Econ 312

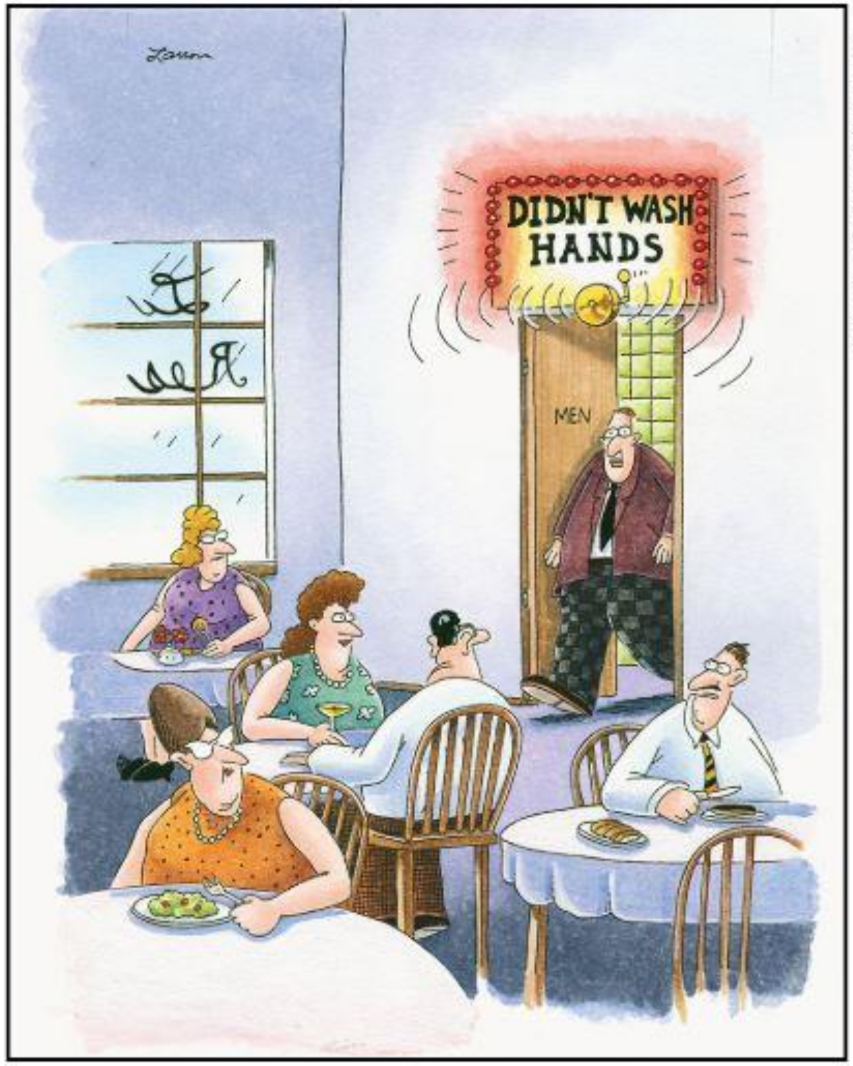
Wednesday, April 29
Missing Data and Quantile Regression

Readings:

- Little, Roderick J. A., "Regression with Missing X's: A Review," *Journal of the American Statistical Association* 87(420), December 1992, 1227-37.
- Koenker, Roger, and Kevin F. Hallock, "Quantile Regression," *Journal of Economic Perspectives* 15(4), Autumn 2001, 143-156.

Class notes: 183 – 186, 193 – 197

Today’s Far Side offering



Timely …

Context and overview

- This class discusses two of the extended topics identified in the last section of the reading list
- **Missing data** are a common problem in econometrics, mostly in cross sections or panel data
  - Sometimes it is acceptable to use the available data that we have
  - Sometimes we can impute values for the missing observations
- **Quantile regression** is a non-parametric procedure that allows us to model not the mean (expected value) of *y*, but the median or any other quantile/percentile of the distribution
  - This can be useful when we are concerned about how *x* affects the overall distribution and the tails, not just the central tendency

When do we have missing data?

- Omitted or invalid responses in surveys
- Attrition in longitudinal surveys
- Data collectors often don’t care: Reed does not require class rank but records it if applicants provide it
- In macro data: inconsistent data-collection procedures over time

Why are the data missing?

- **Missing completely at random** (MCAR)
  - Probability of missingness is unrelated to any variable in our model, including the actual value of the one that is missing
  - This is the most benign situation
- **Missing at random** (MAR)
  - Probability of missingness is unrelated to the actual value of the missing variable
  - May be related to other variables in the equation
- **Not missing at random** (NMAR)
  - Probability of missingness is related to actual value of missing variable
  - This is most problematic case

Complete-case analysis

- This is what Stata does: delete from the sample any observations for which any variable is missing
- We lose information from variables that *are* present
- We are implicitly assuming that the deleted observations fit the model, so residual would be zero and not change anything
- Complete-case analysis does not lead to bias if missingness does not depend on y (or, implicitly, u)
  - This is a standard sample-selection problem
  - We are fine as long as selection into or out of the sample is unrelated to the error term

Available-case analysis

- Note that regression coefficients depend only on the sample variances and covariances of the variables
- Even if one variable has a missing value for an observation, there is still information about the variances and covariances among the other variables
    - We can use this information to help get better estimators
- Awkward and rarely used
    - Uses different groups of observations for each pair of variables
    - No guarantee that $\mathbf{X'X}$ matrix has an inverse

Dummy-variable methods

- Model is $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$
- x_1 is complete; x_2 has missing

$$M_i = \begin{cases} 1 \text{ if } x_2 \text{ is missing} \\ 0 \text{ otherwise} \end{cases}$$

$$x_{2i}^0 = \begin{cases} x_{2i} \text{ if } M_i = 0 \\ 0 \text{ if } M_i = 1 \end{cases}$$

- Two possible estimators

$$y_{i0} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}^0 + \gamma M_i + u_i$$

- Is biased for β_1 because x_1 proxies for unobserved variation in x_2

$$y_i = \beta_0 + \delta_0 M_i + \beta_1 x_{1i} + \delta_1 x_{1i} M_i + \beta_2 x_{2i}^0 + u_i$$

- Is unbiased
- May be difficult to implement with multiple missing variables unless pattern is very simple

Imputation methods

- Most useful with irregular pattern of missingness
- Use values of other variables to impute missing values
- **Unconditional imputation**: replace missing values with mean
  - Bias due to included variables picking up missing variation
  - Not a good idea
- **Conditional on *x* imputation**: predict missing values based on other *x* variables
  - Use complete cases to estimate $x_{2i} = \delta_0 + \delta_1 x_{1i} + v_i$
  - Compute $\tilde{x}_{2i} = \hat{\delta}_0 + \hat{\delta}_1 x_{1i}$ for missing cases
  - Regress $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 \tilde{x}_{2i} + u_i$ with full sample
  - Consistent if MCAR
  - Standard errors are problematic

Multiple imputation with chained equations

- Basic idea: Construct **multiple imputation values** for missing observations by Monte Carlo, including random draws for error term in imputation equation
- Suite of Stata commands under **mi** heading does this
- This gives multiple alternative "samples" with different values for the imputed observations/variables
- Start with **imputation equation** (complete cases) that includes y:

$$x_{2i} = \delta_0 + \delta_1 x_{1i} + \delta_2 y_i + v_i$$

  - Must include y here
  - Can use probit, tobit, or other LDV model if x_2 has limited range

Generating multiple imputations

- Compute imputed samples m = 1, 2, …, M with

$$\tilde{x}_{2im} = \hat{\delta}_0 + \hat{\delta}_1 x_{1i} + \hat{\delta}_2 y_i + v_{im}$$

  - v_{im} is random draw from estimated error distribution of v
- For each sample m, run regression with imputed values

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 \tilde{x}_{2im} + u_i$$

  - Save estimates $\hat{\beta}_{jm}$ and their squared standard errors $\operatorname{var}\left(\hat{\beta}_{jm}\right)$
- **Combine results** of M equations:

$$\hat{\beta}_j = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_{jm} \qquad \operatorname{var}\left(\hat{\beta}_j\right) = \frac{1}{M}\sum_{m=1}^{M}\operatorname{var}\left(\hat{\beta}_{jm}\right) + \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}_{jm} - \hat{\beta}_j\right)^2$$

- Consistent if data are MAR or MCAR

Quantile regression

- Standard regression models conditional mean of *y* as a linear function of *x*
- What about the other properties of the *y* distribution?
  - Sometimes think about conditional variance σ^2
  - If *y* is normal, then we don't need anything else
- If *y* is not normal, then there may be much more information about *y* distribution than is conveyed by mean and variance
- **Quantile regression** models each/any quantiles of the distribution as a function of *x*

Moments and quantiles

- **Mean** of y is μ that minimizes $\sum_{i=1}^{n}(y_i - \mu)^2$
- **Median** of y is m that minimizes $\sum_{i=1}^{n}|y_i - m|$
- **Quantile τ** is ξ_τ that minimizes $\sum_{i=1}^{n}\rho_\tau(y_i - \xi_\tau)$, $\rho_\tau(x) = \begin{cases} \tau x \text{ if } x \geq 0, \\ -(\tau - 1)x \text{ if } x < 0 \end{cases}$
- Can do individual regressions for each desired τ as linear function of x:

$$\min \sum_{i=1}^{n}\rho_\tau(y_i - \mathbf{x}\boldsymbol{\beta}_\tau)$$

Example using Reed GPA data

- Do-file to perform a set of quantile regressions with uggpa:

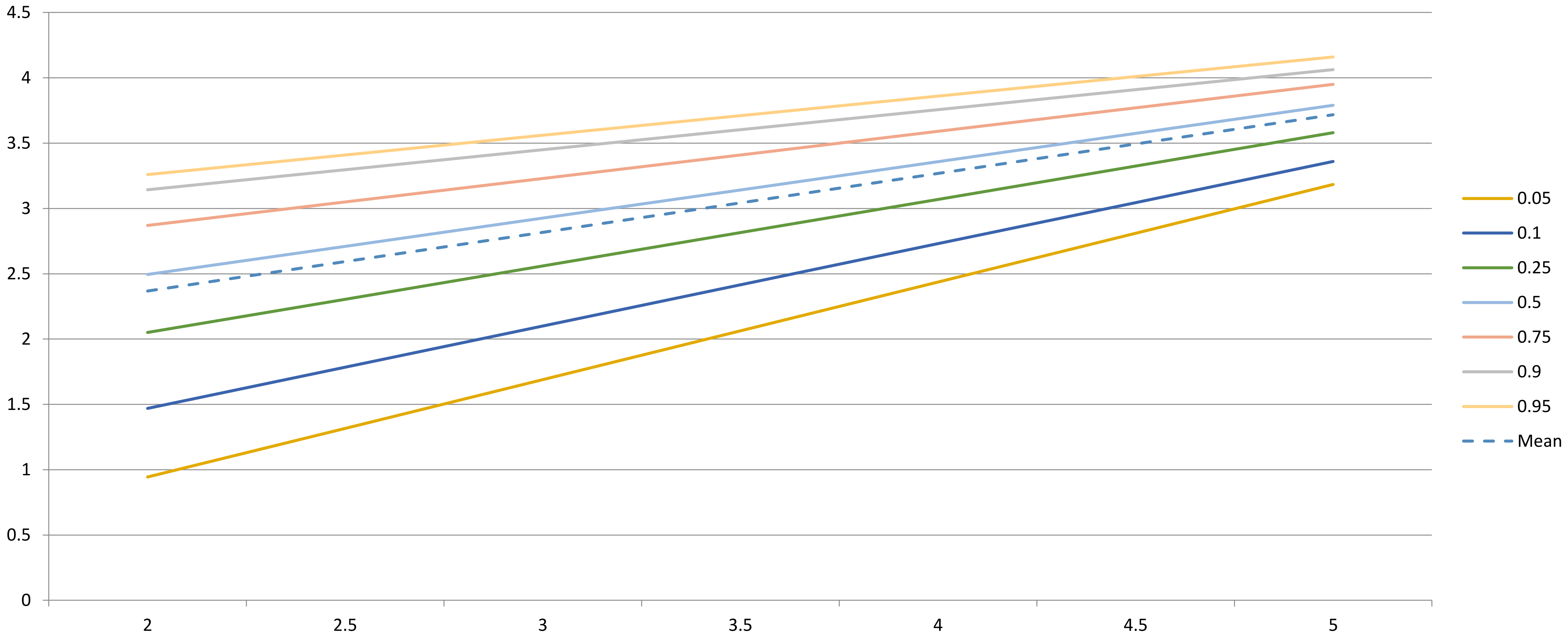
```
foreach qlevel in .05 .1 .25 .5 .75 .9 .95 {
qreg uggpa irdr satm100 satv100 hsgpa female if humfresh , quantile(`qlevel')
outreg using reedqregs , se merge
}
```

- Output is set of regression tables (next slide)
- Can use these to compute fitted values for levels of any *x* and plot (slide after next)

Regression outputs from qreg

	0.05	0.10	0.25	0.50	0.75	0.90	0.95
irdr	0.450	0.377	0.372	0.277	0.223	0.165	0.074
	(0.163)**	(0.072)**	(0.039)**	(0.032)**	(0.028)**	(0.028)**	(0.034)*
satm100	0.116	0.095	-0.012	0.040	0.054	0.068	0.061
	(0.102)	(0.045)*	(0.025)	(0.020)	(0.018)**	(0.018)**	(0.022)**
satv100	-0.064	-0.008	0.067	0.074	0.099	0.090	0.112
	(0.096)	(0.042)	(0.023)**	(0.019)**	(0.017)**	(0.017)**	(0.020)**
hsgpa	0.645	0.538	0.348	0.252	0.212	0.171	0.205
	(0.200)**	(0.088)**	(0.048)**	(0.040)**	(0.034)**	(0.035)**	(0.042)**
female	0.218	0.273	0.141	0.088	0.035	0.034	-0.014
	(0.135)	(0.059)**	(0.033)**	(0.027)**	(0.023)	(0.023)	(0.028)
_cons	-2.321	-1.620	-0.256	0.403	0.771	1.273	1.479
	(0.874)**	(0.385)**	(0.211)	(0.173)*	(0.150)**	(0.152)**	(0.184)**
N	2,230	2,230	2,230	2,230	2,230	2,230	2,230

Plot of predicted quantile values for irdr



Review and summary

- We examined various methods for dealing with **missing data** in our sample
  - Data must be missing at random or missing completely at random to estimate
  - Most flexible and powerful estimate is multiple imputation with chained equations
  - Compute multiple imputed samples and average the results
- **Quantile regression** is a technique for estimating how each *x* affects any part of the *y* distribution, not just the mean
  - Can use this to estimate how each admission variable affects the bottom end of the grade distribution

One, last bad economist joke …

There is a story about the last May Day parade in the Soviet Union. After the tanks and the troops and the planes and the missiles rolled by, there came ten people dressed in black.

"Are they spies?" asked the Russian Premier.

"They are economists," replied the KGB director. "Imagine the havoc they will wreak when we set them loose on the Americans."

--Taken from Jeff Thredgold, *On the One Hand: The Economist's Joke Book*

What’s next?

- Nothing!
- Friday’s class will be devoted to a broad discussion of the econometric implications of the coronavirus pandemic
  - No recorded lecture
  - Just discussion in Zoom conference, which will be recorded for those unable to participate at class time