Simple Linear Regression
Econometrics
Sai Sandeep Chandaluri

Question 1

a. $\sum_{i=1}^{N}(Y_i - \bar{Y}) = 0$

We know that the mean $\bar{Y}$ for a population of size N:

$$\bar{Y} = \frac{Y_1 + Y_2 + ... + Y_N}{N}$$

Substituting the value of $\bar{Y}$, we get:

$$\sum_{i=1}^{N}(Y_i - \bar{Y}) = \sum_{i=1}^{N} Y_i - N \times \bar{Y} = \sum_{i=1}^{N} Y_i - N \times \frac{Y_1 + Y_2 + ... + Y_N}{N} = 0$$

b. $\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{N}(X_i - \bar{X})Y_i$

Expanding the summation $\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$, we get:

$$\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{N}(X_iY_i - X_i\bar{Y} - \bar{X}Y_i + \bar{X}\bar{Y}) = \sum_{i=1}^{N} X_iY_i - \bar{X}\bar{Y} - \bar{X}\bar{Y} + \bar{X}\bar{Y}$$

$$= \sum_{i=1}^{N} X_iY_i - \bar{X}\bar{Y} = \sum_{i=1}^{N}(X_i - \bar{X})Y_i$$

(Note that we have used the equation $\frac{\sum_{i=1}^{N} Y_i}{N} = \bar{Y}$ in the above equations)

Question 2

a. Expectation of a random variable:

Expectation of a random variable, also known as mean, is defined both for continous and discrete random variables as:

If X is a continous random variable with probability density function $f_x(x)$, then expected value of X is

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\ dx$$

If X is a discrete random variable with probability density function $f_x(x)$, then expected value of X is

$$E(X) = \sum_{x} x f_X(x)$$

b. Sample Average:

Sample Average or Sample Mean is the average of the a set of sample data. If n is the size of a sample, then sample average $\bar{x}$ is

$$\bar{x} = \frac{\sum x_i}{n}$$

The sample average is a random variable and its value depends on the random sample that is considered where as expected value is the mean computed over the whole population.

Question 3

a. $E[3X]$

Given $X \sim N(1,2) \implies E[X] = 1$, hence

$$E[3X] = 3E[X] = 3$$

b. $var(3X)$

Given $X \sim N(1,2) \implies var(X) = 2$, hence

$$var(3X) = 3^2 var(X) = 18$$

c. $var(2X - 2Y)$ and $var(2X + 2Y)$

Given $X \sim N(1,2) \implies var(X) = 2$ and $Y \sim N(2,3) \implies var(Y) = 3$, hence

$$var(2X - 2Y) = 2^2 var(X - Y) = 4(var(X) + var(Y)) = 20$$

Similarly,

$$var(2X + 2Y) = 2^2 var(X + Y) = 4(var(X) + var(Y)) = 20$$

d. We get same answer in part(c), whether added or subtracted, because

$$var(aX + bY) = a^2 var(X) + b^2 var(Y) + 2abcov(X,Y)$$

If coefficient of Y is negative ($b < 0$),

$$var(aX + bY) = a^2 var(X) + b^2 var(Y) - 2a|b|cov(X,Y)$$

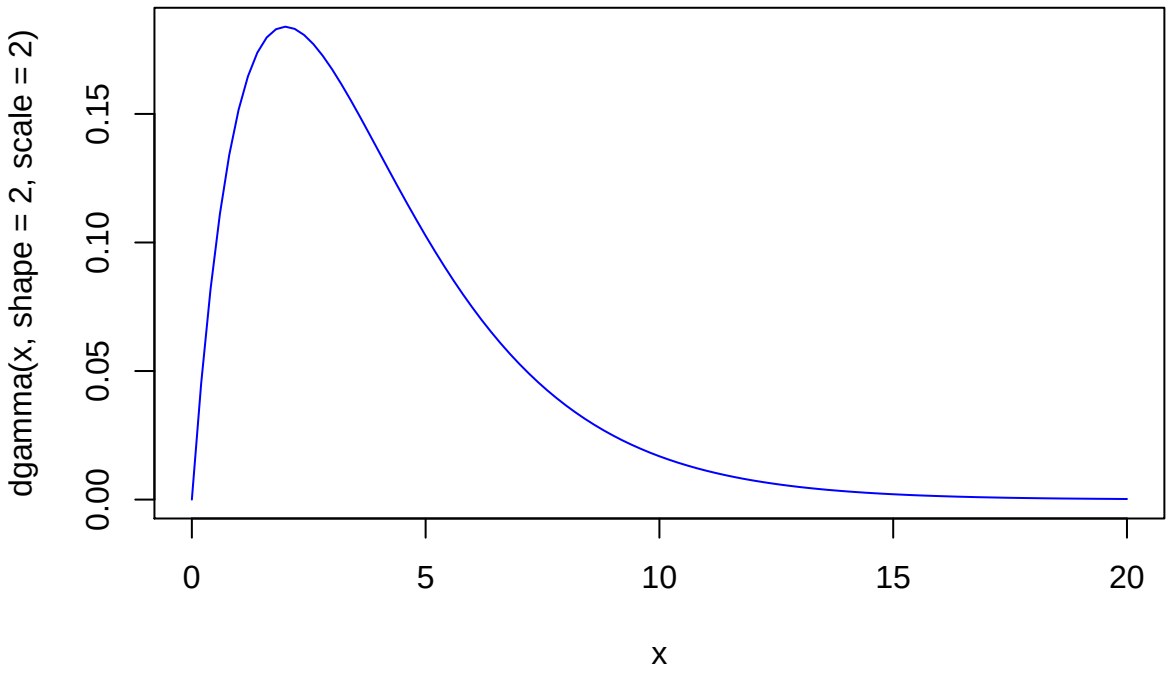
Here, it is given that X and Y are independent, which means $cov(X,Y) = 0$. This relation is important because the above equations give the same value of $var(aX + bY)$, no matter what the sign of coefficient of Y is. If the covariance of X and Y is non-zero, then the coefficient of Y is important in determining the value of $var(aX + bY)$

Question 4

a. Central Limit Theorem states that when multiple independent random variables are added, their sum tend towards a normal distribution, as `n`, the size of the sample increases, irrespective of the parent distribution.

b. Let $X \sim \text{Gamma}(\alpha = 2,\ \beta = 2)$:

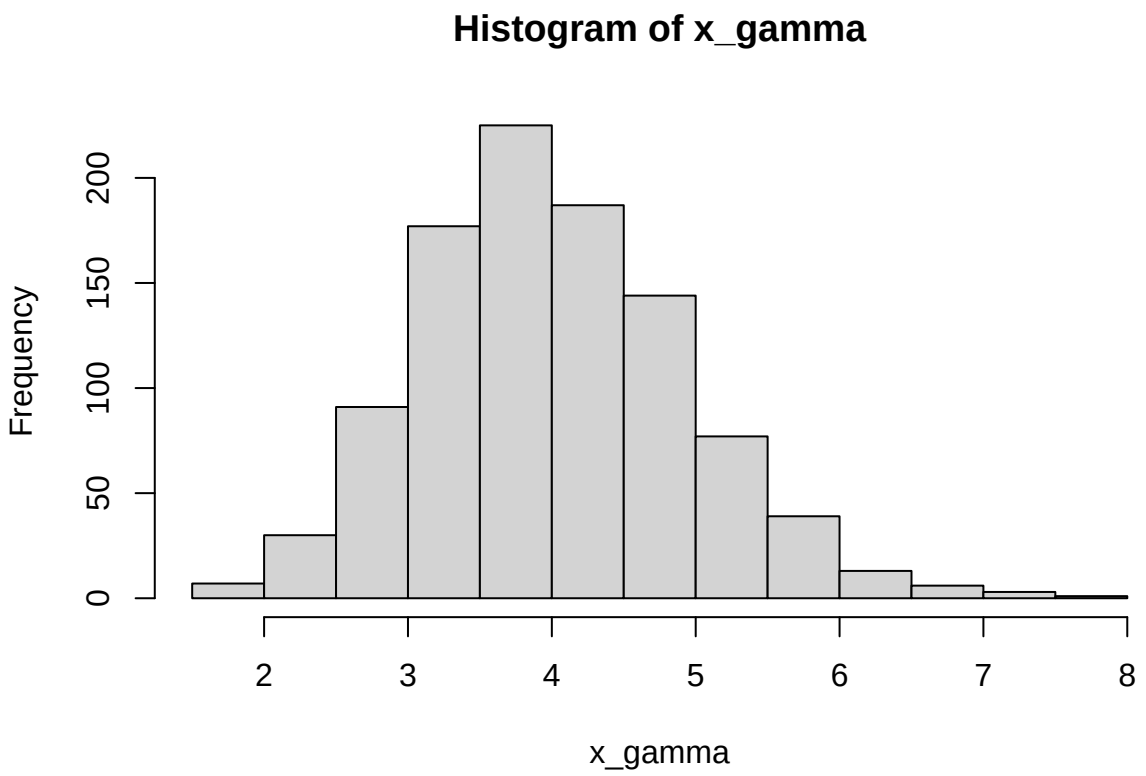
```
curve(dgamma(x, shape=2, scale=2), from = 0, to = 20, col='blue')
```



c. Given n is number of draws from distribution and r is the number of times the process is repeated. The following histogram is obtained when the sample average is calculated for each sample of size 10 is drawn 1000 times.

```
count <- (1:1000)
x_gamma <- c()

for(i in count){
x_gamma[i] <- mean(rgamma(10, shape=2, scale=2))
}
hist(x_gamma)
```

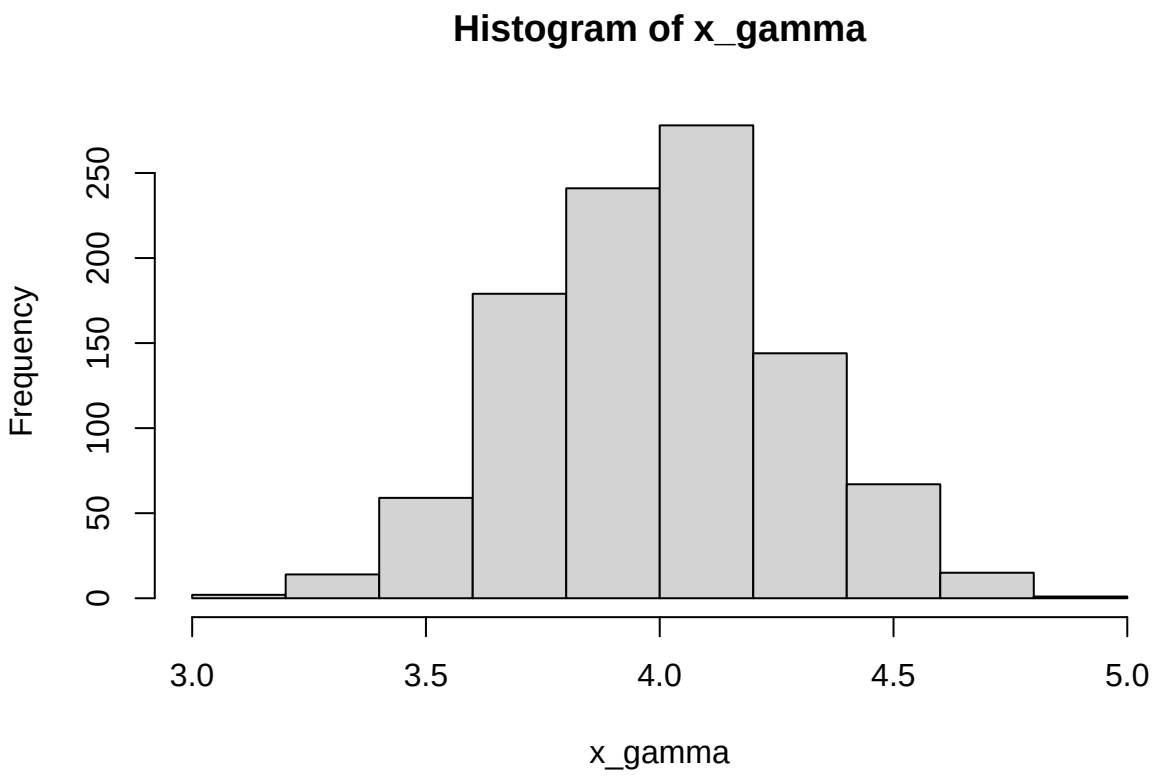


From this histogram, we can observe that the sample mean of Gamma distribution is a random variable that tends to be like Normally distributed. This also describe the Central Limit theorem.

d. Repeating the above with n=100:

```
count <- (1:1000)
x_gamma <- c()

for(i in count){
x_gamma[i] <- mean(rgamma(100, shape=2, scale=2))
}
hist(x_gamma)
```



We can see that as n is increased, the degrees of freedom increases, the skewness in (c) is further reduced and the histogram looks more normal. Hence, as n, the size of the sample increases, the distribution of sample mean becomes more like normally distributed. This also explains the Central Limit Theorem.

e. Given that the dataset is of size 2000 and has 2 variables, height and weight. In this real world example, $n = 2000$ as 2000 is the size of the sample that we are considering and $r = 1$ as we have taken a draw of one sample.

Question 5

A random normal distribution with $N(0, 1)$ and a random t-distribution are taken and are plotted as below:

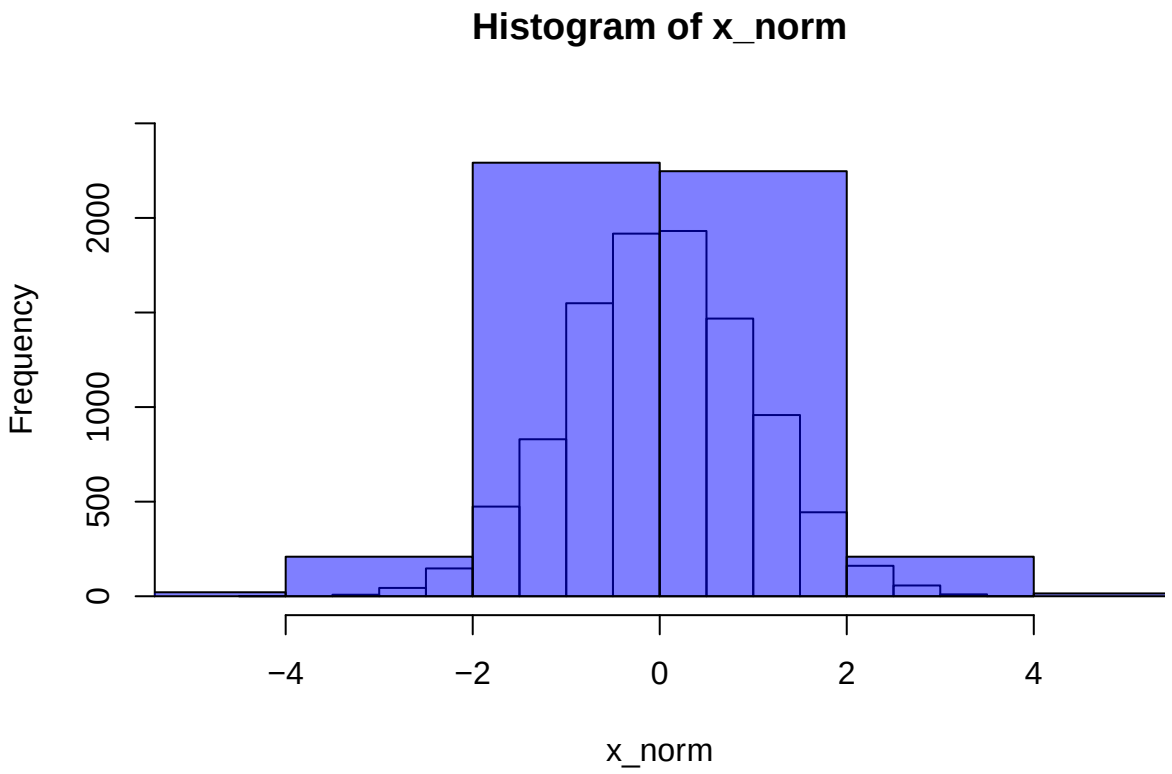
```
library(scales)


x_norm <- rnorm(10000, 0, 1)
x_t <- rt(5000, 5)

hg1 <- hist(x_norm, plot = FALSE)
hg2 <- hist(x_t, plot = FALSE)



plot(hg1, ylim=c(0,2500), xlim = c(-5,5))
plot(hg2, add = T, col=scales::alpha("blue",0.5))
```



We can observe that the normal distribution have much thinner tails than t-distribution.

Question 6

a. The Standard Error of the mean for `VFIAX` index fund return is $0.004 = 0.4\%$

b. The standard error of the mean return of `VFIAX` fund is 0.004, which is almost equal to the mean return of `VFIAX` (0.004). This is definitely a problem for financial analyst to assess the performance of this fund as the confidence interval is very large, much spread and it is tough to predict the return. If we compute the 95% confidence interval, it is approximately $0.004 \pm (2 * 0.004)$, which pretty much overlaps with other funds and it is tough to draw inference about performance of this fund.

c. We know that:

$$StandardError \propto 1/\sqrt{N}$$

Hence, to reduce the standard error of the mean to 1/10th, we need 100 times more data than now.

Size of current sample = 349 - 198 = 151

Hence, size of required sample to reduce Standard Error by 1/10th = 15100.

Question 7

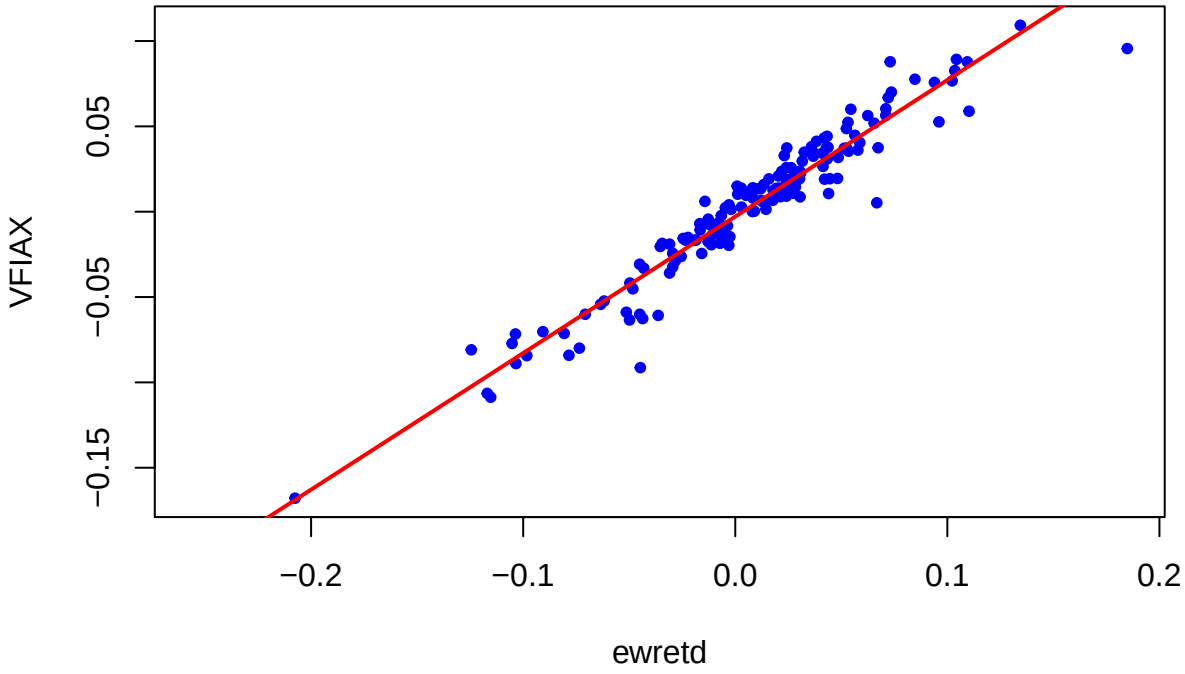
a. Plotting the `VFIAX` index fund return against `ewretd`:

```
library(DataAnalytics)
library(reshape2)
data(Vanguard)

Van=Vanguard[,c(1,2,5)]
```

```
V_reshaped=dcast(Van,date~ticker,value.var="mret")
data(marketRf, package = "DataAnalytics")
Van_mkt = merge(V_reshaped, marketRf, by="date")

plot(VFIAX ~ ewretd, data=Van_mkt, pch=20, col="blue")
out = lm(VFIAX ~ ewretd, data=Van_mkt)
abline(coef(out), col="red", lwd=2)
```



b. Regression output:

```
lmSumm(out)
```

```
## Multiple Regression Analysis:
##     2 regressors(including intercept) and 151 observations
##
## lm(formula = VFIAX ~ ewretd, data = Van_mkt)
##
## Coefficients:
##              Estimate Std Error t value p value
## (Intercept) -0.002855  0.001014   -2.82   0.006
## ewretd       0.799900  0.018520   43.19   0.000
## ---
## Standard Error of the Regression:  0.01231
## Multiple R-squared:  0.926  Adjusted R-squared:  0.926
## Overall F stat: 1865.32 on 1 and 149 DF, pvalue= 0
```