IV. Multiple Regression.

V. Extensions of Multiple Regression

A. Non-Linear Models (Chapter 9)

B. Dummy (Binary) Variables (Chapter 10)

C. Scaling Variables

VI. Problems and Specification Issues

A. Model Selection/Specification

B. Multicollinearity

C. Heteroskedasticity

D. Autocorrelation

VI. Problems and Specification Issues

A. Model Selection/Specification

1. Two Possible Mistakes: omit important independent variables; include variables that don't belong
2. Omitted Variables: estimators are ***biased***
3. Irrelevant Variables: estimators are ***unbiased***, but inefficient – OLS estimators are no longer BLUE
4. Tools and Tests

Compare – effects on hypothesis testing of our two possible mistakes

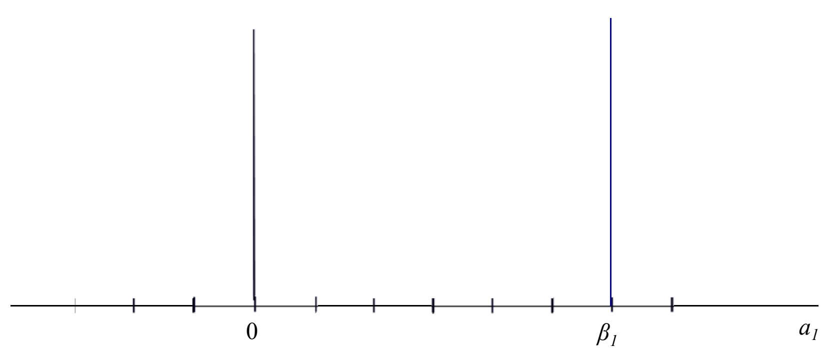
- The correct model is:

  (1) $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$
- But you make a mistake and estimate:

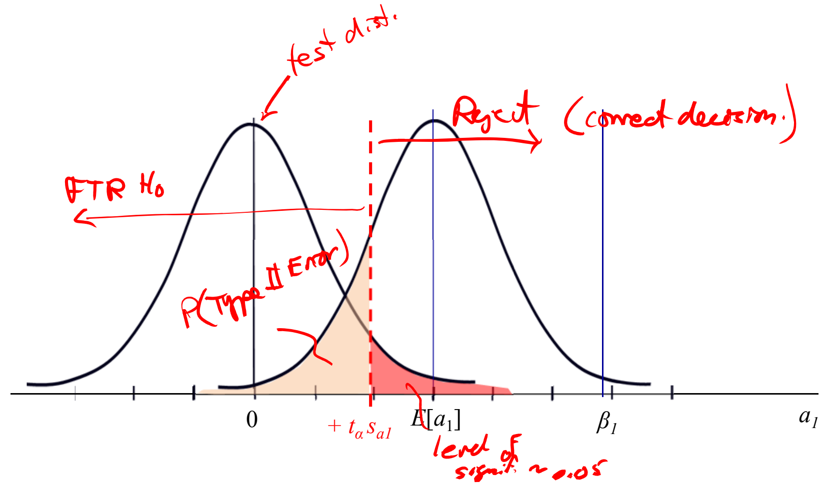
  (2) $Y_i = \alpha_0 + \alpha_1 X_{1i} + v_i$

What are the possible consequences if you test the null hypothesis: $H_0: \beta_1 \le 0$

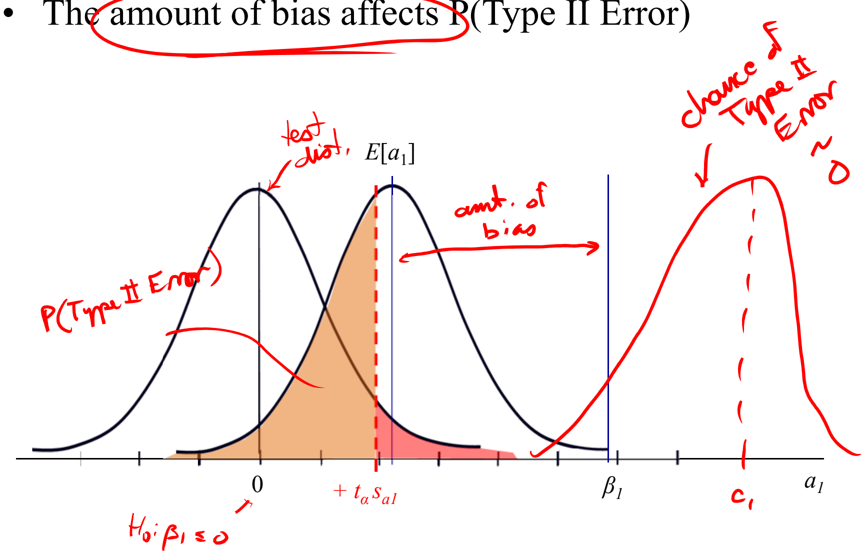
- *Illustrate* the test: H_0: $\beta_1 \le 0$; H_A: $\beta_1 > 0$
- The *null is wrong,* $\beta_1 > 0$
- And, we know $E[a_1] < \beta_1$



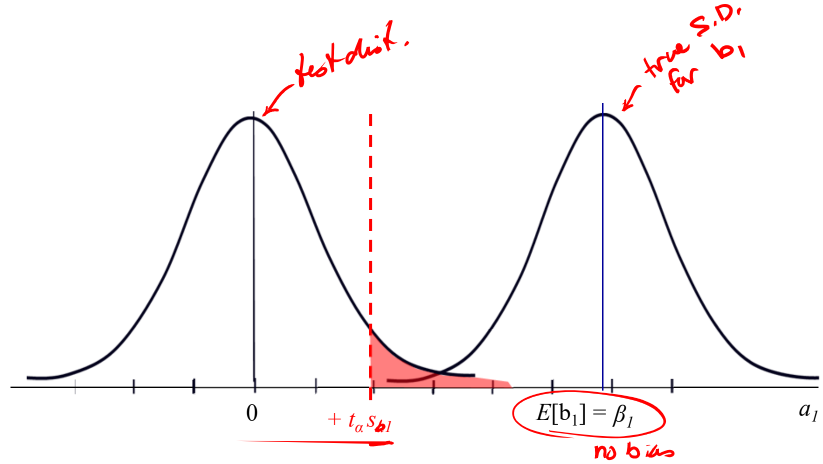
- *Illustrate* the test: H_0: $\beta_1 \le 0$; H_A: $\beta_1 > 0$.
- The *null is wrong,* $\beta_1 > 0$, and
- $E[a_1] < \beta_1$



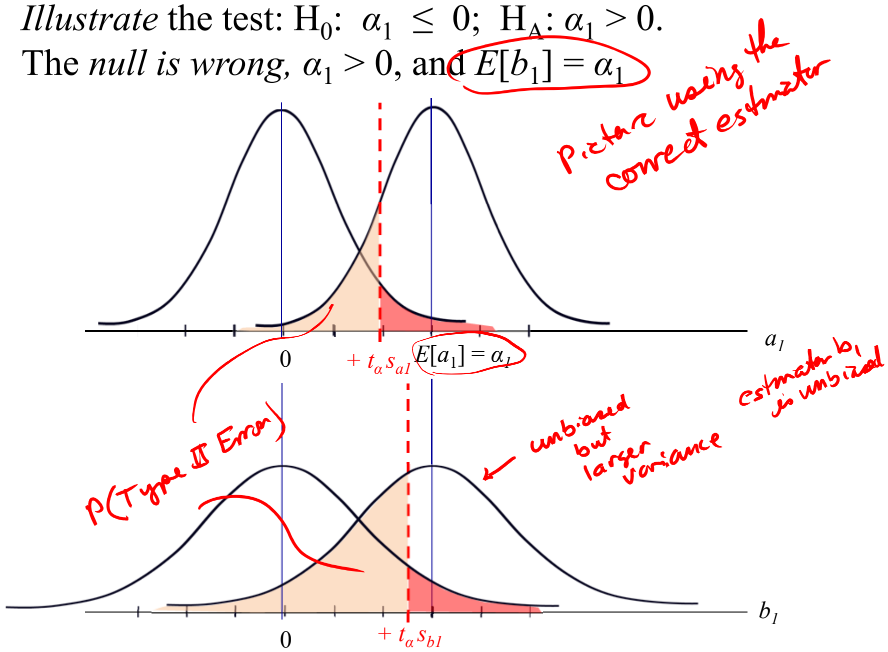
- *Illustrate* the test: H_0: $\beta_1 \le 0$; H_A: $\beta_1 > 0$.
- The *null is wrong,* $\beta_1 > 0$, and $E[a_1] < \beta_1$
- The amount of bias affects P(Type II Error)



- *Illustrate* the test: H_0: $\beta_1 \leq 0$; H_A: $\beta_1 > 0$.
- But the *null is wrong,* $\beta_1 > 0$
- No bias – you use the correct estimator: $E[b_1] = \beta_1$



- Model (2) is correct but you estimate Model (1)
- *Illustrate* the test: H_0: $\alpha_1 \leq 0$; H_A: $\alpha_1 > 0$.
- The *null is wrong,* $\alpha_1 > 0$, and $E[b_1] = \alpha_1$



4. Tools

a. Start with theory and literature review:
- ✓ Do you have all the right variables? What did other researchers find?

b. Check the Adjusted R^2: adding and deleting variable
- ✓ Do additional variables explain variation in Y?

* c. Joint F-tests for additional variables

d. Specification tests:
- ✓ Regression Error Specification Test (RESET)
- ✓ SAS – use the "SPEC" (specification test) option

RESET Test

* No new information - still doing same Xs

i. Estimate: $Y_i = \beta_0 + \beta_0 X_{1i} + \ldots + \beta_K X_{Ki} + u_i$

ii. Save predicted values: $\hat{Y}_i$

* Could pick up nonlinear relationships

iii. Create new variables: $\hat{Y}_i^2$ and $\hat{Y}_i^3$

iv. Estimate again: $Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_K X_{Ki} + \delta_1 \hat{Y}_i^2 + \delta_2 \hat{Y}_i^3 + u_i$

v. Hypothesis test: $H_0: \delta_1 = \delta_2 = 0$
H_a: at least one δ is not zero } F-test

White's Specification Test (SPEC option)

SAS

i. Estimate: $Y_i = \beta_0 + \beta_0 X_{1i} + \ldots + \beta_K X_{Ki} + u_i$

ii. Save the errors and square: $e_i \Rightarrow e_i^2$

iii. Create new variables: $X_1^2, X_2^2 \cdots X_k^2, X_1 \cdot X_2, X_2 \cdot X_3$
$X_1 \cdot X_3, \ldots$

iv. Estimate the auxiliary regression:

$$R^2 \Leftarrow e_i^2 = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_k X_k + c_1 X_1^2 + c_2 X_2^2 + \cdots + c_k X_k^2 + d_{12} X_1 \cdot X_2 + \cdots$$

v. Hypothesis test: $nR^2 \sim \chi^2_{(s)}$ $s = $ # of variables

B. Multicollinearity

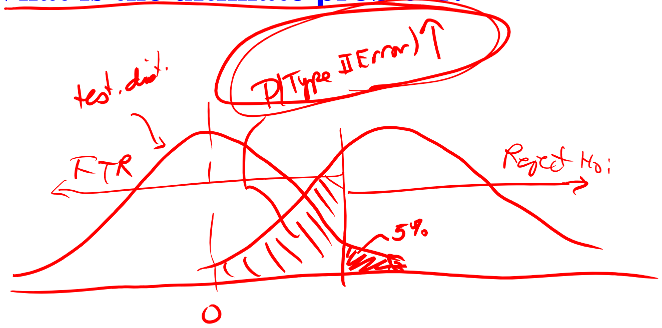
1. **Definition**: The presence of ***linear association*** among independent variables – i.e. linear association or correlation between X_1 and X_2.

$X_{1i} = \lambda X_{2i} + \omega_i$ Not a causal model
no economic theory

- ***Sample Problem*** – the problem lies in your sample data.
- There is ***no causal relationship*** between X_2 and the other independent variables.
- Ie., X_2 does not "cause" X_1 No theory for this relationship

2. Consequences:

- OLS estimators – remain unbiased.
- **Standard errors are inflated.**
- **Calculated t-statistics are deflated.** $\downarrow t_{calc} = \dfrac{b_1 - \beta_1^0}{s_{b_1}\uparrow}$
- **What is the ultimate problem?**



2. Consequences of Multicollinearity

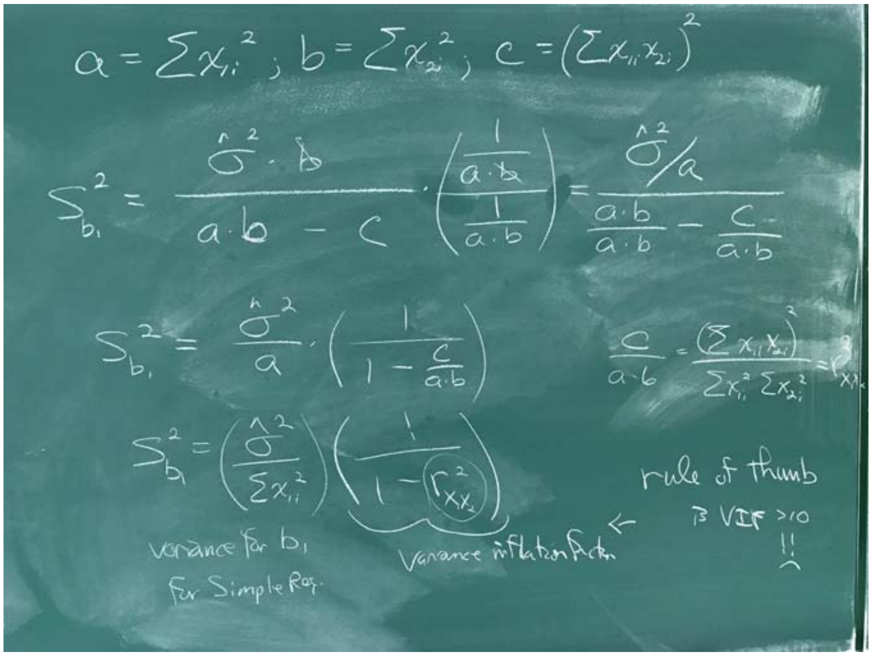
- Multiple reg. standard error – 2 indep. variables:

$$s_{b_1}^2 = \frac{\hat{\sigma}^2 \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i}x_{2i})^2}$$

- If X_1 and X_2 are strongly and linearly associated:

$$r_{x_1x_2}^2 = \frac{\left(\sum x_{1i}x_{2i}\right)^2}{\sum x_{1i}^2 \sum x_{2i}^2} \rightarrow 1$$

$$\sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{1i}x_{2i}\right)^2 \rightarrow 0$$



- Multiple regression variances are inflated – the *variance inflation factor*

1. Multiple regression variance:

$$s_{b_1}^2 = \frac{\hat{\sigma}^2 \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2}$$

2. Multiply by "1":

$$s_{b_1}^2 = \frac{\hat{\sigma}^2 \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2} \cdot \frac{1/\left(\sum x_{1i}^2 \sum x_{2i}^2\right)}{1/\left(\sum x_{1i}^2 \sum x_{2i}^2\right)}$$

3. Rearrange terms and note that some stuff cancels:

$$s_{b_1}^2 = \frac{\hat{\sigma}^2 \cancel{\sum x_{2i}^2} \Big/ \left(\sum x_{1i}^2 \cancel{\sum x_{2i}^2}\right)}{\frac{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} x_{2i})^2}{\sum x_{1i}^2 \sum x_{2i}^2}} = \frac{\hat{\sigma}^2 \Big/ \sum x_{1i}^2}{\frac{\sum x_{1i}^2 \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2} - \frac{(\sum x_{1i} x_{2i})^2}{\sum x_{1i}^2 \sum x_{2i}^2}}$$

4. Simplify the denominator:

$$s_{b_1}^2 = \frac{\hat{\sigma}^2 / \sum x_{1i}^2}{\dfrac{\sum x_{1i}^2 \sum x_{2i}^2}{\sum x_{1i}^2 \sum x_{2i}^2} - \dfrac{(\sum x_{1i}x_{2i})^2}{\sum x_{1i}^2 \sum x_{2i}^2}} = \frac{\boxed{\hat{\sigma}^2 / \sum x_{1i}^2}}{1 - \boxed{\dfrac{(\sum x_{1i}x_{2i})^2}{\sum x_{1i}^2 \sum x_{2i}^2}}}$$

5. Two familiar terms:

Simple regression variance:

$$\boxed{s_{b_1}^2 = \frac{\hat{\sigma}^2}{\sum x_{1i}^2}}$$

Squared correlation coefficient:

$$\boxed{r_{x1x2}^2 = \frac{(\sum x_{1i}x_{2i})^2}{\sum x_{1i}^2 \sum x_{2i}^2}}$$

6. Viola! The variance of the multiple regression estimator is the variance of the simple regression estimator multiplied by the $\boxed{VIF}$

$$s_{b_1}^2 = \frac{\hat{\sigma}^2}{\sum x_{1i}^2} \cdot \left(\frac{1}{1 - r_{x_1x_2}^2} \right)$$

Correlation of X_1 and X_2 causes variance inflation

3. Diagnosis (Multicollinearity)

for model $H_0: \beta_1 = \dots = \beta_k = 0$

a. Classic signs: R^2 and F_{calc} tell you your model is good !! :)

BUT when you simultaneously have low t_{calcs} it suggest a problem

b. Correlation Coefficients

$-1 < r_{x_1x_2} < +1$ or $0 < r^2_{x_1x_2} < 1$ $\Rightarrow$ high correlations of 0.8 to and above

c. Auxilliary Regressions

$X_2 = a + bX_1 + cX_3$ ← estimate aux. reg. check R^2 - is it high?

d. Variance Inflation Factors

related to c $\Rightarrow$ $VIF = \dfrac{1}{1-R^2}$ aux. reg. R^2

or - ask SAS or minitab