Regression, Classification, Model Assessment and Selection

Yuan YAO

Hong Kong University of Science and Technology

Department of Mathematics

February 22, 2019

Regression

- The least squares estimation
- The statistical properties of the least squares estimates.

Classification

- Linear Discriminant Analysis
- Logistic Regression.

Model Assessment

- Cross Validation
- Bootstrap

Feature/Variable Selection

- Subset selection
- Shrinkage methods (Ridge/Lasso)

Outline

Regression
- The least squares estimation
- The statistical properties of the least squares estimates.

Classification
- Linear Discriminant Analysis
- Logistic Regression.

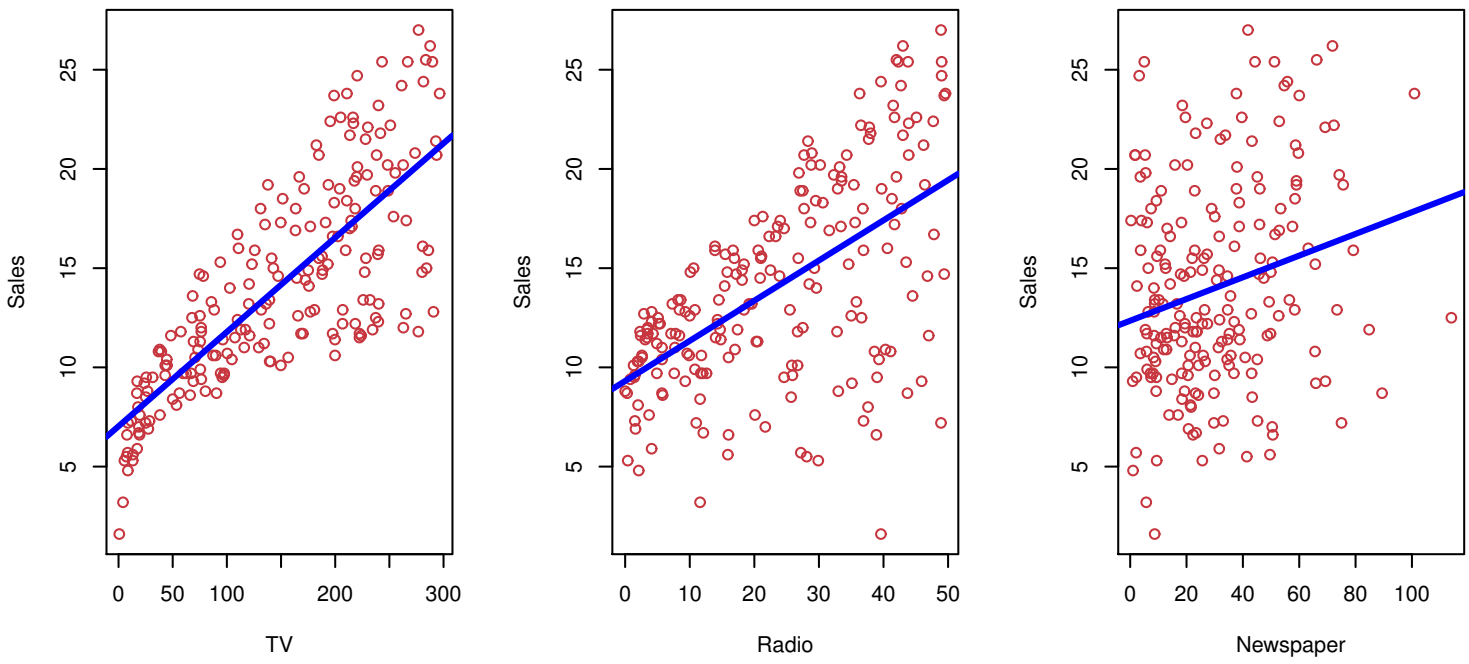
Model Assessment
- Cross Validation
- Bootstrap

Feature/Variable Selection
- Subset selection
- Shrinkage methods (Ridge/Lasso)

Example: Advertising data

The data contains 200 observations.
Sample size: $n = 200$.
Sales: y_i, $i = 1..., n$.
TV (bugdets): $x_{i1}, i = 1, ..., n$.
Radio (budgets): $x_{i2}, i = 1, ..., n$.
Newspaper (budgets): $x_{i3}, i = 1, ..., n$.
Dimensionality: $p = 3$.

Example: Advertising data



Linear models formulation

- Consider linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, ..., n \qquad (2.1)$$

  where $y_i, x_i = (x_{i1}, ..., x_{ip})$ are the i-th observation of the response and covariates.
- Responses are sometimes called dependent variables or outputs;
- covariates called independent variables, or predictors or features or inputs or regressors.
- noise ϵ_i is independent, of zero mean, and fixed but unknown variance, e.g. Gaussian noise $N(0, \sigma^2)$

Example: Advertising data

Now, we consider three covariates: TV, radio and newspapers.
The number of covariates (predictors, or features): $p = 3$.
The linear regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \qquad i = 1, ..., n$$

Estimating the coefficient by the least squares

Minimizing the sum of squares of error (Gauss'1795):

$$\min_{\beta_0,\ldots,\beta_3} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3})^2.$$

Figure: Carl Friedrich Gauss

Notations

With slight abuse of notation, in this chapter, we use

$$\begin{aligned}\mathbf{X} &= \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{1} \vdots \mathbf{x}_1 \vdots \mathbf{x}_2 \vdots \cdots \vdots \mathbf{x}_p \end{pmatrix}.\end{aligned}$$

Here a column of ones, $\mathbf{1}$, is added, which corresponds to the intercept β_0. Then $\mathbf{X}$ is a n by $p+1$ matrix.

Recall that

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad \mathbf{x}_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

The least squares criterion

The least squares criterion is try to minimize the sum of squares:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2.$$

Using matrix algebra, the above sum of squares is

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta).$$

The LSE, fitted values and residuals

By some linear algebra calcuation, the least squares estimator of β is then

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Then

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

is called the fitted values; viewed as the predicted values of the responses based on the linear model.

$$\hat{\epsilon} = \mathbf{y} - \hat{\mathbf{y}}$$

are called residuals. The sum of squares of these residuals

$$RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2.$$

Orthogonality

- The residual $\hat{\epsilon}$ is orthogonal to all columns of $\mathbf{X}$, i.e, all $\mathbf{1}, \mathbf{x}_1, ..., \mathbf{x}_p$. This can be seen by

$$\begin{aligned} & \mathbf{X}^T \hat{\epsilon} = \mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{X} \hat{\beta} \\ = \quad & \mathbf{X}^T \mathbf{y} - \mathbf{X}^\mathbf{T} \mathbf{H} \mathbf{y} = 0. \end{aligned}$$

  where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T = \mathbf{H}^2$ is the projection.
- The residual vector $\hat{\epsilon}$ is orthogonal to the hyperplane formed by vectors $\mathbf{1}, \mathbf{x}_1, ..., \mathbf{x}_p$ in n dimensional real space.

The least squares projection

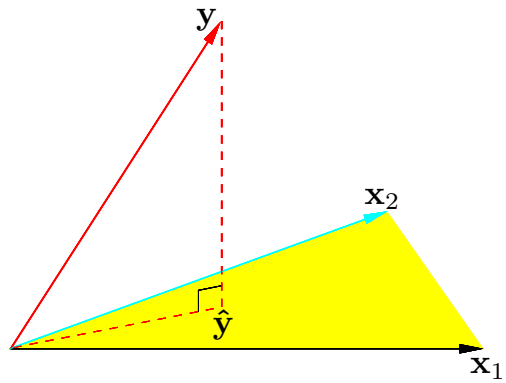


FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector $\mathbf{y}$ is orthogonally projected onto the hyperplane spanned by the input vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. The projection $\hat{\mathbf{y}}$ represents the vector of the least squares predictions*

Result of the estimation

TABLE 3.9. (from ISLR) The advertising data: coefficients of the LSE for the regression on number of units sold on TV, radio and newspaper advertising budgets.

	Coefficient	Std.error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Variable selection

- We may be concerned with a subset of the p variables are irrelevant with the response.
- Let the subset be denoted as $A = \{i_1, ..., i_r\}$, where $r \leq p$. Then, the null hypothesis is

$$H_0 : \beta_{i_1} = \beta_{i_2} = \cdots = \beta_{i_r} = 0,$$

which again is equivalent to

$$H_0 : E(\mathbf{y}) \in \mathcal{L}(A^c),$$

where $\mathcal{L}(A)$ is the linear space in R^n spanned by $\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_r}$, which is of r dimension.

Statistical properties of LSE

Assume $\epsilon_i \sim N(0, \sigma^2)$,

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

$$\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \sim \sigma^2 \chi^2_{n-p-1}$$

$\hat{\beta}$ and RSS are independent

$s^2 = \text{RSS}/(n-p-1)$ unbiased estimate of σ^2

$$\frac{\hat{\beta}_j - \beta_j}{s\sqrt{c_{jj}}} \sim t_{n-p-1}$$

$$\frac{(\hat{\beta}-\beta)^T(\mathbf{X}^T\mathbf{X})(\hat{\beta}-\beta)/p}{s^2} \sim F_{p+1,n-p-1}$$

where $c_{00}, c_{11}, ..., c_{pp}$ are the diagonal elements of $(\mathbf{X}^T\mathbf{X})^{-1}$.

Confidence intervals

For example,

$$\hat{\beta}_j \pm t_{n-p-1}(\alpha/2) s \sqrt{c_{jj}}$$

is a confidence interval for β_j at confidence level $1 - \alpha$. Here $t_{n-p-1}(\alpha/2)$ is the $1 - \alpha/2$ percentile of the t-distribution with degree of freedom $n - p - 1$.

Outline

- Regression
  - The least squares estimation
  - The statistical properties of the least squares estimates.
- Classification
  - Linear Discriminant Analysis
  - Logistic Regression.
- Model Assessment
  - Cross Validation
  - Bootstrap
- Feature/Variable Selection
  - Subset selection
  - Shrinkage methods (Ridge/Lasso)

Examples: The default data

- Simulated data: 10000 individuals.
- Two inputs: income and balance (monthly)
- One output: Default (Yes or No).
- Judge if a trading activity is a fraud or not.

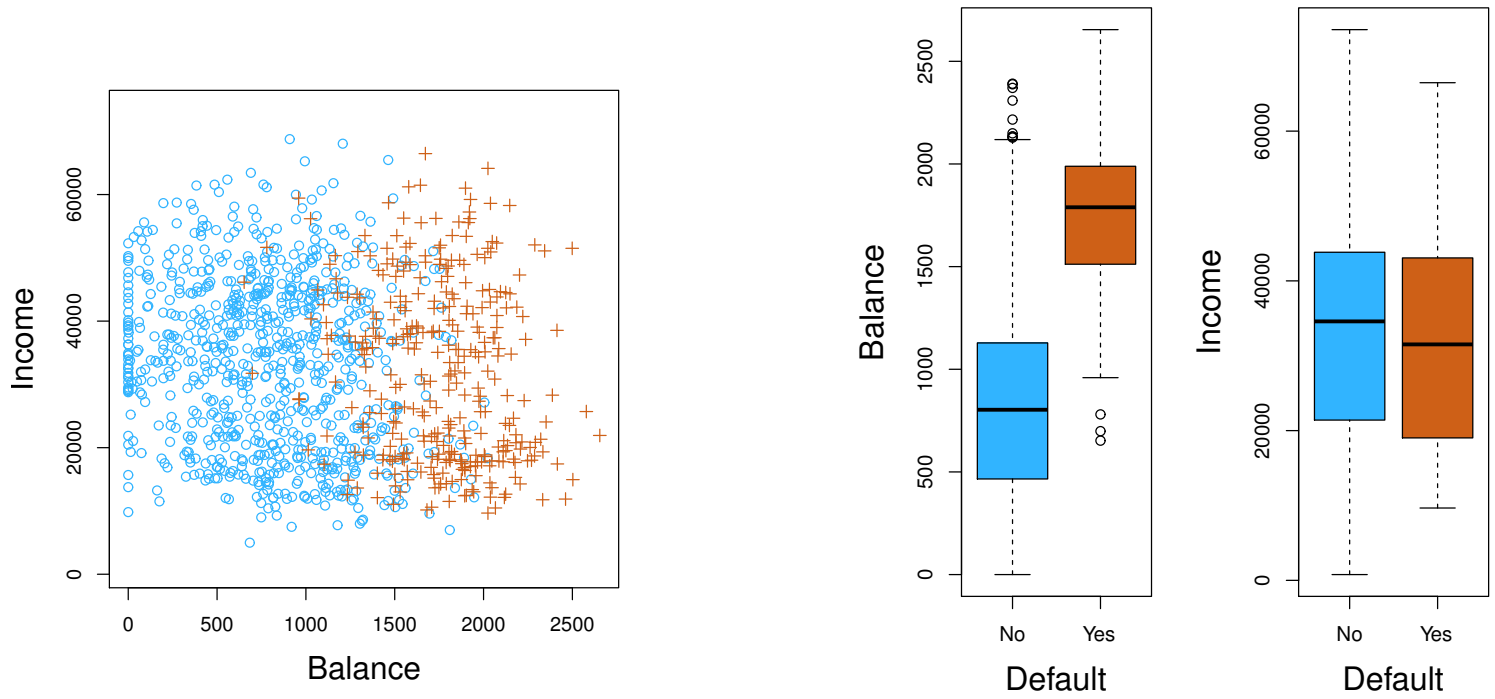


Figure: FIGURE 4.1. The Default data set. Left: The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue. Strong relation between balance and default while a weaker relation between income and default.

The special case of binary response

- Consider the output is binary: two class,
- Code the response into 0 and 1 and apply linear regression produce the same result as linear discriminant anlaysis.
- Not the case for output with more than two classes.

Linear Discriminant Analysis

We choose the k-th class such that the following score is the largest:

$$\hat{\delta}_k(x) = \hat{\mu}_k^T \hat{\Sigma}^{-1} x - \frac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \mu_k + \log \hat{\pi}_k,$$

where $\hat{\mu}_k$ and $\hat{\Sigma}$ and $\hat{\pi}_k$ are sample mean, pooled sample variance, and sample proportion of class k in the data. Specifically, based on data $(x_i, y_i), i = 1, ..., n$,

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i;$$

$$\hat{\Sigma} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T,$$

where n_k is the number of subjects in class k in the data, and

$$\hat{\pi}_k = n_k / n$$

In summary, we classify a subject with input x into class k, if $\hat{\delta}_k$ is the largest.
$\hat{\delta}_k(x)$ is called discrimination function.
$\hat{\delta}_k(x)$ is here a linear function of x. This is why we call it LDA.
The region of values of x being classified into a class has linear boundary, since $\delta_k(x) = \delta_l(x)$ defines a linear hyperplane for x.

Example: The default data

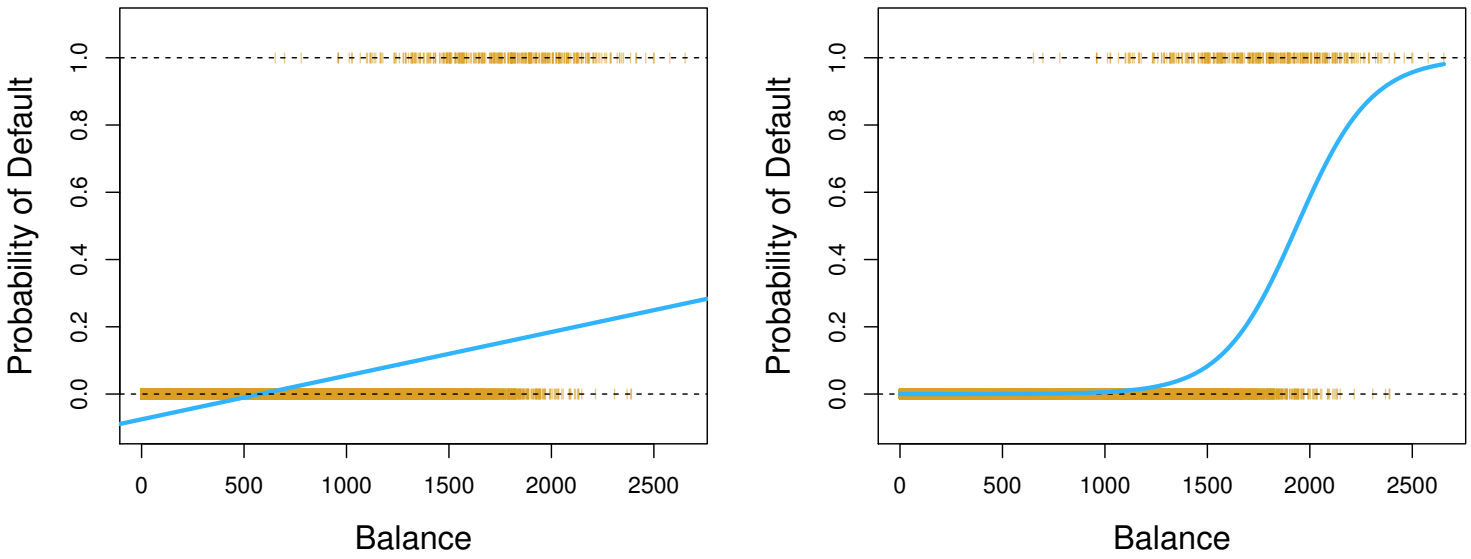


Figure: Left: linear regresion; Right: logistic regression

The setup for binary output

- The training data: $(\mathbf{x}_i, y_i)$: $i = 1, ..., n$.
- $y_i = 1$ for class 1 and $y_i = 0$ for class 0.
- $\mathbf{x}_i = (1, x_{i1}, ..., x_{ip})$ are $p + 1$ vectors with actually p inputs.
- If instead consider linear regression model is

$$y_i = \beta^T \mathbf{x}_i + \epsilon_i$$

  β can be estimated by the least squares, and $\hat{\beta}^T \mathbf{x}_i$ is the predictor of y_i.
- Key idea: should focus on predicting the probability of the classes.
- Using $P(y = 1|\mathbf{x}) = \beta^T \mathbf{x}$ is not appropriate.

The logistic regression model

- Assume

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-\beta^T \mathbf{x})}$$

  As a result,

$$P(y = 0|\mathbf{x}) = 1 - P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(\beta^T \mathbf{x})}$$

- $$\log(\frac{P(y = 1|\mathbf{x})}{P(y = 0|\mathbf{x})}) = \beta^T \mathbf{x}$$

  This is called log-odds or logit. And

$$\frac{P(y = 1|\mathbf{x})}{P(y = 0|\mathbf{x})}$$

  is called odds.
- Interpretation: one unit increase in variable x_j, increases the log-odds of class 1 by β_j.

The maximum likelihood estimation

- Recall that, the likelihood is the joint probability function of joint density function of the data.
- Here, we have independent observations $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$, each follow the (conditional) distribution

$$P(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + \exp(-\beta^T \mathbf{x}_i)} = 1 - P(y_i = 1|\mathbf{x}_i).$$

- So, the joint probability function is

$$\prod_{i=1,...,n;y_i=1} p(y_i = 1|\mathbf{x}_i) \prod_{i=1,...,n;y_i=0} p(y_i = 0|\mathbf{x}_i)$$

  which can be conveniently written as

$$\prod_{i=1}^{n} \frac{\exp(y_i \beta^T \mathbf{x}_i)}{1 + \exp(\beta^T \mathbf{x}_i)}.$$

The likelihood and log-likelihood

- The likelihood function is the same as the joint probability function, but viewed as a function of β.
- The log-likelihood function is

$$\sum_{i=1}^{n}[y_i\beta^T x_i - \log(1+\exp(\beta^T \mathbf{x}_i))]$$

- The maximizor is denoted as $\hat{\beta}$, which is the MLE of β based on logistic model.

Example: the default data with all three inputs: balance, income and student

TABLE 4.3. For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using balance, income, and student status. Student status is encoded as a dummy variable student[Yes], with a value of 1 for a student and a value of 0 for a non-student. In fitting this model, income was measured in thousands of dollars.

	Coefficient	Std.error	t-statistic	p-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
Balance	0.0057	0.0002	24.74	< 0.0001
Income	0.0030	0.0082	0.37	0.7115
Student[Yes]	-0.6468	0.2362	-2.74	0.0062

Outline

Regression
- The least squares estimation
- The statistical properties of the least squares estimates.

Classification
- Linear Discriminant Analysis
- Logistic Regression.

Model Assessment
- Cross Validation
- Bootstrap

Feature/Variable Selection
- Subset selection
- Shrinkage methods (Ridge/Lasso)

Training error is not sufficient enough

- training error easily computable with training data.
- because of possibility of over-fit, it cannot be used to properly assess test error.
- It is possible to "estimate" the test error, by, for example, making adjustments of the training error.
- General purpose method of prediction/test error estimate: validation.

Ideal scenario for performance assessment

- In a "data-rich" scenario, we can afford to separate the data into three parts:
  - training data: used to train various models.
  - validation data: used to assess the models and identify the best.
  - test data: test the results of the best model.
- Usually, people also call validation data or hold-out data as test data.

Cross validation: overcome the drawback of validation set approach

- Our ultimate goal is to produce the best model with best prediction accuracy.
- Validation set approach has a drawback of using ONLY training data to fit model.
- The validation data do not participate in model building but only model assessment.
- A "waste" of data.
- We need more data to participate in model building.

K-fold cross validation

- Divide the data into K subsets, usually of equal or similar sizes (n/K).
- Treat one subset as validation set, the rest together as a training set. Run the model fitting on training set. Calculate the test error estimate on the validation set, denoted as MSE_i, say.
- Repeat the procedures over every subset.
- Average over the above K estimates of the test errors, and obtain

$$\text{CV}_{(K)} = \frac{1}{K}\sum_{i=1}^{K}\text{MSE}_i$$

- Leave-One-Out Cross Validation (LOOCV) is a special case of K-fold cross validation, actually n-fold cross validation.

K-fold cross validation

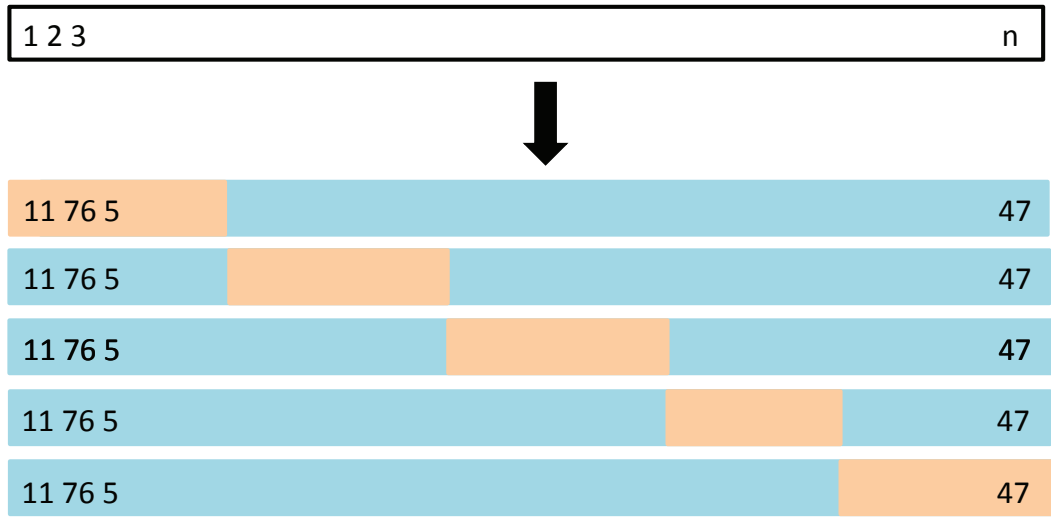


Figure: 5.5. A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.

Bootstrap as a resampling procedure.

- Suppose we have data $x_1, ..., x_n$, representing the ages of n randomly selected people in HK.
- Use sample mean $\bar{x}$ to estimate the population mean μ, the avearge age of all residents of HK.
- How to justify the estimation error $\bar{x} - \mu$? Usually by t-confidence interval, test of hypothesis.
- They rely on normality assumption or central limit theorm.
- Is there another reliable way?
- Just bootstrap:

Boostrap as a resampling procedure.

- Take n random sample (with replacement) from $x_1, ..., x_n$.
- calculate the sample mean of the "re-sample", denoted as $\bar{x}_1^*$.
- Repeat the above a large number M times. We have $\bar{x}_1^*, \bar{x}_2^*, ..., \bar{x}_M^*$.
- Use the distribution of $\bar{x}_1^* - \bar{x}, ..., \bar{x}_M^* - \bar{x}$ to approximate that of $\bar{x} - \mu$.

Example

- X and Y are two random variables. Then minimizer of $\text{var}(\alpha X + (1-\alpha)Y))$ is

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

- Data: $(X_1, Y_1), ..., (X_n, Y_n)$.
- We can compute sample variances and covariances.
- Estimate α by

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$

- How to evalute $\hat{\alpha} - \alpha$, (remember $\hat{\alpha}$ is random and α is unknown).
- Use Bootstrap

Example

- Sample n resamples from $(X_1, Y_1), ..., (X_n, Y_n)$, and compute the sample the sample variance and covariances for this resample. And then compute

$$\hat{\alpha}^* = \frac{(\hat{\sigma}_Y^*)^2 - \hat{\sigma}_{XY}^*}{(\hat{\sigma}_X^*)^2 + (\hat{\sigma}_Y^*)^2 - 2\hat{\sigma}_{XY}^*}$$

- Repeat this procedure, and we have $\hat{\alpha}_1^*, ..., \hat{\alpha}_M^*$ for a large M.
- Use the distribution of $\hat{\alpha}_1^* - \hat{\alpha}, ..., \hat{\alpha}_M^* - \hat{\alpha}$ to approximate the distribution of $\hat{\alpha} - \alpha$.
- For example, we can use

$$\frac{1}{M}\sum_{j=1}^{M}(\hat{\alpha}_j^* - \hat{\alpha})^2$$

  to estimate $E(\hat{\alpha} - \alpha)^2$.
- Use Bootstrap

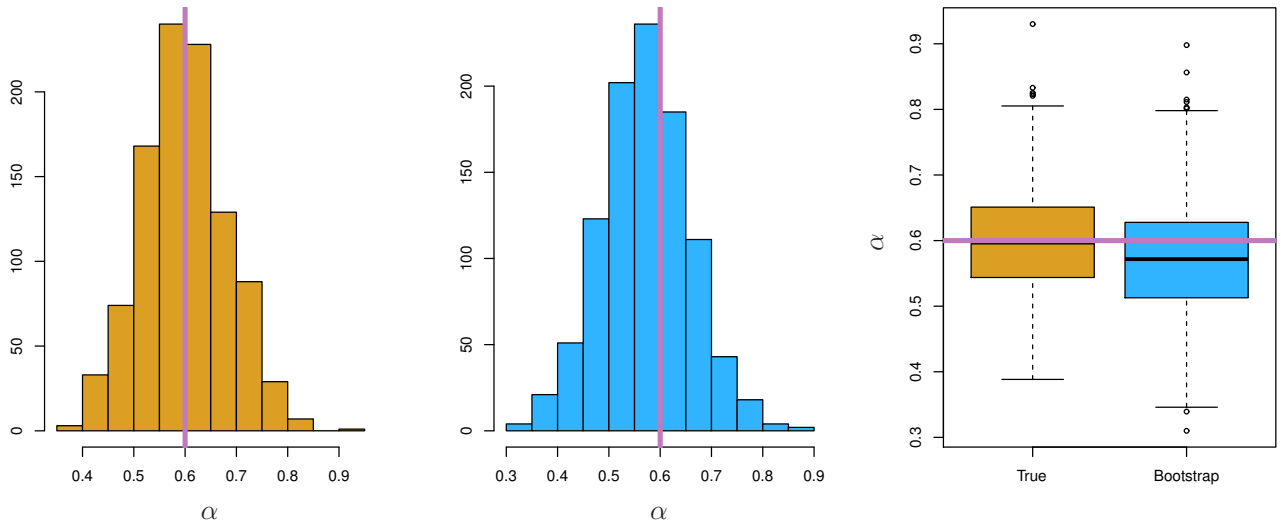


Figure: 5.10. Left: A histogram of the estimates of α obtained by generating 1,000 simulated data sets from the true population. Center: A histogram of the estimates of α obtained from 1,000 bootstrap samples from a single data set. Right: The estimates of α displayed in the left and center panels are shown as boxplots. In each panel, the pink line indicates the true value of α.

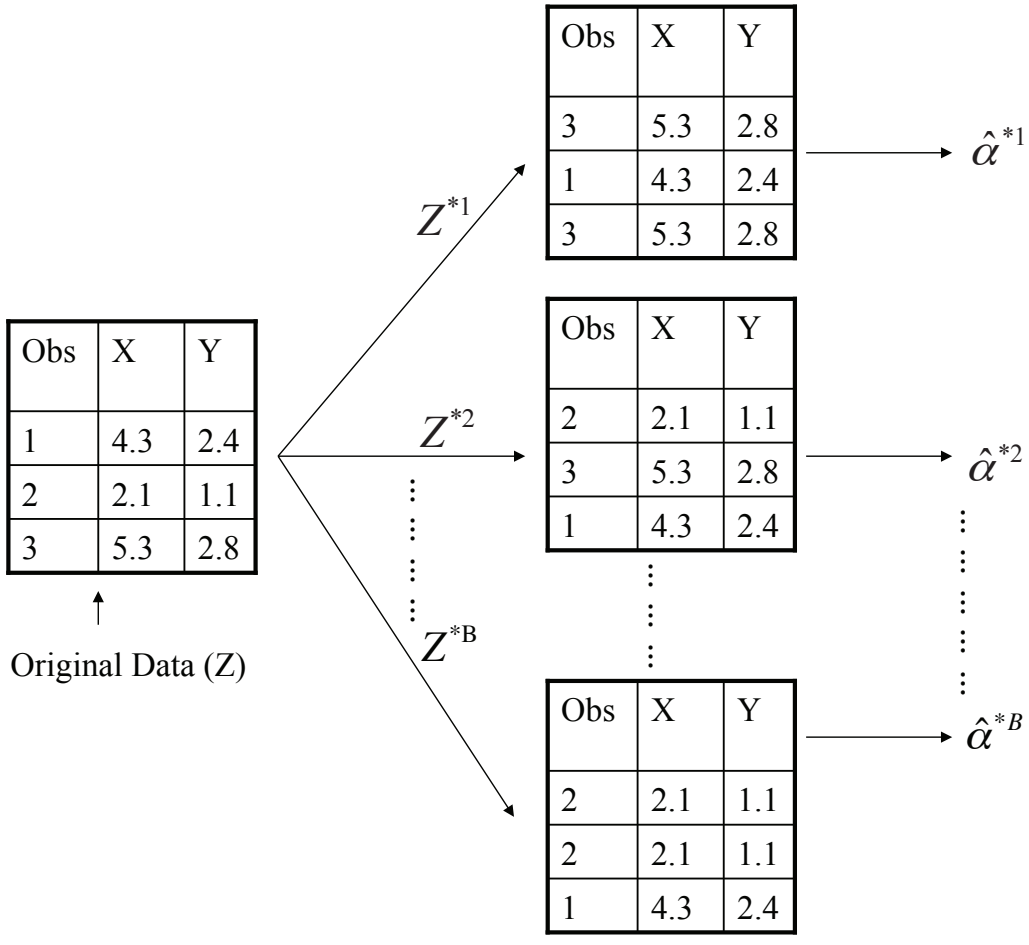

Obs X Y
1 4.3 2.4
2 2.1 1.1
3 5.3 2.8
Original Data (Z)
Z^{*1}
Z^{*2}
Z^{*B}
Obs X Y
3 5.3 2.8
1 4.3 2.4
3 5.3 2.8
Obs X Y
2 2.1 1.1
3 5.3 2.8
1 4.3 2.4
Obs X Y
2 2.1 1.1
2 2.1 1.1
1 4.3 2.4
$\hat{\alpha}^{*1}$
$\hat{\alpha}^{*2}$
$\hat{\alpha}^{*B}$

Figure 5.11. A graphical illustration of the bootstrap approach on a small sample containing $n = 3$ observations. Each bootstrap data set contains n observations, sampled with replacement from the original data set. Each bootstrap data set is used to obtain an estimate of α.

Outline

Regression
- The least squares estimation
- The statistical properties of the least squares estimates.

Classification
- Linear Discriminant Analysis
- Logistic Regression.

Model Assessment
- Cross Validation
- Bootstrap

Feature/Variable Selection
- Subset selection
- Shrinkage methods (Ridge/Lasso)

Interpretability vs. Prediction

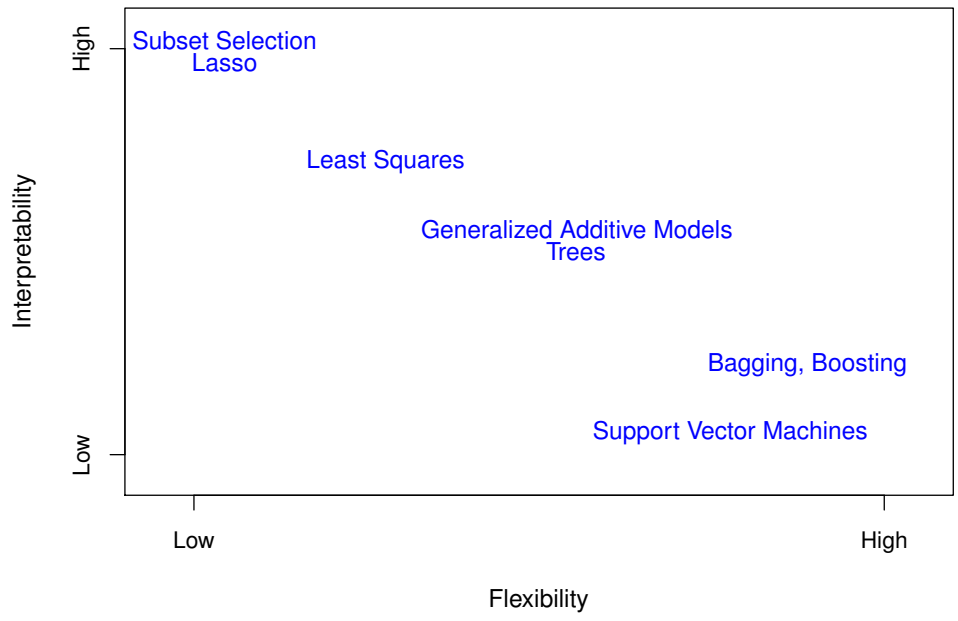


Figure: 2.7. As models become flexible, interpretability drops. **Occam Razor principle**: Everything has to be kept as simple as possible, but not simpler (Albert Einstein).

About this chapter

- Linear model already addressed in detail in Chapter 3.

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

- Model assessment: cross-validation (prediction) error in Chapter 5.
- This chapter is about model selection for linear models.
- The model selection techniques can be extended beyond linear models.
- Details about AIC, BIC, Mallow's C_p mentioned in Chapter 3.

Feature/variable selection

- Not all existing input variables are useful for predicting the output.
- Keeping redundant inputs in model can lead to poor prediction and poor interpretation.
- We consider three ways of variable/model selection:
  1. Subset selection.
  2. Shrinkage/regularization: constraining some regression parameters to 0.
  *3. Dimension reduction: (actually using the "derived inputs" by, for example, principle component approach.)

Best subset selection

- Exhaust all possible combinations of inputs.
- With p variables, there are 2^p many distinct combinations.
- Identify the best model among these models.

Pros and Cons of best subset selection

- Seems straightforward to carry out.
- Conceptually clear.
-
- The search space too large (2^p models), may lead to overfit.
- Computationally infeasible: too many models to run.
- if $p = 20$, there are $2^{20} > 1000,000$ models.

Forward stepwise selection

- Start with the null model.
- Find the best one-variable model.
- With the best one-varialbe model, add one more variable to get the best two-variable model.
- With the best two-varialbe model, add one more variable to get the best three-variable model.
- ....
- Find the best among all these best k-variable models.

Pros and Cons of forward stepwise selection

- Less computation
- Less models ($\sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ models).
- (if $p = 20$, only 211 models, compared with more than 1 million models for best subset selection).
- No problem for first n-steps if $p > n$.
- Once an input is in, it does not get out.

Backward stepwise selection

- Start with the largest model (all p inputs in).
- Find the best $(p-1)$-variable model, by reducing one from the largest model
- Find the best $(p-2)$-variable model, by reducing one variable from the best $(p-1)$-variable model.
- Find the best $(p-3)$-variable model, by reducing one variable from the best $(p-2)$-variable model.
- ....
- Find the best 1-varible model, by reducing one variable from the best 2-variable model.
- The null model.

Pros and Cons of backward stepwise selection

- Less computation
- Less models ($\sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ models).
- (if $p = 20$, only 211 models, compared with more than 1 million models for best subset selection).

- Once an input is out, it does not get in.
- No applicable to the case with $p > n$.

Find the best model based on prediction error.

- General approach by Validation/Cross-Validation (addressed in ISLR Chapter 5).
- Model-based approach by Adjusted R^2, AIC, BIC or C_p (ISLR Chapter 3).

R-squared

- Residue

$$\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij}$$

- Residual Sum of Squares

$$\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2$$

- R-squared

$$R^2 = \frac{\text{SS}_{reg}}{\text{SS}_{total}} = 1 - \frac{\text{SS}_{error}}{\text{SS}_{total}}$$

where $\text{SS}_{error} = \text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2$ and $\text{SS}_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$.

Example: Credit data

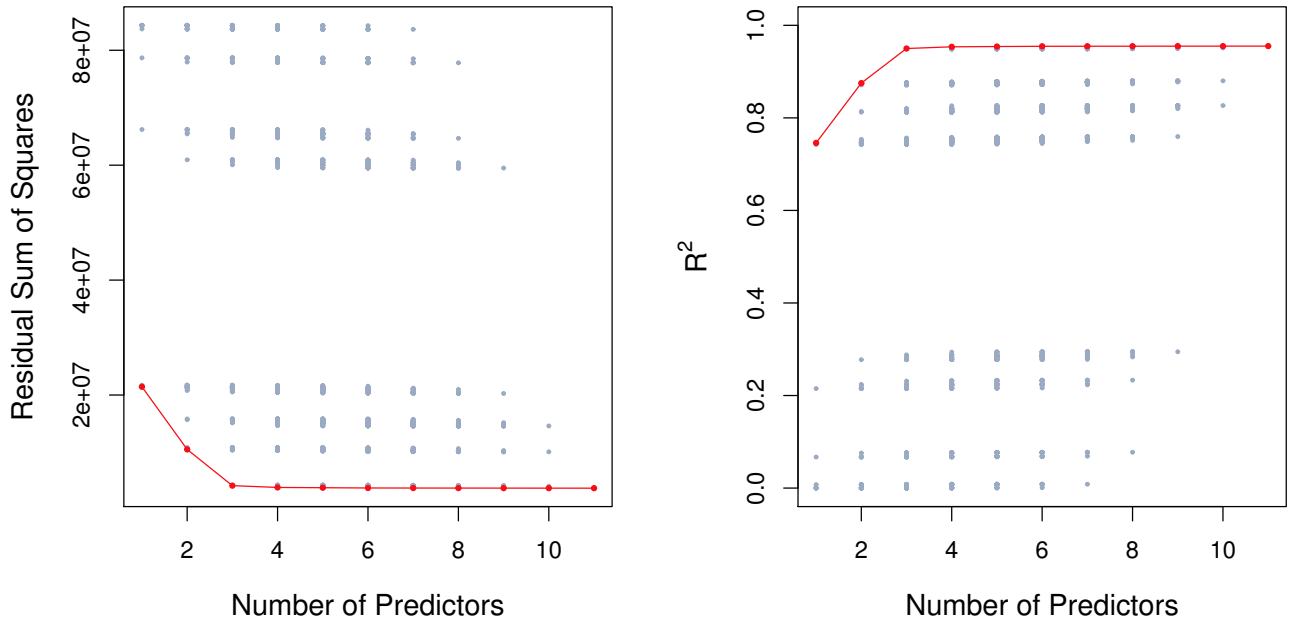


Figure: 6.1. For each possible model containing a subset of the ten predictors in the Credit data set, the RSS and R^2 are displayed. The red frontier tracks the best model for a given number of predictors, according to RSS and R^2. Though the data set contains only ten predictors, the x-axis ranges from 1 to 11, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables.

The issues of R-squared

- The R-squared is the percentage of the total variation in response due to the inputs.
- The R-squared reflects the *training error*.
- However, a model with larger R-squared is not necessarily better than another model with smaller R-squared when we consider *test error*!
- If model A has all the inputs of model B, then model A's R-squared will always be greater than or as large as that of model B.
- If model A's additional inputs are entirely uncorrelated with the response, model A contain more noise than model B. As a result, the prediction based on model A would inevitably be poorer or no better.

a) Adjusted R-squared

- The adjusted R-squared, taking into account of the degrees of freedom, is defined as

$$\begin{aligned}
\text{adjusted } R^2 &= 1 - \frac{\text{MS}_{error}}{\text{MS}_{total}} \\
&= 1 - \frac{\text{SS}_{error}/(n-p-1)}{\text{SS}_{total}/(n-1)} \\
&= 1 - \frac{s^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)}
\end{aligned}$$

  With more inputs, the R^2 always increase, but the adjusted R^2 could decrease since more inputs is penalized by the smaller degree of freedom of the residuals.
- The adjusted R-squared is preferred over the R-squared in evaluating models.

b) Mallow's C_p

Recall that our linear model (2.1) has p covariates, and $s^2 = RSS/(n-p-1)$ is the unbiased estimator of σ^2.
Assume now more covariates are available. Suppose we use only p of the K covariates with $K \geq p$.
The statistic of Mallow's C_p is defined as

$$C_p = \frac{\text{RSS}(k) + 2(k+1)s^2}{n}$$

where RSS(k) is the residual sum of squares for the linear model with k inputs.
The smaller Mallows' C_p is, the better the model is.
The following AIC is more often used, despite that Mallws' C_p and AIC usually give the same best model.

AIC

AIC stands for Akaike information criterion, which is defined as

$$\text{AIC} = \frac{\text{RSS} + 2(p+1)s^2}{ns^2},$$

for a linear model with p inputs, where $s^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2/(n-p-1)$ is the unbiased estimator of σ^2. AIC aims at maximizing the predictive likelihood. The model with the smallest AIC is preferred. The AIC criterion is try to maximize the expected predictive likelihood. In general, it can be roughly derived in the following. Let θ be a parameter of d dimension. $\hat{\theta}$ is the maximum likelihood estimator of θ based on observations $y_1, ..., y_n$. Let θ_0 be the true (unknown) value of θ, and $\mathcal{I}(\theta_0)$ be the Fisher information.

BIC

- BIC stands for Schwarz's Bayesian information criterion, which is defined as

$$\text{BIC} = \frac{\text{RSS} + (1 + p)\log(n)s^2}{ns^2},$$

  for a linear model with p inputs. Again, the model with the smallest BIC is perferred. The derivation of BIC results from Bayesian statistics and has Bayesian interpretation. It is seen that BIC is formally similar to AIC. The BIC penalizes more heavily the models with more number of inputs.

Penalized log-likelihood

- In general AIC/BIC are penalized maximum likelihood, e.g. BIC aims

$$\textbf{minimize} \; -(\text{log likelihood}) + (1+p)\log(n)/n$$

where, the first term is called deviance. In the case of linear regression with normal errors, the deviance is the same as $\log(s^2)$.

Example: credit dataset

Variables	Best subset	Forward stepwise
one	rating	rating
two	rating, income	rating, income
three	rating, income, student	rating, income, student
four	cards, income, student, limit	rating, income, student, limit

TABLE 6.1. 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.

Example

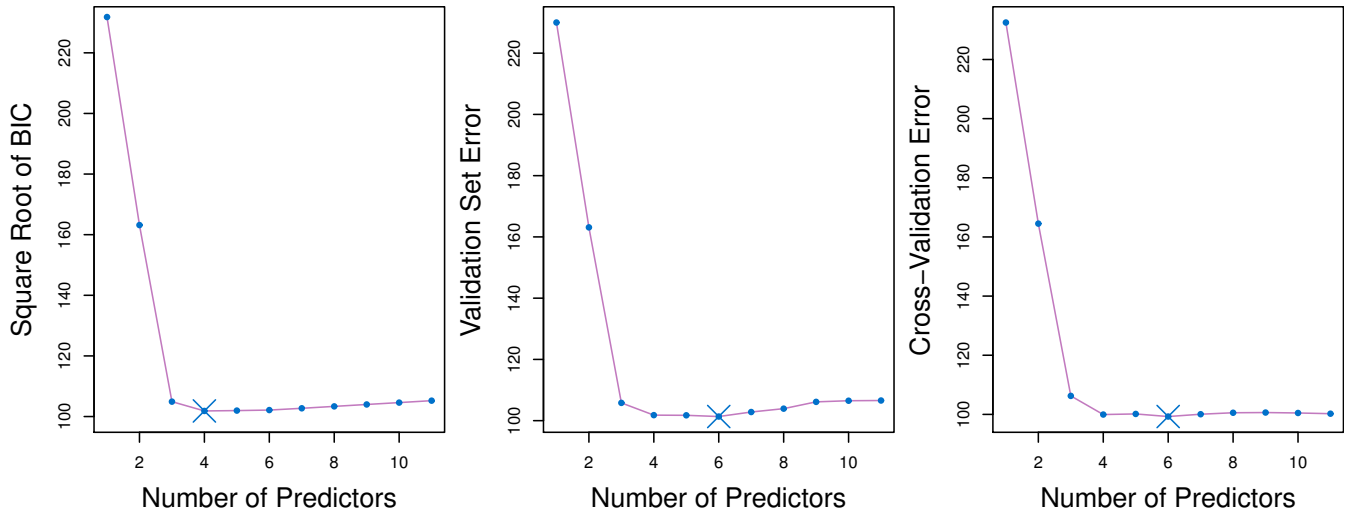


Figure: 6.3. For the Credit data set, three quantities are displayed for the best model containing d predictors, for d ranging from 1 to 11. The overall best model, based on each of these quantities, is shown as a blue cross. Left: Square root of BIC. Center: Validation set errors (75% training data). Right: 10-fold Cross-validation errors.

The one standard deviation rule

- In the above figure, model with 6 inputs do not seem to be much better than model with 4 or 3 inputs.
- Keep in mind the Occam's razor: Choose the simplest model if they are similar by other criterion.

The one standard deviation rule

- Calculate the standard error of the estimated test MSE for each model size,
- Consider the models with estimated test MSE of one standard deviation within the smallest test MSE.
- Among them select the one with the smallest model size.
- (Apply this rule to the Example in Figure 6.3 gives the model with 3 variable.)

Ridge Regression

- The least squares estimator $\hat{\beta}$ is minimizing

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

- The ridge regression $\hat{\beta}_\lambda^R$ is minimizing

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

  where $\lambda \geq 0$ is a tuning parameter.
- The first term measuares goodness of fit, the smaller the better.
- The second term $\lambda\sum_{j=1}^{p}\beta_j^2$ is called shrikage penalty, which *shrinks* β_j towards 0.
- The shrinkage reduces variance (at the cost increased bias)!

Tuning parameter λ.

- $\lambda = 0$: no penalty, $\hat{\beta}_0^R = \hat{\beta}^{LS}$.
- $\lambda = \infty$: infinity penalty, $\hat{\beta}_0^R = 0$.
- Large λ: heavy penalty, more shrinkage of the estimator.
- Note that β_0 is not penalized.

Remark.

- If $p > n$, ridge regression can still perform well by trading off a small increase in bias for a large decrease in variance.
- Ridge regression works best in situations where the least squares estimates have high variance.
- Ridge regression also has substantial computational advantages
- $$\hat{\beta}_{\lambda}^{R} = (\mathbf{X}^{T}\mathbf{X}) + \lambda I)^{-1}\mathbf{X}^{T}\mathbf{y}$$

  where I is $p+1$ by $p+1$ diagonal with diagonal elements $(0, 1, 1, ..., 1)$.

Example: Ridge Regularization Path in Credit data

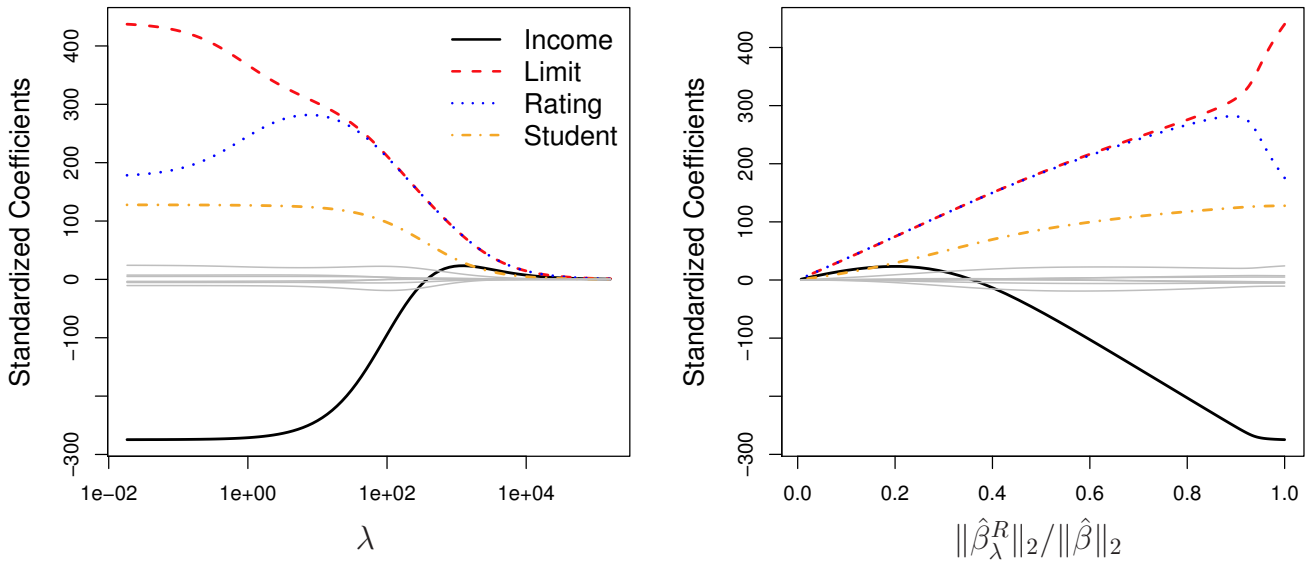


Figure: 6.4. The standardized ridge regression coefficients are displayed for the Credit data set, as a function of λ and $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. Here $\|\mathbf{a}\|_2 = \sqrt{\sum_{j=1}^{p} a_j^2}$.

The Lasso

- Lasso stands for Least Absolute Shrinkage and Selection Operator.
- The Lasso estimator $\hat{\beta}_\lambda^L$ is the minimizer of

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

- We may use $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$, which is the l_1 norm.
- LASSO often shrinks coefficients to be identically 0. (This is not the case for ridge)
- Hence it performs variable selection, and yields sparse models.

Example: Lasso Path in Credit data.

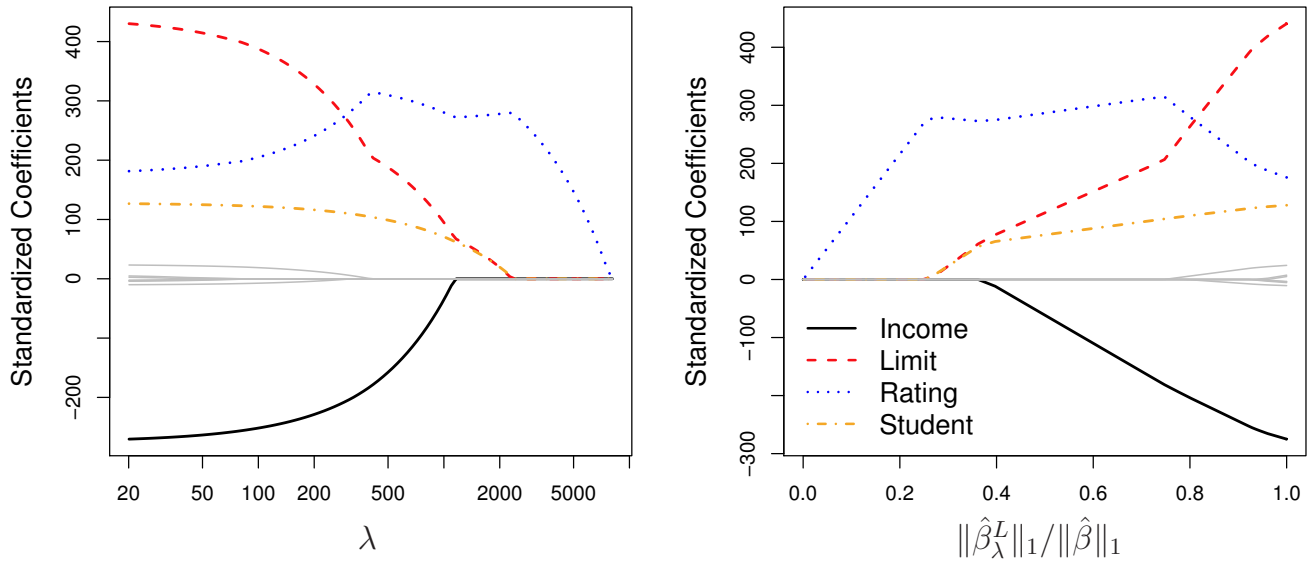


Figure: 6.6. The standardized lasso coefficients on the Credit data set are shown as a function of λ and $\|\hat{\beta}_\lambda^L\|_1/\|\hat{\beta}\|_1$.