Final_677 CASI - Chap 7

Bingtian Ye

2024-05-05

JAMES-STEIN ESTIMATION AND RIDGE REGRESSION

MAIN POINT

What is the James-Stein estimator and in which scenario does it prove to be superior to traditional estimators like the maximum likelihood estimator?

Answer: The James-Stein estimator is known for its ability to reduce the mean squared error more effectively than traditional estimators such as the maximum likelihood estimator, especially in scenarios involving high-dimensional data with multiple parameters and small sample sizes. This shrinkage estimator pulls estimates towards the overall mean, improving accuracy and stability.

How does Ridge regression modify the least squares estimation to address multicollinearity?

Answer: Ridge Regression addresses multicollinearity by adding a penalty to the least squares estimation that is proportional to the square of the coefficient magnitudes (L2 penalty). This modification helps stabilize the regression estimates in the presence of highly correlated predictors, reducing the risk of extreme variance in the estimated coefficients.

What role do empirical Bayes techniques play in the context of James-Stein estimation and Ridge regression?

Answer: In the context of James-Stein estimation and Ridge regression, empirical Bayes techniques estimate the optimal shrinkage factor based on data, allowing for a more adaptive approach that combines frequentist and Bayesian principles. This integration helps enhance the accuracy and stability of the estimators by effectively using prior data to inform the shrinkage process.

What are the practical applications of these statistical methods as demonstrated in the chapter?

Answer: The practical applications discussed in the chapter include enhancing statistical models used in sports analytics, such as predicting baseball batting averages. By applying James-Stein estimators and Ridge regression, statisticians can achieve more accurate predictions, demonstrating how these methods can be effectively utilized in real-world data analysis.

How do these methods impact the interpretability and complexity of the models?

Answer: While these methods, particularly Ridge regression, introduce bias to reduce variance and complexity, they may also affect model interpretability by not reducing coefficients to zero—thus keeping all predictors in the model. This can complicate understanding which variables are most influential. Conversely, methods like the James-Stein estimator simplify interpretations by focusing on significant predictors, although they may introduce some level of bias.

THE MATHEMATICS UNDERLYING THE MATERIAL IN THE CHAPTER

James-Stein Theorem (The James-Stein Estimator)

Suppose that, $x_i \mid \theta_i \sim N(\theta_i, 1)$

independently for $i = 1, 2, ..., N$,with $N >= 4$. Then

$$\mathbb{E}\left[\|\hat{\theta}_{JS} - \theta\|^2\right] < N = \mathbb{E}\left[\|\hat{\theta}_{MLE} - \theta\|^2\right]$$

for all choices of $\mu \in R^N$

The James-Sterin Theorem shows that the expected mean square error of the James-Stein estimator is always less than or equal to the expected mean square error of the maximum likelihood estimator, provided that the number of parameters is at least four. In other words, even though the James-Stein estimator introduces bias, it still has a smaller mean square error (MSE) than the maximum likelihood estimator.This challenges the long-held belief that unbiased estimators are always preferable.

Ridge Regression and Lasso Regression

Assuming we observe an n-dimensional vector $y = (y_1, y_2, ...y_n)^T$ from the linear model,

$$y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2 I)$$

The least squares estimate $\hat{\beta}$ is the minimizer of the total sum of squared errors.

$$\hat{\beta}_{\text{OLS}} = \arg\min_{\beta} \|y - X\beta\|^2$$

We can get unbiased estimator $\hat{\beta}$.

$$\hat{\beta}_{\text{OLS}} = (X^T X)^{-1} X^T y$$

Compare to the least squared estimate, the ridge regression estimate adjusts the least squares method by adding a penalty on the size of the coefficients, which helps to handle multicollinearity and reduce model complexity:

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \left\{\|y - X\beta\|^2 + \lambda\|\beta\|^2\right\}$$

And we can get the following estimator,

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y$$

The Lasso regression, similar to ridge, also modifies the least squares by adding a penalty, but it uses the L_1 norm of the coefficients, encouraging a sparse solution where some coefficient estimates may be exactly zero:

$$\hat{\beta}_{\text{Lasso}} = \arg\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda\|\beta\|_1 \right\}$$

Method	Advantages	Disadvantages
OLS	Simple, interpretable, and efficient under low-dimensional settings with non-collinear variables.	Prone to overfitting, can't handle multicollinearity well, poor performance in high-dimensional spaces.

Method	Advantages	Disadvantages
Ridge	Reduces overfitting, handles multicollinearity by reducing variance of the coefficients, especially suitable when variables outnumber observations.	Does not reduce coefficients to zero, thus does not perform variable selection, retains all variables in the model.
Lasso	Reduces overfitting, can perform variable selection by shrinking some coefficients to zero, helps in model simplicity and interpretability.	May not be stable under scenarios where the number of variables exceeds the number of samples; may randomly select among highly correlated variables.

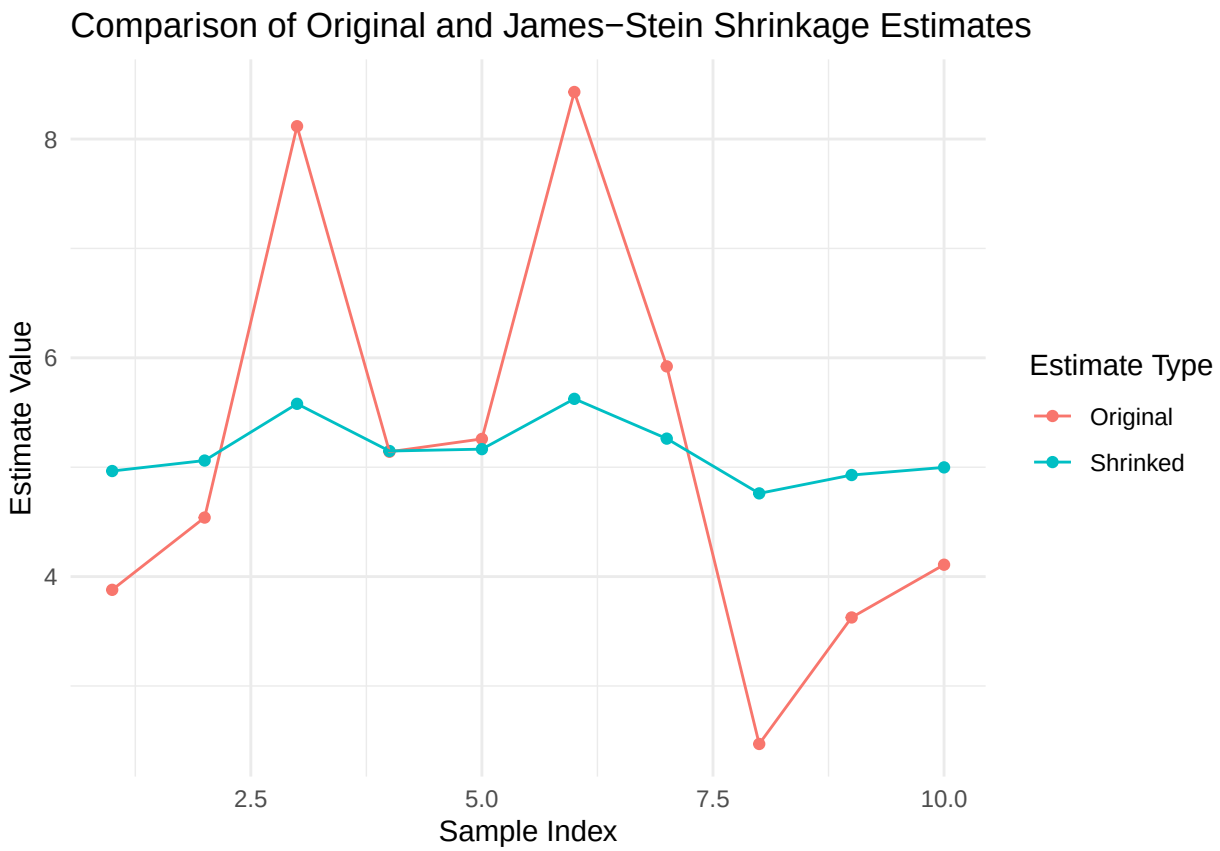
Methods for Selecting the Regularization Parameter λ

1. **Cross-validation**:
   - The most widely used method for selecting λ is k-fold cross-validation. In this approach, the dataset is split into k smaller sets (folds). The model is trained on $k-1$ of these folds, with the remaining part used as a test set to compute a performance metric. This process is repeated k times, with each of the k folds used exactly once as the test set. The λ value minimizing the average test metric across all folds is chosen.
   - For Lasso, typically, the metric is mean squared error (MSE).
2. **Information Criteria**:
   - Criteria like Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) can also be used to select λ. These methods balance model fit and complexity by including a penalty for the number of parameters.
3. **Regularization Path**:
   - Computing the regularization path, which shows the coefficients of the model as a function of different λ values, can help visualize how the sparsity of the model increases with increasing λ. Tools like LARS (Least Angle Regression) or coordinate descent are often used to efficiently compute the entire path.
4. **Domain-Specific Heuristics**:
   - In some fields, there may be established heuristics or empirical methods for selecting λ, based on prior knowledge or specific characteristics of the data.
5. **Through Bayesian Methods**:
   - The paper "Selecting the LASSO Regularization Parameter via Bayesian Principles" investigates the Bayesian interpolation method for estimator selection and applies it to the problem of regularization parameter selection in Ridge and LASSO regressions. It proposes a Bayesian-based criterion for setting λ, which shows comparable or sometimes superior performance compared with other methods like GCV, extended AIC, and extended BIC.
   - Click to read full paper(https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7806091).

ALGORITHE EXAMPLE

Jame-Stein estimator

Assuming y is drawn from a Normal Distribution with mean = 5 and standard deviation = 2".



In the plot, the red line represents the original sample means, which show considerable fluctuations due to being calculated directly from the data without adjustments, thus displaying high variance. In contrast, the teal line depicts the James-Stein shrinkage estimates, which demonstrate more stability as the sample means are adjusted towards the overall mean of all samples. This method of shrinking reduces the influence of extreme values and variability by calculating a shrinkage factor based on the deviation of individual sample means from the overall mean, effectively improving the accuracy and stability of the estimates, especially in scenarios where the number of parameters is close to or exceeds the sample size.

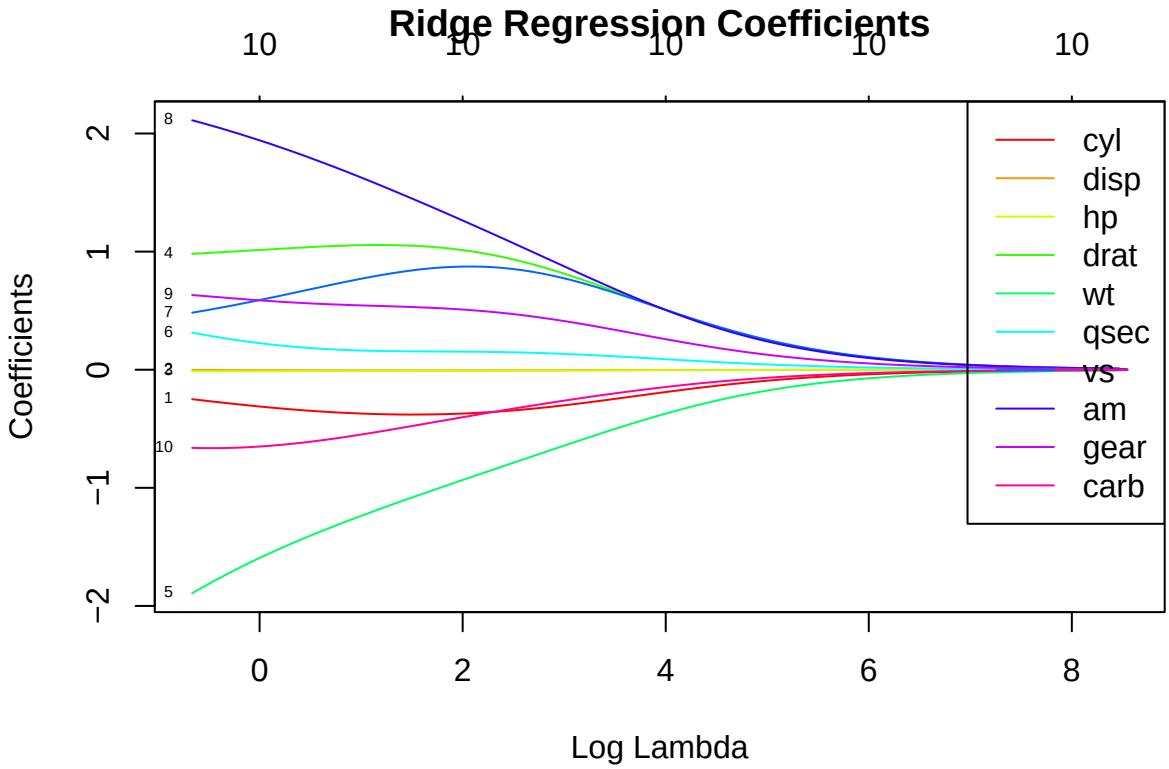
OLS, Ridge Regression and Lasso Regression

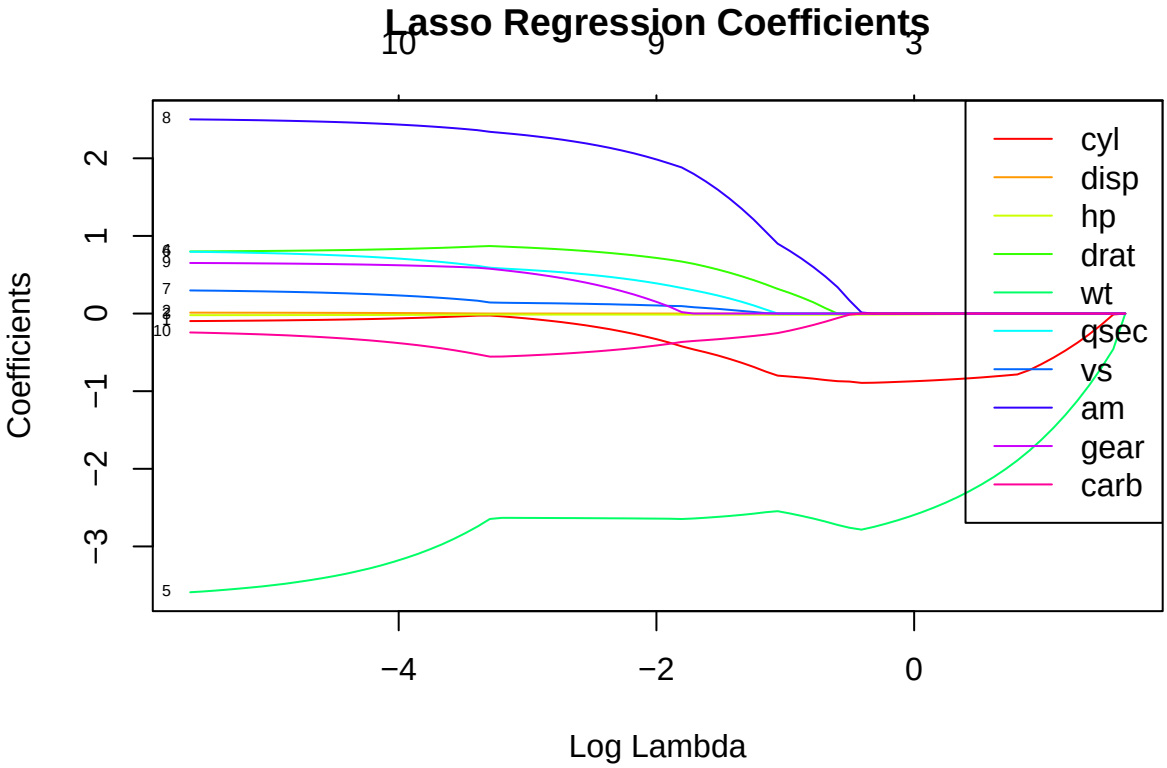
We will use the `mtcars` dataset for our simulation to further explore the differences between three methods: Ordinary Least Squares (OLS), Ridge, and Lasso regression.

Below is a table describing some of the key variables in the `mtcars` dataset:

Variable	Description
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	1/4 mile time
vs	Engine (0 = V-shaped, 1 = straight)
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```





Plots show how the coefficients of the predictors in the `mtcars` dataset change as a function of the regularization parameter λ (lambda) for both Ridge and Lasso regression.

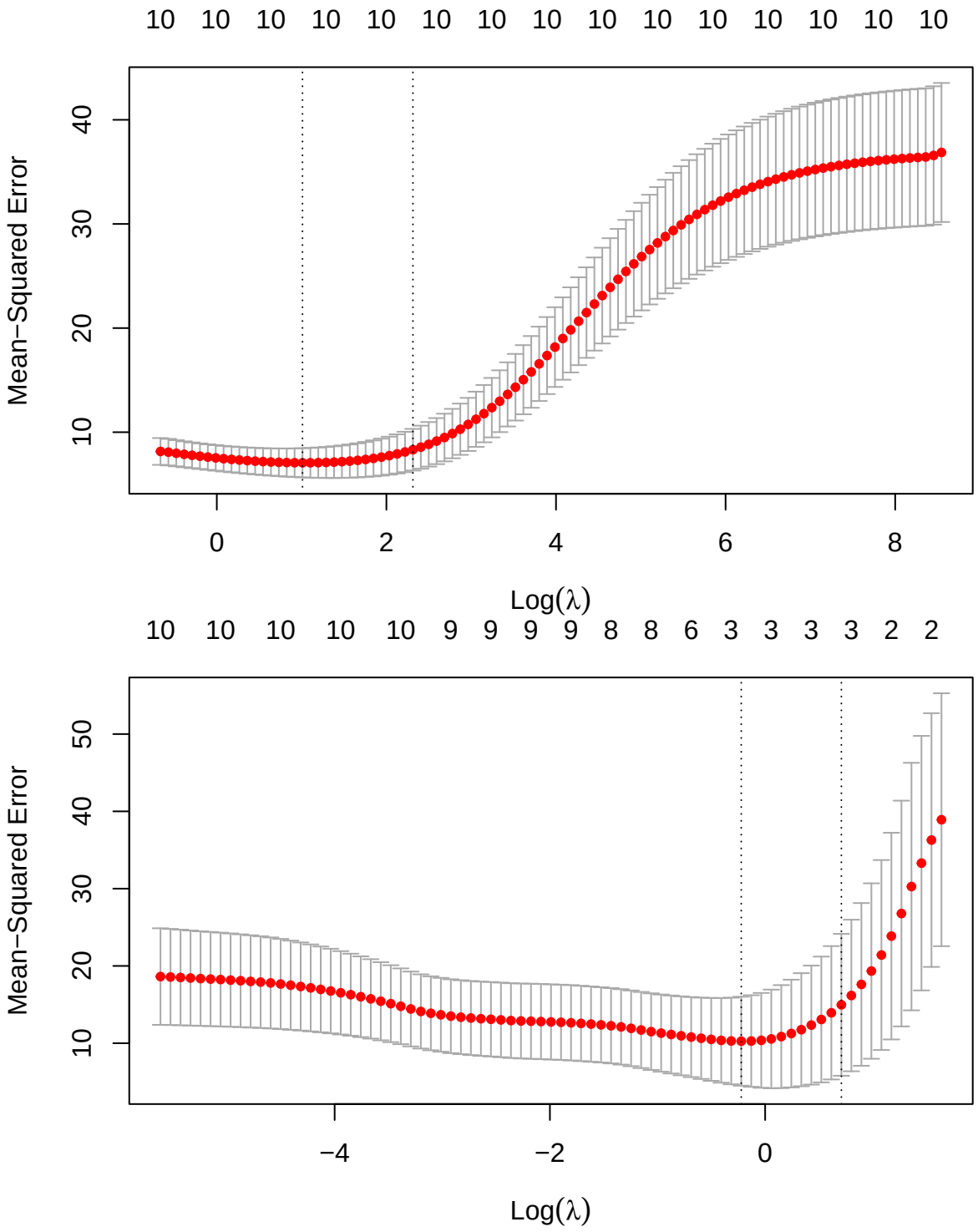
- **Ridge Regression Coefficients**: This plot illustrates that as λ increases, the magnitude of the coefficients decreases smoothly. All coefficients approach zero as λ becomes very large, reflecting Ridge regression's characteristic of shrinking coefficients as part of its regularization approach.
- **Lasso Regression Coefficients**: In contrast to Ridge, the Lasso plot highlights a more abrupt transition of coefficients to zero. This feature of Lasso, known as sparsity, shows that it not only shrinks coefficients but also sets many of them exactly to zero at a certain threshold, effectively performing variable selection.

Both plots are colored to represent each predictor uniquely, making it easy to trace how specific predictors are influenced by changes in λ.

The comparison between OLS, Ridge, and Lasso regression in this exercise demonstrates how regularization (Ridge and Lasso) can affect the coefficients of the predictors. OLS, without any regularization, may fit the data but can suffer from high variance and overfitting, especially when the predictor variables are highly correlated. Ridge and Lasso, through their regularization terms, not only help in reducing overfitting but also in selecting significant predictors (particularly Lasso, which can shrink coefficients to zero, thereby performing variable selection).

Using Cross-Validation selecting the Regularization Parameter λ

Following plots provided illustrate the process of selecting the optimal regularization parameter (λ) for Ridge and Lasso regression applied to the `mtcars` dataset. The mean squared error (MSE) is plotted against log-transformed λ values, providing insights into the balance between model complexity and fit.




```
## [1] "Optimal lambda for Ridge: 2.7467889881046"
```

```
## [1] "Optimal lambda for Lasso: 0.800703565270879"
```

In the provided Ridge regression plot, a U-shaped curve is evident, indicating that the MSE decreases as the regularization parameter λ increases from a very low value, reaching a minimum, and then starts increasing as λ continues to rise. The optimal λ for Ridge regression, identified at approximately 2.747, represents the point where the model achieves the best balance between reducing model complexity and minimizing

prediction error. This optimal point suggests that Ridge regression effectively manages multicollinearity in the dataset while avoiding overfitting through appropriate regularization.

The Lasso regression plot demonstrates a similar trend where MSE decreases with an increase in λ, reaching its lowest point at approximately 0.801 before stabilizing. This optimal λ for Lasso is lower compared to Ridge, which highlights Lasso's characteristic of setting some coefficients to zero, thus effectively simplifying the model while maintaining predictive accuracy. The choice of λ at 0.801 indicates a level of regularization that optimizes the bias-variance trade-off by eliminating non-contributive predictors, enhancing the model's interpretability and generalization ability.

REFERENCE AND HISTORICAL BACKGROUND

historical background

1950s and Early 1960s - Introduction of computer technology to statistics, marking the beginning of computational statistics, which allowed for the development of more complex statistical models previously limited by computational constraints.

1970 - Introduction of Ridge Regression by Hoerl and Kennard to address the limitations of Ordinary Least Squares (OLS) in the presence of multicollinearity. This method stabilized estimates by adding a penalty proportional to the square of coefficients to the OLS loss function.

1980s - Advancements in computational power led to the widespread adoption of computational methods such as bootstrap and cross-validation for model assessment and selection.

1996 - Proposal of Lasso Regression by Robert Tibshirani, which extended the concepts of Ridge by applying an L1 penalty, allowing for both regularization and variable selection by shrinking some coefficients exactly to zero.

Early 21st Century - The rise of Big Data emphasized the importance of methods like Lasso for handling high-dimensional data, especially in fields like genomics, finance, and social sciences.

Mid-2000s to Present - The popularity of Bayesian methods and Empirical Bayes, which are used to set shrinkage parameters in estimators like James-Stein and Ridge Regression. These methods combine the strengths of frequentist and Bayesian approaches, providing powerful tools for modern complex data sets.

reference

1.N. Huri and M. Feder, "Selecting the LASSO regularization parameter via Bayesian principles," 2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE), Eilat, Israel, 2016, pp. 1-5.

2.Draper, Norman R., and R. Craig Van Nostrand. "Ridge regression and James-Stein estimation: review and comments." Technometrics 21.4 (1979): 451-466.

3.James, William, and Charles Stein. "Estimation with quadratic loss." Breakthroughs in statistics: Foundations and basic theory. New York, NY: Springer New York, 1992. 443-460.

4.Saunders, Craig, Alexander Gammerman, and Volodya Vovk. "Ridge regression learning algorithm in dual variables." (1998): 515-521.

5.Efron, Bradley, and Trevor Hastie. "Computer age statistical inference." (2022): 1-xix.

6.ChatGPT for some code and RMD edit.

STATISTICAL PRACTICE IMPLICARION

Improved Estimator Performance: The James-Stein estimator highlights the potential for shrinkage methods to outperform traditional estimators like the maximum likelihood estimator, particularly in situations with small sample sizes or high-dimensional data. This suggests that practitioners should consider shrinkage approaches when dealing with complex data sets where overfitting is a concern.

Handling of Multicollinearity: Ridge Regression introduces a penalty term to the least squares estimation to handle multicollinearity among predictors. This is particularly relevant in scenarios where predictor variables are highly correlated, which can destabilize the estimates from standard regression models. The adoption of Ridge Regression can lead to more reliable and robust models in such cases.

Bias-Variance Trade-off: Both methods introduce bias into the estimates to achieve a reduction in variance. This trade-off is crucial in statistical modeling as it can significantly impact the model's predictive performance. Practitioners must carefully tune the regularization parameter to optimize this trade-off, balancing the model's complexity against its generalizability.

Complex Model Interpretation: While these methods can improve model accuracy, they also complicate the interpretation of the model due to the introduction of bias. For Ridge Regression, all variables remain in the model but their coefficients are shrunk, which can obscure the understanding of variable importance.

Empirical and Bayesian Approaches: The chapter discusses the use of empirical and Bayesian methods to set the regularization parameters in these estimators. This approach can enhance the application of James-Stein and Ridge methods by allowing the data itself to inform the degree of shrinkage, thereby aligning model complexity with the underlying data structure.