Yuchen Huang

MA677 Final Project

Chapter7 James–Stein Estimation and Ridge Regression (CASI)

Main Points, Comments, and Questions

- The James–Stein Estimator
  - The James-Stein estimator: The James-Stein estimator is a seminal empirical Bayes method. $\hat{\mu}_i^{\text{JS}} = \hat{M} + \hat{B}\left(x_i - \hat{M}\right) \qquad \text{for } i = 1, 2, \ldots, N,$ Suppose that $x_i|\mu_i \sim \mathcal{N}(\mu_i, 1)$ independently for i=1, 2, 3, 4, …, N with N >= 4. Then
    $$E\left\{\|\hat{\boldsymbol{\mu}}^{\text{JS}} - \boldsymbol{\mu}\|^2\right\} < N = E\left\{\|\hat{\boldsymbol{\mu}}^{\text{MLE}} - \boldsymbol{\mu}\|^2\right\}$$
    It is known for its counterintuitive property that "shrinking" individual estimates towards a common mean can result in better overall estimation accuracy than treating each estimate separately.
  - Mathematical Basis: The estimator is a mathematical derivation, demonstrating how it outperforms simple unbiased estimators under mean squared error criteria.
  - Comment: It's a game-changer for statistical thinking, pushing forward the development of more advanced shrinkage methods. Think of it as the first step towards smarter statistics.
- The Baseball Players
  - It uses data from Major League Baseball players to illustrate the practical application of the James-Stein estimator, addressing how batting averages are better predicted using this method.
  - Comment: The performance improvements over naive estimates are quantified, showcasing the power of empirical Bayes methods in real-world data scenarios. The example helps clarify the theoretical advantages of shrinkage, particularly in cases with small sample sizes or incomplete information.
- Ridge Regression
  - Ridge regression: It is an extension of the shrinkage idea in the context of regression analysis. This technique addresses multicollinearity issues and improves prediction accuracy. Also, ridge regression can be seen as an empirical Bayes method under certain prior distributions.
  - Ridge estimator: A ridge regression estimate $\hat{\beta}(\lambda)$ is defined, for $\lambda \geq 0$, to be
    $$\hat{\beta}(\lambda) = (\boldsymbol{S} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{S} + \lambda \boldsymbol{I})^{-1}\boldsymbol{S}\hat{\beta}$$
    $\hat{\beta}(\lambda)$ is a shrunken version of $\hat{\beta}$, the bigger λ the more extreme the shrinkage.
  - Lasso estimator: it is a widely used penalty term:
    $$\tilde{\beta}(\lambda) = \arg\min_{\beta}\{\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda\|\beta\|_1\},$$

- Indirect Evidence 2
  - This part talks more sophisticated empirical Bayes methods that incorporate indirect evidence from related datasets or parameters. Examples include studies from genetics and medical statistics where pooling information across different groups leads to more robust inferences.
- Notes and Details
  - Detailed notes provide additional mathematical and statistical insights into the methodologies discussed earlier in the chapter.
  - The limitations and criticisms of empirical Bayes approaches are also discussed, providing a balanced view of their applicability and potential pitfalls.

Application of the topic in R

Ridge Regression(lasso):

Consider the "Tracheostomy" dataset from myMA67 project, which consists of information on patients who have undergone tracheostomy procedures. The dataset includes a binary outcome, where '0' indicates that a patient was not readmitted, and '1' indicates readmission. In this example, I employed lasso regression to develop a model that predicts the likelihood of patient readmission. The training data Tracheostomy is from year 2018 and 2019, and the test data T20 is from year 2020.

~~~~
```{r}

## First, I made data preparation for lasso regression. Here T_x is used for predictors, and T_u is used for binary response variables.

T_x <- data.matrix(Tracheostomy[, c("DIED", "AGE", "FEMALE", "LOS", "TOTCHG", "ZIPINC_QRTL", "RESIDENT", "PAY1", "DMONTH", "Diagnosis")])

T_y <- Tracheostomy$readmit_flag

##  `cv.glmnet` is used to perform cross-validation for the Lasso regression model.

## The parameter alpha = 1 specifies the use of the Lasso penalty (as opposed to Ridge regression or Elastic Net). This function will help in determining the best lambda (regularization parameter) that minimizes prediction error.

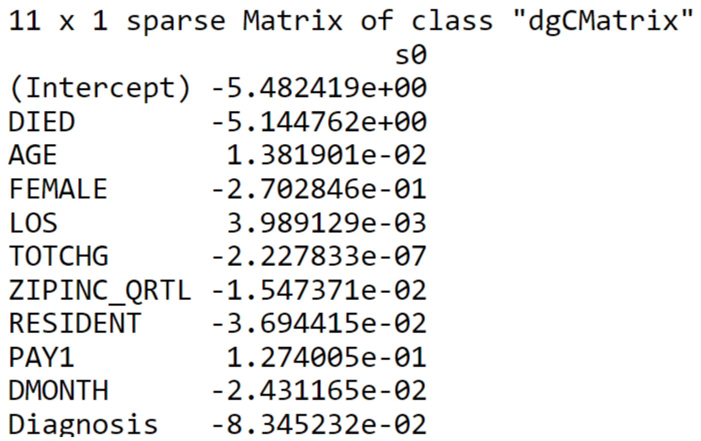
T_cvfit <- cv.glmnet(T_x, T_y, family = "binomial", alpha = 1)

T_bestlambda <- T_cvfit$lambda.min

## Fits the Lasso logistic regression model using the optimal lambda (T_bestlambda) obtained from cross-validation. This model is expected to be the best performing model in terms of balancing bias and variance.

T_lasso <- glmnet(T_x, T_y, family = "binomial", alpha = 1, lambda = T_bestlambda)
~~~~

coef(T_lasso)



```
11 x 1 sparse Matrix of class "dgCMatrix"
                        s0
(Intercept) -5.482419e+00
DIED        -5.144762e+00
AGE          1.381901e-02
FEMALE      -2.702846e-01
LOS          3.989129e-03
TOTCHG      -2.227833e-07
ZIPINC_QRTL -1.547371e-02
RESIDENT    -3.694415e-02
PAY1         1.274005e-01
DMONTH      -2.431165e-02
Diagnosis   -8.345232e-02
```

In this analysis, I used the T20 variable from the dataset to make predictions about patient readmissions. To assess the accuracy of these predictions, I employed a confusion matrix, which provided a clear visualization of the model's performance by showing the number of correct and incorrect predictions.

```
T20_x <- data.matrix(T20[, c("DIED", "AGE", "FEMALE", "LOS", "TOTCHG", "ZIPINC_QRTL", "RESIDENT", "PAY1", "DMONTH", "Diagnosis")])

T20_y <- T20$readmit_flag

T20_predictions <- predict(T_lasso, newx = T20_x, type = "response")

T_roc <- roc(T20_y, T20_predictions)
```

Historical Context

This detailed historical context outlines the origins, development, and broader impact of empirical Bayes methods, particularly through the lens of the James-Stein estimator, illustrating their significant role in shaping contemporary statistical theory and practice

- James-Stein Estimator
  - The James-Stein estimator developed by Charles Stein in 1956 represented a major shift in statistical estimation theory. Stein's initial findings, which were later expanded upon with the contributions of Willard James, showed that in problems involving estimation of multiple means simultaneously, a shrinkage estimator could perform better in terms of mean squared error than separate unbiased estimators for each mean.
  - Initially, Stein's result was met with skepticism because it contradicted long-standing beliefs about the optimality of the sample mean as an unbiased estimator. It took several years and further mathematical substantiation before the statistical community fully acknowledged its advantages.

- Empirical Baye
  - The term "Empirical Bayes" is rooted in the work of Herbert Robbins in the 1950s. Robbins was interested in a form of the Bayes method where the prior distribution is not specified in advance but estimated from the data. This was a radical idea at the time because it blended frequentist and Bayesian viewpoints, proposing a data-driven approach to Bayesian analysis.
  - Robbins' empirical Bayes methods provided a framework for estimating the prior distribution by observing frequencies of past events, thus addressing practical situations where historical data were plentiful but the underlying probabilistic models were complex and not easily understood.
  - Over the decades, empirical Bayes techniques have evolved from relatively simple approaches to sophisticated methods capable of handling high-dimensional data and complex dependency structures. This evolution has been paralleled by advances in computational statistics, which have made more computationally intensive empirical Bayes methods feasible.

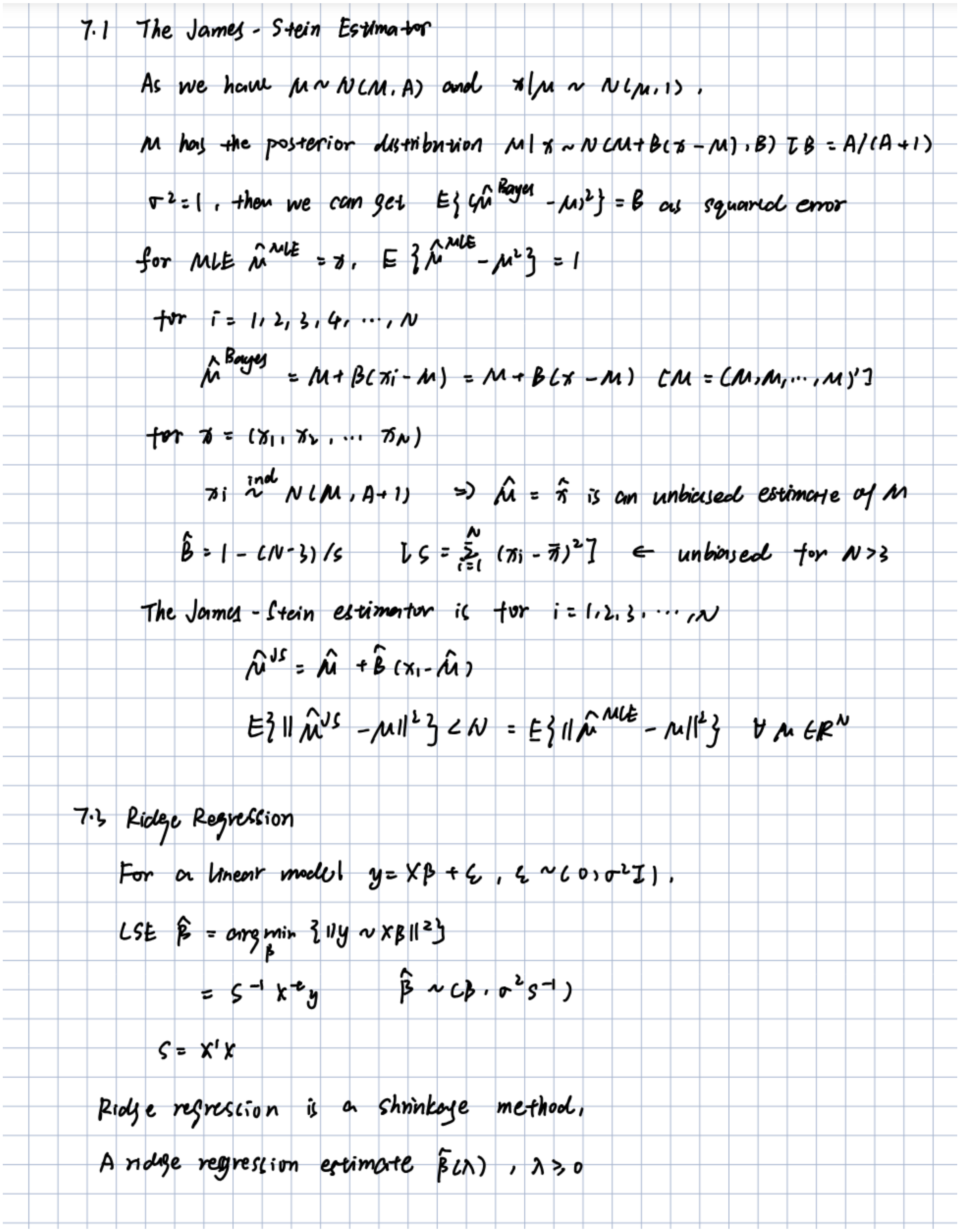
- Broader Statistical Practice
  - The empirical Bayes approach and the development of the James-Stein estimator have profoundly influenced modern statistical practices. These methods have promoted the use of shrinkage techniques in various applications, such as sports statistics (notably in baseball), health sciences, and social sciences.
  - Empirical Bayes methods have been adapted to tackle modern challenges in data science, especially in the fields of machine learning and bioinformatics, where they help in managing large datasets and complex models.
  - The acceptance and integration of these methods mark a cultural shift within the statistical community towards more pragmatically combining frequentist and Bayesian approaches, leading to more robust and effective statistical methodologies.

Statistical Practice Implications

- Rethinking Estimation Strategies
  - Techniques like the James-Stein estimator and ridge regression teach us to appreciate the benefits of shrinkage and regularization. These methods help in reducing overfitting by adjusting our estimates towards a common mean or through penalties, improving model accuracy, especially in complex datasets with many variables.
  - The Bias-Variance Tradeoff approaches encourage us to think critically about the balance between bias (the error introduced by approximating a real-world problem by a simpler model) and variance (the error from sensitivity to small fluctuations in the training set). This balance is crucial in statistical modeling, making our models more reliable and robust.

- Integration of Bayesian and Frequentist Methods
  - Combining principles from Bayesian and frequentist methodologies, empirical Bayes methods use observed data to estimate prior distributions. This hybrid approach enhances model flexibility and robustness, providing a solid framework for integrating prior knowledge with observed data without the strict requirements of traditional Bayesian priors.
  - By utilizing empirical Bayes, we can apply Bayesian thinking more practically. This methodology makes Bayesian techniques more accessible for everyday statistical problems, expanding their use beyond theoretical or highly specialized areas to more general applications.
- Improved Predictive Models
  - Bayes and related methodologies push for models that adapt based on the data itself. This adaptiveness is particularly useful in machine learning and data science, where predicting future events or classifications accurately is often complicated by large amounts of data and numerous influencing factors.
  - As datasets grow in size and complexity, the discussed methods prove invaluable for managing high-dimensional spaces without losing predictive power, often a challenge with traditional statistical methods.
- Advancements in Computational Statistics
  - Modern computational statistics have enabled the practical application of complex models like those involving empirical Bayes methods. These tools help in handling larger datasets and more complex analyses without prohibitive computational costs.
  - The availability of powerful computing resources at lower costs has democratized access to advanced statistical methods, allowing a wider range of researchers and practitioners to implement sophisticated statistical techniques.
- Broad Application Across Disciplines
  - The statistical methods discussed in this chapter are not limited to any specific field but are applicable across a diverse range of disciplines. Whether it's economics, sports, health sciences, or any other field, these methods provide tools for more accurate and reliable data analysis.
  - By applying these techniques, researchers can achieve more scientifically rigorous results, improving the reliability and validity of their findings in various research domains.
- Educational and Theoretical Implications
  - These modern statistical methods are reshaping how statistics is taught in universities and colleges. By integrating these approaches into curricula, educators can provide students with a more comprehensive understanding of how statistical theory can be applied practically.
  - The empirical Bayes methods and related techniques discussed also encourage ongoing theoretical developments in statistics. By exploring when and how these methods work best, researchers can further refine statistical theory, leading to new insights and methodologies.

Mathematics notes and explanation



7.1 The James-Stein Estimator

As we have $\mu \sim N(M, A)$ and $x|\mu \sim N(\mu, 1)$,

μ has the posterior distribution $\mu | x \sim N(M + B(x - M), B)$ [$B = A/(A+1)$

$\sigma^2 = 1$, then we can get $E\{(\hat{\mu}^{Bayes} - \mu)^2\} = B$ as squared error

for MLE $\hat{\mu}^{MLE} = x$, $E\{\hat{\mu}^{MLE} - \mu^2\} = 1$

for $i = 1, 2, 3, 4, \cdots, N$

$$\hat{\mu}^{Bayes} = M + B(x_i - M) = M + B(x - M) \quad [M = (M, M, \cdots, M)']$$

for $x = (x_1, x_2, \cdots x_N)$

$$x_i \overset{ind}{\sim} N(M, A+1) \Rightarrow \hat{M} = \hat{x} \text{ is an unbiased estimate of } M$$

$$\hat{B} = 1 - (N-3)/S \quad [S = \sum_{i=1}^{N} (x_i - \bar{x})^2] \leftarrow \text{unbiased for } N > 3$$

The James-Stein estimator is for $i = 1, 2, 3, \cdots, N$

$$\hat{\mu}^{JS} = \hat{M} + \hat{B}(x_1 - \hat{M})$$

$$E\{\|\hat{\mu}^{JS} - \mu\|^2\} < N = E\{\|\hat{\mu}^{MLE} - \mu\|^2\} \quad \forall \mu \in R^N$$

7.3 Ridge Regression

For a linear model $y = X\beta + \varepsilon$, $\varepsilon \sim (0, \sigma^2 I)$,

LSE $\hat{\beta} = \arg\min_{\beta} \{\|y \sim X\beta\|^2\}$

$$= S^{-1} X^t y \qquad \hat{\beta} \sim (\beta, \sigma^2 S^{-1})$$

$S = X'X$

Ridge regression is a shrinkage method,

A ridge regression estimate $\hat{\beta}(\lambda)$, $\lambda \geq 0$

$$\hat{\beta}(\lambda) = (S + \lambda I)^{-1} X'y$$

$$= (S + \lambda I)^{-1} S\hat{\beta}$$

Bayesian rational for ridge regression,

$\hat{\beta} \sim N_p(\beta, \sigma^2 S^{-1})$

Then the Baysian prior $\beta \sim N_p(0, \frac{\sigma^2}{\lambda} I)$

$\Rightarrow E\{\beta | \hat{\beta}\} = (S + \lambda I)^{-1} S\hat{\beta}$

A widely used penalty term involves "l_1 norm" $\|\beta\|_1 = \sum_1^p |\beta_j|$

$$\hat{\beta}(\lambda) = \arg\min \|y - X\beta\|^2 + \lambda \|\beta_1\|$$

This is the lasso estimator I used in my R code.

If we apply the James-Stein rule to normal model, $\hat{\beta}^{JS}$

$$\hat{\beta}^{JS} = \left[1 - \frac{(p-2)\sigma^2}{\hat{\beta} \cdot S\hat{\beta}}\right]\hat{\beta}$$

Here $\hat{\mu}^{JS} = X\hat{\beta}^{JS}$, $\mu = E[y]$

$\Rightarrow E\{\|\hat{\mu}^{JS} - \mu\|^2\} < p\sigma^2$

Citation

Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, from https://hastie.su.domains/CASI_files/PDF/casi.pdf

Wikipedia contributors. (2024, January 9). James–Stein estimator. In Wikipedia, The Free Encyclopedia. Retrieved [date you accessed the article], from https://en.wikipedia.org/wiki/James%E2%80%93Stein_estimator

Wikipedia contributors. (2024, January 15). Ridge regression. In Wikipedia, The Free Encyclopedia. Retrieved [date you accessed the article], from

https://en.wikipedia.org/wiki/Ridge_regression

OpenAI. (2023). ChatGPT (Version 4.0]). Retrieved 5/5/2024, from

https://www.openai.com/