Regression Shrinkage and Selection via the Lasso[1]

Hsieh Yuan-Hao

2024/10/11

[1] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267-288.

OLS definition Recap

- The variance–covariance matrix of the least squares:

$$Var(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

Typically one estimates the variance σ^2 by

$$\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- it is easy to show that

$$\hat{\beta} \sim \mathcal{N}(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

$$(N-p-1)\hat{\sigma}^2 \sim \sigma^2\chi^2_{N-p-1}$$

- We use these distributional properties to form tests of hypothesis and confidence intervals for the parameters β_j.

Hypothesis test

- To test the hypothesis that a particular coefficient $\beta_j = 0$, we form the standardized coefficient or **Z-score**

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$$

  where v_j is the *jth* diagonal element of $(X^T X)^{-1}$

- Now,we can obtain a $1 - 2\alpha$ confidence interval for β_j:

$$(\hat{\beta}_j - z^{(1-\alpha)} v_j^{\frac{1}{2}} \hat{\sigma}, \hat{\beta}_j + z^{(1-\alpha)} v_j^{\frac{1}{2}} \hat{\sigma})$$

  Here, $z^{(1-\alpha)}$ is the $1 - \alpha$ percentile of the normal distribution.

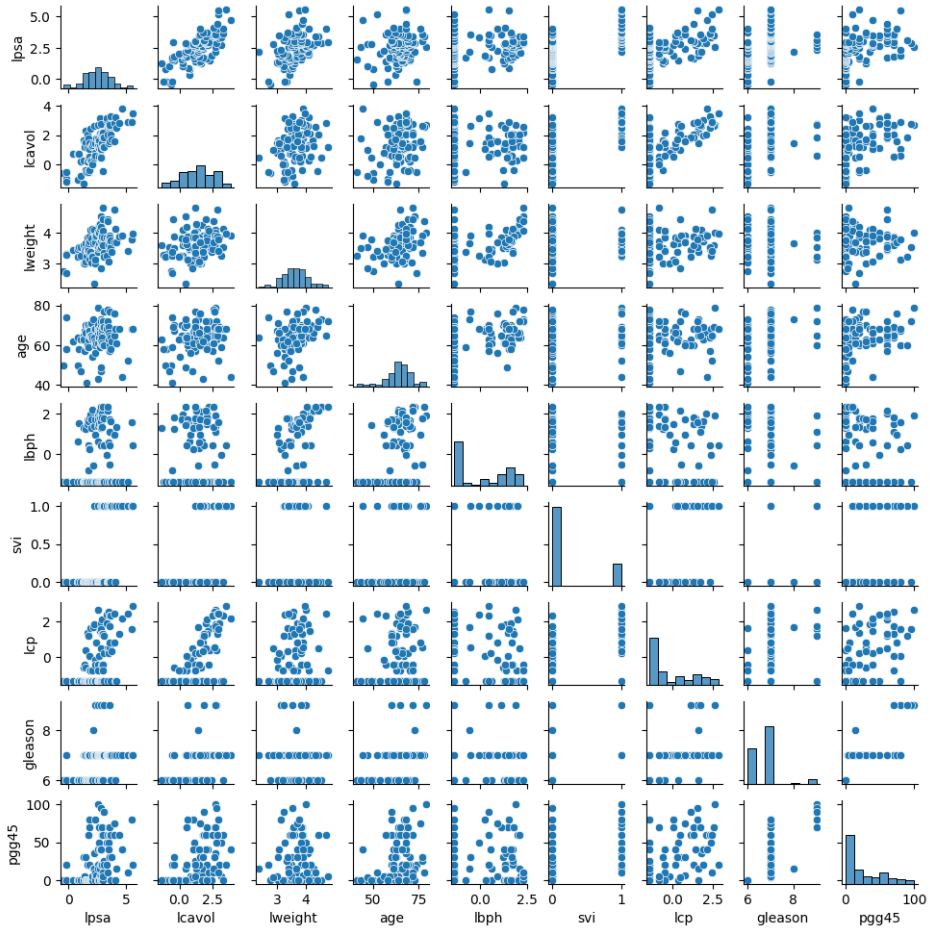
- In a similar fashion we can obtain an approximate confidence set for the entire parameter vector β,

$$C_\beta = \left\{ \beta \;\middle|\; (\hat{\beta} - \beta)^T \mathbf{X}^T \mathbf{X} (\hat{\beta} - \beta) \leq \hat{\sigma}^2 \chi^{2\ (1-\alpha)}_{p+1} \right\}$$

Example - Prostate Cancer Data introduction

- Prostate Cancer dataset[2]

lpsa	log prostate specific antigen
lcavol	log cancer volume
lweight	log prostate weight
age	age
lbph	log of benign prostatic hyperplasia amount
svi	seminal vesicle invasion
lcp	log of capsular penetration
gleason	Gleason score
pgg45	percent of Gleason scores 4 or 5



[2]Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., Yang, N. (1989).

Baseline error rate

- Base Error rate For the ‘DummyRegressor‘ using the mean strategy, the predicted value $\hat{y}_i$ for every observation is the same, equal to the mean of the training set $\bar{y}_{\text{train}}$:

$$\hat{y}_i = \bar{y}_{\text{train}} = \frac{1}{m} \sum_{j=1}^{m} y_j^{\text{train}}$$

  Where:
  - m is the number of observations in the training set.
  - y_j^{train} is the j-th actual target value in the training set.
- The MSE for the ‘DummyRegressor‘ is calculated by comparing each actual value y_i in the test set to the constant prediction $\bar{y}_{\text{train}}$, and then averaging the squared differences:

$$\text{Baseline MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y}_{\text{train}})^2$$

```
Baseline Test Error: 1.057
```

least squares estimation

- The **Coefficient** column provides the estimated effects of each predictor on the log of prostate-specific antigen (lpsa).
- The **Standard Error** column shows the uncertainty in those estimates.
- The **Z-Score** column helps assess the statistical significance of each predictor:
  - A Z-Score larger than 2 (or smaller than -2) in absolute value suggests that the corresponding predictor is **significantly different** from zero at the 0.05 level (roughly).
  - Predictors with Z-scores smaller in magnitude (closer to zero) are likely to be **insignificant** in the model.

Term	Coefficient	Std. Error	Z Score
Intercept	2.46	0.09	27.60
lcavol	0.68	0.13	5.37
lweight	0.26	0.10	2.75
age	-0.14	0.10	-1.40
lbph	0.21	0.10	2.06
svi	0.30	0.12	2.47
lcp	-0.29	0.15	-1.87
gleason	-0.02	0.14	-0.15
pgg45	0.27	0.15	1.74

The mean prediction error on the test data

- The formula for **MSE** is:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  Where:
  - y_i are the true values (from y test).
  - $\hat{y}_i$ are the predicted values (from $\hat{y}$).
- **Standard error** measures the uncertainty in the MSE estimate. A smaller standard error suggests more confidence in the MSE value, while a larger standard error indicates that the MSE estimate is more variable. The formula for **Standard Error (SE)** of the MSE is:

$$\text{SE} = \frac{\text{std}((y_i - \hat{y}_i)^2)}{\sqrt{n}}$$

```
Least Squares Test Error: 0.521
                Std Error: 0.179
```

Comparison of LS, Ridge, and Lasso Regression Coefficients and Errors

- **LS** has the highest test error due to overfitting.
- **Ridge** improves performance by shrinking coefficients, but it still retains all variables.
- **Lasso** offers the best performance by shrinking some coefficients to zero, effectively performing feature selection, which simplifies the model and improves generalization.
- Result:

Term	LS	Ridge	Lasso
Intercept	2.465	2.464	2.468
lcavol	0.676	0.405	0.536
lweight	0.262	0.234	0.187
age	-0.141	-0.042	0.000
lbph	0.209	0.158	0.000
svi	0.304	0.221	0.085
lcp	-0.287	0.010	0.000
gleason	-0.021	0.042	0.000
pgg45	0.266	0.128	0.006
Test Error	0.521	0.492	0.479
Std Error	0.179	0.164	0.164

Introduction to the One-Standard-Error Rule

- **What is the One-Standard-Error Rule?**
  - The **one-standard-error rule** is a model selection criterion used in cross-validation to select the **simplest model** that performs nearly as well as the best-performing model.
  - Instead of picking the model with the absolute minimum error (which can lead to overfitting), we choose the **simpler model** whose performance is within **one standard error** of the minimum.
- **Why Use It?**
  - Helps avoid overfitting by favoring simpler models that generalize better.
  - Selects a model that performs well but uses fewer parameters (or degrees of freedom), making it more interpretable and robust to noise.

How Does the One-Standard-Error Rule Work?(1/2)

1. Cross validation
   - Perform cross-validation to evaluate a set of models over a range of parameter values (e.g., different levels of regularization)
   - For each model, calculate the **average validation error** and the **corresponding standard error (SE)** over the cross-validation folds.
2. Find the Minimum Error:
   - Identify the model with the **lowest cross-validation error**. This model typically has the highest complexity (more parameters or degrees of freedom) and is likely to overfit.
   - Let this minimum error be denoted as $\text{Error}_{\text{min}}$, and its corresponding standard error as SE_{min}.
3. Set a Threshold:
   - Compute a threshold that accounts for the variability in the cross-validation error using the standard error:

$$\text{Threshold} = \text{Error}_{\text{min}} + SE_{\text{min}}$$

   - The **threshold** defines a range of acceptable errors, taking into account the uncertainty in the error estimates.

How Does the One-Standard-Error Rule Work?(2/2)

4. Choose the Simplest Model
   - Among all the models with errors less than or equal to the threshold, **select the simplest model** (the one with the fewest parameters or degrees of freedom).
   - This model may not have the lowest error, but it performs almost as well as the best model while being simpler and more robust.

- **Key Formula:**

$$\text{Threshold} = \text{Error}_{\text{min}} + SE_{\text{min}}$$

Where:

- $\text{Error}_{\text{min}}$ is the minimum cross-validation error (the lowest average error across the folds).
- SE_{min} is the standard error associated with $\text{Error}_{\text{min}}$.

Example of Ridge regression(1/2)

1. Cross-Validation: You evaluate Ridge regression models with different values of α, and for each α, you compute the average cross-validation error and standard error.
   Example error values:
   - $\alpha_1 = 100$: Cross-validation error = 0.25, SE = 0.02
   - $\alpha_2 = 10$: Cross-validation error = 0.22, SE = 0.03
   - $\alpha_3 = 1$: Cross-validation error = 0.21 (minimum), SE = 0.04
   - $\alpha_4 = 0.1$: Cross-validation error = 0.23, SE = 0.05
2. Find the Minimum Error: The model with $\alpha = 1$ has the lowest error (0.21).
3. Set the Threshold

$$\text{Threshold} = 0.21 + 0.04 = 0.25$$

   This means we are willing to accept any model with an error less than or equal to 0.25.

Example of Ridge regression(2/2)

4. Choose the Simplest Model from the evaluated models:
   * $\alpha_1 = 100$: Error = 0.25 (simpler model)
   * $\alpha_2 = 10$: Error = 0.22
   * $\alpha_3 = 1$: Error = 0.21 (most complex)

The model with $\alpha = 100$ is chosen because it is **simpler** (higher regularization, fewer effective parameters) and its error (0.25) is within the acceptable range defined by the threshold.

- **Visual Representation:** In a typical plot, the x-axis represents the model complexity (degrees of freedom or number of parameters), and the y-axis represents the cross-validation error.
  1) Minimum error is at the lowest point in the curve.
  2) One-standard-error rule selects the simplest model to the left of the curve, where the error is still within the one-standard-error threshold.

Some Model Parameters

1. Degrees of Freedom in Ridge Regression:

$$\mathrm{df}(\lambda) = \mathrm{tr}\left[\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\right]$$

Or equivalently:

$$\mathrm{df}(\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

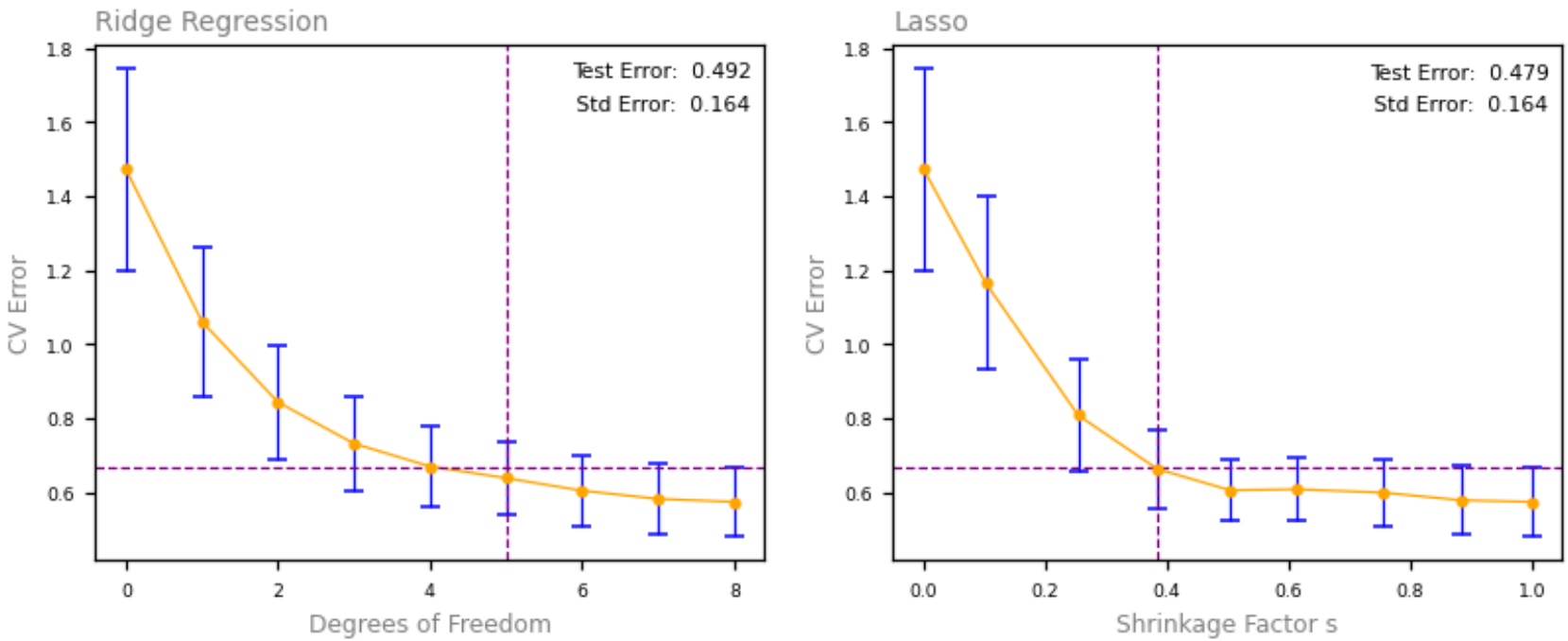
2. Shrinkage Factor s in Lasso Regression:

$$s = \frac{t}{\sum_{j=1}^{p} |\hat{\beta}_j|}$$

Where:

- t is the constraint on the sum of the absolute values of the coefficients $\hat{\beta}_j$.
- $\sum_{j=1}^{p} |\hat{\beta}_j|$ is the sum of the absolute values of the least-squares estimates of the coefficients.

Ridge and lasso CV Error plot



(a) Ridge regression

(b) Lasso regression

Figure 1: Compare between ridge and lasso

Lasso coefficient estimates

- The central 90% percentile intervals all contained the value 0, with the exceptions of those for lcavol and svi.

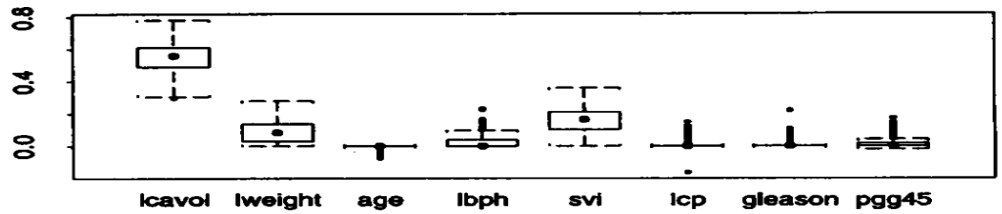
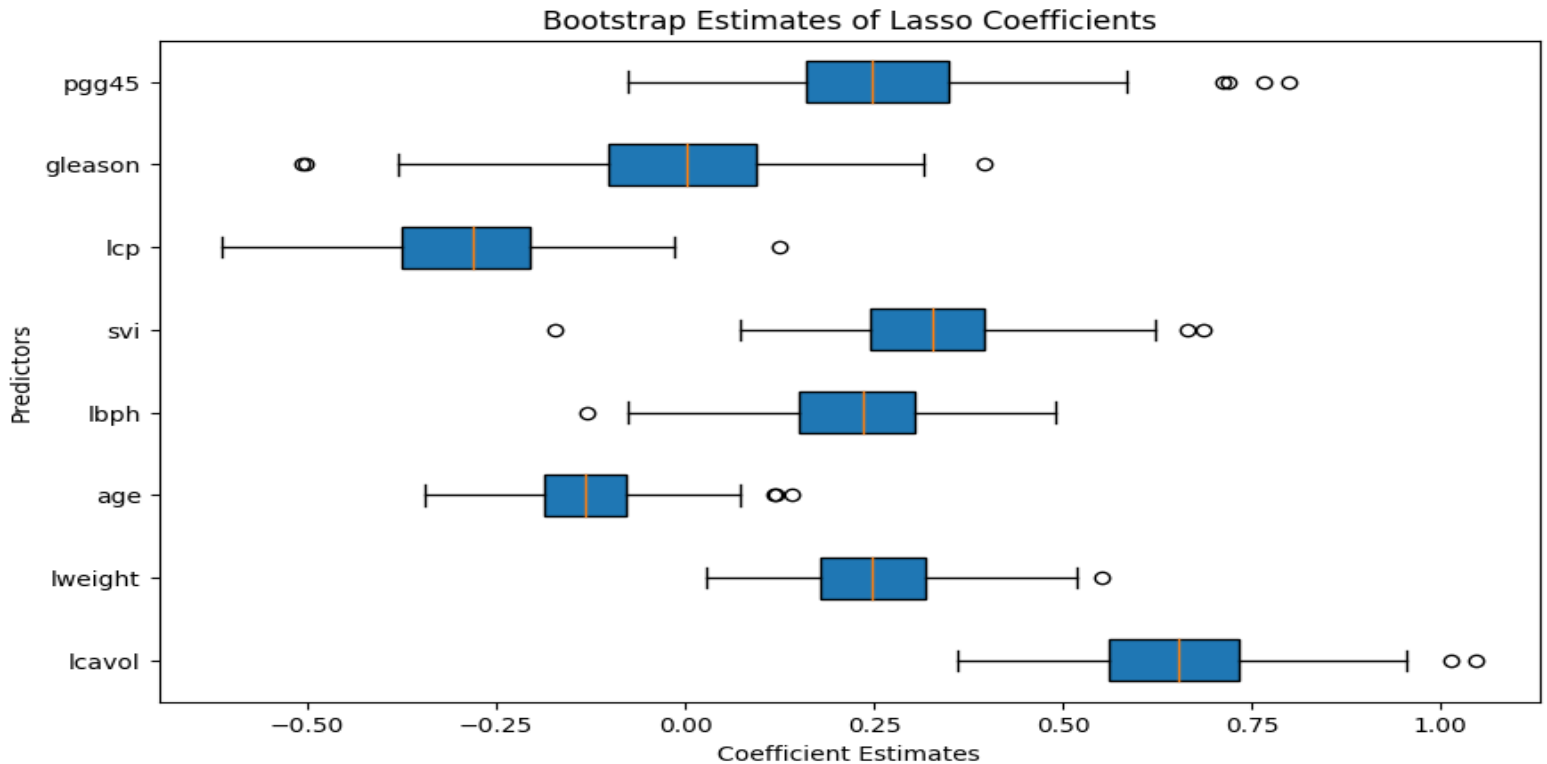


Fig. 6. Box plots of 200 bootstrap values of the lasso coefficient estimates for the eight predictors in the prostate cancer example



Coefficient value plot