BRN Intermediate Methods Session 2

2. Hypothesis Testing for Linear Regression- A Review

Inhwan Ko (Univ. of Washington, Seattle)

July 15th, 2020

Contents

1. Hypothesis Testing- Where Probabilistic Theory Takes Part
2. Linear Regression

- “Ordinary Least Squares?”
- Key Assumptions for BLUE: Gauss-Markov Assumptions
- A Difference from Maximum Likelihood Estimation

The 2nd Limit to Causal Identification: Sample vs. Population

Note from the last session that we cannot use population data for the causal identification. Therefore, we need to *infer* causality with its sample.

All of the quantities of interest (QOIs) inferred from the sample data are now called "population parameter estimates."

The question is: how much are our paremeter estimates correct?

Two Bridges between Sample and Population

We need two bridges that allow us to connect between sample and population:

1. Law of Large Numbers (LLN)
2. Central Limit Theorem

Let's take a look at both concepts.

Law of Large Numbers

Let p be a probability of an event A. When an event A occurs r times out of n times of independent trials, for any real number $\epsilon > 0$:

$$lim_{n \to \infty} P(|\frac{r}{n} - p| > \epsilon) = 0$$

For instance, even though you (mathmatically) know that the probability of having 2 heads out of 3 coins is:

$${}_3\mathsf{C}_2(\frac{1}{2})^2(\frac{1}{2})^1 = \frac{3}{8}$$

But there is a chance that you will *always* have two heads when you throw coins only 10 times.

When you throw it, let say 100,000,000 times, the probability will converge into $\frac{3}{8}$.

Central Limit Theorem

Let's say your sample size is n. You take n observations $(X_1, X_2, ..., X_n)$ from the population and calculate its mean $(\bar{X})$.

Its distribution converges into *normal distribution* as $n \to \infty$.

There is a lot versions of its proof online, so search for it if you are more interested in!

This time I will execute the theorem with multiple simulations.

Central Limit Theorem

Let's generate population data with 10,000 observations, which follow normal distribution with a true mean of 30 and a true standard deviation of 10. Assign this to an object called "pop".

```
library(MASS)

pop <- rnorm(10000, 30, 10)
pop[1:10]
```

```
##  [1] 40.54058 23.08221 12.58148 29.46333 40.88464 43.76221 46.84541 46.00686
##  [9] 30.84084 25.15101
```

Let's take n observations from the population k times and make this a new function.

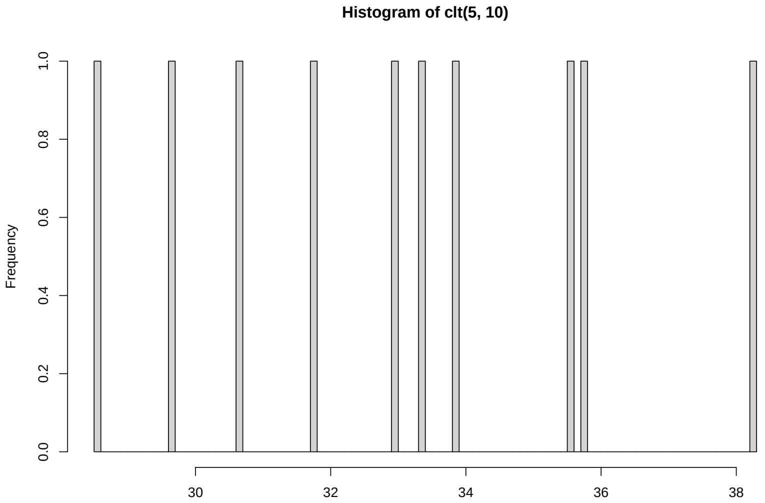
```
samplemean <- NULL

clt <- function(n, k){
  for (i in 1:k) {
  samplemean[i] <- mean(sample(pop, n, replace=T))
  }
  return(samplemean)
}
```

Central Limit Theorem

5 observations, 10 times?

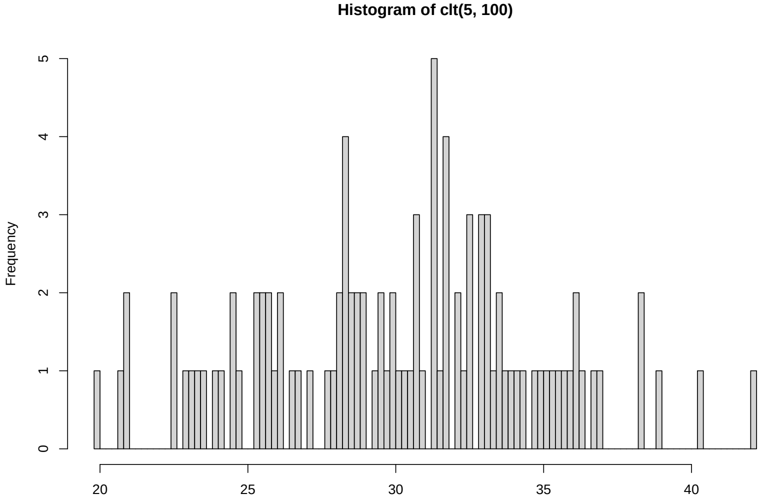
```
hist(clt(5,10), breaks=100)
```



Central Limit Theorem

5 observations, 100 times?

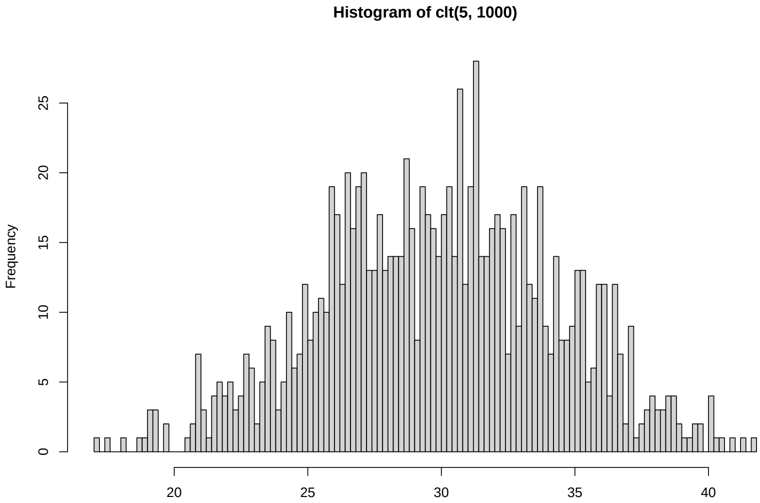
```
hist(clt(5,100), breaks=100)
```



Central Limit Theorem

5 observations, 1000 times?

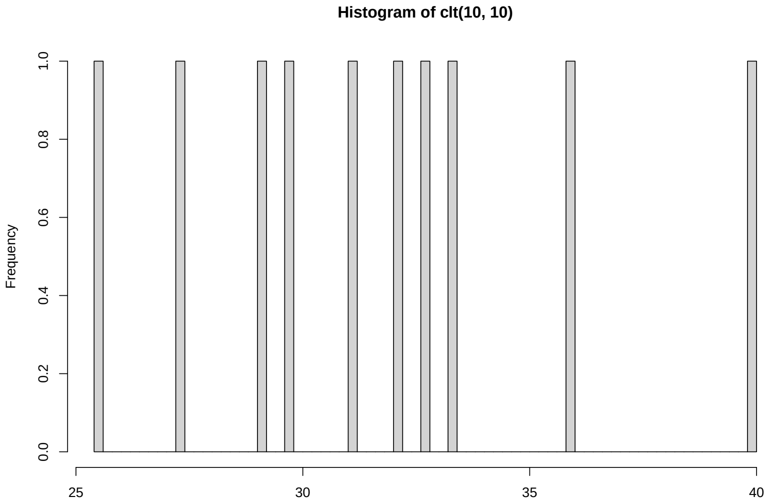
```
hist(clt(5,1000), breaks=100)
```



Central Limit Theorem

10 observations, 10 times?

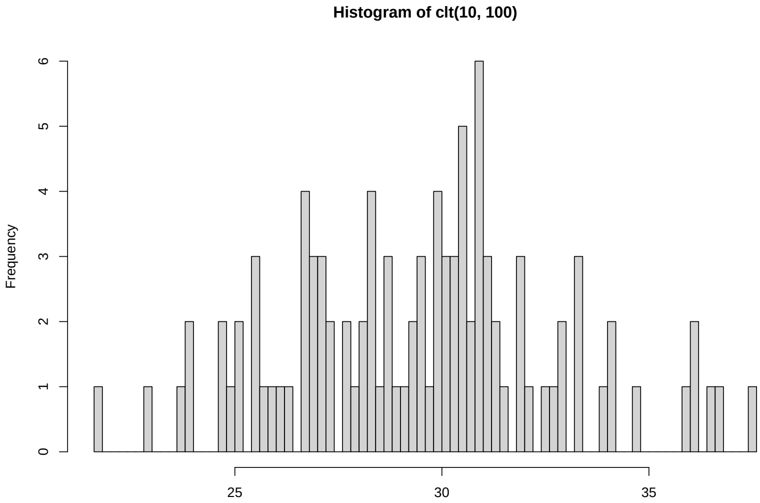
```
hist(clt(10,10), breaks=100)
```



Central Limit Theorem

10 observations, 100 times?

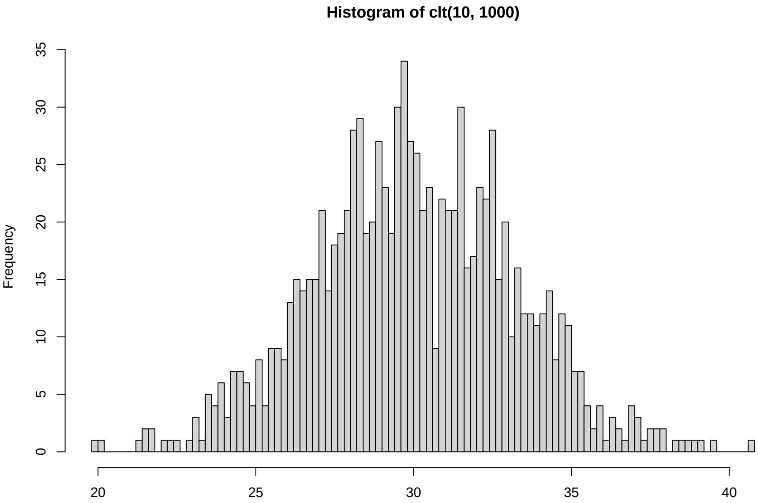
```
hist(clt(10,100), breaks=100)
```



Central Limit Theorem

10 observations, 1000 times?

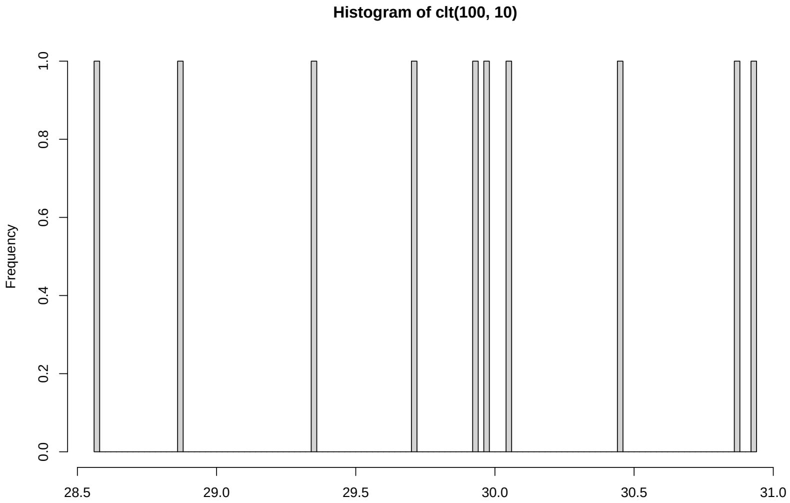
```
hist(clt(10,1000), breaks=100)
```



Central Limit Theorem

100 observations, 10 times?

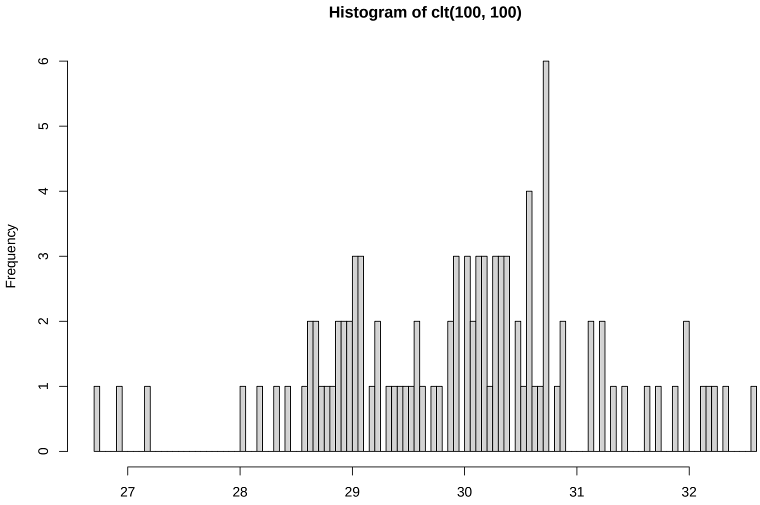
```
hist(clt(100,10), breaks=100)
```



Central Limit Theorem

100 observations, 100 times?

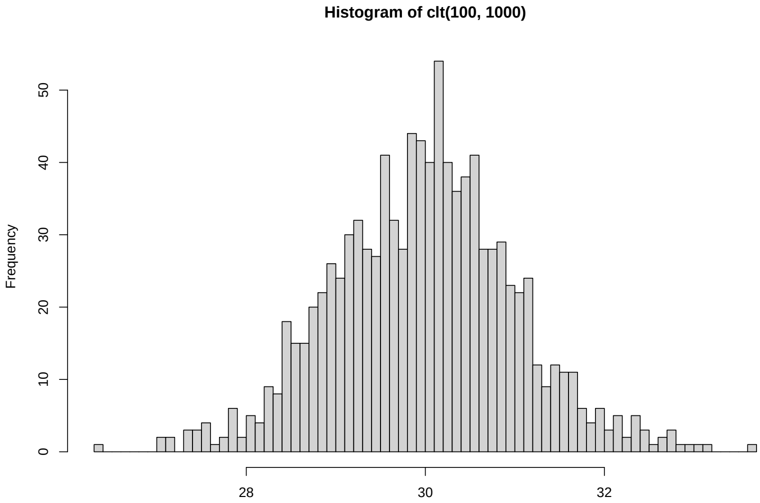
```
hist(clt(100,100), breaks=100)
```



Central Limit Theorem

100 observations, 1000 times?

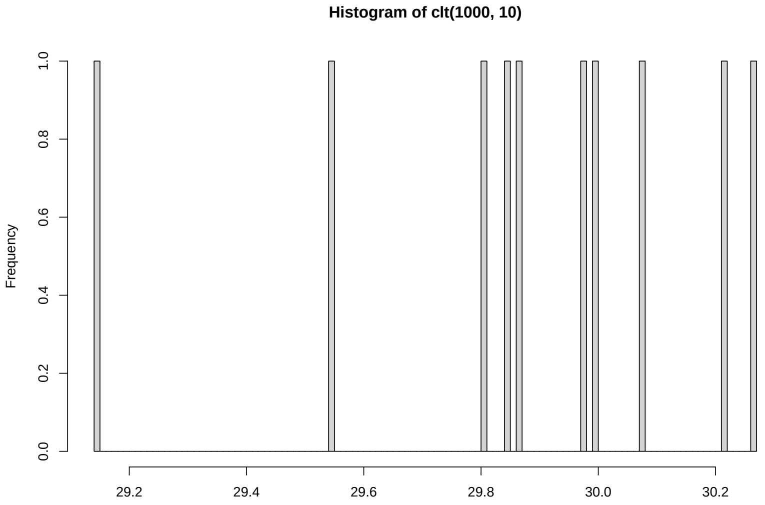
```
hist(clt(100,1000), breaks=100)
```



Central Limit Theorem

1000 observations, 10 times?

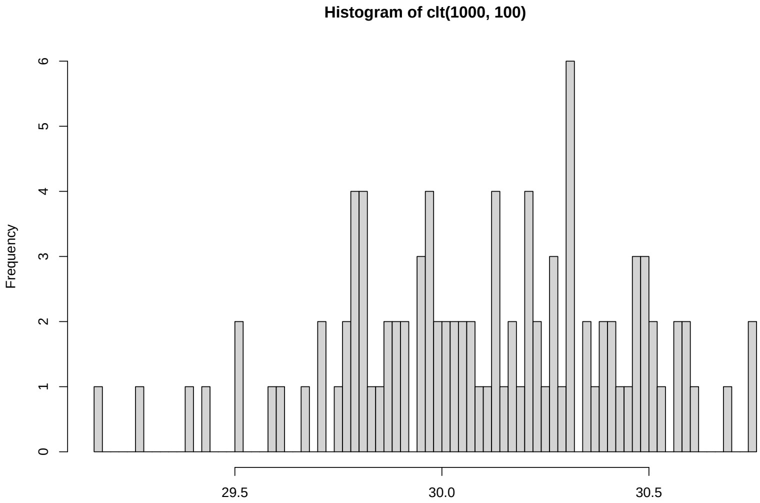
```
hist(clt(1000,10), breaks=100)
```



Central Limit Theorem

1000 observations, 100 times?

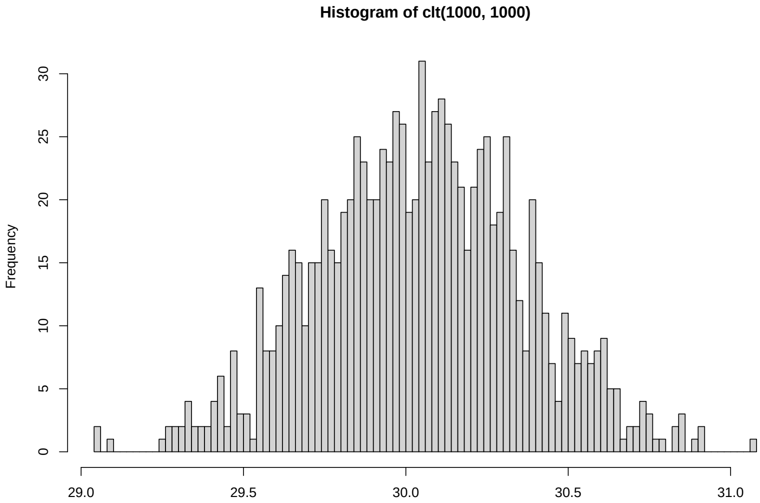
```
hist(clt(1000,100), breaks=100)
```



Central Limit Theorem

1000 observations, 100 times?

```
hist(clt(1000,1000), breaks=100)
```



Central Limit Theorem

Central Limit Theorem (CLT): The sample mean of any population with mean of μ and variance of σ follows $N(\mu, \frac{\sigma^2}{n})$ approximately when its sample size n approaches ∞.

Also, an additional random variable $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$, follows $N(0,1)$ approximately when its sample size n approaches ∞.

We use the latter theorem more often when conducting a hypothesis testing.

Central Limit Theorem

```
clt <- clt(1000,1000)
mean(clt)
```

```
## [1] 30.05191
```

```
sd(clt)^2
```

```
## [1] 0.1005336
```

```
sd(pop)^2/1000
```

```
## [1] 0.0998815
```

P-Value and Hypothesis Testing

Because we can estimate population mean and its variance only with our sample (better if our sample is large enough),

we know whether (and under what probability) the probability distribution function (PDF) of the sample mean contains a certain value.

1. If a certain value is zero, this is our null hypothesis ($H_0 : \bar{X} = 0$), and we are conducting one sample test.
2. If a certain value is another sample mean, this is our null hypothesis $H_0 : \bar{X_1} = \bar{X_2}$, and we are conducting two sample test.

We reject the null hypothesis if, under the null hypothesis, the probability of such a value as that which was actually observed (p-value) is less than or equal to a small, fixed pre-defined threshold value α, which is the level of significance.

Hypothesis Testing- One-Sample Test

```r
sample <- sample(pop, 100)
mean(pop)
```

```
## [1] 30.04778
```

```r
mean(sample)
```

```
## [1] 31.00115
```

```r
t.test(sample)
```

```
##
##  One Sample t-test
##
## data:  sample
## t = 27.558, df = 99, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  28.76904 33.23326
## sample estimates:
## mean of x
##  31.00115
```

Hypothesis Testing- Two-Sample Test

```r
sample1 <- sample(pop, 100)
sample2 <- sample(rnorm(10000, 31, 10), 100)

t.test(sample1, sample2)
```

```
##
##  Welch Two Sample t-test
##
## data:  sample1 and sample2
## t = -2.2256, df = 195.65, p-value = 0.02718
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.1788107 -0.3730453
## sample estimates:
## mean of x mean of y
##  27.47551  30.75144
```

A Pitfall of P-value

In various academic disciplines, p-value is now considered a great source of misreporting a finding of an empirical study:

https://amstat.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108

A few takeaways from the article:

1. By itself, p-value does not tell us anything about the size of the effect of interest. Researchers should bring as much contextual evidence as possible to buttress the claim.

2. A certain threshold (i.e. $p > 0.05$) should not be taken for granted without any reference to the research design or the format of null hypothesis- the p-value can change even without any substantial change to the data or claims but only with those two above (so called p-hacking).

P-value and Linear Regression

The main reason for conducting linear regression is to find out the coefficient, which is basically a conditional mean of X given Y.

We test whether this conditional mean is equal to zero (one-sample, two-tailed test). This is the reason we look at p-value anyways!

But, as explained earlier, be careful of interpreting the p-value of coefficients.

Linear Regression in Scalar Form

Linear regression in scalar form is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{i2} + ... + \beta_k x_{ki} + \varepsilon_i$$

where $\epsilon \sim N(o, \sigma^2)$. This is our stochastic component. There should be no correlation between errors ($\mathbb{E}(\epsilon_i \times \epsilon_j) = 0$ for all $i \neq j$).

Meanwhile, our systematic component writes:

$$\mathbb{E}(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{i2} + ... + \beta_k x_{ki}$$

Standard Error of the Regression

The variance of our error terms, σ^2, can be written as:

$$\begin{aligned}
\sigma^2 &= \mathbb{E}((\varepsilon_i - \mathbb{E}(\varepsilon_i))^2) \\
&= \mathbb{E}((\varepsilon_i - 0)^2) \\
&= \mathbb{E}(\varepsilon_i^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \\
&= \frac{1}{n} RSS
\end{aligned}$$

where RSS is the residual sum of squares. This is referred to as the standard error of the regression.

Square Root of the Mean of the Squared Errors (RMSE)

Also note that:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2}$$

which is the square root of the mean of the squared errors (RMSE). This is how much we expect our observation (y_i) to differ from its expected value $(\mathbb{E}(y_i))$, or our systematic component of the regression.

Linear Regression in Matrix Form

Linear regression in matrix form is:

$$Y = X\beta + \epsilon$$

which can be expanded as below:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Vector of Stochastic Components

Note that we now have a vector of error terms, whose expectation is still zero:

$$\mathbb{E}(\epsilon) = \begin{bmatrix} \mathbb{E}(\varepsilon_1) \\ \mathbb{E}(\varepsilon_2) \\ \vdots \\ \mathbb{E}(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Variance-Covarance Matrix (V-Cov)

Now I introduce (perhaps) the most important concept in matrix linear regression: variance-covariance matrix. It is a $n \times n$ matrix filled with variance and covariances of error terms.

$$\Sigma = \begin{bmatrix} \text{var}(\varepsilon_1) & \text{cov}(\varepsilon_1, \varepsilon_2) & \cdots & \text{cov}(\varepsilon_1, \varepsilon_n) \\ \text{cov}(\varepsilon_2, \varepsilon_1) & \text{var}(\varepsilon_2) & \cdots & \text{cov}(\varepsilon_2, \varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\varepsilon_n, \varepsilon_1) & \text{cov}(\varepsilon_n, \varepsilon_2) & \cdots & \text{var}(\varepsilon_n) \end{bmatrix}$$

$$= \begin{bmatrix} \mathbb{E}(\varepsilon_1^2) & \mathbb{E}(\varepsilon_1 \varepsilon_2) & \cdots & \mathbb{E}(\varepsilon_1 \varepsilon_n) \\ \mathbb{E}(\varepsilon_2 \varepsilon_1) & \mathbb{E}(\varepsilon_2^2) & \cdots & \mathbb{E}(\varepsilon_2 \varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}(\varepsilon_n \varepsilon_1) & \mathbb{E}(\varepsilon_n \varepsilon_2) & \cdots & \mathbb{E}(\varepsilon_n^2) \end{bmatrix}$$

which can be reduced into:

$$\Sigma = \mathbb{E}(\epsilon \epsilon')$$

where ϵ' is a transpose of ϵ.

Variance-Covarance Matrix (V-Cov)

Note that when the assumption $\mathbb{E}(\epsilon_i \epsilon_j) = 0$ for all $i \neq j$ holds, Σ will look like:

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

Beta Coefficient

We want to obtain the value for β which minimizes the residual sum of squares, hence this method being named after “least squares”. Formally:

$$\underset{\beta}{\arg\min} \sum_{i=1}^{n} \epsilon_i^2 = \underset{\beta}{\arg\min}\ \epsilon'\epsilon$$

Although we will not derive the whole process, it is known that β can be shortened as:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

This is our least squares estimator for true β.

Good Beta, Bad Beta

A good beta estimator should have three characteristics:

1. No bias
2. Efficiency
3. Consistency

Let’s take a look at each.

Unbiased Estimator

Bias can be formally written as:

$$\mathbb{E}(\hat{\beta} - \beta)$$

That is: how much is the estimate $\hat{\beta}$ expected to be different from the true parameter β?

The more unbiased an estimator is, the closer it is to the true parameter.

Efficient Estimator

If we were to take multiple times of sample and derive a beta estimate,

how often do we get the estimate that is close to the true parameter?

This is the question of efficiency: remember, "how often". So we need to consider not only the bias but also the variance of estimator.

We calculate efficiency with mean squared error:

$$\text{MSE} = \frac{1}{n}\text{RSS} = \mathbb{E}[(\beta - \hat{\beta})^2] = Var(\hat{\beta}) + \text{Bias}$$

If there is no bias, then MSE reduces to var($\hat{\beta}$).

Consistent Estimator

An estimator is consistent when bias converges to zero as $N \rightarrow \infty$.

Although this sounds unquestionably clear, there are some cases where bias does not coverge to zero even though the sample size approaches infinity.

Yet, not a big concern relative to the two earlier concepts.

Unbiasedness and Efficiency

It is happy to have an unbiased and efficient estimator all the time. But it is not possible always. What do we mean by having (un)biased and (in)efficient estimator?

Let's say there is a sniper trying to practice in a shooting range. The sniper is the best on earth- as long as the rifle has no problem, the sniper will always send the bullets to the center.

But let's say the rifle may have two problems: ACOG(advanced combat optical gunsight) and gun barrel.

ACOG: relates to unbiasedness
Gun barrel: relates to efficiency

If a sniper shoots with good ACOG (unbiased) but bad barrel (inefficient), although the sniper will correctly aim at the bull's eye, the shot group will scatter. Nevertheless, on average a sniper has aimed at the bull's eye.

If a sniper shoots with bad ACOG (biased) but good barrel (efficient), the sniper's aim will be incorrect but the shot group will relatively less scatter.

Gauss-Markov Assumptions for Linear Regression

Even if our sample is representative of the population and the model is correctly specified (meaning we don't have any pre-treatment bias), there are still a few problems for conducting linear regression safely.

They relate to the key assumptions of linear regression to have "Best Linear Unbiased and Efficient" estimate, or BLUE. Those assumptions are:

1. No perfect collinearity
2. Exogenous covariates
3. Mean zero
4. Homoskedasticity
5. No error correlation
6. Non-normal disturbances

Let's take a look at each.

No Perfect Collinearity

Perfect collinearity occurs when $X'X$ is singular.

If this happens, $|X'X| = 0$ (determinant is zero), and β cannot be defined (since $\beta = (X'X)^{-1}X'Y$).

But no worries, usually R or other stat tools will automatically drop a few covariates to way around this issue.

Endogeneity & Exogeneity

Our linear regression in matrix form, $Y = X\beta + \epsilon$ implies that Y is endogenous to X. This means that Y is determined by X as its systematic component.

In other words, X is exogenous to Y as it is assumed to cause Y, not caused by Y. That is, all variance of X must be independent from Y while all variance of Y must be solely dependent upon X.

This second GM assumption is violated when the relationship between X and Y switches- especially when we cannot identify correctly whether X or Y is the cause of the other. $\mathbb{E}(x_i, e_i) \neq 0$

This will give us biased and inconsistent estimator.

Mean Zero

Mean zero means $\mathbb{E}(\epsilon) = 0$. If this does not hold, we are missing a few covariates that are systematic yet not included in our model. In other words, it occurs when we treated something random which in fact is not.

Also happens when there is a measurement error.

Pre-treatment bias takes part in here. Some unknown factor which caused individuals to self-select a treatment, if not included in the systematic component of the regression, will remain in the error term. Its mean will not be zero as some unknown factor has not occured *randomly*.

This will give us biased and inconsistent estimator.

Homoskedasticity

Homoskedasticity means the diagonal elements of Σ are not the same. Therefore:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

This means that how far each individual's observed outcome, y_i, is from the expected outcome predicted from the model, $\hat{y}$, is different across individuals.

If this happens, our linear regression will ignore why each individual has different distance between the predicted line and its observed value and instead **produce an inefficient (but unbiased) estimator**. This situation is called "heteroskedasticity".

No error correlation

Errors should not have correlation with each other. Formally, $\mathbb{E}(\varepsilon_i, \varepsilon_j) = 0$. If this happens along with heteroskedasticity, let's see what happens to our Σ.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

This too will give us unbiased yet inefficient estimator.

Non-Normal Disturbances

Relatively simple, this formally writes: $\epsilon \sim N(0, \sigma^2)$. If this occurs, unless $N \rightarrow \infty$ our standard error will be biased.

$$\mathbb{E}(Y|X)$$