MGMG 522 : Session #2
Learning to Use Regression Analysis
& The Classical Model

(Ch. 3 & 4)

2-1

Steps in Regression Analysis

1. Review the literature and develop the theoretical model
2. Specify the model: choose X's and the functional form
3. Hypothesize the expected signs of the coefficients
4. Gather the data
5. Run the regression and evaluate the regression equation
6. Document the results

2-2

Step 1: Review the literature & develop the theoretical model

- Do not try desperately to find the regression model that best fits your data.
- Rather, you should be able to explain your regression model with some theory.
- Therefore, you need to know the relevant theory by reviewing the literature.
- If there is no theory directly related to your study, try similar or closely related theory.
- If there still is no such closely related theory, you might consult experts in that field or develop your own theory with your story.

Steps 2-3: Specify the model and hypothesize the signs of the coefficients

- The topic you are studying will give you some clues about your dependent variable, Y, to use.
- You must decide what X's to use.
- You must decide the functional form: how X's relate to Y.
- The relationships between X's and Y must be based on some theory, not the data.
- The theory that backs up your model should give you some hints about the expected signs of the coefficients in front of the independent variables.

Steps 4-5: Collect the data, estimate the coefficients, and evaluate

- The model should be specified first, before you go out and start collecting data.
- Verify the reliability of your data. Look out for errors.
- Generally, the more data, the better. [More degrees of freedom]
- Unit of measurement does not matter to the regression analysis, except for the interpretation of the coefficients.
- Use the data you collected, run the OLS regression to obtain the estimates of the coefficients. You might try other estimators and compare their estimates of the coefficients with those from the OLS.
- Then, you must evaluate the estimates of the coefficients. See if the signs are what you expect and significant. See the overall goodness of fit.

2-5

Step 6: Document the results

- When documenting the results, the standard format typically is

$$\hat{Y}_i = 100 + 15X_i$$
$$(0.023)$$

- n = 100, Adj-R^2 = 0.65
- At the minimum, you should include this information.
- Some researchers include more information than that given above.
- You should describe your data and procedures, and explain your results in words.

2-6

Things to Remember

- Do not try to fit a model to your data without some reasonable theory behind your model.
- You do not prove the relationships or theory. But you try to find some evidence that there might be some relationships.
- Model specification is very important.
- Only relevant independent variables should be included in the model.

2-7

The Classical Assumptions

1. The regression model is linear in coefficients, is correctly specified, and has an additive error term.
2. Expected value of the error term is zero. [$E(\varepsilon_i)=0$]
3. All X_i's are uncorrelated with ε_i.
4. Error terms are uncorrelated with each other. [No serial correlation]
5. The error term has a constant variance. [No heteroskedasticity]
6. No independent variable can be written as a linear function of other independent variables. [No perfect multicollinearity]
7. The error term is normally distributed. [Optional, but it is nice to have this assumption.]

2-8

Assumption 1: Linear in coefficients, correctly specified, and an additive error term

- The model must include all relevant independent variables, must have a correct functional form, and has an additive error term.
- The model must be linear in coefficients, but does not have to be linear in variables.

2-9

Assumptions 2-3: $E(\varepsilon_i) = 0$, All X_i's are uncorrelated with ε_i

- The mean of the error term is zero.
- The distribution of ε_i does not have to be a normal distribution, as long as the mean of the distribution is zero.
- OLS regression (with an intercept term) will guarantee that the mean of the error term is zero. [By adjusting the intercept if the actual mean is not zero. Therefore, we will pay little attention to the intercept term.]
- Assumption 3 states that $Cov(X, \varepsilon) = 0$.

2-10

Assumptions 4-5: The error term is uncorrelated with each other, and has a constant variance.

- Error terms are uncorrelated with each other. There should not be any information contained in one error term to predict the other error terms.
- Very important, especially, in the time-series type of analysis.
- If serial correlation exists, it possibly means that you forgot to include some relevant variables in the model.
- To get precise estimates of the coefficients, we also need the error term to have a constant variance.

Assumption 6: No X can be written as a linear function of other X's.

- No perfect multicollinearity such as
  - $X_1 = 2X_2$
  - $X_1 = X_2 + X_3$
  - $X_1 = 5$ (a constant)
- If there is <u>perfect</u> multicollinearity, you will not be able to obtain the estimates of the coefficients.
- Solution: drop one variable from the model.
- The existence of perfect multicollinearity is not a problem, but the existence of <u>imperfect</u> multicollinearity is a problem.

Assumption 7: The error term is normally distributed

- Not required for estimation.
- But important for hypothesis testing.
- The error term can be thought of as the sum of a number of minor influences or errors.
- As the number of these minor influences or errors gets large, the distribution of the error term approaches normal distribution [Central Limit Theorem].
- Intentionally omitting some variables to achieve normality in the error term should never be considered.

Sampling Distribution of Estimators

- Just as the error term has a distribution, the estimates of the coefficients, $\hat{\beta}_k$, have their own distributions too.
- If the error term is normally distributed, the estimates of the coefficients will also be normally distributed because any linear function of a normally distributed variable is itself normally distributed.

Desirable Properties of the Estimators

- We prefer to have an unbiased estimator, the estimator that produces, $\hat{\beta}_k$, where $E(\hat{\beta}_k) = \beta_k$.
- Among all unbiased estimators to choose from, we also prefer an efficient unbiased estimator, the one that produces the smallest variance of the estimates of the coefficients, $VAR(\hat{\beta}_k)$.

2-15

Gauss-Markov Theorem & OLS

If Assumptions 1-6 are met,

- OLS estimator is the minimum variance estimator of all linear unbiased estimators of β_k, k = 0, 1, 2,..., K.
- "OLS is BLUE." [Best Linear Unbiased Estimator]
  - "Best" means that each estimate of β_k has the smallest variance possible.

2-16

If Assumptions 1-7 are met,

- OLS estimator is the minimum variance estimator of all unbiased estimators of β_k, k = 0, 1, 2,..., K.
- "OLS is BUE." [Best Unbiased Estimator]
- OLS estimator is equivalent to the Maximum Likelihood (ML) estimator. That is, coefficients estimated from OLS are identical to those estimated from ML.
- However, ML gives biased estimate of the variance of the error term, while OLS does not.

When All Seven Assumptions Are Met

OLS estimator

1. is unbiased, $E(\hat{\beta}_k) = \beta_k$
2. has the minimum variance among all estimators, $VAR(\hat{\beta}_k)$
3. is consistent, and
4. is normally distributed, $\hat{\beta}_k \sim N(\beta, \sigma^2_{\hat{\beta}})$