Section 14 Advanced Topics

Specification searches

- Experiments vs. non-experiments
    - If we can do random controlled experiments, then we don't need to worry about omitted variables bias because the regressor of interest (treatment effect) is random and uncorrelated with everything that might be omitted.
        - Controlled experiments are becoming more common in economics
        - Development projects may choose villages as treatment or control villages
        - Policies can sometimes be separated randomly into treatment or control groups
            - Is it ethical to withhold "treatment" if we know that it is likely to be beneficial?
        - Of course, experimental economics has long put experimental subjects into controlled settings randomly.
    - Most often, we must use the "fallen fruit" of "natural experiments" or observational data
        - Examples:
            - State policy differences such as the seat-belt law regressions we looked at earlier in the semester
            - Cross-country growth regressions in which countries differ in variables such as initial per-capita income that are supposed to affect growth
        - In these cases, we must worry about selection and omitted-variable bias
            - Can we control for the other variables that are correlated with selection into the "treatment group" (or with the regressor of interest)?
            - If not, our results are biased
- Idealized econometric project
    - Theory tells us exactly which variables should be in the regression as controls
    - All regressors are measured accurately
    - We know about any endogeneity issues and can deal with them using instrumental variables
    - We know the appropriate structure of the error term
    - In this case, we need only do one regression to complete the project
    - None of these conditions is ever fully realized
        - That's why we have tests for the various regression pathologies
        - That's (one reason) why we have tests for significance of regressors

        - That's why we look at our residuals for clues
        - That's why we usually try linear and log-based models
        - That's why we have to experiment with different lag lengths
- In real research, we must deal with what Leamer calls "misspecification error" which, like sampling error, generally causes our results to be imprecise
    - Consider the regressions that you ran with the 254,654 Census observations on further fertility of mothers with two children.
        - How much sampling error is there when $N = 254{,}654$? If all of our assumptions are correct, then our estimates converge with the square root of N, so the standard errors with this sample are divided by 500!
        - What are the likely relative magnitudes of sample error and misspecification error in this exercise?
    - **Null hypotheses and maintained hypotheses**
        - In any statistical test, we make lots of assumptions
        - Some of the assumptions are "givens" such as functional form, structure of the error term, IID (or other assumed nature) of the sample, etc.
            - These are the "maintained hypotheses" that are *assumed* to be true in the test.
            - We usually assume that we have made no misspecification errors as a maintained hypothesis.
        - Some of the assumptions are tested.
            - These are the null hypothesis.
            - We are not sure that these are true; in fact, we usually expect to disprove the null hypothesis.
    - What does a hypothesis test do?
        - It measures the likelihood that such an extreme violation would occur if both the null hypothesis and the maintained hypotheses are true.
        - However, we interpret evidence against this joint set of assumptions as ***invalidating the null hypothesis, not the maintained hypothesis***.
        - In fact, what we have found is evidence that the world is not as the null and maintained hypotheses assume it is.
            - This could be due to the null hypothesis being false with the maintained hypotheses true (which is what we always assume)
            - Or it could be that the maintained hypothesis (or one part of it) is false and the null hypothesis is true (which is the essence of an invalid test: we have made incorrect assumptions underlying the test)
            - Or both could be false (an invalid test that gives the right answer)

        - By separating the assumptions into null and maintained classes, we artificially define which ones we are going to blame for any failure of the data to conform to the collective set of assumptions.
        - If we do this wrong, then we obviously can draw incorrect conclusions.
- Leamer on functional form
    - With a high-enough order polynomial, we can exactly fit the data!

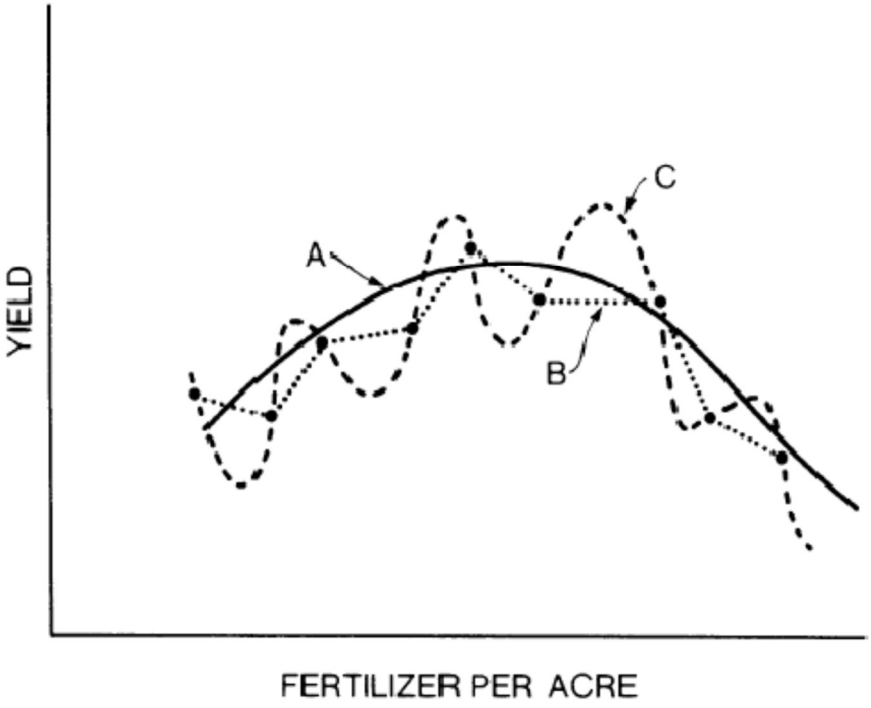


FIGURE 3. HYPOTHETICAL DATA AND THREE ESTIMATED FUNCTIONS

- So what is the right function? A, B, or C?
- We can only answer that question by applying our ***judgment***.
    - All econometric analysis is a combination of calculation and interpretation: this is unavoidable!
    - More so in economics than in "hard" sciences?
- Leamer, page 36, 37:

The false idol of objectivity has done great damage to economic science. Theoretical econometricians have interpreted scientific objectivity to mean that an economist must identify exactly the variables in the model, the functional form, and the distribution of the errors. Given these assumptions, and given a data set, the econometric method produces an objective inference from a data set, unencumbered by the subjective opinions of the researcher.

This advice could be treated as ludicrous, except that it fills all the econometric textbooks. Fortunately, it is ignored by applied econometricians. The econometric art as it is practiced at the computer terminal involves fitting many, perhaps thousands, of statistical models. One or several that the researcher finds pleasing are selected for reporting purposes. This searching for a model is often well intentioned, but there can be no doubt that such a specification search invalidates the traditional theories of inference. The concepts of unbiasedness, consistency, efficiency, maximum-likelihood estimation, in fact, all the concepts of traditional theory, utterly lose their meaning by the time an applied researcher pulls from the bramble of computer output the one thorn of a model he likes best, the one he chooses to portray as a rose. The consuming public is hardly fooled by this chicanery. The econometrician's shabby art is humorously and disparagingly labelled "data mining," "fishing," "grubbing," "number crunching." A joke evokes the Inquisition: "If you torture the data long enough, Nature will confess" (Coase). Another suggests methodological fickleness: "Econometricians, like artists, tend to fall in love with their models" (wag unknown). Or how about: "There are two things you are better off not watching in the making: sausages and econometric estimates."

This is a sad and decidedly unscientific state of affairs we find ourselves in. Hardly anyone takes data analyses seriously. Or perhaps more accurately, hardly anyone takes anyone else's data analyses seriously. Like elaborately plumed birds who have long since lost the ability to procreate but not the desire, we preen and strut and display our *t*-values.

- Leamer is a proponent of "Bayesian" econometrics, which we may study next week if there is interest.
  - In Bayesian model, one specifies a "prior" distribution for the parameter before beginning the analysis
  - Then the evidence from the data is combined with the prior to calculate a "posterior" distribution
  - Criticized because you can get nearly any posterior distribution by varying your prior.
  - Shouldn't your results reflect the evidence from the data and not your opinions?
    - Leamer's point exactly!
    - Conventionally reported results reflect your opinion as much as a Bayesian posterior, but you haven't reported how your opinion conditioned your results.
- How do we solve this problem? Page 38:

> The job of a researcher is then to report economically and informatively the mapping from assumptions into inferences. In a slogan, "The mapping is the message." The mapping does not depend on opinions (assumptions), but reporting the mapping economically and informatively does. A researcher has to decide which assumptions or which sets of alternative assumptions are worth reporting. A researcher is therefore forced either to anticipate the opinions of his consuming public, or to recommend his own opinions. It is actually a good idea to do both, and a serious defect of current practice is that it concentrates excessively on convincing one's self and, as a consequence, fails to convince the general professional audience.

Angrist & Pischke and Leamer's response

- "Credibility revolution" in econometrics?
  - Better and more data
  - "Fewer distractions"
    - Functional form and GLS methods often don't matter
  - Better research design
    - Quasi-experimental methods: IV, D-in-D, RD
    - Randomized trials: STAR and development studies
  - More transparent discussion of research design
  - "Extreme bounds" analysis
    - Has not caught on much

        - Leamer's pet: test all possible specifications and look at the ones that are least favorable to your preferred hypothesis
- Leamer: "Tantalus on the road to Asymptopia"

> We economists trudge relentlessly toward Asymptopia, where data are unlimited and estimates are consistent, where the laws of large numbers apply perfectly and where the full intricacies of the economy are completely revealed. But it's a frustrating journey, since, no matter how far we travel, Asymptopia remains infinitely far away. Worst of all, when we feel pumped up with our progress, a tectonic shift can occur, like the Panic of 2008, making it seem as though our long journey has left us disappointingly close to the State of Complete Ignorance whence we began.

    - Econometricians are still too optimistic about how much they know.
    - Robust standard errors are not a panacea because we still have inefficient estimators.
    - Sensitivity analysis is crucial: show the mapping from assumptions to conclusions!
    - Experiments may be problematic in small samples if we don't observe and control for all the confounding variables:

> Consider first the problem of the additive confounders. We have been taught that experimental randomization of the treatment eliminates the requirement to include additive controls in the equation because the correlation between the controls and the treatment is zero by design and regression estimates with or without the controls are unbiased, indeed identical. That's true in Asymptopia, but it's not true here in the Land of the Finite Sample where correlation is an ever-present fact of life and where issues of sensitivity of conclusions to assumptions can arise even with randomized treatments if the correlations between the randomized treatment and the additive confounders, by chance, are high enough.

    - "Interactive confounders" are variables that affect the effect of our variable of interest on the dependent variable (needing interaction terms). Leaving these out is problematic even in truly randomized experiments.
    - Do "data-generating processes" really exist? Are they stable? Are people rational enough to yield predictable econometric relationships?
    - Modern computers and software have made the actual computations of econometrics trivially easy, but the "thinking" part is just as hard as ever.

> Part of the problem is that we data analysts want it all automated. We want an answer at the push of a button on a keyboard. We know intellectually that thoughtless choice of an instrument can be a severe problem and that summarizing the data with the "consistent" instrumental variables estimate when the instruments are weak is an equally large error.[6] The substantial literature on estimation with weak instruments has not yet produced a serious practical competitor to the usual

instrumental variables estimator. Our keyboards now come with a highly seductive button for instrumental variables estimates. To decide how best to adjust the instrumental variables estimates for small-sample distortions requires some hard thought. To decide how much asymptotic bias afflicts our so-called consistent estimates requires some very hard thought and dozens of alternative buttons. Faced with the choice between thinking long and hard versus pushing the instrumental variables button, the single button is winning by a very large margin.

- This is not going to change! Your generation of econometricians will face ever greater temptation to "push the button" and get results without thinking about the correct underlying assumptions, then publish them if they "look nice."

Lovell's "data mining" experiment

- Stepwise regression
  - How does it work
  - Formally available in Stata
  - Informally practiced by many econometricians
- Lovell's Monte Carlo experiment
  - Create a large set of orthogonal regressors that are unrelated to the dependent variable
  - Regress dependent variable on *c* of these and choose the best 2 *t* statistics
  - Table 1 shows results
  - How often would be expect *k* independent tests to turn up *no* significant *t* statistic at the α level? $(1-\alpha)^k$
  - Lovell's rule of thumb test statistic: choosing best *k* out of *c* candidate regressors give a true significance level of $\alpha = 1-\left(1-\hat{\alpha}\right)^{c/k}$, where $\hat{\alpha}$ is the putative significance level.
  - Second experiment uses macroeconomic regressors that are correlated (and nonstationary) and shows that it is almost always possible to find regressors with "significant" *t* statistics even when the dependent variable is orthogonal. (There is some spurious regression effect here also because his variables are nonstationary.)

Publication bias

- If you look at econometric papers published in journals, most null hypotheses are rejected.
  - The papers published that accept the central null hypothesis tend to fail to reject hypotheses that are widely believed to be false.
- Are all economic hypotheses false? De Long and Lang build simple model to test:
  - Size of test = $\alpha = 0.05 = \Pr[\text{reject} \mid H_0 \text{ true}]$.

- Power of test = $q = \Pr[\text{accept} \mid H_0 \text{ false}]$
- Suppose that the true proportion of true null hypotheses is π

	Fail to reject	Reject	Total
H_0 true	0.95π	0.05π	π
H_0 false	$(1-q)(1-\pi)$	$q(1-\pi)$	$1-\pi$
Total	$(1-q)+(q-0.05)\pi$	$q+(0.05-q)\pi$	

- Let a be a test statistic and let $f(a)$ be its marginal significance level (p value)
- Under the null hypothesis,
  $f(a) \sim U[0,1]$
  $\Pr[f(a) \geq p] = 1-p$
- Under alternative hypothesis, $f(a)$ follows some unknown distribution G so that $\Pr[f(a) \geq p] = 1-G(p)$. We assume $1-G(p) \leq 1-p$.
- Share of test statistics that have p value less than or equal to p (= share of rejected nulls at significance level p) should be

$$\begin{aligned}\Pr[f(a) \geq p] &= \pi(1-p) + (1-\pi)[1-G(p)] \\ &= \pi[(1-p) - (1-G(p))] + (1-G(p))\end{aligned}$$

$$\begin{aligned}\pi &= \frac{\Pr[f(a) \geq p] - [1-G(p)]}{(1-p) - [1-G(p)]} \\ &\leq \frac{\Pr[f(a) \geq p]}{(1-p)}.\end{aligned}$$

- This gives an upper bound for π.
- For example, if $\pi = 1/2$, then at least $\Pr[f(a) \geq 0.80] \geq 0.2 \times 0.50 = 0.10$: at least 10% of actual p values should be in the range (0.80, 1.00).

TABLE 1

DISTRIBUTION OF REPORTED MARGINAL SIGNIFICANCE LEVELS

Marginal Significance Levels	Number of Hypothesis Tests	Estimated Upper Bound on True Nulls/ Unrejected Nulls*
1.0–.9	0	0%
.9–.8	4	23
.8–.7	7	42
.7–.6	7	52
.6–.5	6	54
.5–.4	11	66
.4–.3	11	75
.3–.2	14	86
.2–.1	18	100

* Estimated by the ratio of the number of hypotheses with marginal significance levels in this category or higher to the number of hypothesis tests that should fall in this category or higher, if all null hypotheses or all unrejected null hypotheses were true.

We focus attention on the tighter bounds obtained for values of p^* fixed near one at .9 and .8. The conclusions are striking. In our sample, there are no values of $f(a)$ greater than .9. Among unrejected hypotheses, one-ninth of $f(a)$ values should fall into the range .9–1.0 when the null hypothesis H_0 is true. Our point estimate of the number of unrejected nulls that are in fact true is nine times the number of marginal significance levels that fall between .9 and 1.0. *The implied estimate of π is therefore zero:* no null hypotheses are true.

A less extreme estimate comes from examining the fraction of unrejected null hypotheses with $f(a) > .8$. Two-ninths of unrejected nulls should fall into this category when the null hypothesis is true; we actually find that only four out of the 78 unrejected nulls (and the 276 nulls tested) do so. This produces a point estimate that 23 percent of *unrejected* null hypotheses are true.

- These are point estimates. They can reject the null hypothesis that $\pi = 1/3$ against the alternative that it is $< 1/3$. Thus, they are quite confident that from the evidence of the literature, $\pi < 1/3$.
- Why? They think probably publication bias.

Simulation, Monte Carlo, and bootstrap methods

- In a few special cases, with appropriate assumptions, we know that actual distributions of the test statistics that we use. In some additional cases, we can approximate the asymptotic distributions of these statistics.
    - However, most of the time these assumption are probably dubious.

    - How much of a problem is this? Are the actual expected values and (especially) standard errors of the distributions that different?
    - We can use simulation to determine the properties of our standard test statistics *under the null hypothesis* when the assumptions we usually make fail to hold.
- Simulation methods
    - Monte Carlo analysis is the simulation of the behavior of estimators under controlled conditions that may deviate from the standard assumptions under which it is used.
    - Bootstrap methods apply simulation to a specific sample of data, re-running a regression many times with either parametric or non-parametric error terms to estimate the standard deviation of the test statistic under H_0 (rather than using the conventional standard error as an estimate).
- Generating data for simulations
    - Can use actual variables (as Lovell did in his second data-mining experiment with macro variables) or can generate them "randomly"
    - Error terms are always generated randomly.
    - Random-number generators
        - No computer-generated sequence of numbers is truly random.
        - The way these generators work is to begin with a "seed," then generate new numbers in the sequence based on calculations such as remainders of division by large prime numbers.
            - Same seed implies same sequence of numbers, so if you want to control the process (especially during debugging) you can get the same sequence again.
            - Default seed is usually taken from the seconds of the computer clock or something like that: will not be the same on repeated execution.
            - In Stata: runiform (or uniform) draws a random number from (0, 1). rnormal(mean, std) draws from the normal distribution with mean and standard deviation given.
            - Can generate normal variate as invnorm(runiform())
    - We generate random set of e^* and use them to compute y^* under the null hypothesis about β given the values of x, which may be set to sample values, generated randomly, or something else.
- Implementing Monte Carlo
    - Create repeated samples (how many? 1000? 10000? 100000?) of e^* and y^*.
    - For each sample, calculate the test statistic of interest: $\hat{\beta}$, $\text{se}(\hat{\beta})$, $t(\hat{\beta})$ or anything else.
    - Accumulate the estimates in a new data set.

    - Examine the properties of the estimates:
        - Mean to assess bias
        - Standard deviation to compare to estimated standard error
        - Quantiles to assess critical values or estimate *p* values for your estimates
- Bootstrap standard errors
    - If the assumptions of OLS are not valid for your sample, you can estimate the standard errors of your OLS estimates by using a bootstrap technique
    - Use your actual *x* variables, sample size, etc.
    - Generate a sample of *e** error terms
        - Can use a normal distribution based on the SEE as estimate of standard deviation
        - Can use "re-sampling," assigning random $\hat{u}_i$ values to observation *j*.
    - Calculate sample of *y** values.
    - Run regression of *y** on actual *x* values
    - Save estimated coefficients $\hat{\beta}_k$ for the *k*th replication
    - Repeat *K* times with different randomly generated error terms
    - Examine the distribution of $\hat{\beta}$ by calculating the standard deviation: this is the bootstrap standard error.
    - Look at the 2.5th and 97.5th percentiles of the distribution: these are the critical values for a two-tailed 5% hypothesis test.
- Monte Carlo demonstration of Granger-Newbold spurious regression result
    - Do-file spurious.do:
        - program spurious
        -
        - drop _all
        - set obs 100
        - g id=_n
        - tsset id
        -
        - * Generate y variable as random walk
        -
        - g e=rnormal(0,1)
        - g y=e if id==1
        - replace y=l.y+e if id>1
        -
        - * Generate x variable as independent random walk
        -
        - g a=rnormal(0,1)
        - g x=a if id==1

        - replace x=l.x+a if id>1
        -
        - * Run regression of y on x
        -
        - reg y x
        -
        - end
        -
    - Show single replication
    - What can be retrieved? ereturn list shows available results
    - Command to invoke simulation:
        - simulate b=_b[x] se=_se[x] r2=e(r2) , reps(1000) : spurious
    - Creates data set with 1000 observations with variables b, se, r2
    - Can now use summarize, centile, and histogram to look at behavior of estimates.
    - Contrast dspurious with spurious to see effect of regression on integrated variables.

Methods for coping with missing data

- Missing data problems are very common in econometrics.
    - In surveys, some people omit questions or have undecipherable responses.
    - In longitudinal surveys, attrition usually occurs
    - In databases compiled for other purposes, they often don't care if some variables are missing: Reed database missing class ranks, high-school grades, etc.
    - In macro data, sometimes there is a change in how the series is defined
        - Not exactly missing data, but how do you "splice" the two series?
        - Best way to splice is to regress overlapping observations and use fitted values for shorter series.
        - If not enough overlapping observations to run regression, then can use pivot observation to join series
            - Example: price index changing base years. If value in 2002 is 125 in 1990 dollars and 95 in 2005 dollars, then you can multiply all of the 1990-dollar observations by 95/125 to convert to 2005 dollars.
- Key question that informs missing-data problem: Why are the data missing?
    - Missing completely at random (MCAR): Probability that the observation/variable is missing is unrelated to any variable in the analysis.
    - Missing at random (MAR): Probability that the observation/variable is missing is unrelated to the missing variable, but may be related to other, observed variables.

- Not missing at random (NMAR): Probability that the observation/variable is missing depends on the true value of that variable.

- Methods of dealing with missing data
  - Complete-case analysis
    - This is the default: Stata will simply delete any observations for which one or more variables in the model are missing.
    - We lose information by doing this.
    - Example: Suppose that we are missing one observation on x out of ten and that the coefficients based on the other nine observations are $y = 5 + 10x$. The missing observation has a y value of 25. By omitting this observation, we are implicitly assuming that the x value is 2, so that it will not have a residual and not add to the regression. If the univariate distribution of x in the rest of the sample is such that a value of 2 seems highly unlikely, then we are almost surely missing important information about the relationship by ignoring this observation.
    - Complete-case analysis does not lead to bias if missingness does not depend on y. (This is standard sample-selection problem that we have dealt with before.)
  - Available-case analysis
    - Regression coefficients and standard errors depend only on the sample variances and covariances of the variables.
    - Even if y is missing for an observation, if x_1 and x_2 are available, we can use that observation to contribute to the estimate of the variances of the x variables and to their covariance.
    - This seems to use additional information, but has other problems and it rarely used.
      - Because it uses different groups of observations, there is no guarantee that $X'X$ has an inverse, so it may even be impossible to calculate OLS estimate.
  - Dummy-variable methods
    - $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$
    - X_1 is complete; X_2 has some missing data.
    - Let $M_i = \begin{cases} 1 \text{ if } X_2 \text{ is missing,} \\ 0 \text{ otherwise.} \end{cases}$
    - Let $X_{2i}^0 = \begin{cases} X_{2i} \text{ if } M_i = 0, \\ 0 \text{ if } M_i = 1. \end{cases}$
    - $Y_{i0} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}^0 + \gamma M_i + u_i$ is biased for β_1.
      - β_1 picks up the effect of 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, but is difficult to implement unless pattern of missingness is "block-style."

- Imputation methods
  - If there is an irregular pattern in which several variables have missing observations scattered through the sample (and the same observations do not tend to be missing for all variables), then we have some information about the observations for which a particular variable is missing based on the observed values of other variables.
  - Imputation methods use the values of the other variables (and the pattern of covariance between the observed and missing variables for the part of the sample for which both are observed) to impute estimates of the missing values.
  - Unconditional imputation replaces missing values by the means of the variables.
    - This leads to bias in the coefficients because the other variables that are correlated with the missing one have to carry "extra weight" in predicting *y* for those observations in which the missing *X* is set to its mean.
  - Conditional imputation based on other *X* variables
    - Use complete cases to estimate $X_{2i} = \delta_0 + \delta_1 X_{1i} + v_i$.
      - Could use LDV model if appropriate.
    - Calculate single imputed values for missing observations as $\tilde{X}_{2i} = \hat{\delta}_0 + \hat{\delta}_1 \tilde{X}_{1i}$.
    - Use full sample to estimate $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 \tilde{X}_{2i} + u_i$.
    - This procedure is consistent if data are MCAR.
    - Standard errors are problematic because we don't take account of the imputed nature of the data and the error in measurement that results.
  - Conditional imputation based on other *X* variables and *y*
    - Can include *y* in the imputation regression.
    - Improves quality of imputation if missing *X* is highly correlated with *y*.
    - Leads to bias in OLS regressions of filled-in model.
- Multiple imputation with combined equations (MICE)
  - Instead of replacing missing observation with single conditional expectation based on imputation regression, we construct multiple samples with stochastic imputations: Expected value of missing *X* plus random draw from the error term of the imputation regression, which includes both *X* variables and *y*.

- Use complete cases to estimate $X_{2i} = \delta_0 + \delta_1 X_{1i} + \delta_2 Y_i + \nu_i$.
    - Note that we can (and must) include *y* here when we are using random draws from the distribution rather than expected values.
    - Can use LDV methods if the missing variable is a dummy, ordered, censored, etc.
- Calculate *m* random imputed samples using $\tilde{X}_{2ij} = \hat{\delta}_0 + \hat{\delta}_1 X_{1i} + \hat{\delta}_2 Y_i + \nu_{ij}$, where ν_{ij} is a random draw from the estimated distribution of *v*, usually normal with zero mean and variance equal to the estimated variance of *v* based on residuals.
- For each sample *j*, run the regression using imputed values: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 \tilde{X}_{2i} + u_i$ and get the estimates $\hat{\beta}_{ij}$ and the squared standard errors $\widehat{\text{var}}\left(\hat{\beta}_{ij}\right)$.
- Combine the results of the *m* regressions as follows:
    - $\hat{\beta}_i = \frac{1}{m}\sum_{j=1}^{m} \hat{\beta}_{ij}$
    - $\widehat{\text{var}}\left(\hat{\beta}_i\right) = \frac{1}{m}\sum_{j=1}^{m} \widehat{\text{var}}\left(\hat{\beta}_{ij}\right) + \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{\beta}_{ij} - \hat{\beta}_i\right)^2$
    - The parameter estimate is just the mean of the estimates for the *m* imputed samples.
    - The variance is the mean of the estimated variances in the *m* samples, plus the estimated variance of the parameter estimate across the samples.
        - This last term corrects the standard error for the imputation process, adding variance to account for the fact that the *m* imputations do not all lead to the same answer.
        - Because a highly uncertain imputation process is likely to lead to wide variation in $\hat{\beta}_{ij}$ across samples, this correction to the variance will be high when the imputation process is imprecise.
- Stata 11+ has an implementation of MI models with a "dashboard" to control imputation regressions (which can be OLS, probit, tobit, ordered probit, etc.) and the combined regression using the multiple imputations.
- MICE works with MCAR of MAR data.

- Can also use ML methods to estimate missing-data models (not going to talk about)

Models with varying parameters

- We have talked a lot about Assumption #0: The same model applies to all observations.
  - What if this is false and the model changes from one set of observations (either over time or cross-sectionally) to another?
  - We can model this by allowing some parameters of the model to vary across observations.
  - We have considerable experience with simple, deterministic forms of varying parameters:
    - Dummy variables allow the constant term to differ for the set of observations for which the dummy is turned on.
    - Interaction terms allow the effect of one variable to depend on the magnitude of another (where one or both may be dummies).
    - Splitting samples at recognized breakpoints is another strategy.
  - We now consider models in which the variation in the parameters is at least partially random, especially over time.
- Stationary random parameter models
  - $Y_t = \beta_t X_t + u_t,$
    $\beta_t = \alpha + \delta Z_t + v_t.$
  - Substituting yields $Y_t = \alpha X_t + \delta X_t Z_t + w_t,$
    $w_t \equiv u_t + v_t X_t.$
  - Our usual assumptions are that *u* and *v* are classical error terms that are uncorrelated with one another. In that case, $\text{var}(w_t) = \sigma_u^2 + X_t^2 \sigma_v^2$ and *w* is not serially correlated unless *u* or *v* is.
  - This model is heteroskedastic with variance a proportional to $1 + \lambda x^2$, where $\lambda = \frac{\sigma_v^2}{\sigma_u^2}.$
  - How to estimate?
    - Could use OLS with robust standard errors (did not exist when Maddala wrote his book).
    - Maddala suggests ML with $\ln L = K - \frac{n}{2}\ln\sigma_u^2 - \frac{1}{2}\sum_{i=1}^{n}\ln\left(1+\lambda X_t^2\right) - \frac{1}{2\sigma_u^2}\sum_{i=1}^{n}\frac{\left(Y_t - \alpha X_t - \delta X_t Z_t\right)^2}{1+\lambda X_t^2}$, with *K* an irrelevant constant.
    - Can do this with a two-step procedure:

- For given λ, the α and δ that maximize L are the WLS estimators calculated by applying OLS to
$$\frac{Y_t}{\sqrt{1+\lambda X_t^2}} = \alpha \frac{X_t}{\sqrt{1+\lambda X_t^2}} + \delta \frac{Z_t X_t}{\sqrt{1+\lambda X_t^2}}.$$
- Search over λ to find the value that yields the highest L with $\alpha(\lambda)$ and $\delta(\lambda)$ calculated by WLS/OLS.

- Switching regressions: two (or more) regimes with different parameters
  - We considered the simple case of this with the Quandt likelihood-ratio (QLR) test when we talked about nonstationarity due to breaks in S&W's Chapter 14.
    - The QLR test statistic is the maximum of the Chow-test F statistic considered over possible breakpoints within the middle 70% (or so) of the sample.
    - S&W's Table 14.6 gives the critical values for the QLR test statistic, which does not follow a standard parametric distribution.
  - More interesting case is where model can switch back and forth depending on values of other variables.
    - Example: is economic response to oil-price increases different than oil-price increases? One set of parameters when ΔP_O is positive and a different set when it is negative.
      - This is simple case because there are no unknown parameters in the switching rule.
      - More interesting case is where the switching rule involves unknown parameters.
    - Suppose that the parameters are in regime 1 ($Y_t = \alpha_1 + \beta_1 X_t + u_t$) when $\lambda \equiv \pi_1 Z_1 + \ldots + \pi_k Z_k < c$ and in regime 2 ($Y_t = \alpha_2 + \beta_2 X_t + u_t$) when $\lambda > c$.
    - Error term may also differ between regimes.
    - Can estimate by ML, which is kind of like a regression (to determine the α and β parameters) combined with a probit (to determine which regime governs each observation)
  - Another model of interest is the single-breakpoint model constraining the function to be continuous over time.
    - Example: fitting a trend line to the log of a variable and allowing the trend growth rate to change at some date without allowing the function to jump at that date.
    - Let n_0 be the breakpoint in the sample, so that $Y_t = \alpha_1 + \beta_1 X_t + u_t$ for $1 \le t \le n_0$; $Y_t = \alpha_2 + \beta_2 X_t + u_t$ for $n_0 < t \le N$.

        - Both regression lines must go through the point n_0, so we must impose the restriction $\alpha_1 + \beta_1 X_{n_0} = \alpha_2 + \beta_2 X_{n_0}$ on the estimation. This is a simple linear restriction that can be imposed in OLS by the usual means.
- Adaptive regression: constant term is a random walk.
    - This model was developed before the theory of integrated processes was well understood.
    - The model that they propose has issues with an integrated error term (and dependent variable) that are better handled with differencing and (sometimes) cointegration methods.
    - Can look at the more interesting model where slope is a random walk as well.
        - Cannot estimate all t parameters β_t.
        - Can estimate one of them: suggestion is to estimate the last one (or one after last)
        - For varying constant term: $\alpha_t = \alpha_{t-1} + \nu_t$. Let

$$\begin{aligned} Y_t &= \alpha_T + (\alpha_t - \alpha_T) + \beta X_t + u_t \\ &= \alpha_T + \beta X_t + u_t - \sum_{i=t+1}^{T} \nu_i . \end{aligned}$$

        - When we write the model in terms of α_T, observation $T-1$ has additional variance relative to T because of change in α from $T-1$ to T. Observation $T-2$ has yet more variance because the α is two changes away from α_T. Thus, we end up with a WLS estimator that weights the most recent observations most heavily and earlier observations less.
        - There will also be correlation between the composite error terms because of the accumulation of the parameter changes.
    - This is an intuitively attractive idea for regressions that you are using for forecasting but don't know if the parameters are stable over time.
        - Most recent observations are the most relevant for the forecast, so we weight them the most heavily.
        - Observations in the distant past are not totally irrelevant, but are less important so we include them with lower weights.
- Another class of models is panel-data models in which each cross-sectional unit has a different parameter value:
    - If the varying parameter is the intercept term and the variation is deterministic, then this is the fixed-effects model.
    - If the varying parameter is the intercept term and the variation is random, this is the random-effects model.
    - If the varying parameter is a slope coefficient and the variation is deterministic, then this is a variation on fixed effects where the unit dummies interact with the variable whose coefficient is changing.

        - We lose a lot of degrees of freedom in this model. In the limiting case of all coefficients varying deterministically across units, we are just doing separate time-series regressions for each unit.
    - If the varying parameter is a slope coefficient and variation is random, then we have a variant of the random effects model in which the variance of the "unit-specific error component" for each unit depends on the values of x for that unit.
- When to use varying-parameter models?
    - Can almost always justify it.
    - What do we really gain from modeling the variation in the coefficients rather than putting in the error term?
        - If variation is systematic, then we have a better understanding of how the effect of x on y depends on Z. This is the essence of interaction terms and we know that they can be very useful.
        - If variation is random, then we may not gain too much, although adaptive regression model is appealing and if there are large variations in x, then we might want to take it into account if the coefficient of x varies randomly.

Duration and hazard-rate models

- We have encountered duration problems before: when we considered the censored distribution of unemployment spells in a sample where some are ongoing, for example.
    - In these models, the focus was on what other variables determine the length (duration) of the spell.
- The formal analysis of **hazard (or survival) models** focuses not only on the effects of other variables, but on modeling the probability that a spell will end as a function of its current length.
    - Does it become more or less likely that something will happen when it has not happened for a long time? Earthquakes, divorce, end of a strike, success in job search, survival after events are examples.
- In hazard analysis, we think of a duration event as a sequence of opportunities to end, with a certain probability (hazard) of ending at each time that may depend on other variables and on the current duration of the event.
- Let T be the "spell length" variable with density $f(t)$. (Normal distribution not appropriate because T must be non-negative.)
    - $F(t) = \int_0^t f(s)ds = \Pr[T \le t]$ is the probability that the spell is no longer than t.
    - $S(t) = 1 - F(t) = \Pr[T \ge t]$ is the survival function: the probability that a spell is at least length t.

- The **hazard rate** is defined as the probability that the spell ends now conditional on the fact that it has lasted this long:
  - $\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr[t \le T \le T + \Delta t \mid T \ge t]}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t\, S(t)} = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}.$
    - Note similarity to inverse Mills ratio
  - $\lambda(t) = -\frac{d \ln S(t)}{dt}$ because $f(t)$ is $-S'(t)$
- The **integrated hazard function** is $\Lambda(t) = \int_0^t \lambda(s)\, ds.$
  - $S(t) = e^{-\Lambda(t)}$
  - $\Lambda(t) = -\ln S(t)$
- All of these functions can (obviously) be derived from one another, so f, F, S, λ, and Λ are all equivalent ways to characterize the hazard behavior of the model as a function of current duration t.
- Modeling the hazard rate:
  - Constant hazard rate
    - $\lambda(t) = \lambda,$
    - $\ln S(t) = k - \lambda t,$
      $S(t) = Ke^{-\lambda t} = e^{-\lambda t}$ because $S(0) = 1$
    - With constant hazard rate, $E(t) = 1/\lambda$, so MLE of λ is $1/\bar{t}$
  - Positive or negative **duration dependence**
    - Greene's T25.8 and F25.2 show several common choices for non-constant λ functions

TABLE 25.8 Survival Distributions

Distribution	*Hazard Function, $\lambda(t)$*	*Survival Function, $S(t)$*
Exponential	λ,	$S(t) = e^{-\lambda t}$
Weibull	$\lambda p(\lambda t)^{p-1}$,	$S(t) = e^{-(\lambda t)^p}$
Lognormal	$f(t) = (p/t)\phi[p \ln(\lambda t)]$	$S(t) = \Phi[-p \ln(\lambda t)]$
	[$\ln t$ is normally distributed with mean $-\ln \lambda$ and standard deviation $1/p$.]	
Loglogistic	$\lambda(t) = \lambda p(\lambda t)^{p-1}/[1 + (\lambda t)^p]$,	$S(t) = 1/[1 + (\lambda t)^p]$
	[$\ln t$ has a logistic distribution with mean $-\ln \lambda$ and variance $\pi^2/(3p^2)$.]	

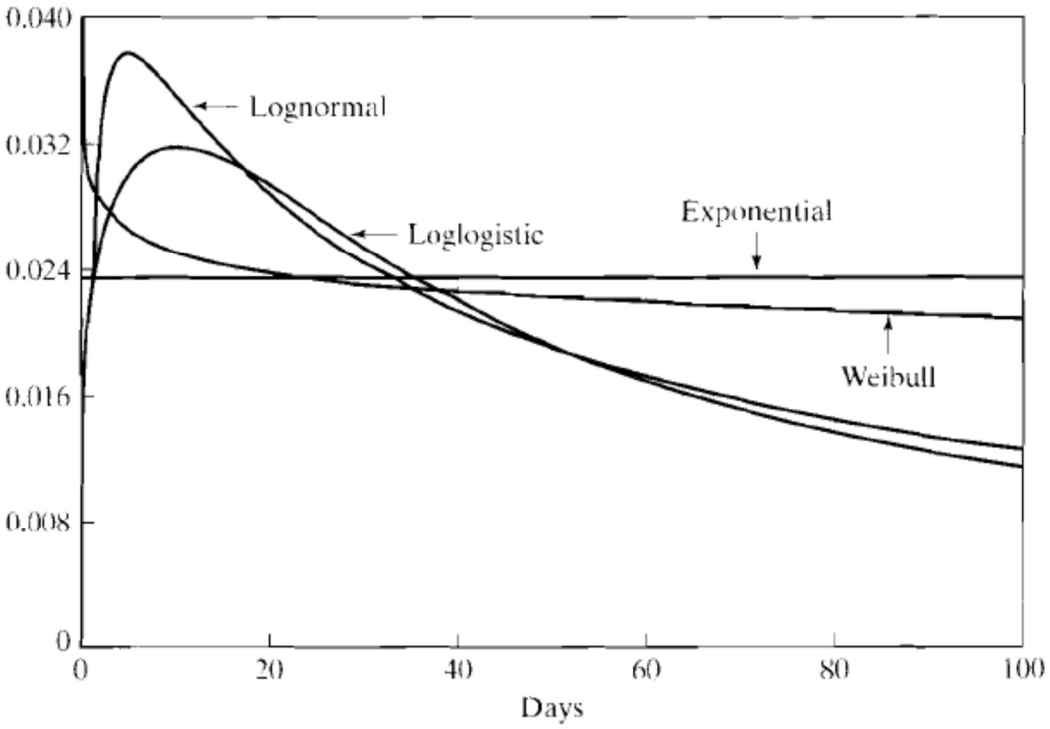


FIGURE 25.2 Parametric Hazard Functions.

- Weibull is a common one because depending on the parameter *p* it can be increasing or decreasing with *t*.

- Estimation of survival models
  - We estimate these models by ML:
  - $$\ln L = \sum_{\substack{uncensored \\ observations}} \ln f(t \mid \theta) + \sum_{\substack{censored \\ observations}} \ln S(t \mid \theta)$$
  - $$= \sum_{\substack{uncensored \\ observations}} \ln \lambda(t \mid \theta) + \sum_{\substack{all \\ observations}} \ln S(t \mid \theta)$$

- Including exogenous variables
  - We usually want other variables to condition the survival/hazard functions
  - One common model: $\lambda_i = e^{-X_i\beta}$ replacing constant λ in Weibull function (or exponential function)
  - Note that *x* must be constant over the spell (such as personal characteristics) or model become more complex. (You would need to know *x* through entire spell in order to model different hazards at different moments during spell.)
- Nonparametric models
  - What do we mean by nonparametric?
    - No assumption of a specific functional form or probability distribution

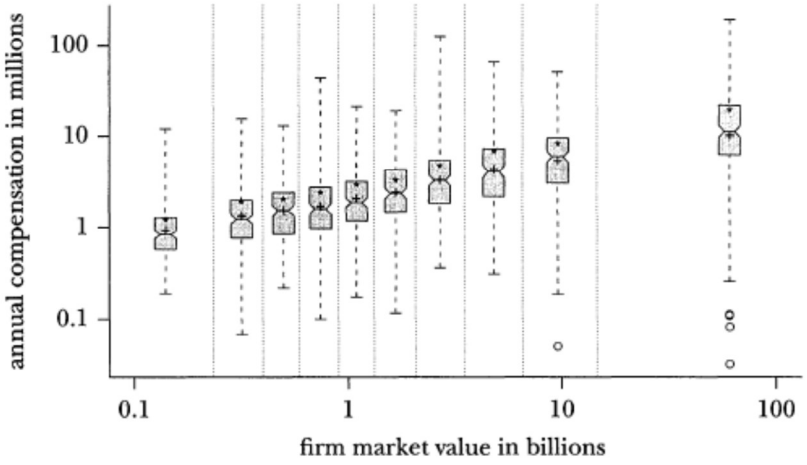
- In case of hazard models, we use the analog of a frequency distribution:
  - What share of spells that lasted two weeks ended in the third week?
  - What share of spells that lasted three weeks ended in the fourth week?
  - Etc.
- Plot these as a function of duration to get empirical hazard function
- Advantages: no distributional assumption, can model unusual shapes
- Disadvantages: does not invoke smoothness assumptions that may be appropriate, difficult to model effects of other variables

Quantile regression

- We get so used to the basic idea of traditional regression analysis that we sometimes forget important details about what we are doing.
  - Standard regression estimates the conditional mean of y as a function of x.
  - What about other properties of the conditional distribution of y?
  - We sometimes talk about the estimated conditional standard deviation (SEE), but rarely about any other attributes of the distribution.
    - If y follows a normal distribution, then we can calculate the whole distribution from the mean and standard deviation.
    - If y is not normal, then we generally don't know all the details of the distribution.
  - There may be much more useful information embodied in the conditional distribution than just the mean.
  - Consider Figure 1 from Koenker & Hallock:
    - Provides: quartiles, range, median, arithmetic and geometric means of CEO compensation for each decile of firm size.
    - What would regression give us?
      - Equivalent of a line connecting the means (either arithmetic or geometric if we used a log function)
    - This is an example of the kind of expanded view of the conditional distribution that we can get from quantile regression, which looks at how the quantiles of the distribution of the dependent variable depend on the regressor.

Figure 1

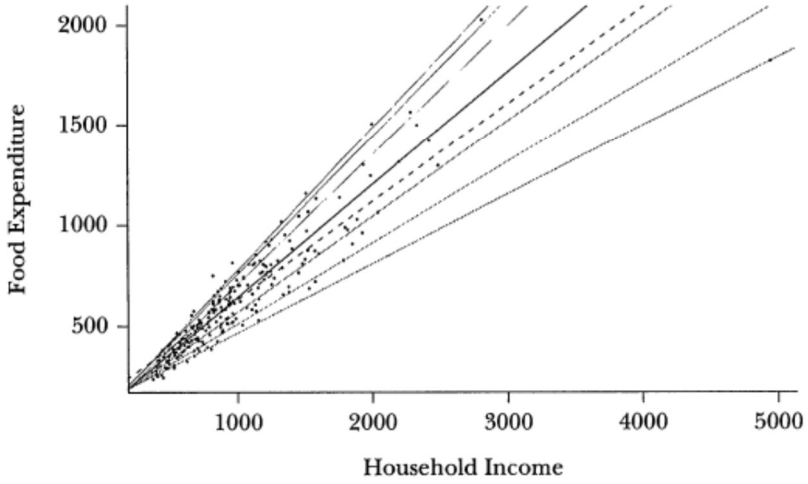
Pay of Chief Executive Officers by Firm Size



- Moments and quantiles as minimization problems:
  - The unconditional mean is the value of μ that minimizes $\sum_{i=1}^{n}(y_i-\mu)^2$
  - Unconditional median is the value of m that minimizes $\sum_{i=1}^{n}|y_i-m|$
  - Unconditional τth quantile is the value of ξ that minimizes $\sum_{i=1}^{n}\rho_\tau(y_i-\xi)$, where ρ_τ is the "tilted absolute value" function $\rho_\tau(x)=\begin{cases}\tau x & \text{if } x\geq 0,\\ -(\tau-1)x & \text{if } x<0.\end{cases}$
- Generalizing to the condition regression situation:
  - In standard parametric regression, we let μ depend on x
  - In quantile regression, we let the τth quantile be a function of x:

    $\min\sum_{i=1}^{n}\rho_\tau(y_i-\xi(x_i,\beta))$, which for the linear case is $\min\sum_{i=1}^{n}\rho_\tau(y_i-x_i\beta)$.
  - Because the ρ function is non-differentiable, we can't use basic calculus methods to solve this, but we can use methods developed for linear programming models to find minimum pretty efficiently.
- Sample output of quantile regression model
  - There will be a separate regression for each quantile that we are interested in.
  - Figure 3 from Koenker & Hallock:

Figure 3
Engel Curves for Food



Notes: This figure plots data taken from Engel's (1857) study of the dependence of households' food expenditure on household income. Seven estimated quantile regression lines for different values of τ {0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95} are superimposed on the scatterplot. The median $\tau = 0.5$ is indicated by the darker solid line; the least squares estimate of the conditional mean function is indicated by the dashed line.

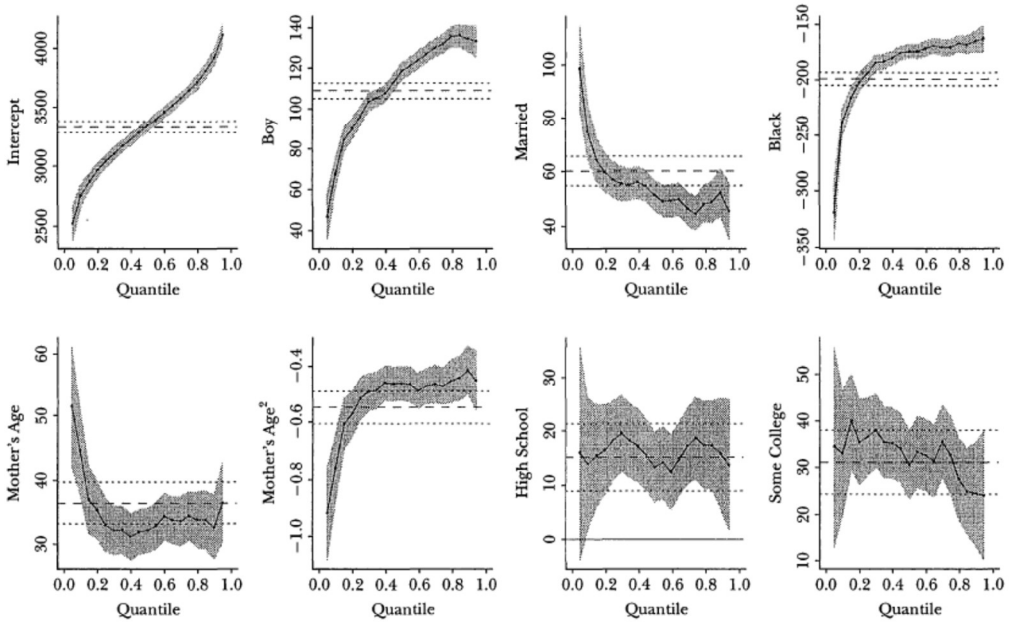
    - Food expenditure as a function of income
    - OLS regression gives us dashed line
    - Condition median of distribution of food expenditure as linear function of income is bold line.
    - Other lines are 0.05, 0.1, 0.25, 0.75, 0.9, 0.95 quantiles of distribution as linear functions of income.
  - Under standard OLS regression, the distribution of food expenditure conditional on income would be assumed to be normal with mean given by the dashed line and constant variance given by SEE^2. (Regressing in log terms would allow variance to be proportional to x.)

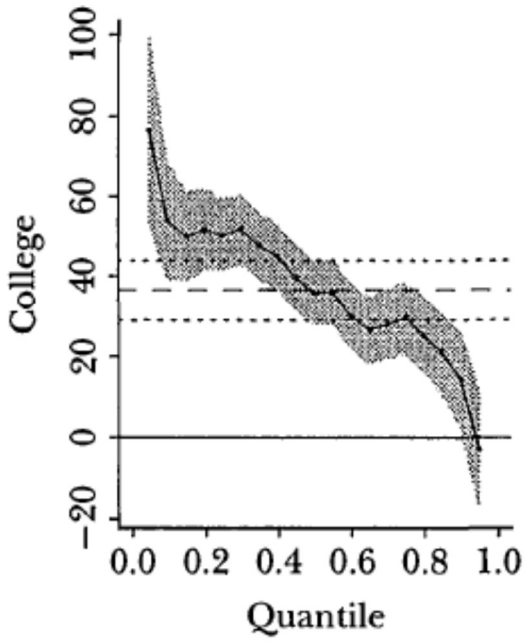
- A multivariate example: Figure 4 shows baby birth weight as a function of mother's variables:
  - Note that each variable can have distinct pattern of effects on different quantiles of the distribution.
  - For example, boy babies tend to be larger, but especially at the top end of the distribution. That suggests that the difference is driven more by really big boys than by really little girls.
  - Can't get that nuance out of an OLS regression.
  - High-school graduation has across-the-board effect on all parts of the distribution.

- Note effect of college graduates: Much less likely to have a very small baby (strong effect at low quantiles) but not much more likely to have a very large baby (little effect at upper quantiles)

Figure 4

Ordinary Least Squares and Quantile Regression Estimates for Birthweight Model





- In Stata: qreg dvar indvars , quantile(0.5) will do 0.5 quantile.

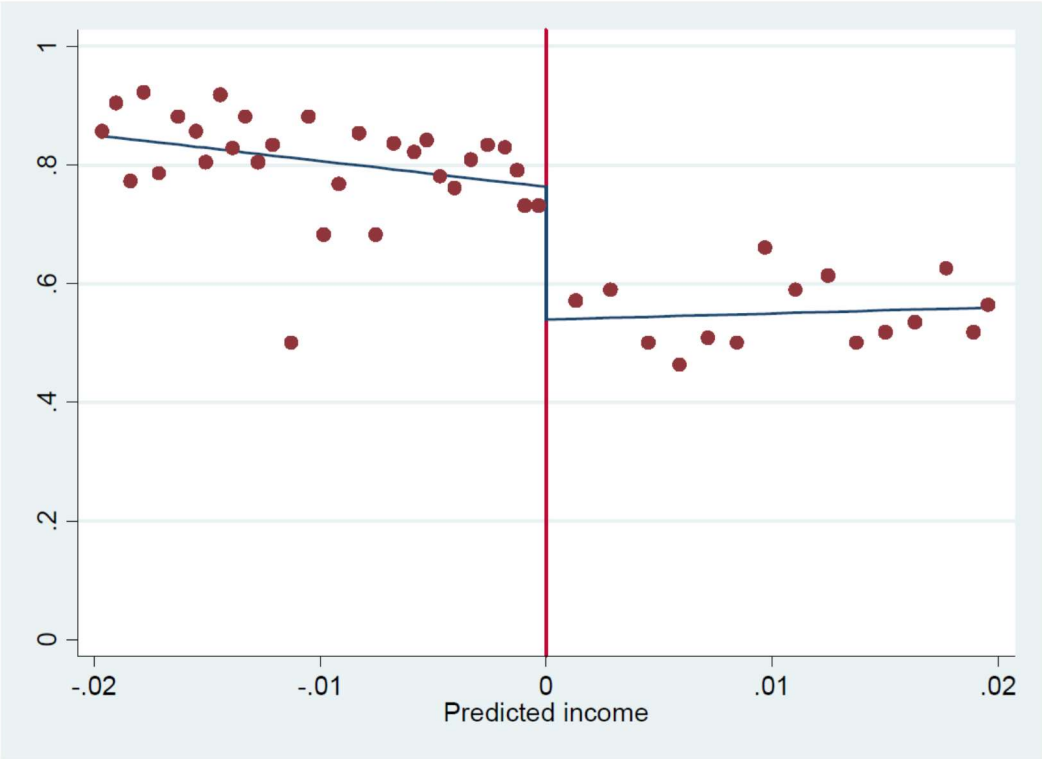
- Example: Reed GPA as dependent variable
    - Reed qreg.dta
    - Show **reg uggpa irdr satm100 satv100 hsgpa female if humfresh**
    - Show qreg with quantile(0.5) for MAD regression estimator
    - Ask: Which of these variables would you expect to impact the top end of the grade distribution more or less than the bottom?
    - Show reedqregs.doc for results of various quantiles
    - Show Reed qregs.xlsx for diagram (old version) of effects of irdr on quantiles of grades

Regression discontinuity models

- Identification is always difficult: endogeneity is always a threat and instruments are rare
    - Randomized experiments are probably the best way to avoid endogeneity, but are not always feasible
    - Sometimes we can find "natural experiments" that allow us to effectively control for the things we cannot measure
        - Example: we can't control people's genetic structure (yet) but we can examine identical twins
- The case that is examined in van der Klaauw (2002): Effect of increased financial aid on probability of enrollment.
    - Intuition says that more aid (or lower cost, in general) → higher likelihood of enrollment, other things held constant
    - If we could measure *all* of the factors that go into the enrollment decision, we could estimate this directly
        - This would include the complete set of colleges to which the student was admitted and the amount of aid/cost at each
        - It would also include all of the relevant characteristics of the student, both "objective" (test scores, high-school grades), "subjective" (essays, interviews, recommendations), and "preferential" (student's preferences about location and characteristics of school, experience during campus visit, etc.)
        - Obviously, we can never measure all of these, so they go into the error term. If they are correlated with the amount of aid, then the estimated effect of aid will be biased.
        - Why would these be correlated with aid?
            - Unmeasured factors would increase aid at College X, but probably also increase aid elsewhere
            - Those who are offered high aid packages at X may be *less* likely to come unless we control for the unobservable factors that affect

aid at X (in the equation) and aid elsewhere (in the error term), which lead to correlation between aid and the error
        - Effect of aid is likely biased downward because of this
    - Could we randomize? Would any school be willing to increase aid for a random selection of students to see which ones come? (Perhaps not)
- The idea of RD is that we sometimes have arbitrary, discrete thresholds where people who are nearly identical but on opposite sides of the threshold are treated differently.
    - This allows us to estimate a "treatment effect" by comparing the nearly identical people on the two sides of the line; we can think of these as natural experiments
    - Examples:
        - Cutoff birthdays for school attendance: September 2 babies are a year older when they start school than August 31 babies: does this affect their outcomes?
        - Laws sometimes have arbitrary cutoff points: If unemployment insurance lasts 26 weeks, is the likelihood of an unemployed worker taking a job higher in the 27th week than the 26th week?
        - van der Klaauw: College X has arbitrary thresholds for awarding discrete levels of aid: are students just above the threshold (who get more aid) more likely to attend than nearly identical students just below the threshold?
        - Manacorda et al.: Welfare program in Uruguay that was designed by economists to be based on "predicted" income to avoid mis-reporting and fluctuations in annual income. Households immediately on both sides of the line are very similar, but one gets transfer and other doesn't. Are household getting transfer more likely to support the government?
- Manacorda result:

Figure 2: Program eligibility and political support for the government



- Classic RD design is illustrated by the van de Klaauw paper's Figure 1:

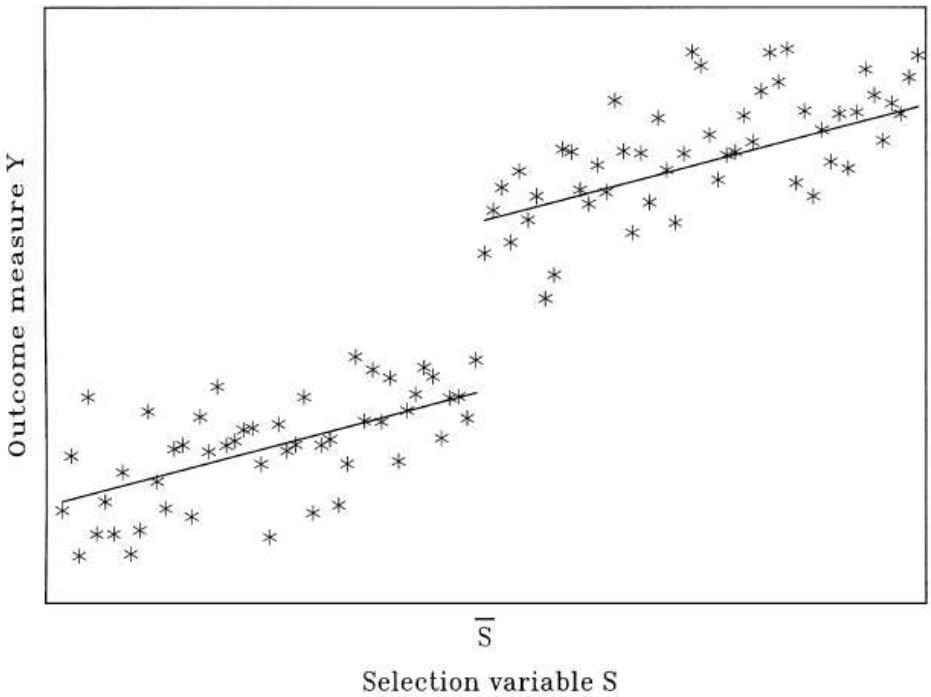


FIGURE 1

REGRESSION–DISCONTINUITY DATA

- The treatment here depends sharply on the value of the fully observable selection variable *S*: people above $\overline{S}$ are in the treatment group and people below in the control group.
- The gap at the selection value $\overline{S}$ is the effect of crossing the threshold, which could be an unbiased measure of the effect of the treatment variable.
- Econometrically, we estimate the two relationships on both sides (which may or may not have the same slope) and then estimate the treatment effect as $\alpha = \lim_{S \downarrow \overline{S}} E[y \mid S] - \lim_{S \uparrow \overline{S}} E[y \mid S]$
- We can estimate this in the simplest case as van der Klaauw's equation (8): $y_i = \beta + \alpha T_i + k(S_i) + \omega_i$, where T is a treatment dummy and $k(S)$ is the general relationship between y and S ignoring the treatment (which could be a linear or nonlinear function, shown as linear here).
- If the relationship between T and S is not sharp, then there may be some people close to $\overline{S}$ who are put into the "wrong" category. This is the "fuzzy" RD design illustrated by the selection criteria in Figure 2 of the paper:

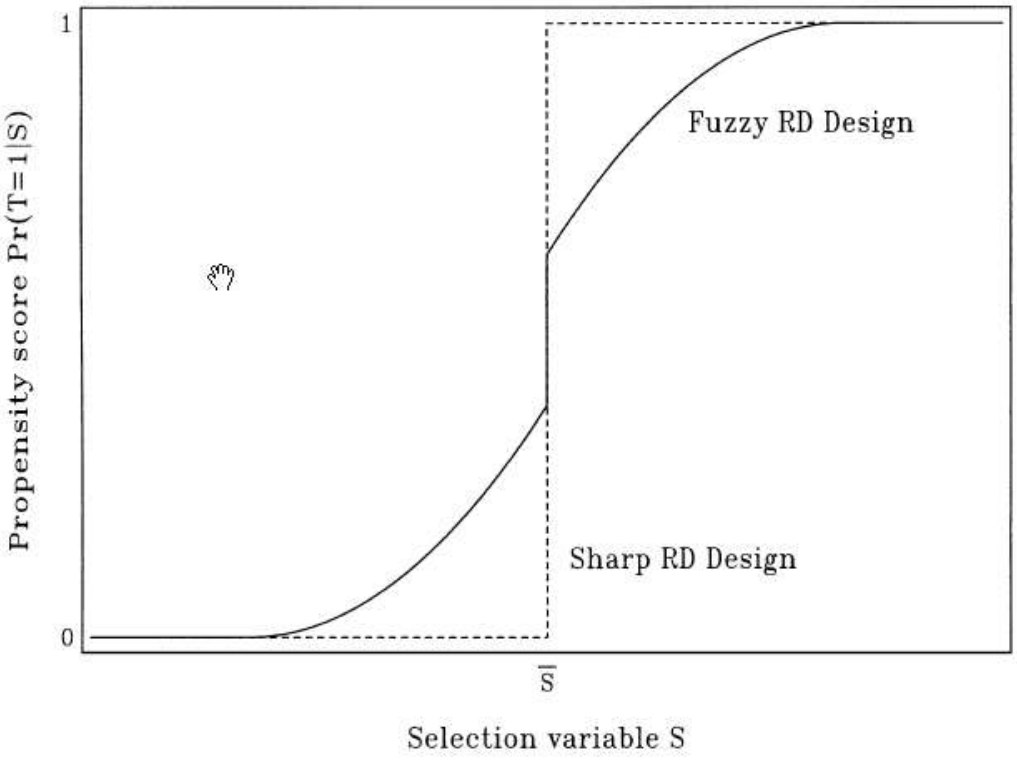


FIGURE 2

ASSIGNMENT IN THE SHARP (DASHED) AND FUZZY (SOLID) RD DESIGN

- In this case, we have to use the "predicted T" rather than the actual T and our identification of α becomes:

$$(9) \qquad \frac{\lim_{S \downarrow S} E[Y \mid S] - \lim_{S \uparrow S} E[Y \mid S]}{\lim_{S \downarrow S} E[T \mid S] - \lim_{S \uparrow S} E[T \mid S]}$$

- Aid offers depend on the thresholds, but also on need for "filers" who applied for need-based aid:

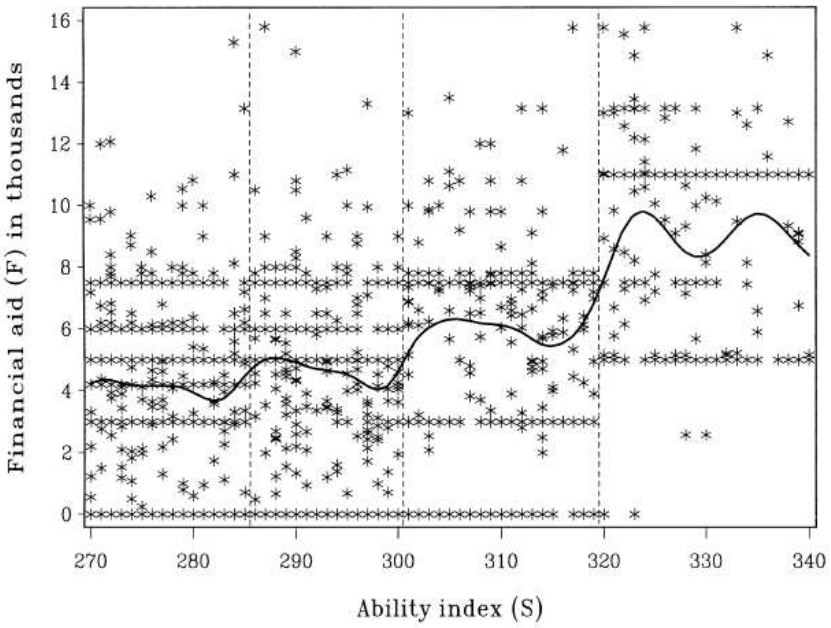


FIGURE 3

FINANCIAL AID OFFERS—FILERS. RAW DATA AND SPLINE SMOOTH (SOLID CURVE)

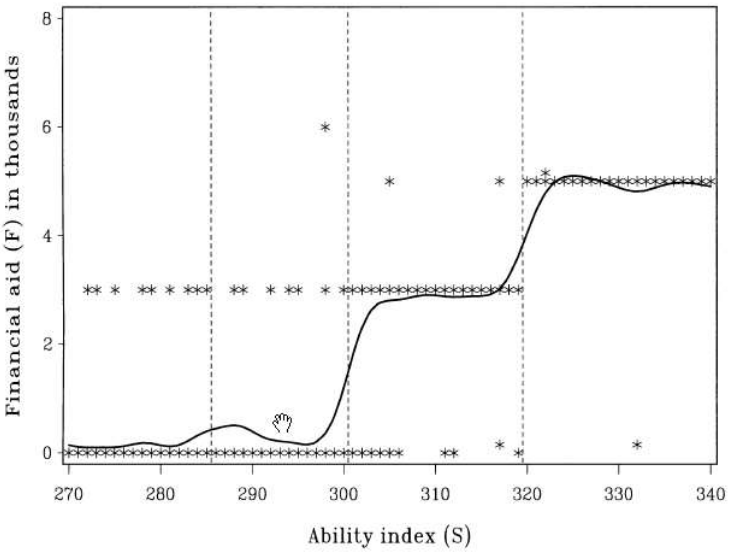


FIGURE 4

FINANCIAL AID OFFERS—NONFILERS. RAW DATA AND SPLINE SMOOTH (SOLID CURVE)

- Clearly the thresholds are important determinants of the amount of aid offered, especially for "non-filers," so this suggests that RD at the thresholds might be a useful way to identify the effect of aid on enrollment.
- For filers, estimated relationship between *S* and enrollment probability is shown in Figure 7:

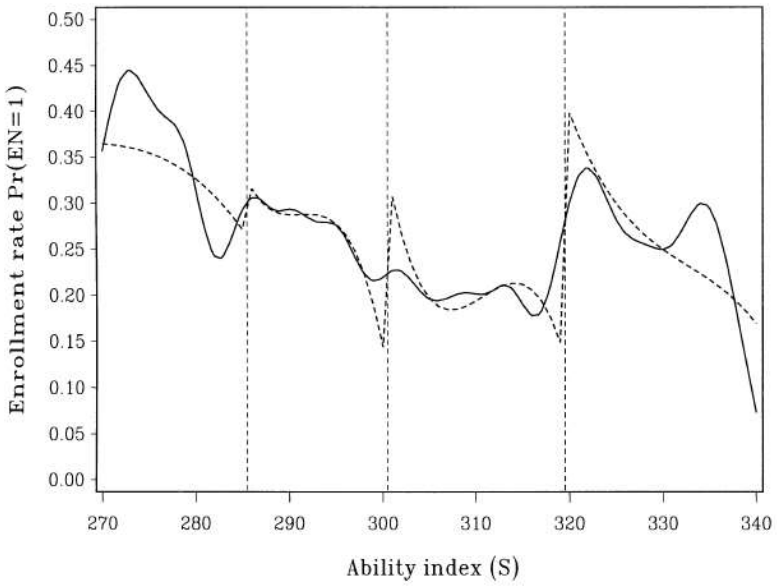


FIGURE 7

ENROLLMENT PROBABILITY—FILERS. PIECEWISE CUBIC REGRESSION (DASHED CURVE) AND NONPARAMETRIC SPLINE SMOOTH (SOLID CURVE)

- Note general downward slope of the relationship between *S* and probability of enrollment: Why?
- Note jumps at the threshold levels: these are the measured effects of the discrete change in aid associated with crossing a threshold: students on opposite sides of the line are very similar except that one group gets more aid than the other.
- Result is a statistically strong and economically significant effect of aid on enrollment: The elasticity for filers is 0.86, which is larger than most estimates in the literature obtained by traditional means (though not of those estimates that have access to additional data such as competing aid offers). The elasticity for non-filers is only 0.13, which is consistent with our expectation (and other evidence).

Section 15 Empirical Research Projects

Starting point: Question and data

- Starting point always must be "What question am I trying to answer?"
    - For thesis: something you can be interested in for a whole year
    - Something that can be answered
- Second consideration: "What data are available to help me find the answer?"
    - Macro data
    - Micro data from existing surveys
    - Collecting your own data from surveys

Methods

- Once you have the question and the data, you can carefully consider what method you should use
- Nature of dependent variable: continuous, limited?
    - Might need to consider LDV models
- What explanatory variables can you measure (and what is omitted)?
- Are there endogeneity concerns?
    - If yes, are appropriate instruments available to allow IV estimation?
- Are there other concerns about the error term?
    - Heteroskedasticity?
    - Autocorrelation?
- Are your data time series, cross section, pooled, or panel?
    - Appropriate models for each, including stationarity concerns
- What is the appropriate specification?
    - Functional form
    - Scaling and/or differencing to make the variables comparable

Estimation, diagnostic testing, re-estimation

- What did you learn from the first regression?
- Are there issues in the residuals or diagnostics based on the coefficients or residuals that suggest that your assumptions are incorrect?
    - Look for outliers and consider why they do not fit
    - (Errors in data)
- Can you test the underlying assumptions formally? Are they OK?

Writing the paper

- Introduction
    - What is the question?
    - How do you go about answering it?
    - What do you conclude?
- Theory section
    - What does economic theory tell us about the question?
    - What variables *should* be in the regression?
    - What considerations does theory suggest about functional form (e.g., CRTS)?
- Literature review
    - May come before theory section
    - Who else has explored this question and what did they find?
- Methods and data section
    - What estimation methods and tests are you proposing to use?
        - Why are these methods appropriate?
    - What data do you have (and not have)?
        - What issues of measurement might be important?
- Results section
    - Regression tables with basic description of results
    - Text must read as a narrative, referring to tables but not relying on them to tell the story.
- Analysis/interpretation/discussion section
    - What do the results mean?
    - Are there simulated experiments using your model that would help the reader understand your results?
    - How strong are the results?
    - Issues of internal and external validity: is it safe to draw conclusions based on your results?
- Conclusion
    - What do you conclude from your analysis?
    - What additional work remains to be done in future research?