Regression Analysis
Chapter 10: Model Selection

Shaobo Jin

Department of Mathematics

Motivation

It is common that we are unsure about which regressors should be used in the regression model and how they should be used.

1. Should I fit $\beta_0 + \beta_1 x_1$ or $\beta_0 + \beta_1 x_1 + \beta_2 x_2$? Nested models
2. Should I fit $\beta_0 + \beta_1 x$ or $\beta_0 + \beta_1 x + \beta_2 x^2$, or even a data-driven $f(x)$?
3. Should I fit $\beta_0 + \beta_1 x_1$ or $\beta_0 + \beta_2 x_2$? Non-nested models

Information Criterion

Suppose that $\boldsymbol{y} \mid \boldsymbol{X}$ follows some distribution with density $f(\boldsymbol{y} \mid \boldsymbol{X})$.

- We want our model to be complex enough to fit the data well.
  - One way to achieve so is to maximize $f(\boldsymbol{y} \mid \boldsymbol{X})$ or $\log f(\boldsymbol{y} \mid \boldsymbol{X})$.
- On the other hand, we don't want the model to be too complex to avoid overfitting.
- We cannot simply choose a complex model in order to have a better fit to the current data.
- An information criterion is often of the form

$$-c \log f(\boldsymbol{y} \mid \boldsymbol{X}) + \text{penality of model complexity}$$

for some constant c.

Kullback-Leibler Divergence

Suppose that the true data generating process is $g(y)$, but we assume $f(y \mid \boldsymbol{x})$, where f and g may not be the same.

- The Kullback-Leibler (KL) divergence is

$$\mathrm{KL}(g, f) = \mathrm{E}\left[\log\left(\frac{g(y)}{f(y \mid \boldsymbol{x})}\right)\right] = \int \log\left(\frac{g(y)}{f(y \mid \boldsymbol{x})}\right) g(y)\, dy.$$

- We can show that $\mathrm{KL}(g, f) \geq 0$.
- A large $\mathrm{KL}(g, f)$ means that f is far away from g.

We can rewrite $\mathrm{KL}(g, f)$ as

$$\mathrm{KL}(g, f) = \mathrm{E}[\log g(y)] - \int \log[f(y \mid \boldsymbol{x})]\, g(y)\, dy.$$

AIC: Minimizing Distance of the Fit from the Truth

- If the unknown parameters are estimated by maximum likelihood, a nearly "unbiased" estimator of

$$\mathrm{KL}(g, f) - \mathrm{E}[\log g(y)] = -\int \log[f(y \mid \boldsymbol{x})]\, g(y)\, dy$$

  is

$$\log[f(\boldsymbol{y} \mid \boldsymbol{X})] - m,$$

  where m is the number of parameters in our model.
- The Akaike information criterion (AIC) is defined as

$$\mathrm{AIC} = -2\log[f(\boldsymbol{y} \mid \boldsymbol{X})] + 2m.$$

- We prefer the model with the smallest AIC or a parsimonious model that has AIC near the minimum.
- In practice, AIC tends to be conservative, in the sense that it tends to select more explanatory variables.

OBS! AIC in Book

The book defines AIC to be

$$n \log\left(\frac{\text{RSS}}{n}\right) + 2p \quad = \quad n \log\left(\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}\right) + 2p.$$

Keep in mind that this expression requires the normality assumption!

Under the normality assumption, the AIC becomes

$$\text{AIC} \quad = \quad n \log\left(\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}\right) + 2\left(p+1\right) + n\log\left(2\pi\right) + n.$$

BIC: Consistent Model Selection

- Bayesian information criterion penalizes a complex model much more than AIC.

$$\text{BIC} \quad = \quad -2\log\left[f\left(\boldsymbol{y} \mid \boldsymbol{X}\right)\right] + \log\left(n\right)m.$$

- We prefer the model with the smallest BIC or a parsimonious model that has BIC near the minimum.
- BIC is consistent in model selection in the sense that

$$P\left(\text{Choose the true model if it is a candidate}\right) \to 1, \text{ as } n \to \infty.$$

- In contrast, AIC is not consistent.

Stepwise Regression Using AIC

If we have $p-1$ covariates, then we can fit in total 2^{p-1} models, which is computationally intensive for a large p.

- Forward selection starts with an intercept model and sequentially adds the term that gives the lowest AIC.
- Backward elimination starts with a model with all explanatory variables and sequentially removes the term that has the greatest decrease in AIC.
- Stepwise regression is a combination of both by checking whether adding one term or deleting one term will yield the smallest AIC.

Cross Validation (CV)

CV focuses on the prediction property.

Algorithm 1: One version of cross validation

Randomly split the data set into K nonoverlapping groups (K-fold CV) or split the data set into n groups (leave-one-out CV, aka jackknife);
for $k = 1$ *in* $1 : K$ **do**
    Take the kth group as test set and the remaining groups as training set ;
    while *for each model* **do**
        Fit it on the training set and evaluate it on the test set ;
        Retain the model performance (e.g., MSE, misclassification error, log-likelihood) ;
    end
end
Summarize the model performance (e.g., average across K groups) ;
Choose the model that performs the best ;
Refit the chosen model using the entire data set ;

More on Cross Validation

Suppose that the true model is one of our candidate models.

- It has been proved that, for leave-one-out CV,

$$\begin{aligned} \lim_{n\to\infty} P\left(\text{Choose a model where a nonzero } \beta \text{ is missing}\right) &= 0, \\ \lim_{n\to\infty} P\left(\text{Choose the most parsimonious true model}\right) &\neq 1. \end{aligned}$$

  Hence, the leave-one-out CV is conservative that select a model of excessive size.

- It has been proved that you need leave-n_v-out CV if you want

$$\lim_{n\to\infty} P\left(\text{Choose the most parsimonious true model}\right) = 1,$$

  where $n_v/n \to 1$ as $n \to \infty$.

More on Leave-One-Out Cross Validation

Suppose now that all candidate models are wrong. Consider

$$L_n \quad = \quad n^{-1} \left(\mathrm{E} \left(\boldsymbol{y} \mid \boldsymbol{X} \right) - \hat{\boldsymbol{y}} \right)^T \left(\mathrm{E} \left(\boldsymbol{y} \mid \boldsymbol{X} \right) - \hat{\boldsymbol{y}} \right).$$

It has been proved that, for any $\epsilon > 0$, we have

$$\lim_{n \to \infty} P \left(\left| \frac{L_n \text{ of the model selected by CV}}{\text{Minimum } L_n \text{ among all candidate models}} - 1 \right| < \epsilon \right) \quad = \quad 1.$$

Cross Validation and AIC

In fact, we can show that leave-one-out CV is asymptotically equivalent to AIC if the model performance is measured by

$$\sum_{i=1}^{n} \log f\left(y_i; \hat{\theta}_{-i}\right),$$

where $\hat{\theta}_{-i}$ is the estimator of θ after removing the ith observation.

Mallows Criterion

The Mallows criterion for linear regression is

$$C_p = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + 2\sigma^2 p,$$

where σ^2 is assumed to be known. If it is unknown, it is replaced by an unbiased estimator.

1. If σ^2 is known, model selections by C_p and AIC are equivalent under the normality assumption.
2. C_p is an unbiased estimator of the out-of-sample prediction mean squared error, if the model is correct.

Some Recommendations

1. The marginality principle means that when you include higher-order terms into the model, always include all lower-order (and same-order terms).

$$
\begin{aligned}
g(\mu_i) =& \beta_0 + (\beta_1 x_{i1} + \beta_2 x_{i2}) + \left(\beta_3 x_{i1}^2 + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i2}^2\right), \\
g(\mu_i) =& \beta_0 + (\beta_1 x_{i1} + \beta_2 x_{i2}) + \beta_4 x_{i1} x_{i2}, \\
g(\mu_i) =& \beta_0 + \beta_3 x_{i1}^2 + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i2}^2.
\end{aligned}
$$

2. For a qualitative predictor with more than two categories, we select the whole variable rather than individual dummy variables created from the qualitative predictor. Otherwise, the result may depend on the choice of reference category.

Ridge Regression

Suppose that we have demeaned both y and x. The ridge regression minimizes the penalized sum-of-squares

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta},$$

where $\lambda > 0$ is a tuning parameter.

1. The penalized sum-of-squares is equivalent to

$$\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \qquad \text{s.t. } \boldsymbol{\beta}^T \boldsymbol{\beta} \leq t,$$

   for some t.

2. The ridge estimator has a closed form expression

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

3. Unfortunately, the ridge regression does not achieve variable selection.

Lasso

Suppose that we have demeaned both y and x. The least absolute shrinkage and selection operator (lasso) minimizes

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_j |\beta_j| .$$

1. The penalized sum-of-squares is equivalent to

$$\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \qquad \text{s.t.} \sum_j |\beta_j| \leq t,$$

   for some t.
2. The lasso estimator will achieve sparsity: the estimates of some parameters can be exactly zero.

Elastic Net

Suppose that we have demeaned both y and x. The elastic net is a hybrid of both ridge and lasso:

$$\frac{1}{n}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)^{T}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)+\lambda\left[\alpha\sum_{j}\left|\beta_{j}\right|+\frac{1-\alpha}{2}\boldsymbol{\beta}^{T}\boldsymbol{\beta}\right],$$

where $\lambda > 0$ is the tuning parameter and $0 \leq \alpha \leq 1$ is also a tuning parameter.

1. $\alpha = 1$: lasso.
2. $\alpha = 0$: ridge.

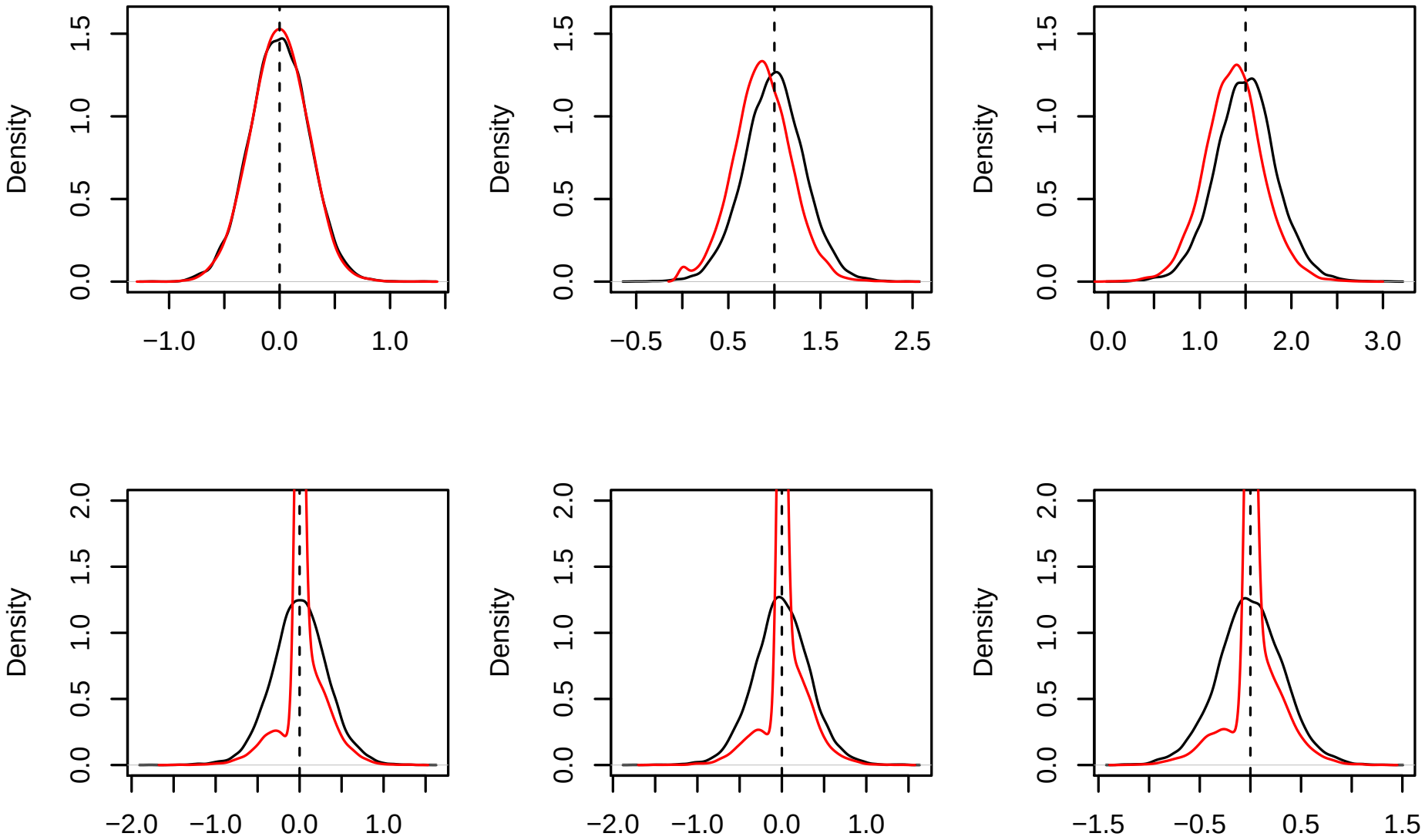
How to Choose λ

- Estimators under penalization are not invariant under rescaling, i.e., a model with meters can be radically different from a model with centimeters. For this reason, covariates are commonly standardized before entering the model (or it is a must to some extent).
- A grid search is commonly used: the optimal λ is chosen from a large number of pre-specified λ values.
- Commonly used criteria include AIC, BIC, CV, etc.

Consequence of Penalization

- bias-variance trade-off : If $\lambda = 0$, the usual estimator is obtained. If $\lambda > 0$, the bias-variance trade-off occurs.
- Shrinkage:
  - $\hat{\boldsymbol{\beta}}$ obtained for a positive λ is shrunk towards zero comparing with the ML estimator with $\lambda = 0$. $\hat{\boldsymbol{\beta}}$ is actually a function of λ.
  - The ridge never produces exact zero estimates, but the lasso and the elastic net can produce exact zero estimates.
- Variable selection: Lasso and elastic net conduct simultaneous parameter estimation and variable selection.
- Highly correlated regressors:
  - Lasso randomly selects one regressor from the set of highly correlated regressors.
  - Ridge and elastic net tend to perform better than lasso.
- High dimensional: They work when $n < p$.

Simulation with $n = 20$ and 5 Covariates



Generalization

- The lasso, ridge, and elastic net estimators are biased. In contrast, SCAD and MCP are supposed to be unbiased if the true values of regression coefficients are large.
- The group lasso can group the variables first and penalize the group of variables.
  - When dummy variables are created for qualitative variables, the dummy variables from the same qualitative variable are penalized together.
  - Variable selection is conducted to the qualitative variable, not only any individual dummy variables.
  - In contrast, the lasso may only include some dummy variables from the same variable, which depends on the choice of the reference group.
  - Lasso does not care about the marginality principle either.