Trường Đại học Kinh tế Quốc dân
Khoa Toán Kinh Tế

Phân tích thống kê nhiều chiều

BÁO CÁO BÀI TẬP GIỮA KÌ

Giảng viên:Nguyễn Mạnh Thế

Họ tên: Trương Ngọc Thuỳ Trang
MSV: 11217004
Lớp: TOKT63

DECEMBER 1, 2023

Mục Lục

Mở đầu

Bài báo cáo này nhằm tập trung vào việc khám phá các phương pháp thống kê quan trọng trong nghiên cứu đa biến, với sự chú ý đặc biệt đến Phân tích Phương sai (ANOVA), Phân tích Thành phần Chính (PCA), Phân tích Phân biệt (DA), và Phân tích Cụm (CA).

Những phương pháp này không chỉ là những công cụ quan trọng trong thống kê và khoa học dữ liệu, mà còn đóng vai trò quan trọng trong việc hiểu rõ sự biến động và mối quan hệ trong dữ liệu đa chiều. Phân tích Phương vai (ANOVA) đã lâu đã là một công cụ mạnh mẽ để kiểm định sự khác biệt giữa các nhóm trong một thí nghiệm. Chúng ta sẽ khám phá cách ANOVA có thể tiếp cận những biến cố phức tạp và đa chiều, đồng thời đi sâu vào các giả thiết và kiểm định liên quan.

Phương pháp Phân tích Thành phần Chính (PCA) sẽ là trọng tâm để hiểu rõ sự tương quan và biểu diễn không gian chiều cao của dữ liệu. Chúng ta sẽ xem xét cách PCA có thể giảm chiều dữ liệu, làm nổi bật các đặc trưng quan trọng và giúp hiểu rõ cấu trúc ẩn của dữ liệu.

Tiếp theo là Phân tích Phân biệt (DA), một phương pháp quan trọng để phân loại quan sát vào các nhóm đã biết trước. Chúng ta sẽ xem xét cách DA có thể giúp hiểu sâu về sự phân biệt giữa các nhóm và áp dụng nó trong các tình huống nghiên cứu cụ thể.

Cuối cùng, Phân tích Cụm (CA) sẽ được đánh giá về khả năng nhóm các quan sát thành các cụm có ý nghĩa, giúp làm rõ sự tương đồng và khác biệt giữa các đối tượng. Chúng ta sẽ chi tiết khám phá cách mỗi phương pháp này có thể đóng góp vào việc hiểu rõ dữ liệu đa biến và ứng dụng của chúng trong nghiên cứu và thực tế.

Bài tập 1: ANOVA

Phương pháp ANOVA (Analysis of Variance) là một kỹ thuật thống kê được sử dụng để kiểm tra sự khác biệt giữa các giá trị trung bình của ba hoặc nhiều nhóm. Nó đánh giá xem có sự khác biệt đáng kể nào giữa các nhóm hay không bằng cách so sánh phương sai giữa các nhóm với phương sai trong các nhóm.

ANOVA là một công cụ mạnh mẽ trong phân tích thống kê và thường được sử dụng trong nghiên cứu để đánh giá sự ảnh hưởng của các biến độc lập là biến định tính lên biến phụ thuộc là định lượng.

Tệp số liệu được sử dụng: "HybridTest.xls" Mục tiêu của nghiên cứu là để so sánh các loại xe ô tô điện và xe xăng truyền thống.

Consumer số liệu sau đây là số đo về lượng tiêu hao năng lượng đo bằng dặm/gallon (MPG) cho hai mẫu xe điện cỡ nhỏ, hai mẫu xe điện cỡ trung, hai mẫu xe điện SUV cỡ nhỏ và hai mẫu xe điện SUV cỡ trung. Tương tự là MPG cho tám mẫu xe xăng thông thường.

```
library(readxl)
HybridTest <- read_excel("Assignment_TKT63/HybridTest.xlsx")
HybridTest
```

```
## # A tibble: 16 × 4
##    `Make/Model`      Class       Type           MPG
##    <chr>             <chr>       <chr>        <dbl>
##  1 Honda Civic       Small Car   Hybrid          37
##  2 Honda Civic       Small Car   Conventional    28
##  3 Toyota Prius      Small Car   Hybrid          44
##  4 Toyota Corolla    Small Car   Conventional    32
##  5 Chevrolet Malibu  Midsize Car Hybrid          27
##  6 Chevrolet Malibu  Midsize Car Conventional    23
##  7 Nissan Altima     Midsize Car Hybrid          32
##  8 Nissan Altima     Midsize Car Conventional    25
##  9 Ford Escape       Small SUV   Hybrid          27
## 10 Ford Escape       Small SUV   Conventional    21
## 11 Saturn Vue        Small SUV   Hybrid          28
## 12 Saturn Vue        Small SUV   Conventional    22
## 13 Lexus RX          Midsize SUV Hybrid          23
## 14 Lexus RX          Midsize SUV Conventional    19
## 15 Toyota Highlander Midsize SUV Hybrid          24
## 16 Toyota Highlander Midsize SUV Conventional    18
```

1.1. Thực hiện một số thống kê mô tả và giả thiết ANOVA

Hybrid ($\bar{x}_{Hybrid} = 29$)				**Conventional ($\bar{x}_{Conventional} = 23.5$)**			
Small car **40.5**	**Midsize car** **29.5**	**Small SUV** **27.5**	**Midsize SUV** **18.5**	**Small car** **30**	**Midsize car** **24**	**Small SUV** **21.5**	**Midsize SUV** **18.5**
37	27	27	23	28	23	21	19
44	32	28	14	32	25	22	18

Với dữ liệu đầu vào là lượng tiêu hao năng lượng MPG được lấy quan sát dựa trên 2 nhân tố là cỡ xe(class) và loại xe(type). Có thể thấy MPG trung bình của loại xe điện(Hybrid) là 29 lớn hơn MPG trung bình của loại xe thường là 23,5. Về cỡ xe thì ở những xe có cỡ nhỏ sẽ tiêu hao năng lượng lớn nhất và những xe SUV cỡ trung tiêu hao năng lượng ít nhất.

Đây là bộ dữ liệu độc lập , biến phụ thuộc MPG là định lượng và 2 nhân tố là loại xe và cỡ xe là 2 biến định đính. Để đo lường sự tác động của 2 nhân tố này lên lượng tiêu hao năng lượng MPG thì phải dùng đến phương pháp phân tích phương sai ANOVA với 2 nhân tố tương tác.

1.1.Các giả thiết ANOVA:

-Mẫu là ngẫu nhiên và độc lập

-Biến phụ thuộc phân phối chuẩn

-Phương sai biến phụ thuộc trong mỗi nhóm là tương đồng

1.2.Kiểm định các giả thiết ANOVA

a.Mẫu là ngẫu nhiên và độc lập

b.Biến phụ thuộc phân phối chuẩn

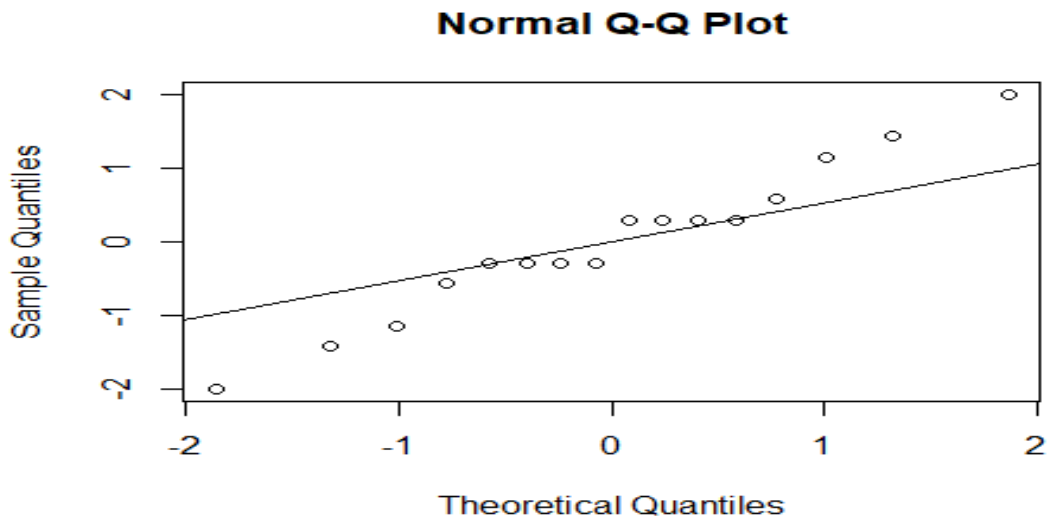
Xét cặp giả thuyết:

H_0: *Biến phân phối chuẩn*

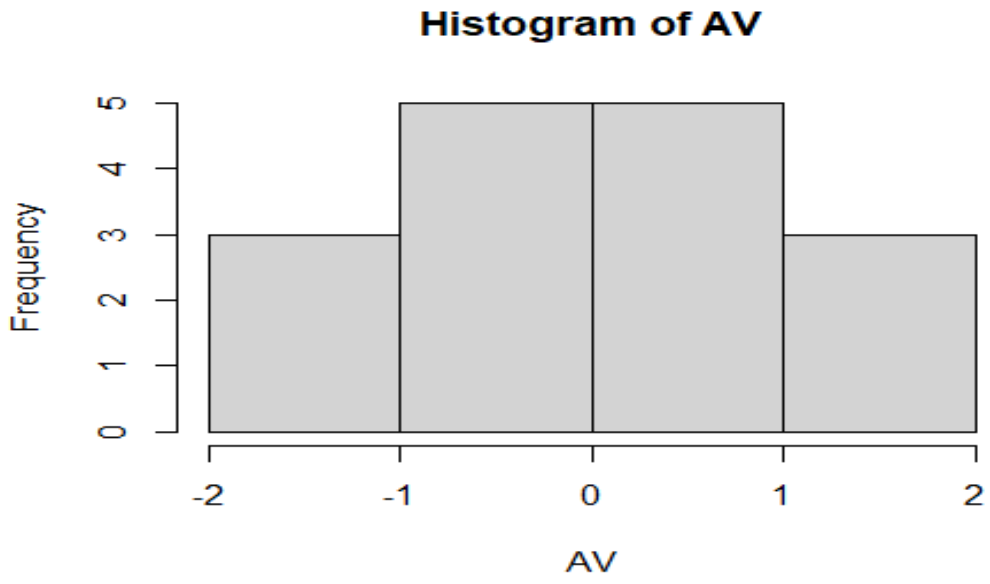
H_1: *Biến không phân phối chuẩn*

```r
library(nortest)
## Kiểm định thông qua sai số
AV=rstandard(aov(MPG~class*type))
shapiro.test(AV)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  AV
## W = 0.9729, p-value = 0.8831
```




```r
# Histogram of residuals
hist(AV)
```



Có thể thấy giá trị p-value = 0.8831 < 0.05 nên chưa có cơ sở Bác bỏ H_0 ở mức ý nghĩa 5%. Do đó có thể nói biến đo lượng tiêu hao năng lượng MPG phân phối chuẩn ở mức ý nghĩa 5%. Đồng thời nhìn vào biểu đồ histogram cũng có thể đánh giá điều này.

c. Phương sai biến phụ thuộc trong mỗi nhóm là tương đồng

Xét cặp giả thiết :H_0: *Phương sai MPG trong mỗi nhóm là tương đồng*

$$H_1: \textit{Phương sai MPG trong mỗi nhóm là khác nhau}$$

```
library(car)
```

```
## Loading required package: carData
```

```
leveneTest(MPG~class*type,data=HybridTest)
```

```
## Warning in anova.lm(lm(resp ~ group)): ANOVA F-tests on an essentially
perfect
## fit are unreliable
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df    F value    Pr(>F)
## group  7 1.0973e+31 < 2.2e-16 ***
##        8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Với mức ý nghĩa 5%,Bác bỏ H_0 do đó phương sai của MPG trong mỗi nhóm không tương đồng. Nhưng ANOVA vẫn có thể dùng khi kích cỡ mẫu của các nhóm là bằng nhau.

1.2. Tác động của loại xe, cỡ xe và tương tác giữa 2 nhân tố lên lượng tiêu hao năng lượng trung bình.

Xét 3 cặp giả thuyết:

-Với nhân tố cỡ xe: $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$

$$H_1: \mu_1^2 + \mu_2^2 + \mu_3^2 + \mu_4^2 \neq 0$$

-Với nhân tố loại xe: $H_0: \mu_H = \mu_C = 0$

$$H_1: \mu_H^2 + \mu_C^2 \neq 0$$

-Với sự tương tác : H_0: *Biến tương tác không ảnh hưởn đến trung bình*

$$H_1: \textit{Biến tương tác ảnh hưởng đến trung bình}$$

```
anova_MPG<-aov(MPG~class+type+class*type,data = HybridTest)
summary(anova_MPG)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## class        3  441.3  147.08  24.014 0.000236 ***
```

```
## type         1  182.2  182.25  29.755 0.000605 ***
## class:type   3   19.3    6.42   1.048 0.422860
## Residuals    8   49.0    6.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Kết quả chỉ ra rằng với mức ý nghĩa 5%, 2 cặp giả thuyết đầu tiên có H_0 bị bác bỏ. Vì vậy nhân tố cỡ xe và loại xe có tác động đến trung bình.

Nhưng tương tác giữa cỡ xe và loại xe thì lại không tác động đến trung bình ở mức ý nghĩa 5% vì chưa có cơ sở bác bỏ H_0.

1.3. Thực hiện các kiểm định so sánh cặp (Multiple Comparison).

Phương pháp phân tích phương sai ANOVA đã xác định rằng có sự tác động đến trung bình của nhân tố cỡ xe và loại xe. Để xác định chính xác sự khác biệt này đến từ cặp nhóm nào ta thực hiện kiểm định so sánh cặp.

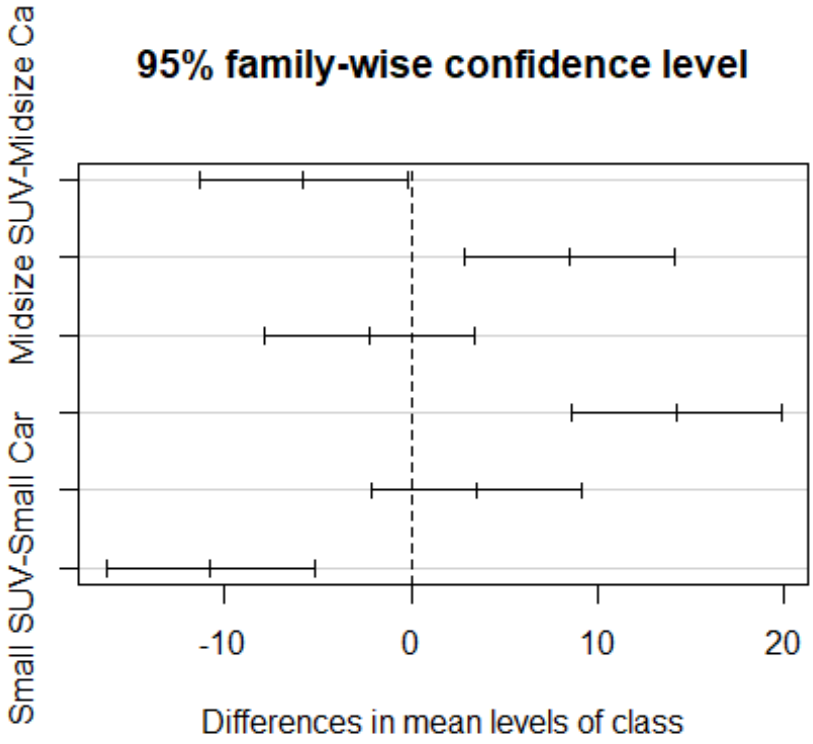
Dùng kiểm định Tukey

H_0:Trung bình của 2 nhóm tương đồng

H_1:Trung bình của 2 nhóm có sự khác biệt

```
library(stats)
TukeyHSD(anova_MPG,c("class"),ordered = FALSE)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = MPG ~ class + type + class * type, data = HybridTest)
##
## $class
##                            diff        lwr       upr     p adj
## Midsize SUV-Midsize Car  -5.75 -11.354116 -0.145884 0.0444736
## Small Car-Midsize Car     8.50   2.895884 14.104116 0.0055113
## Small SUV-Midsize Car    -2.25  -7.854116  3.354116 0.5956797
## Small Car-Midsize SUV    14.25   8.645884 19.854116 0.0001773
## Small SUV-Midsize SUV     3.50  -2.104116  9.104116 0.2640524
## Small SUV-Small Car     -10.75 -16.354116 -5.145884 0.0012450
```



Nhìn vào kết quả ta có thể thấy, với mức ý nghĩa 5%, có sự khác biệt giữa trung bình của 2 nhóm Small Car-Midsize Car, 2 nhóm Small Car-Midsize SUV, 2 nhóm Small SUV-Small Car và 2 nhóm Hybrid-Conventional. Điều này hợp lí với phần thống kê mô tả ở trên khi trung bình MPG của nhóm Small Car lớn nhất và lớn hơn đáng kể so với 3 nhóm còn lại.

Bài tập 2: PCA và phân tích nhân tố

Phương pháp Phân tích Thành phần Chính (PCA) là một công cụ mạnh mẽ trong thống kê và xử lý dữ liệu, với nhiều ứng dụng quan trọng. Một trong những công dụng chính của PCA là giảm chiều của dữ liệu, giúp tối ưu hóa quá trình xử lý và phân tích thông tin. Bằng cách chọn những thành phần chính giữ lại, PCA cho phép giữ lại lượng lớn thông tin quan trọng trong khi loại bỏ sự đa dạng không quan trọng. Ứng dụng đáng chú ý của PCA bao gồm khả năng trực quan hóa dữ liệu trong không gian ít chiều hơn, giúp nhìn thấy rõ cấu trúc và mối quan hệ trong dữ liệu. Ngoài ra, PCA cũng hữu ích trong việc nén dữ liệu, giảm chiều dữ liệu mà vẫn giữ lại sự biểu diễn quan trọng. Điều này không chỉ giúp tiết kiệm dung lượng lưu trữ mà còn tăng tốc các quá trình xử lý và phân tích.

Trong lĩnh vực phân loại và phân cụm, PCA là một công cụ quan trọng. Bằng cách giữ lại các thành phần chính, ta có thể xây dựng các mô hình phân loại hiệu quả hơn và tìm ra những đặc trưng quan trọng trong dữ liệu. PCA cũng thường được sử dụng để xử lý dữ liệu nhiễu và kiểm soát chiều dài dữ liệu trong các lĩnh vực như xử lý ngôn ngữ tự nhiên và chuỗi thời gian. Tóm lại, PCA không chỉ là một công cụ giảm chiều dữ liệu mà còn là một phương pháp mạnh mẽ giúp hiểu rõ và tận dụng thông tin quan trọng trong dữ liệu đa biến.

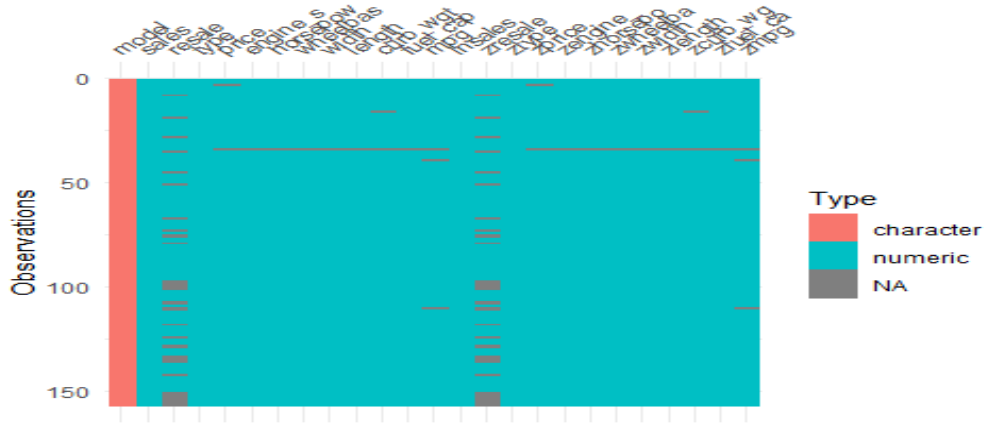
Dữ liệu được sử dụng: “car_sale”. Tệp dữ liệu này chứa các biến về doanh số(sales), giá niêm yết(price) và thông số kỹ thuật cho các kiểu dáng và mẫu xe khác nhau.

Một nhà phân tích muốn dự đoán doanh số bán ô tô từ một tập hợp các yếu tố đã cho. Tuy nhiên, nhiều yếu tố có mối tương quan với nhau và nhà phân tích lo ngại rằng điều này có thể ảnh hưởng xấu đến kết quả phân tích.

```
library(readxl)
car_sales <- read_excel("Assignment_TKT63/car_sales.xlsx")
car_sales<-data.frame(car_sales)
```

Nhận thấy dữ liệu đầu vào bị khuyết giá trị, cột thứ 14 là giá trị ln(sales), từ cột 15 đến cột 25 là các giá trị chuẩn hoá của cột 3 đến cột 13.

2.1. Xử lý dữ liệu



Với bộ dữ liệu đầu vào bị khuyết giá trị có các cách xử lý sau đây:

-Cách 1:Bỏ tất cả các dòng bị thiếu giá trị

```
car_sales1<-na.omit(car_sales)
```

-Cách 2: Thay những giá trị N/A bằng trung bình của biến đó

```
car_salesM<-car_sales[,1:13]
car_sales2=data.frame()
for(i in 2:ncol(car_salesM)) {
  mv<-is.na(car_salesM[,i])
  car_salesM[mv,i]=mean(car_sales[,i],na.rm = TRUE)
}
head(car_salesM)
```

-Cách 3:Hồi quy (regression impution)

-Cách 4:Stochatic regression imputation (Hồi quy có tính đến yếu tố ngẫu nhiên)

-Cách 5: K-Nearest Neighbors:Algorithm

Để đơn giản hoá việc tính toán và phân tích và mẫu cũng đủ lớn nên trong trường hợp này sẽ lấy tập dữ liệu sau khi đã bỏ các dòng có quan sát bị thiếu.

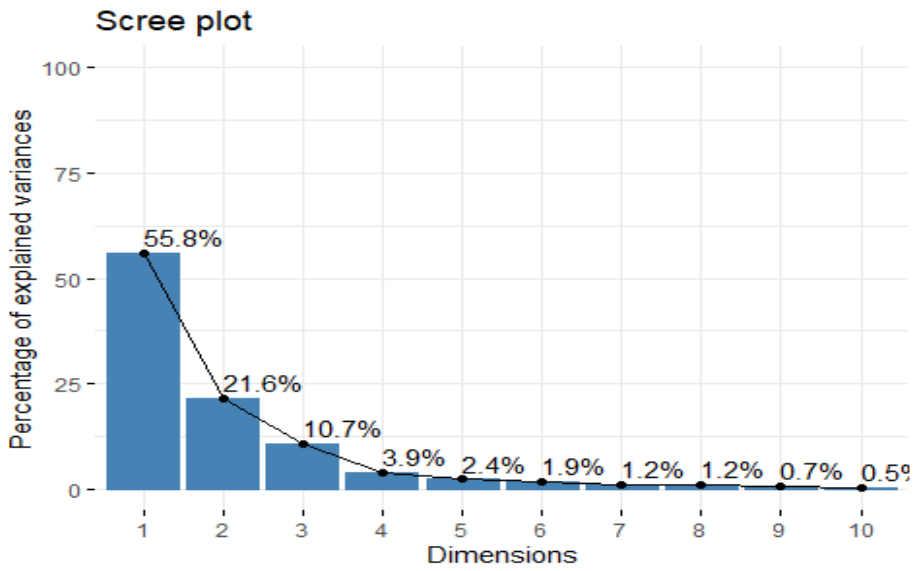
2.2. Phân tích thành phần chính (PCA)

Vì nhà phân tích muốn dự đoán doanh số bán ô tô từ một tập hợp các yếu tố đã cho. Tuy nhiên, nhiều yếu tố có mối tương quan với nhau và nhà phân tích lo ngại rằng điều này có thể ảnh hưởng xấu đến kết quả phân tích. Do đó sẽ sử dụng phân tích thành phần chính PCA để chọn ra các thành phần đại diện quan trọng nhất ảnh hưởng đến giá bán.

```
sales_PCA<-na.omit(car_sales1[,15:25])#lấy từ cột 15 trở đi vì đã được chuẩn
hoá
head(sales_PCA)
```

```
##      zresale      ztype      zprice   zengine_   zhorsepo    zwheelba
## 1 -0.1495606 -0.5926188 -0.41045828 -1.2070012 -0.8103784 -0.82278892
## 2  0.1573356 -0.5926188  0.07032257  0.1331567  0.6887312  0.08019843
## 4  1.0173434 -0.5926188  1.01794859  0.4203334  0.4241825  0.93083869
## 5  0.3651344 -0.5926188 -0.23695910 -1.2070012 -0.6340126 -0.63957410
## 6  0.4786380 -0.5926188  0.45703760 -0.2497456  0.2478166  0.15871907
## 7  1.8271477 -0.5926188  2.41151627  1.0904124  2.1878409  0.72145032
##        zwidth    zlength   zcurb_wg    zfuel_ca       zmpg
## 1 -1.11533688 -1.1125568 -1.1721235 -1.22222719  0.9705266
## 2 -0.24624321  0.4136772  0.2204185 -0.19339977  0.2700372
## 4  0.07242447  0.6891438  0.7485693  0.01236571 -0.4304523
## 5 -0.85460878 -0.6956344 -0.6027356 -0.39916525  0.7370301
## 6  1.43400456  0.3466718  0.2902042  0.14096914 -0.4304523
## 7  0.82563899  0.8082645  0.8310434  1.47844479 -0.6639487
```

```
## Trực quan tỉ lệ các thành phần chính
fviz_eig(pca,addlabels = TRUE,ylim=c(0,100))
```




```
PCA_s=prcomp(sales_PCA)
summary(PCA_s)

## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6
PC7
## Standard deviation     2.4942 1.5596 1.0897 0.66884 0.51948 0.46103
0.35849
## Proportion of Variance 0.5564 0.2175 0.1062 0.04001 0.02413 0.01901
0.01149
## Cumulative Proportion  0.5564 0.7739 0.8801 0.92008 0.94421 0.96322
0.97472
##                            PC8     PC9    PC10    PC11
## Standard deviation     0.35397 0.28091 0.23841 0.14718
## Proportion of Variance 0.01121 0.00706 0.00508 0.00194
## Cumulative Proportion  0.98592 0.99298 0.99806 1.00000
```

Kết quả nhận được có 3 giá trị eigen của 3 nhân tố đầu tiên lớn hơn 1 và 3 nhân tố này chiếm trên 88% tổng phương sai. Do đó số lượng thành phần chính nên giữ lại là 3. Từ dữ liệu ban đầu có 11 biến quan sát có thể dùng để dự báo doanh số bán ô tô nhưng sau khi sử dụng phân tích thành phần chính PCA đã lọc ra được 3 nhân tố chính không tương quan với nhau mà vẫn giải thích đầy đủ được cho doanh số bán ô tô như nhà phân tích mong muốn.

Sau khi đã chọn được 3 thành phần chính được giữ lại sẽ tạo một dữ liệu mới gồm doanh số bán hàng và 3 thành phần chính được giữ lại. Dữ liệu này sử dụng cho các bước phân tích tiếp theo của nhà phân tích và đã giải quyết được nỗi lo tương quan giữa các biến giải thích.

2.3. Đánh giá tác động của các thành phần chính lên doanh số

Sau khi đã loại bỏ được tương quan giữa các biến thu thập được ban đầu và chọn ra được 3 thành phần chính, tiến hành hồi quy doanh số với ba biến giải thích là 3 thành phần chính để đánh giá tác động.

```
library(lmtest)

library(stargazer)

reg = lm(zsale~PC1+PC2+PC3,data=sales_PCA1)
stargazer(reg,type = "text")

##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                  zsale
## ------------------------------------------------
## PC1                               0.038
##                                  (0.034)
##
## PC2                              -0.286***
##                                  (0.054)
##
## PC3                               0.033
##                                  (0.077)
##
## Constant                         -0.000
##                                  (0.083)
##
## ------------------------------------------------
## Observations                       117
## R2                                0.210
## Adjusted R2                       0.189
## Residual Std. Error         0.901 (df = 113)
## F Statistic           9.990*** (df = 3; 113)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Kết quả hồi quy cho ta mô hình ước lượng được:

$$zsale = 0.038 * PC1 - 0.286 * PC2 + 0.033 * PC3$$

Hai nhân tố 1 và 3 có tác động dương lên doanh số bán xe ô tô của cửa hàng trong khi nhân tố 2 lại có tác động ngược chiều. Điều này đúng với thực tế, bởi trong số biến cấu thành nên thành phần chính thứ 2 có bao gồm giá bán. Khi giá tăng thì có thể lượng người mua sẽ giảm đi từ đó dẫn đến doanh số giảm.

Bài tập 3: Phân tích cụm (Cluster Analysis)

Phân tích cụm(cluster analysis), là một phương pháp trong thống kê và học máy được sử dụng để tổ chức các quan sát hoặc biến thành các nhóm hay "cụm" có tính chất tương đồng. Mục tiêu chính của phân tích cụm là xác định cấu trúc tự nhiên hay phân bố ẩn trong dữ liệu mà không cần có sự can thiệp từ người phân tích. Quy trình phân tích cụm thường bao gồm việc đo lường sự tương đồng hoặc không tương đồng giữa các quan sát và sau đó tổ chức chúng thành các nhóm dựa trên các đặc trưng tương đồng nhau. Các phương pháp thường được sử dụng bao gồm K-Means Clustering, Hierarchical Clustering, và Gaussian Mixture Models. Ứng dụng của phân tích cụm rất đa dạng, từ việc phân loại khách hàng trong tiếp thị, phân nhóm quan sát trong nghiên cứu khoa học, đến việc tổ chức văn bản hay hình ảnh vào các nhóm có ý nghĩa. Phân tích cụm giúp hiểu rõ hơn về cấu trúc ẩn trong dữ liệu, giúp tạo ra cái nhìn toàn diện và hỗ trợ quyết định.

Có một tình huống như sau: Các nhà sản xuất ô tô cần có khả năng đánh giá thị trường hiện tại để xác định khả năng cạnh tranh cho xe của họ. Nếu ô tô có thể được nhóm theo dữ liệu có sẵn thì việc này sẽ có thể được thực hiện tự động bằng cách sử dụng phân tích cụm.

Thông tin về các kiểu dáng và nhãn hiệu xe khác nhau có trong tập dự liệu “car_sale” (giống như dữ liệu cho Bài tập 2). Sử dụng quy trình Phân tích cụm phân cấp (hierarchical) để nhóm các loại ô tô bán chạy nhất theo giá cả và đặc tính kĩ thuật của chúng.

3.1. Phương pháp phân cụm khoảng cách liên kết (linkage method) phù hợp

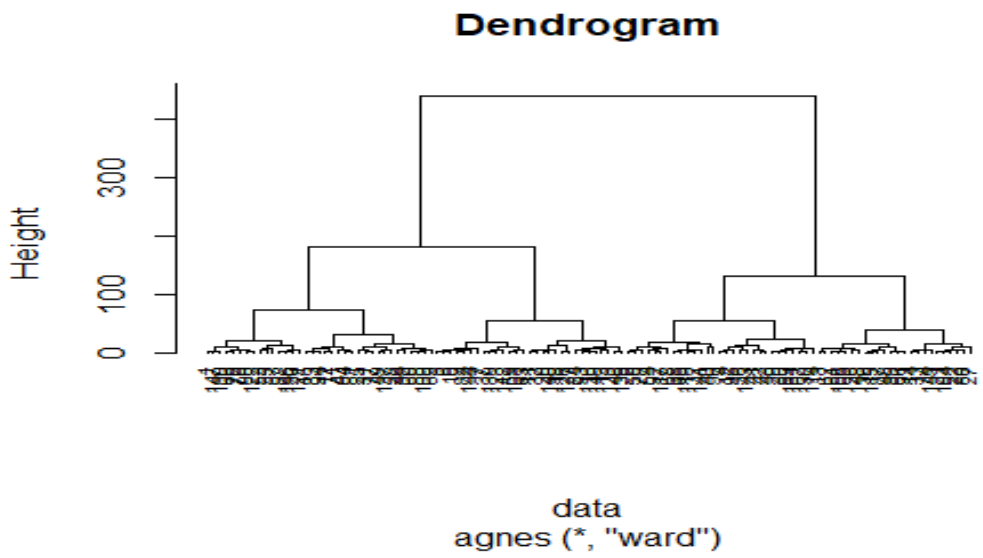
Trong phân tích cụm phân cấp có nhiều phương pháp khác nhau như: average, single, complete, ward…Vậy nên chúng ta cần xác định xem cần chọn phương pháp nào là phù hợp.

```
sapply(m,ac)
```

```
##   average    single  complete      ward
## 0.9426373 0.6219386 0.9698681 0.9921917
```

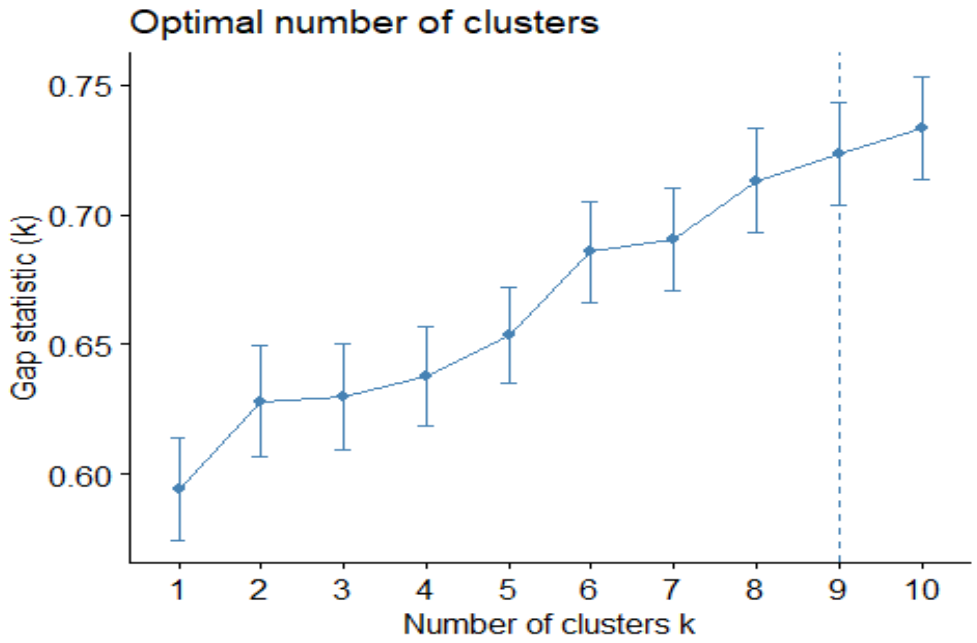
Để đánh giá được điều này thì cần tính đến hệ số tích tụ (agglomerative coefficient).Như kết quả nhận được hệ số tích tụ của phương pháp ward là lớn nhất vậy nên ta chọn phương pháp này.

```
clust<-agnes(data,method="ward")
## produce dendrogram
pltree(clust,cex=0.6,hang=-1,main = "Dendrogram")
```



3.2. Xác định số cụm

```
# Dùng gap statistic để xác định số nhóm
gap_stats<-clusGap(data[,-1],FUN = hcut,nstart=25,K.max = 10,B=50)
fviz_gap_stat(gap_stats)
```



Thông thường khi sử dụng Gap statistic để chọn số cụm thì sẽ chọn số cụm mà có có gap statistic cao nhất. Nhưng với dữ liệu này số cụm được chọn là 6, vì xét trên biểu đồ từ k=6 s đến k=7 đồ thị rất thoải hay ngay cả xét trên Dendrogram thì khoảng cách từ k=6 đến k=7 cũng không đáng kể.

3.3. Tiến hành phân cụm và đánh giá

```
#compute distance matrix (Tính toán ma trận khoảng cách)
d <- dist(data[,-1], method = "euclidean")

#perform hierarchical clustering using Ward's method
final_clust <- hclust(d, method = "ward.D2" )

#cut the dendrogram into 10 clusters
groups <- cutree(final_clust, k=6)
table(groups)
```

```
## groups
##  1  2  3  4  5  6
## 21 36 23  8 15 14
```

Cluster	Sales	Type	Price	Engine	Horsepow	MPG	...
1	55.56	0	13.99	1.85	115.48	**30.29**	...
2	34.95	0	30.13	3.47	210.69	23.61	...
3	72.80	0	18.75	2.48	152.43	25.65	...
4	6.23	0	**66.22**	**4.75**	**321.50**	20.75	...
5	**119.49**	1	24.22	3.54	179.60	19.67	...
6	69.60	1	23.97	3.21	173.43	20.36	...

Sau khi sử dụng các phương pháp hợp lí để tìm ra số cụm K = 6 sẽ tiến hành tính toán khoảng cách và chia dữ liệu về các loại xe thành 6 nhóm. Trong đó nhóm thứ 2 chứa nhiều quan sát nhất và nhóm thứ 4 có ít quan sát nhất.

Nhìn vào kết quả của 6 nhóm, doanh số của nhóm 5 cao nhất là 119.49 mà giá bán và các đặt tính kĩ thuật đều ở tầm trung, trong khi lượng tiêu hao năng lượng lại ở mức thấp nhất trong 6 nhóm. Từ đây có thể đánh giá tâm lý của khách hàng khi mua xe, họ sẽ ưu tiên những loại xe có giá vừa phải, vẫn đáp ứng được về các đặc tính kĩ thuật mà tiêu hao năng lượng thấp để có thể tối thiểu đi chi phí sửa chữa trong quá trình sử dụng.

Ngoài ra, nhóm 4 bao gồm các loại xe có giá thành đắt nhất, các đặc tính kĩ thuật gần như là tốt nhất nhưng lại kén khách hàng mua. Nhóm này phù hợp với phân khúc khách hàng có thu nhập cao.

Có thể thấy từ dữ liệu đầu vào với các quan sát rời rạc, sau khi sử dụng phương pháp phân tích cụm đã có thể gộp lại thành các nhóm có đặc tính giống nhau, dễ dàng hơn trong việc đánh giá và so sánh tổng quát.

Bài tập 4: Phân tích khác biệt (Discriminant Analysis)

Phân tích phân biệt (Discriminant Analysis - DA) là một phương pháp thống kê được sử dụng để phân loại các quan sát vào các nhóm đã được xác định trước. Mục tiêu chính của DA là tìm ra một phương trình hoặc các biến tuyến tính sao cho chúng có thể phân biệt rõ ràng giữa các nhóm.

Quy trình thực hiện phân tích phân biệt thường bắt đầu với việc xác định các nhóm cần phân loại và việc thu thập dữ liệu liên quan. Phương pháp này sau đó xác định các biến quan trọng nhất để tạo ra các phương trình phân biệt. Khi một quan sát mới được đưa vào mô hình, nó sẽ được phân loại vào nhóm mà nó gần với nhất dựa trên giá trị của các biến quan trọng. Công dụng của phân tích phân biệt là đa dạng, bao gồm:

1. ***Phân loại và Dự Báo:*** - DA được sử dụng để phân loại các quan sát vào các nhóm đã biết trước, và cũng có thể được sử dụng để dự báo nhóm cho các quan sát mới.

2. ***Nghiên Cứu Thị Trường và Tiếp Thị:*** - Trong nghiên cứu thị trường, DA giúp xác định đặc điểm quan trọng nhất đối với sự khác biệt giữa các nhóm khách hàng, từ đó tối ưu hóa chiến lược tiếp thị.

3. ***Phân Biệt Đặc Điểm Nhóm:*** - DA giúp hiểu rõ sự khác biệt giữa các nhóm dựa trên các biến quan trọng, có thể áp dụng trong nhiều lĩnh vực như y học, tâm lý học, và giáo dục.

4. ***Kiểm Định Đặc Điểm Nhóm:*** - Phân tích phân biệt được sử dụng để kiểm định xem có sự khác biệt đáng kể giữa các nhóm hay không, cung cấp thông tin quan trọng cho quyết định nghiên cứu. Phân tích phân biệt là một công cụ quan trọng trong thống kê đa biến, đặc biệt hữu ích khi muốn hiểu rõ sự khác biệt giữa các nhóm được xác định trước và áp dụng trong nhiều lĩnh vực nghiên cứu và ứng dụng thực tế.

4.1. Tìm hiểu tình huống

-Một hãng hàng không thu thập dữ liệu về nhân viên của họ đang làm 3 loại công việc:

1) nhân viên dịch vụ khách hàng
2) công nhân kĩ thuật
3) nhân viên điều phối

-Giám đốc Nhân sự của hãng muốn biết liệu ba cách phân loại công việc này có phù hợp với các loại tính cách khác nhau của nhân viên hay không?

-Mỗi nhân viên được thực hiện một loạt bài kiểm tra tâm lý để đánh giá, đo lường về 3 loại tính cách:

(i) tính cách hướng ngoại;

(ii) tính cách hòa đồng;

(iii) tính bảo thủ (khó thay đổi hoặc bị thuyết phục bởi người khác)

-Dự liệu trong tệp “discriminant_example” gồm biến:

JOB:loại công việc được phân loại;

OUTDOOR: Tính hướng ngoại;

SOCIAL: Tính hòa đồng;

CONSERVATIVE: Tính bảo thủ

4.2. Tiến hành phân tích khác biệt

```
library(MASS)
model <- lda(JOB~., data=train1)
model
```

```
## Call:
## lda(JOB ~ ., data = train1)
##
## Prior probabilities of groups:
##         1         2         3
## 0.3190184 0.3987730 0.2822086
##
## Group means:
##       OUTDOOR     SOCIAL CONSERVATIVE
## 1 -0.70426084  0.6838234  -0.43700501
## 2  0.62442898  0.1601706  -0.07579491
## 3 -0.03328216 -1.0359478   0.68745775
##
## Coefficients of linear discriminants:
##                      LD1        LD2
## OUTDOOR       0.5389918 -1.0754960
## SOCIAL       -1.0528910 -0.3321711
## CONSERVATIVE  0.4337316  0.2489320
##
## Proportion of trace:
##    LD1    LD2
## 0.7462 0.2538
```

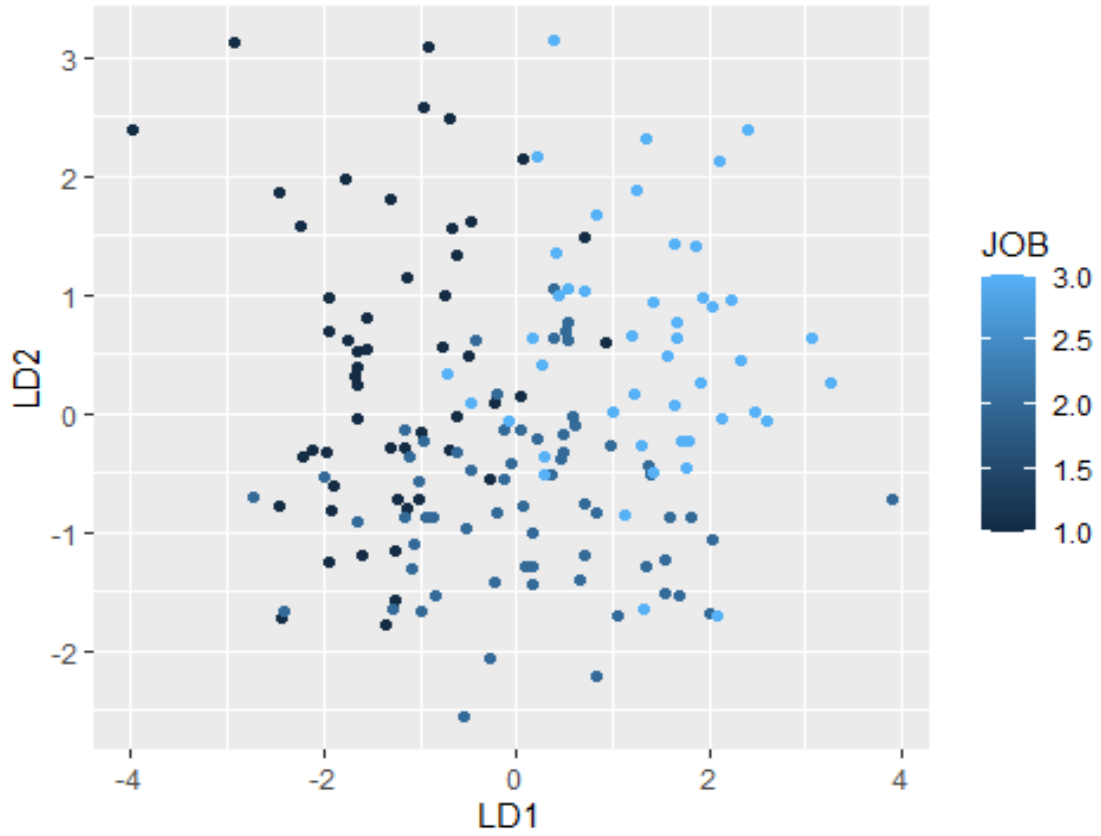
```
predicted <- predict(model,test1 )
print("Tỷ lệ dự báo đúng")
```

```
## [1] "Tỷ lệ dự báo đúng"
```

```
mean(predicted$class==test1$JOB,na.rm=TRUE)
```

```
## [1] 0.7160494
```

```
lda_plot <- cbind(train1, predict(model)$x)
library(ggplot2)
ggplot(data = lda_plot, mapping=aes(LD1,LD2)) + geom_point(aes(color = JOB))
```



Phương pháp phân tích khác biệt (Discriminal Analysis) dùng để phân loại các quan sát vào các nhóm đã biết trước dựa trên đặc điểm của chúng.

Ở tình huống này, dữ liệu đầu vào đã được thu thập gồm các biến JOB, OUTDOOR,SOCIAL và CONSERVATIVE với mục tiêu là liệu số liệu về tính cách của nhân viên này có phù hợp để phân 3 loại công việc.

Với 3 nhóm đã được biết trước, xác suất tiên nhiệm lần lượt của 3 nhóm là 0.32 ; 0.4 ; 0.28 tìm ra được tổ hợp tuyến tính giữa 4 biến giúp phân biệt các nhóm tốt nhất. Hàm phân biệt tuyến tính này đã giải thích được 74.62% sự biến động của tập dữ liệu.

$$JOB = 0.54 * OUTDOOR - 1.05 * SOCIAL + 0.43 * CONSERVATIVE$$

Tài liệu tham khảo

1.PCA-and-FA-commands.docx

2.Discriminant-Analysis-lecture-note.docx

3.R-Code-for-Discriminant-and-Cluster-Analysis.docx

4.An-Intro-to-Applied-Multi-Stat-with-R-by-Everitt-et-al-1.pdf

5.Clustering_Lecture.pdf