

PROBLEM SET 3

Bui Quang Thai
(Thai Bui)

Problem 4

$$y_i = x_i' \beta + \varepsilon_i$$

$$z_i = (z_{i1}, z_{i2}) \quad k = k_1 + k_2$$

$$Z = (Z_1, Z_2)$$

$$P_{Z_1} = Z_1 (Z_1^T Z_1)^{-1} Z_1^T$$

$$\begin{aligned} \rightarrow P_{Z_1} Z_1 &= Z_1 (Z_1^T Z_1)^{-1} Z_1^T Z_1 \\ &= Z_1 \end{aligned}$$

Regressing each column of Z_1 on Z

$$Z_1 = Z\delta + \eta$$

$$\rightarrow \hat{Z}_1 = P_Z Z_1$$

As $Z = (Z_1, Z_2)$, we know that

$$\hat{Z}_1 = Z_1$$

$$\rightarrow Z_1 = P_Z Z_1$$

$$\rightarrow P_Z Z_1 = P_Z P_Z Z_1$$

$$\rightarrow P_Z P_Z = P_Z$$

Under the assumption of correct specification & homoscedasticity, $E[\varepsilon_i | z_i] = 0$ & $E[\varepsilon_i^2 | z_i] = \sigma_\varepsilon^2$

Regressing using Z_1, Z respectively to get the estimators $\hat{\beta}_1, \hat{\beta}_2$

$$\hat{\beta}_1 = (X^T P_{Z_1} X)^{-1} X^T P_{Z_1} Y$$

$$\hat{\beta}_2 = (X^T P_Z X)^{-1} X^T P_Z Y$$

$$E(\hat{\beta}_1) = E(\hat{\beta}_2) = \beta$$

We have the joint distribution

$$\sqrt{n} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \xrightarrow{d} N \left(\begin{bmatrix} \beta \\ \beta \end{bmatrix}, \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \right)$$



$$\begin{aligned} V_{11} &= \lim_{n \rightarrow \infty} E[(\hat{\beta}_1 - \beta)(\hat{\beta}_1 - \beta)^T] \\ &= \sigma_0^2 (X^T P_{Z_1} X)^{-1} X^T P_{Z_1} [(X^T P_{Z_1} X)^{-1} X^T P_{Z_1}]^T \\ &= \sigma_0^2 (X^T P_{Z_1} X)^{-1} X^T P_{Z_1} P_{Z_1} X (X^T P_{Z_1} X)^{-1} \\ &= \sigma_0^2 (X^T P_{Z_1} X)^{-1} \quad (P_{Z_1}^2 = P_{Z_1}, P_{Z_1} = P_{Z_1}^T) \end{aligned}$$

Similarly,

$$V_{12} = \sigma_0^2 (X^T P_{Z_1} X)^{-1} X^T P_{Z_1} [(X^T P_{Z_2} X)^{-1} X^T P_{Z_2}]^T$$

$$V_{21} = \sigma_0^2 (X^T P_{Z_2} X)^{-1} X^T P_{Z_2} [(X^T P_{Z_1} X)^{-1} X^T P_{Z_1}]^T$$

$$V_{22} = \sigma_0^2 (X^T P_{Z_2} X)^{-1}$$

$$\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = \text{Cov}(\hat{\beta}_2, \hat{\beta}_1)$$

$$\rightarrow V_{12} = V_{21}$$

$$V_{21} = \sigma_0^2 (X^T P_{Z_2} X)^{-1} X^T P_{Z_2} P_{Z_1} X (X^T P_{Z_1} X)^{-1}$$

$$P_{Z_2} P_{Z_1} = P_{Z_1}$$

$$\rightarrow V_{21} = \sigma_0^2 (X^T P_{Z_2} X)^{-1} = V_{22}$$

$$\rightarrow V_{22} - V_{12} - V_{21}$$

$$= V_{22} - V_{22} - V_{22} = -V_{22}$$

$$\rightarrow V_{22} - V_{12} - V_{21} = -V_{22}$$

$$\sqrt{n}(\hat{\beta}_1 - \hat{\beta}_2)$$

using delta method

$$\rightarrow \sqrt{n}(\hat{\beta}_1 - \hat{\beta}_2) \xrightarrow{d} N\left(0, \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix}\right)$$

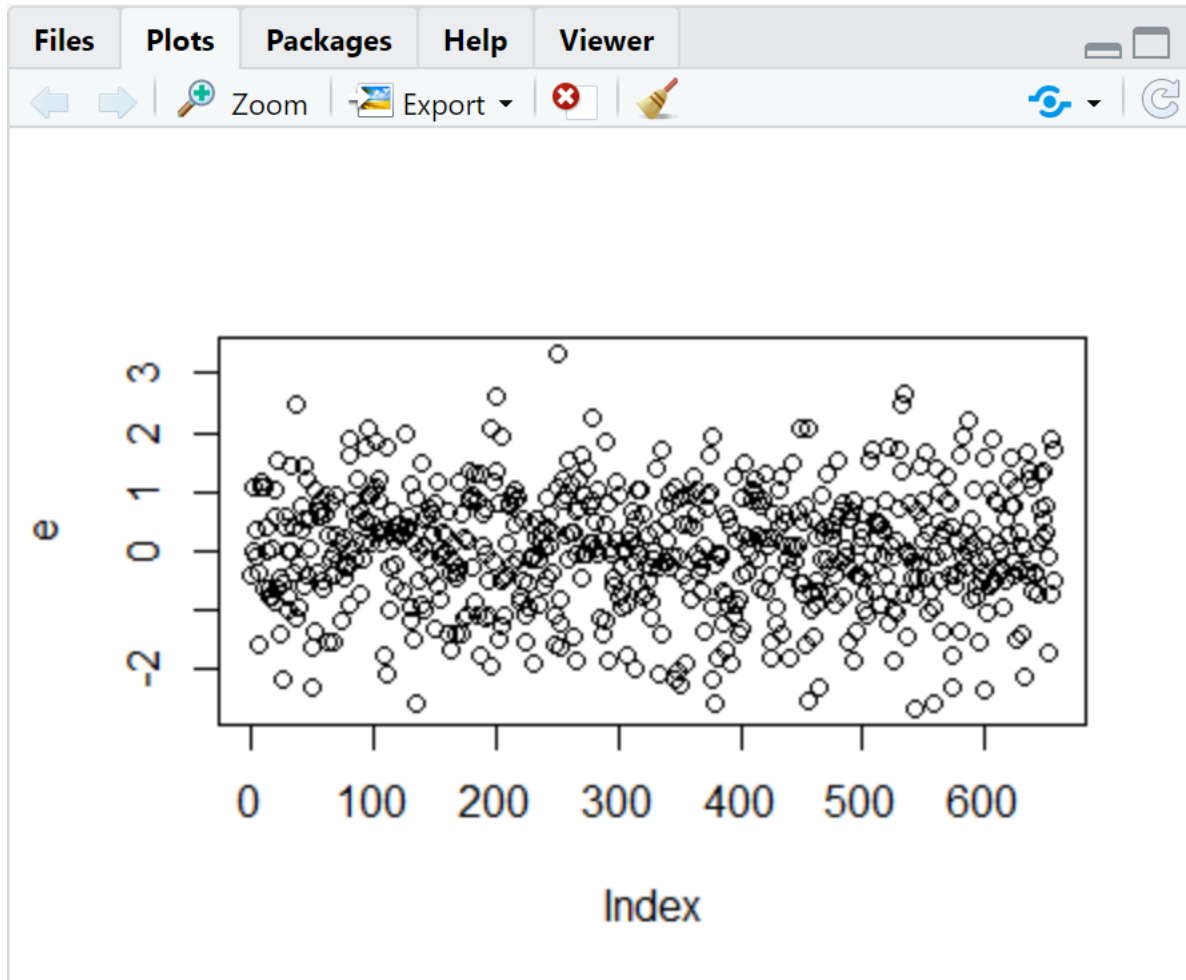
$$\begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$= V_{11} + V_{22} - V_{12} - V_{21} = V_{11} - V_{22}$$

$$\rightarrow \sqrt{n}(\hat{\beta}_1 - \hat{\beta}_2) \xrightarrow{d} N(0, V_{11} - V_{22})$$

Problem 1

```
> library(readxl)
> DataExample <- read_excel("C:/Users/quang/Downloads/DataExample.xls")
New names:
* month -> month...7
* ` ` -> ...10
* month -> month...15
* ` ` -> ...17
> View(DataExample)
> e <- rnorm(656)
> plot(e)
```



```
> y <- 0.01 + -0.2*DataExample$log Pg` + 1*DataExample$log(Income/Pop)`
+ 0.5*DataExample$log Pnc` + 0.4*DataExample$log Puc` + 0.02*Data
Example$month...15 + 0.9*DataExample$log Pd` + e
> model <- lm(y ~ DataExample$log Pg` + DataExample$log(Income/Pop)` + Data
Example$log Pnc` + DataExample$log Puc` + DataExample$month...15 + DataExam
ple$log Pd`)
> summary(model)
```

Call:
`lm(formula = y ~ DataExample$log Pg` + DataExample$log(Income/Pop)` +
DataExample$log Pnc` + DataExample$log Puc` + DataExample$month...15 +
DataExample$log Pd`)`

Residuals:

Min	1Q	Median	3Q	Max
-2.7225	-0.6413	0.0161	0.6806	3.3437

Coefficients:

	Estimate	Std. Error
(Intercept)	1.6947888	3.4579916
DataExample\$`log Pg`	-0.2626595	0.1344636
DataExample\$`log(Income/Pop)`	1.0293078	0.0432408
DataExample\$`log Pnc`	0.4635700	0.0847382
DataExample\$`log Puc`	0.1105878	0.6981711
DataExample\$month...15	0.0200343	0.0003582
DataExample\$`log Pd`	0.9652914	0.1681160

	t value	Pr(> t)
(Intercept)	0.490	0.6242
DataExample\$`log Pg`	-1.953	0.0512
DataExample\$`log(Income/Pop)`	23.804	< 2e-16 ***
DataExample\$`log Pnc`	5.471	6.41e-08 ***
DataExample\$`log Puc`	0.158	0.8742
DataExample\$month...15	55.938	< 2e-16 ***
DataExample\$`log Pd`	5.742	1.44e-08 ***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9831 on 649 degrees of freedom

Multiple R-squared: 0.9267, Adjusted R-squared: 0.926

F-statistic: 1368 on 6 and 649 DF, p-value: < 2.2e-16

> deviance(model)

[1] 627.2366

> summary(model)\$sigma^2

[1] 0.9664662

> vcov(model)

	(Intercept)
(Intercept)	11.9577060704
DataExample\$`log Pg`	-0.1581572684
DataExample\$`log(Income/Pop)`	-0.0050333678
DataExample\$`log Pnc`	-0.0409884510
DataExample\$`log Puc`	-2.3202112738
DataExample\$month...15	-0.0001956087
DataExample\$`log Pd`	-0.0870299849
	DataExample\$`log Pg`
(Intercept)	-1.581573e-01
DataExample\$`log Pg`	1.808047e-02
DataExample\$`log(Income/Pop)`	4.141999e-04
DataExample\$`log Pnc`	2.318386e-04
DataExample\$`log Puc`	1.079526e-02
DataExample\$month...15	2.414563e-05
DataExample\$`log Pd`	1.229190e-02
	DataExample\$`log(Income/Pop)`
(Intercept)	-5.033368e-03
DataExample\$`log Pg`	4.141999e-04
DataExample\$`log(Income/Pop)`	1.869766e-03
DataExample\$`log Pnc`	-1.021985e-04
DataExample\$`log Puc`	2.628604e-04
DataExample\$month...15	2.495750e-06
DataExample\$`log Pd`	4.640905e-04
	DataExample\$`log Pnc`
(Intercept)	-4.098845e-02
DataExample\$`log Pg`	2.318386e-04
DataExample\$`log(Income/Pop)`	-1.021985e-04
DataExample\$`log Pnc`	7.180564e-03
DataExample\$`log Puc`	9.670018e-04
DataExample\$month...15	-4.242139e-08
DataExample\$`log Pd`	-1.200947e-04

```

DataExample$`log Puc`
(Intercept) -2.320211e+00
DataExample$`log Pg` 1.079526e-02
DataExample$`log(Income/Pop)` 2.628604e-04
DataExample$`log Pnc` 9.670018e-04
DataExample$`log Puc` 4.874429e-01
DataExample$month...15 -7.130641e-06
DataExample$`log Pd` -6.154919e-03
DataExample$month...15
(Intercept) -1.956087e-04
DataExample$`log Pg` 2.414563e-05
DataExample$`log(Income/Pop)` 2.495750e-06
DataExample$`log Pnc` -4.242139e-08
DataExample$`log Puc` -7.130641e-06
DataExample$month...15 1.282756e-07
DataExample$`log Pd` 4.904605e-05
DataExample$`log Pd`
(Intercept) -8.702998e-02
DataExample$`log Pg` 1.229190e-02
DataExample$`log(Income/Pop)` 4.640905e-04
DataExample$`log Pnc` -1.200947e-04
DataExample$`log Puc` -6.154919e-03
DataExample$month...15 4.904605e-05
DataExample$`log Pd` 2.826299e-02
> hccm(model)

(Intercept) 13.0937999987
DataExample$`log Pg` -0.1850410433
DataExample$`log(Income/Pop)` 0.0074144070
DataExample$`log Pnc` -0.0446684736
DataExample$`log Puc` -2.5299477159
DataExample$month...15 -0.0002075919
DataExample$`log Pd` -0.1011301832
DataExample$`log Pg`
(Intercept) -1.850410e-01
DataExample$`log Pg` 1.877278e-02
DataExample$`log(Income/Pop)` 6.776260e-05
DataExample$`log Pnc` -2.113661e-05
DataExample$`log Puc` 1.629616e-02
DataExample$month...15 2.231919e-05
DataExample$`log Pd` 1.236422e-02
DataExample$`log(Income/Pop)`
(Intercept) 7.414407e-03
DataExample$`log Pg` 6.776260e-05
DataExample$`log(Income/Pop)` 2.210115e-03
DataExample$`log Pnc` -1.334565e-04
DataExample$`log Puc` -2.060593e-03
DataExample$month...15 2.173185e-06
DataExample$`log Pd` 3.999159e-04
DataExample$`log Pnc`
(Intercept) -4.466847e-02
DataExample$`log Pg` -2.113661e-05
DataExample$`log(Income/Pop)` -1.334565e-04
DataExample$`log Pnc` 7.139557e-03
DataExample$`log Puc` 2.683163e-03
DataExample$month...15 -3.673004e-06
DataExample$`log Pd` -1.242061e-03
DataExample$`log Puc`
(Intercept) -2.529948e+00
DataExample$`log Pg` 1.629616e-02
DataExample$`log(Income/Pop)` -2.060593e-03
DataExample$`log Pnc` 2.683163e-03
DataExample$`log Puc` 5.239525e-01
DataExample$month...15 7.256627e-07

```

```

DataExample$`log Pd` -2.817622e-03
DataExample$month...15 -2.075919e-04
(Intercept) -2.075919e-04
DataExample$`log Pg` 2.231919e-05
DataExample$`log(Income/Pop)` 2.173185e-06
DataExample$`log Pnc` -3.673004e-06
DataExample$`log Puc` 7.256627e-07
DataExample$month...15 1.256675e-07
DataExample$`log Pd` 5.031647e-05
DataExample$`log Pd` -1.011302e-01
(Intercept) -1.011302e-01
DataExample$`log Pg` 1.236422e-02
DataExample$`log(Income/Pop)` 3.999159e-04
DataExample$`log Pnc` -1.242061e-03
DataExample$`log Puc` -2.817622e-03
DataExample$month...15 5.031647e-05
DataExample$`log Pd` 3.016720e-02

```

Variables	Estimate	Std. Error	t value	Asy.Var.
(intercept)	1.6947888	3.4579916	0.490	13.093799999
log Pg	-0.2626595	0.1344636	-1.953	1.877278e-02
log (Income/Pop)	1.0293078	0.0432408	23.804	2.210115e-03
log Pnc	0.4635700	0.0847382	5.471	7.139557e-03
log Puc	0.1105878	0.6981711	0.158	5.239525e-01
Month	0.0200343	0.0003582	55.938	1.256675e-07
log Pd	0.9652914	0.1681160	5.742	3.016720e-02
RSS = 627.2366		Variance of Rediduals = 0.9664662		

```
> density(y, kernel=c("epanechnikov"))
```

Call:

```
density.default(x = y, kernel = c("epanechnikov"))
```

Data: y (656 obs.); Bandwidth 'bw' = 0.8891

```

      x              y
Min.   : 1.501   Min.   :0.000000
1st Qu.: 6.861   1st Qu.:0.008063
Median :12.222   Median :0.051316
Mean    :12.222   Mean    :0.046594
3rd Qu.:17.583   3rd Qu.:0.079077
Max.    :22.943   Max.    :0.096144

```

```
> density(y, kernel=c("gaussian"))
```

Call:

```
density.default(x = y, kernel = c("gaussian"))
```

Data: y (656 obs.); Bandwidth 'bw' = 0.8891

```

      x              y
Min.   : 1.501   Min.   :1.131e-05
1st Qu.: 6.861   1st Qu.:7.858e-03
Median :12.222   Median :5.157e-02
Mean    :12.222   Mean    :4.659e-02
3rd Qu.:17.583   3rd Qu.:8.078e-02
Max.    :22.943   Max.    :9.702e-02

```

```

> for (i in 1:100){
+   e <- rnorm(656)
+   y <- 0.01 + -0.2*DataExample$`log Pg` + 1*DataExample$`log(Income/Pop)`
+     0.5*DataExample$`log Pnc` + 0.4*DataExample$`log Puc` + 0.02*Data
Example$month...15 + 0.9*DataExample$`log Pd` + e
+   est[i,] <- lm(y ~ DataExample$`log Pg` + DataExample$`log(Income/Pop)`
+ DataExample$`log Pnc` + DataExample$`log Puc` + DataExample$month...15 + Da
taExample$`log Pd`)$coefficients
+ }

```

```
> var(est)
```

```

      [,1]      [,2]      [,3]
[1,] 11.7716905319 -2.076512e-01 -6.297044e-03
[2,] -0.2076511818  1.931247e-02  4.214658e-04
[3,] -0.0062970442  4.214658e-04  1.951705e-03
[4,] -0.0153773448 -2.003038e-03 -6.529582e-04
[5,] -2.2468604791  2.048256e-02  1.126025e-03
[6,] -0.0001469062  3.207822e-05  2.585473e-06
[7,] -0.1307373818  1.702383e-02  2.296389e-04
      [,4]      [,5]      [,6]
[1,] -1.537734e-02 -2.246860e+00 -1.469062e-04
[2,] -2.003038e-03  2.048256e-02  3.207822e-05
[3,] -6.529582e-04  1.126025e-03  2.585473e-06
[4,]  7.918649e-03 -2.204037e-03 -3.683855e-06
[5,] -2.204037e-03  4.648495e-01 -2.453666e-05
[6,] -3.683855e-06 -2.453666e-05  1.360384e-07
[7,] -2.102341e-03 -1.378751e-03  5.813064e-05
      [,7]
[1,] -1.307374e-01
[2,]  1.702383e-02
[3,]  2.296389e-04
[4,] -2.102341e-03
[5,] -1.378751e-03
[6,]  5.813064e-05
[7,]  3.341444e-02

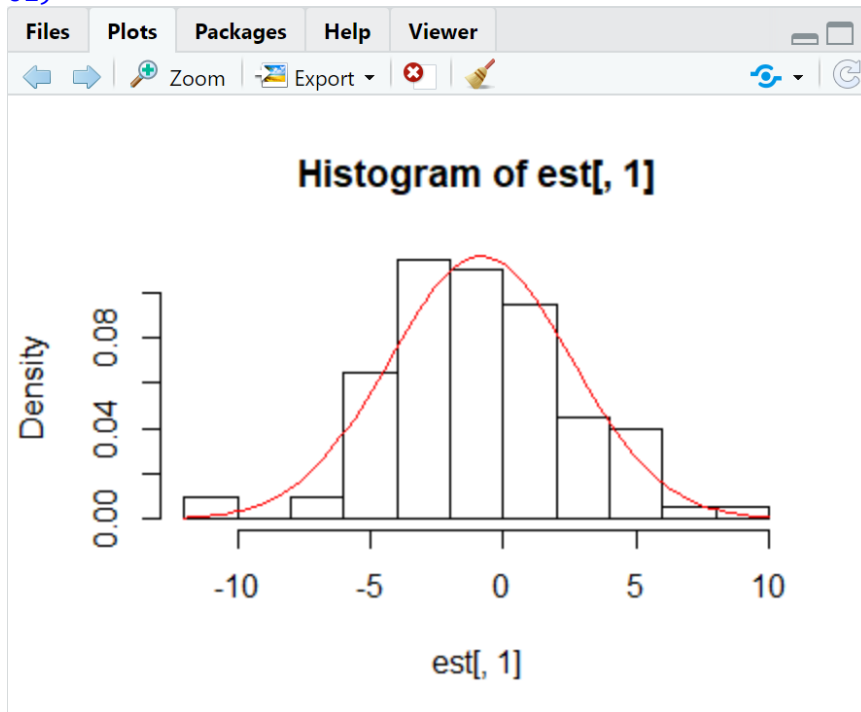
```

Variables	Cross-sample variance
(intercept)	11.7716905319
log Pg	1.931247e-02
log (Income/Pop)	1.951705e-03
log Pnc	7.918649e-03
log Puc	4.648495e-01
Month	1.360384e-07
log Pd	3.341444e-02

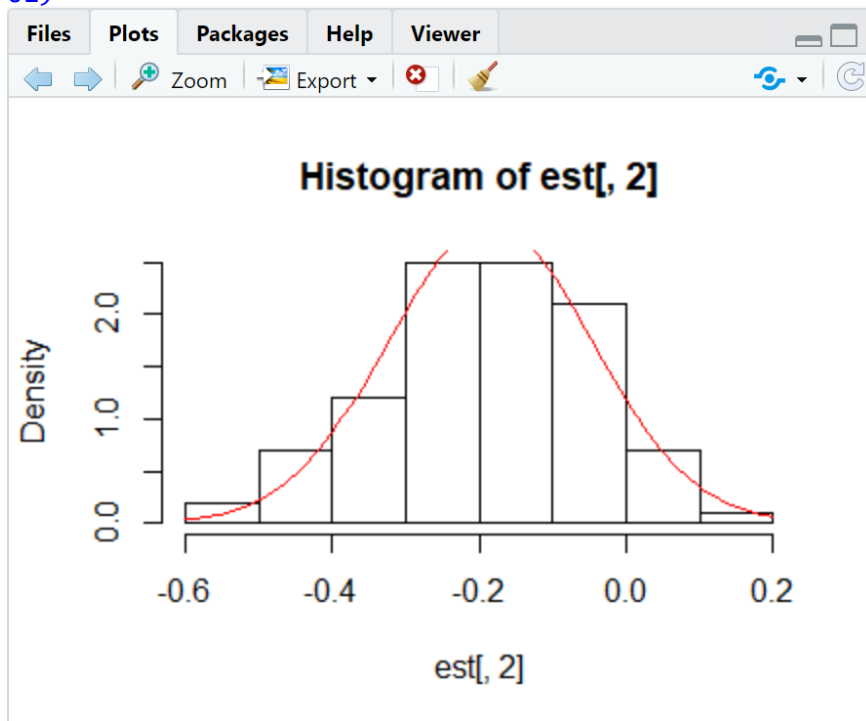
```

> hist(est[,1], prob=TRUE)
> curve(dnorm(x, mean=mean(est[,1]), sd=sqrt(var(est[,1]))),col="red", add=TRUE)

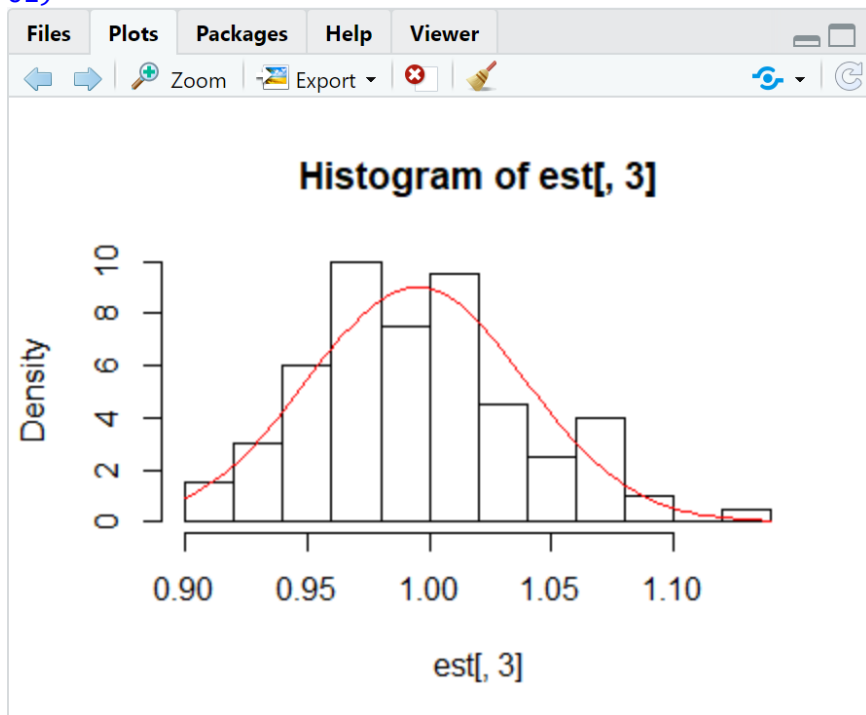
```



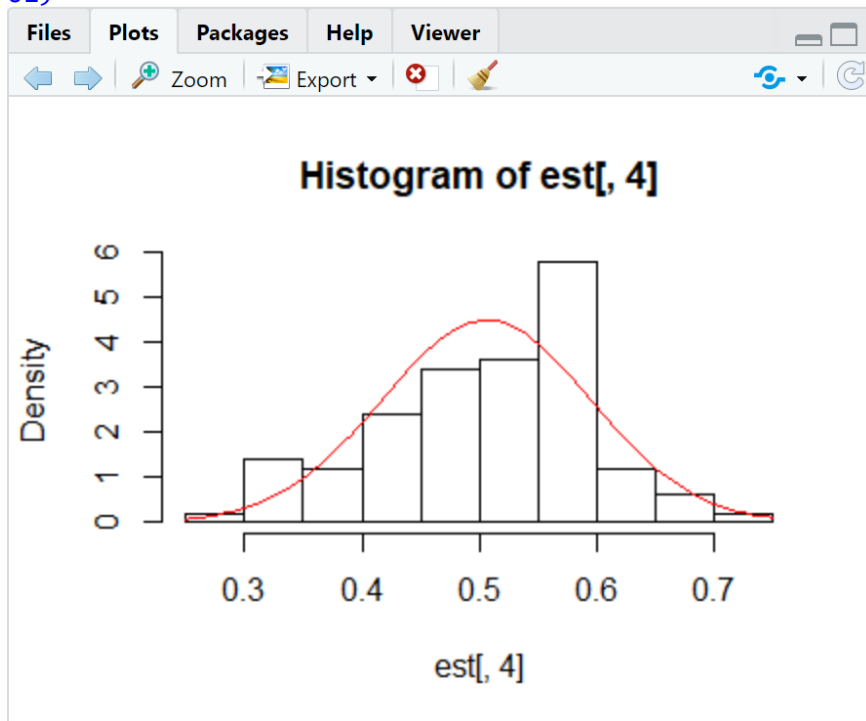
```
> hist(est[,2], prob=TRUE)
> curve(dnorm(x, mean=mean(est[,2]), sd=sqrt(var(est[,2]))),col="red", add=TRUE)
```



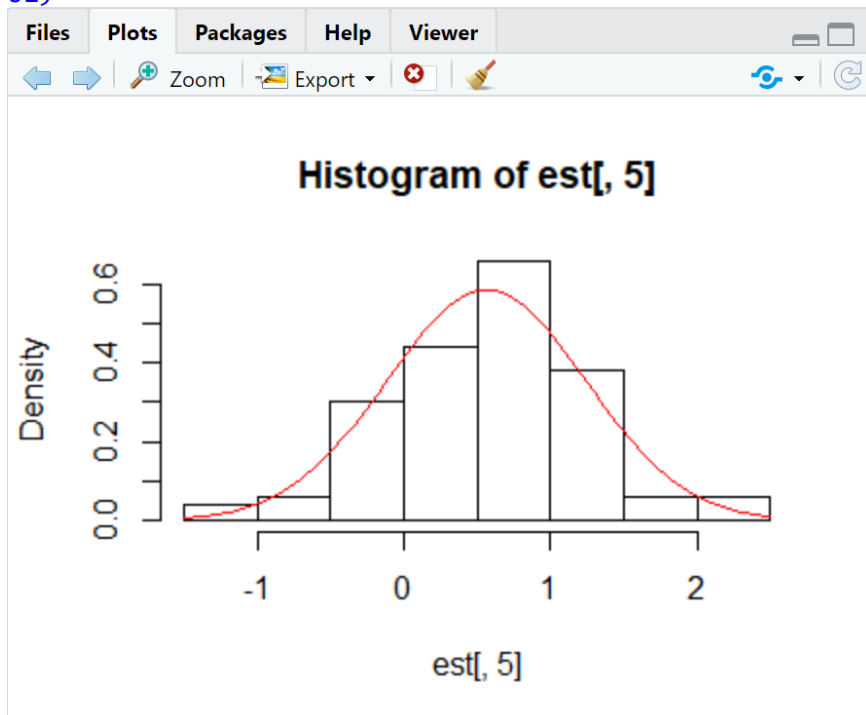
```
> hist(est[,3], prob=TRUE)
> curve(dnorm(x, mean=mean(est[,3]), sd=sqrt(var(est[,3]))),col="red", add=TRUE)
```



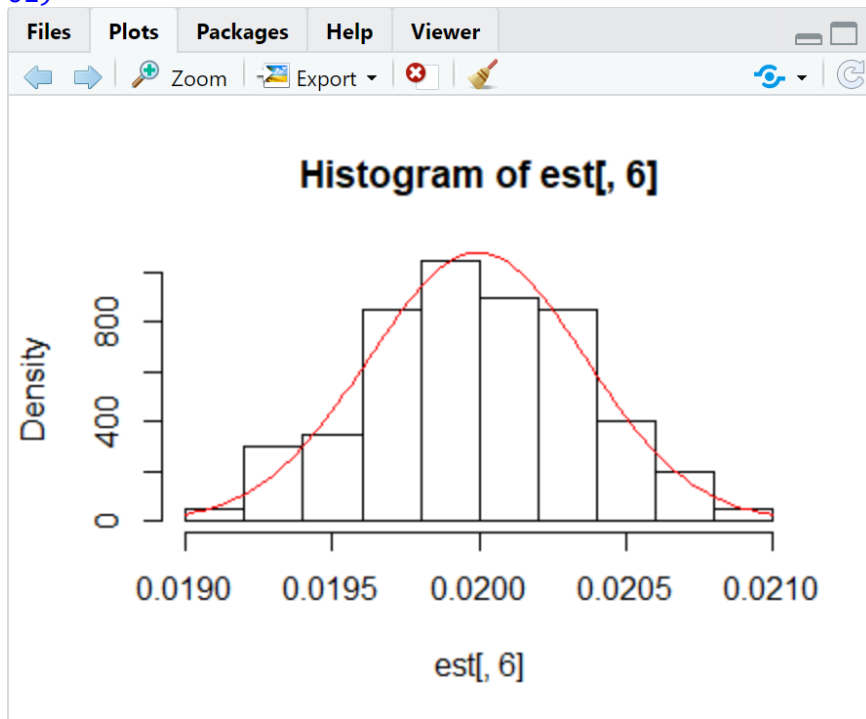
```
> hist(est[,4], prob=TRUE)
> curve(dnorm(x, mean=mean(est[,4]), sd=sqrt(var(est[,4]))),col="red", add=TRUE)
```



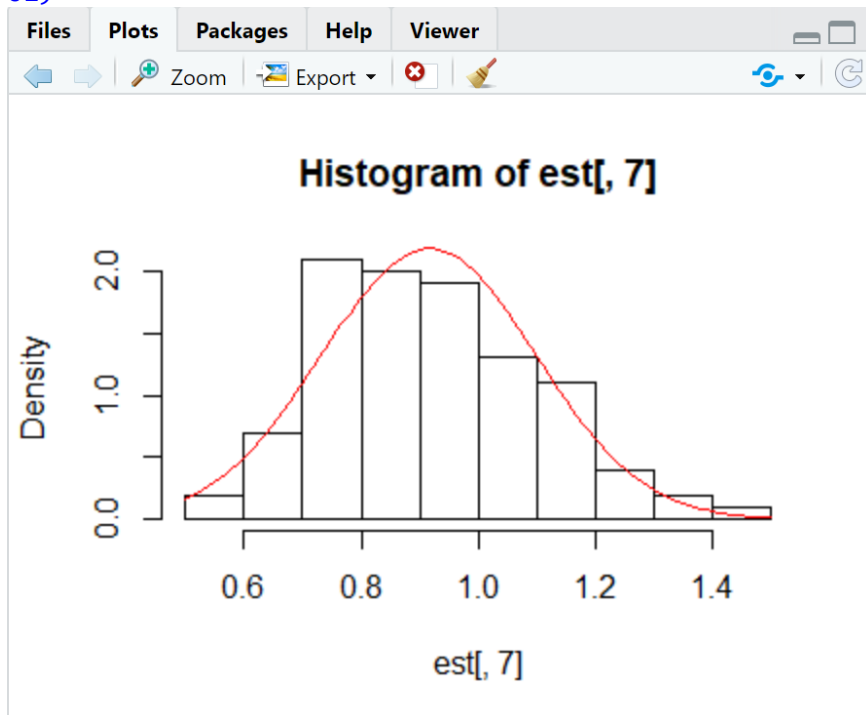
```
> hist(est[,5], prob=TRUE)
> curve(dnorm(x, mean=mean(est[,5]), sd=sqrt(var(est[,5]))),col="red", add=TRUE)
```



```
> hist(est[,6], prob=TRUE)
> curve(dnorm(x, mean=mean(est[,6]), sd=sqrt(var(est[,6]))),col="red", add=TRUE)
```



```
> hist(est[,7], prob=TRUE)
> curve(dnorm(x, mean=mean(est[,7]), sd=sqrt(var(est[,7]))),col="red", add=TRUE)
```



The estimators of OLS regression here should look normal. However, from the histograms above, the estimators don't look exactly normally distributed, but the shapes are similar. Perhaps it would be more normal distribution shaped if there're more regressions.

```
> estexp <- matrix(, ncol = 7, nrow = 100)
> for (i in 1:100){
+   e <- rnorm(656)
+   y <- 0.01 + -0.2*DataExample$log Pg` + 1*DataExample$log(Income/Pop)`
+     0.5*DataExample$log Pnc` + 0.4*DataExample$log Puc` + 0.02*Data
Example$month...15 + 0.9*DataExample$log Pd` + e
+   estexp[i,] <- lm(y ~ DataExample$log Pg` + DataExample$log(Income/Po
p)` + DataExample$log Pnc` + DataExample$log Puc` + DataExample$month...15
+ DataExample$log Pd`)$coefficients
+ }
> View(estexp)
> var(estexp)
```

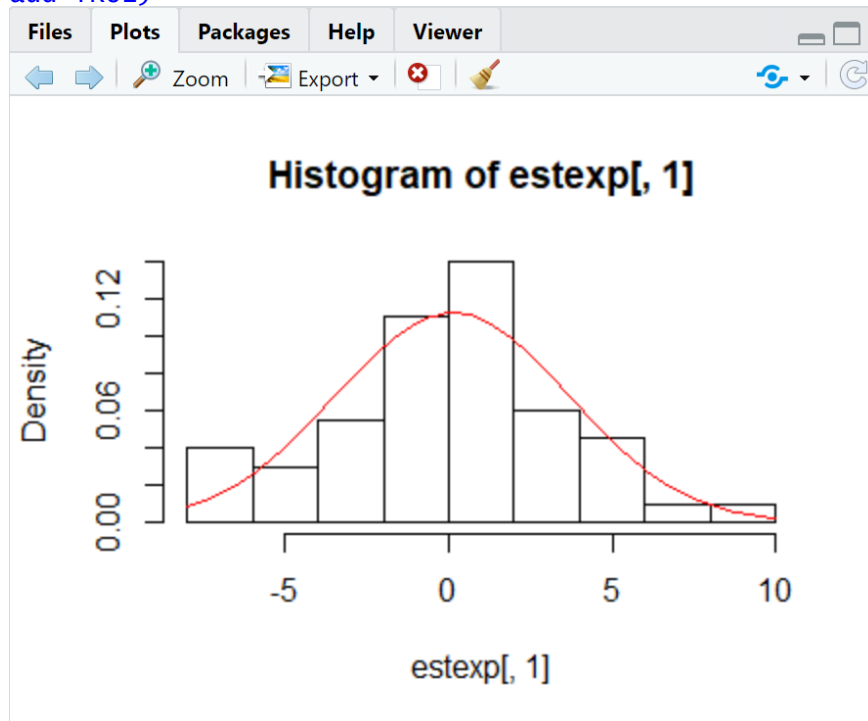
```
      [,1]      [,2]      [,3]
[1,] 1.268417e+01 -1.226888e-01 1.563323e-03
[2,] -1.226888e-01 1.613732e-02 -2.361034e-04
[3,] 1.563323e-03 -2.361034e-04 1.855128e-03
[4,] -4.394387e-02 -2.789946e-04 4.148633e-04
[5,] -2.548899e+00 5.932758e-03 -8.306628e-04
[6,] -5.416539e-05 2.414596e-05 1.771405e-06
[7,] 7.115898e-03 1.159929e-02 -1.634593e-04
      [,4]      [,5]      [,6]
[1,] -4.394387e-02 -2.548899e+00 -5.416539e-05
[2,] -2.789946e-04 5.932758e-03 2.414596e-05
[3,] 4.148633e-04 -8.306628e-04 1.771405e-06
[4,] 6.660836e-03 2.933959e-03 8.822864e-08
[5,] 2.933959e-03 5.493659e-01 -4.440147e-05
[6,] 8.822864e-08 -4.440147e-05 1.519538e-07
[7,] -1.259855e-03 -2.868325e-02 6.355369e-05
      [,7]
[1,] 7.115898e-03
[2,] 1.159929e-02
[3,] -1.634593e-04
[4,] -1.259855e-03
[5,] -2.868325e-02
[6,] 6.355369e-05
[7,] 3.777345e-02
```

variables	Cross-sample variance
(intercept)	1.268417e+01
log Pg	1.613732e-02
log (Income/Pop)	1.855128e-03
log Pnc	6.660836e-03
log Puc	5.493659e-01
Month	1.519538e-07
log Pd	3.777345e-02

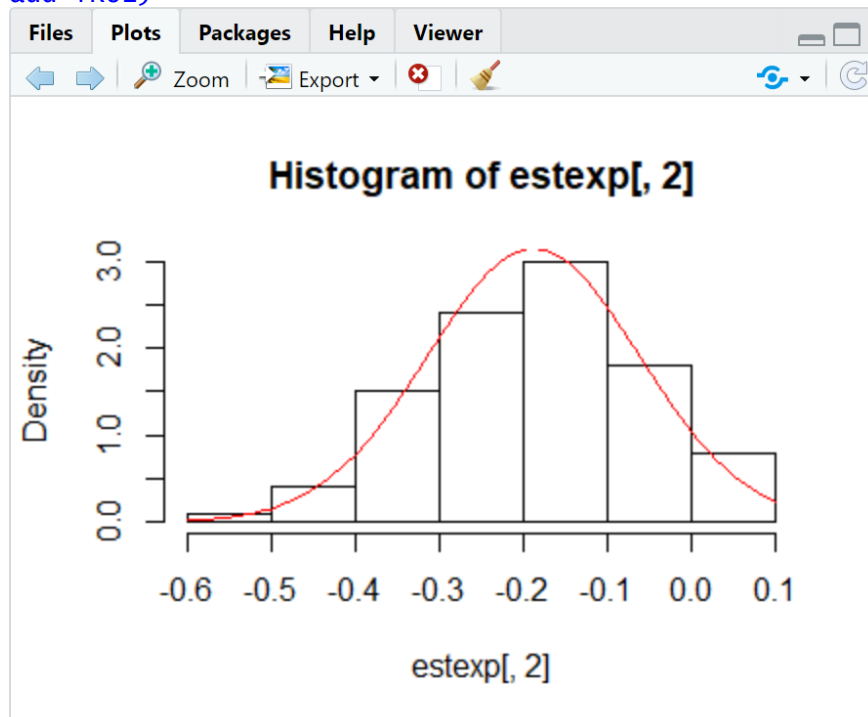
Now with the new distribution of exponential distribution, we have the below histograms with normal distribution curve.

It seems that the distribution of estimator is not as of normal distribution as the last case. However, its shape suggests the convergence to normal distribution. This agrees with the theorem that despite the distribution of disturbances, the estimator will converge into normal distribution.

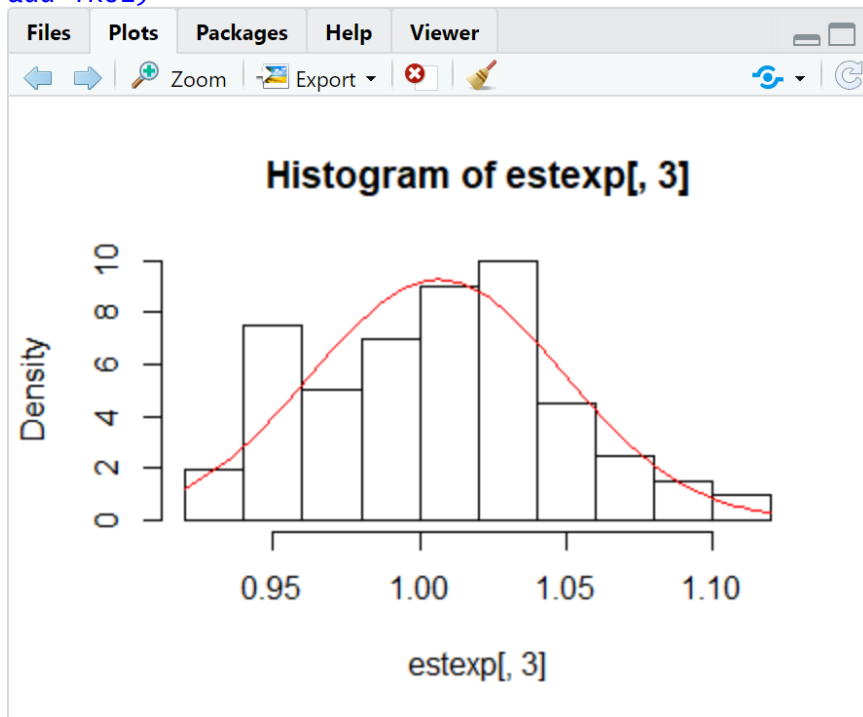
```
> hist(estexp[,1], prob=TRUE)
> curve(dnorm(x, mean=mean(estexp[,1]), sd=sqrt(var(estexp[,1]))),col="red",
add=TRUE)
```



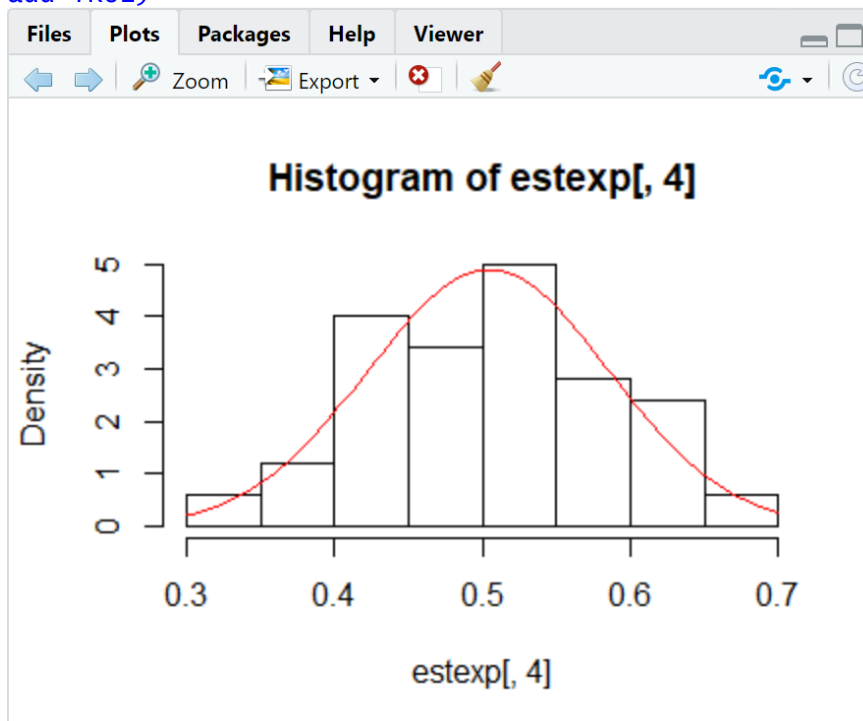
```
> hist(estexp[,2], prob=TRUE)
> curve(dnorm(x, mean=mean(estexp[,2]), sd=sqrt(var(estexp[,2]))),col="red",
add=TRUE)
```



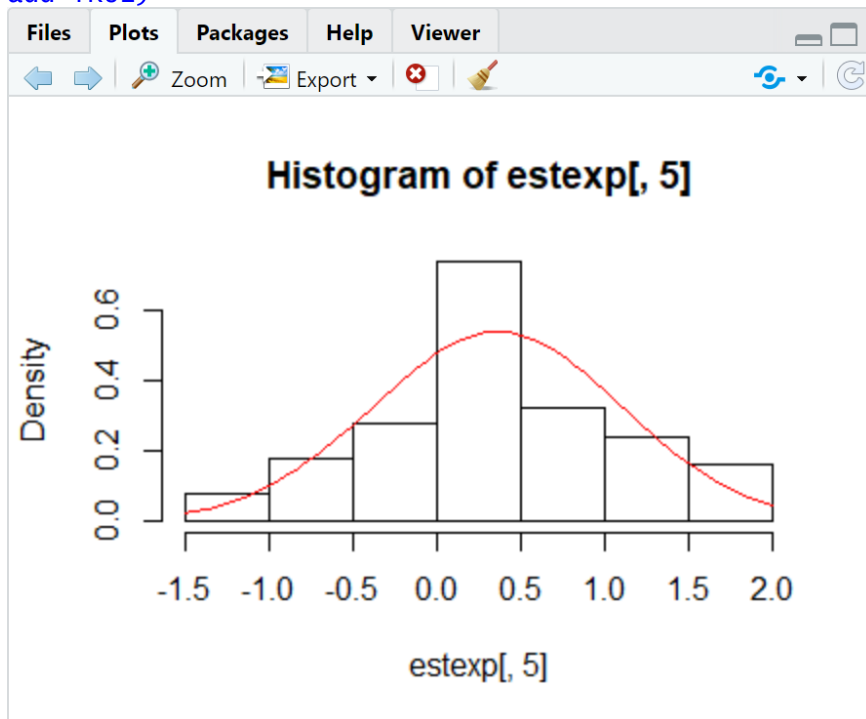
```
> hist(estexp[,3], prob=TRUE)
> curve(dnorm(x, mean=mean(estexp[,3]), sd=sqrt(var(estexp[,3]))),col="red",
add=TRUE)
```



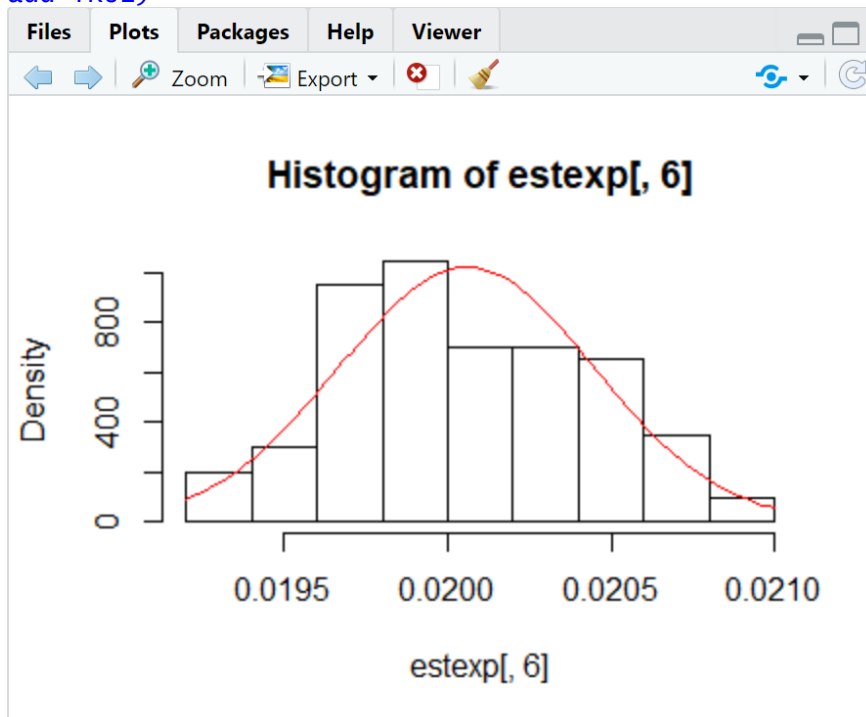
```
> hist(estexp[,4], prob=TRUE)
> curve(dnorm(x, mean=mean(estexp[,4]), sd=sqrt(var(estexp[,4]))),col="red",
add=TRUE)
```



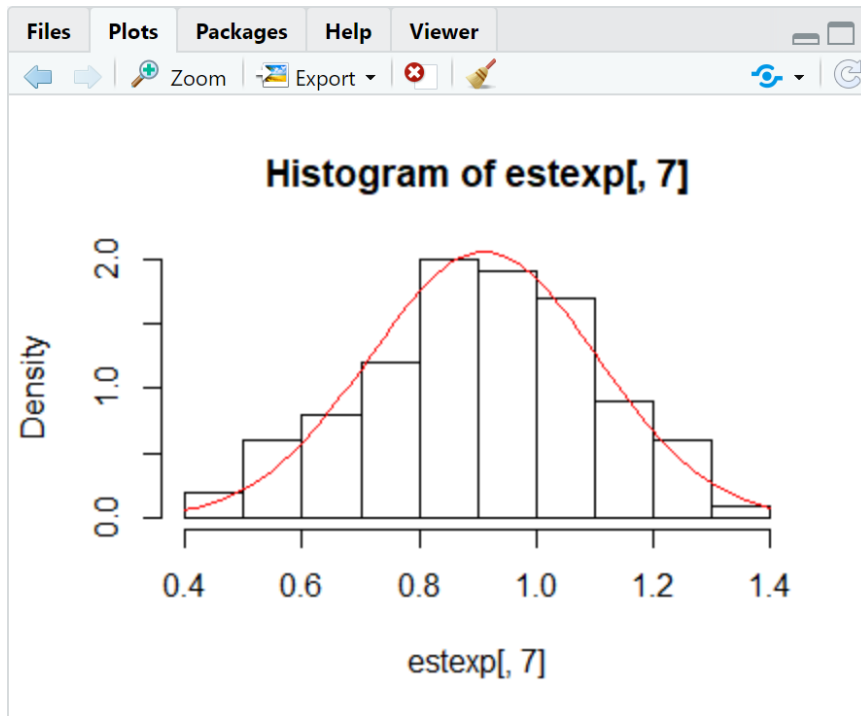
```
> hist(estexp[,5], prob=TRUE)
> curve(dnorm(x, mean=mean(estexp[,5]), sd=sqrt(var(estexp[,5]))),col="red",
add=TRUE)
```



```
> hist(estexp[,6], prob=TRUE)
> curve(dnorm(x, mean=mean(estexp[,6]), sd=sqrt(var(estexp[,6]))),col="red",
add=TRUE)
```



```
> hist(estexp[,7], prob=TRUE)
> curve(dnorm(x, mean=mean(estexp[,7]), sd=sqrt(var(estexp[,7]))),col="red",
add=TRUE)
```



Problem 2

```
> library(readxl)
> datacrime <- read_excel("C:/Users/quang/Downloads/datacrime.xlsx")
> View(datacrime)
> model <- lm(I(log(crmrte))~I(log(prbarr))+I(log(prbconv))+I(log(prbpris))+I(log(avgsen)), data = datacrime, year == 87)
> summary(model)
```

Call:
lm(formula = I(log(crmrte)) ~ I(log(prbarr)) + I(log(prbconv)) + I(log(prbpris)) + I(log(avgsen)), data = datacrime, subset = year == 87)

Residuals:

Min	1Q	Median	3Q	Max
-1.36104	-0.19129	0.07939	0.27754	0.86843

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-4.86792	0.43153	-11.281	< 2e-16
I(log(prbarr))	-0.72397	0.11532	-6.278	1.39e-08
I(log(prbconv))	-0.47251	0.08311	-5.686	1.80e-07
I(log(prbpris))	0.15967	0.20644	0.773	0.441
I(log(avgsen))	0.07642	0.16347	0.467	0.641

(Intercept) ***
I(log(prbarr)) ***
I(log(prbconv)) ***
I(log(prbpris))
I(log(avgsen))

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.429 on 85 degrees of freedom
Multiple R-squared: 0.4162, Adjusted R-squared: 0.3888
F-statistic: 15.15 on 4 and 85 DF, p-value: 2.171e-09

```
> model2 <- lm(I(log(crmrte))~I(log(prbarr))+I(log(prbconv))+I(log(prbpris))+I(log(avgsen))+I(log(pcrmrte)), data = datacrime, year == 87)
> summary(model2)
```

Note that the variable pcrmrte is the crime rates of previous year, which I computed on Excel for convenience.

Call:
lm(formula = I(log(crmrte)) ~ I(log(prbarr)) + I(log(prbconv)) + I(log(prbpris)) + I(log(avgsen)) + I(log(pcrmrte)), data = datacrime, subset = year == 87)

Residuals:

Min	1Q	Median	3Q	Max
-1.28022	-0.08931	0.03055	0.11019	0.32422

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.76663	0.31310	-2.449	0.01643
I(log(prbarr))	-0.18504	0.06276	-2.948	0.00414
I(log(prbconv))	-0.03868	0.04660	-0.830	0.40891
I(log(prbpris))	-0.12669	0.09885	-1.282	0.20351
I(log(avgsen))	-0.15202	0.07829	-1.942	0.05552

```
I(log(pcrmrte)) 0.77981 0.04521 17.248 < 2e-16
```

```
(Intercept) *  
I(log(prbarr)) **  
I(log(prbconv))  
I(log(prbpris))  
I(log(avgsen)) .  
I(log(pcrmrte)) ***  
---
```

Signif. codes:

```
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.2025 on 84 degrees of freedom  
Multiple R-squared: 0.8715, Adjusted R-squared: 0.8638  
F-statistic: 113.9 on 5 and 84 DF, p-value: < 2.2e-16
```

Generally, it seems like a better model considering the huge increase in R^2 . It is intuitively understandable since the crime rate from previous year would be a great indicator for that of next year.

$\log(\text{prbconv})$ now becomes insignificant while $\log(\text{avgsen})$ is now significant. $\log(\text{prbarr})$ seems much smaller.

The reason might be that by adding a meaningful explanatory variable, which was previously considered noise in the estimation, crime rates would not be so elastic to change of some variables. When last year crime rates stays stationary, the change in a deterrent variable won't affect as much comparing to when there's no last year crime rates explanatory variable. The changes in significance can be explained by the fact that the variables might have correlation (or even causal connection) with last year crime rates variable.

```
> library("car")  
> linearHypothesis(model2, c("I(log(prbarr))=0", "I(log(prbconv))=0", "I(log(prbpris))=0", "I(log(avgsen))=0", "I(log(pcrmrte))=0"))  
Linear hypothesis test
```

Hypothesis:

```
I(log(prbarr)) = 0  
I(log(prbconv)) = 0  
I(log(prbpris)) = 0  
I(log(avgsen)) = 0  
I(log(pcrmrte)) = 0
```

Model 1: restricted model

```
Model 2: I(log(crmrte)) ~ I(log(prbarr)) + I(log(prbconv)) + I(log(prbpris))  
+ I(log(avgsen)) + I(log(pcrmrte))
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	89	26.7997				
2	84	3.4447	5	23.355	113.9	< 2.2e-16 ***

Signif. codes:

```
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So the joint significance hypothesis confirms the significance of variables used in the above model. we can reject the hypothesis that all of the coefficients of the explanatory variables are 0.

```
> model3 <- lm(I(log(crmrte))~I(log(prbarr))+I(log(prbconv))+I(log(prbpris))+  
I(log(avgsen))+I(log(pcrmrte))+I(log(wcon))+I(log(wtuc))+I(log(wtrd))+I(log(wfir))+I(log(wser))+I(log(wmfg))+I(log(wfed))+I(log(wsta))+I(log(wloc)), data  
= datacrime, year == 87)  
> summary(model3)
```

Call:

```
lm(formula = I(log(crmrte)) ~ I(log(prbarr)) + I(log(prbconv)) +
  I(log(prbpris)) + I(log(avgsen)) + I(log(pcrmrte)) + I(log(wcon)) +
  I(log(wtuc)) + I(log(wtrd)) + I(log(wfir)) + I(log(wser)) +
  I(log(wmfg)) + I(log(wfed)) + I(log(wsta)) + I(log(wloc)),
  data = datacrime, subset = year == 87)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.08551	-0.08893	0.01248	0.10860	0.35101

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3.79254	1.95747	-1.937	0.0565
I(log(prbarr))	-0.17251	0.06595	-2.616	0.0108
I(log(prbconv))	-0.06836	0.04973	-1.375	0.1733
I(log(prbpris))	-0.21555	0.10240	-2.105	0.0386
I(log(avgsen))	-0.19605	0.08446	-2.321	0.0230
I(log(pcrmrte))	0.74534	0.05303	14.054	<2e-16
I(log(wcon))	-0.28500	0.17752	-1.605	0.1126
I(log(wtuc))	0.06413	0.13433	0.477	0.6344
I(log(wtrd))	0.25371	0.23175	1.095	0.2771
I(log(wfir))	-0.08353	0.19650	-0.425	0.6720
I(log(wser))	0.11276	0.08474	1.331	0.1874
I(log(wmfg))	0.09874	0.11861	0.832	0.4078
I(log(wfed))	0.33613	0.24531	1.370	0.1747
I(log(wsta))	0.03951	0.20721	0.191	0.8493
I(log(wloc))	-0.03698	0.32915	-0.112	0.9108

```
(Intercept)      .
I(log(prbarr))   *
I(log(prbconv))
I(log(prbpris))  *
I(log(avgsen))   *
I(log(pcrmrte)) ***
I(log(wcon))
I(log(wtuc))
I(log(wtrd))
I(log(wfir))
I(log(wser))
I(log(wmfg))
I(log(wfed))
I(log(wsta))
I(log(wloc))
---
```

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1973 on 75 degrees of freedom

Multiple R-squared: 0.8911, Adjusted R-squared: 0.8707

F-statistic: 43.81 on 14 and 75 DF, p-value: < 2.2e-16

```
> linearHypothesis(model3, c("I(log(wcon))=0", "I(log(wtuc))=0", "I(log(wtrd))=0",
  "I(log(wfir))=0", "I(log(wser))=0", "I(log(wmfg))=0", "I(log(wfed))=0", "I(log(wsta))=0",
  "I(log(wloc))=0"))
```

Linear hypothesis test

Hypothesis:

```
I(log(wcon)) = 0
I(log(wtuc)) = 0
I(log(wtrd)) = 0
I(log(wfir)) = 0
I(log(wser)) = 0
I(log(wmfg)) = 0
I(log(wfed)) = 0
```

```
I(log(wsta)) = 0
I(log(wloc)) = 0
```

Model 1: restricted model

Model 2: $I(\log(\text{crmrte})) \sim I(\log(\text{prbarr})) + I(\log(\text{prbconv})) + I(\log(\text{prbpris}))$

+

```
I(log(avgsen)) + I(log(pcrmrte)) + I(log(wcon)) + I(log(wtuc)) +
I(log(wtrd)) + I(log(wfir)) + I(log(wser)) + I(log(wmfg)) +
I(log(wfed)) + I(log(wsta)) + I(log(wloc))
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	84	3.4447				
2	75	2.9198	9	0.52491	1.4981	0.1643

Regressing a model with the added wage variables.

From the hypothesis test result, we can't reject the hypothesis that all the wage variables have their coefficients' significance 0.

Problem 3

```
> library(readxl)
> invest_dat <- read_excel("C:/Users/quang/Downloads/invest.dat.xls")
> View(invest_dat)
> model <- lm(I~Q+C+D,data=invest_dat)
> summary(model)
```

Call:
lm(formula = I ~ Q + C + D, data = invest_dat)

Residuals:

Min	1Q	Median	3Q	Max
-17.482	-3.838	-1.468	2.490	28.633

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.5099	0.3663	20.504	< 2e-16 ***
Q	-0.4213	0.1750	-2.407	0.0164 *
C	7.2293	1.5109	4.785	2.19e-06 ***
D	0.7364	0.4707	1.564	0.1183

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.787 on 561 degrees of freedom
Multiple R-squared: 0.06243, Adjusted R-squared: 0.05741
F-statistic: 12.45 on 3 and 561 DF, p-value: 6.809e-08

```
> vcov(model)
```

	(Intercept)	Q	C
(Intercept)	0.13414800	0.016796183	-0.27710693
Q	0.01679618	0.030640643	-0.22939056
C	-0.27710693	-0.229390557	2.28286300
D	-0.07620369	-0.004942638	0.03318712

	D
(Intercept)	-0.076203690
Q	-0.004942638
C	0.033187119
D	0.221568789

Variables	Std. Error	Left Interval	Right Interval
Intercept	0.3663		
Q	0.1750	-0.4357	-0.4069
C	1.5109	7.1047	7.3539
D	0.4707	0.6976	0.7752

```
> hccm(model)
```

	(Intercept)	Q	C
(Intercept)	0.57117649	-0.3418879	-0.23005729
Q	-0.34188785	0.3447220	-0.36753633
C	-0.23005729	-0.3675363	2.98929567
D	0.02022947	-0.1247757	0.07000167

	D
(Intercept)	0.02022947
Q	-0.12477567
C	0.07000167
D	0.40498252

Variables	Std. Error	Left Interval	Right Interval
Intercept	0.7558		
Q	0.5871	-0.4697	-0.3729
C	1.7290	7.0867	7.3719
D	0.6364	0.6839	0.7889

```
> linearHypothesis(model, c("C=0","D=0"))
Linear hypothesis test

Hypothesis:
C = 0
D = 0

Model 1: restricted model
Model 2: I ~ Q + C + D

   Res.Df  RSS Df Sum of Sq    F    Pr(>F)
1     563 19615
2     561 18788   2    827.09 12.348 5.649e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> linearHypothesis(model, c("Q=0"))
Linear hypothesis test
```

```
Hypothesis:
Q = 0

Model 1: restricted model
Model 2: I ~ Q + C + D

   Res.Df  RSS Df Sum of Sq    F    Pr(>F)
1     562 18982
2     561 18788   1    193.99 5.7923 0.01642 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now according to the joint significance test,
The hypothesis that both variables C & D are insignificant is rejected. At least one out of the two variables must be significant.
The hypothesis that Q is insignificant is also rejected at 5% confidence level. Q is an explanatory variable for I. However, it's not consistent with the theory that it's the only explanator.

```
> QC <- invest_dat$Q*invest_dat$C
> QD <- invest_dat$Q*invest_dat$D
> CD <- invest_dat$C*invest_dat$D
> model2 <- lm(I~Q+C+D+I(Q^2)+I(C^2)+I(D^2)+QC+QD+CD,data=invest_dat)
> summary(model2)
```

```
Call:
lm(formula = I ~ Q + C + D + I(Q^2) + I(C^2) + I(D^2) + QC +
    QD + CD, data = invest_dat)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-13.023  -3.611  -1.296   2.499  28.217
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.93003    0.52614  11.271 < 2e-16 ***
Q              0.40614    0.32644   1.244 0.213981
C              9.95826    2.58738   3.849 0.000133 ***
D              1.73594    0.97909   1.773 0.076775 .
I(Q^2)        -0.10895    0.04562  -2.388 0.017271 *
I(C^2)        -12.62043    3.61622  -3.490 0.000522 ***
I(D^2)         -0.86869    0.53455  -1.625 0.104716
QC              2.09793    0.71644   2.928 0.003548 **
QD             -0.21280    0.32038  -0.664 0.506839
```

```
CD          5.73804    2.93164    1.957 0.050815 .
```

```
---
```

```
Signif. codes:
```

```
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 5.568 on 555 degrees of freedom
```

```
Multiple R-squared: 0.1415, Adjusted R-squared: 0.1276
```

```
F-statistic: 10.16 on 9 and 555 DF, p-value: 1.488e-14
```

```
> linearHypothesis(model2, c("I(Q^2)=0","I(C^2)=0","I(D^2)=0","QC=0","QD=0","CD=0"))
```

```
Linear hypothesis test
```

```
Hypothesis:
```

```
I(Q^2) = 0
```

```
I(C^2) = 0
```

```
I(D^2) = 0
```

```
QC = 0
```

```
QD = 0
```

```
CD = 0
```

```
Model 1: restricted model
```

```
Model 2: I ~ Q + C + D + I(Q^2) + I(C^2) + I(D^2) + QC + QD + CD
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	561	18788				
2	555	17204	6	1584.5	8.5194	7.211e-09 ***

```
---
```

```
Signif. codes:
```

```
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
The hypothesis that all 6 interactions & quadratic variables have the joint significance of 0 is rejected at 0,1% confidence level.
```