Chapter 3

Simple Linear Regression

Flow of study

- Setting up a model of population and sample regression function
- Estimation
- Measuring goodness of fit with coefficient of determination (r^2)
- Assumption underlying Classical Linear Regression Model (CLRM)
- Reporting results
- Interval estimation
- Hypothesis testing
- Other topics

(1) *Population regression function (PRF)*

For this chapter, we are mainly using an income-consumption model for the sake of simplicity. Let's first take a look at your data dimension and let

- X_i : average monthly household income
- Y_i : average monthly household expenditure

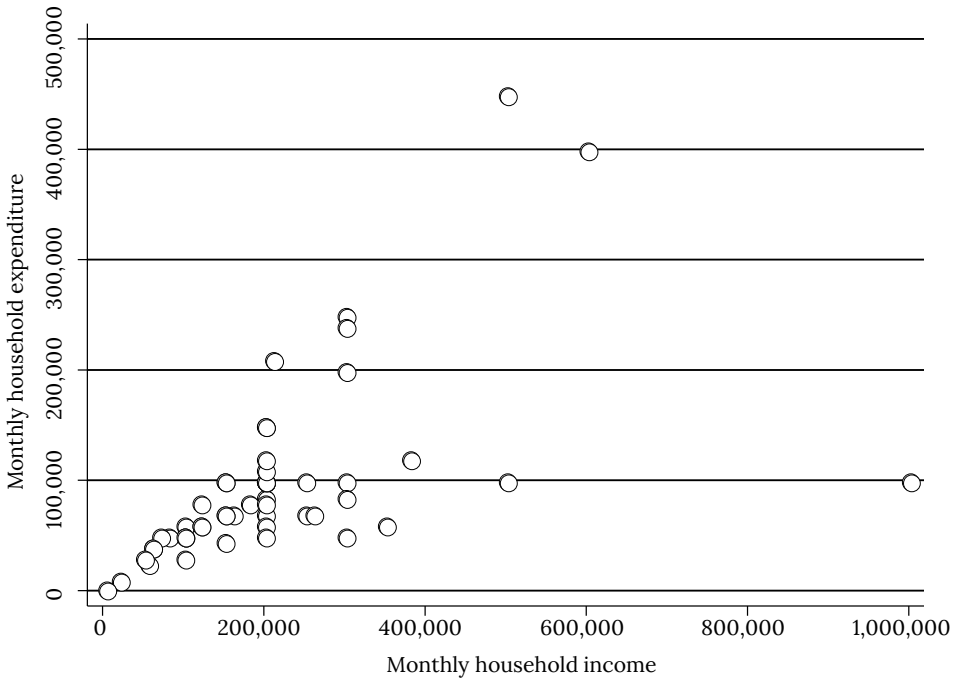
Usually, the income-consumption equation is C = a + bY where C is consumption and Y is income. We are changing this equation into the general form of

- $Y_i = \beta_1 + \beta_2 X_i$

Your data in a table and plotted as a scatter plot look like this.

(1) *Population regression function (PRF)*

Observation	X_i	Y_i
1	200,000	85,000
2	100,000	60,000
3	180,000	80,000
4	60,000	40,000
5	100,000	30,000
6	500,000	100,000
7	.	.
8	.	.
9	.	.
.	.	.
.	.	.



Total observation of this dataset (n) is 52.

(1) *Population regression function (PRF)*

Since we are estimating linear relationship between X and Y, we define a linear **population regression function** as

- $E(Y|X_i) = f(X_i) = \beta_1 + \beta_2 X_i$

So, the β_1 and β_2 are the **parameter** or

As we know that statistics relation is different from mathematical relation, therefore, we introduce a concept of the **stochastic disturbance** or **stochastic error term** as

- $u_i = Y_i - E(Y|X_i)$ or $Y_i = E(Y|X_i) + u_i$

Hence, the stochastic PRF is $Y_i = \beta_1 + \beta_2 X_i + u_i$. Now we have two parts of the PRF which are

- **Systematic** or **deterministic** part : $\beta_1 + \beta_2 X_i$.
- **Random** or **nonsystematic** part : u_i.

(2) *Stochastic specification*

Example: Let's assume that $\beta_1 = 40$ and $\beta_2 = 0.5$, figure out u_i for the following $E(Y|X_i = 180)$

- $Y_1 = 110 = 40 + 0.5(180) + u_1$ then $u_1 =$
- $Y_2 = 120 = 40 + 0.5(180) + u_2$ then $u_2 =$
- $Y_3 = 135 = 40 + 0.5(180) + u_3$ then $u_3 =$

(2) *Stochastic specification*

Why do we always have an error term in our equation? Here are some explanations:

- Vagueness of theory
- Unavailability of data

For example, using family wealth to explain consumption behavior but wealth data are usually not available.

- Core variables and peripheral variables

Variable(s) included in our model might be just peripheral ones. Picking a core variable may contribute to our model more.

- Intrinsic randomness in human behavior

(2) *Stochastic specification*

- Poor proxy variables

Some intrinsic variable cannot be observed, such as intelligence and skills. Most of the time we rely on another variable as a proxy, such as test score, GPAX, work experience, etc.

- Principle of parsimony

It is what it is if there is no strong theory suggesting adding more variable, keep our model as simple as possible and let the error terms be as they are.

- Wrong functional form

(3) *Expected value of the error term*

With the error term included in the PRF, taking the expected through this equation

- $Y_i = E(Y|X_i) + u_i$
-

Since the $E(Y|X_i)$ is a constant, therefore

-

So $E(u_i|X_i) = 0$, or we can say that

- $E(u_i|X_i) = \sum_{i=1}^{n} \left(\frac{u_i|X_i}{n}\right) = 0$

(3) *Expected value of the error term*

Similar to the property of sum of deviation from the mean, we can see the proof here that it is always zero. For instance,

- $\sum_{i=1}^{n}(x_i - \bar{X}) =$

(4) *Sample regression function (SRF)*

In real world scenario, most of the time we cannot collect all the population data. We still can define our **sample regression function** (SRF) as

- $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

The stochastic form of SRF is

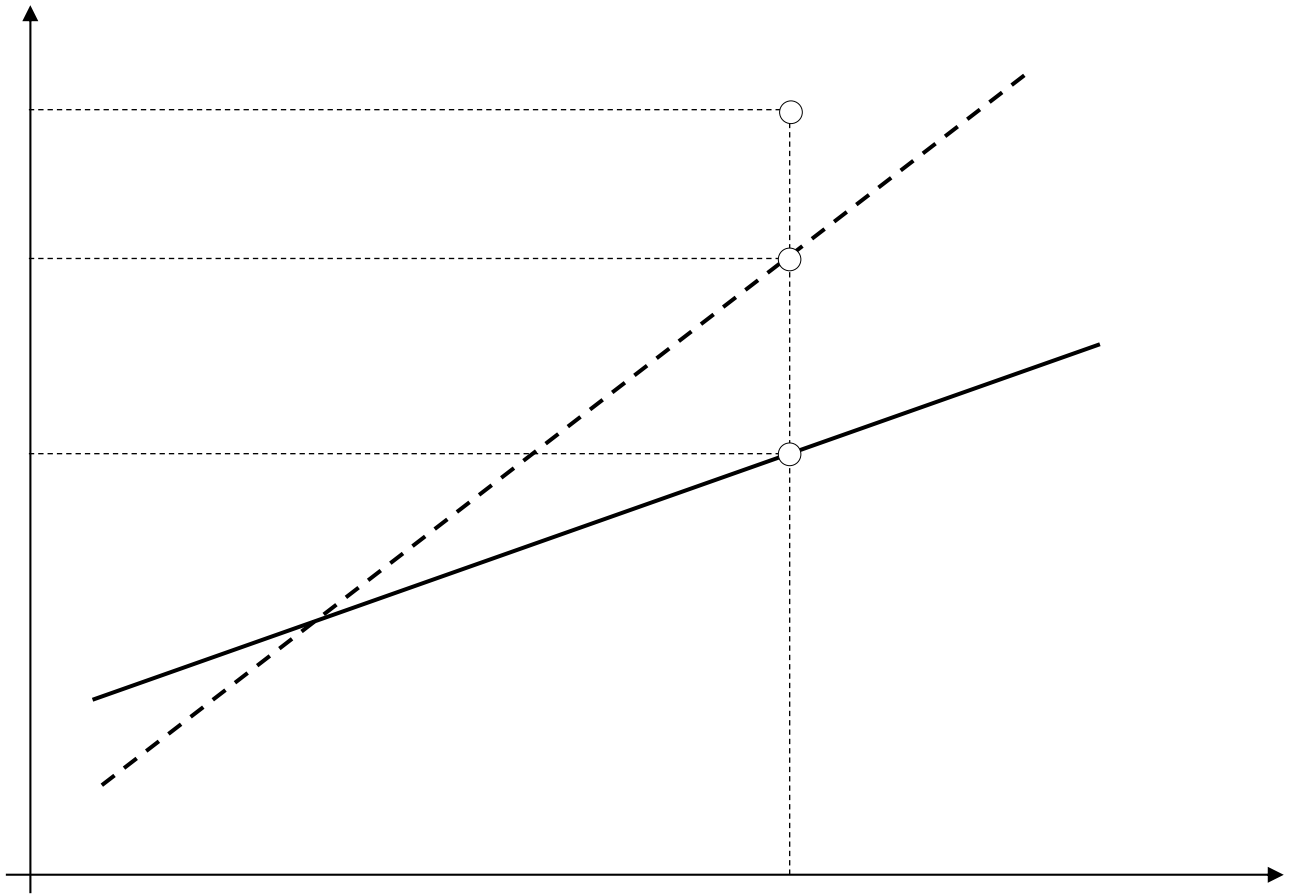
- $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$

where $\hat{Y}_i$ is the estimator of $E(Y|X_i)$

$\hat{\beta}_i$ is the estimator of β_i

$\hat{u}_i$ is the estimator of u_i

(4) *Sample regression function (SRF)*



(4) *Sample regression function (SRF)*

Sum up all the terms we know so far.

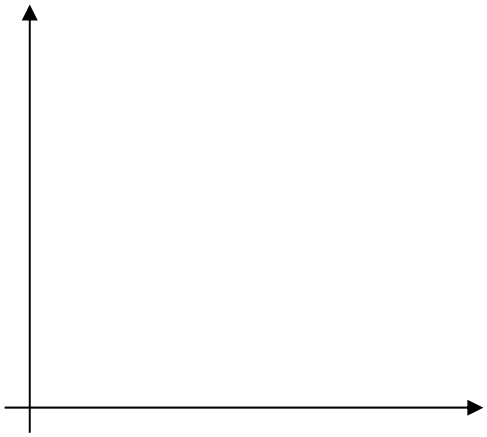
	Population	***Sample***
1. Regression line	$E(Y\|X_i) = \beta_1 + \beta_2 X_i$	$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$
2. Stochastic (data point)	$Y_i = \beta_1 + \beta_2 X_i + u_i$	$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$
3. Parameters	β_1 , β_2 (intercept, slope)	-
4. Estimators	-	$\hat{\beta}_1$, $\hat{\beta}_2$ (intercept, slope)

(5) *Note on linearity*

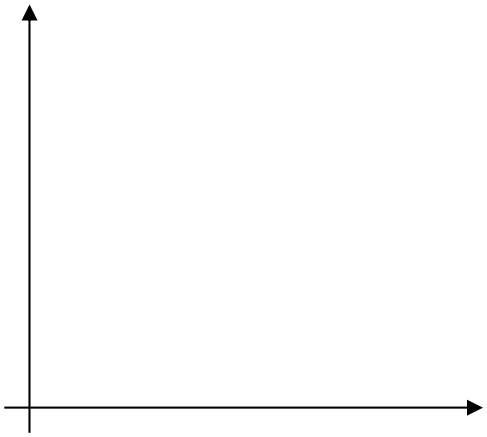
To clear things up, when we mention a linear regression model (**LRM**), we need to specify what are we talking about. There are two types of linearity.

- **Linear in variable**: a function linear or not, considered from the power of X_i.
- **Linear in parameter**: a function consist of linear parameter or not, considered from the power of β_i.

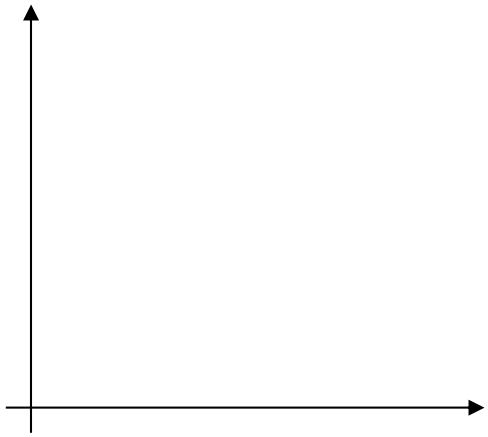
Quadratic



Exponential



Cubic



(1) *Problem statement*

When we obtain a set of data, plot them on a graph, how can we draw a straight line, portraying regression relationship between two variables?

- How to connect each data point with a line that fits best with our data?
- What is/are (a) criteria (ion) that we can rely on?

(2) *Ordinary Least Squares (OLS)*

This method applies for both PRF and SRF. The intuition is we try to draw a linear line that minimize sum of the error terms. From

- $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$

Rearranging the equation, we get

-

Setting up the objective function

-

(2) *Ordinary Least Squares (OLS)*

Solve $\hat{\beta}_1$

(2) *Ordinary Least Squares (OLS)*

Plug $\hat{\beta}_1$ to solve for $\hat{\beta}_2$

(2) *Ordinary Least Squares (OLS)*

$\sum X_i(Y_i - \bar{Y}) = \sum(X_i - \bar{X})(Y_i - \bar{Y})$ because

Eventually, we get

- $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}$
- $\hat{\beta}_2 = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2}$

where $x_i = (X_i - \bar{X})$, $y_i = (Y_i - \bar{X})$, $x_i^2 = (X_i - \bar{X})^2$

(3) *Properties of OLS estimators*

(1) The OLS estimators are expressed solely in terms of the observables.

(2) They are **point estimators**, instead of interval estimators.

(3) They make the SRF passes through the sample mean.

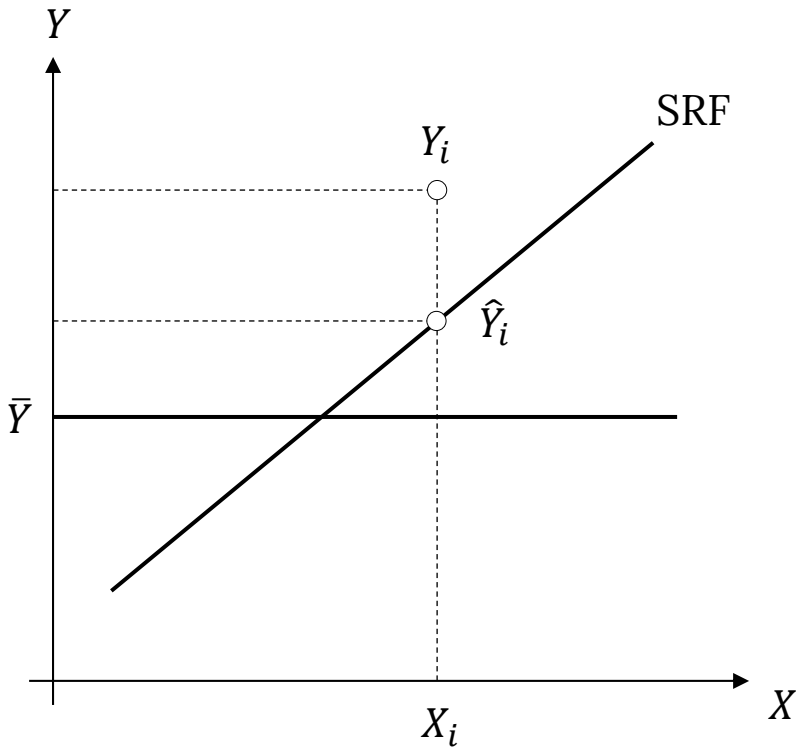
(4) The mean value of $\hat{Y}_i$or $\bar{\hat{Y}} = \bar{Y}$.

(5) The mean value of the residual $\hat{u}_i = 0$.

(6) $\hat{u}_i$ are uncorrelated with both X and $\hat{Y}$.

(1) *Coefficient of determination (r^2)*

The r^2 is determined by how much the is described by the SRF, or the measurement of **'goodness of fit'** of the fitted regression line comparing to an estimator, $\bar{Y}$.



The intuition is that total sum of squares (TSS) is equal to explained sum of squares (ESS) and residual sum of squares (RSS) or

○

(1) *Coefficient of determination (r^2)*

Eventually, we get

- $r^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$ or
- $r^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2}$

There are a few more formulae of r^2 in page 76.

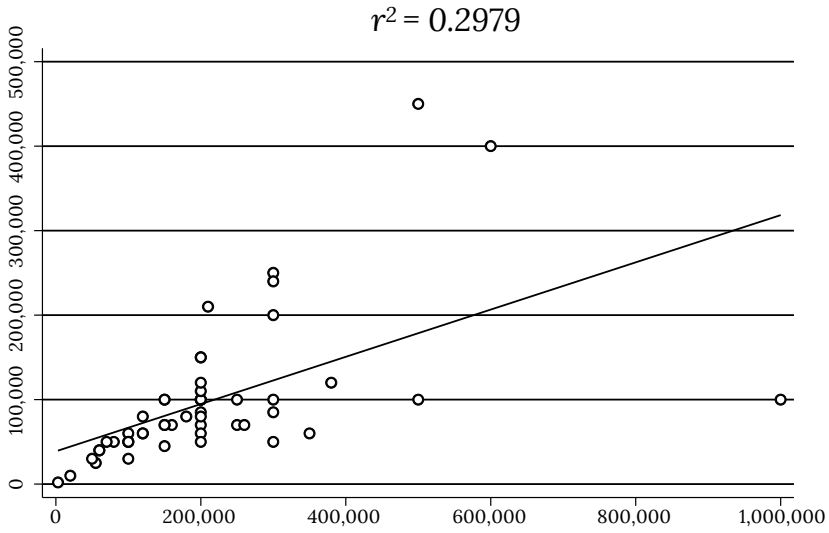
Properties of r^2

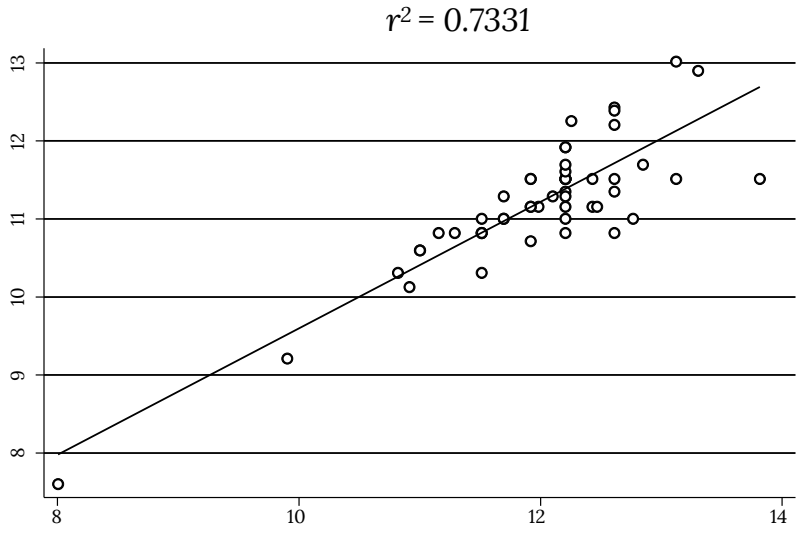
(1) Non-negativity

(2) $0 \leq r^2 \leq 1$

(1) *Coefficient of determination (r^2)*

Example of different r^2 that represent different level of fitting





(2) *Sample coefficient of correlation (r)*

A formula of r^2 is

- $r^2 = \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2}$

We can define **sample coefficient of correlation** (r) easily by

- $r = \pm\sqrt{r^2} = \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}}$

Properties of r

(1) Positive or negative depending on the sign of the term.

(2) $-1 \leq r \leq 1$

(3) Independent of the origin and scale.

(4) If X and Y are statistically independent, $r = 0$, but **not** vice versa.

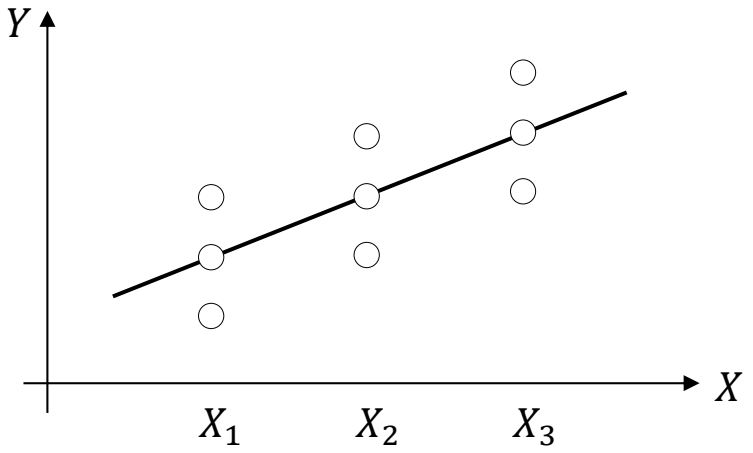
(5) Does not describe non-linear association or causality.

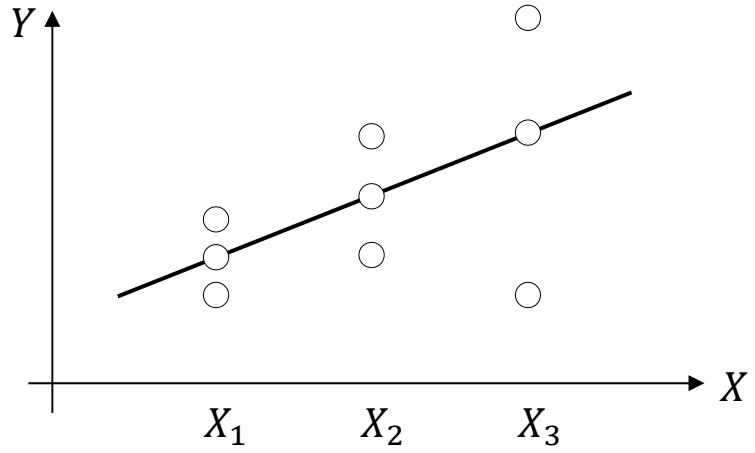
(1) *Linear in parameters*: which is already discussed.

(2) *Exogeneity*: or $cov(X_i, u_i) = 0$ or X_i are independent of the error term. It is an important assumption, but not very relevant in this class since when this assumption is relaxed many statistical properties are still valid.

(3) *Zero mean value of disturbance* u_i: basically, sum of deviation from the mean is always zero, which also implies that there is no **specification error** or **specification bias** or $E(u_i|X_i) = E(u_i) = 0$.

(4) *Homoscedasticity*: constant variance of u_i across X_i or $\text{var}(u_i|X_i) = \sigma^2$.

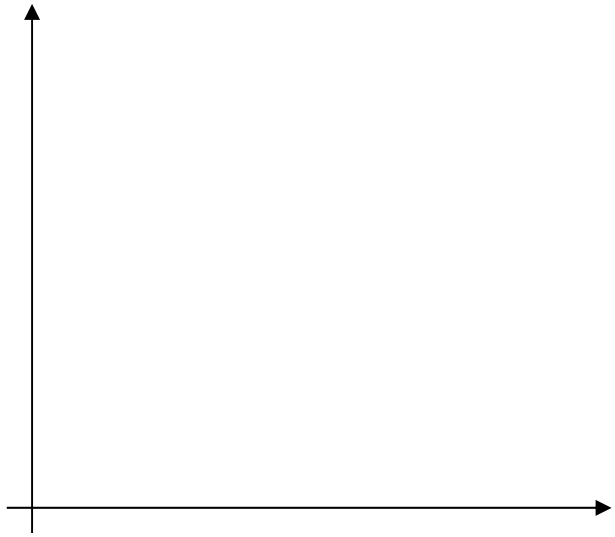




(5) No *autocorrelation*: or serial correlation between disturbances or $cov(u_i, u_j | X_i, X_j) = 0$ or $cov(u_i, u_j) = 0$ if X is non-stochastic.

(6) *Number of observation* $\boldsymbol{n}$ *must be greater than parameter estimated* $\boldsymbol{k}$.

(7) X_i *not be all the same.*



(1) *Traditional method*

Assumed we have estimated the coefficients, we shall consider how an estimation is reported. There can be multiple ways to do so. Let's first define our model, given that

- $consmp_i = \hat{\beta}_1 + \hat{\beta}_2 inc_i + \hat{u}_i$

where $consmp_i$ is household expenditure of student i

inc_i is household income of student i (monthly).

Report the estimators in the equation as follows.

- $\widehat{consmp}_i = 38{,}687.86 + 0.2797 inc_i$

se	(16,223.66)	(0.0607)	$r^2 = 0.2979$
t	(2.38)	(4.61)	$n = 52$
p	(0.021)	(0.000)	$F_{1,50} = 21.22$

(2) STATA *method*

```
      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(1, 50)        =     21.22
       Model |  1.0675e+11         1  1.0675e+11   Prob > F        =    0.0000
    Residual |  2.5157e+11        50  5.0314e+09   R-squared       =    0.2979
-------------+----------------------------------   Adj R-squared   =    0.2839
       Total |  3.5832e+11        51  7.0259e+09   Root MSE        =     70932

------------------------------------------------------------------------------
         exp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         inc |   .2797096   .0607242     4.61   0.000     .1577414    .4016778
       _cons |   38687.86   16223.66     2.38   0.021      6101.68    71274.03
------------------------------------------------------------------------------
```

(3) *Modern method*

Report the estimators and its significance using asteroids. Other information can be put anywhere as you wish.

(Dependent variable: household expenditure) **Independent variables**	
Household income	0.2797*** (0.0607)
constant	38,687.86** (16,223.66)
r^2	0.2979
n	52

Note: */**/*** denotes statistical significance at 90, 95 and 99 percent respectively.

(3) *Modern method*

Table 4: Regression Results

(Dependent variable: hours worked) Independent variables	Estimation part: Labor participation – Probit (Marginal effect)	Estimation part: Labor supply – Truncated regression (Coefficient)
1. Log earned income	1.327** (0.331)	12.389** (1.262)
2. Female # Log earned income	1.344 (1.148)	-5.767** (1.775)
3. Married # Log earned income	-0.913* (0.402)	-5.556** (1.363)
4. Female # married # Log earned income	-2.336 (2.012)	5.352* (2.028)
5. Unearned income x 100 [1]	-0.874** (0.150)	-1.080** (0.276)
6. Female # Unearned income x 100	0.119** (0.026)	0.083 (0.065)
7. Married # Unearned income x 100	-0.064 (0.040)	0.054 (0.065)
8. Female # Married # Unearned income x 100	0.081 (0.058)	-0.060 (0.080)
9. 10th decile # unearned income x 100 [2]	0.684** (0.148)	1.029** (0.274)
Region (base case: Bangkok)		
10. Central	0.722 (0.633)	9.928** (1.028)
11. North	1.683* (0.686)	-1.342 (1.247)
12. Northeast	4.662** (0.681)	-3.690** (1.115)
13. South	3.865** (0.716)	-25.091** (1.239)
Municipal area (base case: municipal)		
14. non-municipal	0.850** (0.329)	-13.054** (1.239)
Sex and marital status (base case: single male)		
15. Female	-6.666** (0.525)	58.419** (16.238)
16. Married	11.202** (0.708)	55.425** (12.674)
17. Female # married	-8.445** (0.736)	-53.717** (18.764)
Individual characteristics		
18. Year of education	0.334** (0.047)	-1.219** (0.095)
19. Age	6.032** (0.075)	0.998** (0.240)
20. Age squared	-0.072** (0.001)	-0.019** (0.003)
21. Number of children aged 0-6 in household	0.308 (2.312)	-1.154 (0.906)
22. Female # Number of children aged 0-6 in household	-4.001** (0.383)	0.337 (1.105)
23. Disability	-36.626** (1.730)	2.895 (3.644)
Constant	-4.697** (0.133)	93.248** (11.467)
Classification / Sigma	83.62	51.482** (0.263)

Note: 1) Since the effect is tiny, unearned income is multiplied with 100.
2) Only the tenth decile interacted with unearned income in a significant manner. Other deciles are controlled, but are not shown here for table concision.
*/** denotes statistical significance at 90 and 95 percent, respectively. # sign refers to the interaction term. No serious collinearity is detected and robust standard error is displayed in parentheses.

(1) *Problem statement*

Multiple samplings yield different estimators, varied values of $\hat{\beta}_1$ and $\hat{\beta}_2$. If we only have a sample, how reliable the point estimation is?

Consider this example, let's first assume that a Socio-Economic Survey dataset from 2019 ***is the population***. The dataset is a national collection of household income and expenditure.

(1) Now let's randomly divide all households into 5 groups. These are assumed ***to be each sample***. There are 45,586 households (observations) in total and each subgroup consists of roughly 9,000 households.

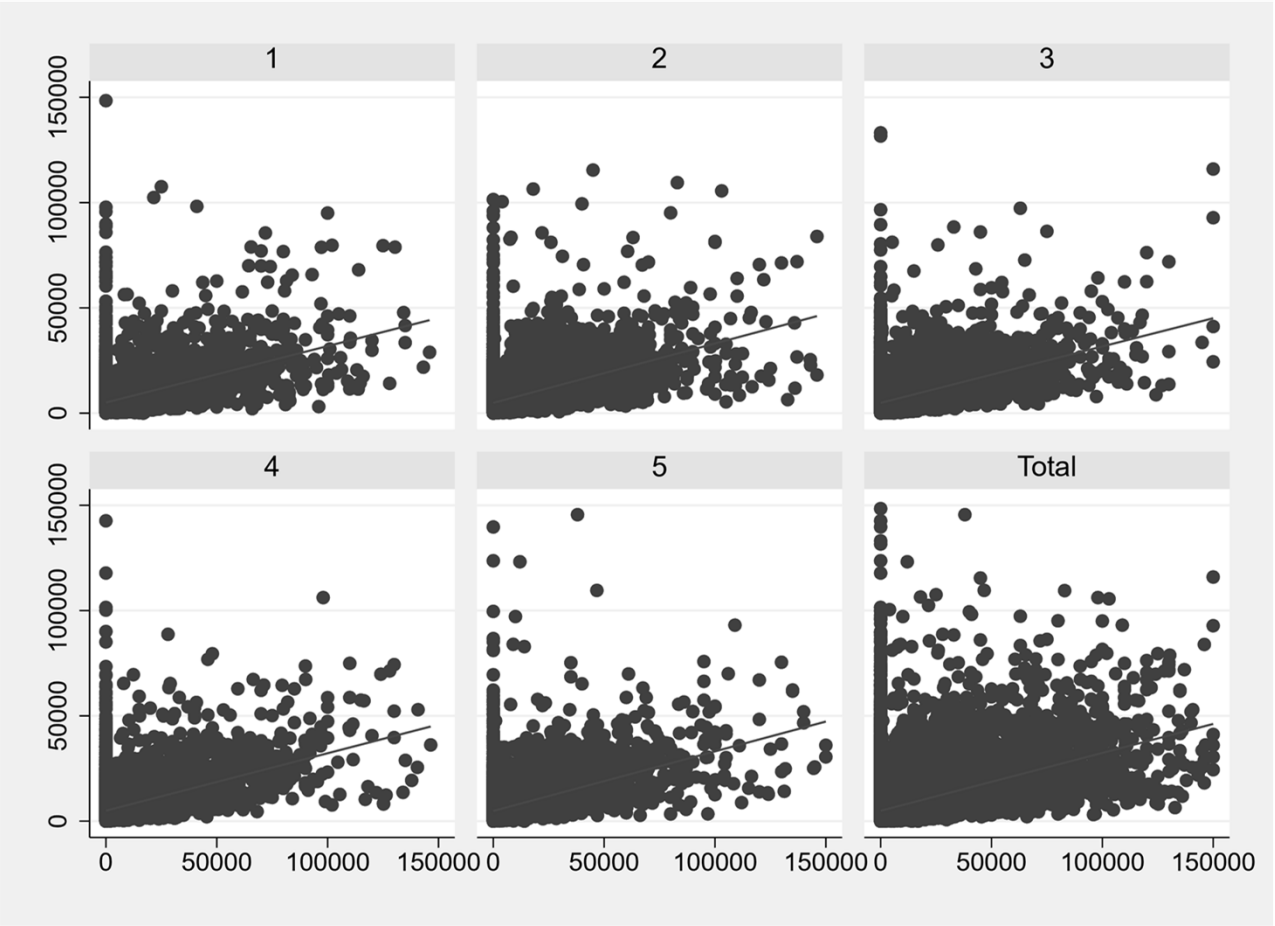
(2) Next, we estimate each subgroup and the whole dataset. We shall see 6 results of $\hat{\beta}_1$ and $\hat{\beta}_2$ in the next page. Note that β_1 and β_2 are used for the whole dataset since we assume that it represents the population.

(1) Problem statement

$\hat{\beta}_1 = 5{,}198.02\ (101.06)$
$\hat{\beta}_2 = 0.2532\ (0.0050)$
$n = 8{,}871$

$\hat{\beta}_1 = 5{,}190.39\ (111.03)$
$\hat{\beta}_2 = 0.2574\ (0.0054)$
$n = 9{,}260$

$\hat{\beta}_1 = 5{,}550.89\ (98.25)$
$\hat{\beta}_2 = 0.2035\ (0.0041)$
$n = 9{,}162$



$\hat{\beta}_1 = 5{,}200.37\ (96.21)$
$\hat{\beta}_2 = 0.2463\ (0.0048)$
$n = 9{,}047$

$\hat{\beta}_1 = 5{,}010.31\ (117.99)$
$\hat{\beta}_2 = 0.2735\ (0.0059)$
$n = 9{,}246$

$\beta_1 = 5{,}271.91\ (47.02)$
$\beta_2 = 0.2424\ (0.0022)$
$n = 45{,}586$

(2) *Further assumptions: Gauss-Markov Theorem*

With certain assumptions imposed, OLS estimators is considered **BLUE** (best linear unbiased estimators)

(1) It is **linear**, that is, a linear function of a random variable, such as the dependent variable Y in the regression model.

(2) It is **unbiased**, that is, its average or expected value, $E(\hat{\beta}_2)$ is equal to the true value β_2 (multiple sampling).

- An estimator is considered unbiased when $E(\hat{\theta}) = \theta$

(3) It has minimum variance in the class of all such linear unbiased estimators: an unbiased estimator with the least variance is known as an **efficient estimator**.

- An estimator $\hat{\theta}_i$ is more efficient than $\hat{\theta}_j$ when $\text{var}(\hat{\theta}_i) < \text{var}(\hat{\theta}_j)$.

(2) *Further assumptions: Normality of* u_i

Assumed that each u_i is distributed normally with

- Mean : $E(u_i) = 0$
- Variance : $E[u_i - E(u_i)]^2 = E(u_i^2) = \sigma^2$
- Covariance : $E(u_i, u_j) = 0$

We can write it shortly with this notation

- $u_i \sim \text{NIID}(0, \sigma^2)$

or the error term (disturbance) is ***normally, identically and independently distributed*** with 0 mean and σ^2 variance.

(2) *Further assumptions: Central Limit Theorem (CLT)*

Let $X_1, X_2, \ldots, X_n$ denote n independent random variables distributed with the same PDF with mean of μ and variance of σ^2, as n increases indefinitely, then

- $\bar{X}_{n\to\infty} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$ regardless of the form of PDF.

We can then transform into a standard normal distribution as

- $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim \mathrm{N}(0,1)$

(2) *Further assumptions: variance of the model*

Sampling will not yield the true value of σ^2, we also have no way to know its true value, therefore we can estimate it from

- $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-k} = \frac{\sum (Y_i - \hat{Y})^2}{n-k}$

where $\sum \hat{u}_i^2$ is residual sum of squares and n-k is the degrees of freedom. k is number of parameters estimated, in this case it is 2 ($\hat{\beta}_1, \hat{\beta}_2$).

This $\hat{\sigma}^2$ can be found in the estimation result from STATA, known as **residual mean squares**. It will later be used for interval estimation and hypothesis testing, as and estimator of true variance.

(3) *Summary of all properties*

For the estimators $\hat{\beta}_1, \hat{\beta}_2$ with all the assumptions imposed, we then have

- Mean : $E(\hat{\beta}_1) = \beta_1$
- Variance : $\sigma^2_{\hat{\beta}_1} = \frac{\sum X_i^2}{n \sum x_i^2} \sigma^2$

Or more compactly $\hat{\beta}_1 \sim \mathrm{N}(\beta_1, \sigma^2_{\hat{\beta}_1})$

- Mean : $E(\hat{\beta}_2) = \beta_2$
- Variance : $\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum x_i^2}$

Or more compactly $\hat{\beta}_2 \sim \mathrm{N}(\beta_2, \sigma^2_{\hat{\beta}_2})$

We can also turn these distributions into standard normal as

- $Z = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{se}_{\hat{\beta}_1}} \sim \mathrm{N}(0,1)$ and $Z = \frac{\hat{\beta}_2 - \beta_2}{\mathrm{se}_{\hat{\beta}_2}} \sim \mathrm{N}(0,1)$

The covariance between these estimators is

- $cov(\hat{\beta}_1, \hat{\beta}_2) = -\bar{X}\left(\frac{\sigma^2}{\sum x_i^2}\right) = -\bar{X} \cdot \mathrm{var}(\hat{\beta}_2)$

Lastly, since σ^2 is not known, we can plug in $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-k}$ instead.

(4) *Confidence interval (CI)*

Once we already set up many assumptions prior to this point, we can utilize them to construct an **interval estimation** by setting up a **confidence interval (CI)**. An example of β_2 is displayed below.

- $P\left[\hat{\beta}_2 - \delta \leq \beta_2 \leq \hat{\beta}_2 + \delta\right] = 1 - \alpha$

where $1 - \alpha$ is confidence coefficient

α is level of significance

$\hat{\beta}_2 - \delta$ is lower limit and $\hat{\beta}_2 + \delta$ is upper limit

For example, if α is 0.05, it means that we are looking for the range of β_2 that includes $1 - \alpha = 1 - 0.05 = 0.95$ or 95% probability.

(4) *Confidence interval (CI)*

Since σ^2 is rarely known, we can t stat instead of Z and replace σ with $\hat{\sigma}$.

- $t = \frac{\hat{\beta}_2 - \beta_2}{\text{se}_{\hat{\beta}_2}}$

Next, replacing this t into the CI, we get

- $P\left[-t_{\frac{\alpha}{2}} \leq \frac{\hat{\beta}_2 - \beta_2}{\text{se}_{\hat{\beta}_2}} \leq t_{\frac{\alpha}{2}}\right] = 1 - \alpha$

Rearranging the term to specify the upper and lower limit

- $P\left[\hat{\beta}_2 - \left(t_{\frac{\alpha}{2}} \cdot \text{se}_{\hat{\beta}_2}\right) \leq \beta_2 \leq \hat{\beta}_2 + \left(t_{\frac{\alpha}{2}} \cdot \text{se}_{\hat{\beta}_2}\right)\right] = 1 - \alpha$

In other words, $100(1-\alpha)\%$ CI for β_2 is

- $\hat{\beta}_2 \pm t_{\frac{\alpha}{2}} \cdot \text{se}_{\hat{\beta}_2}$ and analogously for β_1, $\hat{\beta}_1 \pm t_{\frac{\alpha}{2}} \cdot \text{se}_{\hat{\beta}_1}$

(4) *Confidence interval (CI)*

Consider the same result below as an example.

- $\widehat{consmp}_i = 38{,}687.86 + 0.2797 inc_i$

se	(16,223.66)	(0.0607)	$r^2 = 0.2979$
t	(2.38)	(4.61)	$n = 52$
p	(0.021)	(0.000)	$F_{1,50} = 21.22$

To find a confidence interval of β_1 or β_2, follow these steps.

- ***Step 1***: pick an α.

Usually, acceptable levels of significance are 0.01, 0.05 or 0.1, depending on how many percent that you are going to cover.

For example, if we pick $\alpha = 0.05$, the most common value, it means that the CI that we are going to find will indicate that 95% of the time,
the upper and lower limit will contain the true value of β_2.

(4) *Confidence interval (CI)*

- ***Step 2***: look up for $t_{\frac{\alpha}{2}}$.

Since we assumed that estimators are normally distributed but we do not know the true value of variance, we can use t distribution instead.

Be aware that the degrees of freedom must match our model $(n - k)$.

Also, t value must coincide with selected α. We are constructing the upper and lower limit. Therefore, t is spread symmetrically to the left and the right of the mean.

(4) *Confidence interval (CI)*

- ***Step 3***: calculate the upper and lower limit

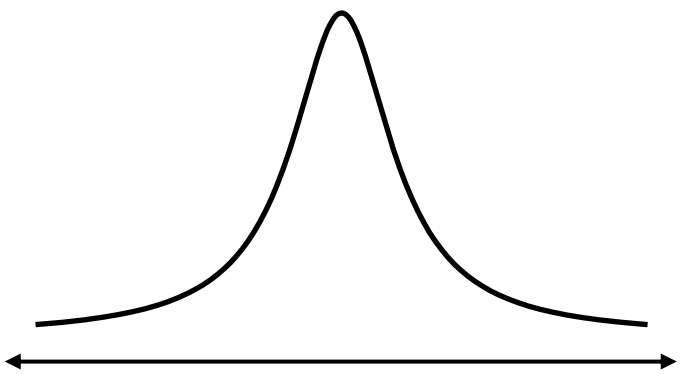
Now we can proceed to calculate the limit, using the formula

- Upper limit : $\hat{\beta}_2 + t_{\frac{\alpha}{2}} \cdot \text{se}_{\widehat{\beta}_2} =$
- Lower limit : $\hat{\beta}_2 - t_{\frac{\alpha}{2}} \cdot \text{se}_{\widehat{\beta}_2} =$

You may probably notice that we use $\hat{\beta}_2$ as a mean since we assumed that $E(\hat{\beta}_2) = \beta_2$. To write it in probability form, on the other hand, we might use this notation.

- $P\left[\hat{\beta}_2 - \left(t_{\frac{\alpha}{2}} \cdot \text{se}_{\widehat{\beta}_2}\right) \leq \beta_2 \leq \hat{\beta}_2 + \left(t_{\frac{\alpha}{2}} \cdot \text{se}_{\widehat{\beta}_2}\right)\right] = 1 - \alpha$
-

(4) *Confidence interval (CI)*



To illustrate the confidence interval, look at this graphical representation on the left.

Our calculation shows that where the upper and lower limit for the true β_2 are.

You may also try practicing finding the confidence interval of β_1.

(4) *Confidence interval (CI)*

We can also find the confidence interval of σ^2 as well. This is, theoretically, important since the standard error for $\hat{\beta}_1$ and $\hat{\beta}_2$ is calculated from $\hat{\sigma}^2$.

Under the normality assumption, we assume that the error term is normally distributed. Hence, their squares are distributed as chi-square.

- $(n-k)\cdot\frac{\hat{\sigma}^2}{\sigma^2}\sim\chi^2$, we can construct the interval for σ^2 as
- $P\left[(n-k)\cdot\frac{\hat{\sigma}^2}{\chi^2_{\frac{\alpha}{2}}}\leq\sigma^2\leq(n-k)\cdot\frac{\hat{\sigma}^2}{\chi^2_{1-\frac{\alpha}{2}}}\right]=1-\alpha$

Be very careful that chi-square ***is not*** a symmetric distribution..

(4) *Confidence interval (CI)*

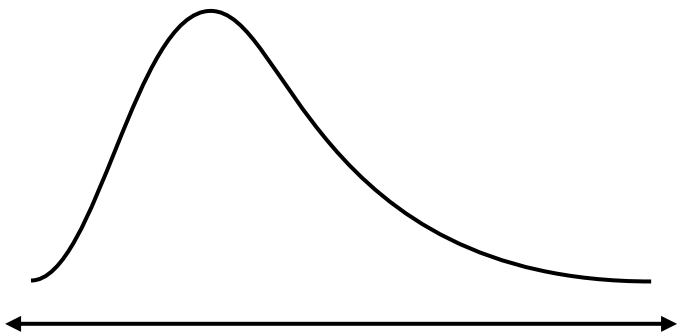
Example: Find the 95% CI of σ^2 **given that** $\hat{\sigma}^2 = 7{,}389{,}147$ and $n = 45$.

- ***Step 1***: pick an α.
- ***Step 2***: look up for $\chi^2_{\frac{\alpha}{2}}$ and $\chi^2_{1-\frac{\alpha}{2}}$ in Chi-square table.
- ***Step 3***: calculate the upper and lower limit

Lower limit: $(n-k) \cdot \dfrac{\hat{\sigma}^2}{\chi^2_{\frac{\alpha}{2}}} =$

Upper limit: $(n-k) \cdot \dfrac{\hat{\sigma}^2}{\chi^2_{1-\frac{\alpha}{2}}} =$

(4) *Confidence interval (CI)*



The confidence interval for σ^2 is the area that the area of selected probability, the true value of σ^2 will be within this area.

(1) *Two tails test*

We, then, apply the concept of concept of confidence interval to test a parameter whether it is, is not, more or less than the value of interest.

For instance, usually we are interested in, almost every parameter, if it is zero or not with a level of significance because if they, or some of them, are zero, it does not represent any statistical relevance between two random variables.

This kind of hypothesis testing is then a *two tails test* since we are not interested in its value. We just want to know if it is zero or not zero.

Note: any parameter that is not zero, we usually say that it is significantly different from zero (or in Thai we say just ***'sig'***).

Let's follow the same example and these steps.

(1) *Two tails test*

- $\widehat{consmp}_i = 38{,}687.86 + 0.2797 inc_i$

se	(16,223.66)	(0.0607)	$r^2 = 0.2979$
t	(2.38)	(4.61)	$n = 52$
p	(0.021)	(0.000)	$F_{1,50} = 21.22$

- ***Step 1***: state a hypothesis.

$H_0: \beta_2 = 0$ – Null hypothesis

$H_a: \beta_2 \neq 0$ - Alternative hypothesis

- ***Step 2***: pick an α.

Our α is a ***level of significance***. In other words, we are trying to test that the parameter β_2 is significantly different from zero or not $1 - \alpha$ of the time that we sample.

(1) *Two tails test*

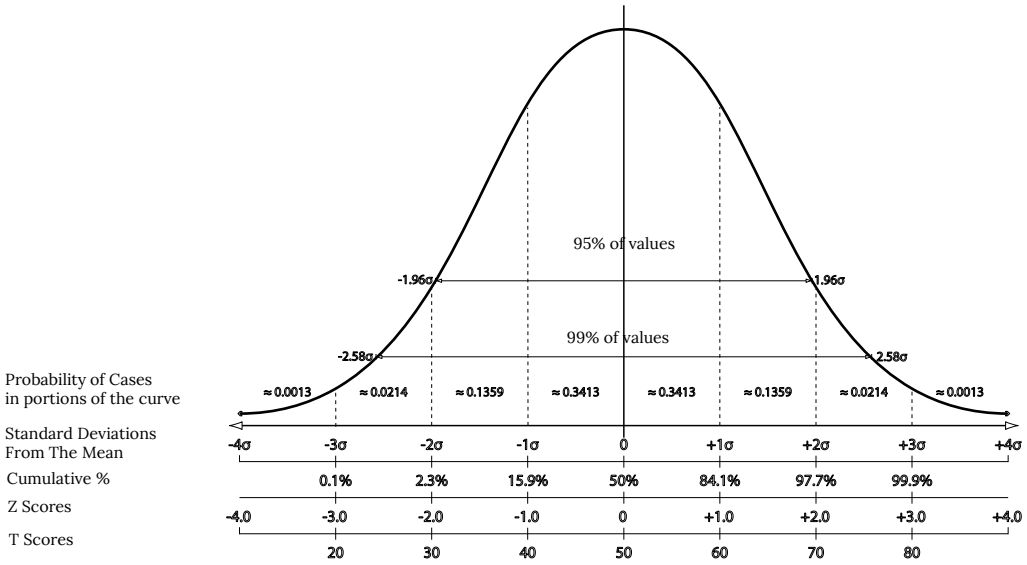
- **Step 3**: calculate test statistics.

$$t_{cal} = \frac{\hat{\beta}_2 - \beta_2}{\text{se}_{\hat{\beta}_2}} =$$

What we are doing here is that we are trying to figure out the distance between the estimator $\hat{\beta}_2$ and our test value β_2 dividing by the standard error of $\text{se}_{\hat{\beta}_2}$. In other words, we are trying to normalize t_{cal} to make this statistics comparable to the distribution of β_2 at test value.

(1) *Two tails test*

Testing against zero is the simplest one because it is very much like normalizing a normal distribution into a standard normal distribution.



(1) *Two tails test*

- ***Step 4***: state the decision rule

Now we pick an α to have an acceptable probability, most of the time we use $\alpha = 0.05$. Use this information to create CI around $\beta_2 = 0$, based on the degrees of freedom, to see how much the distribution covers.

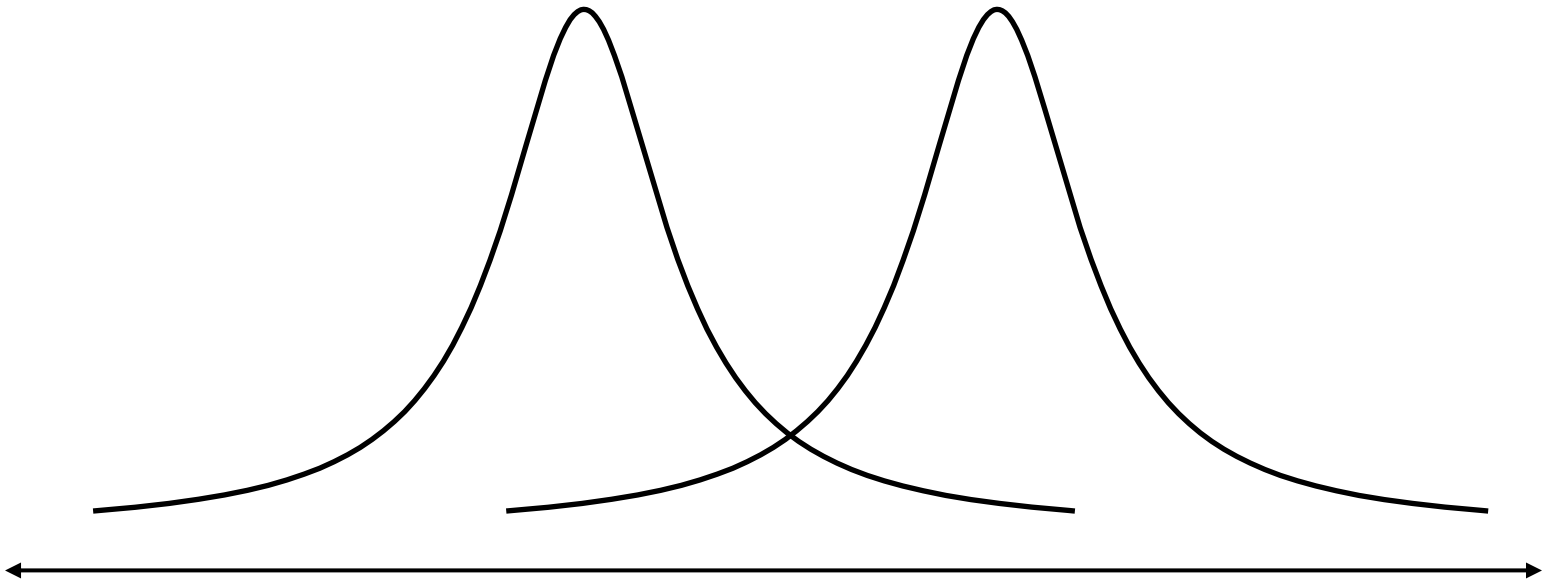
The upper bound : $t_{\frac{\alpha}{2}} =$

The lower bound : $t_{\frac{\alpha}{2}} =$

Note that when we test against zero ($\beta_2 = 0$), the distribution is normalized, therefore, there is no need to transform t from the table value.

(1) *Two tails test*

The test we are performing here is like a comparison between the distribution of the estimator and the test value as illustrated below.



(1) *Two tails test*

- ***Step* 5**: conclude the test.

- If t_{cal} lies **beyond** any boundary of test value CI (critical region), we **can reject the null hypothesis**, at the significance level of 95%.

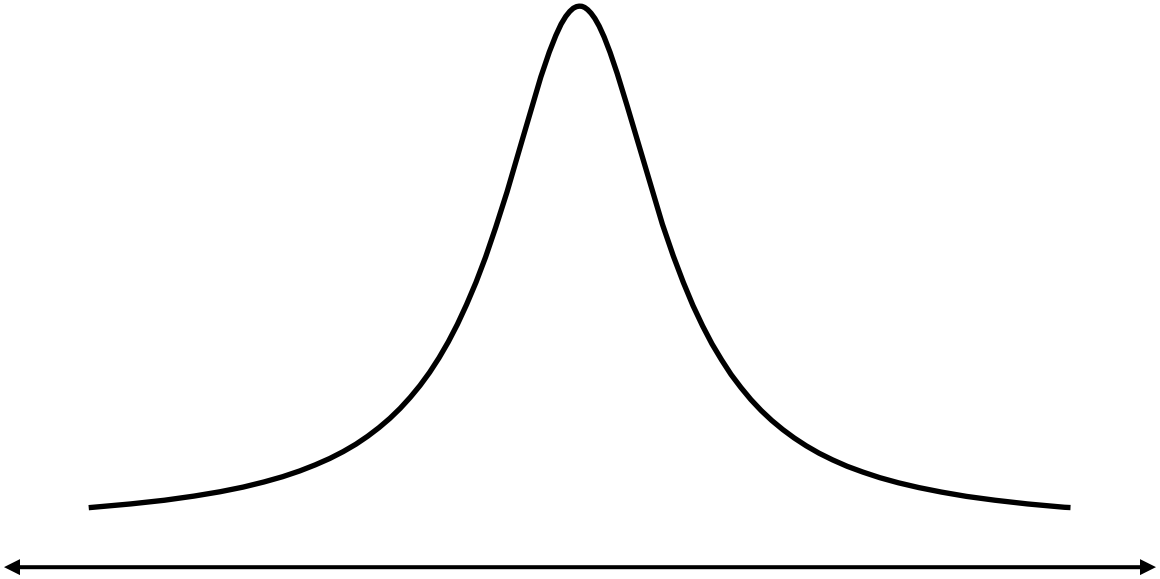
In other words, **we are sure** that β_2 is not zero 95 out of 100 times when we sample.

- If t_{cal} lies **within** any boundary of test value CI (acceptance region), we **cannot reject the null hypothesis**, at the significance level of 95%.

In other words, we **cannot say for sure** that β_2 is not zero 95 out of 100 times when we sample.

(1) *Two tails test*

Conclude this test illustrating the result.



(1) *Two tails test*

Let's now try to test against other value.

- $\widehat{consmp}_i = 38{,}687.86 + 0.2797 inc_i$

se	(16,223.66)	(0.0607)	$r^2 = 0.2979$
t	(2.38)	(4.61)	$n = 52$
p	(0.021)	(0.000)	$F_{1,50} = 21.22$

- **Step 1**: state a hypothesis.

$H_0: \beta_2 = 0.2$ – Null hypothesis

$H_a: \beta_2 \neq 0.2$ - Alternative hypothesis

- **Step 2**: pick an α.

(1) *Two tails test*

- ***Step 3***: calculate test statistics.

$$t_{cal} = \frac{\hat{\beta}_2 - \beta_2}{\text{se}_{\widehat{\beta}_2}} =$$

- ***Step 4***: state the decision rule

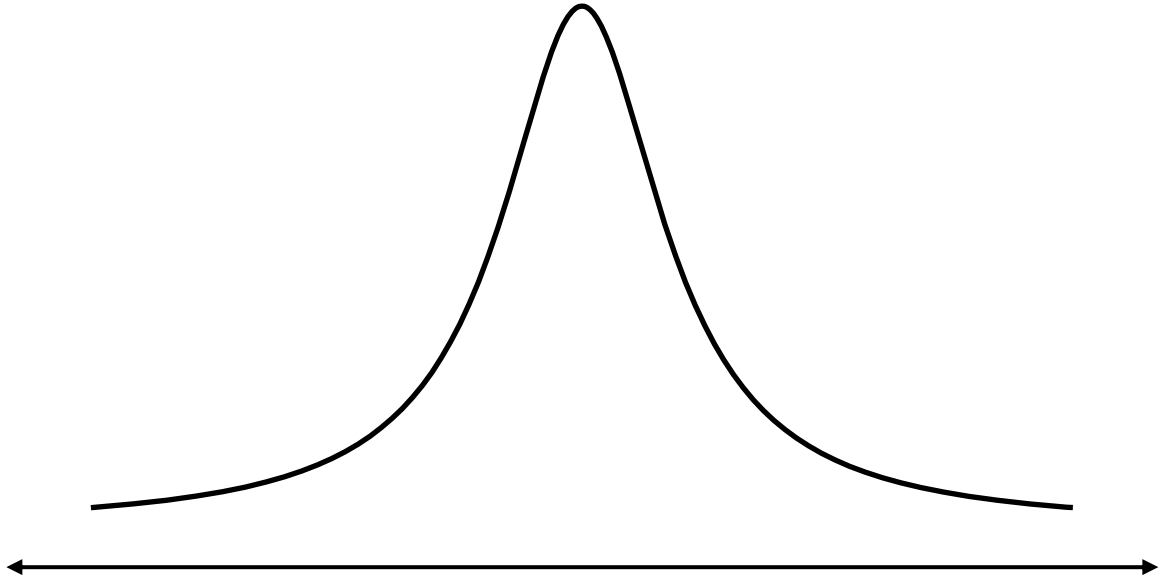
Note that now the test value is not distributed around zero. Therefore, the upper and lower limit must be scaled to distribute around 0.2 instead.

The upper bound : $\beta_2 + t_{\frac{\alpha}{2}} \cdot \text{se}_{\widehat{\beta}_2} =$

The lower bound : $\beta_2 - t_{\frac{\alpha}{2}} \cdot \text{se}_{\widehat{\beta}_2} =$

(1) *Two tails test*

- **Step 5**: conclude the test.



(2) *One tail test*

For a one tail test, we are trying to test whether the parameter is more or less than a specific value of interest or not. There might be some confusion according to flexible hypothesis setup. Consider the same example.

- $\widehat{consmp}_i = 38{,}687.86 + 0.2797 inc_i$

se	(16,223.66)	(0.0607)	$r^2 = 0.2979$
t	(2.38)	(4.61)	$n = 52$
p	(0.021)	(0.000)	$F_{1,50} = 21.22$

Let's imagine that we want to make sure if β_2 is ***less than 0.3*** or not. Note that when we test that β_2 is not zero, we put this in the alternative hypothesis. Therefore, it is quite logical to follow this rule.

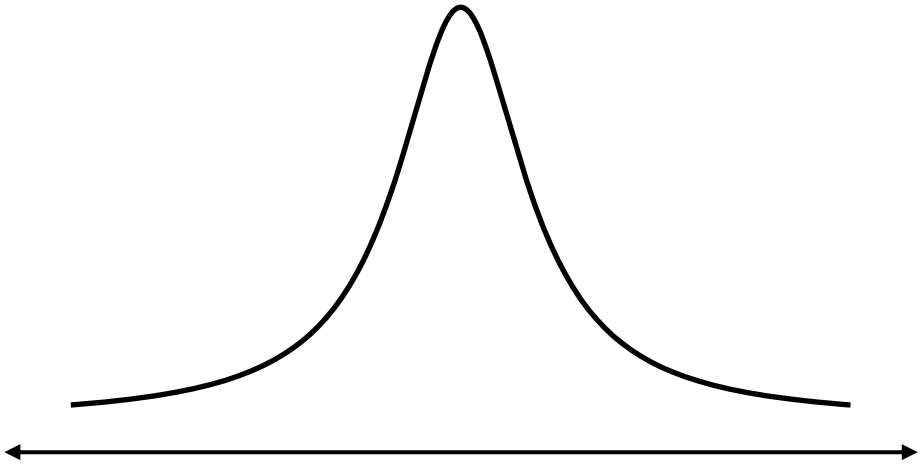
(2) *One tail test*

- ***Step 1***: state a hypothesis.

$H_0: \beta_2 \geq 0.3$ – Null hypothesis

$H_a: \beta_2 < 0.3$ - Alternative hypothesis

As mentioned earlier, we put the result of interest in the alternative hypothesis. We want to make sure if β_2 is ***less than 0.3*** or not. This is called ***left-tail test*** since the rejection area is on the left of the distribution.



(2) *One tail test*

- ***Step 2***: pick an α.
- ***Step 3***: calculate test statistics.

$$t_{cal} = \frac{\hat{\beta}_2 - \beta_2}{\text{se}_{\hat{\beta}_2}} =$$

- ***Step 4***: state the decision rule.

Note that since we are anchoring our thought on the alternative hypothesis, we are looking for a critical region to the left-hand side of the distribution, while the right-hand side is an acceptance region.

The lower bound : $\beta_2 - t_{\frac{\alpha}{2}} \cdot \text{se}_{\hat{\beta}_2} =$

(2) *One tail test*

- ***Step* 5**: conclude the test.

- If t_{cal} lies **beyond** the critical value (in the rejection region), we **can reject the null hypothesis**, at the significance level of 95%.

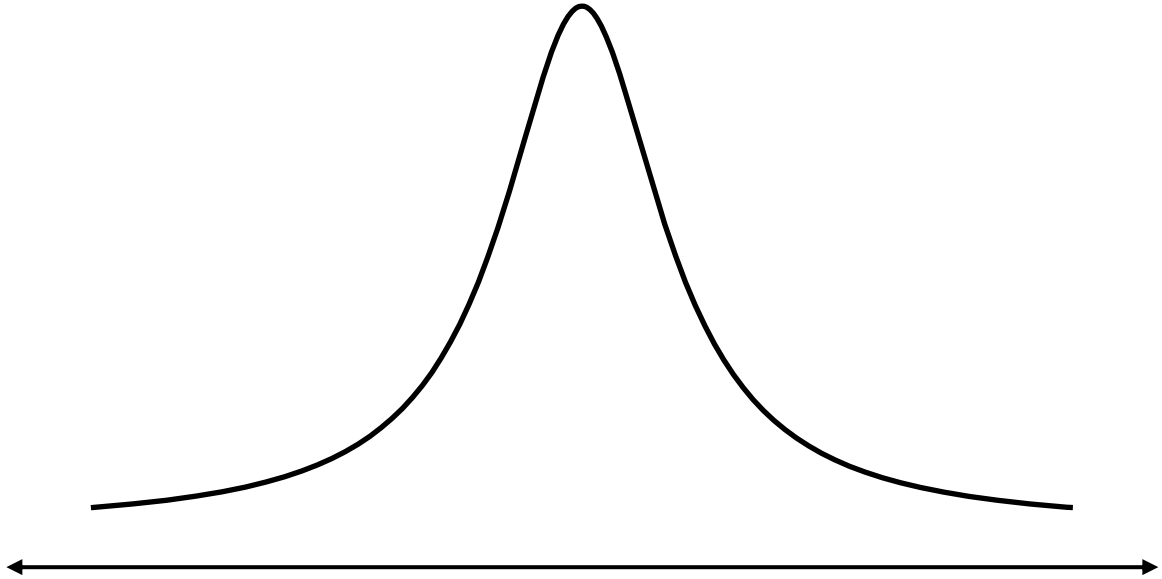
In other words, **we can make sure** that β_2 is less than 0.3 95 out of 100 times when we sample.

- If t_{cal} lies **within** the acceptance region, we **cannot reject the null hypothesis**, at the significance level of 95%.

In other words, we **cannot say for sure** that β_2 is more than 0.3 95 out of 100 times when we sample.

(2) *One tail test*

- ***Step* 5**: conclude the test.



(3) *Additional note*

If we are to make a mistake on our conclusion, there are two types of errors listed here.

- When we **reject** the null hypothesis when it is **true**, we call it a **Type I error.**
- On the other hand, if we **accept** the null hypothesis when it is **false**, we call it a **Type II error**.

	Accept H_0	**Reject H_0**
H_0 is true	Correct conclusion	**Type I error:** reject a true null hypothesis
H_0 is false	**Type II error:** accept a false null hypothesis	Correct conclusion

(3) *Additional note*

According to the STATA report , we can see that P > |t| or P-value is also reported. This is a very useful tool since we do not need to pick a specific level of significance.

- If P > |t| is less than any α, we can make sure that $(1-\alpha)$% of the time β_2 is not zero.

```
      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(1, 50)        =     21.22
       Model |  1.0675e+11         1  1.0675e+11   Prob > F        =    0.0000
    Residual |  2.5157e+11        50  5.0314e+09   R-squared       =    0.2979
-------------+----------------------------------   Adj R-squared   =    0.2839
       Total |  3.5832e+11        51  7.0259e+09   Root MSE        =     70932

------------------------------------------------------------------------------
         exp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         inc |   .2797096   .0607242     4.61   0.000     .1577414    .4016778
       _cons |   38687.86   16223.66     2.38   0.021      6101.68    71274.03
------------------------------------------------------------------------------
```