STAC67 Assignment 1

Jefferson Li, Arib Shaikh

January 20 2021

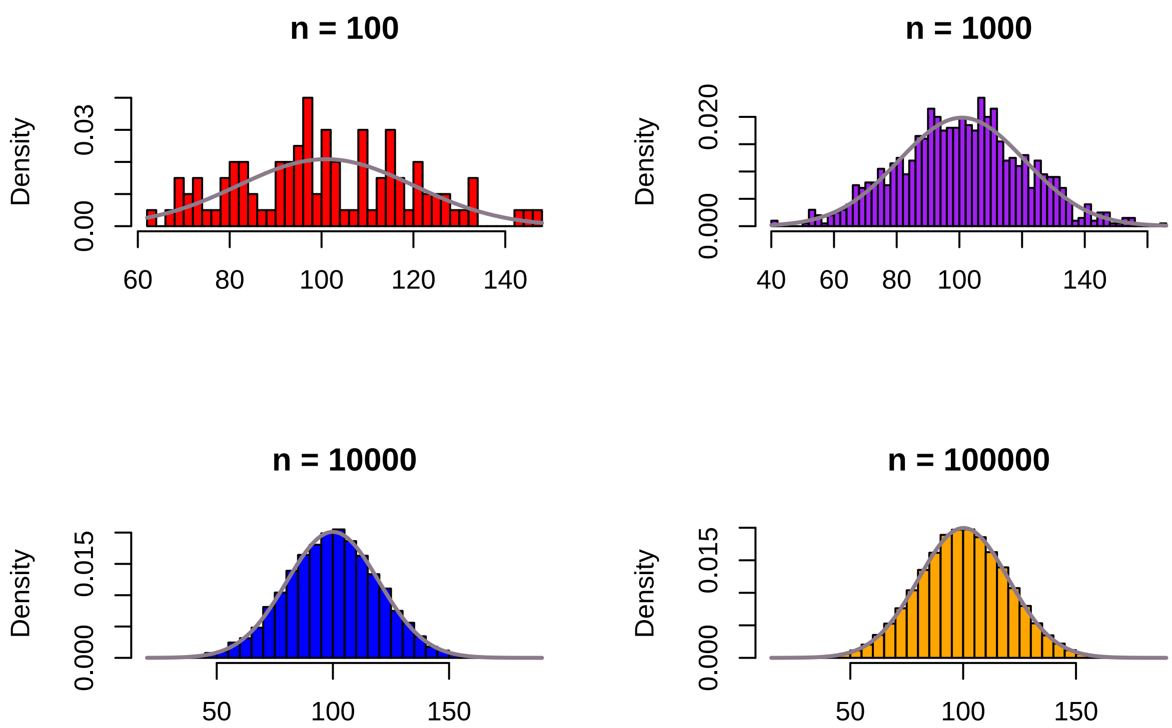
Set Seed

```
set.seed(1005057368)
```

Q. 1

(a) First we generate our random numbers

```
mu = 100
sigma = 20

n3 = rnorm(100, mean = mu, sd = sigma)
n4 = rnorm(1000, mean = mu, sd = sigma)
n5 = rnorm(10000, mean = mu, sd = sigma)
n6 = rnorm(100000, mean = mu, sd = sigma)
```



The distribution of the generated data approach the normal distribution as n increases. In otherwords, the approximate accuracy increases as n increases.

(b) Creating table to compare sample with theoretical

```
summaryTable = matrix(c(
mu, sigma, qnorm(0.025,mu,sigma),qnorm(0.25,mu,sigma),
qnorm(0.5,mu,sigma),qnorm(0.75,mu,sigma),qnorm(0.975,mu,sigma),# theoretical

mean(n3), sd(n3),quantile(n3, probs = c(0.025)),
quantile(n3, probs = c(0.25)),quantile(n3, probs = c(0.50)),
quantile(n3, probs = c(0.75)),quantile(n3, probs = c(0.975)), # n = 100

mean(n4), sd(n4), quantile(n4, probs = c(0.025)), quantile(n4, probs = c(0.25)),
quantile(n4, probs = c(0.50)),
quantile(n4, probs = c(0.75)),quantile(n4, probs = c(0.975)), # n = 1000

mean(n5), sd(n5), quantile(n5, probs = c(0.025)), quantile(n5, probs = c(0.25)),
quantile(n5, probs = c(0.50)),
quantile(n5, probs = c(0.75)),quantile(n5, probs = c(0.975)), # n = 10000

mean(n6), sd(n6), quantile(n6, probs = c(0.025)), quantile(n6, probs = c(0.25)),
quantile(n6, probs = c(0.50)),
quantile(n6, probs = c(0.75)),quantile(n6, probs = c(0.975)) # n = 100000
),ncol=7,byrow=TRUE)
colnames(summaryTable) = c("mean","standard deviation",
```

```r
                            "2.5th percentile",
                            "25th percentile","50th percentile","75th percentile",
                            "97.5th percentile")
rownames(summaryTable) = c("theoretical", "n = 100", "n = 1000", "n = 10000", "n = 100000")

summaryTable
```

```
##                   mean standard deviation 2.5th percentile 25th percentile
## theoretical 100.00000           20.00000         60.80072        86.51020
## n = 100     100.95598           19.12323         68.65721        86.90593
## n = 1000    100.82990           20.09660         62.55115        87.56966
## n = 10000    99.99172           19.82584         60.53259        86.55604
## n = 100000  100.16000           19.97604         61.05341        86.74184
##             50th percentile 75th percentile 97.5th percentile
## theoretical        100.0000        113.4898          139.1993
## n = 100            100.0032        114.4871          137.9545
## n = 1000           100.9495        113.9240          141.1249
## n = 10000          100.1482        113.2960          138.5916
## n = 100000         100.1404        113.6764          139.3164
```

as we can see, every single column gets closer and closer to the theoretical as n increases. This means as we sample more data, the distribution becomes closer and closer to the theoretical normal distribution.

Q. 2

(a)

(i)

Proof. Show that $S_{XX} = \Sigma X_i^2 - n\bar{X}^2$

$$\begin{aligned}
S_{XX} &= \Sigma(X_i - \bar{X})^2 && (1)\\
&= \Sigma(X_i^2 - 2X_i\bar{X} + \bar{X}^2) && (2)\\
&= \Sigma X_i^2 - \Sigma 2X_i\bar{X} + \Sigma\bar{X}^2 && (3)\\
&= \Sigma X_i^2 - 2\bar{X}\Sigma X_i + n\bar{X}^2 && (4)\\
&= \Sigma X_i^2 - 2\bar{X}n\bar{X} + n\bar{X}^2 && (5)\\
&= \Sigma X_i^2 - n\bar{X}^2 && (6)
\end{aligned}$$

□

(ii)

Proof. Show that $S_{XY} = \Sigma X_iY_i - n\bar{X}\bar{Y}$

$$\begin{aligned}
S_{XY} &= \Sigma(X_i - \bar{X})(Y_i - \bar{Y}) && (7)\\
&= \Sigma(X_iY_i - X_i\bar{Y} - \bar{X}Y_i + \bar{X}\bar{Y}) && (8)\\
&= \Sigma X_iY_i - \Sigma X_i\bar{Y} - \Sigma\bar{X}Y_i + \Sigma\bar{X}\bar{Y} && (9)\\
&= \Sigma X_iY_i - \bar{Y}n\bar{X} - n\bar{Y}\bar{X} + n\bar{X}\bar{Y} && (10)\\
&= \Sigma X_iY_i - n\bar{X}\bar{Y} && (11)
\end{aligned}$$

□

(b)

(i)

Proof. Show that $\hat{\beta}_1 = r\frac{s_Y}{s_X}$

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \tag{12}$$
$$= \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{S_{XX}} \tag{13}$$
$$= \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{s_X^2} \tag{14}$$
$$= \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})s_Y}{s_X^2 s_Y} \tag{15}$$
$$= \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{s_X s_Y} \cdot \frac{s_Y}{s_X} \tag{16}$$
$$= r \cdot \frac{s_Y}{s_X} \tag{17}$$

□

(ii)

Proof. Show that $\frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)} = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}}$

$$\frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)} = \frac{r\frac{s_Y}{s_X}}{\sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}} = \frac{r\frac{s_Y}{s_X}}{\sqrt{\frac{\hat{\sigma}^2}{s_X^2}}} = \frac{r\frac{s_Y}{s_X}}{\frac{\hat{\sigma}}{s_X}} = \frac{rs_Y}{\hat{\sigma}} = \frac{r}{\frac{1}{s_Y}\sqrt{\frac{\Sigma e_i^2}{n-2}}} = \frac{r\sqrt{n-2}}{\frac{1}{s_Y}\sqrt{\Sigma e_i^2}} \tag{18}$$

$$= \frac{r\sqrt{n-2}}{\frac{1}{s_Y}\sqrt{\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}} \tag{19}$$

$$= \frac{r\sqrt{n-2}}{\frac{1}{s_Y}\sqrt{\Sigma(Y_i - \bar{Y} + \hat{\beta}_1\bar{X} - \hat{\beta}_1 X_i)^2}} \tag{20}$$

$$= \frac{r\sqrt{n-2}}{\frac{1}{s_Y}\sqrt{\Sigma((Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X}))^2}} \tag{21}$$

$$= \frac{r\sqrt{n-2}}{\frac{1}{s_Y}\sqrt{\Sigma((Y_i - \bar{Y})^2 - 2(Y_i - \bar{Y})\hat{\beta}_1(X_i - \bar{X}) + \hat{\beta}_1^2(X_i - \bar{X})^2)}} \tag{22}$$

$$= \frac{r\sqrt{n-2}}{\frac{1}{s_Y}\sqrt{s_Y^2 - 2\hat{\beta}_1 S_{XY} + \hat{\beta}_1^2 s_X^2}} \tag{23}$$

$$= \frac{r\sqrt{n-2}}{\frac{1}{s_Y}\sqrt{s_Y^2 - 2r\frac{s_Y}{s_X}S_{XY} + \hat{\beta}_1^2 s_X^2}} \tag{24}$$

$$= \frac{r\sqrt{n-2}}{\frac{1}{s_Y}\sqrt{s_Y^2 - 2r\frac{s_Y}{s_X}r \cdot s_X s_Y + \hat{\beta}_1^2 s_X^2}} \qquad \text{Since } S_{XY} = rs_X s_Y \tag{25}$$

$$= \frac{r\sqrt{n-2}}{\frac{1}{s_Y}\sqrt{s_Y^2 - 2r^2 s_Y^2 + \hat{\beta}_1^2 s_X^2}} \tag{26}$$

$$= \frac{r\sqrt{n-2}}{\sqrt{\frac{s_Y^2}{s_Y^2} - 2r^2\frac{s_Y^2}{s_Y^2} + \hat{\beta}_1^2\frac{s_X^2}{s_Y^2}}} \tag{27}$$

$$= \frac{r\sqrt{n-2}}{\sqrt{\frac{s_Y^2}{s_Y^2} - 2r^2 + r^2}} \qquad \text{Since } \hat{\beta}_1\frac{s_X}{s_Y} = r \tag{28}$$

$$= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{29}$$

□

Q.3

(a) Answer:

```
n = 26
XBar = 1613 / n
YBar = 281.9 / n
SXX =  3756.96
SYY = 465.34
SXY = -757.64
```

```
slope = SXY / SXX
intercept = YBar - slope * XBar
```

The slope is -0.201663
The intercept is 23.3531729

(b) Answer:

```
SSE = SYY - slope^2*SXX # From lec 6, where syy = slope^2*sxx + see
sigma2 = ( SSE) / (n - 2) # From lecture
seB0 = sqrt((1/n + XBar^2/SXX)*sigma2)
seB1 = sqrt(sigma2/SXX)
```

$$s.e(\hat{\beta}_1) = 0.0588759$$
$$s.e(\hat{\beta}_1) = 3.7205019$$

(c) Answer:

```
alpha = 0.05

slopeLower =  slope - (qt(1-alpha/2, n-2) * seB1   )
slopeUpper = slope + (qt(1-alpha/2, n-2) * seB1 )

intLower = intercept - (qt(1-alpha/2, n-2) * seB0 )
intUpper = intercept + (qt(1-alpha/2, n-2) * seB0 )
```

The 95% confidence interval for the true slope is

$$[-0.3231768, -0.0801493]$$

The 95% confidence interval for the true intercept is

$$[15.6744343, 31.0319114]$$

(d) Answer:

Since both Confidence Intervals do not contain the value 0, we can conclude that there is a significant linear relationship between age and levels of CBG. With 95% confidence we estimate that the change in CBG decreases by between .32 and .08 for each additional increase of patients. We can also that at age=0, a patient will have CBG in between 15.68 and 31.03 in them.

Q. 4

a)

$$E = \Sigma(Y_i - \beta_1 X_i)^2 = \Sigma(Y_i^2 - 2Y_i\beta_1 X_i + \beta_1^2 X_i^2)$$
$$\frac{\partial E}{\partial \beta_1} = -2\Sigma Y_i X_i + 2\beta_i \Sigma X_i^2$$
$$\frac{\partial E}{\partial \beta_1} = 0 \implies \beta_1 = \frac{\Sigma Y_i X_i}{\Sigma X_i^2} \implies \hat{\beta}_1 = \frac{\Sigma Y_i X_i}{\Sigma X_i^2}$$

b) I cannot conclude that $\Sigma e_i = 0$

Proof. Suppose by contradiction, that $\Sigma e_i = 0$.
Then this holds for

$$X' = X_{1:2} = \{1, 2\}, Y' = Y_{1:2} = \{2, 1\}$$

However,

$$\begin{aligned}
\Sigma e_i &= \Sigma(Y_i - \hat{Y_i}) = \Sigma(Y_i - \frac{\Sigma Y_i X_i}{\Sigma X_i^2} X_i) && (30)\\
(\Sigma e_i)|_{X', Y'} &= Y_1 - \frac{Y_1X_1 + Y_2X_2}{X_1^2 + X_2^2} X_1 + Y_2 - \frac{Y_1X_1 + Y_2X_2}{X_1^2 + X_2^2} X_2 && (31)\\
&= 2 - \frac{2 \cdot 1 + 1 \cdot 2}{1^2 + 2^2} 1 + 1 - \frac{2 \cdot 1 + 1 \cdot 2}{1^2 + 2^2} 2 && (32)\\
&= 2 - \frac{4}{5} + 1 - \frac{4}{5} 2 && (33)\\
&= \frac{2}{5} \neq 0 && (34)\\
& && (35)
\end{aligned}$$

as we can see, there is an counterexample for $\Sigma e_i = 0$, therefore we cannot conclude that

$$\Sigma e_i = 0$$

□

c)

$$\begin{aligned}
s.e(\beta_1) &= Var(\Sigma e_i)|_{\sigma^2 = \hat{\sigma^2}} && (36)\\
&= Var(\frac{\Sigma X_i Y_i}{\Sigma X_i^2})|_{\sigma^2 = \hat{\sigma^2}} && (37)\\
&= \frac{1}{(\Sigma X_i^2)^2} Var(\Sigma X_i Y_i)|_{\sigma^2 = \hat{\sigma^2}} && (38)\\
&= \frac{1}{(\Sigma X_i^2)^2} \Sigma X_i^2 Var(Y_i)|_{\sigma^2 = \hat{\sigma^2}} && (39)\\
&= \frac{\hat{\sigma^2}}{(\Sigma X_i^2)^2} (\Sigma X_i)^2 && (40)\\
&= \frac{\hat{\sigma^2}}{\Sigma X_i^2} && (41)
\end{aligned}$$

$$

d)

$$t^* = \frac{\hat{\beta}_1 - \beta_1}{s.e(\beta_1)} = \frac{\hat{\beta}_1 \Sigma X_i^2}{\hat{\sigma^2}}$$

If $2 \cdot P(t_(n-2) \geq |t^*|) < \alpha$, then we reject the hypothesis, otherwise, we are unable to reject.

e) Note that

$$\text{MLE of } \beta_1 = \hat{\beta}_1 = \frac{\Sigma Y_i X_i}{\Sigma X_i^2}$$

```
numerator = 0
denomerator = 0
data = data.frame(X = c(7,12,4,14,25,30),
                  Y = c(128,213,75,250,446,540))

for(i in seq_len(nrow(data))) { # zip X and Y
  numerator  = numerator + (data[i,1] * data[i,2]) # increment X_i * Y_i
  denomerator = denomerator + data[i,1]^2 # increment X_i^2
}

beta1MLE = numerator / denomerator
```

the MLE of β_1 is 17.9284974

Q. 5

Proof. Show that $Var(\hat{\beta}_1) \leq Var(\hat{\beta}_1^*)$
Note: since $\hat{\beta}_1^*$ is an unbiased estimator,

$$E(\hat{\beta}_1^*) = \Sigma c_1\beta_0 + \beta_1\Sigma c_i X_i = \beta_1 \implies \Sigma c_i = 0 \wedge \Sigma c_i X_i = 1$$

let

$$c_i = k_i - p_i, \text{where } k_i = \frac{X_i - \bar{X}}{\Sigma(X_i - \bar{X})^2}$$

$$\begin{aligned}
Var(\hat{\beta}_1^*) &= \sigma^2\Sigma(k_i - p_i)^2 && (42)\\
&= \sigma^2(\Sigma k_i^2 + 2\Sigma k_i p_i + \Sigma p_i^2) && (43)\\
&\geq \sigma^2(\Sigma k_i^2 + 2\Sigma k_i p_i) && (44)\\
&= \sigma^2(\Sigma k_i^2 + 2\Sigma k_i(k_i - c_i)) \qquad \text{Since } p_i = k_i - c_i && (45)\\
&= \sigma^2(\Sigma k_i^2 + 2\Sigma(k_i^2 - k_i c_i)) && (46)\\
&= \sigma^2(\Sigma k_i^2 + 2\Sigma k_i^2 - 2\Sigma k_i c_i)) && (47)\\
&= \sigma^2(\Sigma k_i^2 + 2\Sigma\frac{1}{\Sigma(X_i - \bar{X})^2} - 2\Sigma\frac{X_i - \bar{X}}{\Sigma(X_i - \bar{X})^2}c_i) && (48)\\
&= \sigma^2(\Sigma k_i^2 + 2\Sigma\frac{1}{\Sigma(X_i - \bar{X})^2} - 2\frac{\Sigma c_i X_i - \Sigma c_i \bar{X}}{\Sigma(X_i - \bar{X})^2}) && (49)\\
&= \sigma^2(\Sigma k_i^2 + 2\Sigma\frac{1}{\Sigma(X_i - \bar{X})^2} - 2\frac{1 - 0}{\Sigma(X_i - \bar{X})^2}) \qquad \text{Since } \Sigma c_i = 0 \wedge \Sigma c_i X_i = 1 && (50)\\
&= \sigma^2(\Sigma k_i^2) && (51)\\
&\geq \sigma^2(\Sigma k_i) = Var(\hat{\beta}_1) && (52)
\end{aligned}$$

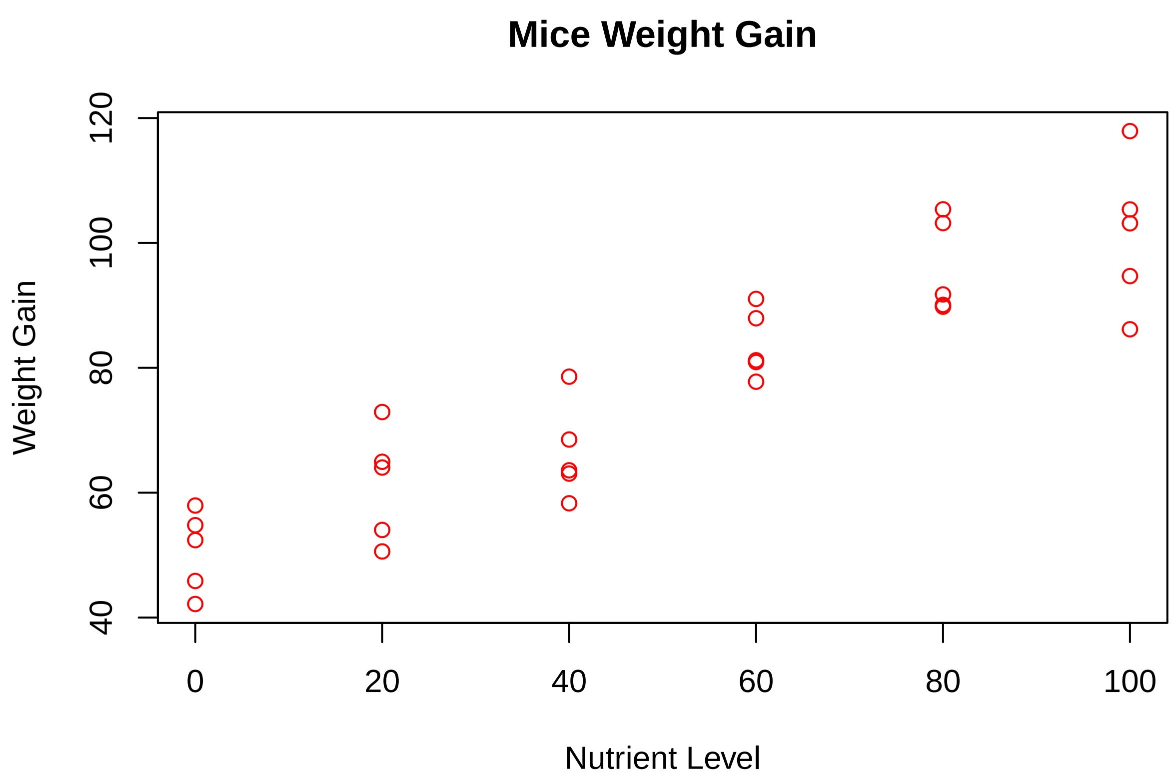
□

Q. 6

```
## Loading required package: readxl
```

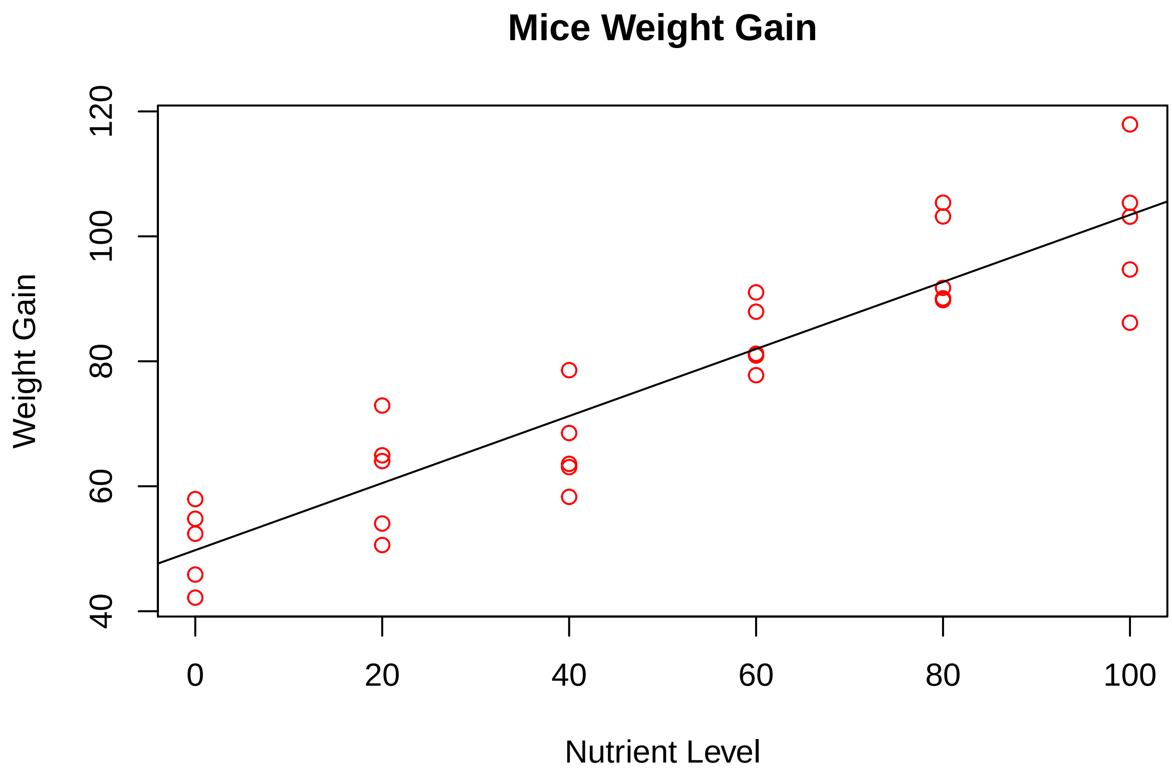
a) Answer :

```
xls_data = read_excel("MiceWeightGain.xls") # import data

plot(xls_data$x,
     xls_data$y,
     type= "p",
     xlab = 'Nutrient Level',
     ylab = 'Weight Gain',
     main="Mice Weight Gain",
     col = "red")
```



b) Answer :

```
plot(xls_data$x,
     xls_data$y,
     type= "p",
     xlab = 'Nutrient Level',
     ylab = 'Weight Gain',
     main="Mice Weight Gain",
     col = "red")

abline(lm(y ~ x, data= xls_data))
```



c) Answer :

```
miceCorrelation = cor(xls_data$x, xls_data$y)
```

There is a strong positive association between weight change and nutrient level as the correlation is 0.9186592

d) Answer :

```
CI = confint(lm(y ~ x, data= xls_data), 'x', level = 0.95)
```

The 95% confidence interval for the mean change in weight as nutrient level is increased by 1 unit is

$$[0.4473474, 0.626033]$$