Discrete analysis_HW1

ID : 111024517 Name : 鄭家豪

```
library(GGally)
library(MASS)
```

Problem 1

(Please read Faraway (2006 or 2016), Chapter 1 before you do this question) The teengamb data has 47 rows and 5 columns. The data was obtained from a survey conducted to study teenage gambling in Britain. This 5 columns in the data are:

sex: 0=male, 1=female

status: Socioeconomic status score based on parents' occupation

income: in pounds per week

verbal: verbal score in words out of 12 correctly defined

gamble: expenditure on gambling in pounds per year

Use **gamble** as the response and the other variables as explanatory variables to perform a data analysis. Your data analysis should consist of:

a. an initial data analysis that explores the numerical and graphical characteristics of the data
b. an exploration of transformations to improve the fit of the model
c. diagnostics to check the assumptions of your model
d. some predictions of future observations for interesting values of the predictors
e. an interpretation of the meaning of the model (parameters) with respect to the particular area of application

Notice that there is always some freedom in deciding which method to use, in what order to apply them, and how to interpret the results. So, there may not be one clear right answer, and good analysts may come up with different models.

Sol.

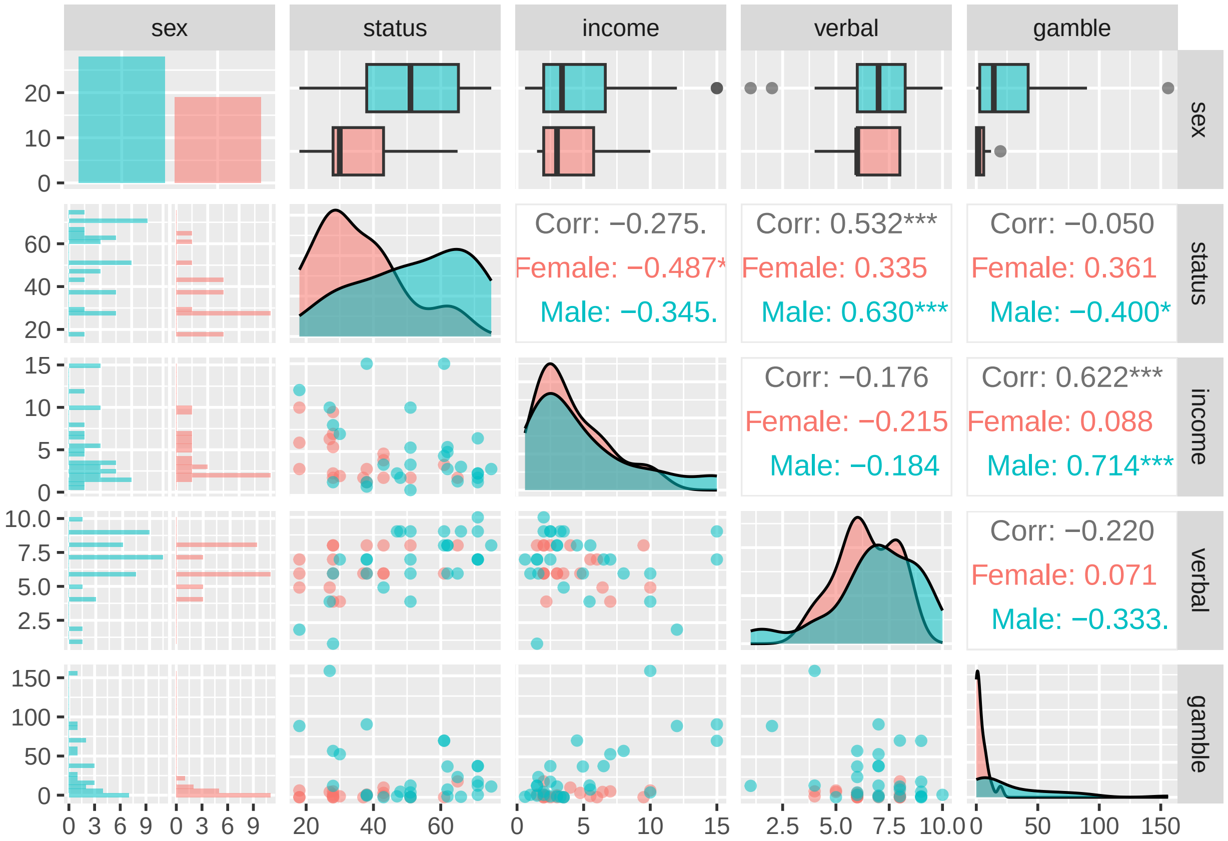
(a)

First, read the teengamb.txt in R console, named it "data1":

```
data1 = as.data.frame(read.table("teengamb.txt",header = TRUE))
data1$sex = as.factor(data1$sex)
```

And use ggpairs from package "GGally" to explore the numerical and graphical characteristics of this data:

```
ggpairs( data1, mapping = aes( color=factor(sex, levels = c(1,0),
                                         labels=c("Female","Male")), alpha=0.3 ) )
```



The following features can be observed from the above plotted chart:

- There is a significant difference in gender numbers (see position (1,1)).
- Males tend to spend more on gambling compared to females (see position (1,5)).
- The variable “gamble” exhibits outliers (see position (1,5)).
- The variable “gamble” has a right-skewed distribution(see position (5,5)). • Income may have a significant effect on gambling behavior (see position (3,5)).
- The correlation coefficient between “verbal” and “status” is slightly higher, there may exist potential collinearity (see position (2,4)).

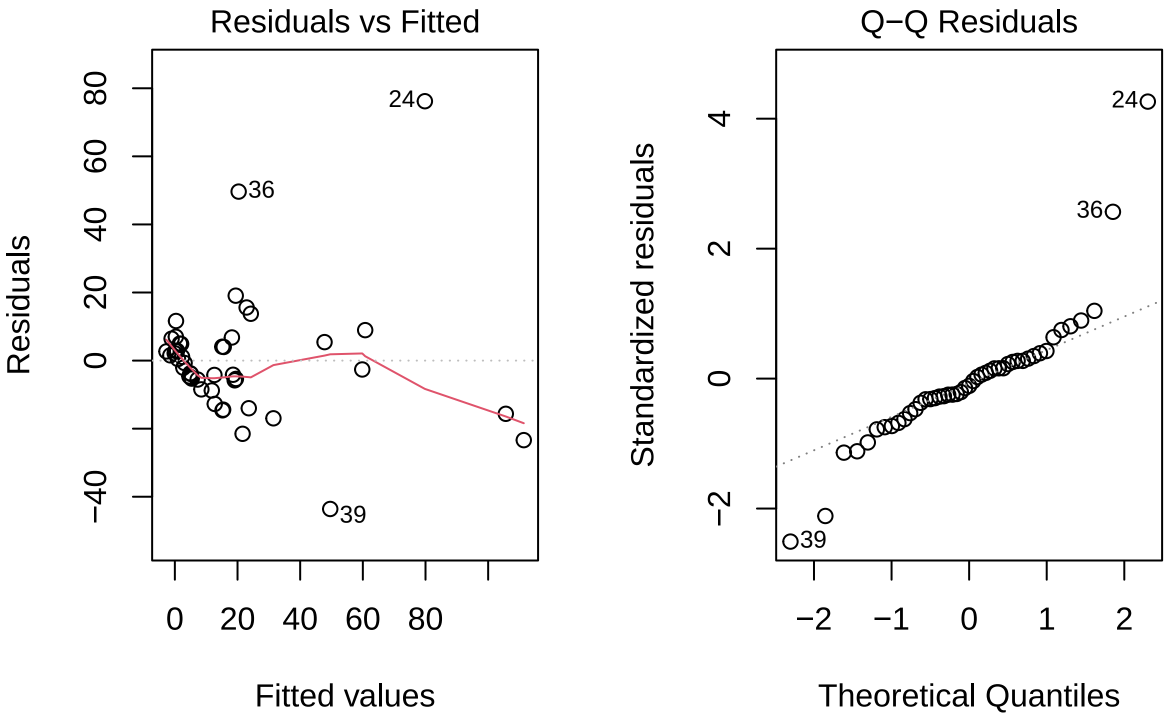
(b)(c)

- Model 1:

gumble ∼ 1 + sex1 + status + income +verbal + sex1:status + sex1:income + sex1:verbal + status:income + status:verbal + income:verbal

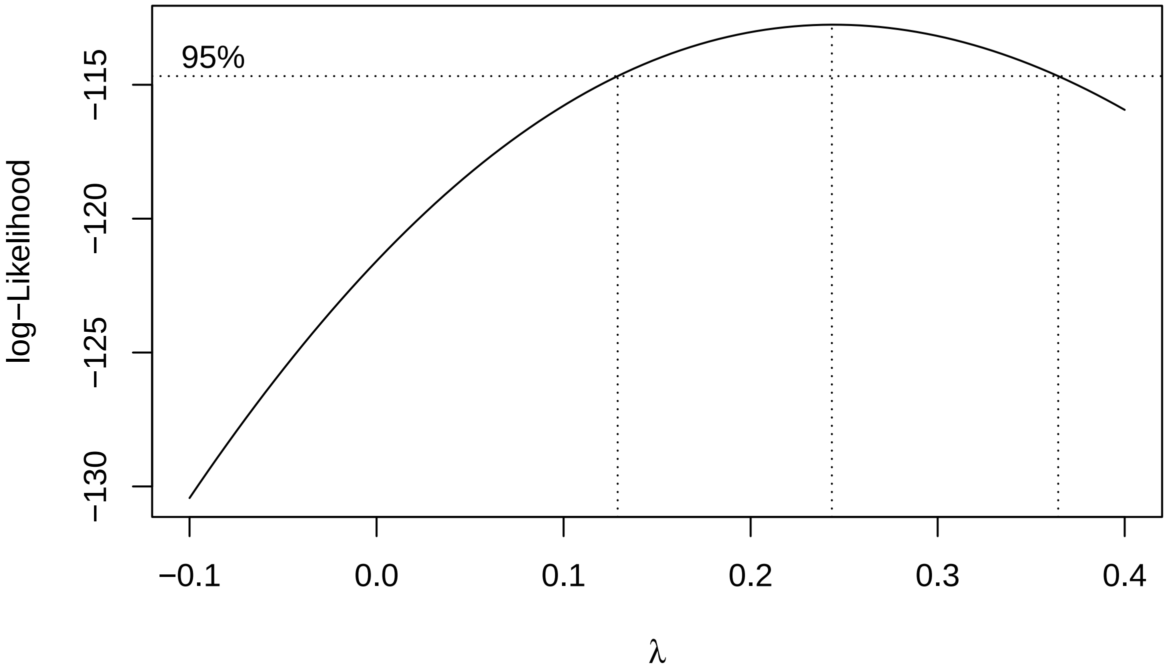
# Residual plot and Q-Q plot (for diagnostics):

```
m1 = lm(gamble ~ (.)^2 , data=data1)
par(mfrow=c(1,2))
plot(m1,1:2)
```



It appears that there is a violation of the assumptions of constant variance and normality for error term. Therefore, choosing Box-Cox transformation for response transformation would be appropriate (for avoiding invalid transformation, shifting gamble right by 0.01) :

```
boxcox(gamble +0.01 ~ (.)^2, data=data1,lambda=seq(-0.1,0.4,0.01), plotit =TRUE)
```

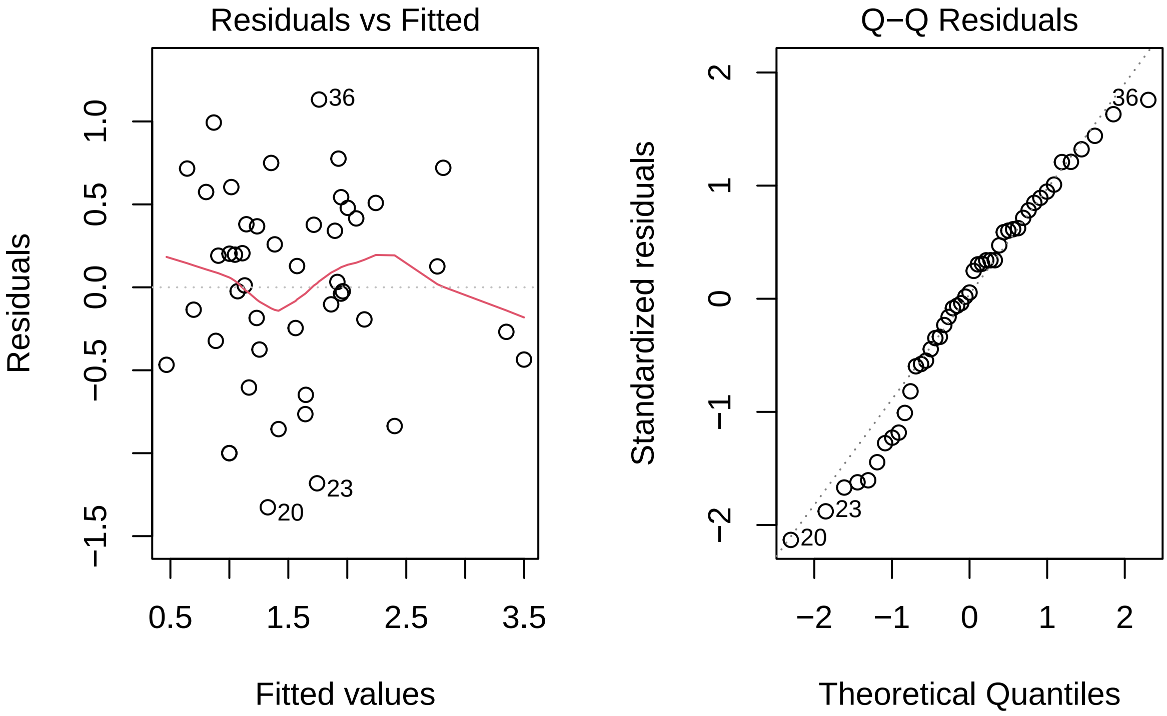


Based on the observation from the plot, selecting lambda = 0.25 to transform the variable “gamble” by taking the fourth root of “gamble,” and then refit Model 2.

- Model 2:

gumble$^{1/4}$ ~ 1 + sex1 + status + income +verbal + sex1:status + sex1:income + sex1:verbal + status:income + status:verbal + income:verbal

```
m2 = lm(gamble^0.25 ~ (.)^2, data=data1)
par(mfrow=c(1,2))
plot(m2,1:2)
```



It appears that the assumptions of constant variance and normality are satisfied. Next, let’s do variables selection (via AIC) by Bidirectional elimination:

```
vs1=step(m2,direction = "both",trace=0)
summary(vs1)
```

```
Call:
lm(formula = gamble^0.25 ~ sex + status + income + verbal + sex:income +
    status:income, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3517 -0.3689  0.1183  0.3783  1.1277

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.794009   0.668753   1.187  0.24211
```

```
sex1          -0.096056   0.373897  -0.257  0.79857   
status         0.027300   0.011439   2.387  0.02183 * 
income         0.262762   0.088544   2.968  0.00505 **
verbal        -0.140544   0.061256  -2.294  0.02710 * 
sex1:income   -0.106351   0.070808  -1.502  0.14096   
status:income -0.002742   0.001830  -1.498  0.14190   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6332 on 40 degrees of freedom
Multiple R-squared:  0.559, Adjusted R-squared:  0.4929 
F-statistic: 8.451 on 6 and 40 DF,  p-value: 6.02e-06
```

There are still multiple non-significant variables present. Try progressively removing variables “sex” with higher p-values:

```r
vs2=lm(gamble^0.25 ~ status + income + verbal + sex:income + status:income ,data=data1)
summary(vs2)
```

```

Call:
lm(formula = gamble^0.25 ~ status + income + verbal + sex:income + 
    status:income, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.3327 -0.3504  0.1544  0.3872  1.1385 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    0.707615   0.571427   1.238 0.222637    
status         0.028531   0.010269   2.778 0.008204 ** 
income         0.273749   0.076644   3.572 0.000923 ***
verbal        -0.141068   0.060521  -2.331 0.024756 *  
income:sex1   -0.120703   0.043009  -2.806 0.007629 ** 
status:income -0.002905   0.001697  -1.712 0.094454 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6259 on 41 degrees of freedom
Multiple R-squared:  0.5583,    Adjusted R-squared:  0.5044 
F-statistic: 10.36 on 5 and 41 DF,  p-value: 1.786e-06
```

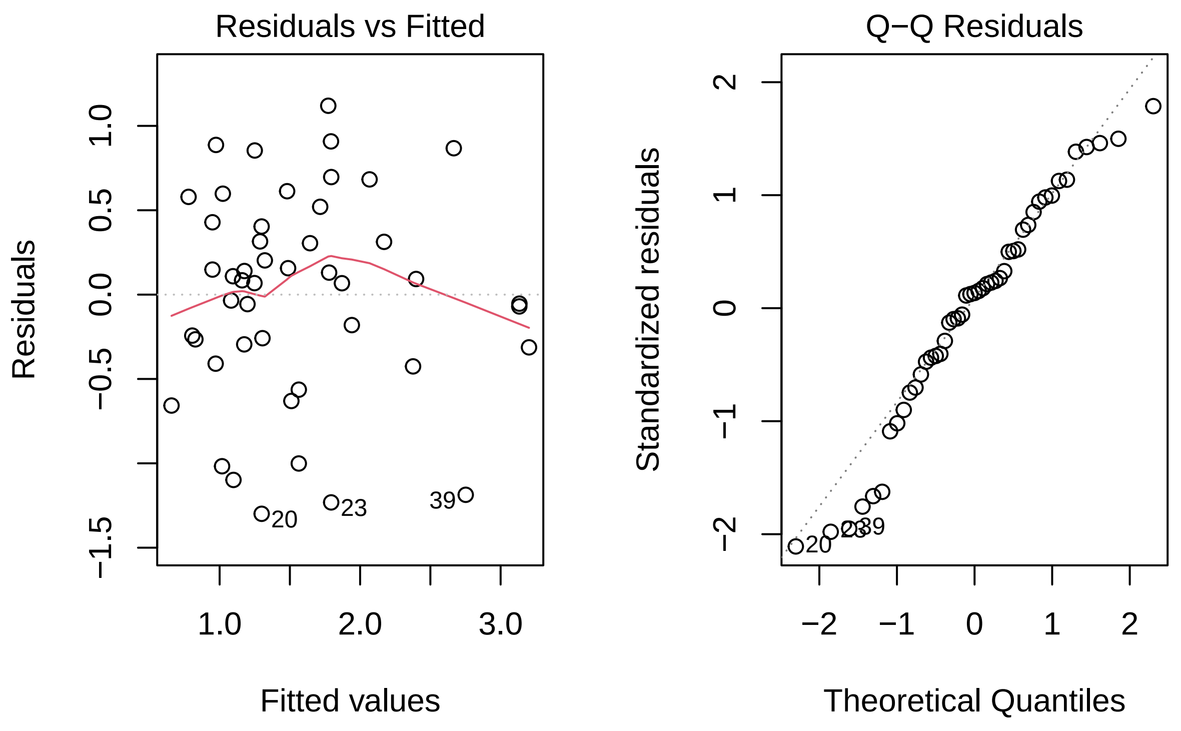
Continuing to remove “status:income” :

```r
vs3 = lm(gamble^0.25 ~ status + income + verbal + sex:income ,data=data1)
summary(vs3)
```

```

Call:
lm(formula = gamble^0.25 ~ status + income + verbal + sex:income, 
    data = data1)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.29910 -0.30377  0.08468  0.41612  1.11924 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.335453   0.448182   2.980  0.00478 ** 
status       0.016620   0.007724   2.152  0.03722 *  
income       0.151017   0.027725   5.447 2.47e-06 ***
verbal      -0.156947   0.061165  -2.566  0.01395 *  
income:sex1 -0.106131   0.043116  -2.461  0.01803 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6402 on 42 degrees of freedom
Multiple R-squared:  0.5267,	Adjusted R-squared:  0.4816 
F-statistic: 11.68 on 4 and 42 DF,  p-value: 1.816e-06
```

```r
par(mfrow=c(1,2))
plot(vs3,1:2)
```



So, $\text{gamble}^{0.25} = 1.335453 + 0.016620\ \text{status} + 0.151017\ \text{income} - 0.156947\ \text{verbal} - 0.106131\ \text{income} \times \text{sex}$ is the fitted model.

(d)(e)

For different gender, let's denote the fitted models as Male and Female, respectively:

$\text{Male}: \ \text{gamble}^{0.25} = 1.335453 + 0.016620\ \text{status} + 0.151017\ \text{income} - 0.156947\ \text{verbal}$

$$\text{Female : gamble}^{0.25} = 1.335453 + 0.016620\ \text{status} + 0.044886\ \text{income} - 0.156947\ \text{verbal}$$

We can observe that for both male and female teenagers :

- There is a weak positive effect on the socioeconomic status of their parents.
- There is a positive effect on the income of their parents, although it is weaker for females.
- There is a negative effect on their verbal scores.

These findings align with general expectations. Using this fitted model, predictions can be made for different combinations of teenager. However, it's crucial to note that if the variables in new data differ from those on which the fitted model depends, it may lead to unstable estimates.

Here are the examples to illustrate:
Example 1 (Interpolation): A male from a high-income family, with parents holding prestigious positions and substantial income, and excelling in verbal scores. (sex = 0, status= 75, income= 15, verbal = 10)

```
ex1 = data.frame(sex=as.factor(0), status=75, income = 15, verbal = 10)
predict(vs3, ex1, interval = "prediction")
```

```
       fit      lwr      upr
1 3.277729 1.756571 4.798886
```

Example 2 (extrapolation): A male from a low-income family, with parents having low-ranking occupations and minimal income, and struggling with verbal scores. (sex = 0, status= 1, income= 0.1, verbal = 0.1)

```
ex2 = data.frame(sex=as.factor(0), status=1, income = 0.1, verbal = 0.1)
predict(vs3, ex2, interval = "prediction")
```

```
      fit        lwr      upr
1 1.35148 -0.2169876 2.919947
```

Problem 2

In the following examples, distinguish between response and explanatory variables.
a. Attitude toward abortion on demand (favor, oppose); gender (male female).
b. Cholesterol level; heart disease (yes, no).
c. Race (white, nonwhite); gender (male, female); vote for President (Republican, Democrat, Other); income.
d. Hospital (A, B); treatment (T1, T2); patient outcome (survive, die).

Sol.

a. response: Attitude toward abortion on demand ; explanatory: gender.

b. response: heart disease; explanatory: Cholesterol level.

c. response: vote for President; explanatory: race, gender, income.

d. response: patient outcome; explanatory: hospital, treatment.

Problem 3

Identify each variable as nominal, ordinal, or interval
a. Political party affiliation (Democrat, Republican, other)
b. Location of hospital in which data collected (London, Boston, Madison, Rochester, Toronto)

c. Highest degree obtained (none, high school, bachelor's, master's, doctorate)
d. Favorite beverage (beer, juice, milk, soft drink, wine, other)
e. Patient condition (good, fair, serious, critical)
f. Patient survival (in number of months)
g. Rating of a movie with 1 to 5 stars, representing (hated it, didn't like it, liked it, really liked it, loved it)

Sol.

● nomial: a, b, d.
- Political party affiliation is nominal variable because they are simply divided into categorical variables.
- The location where data is collected typically falls under the category of nominal variables, as it represents directional information.
- Favorite beverage is considered a nominal variable because the categories (beer, juice, milk, soft drink, wine, other) do not have a natural order or ranking.

● ordinal: c, e, g.
- These three variables all have an inherent order among their levels, therefore they belong to ordinal variables.

● interval: f.