BIOS 6312: Modern Regression Analysis

Andrew J. Spieker, Ph.D.

Assistant Professor of Biostatistics
Vanderbilt University Medical Center

Set 11: Bootstrap Methods

Version: 04/15/2021

Outline

Topics:

- Review of typical inference
- The nonparametric bootstrap

Outline

Topics:

- **Review of typical inference**
- The nonparametric bootstrap

Typical inference

Confidence intervals and inverting the test:

- Consider the following general quantity, which follows a familiar form:

$$S = \frac{\widehat{\theta} - \theta}{\widehat{\text{SE}}(\widehat{\theta})}$$

- When using this quantity to construct CIs, we often rely on two particular properties:
  - S is *pivotal* in large samples, meaning its asymptotic distribution does not depend upon θ.
  - S possesses a distribution that is approximately symmetric about zero in large samples.

Typical inference

Confidence intervals and inverting the test:

- Consider a coefficient, β, from a regression model:

$$\frac{\widehat{\beta} - \beta}{\widehat{\mathsf{SE}}(\widehat{\beta})} \mathrel{\dot{\sim}} t_{df}.$$

- Note that the pivotal property is embedded above. Further,

$$\begin{aligned}
t_{\alpha/2,df} \leq & \quad \tfrac{\widehat{\beta}-\beta}{\widehat{\mathsf{SE}}(\widehat{\beta})} & \leq t_{1-\alpha/2,df} \\
\iff t_{\alpha/2,df}\widehat{\mathsf{SE}}(\widehat{\beta}) \leq & \quad \widehat{\beta} - \beta & \leq t_{1-\alpha/2,df}\widehat{\mathsf{SE}}(\widehat{\beta}) \\
\iff -t_{1-\alpha/2,df}\widehat{\mathsf{SE}}(\widehat{\beta}) \leq & \quad \beta - \widehat{\beta} & \leq -t_{\alpha/2,df}\widehat{\mathsf{SE}}(\widehat{\beta}) \\
\iff \widehat{\beta} - t_{1-\alpha/2,df}\widehat{\mathsf{SE}}(\widehat{\beta}) \leq & \quad \beta & \leq \widehat{\beta} - t_{\alpha/2,df}\widehat{\mathsf{SE}}(\widehat{\beta})
\end{aligned}$$

- From symmetry property, further derive the following:

$$\widehat{\beta} - t_{1-\alpha/2,df}\widehat{\mathsf{SE}}(\widehat{\beta}) \leq \quad \beta \quad \leq \widehat{\beta} + t_{1-\alpha/2,df}\widehat{\mathsf{SE}}(\widehat{\beta})$$

- These properties are the basis for forming symmetric CIs based on large sample theory.

Typical inference

Confidence intervals and inverting the test:

- When no such pivotal quantity exists, confidence intervals can be obtained by directly inverting the test.
- "Find all $\beta^{(0)}$ such that $H_0 : \beta = \beta^{(0)}$ cannot be rejected."

Typical inference

Confidence intervals and inverting the test:

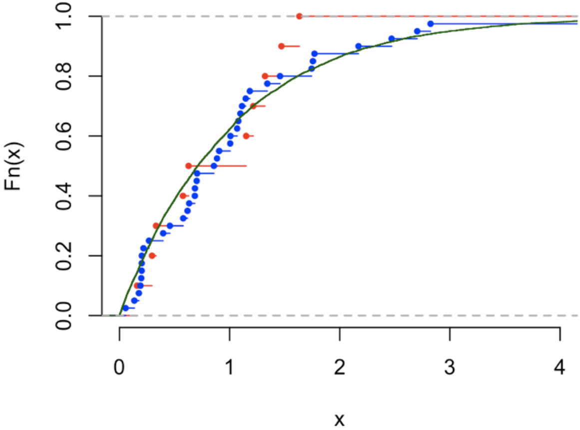
- In linear regression, an *exact* distribution for $\widehat{\beta}$ based on the t-distribution depends upon normality of the errors.
- That distribution is approximately correct for large samples even if normality does not hold.
- In smaller samples, the nonparametric bootstrap can be used to obtain CIs that do not rely on large sample theory.

Outline

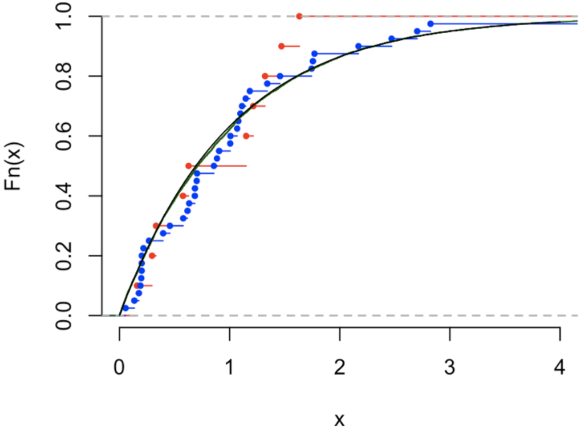
Topics:

- *Review of typical inference*
- **The nonparametric bootstrap**

Preliminaries: $\mathbb{F}_N$, approximates $F(x) = \mathrm{P}(X \leq x)$



Preliminaries: $\mathbb{F}_N$, approximates $F(x) = \mathrm{P}(X \leq x)$



The bootstrap

Main ideas:

- Let F denote cdf for $(\mathbf{X}, Y)$ or $(Y|\mathbf{X})$, depending on context; let $\mathbb{F}_N$ denote empirical cdf based on N observations.
  - $\boldsymbol{\beta} = T(F)$, and hence $\widehat{\boldsymbol{\beta}} = T(\mathbb{F}_N)$.
  - Absent parametric form, $\mathbb{F}_N$ is our best estimate of F.
- Repeat-sample of $\mathbb{F}_N$ with replacement gives information on distribution of $\widehat{\boldsymbol{\beta}}^* = T(\mathbb{F}_N^*)$; asterisk denotes fixed $\mathbb{F}_N$.
- Let $\{\widehat{\boldsymbol{\beta}}_b^*\}_{b=1}^B$ denote the (bootstrap) samples.
- Note two layers of variation:
  - How well $\mathbb{F}_N$ approximates F (better as $N \nearrow \infty$ by Glivencko-Cantelli: $\sup_{t\in[0,1]} |F(t) - \mathbb{F}_N(t)| \longrightarrow_{\text{a.s.}} 0$).
  - How well $\{\widehat{\boldsymbol{\beta}}_b^*\}_{b=1}^B$ approximates $T(\mathbb{F}_N^*)$ (better as $B \nearrow \infty$).
- Which source of variation can we better control?

Estimator-attributed bias:

- Let $\widehat{\boldsymbol{\beta}}^*_b = T(F^*_{N:b})$ denote estimate based on b^{th} bootstrap sample. We may estimate bias as follows:

$$\begin{aligned}\widehat{\text{Bias}} &= \frac{1}{B}\sum_{b=1}^{B}(T(\mathbb{F}^*_{N:b}) - T(\mathbb{F}_N)) \\ &= \frac{1}{B}\sum_{b=1}^{B}\widehat{\boldsymbol{\beta}}^*_b - \widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^* - \widehat{\boldsymbol{\beta}} \approx \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}.\end{aligned}$$

- Note that $\widehat{\boldsymbol{\beta}}^* = \frac{1}{B}\sum_{b=1}^{B}\widehat{\boldsymbol{\beta}}^*_b$ for simplicity.
- Correction won't catch external sources of bias; be warned.

Covariance:

- We may estimate the covariance as well:

$$\widehat{\text{Cov}}\left(\widehat{\boldsymbol{\beta}}\right) = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}}^*)(\widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}}^*)^T$$

- For the k^{th} coefficient, we have:

$$\widehat{v}_k \quad = \quad \widehat{\text{Var}}(\widehat{\beta}_k) = \frac{1}{B}\sum_{b=1}^{B}([\widehat{\boldsymbol{\beta}}_b^*]_k - \widehat{\beta}_k^*)^2$$

Confidence intervals: Normal approximation (bias-correction)

- Symmetric $(1-\alpha)$ CI:

$$(\widehat{\beta}_k - \widehat{\text{Bias}}_k) \pm \sqrt{\widehat{v}_k} z_{1-\alpha/2}$$

- Assumptions:
  - $\widehat{\beta}_k - \beta_k \dot{\sim} \mathcal{N}(\text{Bias}_k, \sigma^2)$, which is symmetric and pivotal.
  - $\widehat{\text{Bias}}_k$ and $\widehat{v}_k$ are good estimates of Bias_k and σ^2.
- Good for cases where N is large enough that normal approximation holds, but no known theoretical formula for asymptotic variance.
- Can use QQ-plots to evaluate departures from normality.

Confidence intervals: Pivot based

- Let $\widehat{\beta}^*_{k(p)}$ denote p^{th} quantile of k^{th} coefficient of $\{\widehat{\beta}^*_k\}_{b=1}^B$.
- Behavior of $\beta_k - \widehat{\beta}_k$ approximately that of $\widehat{\beta}_k - \widehat{\beta}^*_k$:

$$
\begin{aligned}
0.95 &\approx \mathrm{P}\left(\widehat{\beta}^*_{k(\alpha/2)} \leq \widehat{\beta}^*_k \leq \widehat{\beta}^*_{k(1-\alpha/2)}\right) \\
&= \mathrm{P}\left(\widehat{\beta}_k - \widehat{\beta}^*_{k(1-\alpha/2)} \leq \widehat{\beta}_k - \widehat{\beta}^*_k \leq \widehat{\beta}_k - \widehat{\beta}^*_{k(\alpha/2)}\right) \\
&\approx \mathrm{P}\left(\widehat{\beta}_k - \widehat{\beta}^*_{k(1-\alpha/2)} \leq \beta_k - \widehat{\beta}_k \leq \widehat{\beta}_k - \widehat{\beta}^*_{k(\alpha/2)}\right) \\
&= \mathrm{P}\left(2\widehat{\beta}_k - \widehat{\beta}^*_{k(1-\alpha/2)} \leq \beta_k \leq 2\widehat{\beta}_k - \widehat{\beta}^*_{k(\alpha/2)}\right)
\end{aligned}
$$

- Assumptions:
  - $\widehat{\beta}_k - \beta_k$ asymptotically pivotal (not necessarily symmetric).

Confidence intervals:

- There are plenty of other of bootstrap-based confidence intervals. One simple one I did not cover is based on the quantiles of the bootstrap samples.
- The pivot-based confidence interval is generally understood to have better properties.
- See empirical process theory for all kinds of other generalizations, extensions, theoretical results.

Linear regression: Fixed design

- Re-sample residuals $\widehat{\epsilon}_i^*$ from the existing residuals $\{\widehat{\epsilon}_i\}_{i=1}^N$ with replacement.
- Keep $\mathbf{x}_i$ intact and form N new outcomes as $y_i^* = \mathbf{x}_i^T \widehat{\boldsymbol{\beta}} + \widehat{\epsilon}_i^*$ for $i = 1, \ldots, N$.
- Estimate $\widehat{\boldsymbol{\beta}}_b^*$ for $b = 1, \ldots, N$; form estimates/confidence intervals of your choosing from prior methods.
- Assumptions:
  - Homoscedasticity of errors.
  - Correct mean-model.
- Consistent with a designed experiment/randomized trial.
- If $\mathbf{X}$ is discrete, you can simply leave the $\mathbf{x}$'s as they are and resample the outcomes separately within subgroup of $\mathbf{X}$.

Linear regression: Random design

- Re-sample pairs $(\mathbf{x}_i^*, y_i^*)$ from existing observations $\{\mathbf{x}_i, y_i\}_{i=1}^N$ with replacement.
- Estimate $\widehat{\boldsymbol{\beta}}_b^*$ for $b = 1, \ldots, N$; form estimates/confidence intervals of your choosing from prior methods.
- Design changes with each sample.
- Consistent with an observational study.

Linear regression: Fixed vs. random design

- Assume homoscedastic errors.
- If the mean model is correct, either version of the bootstrap should perform well regardless of whether $\mathbf{X}$ is fixed by design or random.
- If $\mathbf{X}$ is fixed by design, mean-model misspecification will tend to result in an overstated variance if you treat $\mathbf{X}$ as random.
- If $\mathbf{X}$ is random by design, mean-model misspecification will tend to result in an understated variance if you treat $\mathbf{X}$ as fixed.

Stata: Example (MRI)

- `regress height age, robust` (recall)
- `regress height age, vce(bs, reps(500))`
- `regress height age, vce(bs, reps(500) nodots)`
- `estat bootstrap, all`

Stata: Example (MRI)

```
. regress height age, robust

Linear regression                               Number of obs     =        735
                                                F(1, 733)         =       9.21
                                                Prob > F          =     0.0025
                                                R-squared         =     0.0120
                                                Root MSE          =     9.6581
```

height	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.1953694	.0643711	-3.04	0.002	-.3217432	-.0689956
_cons	180.3453	4.805937	37.53	0.000	170.9103	189.7804

Stata: Example (MRI)

```
. regress height age, vce(bs, reps(500))
(running regress on estimation sample)

Bootstrap replications (500)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250
..................................................   300
..................................................   350
..................................................   400
..................................................   450
..................................................   500

Linear regression                               Number of obs     =        735
                                                Replications      =        500
                                                Wald chi2(1)      =       8.40
                                                Prob > chi2       =     0.0038
                                                R-squared         =     0.0120
                                                Adj R-squared     =     0.0107
                                                Root MSE          =     9.6581

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
      height |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.1953694   .0674101    -2.90   0.004    -.3274907   -.0632481
       _cons |   180.3453   5.000509    36.07   0.000     170.5445    190.1461
------------------------------------------------------------------------------
```

Stata: Example (MRI)

```
. regress height age, vce(bs, reps(500) nodots)

Linear regression                               Number of obs     =        735
                                                Replications      =        500
                                                Wald chi2(1)      =       8.97
                                                Prob > chi2       =     0.0027
                                                R-squared         =     0.0120
                                                Adj R-squared     =     0.0107
                                                Root MSE          =     9.6581

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
      height |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.1953694    .0652377    -2.99   0.003     -.323233   -.0675058
       _cons |   180.3453    4.874817    37.00   0.000     170.7909    189.8998
------------------------------------------------------------------------------
```

Stata: Example (MRI)

```
. estat bootstrap, all

Linear regression                               Number of obs     =        735
                                                Replications      =        500
```

height	Observed Coef.	Bias	Bootstrap Std. Err.	[95% Conf.	Interval]	
age	-.19536938	-.0014101	.06523773	-.323233	-.0675058	(N)
				-.3367485	-.0664426	(P)
				-.3296939	-.0654481	(BC)
_cons	180.34533	.1100677	4.8748171	170.7909	189.8998	(N)
				170.8138	190.7536	(P)
				170.6618	190.2488	(BC)

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

Stata: Example (MRI)

- `N`: Normal CI
- `P`: Percentile CI
- `BC`: Bias-corrected CI

The bootstrap

Notes:

- There is plenty more to say about the bootstrap.
- You'll have to take advanced regression courses to learn more. Or study empirical process theory if you want to learn it from that angle :).