MATH 564 - Assignment1

—————————**Mohammed Wasim R D(A20497053)**—————————

##Problem 1

Reading the data

```
data<-read.csv("/Users/mohammedwasimrd/Desktop/Excel1.csv")
```

Renaming the coloumns

```
##   Muscle Age
## 1    106  41
## 2     97  47
## 3    113  46
## 4     96  45
## 5    119  41
## 6     92  47
```

Obtaining the regression equation

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
linear<-lm(Muscle ~ Age,data = data)
linear
```

```
##
## Call:
## lm(formula = Muscle ~ Age, data = data)
##
## Coefficients:
## (Intercept)          Age
##     156.224       -1.188
```

```
summary(linear)
```

```
##
## Call:
## lm(formula = Muscle ~ Age, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.121  -6.373  -0.674   6.968  23.455
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.22438    5.68612   27.48   <2e-16 ***
## Age          -1.18820    0.09265  -12.82   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.244 on 57 degrees of freedom
## Multiple R-squared:  0.7426, Adjusted R-squared:  0.7381
## F-statistic: 164.5 on 1 and 57 DF,  p-value: < 2.2e-16
```

Regression function is Yi=156.22438-1.18820Xi+e

Regression fits well as we can see R squared value is 0.7381

```
library(ggplot2)
plot(data$Age, data$Muscle, main = "Muscle Mass versus Age", xlab = "Age", ylab = "Muscle Mass")
abline(linear)
```

(b).(1) : A point estimate of the difference in mean muscle mass for women differing in age in one year is -1.19

```
predict(linear, data.frame(Age = 60), interval = "prediction")
```

```
##        fit      lwr      upr
## 1 84.93239 68.28508 101.5797
```

(b).(2) : A point estimate of the mean muscle mass for women aged X=60 years is 84.93 with prediction between 68.28 and 101.57

(b)(3)Value of the 8th residual is -7.190

```
linear$resid[8]
```

```
##        8
## -7.19079
```

(b)(4)Point estimate od variance is 68%

```r
anova(linear)
```

```
## Analysis of Variance Table
##
## Response: Muscle
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Age        1 11178.3   11178  164.48 < 2.2e-16 ***
## Residuals 57  3873.7      68
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problem 2

Reading the data

```r
data1<-read.table("http://www.cnachtsheim-text.csom.umn.edu/Kutner/Chapter%20%201%20Data%20Sets/CH01PR28
```

Renaming the coloumns

```r
colnames(data1)[1] ="Crime"
colnames(data1)[2]="Percentage"
head(data1)
```

```
##   Crime Percentage
## 1  8487         74
## 2  8179         82
## 3  8362         81
## 4  8220         81
## 5  6246         87
## 6  9100         66
```

Obtaining the regression equation

```r
linear1<-lm(Crime ~ Percentage,data = data1)
linear1
```

```
##
## Call:
## lm(formula = Crime ~ Percentage, data = data1)
##
## Coefficients:
## (Intercept)   Percentage
##     20517.6       -170.6
```

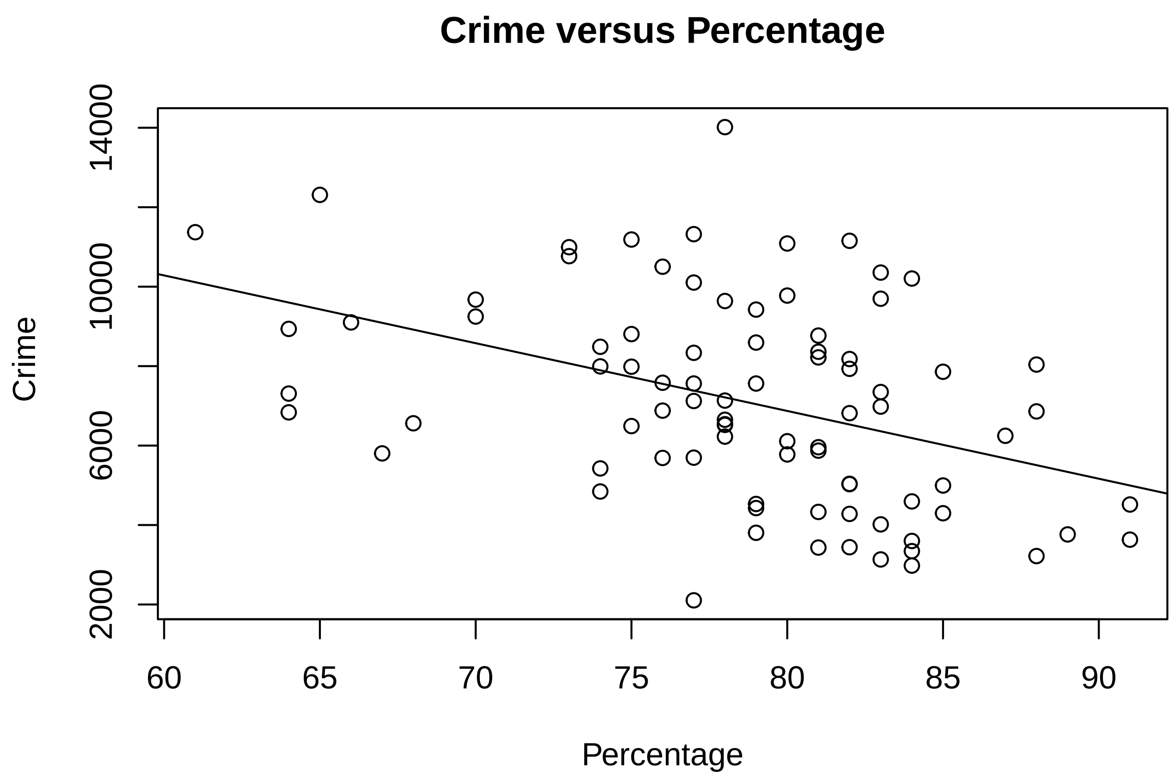
```r
summary(linear1)
```

```
##
## Call:
## lm(formula = Crime ~ Percentage, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5278.3 -1757.5  -210.5  1575.3  6803.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20517.60    3277.64   6.260 1.67e-08 ***
## Percentage   -170.58      41.57  -4.103 9.57e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2356 on 82 degrees of freedom
## Multiple R-squared:  0.1703, Adjusted R-squared:  0.1602
## F-statistic: 16.83 on 1 and 82 DF,  p-value: 9.571e-05
```

Regression function is Yi=20517.60-170.58Xi+e

```
library(ggplot2)
plot(data1$Percentage, data1$Crime, main = "Crime versus Percentage", xlab = "Percentage", ylab = " Cri
abline(linear1)
```



It is not a best fit as data has huge outliers

b.1 : The difference in the mean crime rate for two counties whose high-school graduation rates differ by one percentage point is -170.58

```
predict(linear1, data.frame(Percentage = 80), interval = "prediction")
```

```
##        fit     lwr      upr
## 1 6871.585 2154.92 11588.25
```

b.2A mean crime rate last year is countries with high school graduation percentage X=80 is 6871.585 with Prediction between 2154.92 and 11588.25

```
linear1$resid[10]
```

```
##       10
## 1401.566
```

b.3 point estimate of 10th residual is 1401.566

```
var(data1$Percentage, data1$Crime)
```

```
## [1] -6601.544
```

Point estimate of variance is -6601.544

##Problem 4

```r
data3= read.table("http://www.cnachtsheim-text.csom.umn.edu/Kutner/Chapter%20%201%20Data%20Sets/CH01PR4
```

```r
colnames(data3)[1] ="Tyerror"
colnames(data3)[2]="Manuscript"
head(data3)
```

```
##   Tyerror Manuscript
## 1     128          7
## 2     213         12
## 3      75          4
## 4     250         14
## 5     446         25
## 6     540         30
```

```r
linear2<-lm(Tyerror ~ Manuscript ,data = data3)
linear2
```

```
##
## Call:
## lm(formula = Tyerror ~ Manuscript, data = data3)
##
## Coefficients:
## (Intercept)   Manuscript
##       1.597       17.852
```

```r
summary(linear2)
```

```
##
## Call:
## lm(formula = Tyerror ~ Manuscript, data = data3)
##
## Residuals:
##      1      2      3      4      5      6
##  1.436 -2.825  1.994 -1.530 -1.906  2.832
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.5969     2.0828   0.767    0.486
## Manuscript   17.8524     0.1161 153.727 1.07e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.646 on 4 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998
## F-statistic: 2.363e+04 on 1 and 4 DF,  p-value: 1.074e-08
```

```r
typolikelihood <- function(Y, X, beta1) {
likelihood <- c()
for (i in 1:length(Y)) {
  likelihood[i] <- (1/(sqrt(32 * pi))) * exp(-(1/32) * ((Y[i] - beta1 *X[i])^2))
}
likefunc <- prod(likelihood)
```

```
return(likefunc)
}
```

```
Yi <- c(128, 213, 75, 250, 446, 540)
Xi <- c(7, 12, 4, 14, 25, 30)
typolikelihood(Yi, Xi, 17)
```

```
## [1] 9.45133e-30
```

Likelihood function for b1=17 is 9.45133e-30

```
typolikelihood(Yi, Xi, 18)
```

```
## [1] 2.649043e-07
```

Likelihood function for b1=18 is 2.649043e-07

```
typolikelihood(Yi, Xi, 19)
```

```
## [1] 3.047285e-37
```

Likelihood function for b1=19 is 3.047285e-37

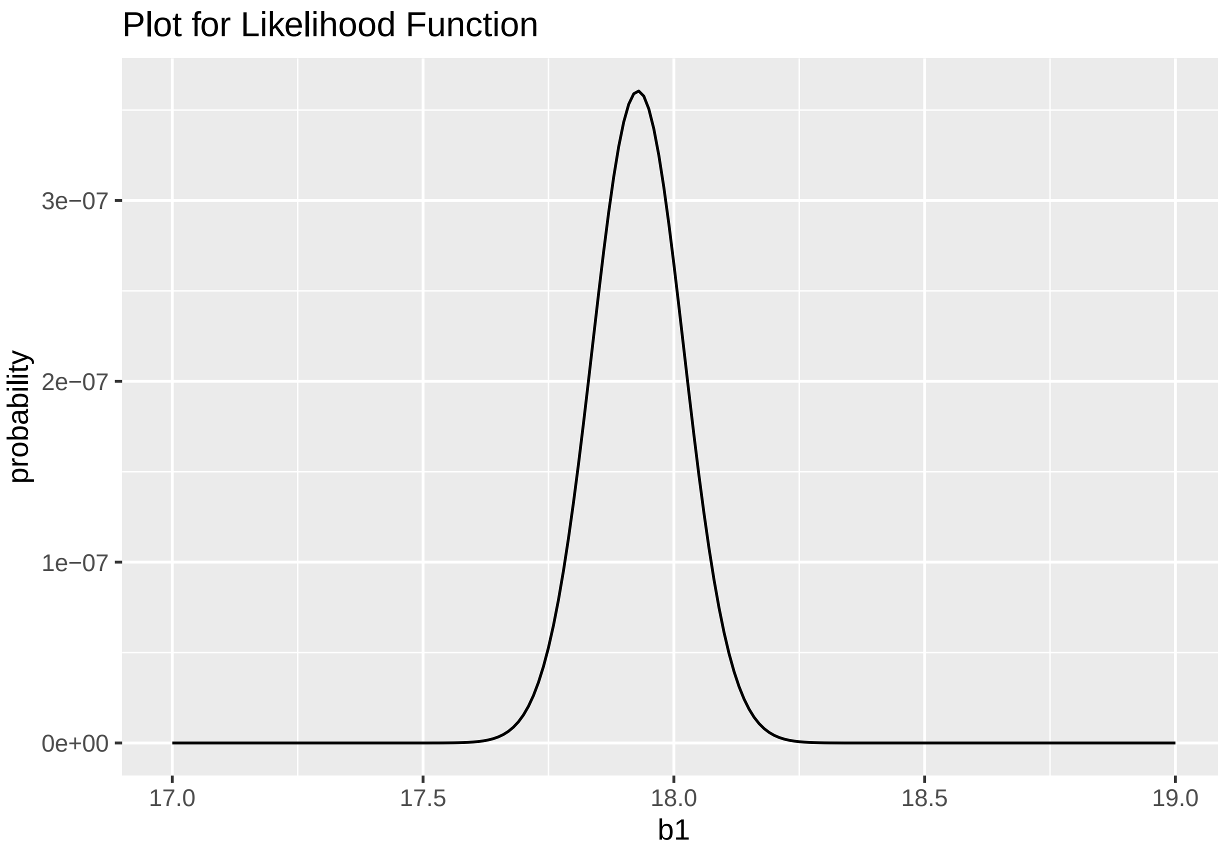
4.b Likelihood function of b1=17 is the highest

```
Sum = 0
res= 0
for ( idx in 1:length(Yi)){
    Sum = Sum + (Xi[idx]*Yi[idx])
    res = res + Xi[idx]*Xi[idx]
}
output = Sum/res
output
```

```
## [1] 17.9285
```

4.c MLE=17.9285

```
library(ggplot2)
b1 <- seq(17, 19, by = 0.01)
typopdf <- c()
for (i in 1:length(b1)) {
  typopdf[i] <- typolikelihood(Yi, Xi, b1[i])
}
typopdf <- data.frame(b1, typopdf)
colnames(typopdf) <- c("b1", "probability")
qplot(b1, probability, data = typopdf, geom = "line") + labs(title = "Plot for Likelihood Function")
```



4.d Yes, MLE seems maximized correspond to the maximum likelihood estimate in part(c), You can clearly see that it's maximized near by 18.0

PROBLEM 3 :-

Let y_1', y_2'' be observations of Y at $X = 5$, y_2' y_2'' be observations of Y at $X = 10$ and y_3' y_3'' be observations of Y at $X = 15$.

$$y = \beta_1 X + \beta_0 + \varepsilon$$

where β_1 β_0 are unknown parameters and ε is error term

By using least square regression method,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \frac{-5(y' - \bar{y}) - 5(y_1'' - \bar{y}) + 5(y_3' - \bar{y}) + 5(y_3'' - \bar{y})}{25 + 25 + 25 + 25}$$

$$= \frac{-5(y_1' + y_1'') + 5(y_3' + y_3'')}{100}$$

$$= \frac{-(y_1' + y_1'') + (y_3' + y_3'')}{20}$$

$$= \frac{2 \times 5 + 2 \times 10 + 2 \times 15}{6} = 10.$$

$$\bar{y} = \frac{y_1' + y_1'' + y_2' + y_2'' + y_3' + y_3''}{6}$$

a) If we have three points $(5, \bar{Y})$ $(10, \bar{-1})$ $(15, \bar{Y})$

$$y = \beta_1' x + \beta_0' + \varepsilon'$$

$$\bar{x} = \frac{5 + 10 + 15}{3} = 10.$$

$$\bar{y} = \frac{\frac{y_1' + y_1''}{2} + \frac{y_2' + y_2''}{2} + \frac{y_3' + y_3''}{2.}}{3}$$

$$= \frac{y_1' + y_1'' + y_2' + y_2'' + y_3' + y_3''}{6}$$

$$\hat{\beta}_1 = \frac{-5(\bar{Y} - \bar{y}) + 5(\bar{Y}_3 - \bar{y})}{25 + 25}$$

$$= \frac{-5 \times \bar{Y}_1 + 5 \times \bar{Y}_3}{50.}$$

$$= \frac{-\bar{Y}_1 + \bar{Y}_3}{10}$$

$$\overline{Y_1} = \frac{y_1' + y_1''}{2}.$$

$$\overline{Y_3} = \frac{y_3' + y_3''}{2}.$$

then,
$$\hat{\beta}_1 = \frac{-\left(\frac{y_1'+y_1''}{2}\right) + \left(\frac{y_3'+y_3''}{2}\right)}{10}.$$

$$= \frac{-(y_1'+y_1'') + (y_3'+y_3'')}{20}.$$

$$= \hat{\beta}_1$$

We also know

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ $\qquad$ $\hat{\beta}_0' = \bar{y} - \beta_1 \bar{x}$

Because $\bar{x}, \bar{y}$ are same for both models,

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

$\therefore$ Least square estimates of both values are same.

b) Since the error variance without the fitting of a regression line is the variance of response error itself where response error ε_i is the difference of the observed and the predicted value that is

$$\varepsilon_i = y_i - \bar{y}$$

Now we know that mean of error responses is 0 which implies

$$Var(\varepsilon) = \frac{1}{n-1}\sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i-\bar{y})^2$$

$= \text{Variance of } Y.$

Problem 4

a) Likelihood function for the six y observations for $\sigma^2 = 16$ is

$$L(\beta_1) = \prod_{i=1}^{6} \frac{1}{\sqrt{32\pi}} \exp\left\{-\frac{1}{32}(Y_i - \beta_1 x_{i})^2\right\}$$

//

PROBLEM - 5

Prove that $\sum_{i=1}^{n} e_i x_i = 0$ and $\sum_{i=1}^{n} e_i = 0$

We know that linear regression equation is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Sum of Squared Error (SSE) is by squaring on both sides

$$S = (\varepsilon_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

minimizing S and deriving w.r.t to β_0 & β_1

$$\frac{d(S)}{d(\beta_0)} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{d(S)}{d(\beta_1)} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0.$$

WKT, $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$

From the above eq^n we can say.

$$\sum_{i=1}^{n} e_i = 0 \text{ and } \sum_{i=1}^{n} e_i x_i = 0$$

Problem 6 :

Prove that $\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2)$

WKT, Simple Linear Regression eq^n is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$y_i \to$ dependent variable , $x_i \to$ independent

β_0 & $\beta_1 \to$ unknown parameters

As $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

$E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ $Cov(\varepsilon_i, \varepsilon_j)$.

Sum of Squares of errors is

$$Q = \sum_{i=1}^{n} \varepsilon_i^2$$

$$= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Taking derivative of SSE we get

$$-2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1) = 0$$

$$\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$-2\sum_{i=1}^{n}(y_i - \beta_0 x_i - \beta_1)x_i = 0$$

$$\sum_{i=1}^{n} y_i x_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} y_i x_i = \hat{\beta}_0 \sum_{i=1}^{n} x$$

Substitute β_0 in the above equation

$$\sum_{i=1}^{n} y_i x_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} y_i x_i = (\bar{y} - \hat{\beta}_1 x)\sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

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

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

$$\therefore \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Find $E(\hat{\beta}_1)$ and $V(\hat{\beta}_1)$:

Estimate of β can be written as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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

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

$$E(\hat{\beta}_1) = E\left[\frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]$$

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

$$= \frac{\beta_0 \sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + \frac{\beta_1 \sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Using $\sum_{i=1}^{n}(x_i-\bar{x})=0$ in above eq^n

$$E(\hat{\beta_1}) = \frac{\beta_1 \sum_{i=1}^{n}(x_i-\bar{x})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\beta_1\left(\sum_{i=1}^{n}x_i^2 - \bar{x}\sum_{i=1}^{n}x_i\right)}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\beta_1\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \beta_1$$

Hence $E(\hat{\beta_1}) = \beta_1$ //

$$V(\hat{\beta_1}) = V\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]$$

$$= \left(\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^2 \sum_{i=1}^{n}(x_i-\bar{x})^2 V(y_i)$$

$$= \left(\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^2 \sum_{i=1}^{n}(x_i-\bar{x})^2 \sigma^2$$

$$V(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Therefore we can conclude that

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right) //.$$