Chapter6_Linear_Model_Selection_and_Regularisation

- Recall the linear model

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

- Despite its simplicity it has distinct advantages in terms of its interpretability and often shows good predictive performance.
- Let us look at ways in which this simple model could be improved, by replacing the ordinary least squares fitting with some alternate fitting procedures.
- **Prediction accuracy:** especially when $p > n$, to control the variance
- **Interpretability:** By removing irrelevant features, that is by setting the corresponding coefficient estimates to zero - we can obtain a model that is more easily interpreted. We will present some approaches for automatically performing **feature selection** .

Three classes of methods

- **Subset Selection:** We identify a subset of the p predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables.
- **Shrinkage:** We fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as **regularization**) has the effect of reducing variance and can also perform variable selection.
- **Dimension Reduction:** We project the p predictors into a M-dimensional subspace, where $M < p$. This is achieved by computing M different **linear combinations**, or **projections**, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares.

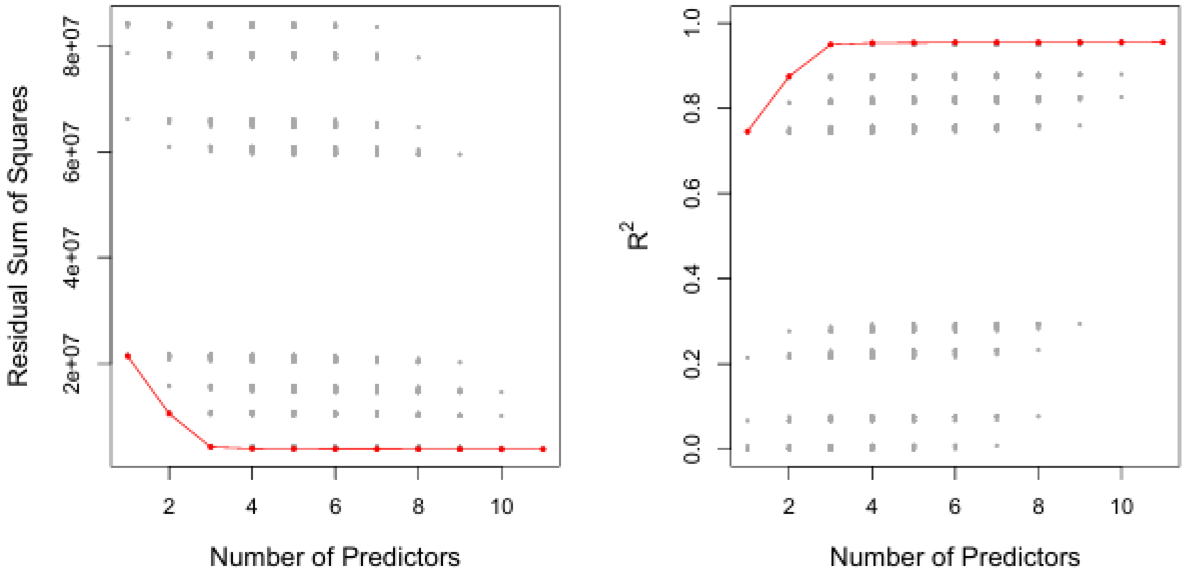
1. Subset Selection

There are two ways: Best Subset Selection and Stepwise Selection

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.



- 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 R2. 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.
  NOTE: Remember that you select the model based on the test error, can't just look at the graph and say that since RSS for model with 8 predictors is less than RSS for a model with 4 predictors because obviously model with more predictors will have less RSS. You can't compare apple with oranges.
- Although we have presented best subset selection for least squares selection, same ideas apply to other type of models such as logistic regression.
- The **deviance** - negative two times the maximized log-likelihood -plays the role of RSS for a broader class of models.

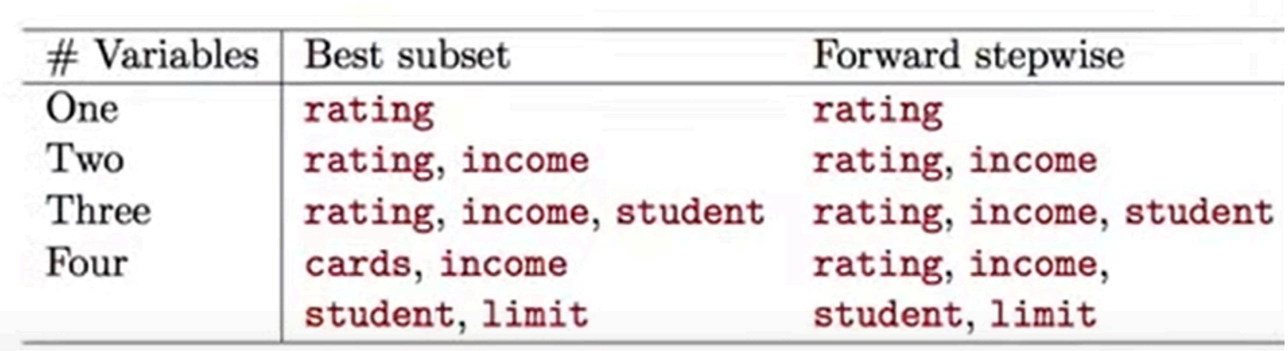
Stepwise Selection:

- For computational reasons, best subset selection can't be applied with very large p
- Best subset selection may also suffer from statistical problems when p is large: larger the search space. the higher the chance of finding good models that look good on training data, even though they might not have any predictive power.

- Thus an enormous search space can lead to overfitting and high variance of the coefficient estimates.
- Hence stepwise is attractive since it explores restricted set of models .

1. **Forward Stepwise Selection**
    - It begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time , until all predictors are in the model.
    - In particular , at each step the variable that gives the greatest additional improvement to the fit is added to the model.
    - In detail:
        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.

- Computational advantage over best subset selection is clear. We consider around p^2 models in forward stepwise selection.
- It's not guaranteed to find the bets possible model out of all 2^p models containing subsets of the p predictors.
- Lets look at the credit data example



# 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

- The first four selected models for best subset selection and forward stepwise selection are given above, notice how the fourth model is different . Best subset gives us the best model with 4 predictors but forward stepwise can only give us a model that already contains rating, income and student. It gonna add limit when we ask best model with 4 variables.
- But just because best subset has a better model on the training data, doesn't mean it's the best model on the test data.
- It can be shown that if there was no co-relation between the variables, you wouldn't have discrepancy between best subset and forward stepwise.

2. **Backward Stepwise Selection**
    - Unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.
    - In detail:
        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.

- Like forward stepwise selection, the backward selection approach searches through only $1 + p(p+1)/2$ models, and so can be applied in settings where p is too large to apply best subset selection
- Like forward stepwise selection, backward stepwise selection is 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.**

Choosing the optimal model

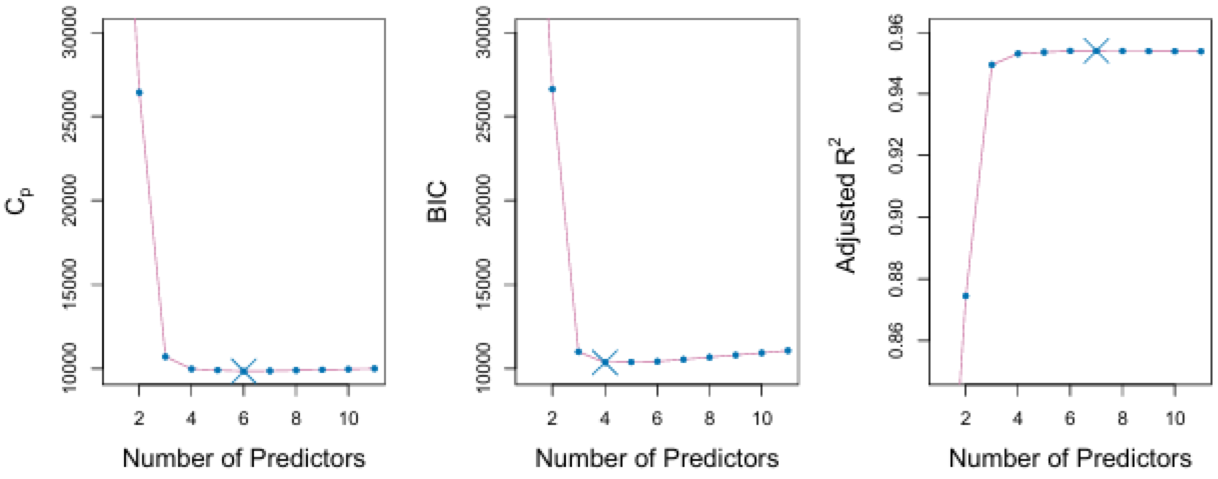
- The model containing all predictors will always have the smallest RSS and the largest R^2 , since these quantities are related to the training error.
- We wish to choose a model with a low test error , not a model with low training error. Recall that training error is usually a poor estimate of test error.
- Therefore, RSS and R^2 aren't suitable for selecting the bets model among a collection of models with different number of predictors.

Estimating Test Errors: two approaches

- We can indirectly estimate test error by making an **adjustment** to the training error for the bias due to overfitting.
- We can directly estimate the test error, using either a validation set approach or a cross-validation approach as discussed.

C_p , AIC ,BIC and Adjusted R^2

- These techniques adjust the training error for the model size, and can be used to select among a set of models with different number of variables.



- The above figure displays the C_p , BIC and Adjusted R^2 for the best model of each size produced by best subset selection on the **Credit** data set.
- Roughly speaking, we want the first two quantities to be as small as possible and the last one to be as large as possible.
- Now for some details:
    1. Mallow's C_p :

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

where d is the total number of parameters used and $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ associated with each response measurement.

2. The AIC criterion is defined for a large class of models fit by maximum likelihood:

$$AIC = -2\log L + 2 \cdot d$$

where L is the maximized value of the likelihood function for the estimated model.

3. Like C_p the BIC will tend to take on a small value for a model with a low test error , and so generally we select the model that has the lowest BIC value.

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

Notice that the BIC replaces the $2d\hat{\sigma}^2$ used by C_p with a $\log{(n)}d\hat{\sigma}^2$ term, where n is the number of observations.
Since $logn > 2$ for any $n > 7$, the BIC statistic generally places places a heavier penalty

on models with many variables , and hence results in the selection of smaller models than C_p

4. For a least squares model with d variables , the adjusted R^2 statistic is calculated as

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

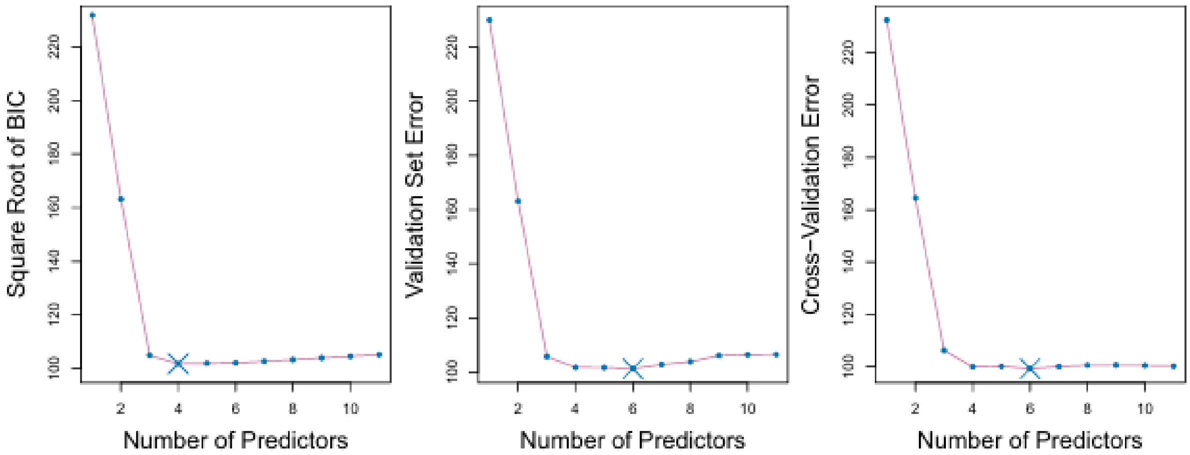
where TSS is the total sum of squares.
For this statistic , a large value of adjusted R^2 indicates a model with a small test error.
While RSS always decreases as the number of variables in the model decreases, $\frac{RSS}{n-d-1}$ may increase or decrease , due to the presence of d in the denominator.
Unlike the R^2 statistic , the adjusted R^2 **pays a price** for the inclusion unnecessary variables in the model.
Another advantage is that you can apply this when $p > n$. Also , unlike other statistics above it doesn't has a theoretical backing and you can't apply it to other models like logistic regression(can't generalize)

- In the case of the linear model with Gaussian errors , maximum likelihood and least squares are the same thing; and C_p and AIC are equivalent.
- One disadvantage of these statistics is that you can't apply it to other models, but methods like cross-validation can be applied to even some crazy models.

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. Also you don't need to know d .
- 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



- 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. Right: Cross-validation errors.
- In the above example, 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.
- In this setting, we can select a model using the **one-standard-error rule**. We first calculate the standard error of the estimated test MSE for each model size, and then select the smallest model for which the estimated test error is within one standard error of the lowest point on the curve.

2. Shrinkage Methods

- The subset selection methods use least squares to fit the 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 *constraints* or *regularizes* the coefficient estimates , or equivalently , that **shrinks** the coefficient estimates towards zero.
- 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 the variance

Ridge Regression

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

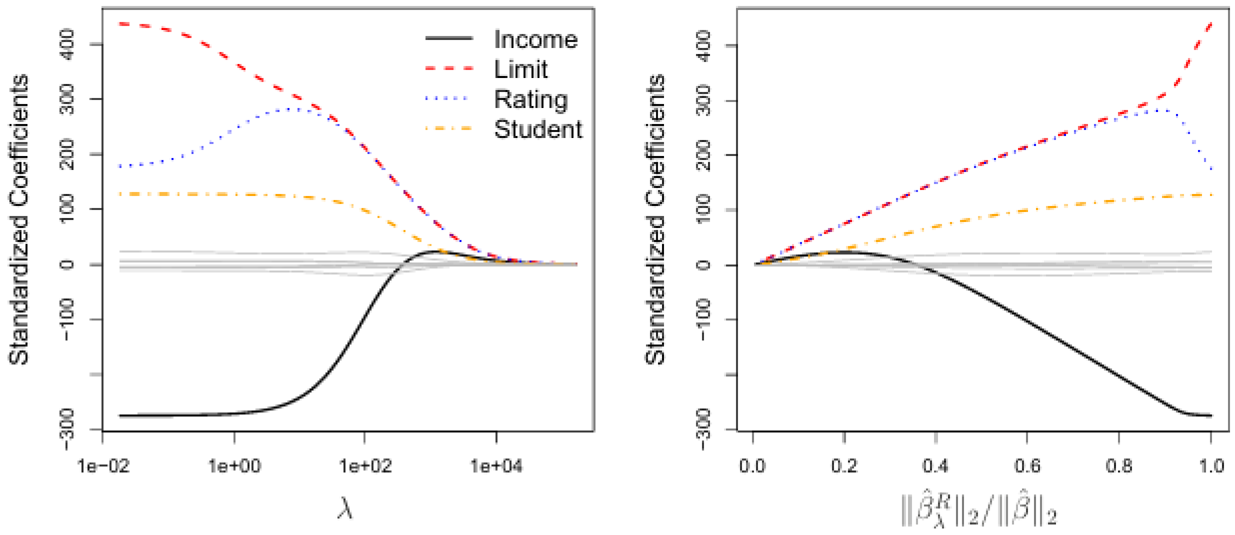
$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

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

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

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

- 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 $\beta_1, \beta_2, \dots, \beta_p$ 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.

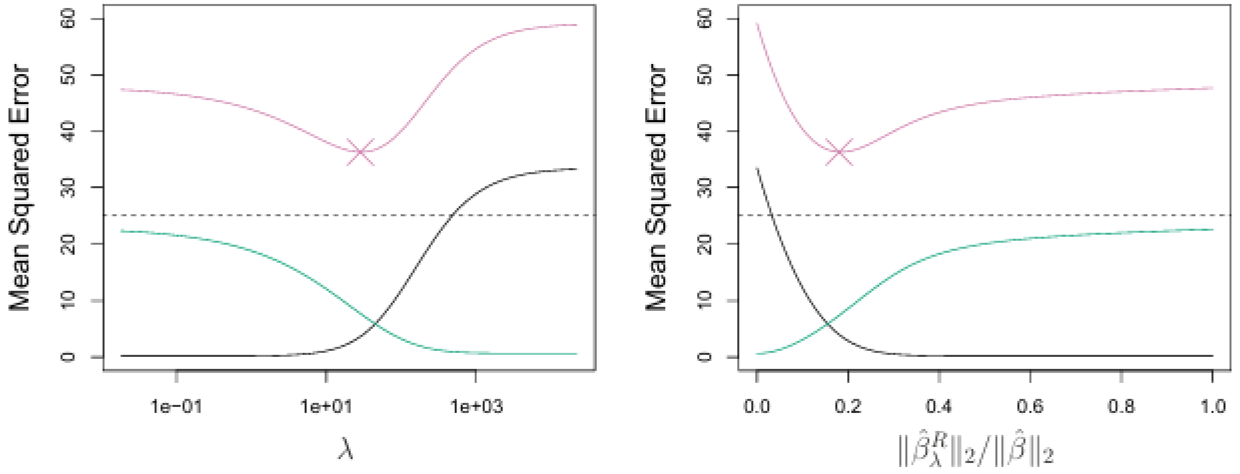


- In the left-hand panel, each curve corresponds to the ridge regression coefficient estimate for one of the ten variables, plotted as a function of λ.
- The right-hand panel displays the same ridge coefficient estimates as the left-hand panel, but instead of displaying λ on the x-axis, we now display $|\hat{\beta}^R|_2/|\hat{\beta}|_2$, 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 $|\beta|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$
- 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

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

Why does Ridge Regression Improve Over Least Squares?



In the above example, we have simulated data with $n = 50$ observations, $p = 45$ predictors, all having non-zero 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 and $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.

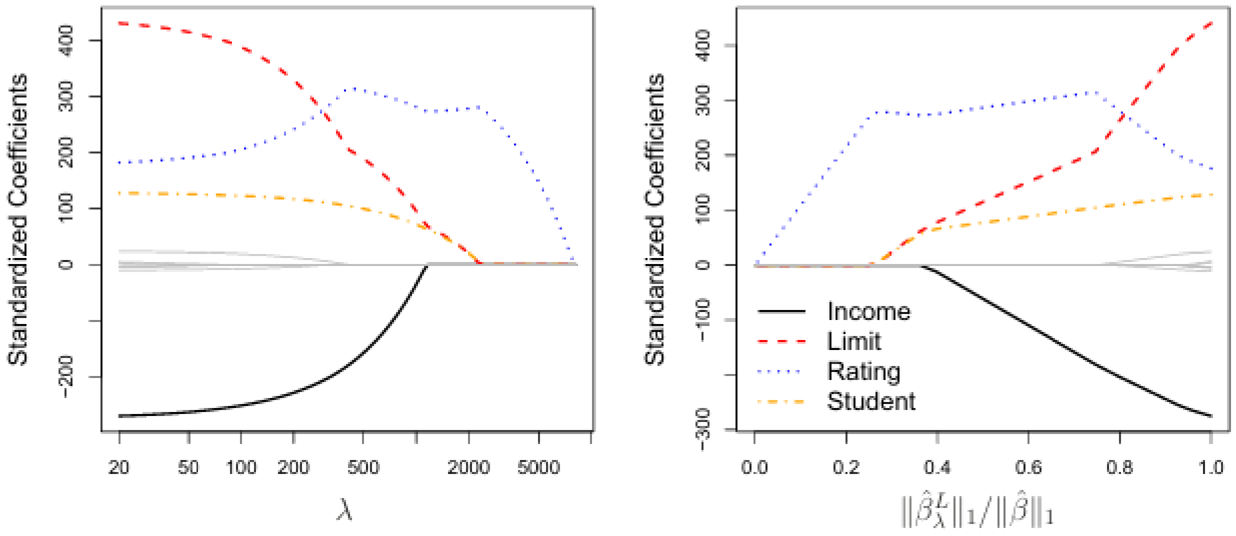
The Lasso

- 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

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

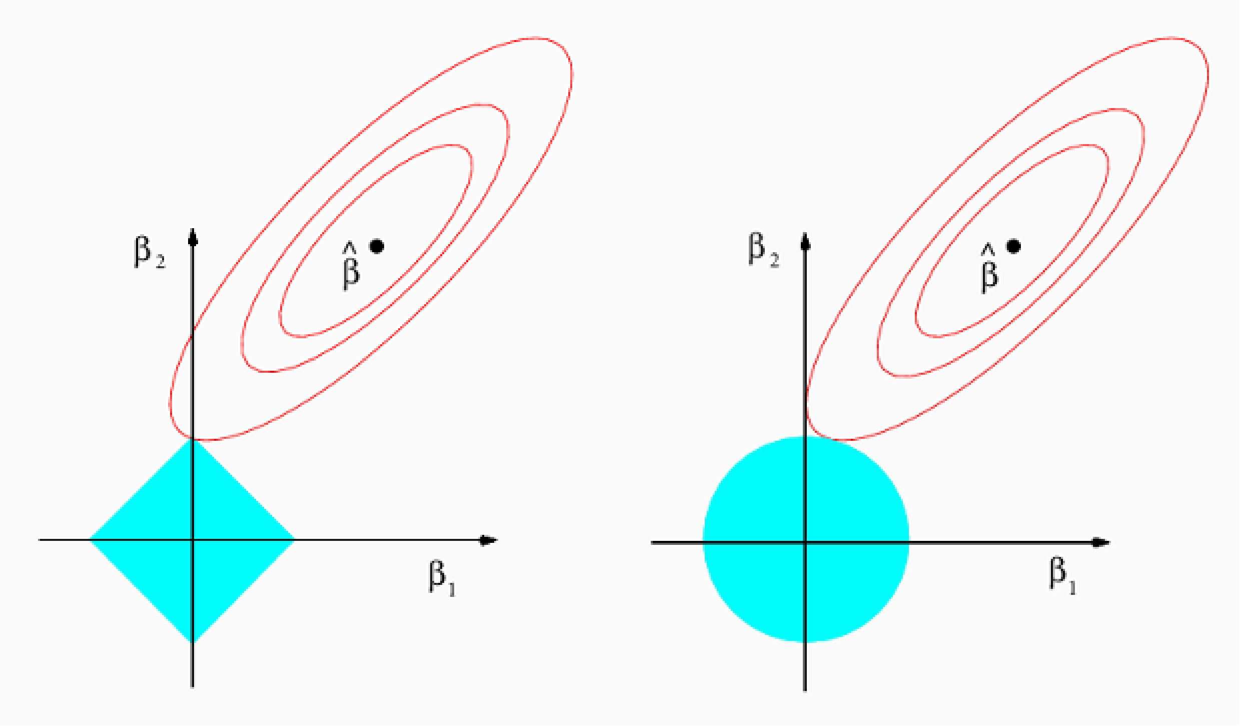
- In statistical parlance, the lasso uses an ℓ_1 (pronounced "ell 1") penalty instead of an ℓ_2 penalty. The ℓ_1 norm of a coefficient vector β is given by $\|\beta\|_1=\sum|\beta_j|$.
- As with ridge regression, the lasso shrinks the coefficient estimates towards zero.
- However, in the case of the 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.



- Above is an example where the standardized lasso coefficients on the Credit data set are shown as a function of λ and $\|\hat{\beta}_\lambda^L\|_1/\|\hat{\beta}\|_1$
- You can see that as we increase the value of λ , we get shrinkage as in ridge regression but something special happens here, some of the coefficients become exactly zero !
- So it's a combination of both shrinkage and selection of variables.

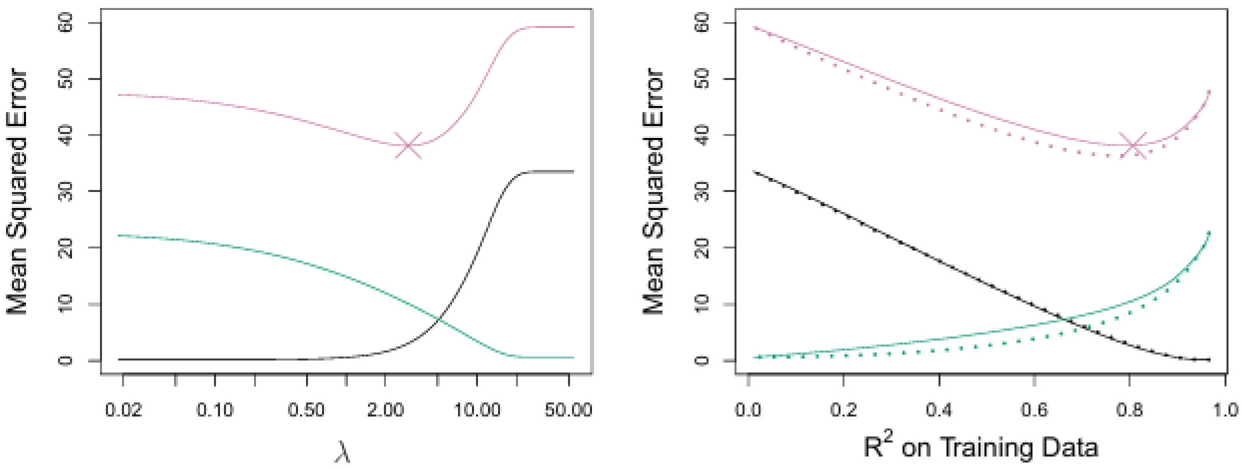
The variable selection property of the Lasso

- Why is it that the Lasso, unlike ridge regression , results in coefficient estimates that are exactly zero?
- One can show that the lasso and ridge regression coefficient estimates solve the problems $\min_\beta \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2$ subject to $\sum_{j=1}^p |\beta_j| \leq s$
  and
  $\min_\beta \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2$ subject to $\sum_{j=1}^p \beta_j^2 \leq s$, respectively.

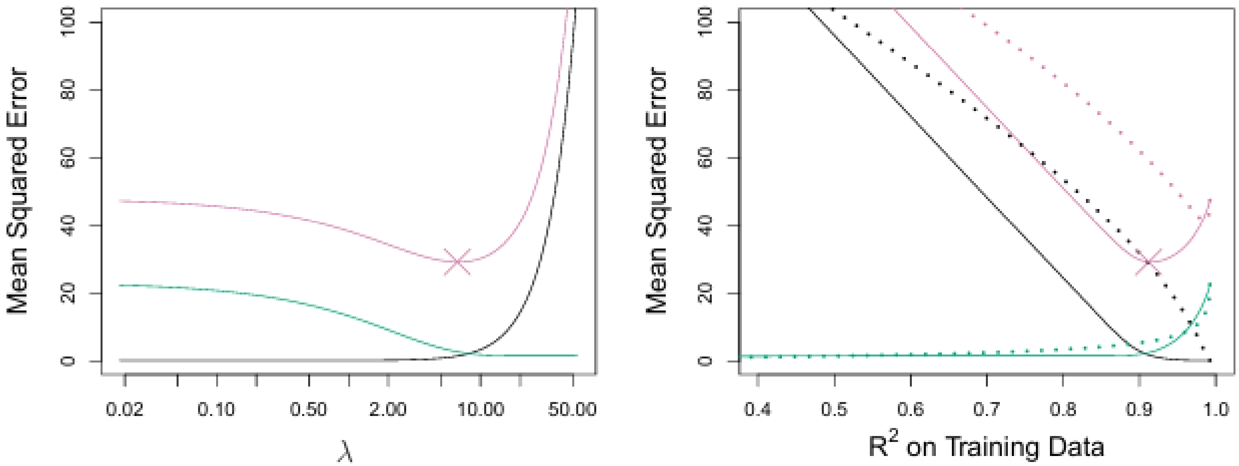


- Shown above are the contours of the error and constraint functions for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$, while the red ellipses are the contours of the RSS.
- Notice that the lasso constraint has corners at each of the axes, and so the ellipse will often intersect the constraint region at an axis. When this occurs, one of the coefficients will equal zero. In higher dimensions, many of the coefficient estimates may equal zero simultaneously.

Comparing the Lasso and Ridge Regression



- Simulated data with $n = 50$ observations , $p = 45$ predictors, all having non-zero coefficients.
- **Left**: Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on a simulated data set.
- **Right**: Comparison of squared bias, variance, and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their R2 on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.
- You can see that ridge performs better than Lasso here. The reason is that the true model isn't sparse but involves 45 predictors(non-zero).



- Here is a similar situation, except that now the response is a function of only 2 out of 45 predictors.
- Left: Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso.
- Right: Comparison of squared bias, variance, and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their R2 on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.

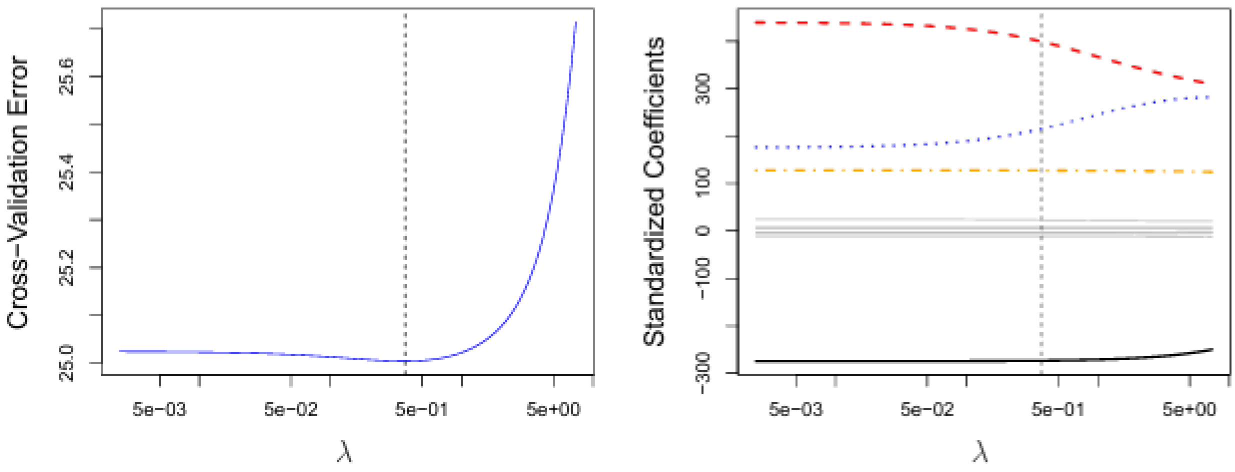
- Now we have a true model which is sparse, so the minimum MSE is for a large value of λ because Lasso wants to make the model sparse. We also see Lasso outperforming Ridge.

Conclusions from the above discussion

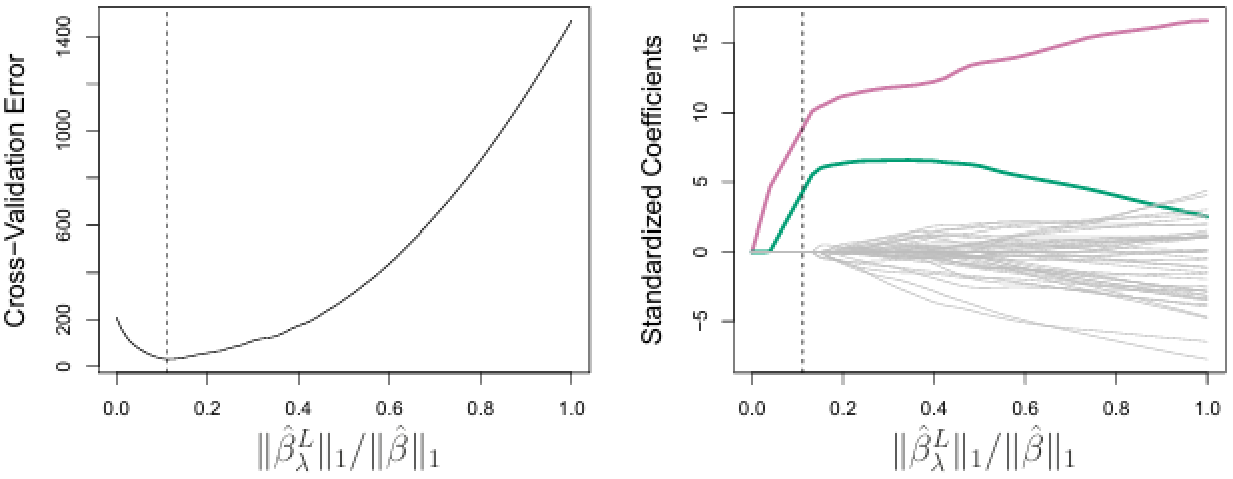
- These two examples illustrates that neither ridge regression nor the lasso will universally dominate the other.
- In general one might expect the lasso to perform better when the response is a function of only a relatively small number of predictors.
- However , the number of predictors that is related to the response is never known a $priori$ for real data sets.
- A technique such as cross-validation can be used in order to determine which approach is better on a particular dataset.

Selecting the tuning parameters for Ridge Regression and Lasso

- As for subset selection, for ridge regression and lasso we require a method to determine which of the models under consideration is best.
- That is , we require a method selecting 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. We choose a grid of λ values and compute the cross-validation error rate for each value of λ
- We then select the tuning parameter value for which the cross-validation error is smallest.
- Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.



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



- Left: Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set with $n = 50$ and two non-zero predictors.
- Right: The corresponding lasso coefficient estimates are displayed. The two signal variables are shown in color, and the noise variables are in gray. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest.

3. Dimension Reduction Methods

The methods that we discussed till now have involved fitting linear regression models, via least squares or a shrunken approach, using the original predictors $X_1, X_2, \ldots, X_p$. We'll now look at a class of approaches that transform the predictors and then fit a least squares model using the transformed variables. We will refer to these techniques as **dimension reduction** methods.
We will see two techniques: **Principal Component Analysis** and **Partial Least Squares.**

Details

- Let $Z_1, Z_2, \ldots, Z_M$ represent $M < p$ **linear combinations** of our original p predictors. That is,

$$Z_m = \sum_{j=1}^{p} \phi_{mj} Xj$$

for some constants $\phi_{m1}, \ldots, \phi_{mp}$.

- We can then fit the linear regression model,

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i$$

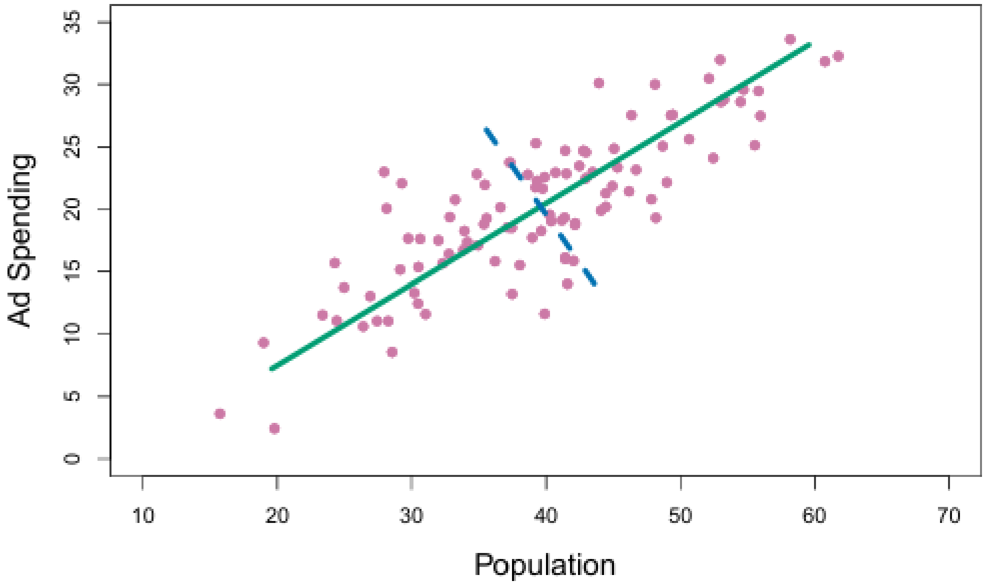
  , $i = 1, \ldots, n$, using ordinary least squares.
- Note that in model (2), the regression coefficients are given by $\theta_0, \theta_1, \ldots, \theta_M$. If the constants $\phi_{m1}, \ldots, \phi_{mp}$ are chosen wisely, then such dimension reduction approaches can often outperform OLS regression.
- It can be shown that

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

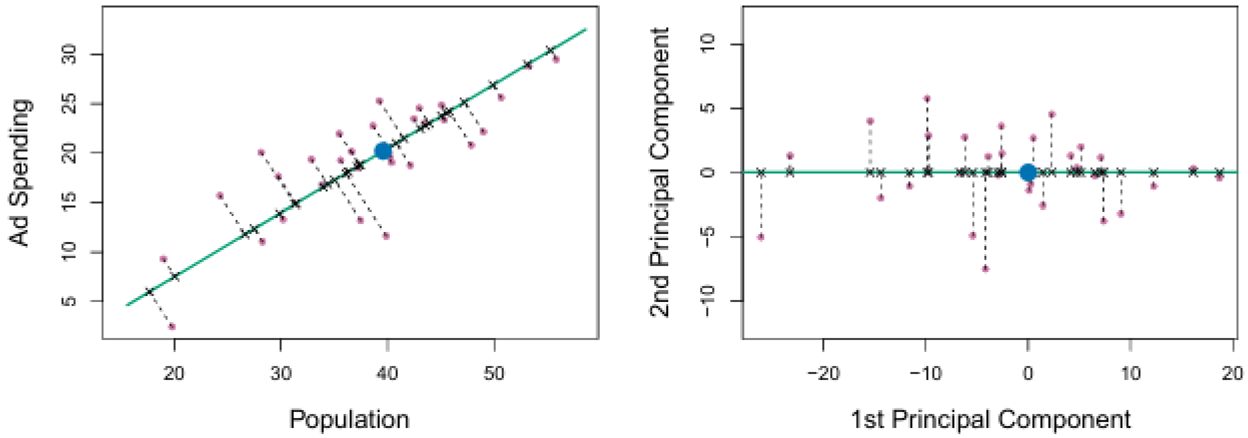
- Hence model(2) can be considered a special case of the original linear regression model.
- Dimension reduction serves to constrain the estimated β_j coefficients, since now they must take the above form.
- This can win in the bias-variance trade off.

Principal Component Regression

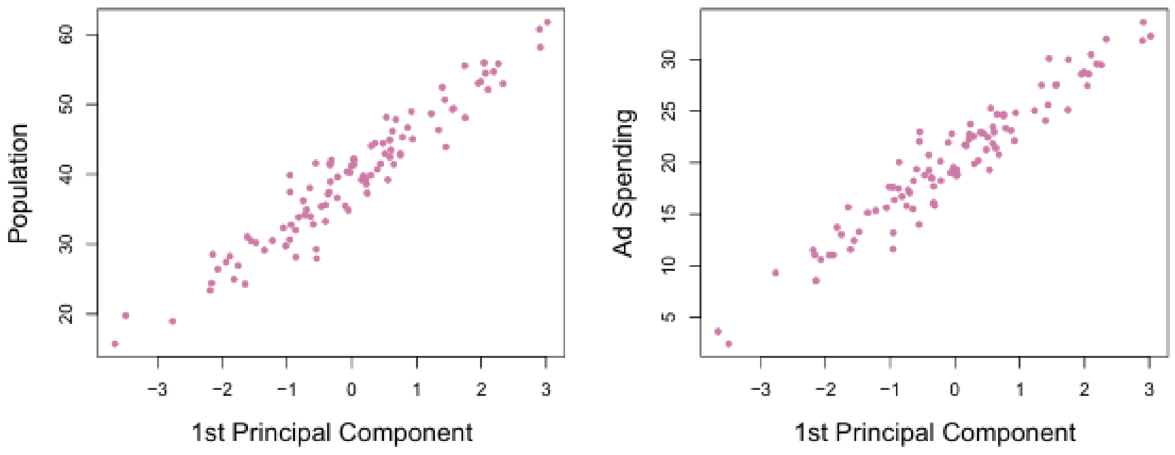
- The most famous technique .
- The first principal component is the (normalized) linear combination of the variables with largest variance.
- The second principal component has largest variance , subject to being uncorrelated with the first.
- And so on.
- Hence with many correlated original variables , we replace them with a small set of principal components that capture their joint variation.



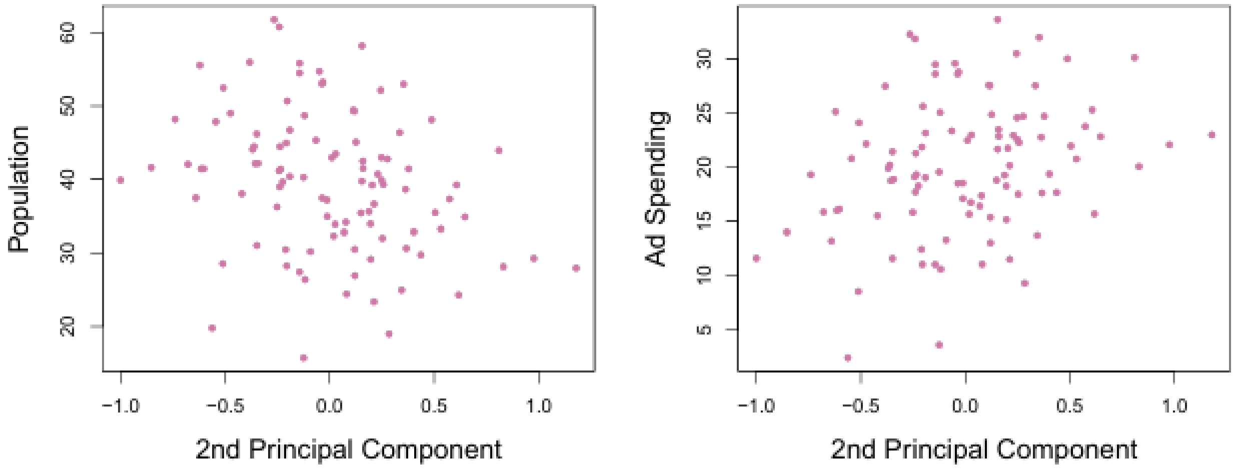
- The population size(pop) and ad spending(ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.



- A subset of the advertising data. The mean pop and ad budgets are indicated with a blue circle.
- Left: The first principal component direction is shown in green. It is the dimension along which the data vary the most, and it also defines the line that is closest to all n of the observations. The distances from each observation to the principal component are represented using the black dashed line segments. The blue dot represents (pop, ad).
- Right: The left-hand panel has been rotated so that the first principal component direction coincides with the x-axis.

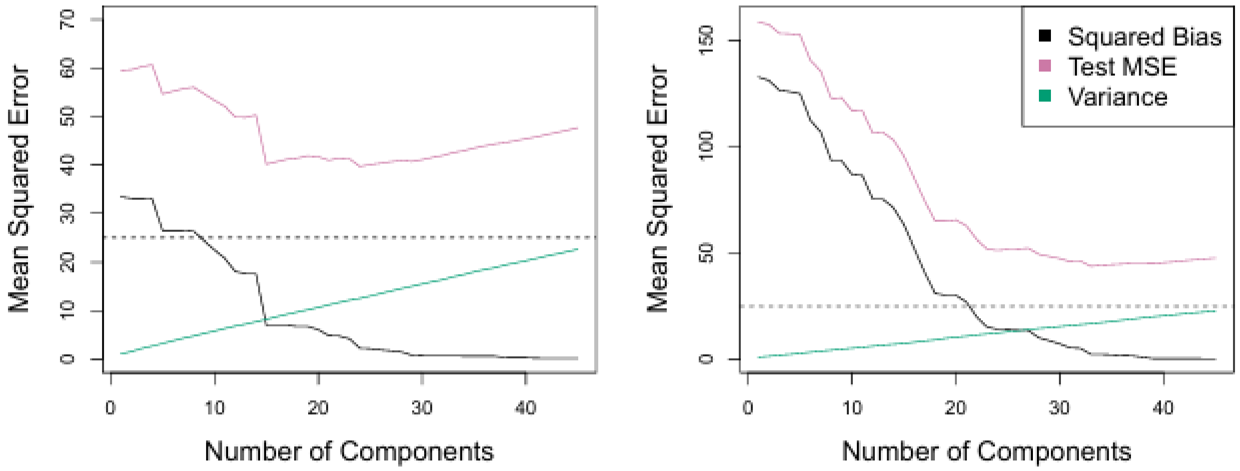


- Plots of the first principal component scores z_{i1} versus pop and ad. The relationships are strong.



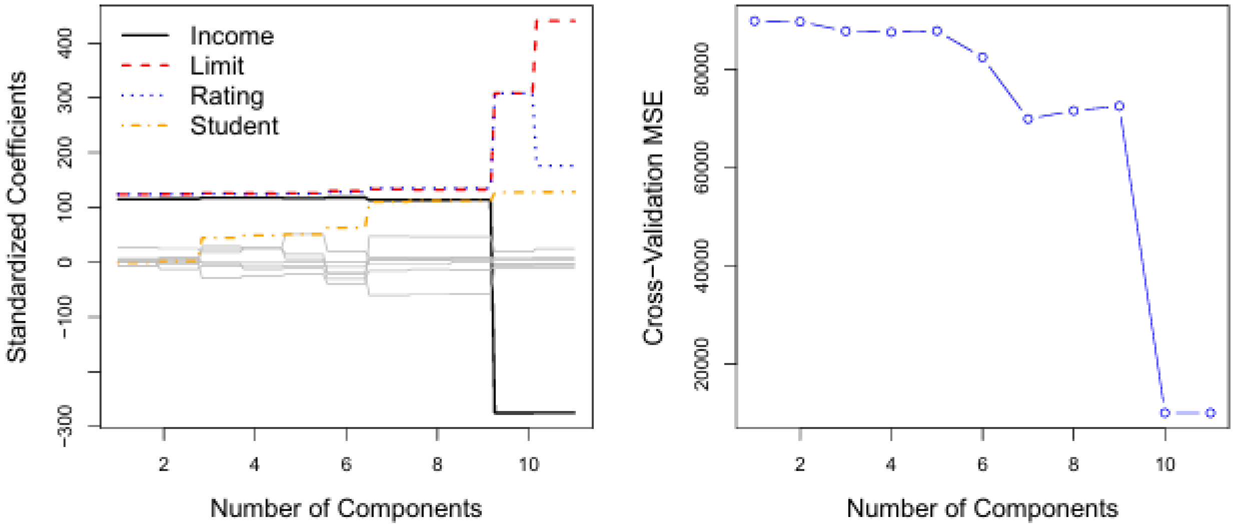
- Plots of the second principal component scores z_{i2} versus pop and ad. The relationships are weak.

Applications to Principal Components Regression



- PCR was applied to two simulated data sets. In each panel, the horizontal dashed line represents the irreducible error.
- Left: Simulated data when $n = 50$ with 45 non-zero predictors.
- Right: Simulated data $n = 50$ with 2 non-zero predictors.

Choosing the number of directions M



- Left: PCR standardized coefficient estimates on the Credit data set for different values of M.
- Right: The ten-fold cross-validation MSE obtained using PCR, as a function of M.
- Prefer using cross-validation to choose M.
- Here you see that Cross-validation error is lowest when $M = 10$, essentially the ordinary least squares cause $p = 10$. This is kind of disappointing as we didn't get any gains but it is contextual and can occur.

Partial Least Squares

- PCR identifies linear combinations, or **directions**, that best represent the predictors $X_1, \ldots, X_p$.
- These directions are identified in an **unsupervised** way, since the response Y is not used to help determine the principal component directions.
- That is, the response does not **supervise** the identification of the principal components.
- Consequently, PCR suffers from a potentially serious drawback: there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.
- Like PCR, PLS is a dimension reduction method, which first identifies a new set of features $Z_1, \ldots, Z_M$ that are linear combinations of the original features, and then fits a linear model via OLS using these M new features.
- But unlike PCR, PLS identifies these new features in a supervised way - that is, 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**.
- Roughly speaking, the PLS approach attempts to find directions that help explain both the response and the predictors.
- **Details:**
  - After standardizing the p predictors, PLS computes the first direction Z_1 by setting each ϕ_{1j} in (1) equal to the coefficient from the simple linear regression of Y onto X_j.
  - One can show that this coefficient is proportional to the correlation between Y and X_j .
  - Hence, in computing $Z_1 = \sum_{j=1}^{p} \phi_{1j} X_j$, PLS places the highest weight on the variables that are most strongly related to the response.
  - Subsequent directions are found by taking residuals and then repeating the above prescription.