Introduction to Statistical Learning

2020-2021 Fall

Linear Model Selection and Regularization

Bias vs Variance

- Variance:
- Variance refers to the amount by which f(X) would change if we estimated it using a different training data set.
- Since the training data are used to fit the statistical learning method, different training data sets will result in a different f(X).
- But ideally the estimate for f should not vary too much between training sets.
- However, if a method has high variance then small changes in the training data can result in large changes in f(X).

Bias vs Variance

- Bias:
- refers to the error that is introduced by approximating a (possibly extremely complicated) real-life problem, by a much simpler model
- linear regression assumes that there is a linear relationship between Y and X1,X2, . . . ,Xp.
- It is unlikely that any real-life problem truly has such a simple linear relationship, and so performing linear regression will undoubtedly result in some bias in the estimate of f.
- If the true f is very close to linear, and so given enough data, it should be possible for linear regression to produce an accurate estimate.
- Generally, more flexible methods result in less bias.

Linear Model Selection and Regularization

- Why people still choose linear models although we have many different ways of predicting?
  - Prediction Accuracy
  - Model Predictability.
- **Prediction Accuracy**
- Provided that the true relationship between the response and the predictors is approximately linear, the least squares estimates will have **low bias.**

Linear Model Selection and Regularization

- **Prediction Accuracy**
- If $n \gg p$,
  - then the least squares estimates tend to also have **low variance**, and hence will perform well on test observations.
- If n is not much larger than p,
  - then there can be a lot of variability in the least squares fit, resulting in overfitting and consequently poor predictions on future observations not used in model training
- If $p > n$,
  - then there is no longer a unique least squares coefficient estimate: the variance is infinite so the method cannot be used at all.

Linear Model Selection and Regularization

- **Prediction Accuracy**
- By ***constraining*** or ***shrinking*** the estimated coefficients,
  - we can often substantially reduce the variance at the cost of a negligible increase in bias.
  - This can lead to substantial improvements in the accuracy with which we can predict the response for observations not used in model training.

Linear Model Selection and Regularization

- **Model Interpretability**
- When the many of the variables used in the regression model, some of them are not associated with the response.
  - Including such irrelevant variables leads to unnecessary complexity in the resulting model.
- By removing these variables (i.e., by setting the corresponding coefficient estimates to zero),
  - We can obtain a model which is more easily interpreted.
  - Least squares is extremely unlikely to yield any coefficient estimates that are exactly zero.

Linear Model Selection and Regularization

- Three important class of methods
  - ***1. Subset selection:***
    - Reduced set of variables obtained. These variables are related to response.
  - ***2. Shrinkage***:
    - Model involves all p varibles, but the estimated coefficients are shunken towards zero relative to the least squares estimates.
    - The shrinkage is also known as ***regularization.***
    - It affects the reducing variance and also it can be used for variable selection
  - ***3. Dimension Reduction:***
    - ***projecting*** 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

Subset Selection – Best Subset Selection

- We fit a separate least squares regression for each possible combination of the p predictors.
  - All p models contains only one variable,
  - All $\binom{p}{2} = p(p-1)/2$ models contain two variables, …
- Then we look at the resulting models, with the goal of identifying the one that is best.
- There are 2^p posibilities to select best model.

Subset Selection – Best Subset Selection

Algorithm 6.1 *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 $k = 1, 2, \ldots p$:

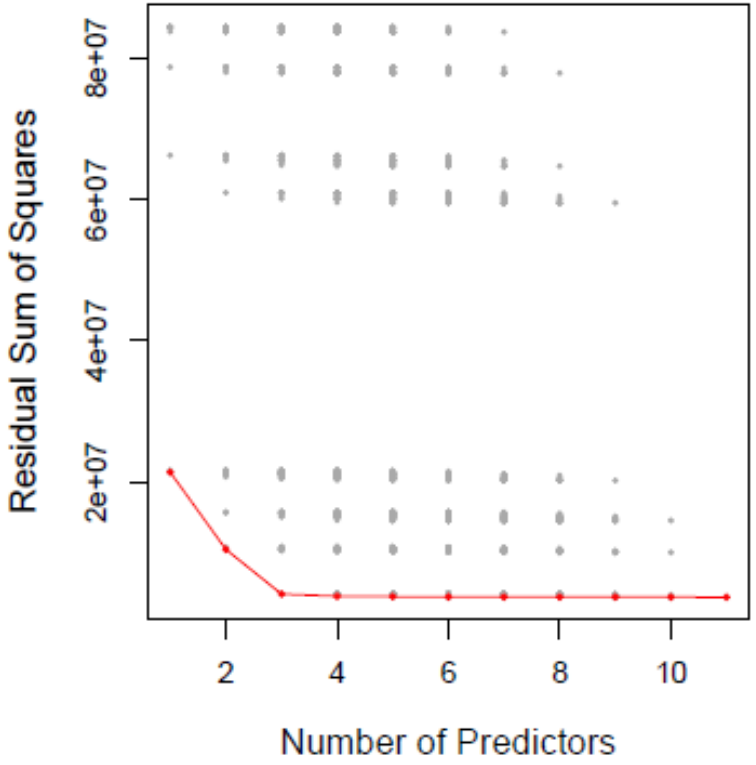
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.

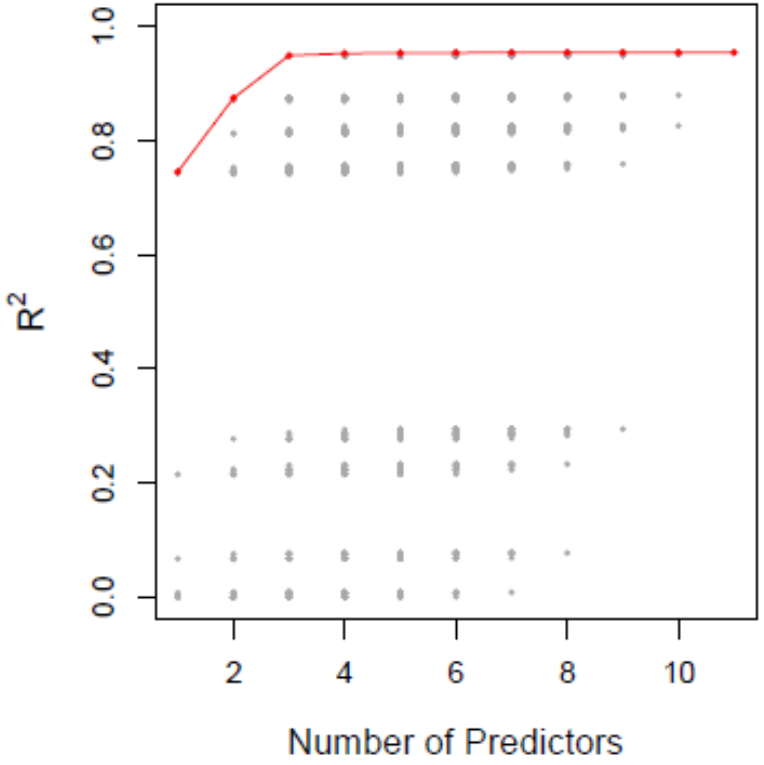
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.

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.

Subset Selection – Best Subset Selection

- Credit Data Set





Subset Selection – Best Subset Selection

- Best subset selection is
  - A simple and conceptually appealing approach,
  - But suffers form *computational limits* (2^p with subsets of p predictors).
    - Consequently, it becomes computationally infeasible for values of p greater than around 40, even with extremely fast modern computers.
  - And also, when p is large, it suffers from statistical problems.
    - The larger the space, the higher the chance finding models that look good on the training data. (overfitting and high variance of coefficient estimates)
  - For this reason alternative methods such as ***stepwise selections*** are used.

Stepwise Selection
Forward Stepwise Selection

- A computationally efficient alternative, contains much less than 2^p variables.
- Forward stepwise selection model begin with no predictors, and then adds predictors to model, one at a time, until all of the predictors in the model.
- A total amount of models

$$1 + \frac{p(p+1)}{2}$$

- At each step the variable that gives the greatest additional improvement to the fit is added to the model.

Stepwise Selection
Forward Stepwise Selection

Algorithm 6.2 *Forward Stepwise Selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) 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.

Stepwise Selection
Forward Stepwise Selection

- **Example**: Credit data set
- Comparison of the best subset selection and forward stepwise selection

# 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

Stepwise Selection
Forward Stepwise Selection

- Forward stepwise selection can be applied even in the high-dimensional setting where n < p,
  - However, in this case, it is possible to construct submodels $M_0, \ldots, M_{n-1}$ only, since each submodel is fit using least squares, which will not yield a unique solution if p ≥ n.

Stepwise Selection
Backward Stepwise Selection

Algorithm 6.3 *Backward Stepwise Selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, 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.
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.

Stepwise Selection
Backward Stepwise Selection

- Backward stepwise selection
  - begins with the full least squares model containing ***all p predictors***, and then iteratively removes the least useful predictor, one-at-a-time.
  - searches $\mathbf{1} + \frac{\boldsymbol{p}(\boldsymbol{p}+\mathbf{1})}{\mathbf{2}}$ models.
  - can be applied in setting where p is too large to apply best subset selection.
  - is **not** also **guranteed** to yield the **best model** containing a subset of the p predictors.
  - requires that $\boldsymbol{n} > \boldsymbol{p}$.
    - In contrast, forward stepwise selection can be used even when $n < p$ (the only viable subset method when p is very large).

Stepwise Selection
Hybrid Approaches

- Hybrid version of forward and backward stepwise selection are available.
  - Variables are added to the model sequentially, in analogy to forward selection.
  - After adding each new variable, the method may also remove any variables that no longer provide an improvement in the model fit.
- Such an approach attempts to more closely mimic best subset selection while retaining the computational advantages of forward and backward stepwise selection.

Choosing Optimal Model

- Best subset selection, forward selection, backward selection
  - To implement these methods, we need to determine which of these models is ***best***.
- The model containing all of the predictors will always have the ***smallest RSS*** and the ***largest*** $\boldsymbol{R^2}$
  - These quantities are related to the training error.
  - But, we want to choose model with a low test error.
  - Morever, the training error can be poor estimate of the test error.
- Therefore, RSS and R^2 are not suitable for selecting the best model among a collection of models with different numbers of predictors.

Choosing Optimal Model

- To select best model, we need to estimate the test error.
- There are two common approaches:
  1. We can *indirectly* estimate test error by ***making an adjustment*** to the training error to account for the bias due to overfitting.
  2. We can *directly* estimate the test error, using either a validation set approach or a cross-validation approach.

C_p, AIC, BIC, and Adjusted R^2

- A number of techniques for adjusting the training error for the model size are available. These approaches can be used to select among a set of models with different numbers of variables:
  - $\boldsymbol{C_p = \frac{1}{n}\left(RSS + 2d\hat{\sigma}^2\right)}$
  - Akaike information criterion (AIC) $\boldsymbol{AIC = \frac{1}{n\hat{\sigma}^2}\left(RSS + 2d\hat{\sigma}^2\right)}$
  - Bayesian information criterion (BIC) $\boldsymbol{BIC = \frac{1}{n}\left(RSS + \log(n)\, d\hat{\sigma}^2\right)}$
  - Adjusted R^2 $\boldsymbol{R^2 = 1 - \frac{RSS}{TSS}}$

Validation and Cross-Validation

- Another alternative approaches is ***directly estimating the test error*** using the validation set and cross-validation methods.
  - Compute the validation set error or the cross-validation error for each model under consideration,
  - Then, select the model for which the resulting estimated test error is smallest.
- In the past, performing cross-validation was computationally prohibitive for many problems with large p and/or large n.
  - Cross-validation is a very attractive approach for selecting from among a number of models under consideration.

2. Shrinkage Models - Ridge

Ridge

Lasso

Shrinkage Methods

- The subset selection method
  - Involves using *least squares* to fit a linear model that contains a subset of the predictors.
- Alternatively, a model containing all p predictors can be fit by using a technique that
  - ***constraints*** or ***regularizes*** the coefficient estimates, or
  - ***shrinks*** the coefficient estimates towards zero.
- Techniques for shrinking the regression coefficients
  - ***Ridge regression***
  - ***Lasso***

Ridge Regression

The ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimizes

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

Shrinkage Penalty

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

The Lasso

- The ***lasso*** is a relatively recent alternative to ridge regression.
- The lasso coefficients, $\hat{\beta}_{\lambda}^{L}$, minimize the quantity

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

Lasso Penalty

- The lasso uses an l_1.
- The l_1 norm of a coefficient vector β is given by $\left||\beta|\right|_1 = \sum|\beta_j|$

Ridge Regression

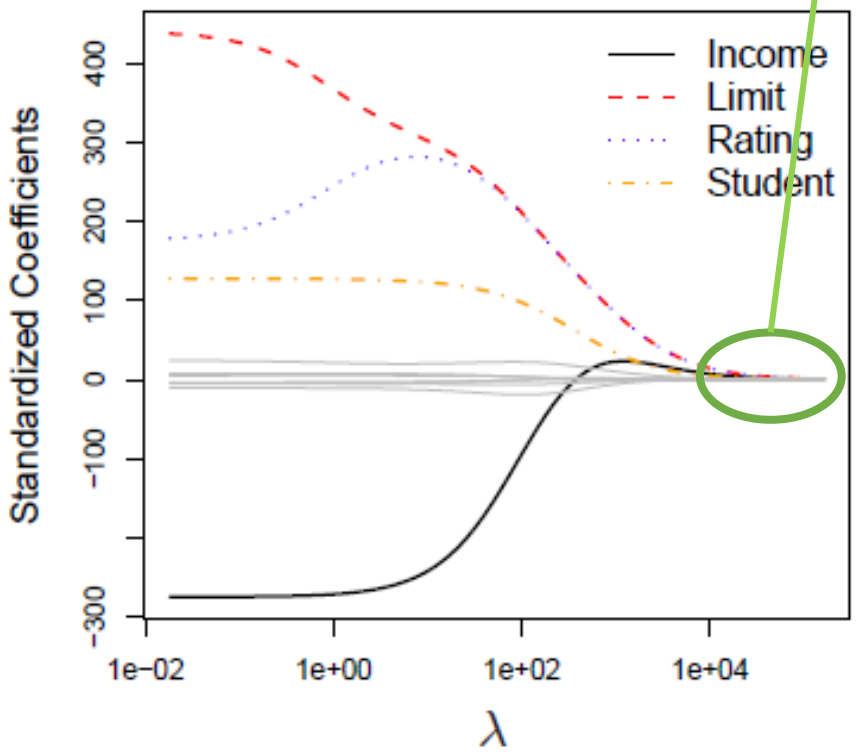
- Tuning Parameter:
  - When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates.
  - As $\lambda \to \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero.

- The shrinkage penalty is applied to $\beta_1, \ldots, \beta_p$ but not to the intercept β_0.
  - We want to shrink the estimated association of each variable with the response

Ridge Regression

Example: Credit Data Set

Null model,
Lambda is extremely large and
coefficient estimates are zero



Ridge Regression

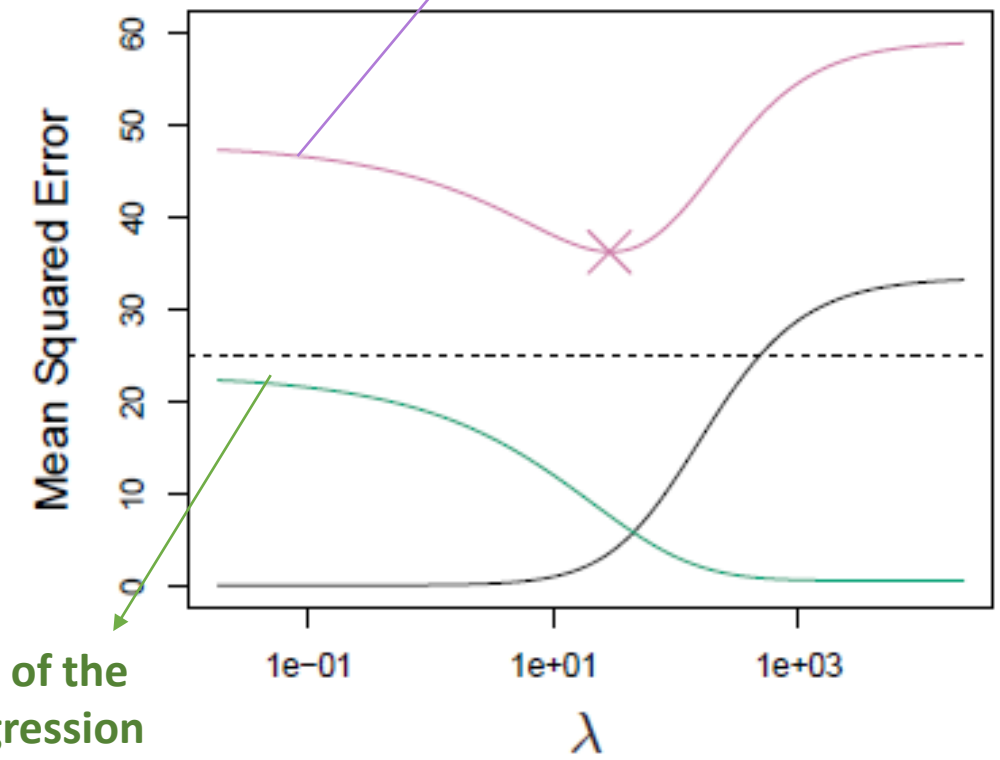
- The ridge regression coefficient estimates can change *substantially* when multiplying a given predictor by a constant.
- $X_j \hat{\beta}_{j,\lambda}^R$ will depend not only on the value of λ, but also on the scaling of the jth predictor.
  - In fact, the value of $X_j \hat{\beta}_{j,\lambda}^R$ may even depend on the scaling of the other predictors!
  - Therefore, it is best to apply ridge regression after *standardizing the predictors*

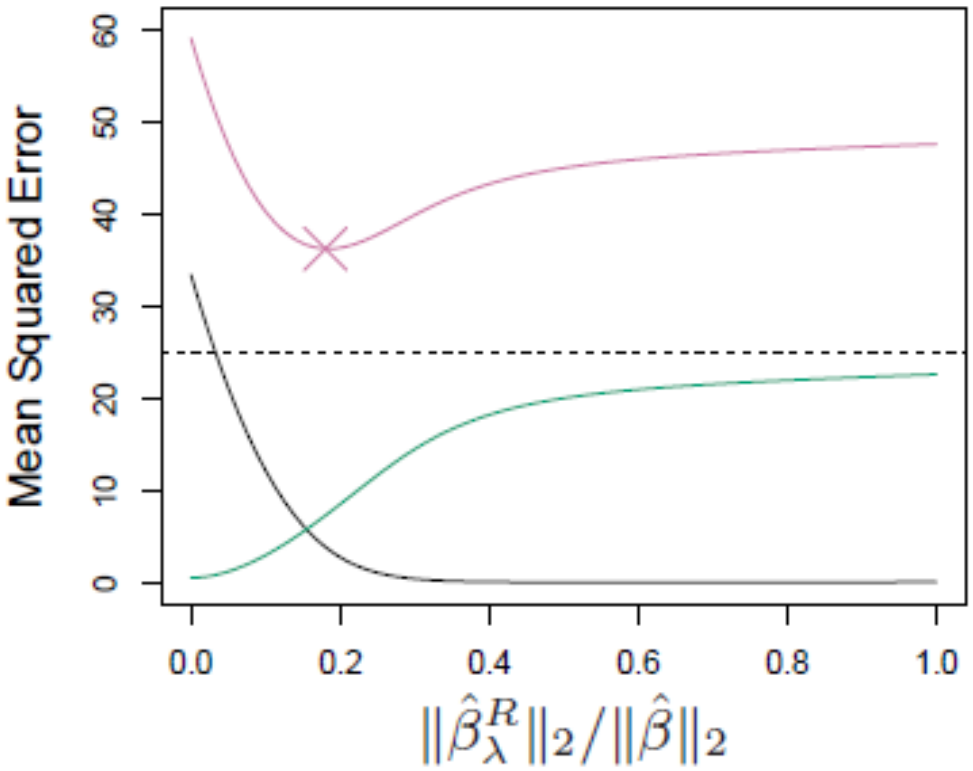
$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}}$$

Ridge Regression

p=45 predictors
n= 50 observations
Simulated Data

Mean Squared Error (MSE)





Variance of the ridge regression predictions

Consequently, the MSE drops considerably as λ increases from 0 to 10.

Ridge Regression

- When the number of variables p is almost as large as the number of observations n, the least squares estimates will be extremely variable.
- And if p > n, then the least squares estimates do not even have a unique solution, whereas ridge regression can still perform well by trading off a small increase in bias for a large decrease in variance.
  - Hence, ***ridge regression works best in situations where the least squares estimates have high variance***.
- Ridge regression also has substantial computational advantages over best subset selection

2. Shrinkage Models - Lasso

Ridge

Lasso

The Lasso

- Ridge regression will include all p predictors in the final model.
- The penalty $\lambda \sum_{j=1}^{p} \beta_j^2$ will shrink all of the coefficients towards zero, but it will not set any of them exactly to zero (unless $\lambda = \infty$).
  - This may not be a problem for prediction **accuracy**,
  - But it can create a challenge in model **interpretation** in settings in which the number of variables p is quite large.
- Increasing the value of λ will tend to reduce the magnitudes of the coefficients, but will not result in exclusion of any of the variables.

The Lasso

- The ***lasso*** is a relatively recent alternative to ridge regression.
- The lasso coefficients, $\hat{\beta}_{\lambda}^{L}$, minimize the quantity

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

Lasso Penalty

- The lasso uses an l_1.
- The l_1 norm of a coefficient vector β is given by $\left||\beta|\right|_1 = \sum|\beta_j|$

The Lasso

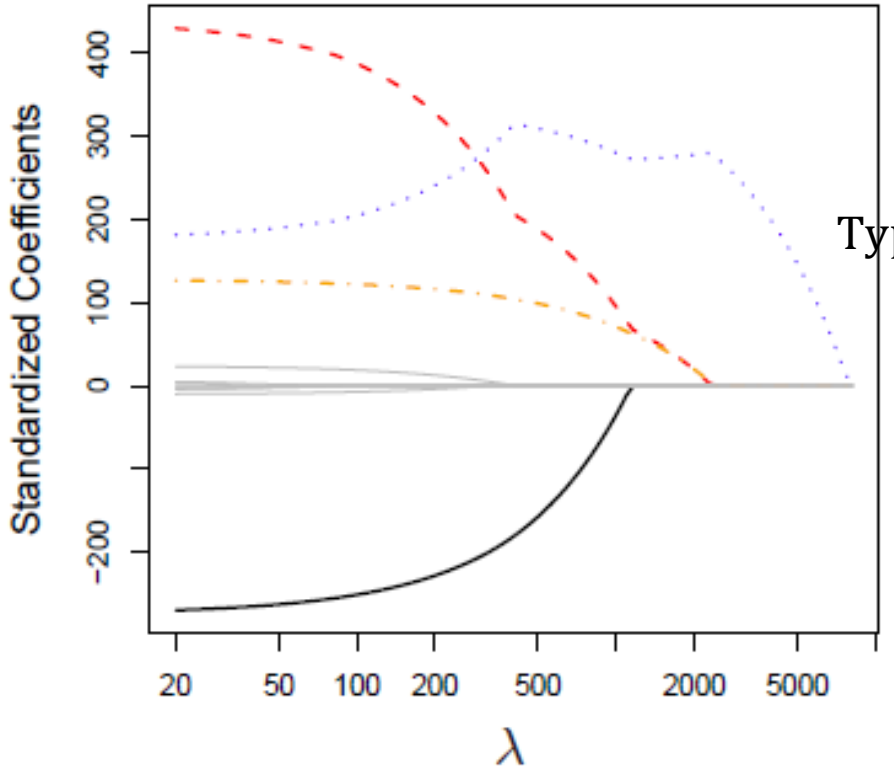
- The lasso also shrinks the coefficient estimates towards zero.

  - Hence, much like best subset selection, the lasso performs variable selection.
  - The lasso are generally much easier to interpret than those produced by ridge regression.
- The lasso yields ***sparse models,*** models that involve only a subset of the variables.

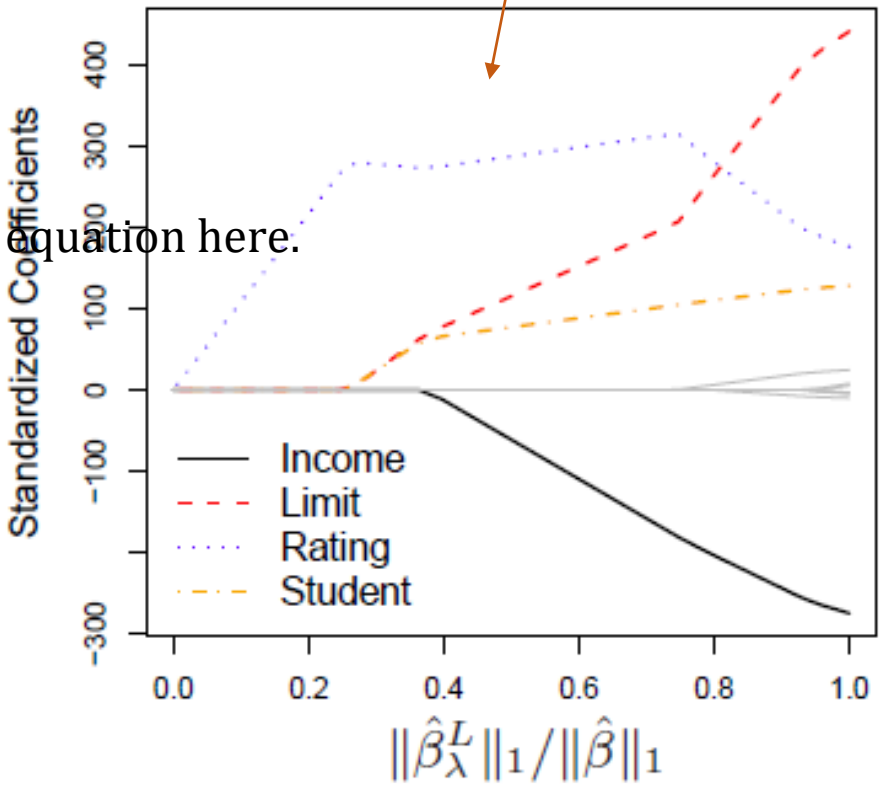
The Lasso

Example: Credit data set

Depending on the value of **λ**, the lasso can produce a model involving any number of variables.



Type equation here.



Another Formulation for Ridge Regression and the Lasso

- Lasso

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

budget

- Ridge

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

Comparing the Lasso and Ridge Regression

- Lasso produces simpler and more interpretable models that involve only a subset of the predictors.
- Neither ridge regression nor the lasso will universally dominate the other.
  - In general, one might expect the lasso to perform better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero.
  - Ridge regression will perform better when the response is a function of many predictors, all with coefficients of roughly equal size.
- A technique such as cross-validation can be used in order to determine which approach is better on a particular data set.

Selecting the Tuning Parameter

- We choose a grid of λ values.
- And compute the cross-validation error for each value of λ.
- 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.

Dimension Reduction Methods

Dimension Reduction Methods

- We now explore squares model using the transformed variables.
  - These techniques are ***dimension reduction methods***.
- Let $Z_1, Z_2, \ldots, Z_M$ represent $M < p$ linear combinations of our original p predictors. That is,

$$Z_m = \sum_{j=1}^{p} \emptyset\, X_j$$

for some constants $\emptyset_{1\mathrm{m}}, \emptyset_{2m}, \ldots, \emptyset_{pm}, m = 1, \ldots, M$.

Dimension Reduction Methods

- We can then fit the linear regression model

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \qquad i = 1, , \dots , n$$

using least squares.

- The term ***dimension reduction*** comes from the fact that this approach reduces the problem of estimating the p+1 coefficients $\beta_0, \beta_1, \dots, \beta_p$ to the simpler problem of estimating the M + 1 coefficients $\theta_0, \theta_1, \dots, \theta_M$, where M < p.

Dimension Reduction Methods

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \emptyset_{jm} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \emptyset_{jm} x_{ij}$$

$$= \sum_{j=1}^{p} \beta_j x_{ij}$$

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

Dimension Reduction Methods

- This constraint on the form of the coefficients has the potential to bias the coefficient estimates.
- However, in situations where p is large relative to n, selecting a value of $M \ll p$ can significantly reduce the variance of the fitted coefficients.
- If $M = p$, and all the Z_m are linearly independent, then the function in the previous slide poses no constraints.

Dimension Reduction Methods

- All dimension reduction methods work in two steps.

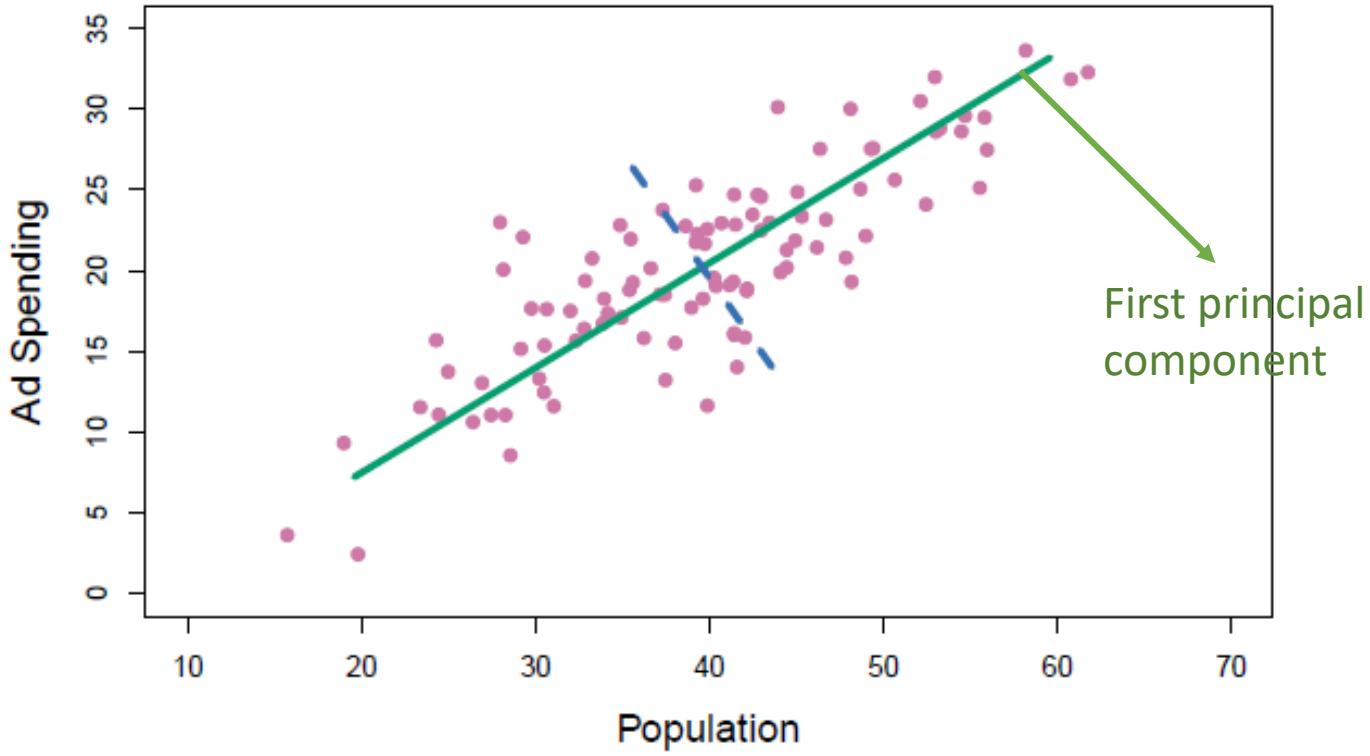
1. the transformed predictors $Z_1, Z_2, \ldots, Z_M$ are obtained.
2. The model is fit using these M predictors.

- The choice of $Z_1, Z_2, \ldots, Z_M$, or equivalently, the selection of the $\emptyset_{jm}$'s, can be achieved in different ways. The approaches:
  - Principal components
  - Partial least squares

Principal Components Regression

- A popular approach for deriving a low-dimensional set of features from a large set of variables.
  - PCA is a technique for reducing the dimension of a $n \times p$ data matrix X.
  - The first principal component direction of the data is that along which the observations vary the most.

Principal Components Regression



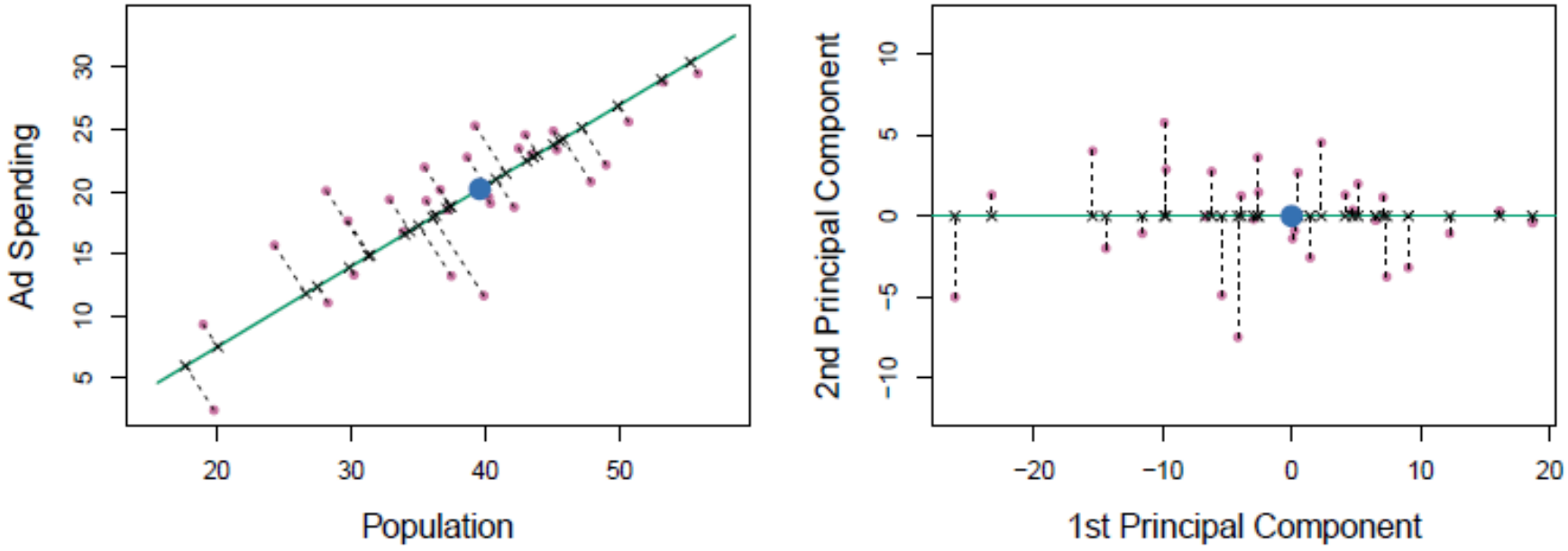
Mathematical represantation of the first principal component:

$$Z_1 = 0.839 \times (pop - \overline{pop}) + 0.544 \times \left(ad - \overline{ad}\right)$$

$$\emptyset_{11} = 0.839 \; ve \; \emptyset_{21} = 0.544$$

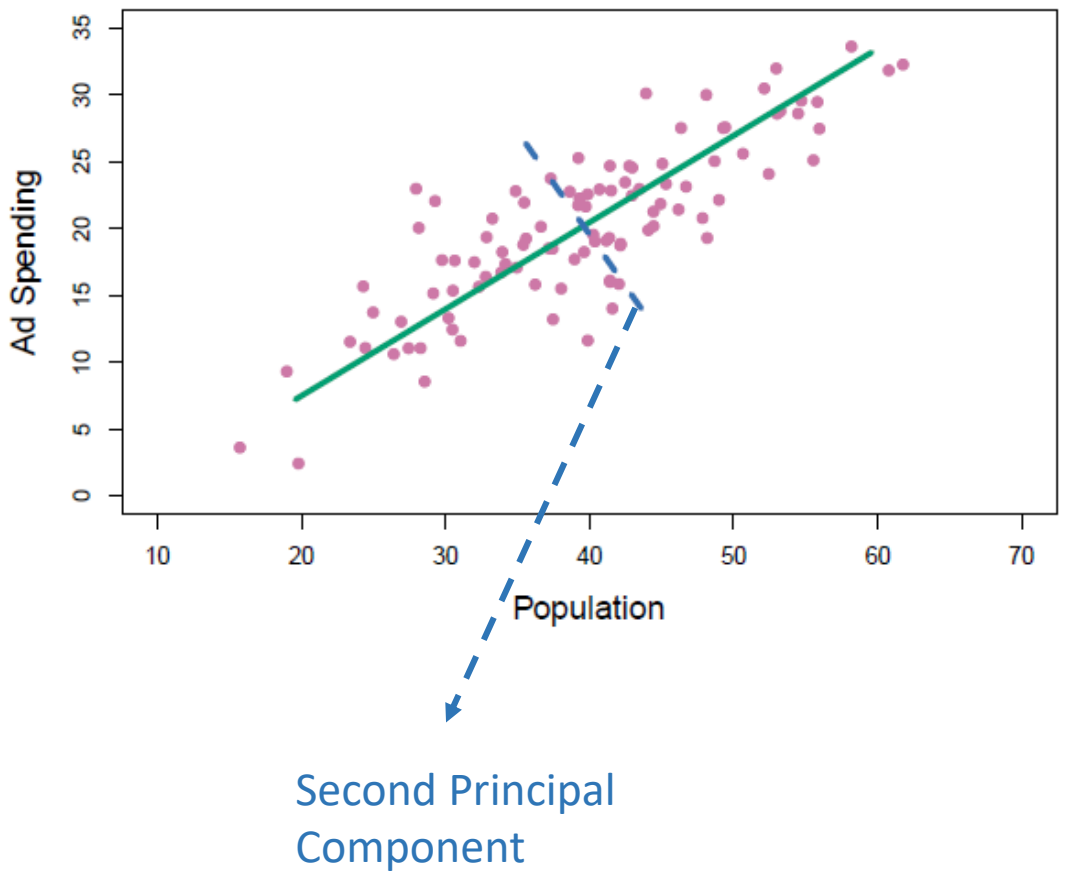
Principal Components Regression

- The crosses represent the projection of each point onto the first principal component line.



Principal Components Regression

- In general, one can construct up to p distinct principal components.
  - The second principal component Z_2 is a linear combination of the variables that is uncorrelated with Z_1, and has largest variance subject to this constraint.

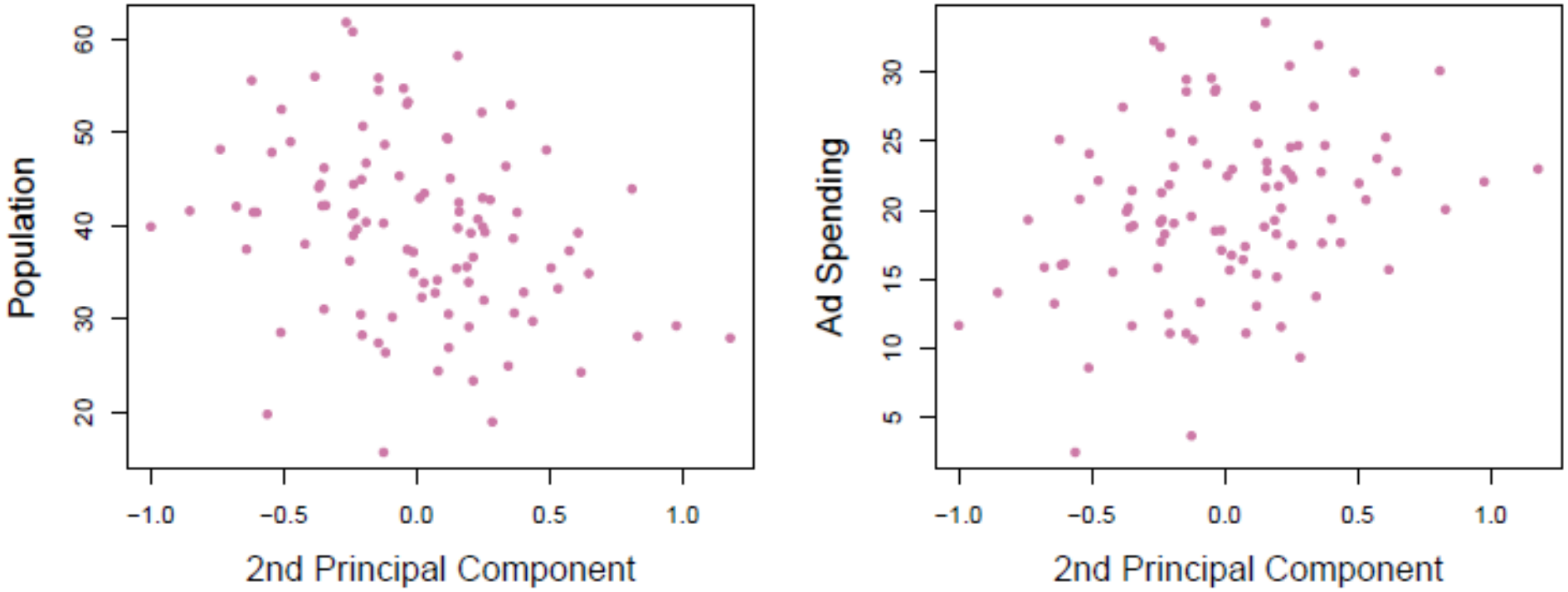


Principal Components Regression

- It turns out that the zero correlation condition of Z_1 with Z_2 is equivalent to the condition that the direction must be ***perpendicular***, or ***orthogonal***, to the first principal component direction.

Principal Components Regression

Second Principal Component Results – Advertising Data



$$Z_2 = 0.544 \times (pop - \overline{pop}) - 0.839 \times (ad - \overline{ad})$$

Principal Components Regression

- With two-dimensional data, we can construct at most two principal components.
- If we had other predictors, then additional components could be constructed.
  - They would successively maximize variance, subject to the constraint of being uncorrelated with the preceding components.

The Principal Components Regression Approach

- The principal components regression (PCR)
  - involves constructing the first M principal components, $Z_1, \ldots, Z_M$,
  - then using these components as the predictors in a linear regression model that is fit using least squares.

- The key idea is that
  - often a small number of principal components suffice to explain most of the variability in the data, as well as the relationship with the response.