ECON 453 - Econometrics
Fall 2023
Exam 1 Study Guide
October 12, 2023

This guide is intended to provide you with a summary of the topics we have covered in the first half of the class. I will also be posting practice problems to provide you with more examples of what you will see on the exam.

1. **Introductory topics/Main themes** – The goal with this section of the course was to introduce some of the main factors that need to be considered in any econometric estimation. These are things you should be able to think about/apply to the different examples of studies we discuss.
    a. Themes to remember throughout semester
        i. The source of data is important in any quantitative analysis
            1. Not really any way statistical methods can fix a bad dataset
            2. Consider issues such as whether data is self-reported (especially for measures like income)
        ii. Don't forget the small stuff
            1. Check data visually before performing advanced statistical analysis
            2. Check descriptive statistics, look for outliers, etc.
            3. Clear presentation of data goes a long way
        iii. Correlation ≠ causation
            1. Finding a strong correlation between x and y does not necessarily imply that x causes y
            2. Need to consider confounding factors, theoretical foundations
            3. Should not be convinced of *causal* relationship based on correlation or simple linear regression
        iv. Regression does not automatically solve our correlation issue
            1. Even when including many control variables, may still have fundamental issue regarding causality
            2. Example from class discussion – do increased health expenditures increase GDP, or does increased GDP lead to increases in health expenditures in a country?
            3. Strong correlation between the two, but traditional OLS regression techniques won't help us figure out which variable is x and which is y.
            4. Look for unique/experimental settings when possible
        v. There is a difference between statistical significance and practical (economic) significance.
            1. Does the relationship we have found have a meaningful impact on real outcomes we would care about?
            2. For example, if a certain treatment has a statistically significant improvement on patient health, ask ourselves other questions such as (1) is it cost effective, (2) does it have a real impact on quality of life?
    b. Types of Data – We have mostly focused on *cross-sectional* studies in class so far. I want you to be familiar with the other types of datasets so that you can discuss when each might be preferable/appropriate.
        i. Cross-sectional data: observations on many individuals at one point in time
        ii. Time-series data: observations on one individual at many points in time
        iii. Panel data: observations that follow many of the same individuals over several points in time
            1. Choosing correct dataset involves weighing:
                a. Type of question you want answered
                b. Practicality of acquiring data

c. Confounding variables issues

i. For example, panel data may help eliminate some omitted variable bias by comparing changes within individuals over time.

d. Statistical problems/issues inherent in each type of data

iv. Note that time-series and panel data models will introduce new statistical issues that we will discuss in more detail in the future

2. **Regression Basics** – In the next part of the class, we discussed some fundamentals about how regression analysis works, what we should pay attention to in the results, and how to use the information to interpret relationships, make predictions, discuss statistical significance, and consider explanatory power.
    a. Characterize the relationship
        i. Which is the dependent (y) variable? Which is the independent (x) variable? What type of relationship between the two do we expect? (linear, negative, strictly platonic, etc.)
    b. Scatterplots – while a basic concept, looking at scatterplots is important for helping to determine the nature of the relationship between variables. They also help us look for outliers/mistakes/measurement issues with the dataset
        i. Should distinguish between explanatory variable (on x-axis) and response variable (on y-axis)
        ii. Each point represents an individual from the dataset
        iii. Interpretation
            1. Form - linear, roughly linear, u-shaped, etc.
            2. Direction – positive, negative, none
            3. Strength – how much scatter?
            4. Unusual Features
                a. Outliers – either in one direction or from the overall pattern of the relationship
                b. Subgroups/clusters
    c. Correlation (r)
        i. Attempt to quantify the strength of the relationship between two variables
        ii. r will be between -1 and 1
        iii. Used to interpret strength and direction of relationship
            1. Positive or negative – from the sign of r
            2. Strength – closer to 1 or -1 indicates strong relationships
    d. Simple Linear Regression
        i. We want to find an equation that best describes the linear relationship between the explanatory and response variables
            1. Gives more information than simple correlation, and can be used for predictions and forecasting
            2. Helps us quantify specific relationship (e.g., impact of cigs on baby weight) so that we can determine policy, make better life choices, etc.
        ii. Ordinary Least Squares (OLS) regression
            1. The method we use to find this equation is to minimize the squared error (how far off the equation predicts y for each individual)
            2. The equation we end up with is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
            3. The coefficients in this equation are found by using:

$$\hat{\beta}_1 = r\frac{s_y}{s_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

a. and generally given directly in the form of computer output

b. You need to be able to interpret coefficients

i. The constant (or intercept) tells us our prediction for y when x is 0

1. This may not always make intuitive sense – it tells us what value of y we predict if all x variables are equal to zero, which may not be realistic for some models

ii. The slope (β_1) is the coefficient we are really interested in

1. Tells us slope – how much will a one-unit increase in our x variable, on average, cause y to increase or decrease by?

2. Example: $\hat{y} = 13 - 2.5x$, the slope coefficient tells us if x increases by one unit, we predict y will decrease by 2.5.

iii. Make sure you are paying attention to units of measurement when interpreting coefficients

1. Percentage vs. percentage point?

a. Has something been logged? No, well then why are you saying "percent"???

2. Income measured in 1000s of $?

3. In some cases, useful to scale to give more intuitive explanation of results

4. R^2 – The coefficient of determination

a. Tells us the percentage of variation in y explained by our regression model

b. Is between 0 and 1, with larger numbers indicating we will get better predictions from our equation.

i. Mathematically can be found by using

1. SST = total variation in y

2. SSE = variation in y explained by regression

3. SSR = Sum of squared residuals, the amount of variation in y left unexplained

4. R^2 = SSE/SST or R^2 = 1-(SSR/SST)

5. Residuals

a. How far off are the predictions from our equation for each individual in the sample?

i. $\hat{\mu}_i = y_i - \hat{y}_i$ (actual value of y – predicted value)

b. Residual plots

i. Error (residual) on y-axis and explanatory on x-axis

ii. Helps check validity of homoskedasticity assumption (see below)

iii. Don't want to see a pattern (like a trumpet), would ideally observe constant variance (homoskedasticity)

iii. Issues with simple linear regression

1. Assumptions need to be satisfied (see list in multiple linear regression section)

2. In particular, simple linear regression generally has a problem with confounding factors biasing our results

a. Violates the "ceteris paribus" condition

b. Example: can't trust coefficient predicting how cigs affect baby birth weight when so many other things not accounted for (age of mother, family income, other health choices, etc.)

c. Even if we only care about one particular x variable and how it impacts y, generally need to add more explanatory variables to control for other important factors.

e. Multiple Linear Regression – As discussed above, simple regression between y and one x-variable often omits too much. In order to both more accurately predict y and more precisely identify the impact of a particular x variable on y, we move to multiple linear regression.

i. Working with multiple linear regression equations

1. Make predictions, find residuals, etc. as we did with simple regression, just need to have values for each x variable
2. Key to interpretation is to remember ceteris paribus. Each coefficient (β) is telling us the impact of that variable on y, holding the others constant.
3. We care about three things with the coefficients

a. Sign (positive or negative)

b. Magnitude (how big of an impact does a change in x have on y?)

c. Significance (is the relationship between this particular x and y statistically significant? See testing section below)

ii. Explanatory power of regression model

1. R^2 tells us percentage of variation in y explained by our model

a. Problem – will always increase as we add more x variables, even if they are not significant

2. *<u>Adjusted R^2</u>* corrects for this by accounting for how many explanatory variables we have in our model

a. Taking out a variable that is not significant will cause R^2 to go down but should not have much of an impact on adjusted R^2.

3. We tend to emphasize adjusted R^2 when reporting results from multiple linear regression models

iii. Need to think about issues with multicollinearity (see next section)

3. **Testing the Validity of our Regressions.** For the next part of the course, we considered how much we can trust the results of our regression models. We discussed some of the common issues that arise, how to detect them, and how to deal with them.

a. Assumptions – in order for our estimation to be unbiased and efficient, we need the following

i. Linear relationship between y and x (as specified). Note that this might mean the relationship between x^2 and y or ln(x) and y is roughly linear

ii. Good dataset (from random sample, etc.)

iii. Each x variable has independent variation observed in our dataset, and there is no perfect linear relationship between x variables

1. i.e. you cannot put rushing yards, passing yards, and total yards in model predicting NFL wins (since rushing + passing = total)
2. Need enough observations so that there is sufficient unique variation in each of the x variables

iv. Expected value of our residuals, conditional on each x, is equal to 0.

1. "Exogeneity" assumption
2. Need the residual (unexplained part of y) to be uncorrelated with the x variables we have in our model

a. If we tend to over or under-predict for individuals with a certain range of x (high-income households, for example) then we have a problem

v. Homoskedastic errors

1. Need the variance of our residuals to be constant for all values of x (i.e. don't want accuracy of predicted spending on clothes to change as income increases)

2. Check visually using residual plots

a. Do not want to see a pattern as value of each x changes

vi. If these 5 assumptions are true, then OLS is BLUE (best linear unbiased estimator)

1. Fun, because it rhymes

2. Unbiased means $E(\hat{\beta}_1) = \beta_1$

3. Best means this gives us the most accurate predictions of any linear model

4. Overall, means we can generally trust the results of hypothesis tests

b. Multicollinearity

i. When we have strong correlation between two or more x variables

ii. Multicollinearity is present to some extent in every multiple linear regression model, it is a question of severity

iii. What problems does it cause?

1. The basic problem is that it is hard for the model to determine which x is causing changes in y if there are two or more x's that are highly correlated

2. In terms of our results, it can cause

a. Incorrect signs on coefficients

b. Just generally incorrect coefficients (biased)

c. Incorrect standard errors

i. In particular, imprecise estimates of coefficients (large standard errors)

ii. Leads to incorrect inference

d. Strange, nonsensical results

i. Really high R^2 values with few or no variables that are significant

ii. Variables we know are important coming up insignificant or with incorrect sign

iv. How do I know if this is a problem?

1. Common sense/theory

a. GDP and poverty in a state are likely to be strongly correlated, including both to try and explain crime rates will be problematic

2. Results are strange

3. Examine correlation tables

a. How high are pair-wise correlations between explanatory variables?

b. No hard rules, but for example, anything above 0.8 or below -0.8 is definitely a concern, anything above 0.5 or below -0.5 might be causing problems

4. Variance inflation factor (VIF)

a. Found by considering how much of each explanatory variable can be predicted by other x variables in the model

i. VIF = $1/(1-R^2_j)$

                - ii. R^2_j is the R^2 from regressing variable x_j on the rest of the x variables in your model
            - b. Each x variable has a VIF
            - c. Values of >5 cause for concern, >10 really concerned, but sometimes smaller values are actually trouble (especially in smaller datasets)
    - v. How do I correct for this problem?
        - 1. Get more observations
        - 2. Drop one or more of the variables that are highly correlated
        - 3. Check various specifications
            - a. How sensitive are key coefficients to changes in set of explanatory variables
        - 4. Cry alone in a dark room
- c. Heteroskedasticity
    - i. What is it?
        - 1. The absence of homoskedasticity
        - 2. Essentially means our errors violate constant variance assumption
            - a. For example: the residuals (how far off our predictions are) get larger as the value of our x variable increases
    - ii. What problems does it cause?
        - 1. Good news: our coefficients are not wrong
        - 2. Bad news: our standard errors are incorrect
            - a. This means t-stats, confidence intervals are incorrect
            - b. Can't trust your findings in terms of statistical significance
            - c. OLS is just LUE now, which is just sad
    - iii. How do I know if this is a problem?
        - 1. Check your residual plots
            - a. Plot errors against each x variable
            - b. Plot errors against predicted y-values
            - c. Want to see random scatter with no pattern
        - 2. Formal tests – several options, we will use White's Test (usually results are same across different tests)
            - a. Basic idea: use residuals from your regression and regress these against some combination of the explanatory variables
            - b. Most statistical packages will perform these (and other) heteroskedasticity tests for you
            - c. In each case hypotheses for test are:
                - i. H_0: Homoskedastic errors (no problem)
                - ii. H_A: Heteroskedastic errors
            - d. Check the p-value from computer output
                - i. Low p-values (below .05) mean reject null, which means we have a problem
    - iv. How do I correct for this problem?
        - 1. Check your model specification
            - a. Perhaps you have omitted an important variable
            - b. Perhaps you should take logs of skewed variables
            - c. Perhaps you can restrict the sample
        - 2. Estimate robust standard errors
            - a. Doesn't change coefficients (they weren't incorrect) but adjusts standard errors to account for hetero problem
            - b. Most statistical packages (gretl, R, Stata, etc.) will be able to do this for you, not something you will do by hand

c. Note: do not want to apply correction in cases where heteroskedasticity is not a problem (don't just automatically use robust errors in all models, they are incorrect in absence of heteroskedasticity).

d. Omitted variable bias (OVB)

i. When we leave important variables out of a model

ii. What problems does it cause?

1. The problem is this might make the estimated coefficients on the variables that are included biased

a. Example: trying to predict income using education. We know innate ability/work ethic should be in there but don't have a way to measure. As a result, the coefficient on education (how much education impacts income) is biased.

iii. How do I know if this is a problem?

1. Ask yourself two questions (silently)

a. (1) Is the omitted variable correlated with y?

b. (2) Is the omitted variable correlated with the explanatory variable I am interested in?

c. If the answer is yes to both, multiplying the signs of the two correlations will tell you the direction of the bias

d. Also factor in the magnitude of these correlations. If you think one or both of these have minimal correlation, bias is unlikely to be a big problem

iv. How do I correct for this problem?

1. Try to find some variable to measure the omitted variable

2. Acknowledge that the bias is present and state that the true coefficient is likely to be lower/higher/etc.

4. **Model Building** – In the next section of the class, we expanded our statistical "tool belts" to include new types and forms of variables in our regression models. As we learn these, it is as if a whole new world of statistical understanding appears before our eyes, and the true joy of econometric modeling enters our hearts. Here we learn about the wonders of logarithmic and quadratic transformations, dummy variables, and interaction terms. One big theme of this section of the course is that we are learning different modeling options that may be useful in different situations. We do not use log transformations in all models, for example, only when it improves our estimation.

a. Logarithmic transformation of variables

i. Useful for various reasons

1. Modeling a relationship that has a theoretically exponential function (such as production or utility functions in Economics)

2. Modeling nonlinear relationships we observe in the data

a. Using a logarithmic transformation can be a better fit to describe relationship (e.g., health expenditures and life expectancy, or law school rankings and salary examples)

b. Dealing with variables that are skewed or have outliers (such as income)

i. Much like the holidays, taking the logarithm helps bring individuals closer together (condenses the distributions) while preserving the same order

ii. This can reduce the influence of outliers/skew on our estimated relationships

3. Interpreting coefficients as percentage changes and elasticities.

a. We can approximate the coefficients on logged variables as telling us what happens when a 1% change occurs, as opposed to a 1 unit change

i. Get this tattooed on your person: *logs = % changes*

b. Note, pay attention to which variables are logged, as the interpretation changes when both X and Y are logged, only X is logged, etc.

b. Quadratic variables

i. Give us a way to allow for relationships that are non-linear (curved relationships).

ii. Example: age of house and price. Generally value drops with age, but once house gets really old, price actually tends to go back up (unless it is haunted).

iii. May include both X_1 and X_1^2 as explanatory variables if we think this is the case

iv. Interpretation of the quadratic term can be difficult, but our predictions and explanatory power of model can be improved

1. The signs of the linear and quadratic terms tell us the general shape of the relationship. For example, if the coefficient on X_1 is positive and the coefficient on X_1^2 is negative, then we would say that as X_1 increases, Y increases, but at a diminishing rate. If X_1 is positive and the coefficient on X_1^2 is also positive, then we would say that as X_1 increases, Y increases at an increasing rate.
2. We tend to focus on how the overall relationship is estimated (diminishing returns, for example) and statistical significance, rather than the interpretations of individual coefficients.
3. When using these equations to predict, need to plug in both X_1 and the squared value of X_1 in the appropriate places (never plug them in inappropriate places)

c. Dummy variables

i. Give us a way to include categorical variables such as race, type of movie, gender, region of country, etc.

ii. Can only do 1 or 0 values. If only 2 categories, need 1 dummy variable, if 4 categories, can include up to 3 dummy variables, etc.

1. Including a dummy for every category provides redundant information and violates one of our OLS assumptions

a. It also violates Dan's trust

iii. Omitted category is the "reference category" that the estimated coefficients on other dummy variables will be compared to

1. In example of life expectancy across countries in lecture, dummy variables for Oceania, Asia, Europe, South America, Africa are included, so North America is the omitted region, coefficients on each of other regions tells us how countries in that region relate to countries in North America in terms of infant mortality rates
2. Can use F-tests, theory to tell us if we need to include all categories, or just certain ones

a. For example, in cross-country health regressions, should I include a dummy for each continent (except North America) or do I just need to include one for Africa (if all others are statistically insignificant)

3. Dummy variables can be a good way to control for omitted variable bias. For example, when doing cross-sectional analysis on health or economy across countries, many factors are hard to control for. Including dummies for region can alleviate some of this concern.

iv. Should I turn a quantitative variable into a dummy variable?

1. This applies to the case where we have a choice
2. For example, could record a variable as whether or not the majority in a state voted Republican (0 if No, 1 if Yes), or could record it as what % of population voted Republican in previous election
3. Quantitative includes more information, more variation, better statistically. Making qualitative (dummy) may make it easier to communicate results to wide audience, and we may mostly care about qualitative outcome (for example: we care who wins the game, but not necessarily by how much).

v. Interaction terms with dummy variables

1. Give us a way to determine if the effects of other explanatory variables differ by category.
2. A dummy variable, by itself, allows the intercept to be different for different groups.
    - a. E.g. do earnings differ by gender?
3. The interaction term allows the slope of the relationship between another X and Y to be different across groups
    - a. E.g. does the effect of a bachelor's degree on earnings differ by gender?
4. Can use our interaction terms/dummy variables to create separate estimation equations for each category
    - a. Example from IC5 of separate equations for the relationship between hours worked and family situation (married/kids) for males and females

5. **Testing our Coefficients** – In this section, we focused on the various hypothesis tests that determine the statistical significance of the relationships we are examining with our models.

a. Hypothesis testing of individual coefficients – is there a relationship between X_1 and Y?

i. Setting up hypotheses

1. Claim we are trying to find evidence of goes in the alternative hypothesis (H_A), opposite (default) goes in the null hypothesis (H_0)
2. Reference the parameter from the population, not the statistic (use beta, not beta-hat)
3. Most commonly, we will set up as:
    - a. $H_0 : \beta_1 = 0$
      $H_A : \beta_1 \neq 0$
    - b. The null hypothesis is that there is no significant relationship between X_1 and Y. The alternative is that a significant relationship exists
    - c. Can also test against any hypothesized value, such as a benchmark, or certain impact we need to surpass to achieve cost-effectiveness:
        - i. For example: do the results indicate that increasing X_1 by 1 unit will increase Y by at least 2 units? Can set up hypotheses as
        - ii. $H_0 : \beta_1 \leq 2$
          $H_A : \beta_1 > 2$
        - iii. This is not very common, but is possible

ii. Finding and stating conclusions

1. Find p-value - practically speaking, both t-statistic and p-value are generally from computer output, not calculated
    - a. Can quickly determine significance (or not) of many variables
        - i. P-value < 0.05 standard benchmark for significance

ii. P-values between 0.05 and 0.10 are in a little bit of a gray area (weak significance)

2. “Reject null” or “Fail to reject null”
   a. These are your two choices for conclusions. If you say something like “accept null”, you and I have to fight
   b. If p-value<α, reject null
      i. For coefficient test, this will mean there is a statistically significant relationship between x and y
   c. If p-value>α, fail to reject null
      i. No significant relationship between x and y
   d. Be sure to state what you found in practical terms, relate back to the question at hand

b. Inference for overall model
   i. F test for overall significance
      1. If there are three explanatory variables in our model, null will be written as: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
      2. Alternative will be that the variables are jointly significant
      3. Tells us if our set of x variables (as a group) has any usefulness in predicting y
      4. F – stat and p-value tell us whether we should reject null or not.
      5. This is the F-stat and p-value presented in standard computer output, you generally do not need to make this calculation on your own
      6. Not a very stringent test of significance. If we fail to reject null, we are saying our regression model has no explanatory power.

c. F-tests for joint significance of a subset of variables
   i. Similar to above, used to test joint significance of some, but not all, variables in the model
      1. Unrestricted model (UR) has all variables in it. Restricted model (R) removes the ones we are testing. The number of variables we are testing is called number of restrictions (q).
   ii. Example from class – was adding the regional dummy variables to our model predicting COVID death rate (PS1) helpful?
      1. Can test joint significance of the regional dummies
   iii. Suppose we have 5 explanatory variables and want to test joint significance of 4th and 5th (should the X_4 and X_5 variables be in the model, or are we better off without them?)
      1. Set up hypotheses: $H_0 : \beta_4 = \beta_5 = 0$
      2. In gretl, we will write as: b[4] =0, b[5] = 0.
      3. Check p-value. If p-value is < 0.05, we will reject null, which says variables X_4 and X_5 jointly add explanatory power to model

d. Wald Test for Equality of Coefficients
   i. Used to test whether the coefficient on one variable is significantly different from the coefficient on another variable in our model
   ii. Generally, hypotheses will be set up as: $H_0 : \beta_1 = \beta_2$
   $H_A : \beta_1 \neq \beta_2$
   iii. Note: this is generally performed as an F-test but requires a complicated standard error. We rely on the computer to do the calculations here.
      1. In gretl, we write as b[1] – b[2] = 0
      2. Low p-values -> reject null (coefficients are significantly different)
   iv. The conclusion for this kind of test is *not* whether the variables add joint significance. The conclusion is whether or not the impact of X_1 (on Y) is significantly different from the impact of X_2. For example, is there a difference

(in terms of negative effects on baby birth weight) for a mother who smokes a pack of cigs per day vs. a mother who drinks a case of Miller High Life (the champagne of beers) per day, (all else equal)?

1. Example from class – does impact of mother's education level on young male's wages differ from impact of father's education level?

e. Standardized Coefficients – a related issue. Give us a way to compare the relative impact of different independent variables with different scales.

   i. Example: in original regression, we are trying to determine if education or IQ is more important in explaining wages, but education is measured in years and IQ is measured in points.

   ii. Tell us how much a one standard deviation increase in X variable will affect Y (in standard deviations)

      a. Calculation: $b_1^* = \hat{\beta}_1 * \frac{Std\ Deviation\ of\ X}{Std\ Deviation\ of\ Y}$

   iii. Not very useful in terms of predicting Y, main function is just to compare magnitudes of effects across explanatory variables