COM 599200 – Statistical Learning

Lecture 6 – Linear Model Selection and Regularization

Y.-W. Peter Hong

Linear Model Selection and Regularization

- In regression, the standard linear model is

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

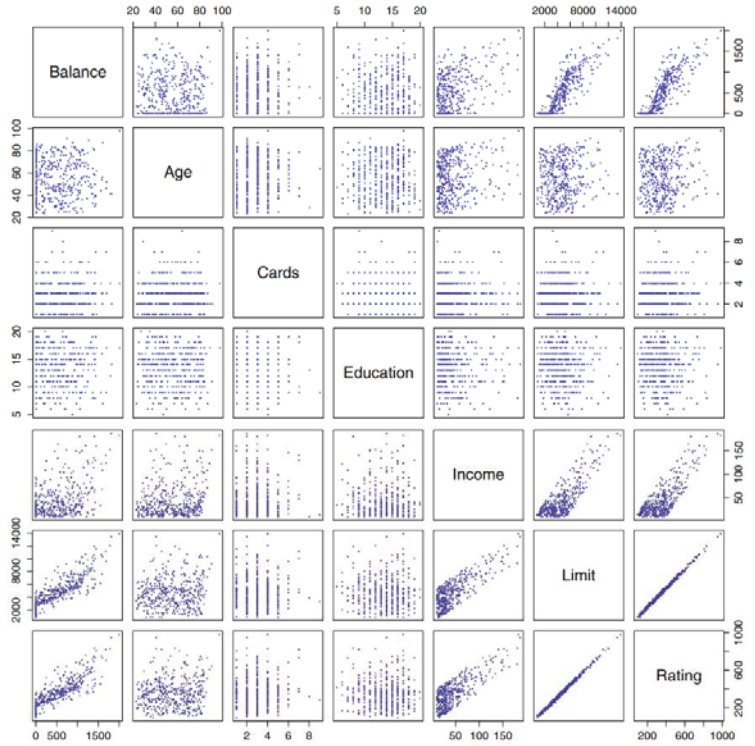
- What if p is close to (or greater) than n?

  ➔ Results in large variance due to overfitting.
- Which are the relevant and irrelevant predictors?

  ➔ Important for model interpretability.
- Alternative or modified least squares approaches:
  - Subset Selection
  - Shrinkage (or Regularization)
  - Dimension Reduction (via Linear Transformation)

Motivating Example

Example (Credit Card Data Set):



Subset Selection

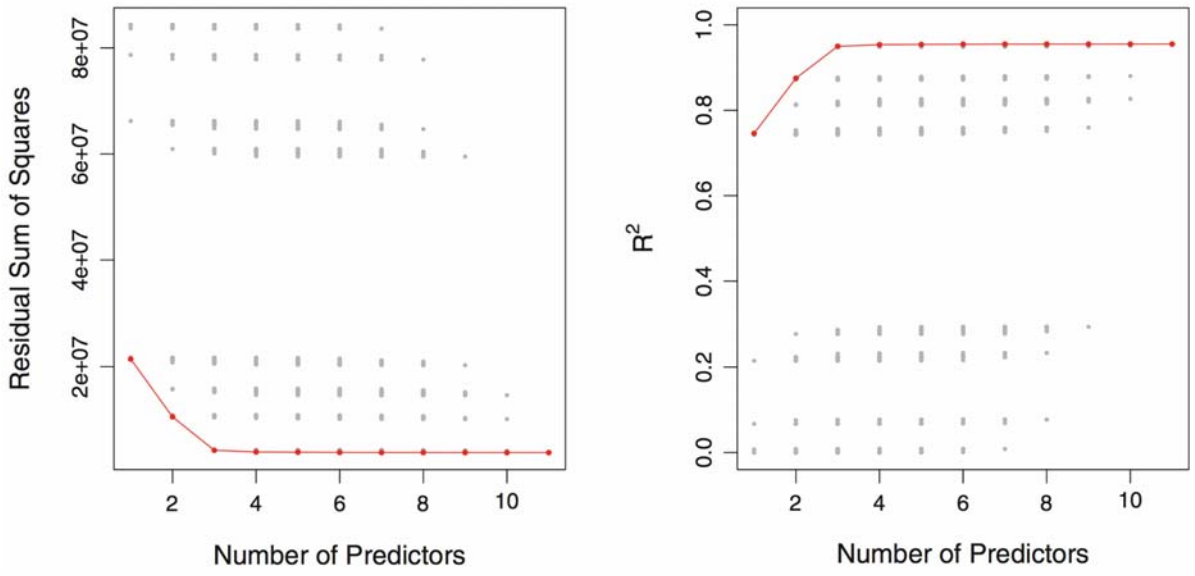
- Subset selection involves identifying a subset of the p predictors that are related to the response.
- **Best Subset Selection:** Try fitting to all 2^p possible models and choose the "best" one.

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.

Example: Credit Card Data Set

Example (Credit Card Balance Prediction):



➔ Same can be applied for logistic regression using *deviance* $-2\log \Pr(\mathbf{y}|\mathbf{X};\beta)$ as the performance measure.

(Forward) Stepwise Selection (1/2)

- **Forward Stepwise Selection:** Start from 0 predictors, and, in each step, add the variable that yields the greatest additional improvement to the fit.

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.

(Forward) Stepwise Selection (2/2)

- Best subset selection requires fitting to 2^p models, whereas forward stepwise selection requires fitting to $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$.

  ➔ For $p=20$, we have $1,048,576(\text{best})$ vs $211(\text{forward})$.

Example (Credit Card Balance):

# 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

Backward Stepwise Selection

- **Backward Stepwise Selection:** Starts with all predictors and, in each step, removes the one that is least useful. (➔ requires $n > p$)

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.

- **Hybrid Forward and Backward Selection**

Choosing the Optimal Model

- Training set MSE (or RSS) and R^2 cannot be used for model selection since they always improve with the addition of input variables.
- Two approaches:
  1) Adjustment to the training error to account for the bias due to overfitting.
  2) Directly estimate test error using cross-validation.
- 4 common approaches for 1) include C_p, *Akaike information criterion (AIC)*, *Bayesian information criterion (BIC)*, and *adjusted* R^2.

Adjustment to Training Error (C_p & AIC)

- Mallow's C_p is defined as

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

where $\hat{\sigma}^2$ is an estimate of the variance of error ϵ (using the full model containing all predictors), and d is the number of predictors selected in the model.

➔ Alternatively, we can define

$$C'_p = \frac{\text{RSS}}{\hat{\sigma}^2} + 2d - n$$

- Akaike Information Criterion (AIC) is defined as

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

Training Error

- Recall that the training error is

$$\overline{\text{err}} = \frac{1}{n}\sum_{i=1}^{n} L\big(y_i, \hat{f}(x_i;\mathcal{D})\big)\left(= \frac{1}{n}\sum_{i=1}^{n} \big(y_i - \hat{f}(x_i;\mathcal{D})\big)^2\right)$$

  where $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is the training data.

➔ $\overline{\text{err}}$ is usually less than test error $\text{Err} = E[L(Y, \hat{f}(X;\mathcal{D}))]$.

In-Sample Error (1/2)

- Let us define the *in-sample error* as

$$\text{Err}_{\text{in}} = \frac{1}{n}\sum_{i=1}^{n} E_{\mathbf{y}}\Big[E_{\mathbf{y}^{\text{new}}}\big[L(y_i^{\text{new}}, \hat{f}(x_i;\mathcal{D}))\big]\Big].$$

- Training error under-estimates in-sample error by

$$\text{Err}_{\text{in}} - E_{\mathbf{y}}[\overline{\text{err}}] =$$

In-Sample Error (2/2)

$$\Rightarrow \ \mathrm{Err}_{\mathrm{in}} = E_{\mathbf{y}}[\overline{\mathrm{err}}] + \frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}(y_i, \hat{y}_i).$$

- For $Y = f(X) + \epsilon$ and $\hat{\mathbf{y}}$ that is obtained by a linear fit with d inputs, i.e., $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. We have

$$\sum_{i=1}^{n}\mathrm{Cov}(y_i, \hat{y}_i) =$$

Thus, $\mathrm{Err}_{\mathrm{in}} = E_{\mathbf{y}}[\overline{\mathrm{err}}] + \frac{2d}{n}\sigma^2.$

➔ Recall that $C_p = \overline{\mathrm{err}} + \frac{2d}{n}\hat{\sigma}^2 = \frac{1}{n}(\mathrm{RSS} + 2d\hat{\sigma}^2)$

Effective Degrees of Freedom

- For a general linear fit $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$, the in-sample error is

$$\mathrm{Err}_{\mathrm{in}} = E_{\mathbf{y}}[\overline{\mathrm{err}}] + \frac{2}{n}\mathrm{tr}(\mathbf{S})\sigma^2.$$

- In this case, we can define

$$d(\mathbf{S}) = \mathrm{tr}(\mathbf{S})$$

  as the effective number of parameters, or the effective degrees of freedom.

- Hence, for a general linear fit, the C_p-statistic can be defined as

$$C_p = \frac{1}{n}(\mathrm{RSS} + 2\mathrm{tr}(\mathbf{S})\hat{\sigma}^2).$$

AIC as Estimate of In-Sample Error (1/2)

- Akaike Information Criterion (AIC) is similarly an estimate of $\mathrm{Err}_{\mathrm{in}}$ when a log-likelihood loss function is considered, i.e., when

$$L(y, \hat{p}(x;\beta)) = -2\log \Pr(y|x;\beta)$$

$$\Big(= -2[y\log\hat{p}(x;\beta) + (1-y)\log(1-\hat{p}(x;\beta))], \text{ for } y \in \{0,1\}\Big).$$

- Asymptotically, as $N \to \infty$, it can be shown that

$$\mathrm{Err}_{\mathrm{in}} = \frac{-2}{n}\sum_{i=1}^{n} E_{\mathbf{y}}\Big[E_{\mathbf{y}^{\mathrm{new}}}\big[\log \Pr(y_i^{\mathrm{new}}|x_i;\hat{\beta}(\mathbf{X},\mathbf{y}))\big]\Big]$$

$$\approx \frac{-2}{n}\sum_{i=1}^{n} E\big[\log \Pr(y_i|x_i;\hat{\beta}(\mathbf{X},\mathbf{y}))\big] + \frac{2d}{n}.$$

AIC as Estimate of In-Sample Error (2/2)

- For linear model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$, we have

Adjustment to Training Error (BIC & Adj. R^2)

- Bayesian Information Criterion (BIC) is defined as

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

➔ Heavier penalty on models with more variables.

- The adjusted R^2 is defined as

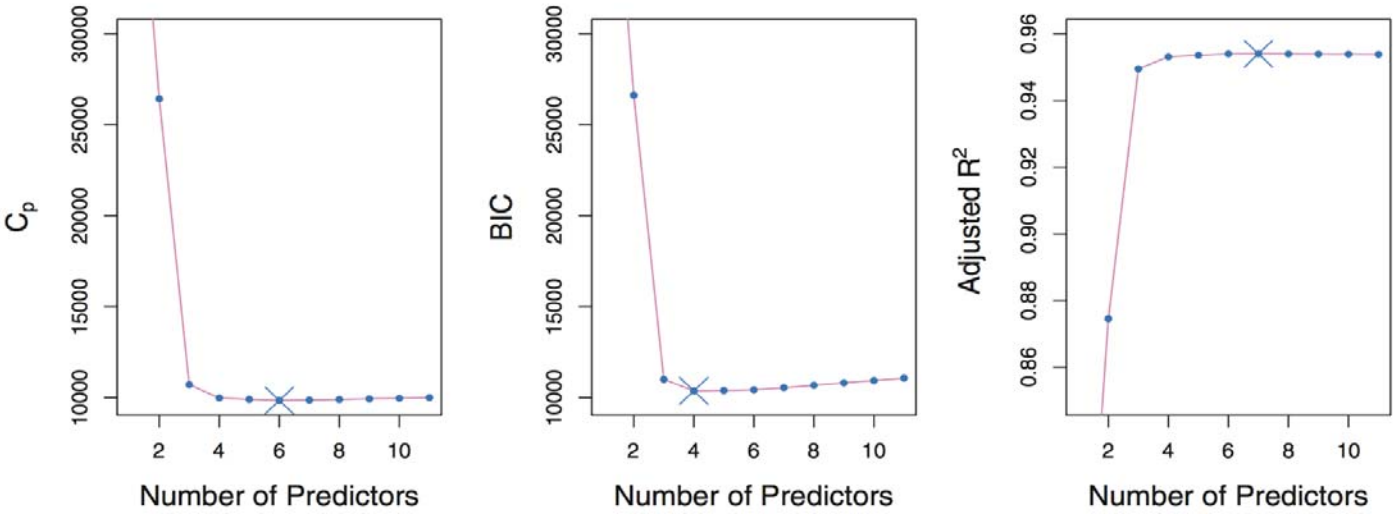
$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}.$$

➔Compared to R^2, adjusted R^2 pays the price for the inclusion of unnecessary variables.

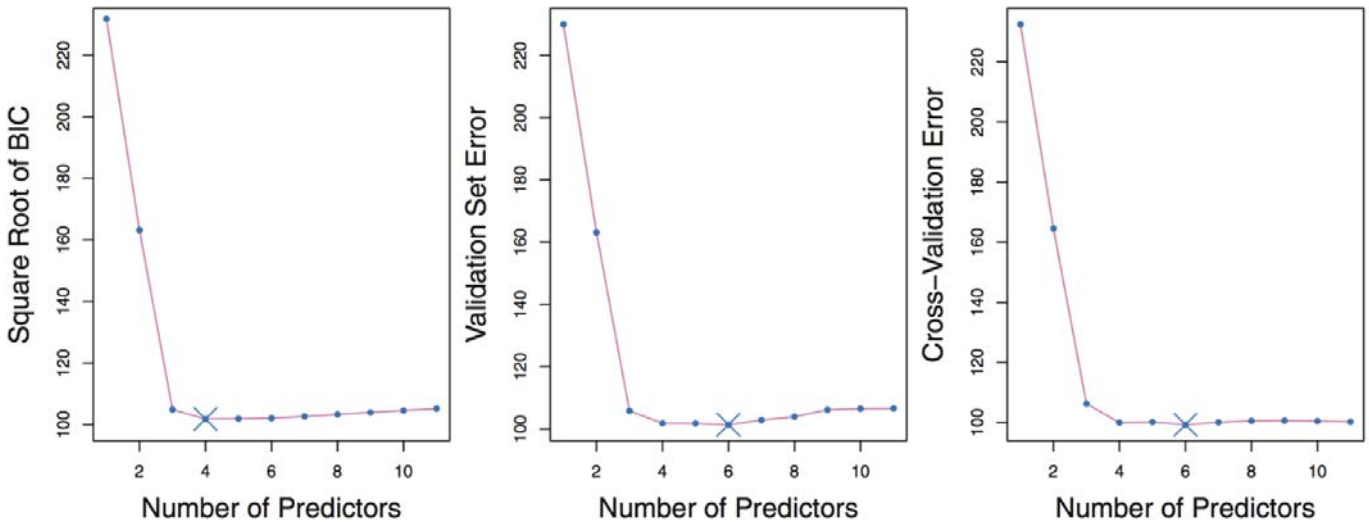
Comparing Different Criteria (1/2)

Example (Credit Card Data Set):

➔ Best subset selection



Comparing Different Criteria (2/2)



➔The one standard error rule (with respect to the estimated test MSE) results in the 3-variable model for validation set and CV approaches.

Shrinkage Methods – Ridge Regression (1/2)

- Recall that the least squares solution minimizes

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

- **Ridge regression** chooses coefficients to minimize

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

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

➔ $\lambda\sum_{j=1}^{p}\beta_j^2$ is the shrinkage penalty for $\beta_1, \ldots, \beta_p$.

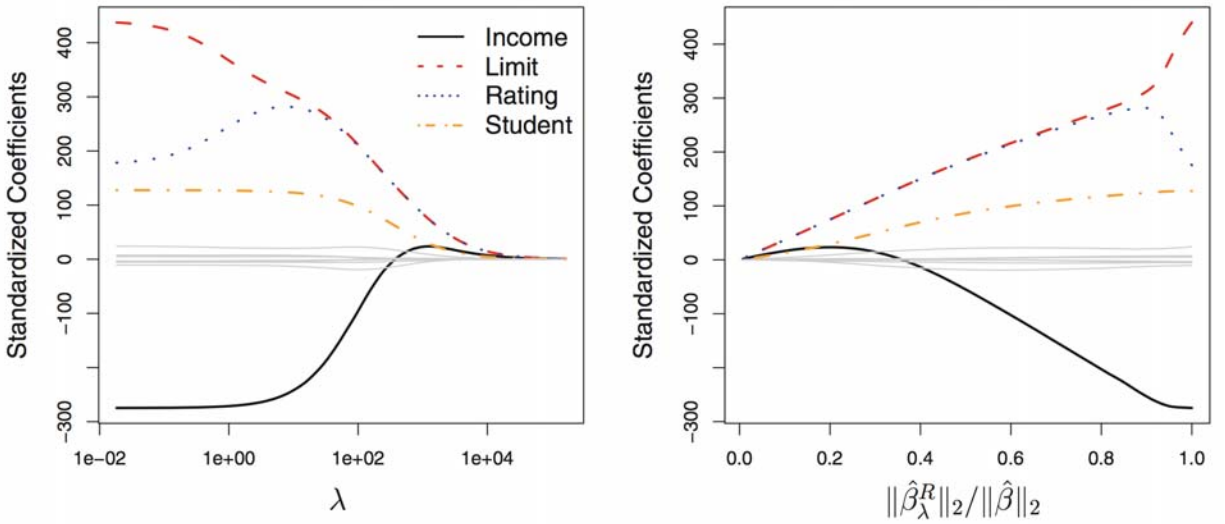
Remark: Ridge regression fit varies with the scale of x_{ij}.

Ridge Regression Example

➔Apply ridge regression to standardized predictors

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

Example: (Credit Card Data Set):



Ridge Regression Solution

- Notice that the solution for β_0 is

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right) = \bar{y} - \sum_{j=1}^{p}\beta_j \bar{x}_j.$$

- Hence, the problem reduces to

$$\sum_{i=1}^{n}\left[(y_i - \bar{y}) - \sum_{j=1}^{p}\beta_j(x_{ij} - \bar{x}_j)\right]^2 + \lambda\sum_{j=1}^{p}\beta_j^2.$$

- By taking centered inputs and outputs, the problem reduces to choosing $\beta = (\beta_1, \ldots, \beta_p)^T$ that minimizes

$$\mathrm{RSS}(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta$$

  where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$.

➔ The ridge regression solution is $\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$.

Examining the Solution (1/3)

- Let us take the singular value decomposition (SVD) of the centered input matrix $\mathbf{X}$ to yield

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

  where $\mathbf{U}$ and $\mathbf{V}$ are $n \times p$ and $p \times p$ orthogonal matrices, and $\mathbf{D}$ is a $p \times p$ diagonal matrix of singular values $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$.

- In this case, the fitted value for $\mathbf{y}$ is

Examining the Solution (2/3)

- By the eigenvalue decomposition

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T,$$

  we can observe that d_j^2 is an eigenvalue of $\mathbf{X}^T\mathbf{X}$ corresponding to eigenvector v_j.

- The projection of $\mathbf{X}$ onto the direction v_j yields coordinate values $\mathbf{z}_j = \mathbf{X}v_j$ that has sample variance

$$\frac{1}{n}\sum_{i=1}^{n} z_{ij}^2 = \frac{1}{n}\mathbf{z}_j^T\mathbf{z}_j = \frac{1}{n}v_j^T\mathbf{X}^T\mathbf{X}v_j = \frac{d_j^2}{n}.$$

  ➔ v_j is the j-th principal component direction of $\mathbf{X}$.

Examining the Solution (3/3)

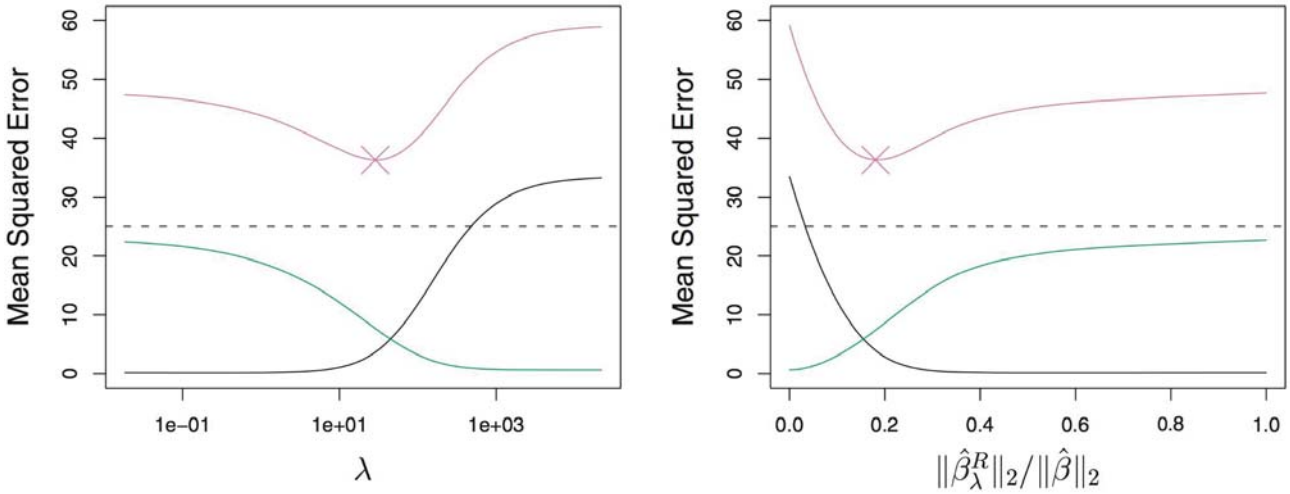
- The effective degrees of freedom is

$$\mathrm{df}(\lambda) = \mathrm{tr}\big(\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\big) = \sum_{j=1}^{p}\frac{d_j^2}{d_j^2 + \lambda}.$$

Bias-Variance Tradeoff for Ridge Regression

Example:

- Simulated data with $p = 45$ predictors, $n = 50$ observations.



➔ Ridge regression works best in cases where the least squares estimates have high variance (e.g., when p is close to or $> n$).

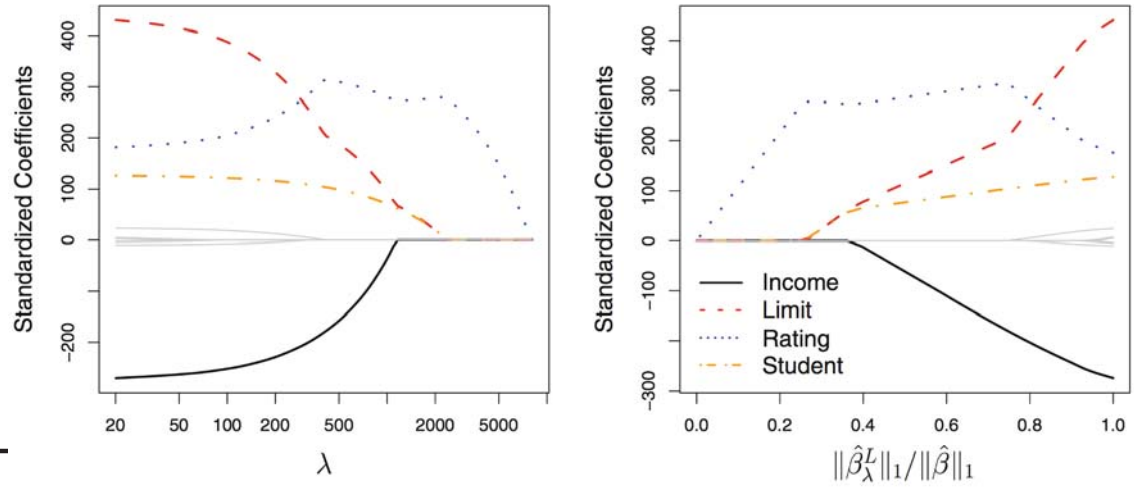
Shrinkage Methods – Lasso

- **The lasso method** chooses coefficients to minimize

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

➔ Lasso uses an ℓ_1-penalty $\lambda\|\beta\|_1$ (instead $\lambda\|\beta\|_2$).

➔ ℓ_1-penalty is more likely to force coefficients to 0.



Alternative Formulation (1/3)

- Ridge regression for a certain $\lambda \geq 0$ is equivalent to solving the following optimization problem

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

for some s. Or, for centered input and outputs

$$\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \text{ subject to } \beta^T \beta \leq s.$$

  - Case 1 ($\hat{\beta}_{\text{lin}}^T \hat{\beta}_{\text{lin}} \leq s$): In this case, $\hat{\beta}_{\text{rid}} = \hat{\beta}_{\text{lin}}$.
  - Case 2 ($\hat{\beta}_{\text{lin}}^T \hat{\beta}_{\text{lin}} > s$): In this case, we must have $\hat{\beta}_{\text{rid}}^T \hat{\beta}_{\text{rid}} = s$. By the method of Lagrange multipliers, $\hat{\beta}_{\text{rid}}$ minimizes

$$\mathcal{L}(\beta) \triangleq \text{RSS}(\beta) + \lambda \beta^T \beta$$

for some $\lambda > 0$.

Alternative Formulation (2/3)

Alternative Formulation (3/3)

- Similarly, the lasso for a certain $\lambda \geq 0$ is equivalent to solving the following optimization problem

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

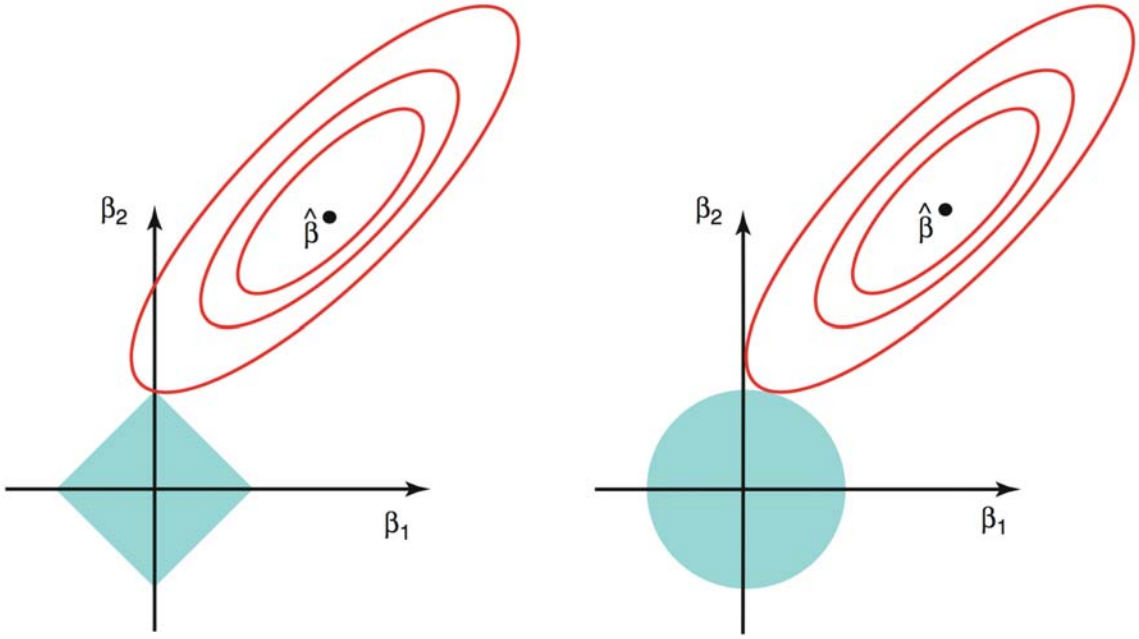
  for some s.

- **Remark:** The best subset selection can also be formulated as

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

Insights for Lasso

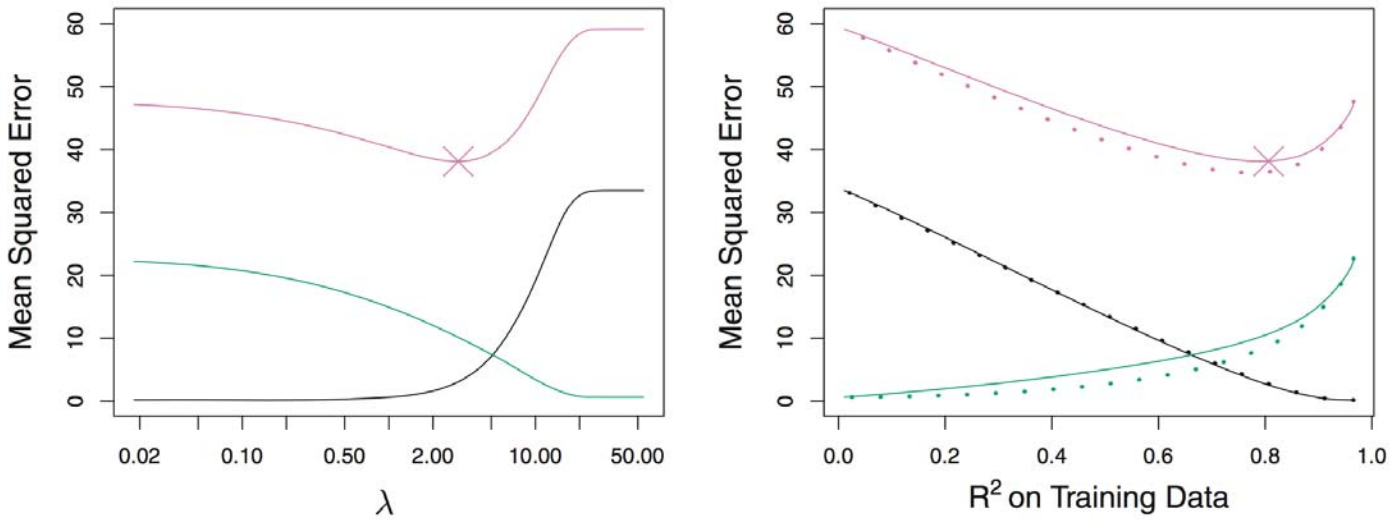
Question: Why is lasso more likely to result in coefficient estimate that are exactly equal to zero?



Ridge Regression vs Lasso (1/2)

Example:

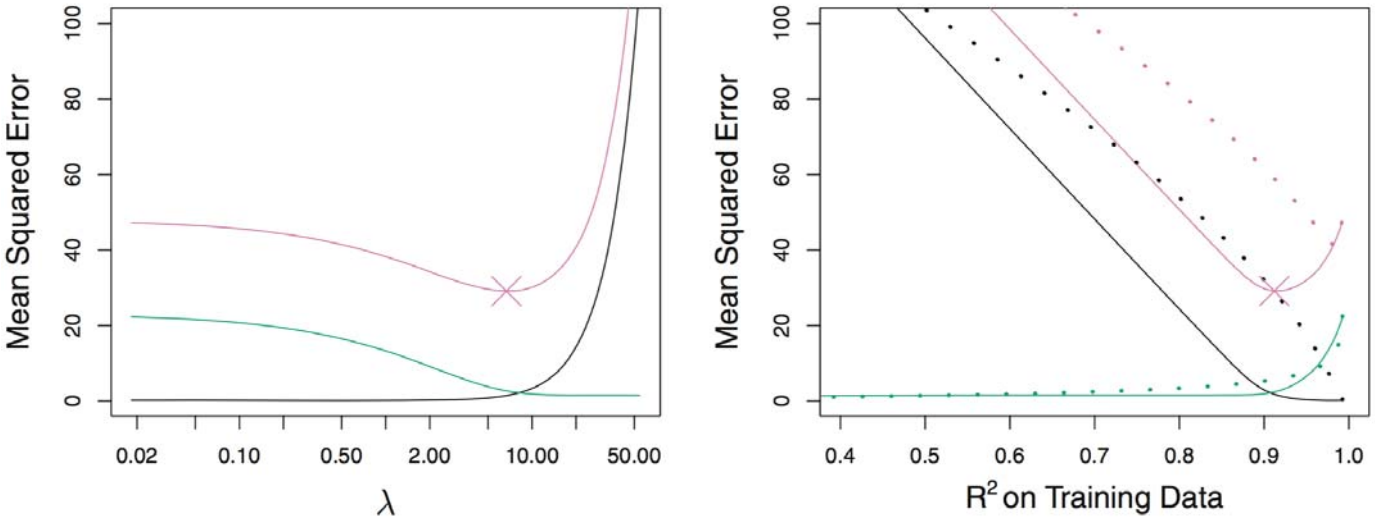
- Data generated using all 45 predictors.



Ridge Regression vs Lasso (2/2)

Example:

- Data generated using only 2 out of 45 predictors.



Bayesian Interpretation (1/3)

- Let $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$ be the underlying linear model, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.
- Also, let $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ (or $\{\mathbf{X}, \mathbf{y}\}$) be the available data set, where

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I}).$$

- Least squares linear regression is equivalent to finding β that maximizes the likelihood function

$$f(\mathbf{y}|\mathbf{X}, \beta) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(\frac{-1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\beta\|^2\right).$$

- That is,

Bayesian Interpretation (2/3)

- Suppose that β has prior distribution $f(\beta)$. Then, we can choose β to maximize the posterior distribution

$$f(\beta|\mathbf{y}, \mathbf{X}) = \frac{f(\mathbf{y}|\mathbf{X}, \beta) f(\beta)}{f(\mathbf{y}|\mathbf{X})}.$$

- For $\beta = (\beta_0, \ldots, \beta_p)$ that are i.i.d. $\mathcal{N}(0, \gamma)$, we have

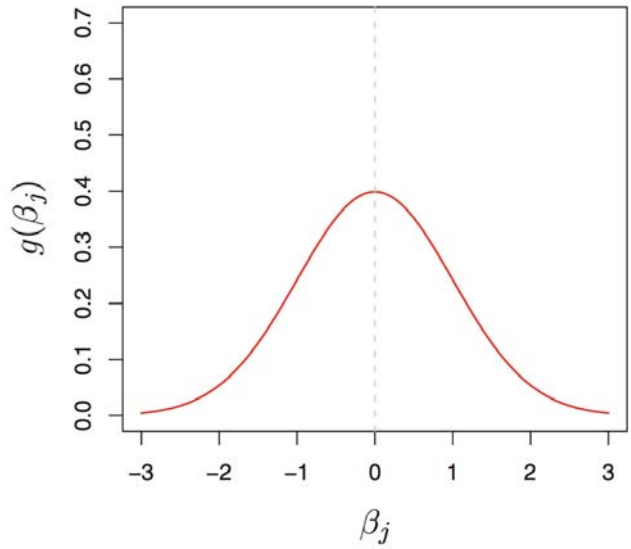
$$\hat{\beta} = \arg\max_{\beta} f(\beta|\mathbf{y}, \mathbf{X})$$

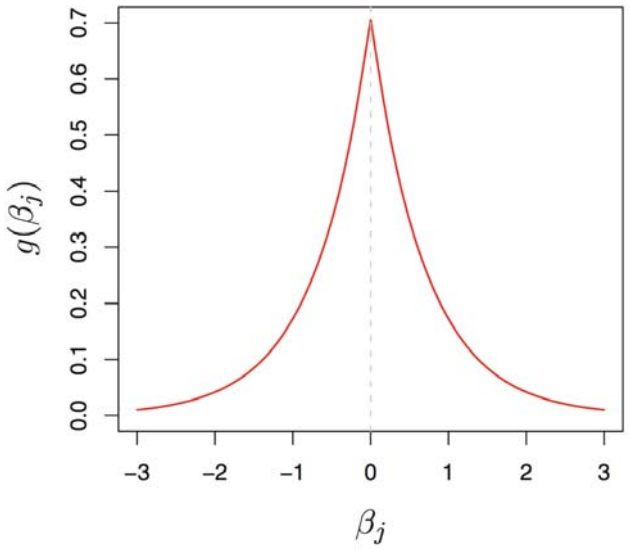
Bayesian Interpretation (3/3)

- For $\beta = (\beta_0, \ldots, \beta_p)$ that are i.i.d. Laplace distributed with mean 0 and scalar parameter γ, we have

$$\hat{\beta} = \arg\max_{\beta} f(\beta|\mathbf{y}, \mathbf{X})$$

Gaussian vs Laplace Distribution

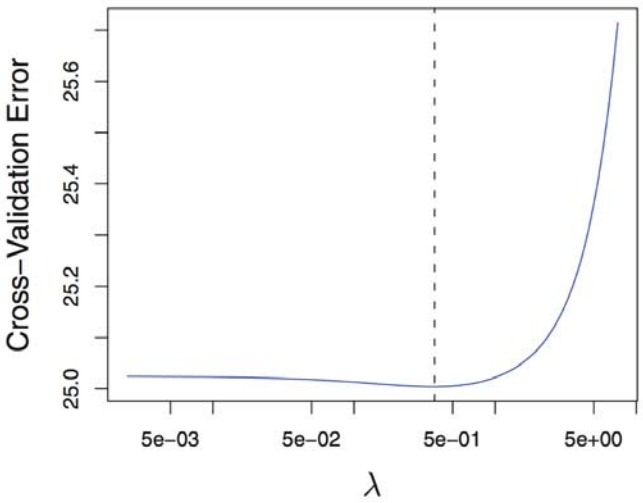


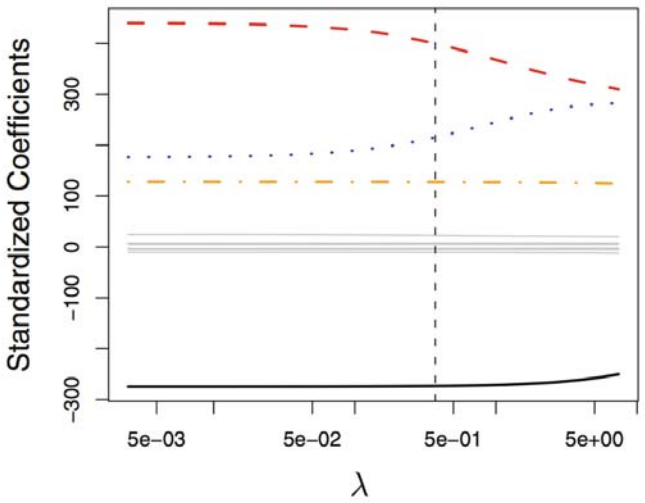


Selecting the Tuning Parameters (1/2)

- Choose a grid of λ values, and compute their CV errors.
- Select the value of λ that yields the smallest CV error.
- Refit model using all observations and the selected parameter.

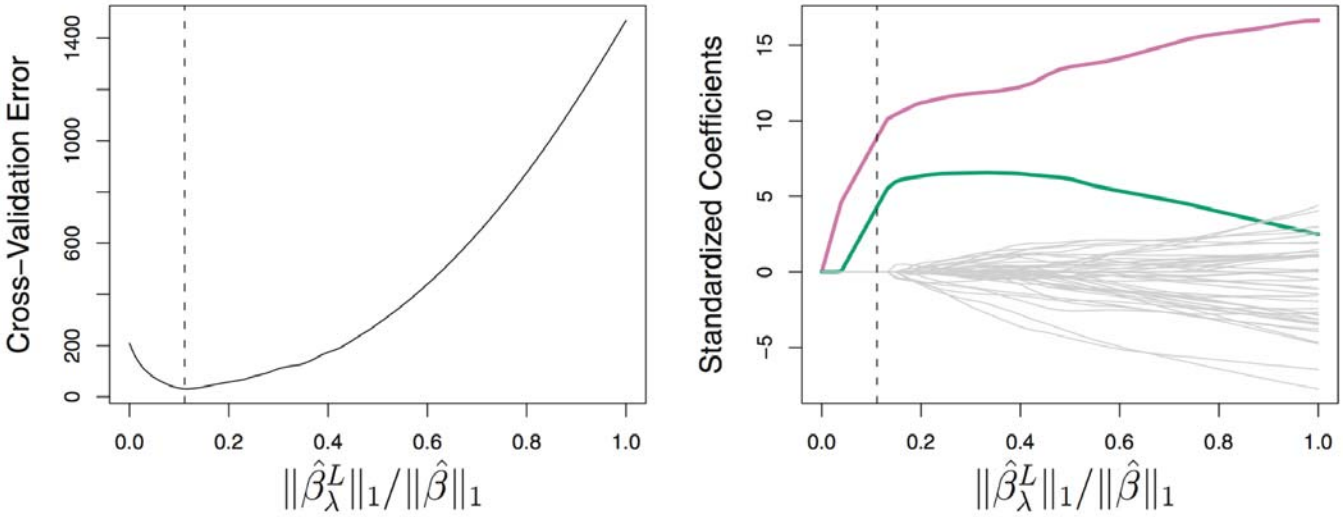
Example (Ridge Regression on Credit Data Set):





Selecting the Tuning Parameters (2/2)

Example (Lasso on Sparse Simulated Data):



Dimension Reduction Methods (1/2)

- Dimension reduction methods fit models to variables transformed from the original predictors $X_1, \ldots, X_p$.
- Let $Z_1, \ldots, Z_M$ be $M(< p)$ linear combinations of original predictors $X_1, \ldots, X_p$, i.e.,

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j, \text{ for } m = 1, \ldots, M.$$

➔ Observations: $z_{im} = \sum_{j=1}^{p} \phi_{jm} x_{ij}, \text{ for } i = 1, \ldots, n.$

- With the transformed predictors, we then find linear regression coefficients $\theta_0, \ldots, \theta_M$ to minimize

$$\sum_{i=1}^{n} \left(y_i - \theta_0 - \sum_{m=1}^{M} \theta_m z_{im} \right)^2.$$

Dimension Reduction Methods (2/2)

- Note that

$$\sum_{m=1}^{M} \theta_m z_{im} =$$

➔Dimension reduction methods are special cases of linear regression with coefficients

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}, \ j = 1, \ldots, p.$$

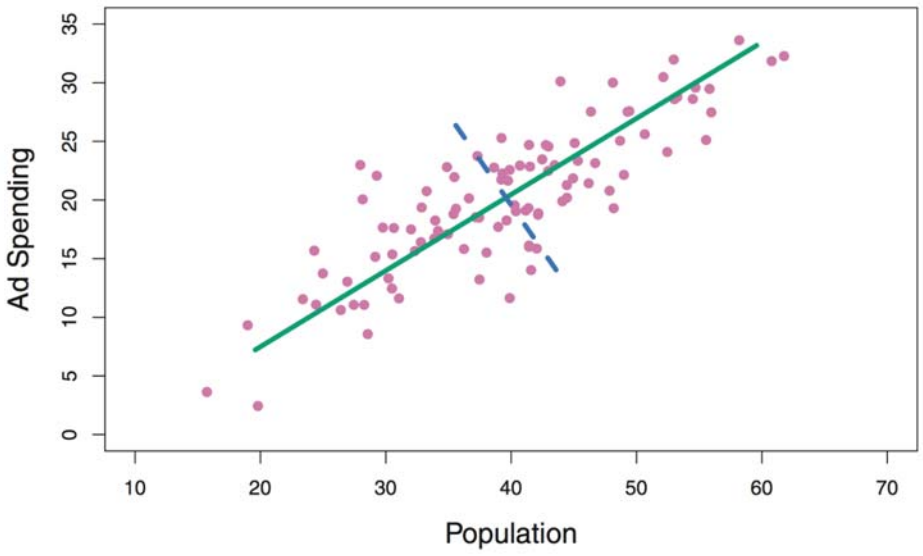
Question: How do we select transformation coefficients $\phi_{jm}, \ j = 1, \ldots, p, \ m = 1, \ldots, M$?

➔Principal Components Analysis (PCA)

➔Partial Least Squares (PLS)

Principal Components Analysis (1/5)

- **Principal components analysis (PCA)** transforms data points to a set of values of linearly uncorrelated variables called *principal components*.



$$\begin{matrix} (x_{11}, x_{12}) \\ (x_{21}, x_{22}) \\ \vdots \\ (x_{n1}, x_{n2}) \end{matrix} \quad \Rightarrow \quad \begin{matrix} (z_{11}, z_{12}) \\ (z_{21}, z_{22}) \\ \vdots \\ (z_{n1}, z_{n2}) \end{matrix}$$

Principal Components Analysis (2/5)

- Let

$$Z_1 = \phi_{11} \underbrace{(\text{pop} - \overline{\text{pop}})}_{X_1} + \phi_{21} \underbrace{(\text{ad} - \overline{\text{ad}})}_{X_2}$$

be the *first principal component*, where $\phi_1 = (\phi_{11}, \phi_{21})^T$ is chosen to preserve the most variability. That is,

$$\phi_1 = \arg \max_{\|\phi_1'\|^2=1} \text{Var}\left(\phi_{11}' X_1 + \phi_{21}' X_2\right)$$

$$\approx \arg \max_{\|\phi_1'\|^2=1} \frac{1}{n} \sum_{i=1}^{n} \left(\phi_{11}' x_{i1} + \phi_{21}' x_{i2}\right)^2.$$

Principal Components Analysis (3/5)

➔Here, $\phi_{11} = 0.839$ and $\phi_{21} = 0.544$.

➔ $z_{11}, z_{21}, \ldots, z_{n1}$, where $z_{i1} = 0.839 \cdot x_{i1} + 0.544 \cdot x_{i2}$ are the *(first) principal component scores*.

Principal Components Analysis (4/5)

- Let $\tilde{X} = (\tilde{X}_1, \tilde{X}_2)^T = X - \phi_1(\phi_1^T X)$ be the residual after subtracting out the first principal component.

➔Notice that $\tilde{X} = (\mathbf{I} - \phi_1\phi_1^T)X$ is the orthogonal projection of X onto the null space of ϕ_1.

- The second principal component is

$$Z_2 = \tilde{\phi}_{21}\tilde{X}_1 + \tilde{\phi}_{22}\tilde{X}_2 \left(= \tilde{\phi}_2^T\tilde{X} = \underbrace{\tilde{\phi}_2^T(\mathbf{I} - \phi_1\phi_1^T)}_{\phi_2^T}X\right)$$

with

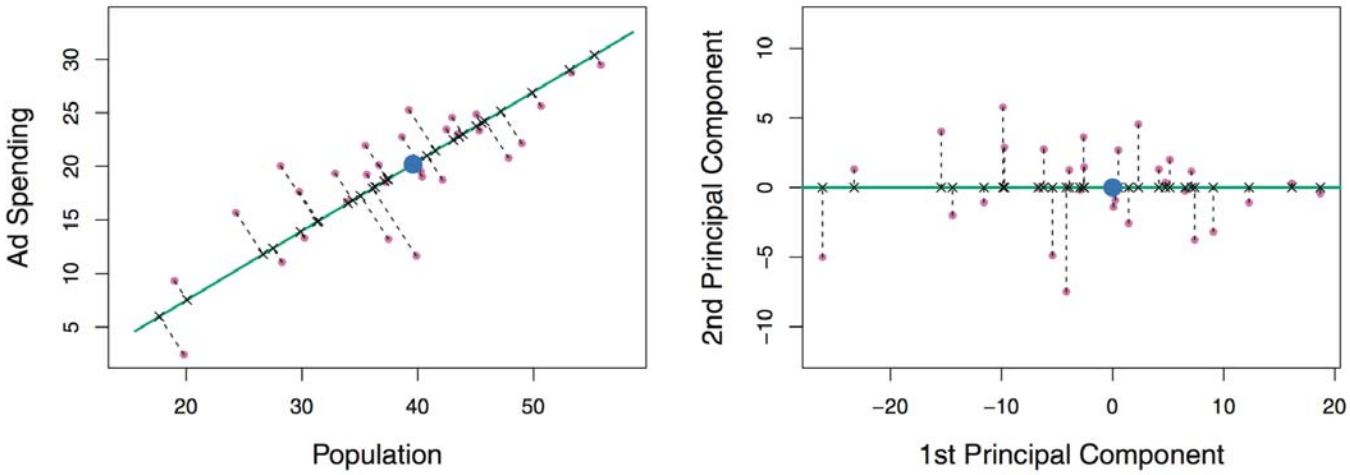
$$\tilde{\phi}_2 = \arg\max_{\|\tilde{\phi}_2'\|^2=1} \mathrm{Var}\left(\tilde{\phi}_{21}'\tilde{X}_1 + \tilde{\phi}_{22}'\tilde{X}_2\right)$$

$$\approx \arg\max_{\|\tilde{\phi}_2'\|^2=1} \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\phi}_{21}'\tilde{x}_{i1} + \tilde{\phi}_{22}'\tilde{x}_{i2}\right)^2.$$

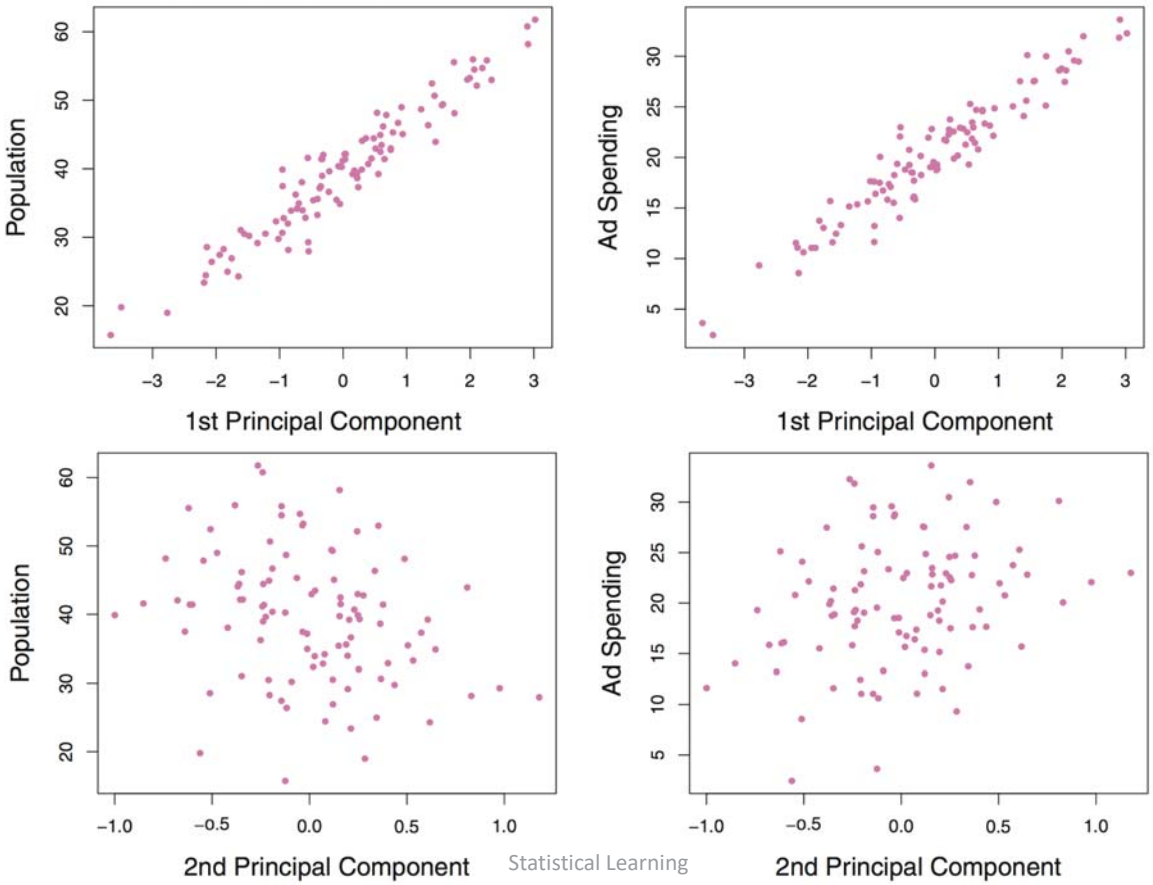
Principal Components Analysis (5/5)

➔ Here, $\phi_{21} = 0.544$ and $\phi_{22} = -0.839$.

Illustration of PCA



First and Second Principal Components



General PCA Procedure

- Let $\phi_\ell = (\phi_{1\ell}, \phi_{2\ell}, \ldots, \phi_{p\ell})^T$, for $\ell = 1, \ldots, k-1$, be the first $k-1$ principal component directions, and let

$$\tilde{X}^{(k)} = (\tilde{X}_1^{(k)}, \ldots, \tilde{X}_p^{(k)})^T = X - \sum_{\ell=1}^{k-1} \phi_\ell (\phi_\ell^T X).$$

- Then, the k-th principal component is

$$Z_k = \tilde{\phi}_k^T \tilde{X}^{(k)} = \tilde{\phi}_k^T \left(\mathbf{I} - \sum_{\ell=1}^{k-1} \phi_\ell \phi_\ell^T \right) X$$

where

$$\tilde{\phi}_k = \arg \max_{\|\tilde{\phi}_k'\|^2=1} \mathrm{Var}\left((\tilde{\phi}_k')^T \tilde{X}^{(k)} \right)$$

$$\approx \arg \max_{\|\tilde{\phi}_k'\|^2=1} \frac{1}{n} (\tilde{\phi}_k')^T (\tilde{\mathbf{X}}^{(k)})^T \tilde{\mathbf{X}}^{(k)} \tilde{\phi}_k'.$$

➔ ***Standardizing the data is important!***

In Summary

- PCA seeks directions that preserve the maximum variance. That is, we find change of coordinate matrix $\mathbf{\Phi} = (\phi_1, \ldots, \phi_p)$ that yields

$$\mathbf{Z} = \mathbf{X}\mathbf{\Phi}$$

whose k-th column preserves the maximum variance in

$$\tilde{\mathbf{X}}^{(k)} = \mathbf{X} - \textstyle\sum_{j=1}^{k-1} \mathbf{X}\phi_j\phi_j^T.$$

- By picking only the first M $(< p)$ principal components, we are modeling

$$\mathbf{X} = \mathbf{Z}^{1:M} (\mathbf{\Phi}^{1:M})^T + \mathbf{E}$$

with $\|\mathbf{E}\|_F$ being minimized.

Methods for Finding Principal Components

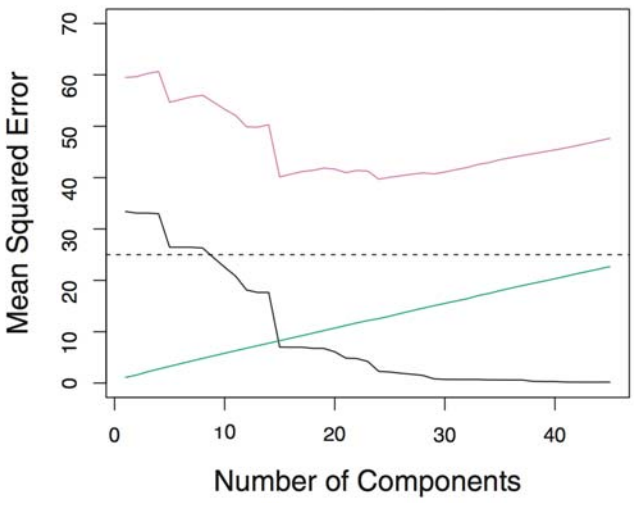
- By singular value decomposition (SVD) $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, we obtain the loading matrix $\mathbf{\Phi} = (\phi_1, \ldots, \phi_p) = \mathbf{V}$ and the principal component scores $\mathbf{Z} = \mathbf{U}\mathbf{D}$.
- By the Nonlinear Iterative PArtial Least Squares (NIPALS) algorithm: (To find the k-th PC using $\tilde{\mathbf{X}}^{(k)}$)
  1) Take a vector $\mathbf{z}_k$ from the columns of $\tilde{\mathbf{X}}^{(k)}$.
  2) Calculate $\phi_k \leftarrow (\tilde{\mathbf{X}}^{(k)})^T \mathbf{z}_k / (\mathbf{z}_k^T \mathbf{z}_k)$.
  3) Normalize $\phi_k \leftarrow \phi_k / \|\phi_k\|$.
  4) Calculate $\mathbf{z}_k \leftarrow \tilde{\mathbf{X}}^{(k)} \phi_k / (\phi_k^T \phi_k)$.
  5) Compare $\mathbf{z}_k$ with that in previous iteration, if the same, then stop; if not, go to Step 2.

Principal Components Regression

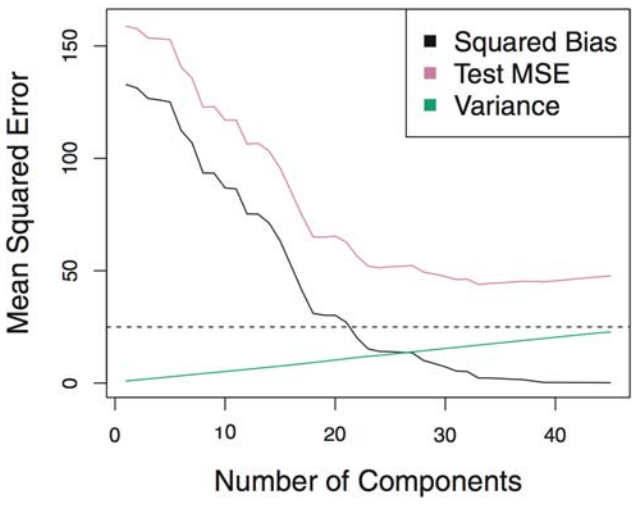
- **Principal components regression (PCR)** utilizes the first M principal components, $Z_1, \ldots, Z_M$, as the predictors to perform least squares linear regression.

Example: (Simulated Data Set from Figs. 6.8, 6.9)

all 45 predictors



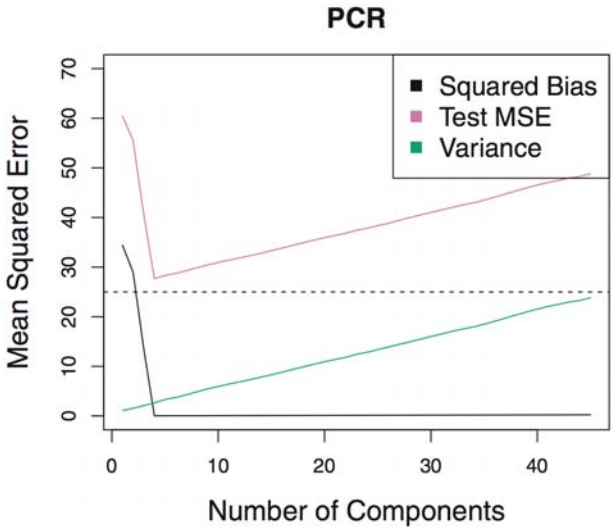
only 2 predictors

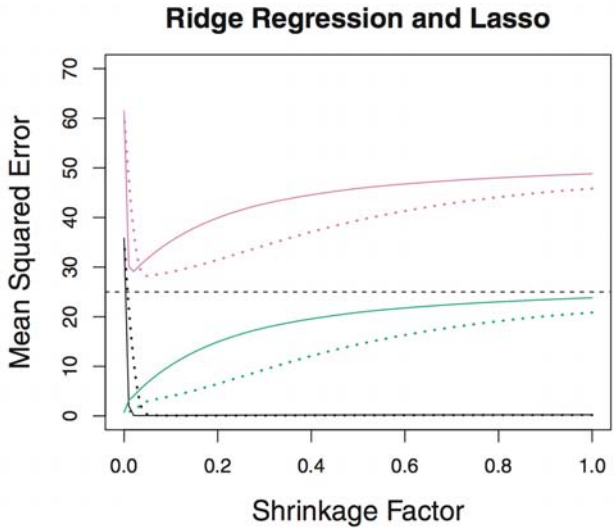


Example

Example:

- Simulated data set where the response depends only on the first 5 principal components.

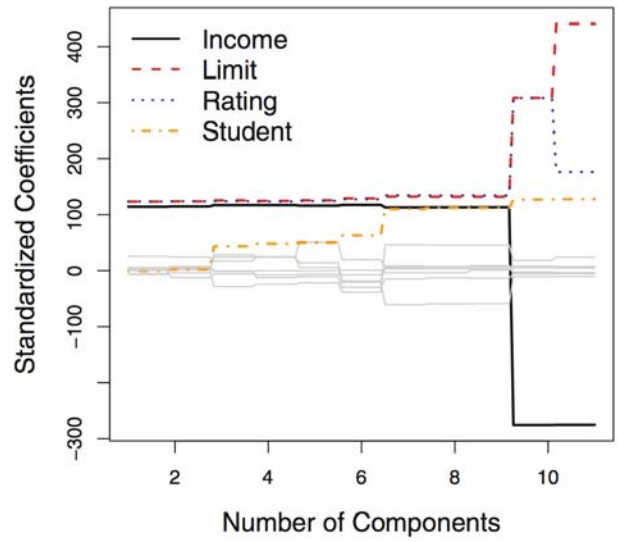


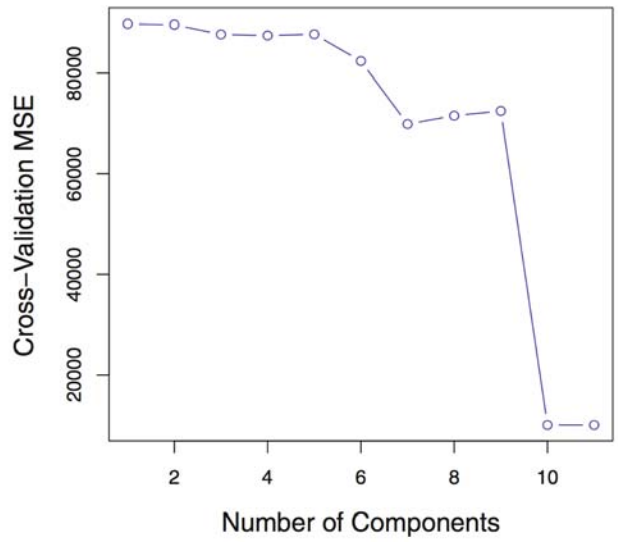


Example

Example (Credit Card Data Set):

- Total of 11 predictors.





Partial Least Squares (1/2)

- Recall that PCR chooses directions that preserve the most information about the predictors $X_1, \ldots, X_p$.

➔No guarantee that the preserved information is relevant to the response Y.

- **Partial least squares (PLS)** is a supervised alternative to PCR that chooses new features $Z_1, \ldots, Z_M$ that approximate $X_1, \ldots, X_p$ well, but are also related to Y.
  - The first feature $Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$ is chosen such that, for each j, the coefficient ϕ_{j1} minimizes

$$E[(Y - \phi_{j1} X_j)^2] \approx \frac{1}{n} \sum_{i=1}^{n} (y_i - \phi_{j1} x_{ij})^2.$$

➔

Partial Least Squares (2/2)

  - Then, adjust each of the variables $X_1, \ldots, X_p$ by regressing each variable on Z_1 and take residuals. That is, compute

$$\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)^T = X - \beta_1 Z_1 \quad \left(\text{or } \tilde{\mathbf{X}} = \mathbf{X} - \mathbf{z}_1 \beta_1^T\right)$$

where $\beta_1 = (\beta_{11}, \ldots, \beta_{p1})^T$ is chosen to minimize

$$E[\|X - \beta_1 Z_1\|^2] \approx \frac{1}{n} \sum_{i=1}^{n} \|x_i - \beta_1 z_{i1}\|^2 = \frac{1}{n} \|\mathbf{X} - \mathbf{z}_1 \beta_1^T\|_F^2.$$

➔

  - The second feature $Z_2 = \sum_{j=1}^{p} \phi_{j2} X_j$ is chosen such that, for each j, the coefficient ϕ_{j2} minimizes

$$E[(Y - \phi_{j2} \tilde{X}_j)^2] \approx \frac{1}{n} \sum_{i=1}^{n} (y_i - \phi_{j2} \tilde{x}_{ij})^2.$$

In Summary

- PCA seeks directions that preserve the maximum variance. This can be done independently for both $\mathbf{X}$ and $\mathbf{Y}$ to get

$$\mathbf{X} = \mathbf{Z}\mathbf{\Phi}^T + \mathbf{E} \quad \text{and} \quad \mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F}.$$

  ➔ Directions are chosen only w.r.t. variability in $\mathbf{X}$.

- In PLS, we seek to find decompositions so that the covariance between $\mathbf{Z}$ and $\mathbf{U}$ is maximized. Then, by regressing $\mathbf{u}_j$ onto $\mathbf{z}_j$, we can obtain approximation $\hat{\mathbf{u}}_j = b_j \mathbf{z}_j$ where $b_j = \mathbf{u}_j^T \mathbf{z}_j / (\mathbf{z}_j^T \mathbf{z}_j)$ and, thus,

$$\hat{\mathbf{Y}} = \mathbf{Z}\mathbf{B}\mathbf{Q}^T$$

  where $\mathbf{B} = \mathrm{diag}(b_1, \ldots, b_M)$.

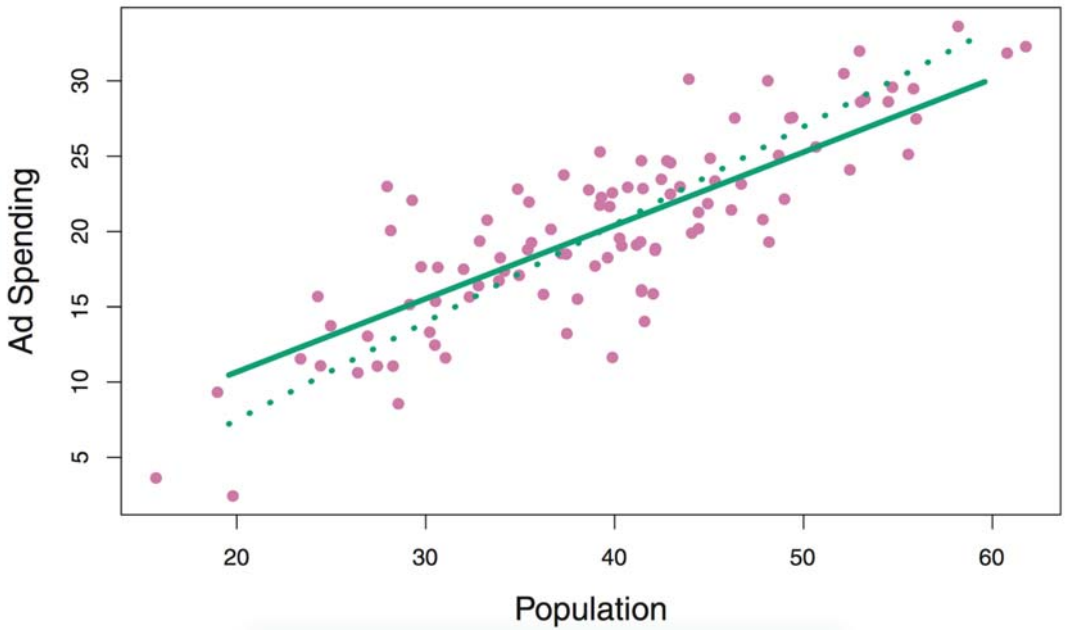
NIPALS for Partial Least Squares

- PLS by the NIPALS algorithm: (To find the k-th PLS score vector for $\tilde{\mathbf{X}}^{(k)}$)
  1) Take a vector $\mathbf{u}_k$ from $\tilde{\mathbf{Y}}^{(k)}$.
  2) Calculate $\phi_k \leftarrow (\tilde{\mathbf{X}}^{(k)})^T \mathbf{u}_k / \|(\tilde{\mathbf{X}}^{(k)})^T \mathbf{u}_k\|$.
  3) Calculate $\mathbf{z}_k \leftarrow \tilde{\mathbf{X}}^{(k)} \phi_k / (\phi_k^T \phi_k)$.
  4) Calculate $q_k \leftarrow (\tilde{\mathbf{Y}}^{(k)})^T \mathbf{z}_k / \|(\tilde{\mathbf{Y}}^{(k)})^T \mathbf{z}_k\|$.
  5) Calculate $\mathbf{u}_k \leftarrow \tilde{\mathbf{Y}}^{(k)} q_k / (q_k^T q_k)$.
  6) Compare $\mathbf{z}_k$ with that in previous iteration, if the same, then stop; if not, go to Step 2.
- Then, with $\beta_k = \tilde{\mathbf{X}}^{(k)} \mathbf{z}_k / (\mathbf{z}_k^T \mathbf{z}_k)$, we have $\tilde{\mathbf{X}}^{(k+1)} = \tilde{\mathbf{X}}^{(k)} - \mathbf{z}_k \beta_k^T$ and $\tilde{\mathbf{Y}}^{(k+1)} = \tilde{\mathbf{Y}}^{(k)} - \alpha_{\mathbf{u}_k | \mathbf{z}_k} \mathbf{z}_k q_k^T$

Example

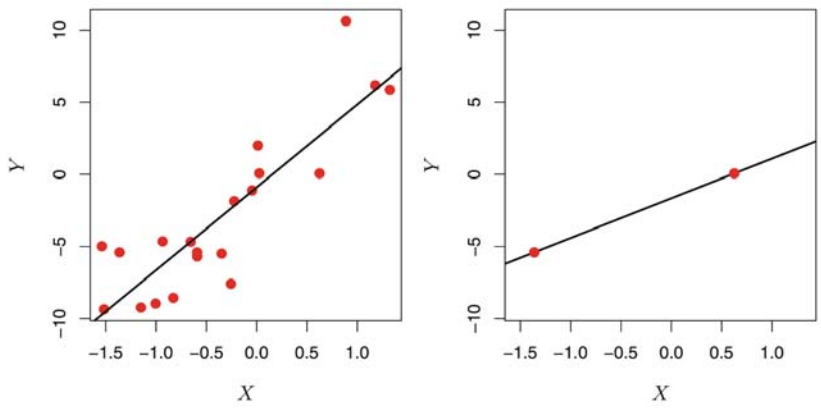
Example (Synthetic Data Set of Sales):

- Response: Sales in each of 100 regions.
- Predictors: Population Size and Advertising Spending



Higher Dimensional Data

- Some case may have p close to or greater than n.
  - E.g., (1) predict blood pressure based on age, gender, BMI, and half a million of SNPs; (2) predict shopping pattern based on search engine history of individuals.
- What goes wrong in higher dimensions?
  - When $p \geq n$, least squares yields a perfect fit.



Regression in High Dimensions

- Methods introduced for fitting less flexible models, such as subset selection, ridge regression, lasso etc, are useful in high dimensional settings.