Bootstrap regression: alternative approach

In an alternative formulation of the bootstrap for linear regression, consider

$$y_i = \boldsymbol{x}_i^\mathsf{T} \beta + \epsilon_i$$

where $\epsilon_i \sim F(.)$, where F is constrained to have mean zero.

If the distribution F is regarded as unknown, but the form of the linear regression is satisfactory (up to unknown parameters), attention can focus on bootstrapping the residual errors $\epsilon_1, \ldots, \epsilon_n$.

Bootstrap regression: alternative approach

In this β can be estimated by least-squares

$$\widehat{\beta} = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{y}$$

and then the residual cdf F estimated by the empirical cdf of the residuals

$$e_i = y_i - \boldsymbol{x}_i^\intercal \widehat{\beta} \qquad i = 1, \ldots, n$$

denoted $\widehat{F}_n$.

The bootstrap data $y_1^{(b)}, \ldots, y_n^{(b)}$ are then computed by producing a bootstrap resample $e_1^{(b)}, \ldots, e_n^{(b)}$, and setting

$$y_i^{(b)} = \boldsymbol{x}_i^\intercal \beta + e_i^{(b)}$$

Bootstrap regression: alternative approach

This then yields the bootstrap estimate

$$\widehat{\beta}^{(b)} = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{y}^{(b)}.$$

In this case, the idealized bootstrap estimate ($B = \infty$) of the standard error is available analytically via

$$\mathbb{V}ar[\widehat{\beta}^{(b)}] = \widehat{\sigma}^2 (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}$$

where

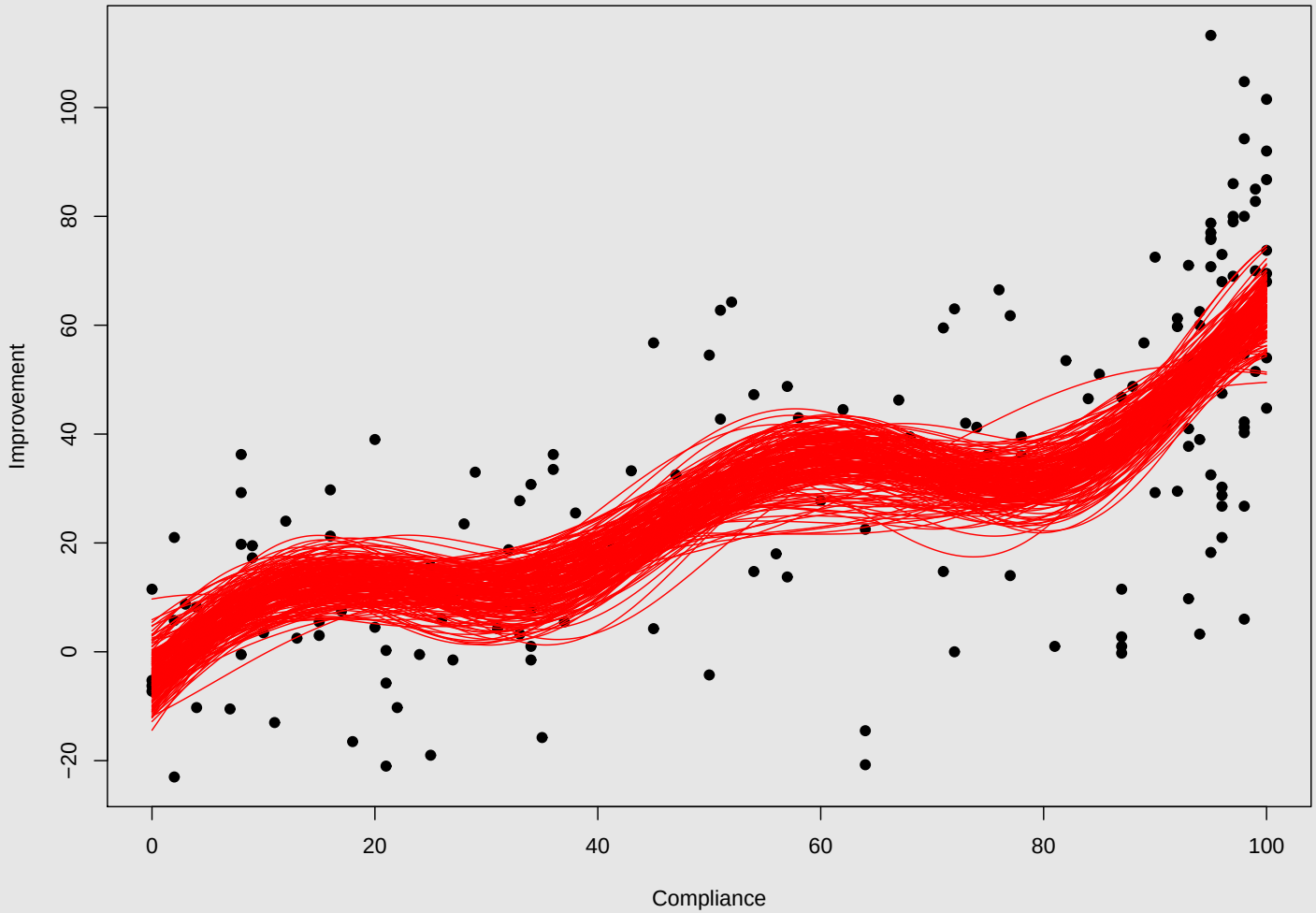
$$\widehat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} e_i^2$$

that is, identical to the usual variance in least-squares.

Bootstrap regression: alternative approach

Example (Cholostyramine Data)

$B = 200$ harmonic regressions with bootstrapped residuals.



Estimating Bias

The bias of an estimator T_n of $\theta(F)$ is

$$b_{T_n}(F) = \mathbb{E}_F[T_n] - \theta(F)$$

which can be estimated by

$$\widehat{b}_{T_n}(\widehat{F}_n) = \mathbb{E}_{\widehat{F}_n}[T_n^{(*)}] - \theta(\widehat{F}_n)$$

where $T_n^{(*)}$ is the bootstrap estimator, derived from bootstrap resamples

$$t_n^{(1)}, \ldots, t_n^{(B)}$$

and $\theta(\widehat{F}_n)$ is the plug-in estimate. Typically $\mathbb{E}_{\widehat{F}_n}[T_n^{(*)}]$ is approximated by

$$\frac{1}{B}\sum_{b=1}^{B} t_n^{(b)}$$

Estimating Bias

If $T_n = \theta(\widehat{F}_n)$, that is, the estimator is the plug-in estimator, then an alternative estimate of bias can be obtained for $B = \infty$.

Denote by $P_j^{(b)}, j = 1, \ldots, n$ denote the proportion of a bootstrap resampled data set $y_1^{(b)}, \ldots, y_n^{(b)}$ that are equal to y_j.

$$P_j^{(b)} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{y_j\}}(y_i^{(b)})$$

The (random) probability distribution $P^{(b)} = (P_1^{(b)}, \ldots, P_n^{(b)})$ (the *resampling vector*) can be used to construct the bootstrap estimate in terms of the original data; for example,

$$\overline{y}^{(b)} = \sum_{j=1}^{n} P_j^{(b)} y_j.$$

Estimating Bias

The bootstrap resampling vectors are thus

$$P^{(1)}, \ldots, P^{(B)}.$$

In general, denote the bootstrap estimator in terms of the

$$T_n^{(b)} = T_n(P^{(b)})$$

Let $P_0 = \frac{1}{n}\mathbf{1}_n$ denote the vector of length n containing identical elements $1/n$. Then, in this notation

$$T_n(P_0) = \theta(\widehat{F}_n)$$

Estimating Bias

Let

$$\overline{P}^{(*)} = \frac{1}{B}\sum_{b=1}^{B} P^{(b)}$$

This allows a new bias estimate

$$\widetilde{b}_{T_n}(\widehat{F}_n) = \frac{1}{B}\sum_{b=1}^{B} t_n^{(b)} - T_n(\overline{P}^{(*)})$$

to be compared with the plug-in estimate

$$\widehat{b}_{T_n}(\widehat{F}_n) = \frac{1}{B}\sum_{b=1}^{B} t_n^{(b)} - T_n(P_0)$$

Estimating Bias

Example (Hormone patch data)

Eight subjects were treated with two different hormone patches, and the subsequent level of hormone in the bloodstream measured.

ID	Placebo	Old Patch	New Patch	old-plac. z	new - old y
1	9243	17649	16449	8406	-1200
2	9671	12013	14614	2342	2601
3	11792	19979	17274	8187	-2705
4	13357	21816	23798	8459	1982
5	9055	13850	12560	4795	-1290
6	6290	9806	10157	3516	351
7	12412	17208	16570	4796	-638
8	18806	29044	26325	10238	-2719

Estimating Bias

Example (Hormone patch data)

The parameter of interest here is

$$\theta = \frac{\mathbb{E}[Y]}{\mathbb{E}[Z]}$$

which is a measure of bio-equivalence. The plug-in estimate is

$$t_n = \theta(\widehat{F}_n) = \frac{\overline{y}}{\overline{z}} = -0.0713.$$

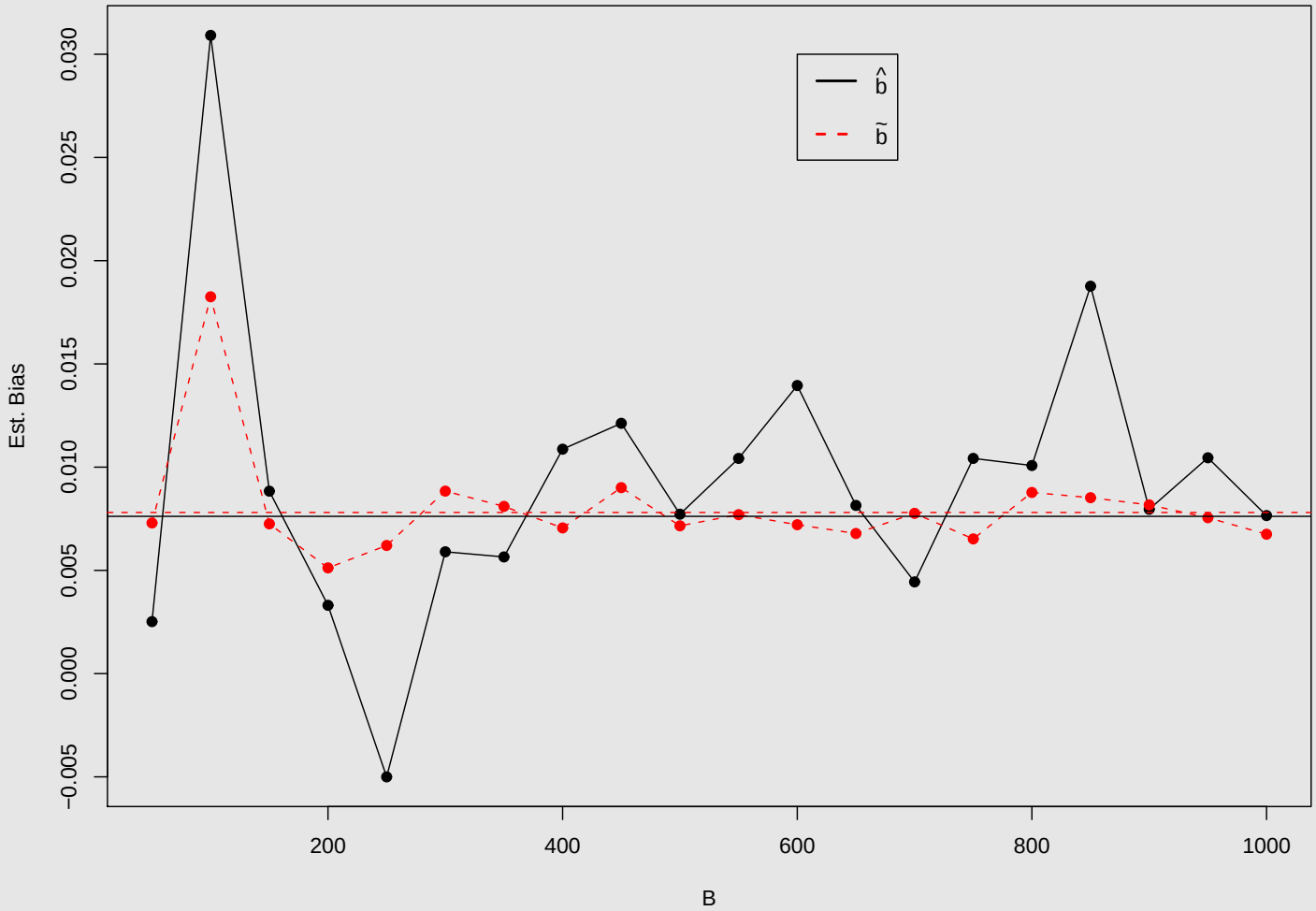
This parameter can be studied using the bootstrap, resampling individual subjects as the i.i.d. units.

For different values of B, we study the bias estimates given by the previous formulae.

Estimating Bias

Example (Hormone patch data)

Bias estimates $\widehat{b}$ and $\widetilde{b}$ for B up to 1000.



Estimating Bias

Example (Hormone patch data)

The bias estimate $\widetilde{b}$ is much less variable than $\widehat{b}$.

The two bias estimates agree approximately when $B = 100000$

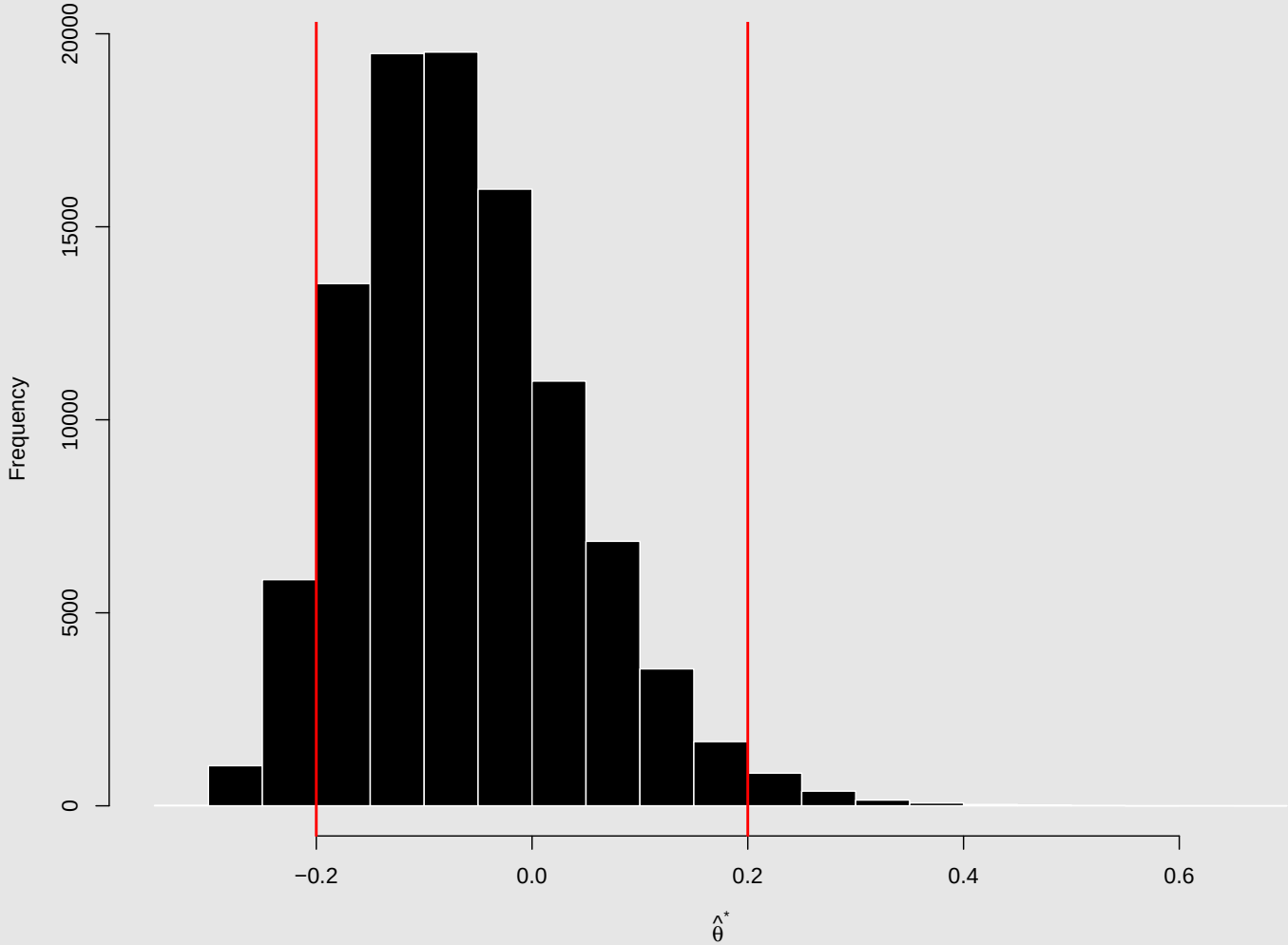
$$\widehat{b}_{T_n}(\widehat{F}_n) = 0.00762 \qquad \widetilde{b}_{T_n}(\widehat{F}_n) = 0.00781$$

but for B up to 1000, $\widetilde{b}$ is very stable, whereas $\widehat{b}$ fluctuates considerably.

Estimating Bias

Example (Hormone patch data)

Sampling distribution of $\widehat{\theta}$: $\mathbb{P}[|\widehat{\theta}| < 0.2] \simeq 0.9158$.



The jackknife

The *jackknife* was introduced as another bias estimation procedure. For $i = 1, \ldots, n$, let

$$\boldsymbol{y}_{(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$$

be the *i-deleted* data set (with the ith datum deleted). Let

$$\widehat{\theta}_{(i)} = t_{n-1}(\boldsymbol{y}_{(i)})$$

be the estimate derived from the i-deleted data, and let

$$\widehat{\theta}_{(.)} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\theta}_{(i)}$$

A bias estimate for a plug-in statistic $\widehat{\theta}$ is then

$$\widehat{b}_{\text{JACK}} = (n-1)(\widehat{\theta}_{(.)} - \widehat{\theta})$$

The jackknife

In this calculation

$$\widehat{\theta} = \theta(\widehat{F}_n) \qquad\qquad \widehat{\theta}_{(i)} = \theta(\widehat{F}_{n(i)})$$

are the plug-in estimators derived from the full and i-deleted data respectively.

A jackknife estimate for the standard error of a statistic is

$$\widehat{\text{se}}_{\text{JACK}} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\widehat{\theta}_{(i)} - \widehat{\theta}_{(.)})^2}$$

This formula applies for *smooth* functions t of the data.

Note: There is no resampling here, so precisely n computations are needed.

The jackknife

Formally, the jackknife is an approximation to the bootstrap that arises from a linear approximation to the statistic of interest, derived from a Taylor expansion.

The above bias/s.e. formulae are derived under the assumption that the identity

$$\widehat{\theta} = t_n(\boldsymbol{y}) = \mu + \frac{1}{n}\sum_{i=1}^{n} \alpha(y_i)$$

holds exactly from some function $\alpha(.)$. The jackknife arises from regarding a first-order Taylor **approximation** for arbitrary t_n.

The jackknife procedure fails when such an expansion fails; for example, for studying the **median**, the jackknife provides an *inconsistent* estimator.