Linear Model Selection and Regularization

1. Subset selection
2. Shrinkage methods
3. Dimension reduction methods (using derived inputs)
4. High dimension data analysis

Feature/variable selection

- NOT all existing input variables are useful for predicting output.

Feature/variable selection

- NOT all existing input variables are useful for predicting output.
- Keeping *redundant* inputs in model can lead to
  - poor prediction,
  - poor interpretation.

Feature/variable selection

- NOT all existing input variables are useful for predicting output.
- Keeping *redundant* inputs in model can lead to
  - poor prediction,
  - poor interpretation.
- We consider three ways of variable/model selection:

Feature/variable selection

- NOT all existing input variables are useful for predicting output.
- Keeping *redundant* inputs in model can lead to
  - poor prediction,
  - poor interpretation.
- We consider three ways of variable/model selection:
  1. **Subset selection**.

Feature/variable selection

- NOT all existing input variables are useful for predicting output.
- Keeping *redundant* inputs in model can lead to
  - poor prediction,
  - poor interpretation.
- We consider three ways of variable/model selection:
  1. **Subset selection**.
  2. **Shrinkage/regularization**.

Feature/variable selection

- NOT all existing input variables are useful for predicting output.
- Keeping *redundant* inputs in model can lead to
  - poor prediction,
  - poor interpretation.
- We consider three ways of variable/model selection:
  1. **Subset selection**.
  2. **Shrinkage/regularization**.
  3. **Dimension reduction**.

Best subset selection

- Exhaust all possible combinations of inputs.

Best subset selection

- Exhaust all possible combinations of inputs.
- With p variables, there are 2^p many distinct combinations.

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.

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.

QUESTION:

- What is the best?
- Which criterion do we use?

The algorithm of best subset selection

- Step 1. Let $\mathcal{M}_0$ be the *null model*, $Y = \beta_0 + \epsilon$. Which contains no predictors.
- Step 2. For $k = 1, 2, ..., p$,
  - Fit all $\binom{p}{k} = p!/(k!(n-k)!)$ models that contain exactly k predictors.
  - Pick the best model, that with largest R^2, among them and call it $\mathcal{M}_k$.
- Step 3. Select a single best model from $\mathcal{M}_0, ..., \mathcal{M}_p$ by cross validation or AIC or BIC or C_p or adjusted R^2.

Recall: Definitions

- 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

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}},$$

where $\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$.

- the model containing all of the predictors will always have the smallest RSS and the largest R^2, since these quantities are related to the training error.

- The model containing all of the predictors will always have the *smallest* RSS and the *largest* R^2, since these quantities are related to the training error.

- The model containing all of the predictors will always have the *smallest* RSS and the *largest* R^2, since these quantities are related to the training error.
- RSS and R^2 are NOT suitable for selecting the best model.

- The model containing all of the predictors will always have the *smallest* RSS and the *largest* R^2, since these quantities are related to the training error.
- RSS and R^2 are NOT suitable for selecting the best model.
- In order to select the best model with respect to test error, there are two common approaches:
  1. We can *indirectly* estimate test error
     - make an adjustment to the training error to account for the bias due to overfitting.
     - C_p, AIC, BIC, and Adjusted R^2.
  2. We can *directly* estimate the test error,
     - using either a validation set approach or a cross-validation approach.
     - Cross-validation.

a). Adjusted R-squared.

R^2 may not be a good criterion when we consider *test error*!

- The R-squared reflects the *training error*.

a). Adjusted R-squared.

R^2 may not be a good criterion when we consider *test error*!

- The R-squared reflects the *training error*.
- The *adjusted* R-squared, taking into account of the *degrees of freedom*, is defined as

$$\begin{aligned} \text{adjusted } R^2 &= 1 - \frac{\text{MS}_{error}}{\text{MS}_{total}} \\ &= 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}. \end{aligned}$$

a). Adjusted R-squared.

R^2 may not be a good criterion when we consider *test error*!

- The R-squared reflects the *training error*.
- The *adjusted* R-squared, taking into account of the *degrees of freedom*, is defined as

$$\begin{aligned} \text{adjusted } R^2 &= 1 - \frac{\text{MS}_{error}}{\text{MS}_{total}} \\ &= 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}. \end{aligned}$$

- With more inputs, the R^2 always not decreasing, but the adjusted R^2 may decrease since more irrelevant inputs are penalized by the smaller degree of freedom of the residuals.
- The adjusted R-squared is preferred over the R-squared in evaluating models.

b). Mallows' C_p.

Recall that our linear model with p covariates,

b). Mallows' C_p.

Recall that our linear model with p covariates,

- $s_p^2 = \hat{\sigma}^2 = \text{RSS}/(n-p-1)$ is the unbiased estimator of σ^2.

b). Mallows' C_p.

Recall that our linear model with p covariates,

- $s_p^2 = \hat{\sigma}^2 = \text{RSS}/(n-p-1)$ is the unbiased estimator of σ^2.
- Suppose we use only d of the p covariates with $d \leq p$, for example, $x_{j_1}, \cdots, x_{j_d}$.
- The statistic of Mallow's C_p is defined as (Page 221)

$$C_p = \frac{\text{RSS}(d)}{s_p^2} + 2d - n, \quad \text{or} \quad \frac{1}{n}\left(\text{RSS}(d) + 2ds_p^2\right).$$

  - $\text{RSS}(d) = \sum_{i=1}^{n}(\beta_0 + \beta_{j_1} x_{ij_1} + \cdots \beta_{j_d} z_{ij_d} - y_i)^2$ is the residual sum of squares for the linear model with d inputs.

b). Mallows' C_p.

Recall that our linear model with p covariates,

- $s_p^2 = \hat{\sigma}^2 = \text{RSS}/(n-p-1)$ is the unbiased estimator of σ^2.
- Suppose we use only d of the p covariates with $d \leq p$, for example, $x_{j_1}, \cdots, x_{j_d}$.
- The statistic of Mallow's C_p is defined as (Page 221)

$$C_p = \frac{\text{RSS}(d)}{s_p^2} + 2d - n, \quad \text{or} \quad \frac{1}{n}\left(\text{RSS}(d) + 2ds_p^2\right).$$

  - $\text{RSS}(d) = \sum_{i=1}^{n}(\beta_0 + \beta_{j_1}x_{ij_1} + \cdots \beta_{j_d}z_{ij_d} - y_i)^2$ is the residual sum of squares for the linear model with d inputs.
  - Mallows' C_p: the smaller it is, the better the model is.

c). AIC.

- AIC stands for *Akaike information criterion*,
- aims at maximizing the predictive likelihood.

$$\text{AIC} = \frac{1}{ns_p^2}\left(\text{RSS}(d) + 2ds_p^2\right),$$

  when Gaussian likelihood is assumed in least square regression.
- The model with the smallest AIC is preferred.

d). BIC.

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

$$\text{BIC} = \frac{1}{ns_p^2}\left(\text{RSS}(d) + ds_p^2(\log n)\right),$$

for a linear model with p inputs.

- the model with the smallest BIC is perfered.
- The derivation of BIC results from Bayesian statistics and has Bayesian interpretation.
- It replaces $2ds_p^2$ in AIC by $(\log n)ds_p^2$, so for $\log n > 2$ or $n > 7$, BIC penalizes more heavily the models with more number of inputs.

Example

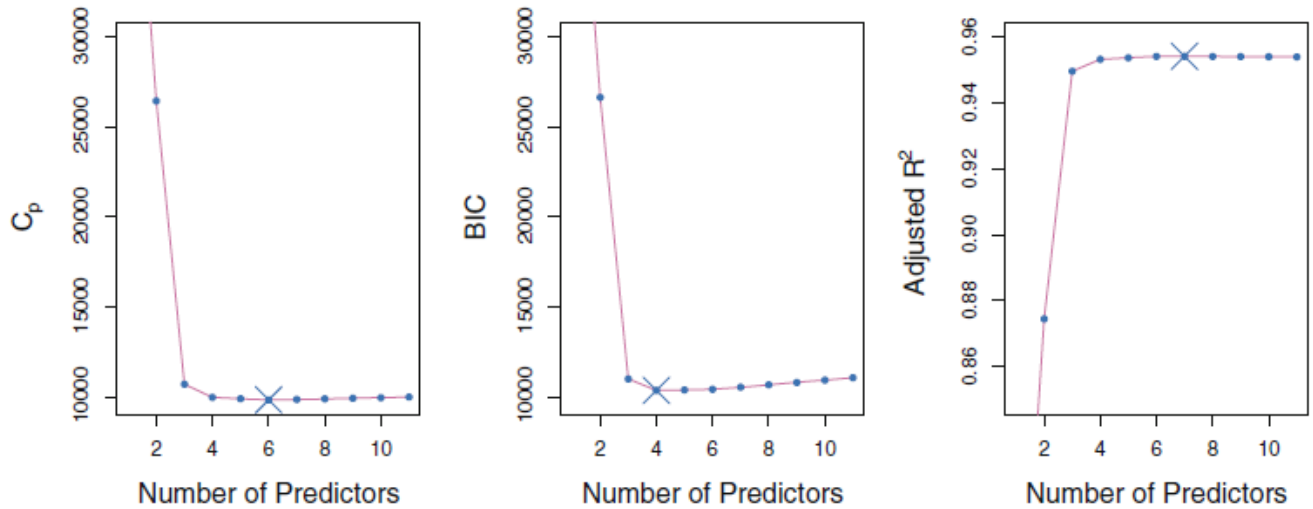


Figure: 6.2. C_p, BIC, and adjusted R^2 are shown for the best models of each size for the Credit data set (the lower frontier in Figure 6.1).

- C_p and BIC are estimates of test MSE.
- In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected.
- The other two plots are rather flat after four variables are included.

Pros and Cons of best subset selection

Advantages:

- Seems straightforward to carry out.
- Conceptually clear.

Pros and Cons of best subset selection

Advantages:

- Seems straightforward to carry out.
- Conceptually clear.

Disadvantages:

- The search space too large (2^p models).
- if $p = 20$, there are $2^{20} > 1000,000$ models.
- Computationally infeasible: too many models to run.

Forward stepwise selection

- Start with the *null* model.

Forward stepwise selection

- Start with the *null* model.
- Find the *best one-variable* model.

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.

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.

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.

The algorithm of forward stepwise selection

- Step 1. Let $\mathcal{M}_0$ be the null model, $Y = \beta_0 + \epsilon$, which contains no predictors.

The algorithm of forward stepwise selection

- Step 1. Let $\mathcal{M}_0$ be the null model, $Y = \beta_0 + \epsilon$, which contains no predictors.
- Step 2. For $k = 0, 1, ..., p-1$,
  - Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
  - 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.

The algorithm of forward stepwise selection

- Step 1. Let $\mathcal{M}_0$ be the null model, $Y = \beta_0 + \epsilon$, which contains no predictors.
- Step 2. For $k = 0, 1, ..., p-1$,
  - Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
  - 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.
- step 3. Select a single best model from $\mathcal{M}_0, ..., \mathcal{M}_p$ by cross validation or AIC or BIC or C_p or adjusted R^2.

The algorithm of forward stepwise selection

- Step 1. Let $\mathcal{M}_0$ be the null model, $Y = \beta_0 + \epsilon$, which contains no predictors.
- Step 2. For $k = 0, 1, ..., p-1$,
  - Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
  - 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.
- step 3. Select a single best model from $\mathcal{M}_0, ..., \mathcal{M}_p$ by cross validation or AIC or BIC or C_p or adjusted R^2.

Attention: You should use a suitable criterion for your problem!!!!!

Pros and Cons of forward stepwise selection

Pros and cons:

- Less computation.

Pros and Cons of forward stepwise selection

Pros and cons:

- 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.

Pros and Cons of forward stepwise selection

Pros and cons:

- 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 in, it does not get out.

Pros and Cons of forward stepwise selection

Pros and cons:

- 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 in, it does not get out.
- Forward stepwise tends to do well in practice, it is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.

Example: credit dataset

Variables	Best subset	Forward stepwise
1	rating	rating
2	rating, income	rating, income
3	rating, income, student	rating, income, student
4	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,
- the fourth models differ.

Backward stepwise selection

- Start with the largest model (all p inputs in).

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.

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.
- For $k = p, p-1, \cdots, 1$:
  - (a) Consider all k models that contain all but one of the predictors in Mk, for a total of $k-1$ predictors.
  - (b) 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.
- Select a single best model from among $\mathcal{M}_0, \cdots, \mathcal{M}_p$ using crossvalidated prediction error, C_p (AIC), BIC, or adjusted R_2.

Algorithm of backward stepwise selection

- Step 1. Let $\mathcal{M}_p$ be the full model.

Algorithm of backward stepwise selection

- Step 1. Let $\mathcal{M}_p$ be the full model.
- Step 2. For $k = p, p-1, ..., 1$,
  - Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$ for a total of $k-1$ predictors
  - 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.

Algorithm of backward stepwise selection

- Step 1. Let $\mathcal{M}_p$ be the full model.
- Step 2. For $k = p, p-1, ..., 1$,
  - Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$ for a total of $k-1$ predictors
  - 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.
- Step 3. Select a single best model from $\mathcal{M}_0, ..., \mathcal{M}_p$ by cross validation or AIC or BIC or C_p or adjusted R^2.

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.
- it is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.

Example

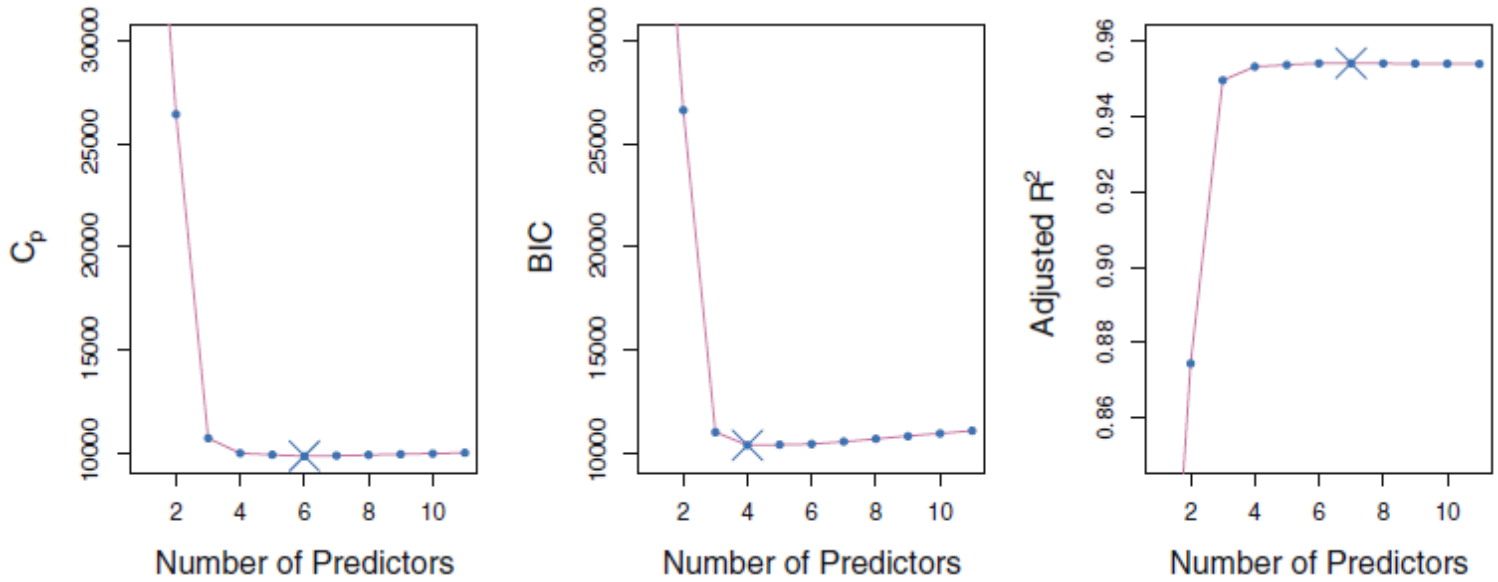


Figure: 6.2. C_p, BIC, and adjusted R^2 are shown for the best models of each size for the Credit data set (the lower frontier in Figure 6.1). C_p and BIC are estimates of test MSE. In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected. The other two plots are rather flat after four variables are included.

Example

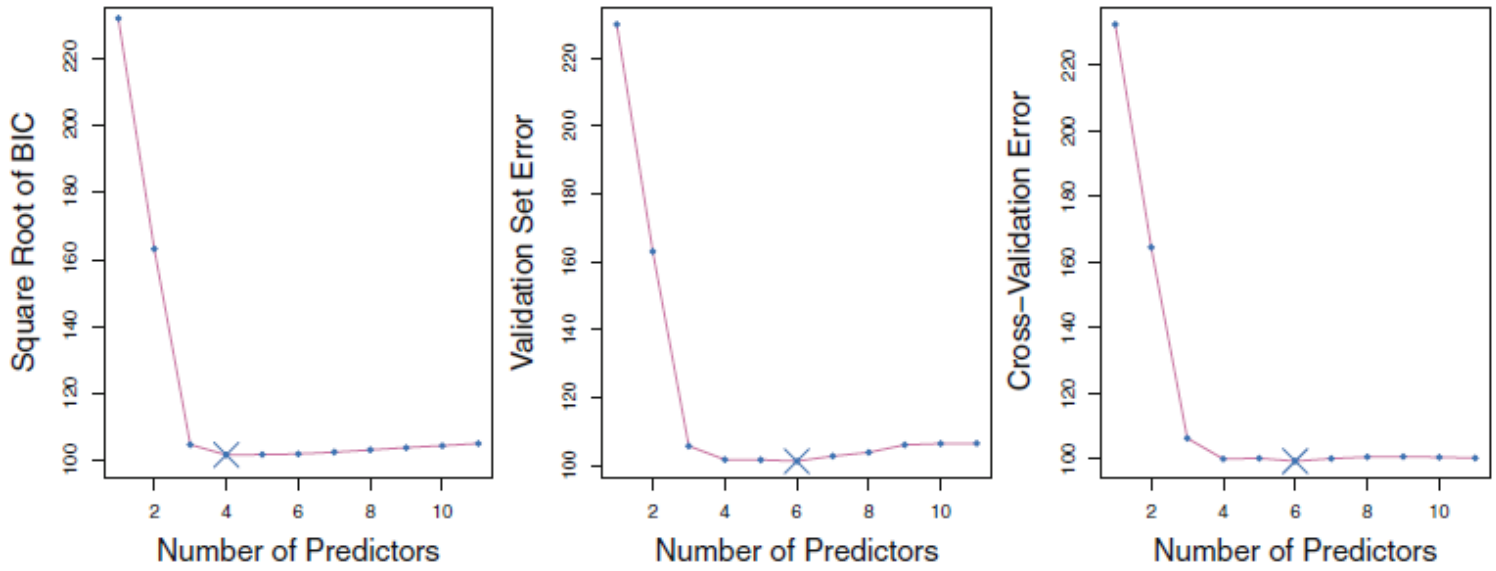


Figure: 6.3. For the Credit data set, three quantities are displayed for the best model containing p predictors, for p ranging from 1 to 11.

- *Left*: Square root of BIC.
- *Center*: Validation set errors (75% training data).
- *Right*: 10-fold Cross-validation errors.
- The overall best model, based on each of these quantities, is shown as a blue cross.

The one standard deviation rule

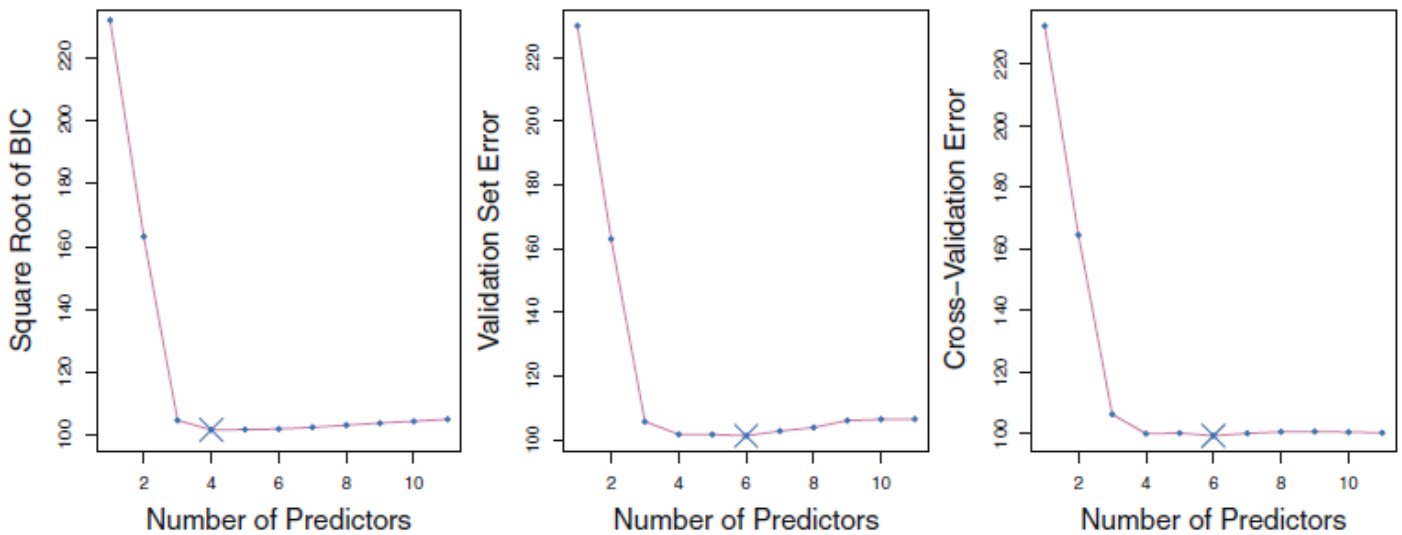
The Occam's razor:

- Choose the simplest model if they are similar by other criterion.
- Among them select the one with the smallest model size.

The one standard deviation rule

The Occam's razor:

- Choose the simplest model if they are similar by other criterion.
- Among them select the one with the smallest model size.



- Model with 6 inputs do not seem to be much better than model with 4 or 3 inputs.

Shrinkage methods

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

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

Shrinkage methods

- 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 Shrinkage method is minimizing

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\mathcal{R}(\beta_1, \cdots, \beta_p).$$

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

Shrinkage methods

- 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 Shrinkage method is minimizing

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\mathcal{R}(\beta_1, \cdots, \beta_p).$$

  where $\lambda \geq 0$ is a tuning parameter.
- The first term measures goodness of fit, the smaller the better.

Shrinkage methods

- 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 Shrinkage method is minimizing

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\mathcal{R}(\beta_1, \cdots, \beta_p).$$

  where $\lambda \geq 0$ is a tuning parameter.
- The first term measures goodness of fit, the smaller the better.
- The second term $\lambda\mathcal{R}(\beta_1, \cdots, \beta_p)$ is called *shrikage penalty*, which shrinks $\beta_j, j = 1, \cdots, p$ towards 0.
- Note that β_0 is NOT penalized.

Ridge

- Ridge: $\mathcal{R}(\beta_1, \cdots, \beta_p) = \sum_{j=1}^{p} \beta_j^2$.

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

- The solution is

$$\hat{\beta}_\lambda^R = (\mathbf{X}^T \mathbf{X} + \lambda I)^{-1} \mathbf{X}^T \mathbf{y},$$

where $I = \text{diag}([0, 1, 1, ..., 1])$ is $(p+1)$-by-$(p+1)$ diagnal matrix.

The lasso

- The lasso $\mathcal{R}(\beta_1, \cdots, \beta_p) = \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$, which is the l_1 norm.
- Lasso stands for *Least Absolute Shrinkage and Selection Operator*.
- The Lasso estimator $\hat{\beta}_\lambda^L$ is the minimizer of

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

which is equivalent to

$$\min_{\beta_0, \cdots, \beta_p, \alpha_1, \cdots, \alpha_p} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\alpha_j| + \frac{\rho}{2} \sum_{i=1}^{p} (\alpha_i - \beta_i)^2.$$

We can solve it by alternating minimization.

- Solve $\beta_0, \cdots, \beta_p$ with $\alpha_1, \cdots, \alpha_p$ fixed:

$$\min_{\beta_0,\cdots,\beta_p} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \frac{\rho}{2}\sum_{i=1}^{p}(\alpha_i - \beta_i)^2,$$

- Solve $\alpha_1, \cdots, \alpha_p$ with $\beta_0, \cdots, \beta_p$ fixed:

$$\min_{\alpha_1,\cdots,\alpha_p} \lambda\sum_{j=1}^{p}|\alpha_j| + \frac{\rho}{2}\sum_{i=1}^{p}(\alpha_i - \beta_i)^2.$$

- For the first one, we can solve by least squares.
- For the second one, we can use soft thresholding.

$$\alpha_i = \begin{cases} \beta_i - \lambda/\rho, & \beta_i \geq \lambda/\rho, \\ 0, & |\beta_i| < \lambda/\rho, \\ \beta_i + \lambda/\rho, & \beta_i \leq -\lambda/\rho, \end{cases}$$

- Solve $\beta_0, \cdots, \beta_p$ with $\alpha_1, \cdots, \alpha_p$ fixed:

$$\min_{\beta_0,\cdots,\beta_p} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \frac{\rho}{2}\sum_{i=1}^{p}(\alpha_i - \beta_i)^2,$$

- Solve $\alpha_1, \cdots, \alpha_p$ with $\beta_0, \cdots, \beta_p$ fixed:

$$\min_{\alpha_1,\cdots,\alpha_p} \lambda\sum_{j=1}^{p}|\alpha_j| + \frac{\rho}{2}\sum_{i=1}^{p}(\alpha_i - \beta_i)^2.$$

- For the first one, we can solve by least squares.
- For the second one, we can use soft thresholding.

$$\alpha_i = \begin{cases} \beta_i - \lambda/\rho, & \beta_i \geq \lambda/\rho, \\ 0, & |\beta_i| < \lambda/\rho, \\ \beta_i + \lambda/\rho, & \beta_i \leq -\lambda/\rho, \end{cases}$$

- Ridge has closed form solution. Lasso generally does not have closed form solution.

Standardize the inputs.

- Least squares is unaffected by the scale of X_j,

$$X_j\hat{\beta}_j = (cX_j)(\hat{\beta}_j/c).$$

- Shrinkage method is affected by λ as well as the scale of the inputs.

Standardize the inputs.

- Least squares is unaffected by the scale of X_j,

$$X_j\hat{\beta}_j = (cX_j)(\hat{\beta}_j/c).$$

- Shrinkage method is affected by λ as well as the scale of the inputs.
- Suggest to apply standardization before trying regression with penalties.
- For j-th input X_j with observations: $(x_{1j}, ..., x_{nj})$, standardize it as

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

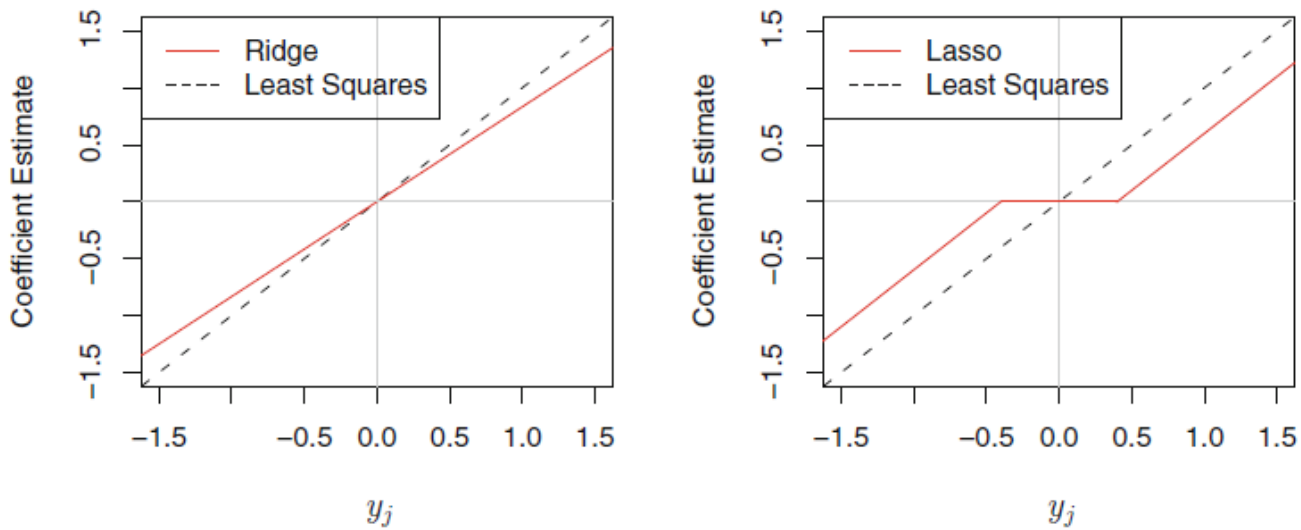
to get rid of the scale of X_j.

Example

- Consider the simple model $y_i = \beta_i + \epsilon_i$, $i = 1, ..., n$ and $n = p$.
  Then,
  The least squares $\hat{\beta}_j = y_j$; the ridge $\hat{\beta}_j^R = y_j/(1 + \lambda)$
  The Lasso $\hat{\beta}_j^L = \text{sign}(y_j)(|y_j| - \lambda/2)_+$.

Example

- Consider the simple model $y_i = \beta_i + \epsilon_i$, $i = 1, ..., n$ and $n = p$. Then,
  The least squares $\hat{\beta}_j = y_j$; the ridge $\hat{\beta}_j^R = y_j/(1+\lambda)$
  The Lasso $\hat{\beta}_j^L = \text{sign}(y_j)(|y_j| - \lambda/2)_+$.



- *Left*: The ridge regression coefficient estimates are shrunken proportionally towards zero, relative to the least squares estimates.
- *Right*: Lasso coefficient estimates are soft-thresholded towards 0.

Another formulation

- For Lasso: Minimize

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

- For Ridge: Minimize

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

- For l_0: Minimize

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

l_0 method penalizes number of non-zero coefficients. A difficult (NP-hard) problem for optimization.

Variable selection property for Lasso

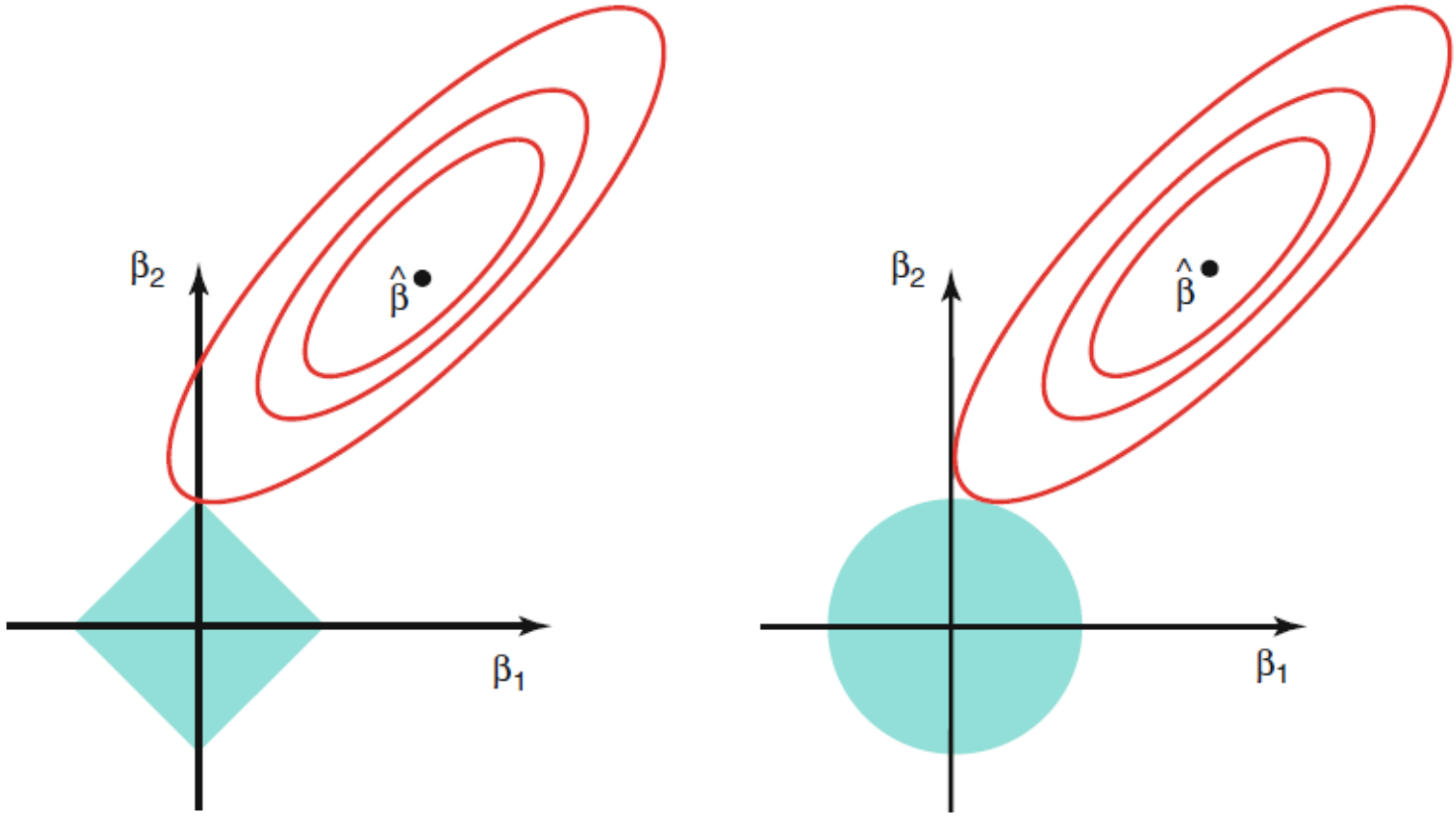


Figure: 6.7. Contours of the error and constraint functions for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$, while the red ellipses are the contours of the RSS.

- 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.
- This is not the case for ridge.
- It performs variable selection, and yields sparse models.

- 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.
- This is not the case for ridge.
- It performs variable selection, and yields sparse models.

Why we set $\mathcal{R}(\beta_1, \cdots, \beta_p)$, which is not related to β_0?

Tuning parameter λ.

- $\lambda = 0$: no penalty, $\hat{\beta}_0^R = \hat{\beta}_0^L = \hat{\beta}$.
- $\lambda = \infty$: infinity penalty, $\hat{\beta}_\infty^R = \hat{\beta}_\infty^L = 0$.
- Large λ: heavy penalty, more shrinkage of the estimator.

Example: ridge.

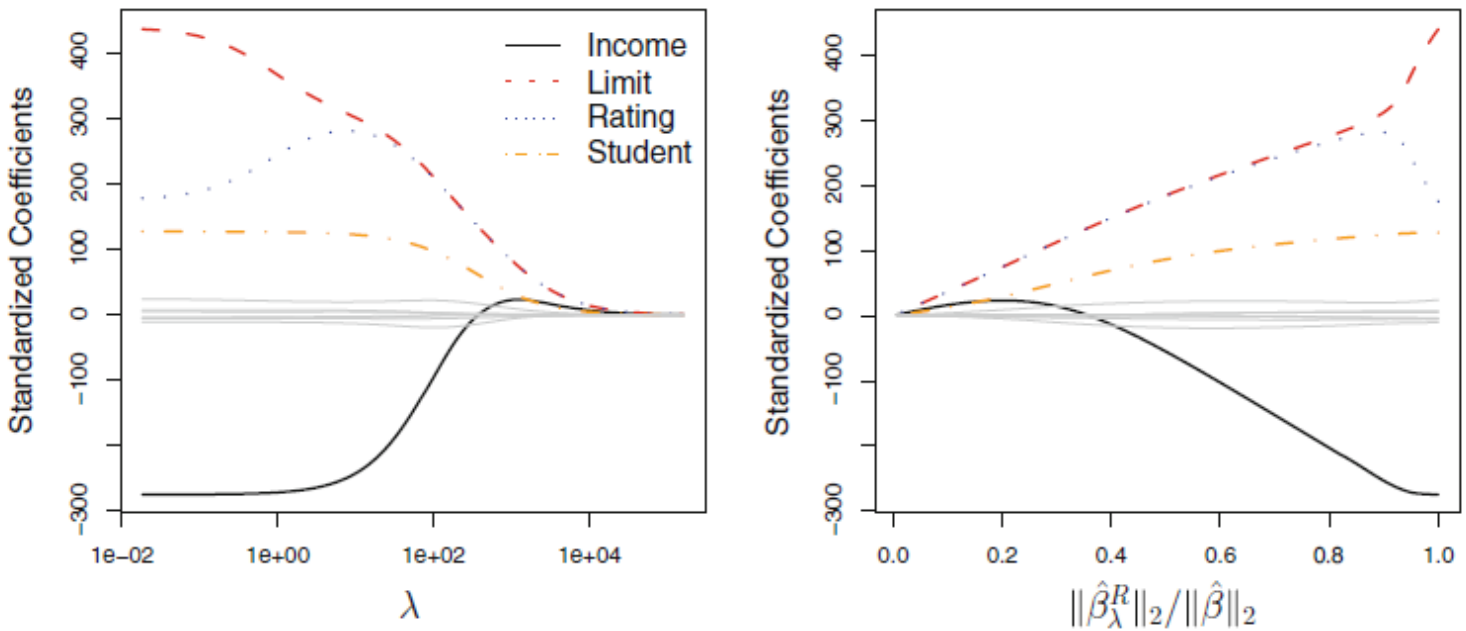


Figure: 6.4. The standardized ridge regression coefficients are displayed for the Credit data set, as a function of λ and $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. Here $\|\mathbf{a}\|_2 = \sqrt{\sum_{j=1}^{p} a_j^2}$.

Example: lasso.

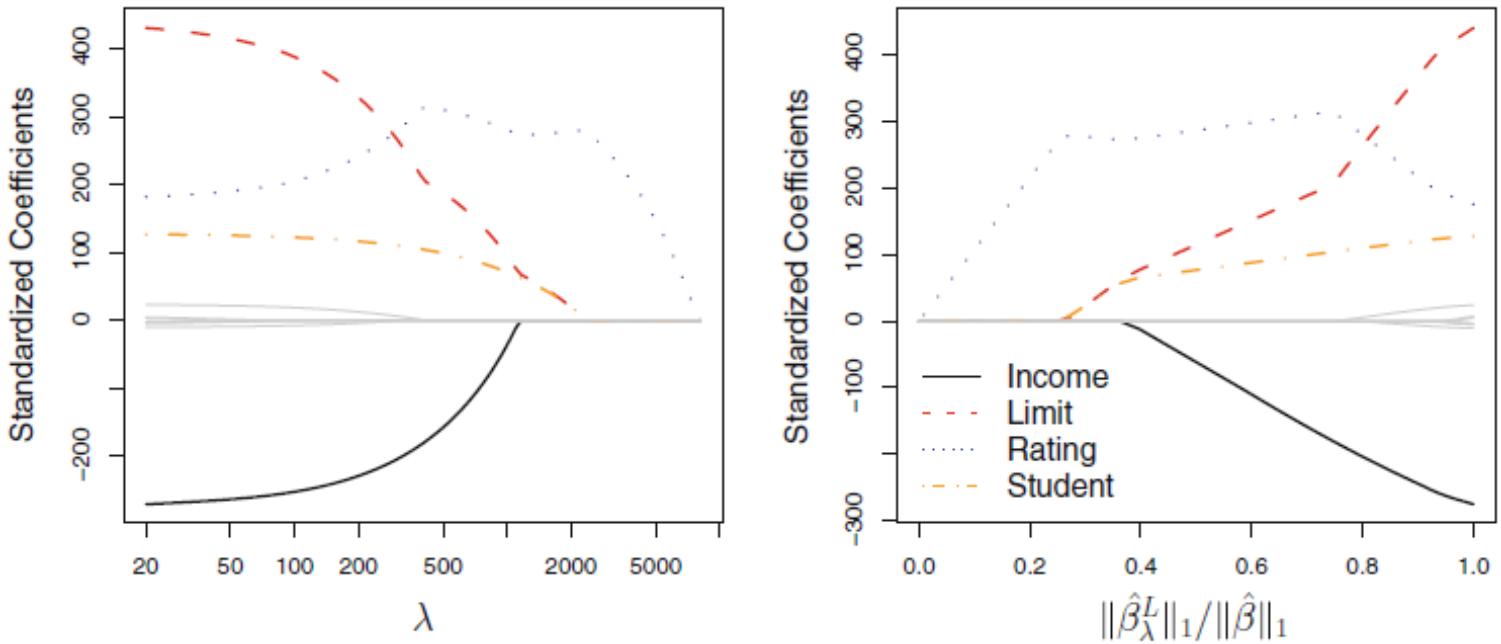


Figure: 6.6. The standardized lasso coefficients on the Credit data set are shown as a function of λ and $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

Bias-variance tradeoff (why ridge improves over LSE)

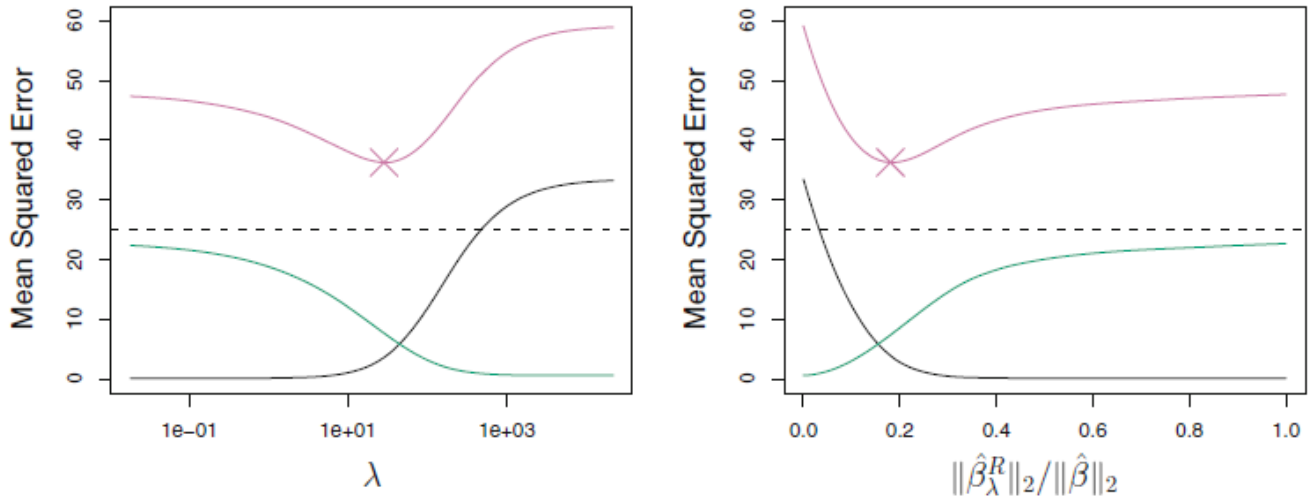


Figure: Simulated data ($p = 45, n = 50$).

- Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions.
- The horizontal dashed lines indicate the minimum possible MSE.
- Purple crosses – ridge regression models with smallest MSE.

Bayesian interpretation

- Suppose $\beta = (\beta_0, ..., \beta_p)$ are random variables with a prior distribution $p(\cdot)$.
- Given β and the input X, Y has conditional density $f(y|X, \beta)$.
- The posterior distribution of the parameter β is

$$p(\beta|X, Y) \propto f(Y|X, \beta)p(\beta|X) = f(Y|X, \beta)p(\beta)$$

The proporionality means a constant (not related with β) multiplier. (β and X are independent.)

Bayesian interpretation

- Now consider the linear regression model, $Y = \beta_0 + X_1\beta_1 + ... + X_p\beta_p + \epsilon$, with ϵ conditioning on X follows $N(0, \sigma^2)$.
- If the β has the normal prior, the prior of β following normal distribution with mean 0 then the posterior mode for β is ridge estimator.
- If the β has the double exponential prior:

$$f(t) = \lambda e^{-\lambda|t|}/2$$

the prior of β following a double-exponential (Laplace) distribution with mean zero, then it follows that the posterior mode for β is the lasso solution.

The Gaussian and double expoential curves

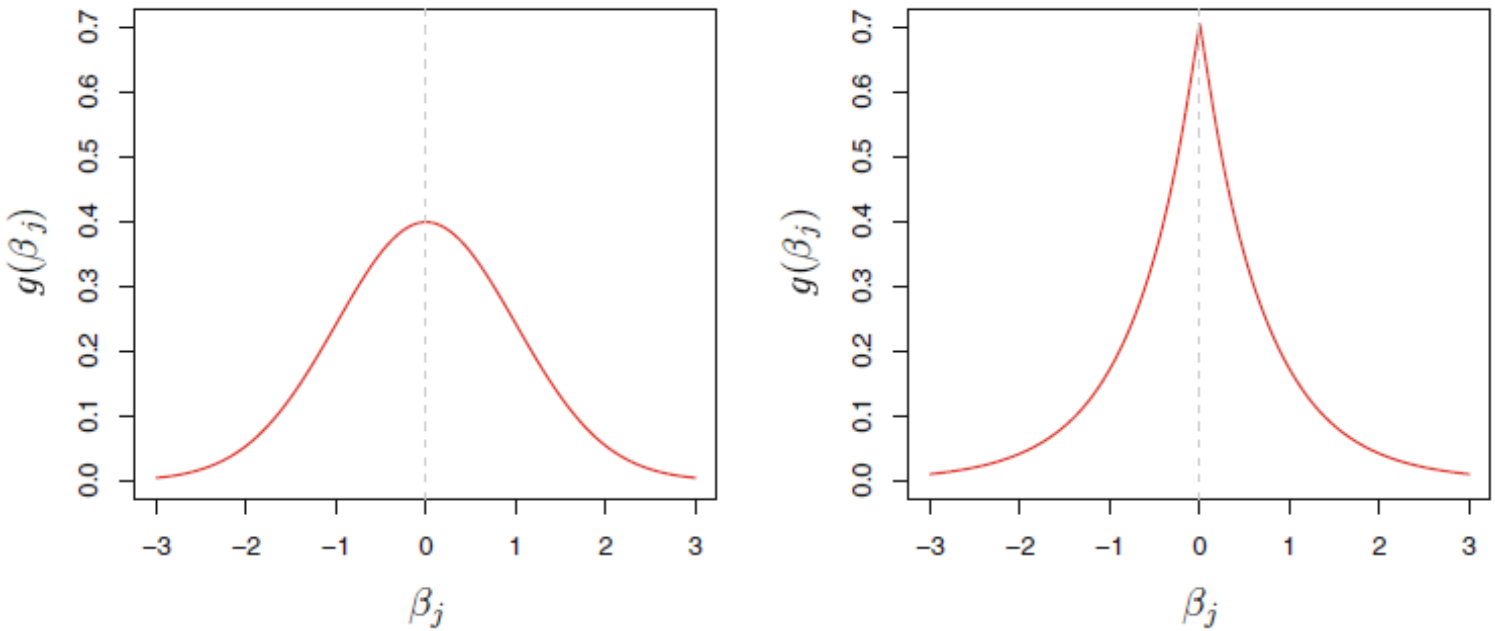


Figure: 6.11. Left: Ridge regression is the posterior mode for β under a Gaussian prior. Right: The lasso is the posterior mode for β under a double-exponential prior.

Tuning parameter selection by cross-validation: Credit data

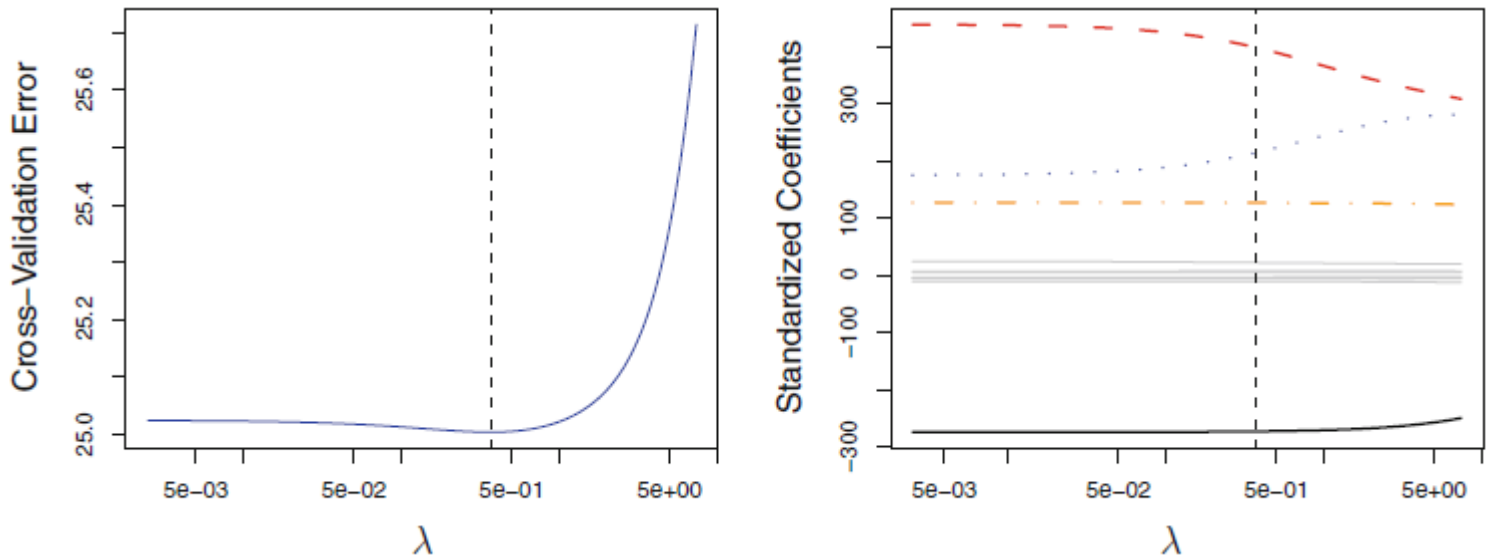


Figure: 6.12. Left: Cross-validation errors that result from applying ridge regression to the Credit data set with various value of λ. Right: The coefficient estimates as a function of λ. The vertical dashed lines indicate the value of λ selected by cross-validation.

Example

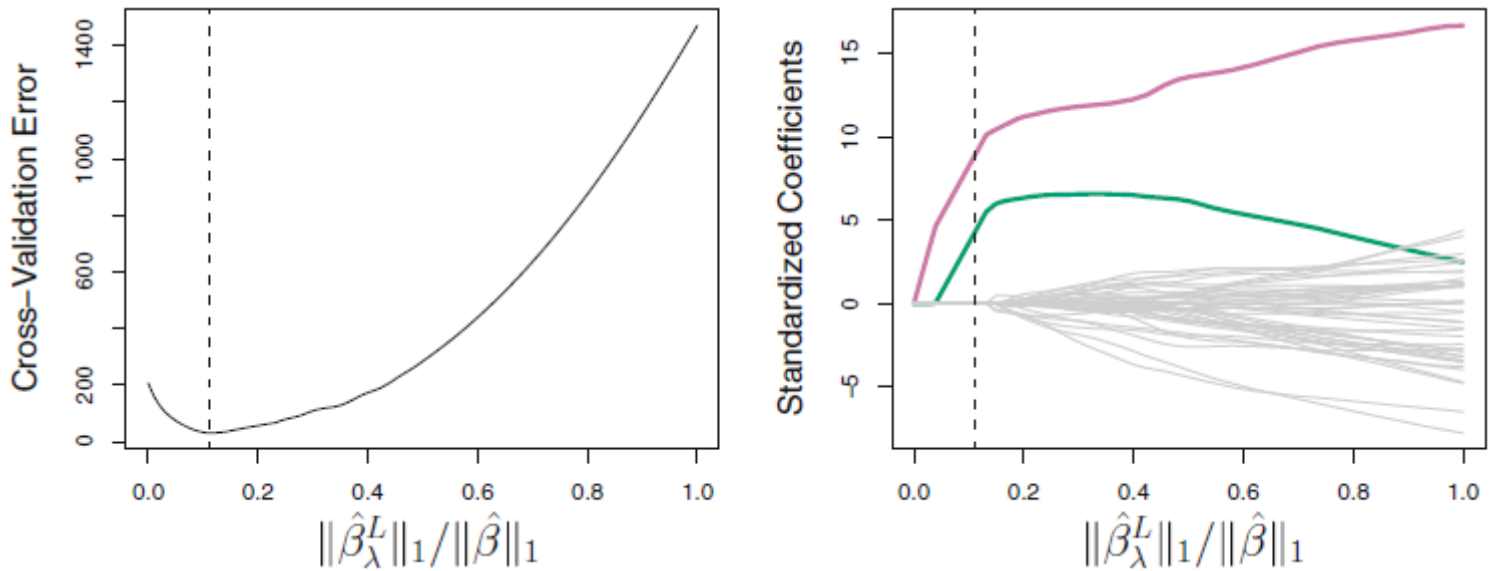


Figure: 6.13. Left: Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set from Figure 6.9. Right: The corresponding lasso coefficient estimates are displayed. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest.

Dimension reduction methods.

- The linear regression

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon_i, \ \ i = 1, \cdots, n.$$

- Our idea: when p is large,
  - transform the predictors
  - fit a least squares model using the transformed variables.
  - it is called *dimension reduction* methods.
- That is,
  - $Z_1, Z_2, \cdots, Z_M$ represent $M \ll p$ linear combinations of $X_1, ..., X_p$.

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j. \tag{2}$$

  - We regress on $Z_1, ..., Z_M$ with $M < p$.

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m Z_m + \epsilon_i, \ \ i = 1, ..., n. \tag{3}$$

Dimension reduction methods.

- The linear regression

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon_i, \ \ i = 1, \cdots, n.$$

- Our idea: when p is large,
  - transform the predictors
  - fit a least squares model using the transformed variables.
  - it is called *dimension reduction* methods.
- That is,
  - $Z_1, Z_2, \cdots, Z_M$ represent $M \ll p$ linear combinations of $X_1, ..., X_p$.

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j. \tag{2}$$

  - We regress on $Z_1, ..., Z_M$ with $M < p$.

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m Z_m + \epsilon_i, \ \ i = 1, ..., n. \tag{3}$$

- Plug (2) into (3) and compare with the original linear regression, we have

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}.$$

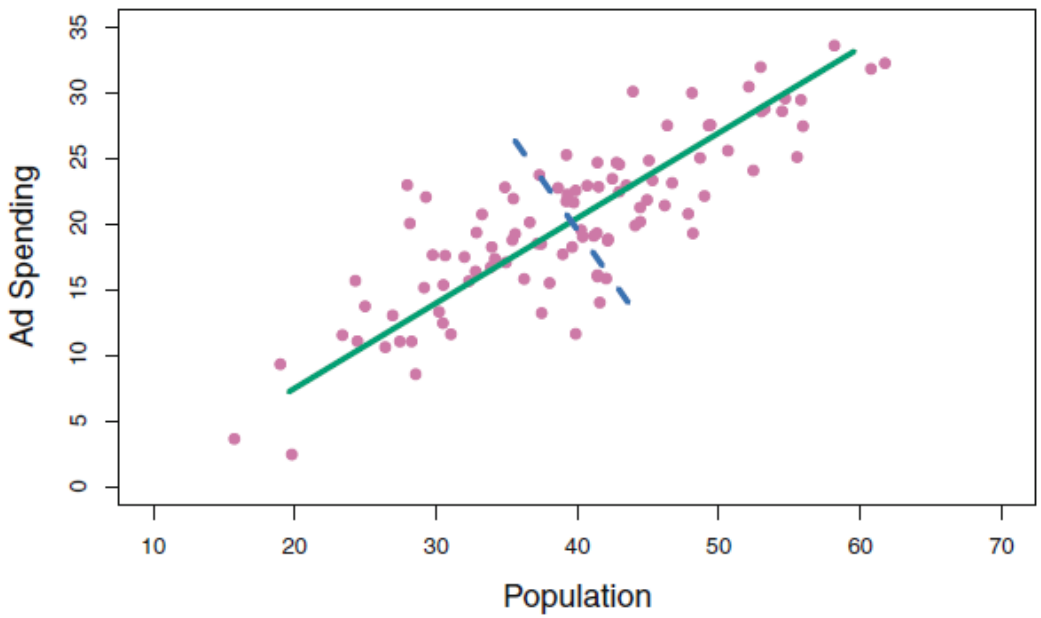
- Dimension reduction serves to constrain the estimated β_j coefficients.
- For Dimension reduction, a key step is how to determine the linear combination.
- Choose $Z_1, .., Z_M$ as the first M principle components.

Principal Component Analysis

- PCA is a technique for reducing the dimension of data matrix.
- The first principal component direction of the data is that along which the observations *vary the most*.
- The second principal component direction of the data is that along which the observations *vary the second most*.
- and so on...

Principal Component Analysis

- PCA is a technique for reducing the dimension of data matrix.
- The first principal component direction of the data is that along which the observations *vary the most*.
- The second principal component direction of the data is that along which the observations *vary the second most*.
- and so on...



- Let X be the random vector (sample) of p dimension,
- $\mathbf{a}$ be the vector the sample varies the most.
- The variation can be measured by $Z_1 = \mathbf{a}^T(X - \mathbb{E}(X))$.
- The variation of a one dimensional random variable X can be quantified by its variance. Then

$$\text{var}(Z_1) = \mathbf{a}^T \text{var}(X)\mathbf{a}.$$

- $\mathbf{a}$ should be the leading eigenvector of $\text{var}(X)$.

- Let X be the random vector (sample) of p dimension,
- $\mathbf{a}$ be the vector the sample varies the most.
- The variation can be measured by $Z_1 = \mathbf{a}^T(X - \mathbb{E}(X))$.
- The variation of a one dimensional random variable X can be quantified by its variance. Then

$$\text{var}(Z_1) = \mathbf{a}^T \text{var}(X)\mathbf{a}.$$

- $\mathbf{a}$ should be the leading eigenvector of $\text{var}(X)$.
- Let the columns of $A \in \mathbb{R}^{p \times M}$ be the M vectors the sample varies the most.
- The variation can be measured by $Z = A^T(X - \mathbb{E}(X))$.

- Let X be the random vector (sample) of p dimension,
- $\mathbf{a}$ be the vector the sample varies the most.
- The variation can be measured by $Z_1 = \mathbf{a}^T(X - \mathbb{E}(X))$.
- The variation of a one dimensional random variable X can be quantified by its variance. Then

$$\text{var}(Z_1) = \mathbf{a}^T \text{var}(X)\mathbf{a}.$$

- $\mathbf{a}$ should be the leading eigenvector of $\text{var}(X)$.
- Let the columns of $A \in \mathbb{R}^{p \times M}$ be the M vectors the sample varies the most.
- The variation can be measured by $Z = A^T(X - \mathbb{E}(X))$.
- For a p-dimension random vector, its variation, fully described by its covariance matrix Σ. Then

$$\text{var}(Z) = A^T \text{var}(X) A = A^T \Sigma A.$$

- A will be the eigenvectors corresponding to the the largest M eigenvalues when we solve A by maximizing $\text{var}(Z)$.

Principal Components as major statistical methodology

- Let X be the random vector of p dimension that we are concerned with and Σ be the covariance matrix of X.

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}, \Sigma = \text{var}(X) = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \vdots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{pmatrix}.$$

  where $\sigma_{kl} = \text{cov}(X_k, X_l)$.
- We assume here $E(X) = 0$ for convenience, since the mean of X plays no role in PCs.

Principal Component Analysis (PCA).

- By matrix singular value decomposition, we know

$$\Sigma = \mathbf{e}\Lambda\mathbf{e}'$$

where

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}, \ \mathbf{e} = (\mathbf{e}_1 \vdots \cdots \vdots \mathbf{e}_p) = \begin{pmatrix} e_{11} & \cdots & e_{1p} \\ \vdots & \vdots & \vdots \\ e_{p1} & \cdots & e_{pp} \end{pmatrix}$$

with $\lambda_1 \geq \cdots \geq \lambda_p > 0$ and $\mathbf{e}\mathbf{e}' = I_p$.

- $(\lambda_k, \mathbf{e}_k)$, $k = 1, ..., p$, are the eigenvalue-eigenvector pairs of the matrix Σ.

- *The first P.C:* $Z_1 = \mathbf{e}_1^T X$ *– the most important.*

$$\text{var}(Z_1) = \lambda_1 = \max\{\text{var}(b'X) : \|b\| = 1,\ b \in R^p\}.$$

The fraction of total variation of X explained by Z_1 is

$$\frac{\text{var}(Z_1)}{\text{var}(Z_1) + \cdots + \text{var}(Z_p)} = \frac{\lambda_1}{\lambda_1 + \cdots + \lambda_p}.$$

- *The first P.C:* $Z_1 = \mathbf{e}_1^T X$ – *the most important.*

$$\text{var}(Z_1) = \lambda_1 = \max\{\text{var}(b'X) : \|b\| = 1,\ b \in R^p\}.$$

The fraction of total variation of X explained by Z_1 is

$$\frac{\text{var}(Z_1)}{\text{var}(Z_1) + \cdots + \text{var}(Z_p)} = \frac{\lambda_1}{\lambda_1 + \cdots + \lambda_p}.$$

- ...
- *The k-th P.C:* $Z_k = \mathbf{e}_k^T X$ – *the k-th important.*

$$\text{var}(Z_k) = \lambda_k = \max_{b \in R^p}\{\text{var}(b'X) : \|b\| = 1,\ b'X \perp Z_i, i = 1, \cdots, k-1\},$$

Here and throughout, $\perp$ means 0 correlation. The fraction of total variation of X explained by Z_k is

$$\frac{\text{var}(Z_k)}{\text{var}(Z_1) + \cdots \text{var}(Z_p)} = \frac{\lambda_k}{\lambda_1 + \cdots \lambda_p}.$$

A summary table of PCs

		eigenvalue (variance)	eigenvector (combination coefficient)	percent of variation explained	P.C.s as linear combination of $X - \mu$
1st	Z_1	λ_1	$u\mathbf{e}_1$	$\lambda_1 / \sum_{j=1}^{p} \lambda_j$	$Z_1 = u_1'(X - \mu)$
2nd	Z_2	λ_2	bfe_2	$\lambda_2 / \sum_{j=1}^{p} \lambda_j$	$Z_2 = u_2'(X - \mu)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p-th	Z_p	λ_p	$\mathbf{e}_p$	$\lambda_p / \sum_{j=1}^{p} \lambda_j$	$Z_1 = u_p'(X - \mu)$

- The population P.C.s are only theoretical,
- in data analysis, we need to work with the sample P.C.s.
- Suppose there are n observations of p variables presented as

$$\mathbf{X} = \left(X_{(1)} \vdots X_{(2)} \vdots \cdots \vdots X_{(p)}\right) = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}_{n \times p}.$$

  Then $X_{(k)}$, an n-vector, contains all n observations of the k-th variable.
- We compute the sample variance matrix, denoted as $\mathbf{S}$ and do eigenvalue decomposition,

$$\mathbf{S} = \hat{\mathbf{e}}\hat{\Lambda}\hat{\mathbf{e}}'.$$

A summary of sample P.C.s

		eigenvalue (variance)	eigenvector (combination coefficient)	percent of variation explained	P.C.s as linear combination of $X - \mu$
1st	$Z_{(1)}$	$\hat{\lambda}_1$	$\hat{\mathbf{e}}_1$	$\hat{\lambda}_1 / \sum_{j=1}^{p} \hat{\lambda}_j$	$Z_{(1)} = \sum_{j=1}^{p} \hat{e}_{j1}(X_{(j)} - \bar{X}_1)$
2nd	$Z_{(2)}$	$\hat{\lambda}_2$	$\hat{\mathbf{e}}_2$	$\hat{\lambda}_1 / \sum_{j=1}^{p} \hat{\lambda}_j$	$Z_{(2)} = \sum_{j=1}^{p} \hat{e}_{j2}(X_{(j)} - \bar{X}_1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p-th	$Z_{(p)}$	$\hat{\lambda}_p$	$\hat{\mathbf{e}}_1$	$\hat{\lambda}_p / \sum_{j=1}^{p} \hat{\lambda}_j$	$Z_{(p)} = \sum_{j=1}^{p} \hat{e}_{jp}(X_{(j)} - \bar{X}_1)$

Principal component regression (PCR).

- Key assumption: a small number of principal components suffice to explain most of the variability in the data, as well as the relationship with the response.
- the directions in which $X_1, \cdots, X_p$ show the most variation are the directions that are associated with Y.

Principal component regression (PCR).

- Key assumption: a small number of principal components suffice to explain most of the variability in the data, as well as the relationship with the response.
- the directions in which $X_1, \cdots, X_p$ show the most variation are the directions that are associated with Y.
- Set $Z_i, i = \mathbf{e}_i, i = 1, \cdots, M$.
- fit a least squares model to $Z_1, \cdots, Z_M$.

Principal component regression (PCR).

- Key assumption: a small number of principal components suffice to explain most of the variability in the data, as well as the relationship with the response.
- the directions in which $X_1, \cdots, X_p$ show the most variation are the directions that are associated with Y.
- Set $Z_i, i = \mathbf{e}_i, i = 1, \cdots, M$.
- fit a least squares model to $Z_1, \cdots, Z_M$.
- If assumption holds, the fitting will lead a good result.

Partial least squares approach

- Principal components are designed to explain variation within X, not the relation of X with Y.
- The *key* assumption with principal components regression may NOT hold.
- Partial least squares approach avoids this shortcoming.

Partial least squares approach (PLS).

Like PCR, PLS is a dimension reduction method,

- first identifies squares a new set of features $Z_1, \cdots, Z_M$ that are linear combinations of the original features,
- fits a linear model via least squares using these M new features.

Partial least squares approach (PLS).

Like PCR, PLS is a dimension reduction method,

- first identifies squares a new set of features $Z_1, \cdots, Z_M$ that are linear combinations of the original features,
- fits a linear model via least squares using these M new features.

Unlike PCR, PLS identifies these new features in a supervised way

- it makes use of the response Y in order to identify new features that not only approximate the old features well,
- but also that are related to the response.

Partial least squares approach

- standardize each input $\mathbf{x}_j$ to have mean 0 and variance 1.
- Set $\hat{\mathbf{y}}^{(0)} = \bar{y}\mathbf{1}$ and $\mathbf{x}_j^{(0)} = \mathbf{x}_j, j = 1, \cdots, M$.
- For $m = 1, 2, \cdots, M$,
  $\mathbf{z}_m = \sum_{j=1}^{p} \hat{\phi}_{mj}\mathbf{x}_j^{(m-1)}$, where $\hat{\phi}_{mj} = \mathbf{y}^T\mathbf{x}_j^{(m-1)}$.
  $\hat{\theta}_m = \mathbf{z}_m^T\mathbf{y}/\mathbf{z}_m^T\mathbf{z}_m$
  $\hat{\mathbf{y}}^{(m)} = \hat{\mathbf{y}}^{(m-1)} + \hat{\theta}_m\mathbf{z}_m$.
  $\mathbf{x}_j^{(m)} = \mathbf{x}_j^{(m-1)} - s_{jm}\mathbf{z}_m$, where $s_{jm} = \mathbf{z}_m^T\mathbf{x}_j^{(m-1)}/\mathbf{z}_m^T\mathbf{z}_m$
- Output the sequence of fitted vectors $\{\hat{y}^{(m)}\}_1^p$. Since the $\{z_l\}_1^m$ are linear in the original $\mathbf{x}_j$, so is $\hat{y}^{(m)} = X\hat{\beta}^{\text{pls}}(m)$. These linear coefficients can be recovered from the sequence of PLS transformations.

Example

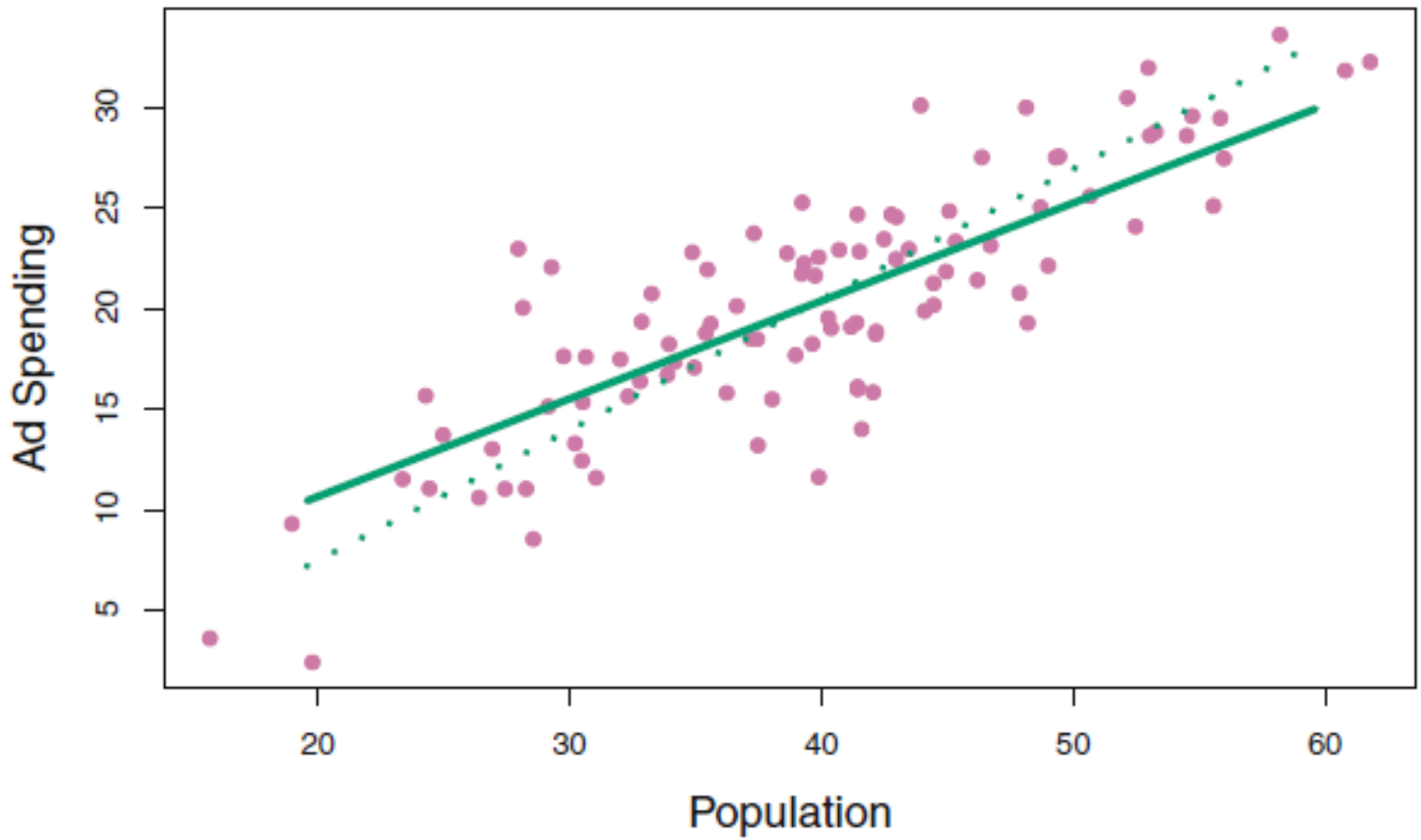


Figure: 6.21. For the advertising data, the first PLS direction (solid line) and first PCR direction (dotted line) are shown.

Partial least squares approach

- Partial least squares puts more weights on the variables with higher correlation with the response.
- It seeks the directions that have high variance and have high correlation with the reposne (while PCR only seeks those direction with high variance.)
- When the relationship between response and predictors is strong, PLS is better.
- Popular in chemometrics.

High dimension data

- Digitization of the society brings big data.
- Many of the datasets contain large number of varaibles.
- It is common that $p \gg n$, which is called the *curse of dimensionality*.
- Example: predition of blood pressure.
  Response: blood pressure.
  Inputs: SNPs; (Individual DNA mutations).
  n may be of hundreds, but p can be of millions.

The trouble

- Large p makes our linear regression model too flexible (or too large).
- If $p > n$, the LSE is not even uniquely determined.
- It can easily lead to overfit.
- A common phenomenon: small training error, but large test error.

Deal with high dimensional data

- Fit less flexible models to avoid overfit.
- feature selction:
  - forward stepwise selection,
  - ridge regression,
  - the lasso,
  - principal components regression.