Homework 1

Madhuri Raman

10/2/2021

1.

```
library(faraway)
library(tidyverse)

df <- data.frame(teengamb)
(dim(df))
```

```
## [1] 47  5
```

There are 47 rows and 5 columns in the `teengamb` data set.

Let's examine the first 6 rows of the data set to understand what each variable looks like.

```
head(df)
```

```
##   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
```

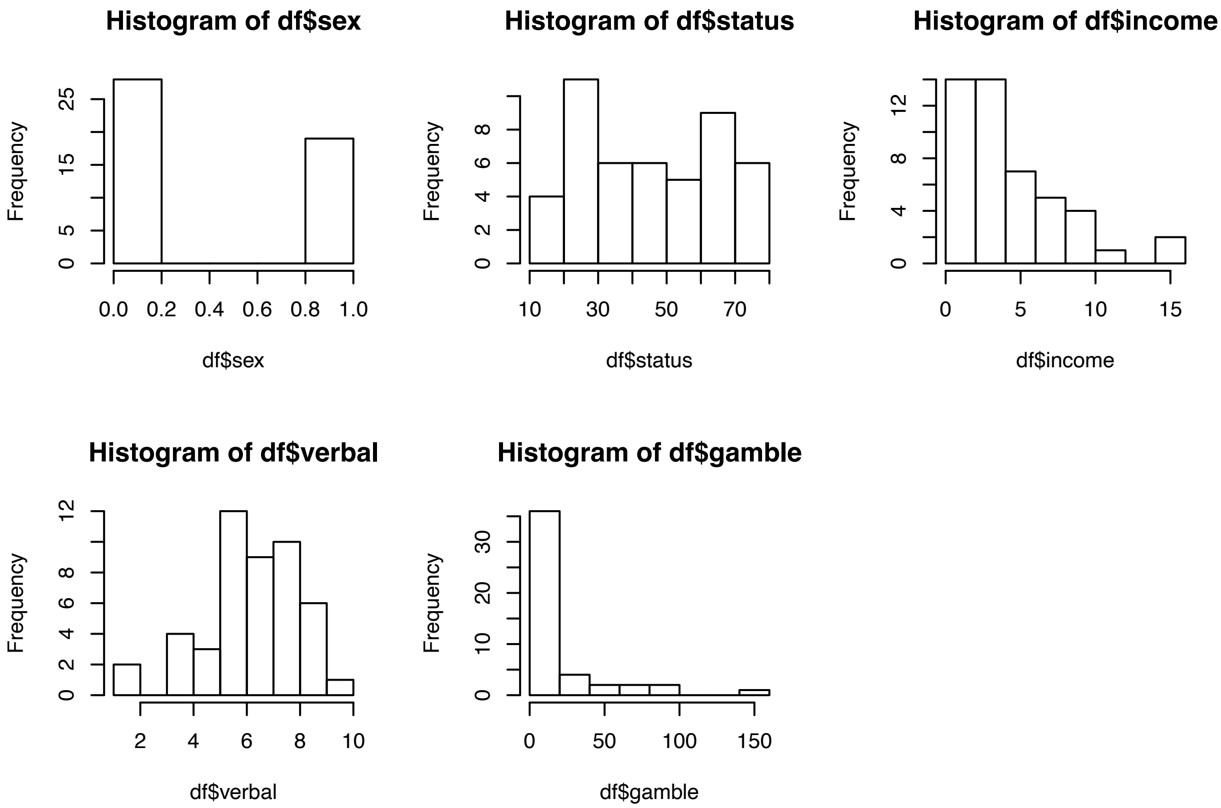
Univariate EDA (Histograms and Summary statistics)

```
par(mfrow = c(2,3))
hist(df$sex)
hist(df$status)
hist(df$income)
hist(df$verbal)
hist(df$gamble)


df %>%
    select(sex, status, income, verbal, gamble) %>%
    map_df(.f = ~ broom::tidy(summary(.x)), .id = "variable")
```

```
## # A tibble: 5 x 7
##   variable minimum    q1 median   mean    q3 maximum
##   <chr>      <dbl> <dbl>  <dbl>  <dbl> <dbl>   <dbl>
## 1 sex            0     0      0  0.404  1          1
```

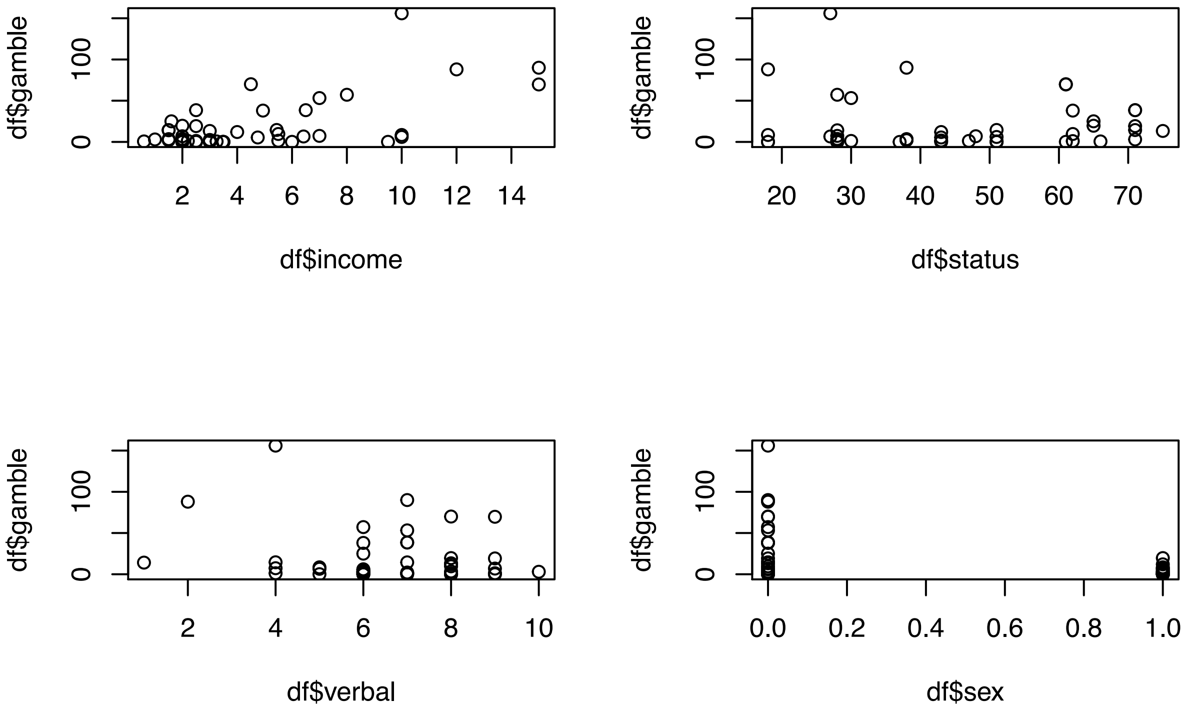
```
## 2 status       18     28     43     45.2    61.5        75
## 3 income        0.6    2      3.25   4.64    6.21       15
## 4 verbal        1      6      7       6.66   8          10
## 5 gamble        0      1.1    6     19.3    19.4       156
```



Above we see the histograms of each variable in the data set. Along with the summary statistics table, these histograms give us a good sense of the distribution of each variable, predictors and responses, in our data. It is interesting to note the distributions of `income` and `gamble` are extremely right-skewed while `verbal` is slightly left-skewed. Specifically, we can identify likely outliers in the right tail of the distribution of `gamble` since the maximum value of 156 is extremely far from the q3 of 19.4 and mean of 19.3 for that variable. Based on the problem statement, it is reasonable to consider `gamble` as the response variable (Y) and `status`, `income`, `verbal`, and `sex` as potential predictors (Xs).

Bivariate EDA (Pairwise scatterplots)

```r
par(mfrow = c(2,2))
plot(x = df$income, y = df$gamble)
plot(x = df$status, y = df$gamble)
plot(x = df$verbal, y = df$gamble)
plot(x = df$sex, y = df$gamble)
```



Based on the pairwise scatterplots of the four potential predictors vs. response `gamble`, we notice a potentially linear relationship between `gamble` and `income` as well as differing distributions of `gamble` for sex=0 males compared to sex=1 females. This suggests that as we proceed with modeling `gamble`, it may be useful to include `income` and `sex` as predictors.

2.

Now we will fit the following regression model:

$$gamble = \beta_0 + \beta_1 * sex + \beta_2 * status + \beta_3 * income + \beta_4 * verbal$$

```
lm.out.2 <- lm(data = df, gamble ~ sex + status + income + verbal)
summary(lm.out.2)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex         -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

b)

```r
sort(lm.out.2$residuals, decreasing = TRUE)[1]
```

```
##       24
## 94.25222
```

The 24th observation has the largest positive residual.

c)

```r
mean(lm.out.2$residuals)
```

```
## [1] -3.065293e-17
```

```r
median(lm.out.2$residuals)
```

```
## [1] -1.451392
```

The mean of the residuals is -3.065293e-17 which is nearly zero. The median of the residuals is -1.451392.

d)

```r
cor(lm.out.2$residuals, lm.out.2$fitted.values)
```

```
## [1] -1.070659e-16
```

The correlation between the residuals and fitted values of this model is -1.070659e-16 which is nearly zero.

e)

```r
cor(lm.out.2$residuals, df$income)
```

```
## [1] -7.242382e-17
```

The correlation between the residuals and `income` is -7.242382e-17 which is nearly zero.

f)

```r
lm.out.2$coefficients[2] # note that females = 1, males = 0
```

```
##       sex
## -22.11833
```

Holding all other predictors constant, the predicted gambling expenditure for males is approximately 22.12 pounds per year higher than for females.

3. $X_1 \ldots X_n$, $Y_1 \ldots Y_n$, $\hat\beta_0, \hat\beta_1, \hat\sigma$ OLS output for this data

$\tilde X_i = c\cdot(X_i + d)$, $c \neq 0$ same

$\hookrightarrow c(X_1+d) \ldots c(X_n+d)$, $Y_1 \ldots Y_n$, $\tilde\beta_0, \tilde\beta_1, \tilde\sigma$ OLS for this.

WTK: $\tilde\beta_0, \tilde\beta_1, \tilde\sigma$ in terms of $\hat\beta_0, \hat\beta_1, \hat\sigma, c, d$.

Well we know that the OLS solutions are:

$$\hat\beta_0 = \bar Y - \hat\beta_1 \bar X \quad \text{where} \quad \bar Y = \frac{1}{n}\sum_{i=1}^n Y_i,$$

$$\hat\beta_1 = \sum_{i=1}^n \frac{(X_i - \bar X)(Y_i - \bar Y)}{(X_i - \bar X)^2} \qquad \bar X = \frac{1}{n}\sum_{i=1}^n X_i$$

$$\hat\sigma^2 = \frac{1}{n-2}\sum_{i=1}^n \hat\epsilon_i^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat\beta_0 - \hat\beta_1 X_i)^2$$

So: $\bar X_{new} = \frac{1}{n}\sum_{i=1}^n c\cdot(X_i + d) = \frac{1}{n} c \sum_{i=1}^n (X_i + d) = \frac{1}{n}\cdot c\sum_{i=1}^n X_i + \frac{c\cdot d\cdot n}{n}$

$\bar Y_{new} = \frac{1}{n}\sum_{i=1}^n Y_i = \bar Y$ (same)

$$= c\cdot \frac{1}{n}\sum_{i=1}^n X_i + c\cdot d$$
$$= c\cdot \bar X + c\cdot d$$
$$= c\cdot(\bar X + d)$$

- $$\begin{aligned}\tilde\beta_0 &= \bar Y - \tilde\beta_1 \bar X_{new} \\ &= \bar Y - \tilde\beta_1 \cdot c\cdot(\bar X + d) \\ &= \bar Y - \frac{1}{\cancel{c}}\cdot \hat\beta_1 \cdot \cancel{c}\cdot(\bar X + d) \\ \tilde\beta_0 &= \bar Y - \hat\beta_1(\bar X + d) \\ &= \bar Y - \hat\beta_1 \bar X - \hat\beta_1 d\end{aligned}$$

$$\Rightarrow \boxed{\tilde\beta_0 = \hat\beta_0 - d\cdot\hat\beta_1}$$

- $$\begin{aligned}\tilde\beta_1 &= \sum_{i=1}^n \frac{(\tilde X_i - \bar X_{new})\cdot(Y_i - \bar Y)}{(\tilde X_i - \bar X_{new})^2} \\ &= \sum_{i=1}^n \frac{\big(c(X_i+d) - c(\bar X + d)\big)\cdot(Y_i - \bar Y)}{\big(c(X_i+d) - c(\bar X+d)\big)^2}\end{aligned}$$

$$= \sum_{i=1}^{n} \frac{c\,(x_i + \cancel{d} - \bar{x} - \cancel{d}) \cdot (y_i - \bar{y})}{\left(c\,(x_i + \cancel{d} - \bar{x} - \cancel{d}) \right)^2}$$

$$= \sum_{i=1}^{n} \frac{c\,(x_i - \bar{x})\,(y_i - \bar{y})}{c^2\,(x_i - \bar{x})^2}$$

$$= \frac{1}{c} \sum_{i=1}^{n} \frac{(x_i - \bar{x})\,(y_i - \bar{y})}{(x_i - \bar{x})^2}$$

$$\Rightarrow \quad \boxed{\tilde{\beta}_1 = \frac{1}{c} \cdot \hat{\beta}_1}$$

- $$\tilde{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \tilde{\epsilon}_i^2$$

$$= \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 \tilde{x}_i \right)^2$$

$$= \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 + d \cdot \hat{\beta}_1 - \frac{1}{c} \cdot \hat{\beta}_1 \cdot \tilde{x}_i \right)^2$$

$$= \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 + \hat{\beta}_1 \left(d - \frac{1}{c} \tilde{x}_i \right) \right)^2$$

$$= \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 + \hat{\beta}_1 \left(d - \cancel{\frac{1}{c}} \cdot \cancel{c}\,(x_i + d) \right) \right)^2$$

$$= \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 + \hat{\beta}_1 \left(\cancel{d} - x_i - \cancel{d} \right) \right)^2$$

$$= \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i \right)^2 = \hat{\sigma}^2$$

$$\Rightarrow \quad \boxed{\tilde{\sigma} = \hat{\sigma}}$$

4. $X \in \{0,1\} \rightarrow X = \begin{cases} 0, & \text{placebo} \\ 1, & \text{treatment} \end{cases}$ $\quad \overline{Y}_P \quad E[Y|X=0]$

$Y \in \mathbb{R}$ blood pressure $\quad \overline{Y}_T \quad E[Y|X=1]$

What are OLS coeffs $\hat{\beta}_0, \hat{\beta}_1$ in terms of $\overline{Y}_P, \overline{Y}_T$?

we know $\hat{\beta}_0 = \overline{Y} + \hat{\beta}_1 \overline{X}$

Model: $E[Y|X=x] = \hat{\beta}_0 + \hat{\beta}_1 x$

so $E[Y|X=0] = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 = \hat{\beta}_0 = \overline{Y}_P$

$E[Y|X=1] = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 = \hat{\beta}_0 + \hat{\beta}_1 = \overline{Y}_T$

so
$$\boxed{\begin{aligned} \hat{\beta}_0 &= \overline{Y}_P \\ \hat{\beta}_1 &= \overline{Y}_T - \overline{Y}_P \end{aligned}}$$

5.

a)

```r
set.seed(819)

totalcover1 <- 0
totalcover2 <- 0

NN <- 1000
xi <- rnorm(n = 100, mean = 0, sd = 1) # simulated dataset Xs
beta0 <- 1
beta1 <- 1
sigmasquared <- 1
ei <- rnorm(n = 100, mean = 0, sd = sqrt(sigmasquared))
Yi <- beta0 + beta1*xi + ei # simulated dataset Ys
df5a <- data.frame(X = c(xi), Y = c(Yi))

lm.out.5a <- lm(Y ~ X, data = df5a)

Xnew1 <- -0.5
Xnew2 <- 2
newXs <- data.frame(X = c(Xnew1, Xnew2))

for (ii in 1:NN){
  # 90% prediction intervals at x = -0.5 and x = 2
  p <- predict(lm.out.5a, newdata = newXs, interval = "predict", level = 0.90)
  Ynew1 <- beta0 + beta1*Xnew1 + rnorm(n = 1, mean = 0, sd = sqrt(sigmasquared))
  Ynew2 <- beta0 + beta1*Xnew2 + rnorm(n = 1, mean = 0, sd = sqrt(sigmasquared))
  cover1 <- ifelse(Ynew1 >= p[1,2] & Ynew1 <= p[1,3], 1, 0)
  cover2 <- ifelse(Ynew2 >= p[2,2] & Ynew2 <= p[2,3], 1, 0)
  totalcover1 <- totalcover1 + cover1
  totalcover2 <- totalcover2 + cover2
}

(coveragerate1 <- totalcover1 / NN)
```

```
## [1] 0.915
```

```r
(coveragerate2 <- totalcover2 / NN)
```

```
## [1] 0.928
```

We observe a coverage rate of about 0.915 for Xnew1 = -0.5 and 0.928 for Xnew2 = 2.

b)

```r
set.seed(819)

totalcover1 <- 0
totalcover2 <- 0

NN <- 1000
```

```r
xi <- rnorm(n = 100, mean = 0, sd = 1) # simulated dataset Xs
beta0 <- 1
beta1 <- 1
sigmasquared <- 1
ei <- rnorm(n = 100, mean = 0, sd = sqrt(sigmasquared))
Yi <- beta0 + beta1*xi + exp(xi) + ei # simulated dataset Ys
df5a <- data.frame(X = c(xi), Y = c(Yi))

lm.out.5a <- lm(Y ~ X, data = df5a)

Xnew1 <- -0.5
Xnew2 <- 2
newXs <- data.frame(X = c(Xnew1, Xnew2))

for (ii in 1:NN){
  # 90% prediction intervals at x = -0.5 and x = 2
  p <- predict(lm.out.5a, newdata = newXs, interval = "predict", level = 0.90)
  Ynew1 <- beta0 + beta1*Xnew1 + exp(Xnew1) + rnorm(n = 1,
                                                     mean = 0,
                                                     sd = sqrt(sigmasquared))
  Ynew2 <- beta0 + beta1*Xnew2 + exp(Xnew2) + rnorm(n = 1,
                                                     mean = 0,
                                                     sd = sqrt(sigmasquared))
  cover1 <- ifelse(Ynew1 >= p[1,2] & Ynew1 <= p[1,3], 1, 0)
  cover2 <- ifelse(Ynew2 >= p[2,2] & Ynew2 <= p[2,3], 1, 0)
  totalcover1 <- totalcover1 + cover1
  totalcover2 <- totalcover2 + cover2
}

(coveragerate11 <- totalcover1 / NN)
```

```
## [1] 0.949
```

```r
(coveragerate22 <- totalcover2 / NN)
```

```
## [1] 0.212
```

Now we observe a coverage rate of 0.949 for Xnew1 = -0.5 and 0.212 for Xnew2 = 2.

c)

In part a we see that the coverage rates for both new x values' prediction intervals were close to 0.90. This is what we expect to see, especially if we run the simulation for more than 1000 iterations because we generate 90% confidence intervals. Specifically, all model assumptions of normality are satisfied, so basing our confidence intervals off of the standard normal distribution makes sense.

However, in part b, we see two very difference coverage rates. Note that in this model we now add an extra $e^x i$ term. This violates the normality assumptions of the model because now $E[Y|Xi]$ is no longer equal to $\beta_0 + \beta_1 * X$ as our OLS regression model assumes. The only reason that the coverage rate for Xnew1 = -0.5 with this model still seems okay (approx. 0.90) is because $e^{-0.5} = 0.6065$ is nearly zero so it does not affect Ynew1 that much. On the other hand, for Xnew2, $e^2 = 7.34$ which means that Ynew2 is 7.34 higher than what it should be under proper model assumptions. This is why the coverage rate for Xnew2 = 2 in part b is very bad compared to part a.