MGMG 522 : Session #4

Choosing the Independent Variables and a Functional Form

(Ch. 6 & 7)

Major Specification Problems

1. Problem with the selection of the independent variables.
2. Problem with the functional form.
3. Problem with the form of the error term.

Problems with the Selection of the Independent Variables

- The choice of independent variables is up to the researcher to decide.
- This freedom does not come without a cost.
- Problems
  1. Omitted variables
  2. Irrelevant variables
- Your underlying theory should give you some hints about what independent variables should be included in your regression model.
- The statistical fit is less important than the underlying theory.

4-3

Case 1: Omitted Variables

- It occurs when you don't include important independent variables in your regression model when you should, either because you don't think of them or you think of them but you can't get the data.
- True model: $Y = \beta_0+\beta_1X_1+\beta_2X_2+\varepsilon$
- Your model: $Y = \beta_0+\beta_1X_1+\varepsilon^*$
  where, $\varepsilon^* = \beta_2X_2+\varepsilon$
- If X_1 and X_2 are not completely independent, ε^* will not be independent of X_1, a violation of the classical assumption #3 (all X's are uncorrelated with ε).

4-4

- Problems:

1. OLS is no longer BLUE.
2. Coefficient estimates are <u>biased</u>, $E(\hat{\beta}_k) \neq \beta_k$.
3. $VAR(\hat{\beta}_k)$, variances of the coefficient estimates decrease. See p. 4-8.

- For a 2-independent variable model, it can be shown that,

  $E(\beta_1) = \beta_1 + \beta_2\alpha_1$

  where α_1 is from: $X_2 = \alpha_0 + \alpha_1 X_1 + u$

  and u is a classical error term
- Coefficient estimates could be unbiased if $\beta_2 = 0$ or $\alpha_1 = 0$. But, that is unlikely.

4-5

$E(\beta_1) = \beta_1 + \beta_2\alpha_1$

- The amount of bias = $\beta_2\alpha_1$.
- Or, the amount of bias = $\beta_2 f(r_{x1,x2})$.
- The direction of the bias can be determined by the signs of β_2 and α_1. For example, (-)(-)=(+) or (-)(+)=(-).
- To correct for the omitted variables problem,

1. Think again about your theory. What other important variables could be missing?
2. If the signs of the included coefficient estimates are unexpected, you could probably tell the direction of the bias and have some clues about the missing variables.

4-6

Case 2: Irrelevant Variables

- It occurs when you have some unnecessary independent variables in your regression model.
- True model: $Y = \beta_0+\beta_1X_1+\varepsilon$
- Your model: $Y = \beta_0+\beta_1X_1+\beta_2X_2+\varepsilon^{**}$
  where, $\varepsilon^{**} = \varepsilon-\beta_2X_2$
- The coefficient estimates will still be unbiased, but $VAR(\hat{\beta}_k)$ increase, lowing the reported t-values.

- For the model, $Y = \beta_0+\beta_1X_1+\beta_2X_2$.

$$VAR(\hat{\beta}_1)=\frac{\sigma^2}{(1-r_{12}^2)\cdot\sum(X_1-\overline{X}_1)^2}$$

- If $r_{12} \neq 0$, the variance will increase.
- If $r_{12} = 0$ or the irrelevant variable is not in the regression model, the variance will stay the same.
- Now, it seems like having extra unnecessary variables is not as serious a problem compared to the omitted variables problem.
- In fact, we want neither one of these problems.

Four Criteria to Help You Choose the Independent Variables

1. Theory
2. t-Test
3. Adj-R^2
4. Bias

- If all four conditions are met, that variable should be in your model.
- If not, that variable doesn't belong in your model.
- If some conditions are met while some are not, use your judgment.

4-9

- When you add an omitted variable, usually
  - Adj-R^2 will rise
  - Coefficient estimates will change
- When you add an irrelevant variable, usually
  - Adj-R^2 will fall
  - Coefficient estimates will not change
  - t-values become less significant
- Don't rely on the Adj-R^2 criterion alone, because it can be shown that Adj-R^2 will rise if you include a variable with t-value > 1 but not significant.
- Adj-R^2 will also rise if you delete a variable with t-value < 1 from your regression model. See an example on pp. 173-176.

4-10

Three Methods to Avoid When Choosing Independent Variables

1. Data Mining
2. Stepwise Regression
3. Sequential Search

- You should specify as few models as possible.
- The more you look, the higher the chance you will find a model that has a good statistical fit with not much theoretical support.
- Do not select a variable based on its t-value, because that technique creates a systematic bias. How?

4-11

Lagged Variable

- Sometimes, the change in Y is not caused by the change in X from the current period, but from the other period.
- The coefficient estimate of a lagged variable measures the change in Y this period as a result of a change in X in the other period.

4-12

Other Specification Criteria

- Besides the four criteria outlined on p. 4-9, there are other specification criteria.

1. Ramsey's Regression Specification Error Test (RESET)
2. Akaike's Information Criterion (AIC)
3. Schwarz Criterion (SC)

Ramsey's Regression Specification Error Test (RESET)

- H_0: There is no specification error.
- H_1: There is a specification error.
- If F-value from the RESET is higher than the critical F-value, we can reject H_0, meaning that there is a specification error. However, RESET doesn't tell how to correct it.
- If F-value from the RESET is lower than the critical F-value, we cannot reject H_0, meaning that we probably have a correct specification.
- RESET is more useful in confirming our model than telling us what's wrong and how to correct our model.

AIC and SC

- AIC and SC are used to compare two regression models.
- Both AIC and SC penalize the addition of another independent variable if it doesn't improve the overall fit significantly.
- Between two regression models, the one with lower AIC and SC values is preferred.

4-15

Choosing a Functional Form

- Now, you have a set of independent variables, you still need to specify a functional form.
- That is, how Y is related to each X.
- About the intercept term:

1. Do not suppress the intercept term even if the theory suggests.
2. Do not rely on the estimate of the intercept term for analysis or inference.

4-16

Different Functional Forms

1. Linear Form: $Y = \beta_0+\beta_1X_1+\varepsilon$
2. Double-Log Form: $\ln Y = \beta_0+\beta_1\ln X_1+\varepsilon$
3. Semilog Form: $\ln Y = \beta_0+\beta_1X_1+\varepsilon$, or
   $Y = \beta_0+\beta_1\ln X_1+\varepsilon$
4. Polynomial Form:
   $Y = \beta_0+\beta_1X_1+\beta_2(X_1)^2+\beta_3(X_1)^3+\varepsilon$
5. Inverse Form: $Y = \beta_0+\beta_1(1/X_1)+\varepsilon$

- Theory usually suggests only the signs of the coefficients, not the functional form.

4-17

Linear Form

- $Y = \beta_0+\beta_1X_1+\beta_2X_2+\ldots+\varepsilon$
- The slope is constant. $\frac{\Delta Y}{\Delta X_1} = \beta_1$
- But, the elasticity of Y with respect to X is <u>not</u> constant. $\frac{\Delta Y/Y}{\Delta X_1/X_1} = \beta_1\left(\frac{X_1}{Y}\right)$
- Unless the theory suggests otherwise, the linear form should be used.

4-18

Double-Log Form

- $\ln Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \ldots + \varepsilon$
- This is another popular form besides the linear form.
- It is also known as the "Log-linear" form.
- The slope is <u>not</u> constant.
- But, the elasticity of Y with respect to X is constant. $\frac{\Delta Y / Y}{\Delta X_1 / X_1} = \frac{\Delta(\ln Y)}{\Delta(\ln X_1)} = \beta_1$
- See p. 212 for more information.

4-19

Semilog Form

- $\ln Y = \beta_0 + \beta_1 X_1 + \beta_2 \ln X_2 + \varepsilon$, or
  $Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 X_2 + \varepsilon$
- Similar to the Double-Log form, except that some variables, but not all, are in log form.
- See p. 214 for more information.

4-20

Polynomial Form

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_1)^2 + \beta_3 (X_1)^3 + \varepsilon$
- Appropriate for a model where changes in X cause Y to increase/decrease over some range and decrease/increase over other range.
- See p. 217 for more information.

4-21

Inverse Form

- $Y = \beta_0 + \beta_1 (1/X_1) + \varepsilon$
- Appropriate for a model where the impact of an independent variable approaches zero as its value gets large.
- See p. 219 for more information.

4-22

Selecting a Functional Form

- Rely on your theory. What does your theory tell you about the relationships?
- Do not compare Adj-R^2 from a linear in variable model with a nonlinear in variable model. Because
  - Adj-R^2 are not comparable when Y is transformed. Use $\text{Quasi} - R^2 = 1 - \frac{\sum\left[Y_i - \text{antilog}(\ln\hat{Y}_i)\right]^2}{\sum\left[Y_i - \bar{Y}\right]^2}$ for comparison instead.
  - Adj-R^2 may look good inside the range of the sample, but could look bad outside the range of the sample.

4-23

Dummy Variables

1. <u>Dummy Intercept</u> takes the form:
   $Y = \beta_0 + \beta_1 X_1 + \beta_2 D + \varepsilon$
2. <u>Dummy Slope</u> takes the form:
   $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1 D + \varepsilon$
3. Both <u>Dummy Intercept</u> and <u>Dummy Slope</u> take the form:
   $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1 D + \beta_3 D + \varepsilon$

** We will discuss the concept and use of dummy variables again in “Panel Data Regression” if we have time.

4-24

Appendix: General F-Test

- Any F-test we've seen so far can be thought of as a special case of the general F-test.
- The general F-test tests more than one coefficient at a time.
- The null hypothesis for the general F-test is what we think is correct.
- We usually want to "accept" H_0.
- This contrasts to the traditional way of hypothesis testing we've learned.

Steps in General F-Test

1. Specify the null and alternative hypotheses.
2. The null hypothesis will be used as a constraint to be put on the equation.
3. Calculate RSSs from the constraint and the unconstraint equations.
4. If the fits of the two equations are not significantly different, we will "accept" H_0. If the fits are significantly different, reject H_0.

- F-statistic: $F = \dfrac{(RSS_C - RSS_U)/M}{RSS_U/(n-K-1)}$
- RSS_C = Residual Sum of Squares from the constraint equation
- RSS_U = Residual Sum of Squares from the unconstraint equation
- M = # of constraints
- K = # of independent variables in the unconstraint equation
- n = # of observations

4-27

Example of General F-Test

- $Y = \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\varepsilon$ ----- (1)
- Suppose you think $\beta_1=\beta_3=\beta_4=0$
- In other words, $Y = \beta_0+\beta_2X_2+\varepsilon$. ----- (2)
- Therefore, your H_0 is $\beta_1=\beta_3=\beta_4=0$.
- And your H_1: The original model fits the data OK.
- You'll run OLS of (1) and obtain RSS_U.
- You'll run OLS of (2) and obtain RSS_C.
- Substitute RSS_U and RSS_C from (1) and (2) into the F-statistic formula.
- Then, compare your F-value with the critical F-value and make the decision whether or not to reject H_0.
- Note for this example, K = 4, M = 3.

4-28

Chow Test

- Chow test is a test whether two data sets can be combined into one data set because the slopes are not statistically different.
- Put differently, there is no structural change in the model between the two data sets (e.g., before and after a war.)
- H_0: Slopes in the two data sets are not different (no structural change).
- H_1: Slopes in the two data sets are different (there is structural change).

4-29

Steps in Chow Test

1. Run two separate OLS regressions with the same specification for each data set and record RSS from each data set. Call these, RSS_1 and RSS_2.
2. Combined the two data sets into one and run OLS with the same specification again, and record RSS. Call it, RSS_T.

$$F = \frac{(RSS_T - RSS_1 - RSS_2)/(K+1)}{(RSS_1 + RSS_2)/(N_1 + N_2 - 2K - 2)}$$

K = # of independent variables
N_1 = # of observations in sample 1
N_2 = # of observations in sample 2

3. Reject H_0 if F-value > critical F-value.

4-30