào biến Operating Airline
- Biến Published Airline IATA Code

- Biến Adjusted Activity Type Code sẽ là biến phụ thuộc vào biến Activity Type Code
- Biến Adjusted Passenger Count sẽ là biến phụ thuộc vào Biến Passenger count

+ Biến độc lập:

- Biến Operating Airline, thể hiện hãng hàng không thực hiện chuyến bay
- Biến Published Airline là biến độc lập, biến Published Airline thể hiện hãng hàng không nào bay gộp với hãng vận hành nên có khả năng sẽ phụ thuộc biến Operating Airline
- Biến GEO Region là biến độc lập có biến GEO Summary và biến Terminal là biến phụ thuộc vào. Biến GEO region thể hiện vùng địa lý mà chuyến bay thực hiện còn biến GEO Summary là biến thể hiện chuyến bay ngoài hoặc trong nước nên sẽ phụ thuộc vào biến GEO region
- Biến Price Category Code sẽ là độc lập thể hiện giá vé
- Biến Activity Type Code sẽ là biến độc lập thể hiện hoạt động thực hiện
- Biến Terminal và biến Boarding Area sẽ là biến độc lập thể hiện ga hàng không
- Biến Passenger count là biến độc lập thể hiện số lượng hành khách
- Biến Year và Month sẽ là 2 biến độc lập

+ Vì biến Adjusted Passenger Count sẽ là biến phụ thuộc vào Biến Passenger count vậy ta có biến Adjusted Activity Type Code cũng sẽ là biến kiểm soát của biến Adjusted Passenger Count và là biến định tính.

# 2. KIẾN THỨC NỀN

## 2.1. Phân tích phương sai (Anova)

### ❖ Mục đích

Mục tiêu của phân tích phương sai là so sánh trung bình của nhiều nhóm (tổng thể) dựa trên các số trung bình của các mẫu quan sát từ các nhóm này và thông qua kiểm định giả thuyết để kết luận về sự bằng nhau của các số trung bình này.

Trong nghiên cứu, phân tích phương sai được dùng như là một công cụ để xem xét ảnh hưởng của một hay một số yếu tố nguyên nhân (định tính) đến một yếu tố kết quả (định lượng).

*Ví dụ:*

- Nghiên cứu ảnh hưởng của phương pháp đánh giá của giáo viên đến kết quả học tập của sinh viên.
- Nghiên cứu ảnh hưởng của bậc thợ tới năng suất lao động.
- Nghiên cứu ảnh hưởng của phương pháp bán hàng, trình độ (kinh nghiệm) của nhân viên bán hàng đến doanh số.

### 2.1.1. Phân tích phương sai một yếu tố

Phân tích phương sai một yếu tố là phân tích ảnh hưởng của một yếu tố nguyên nhân (dạng biến định tính) đến một yếu tố kết quả (dạng biến định lượng) đang nghiên cứu.

Giả sử cần so sánh số trung bình của k tổng thể độc lập. Ta lấy k mẫu có số quan sát là $n_1$, $n_2$… $n_k$: tuân theo phân phối chuẩn.

Trung bình của các tổng thể được ký hiệu là $\mu_1$; $\mu_2$ ….$\mu_k$ thì mô hình phân tích phương sai một yếu tố ảnh hưởng được mô tả dưới dạng kiểm định giả thuyết như sau:

Giả thiết không: $H_o$: $\mu_1 = \mu_2 = \ldots = \mu_k$

Giả thiết đối $H_1$: Tồn tại ít nhất 1 cặp có $\mu_i \neq \mu_j$; $i \neq j$

*Để kiểm định ta đưa ra 3 giả thiết sau:*

1) Mỗi mẫu tuân theo phân phối chuẩn N~(μ, $\sigma_2$)

2) Các phương sai tổng thể bằng nhau

3) Ta lấy k mẫu độc lập từ k tổng thể. Mỗi mẫu được quan sát $n_j$ lần.

***Các bước tiến hành:***

***Bước 1***: Tính các trung bình mẫu và trung bình chung của k mẫu

- Ta lập bảng tính toán như sau:

| TT | k mẫu quan sát | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | ... | k |
| 1 | $X_{11}$ | $X_{12}$ | $X_{13}$ | | $X_{1k}$ |
| 2 | $X_{21}$ | $X_{22}$ | $X_{23}$ | | $X_{2k}$ |
| 3 | $X_{31}$ | $X_{32}$ | $X_{33}$ | | $X_{3k}$ |
| ... | | | | | |
| ... | | | | | |
| j | $X_{j1}$ | $X_{j2}$ | $X_{j3}$ | | $X_{jk}$ |
| Trung bình mẫu | $\overline{x_1}$ | $\overline{x_2}$ | $\overline{x_3}$ | | $\overline{x_k}$ |

Trung bình mẫu $x_1$, $x_2$, $x_k$ được tính theo công thức: $\overline{x_i} = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i} (i = 1,2,..k)$

Trung bình chung của k mẫu được tính theo công thức: $\overline{x} = \frac{\sum_{i=1}^{k} n_i \overline{x_i}}{\sum_{i=1}^{k} n_i} (i = 1,2,..k)$

*Bước 2*: Tính các tổng độ lệch bình phương

Tổng các độ lệch bình phương trong nội bộ nhóm (nội bộ từng mẫu - SSW) được tính theo công thức sau:

| Nhóm 1 | Nhóm 2 | Nhóm k |
|---|---|---|
| $SS_1 = \sum_{j=1}^{n_1} (X_{j1} - \overline{x_1})^2$ | $SS_2 = \sum_{j=1}^{n_2} (X_{j2} - \overline{x_2})^2$ | $SS_k = \sum_{j=1}^{n_k} (X_{jk} - \overline{x_k})^2$ |
| $SSW = SS_1 + SS_2 + ... + SS_k = \sum_{i=1}^{k} \sum_{ij=11}^{n_i} (X_{ij} - \overline{x_i})^2$ | | |

Tổng các độ lệch bình phương giữa các nhóm (SSB):

$$SSB = \sum_{i=1}^{k} n_i (\overline{x_i} - \overline{x})^2$$

Tổng các độ lệch bình phương của toàn bộ tổng thể (SST):

$$SST = SSW + SSB = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{x})^2$$

*Bước 3:* Tính các phương sai (phương sai của nội bộ nhóm và phương sai giữa các nhóm)

Ta ký hiệu k là số nhóm (mẫu); n là tổng số quan sát của các nhóm thì các phương sai được tính theo công thức sau:

| | |
|---|---|
| $MSW = \frac{SSW}{n-k}$ | $MSB = \frac{SSB}{k-1}$ |

MSW: Là phương sai nội bộ nhóm SSB: Là phương sai giữa các nhóm

***Bước 4***: Kiểm định giả thiết

Tính tiêu chuẩn kiểm định F (F thực nghiệm): $F = \frac{MSB}{MSW}$

F > F ((k-1; n-k); α)

Ta bác bỏ giả thuyết $H_0$ cho rằng trị trung bình của k tổng thể bằng nhau

Tìm F lý thuyết (F tiêu chuẩn = F (k-1; n-k; α)):

- F lý thuyết là giá trị giới hạn tra từ bảng phân phối F với k-1 bậc tự do của phương sai ở tử số và n-k bậc tự do của phương sai ở mẫu số với mức ý nghĩa α.
- F lý thuyết có thể tra qua hàm FINV (α, k-1, n-1) trong EXCEL.
- Nếu F thực nghiệm > F lý thuyết, bác bỏ Ho, nghĩa là các số trung bình của k tổng thể không bằng nhau.

Bảng phân tích phương sai 1 yếu tố khi sử dụng máy tính tóm tắt như sau:

Bảng phân tích phương sai tổng quát dịch ra tiếng việt – ANOVA

| Nguồn biến động | Tổng độ lệch bình phương (SS) | Bậc tự do (df) | Phương sai (MS) | F- Tỷ số |
|---|---|---|---|---|
| Giữa các mẫu | SSB | (k-1) | MSB | $F = \frac{MSB}{MSW}$ |
| Trong nội bộ các mẫu | SSW | (n-k) | MSW | |
| Tổng số | SST | (n-1) | | |

### 2.1.2. Phân tích phương sai 2 yếu tố

Phân tích phương sai 2 yếu tố nhằm xem xét cùng lúc hai yếu tố nguyên nhân (dưới dạng dữ liệu định tính) ảnh hưởng đến yếu tố kết quả (dưới dạng dữ liệu định lượng) đang nghiên cứu.

*Ví dụ*: Nghiên cứu ảnh hưởng của loại chất đốt và loại lò sấy đến tỷ lệ vải loại 1 sấy khô.

Phân tích phương sai 2 yếu tố giúp chúng ta đưa thêm yếu tố nguyên nhân vào phân tích làm cho kết quả nghiên cứu càng có giá trị.

Giả sử ta nghiên cứu ảnh hưởng của 2 yếu tố nguyên nhân định tính đến một yếu tố kết quả định lượng nào đó.

Ta lấy mẫu không lặp lại, sau đó các đơn vị mẫu của yếu tố nguyên nhân thứ nhất sắp xếp thành K nhóm (cột), các đơn vị mẫu của yếu tố nguyên nhân thứ hai sắp xếp thành H khối (hàng). Như vậy, ta có bảng kết hợp 2 yếu tố nguyên nhân gồm K cột và H hàng và (K x H) ô dữ liệu. Tổng số mẫu quan sát là n = (K x H).

| Hàng (Khối) | Cột (nhóm) | | | |
|---|---|---|---|---|
| | 1 | 2 | … | K |
| 1 | $X_{11}$ | $X_{21}$ | | $X_{K1}$ |
| 2 | $X_{12}$ | $X_{22}$ | | $X_{K2}$ |
| … | | | | |
| H | $X_{1K}$ | $X_{2K}$ | | $X_{KH}$ |

*Để kiểm định ta đưa ra 2 giả thiết sau:*

1) Mỗi mẫu tuân theo phân phối chuẩn N~(μ, σ2)

2) Ta lấy K mẫu độc lập từ K tổng thể, H mẫu độc lập từ H tổng thể. Mỗi mẫu được quan sát 1 lần không lặp

***Các bước tiến hành***

***Bước 1***: Tính các số trung bình

| Trung bình riêng của từng nhóm (K cột) | Trung bình riêng của từng khối (H hàng) |
|---|---|
| $\overline{X}_i = \frac{\sum_{j=1}^{H} X_{ij}}{H}$ <br> $i = 1, 2 \ldots K$ | $\overline{X}_j = \frac{\sum_{i=1}^{K} X_{ij}}{K}$ <br> $j = 1, 2 \ldots H$ |

**Trung bình chung của toàn bộ mẫu quan sát** $\overline{X} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{H} X_{ij}}{n} = \frac{\sum_{i=1}^{K} \overline{X}_i}{K} = \frac{\sum_{j=1}^{H} \overline{X}_j}{H}$

Bước 2: Tính tổng các độ lệch bình phương

| Diễn giải | Công thức |
|---|---|
| 1. Tổng các độ lệch bình phương chung (SST) *Phản ánh biến động của yếu tố kết quả do ảnh hưởng của tất cả các yếu tố* | $SST = \sum_{i=1}^{K}\sum_{j=1}^{H}(X_{ij} - \overline{X})^2$ |
| 2. Tổng các độ lệch bình phương giữa các nhóm (SSK) *Phản ánh biến động của yếu tố kết quả do ảnh hưởng của yếu tố nguyên nhân thứ nhất (xếp theo cột)* | $SSK = H\sum_{i=1}^{K}(\overline{X}_i - \overline{X})^2$ |
| 3.Tổng các độ lệch bình phương giữa các nhóm (SSH) *Phản ánh biến động của yếu tố kết quả do ảnh hưởng của yếu tố nguyên nhân thứ hai (xếp theo hàng)* | $SSH = K\sum_{j=1}^{H}(\overline{X}_j - \overline{X})^2$ |
| 4. Tổng các độ lệch bình phương phần dư (ERROR) *Phản ánh biến động của yếu tố kết quả do ảnh hưởng của yếu tố nguyên nhân khác không nghiên cứu* | $SSE = SST - SSK - SSH$ |

***Bước 3:*** Tính các phương sai

| Diễn giải | Công thức |
|---|---|
| 1. Phương sai giữa các nhóm (cột) (MSK) | $MSK = \frac{SSK}{K-1}$ |
| 2. Phương sai giữa các khối (hàng) (MSH) | $MSH = \frac{SSH}{H-1}$ |
| 3. Phương sai phần dư (MSE) | $MSE = \frac{SSE}{(K-1)(H-1)}$ |

***Bước 4***: Kiểm định giả thuyết

Tính tiêu chuẩn kiểm định F (F thực nghiệm)

| | |
|---|---|
| $F_1 = \frac{MSK}{MSE}$ | Trong đó: MSK là phương sai giữa các nhóm (cột)<br>MSE là phương sai phần dư<br>F1 dùng kiểm định cho yếu tố nguyên nhân thứ nhất |
| $F_2 = \frac{MSH}{MSE}$ | Trong đó: MSH là phương sai giữa các khối (hàng)<br>MSE là phương sai phần dư.<br>$F_2$ dùng kiểm định cho yếu tố nguyên nhân thứ hai |

Tìm F lý thuyết cho 2 yếu tố nguyên nhân

- Yếu tố nguyên nhân thứ nhất:

F tiêu chuẩn = F (k-1; (k-1)(h-1), α) là giá trị giới hạn tra từ bảng phân phối F với k-1 bậc tự do của phương sai ở tử số và (k-1)(h-1) bậc tự do của phương sai ở mẫu số với mức ý nghĩa α. F lý thuyết có thể tra qua hàm FINV(α, k-1,(k-1)(h-1)) trong EXCEL.

- Yếu tố nguyên nhân thứ hai:

F tiêu chuẩn = F (h-1; (k-1)(h-1), α) là giá trị giới hạn tra từ bảng phân phối F với h-1 bậc tự do của phương sai ở tử số và (k-1)(h-1) bậc tự do của phương sai ở mẫu số với mức ý nghĩa α. F lý thuyết có thể tra qua hàm FINV(α, h-1, (k-1)(h-1)) trong EXCEL.

Nếu F1 thực nghiệm > F1 lý thuyết, bác bỏ Ho, nghĩa là các số trung bình của k tổng thể nhóm (cột) không bằng nhau.

Nếu F2 thực nghiệm > F2 lý thuyết, bác bỏ Ho, nghĩa là các số trung bình của k tổng thể khối (hàng) không bằng nhau.

Bảng phân tích phương sai 2 yếu tố khi sử dụng

Phần mềm xử lý số liệu tóm tắt như sau:

| *Nguồn biến động* | *Tổng độ lệch bình phương (SS)* | *Bậc tự do (df)* | *Phương sai (MS)* | *F- Tỷ số* |
|---|---|---|---|---|
| Giữa các hàng | SSH | (h-1) | MSH | $F_1$ |
| Giữa các cột | SSK | (k -1) | MSK | $F_2$ |
| Phần dư | SSE | (k -1) (h-1) | MSE | |
| Tổng số | SST | (n-1) | | |

## 2.2. Phương pháp hồi quy

### 2.2.1. Khái niệm

"Hồi quy" (regression) là một kỹ thuật được sử dụng để hiểu mối quan hệ giữa một biến phụ thuộc (biến được dự đoán) và một hoặc nhiều biến độc lập (biến giải thích). Mục tiêu của hồi quy là tìm một mô hình toán học hoặc mối quan hệ số lượng giữa các biến này để có thể dự đoán giá trị của biến phụ thuộc khi biết giá trị của các biến độc lập.

### 2.2.2. Phương trình hồi quy tuyến tính bội

Trong hồi quy tuyến tính bội, mô hình cố gắng tìm một mối quan hệ tuyến tính giữa biến phụ thuộc và các biến độc lập bằng cách tạo ra một phương trình có dạng:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_r X_r + \varepsilon$$

Trong đó:

Y là biến phụ thuộc cần dự đoán.

$X_1$, $X_2$,...,$X_r$ là các biến độc lập.

$\beta_0$, $\beta_1$, $\beta_2$,...,$\beta_r$ là các hệ số hồi quy, thể hiện mức độ ảnh hưởng của mỗi biến độc lập lên biến phụ thuộc.

ε là sai số ngẫu nhiên có phân phối chuẩn $N \sim (0, \sigma^2)$

Ta biết rằng dù mô hình có nhiều biến độc lập nhưng vẫn tồn tại những yếu tố tác động đến biến phụ thuộc mà không được đưa vào mô hình vì nhiều lí do (không có số liệu hoặc không muốn đưa vào). Do đó mô hình vẫn tồn tại sai số ngẫu nhiên ε đại diện cho các yếu tố khác ngoài các biến $x_i$ (i=1, 2, 3,…k) có tác động đến Y nhưng không là biến số.

Xét một mẫu ngẫu nhiên với n quan sát cụ thể, ta có hồi quy mẫu như sau:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots\ldots + \hat{\beta}_k x_k + \varepsilon_i$$

Với $\varepsilon_i$ là phần dư tại quan sát I, được tính bởi công thức sau : $\varepsilon_i = Y_i - \hat{Y}_i$

**2.2.3. Ý nghĩa các hệ số hồi quy**

Xuất phát từ hàm hồi quy tổng thể:

$$(Y|x1, x2, \ldots, xk) = \beta 0 + \beta 1 x1 + \beta 2 x2 + \cdots + \beta k x\text{k}$$

Hệ số chặn (intercept - $\beta_0$): Đây là giá trị của biến phụ thuộc (Y) khi tất cả các biến độc lập (X) đều bằng 0. Thông thường, trong một ngữ cảnh thực tế, không phải lúc nào cũng có ý nghĩa thực tế mà nói rằng tất cả các biến độc lập đều bằng 0, do đó, hệ số chặn không luôn có ý nghĩa thực tế.

Hệ số của các biến độc lập (slope coefficients - $\beta_1$, $\beta_2$,...,$\beta_r$): Đây là mức độ thay đổi trung bình trong biến phụ thuộc (Y) khi một biến độc lập (X) tăng lên 1 đơn vị, giữa khi các biến độc lập còn lại không thay đổi. Chúng cho biết độ lớn và hướng của ảnh hưởng của từng biến độc lập lên biến phụ thuộc.

Có 3 khả năng có thể xảy ra với hệ số $\beta_i$:

1) $\beta_i > 0$: khi đó mối quan hệ giữa $Y$ và $x_i$ là thuận chiều, nghĩa là khi $x_i$ (tăng hay giảm) trong điều kiện các biến độc lập khác không thay đổi thì $Y$ cũng tăng (hoặc giảm).
2) $\beta_i < 0$: khi đó mối quan hệ giữa $Y$ và $x_i$ là nghịch chiều, nghĩa là khi $x_i$ (tăng hay giảm) trong điều kiện các biến độc lập khác không thay đổi thì $Y$ cũng giảm (hoặc tăng).
3) $\beta_i$=0: khi đó không có mối quan hệ tương quan giữa $Y$ và $x_i$, cụ thể là Y không phụ thuộc vào $x_i$, hay nói cách khác $x_i$ không ảnh hưởng đến $Y$.

$\varepsilon_i$ là phần sai lệch giữa giá trị của $Y$ trong phương trình và giá trị thực tế của $Y$. Thực chất, mô hình này thường chỉ dự đoán tốt kỳ vọng của $Y$ chứ không phải giá trị của $Y$ trong thực tế, hay nói cách khác $E(Y|_{x_i}\ theo\ các\ i)$= $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$, còn ε là một biến ngẫu nhiên có kì vọng là 0 và phương sai $\sigma^2$.

*Tương quan giữa các biến độc lập:* Nếu có một số biến độc lập liên quan chặt chẽ với nhau, các hệ số của chúng có thể bị ảnh hưởng. Điều này có thể dẫn đến hiện tượng gọi là đa cộng tuyến (multicollinearity), làm giảm khả năng giải thích của mô hình.

### 2.2.4. Xác định giá trị các tham số trong mô hình hồi quy tuyến tính bội

Có nhiều cách để xác định giá trị của các tham số, tuy nhiên, trong số đó, phương pháp bình phương cực tiểu (phương pháp OLS) là phương pháp thường được sử dụng nhất.

Xét mô hình $k$ biến: $(Y) = \beta 0 + \beta 1 x1 + \beta 2 x2 + \cdots + \beta k x k + \varepsilon_i$

Giả sử, có $n$ quan sát và $k$ biến hồi quy thỏa mãn $n > k$, đặt $x_{ij}$ là quan sát thứ i. Số quan sát là: $x_{i1}, x_{i2}, \ldots, x_{ik}$ với i=1,2,3,….,n (n > k)

Ta sẽ sử dụng thông tin từ mẫu để xây dựng các ước lượng cho các hệ số $\beta_i$ (i=0,1,2,…,k).

Tại mỗi quan sát I, hàm hồi quy mẫu được viết thành:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots\ldots\ldots + \hat{\beta}_k x_k + \varepsilon_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_0 x_{ij} + \varepsilon_i$$

Trong đó, $\hat{y}_i$ là các giá trị ước lượng cho $y_i$và sai lệch giữa hai giá trị này là phần dư:

$$\varepsilon_i = (y_i - \hat{y}_i)$$

Tương tự như mô hình hồi quy tuyến tính hai biến, phương pháp bình phương cực tiểu nhằm xác định các giá trị $\hat{\beta}_i$ sao cho tổng bình phương các phần tư là nhỏ nhất:

$$\text{L} = \sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{i=1}^{n} \beta_J x_i) \rightarrow \text{Min}$$

Khí đó, các giá trị $\hat{\beta}_0, \hat{\beta}_1, \ldots\ldots, \hat{\beta}_k$ sẽ là nghiệm của hệ gồm k phương trình sau:

$$\frac{\partial L}{\partial \hat{\beta}_0} = -2\left(\sum_{i=1}^{k} y_i - \hat{\beta}_0 - \sum_{i=1}^{k} \hat{\beta}_0 x_{ij}\right) = 0$$

$$\frac{\partial L}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{k} x_{i1} \left(y_i - \hat{\beta}_0 \sum_{i=1}^{k} \hat{\beta}_i x_{ij}\right) = 0 \quad \frac{\partial L}{\partial \hat{\beta}_k} = -2 \sum_{i=1}^{k} x_i$$

Đơn giản hệ phương trình ta được

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots\ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_1$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} x_{ik} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} x_{ik} + \ldots\ldots\ldots + \hat{\beta}_k \sum_{i=1}^{n} {x_{ik}}^2 = \sum_{i=1}^{n} y_1 x_{ik}$$

## 2.3 Tìm khoảng ước lượng và kiểm định 1 mẫu

### 1.1 Độ tin cậy và khoảng tin cậy:

Giả sử ta tìm được giá trị trung bình của một mẫu là k. Giờ ta muốn dự đoán giá trị trung bình của tổng thể. Xác suất mà giá trị trung bình của tổng thể thuộc khoảng (a;b) là β, ta nói β là độ tin cậy còn (a;b) là khoảng tin cậy. Mức ý nghĩa kí hiệu α và có tính chất α + β = 1.

### 1.2 Tìm khoảng tin cậy (khoảng ước lượng):

| Dạng | Điều kiện | Loại | Khoảng tin cậy |
|---|---|---|---|
| Tỷ lệ | | Trái | $p \in (f - z_\alpha.\frac{\sqrt{f(1-f)}}{\sqrt{n}};1)$ |
| | | Đối xứng | $p \in (f - \varepsilon; f + \varepsilon)$ với $\varepsilon = z_{\frac{\alpha}{2}}.\frac{\sqrt{f(1-f)}}{\sqrt{n}}$ |
| | | Phải | $p \in (0; f + z_\alpha.\frac{\sqrt{f(1-f)}}{\sqrt{n}})$ |
| Trung bình | Biết $\sigma^2$ , Phân phối chuẩn | Trái | $\mu \in (\overline{x} - z_\alpha.\frac{\sigma}{\sqrt{n}}; +\infty)$ |
| | | Đối xứng | $\mu \in (\overline{x} - \varepsilon; \overline{x} + \varepsilon)$ với $\varepsilon = z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}$ |
| | | Phải | $\mu \in (-\infty; \overline{x} + z_\alpha.\frac{\sigma}{\sqrt{n}})$ |
| | Chưa biết $\sigma^2$, tìm được s và n<30, phân phối chuẩn | Trái | $\mu \in (\overline{x} - t_\alpha; n-1.\frac{s}{\sqrt{n}}; +\infty)$ |
| | | Đối xứng | $\mu \in (\overline{x} - \varepsilon; \overline{x} + \varepsilon)$ với $\varepsilon = t_{\frac{\alpha}{2}}.n-1.\frac{s}{\sqrt{n}}$ |
| | | Phải | $\mu \in (-\infty; \overline{x} + t_\alpha; n-1.\frac{s}{\sqrt{n}})$ |
| | Chưa biết $\sigma^2$, tìm được s và n≥30 | Trái | $\mu \in (\overline{x} - z_\alpha.\frac{s}{\sqrt{n}}; +\infty)$ |
| | | Đối xứng | $\mu \in (\overline{x} - \varepsilon; \overline{x} + \varepsilon)$ với $\varepsilon = z_{\frac{\alpha}{2}}.\frac{s}{\sqrt{n}}$ |
| | | Phải | $\mu \in (-\infty; \overline{x} + z_\alpha.\frac{s}{\sqrt{n}})$ |
| Phương sai | Tìm được s | Đối xứng | $\sigma^2 \in (\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};n-1}}; \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};n-1}})$ |

Giải thích một số kí hiệu:

- Độ chính xác của ước lượng có kí hiệu ε (tên gọi khác là bán kính, ngưỡng sai số của ước lượng)

- Chiều dài khoảng ước lượng có kí hiệu 2. E

Cách tính một số biến trong công thức:

- Tính $z_{\frac{a}{2}}$ dùng bảng phân phối chuẩn tra ngược giá trị, sao cho:

$$\Phi\left(z_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$$

- Tính $z_a$ dùng bảng phân phối chuẩn tra ngược giá trị, sao cho:

$$\Phi\left(z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

- Tính $z_{\frac{a}{2}}$. $n-1$ dùng bảng student tra giá trị hàng n-1 và cột $\frac{a}{2}$

- Tính $t_a$. $n-1$ dùng bảng student tra giá trị hàng n-1 và cột α

- Tính $\chi^2_{\frac{\alpha}{2}}$;n-1 dùng bảng chi bình phương tại hàng n-1 và cột $\frac{\alpha}{2}$
- Tính $\chi^2_{1-\frac{\alpha}{2}}$;n-1 dùng bảng chi bình phương tại hàng n-1 và cột 1-$\frac{\alpha}{2}$

**1.3 Định nghĩa về kiểm định 1 mẫu:**

- Một giả thuyết thống kê, hay gọi tắt là giả thuyết, là một phát biểu hay khẳng định về giá trị của một tham số, một vài tham số, hay dạng của phân phối của một quần thể.

- Giả thuyết $Ho$ là một giả thuyết mà ban đầu được giả định là đúng.

- Giả thuyết $H1$ là giả thuyết ngược lại với $Ho$.

- Ở giả thuyết $Ho$ ta thường ưu tiên sử dụng dấu "=" vì nó sẽ giúp giả thuyết trở nên đơn giản hơn khi ta chỉ cần chú ý đến điều kiện của $H1$.

- Một kiểm định thống kê là một phương pháp sử dụng dữ liệu mẫu để kiểm tra xem ta có thể bác bỏ $Ho$ hay không.

- Giả thuyết $Ho$ sẽ bị bác bỏ chỉ khi mẫu cho thấy $Ho$ sai. Hai kết luận có thể có từ một kiểm định thống kê là thác bỏ $Ho$ hay không thể bác bỏ $Ho$.

- RR là miền bác bỏ để xét xem tiêu chuẩn kiểm định có thuộc miền bác bỏ hay không để từ đó có thể bác bỏ $Ho$ hay không.

**1.4 Kiểm định một mẫu:**

**a. Kiểm định tỉ lệ một mẫu:**

Xác định giả thuyết $H1$ và $Ho$

Tùy trường hợp mà sử dụng các công thức sau

| Giả thuyết $H_O$ | Giả thuyết đối $H_1$ | Tiêu chuẩn kiểm định | Miền bác bỏ |
|---|---|---|---|

| | | | |
|---|---|---|---|
| $p = p_0$ | $p \neq p_0$ | $z_{qs} = \frac{f - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ | RR = $(-\infty; -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}; +\infty)$ |
| $p = p_0$ hoặc $p \leq p_0$ | $p > p_0$ | | RR = $(z_\alpha; +\infty)$ |
| $p = p_0$ hoặc $p \geq p_0$ | $p < p_0$ | | RR = $(-\infty; -z_\alpha)$ |

Kết luận:

- Nếu $z_{qs} \in$ RR thì ta có thể bác bỏ giả thuyết $H_O$ và chấp nhận giả thuyết $H_1$ .
- Nếu $z_{qs} \notin$ RR Thì ta không bác bỏ ý kiến $H_O$ (chưa có đủ bằng chứng để bác bỏ $H_O$).

b. Kiểm định trung bình một mẫu:

Xác định giả thuyết $H_1$ và $H_O$

Tùy trường hợp mà sử dụng các công thức sau

| Dạng | Giả thuyết $H_O$ | Giả thuyết đối $H_1$ | Tiêu chuẩn kiểm định | Miền bác bỏ $H_O$ |
|---|---|---|---|---|
| Có phân phối chuẩn và đã biết $\sigma^2$ | $\mu = \mu_0$ | $\mu \neq \mu_0$ | $z_{qs} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ | RR = $(-\infty; -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}; +\infty)$ |
| | | $\mu > \mu_0$ | | RR = $(z_\alpha; +\infty)$ |
| | | $\mu < \mu_0$ | | RR = $(-\infty; -z_\alpha)$ |
| Có phân phối chuẩn và chưa biết $\sigma^2$, n<30 | $\mu = \mu_0$ | $\mu \neq \mu_0$ | $T_{qs} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$ | RR = $(-\infty; -t_{\frac{\alpha}{2}}.n-1) \cup (t_{\frac{\alpha}{2}}.n-1; +\infty)$ |
| | | $\mu > \mu_0$ | | RR = $(t_{\frac{\alpha}{2}}.n-1; +\infty)$ |
| | | $\mu < \mu_0$ | | RR = $(-\infty; -t_{\frac{\alpha}{2}}.n-1)$ |
| Có phân phối tùy ý và chưa biết $\sigma^2$, n≥30 | | $\mu \neq \mu_0$ | $z_{qs} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$ | RR = $(-\infty; -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}; +\infty)$ |
| | | $\mu > \mu_0$ | | RR = $(z_\alpha; +\infty)$ |
| | | $\mu < \mu_0$ | | RR = $(-\infty; -z_\alpha)$ |

Kết luận

- Nếu $z_{qs} \in$ RR thì ta có thể bác bỏ giả thuyết $H_O$ và chấp nhận giả thuyết $H_1$ .
- Nếu $z_{qs} \notin$ RR Thì ta không bác bỏ ý kiến $H_O$ (chưa có đủ bằng chứng để bác bỏ $H_O$).

c. Kiểm định phương sai một mẫu:

Xác định giả thuyết $H_1$ và $H_O$

Tùy trường hợp mà sử dụng các công thức sau

| Giả thuyết $H_O$ | Giả thuyết đối $H_1$ | Tiêu chuẩn kiểm định | Miền bác bỏ |
|---|---|---|---|
| $\sigma^2 = \sigma^2_0$ | $\sigma^2 \neq \sigma^2_0$ | $\chi^2_0 = \frac{(n-1)s^2}{\sigma^2_0}$ | RR = $(0; \chi^2_{1-\frac{\alpha}{2}};n-1) \cup (\chi^2_{1-\frac{\alpha}{2}};n-1; +\infty)$ |
| $\sigma^2 = \sigma^2_0$ hoặc $\sigma^2 \leq \sigma^2_0$ | $\sigma^2 > \sigma^2_0$ | | RR = $(\chi^2_{1-\frac{\alpha}{2}};n-1; +\infty)$ |
| $\sigma^2 = \sigma^2_0$ hoặc $\sigma^2 \geq \sigma^2_0$ | $\sigma^2 < \sigma^2_0$ | | RR = $(0; \chi^2_{1-\frac{\alpha}{2}};n-1)$ |

Kết luận

- Nếu $z_{qs} \in$ RR thì ta có thể bác bỏ giả thuyết $H_O$ và chấp nhận giả thuyết $H_1$ .
- Nếu $z_{qs} \notin$ RR Thì ta không bác bỏ ý kiến $H_O$ (chưa có đủ bằng chứng để bác bỏ $H_O$).

## 2. Kiểm định mẫu

Kiểm định 2 mẫu là một công cụ quan trọng trong thống kê, giúp chúng ta so sánh trung bình hoặc tỉ lệ của hai nhóm dữ liệu để xác định xem liệu có sự khác biệt ý nghĩa giữa chúng hay không. Chúng ta thường sử dụng kiểm định này khi có nhu cầu so sánh hiệu quả của hai điều kiện, nhóm, hoặc biến số khác nhau.

-Giả thiết không $H_0$ : (Null Hypothesis) là giả thiết về yếu tố cần kiểm định của tổng thể ở trạng thái bình thường, không chịu tác động của các hiện tượng liên quan. Yếu tố trong $H_0$ phải được xác định cụ thể.

- Giả thiết đối $H_1$ (Alternative Hypothesis) là một mệnh đề mâu thuẫn với $H_0$ , $H_1$ thể hiện xu hướng cần kiểm định. Vì ta sẽ dựa vào thông tin thực nghiệm của mẫu để kết luận xem có thừa nhận các giả thiết nêu trên hay không nên công việc này gọi là kiểm định thống kê.

-Tiêu chuẩn kiểm định là hàm thống kê G = G( $X_1$ ,$X_2$ , ..,$X_n$ , $\theta_0$ ), xây dựng trên mẫu

ngẫu nhiên W= ( $X_1$ , $X_2$ , .., $X_n$ ) và tham số $\theta_0$ liên quan đến $H_0$ ; Điều kiện đặt ra với thống kê G là nếu H0 đúng thì quy luật phân phối xác suất của G phải hoàn toàn xác định.

- Miền bác bỏ giả thiết RR ( Rejection region) là miền số thực thỏa P(G∈RR /$H_0$ đúng) =α. α là một số khá bé, thường không quá 10% và được gọi là mức ý nghĩa của kiểm định. Một ký hiệu khác của miền bác bỏ được dùng trong bài: $W_\alpha$

- Miền chấp nhận AR: phần bù của miền bác bỏ trong R.

- Quy tắc kiểm định: Từ mẫu thực nghiệm, ta tính được một giá trị cụ thể của tiêu chuẩn kiểm định, gọi là giá trị kiểm định thống kê: gqs = G($x_1$ , $x_2$ , .., $x_n$ , $\theta_2$ ) . Theo nguyên lý xác suất bé, biến cố G ∈ RR có xác suất nhỏ nên với 1 mẫu thực nghiệm ngẫu nhiên, nó

không thể xảy ra.

Do đó:

+ Nếu gqs ∈ RR thì bác bỏ $H_0$ , thừa nhận giả thiết $H_1$ .

+ Nếu gqs ∉ RR : ta chưa đủ dữ liệu khẳng định $H_0$ sai. Vì vậy ta chưa thể chứng minh được $H_1$ đúng.

Công thức kiểm định 2 mẫu

a) Kiểm định tỉ lệ:

| | GT không $H_0$ | GT đối $H_1$ | Tiêu chuẩn kiểm định | Miền bác bỏ với mức ý nghĩa $\alpha$ |
|---|---|---|---|---|
| BT<br>2 mẫu<br>$n_1 \geq 30$<br>$n_2 \geq 30$ | $p_1 = p_2$ | $p_1 \neq p_2$ | $Z_{qs} = \dfrac{f_1 - f_2}{\sqrt{\bar{f}(1-\bar{f})(\frac{1}{n_1}+\frac{1}{n_2})}}$ | $RR = (-\infty, -z_{\frac{\alpha}{2}}) \cup (-z_{\frac{\alpha}{2}}, +\infty)$ |
| | | $p_1 < p_2$ | | $RR = (-\infty, -z_\alpha)$ |
| | | $p_1 > p_2$ | | $RR = (z_\alpha, +\infty)$ |

$$\text{Tỉ lệ mẫu gộp: } \bar{f} = \frac{m_1 + m_2}{n_1 + n_2} = \frac{f_1.n_1 + f_2.n_2}{n_1 + n_2}$$

b) Kiểm định trung bình

| TT | Phân bố của tổng thể | GT $H_0$ | GT $H_1$ | Miền bác bỏ RR | Tiêu chuẩn kiểm định |
|---|---|---|---|---|---|
| a | *2 mẫu độc lập<br>* $X_1;X_2$ có pp chuẩn<br>*Đã biết $\sigma_1^2$ và $\sigma_2^2$ | $\mu_1 = \mu_2$ | $\mu_1 \neq \mu_2$ | $(-\infty, -z_{\frac{\alpha}{2}}) \cup (-z_{\frac{\alpha}{2}}, +\infty)$ | $Z_{qs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ |
| | | | $\mu_1 < \mu_2$ | $(-\infty, -z_\alpha)$ | |
| | | | $\mu_1 > \mu_2$ | $(z_\alpha, +\infty)$ | |
| b | *2 mẫu độc lập<br>* $X_1;X_2$ có pp chuẩn<br>*Chưa biết $\sigma_1^2, \sigma_2^2$; gt $\sigma_1^2 = \sigma_2^2$<br>Dấu hiệu quy ước để nhận biết trương hợp b từ mẫu $\frac{s_1}{s_2} \in \left[\frac{1}{2}; 2\right]$ | $\mu_1 = \mu_2$ | $\mu_1 \neq \mu_2$ | $\left(-\infty, -t_{\frac{\alpha}{2}}(n_1 + n_1 - 2)\right) \cup \left(t_{\frac{\alpha}{2}}(n_1 + n_1 - 2), +\infty\right)$ | $T_{qs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$ |
| | | | $\mu_1 < \mu_2$ | $(-\infty, -t_\alpha(n_1 + n_1 - 2))$ | |
| | | | $\mu_1 > \mu_2$ | $(t_\alpha(n_1 + n_1 - 2), +\infty)$ | |
| c | *2 mẫu độc lập<br>* $X_1;X_2$ có pp chuẩn<br>*Chưa biết $\sigma_1^2, \sigma_2^2$; gt $\sigma_1^2 \neq \sigma_2^2$<br>Dấu hiệu quy ước để nhận biết trương hợp c từ mẫu $\frac{s_1}{s_2} \notin \left[\frac{1}{2}; 2\right]$ | $\mu_1 = \mu_2$ | $\mu_1 \neq \mu_2$ | $\left(-\infty, -t_{\frac{\alpha}{2}}(V)\right) \cup \left(t_{\frac{\alpha}{2}}(V), +\infty\right)$ | $T_{qs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$ |
| | | | $\mu_1 < \mu_2$ | $(-\infty, -t_\alpha(V))$ | |
| | | | $\mu_1 > \mu_2$ | $(t_\alpha(V), +\infty)$ | |

| TT | Phân bố của tổng thể | GT $H_0$ | GT $H_1$ | Miền bác bỏ RR | Tiêu chuẩn kiểm định |
|---|---|---|---|---|---|
| d | *2 mẫu độc lập<br>* $X_1;X_2$ có pp tùy ý<br>*Đã biết $\sigma_1^2$ và $\sigma_2^2$ | $\mu_1 = \mu_2$ | $\mu_1 \neq \mu_2$ | $(-\infty, -z_{\frac{\alpha}{2}}) \cup (-z_{\frac{\alpha}{2}}, +\infty)$ | $Z_{qs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$<br>Nếu chưa biết $\sigma$ thì dùng $s$ |
| | | | $\mu_1 < \mu_2$ | $(-\infty, -z_\alpha)$ | |
| | | | $\mu_1 > \mu_2$ | $(z_\alpha, +\infty)$ | |
| e | *2 mẫu phụ thuộc tương ứng theo cặp<br>* $X_1;X_2$ có pp chuẩn<br>*Chưa biết $\sigma_D^2$ | $\mu_1 = \mu_2$ hay $\mu_D = 0$ | $\mu_1 \neq \mu_2$ | $\left(-\infty, -t_{\frac{\alpha}{2}}(n-1)\right) \cup (t_{\frac{\alpha}{2}}(n-1), +\infty)$ | Đặt $D=X_1-X_2$<br>$T_{qs} = \frac{\bar{D}}{S_D}\sqrt{n}$ |
| | | | $\mu_1 < \mu_2$ | $(-\infty, -t_\alpha(n-1))$ | |
| | | | $\mu_1 > \mu_2$ | $(t_\alpha(n-1), +\infty)$ | |
| f | *2 mẫu phụ thuộc tương ứng theo cặp<br>*2 mẫu có n>30<br>* $X_1;X_2$ có pp chuẩn<br>*D có phân phối tùy ý<br>*Chưa biết hoặc đã biết $\sigma_D^2$ | $\mu_1 = \mu_2$ hay $\mu_D = 0$ | $\mu_1 \neq \mu_2$ | $(-\infty, -z_{\frac{\alpha}{2}}) \cup (-z_{\frac{\alpha}{2}}, +\infty)$ | Đặt $D=X_1-X_2$<br>$Z_{qs} = \frac{\bar{D}}{\sigma_D}\sqrt{n}$<br>Nếu chưa biết $\sigma$ thì dùng $s$ |
| | | | $\mu_1 < \mu_2$ | $(-\infty, -z_\alpha)$ | |
| | | | $\mu_1 > \mu_2$ | $(z_\alpha, +\infty)$ | |

# 3. Tiền xử lí dữ liệu

## 3.1. Đọc dữ liệu

Đọc tên tệp "Air.csv" vào khung dữ liệu (được ghi chú là oridata trong R).

| | index | Activity.Period | Operating.Airline | Operating.Airline.IATA.Code | Published.Airline | Published.Airline.IATA.Code | GEO.Summary | GEO.Region |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 200507 | ATA Airlines | TZ | ATA Airlines | TZ | Domestic | US |
| 2 | 1 | 200507 | ATA Airlines | TZ | ATA Airlines | TZ | Domestic | US |
| 3 | 2 | 200507 | ATA Airlines | TZ | ATA Airlines | TZ | Domestic | US |
| 4 | 3 | 200507 | Air Canada | AC | Air Canada | AC | International | Canada |
| 5 | 4 | 200507 | Air Canada | AC | Air Canada | AC | International | Canada |
| 6 | 5 | 200507 | Air China | CA | Air China | CA | International | Asia |
| 7 | 6 | 200507 | Air China | CA | Air China | CA | International | Asia |
| 8 | 7 | 200507 | Air France | AF | Air France | AF | International | Europe |
| 9 | 8 | 200507 | Air France | AF | Air France | AF | International | Europe |
| 10 | 9 | 200507 | Air New Zealand | NZ | Air New Zealand | NZ | International | Australia / Oceania |
| 11 | 10 | 200507 | Air New Zealand | NZ | Air New Zealand | NZ | International | Australia / Oceania |
| 12 | 11 | 200507 | AirTran Airways | FL | AirTran Airways | FL | Domestic | US |
| 13 | 12 | 200507 | AirTran Airways | FL | AirTran Airways | FL | Domestic | US |
| 14 | 13 | 200507 | Alaska Airlines | AS | Alaska Airlines | AS | Domestic | US |
| 15 | 14 | 200507 | Alaska Airlines | AS | Alaska Airlines | AS | Domestic | US |
| 16 | 15 | 200507 | Alaska Airlines | AS | Alaska Airlines | AS | Domestic | US |

| Region | Activity.Type.Code | Price.Category.Code | Terminal | Boarding.Area | Passenger.Count | Adjusted.Activity.Type.Code | Adjusted.Passenger.Count | Year | Month |
|---|---|---|---|---|---|---|---|---|---|
| | Deplaned | Low Fare | Terminal 1 | B | 27271 | Deplaned | 27271 | 2005 | July |
| | Enplaned | Low Fare | Terminal 1 | B | 29131 | Enplaned | 29131 | 2005 | July |
| | Thru / Transit | Low Fare | Terminal 1 | B | 5415 | Thru / Transit * 2 | 10830 | 2005 | July |
| da | Deplaned | Other | Terminal 1 | B | 35156 | Deplaned | 35156 | 2005 | July |
| da | Enplaned | Other | Terminal 1 | B | 34090 | Enplaned | 34090 | 2005 | July |
| | Deplaned | Other | International | G | 6263 | Deplaned | 6263 | 2005 | July |
| | Enplaned | Other | International | G | 5500 | Enplaned | 5500 | 2005 | July |
| e | Deplaned | Other | International | A | 12050 | Deplaned | 12050 | 2005 | July |
| e | Enplaned | Other | International | A | 11638 | Enplaned | 11638 | 2005 | July |
| alia / Oceania | Deplaned | Other | International | G | 4998 | Deplaned | 4998 | 2005 | July |
| alia / Oceania | Enplaned | Other | International | G | 4962 | Enplaned | 4962 | 2005 | July |
| | Deplaned | Low Fare | International | A | 8055 | Deplaned | 8055 | 2005 | July |
| | Enplaned | Low Fare | International | A | 7984 | Enplaned | 7984 | 2005 | July |
| | Deplaned | Other | International | A | 36641 | Deplaned | 36641 | 2005 | July |
| | Enplaned | Other | International | A | 39379 | Enplaned | 39379 | 2005 | July |
| | Thru / Transit | Other | International | A | 3678 | Thru / Transit * 2 | 7356 | 2005 | July |

*Hình 1: khung dữ liệu ban đầu*

## 3.2. Làm sạch dữ liệu

Như chúng ta thấy, dữ liệu trên có tổng cộng 17 biến, trong đó có 1 số biến không cần thiết. Do vậy, ta cần loại bỏ các biến không cần bằng cách tạo 1 tệp mới bao gồm các biến chính.

| | Operating.Airline.IATA.Code | GEO.Summary | GEO.Region | Price.Category.Code | Terminal | Boarding.Area | Adjusted.Activity.Type.Code | Adjusted.Passenger.Count | Year | Month |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | TZ | Domestic | US | Low Fare | Terminal 1 | B | Deplaned | 27271 | 2005 | July |
| 2 | TZ | Domestic | US | Low Fare | Terminal 1 | B | Enplaned | 29131 | 2005 | July |
| 3 | TZ | Domestic | US | Low Fare | Terminal 1 | B | Thru / Transit * 2 | 10830 | 2005 | July |
| 4 | AC | International | Canada | Other | Terminal 1 | B | Deplaned | 35156 | 2005 | July |
| 5 | AC | International | Canada | Other | Terminal 1 | B | Enplaned | 34090 | 2005 | July |
| 6 | CA | International | Asia | Other | International | G | Deplaned | 6263 | 2005 | July |
| 7 | CA | International | Asia | Other | International | G | Enplaned | 5500 | 2005 | July |
| 8 | AF | International | Europe | Other | International | A | Deplaned | 12050 | 2005 | July |
| 9 | AF | International | Europe | Other | International | A | Enplaned | 11638 | 2005 | July |
| 10 | NZ | International | Australia / Oceania | Other | International | G | Deplaned | 4998 | 2005 | July |
| 11 | NZ | International | Australia / Oceania | Other | International | G | Enplaned | 4962 | 2005 | July |
| 12 | FL | Domestic | US | Low Fare | International | A | Deplaned | 8055 | 2005 | July |
| 13 | FL | Domestic | US | Low Fare | International | A | Enplaned | 7984 | 2005 | July |
| 14 | AS | Domestic | US | Other | International | A | Deplaned | 36641 | 2005 | July |
| 15 | AS | Domestic | US | Other | International | A | Enplaned | 39379 | 2005 | July |
| 16 | AS | Domestic | US | Other | International | A | Thru / Transit * 2 | 7356 | 2005 | July |
| 17 | AS | International | Canada | Other | International | A | Deplaned | 7977 | 2005 | July |

*Hình 2: Khung dữ liệu sau khi lọc*

### 3.3. Kiểm tra dữ liệu khuyết và biến đổi dữ liệu

Để đảm bảo không có vấn đề gì trong tập dữ liệu này, nhóm chúng em triển khai một số chức năng để kiểm tra xem có dữ liệu bị thiếu hay không và biến đổi dữ liệu "month" sang dạng số.

```
> apply(is.na(newdata),2,which)
integer(0)
```

Kết quả là không có dữ liệu khuyết.

Ta đổi cột month sang dạng số để thuận tiện cho việc tính toán.

```
> newdata$Month <- match(newdata$Month, month.name)
```

# 4. Thống kê mô tả

## 4.1. Dữ liệu sau tóm tắt

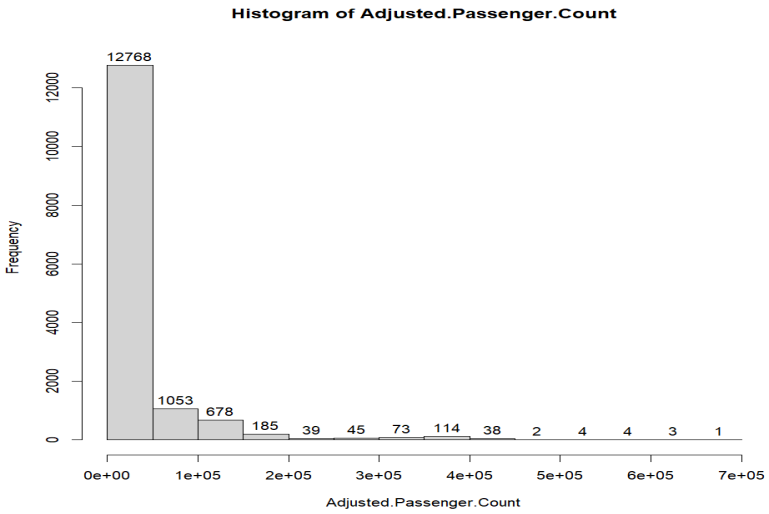
Sau khi thực hiện quá trình dọn dẹp, hiện tại chúng ta đã có một bộ dữ liệu rõ ràng và sạch sẽ trong newdata. Tóm tắt bằng cách sử dụng hàm summary trong R.

```
Adjusted.Passenger.Count      Year          Month
Min.   :     1           Min.   :2005   Min.   : 1.000
1st Qu.:  5496           1st Qu.:2008   1st Qu.: 3.000
Median :  9354           Median :2010   Median : 7.000
Mean   : 29332           Mean   :2010   Mean   : 6.551
3rd Qu.: 21182           3rd Qu.:2013   3rd Qu.:10.000
Max.   :659837           Max.   :2016   Max.   :12.000
```

*Hình 3: dữ liệu sau tóm tắt*

## 4.2. Vẽ các đồ thị biểu diễn

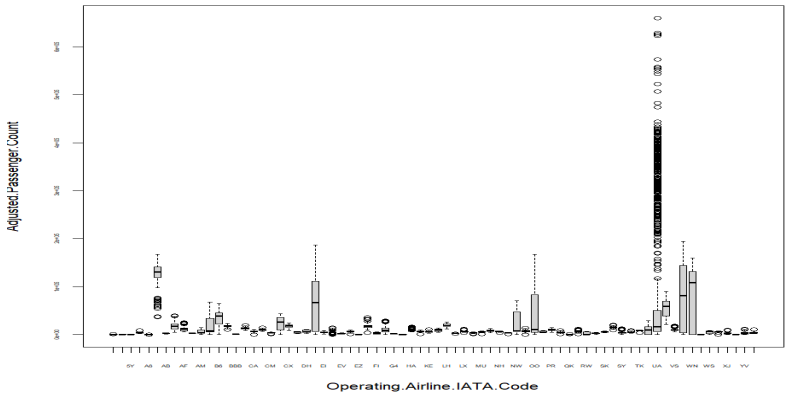
Đồ thị *Hist* thể hiện phân phối của số lượng hàng khách




*Hình 4: phân bố của số lượng hàng khách*

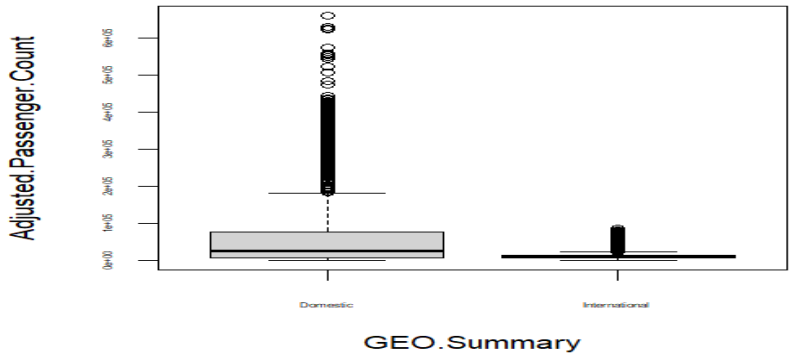
- Nhận xét: đồ thị không tuân theo phân phối chuẩn, có phân bố lệch phải, chứng tỏ có một số hãng bay tiếp nhận khách hàng trong tháng cao bất thường.

Đồ thị *Boxplot* thể hiện phân phối của số lượng hàng khách theo các biến phân loại:



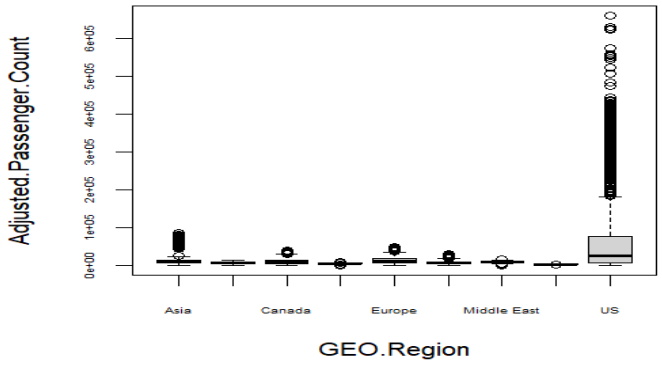

*Hình 5: phân phối số lượng hàng khách theo hãng bay*

- Nhận xét:
    - Hầu hết các hãng bay đều có giá trị ngoại lai
    - Các hãng bay có phân phối không đều



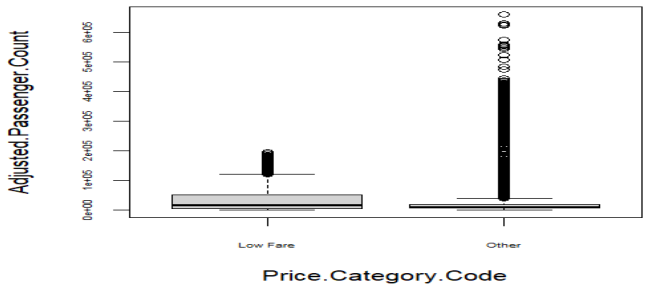

*Hình 6: phân phối số lượng hàng khách theo nhu cầu đi trong nước và quốc tế*

- Nhận xét:
    - Dữ liệu có giá trị ngoại biên.
    - Domestic có phân phối hàng khách không đều trong khi International thì ngược lại.
    - Domestic có phân phối bị lệch phải, có nghĩa có các chuyến bay có lượng hàng khách cao bất thường.
    - Domestic có hộp cao hơn International, chứng tỏ lượng hàng khách cao vượt trội.



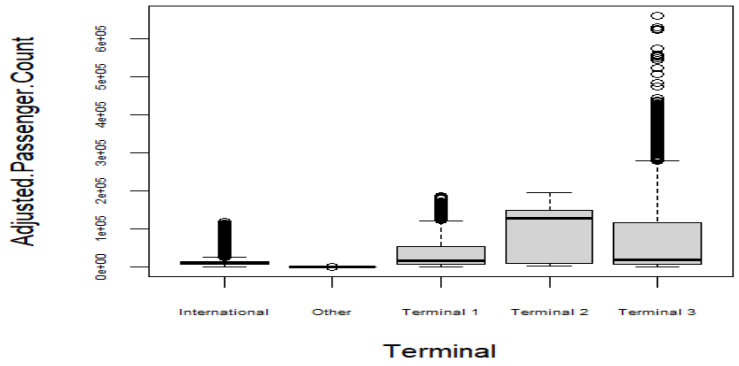

*Hình 7: phân phối số lượng hàng khách theo vùng địa lý*

- Nhận xét:
    - các khu vực đều có đường trung vị bằng nhau
    - khu vực US có hộp cao hơn so với các khu vực khác. Cho thấy khu vực này có lượng hàng khách vượt trội hơn các khu vực.



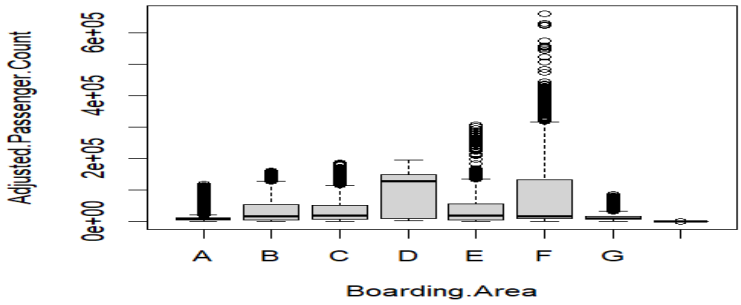

*Hình 8: phân phối số lượng hàng khách theo mã loại giá vé*

- Nhận xét:
    - các loại vé đều có trung vị bằng nhau
    - Low Fare có hộp cao hơn những chứng tỏ lượng hàng khách sử dụng vé Low Fare mà không có ngoại lai cao hơn những loại khác



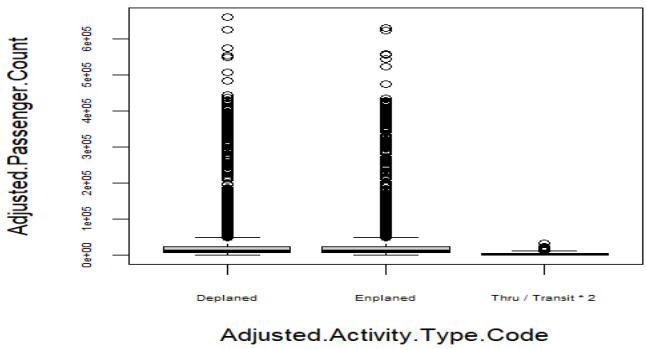

*Hình 9: phân phối số lượng hàng khách theo khu vực chờ*

- Nhận xét:
    - Các hộp có giá trị trung vị gần như bằng nhau, ngoại trừ Terminal 2
    - Terminal 2 có hộp cao nhất



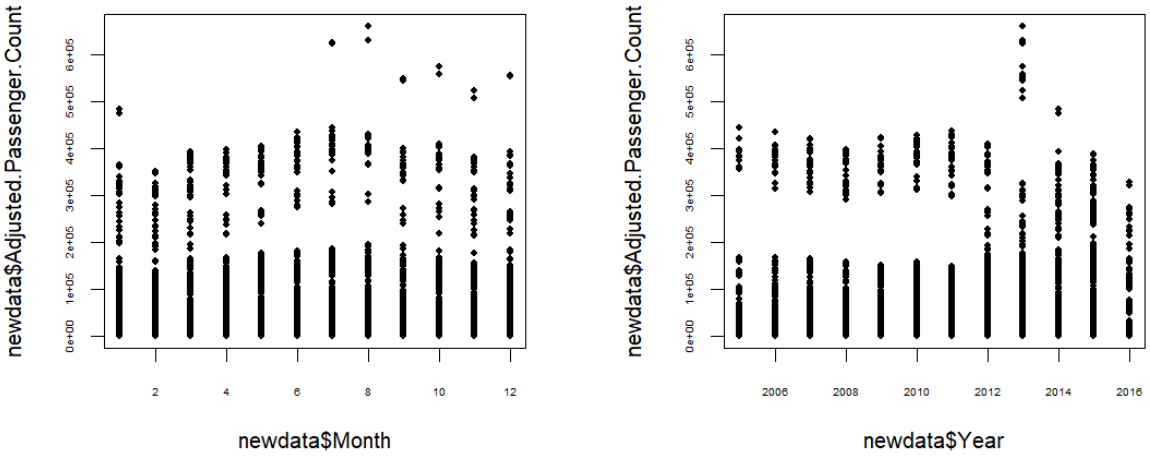

*Hình 10: phân phối số lượng hàng khách theo Boarding Area*

- Nhận xét:
    - Các hộp có giá trị trung vị gần như bằng nhau, ngoại trừ D
    - D có hộp cao nhất




*Hình 11: Phân phối số lượng hàng khách theo mã trạng thái hoạt động*

Đồ thị *plot* thể hiện sự phân tán của số lượng hàng khách theo tháng và năm:




*Hình 12,13: đồ thị phân tán lượng hàng khách theo tháng và năm*

- Nhận xét: ta nhận thấy không có mối quan hệ tuyến tính giữa lượng hàng khách so với tháng và năm.
- Lượng hàng khách theo tháng và năm có phân bố đều.

# 5. Thống kê suy diễn

## 5.1. Xây dựng mô hình anova 1 nhân tố:

Vì lý do tệp dữ liệu tương đối lớn và có tương đối nhiều hãng bay ở các khu vực khác nhau, nên trong bài toán này ta chọn lọc dữ liệu và sử dụng phương pháp anova để so sánh số lượng hành khách trong ở các hãng bay khu vực địa lý Canada.

### 5.1.1. Lọc dữ liệu khu vực Canada:

| | Operating.Airline.IATA.Code | GEO.Summary | GEO.Region | Price.Category.Code | Terminal | Boarding.Area | Adjusted.Activity.Type.Code | Adjusted.Passenger.Count | Year | Month |
|---|---|---|---|---|---|---|---|---|---|---|
| 4 | AC | International | Canada | Other | Terminal 1 | B | Deplaned | 35156 | 2005 | 7 |
| 5 | AC | International | Canada | Other | Terminal 1 | B | Enplaned | 34090 | 2005 | 7 |
| 17 | AS | International | Canada | Other | International | A | Deplaned | 7977 | 2005 | 7 |
| 18 | AS | International | Canada | Other | International | A | Enplaned | 8837 | 2005 | 7 |
| 80 | OO | International | Canada | Other | Terminal 3 | F | Deplaned | 3688 | 2005 | 7 |
| 81 | OO | International | Canada | Other | Terminal 3 | F | Enplaned | 3633 | 2005 | 7 |
| 103 | UA | International | Canada | Other | Terminal 3 | F | Deplaned | 12874 | 2005 | 7 |
| 104 | UA | International | Canada | Other | Terminal 3 | F | Enplaned | 12085 | 2005 | 7 |
| 105 | UA | International | Canada | Other | Terminal 3 | F | Thru / Transit * 2 | 20 | 2005 | 7 |
| 114 | WS | International | Canada | Other | International | A | Deplaned | 2480 | 2005 | 7 |
| 115 | WS | International | Canada | Other | International | A | Enplaned | 2211 | 2005 | 7 |
| 119 | AC | International | Canada | Other | Terminal 1 | B | Deplaned | 34878 | 2005 | 8 |
| 120 | AC | International | Canada | Other | Terminal 1 | B | Enplaned | 34105 | 2005 | 8 |
| 132 | AS | International | Canada | Other | International | A | Deplaned | 8901 | 2005 | 8 |
| 133 | AS | International | Canada | Other | International | A | Enplaned | 9020 | 2005 | 8 |
| 196 | OO | International | Canada | Other | Terminal 3 | F | Deplaned | 3672 | 2005 | 8 |
| 197 | OO | International | Canada | Other | Terminal 3 | F | Enplaned | 3570 | 2005 | 8 |
| 222 | UA | International | Canada | Other | Terminal 3 | F | Deplaned | 13895 | 2005 | 8 |
| 223 | UA | International | Canada | Other | Terminal 3 | F | Enplaned | 12145 | 2005 | 8 |
| 232 | WS | International | Canada | Other | International | A | Deplaned | 2300 | 2005 | 8 |
| 233 | WS | International | Canada | Other | International | A | Enplaned | 2251 | 2005 | 8 |
| 237 | AC | International | Canada | Other | Terminal 1 | B | Deplaned | 24473 | 2005 | 9 |
| 238 | AC | International | Canada | Other | Terminal 1 | B | Enplaned | 26120 | 2005 | 9 |
| 250 | AS | International | Canada | Other | International | A | Deplaned | 6635 | 2005 | 9 |
| 251 | AS | International | Canada | Other | International | A | Enplaned | 7783 | 2005 | 9 |
| 252 | AS | International | Canada | Other | International | A | Thru / Transit * 2 | 322 | 2005 | 9 |

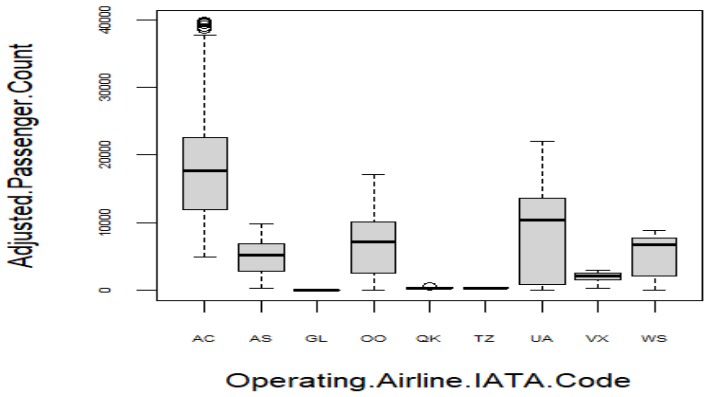
*Hình 14: dữ liệu các hãng bay ở Canada*

Kiểm tra các hãng hàng không hoạt động ở Canada

[1] "AC" "AS" "OO" "UA" "WS" "GL" "TZ" "QK" "VX"

- Nhận xét: có 9 hãng khàng không đang hoạt động tại Canada

### 5.1.2. Vẽ đồ thị Boxplot thể hiện phân phối hàng khách ở các hãng hàng không ở Canada




*Hình 15: biểu đồ phân bố lượng hàng khách ở các hãng bay Canada*

- Nhận xét:
    - các hãng GL, QK, TZ có lượng hàng khách tương đối nhỏ so với các hãng khác.
    - Hãng AC có hộp cao vượt trội so với các hãng khác, chứng tỏ khu vực Canada hãng bay AC được ưa chuộng nhất tại đây.

### 5.1.3. Kiểm tra các giả định

- Giả định 1: Giả định về phân phối chuẩn: số lượng hàng khách ở các hãng hàng không trên các chuyến bay có phân phối chuẩn
- Giả định 2: Tính đồng nhất của phương sai: Phương sai về số lượng hành khách ở các chuyến bay ở các hãng hàng không bằng nhau.

*Kiểm tra giả định 1:*

*$H_0$: số lượng hàng khách tuân theo phân phối chuẩn*

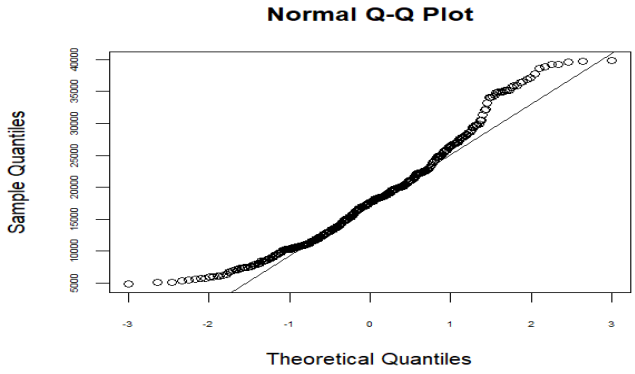
*$H_1$: số lượng hàng khách không tuân theo phân phối chuẩn*

Hãng AC

```
        Shapiro-Wilk normality test

data:  ACdata$Adjusted.Passenger.Count
```

```
W = 0.95796, p-value = 9.85e-09
```



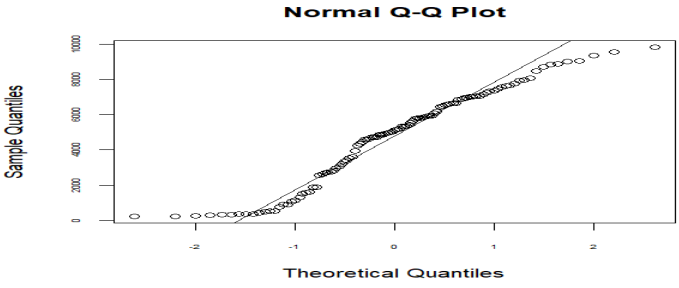

*Hình 16: đồ thị tương quan lượng hàng khách của hãng AC*

- Nhận xét:
    - Có nhiều điểm quan trắc lệch ra khỏi đường thẳng kì vọng phân phối chuẩn.
    - P value = 9.85e-09 rất nhỏ so với mức ý nghĩa 5% nên ta bác bỏ H0. Vì vậy số lượng hàng khách của hãng không tuân theo phân phối chuẩn

Hãng AS

```
Shapiro-Wilk normality test

data:  ASdata$Adjusted.Passenger.Count
W = 0.9504, p-value = 0.000478
```




*Hình 17: đồ thị tương quan lượng hàng khách của hãng AS*

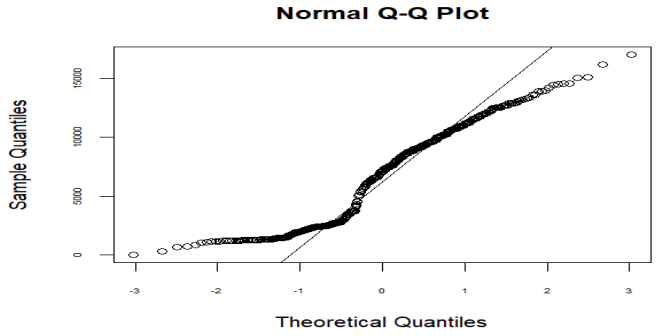
- Nhận xét:

- Có nhiều điểm quan trắc lệch ra khỏi đường thẳng kì vọng phân phối chuẩn.
- P value = 0.000478. Vì vậy số lượng hàng khách của hãng không tuân theo phân phối chuẩn.

Hãng OO

```
Shapiro-Wilk normality test

data:  OOdata$Adjusted.Passenger.Count
W = 0.93564, p-value = 5.125e-12
```



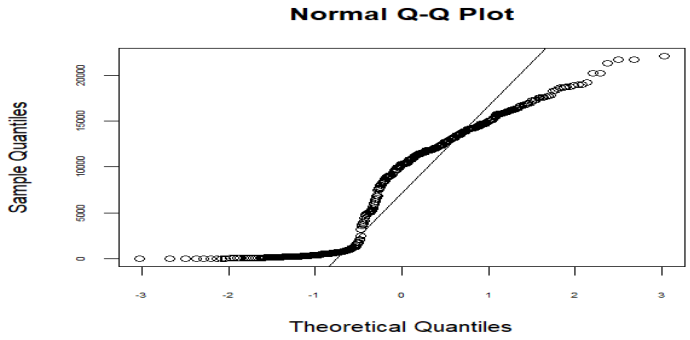

*Hình 18: đồ thị tương quan lượng hàng khách của hãng OO*

- Nhận xét:
  - Có nhiều điểm quan trắc lệch ra khỏi đường thẳng kì vọng phân phối chuẩn.
  - P value = 5.125e-12. Vì vậy số lượng hàng khách của hãng không tuân theo phân phối chuẩn.

Hãng UA

```
Shapiro-Wilk normality test

data:  UAdata$Adjusted.Passenger.Count
W = 0.89825, p-value = 8.133e-16
```



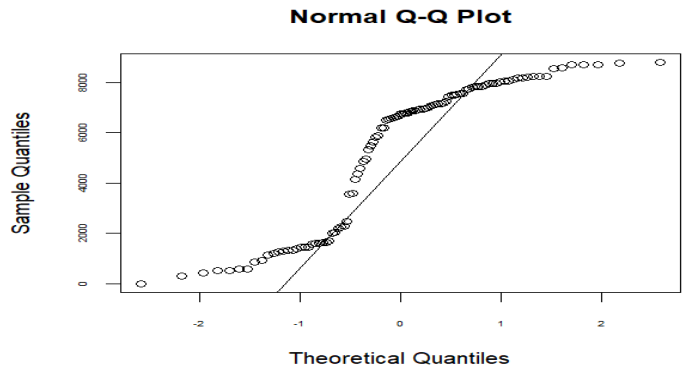

*Hình 19: đồ thị tương quan lượng hàng khách của hãng UA*

- Nhận xét:
    - Có nhiều điểm quan trắc lệch ra khỏi đường thẳng kì vọng phân phối chuẩn.
    - P value = 8.133e-16. Vì vậy số lượng hàng khách của hãng không tuân theo phân phối chuẩn.

Hãng WS

```
Shapiro-Wilk normality test

data:  WSdata$Adjusted.Passenger.Count
W = 0.84677, p-value = 6.219e-09
```




*Hình 20: đồ thị tương quan lượng hàng khách của hãng WS*

- Nhận xét:
    - Có nhiều điểm quan trắc lệch ra khỏi đường thẳng kì vọng phân phối chuẩn.
    - P value = 6.219e-09. Vì vậy số lượng hàng khách của hãng không tuân theo phân phối chuẩn.

Hãng GL

```
Error in shapiro.test(GLdata$Adjusted.Passenger.Count) :
  sample size must be between 3 and 5000
```

Ta kiểm tra lại số lượng biến

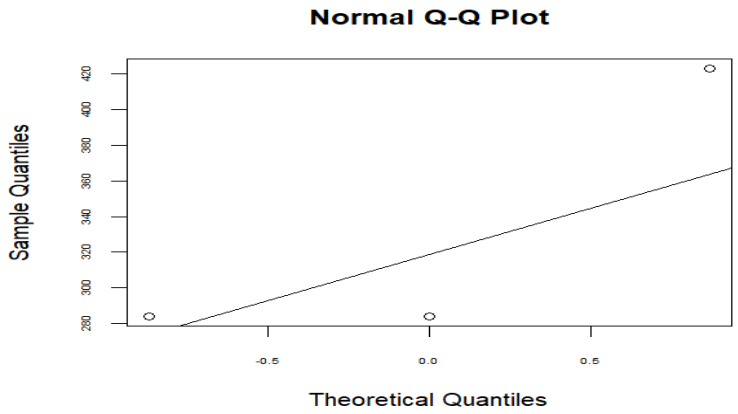
```
> head(GLdata$Adjusted.Passenger.Count)
[1] 28
```

- Nhận xét: Hãng bay GL chỉ có đúng 1 chuyến bay với 28 hàng khách.

Hãng TZ

```
Shapiro-Wilk normality test

data:  TZdata$Adjusted.Passenger.Count
W = 0.75, p-value < 2.2e-16
```



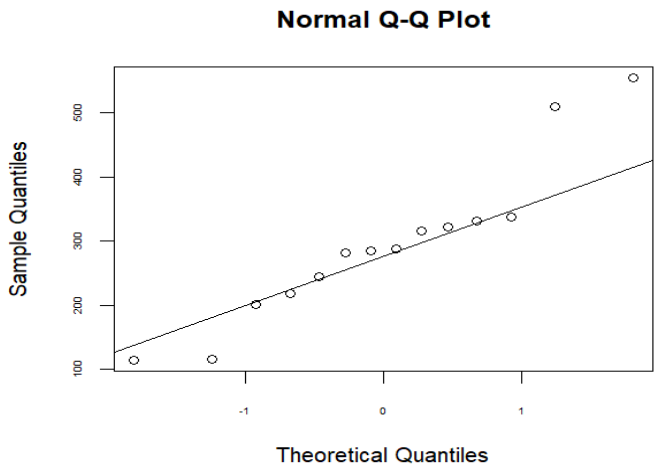

*Hình 21: đồ thị tương quan lượng hàng khách của hãng TZ*

- Nhận xét: số lượng điểm quan trắc rất ít và không tuân theo phân phối chuẩn

Hãng QK

```
Shapiro-Wilk normality test

data:  QKdata$Adjusted.Passenger.Count
W = 0.91503, p-value = 0.1864
```



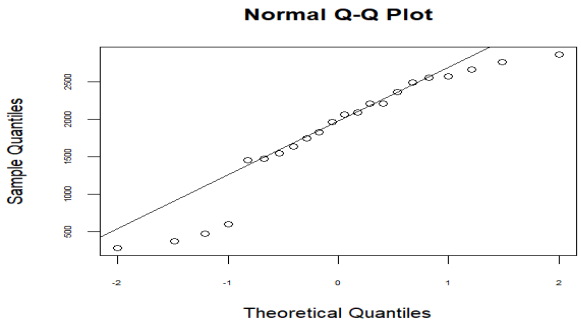

*Hình 22: đồ thị tương quan lượng hàng khách của hãng QK*

- Nhận xét:
  - Số lượng điểm quan trắc khá ít
  - Các điểm dường như nằm trên đường kì vọng
  - P value = 0.1864 lớn hơn mức ý nghĩa 5% nên số lượng hàng khách của hãng tuân theo phân phối chuẩn.

Hãng VX

```
Shapiro-Wilk normality test

data:  VXdata$Adjusted.Passenger.Count
W = 0.90463, p-value = 0.03682
```




*Hình 23: đồ thị tương quan lượng hàng khách của hãng VX*

- Nhận xét:
  - Số điểm quan trắc khá ít
  - Các điểm dường như nằm trên đường kì vọng phân phối chuẩn nên nhìn vào biểu đồ ta có thể dự đoán số lượng hàng khách của hãng tuân theo phân phối chuẩn. Tuy nhiên, p value= 0.03628 nhỏ hơn mức ý nghĩa 5% nên ta bác bỏ giả thuyết $H_0$. Vì vậy, số lượng hàng khách của hãng không tuân theo phân phối chuẩn.

*Kiểm tra Giả định 2:*

Giả thuyết $H_0$: phương sai số lượng hàng khách ở các hãng bằng nhau.

Giả thuyết $H_1$: Có ít nhất 2 hãng có phương sai số lượng hàng khách khác nhau.

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    8  39.342 < 2.2e-16 ***
      1409
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Vì *p-value* = 2.2e-16 < mức ý nghĩa 5% nên ta bác bỏ được giả thuyết $H_0$. Vì vậy phương sai số lượng hàng khách ở các hãng khác nhau.

**Kết luận: 2 giả định ở trên đều không thỏa mãn điều kiện của anova.**

### 5.1.4. Thực hiện Anova 1 nhân tố

Giả sử 2 giả định trên đều thỏa mãn.

```
                               Df    Sum Sq   Mean Sq F value Pr(>F)
Operating.Airline.IATA.Code     8 3.839e+10 4.799e+09   141.1 <2e-16 ***
Residuals                    1409 4.794e+10 3.403e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Giả thuyết $H_0$: Số lượng hành khách trung bình ở các hãng bay khu vực Canada bằng nhau.

Giả thuyết $H_1$: Có ít nhất 2 hãng bay khu vực Canada có lượng hành khách trung bình khác nhau.

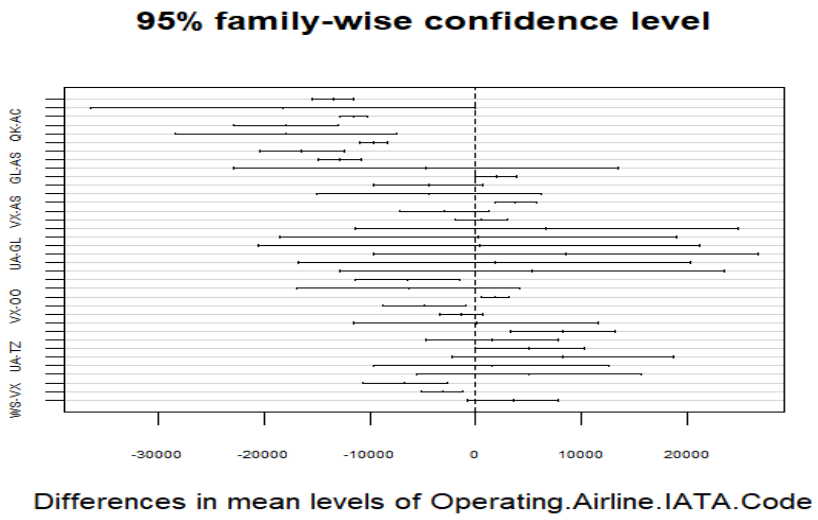
Vì pr(>F) < 2e-16 (rất bé so với mức ý nghĩa 5%) nên ta bác bỏ $H_0$, vậy có sự khác biệt về lượng hành khách trung bình trong các chuyến bay ở các hãng bay khu vực Canada.

### 5.1.5. Thực hiện so sánh bội

```
> TukeyHSD(anova1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Adjusted.Passenger.Count ~ Operating.Airline.IATA.Code, da
ta = Canadadata)
$Operating.Airline.IATA.Code
              diff          lwr          upr     p adj
AS-AC -13477.30323 -15454.63945 -11499.96701 0.0000000
GL-AC -18223.56011 -36369.52212    -77.59809 0.0480676
OO-AC -11542.36214 -12857.90696 -10226.81731 0.0000000
QK-AC -17957.34582 -22892.20479 -13022.48685 0.0000000
TZ-AC -17921.22678 -28426.31057  -7416.14298 0.0000048
UA-AC  -9722.65863 -11028.80726  -8416.51000 0.0000000
VX-AC -16424.96920 -20402.84746 -12447.09094 0.0000000
WS-AC -12913.40477 -14934.63030 -10892.17924 0.0000000
GL-AS  -4746.25688 -22950.41511  13457.90134 0.9965996
OO-AS   1934.94109    -26.20588   3896.08806 0.0564678
QK-AS  -4480.04260  -9624.77508    664.68989 0.1467606
TZ-AS  -4443.92355 -15049.21585   6161.36875 0.9309933
UA-AS   3754.64460   1799.78821   5709.50098 0.0000001
```

```
VX-AS  -2947.66597  -7183.10629   1287.77435 0.4313597
WS-AS    563.89846  -1926.24047   3054.03738 0.9987496
OO-GL   6681.19797 -11463.00707  24825.40300 0.9673581
QK-GL    266.21429 -18491.03368  19023.46225 1.0000000
TZ-GL    302.33333 -20622.25277  21226.91944 1.0000000
UA-GL   8500.90148  -9642.62471  26644.42766 0.8756547
VX-GL   1798.59091 -16729.90157  20327.08339 0.9999980
WS-GL   5310.15534 -12898.82241  23519.13309 0.9926735
QK-OO  -6414.98368 -11343.37814  -1486.58922 0.0018165
TZ-OO  -6378.86464 -16880.91322   4123.18395 0.6229772
UA-OO   1819.70351    538.19535   3101.21166 0.0003769
VX-OO  -4882.60706  -8852.46277   -912.75135 0.0043870
WS-OO  -1371.04263  -3376.43325    634.34799 0.4572219
TZ-QK     36.11905 -11492.76639  11565.00448 1.0000000
UA-QK   8234.68719   3308.79255  13160.58183 0.0000084
VX-QK   1532.37662  -4662.93848   7727.69173 0.9976510
WS-QK   5043.94105   -117.81896  10205.70106 0.0614621
UA-TZ   8198.56814  -2302.30756  18699.44385 0.2707546
VX-TZ   1496.25758  -9656.58429  12649.09944 0.9999755
WS-TZ   5007.82201  -5605.74096  15621.38498 0.8712321
VX-UA  -6702.31057 -10669.06244  -2735.55870 0.0000063
WS-UA  -3190.74614  -5189.98539  -1191.50689 0.0000281
WS-VX   3511.56443   -744.54280   7767.67167 0.2032891
```




*Hình 24: Những sự khác biệt về mức độ hoạt động của các cặp hãng bay.*

Nhận xét:

- Các cặp có khoảng tin cậy đi qua đường thẳng số 0 đều có giá trị trung bình bằng nhau như GL-AS, OO-AS,QK-AS…
- Hãng AC có lượng khách trung bình lớn nhất.

### 5.2. Kiểm định 2 mẫu:

*Kiểm định sự khác nhau giữa 2 khu vực đến (GEO.Region) là Mexico là US với mức ý nghĩa 5%*

*H0: không có sự khác biệt giữa 2 điểm đến.*

*H1: có sự khác biệt giữa 2 địa điểm.*

*Ta dùng t.test với độ tin cậy 95%*

```
data:  arrival1 and arrival2
t = -18.307, df = 30012, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.07185181 -0.05795428
sample estimates:
 mean of x  mean of y
0.07429866 0.13920171
```

Nhận xét:

- Do p-value< mức ý nghĩa 5% nên có thể nói rằng có sự khác nhau giữa 2 khu vực đến được chọn có ý nghĩa thống kê.
- Giả thuyết nghịch (alternative hypothesis) có thể phát biểu là: sự khác nhau của gi á trị trung bình của 2 địa điểm trên không bằng 0 ( true difference in means is not equal to 0 ).
- Với KTC 95% thì sự khác nhau 2 giá trị trung bình là: -0.07185 đến -0.05795
- Trung bình của lần lượt 2 địa điểm là 0.07429 và 0.13920.

### 5.3. Xây dựng mô hình hồi quy tuyến tính:

#### 5.3.1. xây dựng mô hình

Xét mô hình hồi quy tuyến tính gồm biến Passenger.Count là biến phụ thuộc và các biến còn lại trong newdata là biến độc lập.

**Mô hình 1:**

```
Call:
lm(formula = Adjusted.Passenger.Count ~ ., data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-157908   -7495    -746    3553  477252

Coefficients: (6 not defined because of singularities)
                                            Estimate Std. Error
(Intercept)                                -1.348e+06  2.284e+05
Operating.Airline.IATA.Code4T              -1.348e+03  1.148e+04
Operating.Airline.IATA.Code5Y               2.753e+04  4.737e+04
Operating.Airline.IATA.Code9W              -2.293e+04  1.149e+04
Operating.Airline.IATA.CodeA8               3.316e+03  3.954e+04
Operating.Airline.IATA.CodeAA               5.273e+04  6.653e+03
Operating.Airline.IATA.CodeAB              -1.515e+04  8.889e+03
Operating.Airline.IATA.CodeAC               5.798e+04  6.452e+03
Operating.Airline.IATA.CodeAF              -1.014e+03  6.603e+03
Operating.Airline.IATA.CodeAI              -4.065e+04  1.511e+04
Operating.Airline.IATA.CodeAM               1.235e+04  6.725e+03
Operating.Airline.IATA.CodeAS               8.605e+03  6.146e+03
Operating.Airline.IATA.CodeB6               4.413e+04  6.093e+03
Operating.Airline.IATA.CodeBA               5.022e+03  6.603e+03
Operating.Airline.IATA.CodeBBB             -1.083e+04  1.695e+04
Operating.Airline.IATA.CodeBR              -2.748e+04  6.858e+03
Operating.Airline.IATA.CodeCA              -3.373e+04  6.852e+03
Operating.Airline.IATA.CodeCI              -1.835e+04  6.647e+03
Operating.Airline.IATA.CodeCM              -8.755e+03  2.930e+04
Operating.Airline.IATA.CodeCP              -4.691e+04  7.400e+03
Operating.Airline.IATA.CodeCX              -1.109e+04  6.647e+03
Operating.Airline.IATA.CodeCZ              -2.687e+04  9.221e+03
Operating.Airline.IATA.CodeDH               1.920e+04  1.343e+04
Operating.Airline.IATA.CodeDL               1.749e+04  6.392e+03
Operating.Airline.IATA.CodeEI              -1.499e+04  7.272e+03
Operating.Airline.IATA.CodeEK              -2.463e+04  6.833e+03
Operating.Airline.IATA.CodeEV              -6.021e+04  1.035e+04
Operating.Airline.IATA.CodeEY              -2.477e+04  9.071e+03
Operating.Airline.IATA.CodeEZ               2.437e+04  4.741e+04
Operating.Airline.IATA.CodeF9               1.623e+04  6.264e+03
Operating.Airline.IATA.CodeFI              -7.089e+03  1.064e+04
Operating.Airline.IATA.CodeFL               1.129e+04  6.274e+03
Operating.Airline.IATA.CodeG4               1.658e+03  1.126e+04
Operating.Airline.IATA.CodeGL              -4.586e+04  1.133e+04
Operating.Airline.IATA.CodeHA              -5.176e+04  6.452e+03
Operating.Airline.IATA.CodeJL              -2.149e+04  6.645e+03
Operating.Airline.IATA.CodeKE              -2.253e+04  6.647e+03
```

| | | |
|---|---|---|
| Operating.Airline.IATA.CodeKL | -3.382e+03 | 6.603e+03 |
| Operating.Airline.IATA.CodeLH | -5.684e+03 | 6.733e+03 |
| Operating.Airline.IATA.CodeLP | -2.644e+04 | 7.410e+03 |
| Operating.Airline.IATA.CodeLX | -2.057e+04 | 7.088e+03 |
| Operating.Airline.IATA.CodeMQ | -6.498e+04 | 7.351e+03 |
| Operating.Airline.IATA.CodeMU | -2.528e+04 | 7.684e+03 |
| Operating.Airline.IATA.CodeMX | 5.823e+03 | 7.232e+03 |
| Operating.Airline.IATA.CodeNH | -3.421e+04 | 6.858e+03 |
| Operating.Airline.IATA.CodeNK | 1.526e+04 | 9.642e+03 |
| Operating.Airline.IATA.CodeNW | -1.433e+04 | 6.606e+03 |
| Operating.Airline.IATA.CodeNZ | -2.414e+03 | 6.883e+03 |
| Operating.Airline.IATA.CodeOO | -4.077e+04 | 6.230e+03 |
| Operating.Airline.IATA.CodeOZ | -2.231e+04 | 6.647e+03 |
| Operating.Airline.IATA.CodePR | -1.796e+04 | 6.647e+03 |
| Operating.Airline.IATA.CodeQF | 8.187e+03 | 7.168e+03 |
| Operating.Airline.IATA.CodeQK | 3.874e+04 | 1.211e+04 |
| Operating.Airline.IATA.CodeQX | -5.809e+04 | 6.675e+03 |
| Operating.Airline.IATA.CodeRW | -6.353e+04 | 9.952e+03 |
| Operating.Airline.IATA.CodeSE | -1.032e+04 | 9.268e+03 |
| Operating.Airline.IATA.CodeSK | -2.169e+04 | 7.764e+03 |
| Operating.Airline.IATA.CodeSQ | -2.585e+04 | 6.858e+03 |
| Operating.Airline.IATA.CodeSY | 1.120e+04 | 6.069e+03 |
| Operating.Airline.IATA.CodeTA | 7.595e+03 | 2.745e+04 |
| Operating.Airline.IATA.CodeTK | -2.006e+04 | 1.009e+04 |
| Operating.Airline.IATA.CodeTZ | 3.975e+04 | 8.128e+03 |
| Operating.Airline.IATA.CodeUA | 1.265e+04 | 6.108e+03 |
| Operating.Airline.IATA.CodeUS | 3.250e+03 | 6.342e+03 |
| Operating.Airline.IATA.CodeVS | -2.756e+03 | 6.603e+03 |
| Operating.Airline.IATA.CodeVX | 9.880e+04 | 6.091e+03 |
| Operating.Airline.IATA.CodeWN | 1.010e+05 | 6.216e+03 |
| Operating.Airline.IATA.CodeWO | -2.596e+04 | 2.310e+04 |
| Operating.Airline.IATA.CodeWS | 5.508e+04 | 7.147e+03 |
| Operating.Airline.IATA.CodeXE | -5.768e+04 | 9.260e+03 |
| Operating.Airline.IATA.CodeXJ | -6.169e+04 | 8.531e+03 |
| Operating.Airline.IATA.CodeXP | -5.085e+04 | 2.798e+04 |
| Operating.Airline.IATA.CodeYV | -5.987e+04 | 7.101e+03 |
| Operating.Airline.IATA.CodeYX | -6.217e+04 | 7.175e+03 |
| GEO.SummaryInternational | -2.513e+04 | 2.281e+03 |
| GEO.RegionAustralia / Oceania | -3.024e+04 | 2.800e+03 |
| GEO.RegionCanada | -7.880e+04 | 2.582e+03 |
| GEO.RegionCentral America | -3.189e+04 | 2.803e+04 |
| GEO.RegionEurope | -1.561e+04 | 2.775e+03 |
| GEO.RegionMexico | -3.657e+04 | 2.430e+03 |
| GEO.RegionMiddle East | NA | NA |
| GEO.RegionSouth America | NA | NA |
| GEO.RegionUS | NA | NA |
| Price.Category.CodeOther | 6.340e+04 | 3.122e+03 |
| TerminalOther | -5.118e+04 | 3.913e+04 |
| TerminalTerminal 1 | -6.692e+02 | 9.110e+03 |
| TerminalTerminal 2 | 2.566e+04 | 3.107e+03 |
| TerminalTerminal 3 | 1.134e+05 | 2.292e+03 |
| Boarding.AreaB | 1.159e+04 | 8.983e+03 |
| Boarding.AreaC | 1.244e+04 | 9.140e+03 |
| Boarding.AreaD | NA | NA |
| Boarding.AreaE | -9.572e+04 | 2.090e+03 |
| Boarding.AreaF | NA | NA |

```
Boarding.AreaG                                    1.238e+04  1.975e+03
Boarding.AreaOther                                       NA         NA
Adjusted.Activity.Type.CodeEnplaned              -1.876e+03  6.538e+02
Adjusted.Activity.Type.CodeThru / Transit * 2 -6.635e+04  1.489e+03
Year                                              6.651e+02  1.135e+02
Month                                             3.363e+02  9.212e+01
                                                 t value Pr(>|t|)
(Intercept)                                       -5.904 3.63e-09 ***
Operating.Airline.IATA.Code4T                     -0.117 0.906519
Operating.Airline.IATA.Code5Y                      0.581 0.561072
Operating.Airline.IATA.Code9W                     -1.995 0.046028 *
Operating.Airline.IATA.CodeA8                      0.084 0.933168
Operating.Airline.IATA.CodeAA                      7.926 2.42e-15 ***
Operating.Airline.IATA.CodeAB                     -1.705 0.088291 .
Operating.Airline.IATA.CodeAC                      8.986  < 2e-16 ***
Operating.Airline.IATA.CodeAF                     -0.154 0.877922
Operating.Airline.IATA.CodeAI                     -2.691 0.007129 **
Operating.Airline.IATA.CodeAM                      1.837 0.066256 .
Operating.Airline.IATA.CodeAS                      1.400 0.161487
Operating.Airline.IATA.CodeB6                      7.244 4.58e-13 ***
Operating.Airline.IATA.CodeBA                      0.760 0.446986
Operating.Airline.IATA.CodeBBB                    -0.639 0.522860
Operating.Airline.IATA.CodeBR                     -4.007 6.18e-05 ***
Operating.Airline.IATA.CodeCA                     -4.923 8.62e-07 ***
Operating.Airline.IATA.CodeCI                     -2.761 0.005762 **
Operating.Airline.IATA.CodeCM                     -0.299 0.765077
Operating.Airline.IATA.CodeCP                     -6.339 2.38e-10 ***
Operating.Airline.IATA.CodeCX                     -1.669 0.095217 .
Operating.Airline.IATA.CodeCZ                     -2.914 0.003570 **
Operating.Airline.IATA.CodeDH                      1.429 0.152893
Operating.Airline.IATA.CodeDL                      2.737 0.006213 **
Operating.Airline.IATA.CodeEI                     -2.061 0.039291 *
Operating.Airline.IATA.CodeEK                     -3.605 0.000314 ***
Operating.Airline.IATA.CodeEV                     -5.816 6.15e-09 ***
Operating.Airline.IATA.CodeEY                     -2.731 0.006320 **
Operating.Airline.IATA.CodeEZ                      0.514 0.607193
Operating.Airline.IATA.CodeF9                      2.591 0.009569 **
Operating.Airline.IATA.CodeFI                     -0.666 0.505297
Operating.Airline.IATA.CodeFL                      1.800 0.071944 .
Operating.Airline.IATA.CodeG4                      0.147 0.882994
Operating.Airline.IATA.CodeGL                     -4.046 5.23e-05 ***
Operating.Airline.IATA.CodeHA                     -8.023 1.11e-15 ***
Operating.Airline.IATA.CodeJL                     -3.234 0.001224 **
Operating.Airline.IATA.CodeKE                     -3.390 0.000700 ***
Operating.Airline.IATA.CodeKL                     -0.512 0.608592
Operating.Airline.IATA.CodeLH                     -0.844 0.398607
Operating.Airline.IATA.CodeLP                     -3.568 0.000360 ***
Operating.Airline.IATA.CodeLX                     -2.902 0.003717 **
Operating.Airline.IATA.CodeMQ                     -8.839  < 2e-16 ***
Operating.Airline.IATA.CodeMU                     -3.290 0.001004 **
Operating.Airline.IATA.CodeMX                      0.805 0.420749
Operating.Airline.IATA.CodeNH                     -4.988 6.16e-07 ***
Operating.Airline.IATA.CodeNK                      1.583 0.113452
Operating.Airline.IATA.CodeNW                     -2.169 0.030081 *
Operating.Airline.IATA.CodeNZ                     -0.351 0.725799
Operating.Airline.IATA.CodeOO                     -6.544 6.21e-11 ***
```

```
Operating.Airline.IATA.CodeOZ                    -3.356 0.000792 ***
Operating.Airline.IATA.CodePR                    -2.703 0.006888 **
Operating.Airline.IATA.CodeQF                     1.142 0.253415
Operating.Airline.IATA.CodeQK                     3.200 0.001377 **
Operating.Airline.IATA.CodeQX                    -8.703  < 2e-16 ***
Operating.Airline.IATA.CodeRW                    -6.384 1.78e-10 ***
Operating.Airline.IATA.CodeSE                    -1.113 0.265610
Operating.Airline.IATA.CodeSK                    -2.793 0.005224 **
Operating.Airline.IATA.CodeSQ                    -3.769 0.000164 ***
Operating.Airline.IATA.CodeSY                     1.846 0.064896 .
Operating.Airline.IATA.CodeTA                     0.277 0.782048
Operating.Airline.IATA.CodeTK                    -1.987 0.046924 *
Operating.Airline.IATA.CodeTZ                     4.890 1.02e-06 ***
Operating.Airline.IATA.CodeUA                     2.072 0.038304 *
Operating.Airline.IATA.CodeUS                     0.512 0.608332
Operating.Airline.IATA.CodeVS                    -0.417 0.676389
Operating.Airline.IATA.CodeVX                    16.221  < 2e-16 ***
Operating.Airline.IATA.CodeWN                    16.253  < 2e-16 ***
Operating.Airline.IATA.CodeWO                    -1.124 0.261083
Operating.Airline.IATA.CodeWS                     7.707 1.36e-14 ***
Operating.Airline.IATA.CodeXE                    -6.229 4.82e-10 ***
Operating.Airline.IATA.CodeXJ                    -7.231 5.01e-13 ***
Operating.Airline.IATA.CodeXP                    -1.817 0.069205 .
Operating.Airline.IATA.CodeYV                    -8.431  < 2e-16 ***
Operating.Airline.IATA.CodeYX                    -8.664  < 2e-16 ***
GEO.SummaryInternational                        -11.017  < 2e-16 ***
GEO.RegionAustralia / Oceania                   -10.800  < 2e-16 ***
GEO.RegionCanada                                -30.517  < 2e-16 ***
GEO.RegionCentral America                        -1.138 0.255298
GEO.RegionEurope                                 -5.625 1.89e-08 ***
GEO.RegionMexico                                -15.051  < 2e-16 ***
GEO.RegionMiddle East                                NA       NA
GEO.RegionSouth America                              NA       NA
GEO.RegionUS                                         NA       NA
Price.Category.CodeOther                         20.306  < 2e-16 ***
TerminalOther                                    -1.308 0.190880
TerminalTerminal 1                               -0.073 0.941448
TerminalTerminal 2                                8.260  < 2e-16 ***
TerminalTerminal 3                               49.472  < 2e-16 ***
Boarding.AreaB                                    1.291 0.196855
Boarding.AreaC                                    1.361 0.173566
Boarding.AreaD                                       NA       NA
Boarding.AreaE                                  -45.799  < 2e-16 ***
Boarding.AreaF                                       NA       NA
Boarding.AreaG                                    6.270 3.70e-10 ***
Boarding.AreaOther                                   NA       NA
Adjusted.Activity.Type.CodeEnplaned              -2.869 0.004117 **
Adjusted.Activity.Type.CodeThru / Transit * 2 -44.551  < 2e-16 ***
Year                                              5.860 4.73e-09 ***
Month                                             3.651 0.000262 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38660 on 14914 degrees of freedom
Multiple R-squared:  0.5626,   Adjusted R-squared:  0.5599
F-statistic: 208.5 on 92 and 14914 DF,  p-value: < 2.2e-16
```

- Nhận xét: Theo nhóm chúng em tìm hiểu, các giá trị NA xảy ra khi hai hoặc nhiều biến dự đoán có mối tương quan cao với nhau, được gọi là đa cộng tuyến hoàn hả.

Xây dựng mô hình 2 bằng cách loại 2 biến GEO.Region, Boarding.Area ra khỏi mô hình 1.

**Mô hình 2:**

```
Call:
lm(formula = Adjusted.Passenger.Count ~ . - GEO.Region - Boarding.Area,
    data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-130078   -8476    -421    3858  516195

Coefficients:
                                          Estimate Std. Error
(Intercept)                            -1054787.0   247048.2
Operating.Airline.IATA.Code4T             19191.9    12270.5
Operating.Airline.IATA.Code5Y             28113.2    51527.6
Operating.Airline.IATA.Code9W             13662.1    12252.3
Operating.Airline.IATA.CodeA8              2982.6    43006.2
Operating.Airline.IATA.CodeAA             27552.1     6981.3
Operating.Airline.IATA.CodeAB             10359.1     9398.3
Operating.Airline.IATA.CodeAC             11840.4     6692.6
Operating.Airline.IATA.CodeAF             20268.4     6790.1
Operating.Airline.IATA.CodeAI              9356.6    16132.1
Operating.Airline.IATA.CodeAM             13321.5     6995.9
Operating.Airline.IATA.CodeAS              9629.2     6472.9
Operating.Airline.IATA.CodeB6             35376.6     6522.8
Operating.Airline.IATA.CodeBA             26304.4     6790.1
Operating.Airline.IATA.CodeBBB            10136.5    18286.5
Operating.Airline.IATA.CodeBR             21795.6     6790.1
Operating.Airline.IATA.CodeCA             15526.1     6783.8
Operating.Airline.IATA.CodeCI             18536.8     6790.1
Operating.Airline.IATA.CodeCM              9346.0    12862.6
Operating.Airline.IATA.CodeCP            -39736.4     7808.1
Operating.Airline.IATA.CodeCX             25800.6     6790.1
Operating.Airline.IATA.CodeCZ             10685.5     9713.9
Operating.Airline.IATA.CodeDH              9114.7    14560.1
Operating.Airline.IATA.CodeDL             38139.3     6632.0
Operating.Airline.IATA.CodeEI             12510.9     7563.2
Operating.Airline.IATA.CodeEK             17553.3     6983.9
Operating.Airline.IATA.CodeEV            -45642.0    11042.5
Operating.Airline.IATA.CodeEY             12779.7     9544.9
Operating.Airline.IATA.CodeEZ             66842.3    51529.0
Operating.Airline.IATA.CodeF9             27690.9     6601.5
Operating.Airline.IATA.CodeFI             13535.7    11337.2
Operating.Airline.IATA.CodeFL             19846.8     6641.1
Operating.Airline.IATA.CodeG4             12148.2    12138.7
Operating.Airline.IATA.CodeGL            -54075.4    12239.0
Operating.Airline.IATA.CodeHA            -45646.3     6809.3
```

```
Operating.Airline.IATA.CodeJL                    15398.6     6788.0
Operating.Airline.IATA.CodeKE                    14357.7     6790.1
Operating.Airline.IATA.CodeKL                    17901.1     6790.1
Operating.Airline.IATA.CodeLH                    27981.2     6790.1
Operating.Airline.IATA.CodeLP                    10661.8     7666.7
Operating.Airline.IATA.CodeLX                    13442.7     7197.6
Operating.Airline.IATA.CodeMQ                  -101754.6     7636.8
Operating.Airline.IATA.CodeMU                    12158.8     7976.9
Operating.Airline.IATA.CodeMX                    18135.3     7343.7
Operating.Airline.IATA.CodeNH                    15064.8     6790.1
Operating.Airline.IATA.CodeNK                     5361.1    10419.9
Operating.Airline.IATA.CodeNW                    10487.3     6875.8
Operating.Airline.IATA.CodeNZ                    16594.6     6779.6
Operating.Airline.IATA.CodeOO                   -17266.7     6494.1
Operating.Airline.IATA.CodeOZ                    14582.2     6790.1
Operating.Airline.IATA.CodePR                    18927.9     6790.1
Operating.Airline.IATA.CodeQF                    14607.8     7257.3
Operating.Airline.IATA.CodeQK                   -12578.3    13004.8
Operating.Airline.IATA.CodeQX                   -43916.1     7017.4
Operating.Airline.IATA.CodeRW                   -48642.1    10705.1
Operating.Airline.IATA.CodeSE                    11364.4     9803.6
Operating.Airline.IATA.CodeSK                    12526.3     7976.9
Operating.Airline.IATA.CodeSQ                    23425.9     6790.1
Operating.Airline.IATA.CodeSY                    11349.1     6474.9
Operating.Airline.IATA.CodeTA                    13745.5     6790.1
Operating.Airline.IATA.CodeTK                    14299.3    10620.3
Operating.Airline.IATA.CodeTZ                    40144.9     8727.9
Operating.Airline.IATA.CodeUA                    33938.8     6297.0
Operating.Airline.IATA.CodeUS                    16452.9     6666.5
Operating.Airline.IATA.CodeVS                    18526.4     6790.1
Operating.Airline.IATA.CodeVX                    84500.1     6490.8
Operating.Airline.IATA.CodeWN                   111750.7     6548.6
Operating.Airline.IATA.CodeWO                   -13902.8    25070.2
Operating.Airline.IATA.CodeWS                    13308.2     7506.6
Operating.Airline.IATA.CodeXE                   -42928.4     9827.5
Operating.Airline.IATA.CodeXJ                   -46617.0     9012.5
Operating.Airline.IATA.CodeXP                   -57294.1    30402.7
Operating.Airline.IATA.CodeYV                   -45713.5     7496.1
Operating.Airline.IATA.CodeYX                   -84116.7     7494.4
GEO.SummaryInternational                        -67861.3     1298.4
Price.Category.CodeOther                         59890.5     3380.8
TerminalOther                                   -57316.5    42534.0
TerminalTerminal 1                               -9220.5     1633.9
TerminalTerminal 2                               36094.1     3270.5
TerminalTerminal 3                               48180.7     1421.2
Adjusted.Activity.Type.CodeEnplaned               -981.9      710.5
Adjusted.Activity.Type.CodeThru / Transit * 2   -65214.6     1615.6
Year                                               523.5      122.8
Month                                              323.5      100.2
                                                t value Pr(>|t|)
(Intercept)                                      -4.270 1.97e-05 ***
Operating.Airline.IATA.Code4T                     1.564 0.117822
Operating.Airline.IATA.Code5Y                     0.546 0.585352
Operating.Airline.IATA.Code9W                     1.115 0.264841
Operating.Airline.IATA.CodeA8                     0.069 0.944710
Operating.Airline.IATA.CodeAA                     3.947 7.97e-05 ***
```

```
Operating.Airline.IATA.CodeAB                1.102 0.270375
Operating.Airline.IATA.CodeAC                1.769 0.076886 .
Operating.Airline.IATA.CodeAF                2.985 0.002841 **
Operating.Airline.IATA.CodeAI                0.580 0.561924
Operating.Airline.IATA.CodeAM                1.904 0.056906 .
Operating.Airline.IATA.CodeAS                1.488 0.136874
Operating.Airline.IATA.CodeB6                5.424 5.93e-08 ***
Operating.Airline.IATA.CodeBA                3.874 0.000108 ***
Operating.Airline.IATA.CodeBBB               0.554 0.579371
Operating.Airline.IATA.CodeBR                3.210 0.001331 **
Operating.Airline.IATA.CodeCA                2.289 0.022111 *
Operating.Airline.IATA.CodeCI                2.730 0.006342 **
Operating.Airline.IATA.CodeCM                0.727 0.467483
Operating.Airline.IATA.CodeCP               -5.089 3.64e-07 ***
Operating.Airline.IATA.CodeCX                3.800 0.000145 ***
Operating.Airline.IATA.CodeCZ                1.100 0.271339
Operating.Airline.IATA.CodeDH                0.626 0.531320
Operating.Airline.IATA.CodeDL                5.751 9.05e-09 ***
Operating.Airline.IATA.CodeEI                1.654 0.098112 .
Operating.Airline.IATA.CodeEK                2.513 0.011968 *
Operating.Airline.IATA.CodeEV               -4.133 3.60e-05 ***
Operating.Airline.IATA.CodeEY                1.339 0.180624
Operating.Airline.IATA.CodeEZ                1.297 0.194590
Operating.Airline.IATA.CodeF9                4.195 2.75e-05 ***
Operating.Airline.IATA.CodeFI                1.194 0.232528
Operating.Airline.IATA.CodeFL                2.989 0.002808 **
Operating.Airline.IATA.CodeG4                1.001 0.316948
Operating.Airline.IATA.CodeGL               -4.418 1.00e-05 ***
Operating.Airline.IATA.CodeHA               -6.704 2.11e-11 ***
Operating.Airline.IATA.CodeJL                2.268 0.023314 *
Operating.Airline.IATA.CodeKE                2.114 0.034489 *
Operating.Airline.IATA.CodeKL                2.636 0.008389 **
Operating.Airline.IATA.CodeLH                4.121 3.79e-05 ***
Operating.Airline.IATA.CodeLP                1.391 0.164347
Operating.Airline.IATA.CodeLX                1.868 0.061826 .
Operating.Airline.IATA.CodeMQ              -13.324  < 2e-16 ***
Operating.Airline.IATA.CodeMU                1.524 0.127465
Operating.Airline.IATA.CodeMX                2.469 0.013541 *
Operating.Airline.IATA.CodeNH                2.219 0.026527 *
Operating.Airline.IATA.CodeNK                0.515 0.606908
Operating.Airline.IATA.CodeNW                1.525 0.127218
Operating.Airline.IATA.CodeNZ                2.448 0.014388 *
Operating.Airline.IATA.CodeOO               -2.659 0.007850 **
Operating.Airline.IATA.CodeOZ                2.148 0.031765 *
Operating.Airline.IATA.CodePR                2.788 0.005317 **
Operating.Airline.IATA.CodeQF                2.013 0.044149 *
Operating.Airline.IATA.CodeQK               -0.967 0.333457
Operating.Airline.IATA.CodeQX               -6.258 4.00e-10 ***
Operating.Airline.IATA.CodeRW               -4.544 5.57e-06 ***
Operating.Airline.IATA.CodeSE                1.159 0.246392
Operating.Airline.IATA.CodeSK                1.570 0.116361
Operating.Airline.IATA.CodeSQ                3.450 0.000562 ***
Operating.Airline.IATA.CodeSY                1.753 0.079659 .
Operating.Airline.IATA.CodeTA                2.024 0.042954 *
Operating.Airline.IATA.CodeTK                1.346 0.178189
Operating.Airline.IATA.CodeTZ                4.600 4.27e-06 ***
```

```
Operating.Airline.IATA.CodeUA                    5.390 7.17e-08 ***
Operating.Airline.IATA.CodeUS                    2.468 0.013598 *
Operating.Airline.IATA.CodeVS                    2.728 0.006371 **
Operating.Airline.IATA.CodeVX                   13.018  < 2e-16 ***
Operating.Airline.IATA.CodeWN                   17.065  < 2e-16 ***
Operating.Airline.IATA.CodeWO                   -0.555 0.579208
Operating.Airline.IATA.CodeWS                    1.773 0.076272 .
Operating.Airline.IATA.CodeXE                   -4.368 1.26e-05 ***
Operating.Airline.IATA.CodeXJ                   -5.172 2.34e-07 ***
Operating.Airline.IATA.CodeXP                   -1.885 0.059516 .
Operating.Airline.IATA.CodeYV                   -6.098 1.10e-09 ***
Operating.Airline.IATA.CodeYX                  -11.224  < 2e-16 ***
GEO.SummaryInternational                       -52.267  < 2e-16 ***
Price.Category.CodeOther                        17.715  < 2e-16 ***
TerminalOther                                   -1.348 0.177825
TerminalTerminal 1                              -5.643 1.70e-08 ***
TerminalTerminal 2                              11.036  < 2e-16 ***
TerminalTerminal 3                              33.900  < 2e-16 ***
Adjusted.Activity.Type.CodeEnplaned             -1.382 0.166988
Adjusted.Activity.Type.CodeThru / Transit * 2 -40.367  < 2e-16 ***
Year                                             4.264 2.02e-05 ***
Month                                            3.229 0.001243 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42060 on 14923 degrees of freedom
Multiple R-squared:  0.4822,   Adjusted R-squared:  0.4793
F-statistic: 167.4 on 83 and 14923 DF,  p-value: < 2.2e-16
```

Xây dựng mô hình 3 bằng cách loại bỏ biến Terminal từ mô hình 2

**Mô hình 3:**

```
Call:
lm(formula = Adjusted.Passenger.Count ~ . - GEO.Region - Boarding.Area -
    Terminal, data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-114543   -6530    -444    3937  541839

Coefficients:
                                              Estimate Std. Error
(Intercept)                                -1235637.6   256295.8
Operating.Airline.IATA.Code4T                 34553.0    12882.4
Operating.Airline.IATA.Code5Y                -23105.8    32009.5
Operating.Airline.IATA.Code9W                 29215.2    12862.7
Operating.Airline.IATA.CodeA8                -46302.4    11499.0
Operating.Airline.IATA.CodeAA                 75785.4     7039.6
Operating.Airline.IATA.CodeAB                 25656.7     9840.6
Operating.Airline.IATA.CodeAC                 41806.3     6901.6
Operating.Airline.IATA.CodeAF                 35642.5     7068.6
Operating.Airline.IATA.CodeAI                 24273.8    16961.1
Operating.Airline.IATA.CodeAM                 28546.6     7287.5
Operating.Airline.IATA.CodeAS                 16681.6     6683.7
```

| | | |
|---|---|---|
| Operating.Airline.IATA.CodeB6 | 30265.8 | 6826.3 |
| Operating.Airline.IATA.CodeBA | 41678.5 | 7068.6 |
| Operating.Airline.IATA.CodeBBB | 25701.2 | 19236.5 |
| Operating.Airline.IATA.CodeBR | 37169.7 | 7068.6 |
| Operating.Airline.IATA.CodeCA | 30848.1 | 7062.1 |
| Operating.Airline.IATA.CodeCI | 33910.9 | 7068.6 |
| Operating.Airline.IATA.CodeCM | 24247.5 | 13506.7 |
| Operating.Airline.IATA.CodeCP | -29135.5 | 8028.2 |
| Operating.Airline.IATA.CodeCX | 41174.7 | 7068.6 |
| Operating.Airline.IATA.CodeCZ | 25631.5 | 10174.0 |
| Operating.Airline.IATA.CodeDH | 4551.5 | 15318.3 |
| Operating.Airline.IATA.CodeDL | 40508.4 | 6815.6 |
| Operating.Airline.IATA.CodeEI | 27784.9 | 7891.7 |
| Operating.Airline.IATA.CodeEK | 32627.3 | 7275.1 |
| Operating.Airline.IATA.CodeEV | -46546.3 | 11506.3 |
| Operating.Airline.IATA.CodeEY | 27726.1 | 9994.8 |
| Operating.Airline.IATA.CodeEZ | 25217.2 | 32019.4 |
| Operating.Airline.IATA.CodeF9 | 13455.9 | 6730.1 |
| Operating.Airline.IATA.CodeFI | 29323.6 | 11894.7 |
| Operating.Airline.IATA.CodeFL | 6654.8 | 6814.9 |
| Operating.Airline.IATA.CodeG4 | -1904.5 | 12672.4 |
| Operating.Airline.IATA.CodeGL | -45441.6 | 12855.5 |
| Operating.Airline.IATA.CodeHA | -42961.6 | 7069.2 |
| Operating.Airline.IATA.CodeJL | 30759.9 | 7066.4 |
| Operating.Airline.IATA.CodeKE | 29731.8 | 7068.6 |
| Operating.Airline.IATA.CodeKL | 33275.2 | 7068.6 |
| Operating.Airline.IATA.CodeLH | 43355.3 | 7068.6 |
| Operating.Airline.IATA.CodeLP | 25896.8 | 8001.7 |
| Operating.Airline.IATA.CodeLX | 28587.6 | 7502.0 |
| Operating.Airline.IATA.CodeMQ | -45345.9 | 7837.8 |
| Operating.Airline.IATA.CodeMU | 27175.9 | 8330.6 |
| Operating.Airline.IATA.CodeMX | 33766.9 | 7658.9 |
| Operating.Airline.IATA.CodeNH | 30438.9 | 7068.6 |
| Operating.Airline.IATA.CodeNK | 694.4 | 10947.4 |
| Operating.Airline.IATA.CodeNW | 16963.4 | 7116.1 |
| Operating.Airline.IATA.CodeNZ | 31867.0 | 7057.7 |
| Operating.Airline.IATA.CodeOO | 16514.8 | 6632.4 |
| Operating.Airline.IATA.CodeOZ | 29956.3 | 7068.6 |
| Operating.Airline.IATA.CodePR | 34302.0 | 7068.6 |
| Operating.Airline.IATA.CodeQF | 30136.9 | 7566.8 |
| Operating.Airline.IATA.CodeQK | 26772.4 | 13546.1 |
| Operating.Airline.IATA.CodeQX | -45117.2 | 7181.3 |
| Operating.Airline.IATA.CodeRW | -50135.3 | 11136.6 |
| Operating.Airline.IATA.CodeSE | 26350.2 | 10269.8 |
| Operating.Airline.IATA.CodeSK | 27543.4 | 8330.6 |
| Operating.Airline.IATA.CodeSQ | 38800.0 | 7068.6 |
| Operating.Airline.IATA.CodeSY | 2841.7 | 6751.4 |
| Operating.Airline.IATA.CodeTA | 29119.6 | 7068.6 |
| Operating.Airline.IATA.CodeTK | 29224.2 | 11134.2 |
| Operating.Airline.IATA.CodeTZ | 29933.1 | 9105.6 |
| Operating.Airline.IATA.CodeUA | 64460.4 | 6496.7 |
| Operating.Airline.IATA.CodeUS | 12952.6 | 6802.4 |
| Operating.Airline.IATA.CodeVS | 33900.5 | 7068.6 |
| Operating.Airline.IATA.CodeVX | 100429.1 | 6577.1 |
| Operating.Airline.IATA.CodeWN | 96302.2 | 6667.4 |
| Operating.Airline.IATA.CodeWO | -1113.3 | 26396.0 |

```
Operating.Airline.IATA.CodeWS                     28593.8     7831.5
Operating.Airline.IATA.CodeXE                    -43955.7    10207.6
Operating.Airline.IATA.CodeXJ                    -47845.2     9330.9
Operating.Airline.IATA.CodeXP                    -48980.8    32017.5
Operating.Airline.IATA.CodeYV                    -46896.9     7700.0
Operating.Airline.IATA.CodeYX                    -45539.1     7732.4
GEO.SummaryInternational                         -75297.1     1194.3
Price.Category.CodeOther                          46940.9     3540.8
Adjusted.Activity.Type.CodeEnplaned                -347.7      747.1
Adjusted.Activity.Type.CodeThru / Transit * 2    -61478.5     1699.4
Year                                                615.7      127.4
Month                                               342.7      105.5
                                                 t value Pr(>|t|)
(Intercept)                                       -4.821 1.44e-06 ***
Operating.Airline.IATA.Code4T                      2.682 0.007322 **
Operating.Airline.IATA.Code5Y                     -0.722 0.470402
Operating.Airline.IATA.Code9W                      2.271 0.023142 *
Operating.Airline.IATA.CodeA8                     -4.027 5.69e-05 ***
Operating.Airline.IATA.CodeAA                     10.766  < 2e-16 ***
Operating.Airline.IATA.CodeAB                      2.607 0.009137 **
Operating.Airline.IATA.CodeAC                      6.058 1.42e-09 ***
Operating.Airline.IATA.CodeAF                      5.042 4.65e-07 ***
Operating.Airline.IATA.CodeAI                      1.431 0.152409
Operating.Airline.IATA.CodeAM                      3.917 9.00e-05 ***
Operating.Airline.IATA.CodeAS                      2.496 0.012576 *
Operating.Airline.IATA.CodeB6                      4.434 9.33e-06 ***
Operating.Airline.IATA.CodeBA                      5.896 3.80e-09 ***
Operating.Airline.IATA.CodeBBB                     1.336 0.181549
Operating.Airline.IATA.CodeBR                      5.258 1.47e-07 ***
Operating.Airline.IATA.CodeCA                      4.368 1.26e-05 ***
Operating.Airline.IATA.CodeCI                      4.797 1.62e-06 ***
Operating.Airline.IATA.CodeCM                      1.795 0.072638 .
Operating.Airline.IATA.CodeCP                     -3.629 0.000285 ***
Operating.Airline.IATA.CodeCX                      5.825 5.83e-09 ***
Operating.Airline.IATA.CodeCZ                      2.519 0.011768 *
Operating.Airline.IATA.CodeDH                      0.297 0.766374
Operating.Airline.IATA.CodeDL                      5.943 2.85e-09 ***
Operating.Airline.IATA.CodeEI                      3.521 0.000432 ***
Operating.Airline.IATA.CodeEK                      4.485 7.35e-06 ***
Operating.Airline.IATA.CodeEV                     -4.045 5.25e-05 ***
Operating.Airline.IATA.CodeEY                      2.774 0.005543 **
Operating.Airline.IATA.CodeEZ                      0.788 0.430966
Operating.Airline.IATA.CodeF9                      1.999 0.045588 *
Operating.Airline.IATA.CodeFI                      2.465 0.013702 *
Operating.Airline.IATA.CodeFL                      0.977 0.328826
Operating.Airline.IATA.CodeG4                     -0.150 0.880540
Operating.Airline.IATA.CodeGL                     -3.535 0.000409 ***
Operating.Airline.IATA.CodeHA                     -6.077 1.25e-09 ***
Operating.Airline.IATA.CodeJL                      4.353 1.35e-05 ***
Operating.Airline.IATA.CodeKE                      4.206 2.61e-05 ***
Operating.Airline.IATA.CodeKL                      4.707 2.53e-06 ***
Operating.Airline.IATA.CodeLH                      6.134 8.81e-10 ***
Operating.Airline.IATA.CodeLP                      3.236 0.001213 **
Operating.Airline.IATA.CodeLX                      3.811 0.000139 ***
Operating.Airline.IATA.CodeMQ                     -5.786 7.37e-09 ***
Operating.Airline.IATA.CodeMU                      3.262 0.001108 **
```

```
Operating.Airline.IATA.CodeMX                    4.409 1.05e-05 ***
Operating.Airline.IATA.CodeNH                    4.306 1.67e-05 ***
Operating.Airline.IATA.CodeNK                    0.063 0.949423
Operating.Airline.IATA.CodeNW                    2.384 0.017147 *
Operating.Airline.IATA.CodeNZ                    4.515 6.37e-06 ***
Operating.Airline.IATA.CodeOO                    2.490 0.012785 *
Operating.Airline.IATA.CodeOZ                    4.238 2.27e-05 ***
Operating.Airline.IATA.CodePR                    4.853 1.23e-06 ***
Operating.Airline.IATA.CodeQF                    3.983 6.84e-05 ***
Operating.Airline.IATA.CodeQK                    1.976 0.048129 *
Operating.Airline.IATA.CodeQX                   -6.283 3.42e-10 ***
Operating.Airline.IATA.CodeRW                   -4.502 6.79e-06 ***
Operating.Airline.IATA.CodeSE                    2.566 0.010304 *
Operating.Airline.IATA.CodeSK                    3.306 0.000948 ***
Operating.Airline.IATA.CodeSQ                    5.489 4.11e-08 ***
Operating.Airline.IATA.CodeSY                    0.421 0.673837
Operating.Airline.IATA.CodeTA                    4.120 3.82e-05 ***
Operating.Airline.IATA.CodeTK                    2.625 0.008681 **
Operating.Airline.IATA.CodeTZ                    3.287 0.001014 **
Operating.Airline.IATA.CodeUA                    9.922  < 2e-16 ***
Operating.Airline.IATA.CodeUS                    1.904 0.056913 .
Operating.Airline.IATA.CodeVS                    4.796 1.63e-06 ***
Operating.Airline.IATA.CodeVX                   15.269  < 2e-16 ***
Operating.Airline.IATA.CodeWN                   14.444  < 2e-16 ***
Operating.Airline.IATA.CodeWO                   -0.042 0.966359
Operating.Airline.IATA.CodeWS                    3.651 0.000262 ***
Operating.Airline.IATA.CodeXE                   -4.306 1.67e-05 ***
Operating.Airline.IATA.CodeXJ                   -5.128 2.97e-07 ***
Operating.Airline.IATA.CodeXP                   -1.530 0.126084
Operating.Airline.IATA.CodeYV                   -6.090 1.15e-09 ***
Operating.Airline.IATA.CodeYX                   -5.889 3.96e-09 ***
GEO.SummaryInternational                       -63.045  < 2e-16 ***
Price.Category.CodeOther                        13.257  < 2e-16 ***
Adjusted.Activity.Type.CodeEnplaned             -0.465 0.641669
Adjusted.Activity.Type.CodeThru / Transit * 2 -36.178  < 2e-16 ***
Year                                             4.833 1.36e-06 ***
Month                                            3.248 0.001164 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44320 on 14927 degrees of freedom
Multiple R-squared:  0.4249,   Adjusted R-squared:  0.4219
F-statistic: 139.6 on 79 and 14927 DF,  p-value: < 2.2e-16
```

Xây dựng mô hình 4 bằng cách loại bỏ biến Adjusted.Activity.Type.Code từ mô hình 3

**Mô hình 4:**

```
Call:
lm(formula = Adjusted.Passenger.Count ~ . - GEO.Region - Boarding.Area -
    Terminal - Adjusted.Activity.Type.Code, data = newdata)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-110882   -8769    -770    4135  553708

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -2086536.3   266691.5  -7.824 5.47e-15 ***
Operating.Airline.IATA.Code4T    36856.9    13459.0   2.738 0.006180 **
Operating.Airline.IATA.Code5Y   -47701.9    33435.0  -1.427 0.153686
Operating.Airline.IATA.Code9W    38110.8    13436.3   2.836 0.004568 **
Operating.Airline.IATA.CodeA8   -40447.0    12012.3  -3.367 0.000761 ***
Operating.Airline.IATA.CodeAA    82890.7     7352.0  11.275  < 2e-16 ***
Operating.Airline.IATA.CodeAB    33358.5    10278.9   3.245 0.001176 **
Operating.Airline.IATA.CodeAC    49481.9     7207.3   6.866 6.88e-12 ***
Operating.Airline.IATA.CodeAF    43657.0     7381.6   5.914 3.41e-09 ***
Operating.Airline.IATA.CodeAI    30108.3    17719.8   1.699 0.089314 .
Operating.Airline.IATA.CodeAM    35872.1     7610.9   4.713 2.46e-06 ***
Operating.Airline.IATA.CodeAS    15741.6     6982.8   2.254 0.024190 *
Operating.Airline.IATA.CodeB6    28454.0     7131.8   3.990 6.65e-05 ***
Operating.Airline.IATA.CodeBA    49693.0     7381.6   6.732 1.73e-11 ***
Operating.Airline.IATA.CodeBBB   34649.8    20096.2   1.724 0.084694 .
Operating.Airline.IATA.CodeBR    45184.2     7381.6   6.121 9.52e-10 ***
Operating.Airline.IATA.CodeCA    38824.7     7374.7   5.265 1.42e-07 ***
Operating.Airline.IATA.CodeCI    41925.4     7381.6   5.680 1.37e-08 ***
Operating.Airline.IATA.CodeCM    30104.1    14110.5   2.133 0.032904 *
Operating.Airline.IATA.CodeCP   -25155.2     8386.9  -2.999 0.002710 **
Operating.Airline.IATA.CodeCX    49189.2     7381.6   6.664 2.76e-11 ***
Operating.Airline.IATA.CodeCZ    31658.2    10628.1   2.979 0.002899 **
Operating.Airline.IATA.CodeDH     5348.3    16004.2   0.334 0.738249
Operating.Airline.IATA.CodeDL    45561.5     7119.3   6.400 1.60e-10 ***
Operating.Airline.IATA.CodeEI    35341.4     8242.3   4.288 1.82e-05 ***
Operating.Airline.IATA.CodeEK    39824.8     7598.1   5.241 1.61e-07 ***
Operating.Airline.IATA.CodeEV   -37607.9    12018.8  -3.129 0.001757 **
Operating.Airline.IATA.CodeEY    33766.6    10440.9   3.234 0.001223 **
Operating.Airline.IATA.CodeEZ    34867.9    33452.0   1.042 0.297277
Operating.Airline.IATA.CodeF9    12082.5     7031.3   1.718 0.085749 .
Operating.Airline.IATA.CodeFI    39293.9    12424.0   3.163 0.001566 **
Operating.Airline.IATA.CodeFL     5576.3     7119.9   0.783 0.433521
Operating.Airline.IATA.CodeG4    -2383.3    13239.8  -0.180 0.857146
Operating.Airline.IATA.CodeGL   -37453.3    13429.1  -2.789 0.005294 **
Operating.Airline.IATA.CodeHA   -35759.5     7382.9  -4.844 1.29e-06 ***
Operating.Airline.IATA.CodeJL    38540.6     7379.5   5.223 1.79e-07 ***
Operating.Airline.IATA.CodeKE    37746.3     7381.6   5.114 3.20e-07 ***
Operating.Airline.IATA.CodeKL    41289.7     7381.6   5.594 2.26e-08 ***
Operating.Airline.IATA.CodeLH    51369.9     7381.6   6.959 3.56e-12 ***
Operating.Airline.IATA.CodeLP    33274.9     8357.3   3.982 6.88e-05 ***
Operating.Airline.IATA.CodeLX    35560.6     7835.3   4.538 5.71e-06 ***
Operating.Airline.IATA.CodeMQ   -36804.8     8185.2  -4.496 6.96e-06 ***
Operating.Airline.IATA.CodeMU    33549.1     8701.7   3.855 0.000116 ***
Operating.Airline.IATA.CodeMX    42964.2     7997.5   5.372 7.90e-08 ***
Operating.Airline.IATA.CodeNH    38453.4     7381.6   5.209 1.92e-07 ***
Operating.Airline.IATA.CodeNK      959.6    11437.6   0.084 0.933140
Operating.Airline.IATA.CodeNW    20687.3     7434.0   2.783 0.005396 **
Operating.Airline.IATA.CodeNZ    39812.4     7370.2   5.402 6.70e-08 ***
Operating.Airline.IATA.CodeOO    23826.7     6926.3   3.440 0.000583 ***
Operating.Airline.IATA.CodeOZ    37970.8     7381.6   5.144 2.72e-07 ***
Operating.Airline.IATA.CodePR    42316.5     7381.6   5.733 1.01e-08 ***
```

```
Operating.Airline.IATA.CodeQF     38853.2     7901.7   4.917 8.88e-07 ***
Operating.Airline.IATA.CodeQK     36649.7    14149.9   2.590 0.009604 **
Operating.Airline.IATA.CodeQX    -37544.0     7499.8  -5.006 5.62e-07 ***
Operating.Airline.IATA.CodeRW    -43771.9    11633.8  -3.762 0.000169 ***
Operating.Airline.IATA.CodeSE     31122.4    10728.8   2.901 0.003727 **
Operating.Airline.IATA.CodeSK     33916.6     8701.7   3.898 9.75e-05 ***
Operating.Airline.IATA.CodeSQ     46814.5     7381.6   6.342 2.33e-10 ***
Operating.Airline.IATA.CodeSY      1404.5     7053.6   0.199 0.842176
Operating.Airline.IATA.CodeTA     37134.1     7381.6   5.031 4.94e-07 ***
Operating.Airline.IATA.CodeTK     35173.7    11631.5   3.024 0.002499 **
Operating.Airline.IATA.CodeTZ     13834.5     9502.4   1.456 0.145444
Operating.Airline.IATA.CodeUA     58782.3     6785.6   8.663  < 2e-16 ***
Operating.Airline.IATA.CodeUS     19012.6     7104.9   2.676 0.007459 **
Operating.Airline.IATA.CodeVS     41915.0     7381.6   5.678 1.39e-08 ***
Operating.Airline.IATA.CodeVX     98765.8     6871.4  14.373  < 2e-16 ***
Operating.Airline.IATA.CodeWN     74447.4     6938.6  10.729  < 2e-16 ***
Operating.Airline.IATA.CodeWO      6643.0    27576.6   0.241 0.809640
Operating.Airline.IATA.CodeWS     36224.0     8179.3   4.429 9.55e-06 ***
Operating.Airline.IATA.CodeXE    -35573.6    10662.0  -3.336 0.000851 ***
Operating.Airline.IATA.CodeXJ    -40427.0     9746.4  -4.148 3.37e-05 ***
Operating.Airline.IATA.CodeXP    -39997.2    33450.2  -1.196 0.231823
Operating.Airline.IATA.CodeYV    -39250.6     8041.9  -4.881 1.07e-06 ***
Operating.Airline.IATA.CodeYX    -37083.6     8075.0  -4.592 4.42e-06 ***
GEO.SummaryInternational         -76109.6     1247.6 -61.007  < 2e-16 ***
Price.Category.CodeOther          38378.9     3691.4  10.397  < 2e-16 ***
Year                               1039.4      132.6   7.841 4.75e-15 ***
Month                               391.1      110.2   3.548 0.000389 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46300 on 14929 degrees of freedom
Multiple R-squared:  0.3722,  Adjusted R-squared:  0.369
F-statistic: 114.9 on 77 and 14929 DF,  p-value: < 2.2e-16
```

Xây dựng mô hình 5 bằng cách loại bỏ biến Operating.Airline.IATA.Code từ mô hình 4:

**Mô hình 5:**

```
Call:
lm(formula = Adjusted.Passenger.Count ~ . - GEO.Region - Boarding.Area -
    Terminal - Adjusted.Activity.Type.Code - Operating.Airline.IATA.Code,
    data = newdata)

Residuals:
   Min     1Q Median     3Q    Max
-75604 -14967  -4220   4767 589324

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -3431809.1   279647.5 -12.272  < 2e-16 ***
GEO.SummaryInternational   -54266.8      975.0 -55.655  < 2e-16 ***
Price.Category.CodeOther    21386.2     1418.2  15.079  < 2e-16 ***
Year                         1727.5      139.0  12.424  < 2e-16 ***
```

```
Month                          430.6      125.6   3.429 0.000608 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52850 on 15002 degrees of freedom
Multiple R-squared:  0.1779,   Adjusted R-squared:  0.1776
F-statistic: 811.4 on 4 and 15002 DF,  p-value: < 2.2e-16
```

### 5.2.2 So sánh độ hiệu quả giữa các mô hình:

*So sánh độ hiệu quả giữa mô hình 2 và 3:*

*Giả thuyết H0: Mô hình 3 hiệu quả hơn*

*Giả thuyết H1: Mô hình 2 hiệu quả hơn*

```
Analysis of Variance Table

Model 1: Adjusted.Passenger.Count ~ (Operating.Airline.IATA.Code + GEO.Summar
y +
    GEO.Region + Price.Category.Code + Terminal + Boarding.Area +
    Adjusted.Activity.Type.Code + Year + Month) - GEO.Region -
    Boarding.Area
Model 2: Adjusted.Passenger.Count ~ (Operating.Airline.IATA.Code + GEO.Summar
y +
    GEO.Region + Price.Category.Code + Terminal + Boarding.Area +
    Adjusted.Activity.Type.Code + Year + Month) - GEO.Region -
    Boarding.Area - Terminal
  Res.Df        RSS Df  Sum of Sq     F    Pr(>F)
1  14923 2.6397e+13
2  14927 2.9315e+13 -4 -2.918e+12 412.4 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Nhận xét: vì p-value <2.2e-16 nhỏ hơn mức ý nghĩa 5% nên ta bác bỏ giả thuyết $H_0$ => Mô hình 2 hiệu quả hơn.

*So sánh độ hiệu quả giữa mô hình 2 và 4:*

*Giả thuyết $H_0$: Mô hình 4 hiệu quả hơn*

*Giả thuyết $H_1$: Mô hình 2 hiệu quả hơn*

```
Analysis of Variance Table

Model 1: Adjusted.Passenger.Count ~ (Operating.Airline.IATA.Code + GEO.Summar
y +
    GEO.Region + Price.Category.Code + Terminal + Boarding.Area +
    Adjusted.Activity.Type.Code + Year + Month) - GEO.Region -
    Boarding.Area
```

```
Model 2: Adjusted.Passenger.Count ~ (Operating.Airline.IATA.Code + GEO.Summar
y +
    GEO.Region + Price.Category.Code + Terminal + Boarding.Area +
    Adjusted.Activity.Type.Code + Year + Month) - GEO.Region -
    Boarding.Area - Terminal - Adjusted.Activity.Type.Code
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1  14923 2.6397e+13
2  14929 3.2003e+13 -6 -5.6063e+12 528.23 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Nhận xét: vì p-value <2.2e-16 nhỏ hơn mức ý nghĩa 5% nên ta bác bỏ giả thuyết $H_0$ => Mô hình 2 hiệu quả hơn.

*So sánh độ hiệu quả giữa mô hình 2 và 5:*

*Giả thuyết H0: Mô hình 5 hiệu quả hơn*

*Giả thuyết H1: Mô hình 2 hiệu quả hơn*

```
Analysis of Variance Table

Model 1: Adjusted.Passenger.Count ~ (Operating.Airline.IATA.Code + GEO.Summar
y +
    GEO.Region + Price.Category.Code + Terminal + Boarding.Area +
    Adjusted.Activity.Type.Code + Year + Month) - GEO.Region -
    Boarding.Area
Model 2: Adjusted.Passenger.Count ~ (Operating.Airline.IATA.Code + GEO.Summar
y +
    GEO.Region + Price.Category.Code + Terminal + Boarding.Area +
    Adjusted.Activity.Type.Code + Year + Month) - GEO.Region -
    Boarding.Area - Terminal - Adjusted.Activity.Type.Code -
    Operating.Airline.IATA.Code
  Res.Df        RSS  Df   Sum of Sq      F    Pr(>F)
1  14923 2.6397e+13
2  15002 4.1910e+13 -79 -1.5513e+13 111.01 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Nhận xét: vì p-value <2.2e-16 nhỏ hơn mức ý nghĩa 5% nên ta bác bỏ giả thuyết $H_0$ => Mô hình 2 hiệu quả hơn.

Từ việc so sánh các mô hình, ta ra được mô hình 2 là mô hình hiệu quả nhất.

- Nhận xét:
    - Bội $R^2$ của mô hình 2 là 48,44 %, có nghĩa là sự biến thiên của lượng hàng khách phụ thuộc 48,44% là các biến độc lập, còn lại là do các yếu tố khác như sai số hồi quy hoặc các yếu tố độc lập khác chưa đưa vào mô hình.

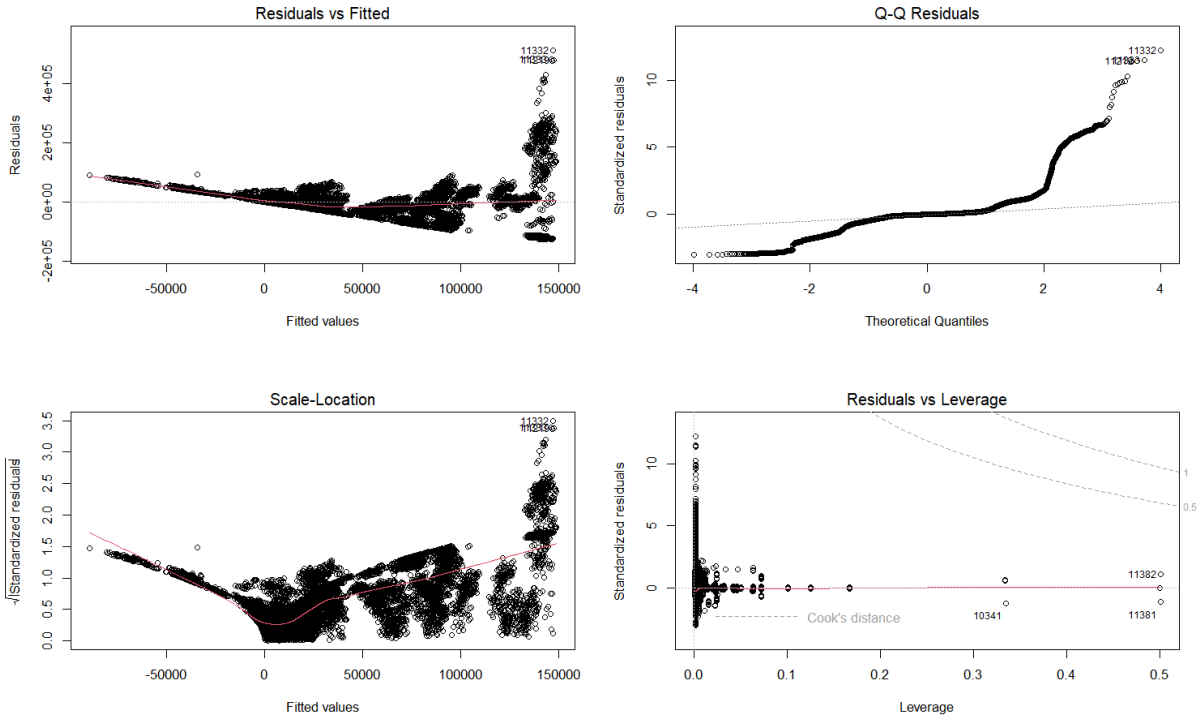
- p−value tương ứng với thống kê F bé hơn 2.2e−16, có ý nghĩa rất cao. Điều này chỉ ra rằng, ít nhất một biến dự báo trong mô hình có ý nghĩa giải thích rất cao đến lượng hàng khách.

### 5.2.3. kiểm tra các giả định của mô hình.

Các giả định của mô hình hồi quy: $Y_i = \beta_0 + X_1B_1 + ... + X_iB_i$, i = 1, ... n

- Giả định 1: Tính tuyến tính của dữ liệu: mối quan hệ giữa biến độc lập và biến phụ thuộc được giả sử là tuyến tính.
- Giả định 2: Sai số có phân phối chuẩn.
- Giả định 3: Phương sai của các sai số là hằng số và có kì vọng bằng 0 $\varepsilon_i \sim N\sim(0,\sigma^2)$
- Giả định 4: Các sai số $\varepsilon_1, ..., \varepsilon_n$ độc lập với nhau.

Ta thực hiện phân tích thặng dư để kiểm tra các gải định của mô hình.




*Hình 25: kết quả khi vẽ các đồ thị phân tích thặng dư*

- Đồ thị đầu tiên (Residuals vs Fitted) là đồ thị vẽ giá trị thặng dư tương ứng với giá trị dự báo. Đồ thị kiểm tra 2 giả định là giả định 1 và giả định 3.
    - Nhìn đồ thị ta thấy đường màu đỏ không phải là đường thẳng nằm ngang nên giữa biến phụ thuộc và biến độc lập của giả định 1 không có quan hệ tuyến tính.
    - Đường màu đỏ không nằm sát ngay đường Residuals= 0 nên giả định sai số có kì vọng bằng 0 là chưa thỏa mãn.
    - Các điểm sai số phân tán tập trung gần đường màu đỏ nên giả định phương sai của sai số là hằng số cũng chưa thỏa mãn.
- Đồ thị thứ 2 (normal Q-Q) vẽ các sai số đã được chuẩn hóa, dùng để kiểm tra giả định 2: giả định sai số có phân phối chuẩn.
    - Nhìn đồ thị ta thấy có rất nhiều điểm bị lệch so với đường phân phối chuẩn, nên giả định sai số có phân phối chuẩn chưa được thỏa mãn.
- Đồ thị thứ 3 (scale-location) vẽ căn bậc hai của các sai số được chuẩn hóa dùng để kiểm tra giả định phương sai của các sai số là hằng số và giả định 4.
    - Nếu như đường màu đỏ trên đồ thị là đường thẳng nằm ngang và các điểm thặng dư phân tán đều xung quanh đường thẳng này thì giả định thứ 3 và 4 được thỏa. Nếu như đường màu đỏ có độ dốc (hoặc cong) hoặc các điểm thặng dư phân tán không đều xung quanh đường thẳng này, thì giả định thứ 3 và 4 bị vi phạm. Ta thấy rằng các giá trị sai số trong đồ thị không phân tán đều xung quanh và đường màu đỏ nằm dốc nên giả định về phương sai của các sai số là hằng số không được thỏa mãn và giả định các sai số độc lập với nhau là không thỏa mãn.
- Đồ thị thứ 4 (residuals vs leverage) dùng để xác định các điểm có ảnh hưởng cao.
    - Nhìn đồ thị ta thấy không có điểm nào vượt ra khỏi đường Cook's distance nên không chứa các điểm có ảnh hưởng cao.

**Kết luận**: Giá trị *p-value* của mô hình < 2.2e-16 cho thấy mô hình hồi quy khá tốt. Tuy nhiên, mô hình vi phạm một số giả định nên kết quả mà mô hình dự

đoán có thể không chính xác. Bên cạnh đó, do giá trị $R^2$ =44,48 là khá thấp nên mô hình hồi quy này chỉ nên dùng để tham khảo, không nên áp dụng trong thực tế.

# 6. Thảo luận và mở rộng

## 6.1. Hạn chế của những phương pháp sử dụng trong bài làm

- Phân tích phương sai (Anova): Phân tích phương sai (ANOVA) là một phương pháp cực kỳ quan trọng trong phân tích dữ liệu khám phá và xác nhận. Thật không may, trong các vấn đề phức tạp, không phải lúc nào cũng dễ dàng thiết lập ANOVA thích hợp.

- Phương pháp phân tích hồi quy: Muốn đạt kết quả nghiên cứu chính xác và có độ tin cậy cao, phải có nhiều mẫu nghiên cứu, tốn kém chi phí và nhiều thời gian.

## 6.2. Thảo luận về ý nghĩa thực tiễn của vấn đề nghiên cứu liên quan đến bộ dữ liệu:

Thông qua nghiên cứu bộ dữ liệu, ta có thể đưa ra mô hình dự đoán, biết được sự ảnh hưởng của các thông số đến lượng khách hàng đi máy bay, là một công cụ hữu ích để hiểu ngành hàng không và lập kế hoạch du lịch. có thể được sử dụng để tạo ra bản đồ nhiệt về mô hình giao thông hàng không, cũng như được sử dụng để nghiên cứu tác động của các yếu tố khác nhau đến số lượng hành khách giao thông hàng không, chẳng hạn như thời gian trong năm hoặc ngày, giá vé máy bay hoặc số lượng chuyến bay do một hãng hàng không cung cấp.

# 7. R code


codebtlxstk.1.R

## 8. Nguồn tài liệu tham khảo

- Nguyễn Tiến Dũng, Nguyễn Đình Huy, XÁC SUẤT – THỐNG KÊ và PHÂN TÍCH SỐ LIỆU, NHÀ XUẤT BẢN ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH

- "Phân tích phương sai một yếu tố bằng R với câu lệnh đơn giản | PT Thống Kê 24 | Learn to do SCIENCE" [https://www.youtube.com/watch?v=kvSuWlyFpOk]

- "Hướng dẫn R làm bài tập lớn xstk" [https://ww.youtube.com/watch?v=2TDZcsbPNTY]