# Multivariate Analysis Exam 1

107508006 歐西四 陳葳芃

## Question 1

Please read in the mtcars dataset in R and name it data. Then, calculate the mean, median, standard deviation, minimum, and maximum values of all variables in the dataset and present them in a table. If there are missing values in the dataset, please mark them.

Sol.:

(a) I use this code to name the mtcars dataset as ***data***

```
>data<-mtcars
```

(b) In ***Table.1.1***, we could see the result of a few descriptive statistics which contains mean, median, standard deviation, minimum, and maximum values

| | mpg | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mean | 20.090625 | 6.187500 | 230.7219 | 146.68750 | 3.5965625 | 3.2172500 | 17.848750 | 0.4375000 | 0.4062500 | 3.6875000 | 2.8125 |
| Median | 19.200000 | 6.000000 | 196.3000 | 123.00000 | 3.6950000 | 3.3250000 | 17.710000 | 0.0000000 | 0.0000000 | 4.0000000 | 2.0000 |
| S.d. | 6.026948 | 1.785922 | 123.9387 | 68.56287 | 0.5346787 | 0.9784574 | 1.786943 | 0.5040161 | 0.4989909 | 0.7378041 | 1.6152 |
| Min. | 10.400000 | 4.000000 | 71.1000 | 52.00000 | 2.7600000 | 1.5130000 | 14.500000 | 0.0000000 | 0.0000000 | 3.0000000 | 1.0000 |
| Max. | 33.900000 | 8.000000 | 472.0000 | 335.00000 | 4.9300000 | 5.4240000 | 22.900000 | 1.0000000 | 1.0000000 | 5.0000000 | 8.0000 |

***Table.1 Descriptive statistics of the mtcars data***

(c) In ***Figure.1.1*** First, I want to check which column of the dataset contains missing value. Second, I try to calculate the amount of NA in each column, with the same result, I find nothing out. The result shows that it seems to no exist any missing value in this dataset, and I do not need to mark anything.

```
> #column names which contains NA
> names(which(colSums(is.na(data))>0))
character(0)
> #amount of missing values in each column
> colSums(is.na(data))
 mpg  cyl disp   hp drat   wt qsec   vs   am gear carb
   0    0    0    0    0    0    0    0    0    0    0
```

***Figure 1.1 The process of checking missing value***

## Question 2

Please calculate the correlation matrix of all variables in the data dataset and present the results in matrix form. If there are missing values in the dataset, please impute them first.

Sol.:

(a) I conduct the following code to try to impute missing values first

```
> data<-na.omit(data)
```

(b) In ***Table 2.1*** is he correlation matrix of all variables of the dataset—mtcars

```
> data_cor
            mpg        cyl       disp         hp        drat         wt        qsec         vs          am       gear        carb
mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594  0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958 -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799 -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479 -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000 -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157  0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953 -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870 -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059 -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000
```

***Table 2.1 Correlation matrix of 11 variables in mtcars dataset***

## Question 3

Please use regression analysis to explore the relationship between fuel economy (mpg) and other variables. First, standardize all variables in the dataset. Then, use regression analysis to predict fuel economy (mpg). Please note that if there are missing values in the dataset, please impute them first. Finally, list the coefficients, intercept, R-squared value, adjusted R-squared value, and other statistical measures of the regression model.

Sol.:

(a) Based on Question 2, we've checked there is no any missing value.

(b) In the beginning, I check the structure of all variables, standardize all variables[1] and present first six samples with head() function.

```
> ## Question 3
> str(data)
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
> data_st<-as.data.frame(scale(data))
> head(data_st)
                          mpg        cyl        disp         hp       drat           wt       qsec         vs         am       gear       carb
Mazda RX4          0.1508848 -0.1049878 -0.57061982 -0.5350928  0.5675137 -0.610399567 -0.7771651 -0.8680278  1.1899014  0.4235542  0.7352031
Mazda RX4 Wag      0.1508848 -0.1049878 -0.57061982 -0.5350928  0.5675137 -0.349785269 -0.4637808 -0.8680278  1.1899014  0.4235542  0.7352031
Datsun 710         0.4495434 -1.2248578 -0.99018209 -0.7830405  0.4739996 -0.917004624  0.4260068  1.1160357  1.1899014  0.4235542 -1.1221521
Hornet 4 Drive     0.2172534 -0.1049878  0.22009369 -0.5350928 -0.9661175 -0.002299538  0.8904872  1.1160357 -0.8141431 -0.9318192 -1.1221521
Hornet Sportabout -0.2307345  1.0148821  1.04308123  0.4129422 -0.8351978  0.227654255 -0.4637808 -0.8680278 -0.8141431 -0.9318192 -0.5030337
Valiant           -0.3302874 -0.1049878 -0.04616698 -0.6080186 -1.5646078  0.248094592  1.3269868  1.1160357 -0.8141431 -0.9318192 -1.1221521
```

***Table 3.1 Structure of variables and the first six samples***

(c) About the variable selection, I would like to perform stepwise regression analysis to decide my model. Both of the methods (forward and backward) compute the same selection of variables. So we use hp + wt to predict mpg.

```
> summary(forward.lm)

Call:
lm(formula = mpg ~ wt + hp + drat, data = train_df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.53854 -0.30710 -0.07147  0.18105  1.01172

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02175    0.07913  -0.275 0.785672
wt          -0.40183    0.14712  -2.731 0.011397 *
hp          -0.39947    0.10480  -3.812 0.000802 ***
drat         0.22101    0.11938   1.851 0.075967 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4247 on 25 degrees of freedom
Multiple R-squared:  0.8384,    Adjusted R-squared:  0.819
F-statistic: 43.23 on 3 and 25 DF,  p-value: 4.836e-10
```

Base on the linear model
>lm(mpg ~ hp + wt, data = train_df[2]),
named *forward.lm* here. We could obtain the statistical measures we need in ***Table 3.2***.

Here are some statistical result:
*Coefficients (refer to the ***Table3.2***)
*Intercept = -0.2175
*R-squared value = 0.8384
*Adjusted R-squared value = 0.819

***Table3.2 Summary of fitting model***

[1] With respect to the response variable, for the prediction need, I decide to use the original value (before standardization) rather than standardized ones. As for independent variables, it will conduct in the model with standardized form.

[2] The name of training data frame. To avoid overfitting, I divide the original data set into training and testing parts. But, the sample size is not large enough when training set takes 70% part of full data. As the result of this, the training set will have 90% portion (n=29) of full data.

## Question 4

Please use regression diagnostics to determine whether the regression model in Question 3 satisfies the assumptions of the regression model. Specifically, perform normality tests, homoscedasticity tests, and independence tests on the residuals of the model. If the model does not meet the assumptions, please explain the reasons and suggest improvements.

Sol.:

To determine the model is satisfies the assumptions of the regression model or not, I will perform the following 3 tests, normality tests, homoscedasticity tests, independence tests separately.

(a) Normality tests

Based on the ***Table 4.1***, at α =0.05, the result of normality tests **won't reject H0** which claim that **the residual is normal distributed** with p-value = 0.05938.

(b) Homoscedasticity tests

Based on the ***Table 4.2***, at α =0.05, the result of homoscedasticity tests **won't reject H0** which claim that **the error is constant along the values of the dependent variable** with p-value = 0.4029.

(c) Independence tests

Based on the ***Table 4.3***, at α =0.05, the result of homoscedasticity tests **won't reject H0** which claim that **there is no relation between the different examples** with p-value = 0.9565.

```
> #perform normality tests
> residuals <- forward.lm$residuals
> qqnorm(residuals)
> qqline(residuals)
> shapiro.test(residuals)

        Shapiro-Wilk normality test

data:  residuals
W = 0.93128, p-value = 0.05938
```

***Table 4.1 The result of normality tests***

```
> #perform homoscedasticity tests
> library(car)
> leveneTest(residuals ~ as.factor(train_df$gear), data = train_df)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  0.9415 0.4029
      26
```

***Table 4.2 The result of homosscedaticity tests***

```
> #perform independence tests
> library(lmtest)
> dwtest(forward.lm)

        Durbin-Watson test

data:  forward.lm
DW = 2.5857, p-value = 0.9565
alternative hypothesis: true autocorrelation is greater than 0
```

***Table 4.3 The result of independence tests***

## Question 4

Please use regression diagnostics to determine whether the regression model in Question 3 satisfies the assumptions of the regression model. Specifically, perform normality tests, homoscedasticity tests, and independence tests on the residuals of the model. If the model does not meet the assumptions, please explain the reasons and suggest improvements.

Sol.:

(d) The reasons might be the relationship between independent variables and dependent is nonlinear, independent variables be strongly collinear, the extremely value exists in the data or the samples is just not large enough.

In case of that the model does not meet the assumptions, we could……
transform the dependent variable (Box-Cox transformation), redefine the dependent variable (y), remove outlier/ highly related independent variables, set the weight to the independent variables .

```
here is a few example code if we choose to do the Box-Cox
> transformation
> library(MASS)
> bc=boxcox(forward.lm, lambda=seq(-3,3))
> best.lm=bc$x[which(bc$y==max(bc$y))]
> new.forward.lm= lm((mpg)^(best.lm)~hp + wt, data=train_df)
> plot(new.forward.lm)
```

# Appendix: R code

```
##### Multivariate Analysis Exam 1 #####
### Student ID: 107508006
### Department: 歐西四
### Name      : 陳葳芃

## Question 1
#read in the mtcars data set in R and name it "data".
data<-mtcars

#present a few descriptive statistics in a table
#order: mean, median, standard deviation, minimum, maximum
rbind("Mean   :"=sapply(data, mean),
      "Median :"=sapply(data, median),
      "S.d.   :"=sapply(data, sd),
      "Min.   :"=sapply(data, min),
      "Max.   :"=sapply(data, max))

#column names which contains NA
names(which(colSums(is.na(data))>0))
#amount of missing values in each column
colSums(is.na(data))


## Question 2
#impute missing values first
data<-na.omit(data)
#calculate and present the correlation matrix of all variables
data_cor<-cor(data)
data_cor

## Question 3
str(data)
data_st<-as.data.frame(scale(data))
data_st[,1] <- data[,1]
set.seed(69)
#amount of missing values in each column
colSums(is.na(data_st))

#切訓練及測試集（因考量到樣本數，故比例調為9 train:1 test）
train.index <- sample(x=1:nrow(data_st),
size=ceiling(0.9*nrow(data_st) ))
train_df <- data_st[train.index, ]
test_df <- data_st[-train.index, ]
m.null = lm(mpg ~ 1, data = train_df)
m.full = lm(mpg ~ ., data = train_df)

#模型變數挑選stepwise regression analysis
```

```
forward.lm = step(m.null, scope=list(lower=m.null, upper=m.full),
direction="forward")
summary(forward.lm)
backward.lm = step(m.full, scope = list(upper=m.full),
direction="backward")
summary(backward.lm)
#兩個模型挑選結果均為lm(mpg ~ hp + wt, data = train_df)

#比較各模型（full model, forward.test, backward.test）與真實值的結果
v1<-test_df[rownames(test_df),1]
Q<-matrix(data=v1,ncol = nrow(test_df),nrow=1,byrow=TRUE)
colnames(Q)=rownames(test_df)
rownames(Q)="Actual value"
Q<-as.table(Q)

predict_result<-rbind( "lm.test" = predict(m.full, test_df),
                       "forward.test" = predict(forward.lm,
test_df),
                       "backward.test" = predict(backward.lm,
test_df),Q)
t(predict_result)


## Question 4
#簡易check回歸假設的方式
plot(forward.lm)
#常態大致符合V(但在head tail的表現不是非常理想)
#變異數同質性V()
#獨立性V(大致在0附近隨機跳動)

#perform normality tests
residuals <- forward.lm$residuals
qqnorm(residuals)
qqline(residuals)
shapiro.test(residuals)
#結果顯示，p值為 0.05938，因此我們無法拒絕常態分配假設

#perform homoscedasticity tests
library(car)
leveneTest(residuals ~ as.factor(train_df$gear), data = train_df)
#結果顯示，p值為0.8669，因此我們無法拒絕變異數同質性假設。

#perform independence tests
library(lmtest)
dwtest(forward.lm)
#結果顯示，p值為0.817，因此我們無法拒絕獨立假設。


#如何解決heteroscedasticity 之問題？
```

```
#調整變數權數、重新定義應變數(y)、對應變數(y)作轉換
#here is an example if we choose to do the Box-Cox transformation
library(MASS)
bc=boxcox(forward.lm, lambda=seq(-3,3))
best.lm=bc$x[which(bc$y==max(bc$y))]
new.forward.lm= lm((mpg)^(best.lm)~hp + wt, data=train_df)
plot(new.forward.lm)
```