Assignment #1

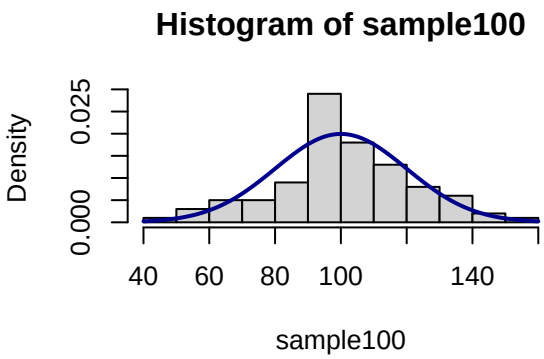
Aly Abdelwahed, Manish Suresh

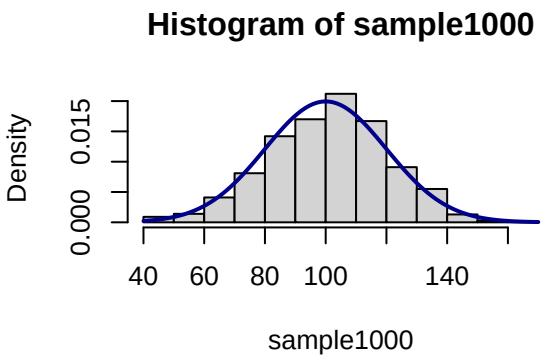
1/21/2021

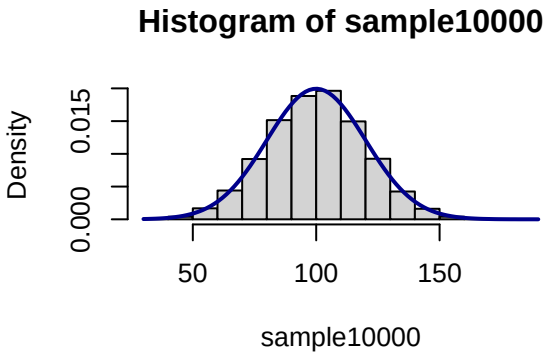
Question 1

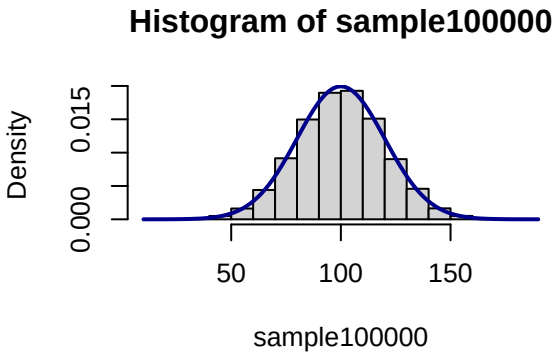
Part A

```
set.seed(1005228013)

x <- seq(20, 160, by = 10)


sample100 <- rnorm(100, mean=100, sd=20)
sample1000 <- rnorm(1000, mean=100, sd=20)
sample10000 <- rnorm(10000, mean=100, sd=20)
sample100000 <- rnorm(100000, mean=100, sd=20)
par(mfrow=c(2,2))

hist(sample100, prob =TRUE)
curve(dnorm(x, mean=100, sd=20),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")

hist(sample1000, prob=TRUE)
curve(dnorm(x, mean=100, sd=20),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")

hist(sample10000, prob=TRUE)
curve(dnorm(x, mean=100, sd=20),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")

hist(sample100000, prob=TRUE)
curve(dnorm(x, mean=100, sd=20),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```









- Accuracy notes
  - In sample 1 (100 samples) we can say it is average accuarcy (60%) because most of the histogram touches the density graph with a few mismatches.
  - In sample 2 (1000 samples) we can say it is fair/somewhat accurate (75%) because the every bar of the histogram touches the density graph but density graph does not pass through the center of the bar indicating that the accuracy is off.
  - In sample 3 (10000 samples) we can say it is pretty accurate (95%) because the density graph of the normal distributuion passes through the middle of every bar on the histogram as seen above but the mean of of the histogram is slightly shifted.
  - In sample 4 (100000 samples) we can say it is very accurate (99%) because the density graph of the normal distributuion passes through the middle of every bar on the histogram as seen above.

Part B

	Theoretical Values	Sample100	Sample1000	Sample10000	Sample100000
Mean	**100**	**100.9533276**	**100.9755801**	**99.9551842**	**100.0683012**
Standard Deviation	**20**	**20.6816993**	**19.7701735**	**19.8340198**	**19.9861063**
2.5 Percentile	**60.8007203**	**56.3801422**	**61.7847628**	**61.1645676**	**60.8534245**
25 Percentile	**86.510205**	**90.3942161**	**87.5306466**	**86.4059925**	**86.527583**
50 Percentile	**100**	**99.2083509**	**101.9617392**	**100.1571274**	**100.1198885**
75 Percentile	**113.489795**	**115.4646455**	**113.9047268**	**113.2598264**	**113.4810559**
97.5 Percentile	**139.1992797**	**139.3279282**	**137.0045589**	**138.7266533**	**139.1497184**

- Comparison (TBD)

Question 2

Part A

Part 1

We know that

$$S_{XX} = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$S_{XX} = \sum_{i=1}^{n} (X_i^2 - 2X_i\bar{X} + \bar{X}^2)$$

$$S_{XX} = \sum_{i=1}^{n} (X_i^2) - \sum_{i=1}^{n} (2X_i\bar{X}) + \sum_{i=1}^{n} (\bar{X}^2)$$

$$S_{XX} = \sum_{i=1}^{n} (X_i^2) - 2\bar{X} \sum_{i=1}^{n} (X_i) + n(\bar{X}^2)$$

$$S_{XX} = \sum_{i=1}^{n} (X_i^2) - 2\bar{X} n\bar{X} + n(\bar{X}^2)$$

$$S_{XX} = \sum_{i=1}^{n} (X_i^2) - 2n\bar{X}^2 + n(\bar{X}^2)$$

$$S_{XX} = \sum_{i=1}^{n} (X_i^2) - n\bar{X}^2$$

as wanted.

Part 2

We know that

$$S_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

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

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

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

$$S_{XY} = \sum_{i=1}^{n}(X_iY_i) - \bar{Y}n\bar{X} - \bar{X}n\bar{Y} + n(\bar{X}\bar{Y})$$

$$S_{XY} = \sum_{i=1}^{n}(X_iY_i) - n(\bar{X}\bar{Y})$$

as wanted

Part B

Part 1

$$r\frac{S_x}{S_y} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{(X_i - \bar{X})(Y_i - \bar{Y}}{S_xS_y}\right)\frac{S_y}{S_x}$$

$$r\frac{S_x}{S_y} = \frac{1}{(n-1)S_xS_y}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\frac{S_y}{S_x}$$

$$r\frac{S_x}{S_y} = \frac{1}{(n-1)S_x}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\frac{1}{S_x}$$

$$r\frac{S_x}{S_y} = \frac{1}{(n-1)S_x^2}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$r\frac{S_x}{S_y} = \frac{1}{(n-1)Var(x)}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$r\frac{S_x}{S_y} = \frac{1}{(n-1)\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$r\frac{S_x}{S_y} = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$r\frac{S_x}{S_y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

as wanted

Part B

Part 2

We want to show

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

Expanding the LHS we get

$$LHS = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \cdot \frac{1}{s.e(\hat{\beta}_1)}$$

$$LHS = \frac{S_{XY}}{S_{XX}} \cdot \frac{\sqrt{S_{XX}}}{\hat{\sigma}}$$

$$LHS = \frac{S_{XY}}{\sqrt{S_{XX}}} \cdot \frac{\sqrt{n-2}}{\sqrt{SSE}}$$

$$LHS = \frac{S_{XY}\sqrt{n-2}}{\sqrt{S_{XX} \cdot SSE}}$$

$$LHS = \frac{S_{XY}\sqrt{n-2}}{\sqrt{S_{XX} \cdot (SST - SSR)}}$$

$$LHS = \frac{S_{XY}\sqrt{n-2}}{\sqrt{S_{XX} \cdot (S_{YY} - \hat{\beta}_1^{\,2} \cdot S_{XX})}}$$

$$LHS = \frac{S_{XY}\sqrt{n-2}}{\sqrt{S_{XX} \cdot S_{YY} - \hat{\beta}_1^{\,2} \cdot (S_{XX})^2}}$$

$$LHS = \frac{S_{XY}\sqrt{n-2}}{\sqrt{S_{XX} \cdot S_{YY} - (S_{XY})^2}}$$

Expanding the RHS we get

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

$$RHS = \frac{1}{n-1}\sum(\frac{Y_i - \bar{Y}}{S_Y})(\frac{X_i - \bar{X}}{S_X}) \cdot \frac{\sqrt{n-2}}{\sqrt{1 - (\frac{1}{n-1}\sum(\frac{Y_i \cdot \bar{Y}}{S_Y})(\frac{X_i \cdot \bar{X}}{S_X}))^2}}$$

$$RHS = \frac{\sum(Y_i - \bar{Y})(X_i - \bar{Y})}{(n-1) \cdot S_X \cdot S_Y} \cdot \frac{\sqrt{n-2}}{\sqrt{1 - (\frac{\sum(Y_i - \bar{Y})(X_i - \bar{Y})}{n-1 \cdot S_X \cdot S_Y})^2}}$$

$$RHS = \frac{S_{XY}}{(n-1) \cdot S_X \cdot S_Y} \cdot \frac{\sqrt{n-2}}{\sqrt{1 - \frac{(S_{XY})^2}{(n-1)^2 \cdot (S_X)^2 \cdot (S_Y)^2}}}$$

$$RHS = \frac{S_{XY}}{(n-1) \cdot S_X \cdot S_Y} \cdot \frac{\sqrt{n-2}}{\sqrt{\frac{(n-1)^2 \cdot (S_X)^2 \cdot (S_Y)^2 - (S_{XY})^2}{(n-1)^2 \cdot (S_X)^2 \cdot (S_Y)^2}}}$$

$$RHS = \frac{S_{XY}}{(n-1) \cdot S_X \cdot S_Y} \cdot \frac{\sqrt{n-2} \cdot \sqrt{(n-1)^2 \cdot (S_X)^2 \cdot (S_Y)^2}}{\sqrt{(n-1)^2 \cdot (S_X)^2 \cdot (S_Y)^2 - (S_{XY})^2}}$$

$$RHS = \frac{S_{XY}}{(n-1) \cdot S_X \cdot S_Y} \cdot \frac{\sqrt{n-2} \cdot (n-1) \cdot S_X \cdot S_Y}{\sqrt{(n-1)^2 \cdot (S_X)^2 \cdot (S_Y)^2 - (S_{XY})^2}}$$

$$RHS = \frac{S_{XY} \cdot \sqrt{n-2}}{\sqrt{(n-1)^2 \cdot (S_X)^2 \cdot (S_Y)^2 - (S_{XY})^2}}$$

$$RHS = \frac{S_{XY} \cdot \sqrt{n-2}}{\sqrt{(n-1)^2 \cdot Var_X \cdot Var_Y - (S_{XY})^2}}$$

$$RHS = \frac{S_{XY} \cdot \sqrt{n-2}}{\sqrt{(n-1)^2 \cdot \frac{S_{XX}}{(n-1)} \cdot \frac{S_{YY}}{(n-1)} - (S_{XY})^2}}$$

$$RHS = \frac{S_{XY} \cdot \sqrt{n-2}}{\sqrt{S_{XX} \cdot S_{YY} - (S_{XY})^2}}$$

We have LHS = RHS as wanted

Question 3

Part A

The least square estimates are:

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X}$$

$$\hat{\beta}_1 = \frac{-757.64}{3756.96} \qquad \hat{\beta}_0 = \frac{281.9}{26} - \hat{\beta}_1$$

$$\hat{\beta}_1 = -0.202 \qquad \hat{\beta}_0 = 10.84 - (-0.202) \cdot 62.04$$

$$\hat{\beta}_0 = 685.151$$

The least square estimated for $\hat{\beta}_1$ is -0.202 and $\hat{\beta}_0$ is 685.151

Part B

We know from Question 2 part b that:

$$\hat{\beta}_1 = r\frac{S_y}{S_x} \qquad \& \qquad \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$\therefore \frac{\frac{r \times S_y}{S_x}}{SE(\hat{\beta}_1)} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} <=> \frac{1}{SE(\hat{\beta}_1)} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \div \frac{r \times S_y}{S_x}$$

$$<=> \frac{1}{SE(\hat{\beta}_1)} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \times \frac{S_x}{r \times S_y} = \frac{S_x\sqrt{n-2}}{S_y\sqrt{1-r^2}}$$

$$<=> SE(\hat{\beta}_1) = \frac{S_y\sqrt{1-r^2}}{S_x\sqrt{n-2}}$$

Now we know that:

$$S_{XX} = (n-1)(S_x)^2 \qquad \& \qquad S_{YY} = (n-1)(S_y)^2$$

$$\therefore 3756.96 = (26-1)(S_x)^2 \qquad \& \qquad 465.34 = (26-1)(S_y)^2$$
$$<=> \frac{3756.96}{25} = (S_x)^2 \qquad \& \qquad \frac{465.34}{25} = (S_y)^2$$

$$<=> \sqrt{\frac{3756.96}{25}} = S_x \qquad \& \qquad \sqrt{\frac{465.34}{25}} = S_y$$
$$<=> S_x = 12.26 \qquad \& \qquad S_y = 4.31$$

$$\therefore \ \hat{\beta}_1 = r\frac{12.26}{4.31}$$
$$\therefore \ -0.202 = 2.84r$$
$$<=> r = -0.0711$$
$$\therefore \ \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} <=> \frac{SE(\hat{\beta}_1)}{\hat{\beta}_1} = \frac{\sqrt{1-r^2}}{r\sqrt{n-2}} <=> SE(\hat{\beta}_1) = \frac{\hat{\beta}_1\sqrt{1-r^2}}{r\sqrt{n-2}}$$
$$\therefore SE(\hat{\beta}_1) = \frac{\hat{\beta}_1\sqrt{1-r^2}}{r\sqrt{n-2}} = \frac{-0.202\sqrt{1-(-0.0711)^2}}{-0.0711\sqrt{26-2}} = 0.6$$

Now, we know that:

$$SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}} <=> 0.6 = \sqrt{\frac{\hat{\sigma}^2}{3756.96}} <=> 0.36 = \frac{\hat{\sigma}^2}{3756.96}$$

$$\therefore \hat{\sigma}^2 = 1352.51$$

We know from the lecture notes that: $SE(\beta_0) = S_{b_0} = \sqrt{(\frac{1}{n} + \frac{\overline{X}}{S_{XX}})\hat{\sigma}^2} = \sqrt{(\frac{1}{26} + \frac{\frac{1613}{26}}{3756.96})1352.51} = 8.62$

$$\therefore SE(\hat{\beta}_0) = 8.62 \ \& \ SE(\hat{\beta}_1) = 0.6$$

Part c

The 95% confidence interval for the slope is:

$$(\hat{\beta}_1 - t_{0.975}(24) \times SE(\hat{\beta}_1), \ \hat{\beta}_1 + t_{0.975}(24) \times SE(\hat{\beta}_1))$$
$$= (-0.202 - (2.064 \times 0.6), \ -0.202 + (2.064 \times 0.6))$$
$$= (-1.44, \ 1.04)$$

The 95% confidence interval for the intercept is:

$$(\hat{\beta}_1 - t_{0.975}(24) \times SE(\hat{\beta}_0), \ \hat{\beta}_0 + t_{0.975}(24) \times SE(\hat{\beta}_0))$$
$$= (685.151 - (2.064 \times 8.62), \ 685.151 + (2.064 \times 8.62))$$
$$= (667.36, \ 702.94)$$

Part d

Interpretation for the slope:

With 95% confidence, we estimate that the mean of the change the levels of cortisol-binding globulin (CBG) changes by between a decrease of 1.44 to an increase 1.04 for each additional increase of the person's age.

Interpretation for the intercept:

With 95% confidence, we estimate that the change the levels of cortisol-binding globulin (CBG) is between an increase of 667.46 to 702.94 when the person is born

Question 4

Part A

We want to find the least square estimate of β_1 which is the element $\hat{\beta_1}$ that minimizes the equation

$$Q = \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2$$

Now,

$$\begin{aligned}
\frac{d}{d\beta_1} Q &= \sum_{i=1}^{n} 2(Y_i - \beta_1 X_i) \times \frac{d}{d\beta_1}[(Y_i - \beta_1 X_i)] \\
&= \sum_{i=1}^{n} 2(Y_i - \beta_1 X_i) \times (-X_i) \\
&= -2 \sum_{i=1}^{n} (Y_i X_i - \beta_1 X_i^2) \\
&= -2(\sum_{i=1}^{n} Y_i X_i \ - \beta_1 \sum_{i=1}^{n} X_i^2) \qquad (*)
\end{aligned}$$

Now, setting the equation $*$ equal to zero yields the following:

$$\begin{aligned}
-2(\sum_{i=1}^{n} Y_i X_i \ - \beta_1 \sum_{i=1}^{n} X_i^2) &= 0 \\
=> \sum_{i=1}^{n} Y_i X_i \ - \beta_1 \sum_{i=1}^{n} X_i^2 &= 0 \\
=> \sum_{i=1}^{n} Y_i X_i \ &= \beta_1 \sum_{i=1}^{n} X_i^2 \\
=> \hat{\beta_1} &= \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}
\end{aligned}$$

Therefore the least square estimator of β_1 is

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}$$

Question 5

We know that the point estimator $b_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$ and that for the normal error regression model,

the sampling distribution of $b_1$is normal, with mean and variance: $E\{b_1\} = \beta_1$ and $\sigma^2\{b_1\} = \frac{\sigma^2}{\sum (X_i - \overline{X})^2}$

We also know that b_1 as a linear combination of the observations Y_i as follows:

$$b_1 = \sum k_i Y_i \text{ where } k_i = \frac{X_i - \overline{X}}{\sum (X_i - \overline{X})^2} (*)$$

We know the following properties from the textbook about the coefficients k_i which are:

$$\sum k_i = 0 \ \& \sum k_i X_i = 1 \ \& \sum k_i^2 = \frac{1}{\sum (X_i - \overline{X})^2}$$

The unbiasedness of the point estimator b_1 :

$$E\{b_1\} = E\{\sum k_i Y_i\} = \sum k_i E\{Y_i\} = \sum k_i(\beta_0 + \beta_1 X_i)$$

$$= \beta_0 \sum k_i + \beta_1 \sum k_i X_i = \beta_0(0) + \beta_1(1) = \beta_1$$

Similarly, The variance of b_i can be derived readily. We only need to remember that Y_i are independent random variables, each with variance σ^2, and that k_i are constants. Therefore:

$$\sigma^2\{b_1\} = \sigma^2\{\sum k_i Y_i\} = \sum k_i^2 \sigma^2\{Y_i\} = \sum k_i^2 \sigma^2 = \sigma^2 \sum k_i^2 = \sigma^2 \frac{1}{\sum (X_i - \overline{X})^2}$$

Now, We can estimate the variance of the sampling distribution of b_1 by replacing the parameter σ^2 with MSE,

$$\text{the unbiased estimator of } \sigma^2\text{is: } s^2\{b_1\} = \frac{MSE}{\sum (X_i - \overline{X})^2}$$

Now, the point estimator $s^2\{b_1\}$ is an unbiased estimator of $\sigma^2\{b_1\}$. Taking the positive square root, we obtain $s\{b_1\}$, the point estimator of $\sigma\{b_1\}$.

Now to show that the least squares estimator of β_1 has the minimum variance among all other linear unbiased estimators of the form: $\hat{\beta}_1 = \sum c_i Y_i$ where $c_i$are arbitrary constants

$\because \hat{\beta}_1$is required to be unbiased, the following must hold:

$$E\{\hat{\beta}_1\} = E\{\sum c_i Y_i\} = \sum c_i E\{Y_i\} = \beta_1$$

Now, we know that: $E\{Y_i\} = \beta_0 + \beta_1 X_i$

$$\therefore E\{\hat{\beta}_1\} = \sum c_i(\beta_0 + \beta_1 X_i) = \sum c_i\beta_0 + c_i\beta_1 X_i) = \sum c_i\beta_0 + \sum c_i\beta_1 X_i) = \beta_0 \sum c_i + \beta_1 \sum c_i X_1 = \beta_1$$

$\therefore$ For the unbiasedness condition to hold, the c_i must follow the restrictions:

$$\sum c_i = 0 \qquad \& \qquad \sum c_i x_i = 1 \qquad (**)$$

Now, we know that: $\sigma^2\{\hat{\beta}_1\} = \sum c_i^2 \sigma^2\{Y_i\} = \sigma^2 \sum c_i^2$ Now, Let $c_i = k_i + d_i$ where k_i are the least squares constants in the equation (*) and d_i are arbitrary constants

$$\therefore \sigma^2\{\beta_1\} = \sigma^2 \sum c_i^2 = \sigma^2 \sum (k_i + d_i)^2$$

$$= \sigma^2 \sum (k_i^2 + d_i^2 + 2k_i d_i) = \sigma^2(\sum k_i^2 + \sum d_i^2 + 2\sum k_i d_i)$$

$$= \sigma^2 \sum k_i^2 + \sigma^2 \sum d_i^2 + 2\sigma^2 \sum k_i d_i) = \sigma^2\{b_1\} + \sigma^2 \sum d_i^2 + 2\sigma^2 \sum k_i d_i)$$

Now,

$$\sum k_i d_i = \sum k_i(c_i - k_i) = \sum c_i k_i - k_i^2 = \sum c_i k_i - \sum k_i^2$$

$$= \sum c_i \big[\frac{X_i - \overline{X}}{\sum (X_i - \overline{X})^2}\big] - \frac{1}{\sum (X_i - \overline{X})^2}$$

$$= \sum \frac{c_i X_i - c_i \overline{X}}{\sum (X_i - \overline{X})^2} - \frac{1}{\sum (X_i - \overline{X})^2}$$

$$= \frac{\sum c_i X_i - \sum c_i \overline{X}}{\sum (X_i - \overline{X})^2} - \frac{1}{\sum (X_i - \overline{X})^2}$$

$$= \frac{\sum c_i X_i - \overline{X} \sum c_i}{\sum (X_i - \overline{X})^2} - \frac{1}{\sum (X_i - \overline{X})^2}$$

$$= \frac{1 - \overline{X}(0)}{\sum (X_i - \overline{X})^2} - \frac{1}{\sum (X_i - \overline{X})^2} \qquad \textbf{by (**)}$$

$$= \frac{1}{\sum (X_i - \overline{X})^2} - \frac{1}{\sum (X_i - \overline{X})^2} = 0$$

$\therefore \sigma^2\{\hat{\beta}_1\} = \sigma^2\{b_1\} + \sigma^2 \sum d_i^2$ **Note that the smallest value of** $\sum d_i^2$ **is zero. Hence, the variance of** β_1 **is at a minimum when** $\sum d_i^2 = 0$**. This can only occur when** $d_i = 0 \ \forall d_i$

$$\therefore c_i \equiv k_i$$

$\therefore$ **The least squares estimator** b_1 **has minimum variance among all unbiased linear estimators as needed**

Question 6

Packages Required

```
library(tidyverse)
library(ggplot2)
```

Reading in the data

```
myData <- read_csv("MiceWeightGain.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   x = col_double(),
##   y = col_double()
## )
```

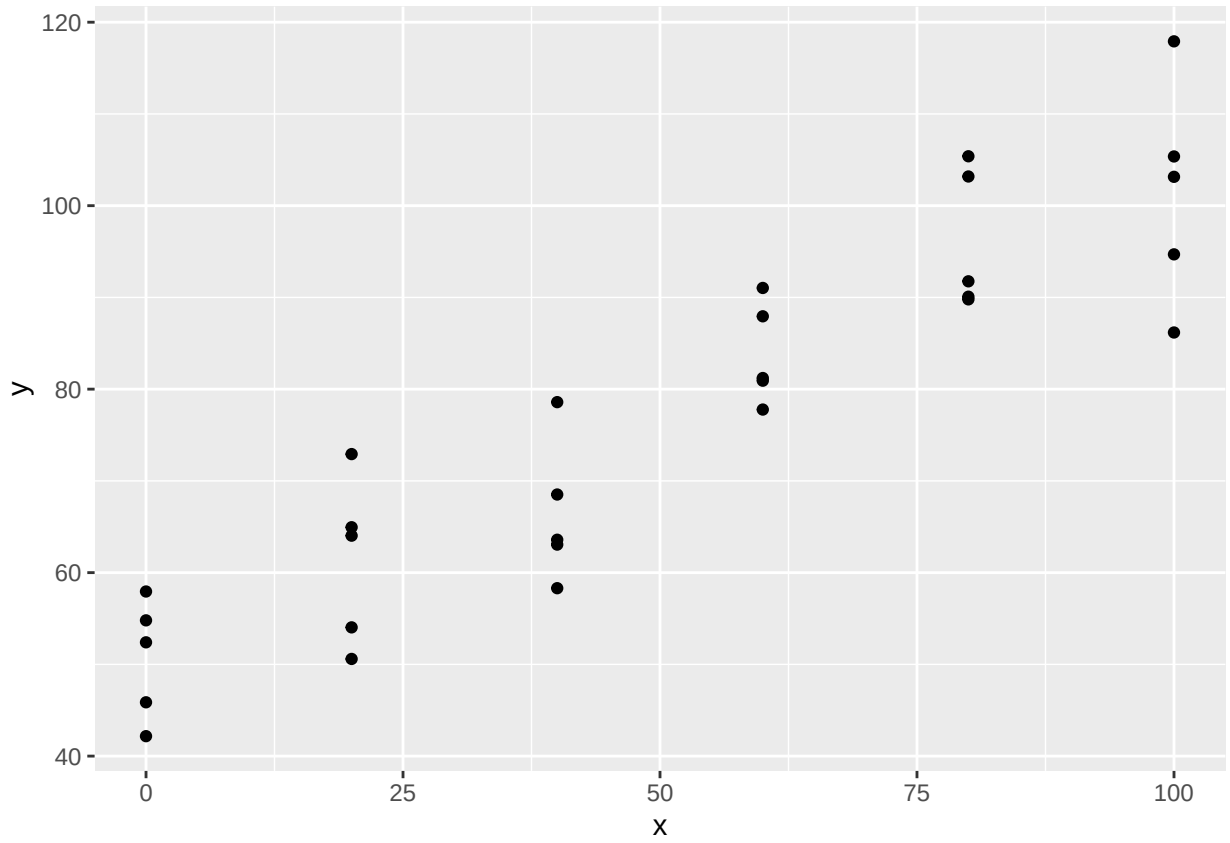
Displaying some data

```
myData
```

```
## # A tibble: 30 x 2
##        x     y
##    <dbl> <dbl>
##  1     0  52.4
##  2     0  45.9
##  3     0  57.9
##  4     0  54.8
##  5     0  42.2
##  6    20  64.0
##  7    20  54.0
##  8    20  72.9
##  9    20  50.6
## 10    20  64.9
## # ... with 20 more rows
```

Part A - Drawing a scatter plot of weight change versus nutrient level

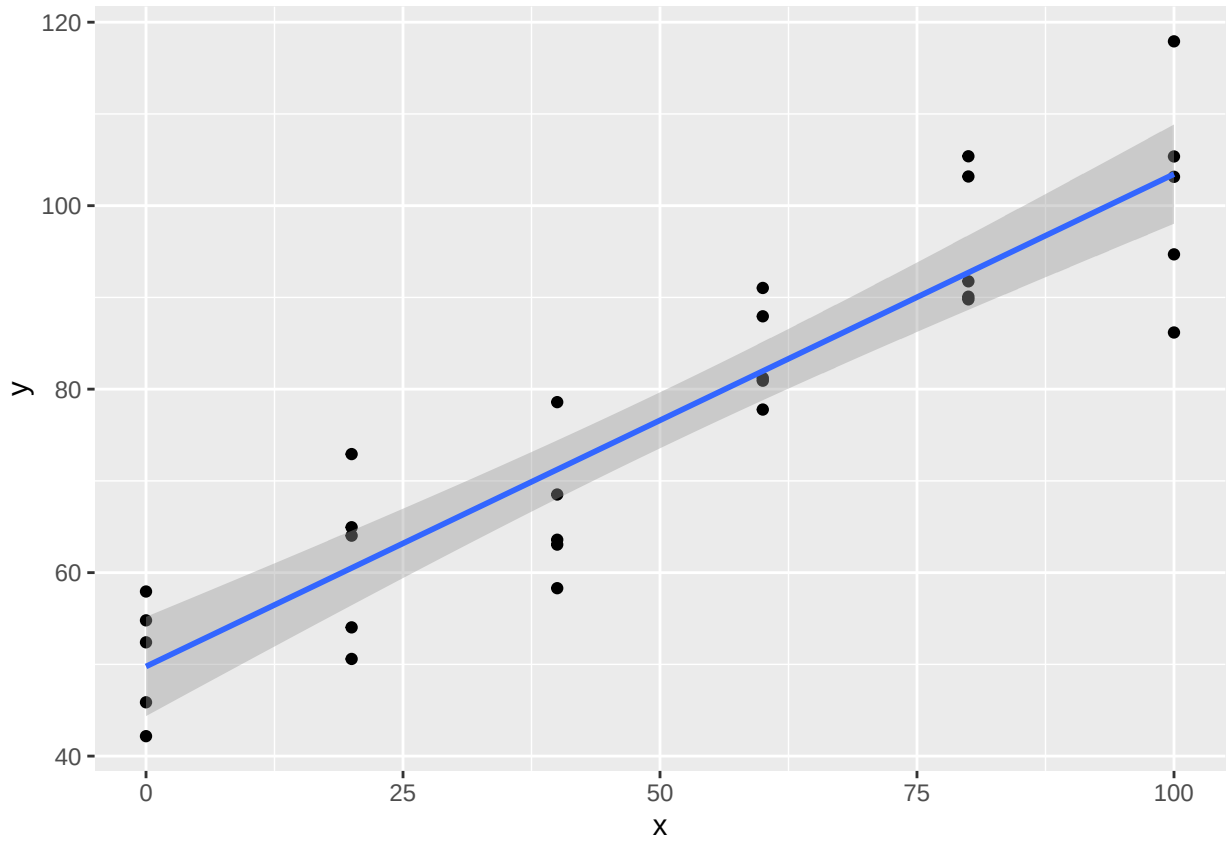
```
ggplot(myData, aes(x = x, y = y)) + geom_point()
```



Part B - Fitting a simple linear regression, relating weight change to nutrient level

```
ggplot(myData, aes(x = x, y=y)) + geom_point() + geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Part C - Testing whether there is a positive association between weight change and nutrient level

Since we are testing whether there is a positive association between weight change and nutrient level, our Null Hypothesis and alternate Hypothesis is

$$H_0 = 0 \;\; H_a > 0$$

Fit the model and get the summary of the regression fit.

```
fit <- lm(y~x, data=myData)

stats <- summary(fit)

stats
```

```
##
## Call:
## lm(formula = y ~ x, data = myData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.2614  -5.9008  -0.8542   5.7353  14.4746
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.76937    2.64107   18.84  < 2e-16 ***
## x            0.53669    0.04362   12.30 8.22e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.16 on 28 degrees of freedom
## Multiple R-squared:  0.8439, Adjusted R-squared:  0.8384
## F-statistic: 151.4 on 1 and 28 DF,  p-value: 8.222e-13
```

We can gather from the summary that the test statistic is 12.304964 and the p-value is $4.1109934 \times 10^{-13}$.

Note: Since this is one sided test the p-value formula is $P(|t^*| > \alpha)$ and not $P(|t^*| > \alpha) * 2$.

Using the p-value approach, since the p-value is less than the level of significance 0.05. We reject the Null Hypothesis.

That means we can say that there is a positive association between weight change and nutrient level.

Part D - A 95% confidence interval for the mean change in weight as nutrient level is increased by 1 unit

We can gather from the summary the following data

- $$\hat{\beta_1} = 0.5366906$$
- $$SE(\hat{\beta_1}) = 0.0436158$$

From the above information and using the formula $\hat{\beta_1} \pm t_{(1-\frac{\alpha}{2})(n-2)} \cdot SE(\hat{\beta_1})$ we can say that the confidence interval is

$$(0.4473477, 0.6260334)$$

That is we estimate with 95% confidence that the mean change in weight increases between **0.4473477** and **0.6260334** as nutrient level is increased by 1 unit