P9120hw1

Question 2

Methodology

Given p independent normal random variables $X_j \sim N(\mu_j, 1)$, we aim to estimate $\mu_{\max} = \max(\mu_1, \ldots, \mu_p)$ based on a sample of $n = 100$ observations. For each scenario, we generate a training sample and apply the bootstrap procedure to compute the 95% confidence intervals for $\mu_{\max}$.

The bootstrap procedure involves the following steps:

1. Estimate $\mu_{\max}$ by computing the maximum of the sample means for each training set.
2. Generate B bootstrap samples from the training set, where each bootstrap sample is obtained by sampling with replacement from the original data.
3. For each bootstrap sample, recompute the estimate of $\mu_{\max}$.
4. Construct confidence intervals using the three bootstrap methods:
    - **Percentile method**: The interval is formed by taking the 2.5th and 97.5th percentiles of the bootstrap distribution.
    - **Basic method**: The interval is calculated as $2 \cdot \hat{\mu}_{\max} -$ (2.5th and 97.5th percentiles of the bootstrap distribution).
    - **BCa method**: The interval adjusts for bias and skewness in the bootstrap distribution using jackknife estimates.

We simulated data for the following six scenarios:

- **(a)** $p = 2$, $\mu_j = 1$ for all j,
- **(b)** $p = 2$, $\mu_j = j$,
- **(c)** $p = 5$, $\mu_j = 1$ for all j,
- **(d)** $p = 5$, $\mu_j = j$,
- **(e)** $p = 10$, $\mu_j = 1$ for all j,
- **(f)** $p = 10$, $\mu_j = j$.

For each scenario, we generated $M = 1000$ training sets and computed 95% confidence intervals for $\mu_{\max}$ using the three bootstrap methods described. The **coverage rate**—the proportion of intervals that contain the true $\mu_{\max}$—is reported for each method and scenario.

Results

The table below shows the coverage rates for the three bootstrap methods across all scenarios:

Scenario	Percentile	Basic	BCa
(a) p=2, $\mu = 1$	0.892	0.948	0.928
(b) p=2, $\mu = j$	0.946	0.953	0.947
(c) p=5, $\mu = 1$	0.309	0.883	0.686
(d) p=5, $\mu = j$	0.946	0.940	0.943
(e) p=10, $\mu = 1$	0.006	0.820	0.182
(f) p=10, $\mu = j$	0.926	0.920	0.924

Table 1: Coverage rates for the Percentile, Basic, and BCa methods across all scenarios.

Discussion

The performance of the bootstrap methods varies significantly across scenarios: For low-dimensional settings with $p = 2$, all three methods generally perform well, with coverage rates close to the nominal 95% level.
As the dimensionality increases ($p = 5$ and $p = 10$), the performance of the Percentile and BCa methods deteriorates, particularly when μ_j is constant (i.e., scenario (e)). The Basic method, however, maintains reasonable performance even in higher dimensions.
The poor performance of the Percentile and BCa methods in scenario (e) suggests that these methods may not be reliable when the number of variables is large and the signal (i.e., the differences between the μ_j) is weak.
These results highlight the importance of carefully choosing the bootstrap method in higher-dimensional settings, particularly when the variables have similar means.

Conclusion

In this simulation study, we explored the performance of three bootstrap methods for constructing confidence intervals for $\mu_{\max}$. While all methods performed well in low-dimensional settings, only the Basic method maintained satisfactory coverage in higher dimensions. The results suggest that the Percentile and BCa methods may require larger sample sizes or more sophisticated techniques to handle higher-dimensional problems effectively.

Q3 Prostate Dataset Analysis

The prostate dataset is used to analyze clinical and laboratory predictors for prostate-specific antigen (PSA) levels in prostate cancer patients. In this report, we evaluate variable selection and prediction performance using best-subset regression and Lasso regression. Specifically, we choose the best model using cross-validation (CV) and Bayesian Information Criterion (BIC).

Data Preparation

The dataset consists of several predictors including clinical factors such as age, Gleason score, and PSA levels. The data is split into training and test sets based on the 'train' column. All variables are scaled to ensure comparability in the models. The following methods are implemented:

- (a) Best-subset regression with the number of variables selected by 5-fold cross-validation.
- (b) Best-subset regression with the number of variables selected by BIC.
- (c) Lasso regression with λ chosen by 5-fold cross-validation.
- (d) Lasso regression with λ chosen by BIC.

Results

(a) Best-subset Linear Regression with Cross-Validation

The 5-fold cross-validation error was computed for models of various sizes. The plot below shows the cross-validation error as a function of the number of variables included in the model.

(b) Best-subset Linear Regression with BIC
We used the Bayesian Information Criterion (BIC) to select the best model. The plot below shows the BIC values for models of different sizes.

(c) Lasso Regression with Cross-Validation

We performed Lasso regression and selected λ using 5-fold cross-validation. The plot of cross-validation error against different values of λ is shown below.

(d) Lasso Regression with BIC

Using BIC, we selected the optimal shrinkage factor. The plot below shows the BIC values for different shrinkage factors in Lasso regression.

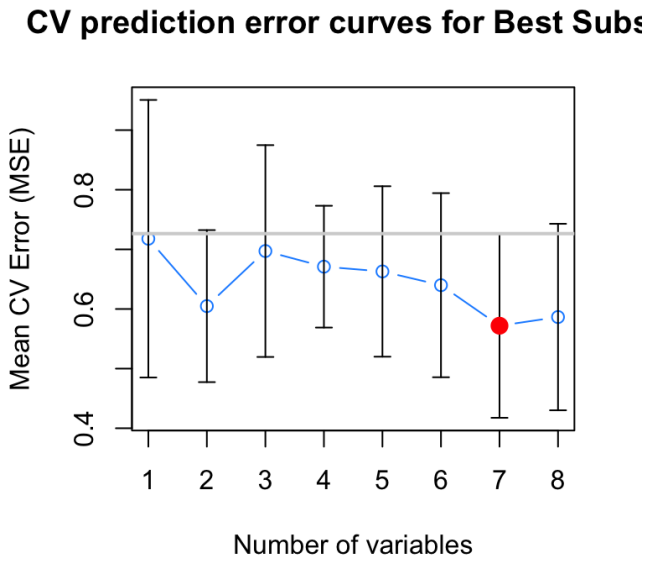

Figure 1: Cross-validation error for Best-subset regression

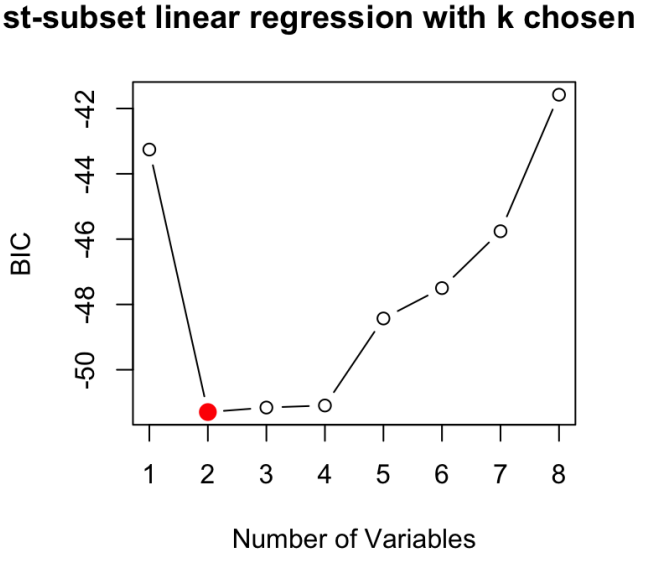

Figure 2: BIC for Best-subset regression

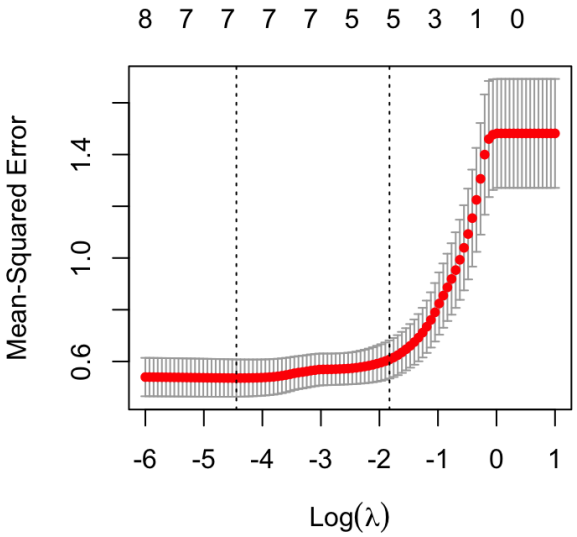

Figure 3: Cross-validation error for Lasso regression

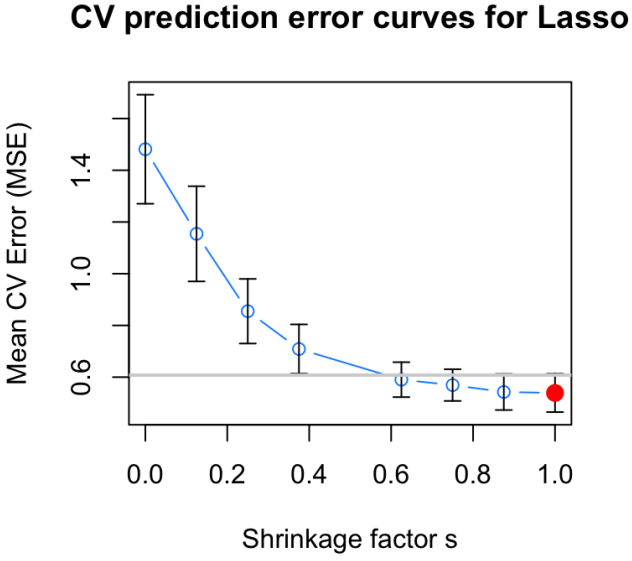

Figure 4: Cross-validation error for Lasso regression

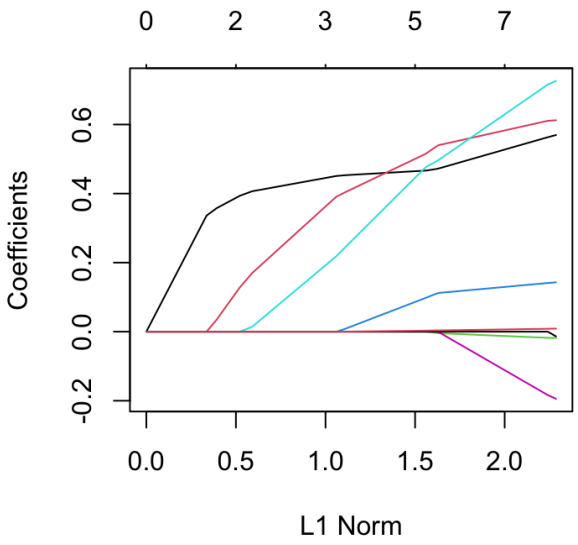

Figure 5: BIC for Lasso regression

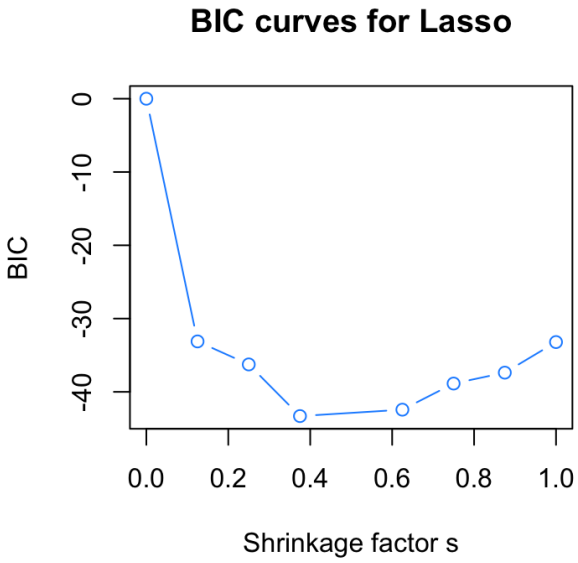

Figure 6: BIC for Lasso regression

Conclusion

In this analysis, we explored two variable selection techniques: best-subset regression and Lasso regression. Both methods were evaluated using cross-validation and BIC to select the optimal model size. Best-subset regression with cross-validation selects a model with 7 variables, while BIC prefers a model with 2 variables. Similarly, Lasso regression with cross-validation selects more complex models, but using the 1se rule or BIC leads to more parsimonious models.

Further investigation could involve comparing the prediction accuracy of the different models on the test dataset to evaluate their out-of-sample performance.