ECON 453 - Econometrics
Fall 2023
Exam 2 Review

This guide is intended to provide you with a summary of the topics we have covered in the second section of the class. I will also be posting practice problems to provide you with more examples of what you will see on the exam. Please note that this exam focuses primarily on the material we have covered in the time since the first exam. This material is discussed in more detail in the review that follows. However, some of the material from the first section of the class naturally carries over as we continue to work with multiple regression. This means that you should be familiar with the following from the first section of the class:

- Interpreting coefficients (sign, magnitude, significance)
- Model strength – (R^2 and adjusted-R^2)
- How to set up a model/characterize relationships
- Basic violations of our assumptions (heteroskedasticity, multicollinearity, OVB, etc.)
- Logarithmic transformations, quadratic terms, dummy variables, interaction terms

1. Limited Dependent Variable Models – Here we discuss what to do if the dependent variable in our model is categorical. Examples we discussed in class include predicting whether or not a person is married and whether or not a person voted.
    a. Linear Probability Model
        i. What is it?
            1. Essentially, we ignore the fact that the dependent variable is categorical (binary) and treat this as any other OLS regression model
        ii. How do we interpret coefficients?
            1. A 1-unit change in x will increase the probability that Y=1 (for example, that the person is married) by $\hat{\beta}_1$ *percentage points*.
        iii. What are some problems that may arise with LPM?
            1. Predictions outside realm of possibility, for example, we see predicted probabilities that estimate the probability a person in married to be less than 0, or greater than 1
            2. Linear models mean no flexibility in interpretations (for example, each year increase in age has same impact on probability of being married, regardless of age level)
            3. Often will have a major problem of heteroskedasticity because of the nature of your y variable, other problems with OLS assumptions are likely as well
        iv. Why do some people use LPM anyway?
            1. Convention in particular fields of research (comparing with previous studies)
            2. Easier to interpret/work with
            3. Results generally end up similar to those from Logit/Probit, despite statistical issues
                a. Interpretations of coefficients from LPM and marginal effects from Logit/Probit often essentially same magnitude
            4. A greater number of casual readers may be able to understand information as compared to more advanced models – LPM choice makes it easier to communicate findings
            5. If using more advanced models (interaction terms, fixed effects, etc.), models such as Logit/Probit may not work appropriately (or require significant extra work to interpret).
    b. Logistic Regression (Logit Model)
        i. What is it?

1. A model that specifies the relationship between binary y and several x variables in a specific, non-linear way
    a. Uses an exponential relationship between y and x variables
    b. $y = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}$
    c. Uses maximum likelihood estimation techniques, rather than OLS methods

ii. What are the benefits of using Logit over LPM?
    1. Limits possible outcome predictions between limits of 0 and 1
    2. Better statistical properties (LPM often violates assumptions needed for inference)
    3. Allows for more flexible interpretations of coefficients
        a. e.g. the effects of getting a year older on the probability of being married can change as the starting age of the person changes

iii. How do we interpret coefficients?
    1. Very carefully
    2. Coefficients from initial computer output have little straightforward interpretation
        a. Change in x will cause change in log-odds ratio of y
        b. Focus on the sign and significance of these coefficients
            i. Use z-stat instead of t-stat, but interpretation of p-values is the same
        c. These "raw" coefficients are needed to estimate predicted values of y (probability of y occurring).
            i. Plug into the exponential equation above to get estimated value of y-hat
    3. Need to compute marginal effects for more meaningful interpretations of results
        a. Note in gretl these are reported in the "slope" column of the results
        b. Most commonly, this means take a partial derivative with respect to one x-variable while holding all others at average level
        c. Increasing x by 1 unit will, holding all other variables constant (at average values), cause the probability of success (y=1) to increase by the estimated marginal effect (in % points)
        d. Can be adjusted to estimate for different levels of all (or some) x-variables
            i. This flexibility is one of the main advantages over LPM, as mentioned earlier
            ii. This is both a blessing and a curse
                1. On the one hand, this is a more realistic fit to what happens in real life
                2. On the other hand, there are infinite combinations of variables to estimate marginal effects for. Need to keep in mind that many non-econometricians might be overwhelmed.

iv. What about significance of our Logit models overall?
    1. Pseudo R^2

a. Since this is not an OLS model, there is not the typical R^2 calculation
b. Gretl and other programs will try to replicate the amount of variation explained, so that this can be interpreted similarly to traditional R^2
c. There are many methods of computing pseudo-R^2, which means we should probably not put too much faith in this statistic.

2. Percent correctly estimated
    a. Another way to assess strength of our model
    b. Compute predicted probabilities for each individual in sample based on model
    c. Choose a benchmark for predicting y =1 (predicting person is married, for example)
        i. Typically if predicted probability (y-hat) is greater than 0.5, consider this as expected success
        ii. Sometimes the benchmark is the mean probability from sample
    d. Compute percentage estimated correctly
    e. Note that we can use this method with the linear probability model as well.
        i. In gretl, we calculate this "by hand" (save fitted values, make a dummy based on cutoff of 0.5, and so on as we did in Problem Set 3).

c. Other types of models
    i. There are many, many variations of limited dependent models. We discussed some of these briefly, but you should be aware of some of the highlights
    ii. Variations on Logit models
        1. Binary
            a. This is the one we focused most of our attention on, most commonly used (did the person get a graduate degree, yes or no?)
        2. Ordered
            a. Used when there are more than 2 categories, but they go in a logical order (1 = bachelor's, 2 = master's, 3 = graduate/professional)
        3. Multinomial
            a. Used when there are more than 2 categories but they do not follow logical order (example, which major do people choose, accounting, economics, marketing, or finance?).
            b. Ordered and multinomial logits can be difficult to work with, but keep in mind the idea of positive/negative effects, significance, and explanatory power. In more advanced models, we tend to rely on these more than specific interpretations of each coefficient.
    iii. Count Data (Poisson Regression)
        1. Useful when we have a discrete variable that has a heavy concentration/skew
            a. Example from class: dependent variable = # of children that people have.
                i. Discrete count: 0, 1, 2, etc.
                ii. Heavy concentration on 0

2. The skewed/categorical nature of the variable means OLS results are often biased.
3. As with discussion of Logit models above, the interpretations can be tricky/not worth trying to communicate
    a. Focus on sign/significance of each variable
    b. Example from class: Economics majors have a negative and significant coefficient in the "kids" Poisson regression, so that means they will tend to have fewer kids.
4. In situations like this, should think about/compare your modeling options.
    a. Turn # of children into binary (has kids/not) and model with LPM or binary Logit
    b. Use Poisson Regression
    c. Break into two different questions:
        i. How does gender/major/etc. influence decision to have children?
        ii. How does gender/major/etc. influence decision of number of children to have, given that there will be some children?
            1. Restrict sample to those with nonzero children

2. Applied Econometric Models – At this point in the class we took a break from our focus on statistical aptitude to think a little bit more about research design in Economics. Given problems like omitted variable bias, what are some methods that are used to help identify the actual relationships between variables?
    a. Endogeneity
        i. The problem we are trying to solve, in a nutshell
        ii. Could be caused by many factors: omitted variable bias, confounding factors, reverse causality, simultaneity, measurement error, etc.
        iii. Means you cannot trust that your estimated coefficients are identified correctly.
        iv. We are imagining a scenario (such as informing decisionmakers on how to create public policy) where identifying correct magnitude of relationships is key.
            1. It is not enough to know that scholarships increase the education level in the population, want to know the numeric relationship so we can decide efficient level of investment
    b. Difference-in-difference (D-in-D)
        i. Idea is to create an experimental setting with a treatment and control group based on naturally occurring events
        ii. Most common application is when a policy change occurs in one area, compare before and after with comparable area(s) that did not have policy change
        iii. Basic method is to find difference (after – before) for treatment group (call it D1) and difference (after – before) for control group (call it D2) and then find difference in the difference D3 = D1 – D2. This D3 is the effect of the policy
            1. Example – did vacancy rates change in buildings near Chicago landmarks in period after 9/11. Compare areas near landmarks (treatment) to areas further away (control) before and after September 11, 2001.
        iv. Can be implemented using dummy variables and interaction terms
            1. Using a regression model, rather than simple averages allows us to:
                a. Control for more factors

b. Test for statistical significance of any effects we find
c. The coefficient on the interaction term (treatment*after) tells us the effect we are looking for.

v. Key assumption that needs to be argued: the control group is the appropriate counterfactual for what would have occurred absent the treatment
1. Want to show, for example, that the two groups were on "parallel paths" before the treatment occurred

c. Instrumental Variable Analysis
i. Idea is to try and use a third variable (Z) to clarify the relationship between the variables of interest (X and Y).
ii. If Z causes a change in X that can be though of as "exogenous", we can use this to identify more clearly how changes in X directly affect Y
1. The I.V. (Z) purges the endogenous variation from X
iii. The instrumental variable (Z) needs to satisfy two main requirements:
1. Needs to significantly impact the X variable of interest
2. Needs to not directly impact the Y variable
iv. Showing the first condition above is relatively straightforward
1. Check the relationship in a regression, how large is the first stage F-statistic?
v. Proving the second condition is tricky
1. If Z affects X, which then affects Y, then Z will be correlated with Y even if it does not have a direct relationship
2. Rely on theoretical arguments here
vi. Estimation can be thought of as two stage process (2 Stage Least Squares (2SLS))
1. Stage 1: regress X on Z, save the fitted values
2. Stage 2: regress Y on the fitted values of X from stage 1
a. These fitted values have been purged of endogenous variation (in theory). They are changes in X that were caused by exogenous shock of Z changing
b. These estimated coefficient in second stage is the one we will use as correctly identifying the relationship between X and Y.

3. Time-Series Models – In the next section of the course, we move on from the comfort of cross-sectional datasets and proceed to analyzing time-series data. The basic nature of these datasets often makes it more challenging to identify the actual relationships between explanatory and response variables. At the same time, time-series data is often the type we need to analyze to answer the most important questions in life, such as: what causes GDP to grow in a country?
a. Basic issues in time-series models
i. Dataset that follows an individual over several points in time. Note that in statistics, an individual could be a person, city, country, etc.
ii. Sample size is often very limited
1. In the case of Macroeconomic variables, for instance, may only be able to observe once per year, and data may not be recorded very far back in time.
iii. Need to be concerned about spurious results
1. If two variables each have an upward trend, it can be very easy to show one has statistically significant impact on the other, when in reality there is no causal relationship.
2. This can often lead to inflated t-statistics on individual variables and very high values for adjusted R^2.
3. Need to interpret results with caution, consider modeling choices.
iv. Timing issues become a big part of our models

1. Deal with trends in the data
    a. Discussed further in next section
2. Lagging variables
    a. May include lagged values of explanatory variables if the relationship between x and y takes time to develop (e.g. relationship between increase in tax credits for having kids and an increase in fertility rates).
    b. May include lagged values of dependent variable to help account for the natural trend in the data, improve accuracy of predictions/forecasting

v. Static Model
1. This is a simple model that essentially treats the data as a cross-section. Does not do anything to account for the nature of the data, and will often have major problems.

vi. Time trends
1. Often the first choice when working with time-series data is to include trend variable
    a. Typically, just a variable that goes from 1 to T, in order, for your dataset, where T is the number of observations
2. Helps account for the natural pattern in many variables over time
3. Note that we can include t^2 and other variations to try and fit for variables that have non-linear trends but may not leave much room for other x variables to be significant.
    a. Goal of most analysis is not to simply describe how y varies over time, but to see which explanatory variables influence y once we account for the trend
4. Occasionally a similar method is utilized in which we de-trend variables in a two-step process
    a. Step 1: estimate a simple model where y-variable is predicted only by time trend.
    b. Step 2: use residuals from step 1 model, which represent "de-trended y" and regress on set of explanatory variables.
    c. One benefit of using this method is that we get a more meaningful value for R^2.
        i. What percentage of variation in y can our x-variables explain, once we have accounted for trend in y?

vii. First-differences model
1. Another way to deal with the trend that occurs in variables over time
2. Our dependent variable is now the change in y from one period to the next, and our explanatory variables are changes in the x values.
3. May alleviate serial correlation concerns present in other time-series models
4. A couple other things to note here:
    a. You will lose one observation from your sample (there is no change in the first period observed)
    b. The R^2 will often be much more reasonable and meaningful as compared with other time-series model choices
        i. Tells us the percentage of the variation of the change in y

viii. Including lagged dependent variable (autoregressive models)
1. In these models we include the lagged value of the y-variable as an explanatory variable. After all, isn't the best predictor of this year's y last year's y?

2. Helps account for the trend in the y-variable specifically
3. May be difficult to find significance in other explanatory variables
4. Allows us to estimate how increase in x this year may affect y for several years to come

ix. Event Study
1. A fancy term for including a dummy variable for certain time periods in our analysis
2. This can be a dummy variable that is 1 after a certain period (such as after President Obama was reelected in our approval ratings example), or something that is "turned on" and then "turned off" again. For example, we could add a dummy variable controlling for the month when Osama Bin Laden was killed (caused a large, temporary increase in Obama's approval rating).
3. A structural break means we have a change in the relationship between our x variables and y variable after a certain point in time.
4. We can check for a structural break based on something we know occurred at a certain point (such as 9/11), or we can decide to check based on looking at the data and noticing a change in trends.

b. Serial Correlation
i. What is it?
1. When our error terms are correlated between observations
2. This is almost exclusively a problem seen in time-series data
a. Errors are correlated between time periods

ii. What problems does it cause?
1. As with heteroskedasticity, serial correlation generally means our coefficients may not be wrong, but the standard errors are incorrect.
a. This means our inference procedures (t-tests and confidence intervals) will give us incorrect conclusions
b. Often serial correlation will greatly inflate the t-stats and R^2 values – making our model too good to be believed

iii. How do I know if this is a problem?
1. Sometimes, you just know
2. Check a plot of the residuals against time
a. Do not want to see a pattern
3. Check a correlogram in gretl (or other program)
4. Durbin-Watson test
a. Null hypothesis: No problem with serial correlation, rejected this null means we have a problem that needs to be dealt with
5. Other tests
a. There are many different variations of tests, usually they will give us the same conclusion, but might check several if results are close
b. Gretl will run several different tests from the results window of a regression using time-series data.

iv. How do I correct for this problem?
1. Try including time trend variable(s)
2. Try including lagged values of dependent variable
a. May cause a bigger problem (with heteroskedasticity) than the problem you started with
3. Run a "first differences" regression where you look at how changes in the x variables explain changes in the y variable

a. $y_t - y_{t-1} = \beta_0 + \beta_1(x_t - x_{t-1}) + \ldots$
b. may solve problem of autocorrelation but also somewhat changes interpretation of your model and results

4. Advanced modeling techniques (mentioned, but not covered in detail in this course): ARIMA, ARCH, GARCH, etc.
5. Statistical correction techniques – Note we did not cover these specifically in this course (examples: Cochrane-Orcutt, Prais-Winsten, etc.)

NOTE: We will not be covering panel data before the exam, so this material (other than the concept of panel data) will not be on the exam. I am including it here for your information.

4. Panel Data - Over the last few lectures of the course we will discuss some of the basics of using panel data to answer some of life's most puzzling questions. These include questions like: how will an individual's income change after the birth of a child, what factors explain differences in crime rates across U.S. cities, and how is it possible that people will pass Econometrics if they have attended 2 or fewer lectures the entire semester? I want you to be able to explain how panel data might help answer some of these questions, why it might be preferred to other types of data, etc. The key factor here is the ability to control for unobserved variables (that are time-invariant) by observing the same individuals over time.
    a. What are the main benefits of panel data?
        i. More variation
            1. Variation both across individuals and within individuals over time
        ii. Better identification – allows us to control for more unobserved heterogeneity
            1. i.e. can control for variables that do not change over time for an individual in fixed-effects estimation
        iii. More observations
        iv. Can provide more general predictions about relationships between variables than cross-sectional (which focus on one particular time period), or time-series (which focus on one particular individual)
    b. What are some potential drawbacks of working with panel data
        i. More complicated, cumbersome to work with
        ii. Can be difficult to pinpoint what exact relationships are causing results in larger models
        iii. May not be answering the question we have in mind. For example, if we want to know how health outcomes affect economic growth in the U.S., is it appropriate to use a panel dataset of several countries to examine this relationship?
        iv. May introduce more problems with errors, and these may be complex problems without well-defined solutions
            1. Correlation in errors both across space and time
        v. Some techniques do not allow for estimation of coefficients on certain important variables
            1. If you put in a dummy variable for each individual for example (fixed-effects), will not be able to estimate coefficient for variables that do not change for individuals over time (such as race).
    c. Modeling Options
        i. Pooled Model
            1. Simplest method (aka amateur hour)
            2. Treats multiple observations from same individual as independent

        a. For example - in crime data, would treat as 92 independent observations, rather than as 2 observations each from 46 cities
        b. Will have a problem in our error terms if there is correlation in variables within a city over time
        c. Really only works in case where you assume no omitted variables are specific to individuals, there is not strong correlation in variables within individuals over time

ii. Fixed-effects estimation
    1. Idea is to control for time-invariant unobserved variables that are unique to each individual
        a. Should not assume this deals with all omitted variable bias issues (some factors change over time)
    2. Differencing variables for each individual across time periods, using dummy variables for each individual will yield equivalent results
        a. Option 1: use the change in crime in each city from one period to next as y, calculate changes in x variables
        b. Option 2: use a set of dummy variables for each city in the sample (except one city)
    3. Cannot estimate coefficients for variables that do not change for an individual over time
        a. For example, region in which each city is located does not change across time, so fixed effects model cannot identify coefficients on regional dummy variables
    4. Coefficients interpreted as effect of x changing within an individual over time on y
        a. For example: how do changes in health care expenditures within a country over time affect GDP growth?
        b. Another example (from class): how do changes in average SAT scores of incoming class for a given university over time affect the freshman retention rate at that university
        c. The reported final coefficient is the average across individuals in your sample, not for any one in particular