Bootstrapping in Stata

Bruce E. Hansen

Econ 706

Introduction: Uses of Bootstrap in Econometrics

- Standard Errors
  - Coefficient estimate
  - Function of estimates
- Confidence Intervals
  - Normal-based
  - Percentile
  - Bias-Corrected (BC)
  - Accelerated and Bias-Corrected (BC_a)
  - Percentile-t
- Joint Tests
- Bootstrap for Quantile Regression
- Number of bootstrap replications

Example: Probit Model for Marriage

Sample: March 2009 CPS
Population: U.S. Black women in Midwest (n=433)
Percent Married: 37%
Probit for *married* as a function of

- *age*, *age*^{2}, *education*, .

.probit mar age age2 education if bf, r

- This calculates (robust) asymptotic standard errors

```
Probit regression                               Number of obs   =        433
                                                Wald chi2(3)    =      26.03
                                                Prob > chi2     =     0.0000
Log pseudolikelihood = -271.96692               Pseudo R2       =     0.0465
```

mar	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.1380565	.0507529	2.72	0.007	.0385825	.2375304
age2	-.0014209	.0006022	-2.36	0.018	-.0026013	-.0002406
education	.0898554	.0272241	3.30	0.001	.0364971	.1432137
_cons	-4.714038	1.103265	-4.27	0.000	-6.876397	-2.551678

STATA Command for Bootstrap Standard Error

.probit mar age age2 education if bf, vce(bootstrap, reps(1000))
or
.bootstrap, reps(1000): probit mar age age2 education if bf

- The vce(bootstrap) option specifies to use the bootstrap for variance-covariance estimation (vce)
- If reps(#) is omitted, the default bootstrap replications is $R = 50$.
- The **vce(bootstrap)** option works with many estimation commands.
  - STATA recommends **vce(bootstrap)** over **bootstrap** as the estimation command handles clustering and model-specific details
- **bootstrap** works more broadly, including non-estimation and user-written commands, or functions of coefficients

Bootstrap Standard Error Calculation

- Computes the coefficient estimates $\hat{\beta}$ on the estimation sample
- Draws n observations at random from the estimation sample
- Computes the estimates $\hat{\beta}^*$ on this simulated sample
- Repeats this R times, obtaining $\hat{\beta}^*_b$, $b = 1, ..., R$

$$\begin{aligned} \bar{\beta}^* &= \frac{1}{R}\sum_{b=1}^{R}\hat{\beta}^*_b \\ \widehat{se(\hat{\beta})} &= \sqrt{\frac{1}{R-1}\sum_{b=1}^{R}\left(\hat{\beta}^*_b - \bar{\beta}^*\right)^2} \end{aligned}$$

.probit mar age age2 education if bf, vce(bootstrap, reps(1000))

```
Probit regression                               Number of obs     =        433
                                                Replications      =      1,000
                                                Wald chi2(3)      =      25.18
                                                Prob > chi2       =     0.0000
Log likelihood = -271.96692                     Pseudo R2         =     0.0465
```

mar	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
age	.1380565	.0536625	2.57	0.010	.0328798	.2432331
age2	-.0014209	.0006385	-2.23	0.026	-.0026724	-.0001695
education	.0898554	.0279496	3.21	0.001	.0350752	.1446356
_cons	-4.714038	1.159029	-4.07	0.000	-6.985693	-2.442382

Bootstrap standard errors.

Function of Coefficient

In the equation

$$\Pr\left(married = 1\right) = \Phi\left(\beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 edu\right)$$

The age at which the probability is maximized (if $\beta_1 > 0$ and $\beta_2 < 0$) is

$$\theta = \frac{-\beta_1}{2\beta_2}$$

This is easily estimated by

$$\hat{\theta} = \frac{-\hat{\beta}_1}{2\hat{\beta}_2}.$$

We could obtain standard errors by the Delta method.
Because of the nonlinearity, bootstrap standard errors will be more reliable.

Delta Method

.nlcom (theta: -_b[age]/(_b[age2]*2))

```
. nlcom (theta: -_b[age]/(_b[age2]*2))

       theta:  -_b[age]/(_b[age2]*2)
```

mar	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
theta	48.57935	3.873624	12.54	0.000	40.98719	56.17151

The output shows

- The estimated coefficient $\hat{\theta} = \frac{-\hat{\beta}_1}{2\hat{\beta}_2} = 49$
  - age of maximum probability of marriage
- Its asymptotic standard error
- Normal-based confidence interval

Bootstrap Standard error

.bootstrap theta=(-_b[age]/(_b[age2]*2)), reps(1000): probit mar age age2 education if bf

$$\hat{\theta}_b^* = \frac{-\hat{\beta}_{1b}^*}{2\hat{\beta}_{2b}^*}$$

$$\widehat{se(\hat{\theta})} = \sqrt{\frac{1}{R-1}\sum_{b=1}^{R}\left(\hat{\theta}_b^* - \bar{\theta}^*\right)^2}$$

```
Probit regression                               Number of obs      =       433
                                                Replications       =     1,000

      command:  probit mar age age2 education
        theta:  -_b[age]/(_b[age2]*2)
```

	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
theta	48.57935	17.42398	2.79	0.005	14.42898	82.72973

Bootstrap Confidence Intervals

There are several methods

- Normal-based
- Percentile
- Bias-corrected (BC)
- BC_a : Bias-corrected and accelerated
- Percentile-t

Choices

- STATA reports normal-based intervals in default tablle
  - Least desireable
- Percentile and BC intervals are easy to obtain
  - BC preferred to percentile
- The BC_a is expected to perform better, but can be computationally costly in large data sets and/or non-linear estimation
- The percentile-t require more programming and requires standard errors, but can perform well

Normal Confidence interval

Stata reports a "Normal-based 95% confidence interval":

$$\hat{\beta} \pm 1.96\widehat{se(\hat{\beta})}$$

- It uses the bootstrap standard error
- And the asymptotic normal approximation.
- Not a good choice

Percentile Interval

- Let Q^*_α be the α'th quantile of $\hat{\beta}^*$
- Estimated as the α'th empirical quantile of the simulated $\hat{\beta}^*_b$
- The $1-2\alpha$ percentile interval for β is $\left[Q^*_\alpha, Q^*_{1-\alpha}\right]$
- If $\hat{\beta}$ is biased, the percentile interval will be even more biased than asymptotic interval.
- Not recommended

Bias-Corrected Percentile Interval:

Define a measure of bias and its bootstrap estimator

$$\begin{aligned} p &= \Pr\left(\hat{\beta} < \beta\right) \\ \hat{p} &= \frac{1}{R}\sum_{b=1}^{B}\left\{\hat{\beta}_b^* < \hat{\beta}\right\} \end{aligned}$$

Transform $\hat{p}$ into normal units

$$\hat{c} = \Phi^{-1}\left(\hat{p}\right)$$

Define the shifted percentage

$$\lambda(\alpha) = \Phi\left(2\hat{c} + Z_\alpha\right)$$

where Z_α is the $\alpha' th$ quantile of $N(0,1)$.

The BC percentile interval for β is $\left[Q^*_{\lambda(\alpha)}, Q^*_{\lambda(1-\alpha)}\right]$

If $\hat{\beta}$ is unbiased then $\hat{p} = .5$, $\hat{c} = 0$, $\lambda(\alpha) = \alpha$ and $Q^*_{\lambda(\alpha)} = Q^*_\alpha$

When $\hat{\beta}$ is negatively biased then $\hat{p} > .5$, $\hat{c} > 0$, $\lambda(\alpha) > \alpha$ and $Q^*_{\lambda(\alpha)} > Q^*_\alpha$

If $\hat{\theta} \sim N(\theta - b, \sigma^2)$ is normal but biased, the BC interval will be exact.

Bias-Corrected and Accelerated Interval

BC_a is similar to BC

$$\lambda(\alpha) = \Phi\left(\hat{c} + \frac{\hat{c} + Z_\alpha}{1 - a\left(\hat{c} + Z_\alpha\right)}\right)$$

When $a = 0$, then BC and BC_a coincide.

The optimal a is called the "acceleration" because it refers to the rate of change of the standard error of $\hat{\beta}$ with respect to β, and is a scale of the skewness of $\hat{\beta}$.

a is estimated in STATA by the jackknife, which requires n re-estimations.

If n is large and estimation nonlinear this is costly.

Computation

For BC interval

.probit mar age age2 education if bf, vce(bootstrap, reps(1000))
.estat bootstrap

- **estat** reports the bootstrap estimate of bias and the BC percentile interval
- A postestimation command:
  - Needs to follow estimation with bootstrap standard errors
- **estat bootstrap, all**

```
. estat bootstrap

Probit regression                               Number of obs      =       433
                                                Replications       =      1000
```

mar	Observed Coef.	Bias	Bootstrap Std. Err.	[95% Conf.	Interval]	
age	.13805646	.0092831	.05612334	.0481732	.2603911	(BC)
age2	-.00142094	-.0001115	.00066532	-.0029528	-.0003454	(BC)
education	.0898554	.0007389	.02830316	.0320437	.1455764	(BC)
_cons	-4.7140375	-.1992479	1.2492997	-7.535929	-2.723436	(BC)

(BC) bias-corrected confidence interval

```
. estat bootstrap, all

Probit regression                               Number of obs      =       433
                                                Replications       =      1000
```

mar	Observed Coef.	Bias	Bootstrap Std. Err.	[95% Conf.	Interval]	
age	.13805646	.0092831	.05612334	.0280567	.2480562	(N)
				.0568253	.2738451	(P)
				.0481732	.2603911	(BC)
age2	-.00142094	-.0001115	.00066532	-.0027249	-.0001169	(N)
				-.0030192	-.0004481	(P)
				-.0029528	-.0003454	(BC)
education	.0898554	.0007389	.02830316	.0343822	.1453286	(N)
				.0344849	.1462973	(P)
				.0320437	.1455764	(BC)
_cons	-4.7140375	-.1992479	1.2492997	-7.16262	-2.265455	(N)
				-7.59685	-2.782275	(P)
				-7.535929	-2.723436	(BC)

```
(N)    normal confidence interval
(P)    percentile confidence interval
(BC)   bias-corrected confidence interval
```

For BC_a interval

.probit mar age age2 education if bf, vce(bootstrap, reps(1000) bca)

.estat bootstrap, bca

or

.estat bootstrap, all

- The **bca** option in **vce** tells STATA to calculate the acceleration a
- This is done by the jackknife and can be computationally costly
- The bca option in estat tells STATA to report the BC_a interval instead of the BC

```
. estat bootstrap, bca

Probit regression                              Number of obs      =       433
                                               Replications       =      1000
```

mar	Observed Coef.	Bias	Bootstrap Std. Err.	[95% Conf.	Interval]	
age	.13805646	.0085802	.05474557	.0327771	.2320904	(BCa)
age2	-.00142094	-.0000992	.00065104	-.0025545	-.0002154	(BCa)
education	.0898554	.0033017	.02714671	.0373695	.1456622	(BCa)
_cons	-4.7140375	-.2223964	1.1774249	-6.832892	-2.293183	(BCa)

```
(BCa)  bias-corrected and accelerated confidence interval
```

```
. estat bootstrap, all

Probit regression                               Number of obs      =       433
                                                Replications       =      1000

------------------------------------------------------------------------------
             |    Observed               Bootstrap
         mar |       Coef.       Bias    Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .13805646   .0085802   .05474557    .0307571   .2453558   (N)
             |                                       .0540528    .261395   (P)
             |                                          .0467   .2478884  (BC)
             |                                       .0327771   .2320904 (BCa)
        age2 |  -.00142094  -.0000992   .00065104    -.002697  -.0001449   (N)
             |                                      -.0029675  -.0004549   (P)
             |                                      -.0027947  -.0003734  (BC)
             |                                      -.0025545  -.0002154 (BCa)
   education |    .0898554   .0033017   .02714671    .0366488    .143062   (N)
             |                                       .0406278   .1490357   (P)
             |                                       .0374336   .1466715  (BC)
             |                                       .0373695   .1456622 (BCa)
       _cons |  -4.7140375  -.2223964   1.1774249   -7.021748  -2.406327   (N)
             |                                      -7.310817  -2.935916   (P)
             |                                      -7.129984  -2.674825  (BC)
             |                                      -6.832892  -2.293183 (BCa)
------------------------------------------------------------------------------
(N)    normal confidence interval
(P)    percentile confidence interval
(BC)   bias-corrected confidence interval
(BCa)  bias-corrected and accelerated confidence interval
```

Function of Coefficient

- Works the same
- **.bootstrap theta=(-_b[age]/(_b[age2]*2)), reps(1000): probit mar age age2 education if bf**
- **.estat bootstrap**

```
. estat bootstrap

Probit regression                                 Number of obs      =       433
                                                  Replications       =      1000

      command:  probit mar age age2 education
        theta:  -_b[age]/(_b[age2]*2)
```

	Observed Coef.	Bias	Bootstrap Std. Err.	[95% Conf. Interval]		
theta	48.579351	.9110359	11.907861	43.86382	66.01958	(BC)

(BC) bias-corrected confidence interval

Alternative Syntax

.bootstrap "probit mar age age2 education if bf" _b, reps(1000)
or
.bootstrap "probit mar age age2 education if bf" _b _se, reps(1000)
contrast **.bootstrap, reps(1000): probit mar age age2 education if bf**

- Computes BC percentile intervals with one command
- Requires expression list (**_b** and/or **_se**) to specify statistics
- Works on many STATA operations
- Reports
  - bias
  - standard error
  - normal confidence interval
  - percentile interval
  - bias-corrected interval

.bootstrap "probit mar age age2 education if bf" _b, reps(1000)

```
Bootstrap statistics                              Number of obs     =
                                                  Replications      =
```

Variable	Reps	Observed	Bias	Std. Err.	[95% Conf.	Interval]
b_age	1000	.1380565	.0145857	.0538347	.0324143	.2436986
					.0603044	.2668444
					.0486678	.2495679
b_age2	1000	-.0014209	-.0001774	.0006448	-.0026863	-.0001556
					-.0029643	-.0005232
					-.002781	-.0003897
b_education	1000	.0898554	.0010376	.0272817	.0363194	.1433914
					.0394318	.1495175
					.0381109	.1486244
b_cons	1000	-4.714037	-.2999402	1.151227	-6.973137	-2.454938
					-7.56545	-3.009196
					-6.978446	-2.75727

Percentile-t Intervals

- Let $se(\hat{\beta})$ be standard error for $\hat{\beta}$
  - Best if "robust" standard error
- Let $\hat{\beta}^*$, $se(\hat{\beta}^*)$ be bootstrap statistics.
- Define bootstrap t-statistics

$$t^* = \frac{\hat{\beta}^* - \hat{\beta}}{se(\hat{\beta}^*)}$$

- Let Q^*_α be the α'th quantile of t^*
- The $1 - 2\alpha$ percentile-t interval for β is $[\hat{\beta} - se(\hat{\beta})Q^*_{1-\alpha}, \hat{\beta} - se(\hat{\beta})Q^*_\alpha]$

Computation of percentile-t

- Not pre-programmed in STATA, but can be computed without programming
- Method described in Poi(STATA Journal, 2004)
- **.bootstrap "probit mar age age2 education if bf,r" _b _se, reps(1000) saving(bsdata) replace**
- In addition to earlier calculations, this stores the $R \times 1$ bootstrap statistics $\hat{\beta}^*$ and $se(\hat{\beta}^*)$ in **_b** and **_se** in the file *bsdata.dta*
- Be careful to specify the method for calculation of asymptotic standard errors
  - Here I use the robust option
- Next re-estimate model using original dataset using same command
- **.probit mar age age2 education if bf,r**

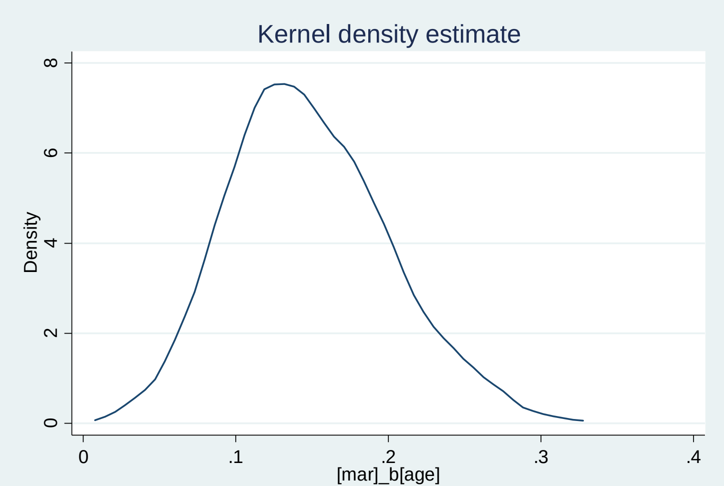
- Load save coefficients and standard errors into memory
  - **.use bsdata**
- create bootstrap t-ratios
  - **.gen t_age=(b_age - _b[age]) / se_age**
- calculate quantiles of t-ratio (e.g. 2.5% and 97.5%)
  - **._pctile t_age, p(2.5, 97.5)**
- display quantiles (check results) and confidence endpoints

```
. dis r(r2)
2.4759161

. dis r(r1)
-2.0422164

. dis _b[age] - _se[age]*r(r2)
.01239647

. dis _b[age] - _se[age]*r(r1)
.24170492
```

Graphical

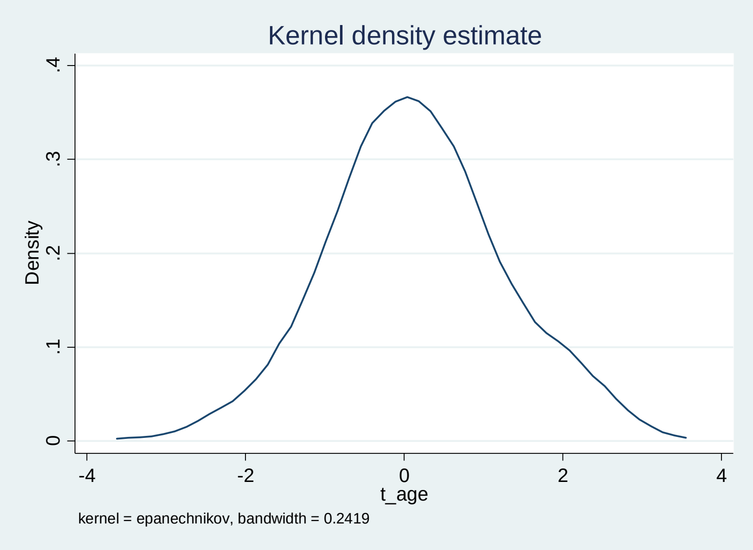
You can display density plots of the bootstrap distributions

- **.use bsdata**
- **.kdensity b_age**



t-ratio

- **.kdensity t_age**



Applying bootstrap to program output

- You can apply the bootstrap to complicated (multi-step) estimators
- You may need to write a program
- Apply the command **bootstrap** to the program
  - **.bootstrap "probit mar age" _b, reps(1000)**
  - **.bootstrap, reps(1000): probit mar age**
- Caveat: In many cases
  - STATA does not know the estimation sample
  - STATA will implement the bootstrap by drawing from all observations
  - For example, we had 207,921 total observation but only 433 in the estimation sample
  - It is best to first drop all observations which are not in the estimation sample, before running the bootstrap.
  - Otherwise, STATA will create samples of size 207,921

Joint Tests

- Wald Tests

$$W = \left(\hat{\theta} - \theta_0\right)' \hat{V}_\theta^{-1} \left(\hat{\theta} - \theta_0\right)$$

  where $\hat{V}_\theta$ is an estimate of $var(\hat{\theta})$

  - Conventional Wald test uses estimate of asymptotic variance matrix
  - STATA test after bootstrap estimation uses bootstrap estimate of variance matrix
  - STATA does not implement correct bootstrap joint test
    - p-values based on distribution of

$$W^* = \left(\hat{\theta}^* - \hat{\theta}\right)' \hat{V}_\theta^{*-1} \left(\hat{\theta}^* - \hat{\theta}\right)$$

    - Could be done via programming
    - Note that the bootstrap specifies the null values of θ to be $\hat{\theta}$, not θ_0

Quantile Regression

- Quantile Regression has its own bootstrap syntax
- Example: conditional median of hourly wage
  - function of *age*, *education*, *age*^{2}, *education*^{2}
  - same sample: black women, non-south

.qreg hrwage age age2 education if bf

```
Median regression                                   Number of obs =        433
  Raw sum of deviations 1689.071 (about 15.264423)
  Min sum of deviations 1334.419                    Pseudo R2     =     0.2100
```

hrwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.5230953	.2507331	2.09	0.038	.0302772	1.015913
age2	-.0036805	.0029256	-1.26	0.209	-.0094307	.0020697
education	2.330236	.1704213	13.67	0.000	1.995271	2.6652
_cons	-30.9965	5.583125	-5.55	0.000	-41.97019	-20.02282

Bootstrap standard errors

.bsqreg hrwage age age2 education if bf, reps(1000)

```
Median regression, bootstrap(1000) SEs              Number of obs =        433
  Raw sum of deviations 1689.071 (about 15.264423)
  Min sum of deviations 1334.419                    Pseudo R2     =     0.2100
```

hrwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.5230953	.26829	1.95	0.052	-.0042312	1.050422
age2	-.0036805	.0033228	-1.11	0.269	-.0102116	.0028505
education	2.330236	.2221715	10.49	0.000	1.893555	2.766916
_cons	-30.9965	5.809094	-5.34	0.000	-42.41433	-19.57868

Functions of parameters

.bootstrap theta=(-_b[age]/(_b[age2]*2)),reps(1000):qreg hrwage age age2 education if bf

```
Bootstrap results                               Number of obs     =        433
                                                Replications      =      1,000

      command:  qreg hrwage age age2 education
        theta:  -_b[age]/(_b[age2]*2)
```

	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf. Interval]	
theta	71.0625	1553345	0.00	1.000	-3044429	3044571

BC Percentile Intervals

estat bootstrap does not work with **qreg**

Instead use

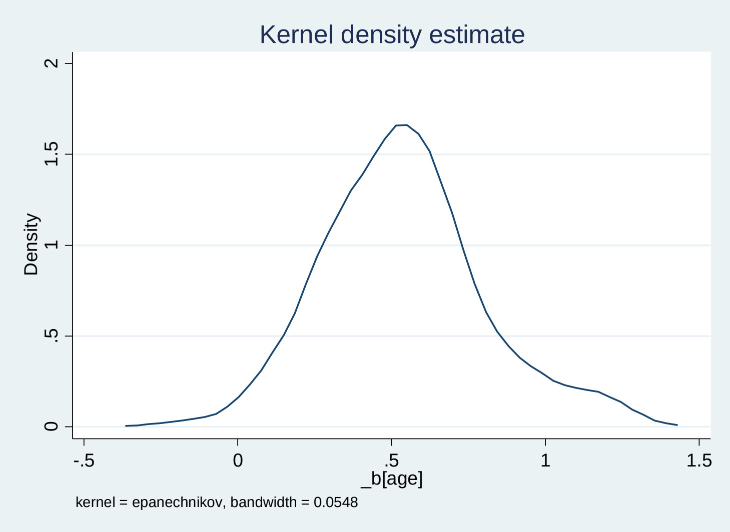
.bootstrap "qreg hrwage age age2 education if bf" _b, reps(1000) saving (bsdata) replace

```
Bootstrap statistics                          Number of obs    =       433
                                              Replications     =      1000

------------------------------------------------------------------------------
Variable     |  Reps  Observed       Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
       b_age |  1000  .5230953   .0135267  .2661738   .0007715   1.045419   (N)
             |                                        .0582089   1.161127   (P)
             |                                        .0525522   1.156968  (BC)
      b_age2 |  1000 -.0036805   -.000395  .0032746  -.0101064   .0027454   (N)
             |                                       -.0119208    .001881   (P)
             |                                       -.0110044   .0024124  (BC)
 b_education |  1000  2.330235  -.0193944  .2206628    1.89722   2.763251   (N)
             |                                        1.862233   2.706177   (P)
             |                                         1.86039   2.705699  (BC)
      b_cons |  1000  -30.9965   .4477234  5.793506  -42.36534  -19.62766   (N)
             |                                       -42.95274    -19.603   (P)
             |                                       -44.16151  -20.38403  (BC)
```

Plot bootstrap distributions

.use bsdata, replace

.kdensity b_age



Number of Bootstrap Replications

- Early literature suggested that small R (e.g. $R = 50$) is sufficient
- This advice is stated in STATA manual
- Recent research (Andrews & Buchinsky) says that this is far from sufficient
- $R = 1000$ is a minimum for most calculations
- $R > 3000$ is often necessary
- I suggest using $R = 10,000$ for final calculations if possible (for submission/publication)
- Andrews-Buchinsky derive methods for determining R
    - Poi (2004) describes STATA implementation
    - Might be easier to just set R large and be patient

Clustered Samples

- Bootstrap methods treat a cluster as an observation
- Resample entire clusters to create bootstrap data sets
- Example: Duflo, Dupas and Kramer (2011) investigate the impact of *tracking* on *testscores* in elementary schols in Kenya.
- 111 schools (clusters)

Linear regression, clustered by schoolid

.reg testscore tracking, cluster(schoolid)

```
. reg testscore tracking, cluster(schoolid)

Linear regression                               Number of obs     =      5,269
                                                F(1, 110)         =       3.65
                                                Prob > F          =     0.0586
                                                R-squared         =     0.0053
                                                Root MSE          =     .99743

                               (Std. Err. adjusted for 111 clusters in schoolid)
------------------------------------------------------------------------------
             |               Robust
   testscore |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    tracking |   .1469204   .0768674     1.91   0.059    -.0054128    .2992536
       _cons |  -.0824249   .0535249    -1.54   0.126    -.1884986    .0236488
```

Bootstrap commands

.bootstrap, reps(1000): reg testscore tracking, cluster(schoolid)
.reg testscore tracking, cluster(schoolid) vce(bootstrap, reps(1000) bca)

(Replications based on 111 clusters in schoolid)

testscore	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
tracking	.1469204	.0788651	1.86	0.062	-.0076523	.3014931
_cons	-.0824249	.0565709	-1.46	0.145	-.1933019	.0284521

.estat bootstrap, all

(Replications based on 111 clusters in schoolid)

testscore	Observed Coef.	Bias	Bootstrap Std. Err.	[95% Conf.	Interval]	
tracking	.14692041	-.0016042	.07886507	-.0076523	.3014931	(N)
				-.0095893	.2978102	(P)
				-.0105832	.2966537	(BC)
				-.0108045	.2962839	(BCa)
_cons	-.08242489	.0018846	.05657095	-.1933019	.0284521	(N)
				-.1935774	.0300271	(P)
				-.2031856	.0229642	(BC)
				-.2010211	.0251257	(BCa)

(N) normal confidence interval
(P) percentile confidence interval
(BC) bias-corrected confidence interval
(BCa) bias-corrected and accelerated confidence interval