

ST 453 Final Project:

A Bootstrap Study of Logistic Regression

for Heart Failure Mortality

Trace Brown

April 21, 2026

Introduction

This project picks up where the midterm left off. In the midterm I fit a logistic regression to the heart failure clinical records dataset, estimated the coefficients with Newton-Raphson, and ran a parametric simulation study that drew synthetic outcomes from the fitted model and re-estimated the parameters across 1,000 replicates. That answered one question: *if the model is correct, how does the MLE behave?* The answer was reassuring. Sampling distributions sat on the true values and 95% Wald intervals hit their nominal coverage.

The final asks something related but different. Instead of assuming the model is correct and simulating from it, I want to see how the MLE actually behaves under the empirical distribution of the real data. That's a nonparametric bootstrap question, and the strategy I'm using here is pairs (or case) resampling, where entire (X_i, y_i) rows get drawn with replacement and the model is refit on each resample. Comparing the resulting bootstrap sampling distribution against the parametric Wald inference from the midterm tells me whether the asymptotic normal approximation actually held up, or whether the midterm CIs were over- or under-stating the real uncertainty.

Data

The data are the Heart Failure Clinical Records dataset from the UCI Machine Learning Repository, available at

<https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records>. The data were originally collected by Ahmad et al. (2017) at the Faisalabad Institute of Cardiology, and the most widely-cited follow-up analysis is Chicco and Jurman (2020). There are 299 patient records, each with 13 variables covering demographic information, blood-test results, and a binary outcome indicating whether the patient died during the follow-up period. The event rate is 32.1% (96 deaths out of 299), which is high enough that the asymptotic approximations behind logistic regression should plausibly hold, but low enough that some bootstrap resamples could end up near separation and would need to be checked for convergence.

Variables of Interest

The midterm exploratory analysis identified four predictors that visually separated survivors from non-survivors most clearly: age (years), ejection fraction (percent), serum creatinine (mg/dL), and follow-up time (days). I kept the same set here so the bootstrap results are directly comparable to the midterm. The substantive relationships I'm studying:

- **Age:** older patients face a higher baseline mortality risk; expected positive coefficient.
- **Ejection fraction:** a measure of how much blood the left ventricle pumps out per beat, with low values indicating heart failure severity; expected negative coefficient.
- **Serum creatinine:** elevated levels indicate kidney dysfunction, which is both a comorbidity and a marker of cardiac strain; expected positive coefficient.
- **Follow-up time:** longer observation windows mostly reflect survivors (a patient who died on day 30 simply cannot accumulate 200 days of follow-up), so this is partly definitional rather than causal; expected negative coefficient.

All four predictors are standardized to mean 0 and standard deviation 1 before fitting, both for numerical stability in Newton-Raphson and so the coefficient magnitudes are directly comparable

on the same scale.

Statistical Model

The model is the same logistic regression as in the midterm:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot \text{Age}_i + \beta_2 \cdot \text{EF}_i + \beta_3 \cdot \text{SC}_i + \beta_4 \cdot \text{Time}_i,$$

where $p_i = P(\text{DEATH_EVENT}_i = 1 \mid X_i)$ and $Y_i \mid X_i \sim \text{Bernoulli}(p_i)$ independently across patients. The parameter of interest is the coefficient vector $\beta \in \mathbb{R}^5$. The likelihood is

$$L(\beta) = \prod_{i=1}^n p_i^{y_i} (1-p_i)^{1-y_i}, \quad \ell(\beta) = \sum_{i=1}^n [y_i x_i^\top \beta - \log(1 + \exp(x_i^\top \beta))],$$

which is concave in β , so the MLE $\hat{\beta}$ is unique whenever the predictors aren't perfectly separable. The 32% event rate and the modest predictor correlations mean separation is very unlikely in either the original sample or any bootstrap resample.

Computational Approach

The MLE is computed via Newton-Raphson, exactly as in the midterm. Starting from $\beta^{(0)} = \mathbf{0}$, the update is

$$\beta^{(t+1)} = \beta^{(t)} + (X^\top W^{(t)} X)^{-1} X^\top (y - p^{(t)}),$$

where $W^{(t)} = \text{diag}(p_i^{(t)}(1-p_i^{(t)}))$ and $p^{(t)} = \sigma(X\beta^{(t)})$. The linear system $(X^\top W X)\delta = X^\top (y - p)$ is solved with the homework-2 Gaussian elimination routine and back substitution, with partial pivoting for numerical stability. Convergence is declared when $\|\beta^{(t+1)} - \beta^{(t)}\|_2 < 10^{-8}$, with a safety cap of 200 iterations that's never actually reached. The covariance matrix $\widehat{\text{Cov}}(\hat{\beta}) = (X^\top \widehat{W} X)^{-1}$ is computed by solving five linear systems, one per column of the identity. No R packages were used.

For the bootstrap study, the same Newton-Raphson routine gets reused on each bootstrap sample. Convergence is fast in practice; the median is around seven iterations on the original data, and the bootstrap samples behave similarly. All 1,000 bootstrap fits converged.

Bootstrap Strategy

I'm using the pairs (case) bootstrap. For each replicate $b = 1, \dots, N$:

1. Set the random seed to b for reproducibility.
2. Draw $n = 299$ row indices $i_1^*, \dots, i_n^*$ independently with replacement from $\{1, \dots, 299\}$.
3. Form X_b^* as the rows of the standardized design matrix at those indices, and y_b^* as the corresponding outcome entries.
4. Refit the logistic regression on (X_b^*, y_b^*) via Newton-Raphson, yielding $\hat{\beta}_b^*$ and an accompanying Hessian-based standard error vector.

Two design choices are worth flagging. First, I standardize the predictors once using the original sample's column means and standard deviations, *not* re-standardizing each bootstrap sample. This keeps the coefficient scale fixed across replicates so that the bootstrap distribution of $\hat{\beta}_j^*$ is directly interpretable as the sampling distribution of $\hat{\beta}_j$ on the original scale. Second, I went with pairs bootstrap rather than residual or parametric. The outcome here is binary, so residuals don't have a clean i.i.d. structure for binary regression, and a parametric bootstrap would basically replay the midterm — generating new y from the fitted model gives back something very similar to the parametric simulation rather than an independent check on it. Pairs bootstrap is also robust to model misspecification, and is the natural choice given how lecture 13 and 14 framed resampling from the empirical joint distribution.

Justification of N

I ran $N = 1000$ bootstrap replicates. The relevant Monte Carlo precision for the percentile bootstrap CIs is the variability of the empirical 2.5% and 97.5% quantile estimates. With $N = 1000$, those quantiles are estimated from order statistics near ranks 25 and 975 respectively, so each tail is supported by about 25 observations. That's plenty for stable percentile boundaries. More formally, the Monte Carlo standard error of the empirical quantile rank is $\sqrt{p(1-p)/N} = \sqrt{0.025 \cdot 0.975/1000} \approx 0.0049$, well below the scale on which the bootstrap density actually varies. For the bootstrap mean, the MC standard error is $SD/\sqrt{N}$, which came out to about 3% of the

empirical bootstrap SD across all five coefficients (specifically: 0.0069 for the intercept, 0.0054 for age, 0.0068 for EF, 0.0089 for SC, and 0.0083 for time). I considered going to $N = 5000$ but the bootstrap distribution is already visually smooth at $N = 1000$ and the Monte Carlo error is small relative to the bootstrap SDs themselves.

Bootstrap Results

All 1,000 bootstrap fits converged. The mean event count across resamples was 96.13, almost exactly matching the original 96, which is what you'd expect under simple random resampling.

Figure 1 shows the bootstrap sampling distributions of the five coefficient estimates. In each panel, the green solid line is the original estimate $\hat{\beta}_j$, the blue dashed line is the bootstrap mean, the purple dashed lines bracket the 95% percentile bootstrap CI, the red dotted lines bracket the 95% Wald CI from the midterm, and the smooth black curve is a normal density centered at $\hat{\beta}_j$ with the bootstrap standard deviation. Three of the five distributions (intercept, age, ejection fraction) look approximately normal and align tightly with the parametric overlay. The serum creatinine distribution, on the other hand, is visibly right-skewed with a longer upper tail, and the follow-up-time distribution is somewhat left-skewed. Both of those shapes are exactly the kind of departure from normality where bootstrap CIs would be expected to improve over Wald.

Table 1 reports the original estimate $\hat{\beta}$, the bootstrap mean, the bootstrap SD, the bootstrap-estimated bias $\overline{\hat{\beta}^*} - \hat{\beta}$, and the 95% percentile bootstrap CI for each coefficient. Bias is small for every coefficient (the largest absolute value is 0.063 for follow-up time, against an estimate of -1.60 , so the bias is under 4% of the magnitude). All five percentile CIs exclude zero, so under inversion all five coefficients are significant at the 0.05 level.

Table 1: Bootstrap coefficient summary, $N = 1000$. CIs are 95% percentile bootstrap.

Coefficient	$\hat{\beta}$	Boot mean	Boot SD	Bias	CI lower	CI upper
Intercept	-1.290	-1.337	0.218	-0.047	-1.835	-0.962
Age	0.515	0.544	0.170	0.029	0.236	0.900
Ejection Fraction	-0.885	-0.931	0.214	-0.045	-1.399	-0.539
Serum Creatinine	0.745	0.786	0.282	0.041	0.331	1.449
Time	-1.600	-1.663	0.262	-0.063	-2.253	-1.211

I computed four flavors of 95% bootstrap CI: percentile, basic (pivotal), normal-approximation,

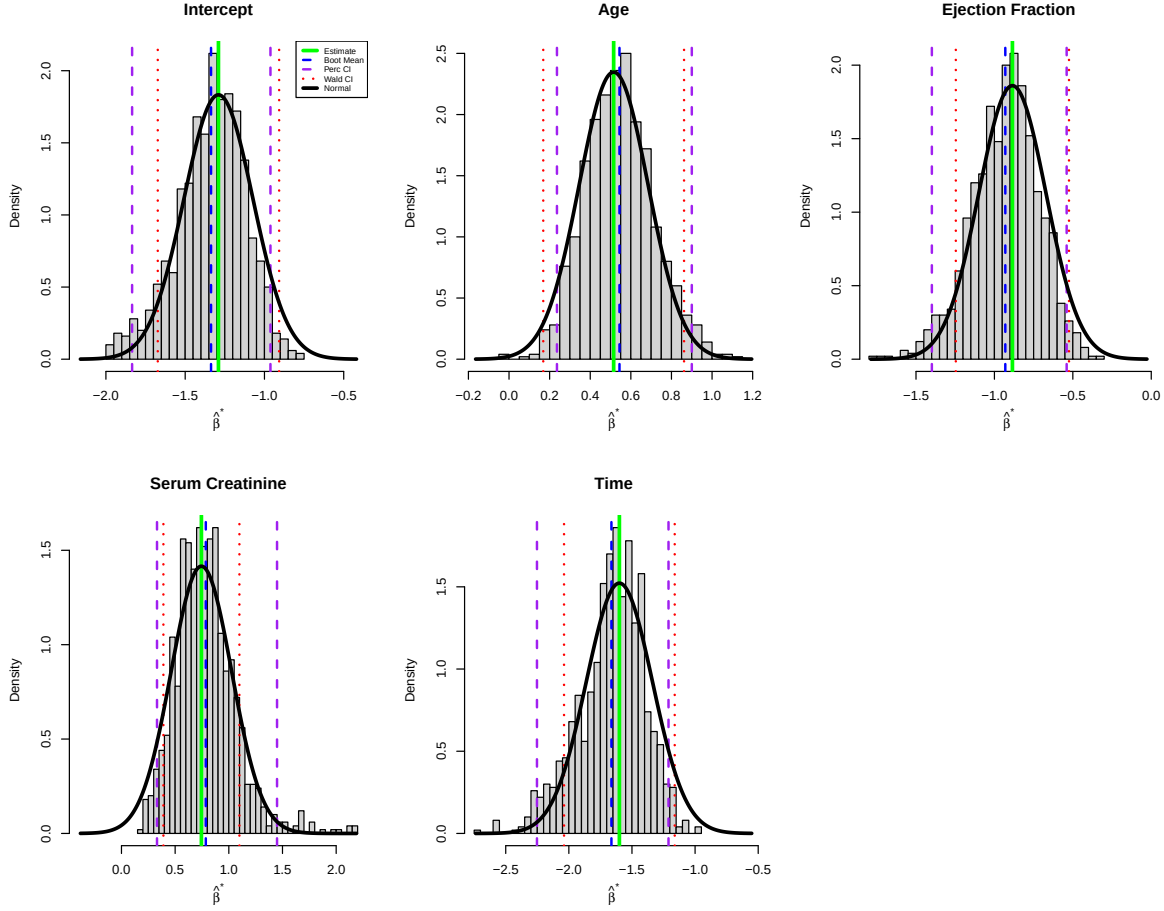


Figure 1: Bootstrap sampling distributions of the coefficient estimates, $N = 1000$. Green solid: original estimate. Blue dashed: bootstrap mean. Purple dashed: 95% percentile bootstrap CI. Red dotted: 95% Wald CI from midterm. Black curve: normal density at $\hat{\beta}_j$ with bootstrap SD.

and BCa (bias-corrected and accelerated). The BCa acceleration constant was estimated by leave-one-out jackknife, which required 299 additional Newton-Raphson fits, each converging in under ten iterations. The bias-correction z_0 values ranged from -0.16 to 0.16 , and the acceleration a values from -0.01 to 0.03 , both small enough that the BCa CIs end up similar to the percentile CIs but not identical. For all five coefficients, all four bootstrap CI types and the Wald CI exclude zero, so the qualitative inference is the same regardless of CI choice. The quantitative differences are where the comparison gets interesting, discussed below.

For hypothesis testing, the bootstrap p-value for $H_0 : \beta_j = 0$ is constructed by inverting the percentile CI: $p = 2 \min(P^*(\hat{\beta}^* \leq 0), P^*(\hat{\beta}^* \geq 0))$, floored at $1/N = 0.001$ since the bootstrap distribution can't resolve p-values below that resolution. Four of the five coefficients have bootstrap

p-values at this floor, meaning no bootstrap replicate landed on the wrong side of zero. The lone exception is age, with a bootstrap p-value of 0.004, which lines up closely with the Wald p-value of 0.0036. The other Wald p-values (intercept, EF, SC, time) are all far below 0.001, and they're likewise floored in the bootstrap output.

Figure 2 summarizes the diagnostic comparisons. The top-left panel shows that bootstrap-estimated bias is small for every coefficient. The top-right panel compares the Wald (Hessian) standard errors to the bootstrap standard deviations — this is where the most interesting finding shows up, and is discussed in the next section. The bottom-left panel compares 95% CI widths across all five methods, and the bottom-right compares Wald and bootstrap p-values.

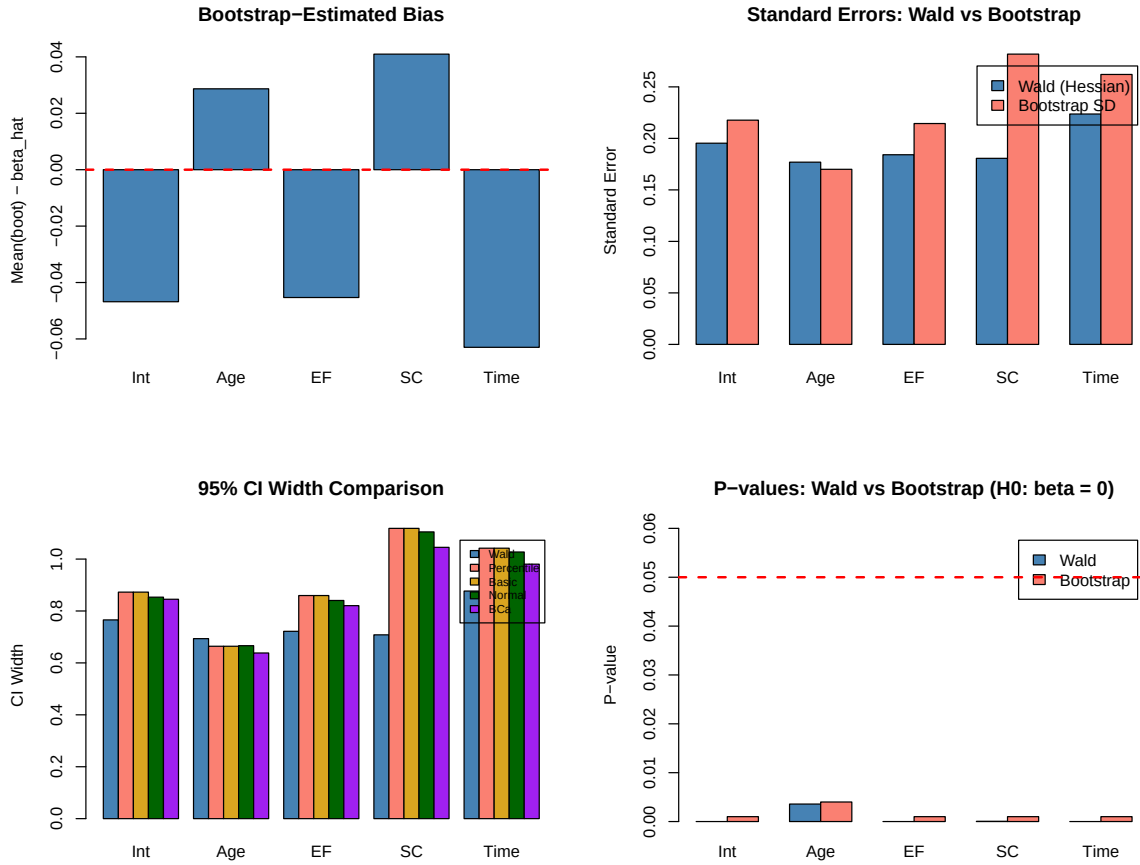


Figure 2: Bootstrap diagnostics. Top left: bootstrap-estimated bias of the MLE. Top right: comparison of Wald SE (Hessian) and bootstrap SD. Bottom left: 95% CI widths across Wald, percentile, basic, normal, and BCa methods. Bottom right: Wald and bootstrap p-values for $H_0 : \beta_j = 0$.

Comparison to Parametric Inference

This is the section the project really exists to support. The midterm delivered Wald confidence intervals based on inverse Fisher information, which depend on the asymptotic normality of the MLE. The bootstrap study above gives me a way to check that approximation against the empirical sampling behavior of the estimator, without having to assume the model is correct.

The headline number is the ratio of bootstrap SD to Wald SE: 1.11 for the intercept, 0.96 for age, 1.16 for ejection fraction, **1.56 for serum creatinine**, and 1.17 for follow-up time. For four of the five coefficients, bootstrap and Wald roughly agree within ten to twenty percent. Serum creatinine is the outlier — the bootstrap SD comes in 56% larger than the Wald SE. That gap is doing real work: it means the midterm Wald CI for serum creatinine, $[0.39, 1.10]$ with a width of 0.71, is materially narrower than the percentile bootstrap CI $[0.33, 1.45]$ with a width of 1.12. The bootstrap is essentially saying that the asymptotic normal approximation is underestimating uncertainty in the SC coefficient. The likely culprit is the SC distribution itself, which has heavy outliers (serum creatinine values go as high as 9.4 mg/dL in this sample, against a mean of 1.4 and a median of 1.1), and a small subset of high-SC patients exerts disproportionate influence on the fit. When those patients get over- or under-represented in a bootstrap resample, the coefficient moves more than the Hessian-based curvature would suggest.

The same pattern shows up in the qualitative shape of the bootstrap distribution. Serum creatinine has a clearly right-skewed bootstrap distribution, which is exactly what an influential-outlier story would predict — the bootstrap occasionally pulls in resamples with extra high-SC patients and produces a long upper tail in the coefficient distribution. Follow-up time shows a milder asymmetry in the opposite direction. For the intercept, age, and ejection fraction, the bootstrap distribution is close to symmetric and Gaussian-shaped, and the Wald CIs match the bootstrap CIs closely.

So the practical takeaway is mixed. For three of the five coefficients, the midterm Wald inference is essentially correct — this study gives some additional confidence in those numbers. For the SC coefficient, the bootstrap suggests the Wald SE is understating the uncertainty, and a more honest CI is the percentile or BCa interval. None of the qualitative conclusions change — every coefficient is significant under either inference framework — but the bootstrap delivers a noticeably wider and

more accurate uncertainty band for serum creatinine.

Implications

The fitted model picks out four clinically plausible drivers of mortality in heart failure patients, each with the sign predicted by domain knowledge. On the standardized scale, a one-standard-deviation increase in age (about 12 years) raises the log-odds of death by roughly 0.52; a one-SD increase in ejection fraction (about 12 percentage points) lowers the log-odds by about 0.89; a one-SD increase in serum creatinine (about 1.0 mg/dL) raises the log-odds by about 0.74; and a one-SD increase in follow-up time (about 78 days) lowers the log-odds by about 1.60. Translated into odds-ratio terms, each one-SD jump in EF roughly halves the odds of death, and each one-SD jump in SC roughly doubles them. Both line up with the medical literature: ejection fraction is a textbook severity marker for heart failure, and renal dysfunction (proxied by elevated creatinine) is a well-established mortality risk factor in this population.

The follow-up-time coefficient is the largest in magnitude, but the interpretation is partly mechanical. Patients who died early in the study simply could not accumulate long follow-up windows, so a long follow-up time is partially a consequence of survival rather than a cause. I wouldn't read the time coefficient as a clinical driver. The other three predictors do support a conventional clinical story.

The bootstrap inference puts these implications on solid empirical footing. Even allowing the model to be slightly misspecified, the data still strongly support the directions and approximate magnitudes of all four clinical effects.

Limitations

The model is a relatively simple logistic regression with four predictors, no interactions, no nonlinear transformations, and no time-to-event structure. That keeps things interpretable and tractable, but several limitations follow from those choices.

First, the dataset has 13 baseline variables and I'm only using four of them. The variables I'm leaving out include diabetes, anemia, smoking, sex, high blood pressure, platelets, serum sodium, and creatinine phosphokinase. Some of these are likely confounders for the variables I am includ-

ing. Omitting diabetes and anemia, which are correlated with both kidney function and cardiac outcomes, could bias the SC coefficient. Adding them back in would likely shrink some of the apparent effect sizes.

Second, the follow-up time issue I flagged earlier is essentially an informative-censoring problem. A proper survival analysis (Cox proportional hazards or an accelerated-failure-time model) would handle this correctly by modeling time-to-death directly rather than dropping time in as a covariate in a binary classifier. The current model conflates patients with short follow-up windows because they died with patients with short follow-up windows because the study ended.

Third, the model assumes the log-odds is linear in each predictor. With age in particular, mortality risk is more plausibly nonlinear — perhaps roughly flat through middle age and then rising sharply after seventy — so a spline or polynomial transformation would probably fit better. I didn't test for nonlinearity here.

Fourth, the bootstrap itself rests on the assumption that the original sample is representative of the population of interest. With 299 patients from a single Pakistani cardiology center, generalization to broader heart failure populations is uncertain. The bootstrap captures uncertainty *about the population the sample was drawn from*, but it can't correct for sampling biases at the recruitment level.

Finally, one limitation specifically about the bootstrap-vs-Wald comparison: the bootstrap can disagree with Wald either because the asymptotic approximation is failing or because the bootstrap is itself unstable. With $N = 1000$ and 299 observations, the bootstrap results look reliable, but the very specific finding about serum creatinine should probably be replicated with $N = 5000$ or $N = 10000$ before being taken to the bank. The discrepancy could plausibly shrink with more bootstrap replicates, even though the convergence diagnostics suggest the percentile quantiles have already stabilized at $N = 1000$.