Problem 1

(A) $Y = X\beta + \varepsilon$

$$Y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ y_{31} \\ \vdots \\ y_{3n_3} \end{pmatrix} \quad X = \begin{bmatrix} 1 & 1 & 0 & 0 & x_{11} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & x_{1n_1} \\ 1 & 0 & 1 & 0 & x_{21} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & x_{2n_2} \\ 1 & 0 & 0 & 1 & x_{31} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 & x_{3n_3} \end{bmatrix} \quad \beta = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta \end{pmatrix} \quad \varepsilon = \begin{pmatrix} e_{11} \\ \vdots \\ e_{1n_1} \\ e_{21} \\ \vdots \\ e_{2n_2} \\ e_{31} \\ \vdots \\ e_{3n_3} \end{pmatrix}$$

(B) No, model matrix X is not full column rank because the first column is a linear combination of columns 2, 3, and 4.

Problem 2

(A) The dataset teengamb has five variables: four quantitative (status, income, verbal, gamble) and one categorical (sex, even though represented by numbers). Without given units, it is difficult to interpret too much from each variable; however, from the summary statistics and graphs, the variables status and verbal appear approximately symmetric. The variables income is slightly skewed right, while gamble is heavily skewed right.

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

Figure 1: Summary Statistics

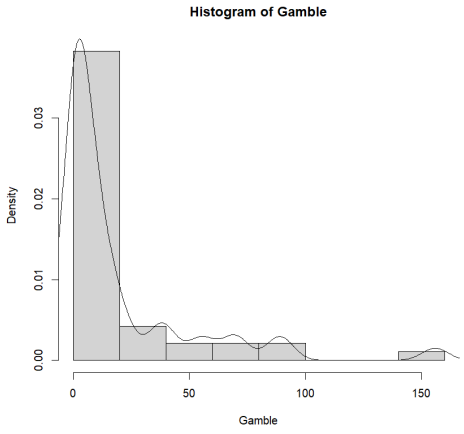


Figure 2: Histogram of Gamble

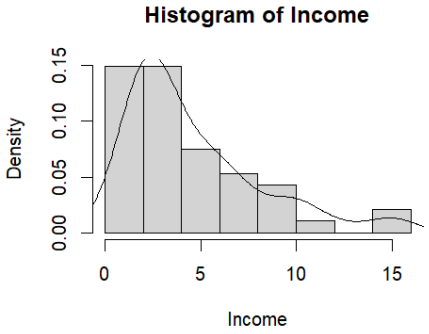


Figure 3: Histogram of Income

(B)

```
#Getting linear fit for gamble vs. income
out <- lm(gamble ~ income, data=teengamb)
out$coefficients
```

```
> out$coefficients
(Intercept)      income
  -6.324559    5.520485
```

$\hat{y} = -6.324559 + 5.520485x$ is the linear regression model found by the lm() function in R.

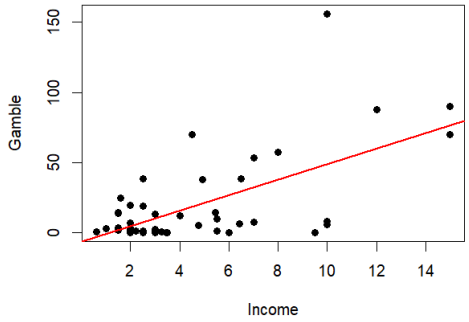


Figure 4: Scatterplot of Gamble vs. Income

(C) $Y = X\beta + \varepsilon$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

(D) Comparing the means and standard deviations of income and gamble for males (X = 0) and females (X = 1).

```
> sapply(split(teengamb$income, teengamb$sex), mean)
       0        1
4.976071 4.149474
> sapply(split(teengamb$income, teengamb$sex), sd)
       0        1
4.086625 2.598240
> sapply(split(teengamb$gamble, teengamb$sex), mean)
       0        1
29.775000  3.865789
> sapply(split(teengamb$gamble, teengamb$sex), sd)
       0        1
37.32418  5.15073
```

Figure 5: Finding Mean and StDev based on Gender

The mean incomes of males and females are relatively close to one another, but the males do boast a slightly higher mean, and larger standard deviation. For the gamble variable, the mean for the males is almost 8 times larger than the mean of the females. The standard deviation for gamble produces similar results being over 7 times larger for males than females. These seems to indicate that males are more likely to spend a lot on gambling than females.

(E)

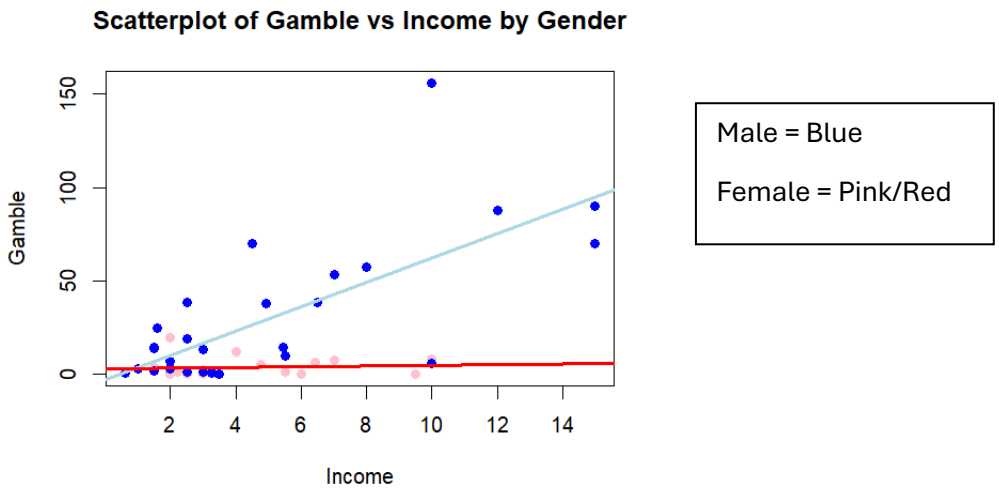
```
> model_male <- lm(gamble ~ income, data = subset(teengamb, sex == 0))
> model_female <- lm(gamble ~ income, data = subset(teengamb, sex == 1))
> model_male$coefficients
(Intercept)      income
  -2.659629    6.518120
> model_female$coefficients
(Intercept)      income
  3.1399737   0.1749176
```

Figure 6: Models for both Male and Female

Male Regression Line: $\hat{y} = -2.659629 + 6.518120x$

Female Regression Line: $\hat{y} = 3.1399737 + 0.1749176x$

(F)



Females are shown to not gamble nearly as much as males. Males tend to gamble more if as their income increases, shown by the positive slope in the linear regression model. Females do not tend to gamble regardless of income status.

Problem 3

(A) $Y = X\beta + \varepsilon$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

(B) $$X^TX = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} 1 & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_2$$

$$\hat{B} = (X^TX)^{-1}X^TY = X^TY = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

Appendix (R Code)

```
#open faraway library
library(faraway)

#summary stats for each variable
summary(teengamb)

#histogram of gamble and income
hist(teengamb$gamble, freq = FALSE,
   main = "Histogram of Gamble", xlab = "Gamble")
lines(density(teengamb$gamble))

hist(teengamb$income, freq = FALSE,
   main = "Histogram of Income", xlab = "Income")
lines(density(teengamb$income))

#scatterplots for all variables except gender which was just 0s and 1s
pairs(teengamb[,c(2,3,4,5)], pch = 19)

#Getting linear fit for gamble vs. income
out <- lm(gamble ~ income, data=teengamb)
out$coefficients

#plotting gamble vs income with the regression line
plot(teengamb$income, teengamb$gamble,
   xlab = "Income", ylab = "Gamble",
   pch = 19)
abline(out, col = "red", lwd = 2)

#looking at summary of gamble and income variables based on gender
sapply(split(teengamb$income, teengamb$sex), mean)
sapply(split(teengamb$income, teengamb$sex), sd)
sapply(split(teengamb$gamble, teengamb$sex), mean)
sapply(split(teengamb$gamble, teengamb$sex), sd)

# Fitting a linear model for each gender separately
model_male <- lm(gamble ~ income, data = subset(teengamb, sex == 0))
model_female <- lm(gamble ~ income, data = subset(teengamb, sex == 1))

model_male$coefficients
model_female$coefficients
```

```
#plotting by color
plot(teengamb$income, teengamb$gamble, col = ifelse(teengamb$sex == 0, "blue", "pink"),
   xlab = "Income", ylab = "Gamble", main = "Scatterplot of Gamble vs Income by Gender", pch = 19)

abline(model_male, col = "lightblue", lwd = 3)
abline(model_female, col = "red", lwd = 3)
```