Survival Prediction After Heart Failure

Li Shandross and Scott Hebert

2022-12-02

Abstract

[maximum 200 words]

Introduction

Cardiovascular disease (CVD) accounts for roughly 17,000,000 millions deaths worldwide annually. Current evaluation of the disease progression in various CVDs currently is lacking. Specifically, one example is heart failure – heart failure outcome prediction is of vital importance in clinical practice throughout the world, yet such prediction has not yet yielded promising results. A study by Chicco & Jurman highlighted the potential of machine learning methods to provide physicians with better tools to predict heart failure patient outcomes. The study used data of 299 patients with heart failure. The study included 10 machine learning methods, of which Random Forests performed the best. Afterward, the study authors performed feature selection, which found that serum creatinine and ejection fraction were the best predictors, and a Random Forests model (and one other machine learning model) using only those two predictors outperformed models with all available predictors, which also included age, anaemia, high blood pressure, blood creatinine phosphokinase, diabetes, blood platelets, sex, serum sodium, and smoking status.

This project utilizes the same dataset as Chicco & Jurman, but it utilizes Bayesian logistic regression methods in order to compare those Bayesian methods to the machine learning methods from the reference paper.

Models

First, a Bayesian logistic regression model with all predictors was formulated:

$y_i \sim Bern(\theta_i)$
$logit(\theta_i) = \beta_0 + \beta_1 a_i + \beta_2 m_i + \beta_3 h_i + \beta_4 k_i + \beta_5 d_i + \beta_6 e_i + \beta_7 p_i + \beta_8 x_i + \beta_9 c_i + \beta_{10} s_i + \beta_{11} g_i$

Then, a model with only the two predictors mentioned by the reference paper to be most important (known henceforth as the "reduced model") was created:

$y_i \sim Bern(\theta_i)$
$logit(\theta_i) = \beta_0 + \beta_1 e_i + \beta_2 c_i$

An intercept-only model was created for reference:

$y_i \sim Bern(\theta_i)$
$logit(\theta_i) = \beta_0$

Last, a model with horseshoe priors was formulated as a method of variable selection:

$y_i \sim Bern(\theta_i)$
$logit(\theta_i) = \beta_0 + \beta_1 a_i + \beta_2 m_i + \beta_3 h_i + \beta_4 k_i + \beta_5 d_i + \beta_6 e_i + \beta_7 p_i + \beta_8 x_i + \beta_9 c_i + \beta_{10} s_i + \beta_{11} g_i$
$\beta_0 \sim N(0, 1)$
$\beta_j | \lambda_j, \tau \sim N(0, \lambda_j \tau)$
$\lambda_j \sim C^+(0, 1)$, $j = 1, \cdots, P$
$\tau \sim C^+(0, \tau_0)$ where $\tau_0 = \frac{p_0}{P - p_0} \frac{\sigma}{\sqrt{n}}$
We approximate σ with pseudo variance $\tilde{\sigma}^2 = 1/\mu(1 - \mu)$ for non-gaussian link

where:

ai refers to patient age in years,

mi refers to the presence of anaemia,

hi refers to the presence of high blood pressure,

ki refers to blood creatinine phosphokinase level in mcg/L,

di refers to the presence of diabetes,

ei refers to ejection fraction (percentage of blood leaving the heart upon each contraction),

pi refers to blood platelets in kiloplatelets/mL,

xi refers to sex (M/F),

ci refers to serum creatinine in mg/dL,

si refers to serum sodium in mEq/L,

gi refers to whether the patient smokes

Methods

First, all models were run using samples of the full data set in order to tune the models. Certain aspects needed to be controlled to avoid divergence issues, including having 1,000 iterations in the models and setting the maximum tree depth to 20 (increased from the default value of 10). Then, the models were run on the full data set.

Full model

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: DEATH_EVENT ~ age + anaemia + creatinine_phosphokinase + diabetes + ejection_fraction + high
##    Data: heart (Number of observations: 299)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Population-Level Effects:
##                           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## Intercept                    10.78      5.94    -0.87    22.58 1.00     3127
## age                           0.05      0.02     0.02     0.08 1.00     3530
## anaemia                      -0.01      0.37    -0.72     0.72 1.00     3509
## creatinine_phosphokinase      0.00      0.00    -0.00     0.00 1.00     3426
## diabetes                      0.15      0.36    -0.53     0.86 1.00     3144
## ejection_fraction            -0.08      0.02    -0.12    -0.05 1.00     3148
## high_blood_pressure          -0.12      0.37    -0.85     0.61 1.00     3679
## platelets                    -0.00      0.00    -0.00     0.00 1.00     3876
```

```
## serum_creatinine             0.72      0.20     0.33     1.13 1.00     3513
## serum_sodium                -0.07      0.04    -0.15     0.01 1.00     3159
## sex                         -0.57      0.43    -1.44     0.26 1.00     3013
## smoking                     -0.02      0.44    -0.88     0.82 1.00     2920
## time                        -0.02      0.00    -0.03    -0.02 1.00     3660
##                           Tail_ESS
## Intercept                     2442
## age                           2595
## anaemia                       2902
## creatinine_phosphokinase      1911
## diabetes                      2762
## ejection_fraction             2499
## high_blood_pressure           2709
## platelets                     2893
## serum_creatinine              2846
## serum_sodium                  2618
## sex                           2490
## smoking                       2317
## time                          2969
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Reduced model

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: DEATH_EVENT ~ ejection_fraction + serum_creatinine
##    Data: heart (Number of observations: 299)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Population-Level Effects:
##                   Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept             0.36      0.54    -0.70     1.41 1.00     4017     3055
## ejection_fraction    -0.06      0.01    -0.09    -0.03 1.00     3236     2833
## serum_creatinine      0.78      0.18     0.45     1.17 1.00     4018     2628
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Intercept-only model

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: DEATH_EVENT ~ 1
##    Data: heart (Number of observations: 299)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Population-Level Effects:
```

```
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept    -0.75      0.13    -1.01    -0.51 1.00     1421     1638
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Horseshoe model

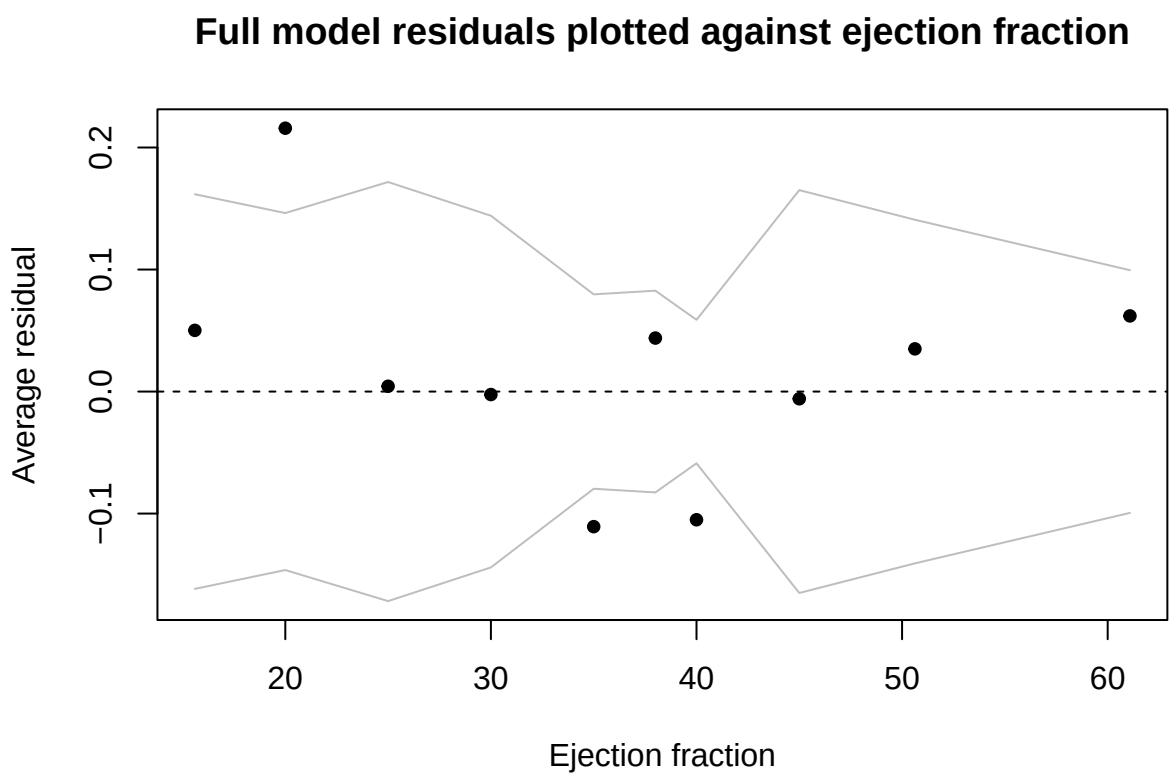
```
##
## Model Info:
##  function:     stan_glm
##  family:       binomial [logit]
##  formula:      DEATH_EVENT ~ .
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 299
##  predictors:   13
##
## Estimates:
##                            mean   sd   10%   50%   90%
## (Intercept)                5.1    4.9  0.1   3.8  12.2
## age                        0.0    0.0  0.0   0.0   0.1
## anaemia                    0.0    0.1 -0.1   0.0   0.1
## creatinine_phosphokinase   0.0    0.0  0.0   0.0   0.0
## diabetes                   0.0    0.1  0.0   0.0   0.1
## ejection_fraction         -0.1    0.0 -0.1  -0.1   0.0
## high_blood_pressure        0.0    0.1 -0.1   0.0   0.1
## platelets                  0.0    0.0  0.0   0.0   0.0
## serum_creatinine           0.6    0.2  0.3   0.6   0.9
## serum_sodium               0.0    0.0 -0.1   0.0   0.0
## sex                        0.0    0.1 -0.1   0.0   0.0
## smoking                    0.0    0.1 -0.1   0.0   0.0
## time                       0.0    0.0  0.0   0.0   0.0
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 0.3    0.0  0.3   0.3   0.4
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##                            mcse Rhat n_eff
## (Intercept)                0.1  1.0  1979
## age                        0.0  1.0  2481
## anaemia                    0.0  1.0  3786
## creatinine_phosphokinase 0.0  1.0  2688
## diabetes                   0.0  1.0  3782
## ejection_fraction          0.0  1.0  3475
## high_blood_pressure        0.0  1.0  3604
## platelets                  0.0  1.0  2950
## serum_creatinine           0.0  1.0  1662
```

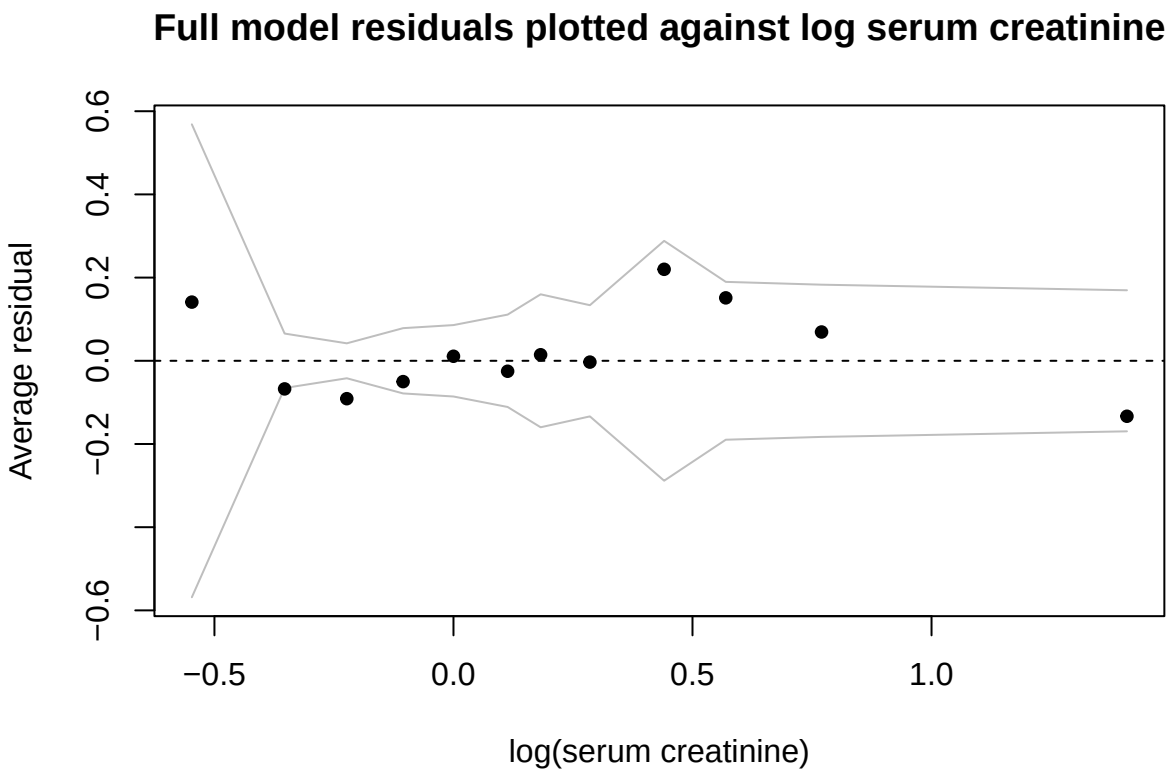
```
## serum_sodium             0.0  1.0  2045
## sex                      0.0  1.0  2458
## smoking                  0.0  1.0  3182
## time                     0.0  1.0  4347
## mean_PPD                 0.0  1.0  3545
## log-posterior            0.3  1.0   527
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

After the models were tuned and finalized, in-sample checks were completed on them. This included binned residual plots plotted against ejection fraction and serum creatinine (the two variables deemed most important by the reference paper the horseshoe prior model). A log transformation was completed on the serum creatinine variable in these plots for readibility.

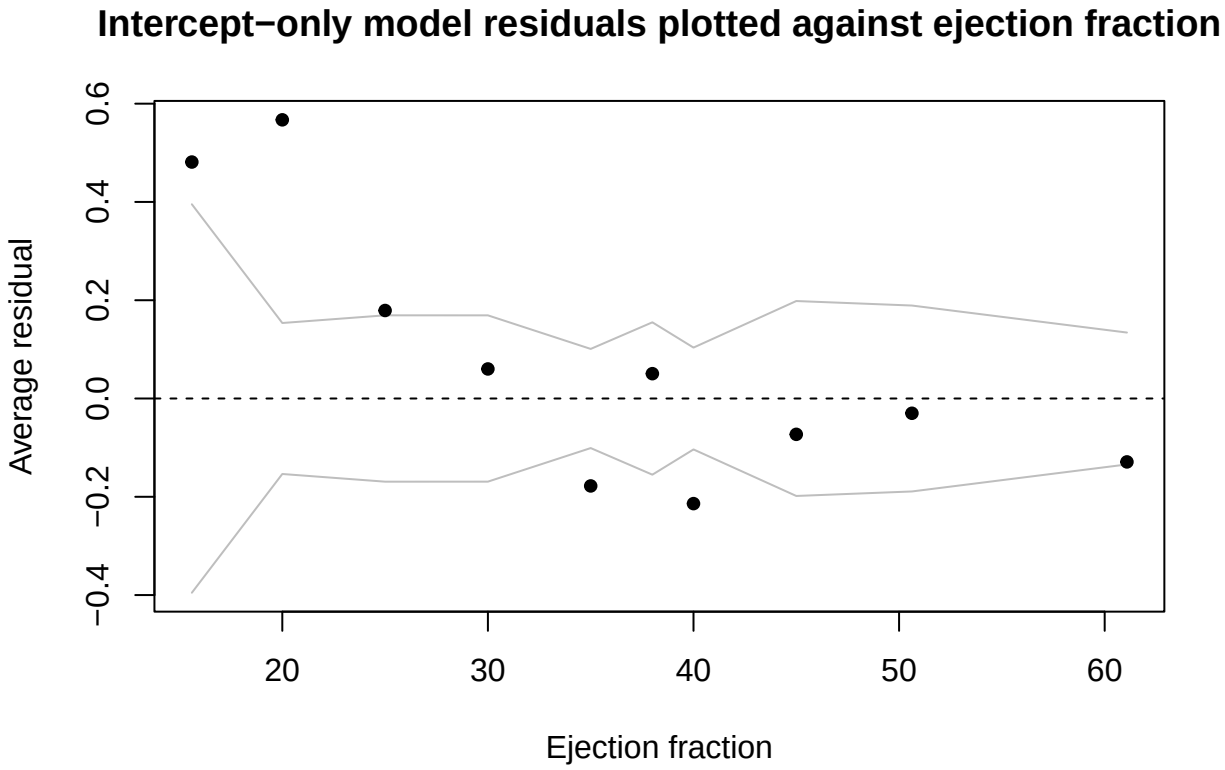
```
# Posterior prediction
ynew_si <- posterior_predict(hs_mod)
# Point estimates for all models
ytilde <- apply(ynew_si, 2, mean)
# Residuals for all models
res <- heart$DEATH_EVENT - ytilde
```

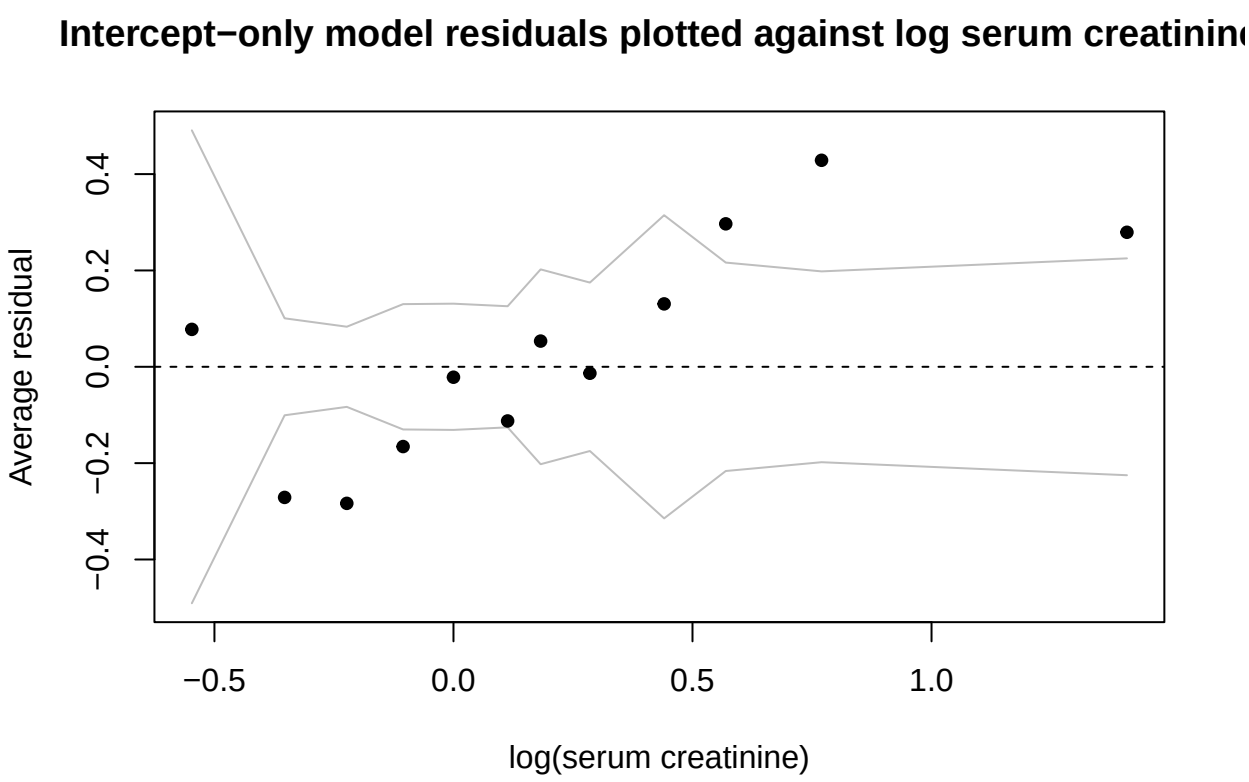
Full model



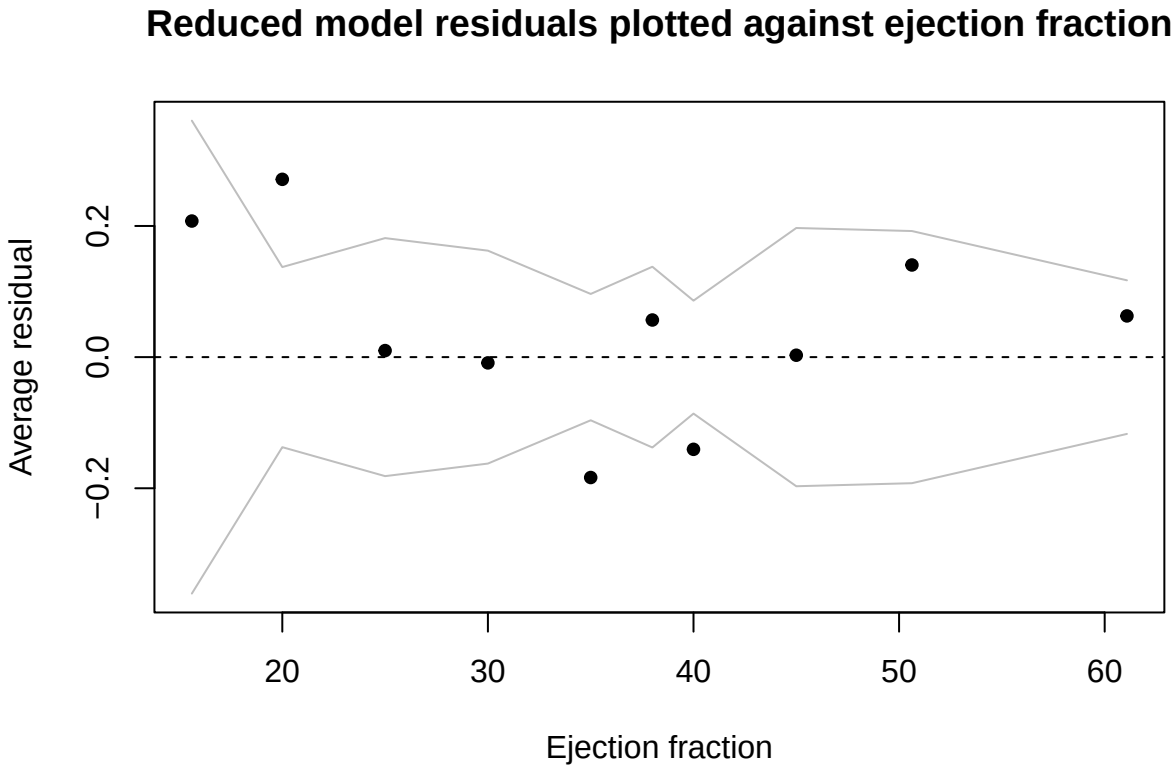

Full model residuals plotted against log serum creatinine
Average residual
0.6
0.4
0.2
0.0
−0.2
−0.6
−0.5
0.0
0.5
1.0
log(serum creatinine)

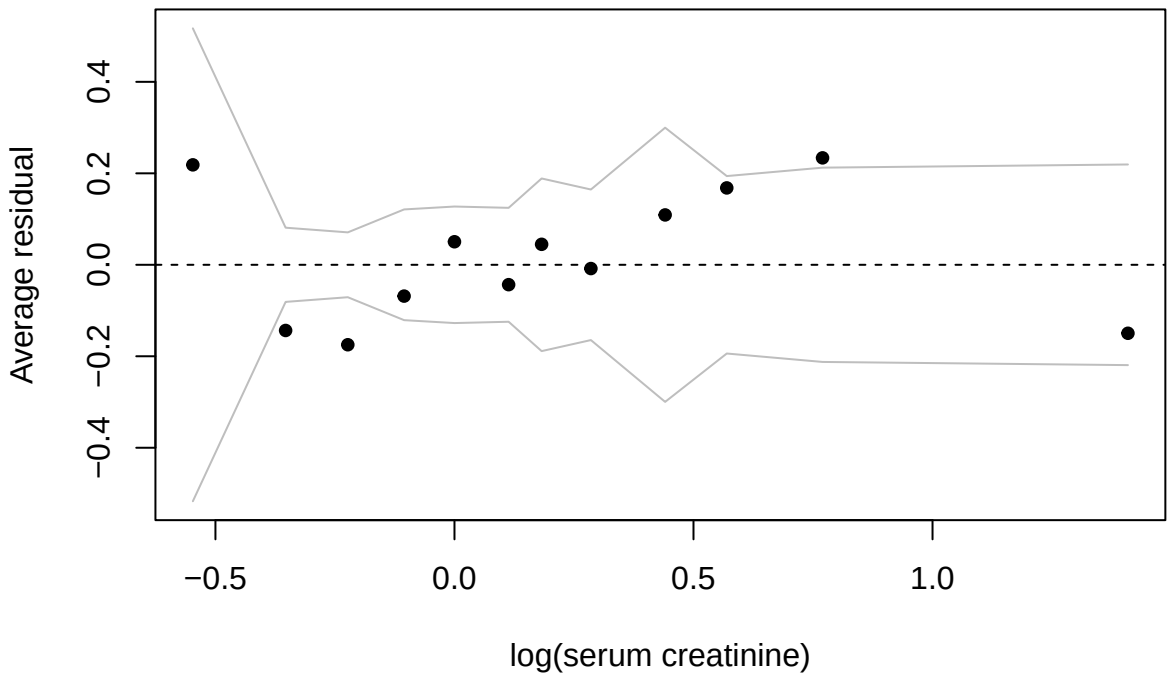
Intercept-only model



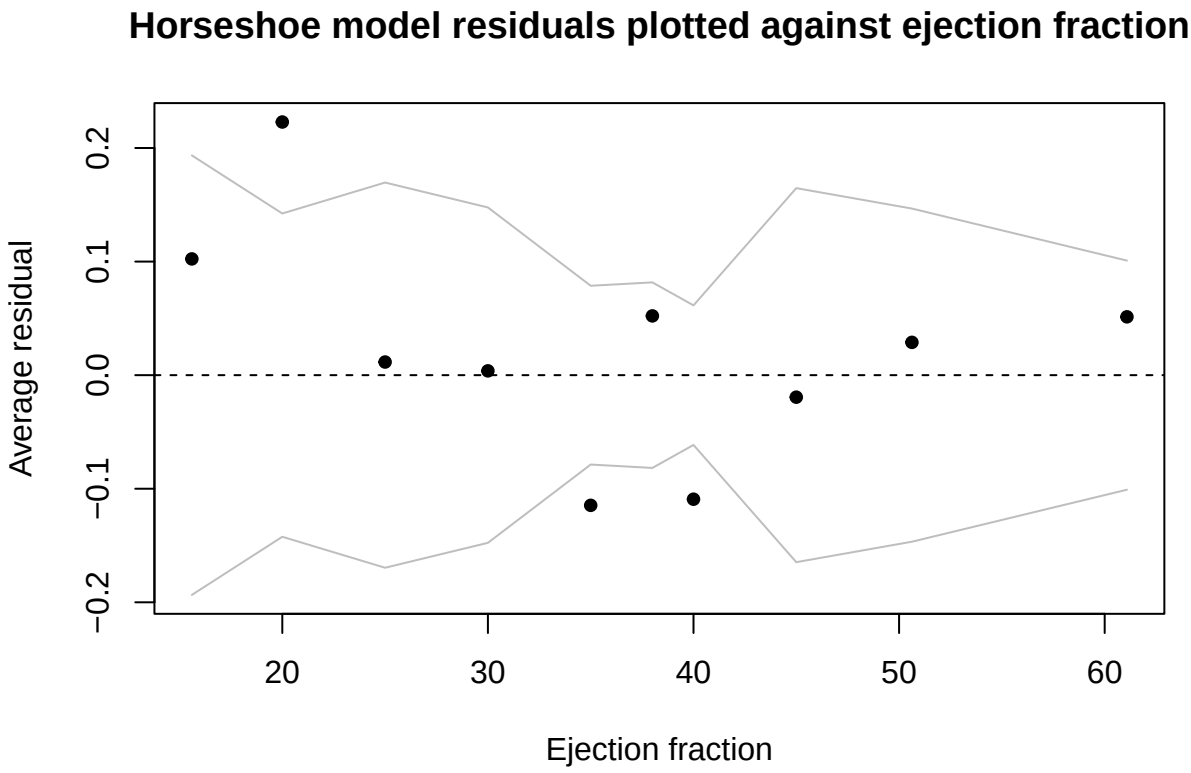

Intercept−only model residuals plotted against log serum creatinin
Average residual
0.4
0.2
0.0
−0.2
−0.4
−0.5
0.0
0.5
1.0
log(serum creatinine)

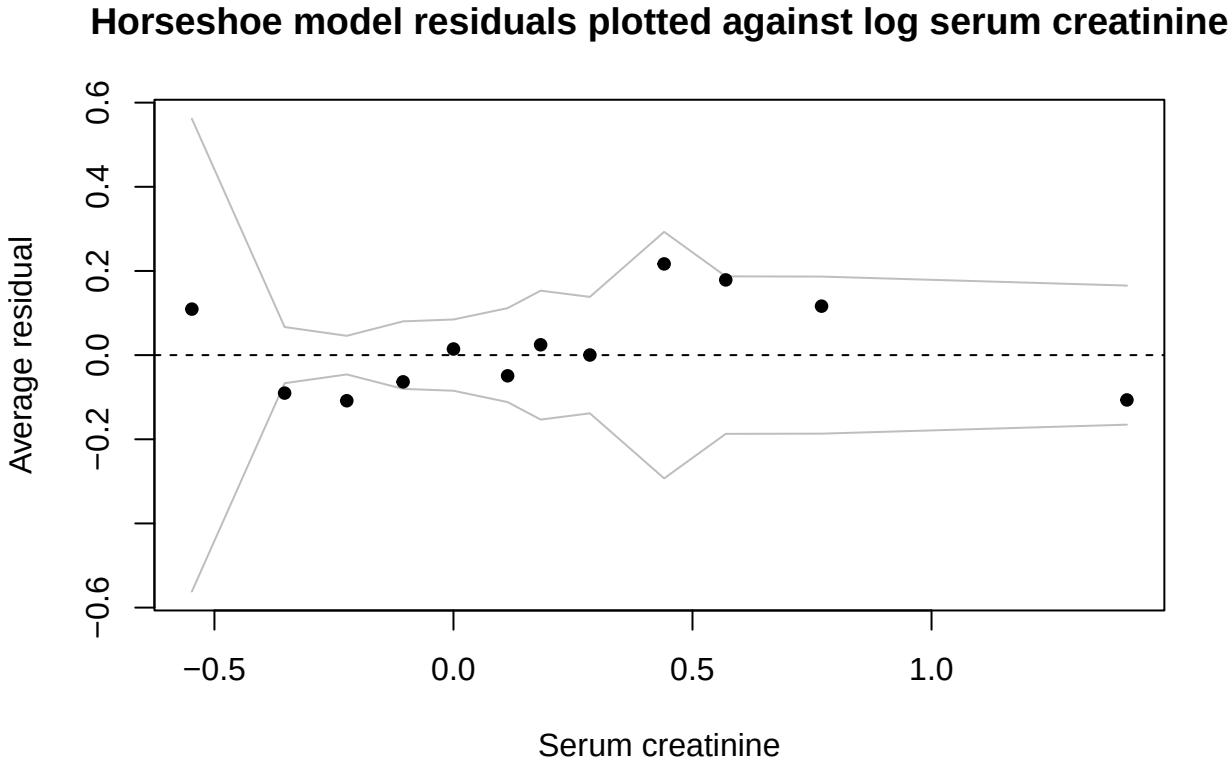
Reduced model




Reduced model residuals plotted against log serum creatinine
Average residual
0.4
0.2
0.0
−0.2
−0.4
−0.5
0.0
0.5
1.0
log(serum creatinine)

Horseshoe model



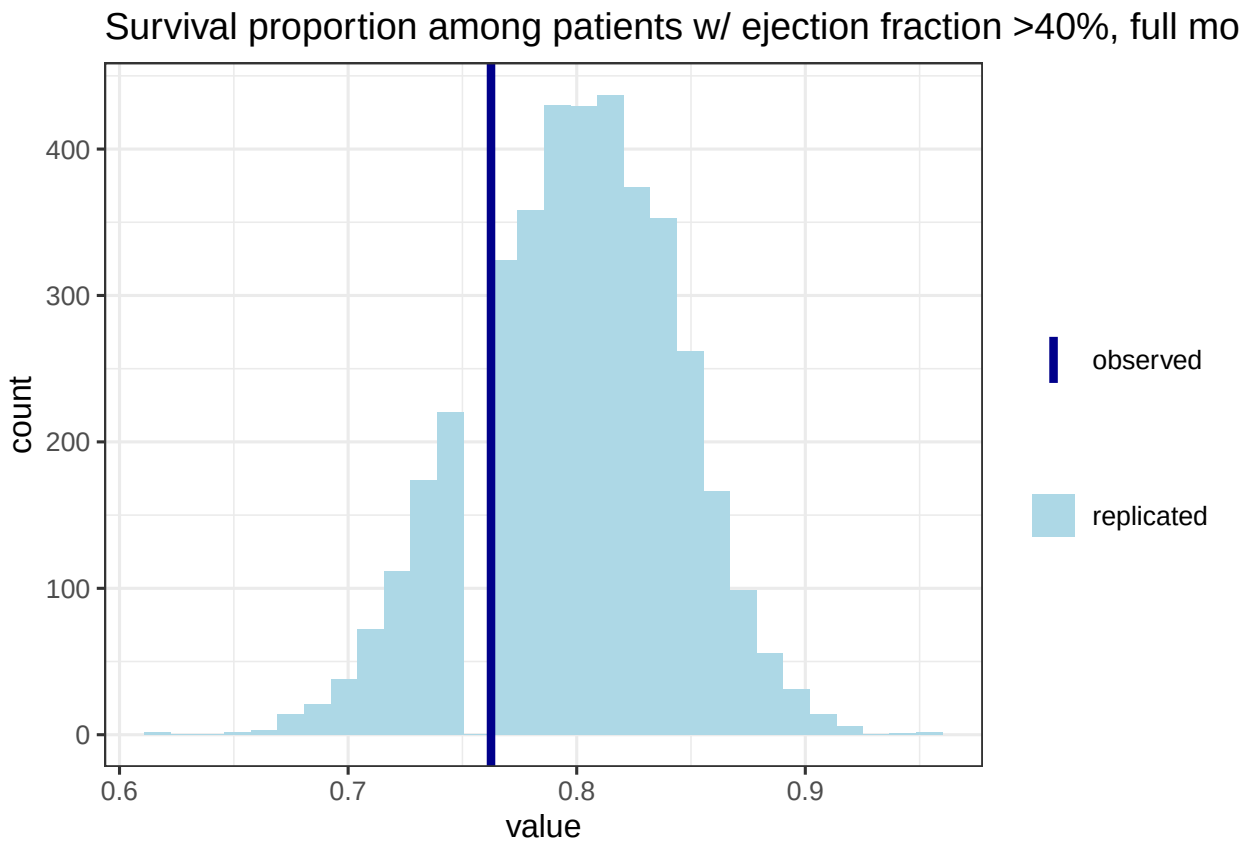


Then, test statistics were created and assessed for all of the models. These statistics included:

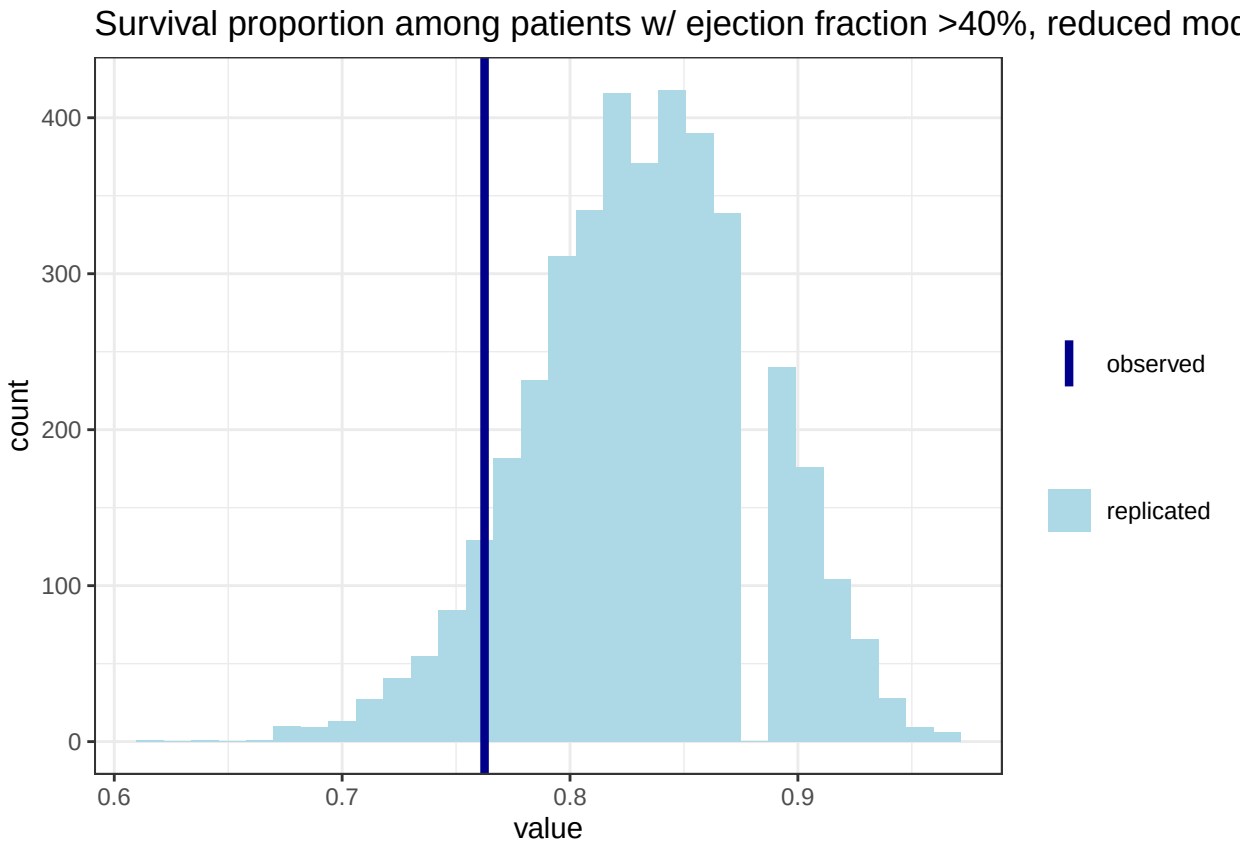
- T1: Proportion of survival
- T2: Proportion of survival among patients with an ejection fraction $> 40\%$ (based on a healthy range discussed in Chicco & Jurman)
- T3: Proportion of survival among patients with serum creatinine < 1.2 mg/dL

Shown here is T2 for all of the models. For that test statistic, the full model has a predictive p-value of 0.836, the reduced model 0.940, the intercept-only model 0.094, and the horseshoe model 0.766. The horseshoe model performed the best in the case of patients with ejection fraction $> 40\%$.

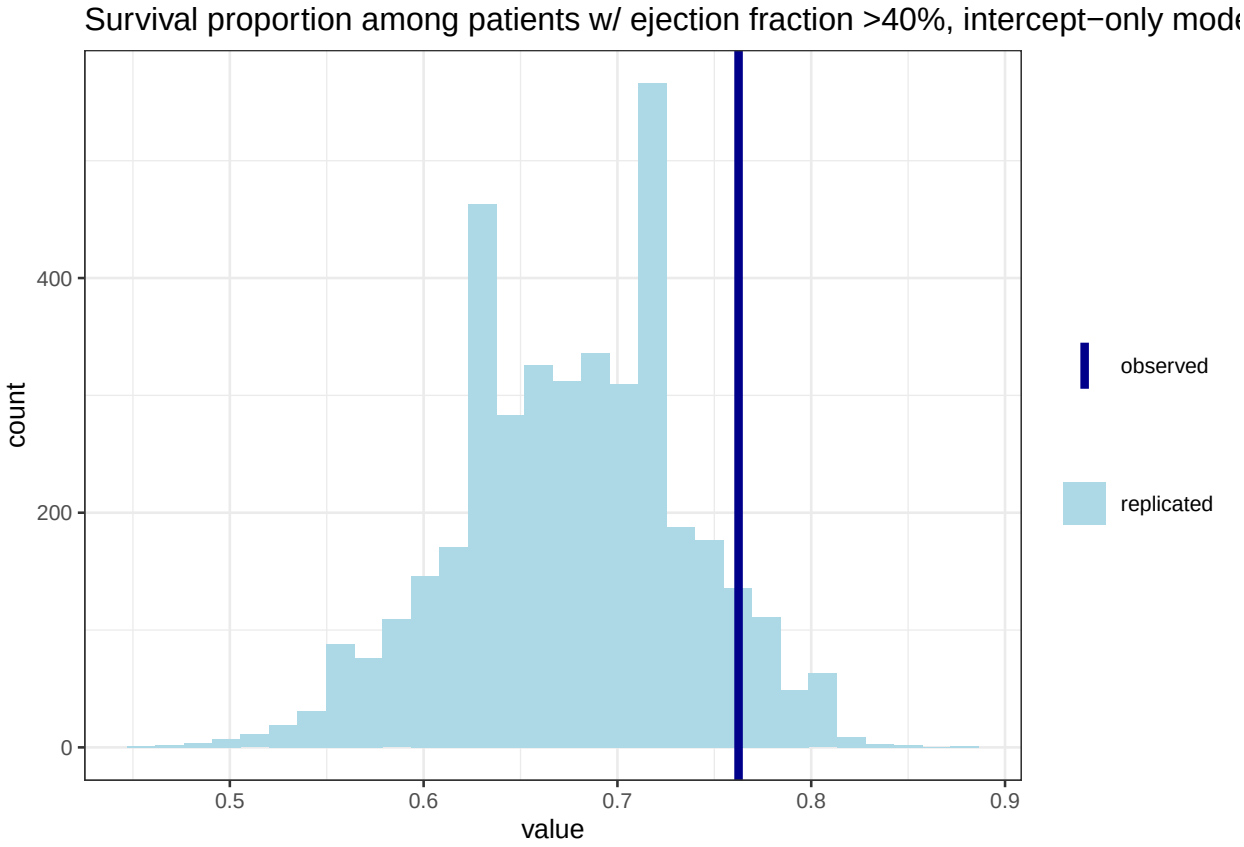
Full model



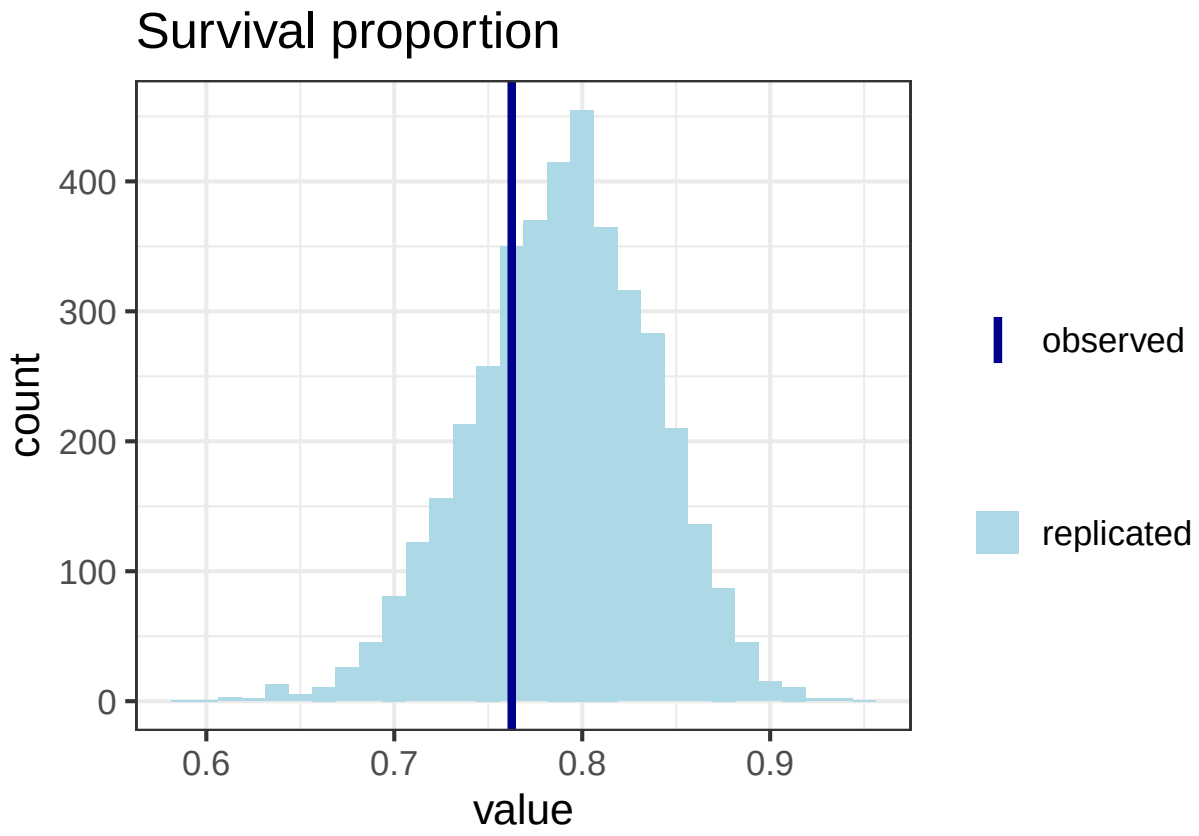
Reduced model



Intercept-only model

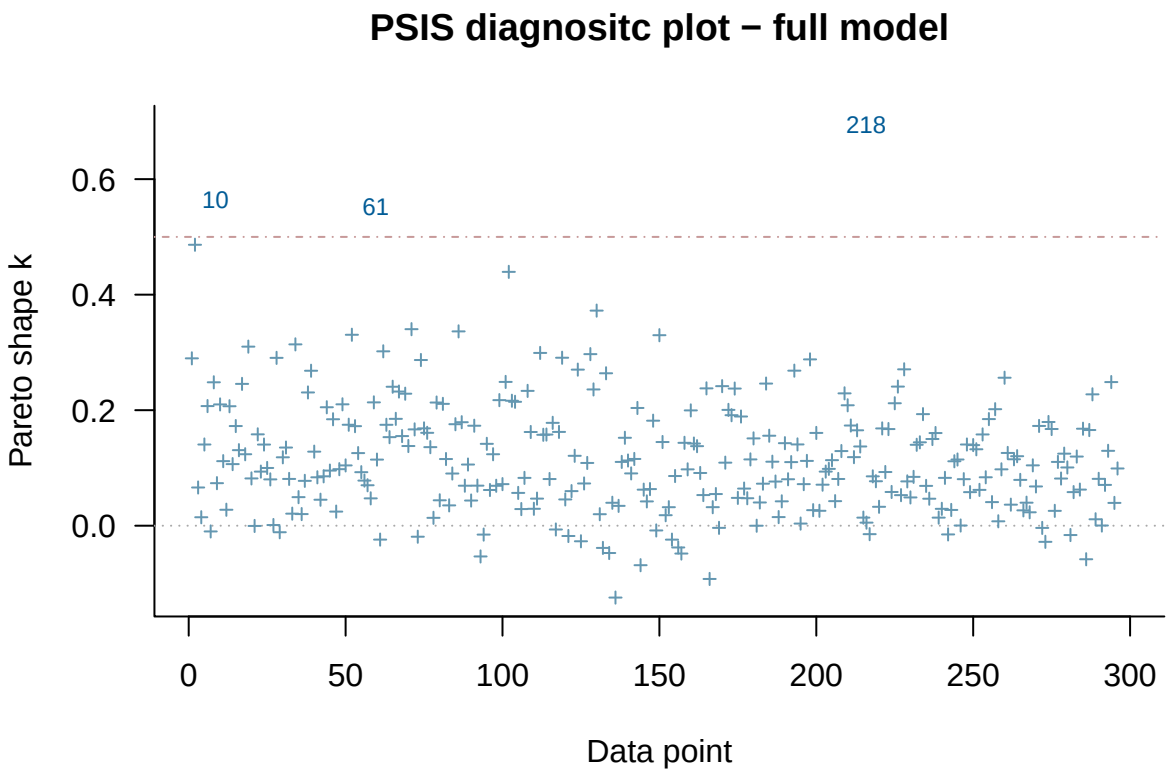


Horseshoe model

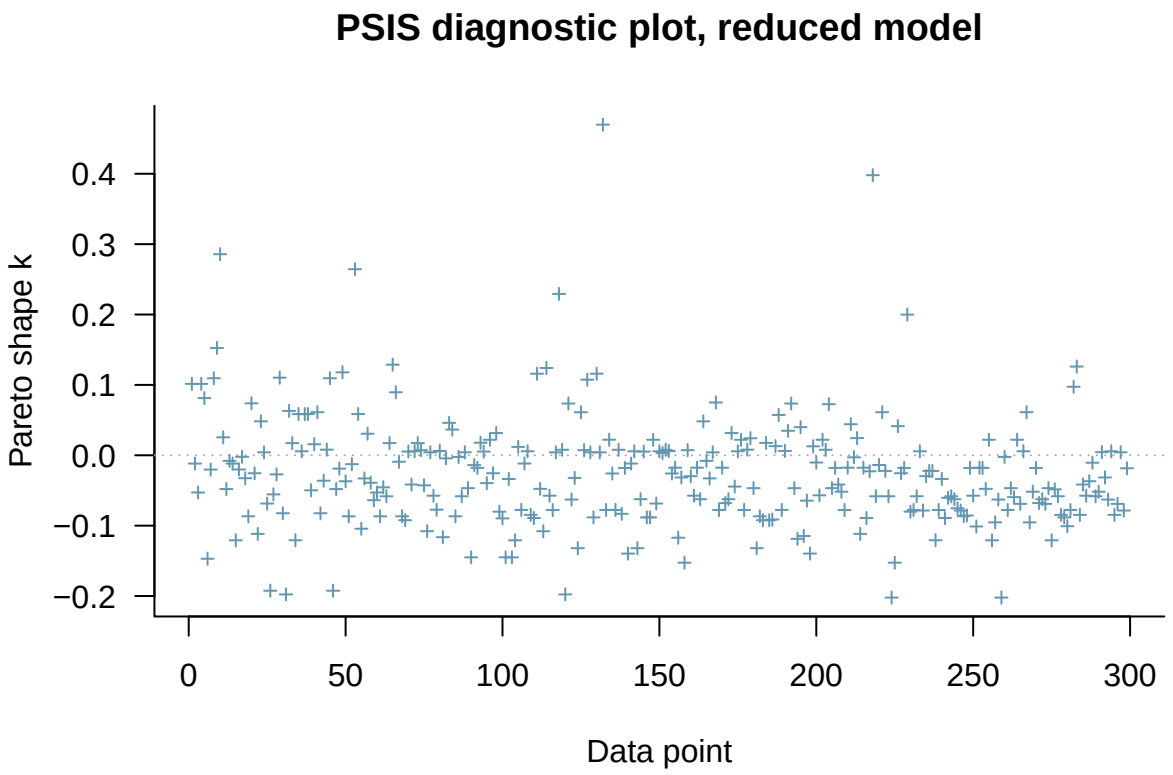


Afterward, out-of-sample model checks were performed. Leave-One-Out (LOO) cross-validation was performed on all of the models. Aside from 3 out of the 299 examples in the full model and 1 example in the horseshoe model which were defined as "okay" (< 0.7), all of the validations ended up with every value being "good" (< 0.5).

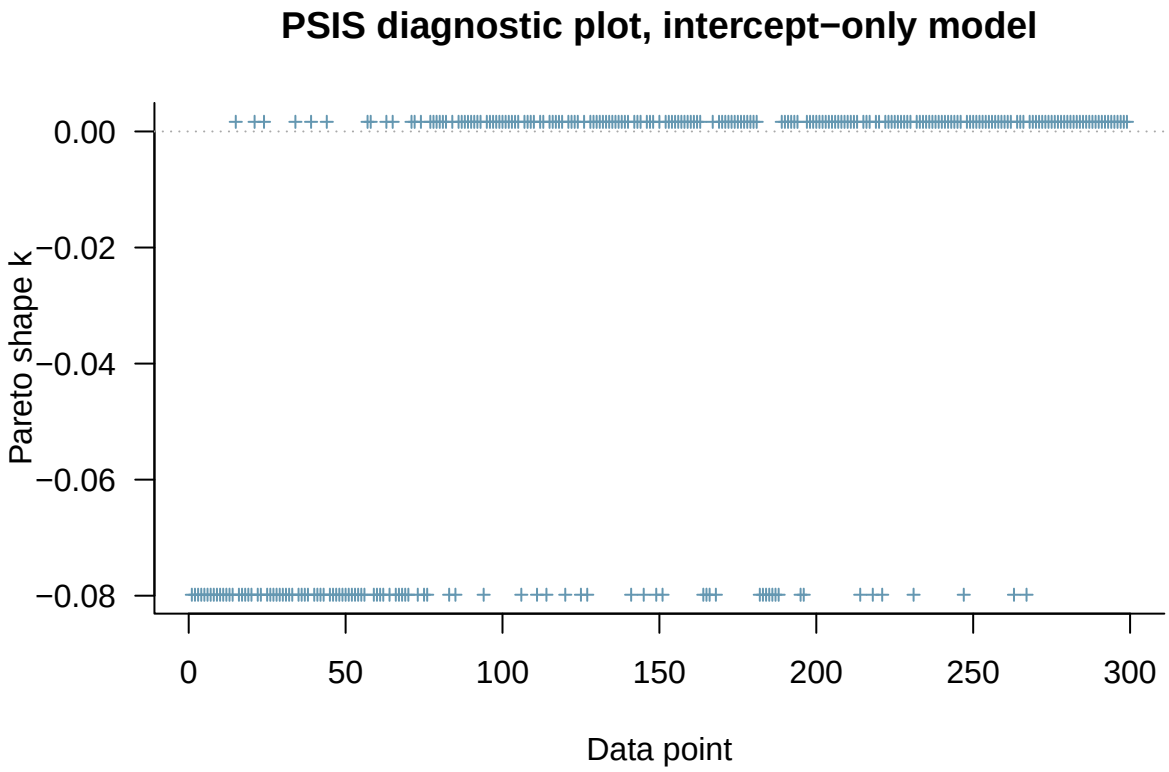
Full model


PSIS diagnositc plot − full model
218
10
61
0.6
0.4
0.2
0.0
Pareto shape k
0
50
100
150
200
250
300
Data point

Reduced model



Intercept-only model



Horseshoe model