Section 7: The Delta Method and the BOOtstrap

Valentine Gilbert

October 30, 2020

Overview

Nonlinear Hypotheses 1/3: Motivation

- So far we've learned how to test hypotheses about:
  - Single parameters (e.g. $H_0 : \beta_1 = 0$)
  - Linear combinations of parameters (e.g. $H_0 : 6\beta_2 + 72\beta_3 = 0$)
  - Joint linear hypotheses (e.g. $H_0 : \beta_1 = 0$ AND $6\beta_2 + 72\beta_3 = 0$)
- But many estimands of interest are **nonlinear** functions of parameters or are **nonparametric**
- Examples:
  - The change in *levels* of outcome as a function of parameters from a log-linear regression
  - The extent of bunching at kinks in the tax schedule bunching
- Two questions:
  - How do we estimate these?
  - How do we test hypotheses/quantify uncertainty around estimates of these?

Nonlinear Hypotheses 2/3: Estimation

Estimation of nonlinear estimands is straightforward:

- Let the estimand be $\theta \equiv g(\beta)$, where g is a smooth function and β is a vector of parameters
- Thanks to Slutsky's theorems, we know that $\hat{\theta} \equiv g(\hat{\beta})$ is a consistent estimator of $g(\beta)$ as long as $\hat{\beta}$ is a consistent estimator of β.
- Example: Suppose you want to estimate the average effect of going from 11 to 12 years of schooling on the *level* of earnings by regressing log earnings on years of schooling
    - $\log(\text{Earn}_i) = \beta_0 + \beta_1 \text{Edu}_i + u_i$
    - $\mathrm{E}(\text{Earn}_i | \text{Edu}_i = x) = e^{\beta_0 + \beta_1 x + \sigma^2/2}$ (assuming $u_i | \text{Edu}_i \sim \mathcal{N}(0, \sigma^2)$)
        - $\mathrm{E}(u_i | \text{Edu}_i) = 0$, so why isn't $\mathrm{E}(\text{Earn}_i | \text{Edu}_i = x) = e^{\beta_0 + \beta_1 x + 0}$? Jensen's Inequality!
    - Therefore $\hat{\theta} = e^{\hat{\beta}_0 + 12\hat{\beta}_1 + \hat{\sigma}^2/2} - e^{\hat{\beta}_0 + 11\hat{\beta}_1 + \hat{\sigma}^2/2}$

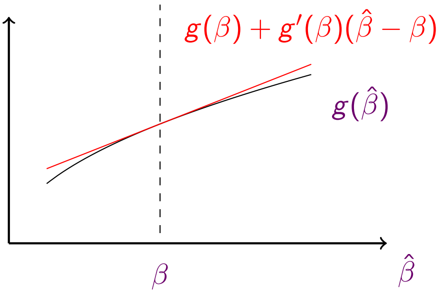
Nonlinear Hypotheses 3/3: Inference

Inference about nonlinear estimates is scarier

- For linear functions of parameter estimates, we could rely on simple rules about the variance operator
- For example:
  - $\mathrm{Var}(6\hat{\beta}_1 + 72\hat{\beta}_3) = 6^2\mathrm{Var}(\hat{\beta}_1) + 72^2\mathrm{Var}(\hat{\beta}_3) + 2 * 6 * 72\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_3)$
  - Then just plug in elements of the variance/covariance matrix to estimate this
- But how do we simplify $\mathrm{Var}(e^{\hat{\beta}_0+12\hat{\beta}_1+\hat{\sigma}^2/2} - e^{\hat{\beta}_0+11\hat{\beta}_1+\hat{\sigma}^2/2})$!?!?
- To estimate standard errors, we have to use the delta method or the bootstrap

The Delta Method 1/5: Intuition

- As we just saw, it's really hard to simplify the variance of a complicated nonlinear function of parameter estimates
- We get around this by using a **linear approximation** of our complicated nonlinear function
- Specifically, use a first-order Taylor-series approximation around the true values of our parameters: $g(\hat{\beta}) \approx g(\beta) + g'(\beta)(\hat{\beta} - \beta)$



The Delta Method 2/5: Estimating Variance

- Once we've taken a linear approximation of our nonlinear function, taking the variance of our estimate is just like in the linear case
- With one parameter:

$$g(\hat{\beta}) \approx g(\beta) + g'(\beta)(\hat{\beta} - \beta)$$
$$\text{Var}(g(\hat{\beta})) \approx g'(\beta)^2 \text{Var}(\hat{\beta})$$
$$\widehat{\text{Var}}(g(\hat{\beta})) = g'(\hat{\beta})^2 \widehat{\text{Var}}(\hat{\beta})$$

- With two parameters:

$$g(\hat{\beta}_0, \hat{\beta}_1) \approx g(\beta_0, \beta_1) + \frac{\partial g(\beta_0, \beta_1)}{\partial \beta_0}(\hat{\beta}_0 - \beta_0) + \frac{\partial g(\beta_0, \beta_1)}{\partial \beta_1}(\hat{\beta}_1 - \beta_1)$$
$$\text{Var}(g(\hat{\beta}_{,0}\, \hat{\beta}_1)) \approx \left(\frac{\partial g(\beta_0, \beta_1)}{\partial \beta_0}\right)^2 \text{Var}(\hat{\beta}_0) + \left(\frac{\partial g(\beta_0, \beta_1)}{\partial \beta_1}\right)^2 \text{Var}(\hat{\beta}_1)$$
$$+ 2\frac{\partial g(\beta_0, \beta_1)}{\partial \beta_0}\frac{\partial g(\beta_0, \beta_1)}{\partial \beta_1}\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$$

- Again, estimate this by plugging in sample estimates for $\beta_0, \beta_1, \text{Var}(\cdot)$, and $\text{Cov}(\cdot)$

The Delta Method 3/5: Simple Example

Consider the estimand $\beta_0\beta_1$. The estimator is $\hat{\beta}_0\hat{\beta}_1$. How do we estimate the variance of this with the delta method?

- $\hat{\beta}_0\hat{\beta}_1 \approx \beta_0\beta_1 + \beta_1(\hat{\beta}_0 - \beta_0) + \beta_0(\hat{\beta}_1 - \beta_1)$
- So $\mathrm{Var}(\hat{\beta}_0\hat{\beta}_1) \approx \beta_1^2\mathrm{Var}(\hat{\beta}_0) + \beta_0^2\mathrm{Var}(\hat{\beta}_1) + 2\beta_0\beta_1\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)$
- $\widehat{\mathrm{Var}}(\hat{\beta}_0\hat{\beta}_1) = \hat{\beta}_1^2\widehat{Var}(\hat{\beta}_0) + \hat{\beta}_0^2\widehat{\mathrm{Var}}(\hat{\beta}_1) + 2\hat{\beta}_0\hat{\beta}_1\widehat{\mathrm{Cov}}(\hat{\beta}_0, \hat{\beta}_1)$

How can we test the hypothesis that $\beta_0\beta_1 = 1$?

- Use the F-statistic: $F = (\hat{\beta}_0\hat{\beta}_1 - 1)^2/\widehat{\mathrm{Var}}(\hat{\beta}_0\hat{\beta}_1) \xrightarrow{d} \chi_1^2$

The Delta Method 4/5: Trickier Example

Now let's return to the example of estimating the average effect of an increase in schooling from 11 to 12 years on the level of earnings

- We said that $\theta = e^{\beta_0+12\beta_1+\sigma^2/2} - e^{\beta_0+11\beta_1+\sigma^2/2}$
- So we can write $\theta \equiv g(\beta_0, \beta_1, \sigma^2)$
  - $\partial g(\cdot)/\partial\beta_0 = e^{\beta_0+12\beta_1+\sigma^2/2} - e^{\beta_0+11\beta_1+\sigma^2/2}$
  - $\partial g(\cdot)/\partial\beta_1 = 12e^{\beta_0+12\beta_1+\sigma^2/2} - 11e^{\beta_0+11\beta_1+\sigma^2/2}$
  - $\partial g(\cdot)/\partial\sigma^2 = .5e^{\beta_0+12\beta_1+\sigma^2/2} - .5e^{\beta_0+11\beta_1+\sigma^2/2}$
- Then plug these into:

$$\begin{aligned}\mathrm{Var}(\hat\theta) \approx &\left(\frac{\partial g(\cdot)}{\partial\beta_0}\right)^2\mathrm{Var}(\hat\beta_0) + \left(\frac{\partial g(\cdot)}{\partial\beta_1}\right)^2\mathrm{Var}(\hat\beta_1) + \left(\frac{\partial g(\cdot)}{\partial\sigma_2}\right)^2\mathrm{Var}(\hat\sigma^2)\\ &+ 2\frac{\partial g(\cdot)}{\partial\beta_0}\frac{\partial g(\cdot)}{\partial\beta_1}\mathrm{Cov}(\hat\beta_0, \hat\beta_1) + 2\frac{\partial g(\cdot)}{\partial\beta_0}\frac{\partial g(\cdot)}{\partial\sigma^2}\mathrm{Cov}(\hat\beta_0, \hat\sigma^2) + 2\frac{\partial g(\cdot)}{\partial\beta_1}\frac{\partial g(\cdot)}{\partial\sigma_2}\mathrm{Cov}(\hat\beta_1, \hat\sigma^2)\end{aligned}$$

- This is equivalent to $\frac{\partial g}{\partial\gamma}(\gamma)'\Omega\frac{\partial g}{\partial\gamma}(\gamma)/N$, where $\frac{\partial g}{\partial\gamma}(\gamma)$ is the vector of partial derivatives and Ω is the asymptotic variance/covariance matrix for my parameters

The Delta Method 5/5: Example Stata Code

```
*Estimate regression
reg log_earn edu, hc2

*Define dg_dtheta = (dg/db1, dg/db0, dg/dgsigma^2)
*Note that constant goes after all the other covariates to conform with variance-covariance matrix
local dg_db0 = exp(_b[_cons]+_b[edu]*12+(e(rmse)^2)/2) - exp(_b[_cons]+_b[edu]*11+(e(rmse)^2)/2)
local dg_db1 = (12*exp(_b[_cons]+_b[edu]*12+(e(rmse)^2)/2)) - (11*exp(_b[_cons]+_b[edu]*11+(e(rmse)^2)/2))
local dg_sigma2 = (0.5*exp(_b[_cons]+_b[edu]*9+(e(rmse)^2)/2)) - (0.5*exp(_b[_cons]+_b[edu]*8+(e(rmse)^2)/2))
matrix define gradient = (`dg_db1' \ `dg_db0' \ `dg_sigma2')
matrix list gradient

*Define Omega
*Multiply var-cov matrix by sample size (not strictly necessicary, later will divide by N)
matrix define V = 929*e(V)
matrix list V

*Add variance of sigma-hat^2 to V and call that Omega
local res_var = 2*(e(rmse)^4)
matrix define omega = V\(0,0)
matrix define omega = omega, (0 \ 0 \ `res_var')
matrix list omega

*Caclulate (dg/dtheta)' * omega * (dg/dtheta)
matrix define variance = (gradient')*omega*gradient
matrix list variance
local variance = variance[1,1]

local delta_se = sqrt(`variance'/929)

***Display Delta Method results
display "estimate = " `theta1'
display "variance = " `variance'
display "standard error =" `delta_se'
```

The Non-Parametric Bootstrap: Intuition

- Recall that Fisher's exact test quantifies design-based uncertainty by drawing counterfactual randomization vectors to generate a distribution of counterfactual test statistics
- **The non-parametric bootstrap** quantifies sampling uncertainty by simulating new samples from the existing data, generating a distribution of parameter estimates
- Ideally, we would draw new samples from the population of interest to quantify sampling uncertainty
- But this is infeasible: We only have access to the sample at hand
- Solution is to draw observations from the sample at hand *with replacement*
- The empirical CDF is an approximation for the true underlying CDF of the data

The Non-Parametric Bootstrap: Simple Example

- Suppose you run the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$ on N observations
- You're interested in the standard error of $\hat{\beta}_0\hat{\beta}_1$ but you don't like the delta method
- Bootstrap your standard errors as follows:
  - Draw a new sample of size N from your data by randomly drawing observations with replacement
  - Rerun the regression and save your estimate, $\hat{\beta}_{0b}\hat{\beta}_{1b}$, where b indexes the bootstrap sample
  - Do this B times, where B is large (ideally $\geq$100,000 times for papers, $\leq$ 10,000 for psets)
  - Take the variance of $\hat{\beta}_{0b}\hat{\beta}_{1b}$. This is the estimated variance of your estimator!
- The same procedure works for much more complicated estimands, including non-parametric ones!

The Parametric Bootstrap

- The **parametric bootstrap** can give more precise estimates of $\widehat{var}$ than the non-parametric bootstrap if the necessary assumptions hold
- Those assumptions are that the errors are independent of X_i and are homoskedastic
- To estimate standard errors with the parametric bootstrap, reassign residuals to observations:
    - Calculate $\hat{Y}_i$ and $\hat{u}_i$ *in the original sample*
    - Sample from $\{\hat{u}_j\}$ with replacement to get a new bootstrapped sample of residuals, $\{\hat{u}_{b,j}\}$
    - Define $Y_{b,i} = \hat{Y}_i + \hat{u}_{b,i}$
    - Regress the bootstrapped $Y_{b,i}$ on the original X_is to estimate $\hat{\beta}_b$
    - Repeat B times and take the variance of your estimate

The Wild Bootstrap

- The **wild bootstrap** allows for heteroskedasticity, but assumes errors are mean independent: $E[u_i|X_i] = 0$
- Can be especially useful in RCTs with few treated groups
- To estimate standard errors with the wild bootstrap, randomly add or subtract residuals:
  - Calculate $\hat{Y}_i$ and $\hat{u}_i$ *in the original sample*
  - Let $k = (1-\sqrt{5})/2$, $p = (1+\sqrt{5})/2\sqrt{5}$. Define outcome for the ith observation in bootstrapped sample b as

$$Y_{b,i} = \begin{cases} \hat{Y}_i + k\hat{u}_i & \text{probability } p \\ \hat{Y}_i + (1-k)\hat{u}_i & \text{with probability } 1-p \end{cases}$$

  - Regress the bootstrapped $Y_{b,i}$ on the original X_is to estimate $\hat{\beta}_b$
  - Repeat B times and take the variance of your estimate

The Bootstrap: Practical Advice

- Use “bsample” in Stata and the “boot” package in R.
- Fixed versus variable X_i’s:
  - On the problem set, you’re asked to estimate the standard error of an estimated policy impact, $\hat{\theta}$, that is a function of regression parameters *and* the X_i’s in the data
  - You’re asked to bootstrap the standard error but, for each bootstrapped sample, estimate θ_b using the X_i’s that correspond to the original sample
  - This means you should estimate the regression parameters from the bootstrapped sample, then use those estimates to predict $\hat{\theta}$ in the *original* sample
  - This corresponds to inference about the policy impact on your original sample, rather than on some superpopulation

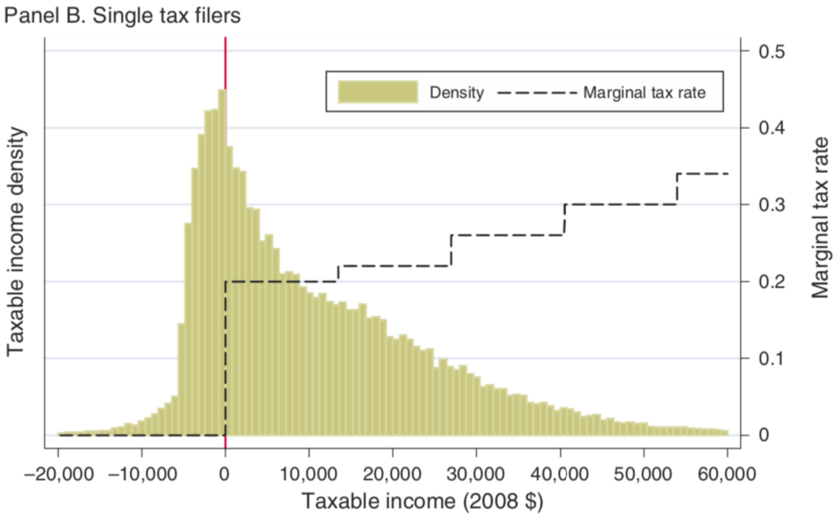


FIGURE 6. TAXABLE INCOME DENSITY, 1960–1969: BUNCHING AROUND FIRST KINK

back