OLS Asymptotics

Taro Mieno

AECN 896-003: Applied Econometrics

OLS Asymptotics (Large Sample Properties)

What is it?

- Properties of OLS that hold only when the sample size is infinite (very large)

OLS Asymptotics (Large Sample Properties)

What is it?

- Properties of OLS that hold only when the sample size is infinite (very large)
- (loosely put) How OLS estimators behave when the number of observations goes infinite (really large)

OLS Asymptotics (Large Sample Properties)

What is it?

- Properties of OLS that hold only when the sample size is infinite (very large)
- (loosely put) How OLS estimators behave when the number of observations goes infinite (really large)

Small sample properties

Under certain conditions,

- Unbiasedness of OLS estimators
- Efficiency of OLS estimators

hold whatever the sample size is (including infinite numbers of observations).

Consistency

Consistency

Verbally (and very loosely),

An estimator is consistent if the probability that the estimator produces the true parameter is 1 when sample size is infinite.

Consistency

Verbally (and very loosely),

An estimator is consistent if the probability that the estimator produces the true parameter is 1 when sample size is infinite.

Example:

OLS estimator of the coefficient on x in the following model with all $MLR.1$ through $MLR.4$ satisfied:

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

with all the conditions necessary for the unbiasedness property of OLS satisfied.

MC simulations: consistency of OLS estimators

Conceptual steps of MC simulations

- generate data (N observations) according to $y_i = \beta_0 + \beta_1 x_i + u_i$
- run on the generated data
- store the coefficient estimate
- repeat the above experiment 1000 times
- examine how the coefficient estimates are distributed

MC simulations: consistency of OLS estimators

Conceptual steps of MC simulations

- generate data (N observations) according to $y_i = \beta_0 + \beta_1 x_i + u_i$
- run on the generated data
- store the coefficient estimate
- repeat the above experiment 1000 times
- examine how the coefficient estimates are distributed

What you should see is

As N gets larger (more observations), the distribution of $\hat{\beta}_1$ get more tightly centered around its true value (here, 1)

Consistency

R code: $N = 100$, 1000, and 10000

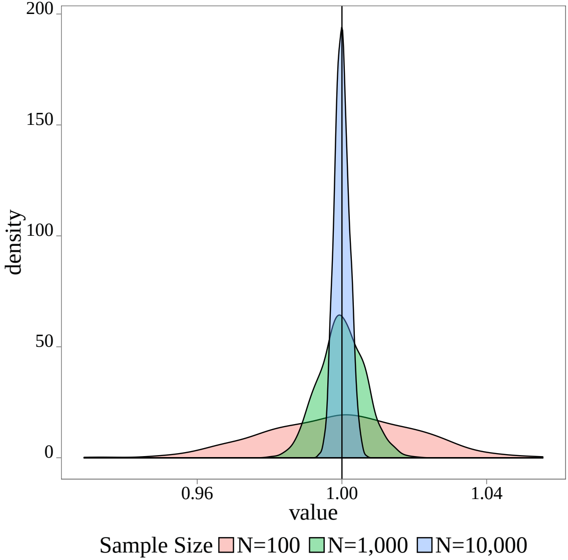
```
#--- Preparation ---#
B <- 1000 # the number of iterations
N_list <- c(100,1000,10000) # sample size
N_len <- length(N_list)
estimate_storage <- matrix(0,B,3) # estimates storage

for (j in 1:N_len){
        temp_N <- N_list[j]
        for (i in 1:B){
        #--- generate data ---#
        x <- rnorm(temp_N) # indep var 1
        u <- rnorm(temp_N)*0.2 # error
        y <- 1 + x + u # dependent variable 1
        data <- data.table(y=y,x=x)

        #--- OLS ---#
        reg <- lm(y~x,data=data) # OLS

        #--- store coef estimates ---#
        estimate_storage[i,j] <- reg$coef[2]
        }
}
```

Consistency

R code: Visualize

```r
#--- wide to long format ---#
plot_data <- melt(data.table(estimate_storage))

#--- create a figure ---#
g_co_ex <- ggplot(data=plot_data) +
        geom_density(aes(x=value,fill=variable),alpha=0.4) +
        geom_vline(xintercept=1) +
        scale_fill_discrete(
                name='Sample Size',
                labels = c('N=100 ', 'N=1,000 ','N=10,000')
                ) +
        theme(
                legend.position='bottom'
                )
```

Consistency



Consistency

Consistency of OLS estimators

Under $MLR.1$ through $MLR.4$, OLS estimators are consistent

MC simulations: Inconsistency of OLS estimators

Conceptual steps of MC simulations

- generate data (N observations) according to $y_i = \beta_0 + \beta_1 x_i + u_i$ with $E[u_i|x_i] \neq 0$
- run on the generated data
- store the coefficient estimate
- repeat the above experiment 1000 times
- examine how the coefficient estimates are distributed

MC simulations: Inconsistency of OLS estimators

Conceptual steps of MC simulations

- generate data (N observations) according to $y_i = \beta_0 + \beta_1 x_i + u_i$ with $E[u_i|x_i] \neq 0$
- run on the generated data
- store the coefficient estimate
- repeat the above experiment 1000 times
- examine how the coefficient estimates are distributed

What should you see?

Would the bias disappear as N gets larger?

Inconsistency

R code: $N = 100$, 1000, and 10000

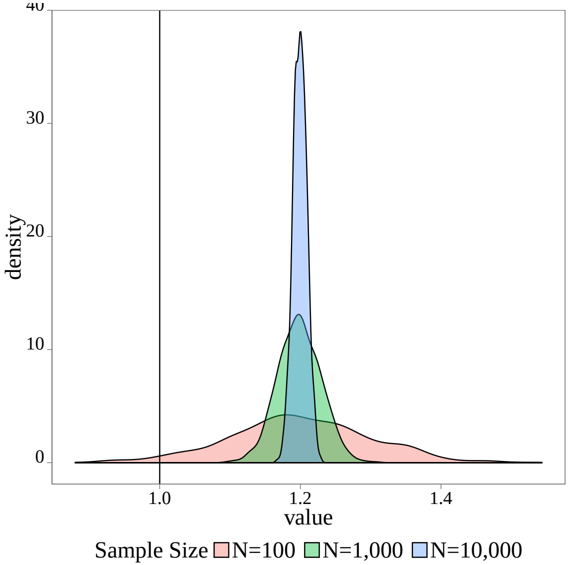
```
#--- Preparation ---#
N_list <- c(100,1000,10000) # sample size
N_len <- length(N_list)
estimate_storage <- matrix(0,B,3) # estimates storage

for (j in 1:N_len){
        temp_N <- N_list[j]
        for (i in 1:B){
        #--- generate data ---#
        mu <- rnorm(temp_N) # shared term between x and u
        x <- rnorm(temp_N) + 0.5*mu
        u <- rnorm(temp_N) + 0.5*mu
        y <- 1 + x + u # dependent variable
        data <- data.table(y=y,x=x)

        #--- OLS ---#
        reg <- lm(y~x,data=data) # OLS

        #--- store coef estimates ---#
        estimate_storage[i,j] <- reg$coef[2]
        }
}
```

Inconsistency

R code: Visualize

```
#--- wide to long format ---#
plot_data <- melt(data.table(estimate_storage))

#--- create a figure ---#
g_inco_ex <- ggplot(data=plot_data) +
  geom_density(aes(x=value,fill=variable),alpha=0.4) +
  geom_vline(xintercept=1) +
  scale_fill_discrete(
    name='Sample Size',
    labels = c('N=100 ', 'N=1,000 ','N=10,000')
    ) +
  theme(
    legend.position='bottom'
    )
```

Inconsistency



Inconsistency of OLS estimators

Important

Bias due to the violation of any of the $MLR.1$ through $MLR.4$ would not go away even if you increase the number of observations.

Asymptotic Normality

Inference

$MLR.6$: Normality

The population error u is independent of the explanatory variables $x_1, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2:

$$u \sim Normal(0, \sigma^2)$$

Remember

- If $MLR.6$ are violated, t-statistic and F-statistic we constructed before are no longer distributed as t-distribution and F-distribution, respectively
- So, whenever $MLR.6$ is violated, our t- and F-tests are invalid

Inference

Fortunately,

You can continue to use t- and F-tests because (slightly transformed) OLS estimators are approximately normally distributed when the sample size is large enough.

Central Limit Theorem (CLT)

Central Limit Theorem (Lindberg-Levy)

Suppose $\{x_1, x_2, \ldots\}$ is a sequence of identically independently distributed random variables with $E[x_i] = \mu$ and $Var[x_i] = \sigma^2 < \infty$. Then, as n approaches infinity,

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Verbally: sample mean less its expected value multiplied by $\sqrt{n}$ is going to be distributed as standard Normal distribution as n goes infinity.

CLT

$x_i \sim Bern[p = 0.3]$

1 with probability p and 0 with probability $1 - p$.

- $E[x_i] = p = 0.3$
- $Var[x_i](\sigma^2) = p(1-p) = 0.21$

According to CLT

$$\left(\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu) \xrightarrow{d} N(0, \sigma^2)\right)$$

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} x_i - 0.3) \xrightarrow{d} N(0, 0.21)$$

MC simulations: CLT

Conceptual steps of the MC simulation

- draw n observations from $x_i \sim Bern(0.3)$
- find its mean, subtract the expected value (here, $E[x_i] = 0.3$), multiply by $\sqrt{n}$ ($\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu)$)
- store the calculated value
- repeat the above experiment 1000 times
- examine how the calculated values are distributed

MC simulations: CLT

Conceptual steps of the MC simulation

- draw n observations from $x_i \sim Bern(0.3)$
- find its mean, subtract the expected value (here, $E[x_i] = 0.3$), multiply by $\sqrt{n}$ ($\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu)$)
- store the calculated value
- repeat the above experiment 1000 times
- examine how the calculated values are distributed

What you should see is

As N gets larger (more observations), the distribution of $\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu)$ looks more and more like $N(0, Var(x_i)$

CLT

R code: CLT

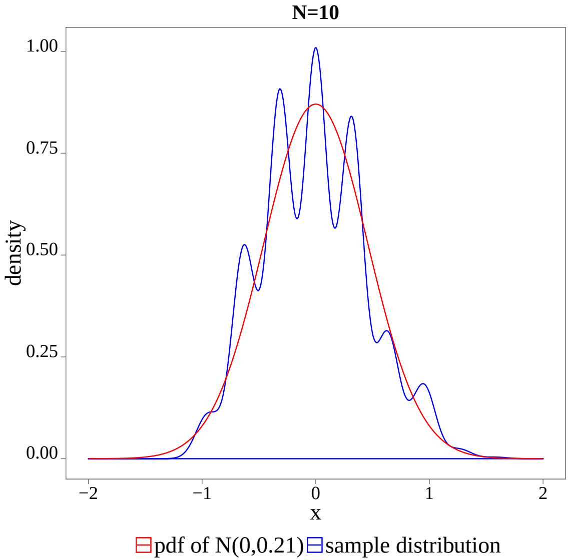
```
set.seed(893269)
#--- the number of observations ---#
# this is what we change
N <- 10 # number of observations
B <- 1000 # number of iterations
p <- 0.3 # mean of the Bernoulli distribution
storage <- rep(0,B)

for (i in 1:B){
  #--- draw from Bern[0.3] (x distributed as Bern[0.3]) ---#
  x_seq <- runif(N)<=p

  #--- sample mean ---#
  x_mean <- mean(x_seq)

  #--- normalize ---#
  lhs <- sqrt(N)*(x_mean-p)

  #--- save lhs to storage ---#
  storage[i] <- lhs
}
```

CLT Visualization: $N = 10$

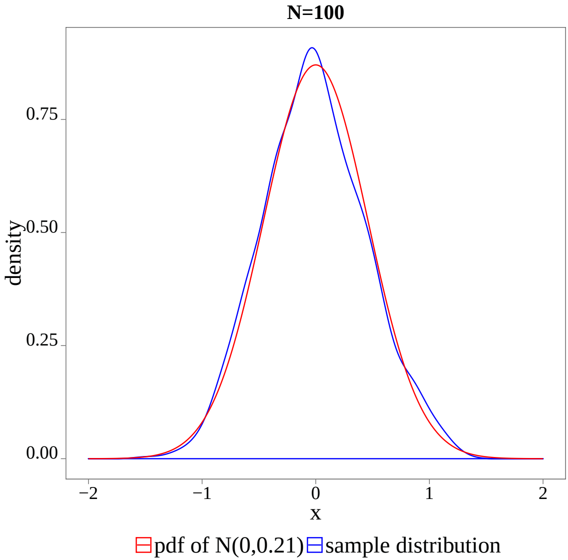
R code: CLT visualize

```
data_pdf <- data.table(
    x = seq(-2,2,length=1000),
    y = dnorm(seq(-2,2,length=1000),sd=sqrt(p*(1-p)))
    )
g_N_10 <- ggplot() +
  geom_density(
    data=data.table(x=storage),
    aes(x=x,color='sample distribution')
    ) +
  geom_line(
    data=data_pdf,
    aes(y=y,x=x,color='pdf of N(0,0.21)')
    ) +
  scale_color_manual(
    values=c('sample distribution'='blue','pdf of N(0,0.21)'='red'),
    name='') +
  theme(
    legend.position='bottom'
    ) +
  ggtitle('N=10')
```

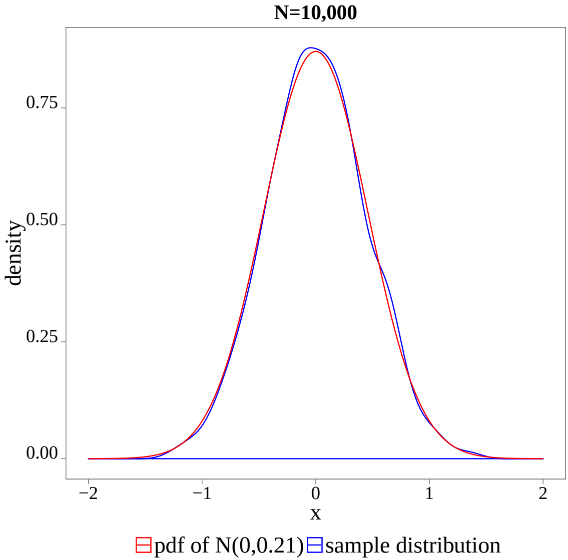
CLT Visualization: $N = 10$



CLT Visualization: $N = 100$



CLT Visualization: $N = 10,000$



CLT

Important

CLT holds for any distribution of x_i as long as it has a finite expected value and variance.

Asymptotics

Under assumptions $MLR.1$ through $MLR.5$ ($MLR.6$ not necessary!!),

Asymptotic Normality of OLS

$$\sqrt{n}(\hat{\beta}_j - \beta_j) \xrightarrow{a} N(0, \sigma^2/\alpha_j^2)$$

where $\alpha_j^2 = plim(\frac{1}{n}\sum_{i=1}^{n} r_{i,j}^2)$, where $r_{i,j}^2$ are the residuals from regressing x_j on the other independent variables.

Consistency of $\hat{\sigma}^2 \equiv \frac{1}{n-k-1}\sum_{i=1}^{n} \hat{u}_i^2$

$\hat{\sigma}^2$ is a consistent estimator of σ^2 ($Var(u)$)

Further,

For each j,

- $(\hat{\beta}_j - \beta_j)/sd(\hat{\beta}_j) \xrightarrow{a} N(0,1)$
- $(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j) \xrightarrow{a} N(0,1)$, where
  $$se(\hat{\beta}_j) = \sqrt{\frac{\hat{\sigma}^2}{SST_j(1-R_j^2}})$$

Small vs. Large Sample

Small sample (any sample size)

Under $MLR.1$ through $MLR.5$ and $MLR.6$ $(u_i \sim N(0, \sigma^2))$,

$$(\hat{\beta}_j - \beta_j)/sd(\hat{\beta}_j) \sim N(0,1)$$
$$(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j) \sim t_{n-k-1}$$

Large sample (when n goes infinity)

Under $MLR.1$ through $MLR.5$ without $MLR.6$,

$$(\hat{\beta}_j - \beta_j)/sd(\hat{\beta}_j) \xrightarrow{a} N(0,1)$$
$$(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j) \xrightarrow{a} N(0,1)$$

Testing under large sample

It turns out,

You can proceed exactly the same way as you did before (practically speaking)!!

1. calculate $(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j)$
2. check if the obtained value is greater than (in magnitude) the critical value for the specified significance level under t_{n-k-1}

Testing under large sample

It turns out,

You can proceed exactly the same way as you did before (practically speaking)!!

1. calculate $(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j)$
2. check if the obtained value is greater than (in magnitude) the critical value for the specified significance level under t_{n-k-1}

But,

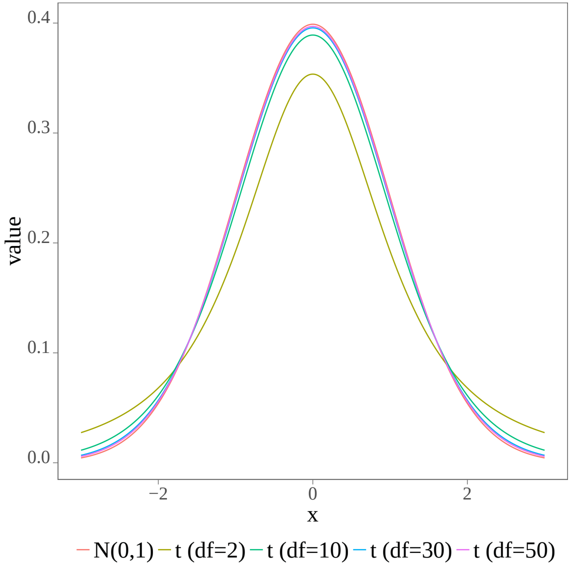
Shouldn't we use $N(0,1)$ when you find the critical value?

Testing under large sample

R code: t vs N distributions

```
x <- seq(-3,3,length=1000)
y_norm <- dnorm(x) # pdf of N(0,1)
y_t_2 <- dt(x,df=2) # pdf of t_{2}
y_t_10 <- dt(x,df=10) # pdf of t_{10}
y_t_30 <- dt(x,df=30) # pdf of t_{30}
y_t_50 <- dt(x,df=50) # pdf of t_{50}

plot_data <- data.table(
  x=x,
  'N(0,1)'=y_norm,
  't (df=2)'=y_t_2,
  't (df=10)'=y_t_10,
  't (df=30)'=y_t_30,
  't (df=50)'=y_t_50
  ) %>%
  melt(id.var='x')

g_t_vs_N <- ggplot(data=plot_data) +
  geom_line(aes(y=value,x=x,color=variable)) +
  scale_color_discrete(name='') +
  theme(
    legend.position='bottom'
    )
```

t vs Normal distributions



Testing under large sample

Since t_{n-k-1} and $N(0,1)$ are almost identical when n is large, there is very little error in using t_{n-k-1} instead of $N(0,1)$ to find the critical value.

Homoskedasticity

Important

The asymptotic normality of OLS does require homoskedasticity assumption ($MLR.5$)!!

- the usual t-statistics and confidence intervals are invalid no matter how large the sample size is if error is heteroskedastic
- we talk extensively about how we should deal with heteroskedasticity