Linked bootstrap

Wesley Brooks

Introduction

Local polynomial regression

Inference

Parametric bootstrap

Naive parametric bootstrap

We require a parametric model from which to make parametric bootstrap draws. For local polynomial regression, the model used to make the draws is the one that uses all available covariates. An independent model is generated at each observation location.

$$\left\{ f(\boldsymbol{s}) h^2 n \right\}^{1/2} \left[\hat{\boldsymbol{\beta}}_{(a)}(\boldsymbol{s}) - \boldsymbol{\beta}_{(a)}(\boldsymbol{s}) - (2\kappa_0)^{-1}\kappa_2 h^2 \left\{ \nabla_{uu}^2 \boldsymbol{\beta}_{(a)}(\boldsymbol{s}) + \nabla_{vv}^2 \boldsymbol{\beta}_{(a)}(\boldsymbol{s}) \right\} \right]$$
$$\xrightarrow{d} N \left(0, \kappa_0^{-2} \nu_0 \sigma^2 \Psi_{(a)}(\boldsymbol{s})^{-1} \right),$$

where $\left\{ \nabla_{uu}^2 \boldsymbol{\beta}_{(a)}(\boldsymbol{s}) + \nabla_{vv}^2 \boldsymbol{\beta}_{(a)}(\boldsymbol{s}) \right\} = \left(\nabla_{uu}^2 \boldsymbol{\beta}_1(\boldsymbol{s}) + \nabla_{vv}^2 \boldsymbol{\beta}_1(\boldsymbol{s}), \dots, \nabla_{uu}^2 \boldsymbol{\beta}_{p_0}(\boldsymbol{s}) + \nabla_{vv}^2 \boldsymbol{\beta}_{p_0}(\boldsymbol{s}) \right)^T$.

Denote bootstrap draws from the distribution for $\hat{\boldsymbol{\beta}}(\boldsymbol{s})$ by $\boldsymbol{\beta}_k^*(\boldsymbol{s})$ for $k = 1, \dots, B$. Then the corresponding response values are

$$y_k^*(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})\boldsymbol{\beta}_k^*(\boldsymbol{s}) + \varepsilon(\boldsymbol{s})$$
$$\text{for } i = 1, \dots, B$$

Now we have the resampled responses at $\boldsymbol{s}$: $y_1^*(\boldsymbol{s}), \dots, y_B^*(\boldsymbol{s})$ from the correct marginal distribution, but the draws at $\boldsymbol{s}$ are unconditional on the bootstrap draws at location $\boldsymbol{t}$.

In truth, the

Linked parametric bootstrap

The linked parametric bootstrap is used to generate draws of the coefficient functions with dependence between locations. This is accomplished by generating a single random vector, and multiplying it by the coefficient covariance matrix at each location.

$$\hat{\boldsymbol{\beta}}(\boldsymbol{s}) \sim N(\hat{\boldsymbol{\mu}}(\boldsymbol{s}), \Sigma(\boldsymbol{s}))$$
$$\boldsymbol{b}_k \sim N(0, I_p)$$
$$\boldsymbol{\beta}_k^{**}(\boldsymbol{s}) = \hat{\boldsymbol{\mu}}(\boldsymbol{s}) + \Sigma(\boldsymbol{s})^{1/2}\boldsymbol{b}_i$$
$$\text{for } k = 1, \dots, B$$

So that $\boldsymbol{\beta}^{**}(\boldsymbol{s})$ has the same marginal distribution as $\boldsymbol{\beta}_k^*(\boldsymbol{s})$ but the resampled functions are smoother than under the naive parametric bootstrap.

Example

The method is illustrated by an example where the data arises from a varying coefficient model with a one-dimensional effect-modifying parameter, $t \in [0, 1]$. The data is generated by:

$$
\begin{aligned}
y(t) &= \beta_0(t) + x(t)\beta_1(t) + \varepsilon(t) \\
\beta_0(t) &= 1 - 4(t - 2) + 0.7(t - 1)^2 + t^3 - 0.2t^4 \\
\beta_1(t) &= \cos t \\
X &\sim^{iid} N(3, 2) \\
\varepsilon &\sim^{iid} N(0, 1)
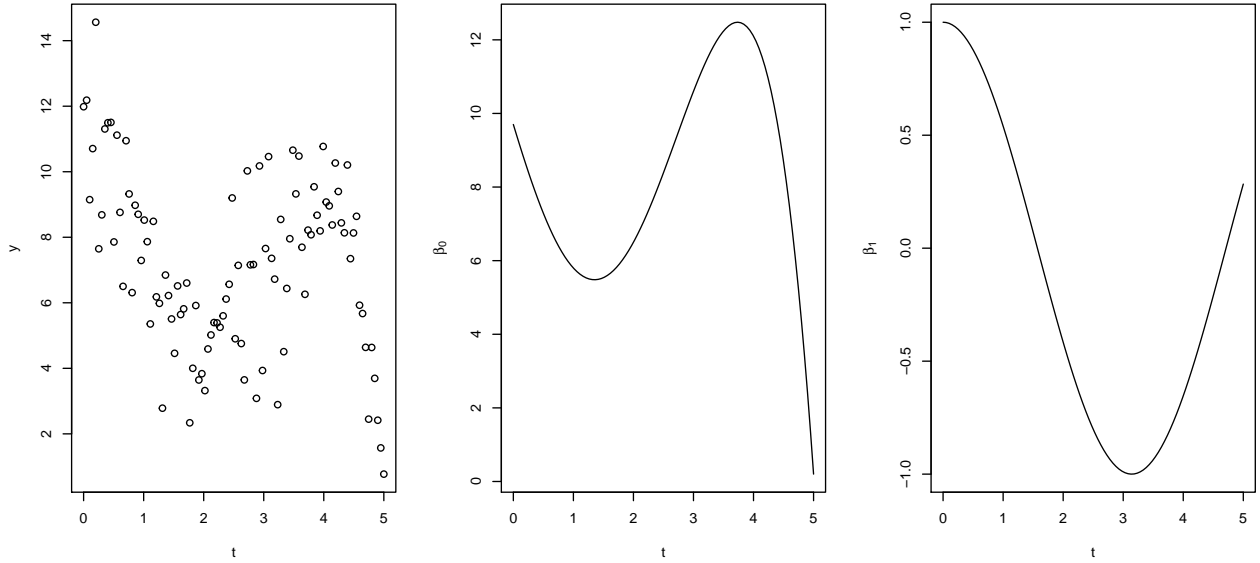\end{aligned}
$$

The response and the coefficients are plotted below:

Figure 1: From left to right, the simulated response, the intercept function and the coefficient function used to simulate the data.

```
## Note: method with signature 'symmetricMatrix#missing' chosen for function 'Schur',
##  target signature 'dsyMatrix#missing'.
##  "dsyMatrix#ANY" would also be valid
```

The parametric bootstrap is used to simulate the distribution of the estimated coefficient functions. The simulated response y is plotted in Figure **??** along with a bootstrap resample y^* from the NPB and a resample y^{**} from the LPB. There is no apparent consistent difference between the two kinds of bootstrap resamples y^* and $y * **$ - note that the resamples fall near each other with neither one consistently greater than the other in the left-hand lot, and generally along the $1 - 1$ line in the right-hand plot. There is an apparent tendency for the resampled data to "trim the peaks" and "fill the valleys" of the response, which we will return to.

In Figure **??**, the true $\beta_0(t)$ and $\beta_1(t)$ coefficient surfaces that were used to simulate the data are plotted along with example draws of $\beta_0^*(t)$, $\beta_0^{**}(t)$, $\beta_1^*(t)$, and $\beta_1^{**}(t)$. It is immediately obvious that the linked bootstrap draws are smoother than the naive bootstrap draws, and the linked bootstrap draws appear to better resemble the truth.
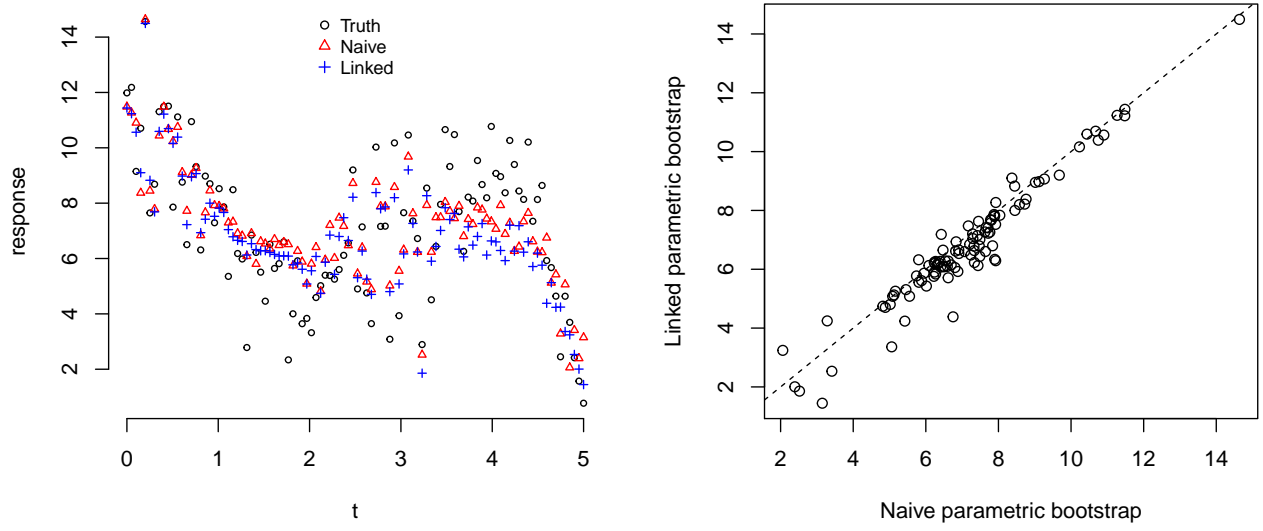
Figure 2: The simulated response, and the response as resampled by the naive parametric bootstrap and the linked parametric bootstrap.
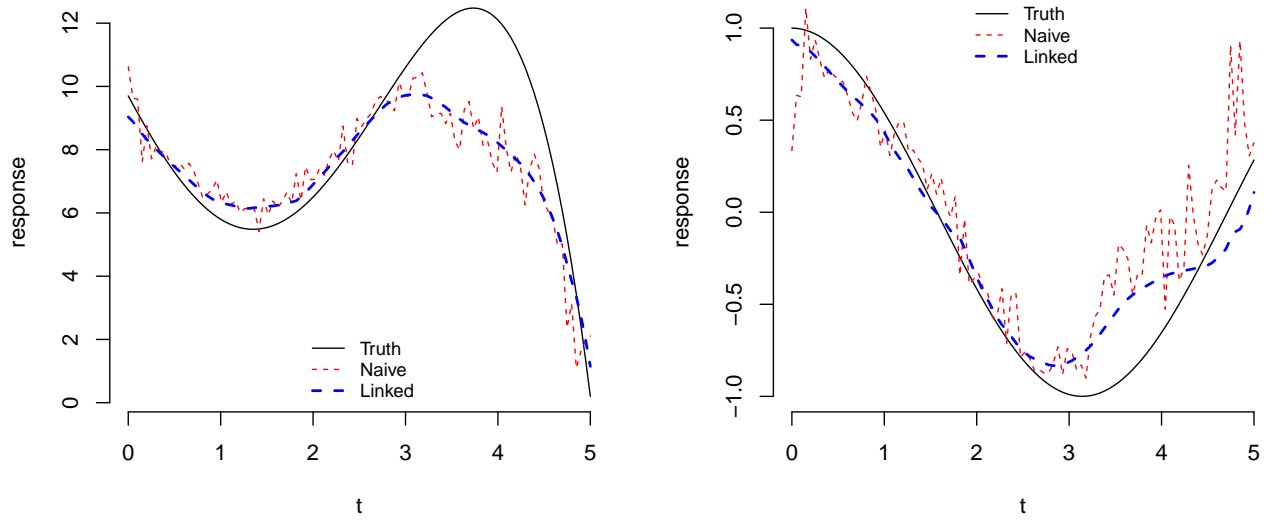

Figure 3: At left, the true coefficient function $\beta_0(t)$ (black solid line) and the resampled $\beta_0^*(t)$ from the naive parametric bootstrap (red) and $\beta_0^{**}(t)$ from the linked parametric bootstrap (blue). At right, the true coefficient function $\beta_1(t)$ (black solid line) and the resampled $\beta_1^*(t)$ from the naive parametric bootstrap (red) and $\beta_1^{**}(t)$ from the linked parametric bootstrap (blue).
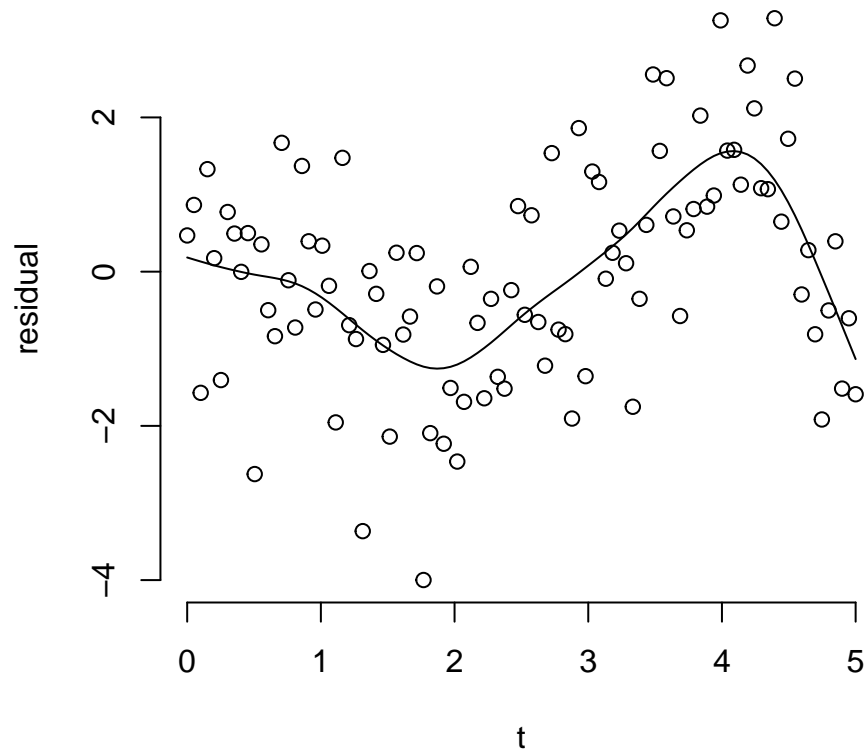
Figure 4: Smoothing spline used to smooth the residuals.

Because the local polynmial model tends to oversmooth the response, we use a spline model to estimate the local bias of the bootstrap response draws, in order to reduce the bias in the resampled data. The residuals from the original model fit are plotted in Figure **??** along with the spline used to smooth the residuals.