PSC 508

Jim Battista

Univ. at Buffalo, SUNY

Some Simple Problems

A few simple problems

- How to diagnose them
- Sometimes, how to help cure them
- BUT REMEMBER
  - Theory is key
  - Experience in the research area is key
  - Both will tell you when to expect certain problems
  - Overcorrecting for problems is itself a real problem

Collinearity

- AKA multicollinearity
- Say you're predicting inflation with the money supply but there are multiple competing measures of the money supply
  - Naive response: I'll include them all and let the regression pick the most bestest!
- You find: ALL of the money supply variables are REALLY REALLY non-significant
- Likely problem: collinearity

Collinearity

- Collinearity is when two or more IVs are highly correlated
- They're so tightly related that there's not enough independent variation between them for OLS to "understand" the separate effects of each of them

Diagnosing collinearity

- Coefficients are "wrong" or backwards counter to very strong theory or longstanding expectations (ie, Republicans are more likely to vote for Obama).
- IVs are correlated with each other
  - No firm guide as to how correlated is *too* correlated, sorry
  - Doesn't fully address multi in multicollinearity
- High "variance inflation factors" or VIFs (above 10)
- Nothing is significant, but great F statistic and high R2

Generating correlation matrices

- To check correlations between your IVs
  - Stata: `corr iv1 iv2 iv3 iv4`
  - R: `output<-cor(dataobject,use="complete.obs")`
  - R: this will compute matrix for ALL variables in dataset!

Checking VIFs

- Stata: after regression, type "vif"
- R: install package "HH" or package "car" (might differ on pc, mac)
- R: run model as `output<-lm(dv~iv1+iv2+iv3,x=TRUE)`
- R: `vif(output)`

Dealing with collinearity

- Did you include multiple versions of the same variable, or almost the same variable?
  - WELL STOP THAT
  - Or at least combine into index (many ways to do this)
- But you can have collinearity even if you do everything right
- It can be just a feature of the data
- What then?
  - Congratulations! You lose!
    - You really can't distinguish the effects of iv1 and iv2 if they're too collinear
    - Admit it, if only to yourself as you weep, and move on
  - If collinear variables are theoretically related, combine into index
  - Get more data and hope it provides enough leverage for OLS to separate them
  - Run for help

Residual plots

- Residuals: the "mistakes" in your regression
- Actual values minus predicted values
- You predicted DV should be 5, but it's 4, that residual is -1
- You predicted 50, it's 60, that residual is 10
- Plotting residuals (versus some IV) can tell us useful things

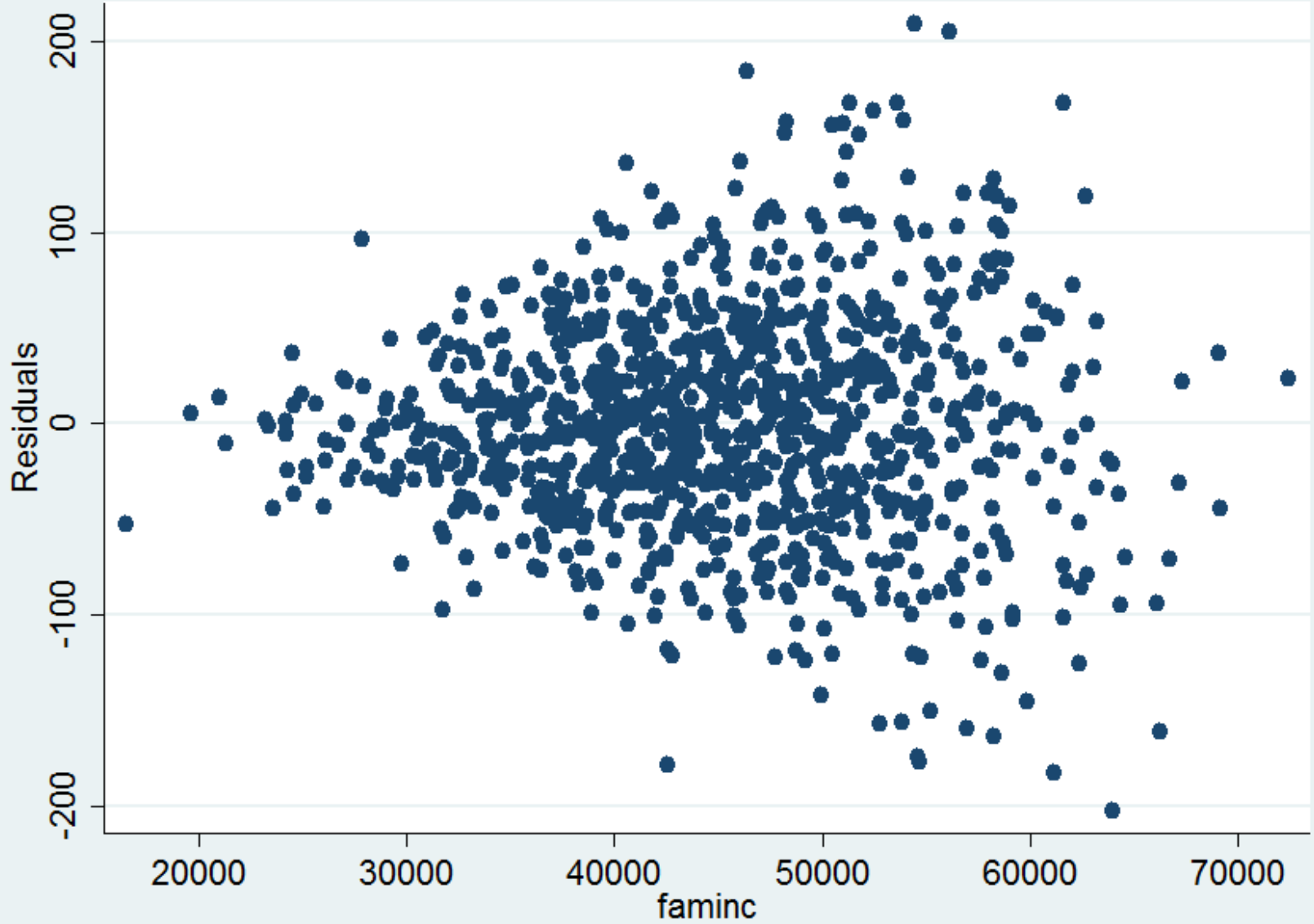
Heteroskedasticity

- Hetero=different, skedasticity=variation
- Variance of errors changes along regression line
- Typical example: errors aren't ± 10, they're $\pm 10\%$ of iv2
  - So larger values of iv2 have larger errors

Diagnosing heteroskedasticity

- Residual plot!
- Stata: `reg dv iv1 iv2 iv3`
- Stata: `rvpplot iv1`
- R: `output<-lm(dv~iv1+iv2+iv3,na.action=na.exclude)`
- R: `residuals<-resid(output)`
- R: `plot(iv1,residuals)`

What does "classical" heteroskedasticity look like? (fake)



But

- Heteroskedasticity doesn't have to be "classical"
- Common test: Breusch-Pagan test
- Stata: run regression, then `hettest`
- R: install `car` package, run regression, then:
- `ncvTest(regressionobject)`
- Either way will report p-value against a null hypothesis of constant variance
- As always, low p implies rejection of that null

Dealing with heteroskedasticity

- Schemes to weight data
  - Run for help
- Robust or sandwich estimators
  - Compute SEs differently to try to deal w/ heteroskedasticity
  - Stata: `regress dv iv1 iv2 iv3, robust`
  - R: Gotta be some way

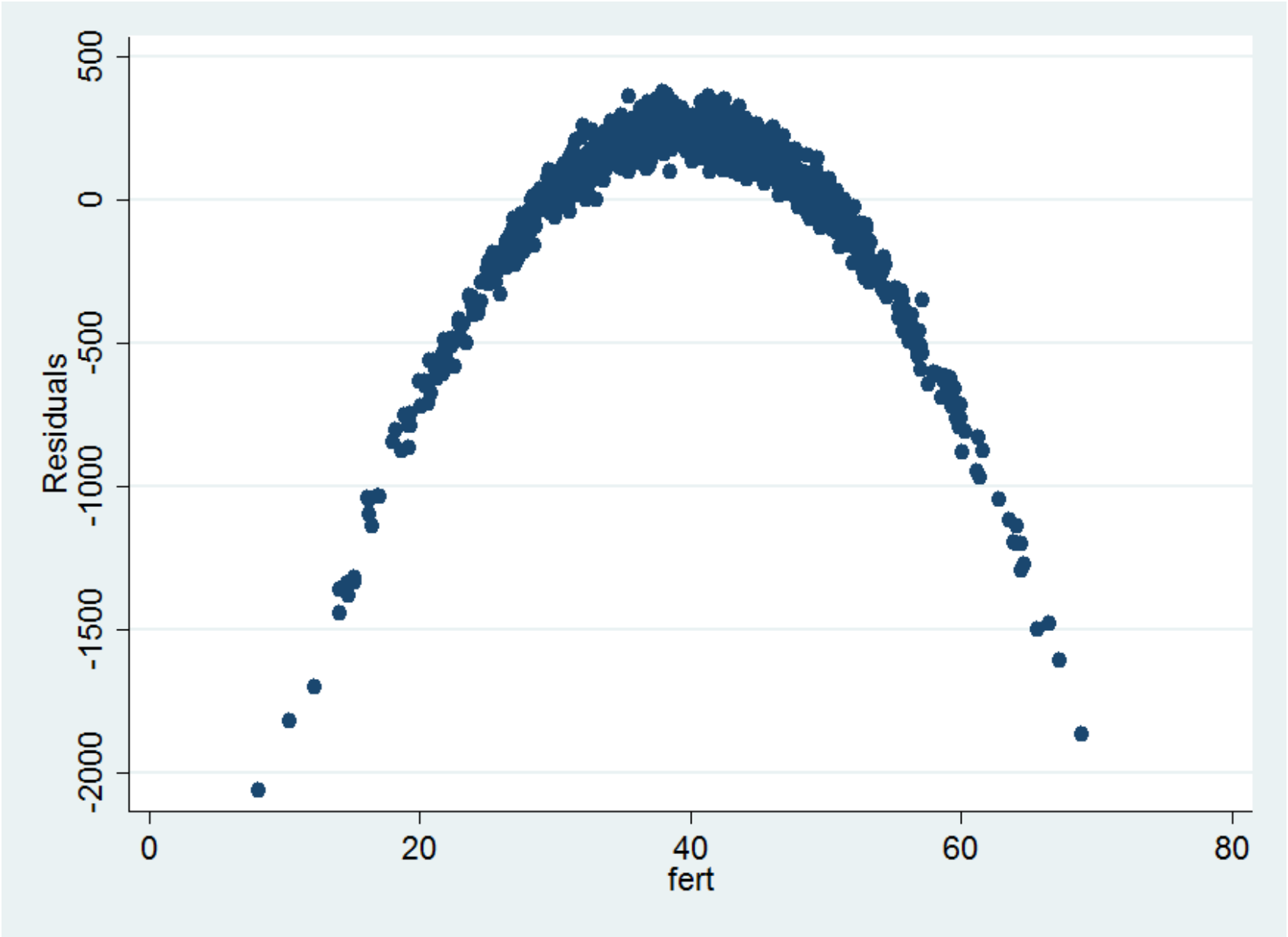
Nonlinearity

- All regressions assume perfect linearity, and fail
- Not a big deal unless "severe"
- Can diagnose some with residual plots
- Usually will know when to expect them

Diminishing returns

- Predict yield with rainfall, fertilizer
- But effects of fertilizer might be nonlinear
  - First little bit helps a lot
  - Next little bit helps less
  - Eventually, some little bit doesn't matter or even hurts

A fake example!



Common solution: squared term

- A common way to deal with these problems is by adding a "squared term"
- In this example, generate fertilizer-squared and add it to the regression
- Stata: `gen fert2=fert^2`, then
  `reg yield rainfall fert fert2`
- R: `fert2<-1*(fert^2)`, then `lm(yield`drainfall+fert+fert2)!
- MAKES INTERPRETING THE EFFECT OF FERTILIZER A GRADE-A PAIN IN THE ASS
- Effect isn't just coefficient on fert, it's also the coefficient on fert2
- DON'T ADD THESE THINGS WILLY-NILLY

Logging

- Sometimes don't want to run regression with "raw" DV or IV
- Expect that effects are nonlinear in a particular way
- Say using city population to predict something
  - Raw IV in regression
    - Coef means that difference between 10,000 and 20,000 is the same as the difference between 1,000,000 and 1,010,000
- "log" the IV
  - Take logarithm of population, include that instead
    - Stata: `gen logx1=log(x1)`
    - R: `logx1<-log(x1)`
  - Now difference between 10,000 and 20,000 is the same as the difference between 1M and 2M

Logging

- Might log DV or IV
- Say y=constant+ b1(x1)
- Logged DV – log(y) instead of y
  - Half-assed interpretation:
  - Increasing x1 by 1 assoc. with b1 times 100 percent change in y
- Logged IV – log(x1) instead of x1
  - Half-assed interpretation:
  - A one percent (NOT PERCENTAGE POINT) increase in X assoc. with a b1/100 units change in y
- Both logged – elasticities!
  - Half-assed interpretation:
  - A one percent increase in x1 assoc. with a b1 percent change in y
- Real interpretations: more complex stuff with e^x scattered about