7. Multiple Regression Analysis: The Problem of Analysis

Three-Variable Model: Notation and Assumptions

Let us consider the following three-variable PRF as:

Regressand: Y_i — Regressors: X_{2i}, X_{3i}

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \qquad (k=3)$$

where

- Y_i is the dependent variable (regressand)
- X_{2i} and X_{3i} are the regressors or the explanatory variables
- u_i is the stochastic disturbance term

Remark: the subscript i is denoted the observation i from our sample data.

In case our data are time series, the subscript t will denote the t observation.

β_1 means the average value of Y when X_2 and X_3 are set equal to zero (so)

β_2 and β_3 are called the partial regression coefficients.

We will talk about the meaning of β_1 and β_2 shortly after knowing the assumptions of the classical linear regression model (CLRM)

i	Y_i	X_{2i}	X_{3i}
1			
2			
3			
4			
⋮			

$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$

t	Y_t	X_{2t}	X_{3t}
1970			
1971			
⋮			
2017			

(k parameters)

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + u_i$$

# of Xs = $k-1$ terms

Under the CLRM, we assume:

1. Zero mean value of u_i

$$E(u_i | X_{2i}, X_{3i}) = 0$$

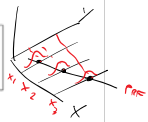
Mean value of Residuals = 0

2. No serial correlation

$$Cov(u_i, u_j) = 0 \quad \text{for all } i \neq j$$

3. Homoscedasticity

$$Var(u_i) = \sigma^2 \rightarrow \text{constant variance of } u_i$$



4. Zero covariance between u_i and each X variable, or

$$Cov(u_i, X_{2i}) = 0 \quad , \quad Cov(u_i, X_{3i}) = 0$$

5. No specification bias or

The model is correctly specified.

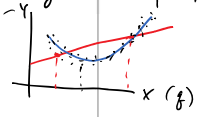
specification bias comes from
(1) add redundant variable(s)
or (2) omit important variable(s) (Exclusion Errors)
or (3) use "wrong functional form"

6. No exact collinearity between the X variables or

Example of exact linear relationship

suppose $X_{3i} = 2X_{2i}$

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$
$$= \beta_1 + \beta_2 X_{2i} + \beta_3 [2X_{2i}] + u_i$$
$$= \beta_1 + [\beta_2 + 2\beta_3] X_{2i} + u_i$$



Partial Effect of X_{2i} on Y and of X_{3i} only cannot be estimated when X_{2i} and X_{3i} have an exact linear relationship.

EXACT LINEAR RELATIONSHIP

X_{2i}	X_{3i}
2	10
4	20
6	30
8	40

$X_{3i} = 5X_{2i}$

INEXACT LINEAR RELATIONSHIP

X_{2i}	X_{3i}	$X_{3i} = 5X_{2i} + u_i$
2	10	$10 = 5.2 + 0$
4	22	$22 = 5.4 + 2$
6	28	$28 = 5.6 - 2$
8	40	$40 = 5.8 + 0$

be estimated when X_{2i} and X_{3i} have an exact linear relationship.

so called "Perfect Multicollinearity"

2	10	10	= 5·2	+ 0
4	22	22	= 5.4	+ 2
6	28	28	= 5·6	− 2
8	40	40	= 5·8	+ 0

4	20
6	30
8	40

$X_{3i} = 5X_{2i}$

$5X_{2i} - X_{3i} = 0$

In general $\lambda_2 X_{2i} + \lambda_3 X_{3i} = 0$ ($\lambda_2 = 5$, $\lambda_3 = -1$)

If λ_2, λ_3 are found to make the above equation true, X_{2i} and X_{3i} has "EXACT" linear relationship.

w/ inexact linear relationship (Imperfect multicollinearity)

β_2 and β_3 can still be estimated.

By the above assumptions, we can find out the conditional expectation of Y_i: $E[Y \mid X_{2i}, X_{3i}]$

Take $E[\cdot \mid X_{2i}, X_{3i}]$ to the PRF:

$$E[Y \mid X_{2i}, X_{3i}] = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + E[u_i \mid X_{2i}, X_{3i}]$$

($E[u_i \mid X_{2i}, X_{3i}] = 0$)

The meaning of partial coefficients:

β_2

$$\frac{d\,E[Y \mid X_{2i}, X_{3i}]}{d\,X_{2i}} = \beta_2$$

on average, when X_{2i} changes by 1 unit, $E[Y \mid X_{2i}, X_{3i}]$ will change by β_2 units, holding X_{3i} constant.

So β_2 = partial effect of X_2 on $E[Y \mid X_{2i}, X_{3r}]$, holding X_{3i} constant.

β_3

D − I − Y

$\hat{\beta}_2$, $\hat{\beta}_3$

7.1 OLS Estimation of the Partial Regression Coefficients

In order to find the OLS estimators, we need to write down the sample regression function (SRF) corresponding to the PRF:

ACTUAL Y, ESTIMATED Y, RESIDUALS

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i$$

GOAL: FIND $\hat{\beta}_2$ AND $\hat{\beta}_3$ (ALSO $\hat{\beta}_1$)

OLS TRICK: MIN $\sum \hat{u}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i})^2$ ($\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$)

F.O.C $\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_1} = 2\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i})(-1) = 0$ ($= \hat{u}_i$)

$$\sum \hat{u}_i = 0 \quad -(1)$$

$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_2} = 2\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i})(-X_{2i}) = 0$ ($= \hat{u}_i$)

$$\sum \hat{u}_i X_{2i} = 0 \quad -(2)$$

$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_3} = 2\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i})(-X_{3i}) = 0$ ($= \hat{u}_i$)

$$\sum \hat{u}_i X_{3i} = 0 \quad -(3)$$

3 Equations, 3 unknown, we can solve for $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3 \ldots$ ☺

POP → PRF → β_2, β_3 true but unknown

SAMPLE → SRF → $\hat{\beta}_2, \hat{\beta}_3$

i	Y_i	$\hat{Y}_i$	$\hat{u}_i = Y_i - \hat{Y}_i$
1			
2			
3			
4			
⋮			
500			

From the FOC, we then get the normal equations:

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3$$
$$\sum Y_i X_{2i} = \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{2i} X_{3i}$$
$$\sum Y_i X_{3i} = \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2$$

We therefore get:

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3$$
$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$
$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$

solutions

Variance and Standard Errors of OLS Estimators

$$var(\hat{\beta}_1) = \left[\frac{1}{n} + \frac{\bar{X}_2^2 \sum x_{3i}^2 + \bar{X}_3^2 \sum x_{2i}^2 - 2\bar{X}_2 \bar{X}_3 \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} x_{3i})^2}\right] * \sigma^2$$
$$se(\hat{\beta}_1) = +\sqrt{var(\hat{\beta}_1)}$$

$E(\hat{\beta}_1) = \beta_1$

$$var(\hat{\beta}_2) = \frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} * \sigma^2$$
$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}$$
$$se(\hat{\beta}_2) = +\sqrt{var(\hat{\beta}_2)}$$

$E(\hat{\beta}_2) = \beta_2$

$$var(\hat{\beta}_3) = \frac{\sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} * \sigma^2$$
$$var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)}$$
$$se(\hat{\beta}_3) = +\sqrt{var(\hat{\beta}_2)}$$

$E(\hat{\beta}_3) = \beta_3$

by Gauss-Markov Theorem, $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ ARE B-L-U-E

smallest variance among all classes of unbiased estimators

$E(\hat{\beta}_i) = \beta_i$

$\hat{\beta}_i$, OLS

$$\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23}\sigma^2}{(1-r_{23}^2)\sqrt{\sum x_{2i}^2}\sqrt{\sum x_{3i}^2}}$$

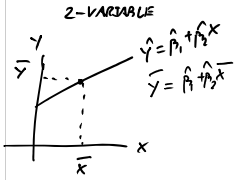
$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-3}$$ → unbiased estimator of σ^2 !!!

Here, $k = 3$ → NUMBER OF EXPLANATORY VARIABLES + INTERCEPT.

7.2 Properties of OLS Estimators : $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$

① The three-variable regression plane passes through the Means $\bar{Y}, \bar{X}_2, \bar{X}_3$

So, $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3$



② The mean value of the estimated Y_i ($\hat{Y}_i$) is equal to the mean value of actual Y_i.

Proof: $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$

we have $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3$.

So $\hat{Y}_i = (\bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3) + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$

$\hat{Y}_i = \bar{Y} + \hat{\beta}_2 X_{2i} - \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 X_{3i} - \hat{\beta}_3 \bar{X}_3$

$= \bar{Y} + \hat{\beta}_2 (X_{2i} - \bar{X}_2) + \hat{\beta}_3 (X_{3i} - \bar{X}_3)$

$\hat{Y}_i = \bar{Y} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}$ where $x_{2i} = X_{2i} - \bar{X}_2$, $x_{3i} = X_{3i} - \bar{X}_3$.

Take summation operators throughout and then divide w/ N:

$\bar{\hat{Y}}_i = \bar{Y}$. #

③ $\hat{y}_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}$

actual = estimated + residuals

$y_i = \hat{y}_i + \hat{u}_i$

$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i$ → SRF in Deviation Form.

$\hat{u}_i = y_i - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}$

$\sum \hat{u}_i^2 = \sum \hat{u}_i (y_i - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i})$

Properties of OLS Estimators (Cont:)

$\sum \hat{u}_i^2 = \sum \hat{u}_i y_i - \hat{\beta}_2 \sum \hat{u}_i x_{2i} - \hat{\beta}_3 \sum \hat{u}_i x_{3i}$ (the last two terms $= 0$)

$\sum \hat{u}_i^2 = \sum (y_i - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}) \cdot y_i$

RSS Residual Sum of Squares

$= \sum y_i^2 - \hat{\beta}_2 \sum x_{2i} y - \hat{\beta}_3 \sum x_{3i} y_i$ → we will make use of this soon...

④ $\sum \hat{u}_i = 0$ and $\frac{\sum \hat{u}_i}{n} = \bar{\hat{u}}_i = 0$

⑤ $\hat{u}_i$ are uncorrelated with X_{2i} : $\sum \hat{u}_i X_{2i} = 0$.

$\hat{u}_i$ ——— X_{3i} : $\sum \hat{u}_i X_{3i} = 0$.

⑥ $\hat{u}_i$ are uncorrelated with $\hat{Y}_i$: $\sum \hat{u}_i \hat{Y}_i = 0$.

Proof: $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$

$\sum \hat{u}_i \hat{Y}_i = \sum \hat{\beta}_1 \hat{u}_i + \sum \hat{\beta}_2 X_{2i} \hat{u}_i + \sum \hat{\beta}_3 X_{3i} \hat{u}_i$

$= 0$ $= 0$ $= 0$

So $\sum \hat{u}_i \hat{Y}_i = 0$.

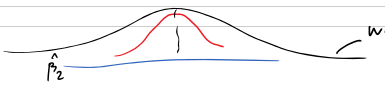
*** ⑦ Recall that $\text{var}(\hat{\beta}_2) = \frac{1}{\sum x_{2i}^2} \cdot \frac{\sigma^2}{(1 - r_{23}^2)}$.

$\text{var}(\hat{\beta}_3) = \frac{1}{\sum x_{3i}^2} \cdot \frac{\sigma^2}{(1 - r_{23}^2)}$.

Given $\sum x_{2i}^2$ and σ^2, if r_{23} ↑ → $\text{var}(\hat{\beta}_2)$ ↑ → Reduce Accuracy of Estimation!

If $r_{23} = 1$, then $\text{var}(\hat{\beta}_2)$ becomes INFINITE!

In this case, $\hat{\beta}_2$ cannot be estimated [PERFECT MULTICOLLINEARITY]



with high r_{23}, $\text{var}(\hat{\beta}_2)$ BIG.

Properties of OLS Estimators (Cont:)

with high r_{23}, $var(\hat{\beta}_2)$ BIG.

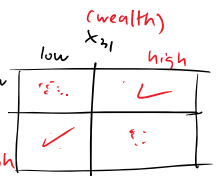
Properties of OLS Estimators (Cont:)

- Given r_{23} and σ^2, if $\sum x_{2i}^2 \uparrow$, $var(\hat{\beta}_2)$ will become lower

  $\sum x_{2i}^2 = \sum_{i=1}^{n} (X_{2i} - \bar{X})^2$

  ↓

  improve accuracy of estimation.



Advice: get more observations as much as possible! This will help to increase power of estimation!

- Given r_{23} and $\sum x_{2i}^2$, if $\sigma^2 \uparrow$, $var(\hat{\beta}_2) \uparrow$ ∴

(8) $\hat{\beta}_2$ and $\hat{\beta}_3$ are Best Linear Unbiased Estimator.

The Multiple Coefficient of Determination R^2 and the Multiple Coefficient of Correlation R

In this section, we will study how to measure the proportion of the variation in Y explained by the variables X_2 and X_3 jointly. This is the same concept of r^2 that we have learned before.

The quantity that gives this information is known as the **the multiple coefficient of determination** and is denoted by R^2.

To derive R^2, we firstly write down the following equation:

$$\begin{aligned} Y_i &= \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i \\ &= \hat{Y}_i + \hat{u}_i \end{aligned} \quad (7.1)$$

→ STANDARD FORM OF SRF

where $\hat{Y}_i$ is the estimated value of Y_i from the fitted regression line and is an estimator of true $E(Y_i|X_{2i}, X_{3i})$.

7.1 may be written as

$$\begin{aligned} y_i &= \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i \\ &= \hat{y}_i + \hat{u}_i \end{aligned} \quad (7.2)$$

→ DEVIATION FORM OF SRF

Squaring 7.2 on both sides and summing over the sample values, we obtain

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{u}_i^2 + 2\sum \hat{y}_i \hat{u}_i = \sum \hat{y}_i^2 + \sum \hat{u}_i^2 \quad (7.3)$$

TSS = ESS + RSS

RSS: variation in y_i that x_2 and x_3 cannot help to explain.

ESS: variation in y_i that could be explained by the helps of adding x_2 and x_3.

TSS: Baseline errors when we simply use sample mean to estimate y_i

Recall that $TSS = ESS + RSS$

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

R^2 → Proportion Variation in Y that can be explained by the regression (by using x_2 and x_3)

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

→ RESIDUAL SUM OF SQUARE (RSS)

→ TOTAL SUM OF SQUARES (TSS)

FROM $\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2 \sum x_{2i} y_i - \hat{\beta}_3 \sum x_{3i} y_i$,

$\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{u}_i^2$

$\sum y_i^2 = \sum \hat{y}_i^2 + \sum y_i^2 - \hat{\beta}_2 \sum x_{2i} y_i - \hat{\beta}_3 \sum x_{3i} y_i$

$\therefore \sum \hat{y}_i^2 = \hat{\beta}_2 \sum x_{2i} y_i + \hat{\beta}_3 \sum x_{3i} y_i$

↓

ESS

$$R^2 = \frac{ESS}{TSS} = \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2} \quad \rightarrow R^2 \text{ in another form.} \tag{7.4}$$

$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \ldots + \hat{\beta}_k X_{ki}$

Regressors = $k-1$ variables

R_j^2 is the R^2 in the regression of X_j on the remaining regressors ($k-2$).

$X_2 = f(X_3)$

The three-or-more-variable analogue of r is the coefficient of multiple correlation, denoted by R, and it is a measure of the degree of association between Y and all the explanatory variables jointly. Although r can be positive or negative, R is always taken to be positive.

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2} \left(\frac{1}{1 - R_j^2} \right)$$

7.2.1 R^2 and the Adjusted R^2

It should be noted that the R^2 is a nondecreasing function of the number of explanatory variables. Thus, when the number of regressors increases, R^2 almost invariably increases and never decreases. **In other words, an additional X variable will not decrease R^2!**

To explain this fact, let we write down the definition of R^2 again:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2} \tag{7.5}$$

Therefore, in comparing two regression models **with the same dependent variable but differing number of X variables**, one should be very wary of choosing the model with the highest R^2.

In light of comparing two R^2 terms, we have to take into account the number of X variables present in the model. To achieve this goal , we can consider the alternative coefficient of determination, which is as follows:

To compare between the two models, it is advised to compare adjusted R^2 instead of R^2.

$$\bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2/(n-3)}{\sum y_i^2/(n-1)} \quad \text{For } k = 3$$

General form:

$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i$

$$\bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2/(n-k)}{\sum y_i^2/(n-1)}$$

k= the number of parameters in the model including the intercept term.

n= the number of observations in the sample data.

The above equation is known as **the adjusted R^2**, denoted by $\bar{R}^2$. The term adjusted means adjusted for the df associated with the sums of squares entering into 7.5.

We can rewrite the the adjusted R^2 as:

$$\bar{R}^2 = 1 - \frac{\hat{\sigma}^2}{S_Y^2}$$

where $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-k}$ → unbiased estimator of σ^2 (residual variance)

$S_Y^2 = \frac{\sum y_i^2}{n-1}$ → sample variance of Y

We can also get the equation which shows the relationship between $\bar{R}^2$ and R^2:

$$\bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2} \cdot \frac{(n-1)}{(n-k)}$$

$$\bar{R}^2 = 1 - [1 - R^2] \cdot \frac{(n-1)}{(n-k)}$$

As $k > 1 \rightarrow \frac{(n-1)}{(n-k)} > 1$

$\bar{R}^2 < R^2$

Note: $\bar{R}^2$ CAN BE NEGATIVE.
Ex: when $R^2 = 0$. (verify this)

Besides R^2 and $\bar{R}^2$ as goodness of fit measures, other criteria are often used to judge the adequacy of a regression model. Two of these are **Akaike's Information criterion and Amemiya's Prediction criteria**, which are used to select between competing models. We will discuss these criteria in greater detail later.

* When comparing two R^2 values, make sure that
  ① Dependent variables of the two models are the same,
  ② sample size must be the same.

① $\ln Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}$
② $Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i}$

① AND ② CANNOT BE COMPARED DIRECTLY BY USING R^2 AS THE INDEPENDENT VARIABLES ARE NOT THE SAME.