The Bootstrap: Further Examples

In the case of i.i.d. univariate sampling, the bootstrap extends easily to more general settings. For example, extension to the i.i.d. vector case is straightforward.

1. A two-sample problem.
2. The Pearson correlation coefficient.
3. Principal components analysis.

The Bootstrap: Two-sample Estimation

Suppose that two random samples are available from different distributions

$$\begin{aligned} F_1 &: y_{11}, \ldots, y_{1n_1} \\ F_2 &: y_{21}, \ldots, y_{2n_2} \end{aligned}$$

and the parameter of interest is the difference in means

$$\mu_1 - \mu_2.$$

In the normal case, under equal variances for the two populations, the standard error of the estimator

$$\overline{y}_1 - \overline{y}_2$$

can be studied analytically; more generally, outside of the normal case and for other estimands, it cannot.

The Bootstrap: Two-sample Estimation

The bootstrap operates in this case by

- resampling $y_{11}^{(b)}, \ldots, y_{1n_1}^{(b)}$ (size n_1) from the first sample,
- resampling $y_{21}^{(b)}, \ldots, y_{2n_2}^{(b)}$ (size n_2) from the second sample,
- forming the difference

$$\overline{y}_1^{(b)} - \overline{y}_2^{(b)}$$

for $b = 1, \ldots, B$. Other estimators can be studied in this way.

The Bootstrap: Pearson correlation

Suppose that pairs of observations $(x_i, y_i), i = 1, \ldots, n$ are available. The Pearson product-moment correlation is

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \overline{xy})}{\sqrt{\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right\}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2\right\}}}$$

which measures the linear correlation between the two variables in the sample.

r is the moment-based estimator of the population correlation ρ.

The Bootstrap: Pearson correlation

The bootstrap operates in this case by

- resampling **pairs** $((x_1^{(b)}, y_1^{(b)}), \ldots, (x_n^{(b)}, y_n^{(b)}))$ (size n)
- forming the resampled value of r

$$r^{(b)}$$

for $b = 1, \ldots, B$.

Note that again it is the i.i.d. units that are being resampled.

The Bootstrap: Pearson correlation

Example (LSAT vs GPA)

The relationship between GPA (x) and LSAT score (y) in $n = 15$ law school students is to be studied.

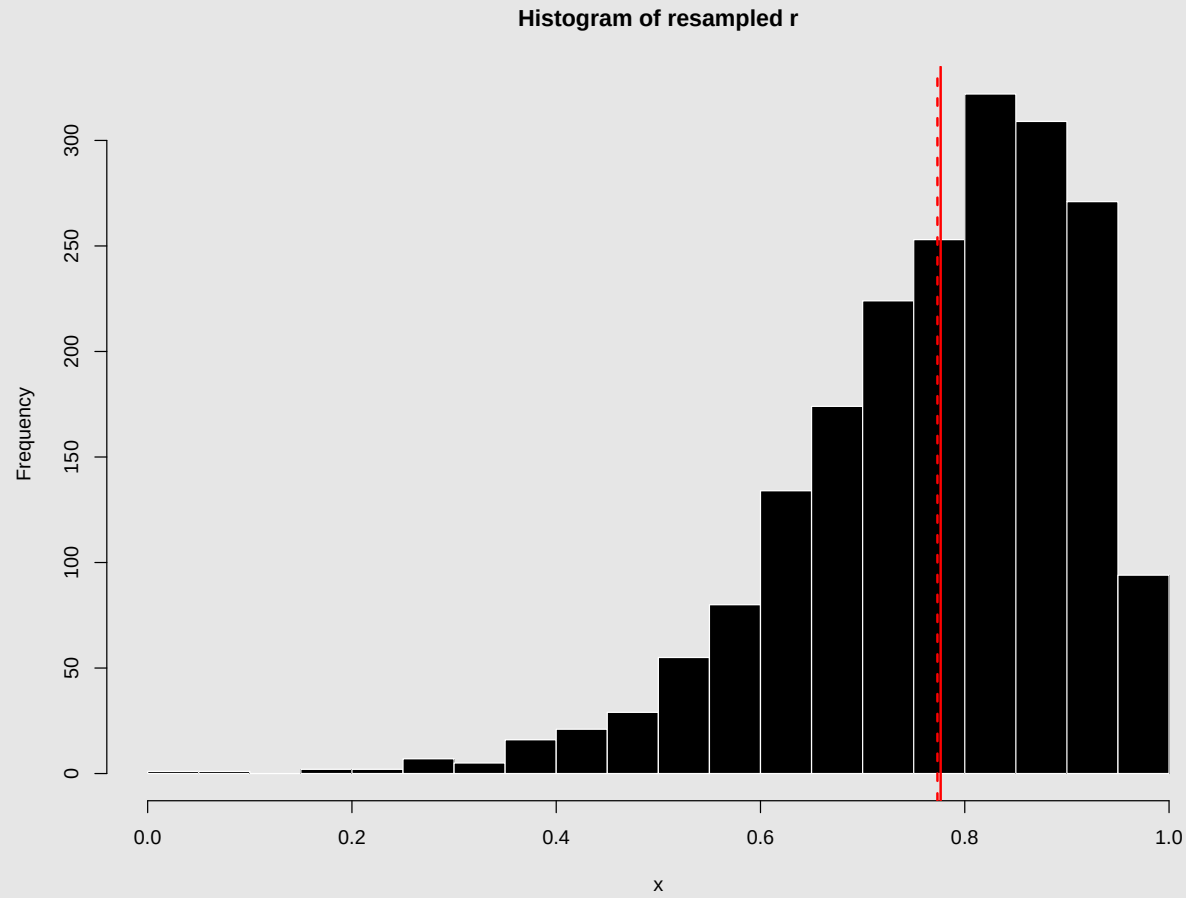
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
GPA	339	330	281	303	344	307	300	343	336	313	312	274	276	288	296
LSAT	576	635	558	578	666	580	555	661	651	605	653	575	545	572	594

For these data $r = 0.776$.

The Bootstrap: Pearson correlation

Example (LSAT vs GPA)

2000 (resampled) sample correlation values, with standard deviation 0.1382.



The Bootstrap: Pearson correlation

Example (LSAT vs GPA)

A common transformation to approximate normality (for large n) uses

$$Z_n = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$$

for which

$$\sqrt{n-3}Z_n \xrightarrow{d} Z \sim \mathcal{N}(\mu, 1)$$

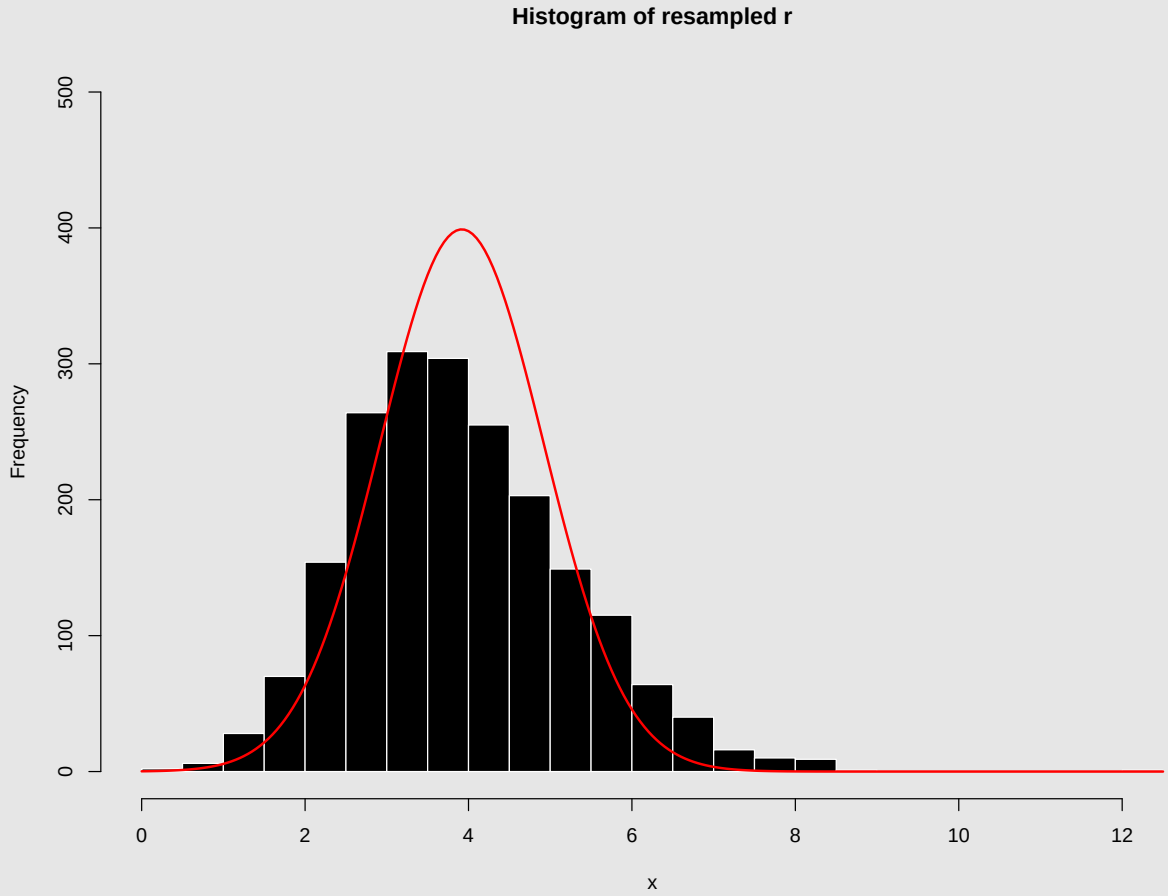
where

$$\mu = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$$

The Bootstrap: Pearson correlation

Example (LSAT vs GPA)

2000 Z_n values, and normal approximation: skewness is apparent, and the bootstrap variance is 1.793, rather than 1.



The Bootstrap: Principal components

Suppose that vectors of observations of length p, $\boldsymbol{x}_i, i = 1, \ldots, n$, are available. Let X denote the $n \times p$ matrix formed by row-binding the n i.i.d. vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, after *column standardization*

- each column of X has its mean subtracted.

Then the sample covariance matrix is

$$S = \frac{1}{n} X^\intercal X$$

This is a positive-definite matrix, that can be decomposed using the eigen-decomposition

$$S = VDV^\intercal$$

where D is the diagonal matrix of eigenvalues, and V is the corresponding eigenvector matrix with orthonormal columns.

The Bootstrap: Principal components

The diagonal elements of D are the sample eigenvalues

$$\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_p$$

and the scaled quantities

$$\widehat{\theta}_j = \frac{\widehat{\lambda}_j}{\sum_{l=1}^{p} \widehat{\lambda}_l} \qquad j = 1, \ldots, p.$$

is an estimate of the proportion of variation explained by the first eigenvector.

The statistical properties of $\widehat{\theta}_j$ are hard to study analytically.

The Bootstrap: Pearson correlation

The bootstrap operates in this case by

- resampling **vectors** $(\boldsymbol{x}_1^{(b)}, \ldots, \boldsymbol{x}_n^{(b)})$ (size n)
- forming the resampled matrix $X^{(b)}$,
- forming the resampled covariance matrix $S^{(b)}$,
- computing eigenvalues $\widehat{\lambda}_1^{(b)}, \ldots, \widehat{\lambda}_p^{(b)}$ and scaling

for $b = 1, \ldots, B$.

Note that again it is the i.i.d. units that are being resampled.

The Bootstrap: Principal components

Example (Education test scores)

88 students each took five tests on mechanics, vectors, algebra, analysis and statistics. This yielded the S matrix

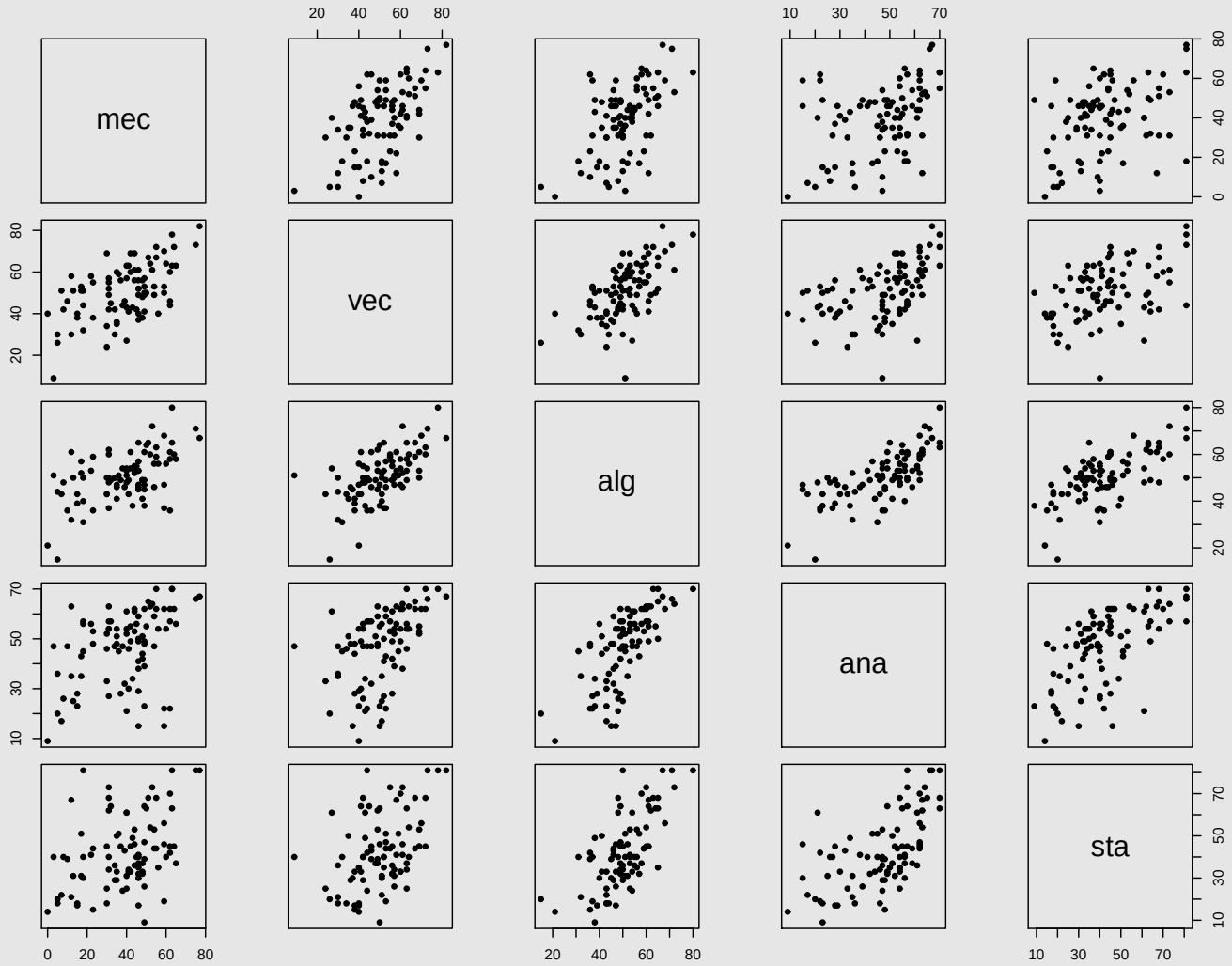
$$S = \begin{bmatrix} 302.29 & 125.78 & 100.43 & 105.07 & 116.07 \\ 125.78 & 170.88 & 84.19 & 93.60 & 97.89 \\ 100.43 & 84.19 & 111.60 & 110.84 & 120.49 \\ 105.07 & 93.60 & 110.84 & 217.88 & 153.77 \\ 116.07 & 97.89 & 120.49 & 153.77 & 294.37 \end{bmatrix}$$

and eigenvalues

$$\widehat{\boldsymbol{\lambda}} = (679.18, 199.81, 102.57, 83.67, 31.79)$$

The Bootstrap: Principal components

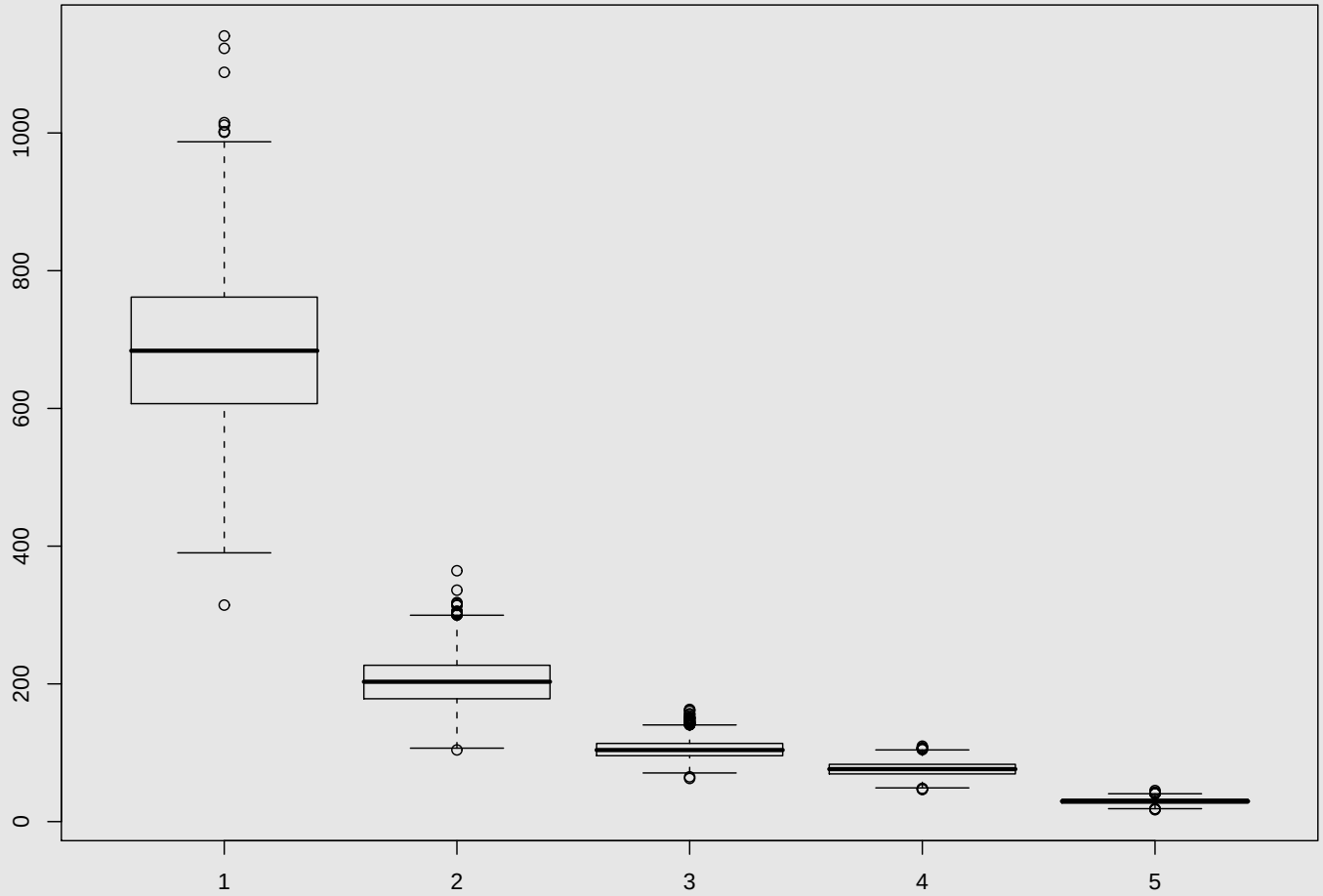
Example (Education test scores)



The Bootstrap: Principal components

Example (Education test scores)

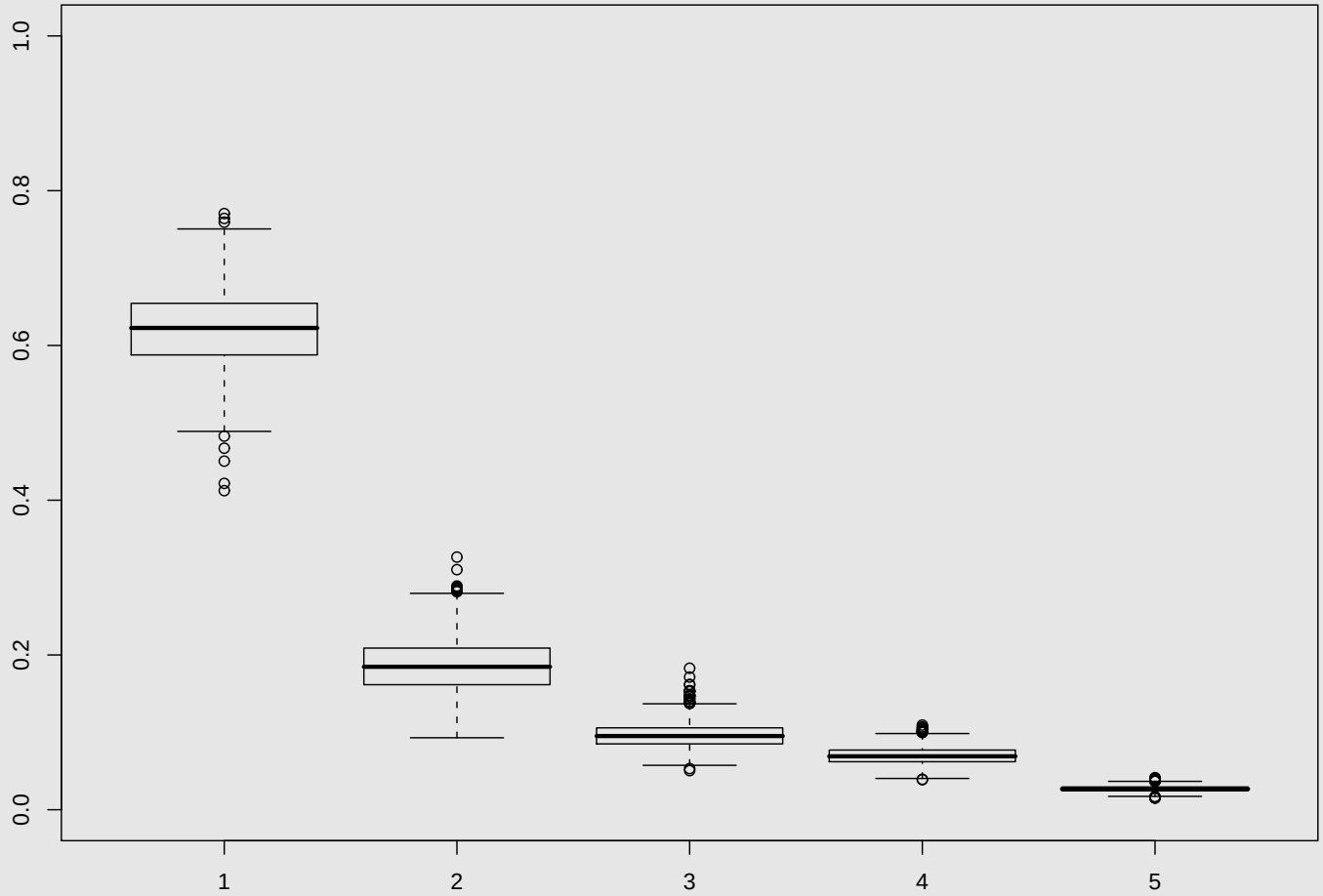
2000 resampled eigenvalues, $\lambda_j, j = 1, \ldots, 5$



The Bootstrap: Principal components

Example (Education test scores)

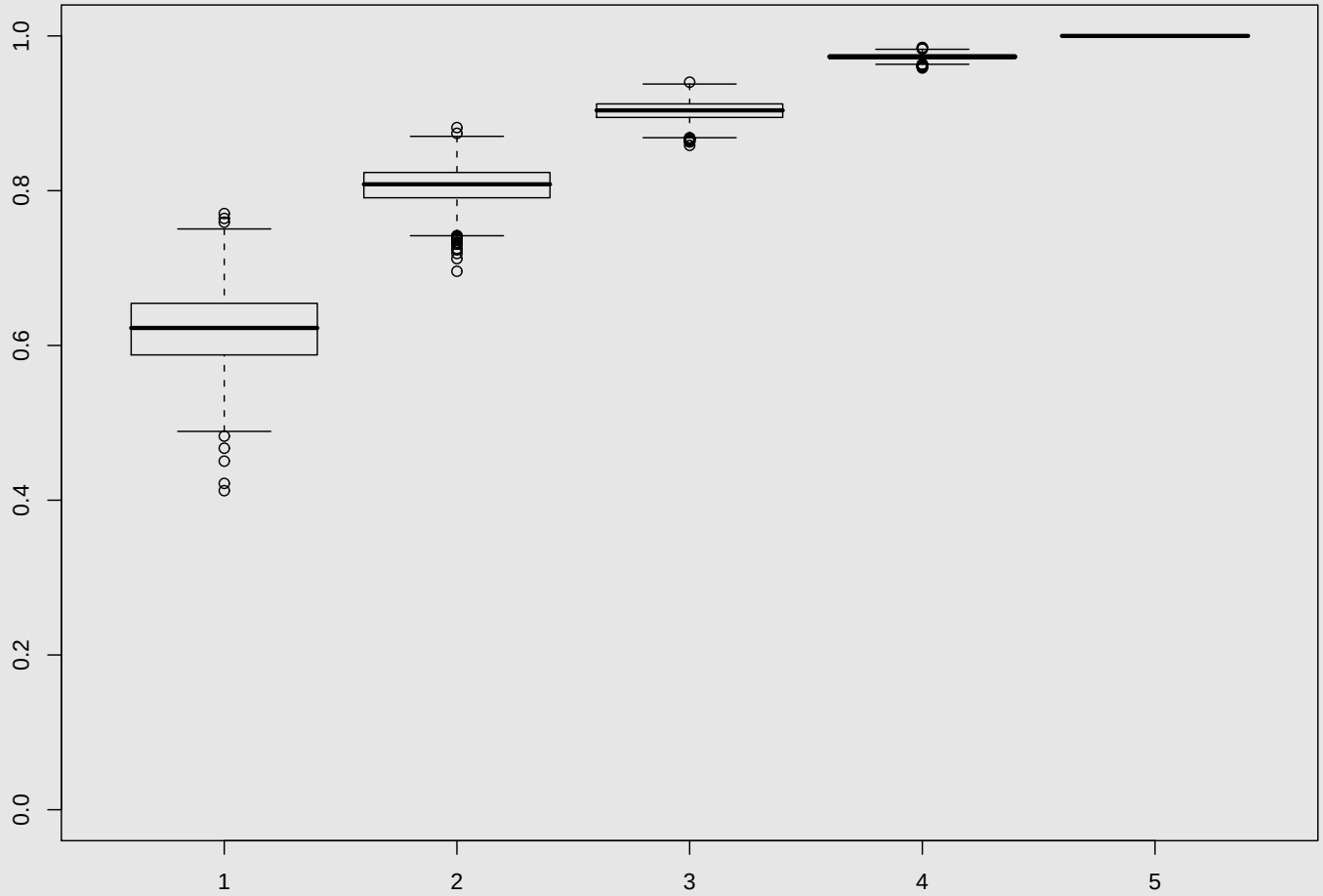
2000 resampled scaled eigenvalues, $\theta_j, j = 1, \ldots, 5$



The Bootstrap: Principal components

Example (Education test scores)

2000 resampled scaled eigenvalues, cumulative:



The Bootstrap: Principal components

Example (Education test scores)

We might conclude from this that really there are only two (orthogonal) dimensions of variation (rather than five) that explain the majority variation in the response.

The *principal components* are the vectors

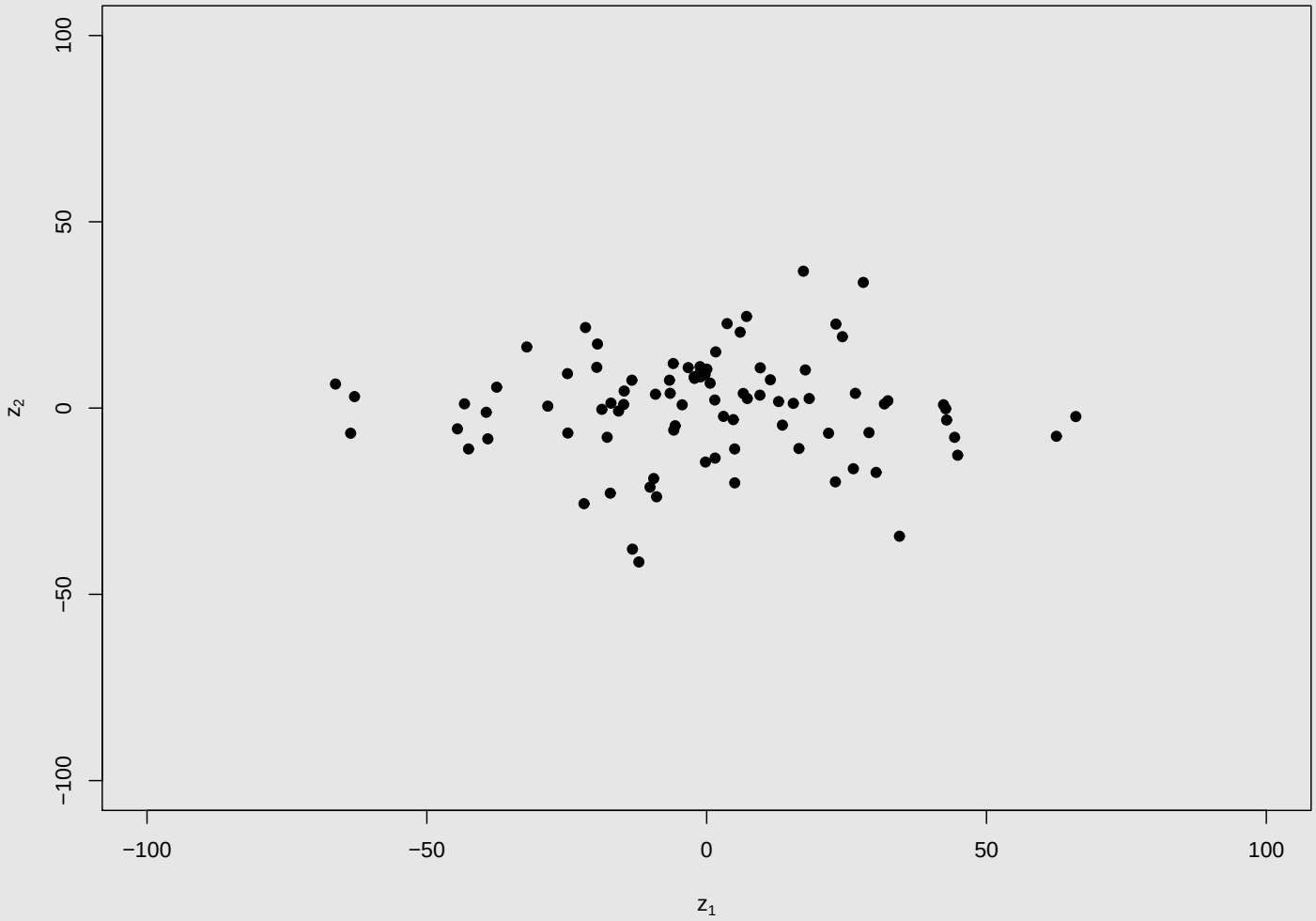
$$z_j = Xv_j \qquad j = 1, \ldots, p$$

where v_j is the jth column of V. The first two principal components carry greater than 80% of the variation observed.

The Bootstrap: Principal components

Example (Education test scores)

Principal components



The Bootstrap: Regression

The bootstrap can also be applied in a regression setting. Suppose pairs of data $(x_i, y_i), i = 1, \ldots, n$ are available, and the relationship is modelled as

$$y_i = g(x_i; \beta) + \epsilon_i \qquad \mathbb{E}[\epsilon_i | x] = 0$$

for some regression function g.

If g is taken to be linear in parameters β,

$$g(x; \beta) = \beta_0 + \sum_{j=1}^{p} \beta_j g_j(x)$$

and the residual errors are presumed Gaussian, then estimation can be carried out using least-squares in the usual way.

In this case, *pointwise* confidence intervals are easy to obtain. However, *simultaneous* confidence intervals are more complicated to compute.

The Bootstrap: Regression

In *harmonic regression*, the g_js are taken to be sin and cos functions. Assuming $p = 2q$ say, and $j = 1, \ldots, p$

$$g_j(x) = \begin{cases} \cos(2\pi x \phi_l) & j = 2l - 1 \\ \sin(2\pi x \phi_l) & j = 2l \end{cases}$$

for known constants $\phi_1, \ldots, \phi_q$.

For example, a typical choice is $q = 2$, with

$$\phi_1 = \frac{1}{2} \qquad \phi_2 = \frac{1}{4}$$

The Bootstrap: Regression

In such a model, the ML estimator is

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

where $\boldsymbol{X}$ is the $n \times (p+1)$ matrix with ith row

$$[1, \cos(2\pi x_i \phi_1), \sin(2\pi x_i \phi_1), \cos(2\pi x_i \phi_2), \sin(2\pi x_i \phi_2)]$$

The variance of $\widehat{\beta}$ is

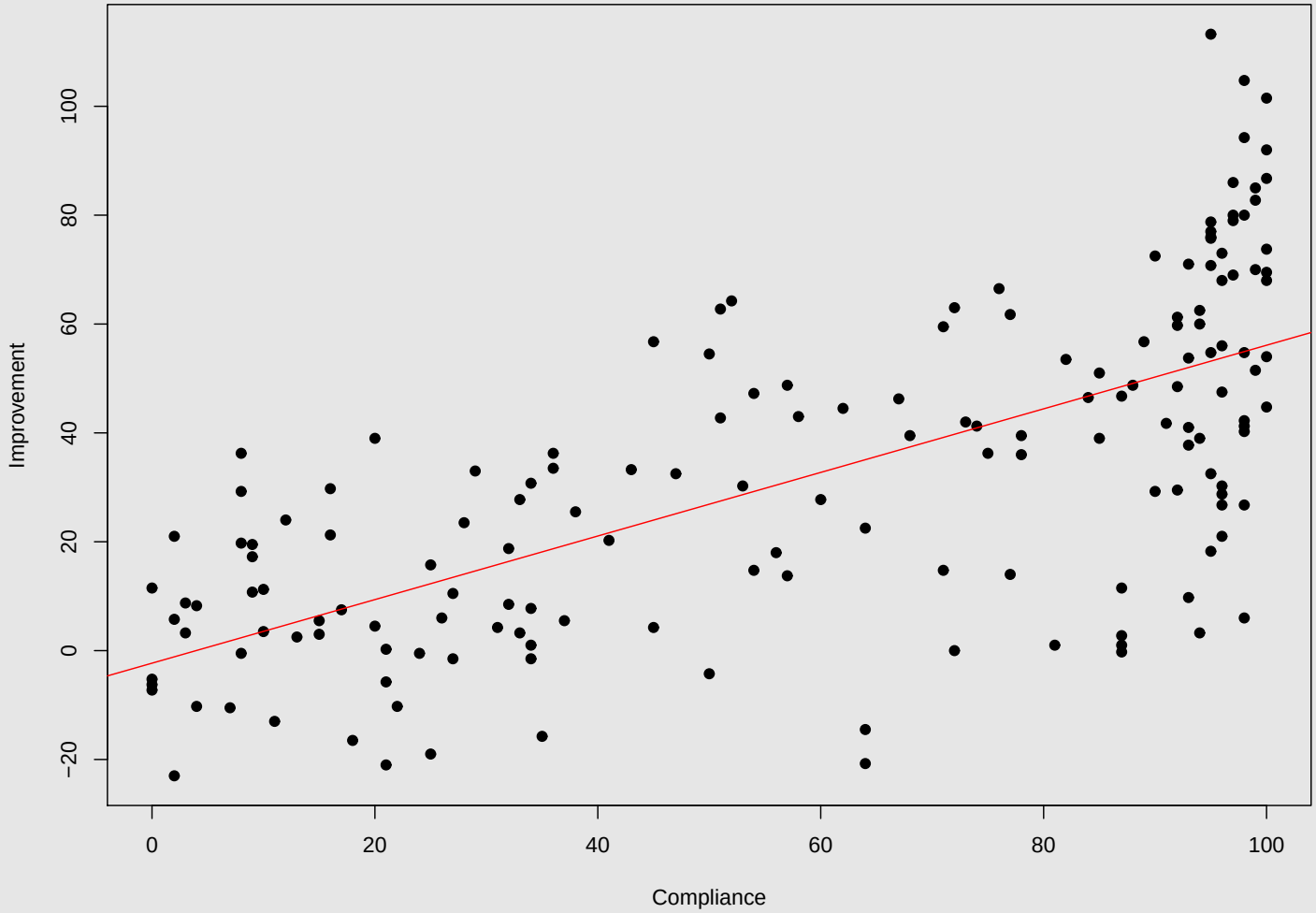
$$\widehat{\sigma}^2 (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}$$

and pointwise confidence intervals can be obtained in the usual way.

The Bootstrap: Regression

Example (Cholostyramine Data)

$n = 164$ men took part in an experiment to see if the drug cholostyramine lowered blood cholesterol levels.



The Bootstrap: Regression

Example (Cholostyramine Data)

Consider the model

$$g(x;\beta) = \beta_{00} + \beta_{01}x + \sum_{j=1}^{q}\left[\beta_{j0}\cos(2\pi x\phi_l) + \beta_{j1}\sin(2\pi x\phi_l)\right]$$

on the range $[0, 100]$ with

$$\phi_1 = \frac{1}{50} \qquad \phi_2 = \frac{1}{66} \qquad \phi_3 = \frac{1}{75}$$

and the three fits obtained taking

- $q = 0$
- $q = 1$
- $q = 2$

The Bootstrap: Regression

The bootstrap operates in this case by

- resampling **pairs** $((x_1^{(b)}, y_1^{(b)}), \ldots, (x_n^{(b)}, y_n^{(b)}))$ (size n) and forming the boostrapped design matrix $\boldsymbol{X}_{(b)}$.
- fitting the regression to the resampled data obtaining

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

- recording the fitted values

$$\hat{y}^{(b)} = \boldsymbol{X}_{(b)}\widehat{\beta}^{(b)}$$

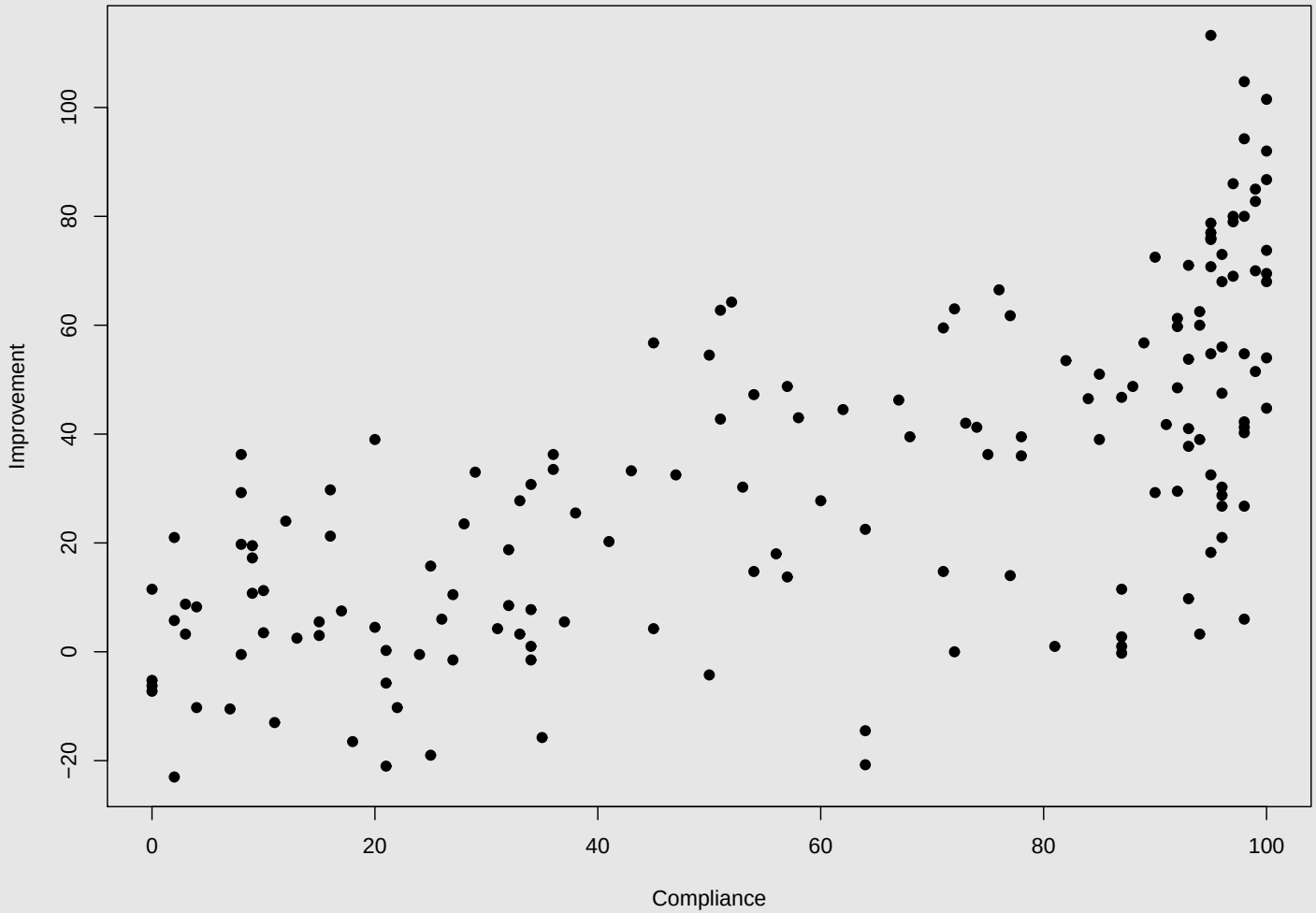
for $b = 1, \ldots, B$.

Again it is the i.i.d. units that are being resampled.

The Bootstrap: Regression

Example (Cholostyramine Data)

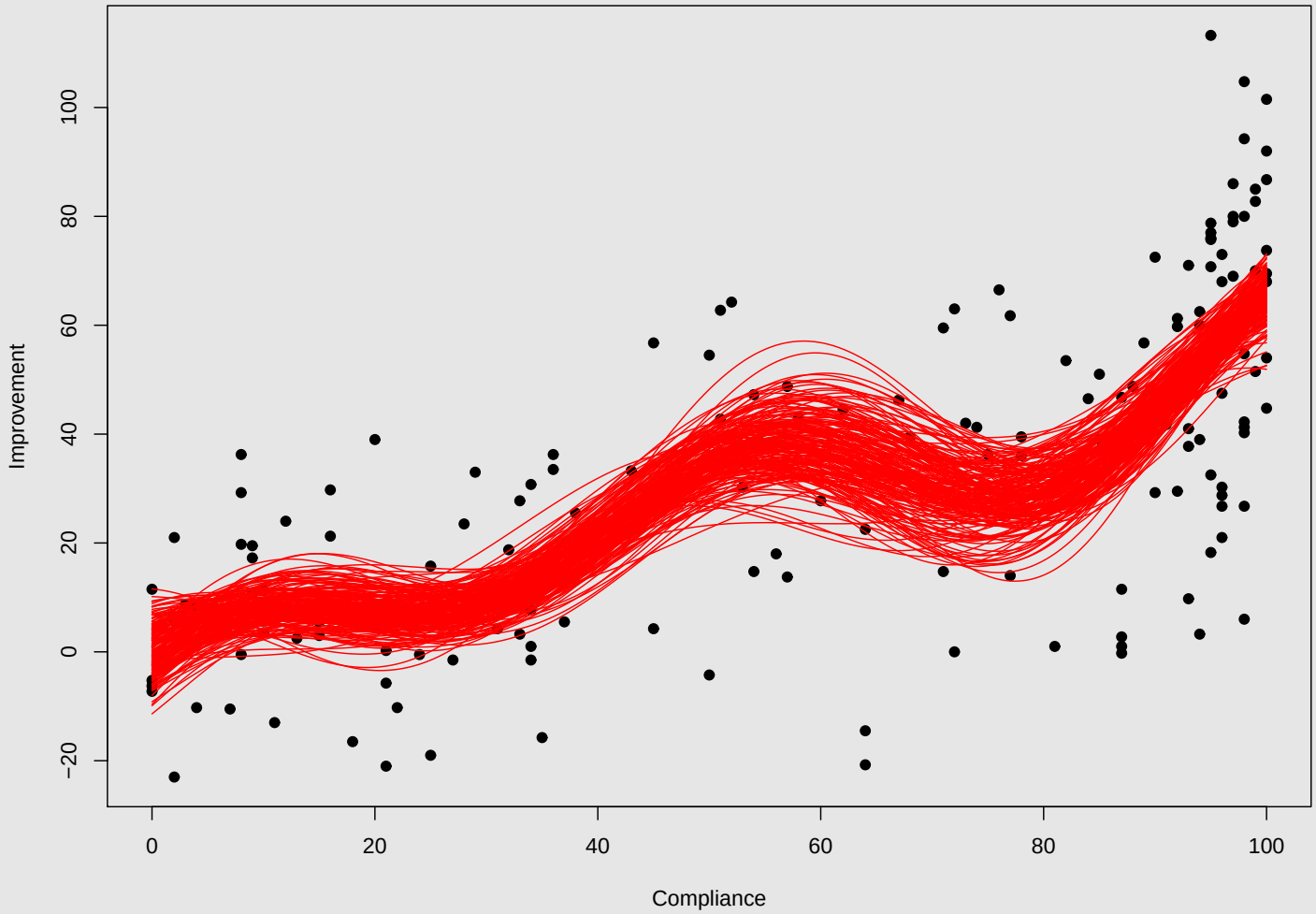
$B = 200$ bootstrapped curves with $q = 2$.



The Bootstrap: Regression

Example (Cholostyramine Data)

$B = 200$ bootstrapped curves with $q = 2$.



The Bootstrap: Loess

A more flexible model is based on *local smoothing* using *loess*.

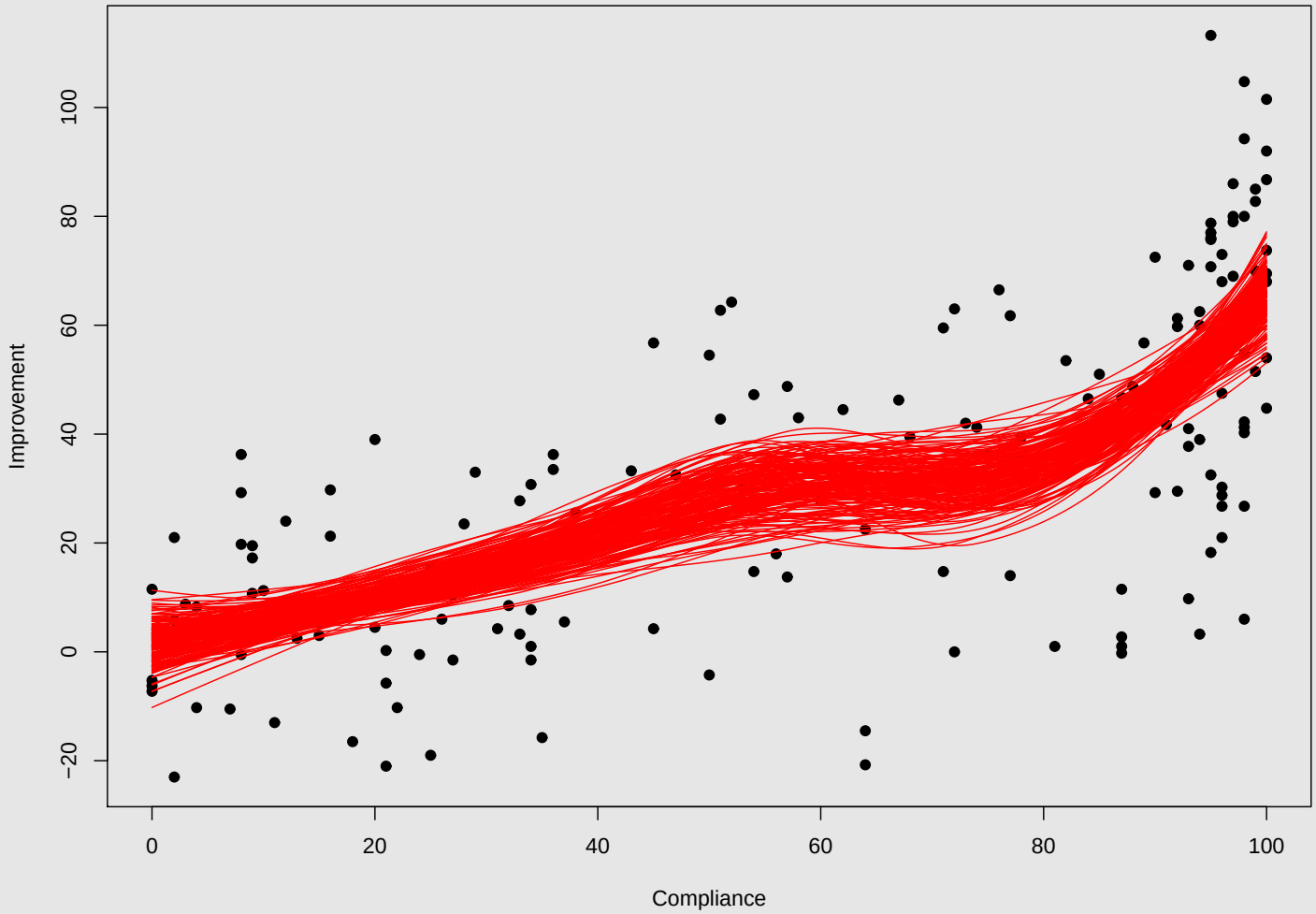
The loess approach fits a series of weighted local linear regression models restricted to a rolling neighbourhood of data points. The size of the neighbourhood (the *span*) determines the smoothness of the resulting fitted curve.

The statistical properties of the resulting fitted curve are hard to study analytically, but can be studied by the bootstrap.

The Bootstrap: Loess

Example (Cholostyramine Data)

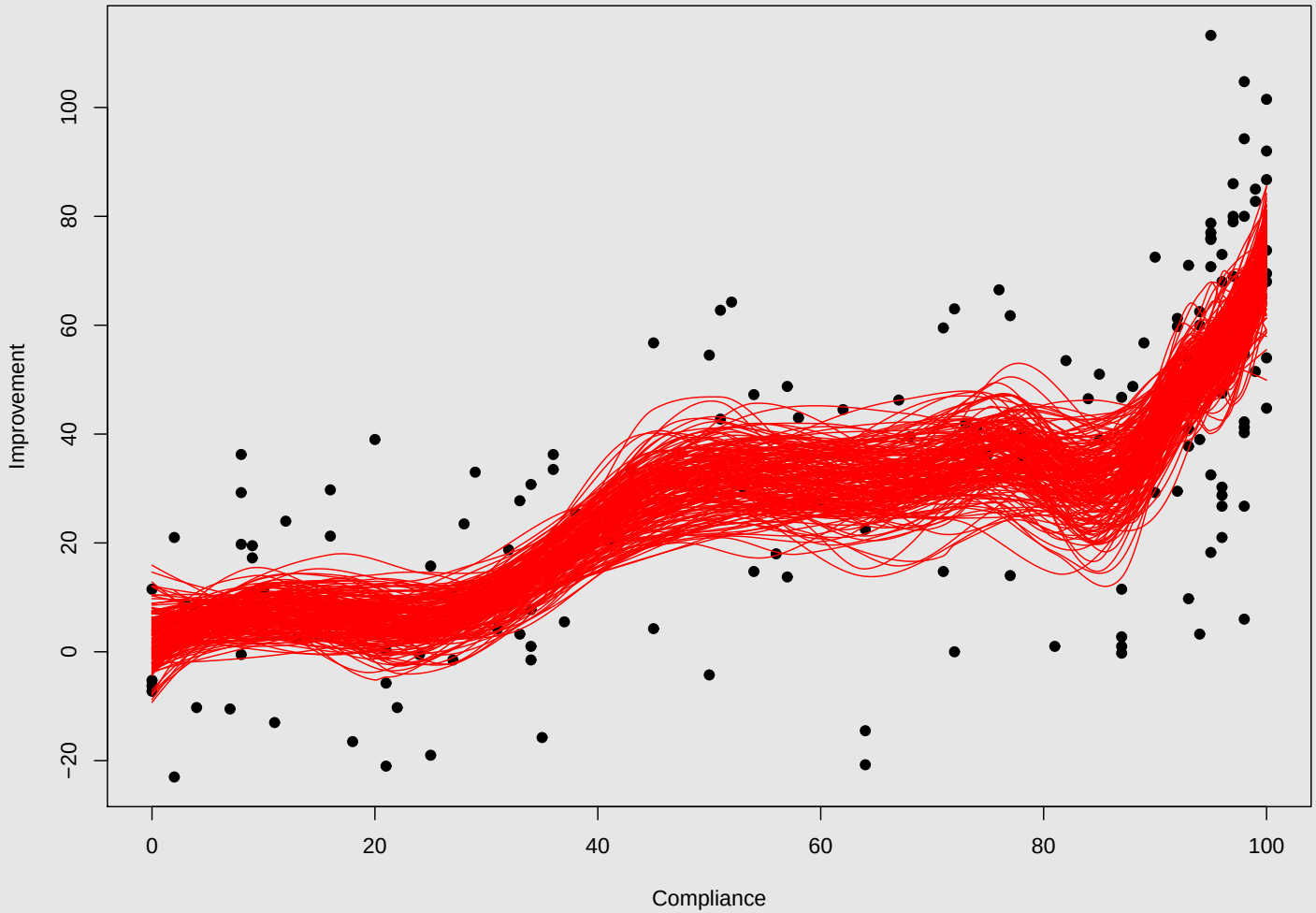
$B = 200$ bootstrapped loess curves with $q = 2$: span is 0.75.



The Bootstrap: Loess

Example (Cholostyramine Data)

$B = 200$ bootstrapped loess curves with $q = 2$: span is 0.375.



The Bootstrap: Loess

The choice of the span can be determined by K-fold *cross-validation*:

1. split data at random into K equal portions of size m.
2. fit the loess model using $K-1$ *training* portions using a fixed value of the span.
3. use the fitted model to predict the values in the remaining *test* portion; record the prediction mean square error

$$\frac{1}{m}\sum_{i=1}^{m}(\widehat{y}_i - y_i)^2$$

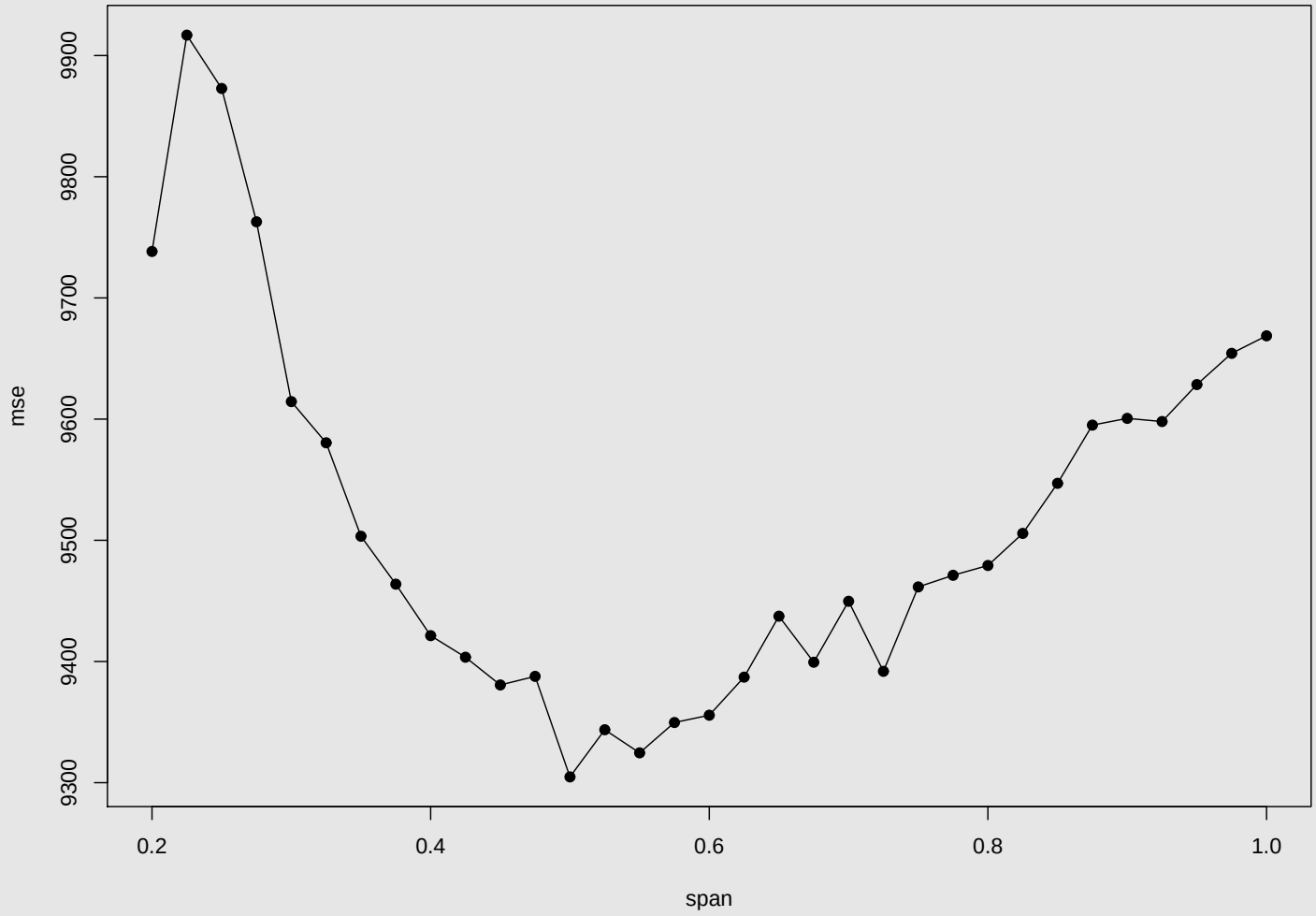
4. repeat 1. - 3. over different random splits, taking the average prediction MSE.

and repeat this whole process for different values of the span, recording the MSE every time. Choose the optimal span to minimize the prediction MSE.

The Bootstrap: Loess

Example (Cholostyramine Data)

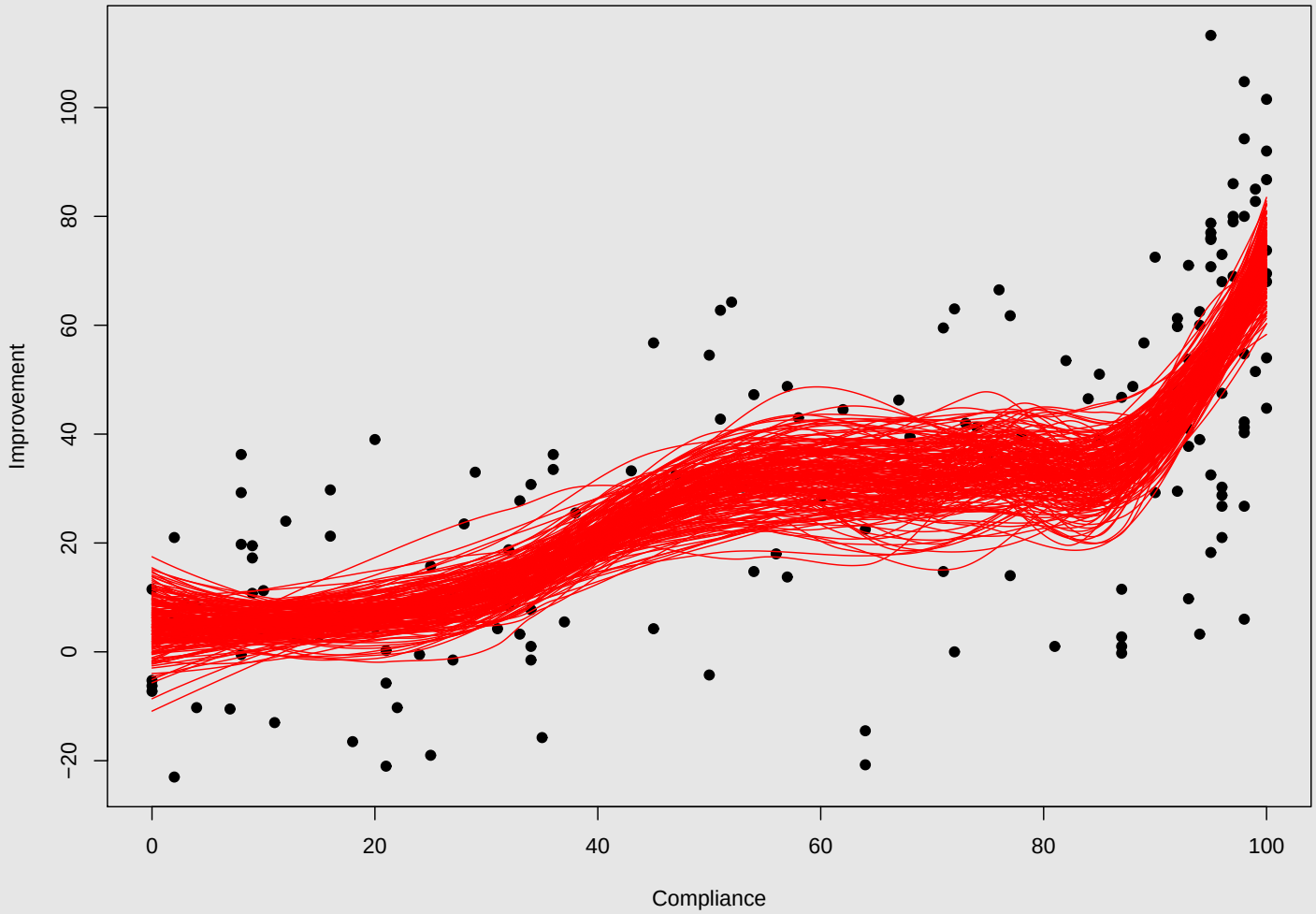
8-fold cross-validation: average prediction MSE from 20000 random splits.



The Bootstrap: Loess

Example (Cholostyramine Data)

$B = 200$ bootstrapped loess curves with $q = 2$: span is 0.50.



The Bootstrap: Loess instability

Note: in this *semi-parametric* setting, the cross-validation can become unstable. A specific random split may disrupt the local structure of the data so that a local fit at any given x is not possible.

This instability can potentially occur in the parametric setting as well; if the x values contain ties (as for the cholostyramine data) then, by chance we may resample an $\boldsymbol{X}_{(b)}$ such that

$$\boldsymbol{X}_{(b)}^{\mathsf{T}}\boldsymbol{X}_{(b)}$$

is singular.

The Bootstrap: Problems

Suppose $Y_1, \ldots, Y_n \sim \mathcal{N}(0,1)$. Consider the statistic

$$T_n = \max\{Y_1, \ldots, Y_n\}$$

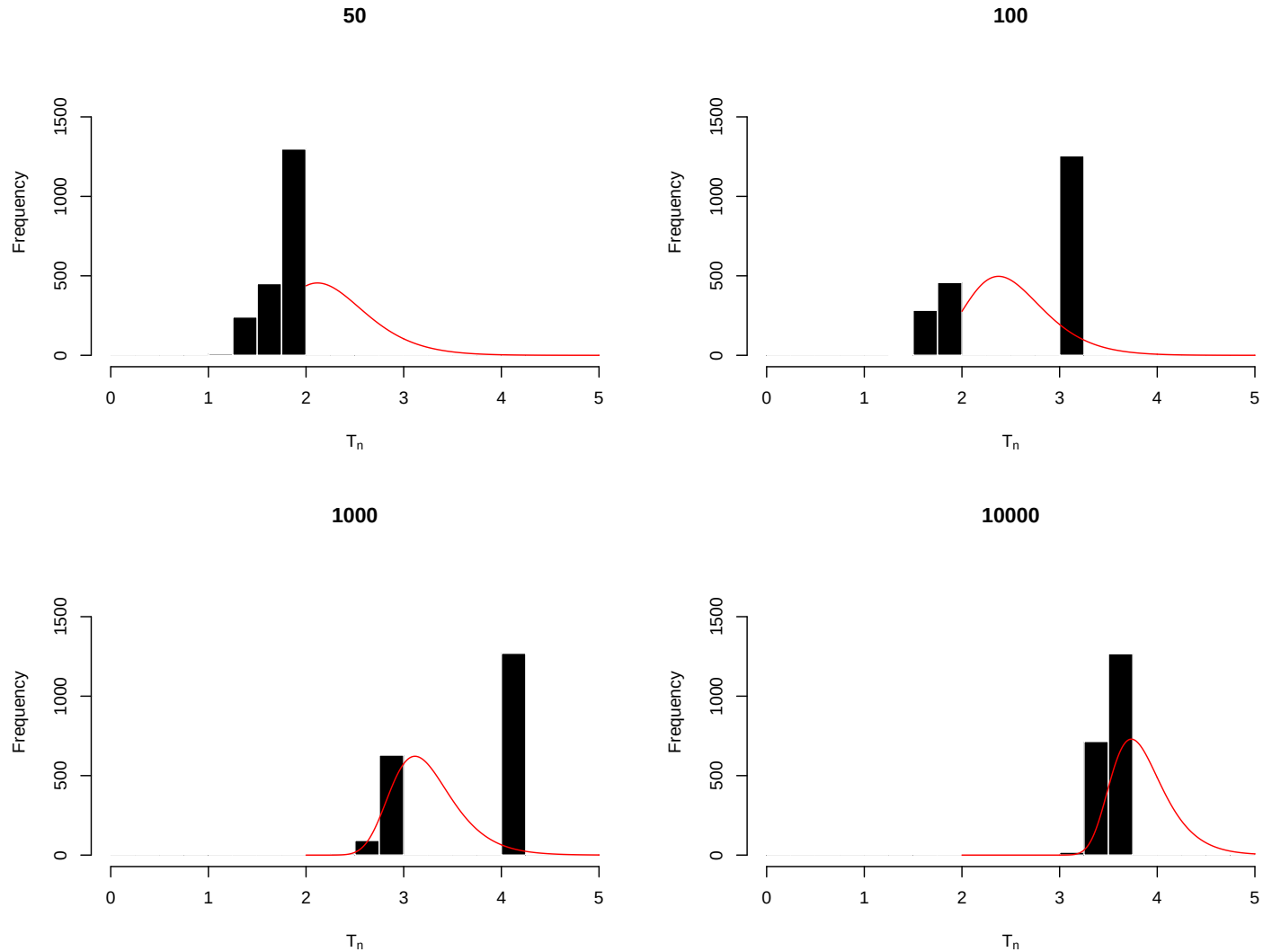
We know that for $t \in \mathbb{R}$

$$F_{T_n}(t) = \{\Phi(t)\}^n \qquad\qquad f_{T_n}(t) = n\phi(t)\{\Phi(t)\}^{n-1}$$

Can we study this statistic via the bootstrap ?

The Bootstrap: Problems

$B = 2000$ bootstrapped maxima for $n = 50, 100, 1000, 10000$.



The Bootstrap: Problems

The maximum order statistic cannot be studied in this setting using the bootstrap.

In this case, T_n has no limiting distribution, that is, as $n \longrightarrow \infty$

$$F_{T_n}(t) \longrightarrow 0$$

for all t. Is this the source of the problem ?

The Bootstrap: Problems

Suppose $Y_1, \ldots, Y_n \sim \mathit{Uniform}(0, \theta)$. Consider again the statistic T_n, which in this case is the ML estimator of θ:

$$\widehat{\theta}_n = T_n = \max\{Y_1, \ldots, Y_n\}.$$

We know that for $t \in \mathbb{R}$

$$F_{T_n}(t) = \left\{\frac{t}{\theta}\right\}^n \qquad\qquad f_{T_n}(t) = n\frac{1}{\theta}\left\{\frac{t}{\theta}\right\}^{n-1}$$

The same problem arises in this setting.

The Bootstrap: Problems

For the bth bootstrap resampled data set

$$\begin{aligned}\mathbb{P}[\widehat{\theta}^{(b)} = \widehat{\theta}] &= \mathbb{P}[\max\{Y_1^{(b)}, \ldots, Y_n^{(b)}\} = \max\{Y_1, \ldots, Y_n\}] \\ &= 1 - \left(1 - \frac{1}{n}\right)^n\end{aligned}$$

and as $n \longrightarrow \infty$

$$1 - \left(1 - \frac{1}{n}\right)^n \longrightarrow 1 - e^{-1} \doteq 0.632$$

Therefore, even for large n, there is a probability mass at $\widehat{\theta}$ in the sampling distribution.

The Bootstrap: Problems

Other problems can occur for the bootstrap when resampling can fundamentally change the nature of the data set:

- **Mixture distributions:** if the Y_i follow a mixture distribution with K components, then resampling may only yield observations from $L < K$ of the components. This changes the statistical behaviour of some estimators.
- **Logistic Regression (or classification):** For binary data, resampling may eliminate **all** of the data with response value 0 (or 1), in which case the logistic regression coefficients are not defined.

The Parametric Bootstrap

The *parametric* bootstrap replaces **resampling** from $\widehat{F}_n$ with **sampling** from a parametric approximation $F(.; \widehat{\vartheta}_n)$, where $\widehat{\vartheta}_n$ is the parameter estimate derived from the original sample.

Essentially, this is a *Monte Carlo* technique, where $\widehat{s}(\widehat{\theta}^{(*)})$ is approximated using B random samples from $F(.; \widehat{\vartheta}_n)$.

This can be useful if a good (and tractable) parametric approximation exists.

The Parametric Bootstrap

Suppose $Y_1, \ldots, Y_n \sim \mathit{Uniform}(0, \theta)$.

- **Nonparametric bootstrap:** resampling from $\{y_1, \ldots, y_n\}$ leads to problems as

$$\mathbb{P}[\widehat{\theta}^{(b)} = \widehat{\theta}] > 0$$

- **Parametric bootstrap:** sampling from $\mathit{Uniform}(0, \widehat{\theta})$ yields

$$\mathbb{P}[\widehat{\theta}^{(b)} = \widehat{\theta}] = 0$$

  and a reasonable approximation to the distribution of $\widehat{\theta}$.

The Parametric Bootstrap

Non-parametric (left) vs Parametric (right)

