Reminders for self

* None

Questions:

* 1) Does explanatory power still matter when the dependent variable measured is logged? I would assume so but how so because you are predicting percentages?
* a. Likewise does the constant matter and what does it mean in our regression where the dependent variable is logged?
* 2) Comparing the p-values the regression at the beginning of 5 where economics is the reference category and to the p-values in the test for equality where it is between economics and accounting. They are very similar p-values but slightly different. Does this mean anything? Should they be the same or similar at all?
* Example of when 0.057 p-value is okay
* Interpreting graphs for heteroskedasticity
* Would you ever group comparison states together if very similar especially when it comes to policy
  + Related to in class assignment#8

Notes:

9/22/2023

1. Do the order of the regressors matter?
   1. A: No, they should not because the regressors are being added to give y-hat. The order of addition should not matter.
2. P-val clarification
   1. Meaning
      1. A: Indicates if a relationship between the regressor(s) and dependent variable exists
      2. A: Is the probability that the null hypothesis is rejected?
      3. A: If low, it is likely that there is a relationship between the dependent variable and the regressor.
         1. Means there is a slope to their relationship that is not zero.
            1. The slope can be positive or negative.
      4. A: There are P values for if the constant is equal to 0 and if the slope is equal to zero
         1. The one in gretl in the OLS regression are p-values for the coefficients which is slope.
   2. If multicollinearity exists which regressor will have the lowest p-value
      1. A: The one that is least likely to have a relationship between the dependent variable and itself.
3. Is there a general tradeoff between p-val and R^2?
   1. A: There shouldn’t necessarily be although there can be.
4. R^2
   1. Is there ever much of a difference between R^2 and adjusted R^2 if there is only 1 regressor
      1. A: There can be. The difference between R^2 and adjusted R^2 is tinier the closer n is to k in n > k where n is the number of data points in our sample for the regression and k is the number of regressors in the regression.
      2. If there is
         1. What does this mean?
         2. What would be an example of this?
            1. A: If n >>> k then you may potentially see this.
5. Changing out a binary dummy variable
   1. Equal and opposite magnitude
      1. A: Yes
   2. Coefficients of other variables stayed the same or changed – Question 2bii problem set 1
      1. A: They actually all stay the same except the constant because the coefficient of the regressor(s) affected by this were added or subtracted from the constant to account for this affect. The constant is essentially the same.
      2. Democrat vs Republican governor
         1. Constant changes but other regressors stay the same
            1. A: They actually all stayed constant because the coefficient of the regressor(s) affected by this were added or subtracted from the constant to account for this affect. The constant is essentially the same.
6. Are negative coefficients problematic in a regression?
   1. A: NO! I was confusing this; it is that you can’t prove a negative. This relates to the p-values. If p-values are high, there might still be a relationship because high p-values don’t prove that a relationship doesn’t exist – just based on p-values in the model here, we can not support the evidence of relationship. The coefficients can still be negative.
7. Are broad groupings of the reference category problematic?
   1. Black vs everyone else

9/\*/2023

NEED TO ASK:

1. Correlation matrix
2. Why not just look at 1 variable if high adjusted R^2
   1. Why look at multiple variables

Notes:

9/5/2023

* Regression
  + Trying to get the most accurate representation of the relationship
    - High R-squared
    - Low P-value
  + Regression Equation types
    - Linear
      * y = b0 + b1x1
      * Unit to unit change
    - Quadratic
      * Unit to unit change
    - Logarithmic
      * ln(y) = b0 + b1x1 +b2x2
      * Unit to percentage change
      * Questions to ask when considering log
        + Do I want to measure by percentage
        + Is there a skew in our variable

Argument for logarithm

* + - * + Are there lots of 0’s as entries in the data

Log transformation will turn them into missing values

9/7/2023

* Make sure you clean the data before you use it
  + Clear missing values
    - !missing(variable)
  + Remember to clear restrictions if you intend to
* Things to look at when you run a regression
  + Number of observations in the regression
  + Coefficients of variables
  + R^2
  + Pval
* R^2
  + Between 1 and 0
    - The closer it is to 1 the more explanatory power our model has

9/12/2023

* Lecture notes - Working with regression
  + Expected to be able to
    - Interpret
      * Coefficients
      * Variables
    - Predict
    - Significance
    - Explanatory power
  + Modeling options
    - Linear
    - Quadratic
    - Log
    - Dummy variables
    - Interactions
      * Multiple variables interacting I believe
        + I think done through multiplication
      * Check to make sure
      * Leaving out all related dummy variables except the interaction
        + It is comparing the interaction category to all the related left out categories which are in the reference category
  + Validity
    - Diagnosing problems
    - Corrections
      * Not typically straightforward
  + Example
    - Cigs per day (x) and birth weight (y)
      * Scatter plot regression is linear y = mx + b
        + Here x is 0 > x > 1
      * True formula
        + Y = B0 + B1X + mu
      * Estimate
        + Y-hat = B0-hat + B1-hatX
      * Residuals
        + MUi-hat = yi - yi-hat
  + Ols regression
    - Min squared residuals
      * Choose B0-hat, B1-hat to do so
      * B1-hat = r \* Sy / Sx
  + Assumptions
    - 1) Linear in parameters
      * Linear relationship between y and x, as specified linear relationship between (ln y) and x
    - 2) Data from random sample
    - 3) No perfect collinearity
    - 4) Exogeneity assumptions
      * E(mu | x) = 0
        + E = expected value
      * Residuals are correlated with x
      * Zero conditional mean
      * When looking at a scatter plot you want to see and equal range for x above and below the regression line
      * Examples
        + 1) If you saw this with the cigs per day (x) and birth weight (y) example

Likely cause is omitted variable correlated with x (income is correlated with smoking)

* + - * + 2) Effect of education on earnings for 25-year-olds

Two groups

In first group, points are mostly above the line and in second group, points are mostly below the line

Endogenous errors

The reason for these errors is there is a missing variable with explanatory power in our model (here it is actually experience)

* + - 5) Homoskedastic errors
      * Var(u | x1, x2, …, xu) = sigma^2
        + Var = variance
        + u = mu
      * Need variance in errors to be similar or x changes
      * Heteroskedastic
        + Data points look like an inequality
        + Trumpeting of errors

Where the data fans out

* + - * Concern about inference
        + P-values

Messes with them?

Sometimes this is all we care about

* + No perfect collinearity
    - There needs to be variation each x
      * (Sx = 0) then B1-hat is undefined
    - You can have n - 1 variables so that there is no perfect collinearity in a regression
    - Extreme collinearity is a concern
      * Example of perfect colinearity
        + Y-hat = B0-hat + B1-hatcigs + B2-hat-male +B3-hat-female
    - Independent variable after controlling other fators
    - Minimum rule concerning collinearity for running a regression
      * n > k
        + n = number of observations
        + k = number of parameters
        + This is a requirement
      * n >>> k
        + This is what we should be shooting for
        + n is way way larger than k
* In class work notes
  + When omitting a variable and choosing one to be a reference to avoid perfect collinearity
    - If you choose the lowest one as a reference, then all of the coefficients of your related variables are positive
      * Some say this is the best
  + Residuals
    - * You want to have the independent variable as x and the residuals (mu-hat)
      * You want the graph of the residuals against x to be a horizontal scatter plot (line slope = 0)

9/14/2023

* Ordinary Least Squares (OLS) Regression Assumptions
  + OLS
    - Choose the line in the data that minimizes the variation in the y
      * Squares the residuals
      * Min SUM(yi - yi-hat)^2
  + 1) Linear in parameters (as specified)
    - Y = B0 + B1X1 + B2X2
  + 2) Data from random sample
    - Ex: D8 vs G8 Healthcare Econ performance thesis student
      * The problem was that those weren’t a random sample
        + He had 200 countries of data to draw upon

Why not use it

* + - * + The D8 and G8 were an intentionally selected sample not random

Would have been better if random

* + 2) No perfect collinearit
    - Independent variation in each variable
      * Ex: Y = B0 + B1X1 + B2X2 + B3X3
        + Y = birthwaight
        + X1 = cigs
        + X2 = income
        + X3 = Age
        + X4 = weight if mom
        + X5 = weight of dad
        + Do we have enough variable data to explain these variables – need a variety
  + 4) E(u | x) = 0
    - u = mu
    - E = error term
    - Exogeneity
  + If 1 - 4 hold, OLS estimates are unbiased
    - E(B1-hat) = B1
      * If biased should be E(B1-hat) = B1 + some bias term
  + 5) Homoskedastic errors
    - Var(u | x1, x2, …, xu) = sigma^2
      * Var = variance
      * u = mu
    - Errors don’t vary in range as x values change
      * Points are spread out
  + If 1 – 5 are true, then OLS is BLUE
    - Best
    - Linear
    - Unbiased
    - Estimator
  + Modeling Issues
    - 1) Too few variables is a problem (omitted variable bias)
    - 2) Too many variables is also a problem (multicolinearity)
      * Multicolinearity
        + There is not enough variation between the variables

Add certain variables are redundant to the explanatory power

* + - 3) What to do with insignificant variables
      * No significance is still worth while knowing
      * If two variables are heavily correlated, together they might not be significant in the model, but on their own they might be

9/19/2023

* Multicolinearity
  + Detection
    - How do we know we have a problem? These might point to one
      * Unexpected results
      * Odd combinations of statistical significance/explanatory power
        + High R^2 but low p-values
      * Sensitivity of coefficients to model specification
        + If they change a lot depending on if new variables are added to the model then multicollinearity might be a problem
        + If there is a sensitivity between variables
        + Then include the combination of variables in your model with the highest R^2 while also including all the variables your looking for

Look at example in lecture slides

* + - We will have collinearity to some extent
      * All variables are correlated to some extent
        + Is it too much?

Should we start to deal with it?

* + - * Sometimes you just can’t have two or more variables together
        + Redundant variables
  + Testing
    - Correlation matrices
      * If there is high correlation values between two variables might be problematic
        + Look at example in lecture slides
      * Not perfect tests
    - Variance inflation factors (VIFs)
      * Benchmarks
        + Values above 5 = probably a problem
        + Values above 10 = probably a problem
      * Not a perfect tests
      * VIF = 1 / (1 – R^2)
      * Look at example in lecture slides
  + Correction
    - If we have a problem, you can probably
      * Increase sample size
      * Change model specification
        + Probably most likely route
        + Change to some nonlinear regression?
        + Create some categorical variables
        + Address colinearity
      * \*\*\*Carefully\*\*\* interpret coeeficients
        + Example from slides:

Health matters, water matters, GDP doesn’t matter if you’ve accounted for the previous two variables

NOT:

Health matters, water matters, GDP doesn’t matter

* + Which combination of variables should you go when the variables are sensitive with collinearity
    - Priorities
      * 1) If you are trying to understand a specific variable, include it in the model
      * 2) Which one has the best combination of explanatory power and statistical significance
      * 3) Which one conceptually makes the most sense o include in the model
* Law school problem
  + File on canvas
  + Variables
    - School rank
    - GPA
    - LSAT
  + Constant is negative
    - Reasons
      * LSAT
        + On a scale of 120 to 180

Could adjust scale so that its from 0 to 60 to make constant more positive

* + Collinearity
    - Removing GPA
      * R^2 is near the same
      * GPA doesn’t contribute to much too explanatory power in the model with all 3 variables
      * LSAT nearly doubles
    - Removing LSAT
      * R^2 is near the same
      * LSAT doesn’t contribute to much too explanatory power in the model with all 3 variables
      * GPA goes up
* Adjusted R^2
  + Adjusts R^2 as we add more explanatory power
  + SSR = residual sum of squares
  + SST = total sum of squares
  + Adding more variables is always going to increase R^2
    - But are those variables meaningful
      * This is when we look at adjusted R^2
        + Penalizes for adding meaningless variables
  + What it means
    - Indicator of add variables if meaningful
  + Adjusted R^2 increases as you add meaningful variables
  + If there is little difference between R^2 and adjusted R^2, means all variables are pretty meaningful

9/21/2023

* Two main reasons to run regressions
  + Be able to forecast/predict the y-variable
    - We might not care about collinearity here and instead just care about explanatory power (being able to predict y variable)
      * What will Amazon’s stock price close at at the end of the day
  + Investigate relationships between variables
    - How was Amazon;s stock price was affected by stay at home orders
* Detection
* Standard Error
  + Matters because it is comes from the coefficients and it is used to calculate the t-value which is then used to calculate the p-value
* Heteroskedasticity
  + Detection
  + Testing
    - We’ll know with this if ther’s heteroskedasticity
    - Regress residuals on different forms of explanatory variables
      * White’s Test
        + Most common used
      * Breusch-Pagan
      * If there is a low p-value means we need to deal with heteroskedastic errors
        + At or below 0.05 is problematic
        + Null hypothesis is that there are no heteroskedastic problems
  + Correction
    - We have good solutions to solve the problem
    - If there is a problem means, there is some factor we’re not considering
      * Address by changing model specification
        + Non-linearities
        + Subsamples

Omit observations

* + - * + Omitted variables
    - Correct standard erros
      * Calculated using a matrix
      * Robust standard errors
        + Only check the box when running the regression if you know there is a heteroskedasticity problems
        + Coefficients are not changed by this

Standard errors, t-ratio, and p-value are affected by this

They should be more accurate if there was a heteroskedasticity problem

* + - * What does it mean to expand the standard error
        + Means to account for more variance in the model

Will lower standard errors, t-ratio, and p-value

* + We care about the standard error here
    - If there are heteroskedastic problems we really shouldn’t trust the standard error
* Quantitative variables
  + Represented by numbers and percentages
* Categorical variables
  + Yes or no
  + Present or not
* Dummy variables
  + If line is not a good fit to the data you can consider dummy variables to break up the data
  + Binary variables that take on values of 0 or 1
  + Categories
    - A category is left out of the model
  + Useful in regression models in numerous ways
    - Incorporating categorical information
  + Specifying non -linear relationships with quantitative variables
  + Simplifying our message
  + Interactions?
    - How they interact with other variables
  + Dependent variable?
    - When the dummy is the dependent variable
  + Example of baby weight dependent on cigarette smoking
  + Interpretations
    - Key is to remember the idea of reference category
    - Compare results between the dummies

9/26/2023

* Outliers
  + When to throw them out
    - Are the data points we’re considering to be outliers being driven by the relationship we’re trying to explore
      * Maybe they’re driven by other things
    - Are you unfairly throwing out data points
      * Are they a crucial part of the story
        + Maybe the relationship you are exploring is causing the outliers
    - Example:
      * Unemployment 2020 of states being driven by stay at home orders and unemployment in 2019 in the same states
        + Hawaii and Nevada were outliers in 2020 unemployment by stay-at-home days

Can be explained by the loss of tourism during the pandemic

Should you consider throwing out because stay at home days might not be driving the outliers

* When omitting a dummy variable from a collection
  + Generally, omit the highest or lowest
* Testing equality of coefficients
  + Are two variables significantly different from one another
  + Look at linear restrictions
    - Can compare the beta of two variables together
    - Is the difference between the two variables statistically significant
      * H0: Beta(x1) = Beta(x2)
      * Beta(x1) - Beta(x2) = 0
        + If p > 0.05 than they are not really different from each other, and one can be used to represent both
  + Look at slides for today for examples

9/28/2023

* Statistically significant
  + Can’t tell if two variables are significantly different from one another if they aren’t in the same model
    - Once in the same model, you can do a linear test
      * Ex:
        + H0: b\_eur = b\_na
        + b[4] – b[6] = 0
  + If looking at if more than two variables are statistically different from one another
    - Look at f-test then
      * Example in gretl linear test:
        + b[3] = 0
        + b[4] = 0
        + b[6] = 0
        + b[7] = 0
        + Low p-value means the variables are statistically different from one another
        + High p-value means we should take out the other variables as they aren’t statistically different
  + Interaction Term
    - I think this will indicate if they are statistically different by variables in interaction
      * Ex: If Immunized, Africa, and Immunized\*Africa are in the model
        + Then we will see how immunizations differ by Africa and the rest the world
    - Ex:
      * Y = 67.06 + 0.5GDP – 7.6Africa + 0.42(Africa\*GDP)
        + “Other” (Africa = 0): y = 67.06 + 0.5GDP
        + Africa (Africa = 1): y = 59.46 + 0.92GDP
    - Ex: Weekend\*Christmas
      * Variable for if Christmas and weekend are both true
      * If significant, the drop in births is meaningful due to the combination
    - Ex: School\_Rank\*GPA
      * The effect of increasing GPA decreases as the school rank increases.
* Testing coefficients
  + Standard test of significance
  + Testing joint significance of a set of coefficients
  + Testing equality if coefficients
  + Testing around a benchmark value

10/3/2023

* Semester Project
  + Theme
    - Topic
      * What influences the outcomes
        + Economic?
        + Cultural?
        + Regional?
        + Religious?
        + Policy?
  + Sample
    - How are you getting your data
    - How it started and how you altered it
      * Describe a lot here
  + What are you trying to explore
    - What do you expect
    - What regressions are you covering
    - How does what your exploring compare to the surrounding context
      * To the world?
* Testing
  + Standard test of significance
    - H0: B1 = 0
  + Testing Joint Significance
    - H0: B1 = B2 = B3 = B = 0
      * Look at F statistic
        + Same as p-value cutoff
        + Look at formula
  + Testing Equality of coefficients
    - Use linear restrictions
    - Same if pvalue > 0.05
    - H0: BSouth = BWest
      * b[3] – b[4] = 0
    - H0: BSouth = BNortheast = bWest
      * b[3] – b[4] = 0
      * b[4] – b[5] = 0
  + Look at slides for example
  + When comparing whether or not one is more impactful
    - You could look at the increases after one standard deviation
      * Use sols
      * Go to file > Function packages > Functions on server
        + To install
      * Go to file > Function packages > Functions on local machine
        + To use

10/3/2023

* Omitted variable bias
  + Unbiased
    - On average our expected value of our coefficients reflect the true population
    - Delta is the coefficient when leaving out the coefficient of x variable that isn’t omitted and the coefficient of y variable that is omitted
  + Look at example online in slides
    - Leaving out IQ will inflate impact of education
    - Look at this when you don’t have data for a variable
      * In this instance it is IQ

10/10/2023

* Review
  + Regression Basics
    - Interpretations
    - Predictions
      * Plug and chug from regression equation
        + Use all of the coefficients in the equation, not just the statistically significant coefficients
    - Residuals
      * Residual = Actual – Predicted
    - Significance of coefficients
      * Be smart in how you interpret p-values
        + Is 0.057 p-value statistically significant

Standard cut-off is 0.05

However, it might be okay in some circumstances

* + - Explanatory power of models
      * Some areas of data make it more difficult to get high R^2 in our regression
        + Country level data will have high R^2
        + Predicting incomes will likely have low R^2
  + Review Topics
    - Multicollinearity
      * Diagnosing
        + VIF

Correlation between explanatory variables

VIF that is high indicates collinearity.

X>5 might have a problem

X>10 have a problem

* + - * + Correlation Matrix
        + Model Sensitivity
    - Heteroskedasticity
      * Diagnosing
        + Whites test

Low problem means heteroskedastic.

* + - * Solutions
        + Model specs
        + Robust standard errors
    - Modeling Options
      * Log Transformations
        + Converts coefficient units to percentages
      * Quadratic Terms
        + Look at both linear term and quadratic term

Helps you imagine what the graph looks like

* + - * Categorical Variables
        + Reference Category

What variables are in the reference category

Can be 1 or more and all the the dummy variables of the category are referencing the respective reference variable

* + - * + Quantitative -> Categorical?
      * Interaction Terms
    - Other Stuff
      * Testing coefficients
        + I think this deals with things like omitted variable bias
      * Standardizing Coefficients
      * Model Sensitivity
      * F-test
        + H0: b1 = b2 = 0

Joint significance

Gretl

b1=0

b2=0

Nothing going on, x1 and x2 do not predict y at all

As a group they add nothing

Different than H0: b1 = b2

Equality of coefficients

Are x1 and x2 having different effects in the model

Gretl

b1-b2=0

* + - * + Low p value indicates our model is better than nothing

10/17/2023

* Categorical dependent variable
  + Example
    - Which factors predict whether someone votes in an election
* Linear Probability Model
  + Involves a yes or no hypothesis question
* Logit/probit models
  + Involves a yes or no hypothesis question
  + Tradeoff between statistical properties and ease of interpretation
  + Options
    - Binary
      * Yes or no
    - Ordered
      * What level of education did you have, order to it
    - Multinomial
      * What major did you choose
  + What does it do
    - Bounds the range two the categories
      * If binary range is 0 to 1
    - Formula: y-hat = (e^(B0 + B1X))/(1 + e^(B0 + B1X))
  + Pros
    - Better fit to the data
      * Errors are more correct
  + Cons
    - More complicated mathematically to explain
    - May not work well for all modeling options
  + How to evaluate the model
    - The percentage your prediction was correct
  + Marginal effects
  + % Correctly Predicted
    - What percentage we correctly predicted yes and/or no
  + Comparing results
    - If results are similar between logit and ols then you might default to using OLS because it is easier to manipulate

10/19/2023

* Exam 1
  + Make sure when you using explanation in interaction terms, controlled variables don’t count as an explanation for explaining interaction terms
    - Controlled would be if its in the interaction model and not one of the categorical variables that could be in the combination
      * IE hours in the male kids question is a controlled variable
* Logit and probit
  + Stuff to pay attention to in output
    - Slope
      * These are marginal effects
      * This is what you look at when interpreting variables
      * Interpreted at +1 unit from the average I think
    - Coefficient
      * This is what you use when predicting values
    - Predict vs actual table below
  + When the results of the logit/probit are quite different than the OLS
    - This is because OLS is problematic to use as a model

10/24/2023

* When exploring and doing statistical analysis on a sample
  + Think in the context of your sample
  + If a category in your sample has quite a low range of values when considering he possible ranges
    - Then you might want to consider combining categories
  + OLS vs Logit
    - If tons of prediction data points are above 1 in OLS model, then you will probably want to use a logit model
      * Groups the predicted data points past 1 to the group right before 1
  + Linear probability model (LPM) is OLS

10/26/2023

* Logit/probit
  + Nonlinear model
    - Can calculate at any level
  + Tradeoffs
    - Although more statistically accurate
      * When working with interaction terms in logit model, it can be problematic
        + Programs like stata can correct for this but gretl is not that careful
  + Trying to model the best way
    - If you think there is nonlinearity in the situation
      * Could do LPM with categories
        + Look at frequency distribution wide gaps in the middle could be bad
      * Or could do logit model
* LPM with categories to model nonlinearity
  + Could do variable = to a specific constant
  + Could do variable chain where inequalities ar used (<,>)
    - Adding incrementally
    - From examples
      * “At least some college”
* When explaining coefficients of variables, don’t want to say the variables already accounted for in the model are the cause
* Frequency distribution for OLS
  + Having a skewed distribution is not ideal for running OLS – count data
    - This is called “count data”
    - What are options
      * Change the formulation of our question
        + Binary dependent variables

Y = kids or not?

Y = at least 2 kids

* + - * + Restrict sample to those that have kids
      * Try a different modeling option for count data
        + Normally we assume a distribution is normal but these distributions don’t

Tobit Model

Poisson regression

Left or right skew in data distribution

10/31/2023

* Missed – alarm didn’t wake me up

11/2/2023

* Recap – applied econometrivs
  + Log
    - Typically used to account for skew, to measure in percentage changes,
    - Also used to recover variables in production function
  + Exogeneity
    - Error terms
      * Omitted variable bias
  + Difference in difference
    - Got to have a good comparison group
      * For example:
        + If examining people and their geography and a specific effect after a policy change

Then a potential thing that could make the comparison between groups bad is if people are moving a lot in between groups especially because of the policy

We want to control for this with a better group if that’s the case

Think of situations like these when looking for a good control group

* + - Experimental and control group are on the same trend
      * If experimental group didn’t receive treatment then its assumed that they would experience the same trend
      * You measure the difference between the baseline and actual
    - To measure if a policy was significant
      * Measure using an interaction
        + Between the shock and the difference in effects

11/7/2023

* Example Problem:
  + State wants to know if giving people a state college scholarship if they stayed in the state afterwards or left
    - Two options
      * Option 1
        + Restrict sample to college educated
        + Treatment: Georgia
        + Control: other states
      * Option 2
        + Restrict sample to Georgia
        + Treatment: college educated
        + Control: non-college educfiated
    - Look at nov 7 dataset

11/14/2023

* 2 Stage Least Squared (2SLS)
  + Get rid of endogenous varable and then measure statistic
  + Steps
    - 1) Regress x on z and get fitted x
    - 2) Regress fitted x on y
      * Using fitted x has already had endogenous effect purged
    - If either step is wrong, the whole model is wrong
  + Look at In Class 9 example
  + Technically should correct standard errors but this is difficult to do manually
    - Gretl has models that already do this so these models are technically more correct
* F statistic if under 10 is bad

11/16/2023

* Types of datasets
  + Cross sectional
    - Many countries at one point in time
    - Concerns
      * Endogeneity, omitted variable bias
  + Time series
    - One country at many points in time
    - Concerns
      * Spurious correlation, endogeneity, serial correlation
      * Spurious correlation
        + Something that looks correlated but isn’t
        + Serial correlation is the time series version of this
      * Serial correlation
        + Sometimes referred to as autocorrelation function (ACF) plot
        + Look for patterns with residuals

Is there correlation between observations

Look at graph of residuals

Correlogram

If p-value is very small, then this problem is likely present

Durbin Watson test

0 and 4 are bad and 2 is good

1.5 to 2.5 is acceptable

Serial correlation can be visualized by regressing residuals against the time trend

The visualization should match the results of the Durbin Watson test

* + - New modeling options
  + Panel data
    - Many countries at many points in time
    - Concerns
      * Both?

\*/\*/2023

* Time Series
  + Issues
    - Spurious correlation
    - Serial correlation
      * Be aware of suspiciously high R^2
    - Non-stationary
  + Potential solutions
    - Trend regression model
      * Account for the trend by adding a time variable to the variable
        + R^2 may be very inflated
        + Doing this helps to more accurately estimate the impact of the other variables in the model
        + Example of where this is useful I the relationship between GDP and unemployment
        + Might use a quadratic trend model rather than a linear trend model

If the quadratic one has a higher R^2

* + - * Can add time variables which focuses on the timing of political events to account for them so that we can more accurately estimate other variables in the model
    - First differences model
      * Look at the changes in a variable for each variable between each two points in time and use these in the model
        + From previous example

Difference in GDP

Difference in unemployment

Difference in bachelors

* + - Can combine a time regression model a first differences model
    - Autoregressive model
      * In OLS regression in gretl click on the lags box and add a lag to the dependent variable
        + Durbin Watson can’t that be used to test statistical significance

Durbin’s h is used

0 is good and above or below 2 is bad

12/5/2023

* Writing the paper
  + Introduction
    - Preview findings
      * Heres my finding, heres why they’re interesting, heres the relevance
  + Data and descriptive statistics
  + Empirical Strategy
    - Hypothesis
    - Setting up for how you found the results
    - This could be a part of the introduction
  + Results
  + Conclusions
    - What can be done next
      * Verification of results
* Points of emphasis for paper
  + Justification
    - Why you made the choices you did
    - It doesn’t matter if something is problematic as long as you are addressing it
    - Think about what gives more context
  + Presentation
  + Reflection
* Panel data
  + Type of variations
    - Between an individuals or within individuals over time
      * Golf shots and prize money
        + Mesured within an individual
  + Pooled cross sections
    - Pooled OLS
      * Is a pooled cross sectional regression
    - Include time dummies when running regression
  + First differences acknowledges they are the same objects we’re observing in each regression
  + Fixed effects
  + Can add
    - Unit dummies
    - Time dummies
* For first differences you can’t measure variations in variables that don’t change
  + Difference between regions a city is in between years
    - It doesn’t change