SGupta_HW01

ST503 Assignment 1

Running Code

(A) Write a brief description of the dataset. Produce some numerical and graphical summaries of the dataset.

```
#install.packages('faraway')
```

```
    library(faraway)
    data(teengamb)
    str(teengamb)
```

```
'data.frame':   47 obs. of  5 variables:
 $ sex   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ status: int  51 28 37 28 65 61 28 27 43 18 ...
 $ income: num  2 2.5 2 7 2 3.47 5.5 6.42 2 6 ...
 $ verbal: int  8 8 6 4 8 6 7 5 6 7 ...
 $ gamble: num  0 0 0 7.3 19.6 0.1 1.45 6.6 1.7 0.1 ...
```

```
    print("Print Summary")
```

```
[1] "Print Summary"
```

```
    summary(teengamb)
```

```
      sex              status          income           verbal
 Min.   :0.0000   Min.   :18.00   Min.   : 0.600   Min.   : 1.00
 1st Qu.:0.0000   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00
 Median :0.0000   Median :43.00   Median : 3.250   Median : 7.00
 Mean   :0.4043   Mean   :45.23   Mean   : 4.642   Mean   : 6.66
 3rd Qu.:1.0000   3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00
 Max.   :1.0000   Max.   :75.00   Max.   :15.000   Max.   :10.00
     gamble
 Min.   :  0.0
 1st Qu.:  1.1
 Median :  6.0
 Mean   : 19.3
 3rd Qu.: 19.4
 Max.   :156.0
```

```
    print("Print top 5 values")
```

```
[1] "Print top 5 values"
```

```
    head(teengamb)
```

```
  sex status income verbal gamble
1   1     51   2.00      8    0.0
2   1     28   2.50      8    0.0
3   1     37   2.00      6    0.0
4   1     28   7.00      4    7.3
5   1     65   2.00      8   19.6
6   1     61   3.47      6    0.1
```

The dataset contains 47 observations of 5 variables:

sex: Binary variable (0 = Male, 1 = Female)

status: ranges from 18 to 75, with a mean of 45.23.

income: ranges from 0.6 to 15, with a mean of 4.64.

verbal: ranges from 1 to 10, with a mean of 6.66.

gamble: ranges from 0 to 156, with a mean of 19.3.

(B) Fit a linear model using the lm() function with gample variable as response, and the income variable as predictors, and report the regression coefficients

```
model <- lm(gamble ~ income, data = teengamb)
summary(model)$coefficients
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -6.324559   6.029874 -1.048871 2.998383e-01
income       5.520485   1.035772  5.329824 3.045433e-06
```

(C) Write the mathematical form of the model you fit in part (B). Clearly define each component in your model.

The fitted model: y hat = beta 0 + beta 1 * income

Beta 0 is the intercept

Beta 1 is the slope

The regression coefficients are :

The intercept is -6.324559

And the slope coefficient is 5.520485

The required regression equation is y hat = -6.324559 + 5.520485 * income

##(D) Further numerical investigation: compute the mean and standard deviation of gamble and income for males (sex=0) and females (sex = 1) separately. Comment on the results.

```
male_dataset <- teengamb[teengamb$sex == 0, ]
female_dataset <- teengamb[teengamb$sex == 1, ]

mean_sd <- data.frame(
  combined_dataset = c("Males", "Females"),
  mean_gamble = c(mean(male_dataset$gamble), mean(female_dataset$gamble)),
  sd_gamble = c(sd(male_dataset$gamble), sd(female_dataset$gamble)),
  mean_income = c(mean(male_dataset$income), mean(female_dataset$income)),
  sd_income = c(sd(male_dataset$income), sd(female_dataset$income))
)
print(mean_sd)
```

```
  combined_dataset mean_gamble sd_gamble mean_income sd_income
1            Males   29.775000  37.32418    4.976071  4.086625
2          Females    3.865789   5.15073    4.149474  2.598240
```

Gambling Behavior:

Males have a significantly higher average gambling expense (29.78) compared to females (3.87)

Males have a significantly higher Gambling variability (SD) (37.32) than for females (5.15).

Income:

Males have a slightly higher average income (4.98) compared to females (4.15).

Income variability is also greater for males (4.09) than for females (2.60).

Males show both higher and more variable gambling expense and income compared to females.

(E) Fit the same linear regression as in part (B), but separately for male and females. Report the regression coefficients.

```r
male_dataset <- teengamb[teengamb$sex == 0, ]
female_dataset <- teengamb[teengamb$sex == 1, ]

male_model <- lm(gamble ~ income, data = male_dataset)
female_model <- lm(gamble ~ income, data = female_dataset)

summary(male_model)
```

```
Call:
lm(formula = gamble ~ income, data = male_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-56.522 -16.402  -2.342   7.901  93.478

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.660      8.021  -0.332    0.743
income         6.518      1.255   5.195 2.01e-05 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.64 on 26 degrees of freedom
Multiple R-squared:  0.5093,    Adjusted R-squared:  0.4905
F-statistic: 26.99 on 1 and 26 DF,  p-value: 2.01e-05
```

```
    summary(female_model)
```

```
Call:
lm(formula = gamble ~ income, data = female_dataset)

Residuals:
   Min     1Q Median     3Q    Max
-4.702 -3.527 -1.790  1.883 16.110

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.1400     2.3273   1.349    0.195
income        0.1749     0.4789   0.365    0.719

Residual standard error: 5.279 on 17 degrees of freedom
Multiple R-squared:  0.007786,  Adjusted R-squared:  -0.05058
F-statistic: 0.1334 on 1 and 17 DF,  p-value: 0.7194
```

The regression coefficients for males

The intercept is -2.660

And the slope coefficient is 6.518

The required regression equation for male is y hat = -2.660 + 6.518 * income

The regression coefficients for females

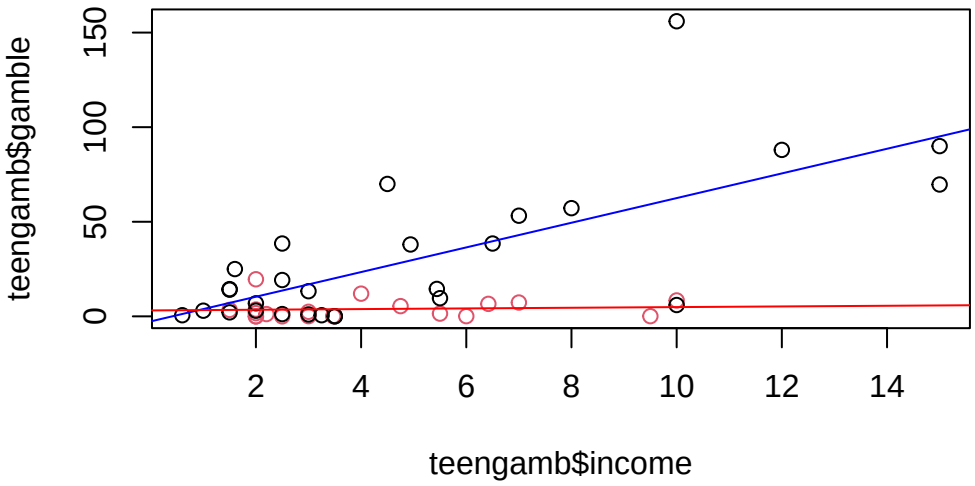
The intercept is

And the slope coefficient is .

The required regression equation for female is, y hat = 3.1400 + 0.1749 * income

(F) Create a scatterplot between gamble (in y axis) and income (x axis), and color the points by sex. Then add two fitted regression lines from part (E) to the plot. Comment on the results.

```
plot(teengamb$income, teengamb$gamble, col = teengamb$sex + 1)
abline(male_model, col = "blue")
abline(female_model, col = "red")
```



If there is an upward trend rising from left-hand corner to upper right-hand corner, the correlation also looks like positive. If there is a downward trend, the correlation is negative.

The males data show a strong positive correlation in the gambling expense as the income also increases.

While females are little bit consistent and show less gambling behavior, irrespective of income.