HKU BUSINESS SCHOOL 港大經管學院

Business Statistics

Lecture 3: Model Selection and Regularization

Zhanrui Cai

Assistant Professor in Analytics and Innovation

ISLR Chapter 5, 6

Linear Model Selection

- In linear model

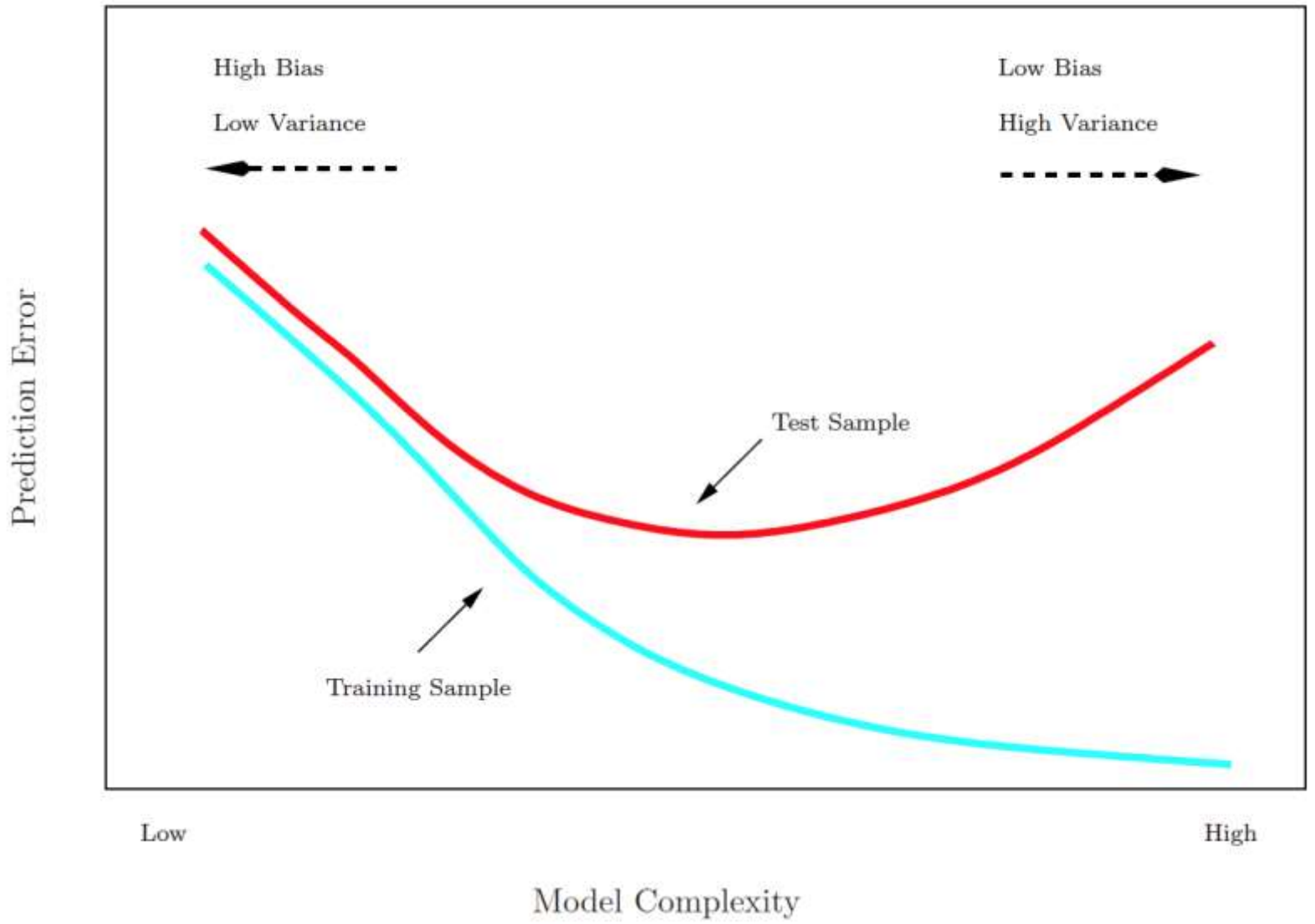
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

- Model selection – select the relevant X features.
  - Number of X variables
  - Which X variables

- Prediction Accuracy: especially when $p > n$, to control the variance and enable model fitting.

- Model Interpretability: by removing irrelevant features through setting the corresponding coefficients to be zero.

Model Complexity & Prediction Error



Estimating Test Error: Two Approaches

1. Indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting.

2. Directly estimate the test error, using either a validation set approach or a cross-validation approach.

1. Adjustment to the Training Error

- Adjusted RSS and R^2 (proposed in 1920s)
- Mallow's C_p (proposed in 1973)
- Akaike Information Criterion (AIC) (proposed in 1973)
- Bayesian Information Criterion (BIC) (proposed in 1973)

Adjusted R^2

- For a least squares model with p variables, the Adjusted R^2 statistic is calculated as

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}$$

  where TSS is the total sum of squares.

- A large Adjusted R^2 indicates a model with a small test error.

Adjusted R^2

- Maximize the Adjusted R^2 = Minimize $\frac{\text{RSS}}{n-p-1}$.
-
- Reason: Unlike R^2, Adjusted R^2 pays a price for the inclusion of unnecessary variables in the model.

Mallow's C_p

- Mallow's C_p:

$$C_p = \frac{1}{n}(\text{RSS} + 2p\hat{\sigma}^2).$$

- $\hat{\sigma}^2$ is the estimated variance of ϵ, from the fitted model.
- p : # of predictors in the model.
- Select the model with the smallest C_p.
- Balance between residuals and model sizes.

Akaike Information Criterion

- The AIC is defined as:

$$AIC = \frac{1}{n}(\text{RSS} + 2\,p\hat{\sigma}^2)$$

- Select the model with the smallest AIC.

- For linear regression models
  - C_p and AIC: equivalent in linear model

Bayesian Information Criterion

- The BIC is defined as

$$BIC = \frac{1}{n}(\text{RSS} + \log(n)p\hat{\sigma}^2).$$

- Select the model with the lowest BIC value.

- Notice that BIC replaces the $2p\hat{\sigma}^2$ used by C_p with a $\log(n)p\hat{\sigma}^2$ term, where n is the number of observations.

- Since $\log(n) > 2$ for any $n > 7$, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than C_p.

Subset Selection

Best subset and stepwise model selection procedures

Best Subset Selection

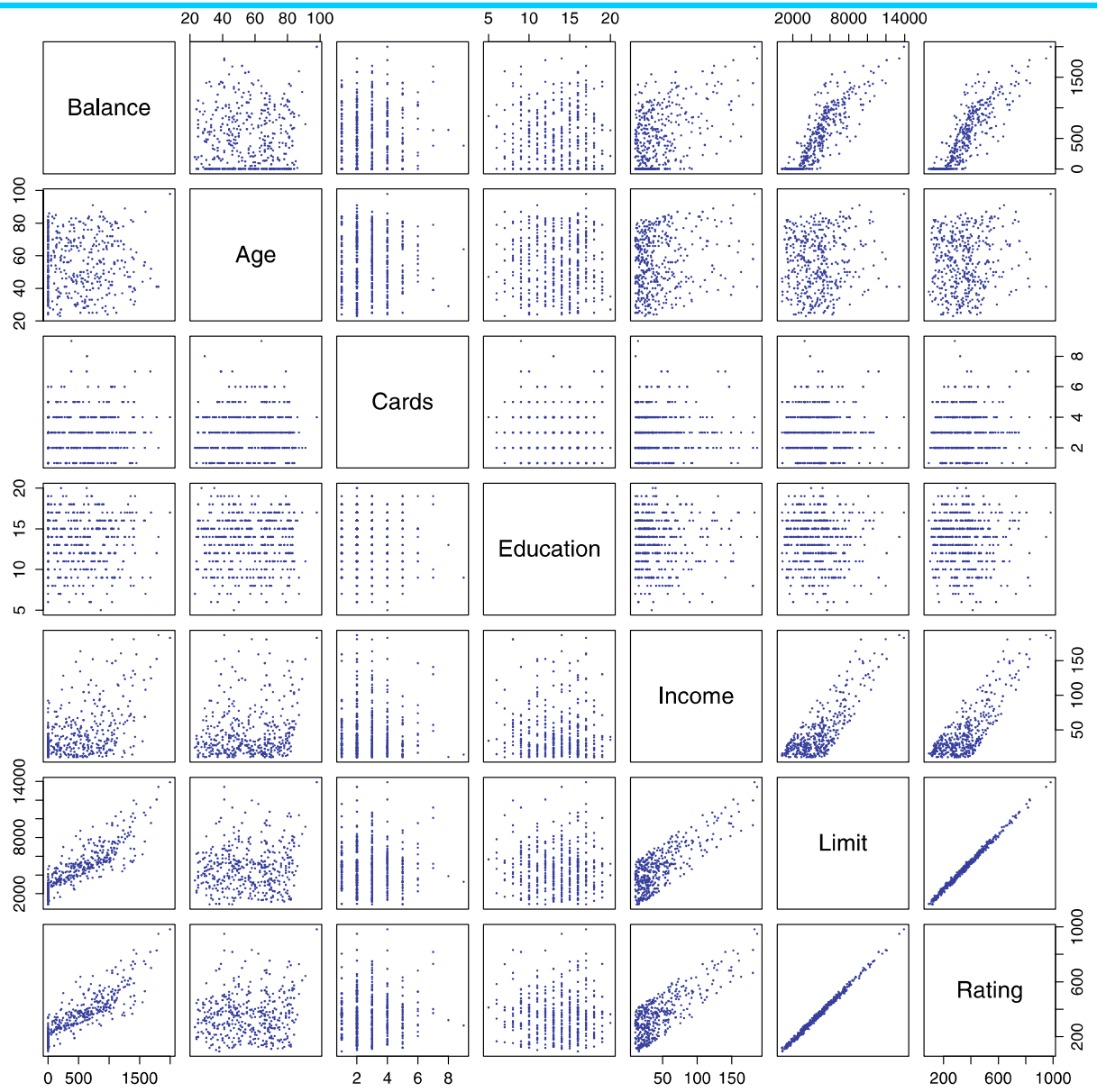
1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Example – Credit Data Set (Page 83)

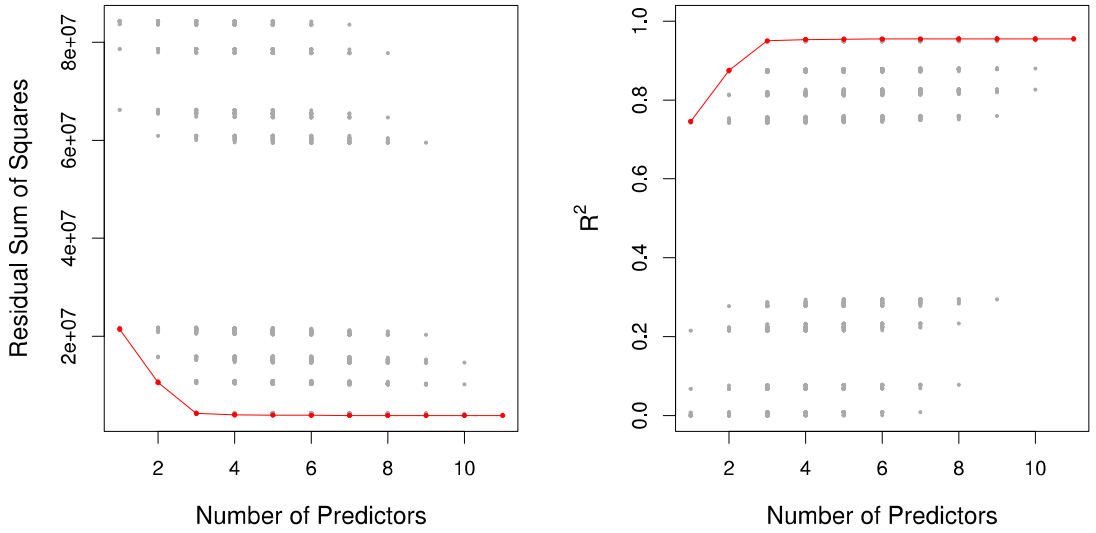
- Balance: average credit card debt for a number of individuals

- Age, Gender, Ethnicity (Caucasian, African American or Asian)
- Cards: number of credit cards
- Education: years of education
- Income (in thousands of dollars)
- Limit: credit limit
- Rating: credit rating
- Student (student status), Status (marital status), Own (house ownership), region (East, West or South)

Scatter Plot Matrix

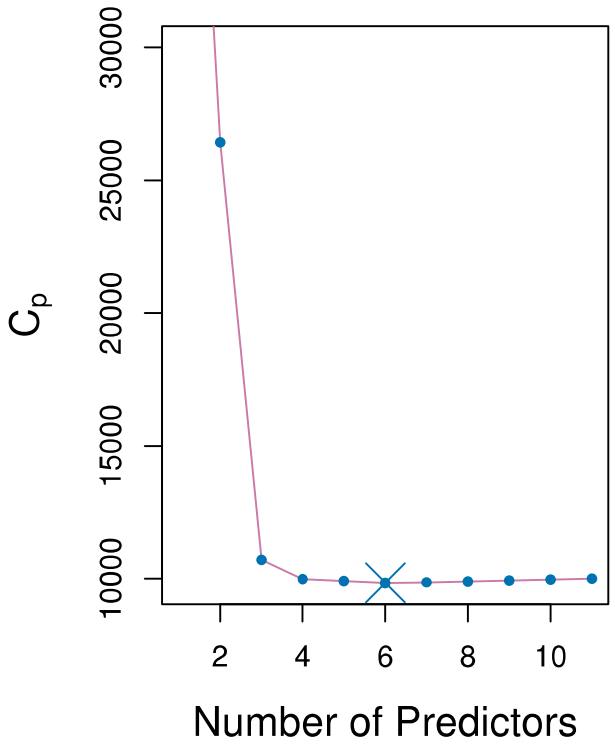


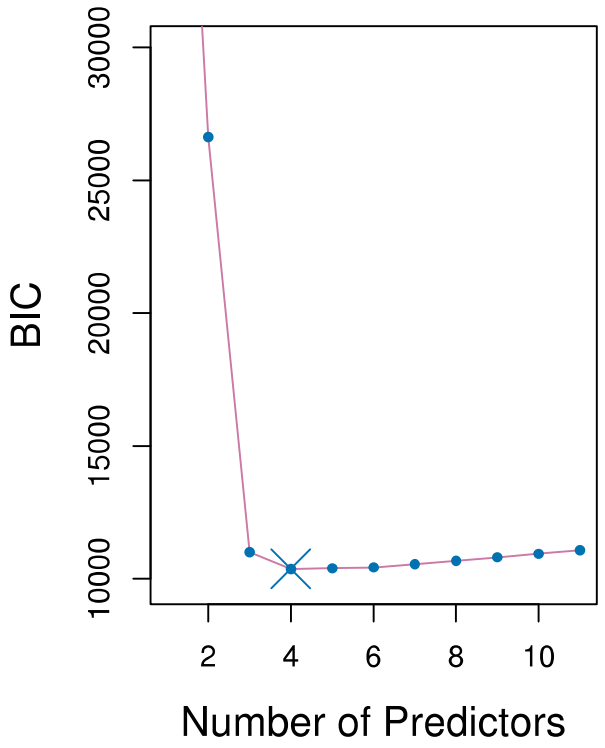
Example – Credit Data Set

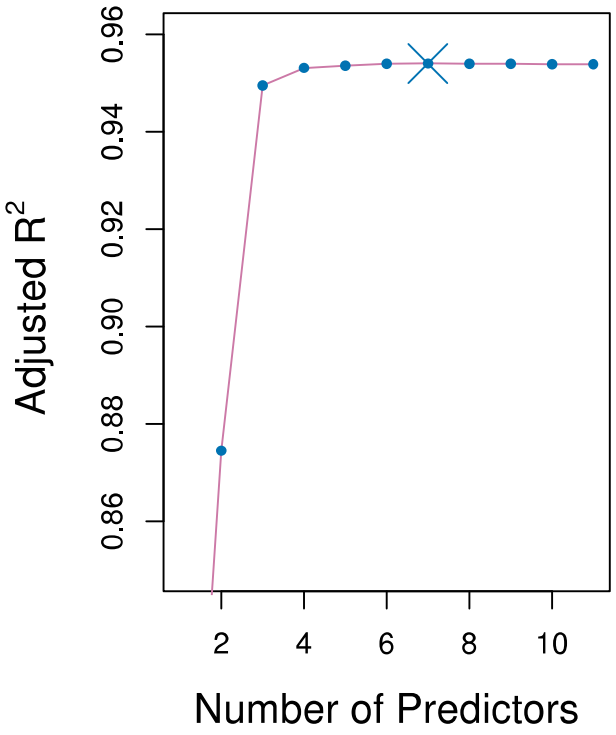


For each possible model containing a subset of the ten predictors in the `Credit` *data set, the RSS and* R^2 *are displayed. The red frontier tracks the best model for a given number of predictors, according to RSS and* R^2*. Though the data set contains only ten predictors, the* x*-axis ranges from* 1 *to* 11*, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables*

Credit Data Example: Best Subset







Subset Selection: Read Function Summary

Only x1, x2, x3 are useful for predicting y.

```
> ### perform subset selection
> model_all <- regsubsets(y ~ ., data = data, nvmax = p, method = "exhaustive")
> summary(model_all)
Subset selection object
Call: regsubsets.formula(y ~ ., data = data, nvmax = p, method = "exhaustive")
6 Variables  (and intercept)
   Forced in Forced out
x1     FALSE      FALSE
x2     FALSE      FALSE
x3     FALSE      FALSE
x4     FALSE      FALSE
x5     FALSE      FALSE
x6     FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: exhaustive
         x1  x2  x3  x4  x5  x6
1  ( 1 ) "*" " " " " " " " " " "
2  ( 1 ) "*" "*" " " " " " " " "
3  ( 1 ) "*" "*" "*" " " " " " "
4  ( 1 ) "*" "*" "*" " " "*" " "
5  ( 1 ) "*" "*" "*" "*" "*" " "
6  ( 1 ) "*" "*" "*" "*" "*" "*"
```

Extension to Other Models

- Although we have presented best subset selection here for least squares regression, the same ideas apply to other types of models, such as logistic regression (to be discussed).

- The loss functions for other models play the role of RSS for a broader class of models.

Stepwise Selection

- For computational reasons, best subset selection cannot be applied with large *p*.
- Best subset selection may also suffer from statistical problems when *p* is large:
  - The larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data.
- An enormous search space can lead to overfitting and high variance of the coefficient estimates.
- For both reasons, stepwise methods, which explore a far more restricted set of models, are attractive alternatives.

Forward Stepwise Selection

- Forward stepwise selection begins with a model with no predictors, then adds predictors one-at-a-time, until all of the predictors are in the model.
- At each step, the variable that gives the greatest additional improvement is added to the model.
- Computational advantage over best subset selection is clear.
- However, not guaranteed to find the best possible model compared to the subset selection.

Forward Stepwise Selection: Details

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   - 2.1 Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   - 2.2 Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Example: Credit Data (Page 231)

# Variables	Best subset	Forward stepwise
One	`rating`	`rating`
Two	`rating`, `income`	`rating`, `income`
Three	`rating`, `income`, `student`	`rating`, `income`, `student`
Four	`cards`, `income`	`rating`, `income`,
	`student`, `limit`	`student`, `limit`

The first four selected models for best subset selection and forward stepwise selection on the `Credit` *data set. The first three models are identical but the fourth models differ.*

Backward Stepwise Selection

- Like forward stepwise selection, backward stepwise selection is another efficient alternative to best subset selection.

- Start with the full least squares model containing all p predictors, then iteratively remove the least useful predictor, one at a time.

Backward Stepwise Selection: Details

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   - 2.1 Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   - 2.2 Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

More on Backward Stepwise Selection

- Both forward & backward selection searches only $1 + p(p + 1)/2$ models, much smaller than the subset selection 2^p models.

- Both forward & backward selection are not guaranteed to yield the best model containing a subset of the *p* predictors.

- Backward selection requires that the number of samples *n* is larger than the number of variables *p* (so that the full model can be fit). In contrast, forward stepwise can be used even when $n < p$, and so is the only viable subset method when *p* is very large.

Extensions

- Hybrid-stepwise selection considers both forward and backward moves at each step, and selects the best of the two.
  - Pros: computationally efficient, error made at an earlier stage can be corrected later.
  - Need a criterion to decide whether to add or drop at each step. e.g. AIC takes proper account of both the number of parameters and how good the model fits.

Implementation

- Download the code subset & stepwise.R and run the lines by yourself.
- OR Ask GPT:
  - How to perform forward and backward regression in R?

When n<p

- When the sample size n is smaller than the number of predictors p, the OLS (ordinary least square) method will fail!
- When p>n, the model can perfectly fit the training data.
- This leads to a model that captures noise rather than the underlying relationship, resulting in poor generalization of new data.
- Forward regression will work, i.e., add one variable at a time, until you have enough predictors.

2. Validation Set and Cross-Validation:

Directly Estimate Testing Error

Validation and Cross-Validation

- Each of the procedures returns a sequence of models $\mathcal{M}_k$ indexed by model size $k = 0, 1, 2, \ldots$. Our job here is to select $\hat{k}$. Once selected, we will return model $\mathcal{M}_{\hat{k}}$
- We compute the validation set error or the cross-validation error for each model $\mathcal{M}_k$ under consideration, and then select the k for which the resulting estimated test error is smallest.
- This procedure has an advantage relative to AIC, BIC, C_p, and adjusted R^2, in that it provides a direct estimate of the test error, and *doesn't require an estimate of the error variance* σ^2.
- It can also be used in a wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance σ^2.

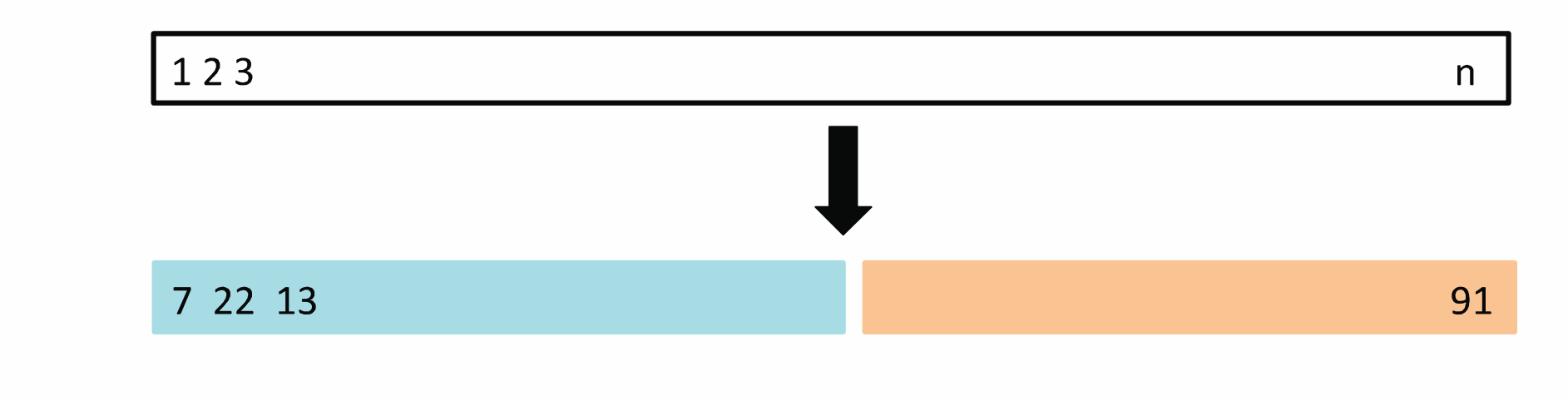
Validation-set Approach

- Goal: find the model with the smallest testing error.
- But no test data is available at the training stage!
- Create artificial testing data from training data!
- Key idea: train & evaluate model on different datasets.

Validation-set Approach

- We randomly divide the available set of samples into two parts: a training set and a validation or hold-out set.
- The model is fit on the training set, and then used to predict the responses for the observations in the validation set.
- The resulting validation-set error provides an estimate of the test error. Use MSE.

The Validation Process



A random splitting into two halves: left part is training set, right part is validation set

Drawbacks of the Validation Set Approach

- Only a subset of the observations are used to fit the model.
- The validation estimate of the test error can be highly variable, depending on the split of the raw data.
- The validation test error may tend to overestimate the test error for the model fit on the entire data set.
- Cross-validation (CV) to the rescue!

K-fold Cross-Validation (CV)

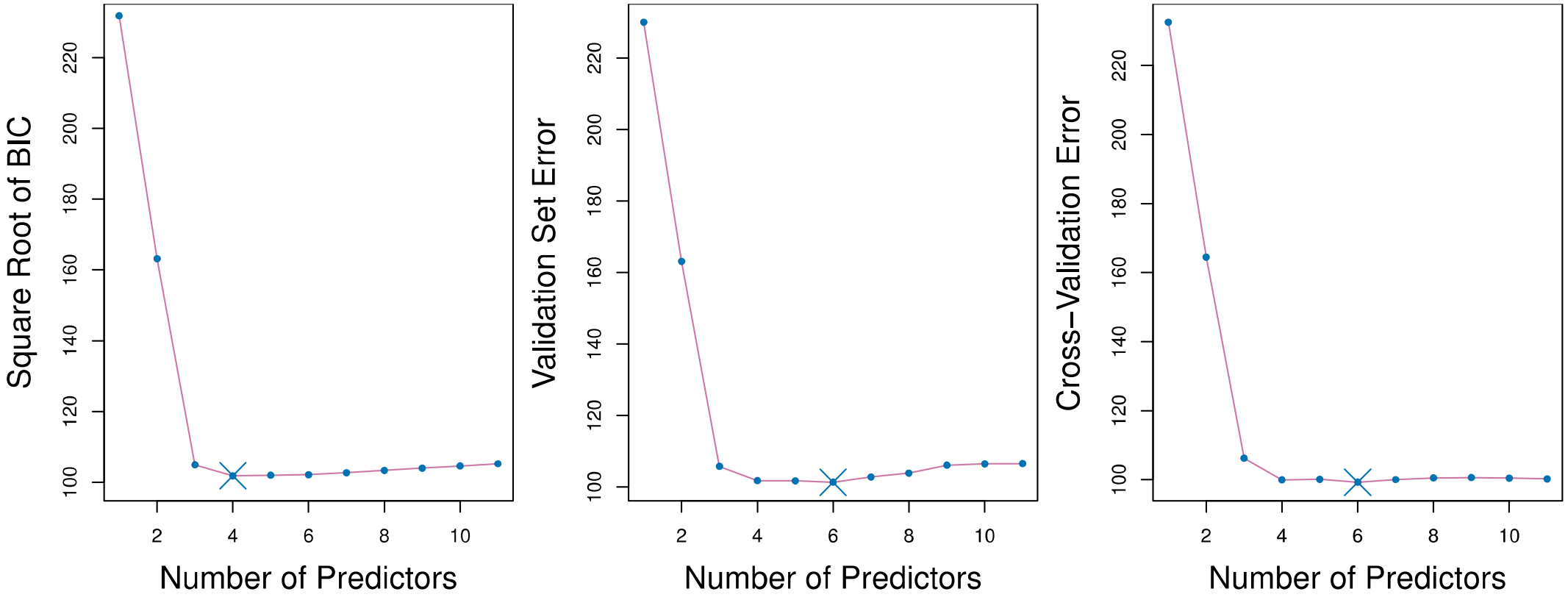
- *Widely used approach* for estimating test error.
- Estimates can be used to select best model, and to give an idea of the test error of the final chosen model.
- Idea is to randomly divide the data into K equal-sized parts. We leave out part k, fit the model to the other $K-1$ parts (combined), and then obtain predictions for the left-out kth part.
- This is done in turn for each part $k = 1, 2, \ldots K$, and then the results are combined.

K-fold CV Illustration

Divide data into K roughly equal-sized parts ($K = 5$ here)

1	2	3	4	5
Validation	Train	Train	Train	Train

Credit Data Example: Validation



Details of Previous Figure

- The validation errors were calculated by randomly selecting three-quarters of the observations as the training set, and the remainder as the validation set.
- The cross-validation errors were computed using $k = 10$ folds.
- In this case, the validation and cross-validation methods both result in a six-variable model.
- However, all three approaches suggest that the four-, five-, and six-variable models are roughly equivalent in terms of their test errors.

Shrinkage Methods

Shrinkage Methods

- The subset selection methods use Least Squares to fit a linear model that contains a subset of the predictors.
- As an alternative, we can fit a model containing all *p* predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero.
- Ridge regression and Lasso regression
- It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance.

Ridge Regression

- Recall that the least squares fitting procedure estimates $\beta_0, \beta_1, \ldots, \beta_p$ using the values that minimize

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

- In contrast, the ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize

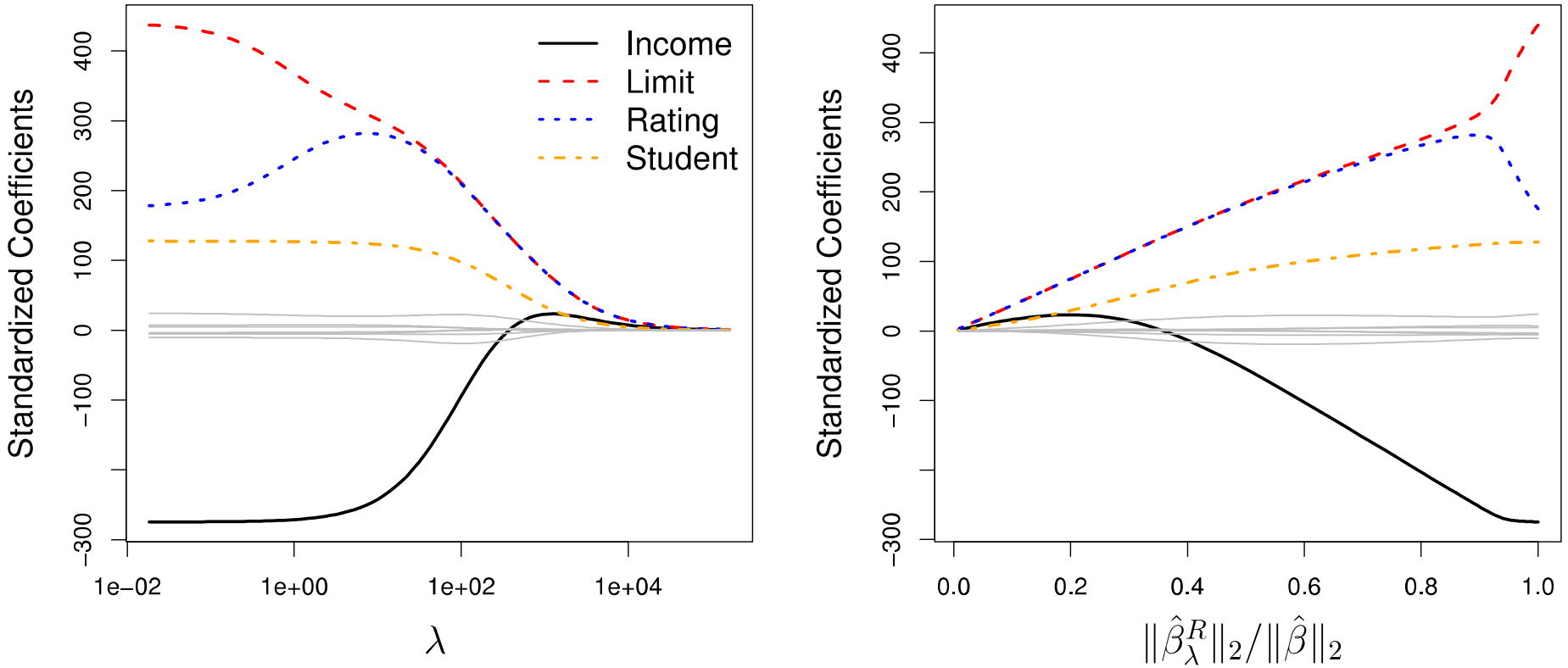
$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\lambda \geq 0$ is a *tuning parameter*, to be determined separately.

Ridge Regression Continued

- As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small.
- However, the second term, $\lambda \sum_j \beta_j^2$, called a *shrinkage penalty*, is small when $\beta_1, \ldots, \beta_p$ are close to zero, and so it has the effect of *shrinking* the estimates of β_j towards zero.
- The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates.
- Selecting a good value for λ is critical; cross-validation is used for this.

Example: Credit Data



Grey lines: unrelated variables, i.e. noise variables, in contrast with signal variables (colored ones).

Details of Previous Figure

- In the left-hand panel, each curve corresponds to the ridge regression coefficient estimate for one of the ten variables, plotted as a function of λ.
- The right-hand panel displays the same ridge coefficient estimates as the left-hand panel, but instead of displaying λ on the x-axis, we now display $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$, where $\hat{\beta}$ denotes the vector of least squares coefficient estimates.
- The notation $\|\beta\|_2$ denotes the ℓ_2 norm (pronounced "ell 2") of a vector, and is defined as $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} {\beta_j}^2}$.

Ridge Regression: Scaling of Predictors

- The standard least squares coefficient estimates are *scale equivariant*: multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of $1/c$. In other words, regardless of how the jth predictor is scaled, $X_j\hat{\beta}_j$ will remain the same.
- In contrast, the ridge regression coefficient estimates can change *substantially* when multiplying a given predictor by a constant, due to the sum of squared coefficients term in the penalty part of the ridge regression objective function.
- Therefore, it is best to apply ridge regression after *standardizing the predictors*, using the formula

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}}$$

The Bias-Variance Decomposition*

- Assume that

$$Y = f(X) + \varepsilon$$

where E(ε)=0 and Var(ε)=σ_ε^2.

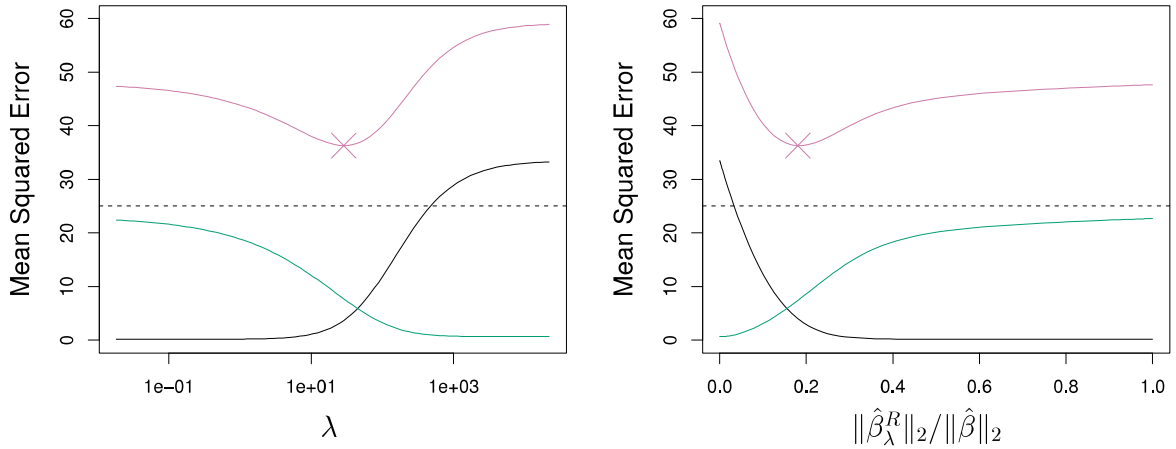
- At an input point $X = x_0$, the expected squared prediction error is

$$\begin{aligned}
\mathrm{Err}(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\
&= \sigma_\varepsilon^2 + [\mathrm{E}\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - \mathrm{E}\hat{f}(x_0)]^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0)) \\
&= \text{Irreducible Error} + \mathrm{Bias}^2 + \text{Variance}.
\end{aligned}$$

- The more complex the model, the lower the bias but the higher the variance.

Ridge Regression Improves Over Least Squares*

The Bias-Variance tradeoff



Simulated data with $n = 50$ observations, $p = 45$ predictors, all having nonzero coefficients. Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.

The Lasso

- Ridge regression does have one obvious disadvantage: unlike subset selection, which will generally select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model
- The *Lasso* is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficients, $\hat{\beta}_\lambda^L$, minimize the quantity

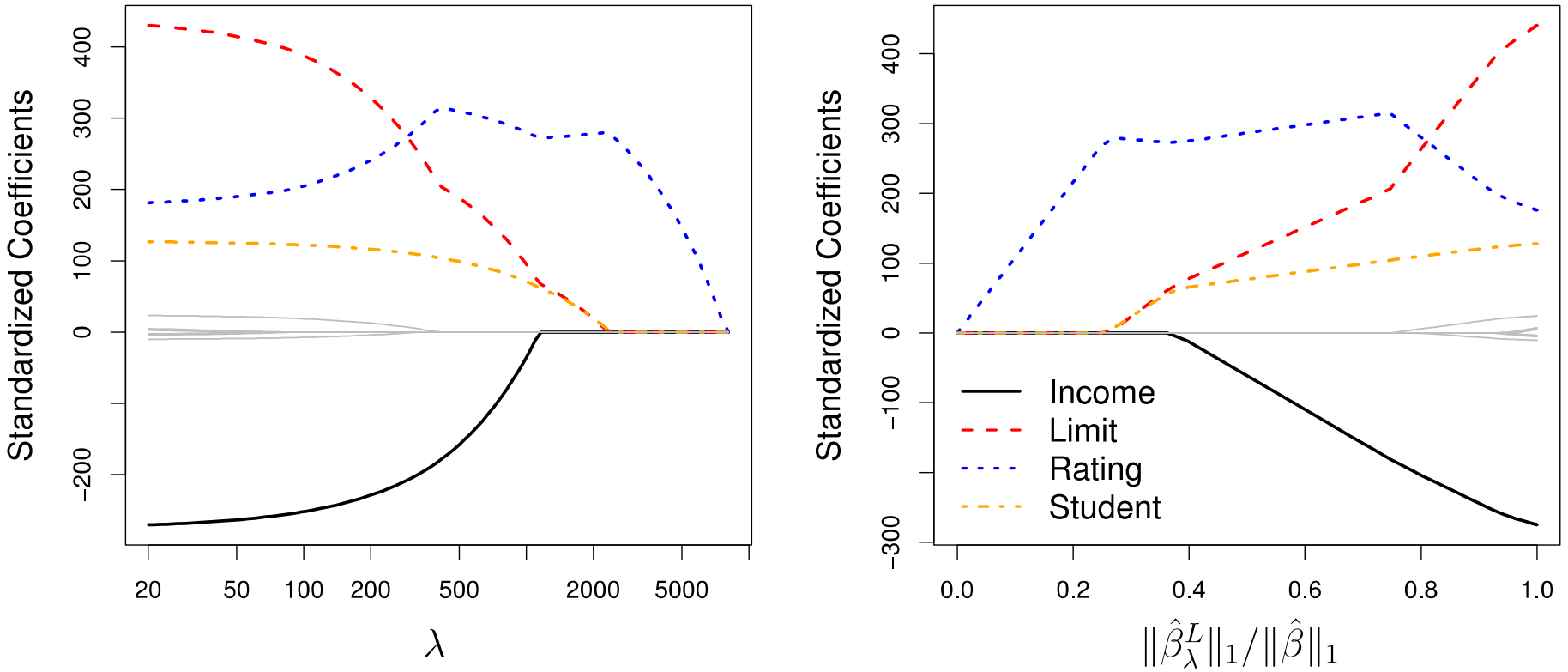
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \;= \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

- In statistical parlance, the lasso uses an ℓ_1 (pronounced "ell 1") penalty instead of an ℓ_2 penalty. The ℓ_1 norm of a coefficient vector β is given by $\|\beta\|_1 = \sum|\beta_j|$.

The Lasso Continued

- As with ridge regression, the lasso shrinks the coefficient estimates towards zero.
- However, in the case of the lasso, the ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.
- Hence, much like best subset selection, the lasso performs *variable selection*.
- We say that the lasso yields *sparse* models — that is, models that involve only a subset of the variables.
- As in ridge regression, selecting a good value of λ for the lasso is critical; cross-validation is again the method of choice.

Example Credit Data



The Variable Selection Property of the Lasso

Why is it that the lasso, unlike ridge regression, results in coefficient estimates that are exactly equal to zero?

One can show that the lasso and ridge regression coefficient estimates solve the problems

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

and

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s,$$

respectively.

Best Subset, Lasso, and Ridge

- Best Subset

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s.$$

- Lasso

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

- Ridge

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s$$

Tuning Parameter Selection

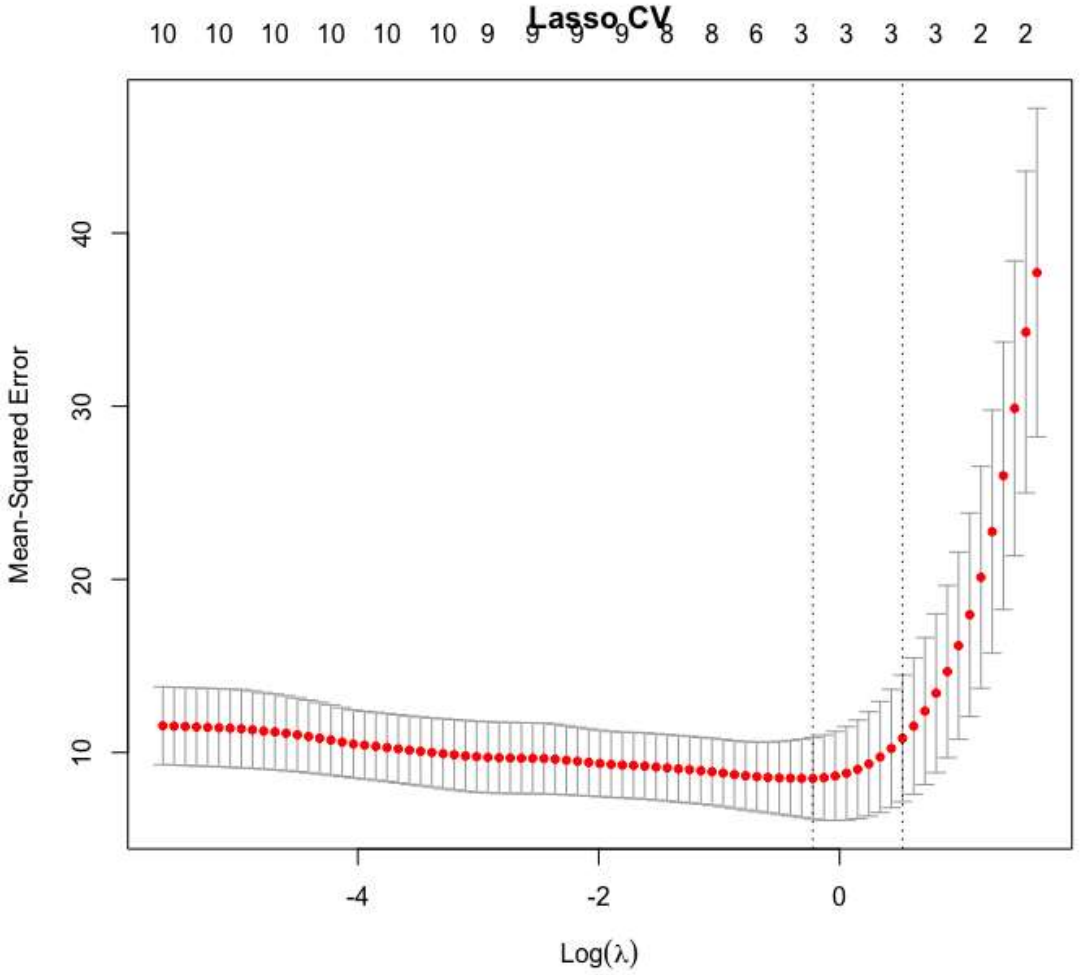
- Similar to subset selection, for ridge regression and lasso, need a method to determine which of the models under consideration is best.
  - Method to select a value for the tuning parameter λ or equivalently, the value of the constraint s.
- Cross-validation provides a simple way to tackle this problem.
  - Choose a grid of λ values, and compute the CV error rate for each value of λ.
  - Select the value of λ for which the CV error is the smallest.
  - Refit the model using all available observations and the selected value of λ.

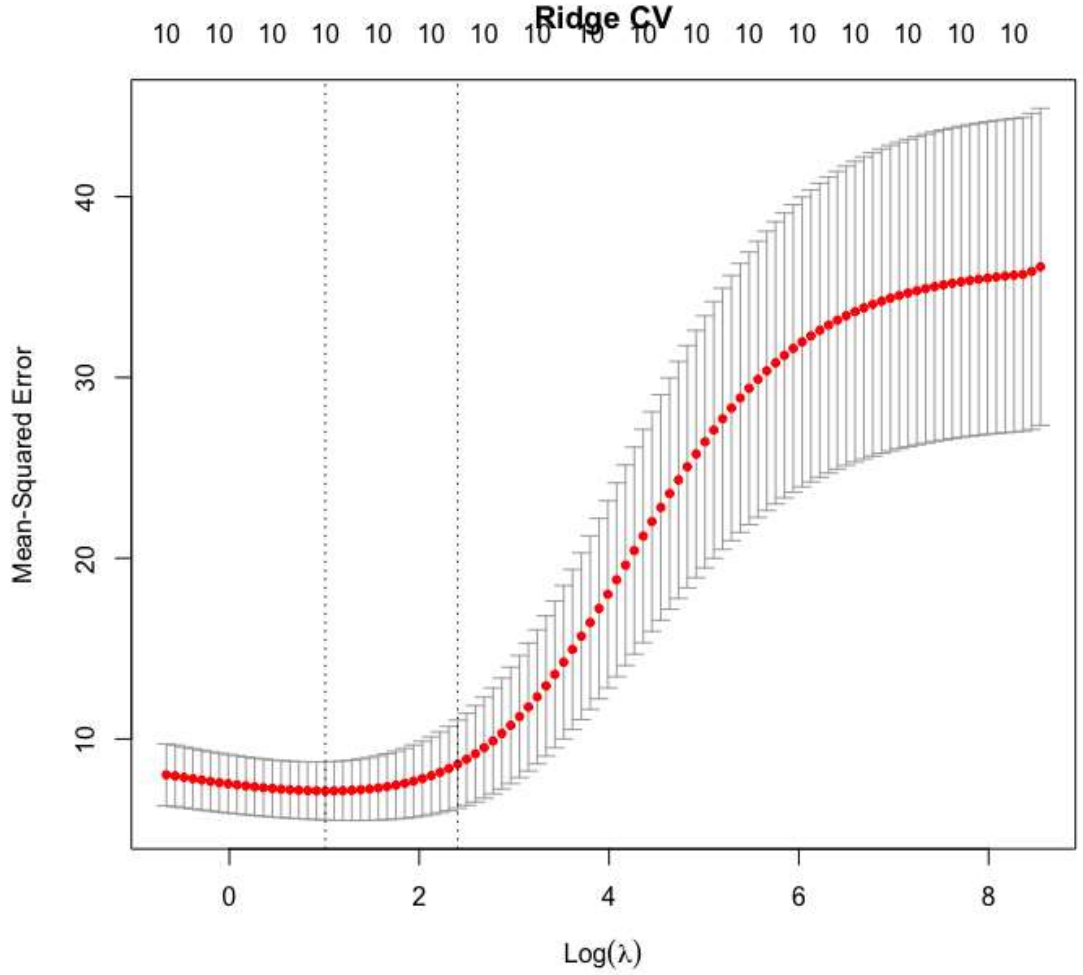
mtcars

- The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

- Response: mpg (miles per gallon)
- Covariates:
  - Cyl: number of cylinders
  - Hp: Gross horsepower
  - Am: Transmission (0 = automatic, 1 = manual)
  - …
  - …
  - 10 predictors in total

Tuning Parameter Selection





Implementation

- Download the code mtcars.R and run the lines by yourself.
- Ask GPT: How to use CV to choose the tunning parameter in Lasso and Ridge regression in R? Use the mtcars data in R to demonstrate it.
- Package: glmnet

Compare Lasso and Ridge

- Lasso
  - Variable selection
  - Model Interpretability
  - Most useful when only a small number of predictors is really influencing the response.

- Ridge
  - Handles collinearity among the covariates.
  - Simple closed form solutions for linear regression (optional).
  - Better testing error especially when many covariates are useful.
  - Can not perform variable selection.

Comparison Conclusions

- Neither ridge nor the lasso will universally dominate the other.
- In general, one might expect the lasso to perform better when the response is a function of only a relatively small number of predictors.
- However, the number of predictors that is related to the response is never known a priori for real data sets.
- Cross-validation can be used in order to determine which approach is better on a particular data set.

Post Model Selection

- How to get p-values for the selected variables?
- Fit an OLS on the selected model.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
cyl         -0.94162    0.55092  -1.709 0.098480 .
hp          -0.01804    0.01188  -1.519 0.140015
wt          -3.16697    0.74058  -4.276 0.000199 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared:  0.8431,    Adjusted R-squared:  0.8263
F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11
```

Elastic Net

- Recall the Lasso
- $$RSS + \lambda||\beta||_1$$
- And the Ridge
- $$RSS + \lambda||\beta||_2$$
- Elastic net: combination
- $$RSS + \lambda(\alpha||\beta||_1 + (1-\alpha)||\beta||_2)$$

Elastic Net

- When $\alpha = 1$, elastic net is equivalent to the Lasso regression.
- When $\alpha = 0$, elastic net is equivalent to the ridge regression.

- Advantage:
  - Flexible
  - Combine the strength of Lasso and Ridge regression.

Elastic Net

- Can you try to apply the elastic net to the mtcars dataset, and find the optimal α that has the smallest CV error?

Summary

- Model selection methods are an essential tool for data analysis, especially for big datasets involving many predictors.
- Research into methods that give sparsity, such as the lasso is an especially hot area.
- There are other sparsity related approaches such as the elastic net.
  - Regularization via L_0, L_1, L_2, L_q, or combinations of them
  - Elastic net: $L_1 + L_2$