Fast Estimation and Choice of Confidence Interval Methods for Step Regression

Shuangcheng Hua  
Fred Hutchinson Cancer Research Center

Youyi Fong (youyifong@gmail.com)  
Fred Hutchinson Cancer Research Center

Jarrod Kath

Research Article

Keywords: dynamic programming, symmetric percentile bootstrap confidence intervals, subsampling, ecology

Posted Date: August 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1946081/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Fast Estimation and Choice of Confidence Interval
Methods for Step Regression

Shuangcheng Hua · Youyi Fong · Jarrod Kath

Received: date / Accepted: date

Abstract In this paper we propose a new fast grid search algorithm for finding the least square estimators of a step regression model. This algorithm makes it practical to compute resampling-based confidence intervals for step regression models. We introduce five data generating models, including one where the mean model is a step model (model correctly specified) and four where the mean models are not step models (model misspecified), and use them to study the coverage probabilities of two new types of resampling-based confidence intervals for step regression: symmetric percentile bootstrap confidence intervals and subsampling confidence intervals using a new set of rules-of-thumb to select block size. Our results show that when the model is correctly specified, the symmetric percentile Efron bootstrap confidence intervals provide close-to-nominal coverage and have shorter intervals than the subsampling methods; when the model is misspecified, the subsampling method using the rules-of-thumb provides good coverage and shorter confidence intervals than the symmetric percentile Efron bootstrap method and the subsampling method using a double bootstrap-like procedure for block size selection. Finally, we apply the proposed methods to a real world environmental dataset on the relationship between grassland productivity, soil moisture anomalies and other hydro-climatic and land use variables to provide inference for the threshold in soil moisture anomalies, across which there is a jump in grassland productivity.

Shuangcheng Hua
Department of Biostatistics, University of Washington

Youyi Fong
Department of Biostatistics, University of Washington
Correspondence: E-mail: youyifong@gmail.com

Jarrod Kath
University of Southern Queensland
1 Introduction

Step regression, also known as discontinuous threshold regression, is a type of non-regular regression model where the mean of an outcome variable is a step function of the covariate of interest, e.g.

$$Y = \alpha + \alpha^T z + \beta I(x > e) + \epsilon,$$

where $Y$ is the response, $x$ is the predictor with threshold effect, $z$ is a covariate vector with $p-1$ dimensions, and $\epsilon$ denotes an error term with bounded variance that is independent between observations, $e$ is the threshold parameter, $\beta$ denotes the size of the jump at the threshold, $\alpha$ is the intercept, and $\alpha_z$ denotes the coefficients corresponding to $z$. This model is often used in practice because it provides a simple but elegant and interpretable way to model certain kinds of threshold-dependent relationships between an outcome and a predictor. Despite the simplicity of step regression models, the presence of threshold parameters changes both estimation and inference for the models in profound ways. For example, it is well understood that the asymptotic distribution of the threshold parameter is nonstandard: while the least square estimator of the change point or threshold converges at a rate of $n^{1/2}$ when both the data generating model and the working model are step linear regression models, it converges at a rate of $n^{1/3}$ when the true underlying model does not actually follow a step regression model (e.g. Bühlmann and Yu, 2002; Pons, 2003; Banerjee and McKeague, 2007; R. and Kosorok, 2008; Song et al., 2016).

The use of step regression in data analysis is currently limited by two factors: (i) Existing methods for step regression model estimation either produce locally optimal instead of globally optimal solutions or are too slow. (ii) The performance of resampling-based confidence interval methods depends critically on block size selection, but there is a dearth of studies on how to select block sizes when the data generating model and the working model do not match. Here we address these two limitations. In Section 2 we propose a fast grid search algorithm for step linear regression that seeks globally optimal solutions and apply the method to empirically study the convergence rates under five different simulation scenarios. In Section 3, we study symmetric percentile Efron bootstrap confidence intervals and subsampling confidence intervals both when the data generating model and the working model match and when they do not match. We further propose new rules for selecting block size in subsampling that compare favorably with the existing double bootstrap-like procedure. In Section 4, we apply the proposed methods to a real data example from a study of hydro-climatic drivers of agricultural grassland productivity under extreme drought and rainfall. We end with a discussion in Section 5.
2 Fast grid search-based estimation of step linear regression models

Due to the threshold parameter, the likelihood function of a step threshold model is non-smooth and non-convex. A natural approach to maximize the likelihood is to approximate the step function in the likelihood by a smooth function (e.g. Gallant and Fuller, 1973; Tishler and Zang, 1981; Pastor and Guallar, 1998; Muggeo, 2003; Fong et al., 2017). However, as shown in Fong (2019), such approximation may lead to inaccurate coverage of bootstrap confidence intervals in threshold linear regressions because the criterion functions are still non-convex and the approach finds locally optimal solutions. Conceptually, the grid search approach (Friedman and Silverman, 1989) is a two-stage process: in the first stage we optimize a series of submodels conditional on a grid of candidate threshold values (typically the realized covariate values in the dataset); in the second stage we simply find the maximum in the set of likelihoods computed in the first stage and take the corresponding threshold value as the estimated threshold. Since conditioning on the threshold values removes both non-smoothness and non-convexity associated with the presence of the threshold parameter, the grid search approach yields globally optimal solutions. The main disadvantage of the grid search approach is its computational burden.

We propose a fast grid search-based estimation method for step linear regression models. The method uses the same strategy from Elder and Fong (2019) and Son and Fong (2020), which deals with continuous two-phase regression models. We sketch the outline of the algorithm here; further details can be found in Section B of the Supplementary Materials. Consider the model given in the introduction, the least squares estimator for the model parameters minimizes the sum of squares of the residuals, which equals $Y^T H e Y = Y^T H Y + (v^T_e r)^2 / (v^T_e v_e - v^T_e H v_e)$, where

$$Y^T H e Y = Y^T H Y + (v^T_e r)^2 / (v^T_e v_e - v^T_e H v_e), \tag{1}$$

where $H \equiv X (X^T X)^{-1} X^T$ and $r \equiv (H - I) Y$. By the QR decomposition, we can further write $v^T_e H v_e = (Q^T_1 v_e)^T Q^T_1 v_e$, where $Q_1$ is the first $p$ columns of the orthogonal matrix $Q$ from the QR decomposition.

Consider two successive (different) values of $e$: $e_t$ and $e_{t+1}$. Suppose $e_t$ and $e_{t+1}$ correspond to the $t^{th}$ and $t + 1^{th}$ ascending ordered values of $x$. Let
\( \mathbf{v}_{t+1} = \mathbf{v}_t - \delta_t \). We find that:

1. \( \mathbf{v}_{t+1}^\top \mathbf{v}_{t+1} = \mathbf{v}_t^\top \mathbf{v}_t - 1 \) \hspace{1cm} (2)
2. \( \mathbf{v}_{t+1}^\top \mathbf{Q}_1 = \mathbf{v}_t^\top \mathbf{Q}_1 - \delta_t^\top \mathbf{Q}_1 \) \hspace{1cm} (3)
3. \( \mathbf{v}_{t+1}^\top \mathbf{r} = \mathbf{v}_t^\top \mathbf{r} - \delta_t^\top \mathbf{r} \) \hspace{1cm} (4)

Here we use the fact that \( \delta_t \) is a vector of size \( n \) with the \((t + 1)\)th entry equal to 1 and 0 everywhere else. \( \delta_t^\top \mathbf{Q}_1 \) is the \( t + 1 \)th row vector of \( \mathbf{Q}_1 \), and \( \delta_t^\top \mathbf{r} \) is the \( t + 1 \)th element of \( \mathbf{r} \). Hence, these update steps can be done very quickly. The full algorithm is described more formally below. The results of some benchmarking experiments are shown in Table 1.

**Algorithm 1** Fast grid search algorithm for step linear regression model

1. Sort the samples by the ascending order of \( x_i \)
2. Compute and store the initial values: \( \mathbf{v}_1^\top \mathbf{v}_1, \mathbf{v}_1^\top \mathbf{Q}_1, \mathbf{v}_1^\top \mathbf{r} \)
3. Compute and store the initial value of \( \mathbf{Y}_1^\top \mathbf{H}_1 \mathbf{Y} \)
4. For \( t \) in 1 to \( n - 1 \):
   - update \( \mathbf{v}_{t+1}^\top \mathbf{v}_{t+1}, \mathbf{v}_{t+1}^\top \mathbf{Q}_1, \mathbf{v}_{t+1}^\top \mathbf{r} \) based on (2), (3), and (4)
   - update \( \mathbf{Y}_t^\top \mathbf{H}_{t+1} \mathbf{Y} \) based on (1)
   - update \( \hat{e} = e_{t+1} \) if \( Y_t^\top \mathbf{H}_{t+1} \mathbf{Y} > Y_t^\top \mathbf{H}_t \mathbf{Y} \)

| \( n \) = 10^3, \( p = 2 \) | 2604 | 0.40 |
| \( n = 10^4, p = 2 \) | - | 3.80 |
| \( n = 10^5, p = 2 \) | - | 42.7 |
| \( n = 10^6, p = 2 \) | - | 465 |
| \( n = 10^6, p = 10 \) | - | 1223 |

Table 1: Run time (sec) for fitting step linear regression models on a single Intel(R) Xeon(R) CPU E5-2667 v4 @ 3.20GHz, averaged over 20 Monte Carlo datasets.

With the fast grid search algorithm, we can more easily study the convergence rates of the parameter estimates empirically. To compare these converge rates to the asymptotic rates, we use five data generating models: (i) Step: The mean of \( Y \) is a step function of \( x \); (ii) Sig,\( \gamma \) for \( \gamma \in \{1, 5, 15\} \): The mean of \( Y \) is a sigmoid function of \( x \). Sig,15 more closely resembles a step function than Sig,1; (iii) Quad: The mean of \( Y \) is a quadratic function of \( x \):

\[
E(Y|Z, X) = \alpha + \alpha_Z Z + \beta I(X > e) \quad \text{(Step)}
\]
\[
E(Y|Z, X) = \alpha + \alpha_Z Z + \beta \frac{\gamma e^{(X-4.7)}}{1 + \gamma e^{(X-4.7)}} \quad \text{(Sig)}
\]
\[
E(Y|Z, X) = \alpha + \alpha_Z Z + \gamma_1 X + \gamma_2 X^2 \quad \text{(Quad)}
\]
The model parameters and the covariate distributions are given in Section A of the Supplementary Materials. The step linear regression model is correctly specified under the Step model and misspecified under the Sig-γ models and the Quad model. Under the Sig-γ models, the step linear regression model can be seen as an approximation of the data generating models because they have the same overall shape; under the Quad model, the interpretation of the step model fit needs to be more carefully considered, e.g. the limits of the step linear regression model parameters depend heavily on the distribution of $X$.

To obtain the limits of the step linear regression model parameters when the model is misspecified or when the limits cannot be inferred from symmetry, we fit the step linear regression model $Y = \alpha + \alpha Z + \beta I(X > e)$ to datasets with sample size $n = 10^6$ and take the average over ten Monte Carlo replicates, the results of which are listed in Tables A.1-A.4 of the Supplementary Materials.

To study the empirical rate of convergence over a spectrum of sample sizes, we consider two sets of sample sizes. The first set has sample sizes of 500, 1,000, 1,500, and 2,000, and the second set has sample sizes of 1,024, 2,048, 4,096, and 8,192,000. At each sample size, we estimate the variability of the estimator by conducting $10^6$ Monte Carlo runs and computing the standard deviation of the parameter estimates across the Monte Carlo replicates, which we denote by $\sigma_n$. To estimate the rate of convergence, we fit the model $\sigma_n = a \times n^b$ by fitting a straight line through $(n, log(\sigma_n))$ so that the slope $\hat{b}$ of the fitted line gives an estimated convergence rate.

The results for the two sets of sample sizes are summarized in Table 2 and Table 3, respectively. When the model is correctly specified, as shown in both Table 2 and Table 3, the estimated convergence rates are close to the asymptotic rates: $n$ for the threshold parameter $e$ and $n^{1/2}$ for the slope parameters. When models are misspecified, the results are complex. We discuss the results one parameter at a time. In larger datasets, $\hat{e}$ is $n^{1/3}$-convergent. In smaller datasets, the convergence rate is between $n^{1/2}$ and $n^{1/3}$, and among the three sigmoid models, the closer the model is to the step model, the faster it converges.

The estimates of $\alpha$ and $\alpha + \beta$, which correspond to the contribution of $x$ to the mean when $x \leq e$ and $x > e$, respectively, converge at a rate between $n^{1/2}$ and $n^{1/3}$. The convergence rate is closer to $n^{1/3}$ when the sample size is larger, and in the series of sigmoid models, closer to $n^{1/3}$ when the true model is further away from the step model.

The behaviors of $\hat{\beta}$, the estimated jump at the threshold, are more interesting. For the quadratic model, the convergence rate is close to $n^{1/2}$ in smaller datasets but tends towards the asymptotic rate of $n^{1/3}$ in larger datasets. For all three sigmoid models, however, the convergence rate is $n^{1/2}$ in both smaller and larger datasets, suggesting that the asymptotic rate of convergence is $n^{1/2}$ and not $n^{1/3}$. The reason for this can be seen from Theorem 2.1 of Banerjee and McKeague (2007). The limiting distribution of $\beta$, when scaled by $n^{1/3}$, equals to $n^{1/3}(c_1 - c_2) \arg \max_t Q(t)$, but due to the symmetry of the these sigmoid models and the choice of the distribution of $X$, $c_1$ equals to $c_2$ in this special case causing this limiting distribution to be degenerate.
Finally, the estimates $\alpha_z$, the slope associated with $Z$, converges at the regular $\sqrt{n}$ rate in both smaller and larger datasets.

<table>
<thead>
<tr>
<th>Step</th>
<th>$c$</th>
<th>$\beta$</th>
<th>$\alpha_x$</th>
<th>$\alpha_x + \beta$</th>
<th>$\alpha_z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sig_15</td>
<td>-1.03 (-1.07, -1.00)</td>
<td>-0.50 (-0.50, -0.49)</td>
<td>-0.51 (-0.52, -0.51)</td>
<td>-0.51 (-0.52, -0.50)</td>
<td>-0.51 (-0.51, -0.50)</td>
</tr>
<tr>
<td>Sig_5</td>
<td>-0.45 (-0.53, -0.38)</td>
<td>-0.50 (-0.54, -0.46)</td>
<td>-0.51 (-0.52, -0.50)</td>
<td>-0.50 (-0.52, -0.47)</td>
<td>-0.51 (-0.51, -0.49)</td>
</tr>
<tr>
<td>Sig_1</td>
<td>-0.39 (-0.45, -0.34)</td>
<td>-0.50 (-0.54, -0.46)</td>
<td>-0.50 (-0.51, -0.48)</td>
<td>-0.47 (-0.51, -0.44)</td>
<td>-0.51 (-0.51, -0.49)</td>
</tr>
<tr>
<td>Quad</td>
<td>-0.35 (-0.39, -0.30)</td>
<td>-0.49 (-0.52, -0.46)</td>
<td>-0.51 (-0.52, -0.48)</td>
<td>-0.44 (-0.48, -0.41)</td>
<td>-0.51 (-0.52, -0.49)</td>
</tr>
</tbody>
</table>

Table 2: Estimated convergence rates and their 95% CIs based on $n \in \{500, 1000, 1500, 2000\}$ from $10^4$ Monte Carlo replicates. $c$: threshold; $\beta$: jump; $\alpha_z$: slope of $z$; $\alpha$: “lower step”; $\alpha + \beta$: “upper step”.

<table>
<thead>
<tr>
<th>Step</th>
<th>$c$</th>
<th>$\beta$</th>
<th>$\alpha_x$</th>
<th>$\alpha_x + \beta$</th>
<th>$\alpha_z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sig_15</td>
<td>-0.99 (-1.01, -0.98)</td>
<td>-0.50 (-0.50, -0.49)</td>
<td>-0.50 (-0.50, -0.49)</td>
<td>-0.50 (-0.50, -0.49)</td>
<td>-0.50 (-0.50, -0.49)</td>
</tr>
<tr>
<td>Sig_5</td>
<td>-0.33 (-0.34, -0.32)</td>
<td>-0.50 (-0.52, -0.47)</td>
<td>-0.42 (-0.44, -0.40)</td>
<td>-0.42 (-0.44, -0.41)</td>
<td>-0.50 (-0.52, -0.48)</td>
</tr>
<tr>
<td>Sig_1</td>
<td>-0.34 (-0.34, -0.33)</td>
<td>-0.50 (-0.52, -0.48)</td>
<td>-0.37 (-0.38, -0.36)</td>
<td>-0.38 (-0.40, -0.35)</td>
<td>-0.50 (-0.52, -0.48)</td>
</tr>
<tr>
<td>Quad</td>
<td>-0.34 (-0.36, -0.33)</td>
<td>-0.42 (-0.43, -0.40)</td>
<td>-0.36 (-0.38, -0.34)</td>
<td>-0.37 (-0.39, -0.35)</td>
<td>-0.51 (-0.52, -0.49)</td>
</tr>
</tbody>
</table>

Table 3: Estimated convergence rates and their 95% CIs based on $n \in \{1024000, 2048000, 4096000, 8192000\}$ from $10^4$ Monte Carlo replicates. $c$: threshold; $\beta$: jump; $\alpha_z$: slope of $z$; $\alpha$: “lower step”; $\alpha + \beta$: “upper step”.

3 Empirical studies of resampling-based confidence interval methods

3.1 Symmetric percentile Efron bootstrap confidence intervals

Efron bootstrap confidence intervals (Carpenter and Bithell, 2000) are widely used for making inference in $\sqrt{n}$-convergence problems. The percentile Efron bootstrap confidence interval is well known to be inconsistent for non-regular problems, including step regression (e.g. Bühlmann and Yu, 2002; Seijo et al., 2011; Yu, 2014). The theoretical properties of other types of general purpose Efron bootstrap confidence intervals (Carpenter and Bithell, 2000), such as inverse percentile and symmetric percentile, are largely unknown. A type of bootstrap method specific for step regression, smoothed percentile bootstrap confidence intervals (Seijo et al., 2011), has been proposed, but it involves a tuning parameter and can be hard to apply in practice (Yu, 2014) and only works when the model is correctly specified. In this section, we compare the coverage of two types of Efron bootstrap confidence intervals under each of the five data generating models introduced in the previous section. In addition to the percentile method, we focus on the symmetric percentile (“symmetric”) method (Hansen, 2017). The symmetric method can be seen as a compromise
between the percentile and inverse percentile methods. For example, for \( \alpha \) the 95% confidence interval is defined as \( \hat{\alpha} \pm q^* \), where \( q^* \) is the 95\% quantile of the bootstrap distribution of \( |\hat{\alpha} - \bar{\alpha}| \).

<table>
<thead>
<tr>
<th>n</th>
<th>( \epsilon )</th>
<th>( \beta )</th>
<th>( \alpha )</th>
<th>( \alpha_x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step</td>
<td>250</td>
<td>500</td>
<td>1,000</td>
<td>2,000</td>
</tr>
<tr>
<td></td>
<td>83.0 (0.5)</td>
<td>83.1 (0.4)</td>
<td>82.7 (1.2)</td>
<td>92.9 (0.86)</td>
</tr>
<tr>
<td></td>
<td>96.7 (0.61)</td>
<td>96.8 (0.30)</td>
<td>96.3 (0.14)</td>
<td>96.3 (0.07)</td>
</tr>
<tr>
<td></td>
<td>94.9 (0.15)</td>
<td>94.9 (0.10)</td>
<td>94.3 (0.07)</td>
<td>94.8 (0.05)</td>
</tr>
<tr>
<td></td>
<td>95.1 (0.15)</td>
<td>95.1 (0.10)</td>
<td>94.7 (0.04)</td>
<td>94.9 (0.04)</td>
</tr>
<tr>
<td></td>
<td>95.1 (0.11)</td>
<td>94.7 (0.08)</td>
<td>94.7 (0.03)</td>
<td>94.9 (0.03)</td>
</tr>
<tr>
<td></td>
<td>94.6 (0.08)</td>
<td>94.9 (0.08)</td>
<td>94.9 (0.04)</td>
<td>94.9 (0.04)</td>
</tr>
<tr>
<td></td>
<td>94.8 (0.08)</td>
<td>94.9 (0.08)</td>
<td>94.8 (0.05)</td>
<td>94.9 (0.05)</td>
</tr>
<tr>
<td></td>
<td>94.9 (0.08)</td>
<td>94.9 (0.08)</td>
<td>94.9 (0.05)</td>
<td>94.9 (0.05)</td>
</tr>
<tr>
<td></td>
<td>95.0 (0.01)</td>
<td>95.0 (0.01)</td>
<td>95.0 (0.01)</td>
<td>95.0 (0.01)</td>
</tr>
</tbody>
</table>

Table 4: Simulation results from 10,000 Monte Carlo runs. Estimated coverage and mean width of percentile and symmetric Efron bootstrap CIs.

As shown in Table 4, when the data is generated from a step model, both percentile and symmetric Efron bootstrap confidence intervals have coverage probabilities close to the nominal level 0.95 for the \( n^{1/2} \)-convergent slope parameters \( \beta, \alpha_x \) and \( \alpha \). As for the threshold parameter \( \epsilon \), the percentile method suffers from undercoverage. Importantly, this does not improve with larger sample sizes, providing empirical evidence that the Efron bootstrap is inconsistent for the \( n \)-convergent \( \hat{\epsilon} \). Interestingly though, the symmetric method produces reasonable coverage with only a small increase in the width of the confidence intervals.

In contrast with the results under correct model specification, under model misspecification both the percentile and symmetric methods show a small amount of over-coverage for the threshold parameter \( \epsilon \). For the slope parameter \( \beta \), even though it is \( n^{1/2} \)-convergent under the Sig.1 model, the percentile method shows substantial undercoverage when the sample size is small-to-
moderate; the symmetric method also under-covers, but to a much smaller extent. For the \( n^{1/3} \)-convergent \( \alpha \), both methods provide reasonable coverage, but show some over-coverage when the sample size increases. Moreover, the amount of over-coverage also increases from Sig.15 to Sig.1. These results, together with the results in Tables 2 and 3, suggest that the more the “effectual” convergence rates move from \( n^{1/2} \) towards \( n^{1/3} \), the more over-coverage there will be for both bootstrap confidence interval methods.

To better understand the reasons behind the difference in performance between the percentile and symmetric methods, we divide the \( 10^4 \) Monte Carlo replicates into three categories based on the skewness of the bootstrap sampling distributions of \( \hat{c} \), as measured by the moment coefficient of skewness (Joanes and Gill, 1998). We regard those calculated to be less than \(-0.5\), between \(-0.5\) and 0.5, and greater than 0.5 as left-skewed, not-skewed, and right-skewed, respectively (Bulmer, 1966). We find that 40% of the bootstrap distributions are left-skewed, 20% are not-skewed, and 40% are right-skewed. While the percentile method covers the truth 71%, 89% and 89% of the time for left-skewed, not-skewed, and right-skewed bootstrap distributions, respectively, the symmetric method covers 96%, 95% and 95% of the time, respectively.

To summarize, these Monte Carlo experiment results show that the symmetric Efron bootstrap confidence interval method may correct for the under-coverage of the \( n \)-convergent threshold parameter exhibited by the percentile Efron bootstrap method when the step regression model is correctly specified. Moreover, the symmetric Efron bootstrap method provides reasonable coverage for the \( n^{1/3} \)-convergent model parameters when the step regression model is misspecified and the sample size is small to moderate.

3.2 Subsampling with a double bootstrap-like procedure for block size selection

Two types of resampling-based methods have been proposed for making inference for non-regular problems: m-out-of-n bootstrap, which resamples with replacement fewer than \( n \) observations (e.g. Bickel et al., 1997), and subsampling, which resamples without replacement fewer than \( n \) observations (e.g. Politis and Romano, 1994; Bertail et al., 1999). Seijo et al. (2011) showed that m-out-of-n bootstrap is valid in a step linear regression model when the mean model is correctly specified. However, it is not clear whether the method is still consistent when the mean model is misspecified. In our empirical studies (results not shown) the m-out-of-n bootstrap confidence intervals provided close-to-nominal coverage with appropriately chosen block sizes when the mean model is correctly specified, but over-covered for all block sizes when the mean model is misspecified. We thus focus our attention on subsampling.

An important consideration for using the subsampling method is the choice of the block size \( m_n \), or \( m \) for short (Politis et al., 1999). A common approach in all blocking methods (e.g. Delgado et al., 2001; Gonzalo and Wolf,
2005; Chakraborty et al., 2013) is nested resampling. This leads to a double bootstrap-like (DBL) procedure (Delgado et al., 2001), in which the original dataset is bootstrapped and subsampling is performed on each bootstrap dataset at a grid of candidate block sizes to look for the block size that provides a close to nominal estimated coverage. The procedure can be described more formally as follows:

1. Draw $B_1$ first-level samples by Efron bootstrap from the data, and calculate the first-level estimates of the parameter of interest $\hat{e}_{b_1}$, $b_1 = 1, ..., B_1$.

2. For each first-level bootstrap sample ($b_1 = 1, ..., B_1$):
   
   (a) Draw $B_2$ second-level samples by sampling $m$ subjects without replacement and calculate the second-level parameter estimates $\hat{e}_{b_1,b_2}$, $b_2 = 1, ..., B_2$ for the smallest $m$ in a grid of 25 values evenly spaced between 0.05$n$ and 0.8$n$.

   (b) Denote the $(α/2) \times 100$ and the $(1 − α/2) \times 100$ percentiles of $\hat{e}_{b_1,b_2}$ ($b_2 = 1, ..., B_2$) by $\hat{e}_{b_1,b_2}^{(α/2)}$ and $\hat{e}_{b_1,b_2}^{(1−α/2)}$, respectively. Construct the subsampling confidence interval for a first-level sample as $(\hat{e}_{b_1}^{(α/2)}, \hat{e}_{b_1}^{(1−α/2)})$.

3. Estimate the coverage probabilities at each $m$ by $\frac{1}{B_1} \sum_{i=1}^{B_1} I\{\hat{e}_{b_1}^{(α/2)} \leq \hat{e}_n \leq \hat{e}_{b_1}^{(1−α/2)}\}$, where $\hat{e}_n$ is the estimate of the parameter of interest from the original dataset.

4. Increment the block size $m$ from the smallest to the next higher value in the grid and repeat steps 2 and 3 until the estimated coverage probability goes above the nominal level.

5. Find $m$ that corresponds to the nominal coverage level by linear interpolation between the two block sizes whose estimated coverage probabilities bracket the nominal level.

For the choice of $B_1$, which affects how well we can estimate the coverage probabilities under different block sizes, and $B_2$, which affects how well we can estimate the subsampling confidence intervals, we experiment with different combinations of $B_1$ and $B_2$ with a fixed $B_1 \times B_2$. The results, shown in Table C.3 of the Supplementary Materials, suggest that the performance of the procedure is not overly sensitive to the choice of $B_1$, but when $B_2$ is too small (50), it leads to small selected block sizes, wide confidence intervals, and over-coverage. For the remainder of the paper, we let $B_1 = 200$ and $B_2 = 200$.

The block size $m$ selected by the double bootstrap-like procedure and the corresponding coverage probabilities and widths of subsampling confidence intervals for the threshold parameter $\epsilon$ are shown in Table 5. When the model is correctly specified, the DBL confidence intervals appear to over-cover; comparing with Table 4, we see that the DBL confidence intervals are on average longer than the symmetric Efron bootstrap confidence intervals for the threshold parameter, e.g. at $n = 1000$, the mean confidence interval width is 0.14 and 0.20 for symmetric Efron bootstrap and DBL, respectively. When the model is misspecified, the DBL confidence intervals also over-cover, but the degree of
Table 5: The performance of the double bootstrap-like procedure with $B_1 = 200$ and $B_2 = 200$. $m$: average selected block size; cvg: coverage probabilities of $e$ estimated from $10^4$ Monte Carlo replicates; width: average width of subsampling confidence intervals.

<table>
<thead>
<tr>
<th>n</th>
<th>m</th>
<th>cvg (width)</th>
<th>m</th>
<th>cvg (width)</th>
<th>m</th>
<th>cvg (width)</th>
<th>m</th>
<th>cvg (width)</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>109</td>
<td>98.0 (.77)</td>
<td>121</td>
<td>97.4 (.76)</td>
<td>129</td>
<td>96.6 (.39)</td>
<td>134</td>
<td>95.3 (2.62)</td>
</tr>
<tr>
<td>500</td>
<td>212</td>
<td>98.2 (.39)</td>
<td>248</td>
<td>97.6 (.43)</td>
<td>255</td>
<td>96.8 (.67)</td>
<td>267</td>
<td>95.3 (1.99)</td>
</tr>
<tr>
<td>1000</td>
<td>404</td>
<td>98.2 (.20)</td>
<td>502</td>
<td>97.1 (.27)</td>
<td>518</td>
<td>96.1 (.47)</td>
<td>530</td>
<td>96.2 (1.54)</td>
</tr>
<tr>
<td>1500</td>
<td>605</td>
<td>98.2 (.13)</td>
<td>759</td>
<td>96.9 (21)</td>
<td>778</td>
<td>96.2 (.40)</td>
<td>793</td>
<td>95.5 (1.34)</td>
</tr>
<tr>
<td>2000</td>
<td>787</td>
<td>98.4 (.10)</td>
<td>1037</td>
<td>96.7 (.18)</td>
<td>1047</td>
<td>96.1 (.36)</td>
<td>1055</td>
<td>95.9 (1.21)</td>
</tr>
</tbody>
</table>

over-coverage decreases as the sample size increases. Comparing with Table 4, we see that the DBL confidence intervals are on average shorter than the symmetric Efron bootstrap confidence intervals for the threshold parameter, e.g. at $n = 1000$, the mean confidence interval width is 0.32 and 0.28 for symmetric Efron bootstrap and DBL, respectively, when the data is simulated from the quadratic model.

3.3 Subsampling with simple rules-of-thumb for block size selection

The heavy computational burden of the double bootstrap-like procedure motivates us to develop alternative methods for block size selection. We start by determining the optimal block sizes for each of the five data generating models we have studied. To do this, we estimate the coverage probabilities of subsampling confidence intervals using $10^4$ Monte Carlo replicates for each $m$ in a grid of block sizes. We then find the $m$ that corresponds to the nominal coverage level by linear interpolation between the two block sizes whose estimated coverage probabilities bracket the nominal level.

The selected block sizes are listed in Table C.1 in the Supplementary Materials. We use a linear model to model the relationship between the Monte Carlo-selected block sizes and the sample sizes, both on the log scale, under correct model specification, and we use a linear mixed effects model to model their relationship under the four misspecified models (Figure 1. We obtain the following relationships:

$$
\log(m) = -0.9207 + 0.9804 \times \log(n), \text{ under correct model specification} \quad (5)
$$

$$
\log(m) = -0.5565 + 0.9961 \times \log(n), \text{ under model misspecification} \quad (6)
$$

Table 6 shows the performance of the subsampling confidence interval method when these rules-of-thumb are used to select block sizes. When the model is correctly specified, the block sizes selected by the rule-of-thumb are on average smaller than those selected by the DBL procedure, which leads the rule-of-thumb confidence intervals to be longer; however, the coverages provided by the rule-of-thumb confidence intervals are actually smaller than those...
Fig. 1: Fitted linear regression lines (linear model for step model and linear mixed effects model with random intercept for non-step models) of the relationship between the sample size $n$ and the block size $m$. The block size $m$ is derived by letting the coverage probabilities be closest to 0.95 for threshold parameter $e$ by subsampling bootstrap.

provided by the DBL procedure. If we restrict to the Monte Carlo replicates for which the DBL method covers and the rule-of-thumb method does not cover, the rule-of-thumb confidence intervals are shorter, but among the Monte Carlo replicates for which both methods cover, the rule-of-thumb confidence intervals are longer. When the model is misspecified, the block sizes selected by the rule-of-thumb are on average bigger. This leads both the confidence intervals to be shorter and the coverage probabilities to be smaller, as we would expect.

3.4 Additional Monte Carlo studies

These simulation results suggest that the symmetric Efron bootstrap confidence interval and the two subsampling confidence intervals using either a double bootstrap-like procedure or a rule-based procedure to select block size all provide reasonable coverage. To investigate the robustness of their performance under a wider variety of scenarios, we conduct two additional Monte Carlo studies. First, we shift the distribution of $X$ so that the threshold $e_0$ is not always at the center of the distribution. The results, summarized in Supplementary Material Section E.1, show that all three methods perform similarly as in the original Monte Carlo study. Second, we let the noise term $\epsilon$ be distributed as a Student’s $t$ distribution with four degrees of freedom instead of a normal distribution. The results, summarized in Supplementary Material Section E.2, show that while the symmetric Efron bootstrap method and the subsampling method using the double bootstrap-like procedure to select block
size perform similarly as in the original Monte Carlo study, the subsampling method with the rule-based method to select block size under-covers when the data is generated from a step model.

Taken together, these results suggest that both the symmetric Efron bootstrap method and the subsampling method with a double bootstrap-like procedure to select block size provide good coverage and can be recommended. These two methods perform similarly when the model is misspecified, but the symmetric bootstrap method performs better, with narrower confidence intervals and closer-to-nominal coverage, when the model is correctly specified. The symmetric bootstrap method is also computationally more efficient than the subsampling method; the subsampling method, on the other hand, is better understood theoretically. The subsampling method with a rule-based procedure to select block size provides a useful alternative when a faster subsampling method is desired, but its coverage may be insufficient when the model is correctly specified and the noise distribution is heavy-tailed.

<table>
<thead>
<tr>
<th></th>
<th>Step</th>
<th>Sig_{15}</th>
<th>Sig_{5}</th>
<th>Sig_{1}</th>
<th>Quad</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>m</td>
<td>cvg (width)</td>
<td>m</td>
<td>cvg (width)</td>
<td>m</td>
</tr>
<tr>
<td>250</td>
<td>89</td>
<td>95.4 (.89)</td>
<td>140</td>
<td>91.3 (.52)</td>
<td>140</td>
</tr>
<tr>
<td>500</td>
<td>176</td>
<td>94.9 (.43)</td>
<td>279</td>
<td>93.3 (.32)</td>
<td>279</td>
</tr>
<tr>
<td>1000</td>
<td>347</td>
<td>95.3 (.21)</td>
<td>557</td>
<td>94.7 (.22)</td>
<td>557</td>
</tr>
<tr>
<td>1500</td>
<td>517</td>
<td>95.0 (.14)</td>
<td>835</td>
<td>94.8 (.18)</td>
<td>835</td>
</tr>
<tr>
<td>2000</td>
<td>686</td>
<td>94.6 (.10)</td>
<td>1112</td>
<td>95.0 (.16)</td>
<td>1112</td>
</tr>
</tbody>
</table>

Table 6: Performance of the prediction rules. m: block size selected by the proposed rules-of-thumb; cvg: coverage probabilities of $e$ estimated from $10^4$ Monte Carlo replicates; width: average width of subsampling confidence intervals.

4 Hydro-climatic drivers of agricultural grassland productivity under extreme drought and rainfall

Efficient methods for estimating thresholds are important in environmental sciences, where complex and large datasets are frequent and abrupt non-linear threshold responses are common. Eutrophication of lake ecosystems Carpenter and Lathrop (2008), fire mediated vegetation transitions in forests and woodlands Adams (2013) and algal responses to light in polar ecosystems Clark et al. (2013) all show some evidence for step threshold responses. However, despite threshold responses being a common feature of natural systems, accurate and efficient techniques for the estimation of thresholds are lacking.

To demonstrate the utility of the fast grid search algorithm developed here for step regression on a real world dataset, we used a remotely sensed derived dataset ($n = 2549$) on grassland productivity responses to soil moisture changes during drought on the Darling Downs, eastern Australia (Plant et al.,
Grasslands and forests are good model systems for investigating thresholds because the roots of vegetation have a discrete physiological limit to the amount of soil moisture they can access. Once moisture levels decline below a critical level, plants can no longer access water and so a rapid decline in plant biomass (possibly leading to plant death) occurs. Given the rapid nature of this change, it is likely to occur as a step threshold (Elmore et al., 2006; Kath et al., 2014). In line with these expectations, Plant et al. (2021) applied Bayesian additive regression trees (BART) to model grassland productivity responses, using the Darling Downs dataset, to soil moisture. Their results suggested a clear step-like response of the rate of grassland productivity change with soil moisture anomalies. The BART approach, while allowing the subjective visualization of a threshold, does not provide an estimate of where that step threshold occurs, nor its uncertainty.

To provide an estimate and quantify the uncertainty of a potential threshold response of grassland productivity change to soil moisture anomalies, we fit a step threshold using the fast grid approach developed here. We use the rule-of-thumb assuming correct model specification (5) because both the regression tree modelling result (Plant et al., 2021, Figure 8) and a three-phase segmented model fit (Figure 2(a) left panel) suggest a very sharp transition.

We account for the influence of 11 other hydro-climatic (e.g. soil moisture and evaporation) and land use (e.g. proportion of woody vegetation and agriculture in the landscape) predictors that may influence the relationship between soil moisture and grassland productivity (Table D.1 in the Supplementary Materials). To account for the nonlinear associations between grassland productivity and these variables, we allow each variable 1-9 degrees of freedom as selected by cross-validated generalized additive models (Wood, 2017).

The right panel of Figure 2(a) shows the step model fit. The change point estimate by the step model is -0.26 with an estimated jump of $2.5 \times 10^{-4}$ (95% symmetric Efron bootstrap CI: $1.9 \times 10^{-4}$, $3.2 \times 10^{-4}$) (Table 7):

$$E(\text{EVI trend}) = \alpha + \alpha_T z + 0.00025 \times (\text{mid soil moisture layer anomaly} + 0.26)$$

The step model therefore quantifies the step threshold shift to a greater rate of grassland productivity decline under drought once soil moisture anomalies exceed a threshold value of $-0.26$ (or a $-26\%$ anomaly in soil moisture). Consistent with the simulation study results under Sig_15, symmetric Efron bootstrap produces slightly shorter 95% confidence interval than subsampling with a DBL procedure for block size selection.

For illustration, we also fit a step model between grassland productivity change to soil moisture anomalies without adjusting for any other covariates. We use the rule-of-thumb assuming model misspecification (6) because a three-phase segmented model fit (Figure 2(b) left panel) clearly shows that the true model is unlikely to be a step model. The change point estimate by the step model is -0.29 with an estimated jump of $6.5 \times 10^{-4}$ (95% symmetric Efron bootstrap CI: $5.8 \times 10^{-4}$, $7.2 \times 10^{-4}$) (Table 7). Consistent with the simulation study results under Sig_1, symmetric Efron bootstrap produces slightly longer
95% confidence interval than subsampling with a DBL procedure for block size selection.

Fig. 2: The grassland example. The solid lines are the fitted curves. The dashed lines are pointwise 95% confidence bands. Subsampling-d and subsampling-r use the double bootstrap-like procedure and the proposed rule-of-thumb, respectively, to select block sizes. In panel (a), “EVI trend partial response” is defined as the observed EVI trend minus the predicted value based on the adjusted covariates.

5 Discussion

Our proposed fast grid search algorithm finds globally optimal solutions to a non-convex, non-smooth problem. It is several orders of magnitude faster than the brute-force grid search algorithm and makes it feasible to analyze large
Table 7: For the grassland example, point estimates and confidence intervals for the threshold $\epsilon$ and jump $\beta$. Subsampling-d and subsampling-r use the double bootstrap-like procedure and the proposed rules of thumb, respectively, to select block size.

<table>
<thead>
<tr>
<th></th>
<th>With covariates adjustment</th>
<th>Without covariates adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\epsilon$</td>
<td>$\beta$</td>
</tr>
<tr>
<td>Point estimate</td>
<td>-0.26</td>
<td>0.00025</td>
</tr>
<tr>
<td>symmetric bootstrap</td>
<td>(-0.29,-0.23)</td>
<td>(0.00019,0.00032)</td>
</tr>
<tr>
<td>Subsampling-dbl</td>
<td>(-0.32,-0.25)</td>
<td>(-0.31,-0.26)</td>
</tr>
<tr>
<td>Subsampling-rule</td>
<td>(-0.32,-0.24)</td>
<td>(-0.30,-0.26)</td>
</tr>
</tbody>
</table>

datasets. We illustrated the use of step regression with a dataset on grassland productivity. Since thresholds are often used to inform targets around which to base environmental management decisions Simmonds et al. (2019), step regression could be applied in a range of natural settings to inform environmental guidelines and regulations (e.g. safe water quality thresholds that are set in freshwater systems, forest restoration targets, etc.). The proposed methods are implemented in the R package *chngpt*, which is hosted on the Comprehensive R Archive Network. The R scripts for simulation studies and real data analysis are available on the Github code repository youyi-fong/StepModelSearchBootstrap.

We studied three resampling-based confidence interval methods under a variety of data generating models. The results support recommendation of the symmetric Efron bootstrap method (symmetric bootstrap) and the subsampling method with a double bootstrap-like procedure (subsampling-dbl) for selecting block size. Both methods have nominal or conservative coverage. The symmetric bootstrap is faster, and produces substantially narrower confidence intervals than subsampling-dbl when the model is correctly specified. We designed our simulation studies to include three models, Sig$_1$, Sig$_5$, and Sig$_{15}$, which increasingly resemble a step model. Interestingly, when the sample size is small ($n = 250$), symmetric bootstrap confidence intervals are also narrower under Sig$_{15}$; this difference decreases under Sig$_5$ and disappears under Sig$_1$. When the sample size is large enough ($n = 2000$), the two methods produce confidence intervals of similar width under all three sigmoid models. On the other hand, the symmetric bootstrap is not as well understood theoretically as subsampling-dbl and may not be as general as subsampling-dbl.

We focused on the threshold parameter in the development of subsampling-based confidence interval methods. Using the block size selected to provide good coverage for the threshold parameter $\epsilon$ is not guaranteed to work for other parameters. This is because different parameters converge at different rates and even parameters with the same convergence rates may require different sample sizes for the asymptotics to kick in. Table C.2 of the Supplementary Materials shows the coverage of all model parameters when the subsampling block size is chosen to optimize the coverage of $\epsilon$. The results show that the coverage for $\beta$ in the Sig$_1$ model is well below the nominal level. To achieve good coverage for parameters other than $\epsilon$, we recommend symmetric Efron bootstrap, which
is fast and provides a coverage between 90-95% depending on the sample size
(Table 4). Alternatively, for a specific parameter, e.g., \( \beta \), subsampling with
a modified double bootstrap-like procedure for targets \( \beta \) may provide improved
coverage at the cost of a higher computational burden.

When there are multiple covariates with threshold effects in a step regression
model, the fast grid algorithm can be generalized by searching through a
multi-dimensional grid of candidate thresholds. A detailed description of the
algorithm for two covariates with threshold effects is shown in Section F of the
Supplementary Materials. It is worth noting that as the number of candidate
thresholds increases, exhaustive search using even the type of fast grid search
algorithm proposed here quickly becomes impractical due to the exponentially
increasing computational burden, which makes heuristic search a necessity.
However, the fast grid search algorithm developed here could become part of
the heuristic search algorithm as an efficient way for providing high quality
solutions to sub-problems.

In this work we have operated under a step regression model that assumes
the observations are independent, the error term and the predictors are in-
dependent of each other, and the predictors are without measurement errors.
The simplicity of this model allows the least squares estimator for the sub-
model with a fixed threshold to have a closed-form solution, which provides
the target for our optimization. Whether the proposed approach to accelerat-
ing computation can be extended to more complex regression models warrants
future research.

Acknowledgment

We are grateful to the Editor and four anonymous referees for their construc-
tive comments. We are grateful to Lindsay N. Carpp for her editorial support.
We thank Professor Jon A Wellner for discussion on the convergence beha-
vior of \( \hat{\beta} \) in the simulation studies. This work was supported by the National
Institutes of Health (R01-AI122991; UM1-AI068635; S10OD028685).

References

Adams, M. A. (2013). Mega-fires, tipping points and ecosystem services: Man-
aging forests and woodlands in an uncertain future. *Forest Ecology and
Management* 294, 250–261.


Bertail, P., D. N. Politis, and J. P. Romano (1999). On subsampling estimators
with unknown rate of convergence. *Journal of the American Statistical
Association* 94(446), 569–579.

n observations: gains, losses, and remedies for losses. *Statistica Sinica* 7(1),
1–31.


Kath, J., K. Reardon-Smith, A. Le Brocque, F. Dyer, E. Dafny, L. Fritz, and M. Batterham (2014). Groundwater decline and tree change in floodplain...


Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supp.pdf