Pima Indians Diabetes Database

Reference: https://www.kaggle.com/uciml/pima-indians-diabetes-database

Name: Solayman Hossain Emon

Student ID: 80744292

In [1]:
```
using CSV, DataFrames, Statistics, StatsBase, Random, Bootstrap, Gadfly, StatsPlots
```

In [2]:
```
diabetes= CSV.read("diabetes.csv", DataFrame);
```

In [3]:
```
Random.seed!(20210909)
```

Out[3]:
```
TaskLocalRNG()
```

In [4]:
```
sz = size(diabetes)
```

Out[4]:
```
(768, 9)
```

In [5]:
```
fnames = names(diabetes)
```

Out[5]:
```
9-element Vector{String}:
 "Pregnancies"
 "Glucose"
 "BloodPressure"
 "SkinThickness"
 "Insulin"
 "BMI"
 "DiabetesPedigreeFunction"
 "Age"
 "Outcome"
```

In [6]:
```
describe(diabetes)
```

Out[6]: 9 rows × 7 columns

	variable	mean	min	median	max	nmissing	eltype
	Symbol	Float64	Real	Float64	Real	Int64	DataType
1	Pregnancies	3.84505	0	3.0	17	0	Int64
2	Glucose	120.895	0	117.0	199	0	Int64
3	BloodPressure	69.1055	0	72.0	122	0	Int64
4	SkinThickness	20.5365	0	23.0	99	0	Int64
5	Insulin	79.7995	0	30.5	846	0	Int64
6	BMI	31.9926	0.0	32.0	67.1	0	Float64
7	DiabetesPedigreeFunction	0.471876	0.078	0.3725	2.42	0	Float64
8	Age	33.2409	21	29.0	81	0	Int64
9	Outcome	0.348958	0	0.0	1	0	Int64

In [7]:
```
x = diabetes[!,:Pregnancies]
n = length(x)
```

Out[7]: 768

In [8]:
```
n = length(x)
```

Out[8]: 768

1. Test Statistic

T-Test

In [9]:
```
using HypothesisTests
```

```
x=diabetes[!,:Pregnancies]
t1 = OneSampleTTest(x, 8)
```

Out[9]:
```
One sample t-test
-----------------
Population details:
    parameter of interest:   Mean
    value under h_0:         8
    point estimate:          3.84505
    95% confidence interval: (3.606, 4.084)

Test summary:
    outcome with 95% confidence: reject h_0
    two-sided p-value:           <1e-99

Details:
    number of observations:   768
    t-statistic:              -34.17202159111878
    degrees of freedom:       767
    empirical standard error: 0.12158917509716566
```

In [10]:
```
pvalue(t1, tail = :both)
```

Out[10]:
```
2.9446570372756663e-156
```

In [11]:
```
pvalue(t1, tail = :left)
```

Out[11]:
```
1.4723285186378331e-156
```

In [12]:
```
pvalue(t1, tail = :right)
```

Out[12]:
```
1.0
```

- We can't conclude that the mean of the population from which the sample was drawn is greater than 8 as we haven't enough evidence to claim that. It is also verified by the Hypothesis Testing *(T-test)*

2. Bootstraping to find the sampling distribution

In [13]:
```
B = 500
x_bar = mean(x)
BMean = zeros(B)
tobs=zeros(B)
atobs=x_bar/(std(x)/sqrt(n))

for i in 1:B
    data = sample(x, n, replace=true, ordered=false)
    BMean[i] = mean(data)
    tobs[i] = mean(data)/(std(data)/sqrt(n))
end
```

In [18]:
```
# confidence interval
cp = percentile(BMean, [2.5,97.5])

# p-value
mean(BMean .>= x_bar)
```

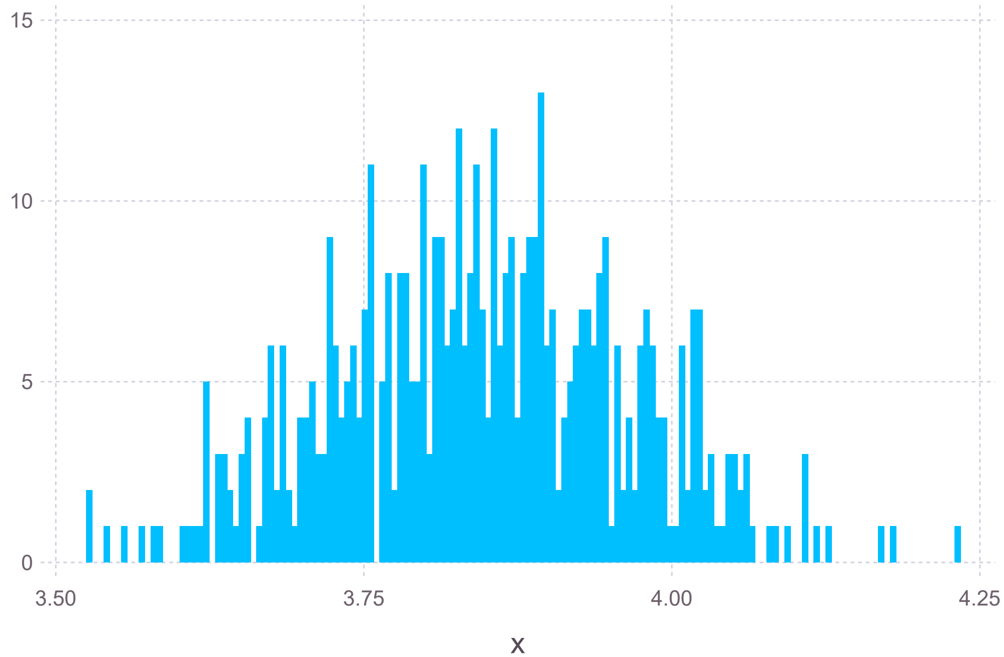
Out[18]:
```
0.504
```

In [19]:
```
mean(tobs .>= atobs)
```

Out[19]:
```
0.536
```

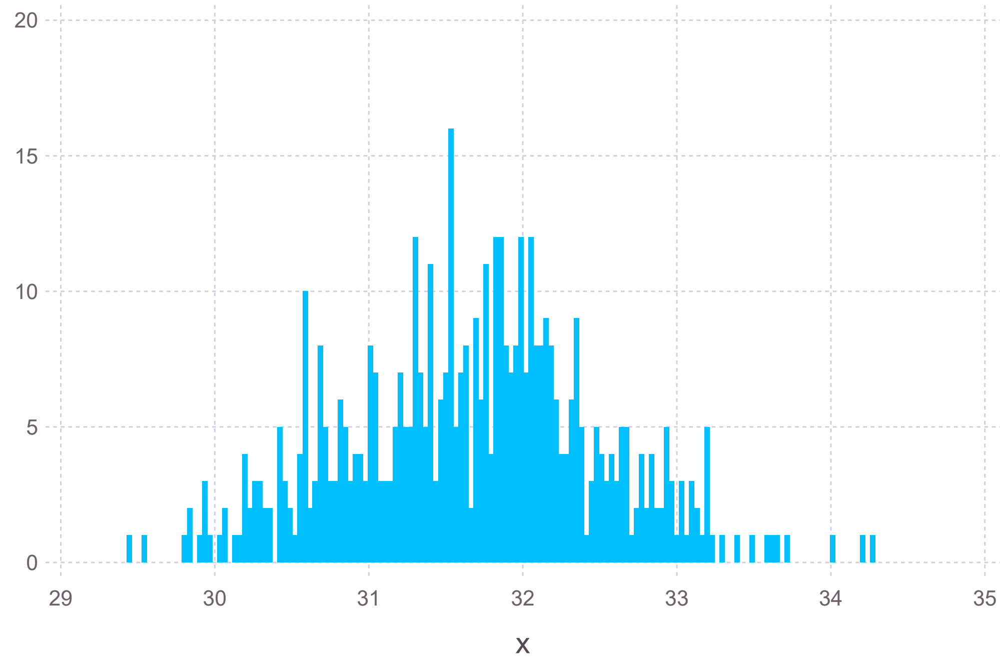
In [17]:
```
Gadfly.plot(x=BMean, Geom.histogram)
```

Out[17]:



In [20]:
```
Gadfly.plot(x=tobs,Geom.histogram)
```

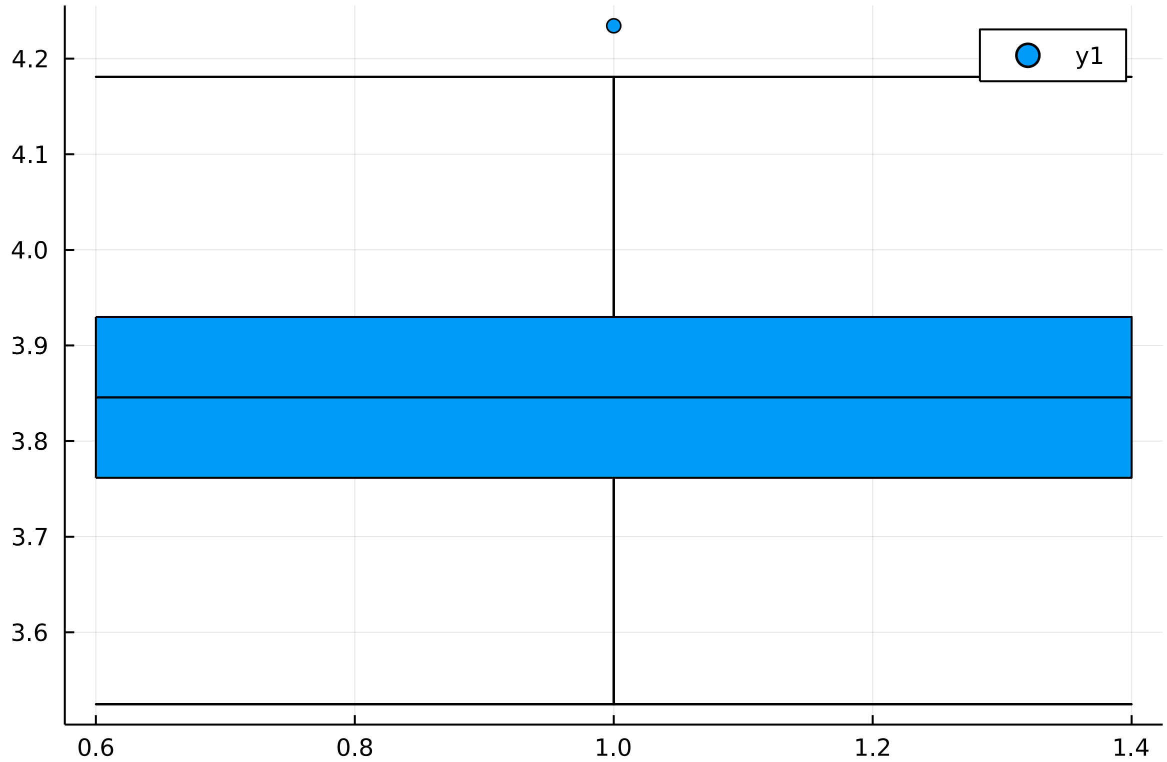
Out[20]:



In [21]:

```
StatsPlots.boxplot(BMean)
```

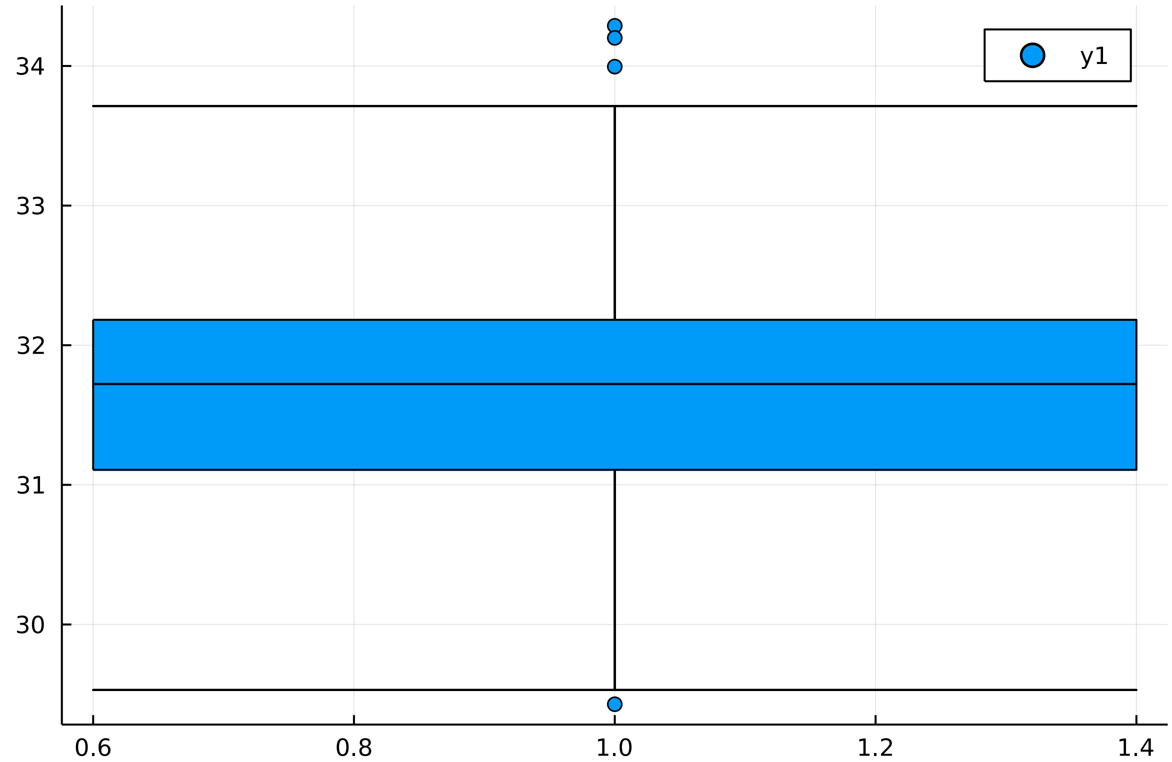
Out[21]:



In [22]:

```
StatsPlots.boxplot(tobs)
```

Out[22]:



Bootstrap Mean

In [23]:
```
mean(BMean)
```

Out[23]:
```
3.8460625
```

Sample Mean

In [24]:
```
x_bar = mean(x)
```

Out[24]:
```
3.8450520833333335
```

3. Claim about the Proportion

In [25]:
```
out = diabetes[diabetes.Outcome .== 1, :Outcome]
typeof(out)
```

Out[25]:
```
Vector{Int64} (alias for Array{Int64, 1})
```

In [26]:
```
samp_prop = length(out)/n
```

Out[26]:
```
0.3489583333333333
```

Z-test

In [27]:
```
assm_prop = 0.5
nume = samp_prop - assm_prop
deno = assm_prop*(1-assm_prop)
Z_test = nume / sqrt(deno/length(out))
```

Out[27]:
```
-4.945317299672939
```

- By observing the *Z-Test*, we have enough evidence to claim that the proportion of diabetes women in the population is different from 0.5

4. Repeat Procedure for the "Outcome" Variable

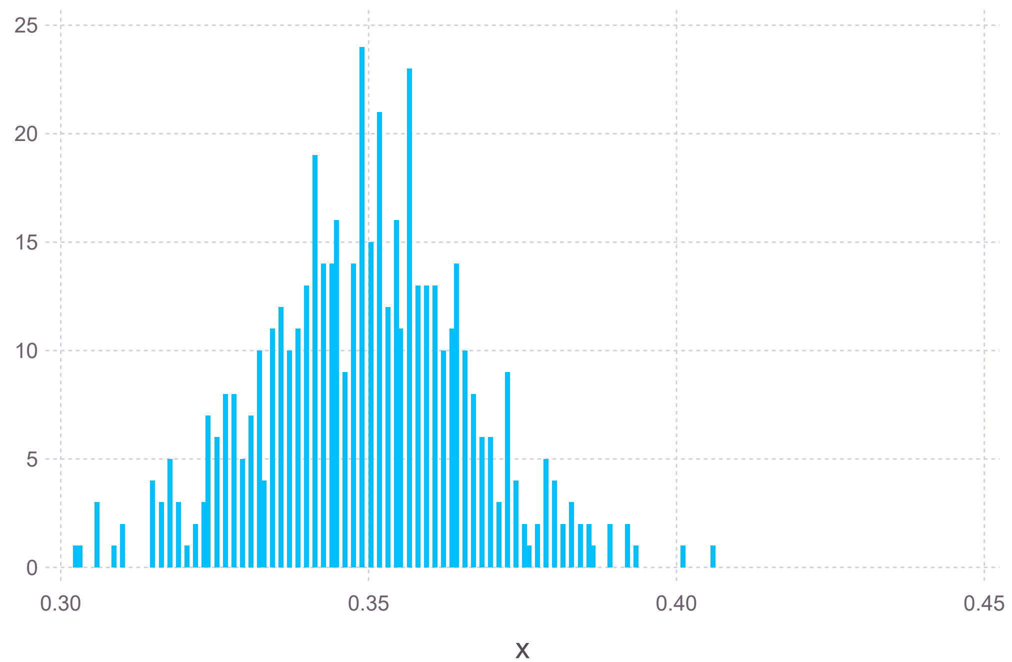
In [28]:
```
Outcome = diabetes[!,:Outcome]

B = 500
obs = zeros(B)
sel_obs = zeros(B)
prop = zeros(B)

for i in 1:B
    obs = sample(Outcome, n, replace=true, ordered=false)
    sel_obs = obs[obs .== 1]
    prop[i] = length(sel_obs) / n
end
```

In [29]:
```
prop
```

Out[29]:
```
500-element Vector{Float64}:
 0.328125
 0.359375
 0.35546875
 0.3203125
 0.3815104166666667
 0.359375
 0.3424479166666667
 0.3541666666666667
 0.3580729166666667
 0.3268229166666667
 0.3385416666666667
 0.3463541666666667
 0.3658854166666667
 ⋮
 0.3190104166666667
 0.3619791666666667
 0.3515625
 0.3606770833333333
 0.3541666666666667
 0.33203125
 0.3151041666666667
 0.3059895833333333
 0.34375
 0.3541666666666667
 0.3658854166666667
 0.3502604166666667
```

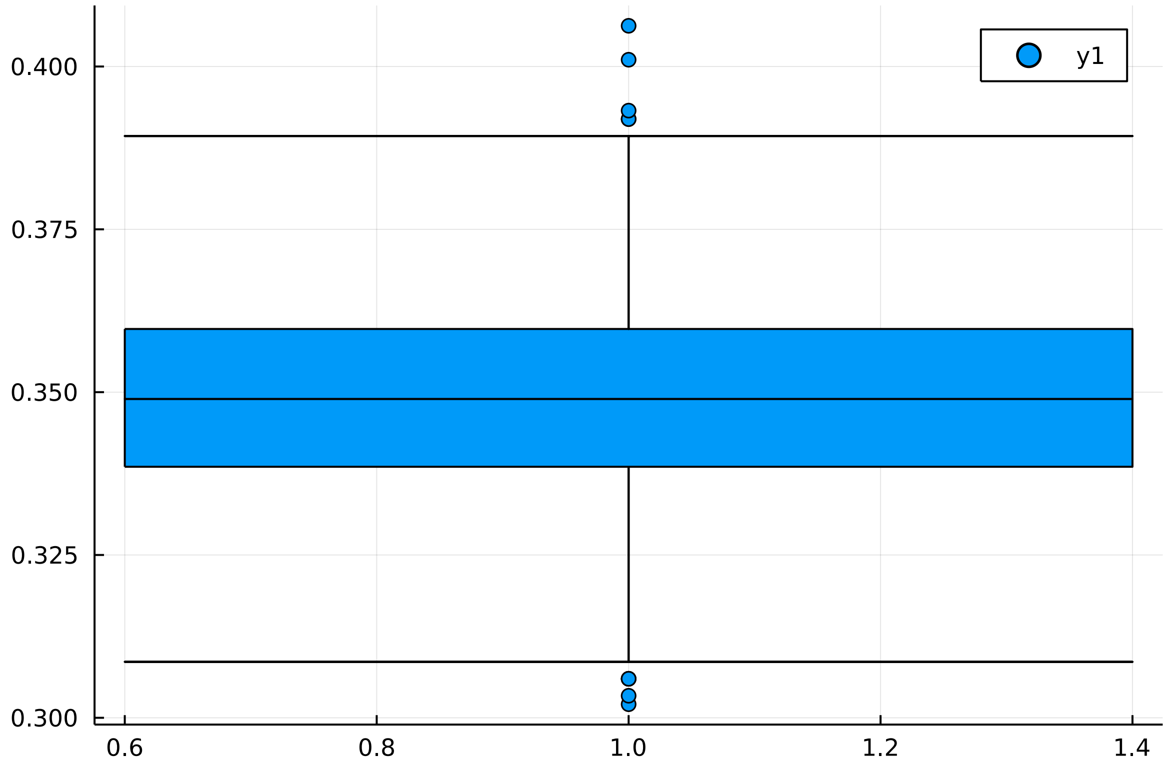
In [30]:
```
Gadfly.plot(x=prop, Geom.histogram)
```

Out[30]:



In [31]:
```
StatsPlots.boxplot(prop)
```

Out[31]:



In [32]:

```
assm_prop = 0.5
nume = mean(prop) - assm_prop
deno = assm_prop*(1-assm_prop)
Z_test = nume / sqrt(deno/n)
```

Out[32]: -8.348773567616584

In [33]:

```
mean(prop)
# confidence interval
cp=percentile(prop, [2.5,97.5])
# p-value
mean(prop .>= samp_prop)
```

Out[33]: 0.546

{Solayman Hossain Emon}

(II)

observe data (Y_i, X_i, Z_i), $i = 1, \dots, n$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \epsilon_i, \quad i = 1, \dots, n$$

$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 Z_1 + \epsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_2 + \beta_2 Z_2 + \epsilon_2$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$Y_n = \beta_0 + \beta_1 X_n + \beta_2 Z_n + \epsilon_n$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \epsilon_i$$

$$\Rightarrow \epsilon_i = Y_i - \beta_0 - \beta_1 X_i - \beta_2 Z_i$$

$$\Rightarrow \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_i - \beta_2 Z_i)^2$$

Using least squares method we get,

$$\frac{\partial}{\partial \beta_0} \left(\sum_{i=1}^{n} \epsilon_i^2 \right) = 0$$

$$\Rightarrow 2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_i - \beta_2 Z_i)(-1) = 0$$

Scanned with CamScanner

$$\Rightarrow \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 X_i - \beta_2 Z_i\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i + \beta_2 \sum_{i=1}^{n} Z_i$$

$$\Rightarrow \beta_0 = \frac{1}{n} \sum_{i=1}^{n} y_i - \frac{\beta_1}{n} \sum_{i=1}^{n} X_i - \frac{\beta_2}{n} \sum_{i=1}^{n} Z_i \qquad \text{——— (i)}$$

Similarly,

$$\frac{\partial}{\partial \beta_1} \left(\sum_{i=1}^{n} \epsilon_i^2 \right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} X_i \left(y_i - \beta_0 - \beta_1 X_i - \beta_2 Z_i\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2 + \beta_2 \sum_{i=1}^{n} X_i Z_i$$

$$\Rightarrow \sum_{i=1}^{n} X_i Y_i = \left(\frac{1}{n} \sum_{i=1}^{n} y_i - \frac{\beta_1}{n} \sum_{i=1}^{n} X_i - \frac{\beta_2}{n} \sum_{i=1}^{n} Z_i \right) \sum X_i + \beta_1 \sum_{i=1}^{n} X_i^2 + \beta_2 \sum_{i=1}^{n} X_i Z_i$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right) - \frac{\beta_1}{n}\left(\sum x_i\right)^2 - \frac{\beta_2}{n}\left(\sum x_i\right)\left(\sum z_i\right) + \beta_1 \sum_{i=1}^{n} x_i^2 + \beta_2 \sum_{i=1}^{n} x_i z_i$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = \beta_1\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right) + \beta_2\left(\sum_{i=1}^{n} x_i z_i - \frac{1}{n}\left(\sum(x_i)\left(\sum z_i\right)\right)\right)$$

$$\Rightarrow \hat{\beta}_1 = \frac{\left\{\sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)\right\} - \beta_2\left\{\sum x_i z_i - \left(\sum x_i\right)\left(\sum z_i\right)\right\}}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \quad \text{——— (B)}$$

So, From equation (i) we get.

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1/n\left(\sum x_i\right)\left\{\sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)\right\}}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} + \frac{\beta_2/n \sum_{i=1}^{n} x_i\left\{\sum x_i z_i - \left(\sum x_i\right)\left(\sum z_i\right)\right\}}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2} - \frac{\beta_2}{n}\sum z_i \quad \text{——— (A)}$$

Again,

$$\frac{\partial}{\partial \beta_2}\left(\sum \epsilon_i^2\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} z_i \left(y_i - \beta_0 - \beta_1 X_i - \beta_2 z_i\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} z_i Y_i = \beta_0 \sum_{i=1}^{n} z_i + \beta_1 \sum_{i=1}^{n} X_i Z_i + \beta_2 \sum_{i=1}^{n} z_i^2$$

$$\Rightarrow \sum_{i=1}^{n} z_i Y_i = \frac{1}{n}\left(\sum Y_i\right)\left(z_i\right) - \frac{1/n \left(\sum X_i\right)\left(\sum z_i\right)\left\{\sum X_i Y_i - \frac{1}{n}\sum X_i \sum Y_i\right\}}{\sum x_i^2 - \frac{1}{n}\left(\sum X_i\right)^2}$$

$$+ \frac{\beta_2/n \left(\sum X_i\right)\left(\sum z_i\right)\left\{\sum x_i z_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum z_i\right)\right\}}{\sum X_i^2 - \frac{1}{n}\left(\sum X_i\right)^2}$$

$$- \frac{\beta_2}{n}\left(\sum z_i\right)^2 + \frac{\left(\sum X_i z_i\right)\left\{\sum x_i Y_i - 1/n \left(\sum x_i\right)\left(\sum Y_i\right)\right\}}{\sum_{i=1}^{n} X_i^2 - 1/n \left(\sum X_i\right)^2}$$

$$- \frac{\beta_2 \left(\sum X_i z_i\right)\left\{\sum x_i z_i - \left(\sum X_i\right)\left(\sum z_i\right)\right\}}{\sum x_i^2 - 1/n \left(\sum x_i\right)^2}$$

$$+ \beta_2 \sum z_i^2$$

Scanned with CamScanner

$$\Rightarrow \sum_{i=1}^{n} z_i y_i - n\bar{y}\bar{z} = -\frac{n\bar{x}\bar{z}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)}{\sum x_i^2 - n\bar{x}^2}$$

$$+ \frac{n\beta_2 \bar{x}\bar{z}\left(\sum x_i z_i - n\bar{x}\bar{z}\right)}{\sum x_i^2 - n\bar{x}^2} - n\beta_2 \bar{z}^2$$

$$+ \frac{\sum x_i z_i \left(\sum x_i y_i - n\bar{x}\bar{y}\right)}{\sum x_i^2 - n\bar{x}^2}$$

$$- \frac{\beta_2 \sum x_i z_i \left(\sum x_i z_i - n\bar{x}\bar{z}\right)}{\sum x_i^2 - n\bar{x}^2}$$

$$+ \beta_2 \sum z_i^2$$

$$\Rightarrow \sum_{i=1}^{n} z_i y_i - n\bar{y}\bar{z} = \frac{\left(\sum x_i z_i - n\bar{x}\bar{z}\right)\left(\sum x_i y_i - n\bar{x}\bar{y}\right)}{\sum x_i^2 - n\bar{x}^2}$$

$$- \frac{\beta_2 \left(\sum x_i z_i - n\bar{x}\bar{z}\right)^2}{\sum x_i^2 - n\bar{x}^2}$$

$$+ \beta_2 \left(\sum z_i^2 - n\bar{z}^2\right)$$

$$\Rightarrow \hat{\beta}_2 = \frac{(\sum z_i y_i - n\bar{y}\bar{z}) - \left\{ \frac{(\sum x_i z_i - n\bar{x}\bar{y})(\sum x_i y_i - n\bar{x}\bar{y})}{\sum x_i^2 - n\bar{x}^2} \right\}}{\left\{ (\sum z_i^2 - n\bar{z}^2) - (\sum x_i z_i - n\bar{x}\bar{z})^2 \right\}} \quad \text{--- (C)}$$

Here, Equation (A), (B), (C) are the required expression for $\hat{\beta}_0$, $\hat{\beta}_1$ & $\hat{\beta}_2$.

Given that,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i$$

$$\sum \epsilon_i^2$$

$$= \sum_{i=1}^{n} \left\{ Y_i - (\beta_0 + \beta_1 X_i + \beta_2 Z_i) \right\}$$

$$= 0$$

Since algebric sum of Deviation from mean is zero.

Similarly,

$$\sum_{i=1}^{n} \hat{\epsilon}_i$$

$$= \sum_{i=1}^{n} \left\{ y_i - \left(\hat{\beta}_0 + \beta_1 X_i + \beta_2 Z_i \right) \right\}$$

$$= 0$$

Therefore, sum of residuals and sum of errors are equal to zero is

[Showed].