Data Analysis Class Test 1

2700298

1 Report: Gambling among teenagers in Britain

1.1 Introduction

The study was conducted into the gambling habits of teenagers living in Britain which included 28 males and 19 females aged 16-19.The respondents were asked how much they spend on gambling in pounds per year.The main purpose of this study is to examine the expenditure of teenagers in Britain on gambling and determine whether there is a difference in spending habits between males and females and try to find the relationship between gambling expenditure and gender by linear model.

Section 1.2 consists of an exploratory data analysis of gambling expenditure and explores the potential relationship between gambling expenditure and gender. Section 1.3 contains the results from fitting a linear regression model to the data, as well as the assessment of the model assumptions. Concluding remarks are given in Section 1.4.

1.2 Exploratory Analysis

Table 1 contains summary statistics on the gambling expenditures of the 28 males and 19 females aged from 16-19. First, we notice that the missing value for both males and females are 0 which is very good.Secondly,in terms of female respondents,we see that the middle 50% of gambling expenditures lie between 0.1 and 6 , with an average expenditure of 3.87. In addition,if we look at male respondents,we could find that the middle 50% of gambling expenditures lie between 2.78 and 42.17 , with an average expenditure of 29.77.What's more,the maximum expenditure for male is 156 pounds which is much greater than female's expenditure (19.6 pounds) while the minimum expenditure is close between two groups.

Table 1: Summary statistics on gambling expenditure by sex

Variable	Sex	Missing	Complete	Mean	SD	Minimum	1st quartile	Median	3rd quartile	Maximum
spent	Female	0	1	3.87	5.15	0	0.10	1.70	6.00	19.6
spent	Male	0	1	29.77	37.32	0	2.78	14.25	42.17	156.0

Then we prefer to draw boxplot of gambling expenditure by sex

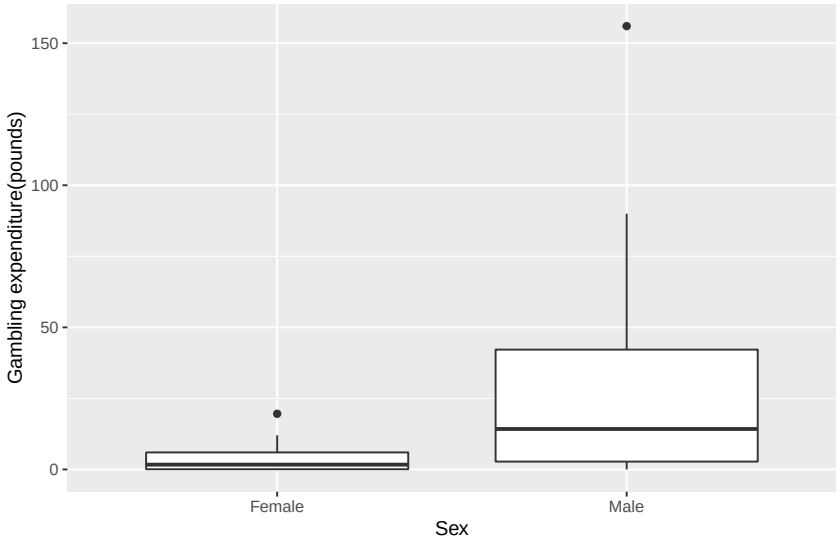

Figure 1: Boxplot of Gambling expenditure by sex

Figure 1 displays a boxplot of the male's gambling expenditure as well as female's gambling expenditure. Here, There is more variability in male's expenditure than female's expenditure, as seen from box range of two groups,showing that there is a big difference in male's interest in gambling.From Figure 1, we see that the mean expenditure for male is a little bit higher than female group which indicates that male would spend more money on gambling compared with female generally speaking.Also,there were outliers in both male group and female group which should be ignored. A linear regression model will now be fitted to assess the relationship between gambling expenditure and gender.

1.3 Formal Analysis

The linear regression model that will be fitted to the data is as follows:

$$\widehat{\text{spent}} = \hat{\alpha} + \hat{\beta}_{\text{Male}} \cdot \mathbb{I}_{\text{Male}}(x)$$

where

- the intercept $\hat{\alpha}$ is the mean expenditure for the baseline category (females);
- $\hat{\beta}_{\text{Male}}$ is the difference in the mean expenditure of males relative to the baseline category (females);
- $\mathbb{I}_{\text{Male}}(x)$ is an indicator function such that

$$\mathbb{I}_{\text{Male}}(x) = \begin{cases} 1 & \text{if the gender of } x\text{th observation is Male,} \\ 0 & \text{Otherwise.} \end{cases}$$

Table 2 displays the estimated intercept and slope parameters of the best-fitting line from the regression model.

Table 2: Estimates of the intercept and slope from the fitted linear regression model.

term	estimate
intercept	3.866
sex: Male	25.909

Hence, the best-fitting line is given as:

$$\widehat{\text{gambling}}\ \text{spent} = 3.866 + 25.909 \cdot \mathbb{I}_{\text{Male}}(x)$$

That is, the mean gambling expenditure for female is simply equal to the intercept term 3.866,then the mean gambling expenditure for male is 3.866+25.909.

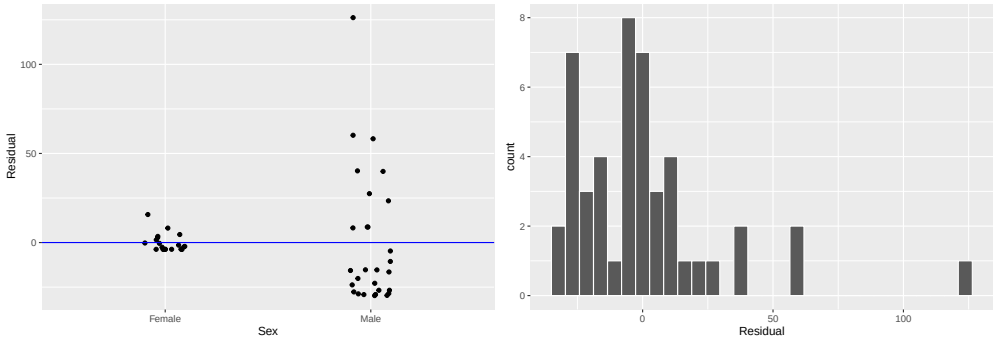

Figure 2: Scatterplot of the residuals against sex (left) and a histogram of the residuals (right).

Figure 2 displays a scatterplot of the residuals against the gender and a histogram of the residuals. From the scatterplot,we see that there is an even spread of the residuals above and below the zero line for both female group and male group after ignoring the outliers existed in each group,and hence our assumption that the residuals have mean zero appears valid. The histogram appears to be slightly right-skewed,however,it appears to be relatively bell-shaped and centred around zero. Hence, the assumption of normally distributed errors appears to hold for the fitted regression model.

1.4 Conclusions

The spending habits between males and females youths living in Britain seem to be quite different.Male youths spend 29.77 pounds on average per year while female youths spend 3.87 pounds on average per year which indicates that male youths are more fond of spending money on gambling.In addition,the relationship between gambling expenditure and gender is

$$\widehat{\text{gambling expenditure}} = 3.866 + 25.909 \cdot \mathbb{I}_{\text{Male}}(x)$$

which is observed from the fitted regression model,so the difference between male youths expenditure and female youths expenditure is 25.909 pounds on average per year.

2 Further Question 1

2.1 (a)

```
FQ1<-read_csv("FQ1.csv")

ggplot(data=FQ1,mapping=aes(x=X,y=Y))+
  geom_point()+
  geom_jitter(width = 0.1, height = 0.1)+
  labs(x="X",y="Y")+
  geom_smooth(method = "lm", se = FALSE)
```

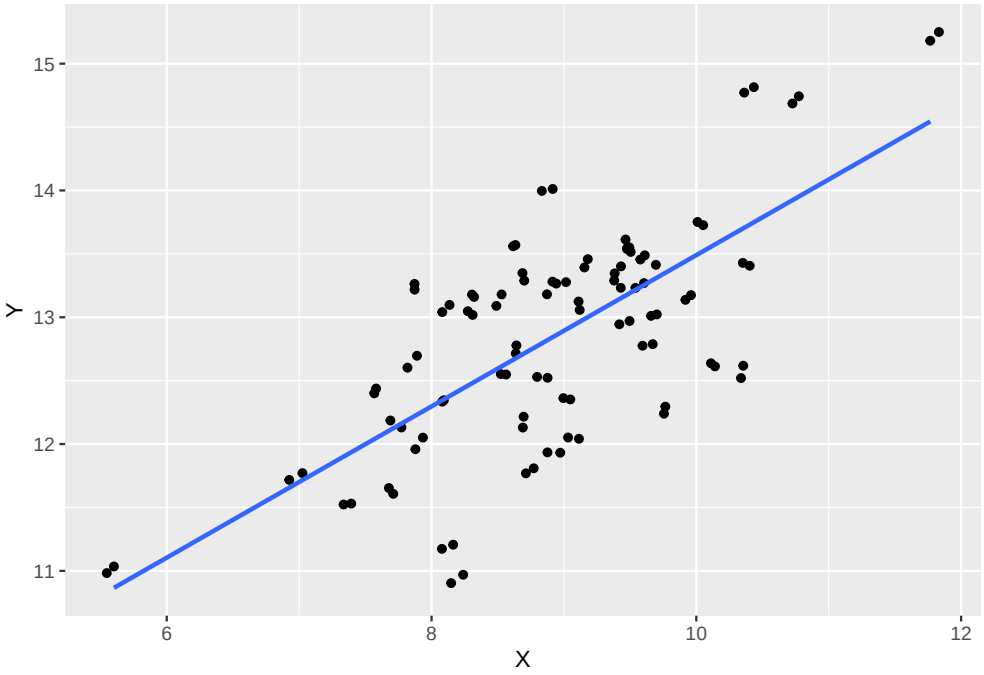

Figure 3: Relationship between X and Y

Figure 3 clearly shows that there is a positive relationship between X and Y,in order to specifiy how strong the positive correlation is,we could use get_correlation() function to calculate the correlation

```
library(moderndive)
FQ1%>%
  get_correlation(formula = Y~X)
```

```
# A tibble: 1 x 1
    cor
  <dbl>
1 0.705
```

The correlation coefficient is around 0.71,so we could say there is a strong positive correlation between X and Y.

2.2 (b)

corr.func() function is written below:

```
corr.func<-function(data){
  n=nrow(data)
  a<-data[,1]
  b<-data[,2]
  a_mean<-sum(a)/n
  b_mean<-sum(b)/n
  result1<-sum((a-a_mean)*(b-b_mean))
  result2<-sum((a-a_mean)^2)
  result3<-sum((b-b_mean)^2)
  result4<-sqrt(result2*result3)
  correlation<-result1/result4
  return(correlation)
}
corr.func(FQ1)
```

```
[1] 0.7053115
```

Then check if it gives the same result with cor() function:

```
cor(FQ1)
```

```
          X         Y
X 1.0000000 0.7053115
Y 0.7053115 1.0000000
```

So corr.func() function has been correctly written.

3 Further Question 2

```
FQ2<-read_csv("FQ2.csv")
comp.data<-function(data){
  for(i in 1:nrow(data)){
    na.col_num<-which(is.na(data[i,])=="TRUE")
    data[i,na.col_num]<-0
    data[i,na.col_num]<-1-rowSums(data)[i]
  }
  return(data)
}
```

To check whether comp.data() function work or not,we would apply FQ2 data into this function

```
FQ2_new<-comp.data(FQ2)
FQ2_new
```

```
# A tibble: 10 x 3
      V1    V2    V3
   <dbl> <dbl> <dbl>
 1  0.5   0.2   0.3
 2  0.7   0.2   0.1
 3  0.25  0.35  0.4
 4  0.5   0.2   0.3
 5  0.63  0.17  0.2
 6  0.13  0.33  0.54
 7  0.7   0.1   0.2
 8  0.45  0.2   0.35
 9  0.38  0.22  0.4
10  0.4   0.48  0.12
```

So all of the missing values(NA's) has been replaced by its relative proportion