

A CLASS OF BOOTSTRAP BASED RESIDUALS FOR COMPOSITIONAL DATA

Gustavo H. A. Pereira^{1,2} and Jianwen Cai²

¹ *Federal University of São Carlos*, ² *University of North Carolina at Chapel Hill*

Supplementary Material

S1 Proof of theorem 1

In this Appendix, we prove Theorem 1 using the definitions presented in Section 2. Under the correct model, if τ_i is consistently estimated, $y_i^{\{b\}}$, for $b = 1, 2, \dots, B$, and y_i have the same asymptotic distribution. As a result $l_i^{\{b\}}$ and l_i have also the same asymptotic distribution and

$$a_i \xrightarrow{D} U_D(0; B), \tag{S1.1}$$

where $U_D(0; B)$ denotes the discrete uniform distribution with parameters 0 and B .

If $0 < k < 1/(B+1)$, then

$$\begin{aligned} P(u_i \leq k) &= P(u_i \leq k, a_i = 0) = P(u_i \leq k | a_i = 0) P(a_i = 0) \\ &= P(U^{\{1\}} \leq k) P(a_i = 0), \end{aligned}$$

where $U^{\{1\}}$ is a uniformly distributed random variable with parameters 0 and $1/(B+1)$. Therefore, considering (S1.1),

$$P(u_i \leq k) \xrightarrow{P} \frac{k}{1/(B+1)} \frac{1}{B+1} = k. \quad (\text{S1.2})$$

If $1/(B+1) \leq k < 1$, then

$$P(u_i \leq k) = \sum_{j=0}^{\lambda-1} P(a_i = j) + P\left(\frac{\lambda}{B+1} < u_i \leq k, a_i = \lambda\right),$$

where λ is the largest integer less than or equal to $k(B+1)$. Then,

$$\begin{aligned} P(u_i \leq k) &= \sum_{j=0}^{\lambda-1} P(a_i = j) + P\left(\frac{\lambda}{B+1} < u_i \leq k | a_i = \lambda\right) P(a_i = \lambda) \\ &= \sum_{j=0}^{\lambda-1} P(a_i = j) + P(U^{\{2\}} \leq k) P(a_i = \lambda), \end{aligned}$$

where $U^{\{2\}}$ is a uniformly distributed random variable with parameters $\lambda/(B+1)$ and $(\lambda+1)/(B+1)$. Therefore, considering (S1.1),

$$P(u_i \leq k) \xrightarrow{P} \frac{\lambda}{B+1} + \frac{k - \lambda/(B+1)}{1/(B+1)} \frac{1}{B+1} = \frac{\lambda}{B+1} + \frac{k(B+1) - \lambda}{B+1} = k. \quad (\text{S1.3})$$

From (S1.2) and (S1.3), $P(u_i \leq k) \xrightarrow{P} k$, for all $0 < k < 1$ and so

$$u_i \xrightarrow{D} U(0; 1). \quad (\text{S1.4})$$

Using (S1.4) and the Sverdrup's theorem (Sen et al., 2010, page 180),

$$\Phi^{-1}\{u_i\} \xrightarrow{D} \Phi^{-1}\{U(0; 1)\},$$

and so

$$s_i \xrightarrow{D} N(0; 1).$$

S2 Additional tables for Section 3.1

S3 Additional details for Section 3.2

S3.1 Detecting misspecification in the distribution of the response

For evaluating if a parametric regression model properly fits a certain response variable, a normal probability plot with simulated envelope (Atkinson, 1981) is commonly used. If all residuals are inside the envelope or if only a few points are outside the envelope and they are all close to one of the bounds of the envelope, there is no evidence that the model is inadequate. We use a normal probability plot with a simulated envelope instead of performing the AD test, because the residuals computed from a single fitted model are not independent.

There is no general rule in the literature to define how far a residual

Table S1: Average distributional measures for the residuals - $n = 50$.

1	Mean	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	Variance	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Skewness	0.01	0.01	-0.01	0.00	0.02	0.01	0.01	0.01
	Kurtosis	3.00	3.00	3.00	3.01	3.00	2.99	3.00	2.99
	A-D Statistics	1.01	1.01	0.95	0.95	0.86	0.85	0.78	0.78
2	Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Variance	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Skewness	0.00	0.00	-0.01	-0.01	-0.01	-0.02	0.01	0.01
	Kurtosis	2.99	2.99	2.99	2.98	3.00	2.99	2.99	2.99
	A-D Statistics	1.03	1.04	1.01	1.01	1.01	1.01	0.87	0.88
3	Mean	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00
	Variance	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Skewness	0.00	0.00	-0.01	-0.01	0.00	0.00	0.00	0.00
	Kurtosis	3.00	3.01	3.01	3.00	2.99	2.99	2.99	2.98
	A-D Statistics	1.11	1.11	0.83	0.85	1.23	1.24	1.09	1.09
4	Mean	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	Variance	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Skewness	0.01	0.01	0.00	0.00	0.01	0.01	-0.01	0.00
	Kurtosis	3.00	2.99	3.02	3.00	3.00	3.00	3.00	2.99
	A-D Statistics	1.02	1.03	1.10	1.11	0.89	0.89	0.94	0.94
5	Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Variance	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Skewness	0.01	0.00	0.00	0.00	0.01	0.01	-0.01	-0.01
	Kurtosis	2.98	2.97	2.99	2.98	3.00	2.99	2.99	2.99
	A-D Statistics	1.14	1.15	1.06	1.09	1.08	1.07	1.14	1.15

S3. ADDITIONAL DETAILS FOR SECTION 3.2

Table S2: Simulation results Scenario 1a - $n = 20$ and $k = 5$.

i	Mean				Variance				Skewness				Kurtosis				AD Statistic			
	$s_i^{\{a1\}}$	$s_i^{\{q1\}}$	$s_i^{\{a2\}}$	$s_i^{\{q2\}}$	$s_i^{\{a1\}}$	$s_i^{\{q1\}}$	$s_i^{\{a2\}}$	$s_i^{\{q2\}}$	$s_i^{\{a1\}}$	$s_i^{\{q1\}}$	$s_i^{\{a2\}}$	$s_i^{\{q2\}}$	$s_i^{\{a1\}}$	$s_i^{\{q1\}}$	$s_i^{\{a2\}}$	$s_i^{\{q2\}}$	$s_i^{\{a1\}}$	$s_i^{\{q1\}}$	$s_i^{\{a2\}}$	$s_i^{\{q2\}}$
1	-0.01	-0.01	0.01	0.00	1.02	1.02	1.01	1.02	-0.05	-0.02	0.00	0.01	2.96	3.01	2.92	3.03	0.23	0.24	0.23	0.36
2	0.01	0.00	-0.02	-0.02	1.00	1.00	0.99	0.99	0.05	0.05	0.00	0.02	3.10	3.06	3.05	3.10	0.62	0.46	0.96	1.07
3	0.00	0.00	0.02	0.01	1.01	1.02	1.01	1.02	-0.02	-0.01	-0.06	-0.09	2.97	3.00	2.97	2.96	0.24	0.23	0.80	0.76
4	0.01	0.01	0.02	0.02	1.02	1.00	1.03	1.01	-0.02	0.01	-0.04	-0.05	2.88	2.87	2.95	2.90	0.67	0.50	1.07	0.82
5	0.00	0.00	0.01	0.01	1.01	1.00	1.02	1.01	0.08	0.04	-0.11	-0.11	3.15	2.94	3.05	3.01	0.27	0.28	0.65	0.61
6	0.00	0.00	0.02	0.03	0.99	0.98	1.00	0.98	0.05	0.06	-0.02	-0.02	2.99	3.07	3.07	2.98	0.30	0.28	0.97	1.11
7	0.02	0.02	-0.02	-0.02	0.96	0.96	0.95	0.96	-0.04	0.00	-0.08	-0.08	3.13	3.01	3.01	3.05	0.78	0.82	0.81	0.60
8	0.00	0.00	0.02	0.02	0.96	0.95	0.97	0.95	-0.10	-0.04	0.03	0.03	3.17	3.18	3.07	3.13	1.29	1.27	0.61	0.93
9	0.02	0.02	0.01	0.01	0.99	1.00	0.99	1.00	0.01	-0.04	-0.06	-0.08	3.11	3.22	3.06	3.19	0.87	1.10	0.66	0.64
10	-0.02	-0.01	-0.01	-0.02	1.00	1.00	1.00	0.99	-0.04	-0.01	0.04	-0.01	2.90	2.88	2.82	2.84	0.61	0.58	0.65	0.53
11	-0.01	-0.02	0.03	0.03	0.94	0.95	0.95	0.95	-0.08	-0.10	0.00	0.01	2.96	3.05	3.00	3.00	0.84	1.01	1.35	1.34
12	0.04	0.04	0.00	0.00	1.02	1.01	1.03	1.03	-0.01	-0.07	-0.14	-0.15	3.19	3.02	3.01	2.96	2.09	2.43	0.75	0.97
13	-0.01	-0.01	0.01	0.02	1.00	0.99	1.01	0.99	-0.04	0.01	0.07	0.08	2.97	2.89	3.01	2.93	0.43	0.41	0.62	0.62
14	-0.01	-0.01	0.01	0.01	1.04	1.04	1.04	1.04	0.01	0.01	-0.15	-0.15	3.07	3.10	3.00	3.07	0.45	0.57	1.50	1.64
15	0.00	0.00	-0.03	-0.03	1.07	1.07	1.07	1.07	0.00	0.02	-0.03	0.00	2.79	2.82	2.83	2.87	1.42	1.29	2.18	1.98
16	0.06	0.06	-0.01	-0.01	0.93	0.94	0.93	0.95	0.02	0.05	0.02	0.00	3.02	3.07	2.93	3.11	4.30	4.66	0.94	1.25
17	0.04	0.04	-0.02	-0.02	1.00	1.02	1.01	1.02	-0.04	-0.05	0.09	0.12	2.87	2.98	2.86	2.89	2.27	2.54	1.19	1.43
18	0.05	0.04	-0.02	-0.02	0.98	1.00	0.98	0.99	-0.02	-0.03	-0.04	-0.05	2.98	2.97	2.95	2.86	2.30	2.28	0.51	0.58
19	-0.05	-0.05	0.01	0.02	0.99	0.99	0.98	0.99	-0.02	0.00	-0.02	0.00	3.16	2.98	3.03	3.02	2.52	2.83	0.39	0.56
20	-0.01	-0.01	-0.03	-0.02	1.04	1.04	1.05	1.03	-0.05	-0.04	0.00	0.01	2.96	2.99	3.11	2.87	0.59	0.78	1.40	1.37
Mean	0.01	0.01	0.00	0.00	1.00	1.00	1.00	1.00	-0.02	-0.01	-0.03	-0.03	3.02	3.01	2.99	2.99	1.15	1.23	0.91	0.96
SD	0.03	0.03	0.02	0.02	0.03	0.03	0.04	0.03	0.04	0.04	0.06	0.07	0.11	0.10	0.08	0.10	1.05	1.16	0.45	0.43

Table S3: Simulation results Scenario 1a - $n = 20$ and $k = 9$.

i	Mean				Variance				Skewness				Kurtosis				AD Statistic			
	$s_i^{\{a1\}}$	$s_i^{\{q1\}}$	$s_i^{\{a2\}}$	$s_i^{\{q2\}}$	$s_i^{\{a1\}}$	$s_i^{\{q1\}}$	$s_i^{\{a2\}}$	$s_i^{\{q2\}}$	$s_i^{\{a1\}}$	$s_i^{\{q1\}}$	$s_i^{\{a2\}}$	$s_i^{\{q2\}}$	$s_i^{\{a1\}}$	$s_i^{\{q1\}}$	$s_i^{\{a2\}}$	$s_i^{\{q2\}}$	$s_i^{\{a1\}}$	$s_i^{\{q1\}}$	$s_i^{\{a2\}}$	$s_i^{\{q2\}}$
1	0.04	0.04	0.02	0.02	0.96	0.99	0.96	0.99	-0.04	-0.09	0.07	0.08	3.16	2.96	3.00	2.90	2.21	1.72	0.58	0.38
2	0.03	0.03	0.00	0.01	0.98	0.97	0.98	0.97	0.01	0.02	-0.03	0.02	2.94	2.97	2.91	2.97	0.70	0.91	0.19	0.31
3	0.00	0.00	-0.03	-0.04	1.03	1.04	1.03	1.03	-0.11	-0.09	0.04	-0.02	2.97	3.06	2.97	3.06	0.89	0.86	0.95	1.39
4	-0.02	-0.02	0.03	0.03	0.97	0.97	0.98	0.97	-0.06	-0.07	0.02	0.04	2.94	2.83	2.95	2.84	0.55	0.43	0.83	0.91
5	0.02	0.02	-0.02	-0.02	1.06	1.05	1.06	1.05	-0.02	-0.06	-0.05	-0.03	2.96	2.84	2.97	2.85	0.99	1.20	0.97	1.25
6	0.00	0.00	0.02	0.02	0.96	0.96	0.97	0.96	0.03	0.04	-0.12	-0.13	3.01	3.09	3.02	3.01	0.70	0.65	1.14	1.18
7	-0.01	0.01	0.01	0.01	0.98	0.99	0.98	0.99	0.03	0.05	-0.05	-0.02	3.05	2.96	3.02	3.06	0.51	0.51	0.51	0.60
8	0.03	0.04	0.00	0.01	0.95	0.96	0.96	0.97	-0.02	0.00	-0.06	-0.02	2.93	2.82	2.93	2.83	1.27	1.57	0.59	0.47
9	0.01	0.01	0.00	0.00	0.95	0.96	0.97	0.97	-0.16	-0.19	0.04	0.02	3.09	3.20	3.08	3.07	1.59	1.60	0.31	0.30
10	0.03	0.04	-0.02	-0.02	1.08	1.05	1.09	1.06	0.02	0.03	0.01	0.02	3.04	2.95	3.07	3.06	1.07	1.22	1.48	0.92
11	0.02	0.02	-0.05	-0.06	0.99	0.98	0.99	0.98	-0.06	-0.07	-0.02	-0.02	3.00	2.89	2.92	2.81	1.07	0.80	1.93	2.31
12	-0.01	0.00	0.03	0.03	0.95	0.93	0.95	0.93	0.02	0.01	-0.02	0.01	3.17	2.99	3.06	3.06	0.70	0.74	1.55	1.70
13	0.04	0.04	0.00	0.02	1.01	1.01	1.02	1.02	-0.04	-0.08	-0.05	-0.02	2.91	2.95	2.98	2.94	1.62	1.62	0.38	0.42
14	0.02	0.01	0.01	0.01	1.10	1.08	1.09	1.09	0.01	-0.02	-0.08	-0.09	3.06	2.97	3.01	2.97	1.36	1.18	1.55	1.44
15	0.03	0.03	-0.02	-0.01	1.03	1.03	1.04	1.03	0.00	0.01	-0.02	0.00	2.87	2.83	2.92	2.89	1.51	1.22	1.07	0.67
16	0.01	0.02	-0.01	0.00	0.95	0.93	0.95	0.94	-0.07	-0.03	0.01	0.03	3.35	3.09	3.13	3.11	1.25	1.22	0.81	0.77
17	0.00	0.00	-0.03	-0.03	1.03	0.99	1.03	1.00	-0.01	0.01	-0.07	-0.06	3.16	3.00	3.03	2.94	0.41	0.30	0.67	0.68
18	0.00	0.01	-0.01	-0.01	1.00	0.98	1.00	1.00	-0.07	-0.05	0.05	0.02	3.00	2.93	3.02	3.03	0.33	0.41	0.26	0.39
19	0.01	0.01	0.00	-0.01	1.01	0.99	1.01	0.99	0.03	0.03	-0.05	-0.03	3.07	2.97	2.98	2.96	0.30	0.25	0.22	0.23
20	-0.10	-0.10	0.02	0.02	0.99	0.99	1.00	1.00	-0.06	-0.04	0.03	0.05	3.14	3.18	3.10	3.17	7.11	7.11	0.67	0.64
Mean	0.01	0.01	0.00	0.00	1.00	0.99	1.00	1.00	-0.03	-0.03	-0.02	-0.01	3.04	2.97	3.00	2.98	1.31	1.28	0.83	0.85
SD	0.03	0.03	0.02	0.02	0.04	0.04	0.04	0.04	0.05	0.06	0.05	0.05	0.12	0.11	0.06	0.10	1.46	1.45	0.50	0.55

should be from the closest bound of the envelope to lead to rejection of the model. However, for a proper comparison of the ability of the residuals to detect model misspecification, it is important to define a rule for model rejection. Let v be a constant for which, when the correct model is fitted, the probability of at least one residual is outside the envelope and is more than v units away from the closest bound of the envelope is 5%. In practice, with a given data set, v will depend on the sample size n and the dispersion of the residuals.

Consider that g datasets of size n were generated from a certain regression model. Consider also that for each of these datasets, the correct model was fitted, a certain residual r_i was calculated for the n observations, and a normal probability plot with simulated envelope was made. Let e_j be zero if all residuals of the j th datasets are inside the simulated envelope and be the distance from the residual that is farthest from the envelope to the closest bound of the envelope. Using this definition, a natural estimator of v is given by $(e_{[0.95g]} + e_{[0.95g+1]})/2$, where $e_{[j]}$ denotes the j th lowest value of e_j across the g datasets.

To estimate v for the residuals considered in this work, we generated 500 datasets based on model (2.2) with $n = 50$ and $k = 3$. The two covariates were generated from the binomial distribution with parameters n and 0.5

and the Dirichlet distributed response variable was generated according to the set 1 of Distributions presented on Table S4.

Table S4: Probability distributions used in the misspecification analysis.

x_{i22}	x_{i23}	Set 1 of distributions	Set 2 of distributions
1	0	Dirichlet(0.80, 0.15, 0.05, $e^{4.6}$)	Dirichlet(0.40, 0.35, 0.25, $e^{4.6}$)
0	1	Dirichlet(0.05, 0.80, 0.15, $e^{4.6}$)	Dirichlet(0.25, 0.40, 0.35, $e^{4.6}$)
1	1	Dirichlet(0.15, 0.05, 0.80, $e^{4.6}$)	Dirichlet(0.35, 0.25, 0.40, $e^{4.6}$)
0	0	Dirichlet(0.35, 0.35, 0.30, $e^{4.6}$)	Dirichlet(0.30, 0.30, 0.35, $e^{4.6}$)

For each of the datasets, we fitted the correct model, calculated the residuals presented in Section 2, and made normal probability plots with simulated envelope. As the residuals of the proposed class have a similar variance as presented in Table 4, we used the same v for these four residuals. To estimate it, we used the plots for these four residuals and denoted its natural estimator by $v^{\{1\}}$. Residuals (2.3) and (2.4) are a sum of squared univariate residuals, and they have the same variances but are different from those from the proposed class. Therefore, we used the same v for these two residuals and denoted its estimator as $v^{\{2\}}$.

Using the 500 generated datasets, Figure S1 presents a bar plot for the number of points outside the simulated envelope and $v^{\{j\}}$ units or more away from the closest envelope bounds, where $j = 1$ for the residuals of

the proposed class and $j = 2$ for $r_i^{\{q3\}}$ and $r_i^{\{q4\}}$. As expected, the plots are similar and, for the six residuals, about 95% of the datasets do not have any point outside the simulated envelope and $v^{\{j\}}$ units or more away from the closest envelope bounds.

To compare the ability of the residuals to detect model misspecification, we generated 500 additional datasets with $k = 3$ and $n = 50$. The covariates were generated in the same way as in the previous datasets, and the response variable was generated as a mixture of the set of Distributions 1 and 2 (Table S4) with probabilities 0.7 and 0.3, respectively. For each of these datasets, we incorrectly fitted model (2.2). Note that this model is incorrect because the true distribution of the response variable is a mixture of Dirichlet distributions and model (2.2) assumes that the response variable is Dirichlet distributed. For each generated dataset, we also calculated the residuals presented in Section 2, and made normal probability plots with simulated envelope.

Figure S2 presents a bar plot for the number of points outside the simulated envelope and $v^{\{j\}}$ units or more away from the closest envelope bounds when the incorrect model was fitted. The residual $s_i^{\{a1\}}$ of the proposed class can detect model misspecification due to an incorrect response distribution more effectively than the other residuals considered in this work. For this

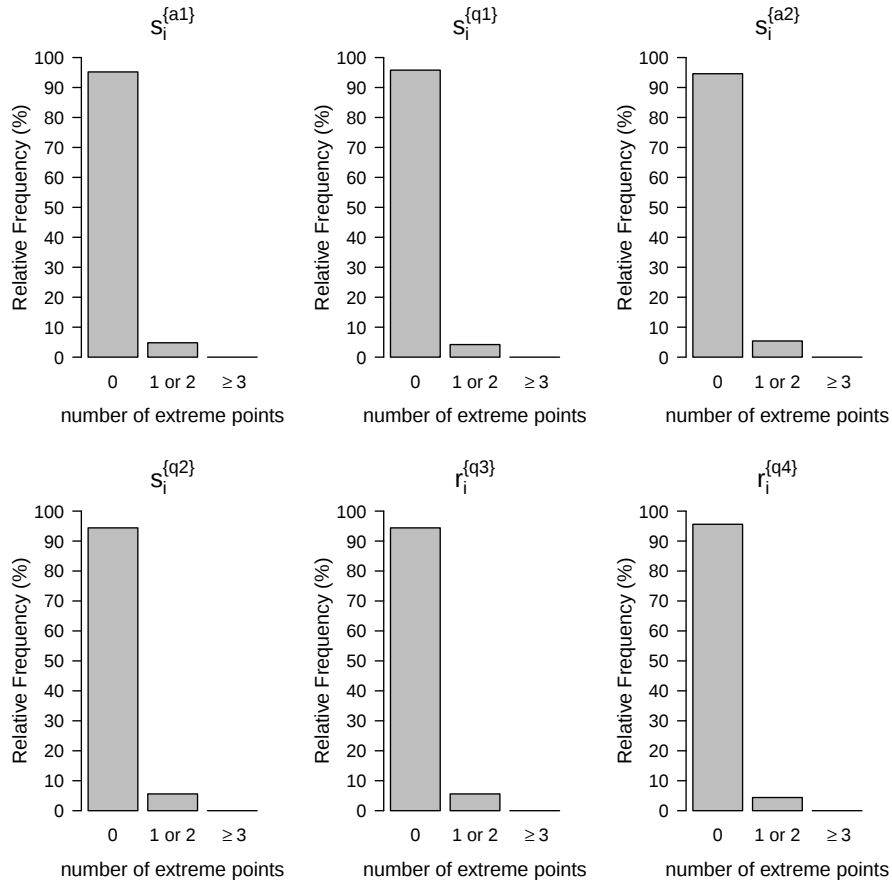


Figure S1: Bar plot for the number of points outside the simulated envelope and $v^{\{j\}}$ units or more away from the closest envelope bounds - Correct model.

residual, only 11% of the datasets do not have any point outside the envelope and distant from the bounds. In addition, using $s_i^{\{a1\}}$, more than 70% of the datasets have at least 3 points outside the envelope and far from the bounds. On the other hand, the other residuals could not detect this kind of model misspecification for more than 20% of the datasets.

S3.2 Detecting misspecification in the linear predictor

To evaluate whether a covariate is missing in the linear predictor, a plot of residuals against the covariate is usually used. Here, we generated 500 samples considering model (2.2) with $n = 50$, $k = 3$, and the same parameter values as in scenario 1b presented in Section 3. For each sample, we fitted the Dirichlet regression model without the second covariate and calculated the values of the six residuals considered in this work.

In practice, we make a plot of the chosen residual against the covariate. However, in this analysis we have 500 samples, and so we need an automatic way to check whether there is a systematic pattern in the plot. Therefore, for each sample and for each of the six residuals, we fitted a simple linear regression in which the response is one of the six residuals and the omitted covariate is the predictor variable. Then we tested whether the parameter associated with the covariate is equal to zero using $\alpha = 5\%$ and the

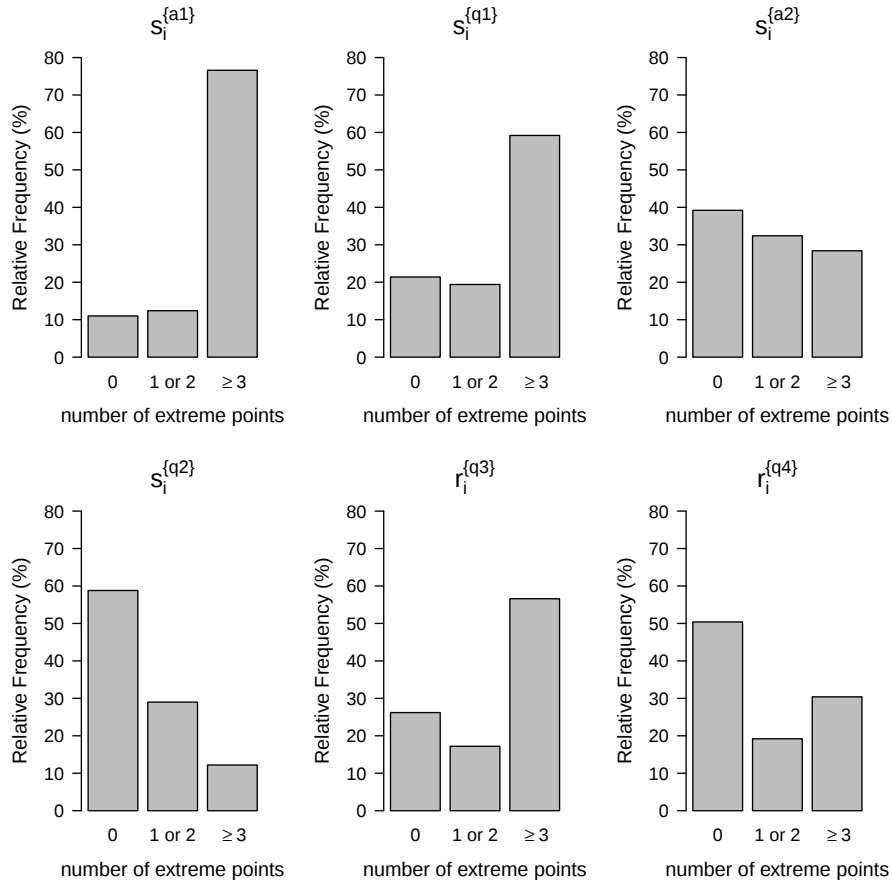


Figure S2: Bar plot for the number of points outside the simulated envelope and $v^{\{j\}}$ units or more away from envelope bounds - Wrong model.

Wald test. A rejection of the null hypothesis in this test indicates that the residual plot has a systematic pattern and that the residual detects model misspecification.

For each residual, Table S5 presents the percentage of samples with a systematic pattern in the residual plot. In the considered scenario, the residuals $s_i^{\{a2\}}$ and $s_i^{\{q2\}}$ detected model misspecification in more than 97% of the samples. On the other hand, the other studied residuals of the proposed class and existing residuals identified model misspecification in less than 45% of the samples. Since $s_i^{\{a2\}}$ and $s_i^{\{q2\}}$ use the same sign function $h_i^{\{2\}}$, this suggests that residuals with this sign function detect much better when an important covariate was omitted from the model than the other studied residuals of the proposed class and the existing residuals.

Table S5: Percentage of samples with a systematic pattern in the residual plot.

Residual	Samples with systematic pattern (%)
$s_i^{\{a1\}}$	42.8
$s_i^{\{q1\}}$	44.2
$s_i^{\{a2\}}$	97.8
$s_i^{\{q2\}}$	97.8
$r_i^{\{q3\}}$	10.0
$r_i^{\{q4\}}$	6.6

Bibliography

Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika* 68(1), 13–20.

Sen, P. K., J. M. Singer, and A. C. P. de Lima (2010). *From finite sample to asymptotic methods in statistics*, Volume 28. Cambridge University Press.