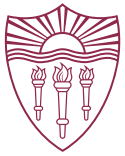


SUPERVISED LEARNING

REGRESSION + REGULARIZATION

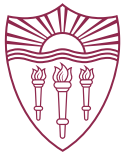
Kristina Lerman

USC Information Sciences Institute

DSCI 552 – Spring 2021

[Slides courtesy of Nathan Bastian, NWU]

Topics this week

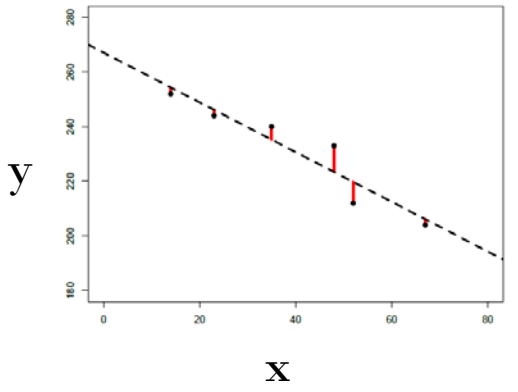


- Linear models
  - Model and feature selection
- Regularization
  - Understand and know how to perform ridge regression and the lasso as shrinkage (regularization) methods.
  - Understand and know how to perform principal components regression and partial least squares as dimension reduction methods.

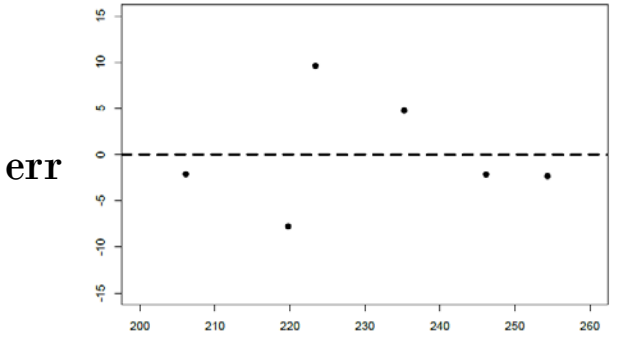
Linear models

- Outcome as a weighted sum of features: flexible, computable and interpretable models
  - Linearity (check residuals) → transform features
  - Homoscedasticity (check residuals) → consider transforming data
  - Normality (check residuals): p-values may be invalid
  - Correlated features → dimensionality reduction
  - Independence of observations → mixed effects models
  - Outliers

Errors (residuals)



Residual plot



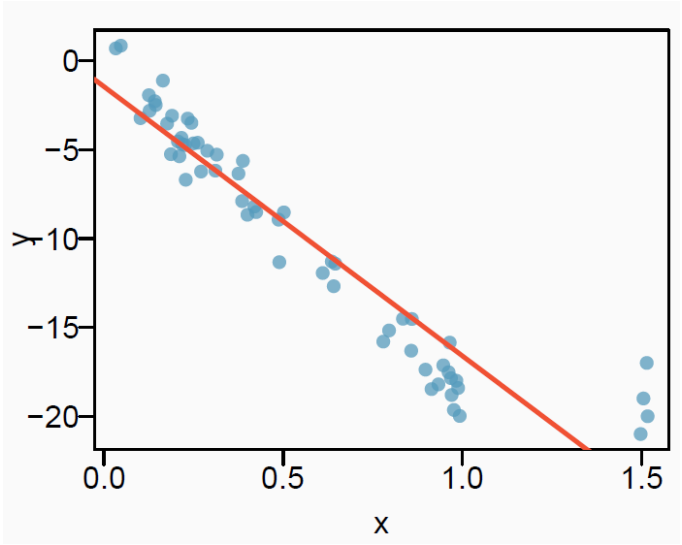
Outliers

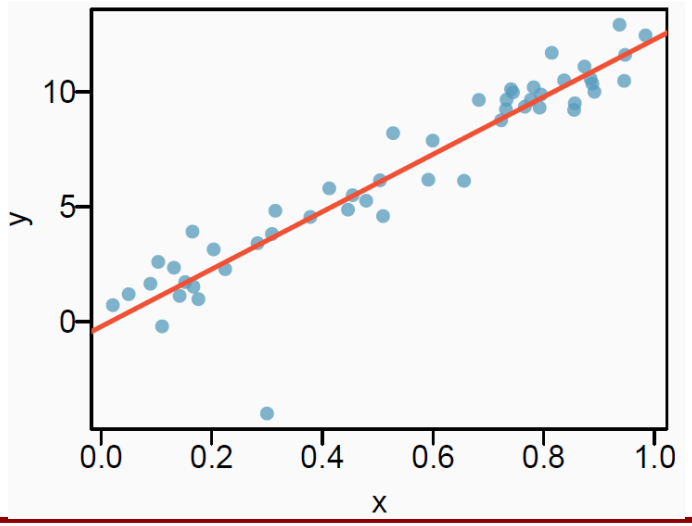
- Influential points: does leaving one data point out substantially change regression coefficients?
- Outliers that lie away from the center of the cloud in the x-direction are called high leverage points.
- A point is influential if including or excluding the point would considerably change the slope of the regression line.

- Do not exclude them from analysis, unless there is an obvious error with the data

Types of outliers

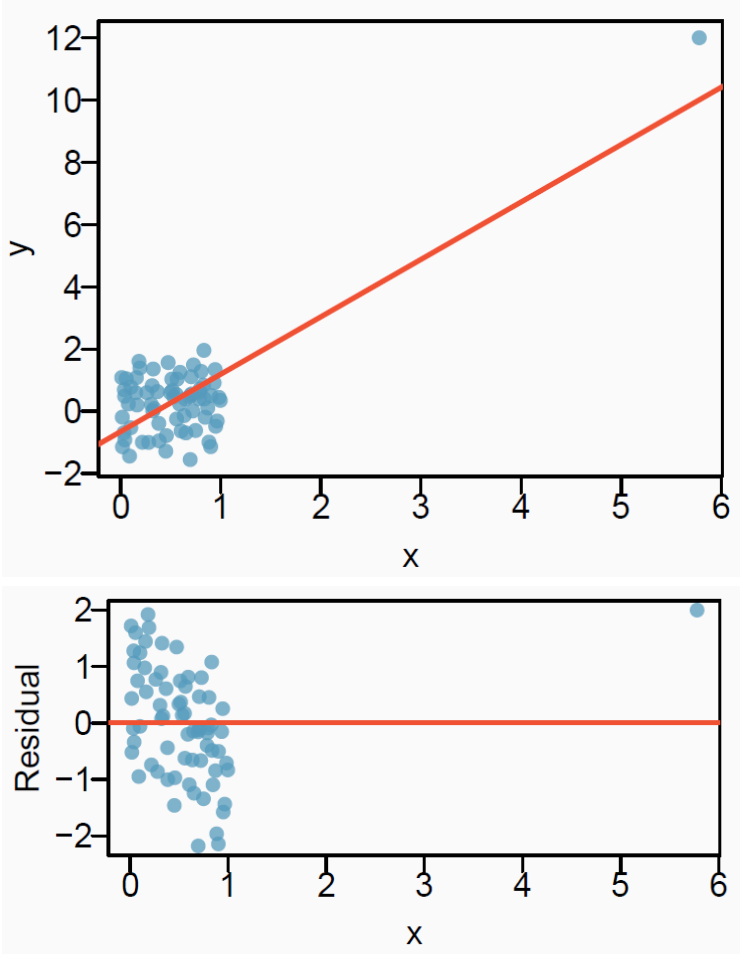
- How do outliers affect the regression line in this plot?
  - To answer, consider what the regression line would be without the outliers.
  - (top) Outliers pull the regression line away from the observations in the larger group of data.
  - (bottom) This outlier does not influence the regression line



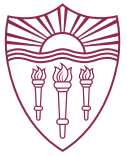


Types of outliers: influential points

- How do outliers affect the regression line in this plot?
  - Without the outlier, there is no observable relationship between x and y.
- Influential points: does leaving one data point out substantially change regression coefficients?
- Do not exclude them from analysis, unless there is an obvious error with the data



Improving the Linear Model



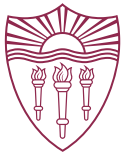
- Recall that standard linear models

$$\mathbf{y} = \beta_0 + \beta_1 \boldsymbol{x}_1 + \ldots + \beta_p \boldsymbol{x}_p$$

- We may want to improve the simple linear model by replacing OLS estimation with some other fitting procedure.

- Why use an alternative fitting procedure?
  - Prediction Accuracy
  - Model Interpretability

Prediction Accuracy

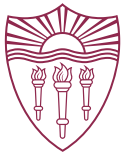


- The OLS estimates have relatively low bias and low variability especially
  - In large data sets; *n* (data size) >> *p* (model complexity)
  - When relationship between the response and predictors is linear.

- If *n* ~ *p*, then the OLS fit can have high variance and may result in overfitting and poor estimates on test data.

- If *p* > *n*, then the variability of the OLS fit increases dramatically, and the variance of these estimates in infinite.

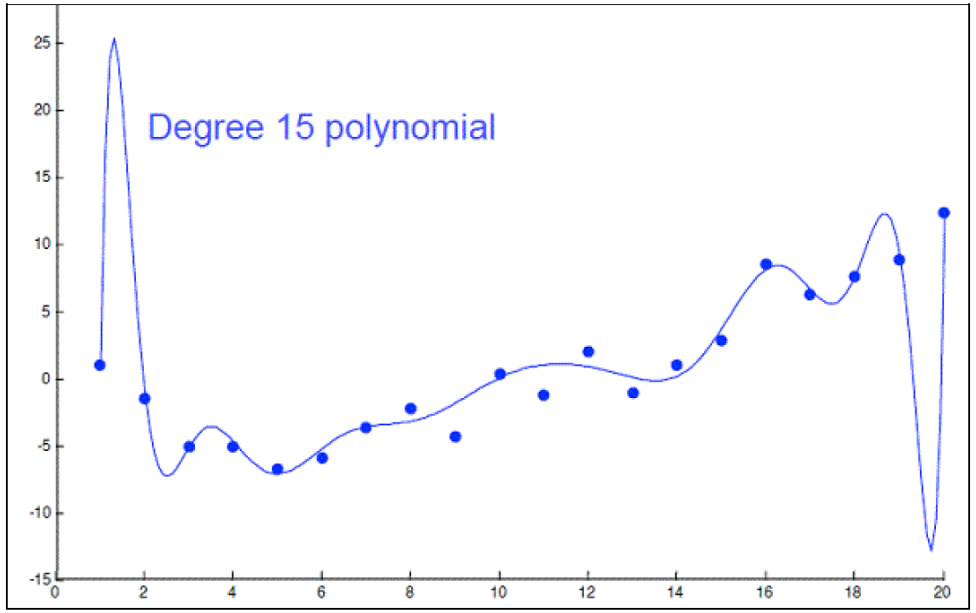
Model Interpretability



- When we have a large number of predictors/features in the model, there will generally be many that have little or no effect on the response.
- Including such irrelevant variable leads to unnecessary complexity.
- Leaving these variables in the model makes it harder to see the effect of the important variables.
- The model would be easier to interpret by removing (i.e. setting the coefficients to zero) the unimportant variables.

Feature Selection

- Carefully selected features can improve model accuracy, but adding too many can lead to overfitting.
  - Overfitted models describe random error or noise instead of any underlying relationship.
  - They generally have poor predictive performance on test data.

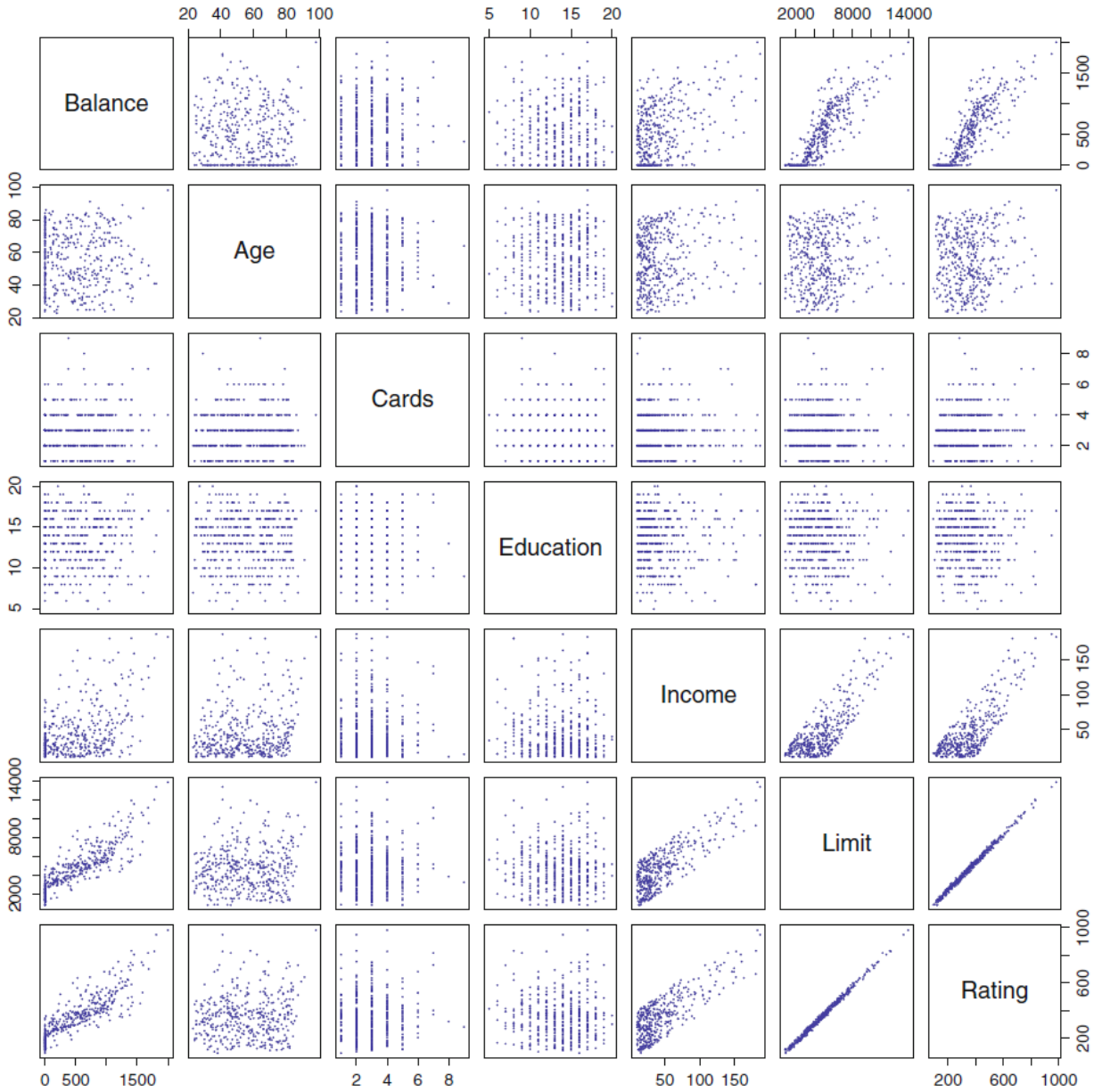


- For instance, we can use a 15-degree polynomial function to fit the following data so that the fitted curve goes nicely through the data points.
- However, a brand new dataset collected from the same population may not fit this particular curve well at all.

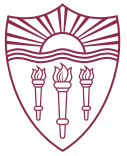
Credit data

Scatter plots

- Outcome:
  - Credit card **balance**
- Features:
  - **Age**,
  - number of **Cards**,
  - years of **Education**,
  - **Income**,
  - credit **Limit**
  - credit **Rating**,

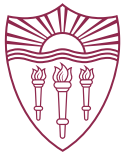


Feature Selection



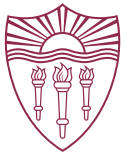
- Subset Selection
  - Identify a subset of features most related to the outcome; fit a model using OLS on the reduced set.
  - Methods: forward feature selection, MRMR, …
- Shrinkage (Regularization)
  - Involves shrinking the estimated coefficients toward zero relative to the OLS estimates; has the effect of reducing variance and performs variable selection.
  - Methods: ridge regression, lasso
- Dimension Reduction
  - Involves projecting the p predictors into a M-dimensional subspace, where $M < p$, and fit the linear regression model using the M projections as predictors.
  - Methods: principal components regression, partial least squares

Forward/Backward Feature Selection



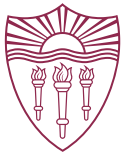
- **Forward Selection**: Forward selection is an iterative method, starting with no features in the model
  - In each iteration, add the best performing feature to the model
    - select the feature with the lowest p-value
  - Continue until adding a new feature does not improve the performance
- **Backward Elimination**: starting with all the features
  - In each iteration, remove the least significant feature
    - feature with the largest insignificant p-value
  - Continue until all features with insignificant p-values are removed

Feature selection & Multi-collinearity



- Reduce the number of colinear features by eliminating un-informative features
- **Variance Inflation Factor** - quantifies the severity of multicollinearity and measures how much the variance of an estimated regression coefficient is increased because of collinearity.
- **Minimum redundancy maximum relevance** (mRMR) – identifies features that are highly correlated with the outcome (relevance), but uncorrelated with each other (redundancy)

Variance Inflation Factor

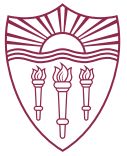


- The VIF is the ratio of the variance of regression coefficient *j* when fitting the full model divided by the variance of ˆβj if fit on its own.

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

- R2 is from regression of data on all features besides j
  - For the Credit data regression of balance on age, rating, and limit indicates that the features have VIF values of 1.01, 160.67, and 160.59.
  - Drop rating. Without this feature, the R2 drops from 0.754 to 0.750.
  - **Rule of thumb** – calculate VIF for each feature and eliminate features with VIF>5.

Choosing the Optimal Model

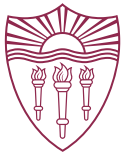


- The model containing all the features/predictors will always have the smallest RSS and the largest R^2, since these quantities are related to the **training error**.

- We wish to choose a model with low **test error**, not a model with low training error. Recall that training error is usually a poor estimate of test error.

- Thus, RSS and R^2 are not suitable for selecting the *best* model among a collection of models with different numbers of predictors.

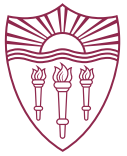
Estimating Test Error



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.

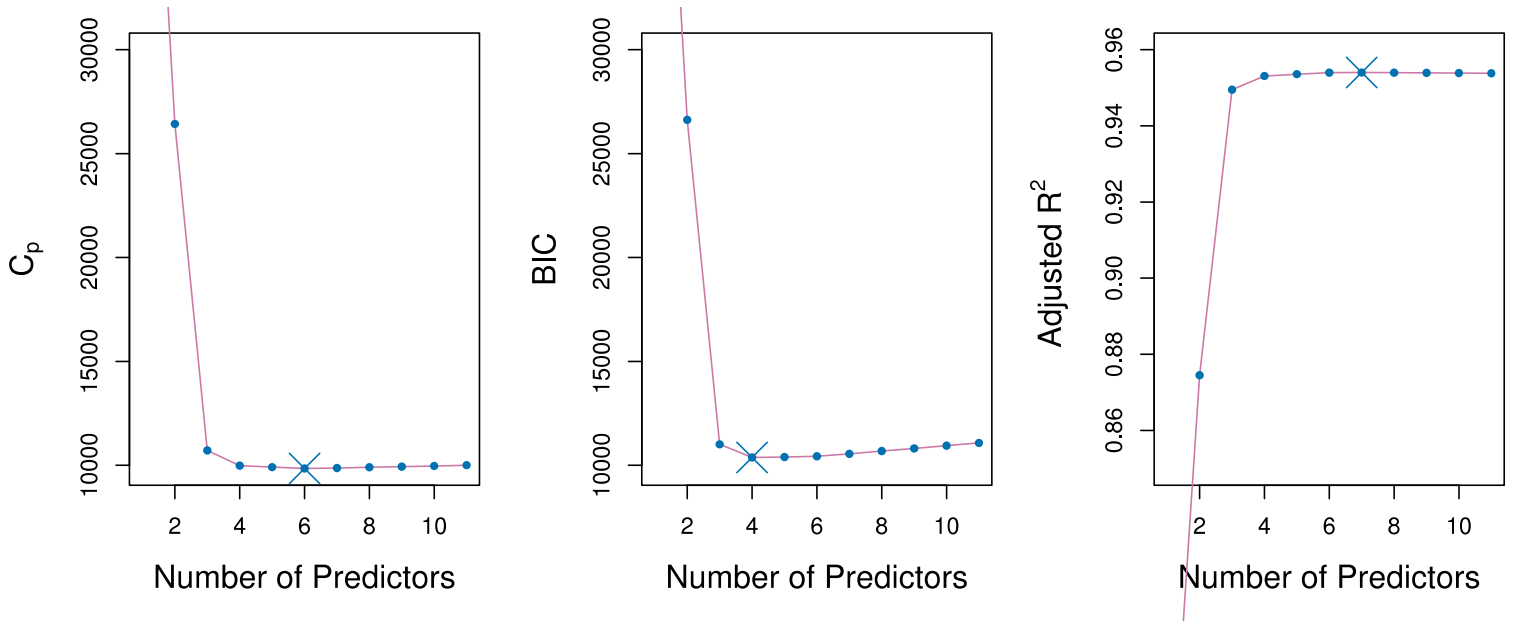
Other Measures of Comparison



- To compare different models, we can use other approaches:
  - Adjusted R^2
  - AIC (Akaike information criterion)
  - BIC (Bayesian information criterion)
  - Mallow's C_p (equivalent to AIC for linear regression)
- These techniques adjust the training error for the model complexity and can be used to select among a set of models with different numbers of variables.
- These methods add penalty to RSS for the number of predictors in the model.

Credit Data: C_p, BIC, and Adjusted R^2

- A small value of C_p and BIC indicates a low error, and thus a better model.
- A large value for the Adjusted R^2 indicates a better model.



Mallow's C_p

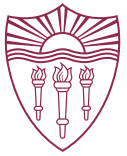
- For a fitted OLS model containing *d* predictors, the C_p estimate of test MSE:

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

where $\hat{\sigma}^2$ is an estimate of the variance of the error ε associated with each response measurement.

- Here, a penalty is added to the training RSS in order to adjust for the fact that the training error tends to underestimate the test error.

Akaike Information Criterion (AIC)



- Defined for a large class of models fit by maximum likelihood.

$$\mathrm{AIC} = -2\log L + 2 \cdot d$$

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

- In the case of the linear model with Gaussian errors, MLE and OLS are the same things; thus, C_p and AIC are equivalent.

Bayesian Information Criterion (BIC)

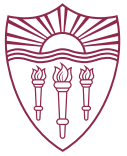
- BIC will tend to take on a small value for a model with a low test error, and so generally we select the model that has the lowest BIC value.

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

- Since log *n* > 2 for an *n* > 7, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than C_p.

- Notice that 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.

Adjusted R^2



- For an OLS model with *d* variables, the adjusted R^2 is calculated:

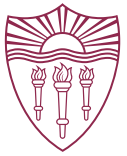
$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

where TSS is the total sum of squares.

- Unlike the other statistics, a large value of adjusted R^2 indicates a model with a small test error.

- The adjusted R^2 statistics *pays a price* for the inclusion of unnecessary variables in the model.

Shrinkage (Regularization) Methods

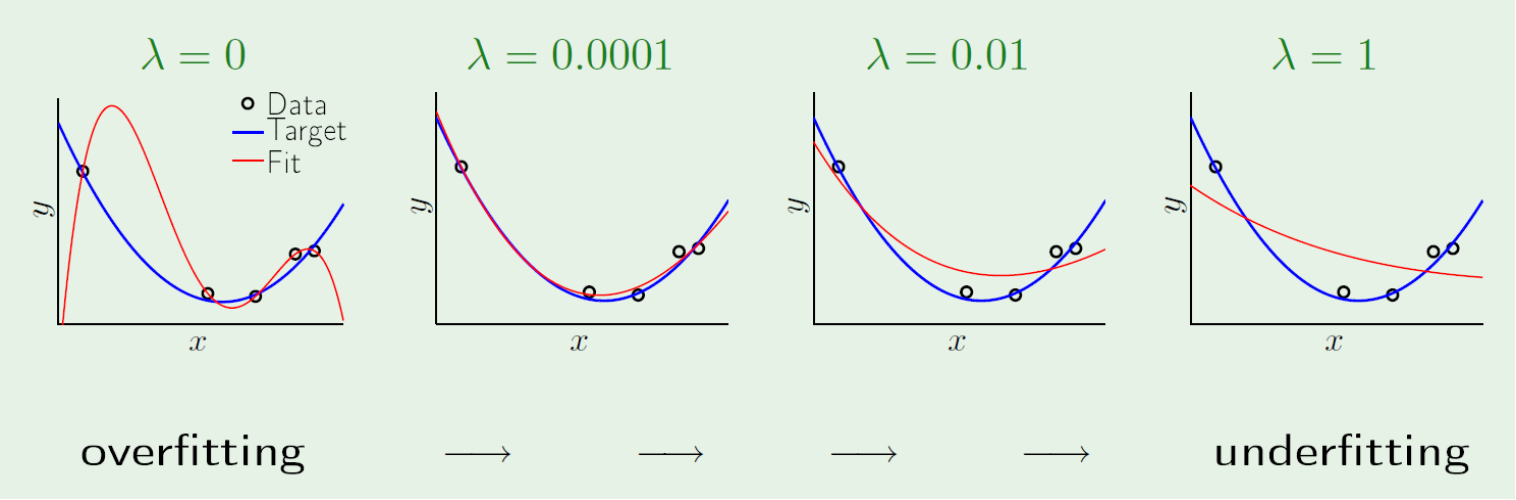


- As an alternative to feature selection, we can fit a model containing all *p* predictors using a technique that constrains or *regularizes* the coefficient estimates
  - To *regularize* means to *shrink* the coefficient estimates towards zero

- 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 their variance.

Shrinkage (Regularization) Methods

- Regularization is our first weapon to combat overfitting.
- It constrains the machine learning algorithm to improve out-of-sample (test) error, especially when noise is present.
- Look at what a little regularization can do:



Ridge Regression

- Recall that the OLS fitting procedure estimates the beta coefficients using the values that minimize:

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

- Ridge regression is similar to OLS, except that the coefficients are estimated by minimizing a slightly different quantity:

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

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

Ridge Regression

- Note that $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage.
- The idea of penalizing by the sum-of-squares of the parameters is also used in neural networks, where it is known as *weight decay,* and also in adversarial learning
- An equivalent

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2$$

$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 \leq t,$$

Ridge Regression

- The effect of this equation is to add a shrinkage penalty of the form

$$\lambda \sum_{j=1}^{p} \beta_j^2,$$

where the tuning parameter λ is a positive value.

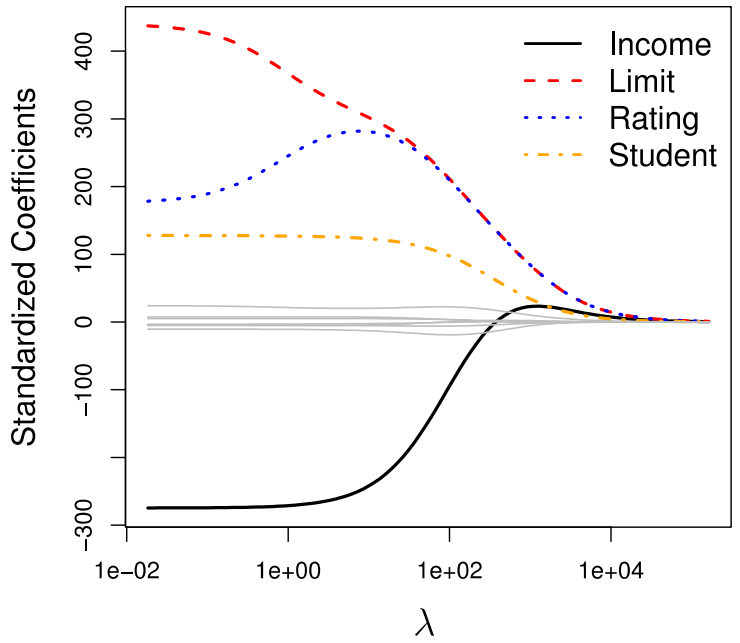
- This has the effect of shrinking the estimated beta coefficients towards zero. It turns out that such a constraint should improve the fit, because shrinking the coefficients can significantly reduce their variance.

- Note that when λ = 0, the penalty term as no effect, and ridge regression will procedure the OLS estimates. Thus, selecting a good value for λ is critical (can use cross-validation for this).

Ridge Regression

- As λ increases, the standardized ridge regression coefficients shrinks towards zero.

- Thus, when λ is extremely large, then all of the ridge coefficient estimates are basically zero; this corresponds to the *null model* that contains no predictors.



Ridge Regression

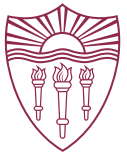
- The standard OLS coefficient estimates are *scale equivariant*.

- However, 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.

- Thus, it is best to apply ridge regression after *standardizing* or *normalizing the predictors*

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

Ridge Regression



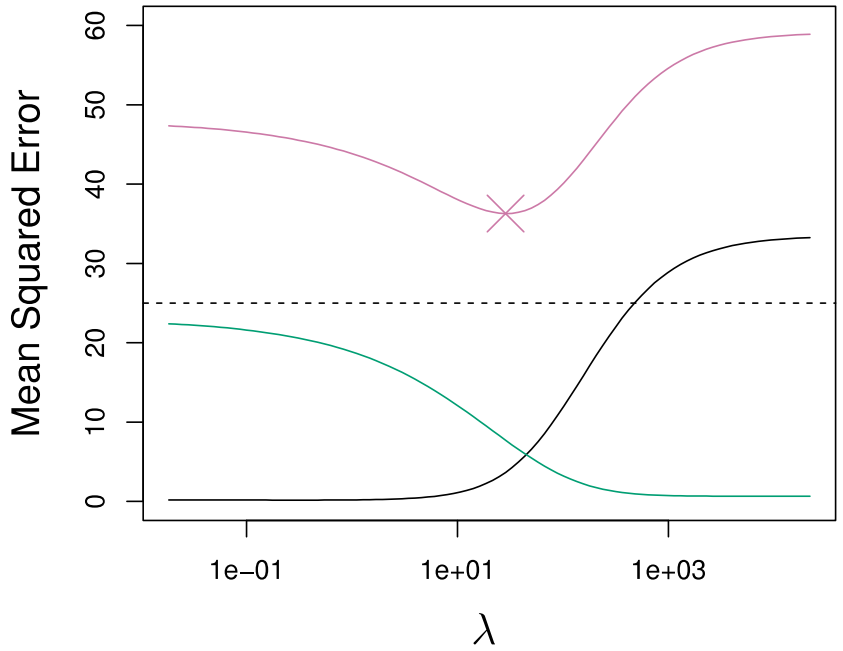
Why does Ridge Regression fitting improve over OLS?

- It turns out that the OLS estimates generally have low bias but can be highly variable. In particular when n and p are of similar size or when $n < p$, then the OLS estimates will be extremely variable

- The penalty term makes the ridge regression estimates *biased* but can also substantially reduce variance; i.e., as λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.

- As a result, there is a bias/variance trade-off.

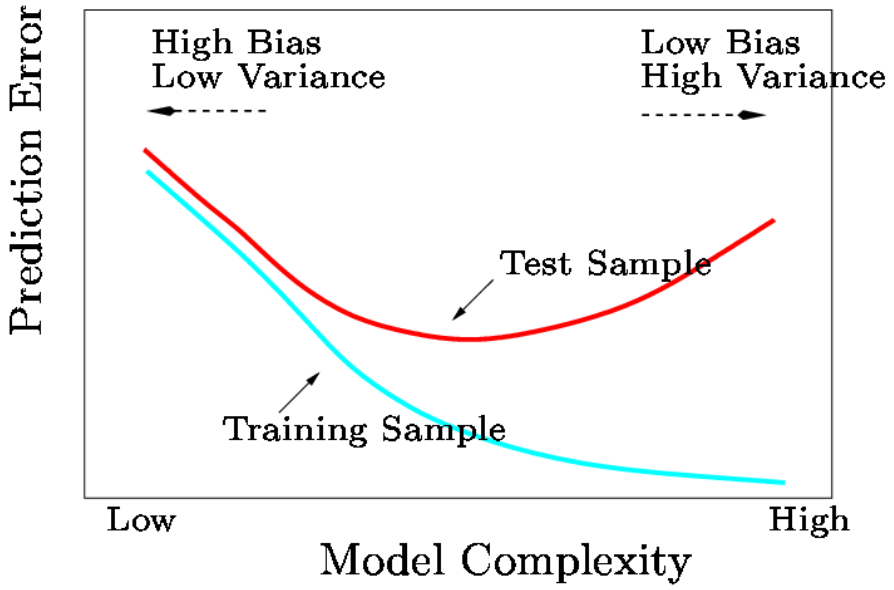
Ridge Regression



- Black = Bias
- Green = Variance
- Purple = MSE

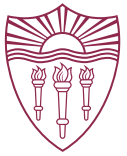
- Increased λ leads to increased bias but decreased variance

Ridge Regression



- In general, the ridge regression estimates will be more biased than the OLS ones but have lower variance.
- Ridge regression will work best in situations where the OLS estimates have high variance.

Ridge Regression



Computational Advantages of Ridge Regression

- If p is large, then finding the best subset of features requires searching through enormous numbers of possible models.

- With ridge regression, for any given λ we only need to fit one model and the computations turn out to be very simple.

- Ridge regression can even be used when $p > n$, a situation where OLS fails completely (i.e. OLS estimates do not even have a unique solution).

Ridge Regression (a mathematical digression)

- In matrix form:

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

the ridge regression solutions are easily seen to be

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

- The solution adds a positive constant to the diagonal of $\mathbf{X}^T\mathbf{X}$ before inversion (making the problem non-singular).

- The *singular value decomposition* (SVD) of the centered matrix **X** gives us some additional insight into the nature of ridge regression.

Ridge Regression (a mathematical digression)

- The SVD of the *N* x *p* matrix **X** has the form $\mathbf{X} = \mathbf{UDV}^T$
- Here, **U** and **V** are *N* x *p* and *p* x *p* orthogonal matrices, with the columns of **U** spanning the column space of **X**, and the columns of **V** spanning the row space.
- **D** is a *p* x *p* diagonal matrix, with diagonal entries $d_1 \geq d_2 \geq \cdots d_p \geq 0$ called singular values of **X**.
- If one or more values $d_j = 0$, **X** is singular.

Ridge Regression (a mathematical digression)

- Using SVD, we can write the OLS fitted vector as:

$$\begin{aligned}\mathbf{X}\hat{\beta}^{\text{ls}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{U}\mathbf{U}^T\mathbf{y},\end{aligned}$$

- The ridge regression solutions are:

$$\begin{aligned}\mathbf{X}\hat{\beta}^{\text{ridge}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{U}\,\mathbf{D}(\mathbf{D}^2+\lambda\mathbf{I})^{-1}\mathbf{D}\,\mathbf{U}^T\mathbf{y} \\ &= \sum_{j=1}^{p}\mathbf{u}_j\frac{d_j^2}{d_j^2+\lambda}\mathbf{u}_j^T\mathbf{y},\end{aligned}$$

where $\mathbf{u}_j$ are the columns of $\mathbf{U}$.

Ridge Regression (a mathematical digression)

- Like linear regression, ridge regression computes the coordinates of **y** with respect to the orthonormal basis **U**.

- It then *shrinks* these coordinates by the factor $\frac{d_j^2}{d_j^2+\lambda}$.

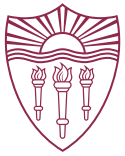
- This means that a greater amount of shrinkage is applied to the coordinates of basis vectors with smaller d_j^2.

- The SVD of the centered matrix **X** is another way of expressing the *principal components* of the variables in **X**.

Ridge Regression (a mathematical digression)

- Thus, we have $\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$, which is the *eigen decomposition* of $\mathbf{X}^T\mathbf{X}$.
- The eigenvectors v_j (columns of **V**) are also called the *principal components* directions of **X**.
- The first principal component direction v_1 has the property that $\mathbf{z}1 = \mathbf{X}v_1$ has the largest sample variance amongst all normalized linear combinations of the columns of **X**.
- The small singular values d_j correspond to directions in the column space of **X** having small variance, and ridge regression shrinks these directions the most.

The Lasso



- One significant problem of ridge regression is that the penalty term will never force any of the coefficients to be exactly zero.
- Thus, the final model will include all *p* predictors, which creates a challenge in model interpretation
- A more modern machine learning alternative is the *lasso*.
- The lasso works in a similar way to ridge regression, except it uses a different penalty term that shrinks some of the coefficients exactly to zero.

Lasso regression

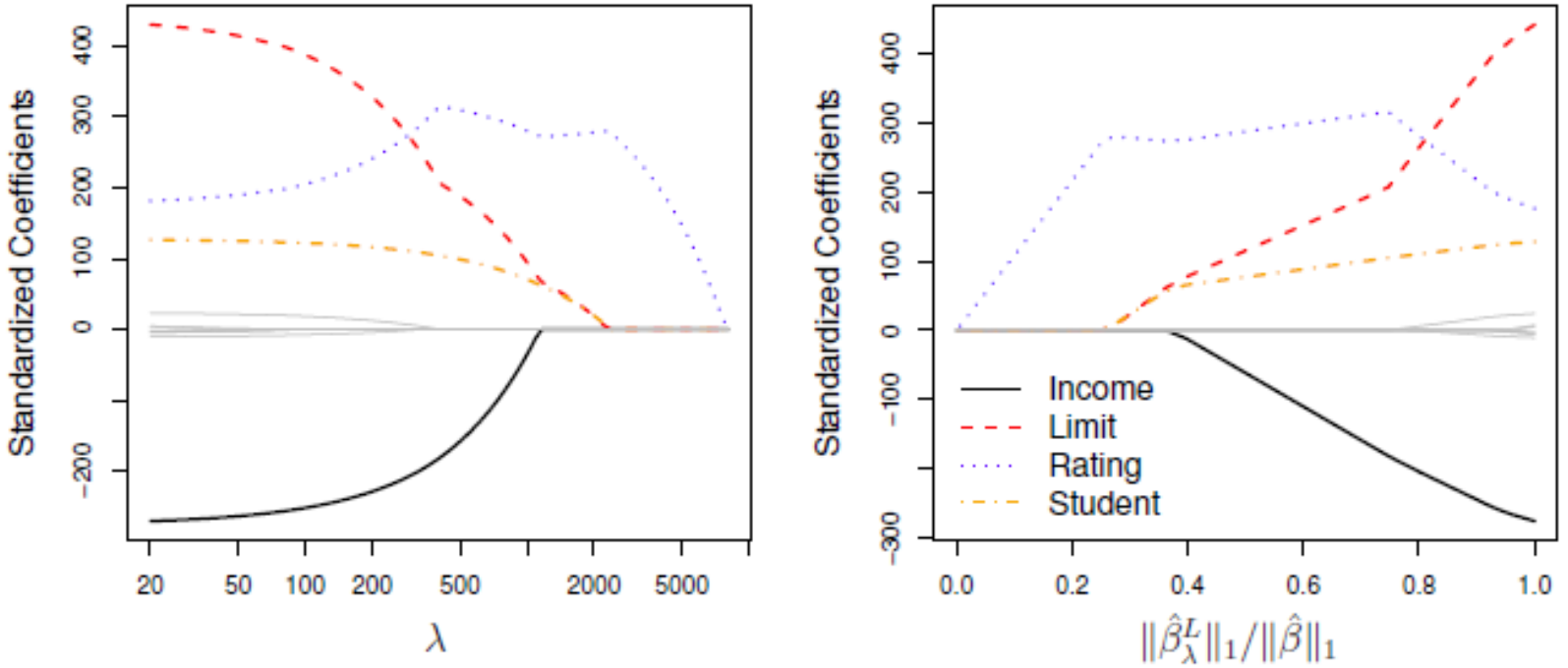
- The lasso coefficients minimize the quantity:

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

- The key difference from ridge regression is that the lasso uses an ℓ_1 penalty instead of an ℓ_2, which has the effect of forcing some of the coefficients to be exactly equal to zero when the tuning parameter λ is sufficiently large.

- Thus, the lasso performs variable/feature selection.

Lasso regression



- When λ = 0, then the lasso simply gives the OLS fit.
- When λ becomes sufficiently large, the lasso gives the null model in which all coefficient estimates equal zero.

Lasso regression

- One can show that the lasso and ridge regression coefficient estimates solves the problems:

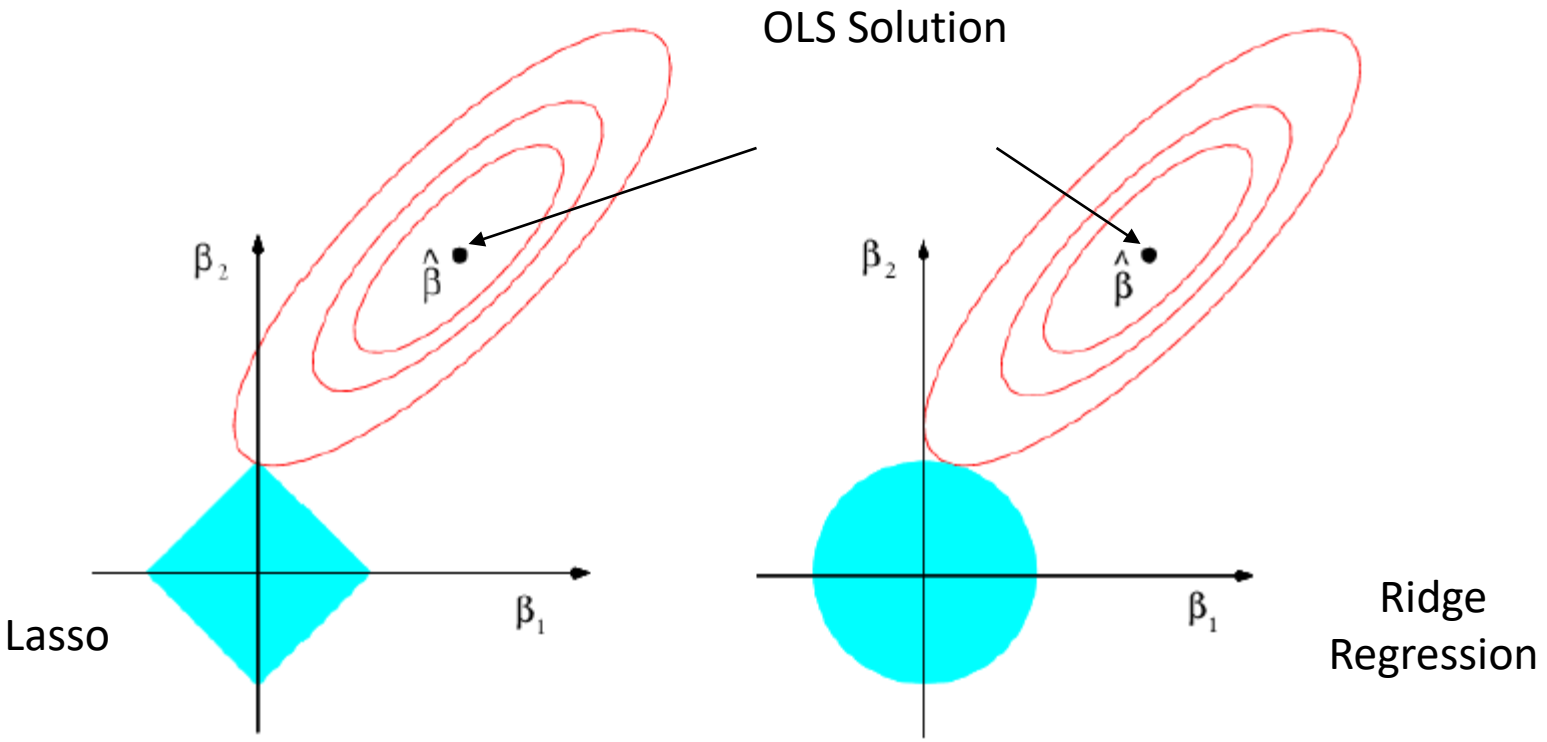
$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

and

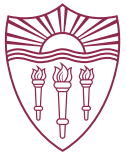
$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s,$$

Lasso regression

- The lasso and ridge regression coefficient estimates are given by the first point at which an ellipse contacts the constraint region.

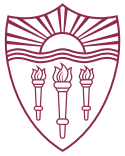


Lasso vs. Ridge Regression



- The lasso has a major advantage over ridge regression, in that it produces simpler and more interpretable models that involved only a subset of predictors.
- The lasso leads to qualitatively similar behavior to ridge regression, in that as λ increases, the variance decreases and the bias increases.
- The lasso can generate more accurate predictions compared to ridge regression.
- Cross-validation can be used in order to determine which approach is better on a particular data set.

Selecting the Tuning Parameter λ

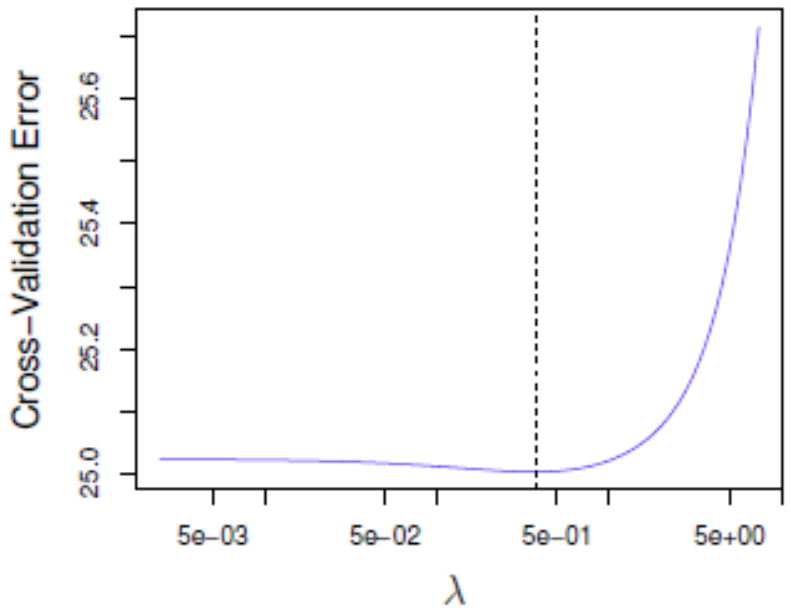


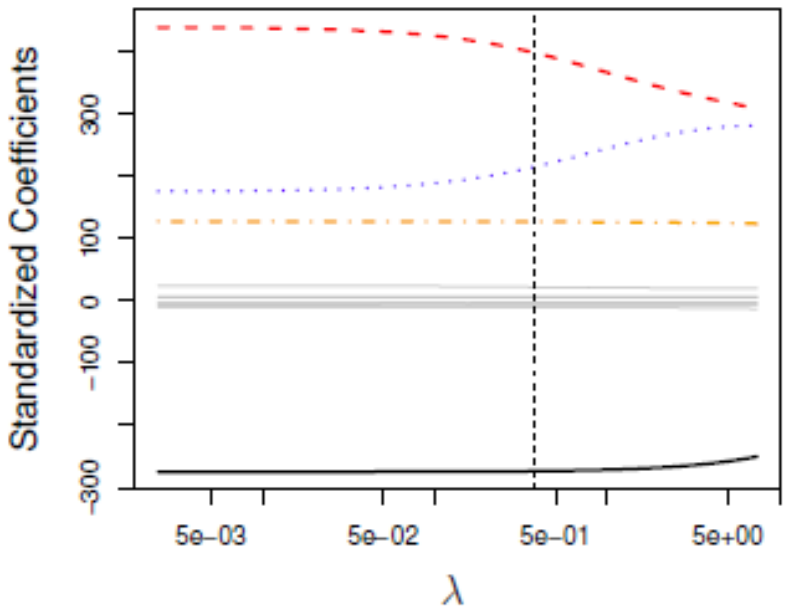
- What is the best value for the tuning parameter λ or equivalently, the value of the constraint *s*?

- Select a grid of potential values; use **cross-validation** to estimate the error rate on test data (for each value of λ) and select the value that gives the smallest error rate.

- Finally, the model is re-fit using all of the variable observations and the selected value of the tuning parameter λ.

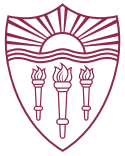
Selecting the Tuning Parameter λ: Credit Data Example





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 indicates the value of* λ *selected by cross-validation.*

Dimension Reduction



- The methods we have discussed so far have involved fitting linear regression models, via OLS or a shrunken approach, using the original predictors.

- We now explore a class of approaches that *transform* the predictors and then fit an OLS model using the transformed variables.

- We refer to these techniques as *dimension reduction* methods.

Dimension Reduction

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

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

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

- We can then fit an OLS linear regression model,

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

Dimension Reduction

- If the constants $\phi_{m1}, \ldots, \phi_{mp}$ are chosen wisely, then such dimension reduction approaches can outperform OLS regression.

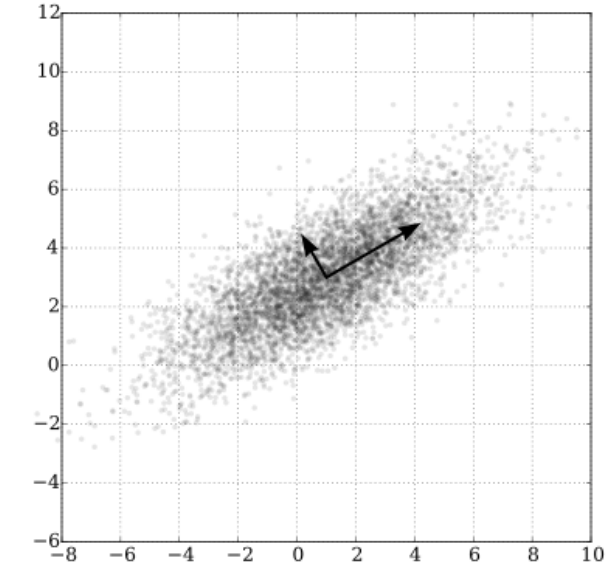
- The term *dimension reduction* comes from the fact that this approach reduces the problem of estimating the $p + 1$ coefficients $\beta_0, \ldots, \beta_p$ to the simpler problem of estimating the $M + 1$ coefficients $\theta_0, \ldots, \theta_M$, where $M < p$.

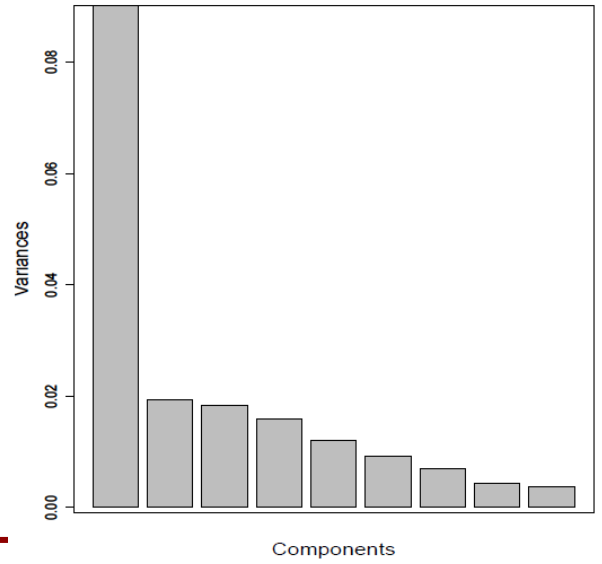
$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{mj} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{mj} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}$$

- This method serves to constrain the estimated β_j coefficients $\beta_j = \sum_{m=1}^{M} \theta_m \phi_{mj}$

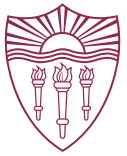
Principal Component Analysis - PCA

- Unsupervised method that identifies the internal structure of high-dimensional data that best explains its variance
- Embeds the data in a new lower-dimensional space, such that
  - The *first component* is that (normalized) linear combination of the variables with the largest variances.
  - The *second principal component* has largest variance, subject to being uncorrelated with the first….etc.
- Thus, with many correlated variables, we replace them with a small set of principal components that capture their joint variation.





Principal Components Regression



- The *principal components regression* (PCR) approach involves constructing the first *M* principal components, and then using these components as the predictors in an OLS linear regression model.

- Key idea: a small number of principal components *explains* most of the variability in the data, as well as the relationship with the response.

- We assume that the directions in which $X_1,\ldots,X_p$ show the most variation are the directions that are associated with *Y*.

- When performing PCR, predictors should be *standardized* prior to generating the principal components.

Principal Components Regression

- PCR forms the derived input columns $\mathbf{z}_m = \mathbf{X}v_m$, and then regresses $\mathbf{y}$ on $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_M$ for some $M \leq p$. Since the $\mathbf{z}_m$ are orthogonal, this regression is just a sum of univariate regressions:

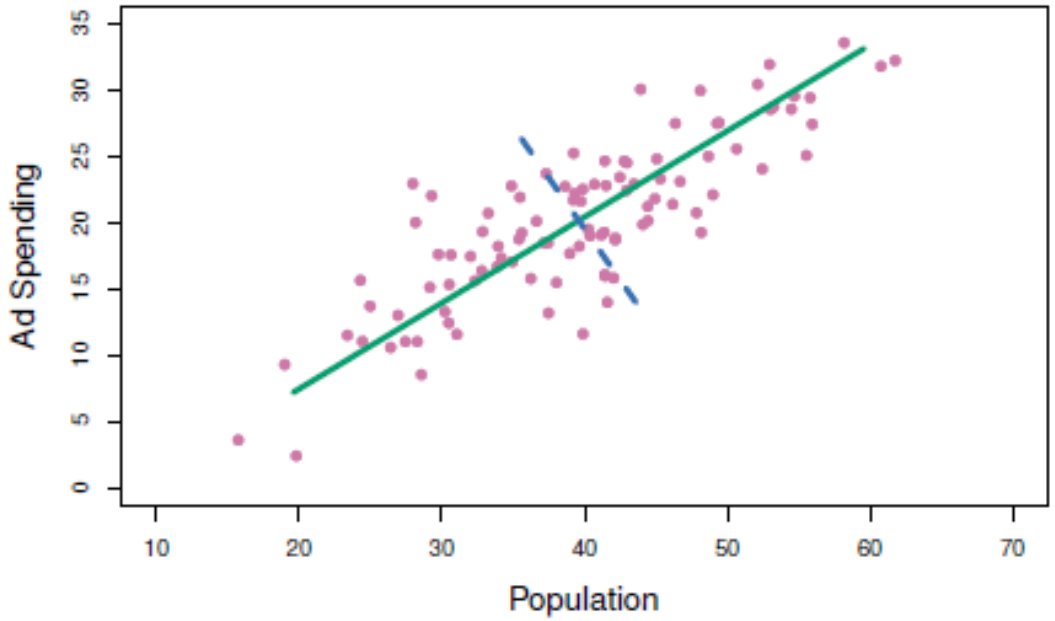
$$\hat{\mathbf{y}}^{\text{pcr}}_{(M)} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m \mathbf{z}_m$$

where $\hat{\theta}_m = \langle \boldsymbol{z}_m, \boldsymbol{y} \rangle / \langle \boldsymbol{z}_m, \boldsymbol{z}_m \rangle$. Since $\mathbf{z}_m$ are each linear combinations of the original $\mathbf{x}_j$, we can express the solution in terms of coefficients of the $\mathbf{x}_j$.

$$\hat{\beta}^{\text{pcr}}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m$$

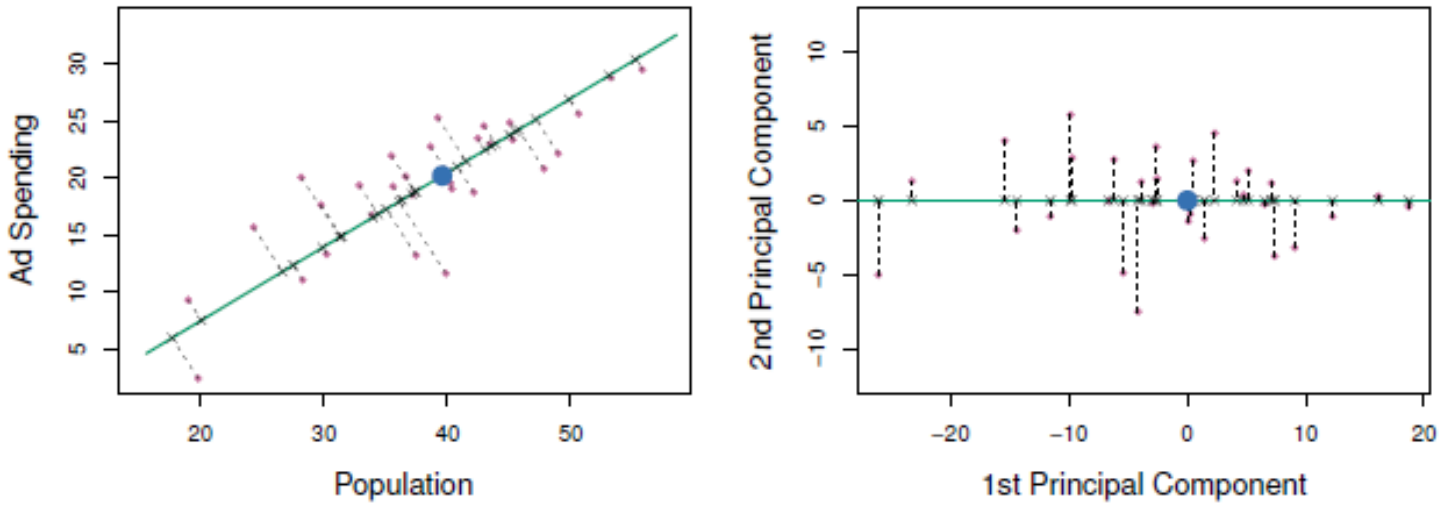
- PCR discards the $p - M$ smallest eigenvalue components.
- PCR is very similar to ridge regression in a certain sense.

Principal Components Regression



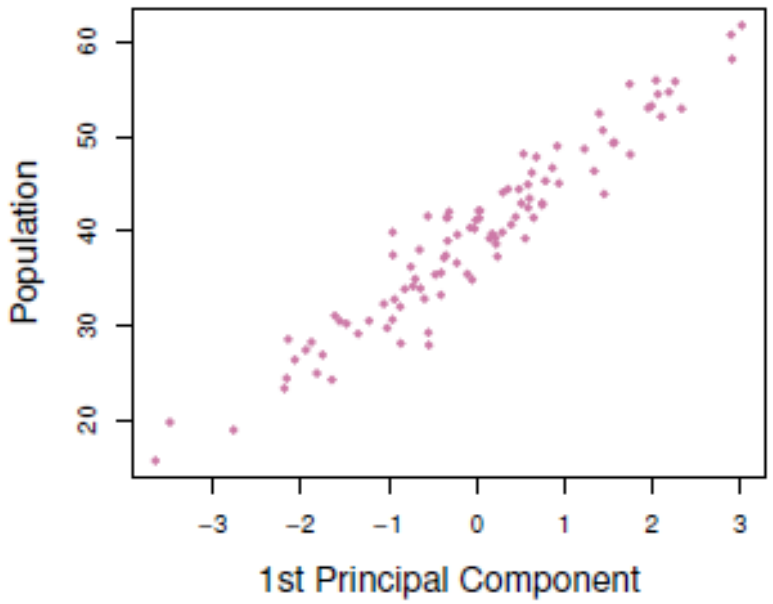
The population size (pop) and ad spending (ad) for 100 *different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

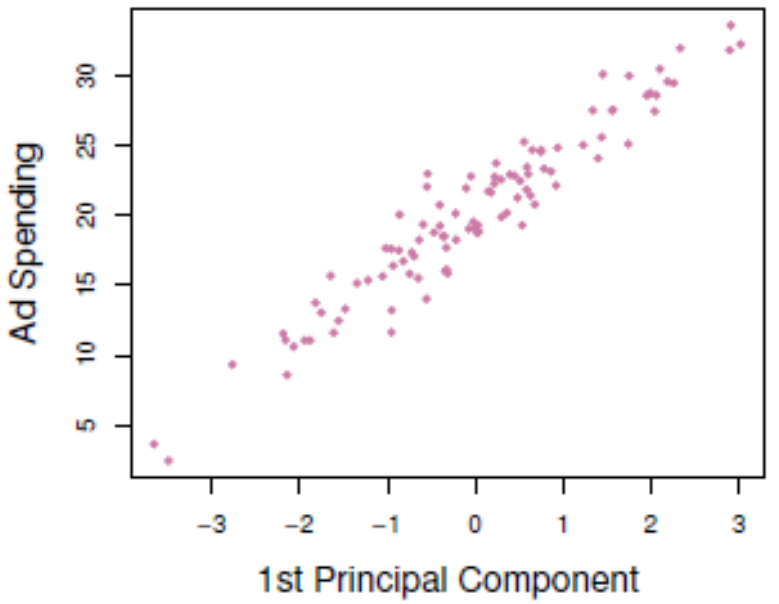
Principal Components Regression (cont.)



A subset of the advertising data. Left: The first principal component, chosen to minimize the sum of the squared perpendicular distances to each point, is shown in green. These distances are represented using the black dashed line segments. Right: The left-hand panel has been rotated so that the first principal component lies on the x-axis.

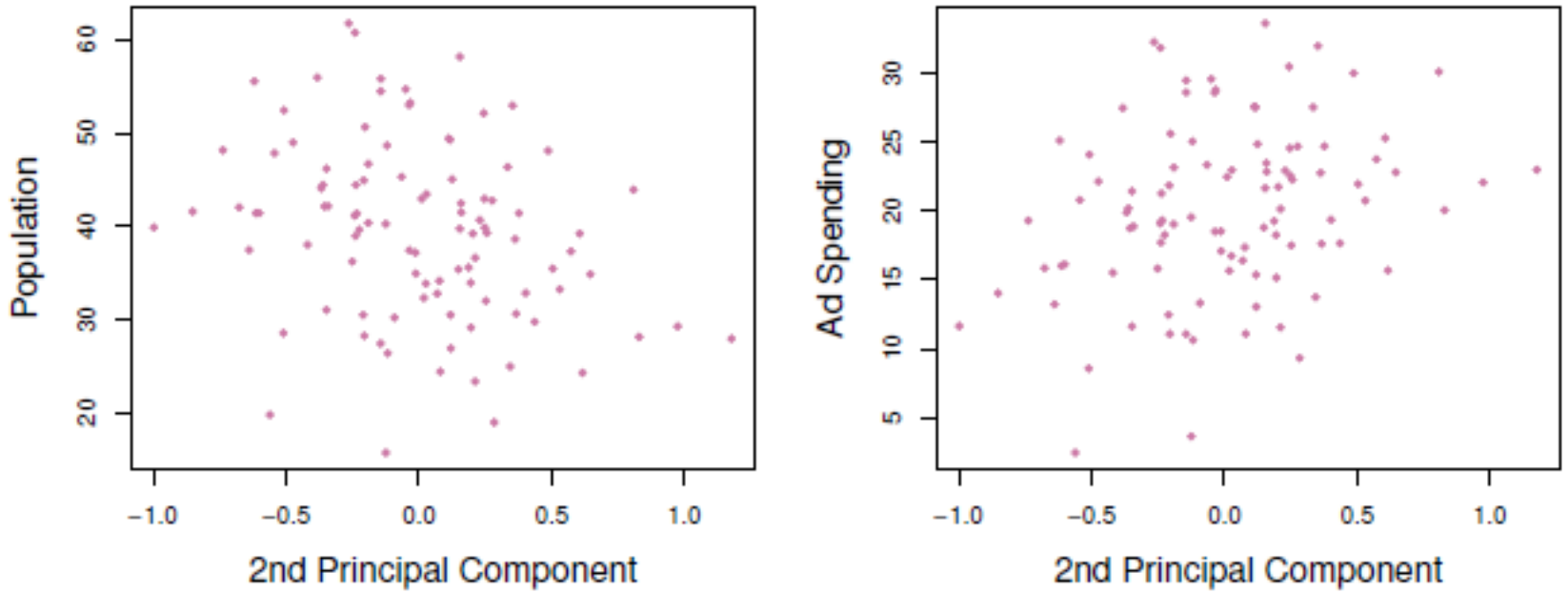
Principal Components Regression





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

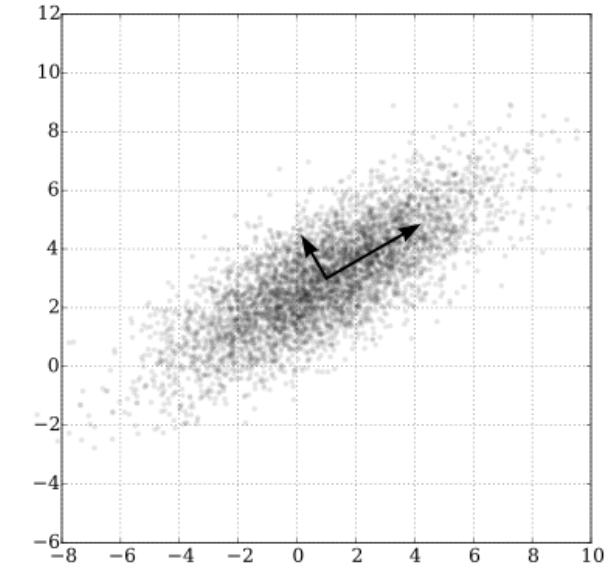
Principal Components Regression

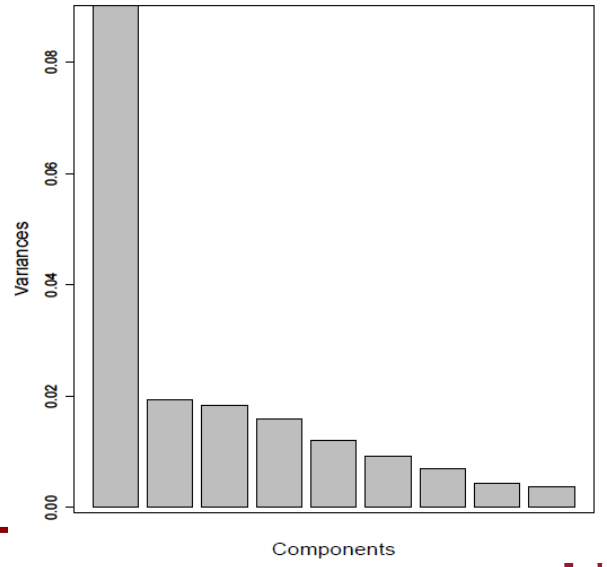


Plots of the second principal component scores z_{i2} versus pop *and* ad. *The relationships are weak.*

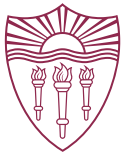
How many components?

- Each component is a linear combination of the existing features
  - Not interpretable
  - Computation:
    - deep mathematics with eigen decomposition
    - Efficient algorithms
- How many components to use?
  - Ignore less significant components
  - Cross validation
  - Information loss



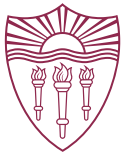


Principal Components Regression



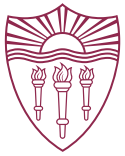
- As more principal components are used in the regression model, the bias decreases but the variance increases.
- PCR will tend to do well in cases when the first few principal components are sufficient to capture most of the variation in the predictors as well as the relationship with the response.
- We note that even though PCR provides a simple way to perform regression using $M < p$ predictors, it *is not* a feature selection method.
- In PCR, the number of principal components is typically chosen by cross-validation.

Partial Least Squares



- PCR identifies linear combinations, or *dimensions*, that best represents the predictors.
- These dimensions are identified is 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.
- PCR suffers from a potentially serious drawback: there is no guarantee that the dimensions that best explain the predictors will also be the best dimensions to use for predicting the response.

Partial Least Squares



- Like PCR, *partial least squares* (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.
- Then PLS fits an OLS linear model using these *M* new features.
- Unlike PCR, PLS identifies these new features in a *supervised* way; PLS 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*.
- The PLS approach attempts to find dimensions that help explain both the response and the predictors.

Partial Least Squares

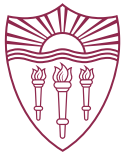
- First, standardize/normalize *p* predictors. Then, computes the first partial least squares dimension Z_1 by setting each ϕ_{1j} in

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

equal to the coefficient from the simple linear regression of *Y* onto X_j.

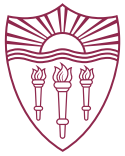
- One can show that this coefficient is proportional to the correlation between *Y* and X_j.

Partial Least Squares



- By 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 dimensions are found by taking residuals and then repeating the above prescription.
- As with PCR, the number *M* of PLS directions used in PLS is a tuning parameters that is typically chosen by cross-validation.
- While the supervised dimension reduction of PLS can reduce bias, it also has the potential to increase variance.

Considerations in High Dimensions

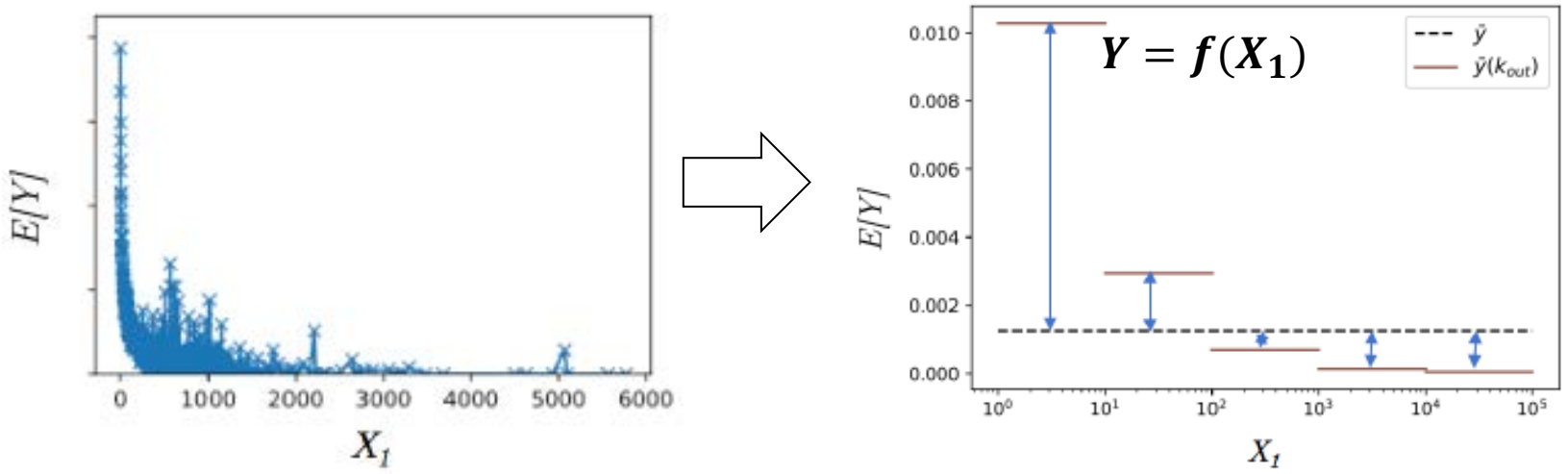


- While *p* can be extremely large, the number of observations *n* is often limited due to cost, sample availability, etc.
- Data sets containing more features than observations are often referred to a *high-dimensional*.
- When the number of features *p* is as large as, or larger than, the number of observations *n*, OLS should not be performed.
  - It is too *flexible* and hence overfits the data.
- Forward stepwise selection, ridge regression, lasso, and PCR are particularly useful for performing regression in the high-dimensional setting.

Structured Sum of Squares Decomposition* (S3D)

- Supervised non-linear feature selection
  - Successively picks features that collectively "best" explain an outcome variable
  - Creates a non-linear model of data

$$R^2 = \frac{\sum_{g=1}^{G} N_g \left(\bar{y}_g - \hat{y}\right)^2}{\sum_{i=1}^{N} (y_i - \hat{y})^2}$$



*https://epjds.epj.org/articles/epjdata/abs/2019/01/13688_2019_Article_201/13688_2019_Article_201.html

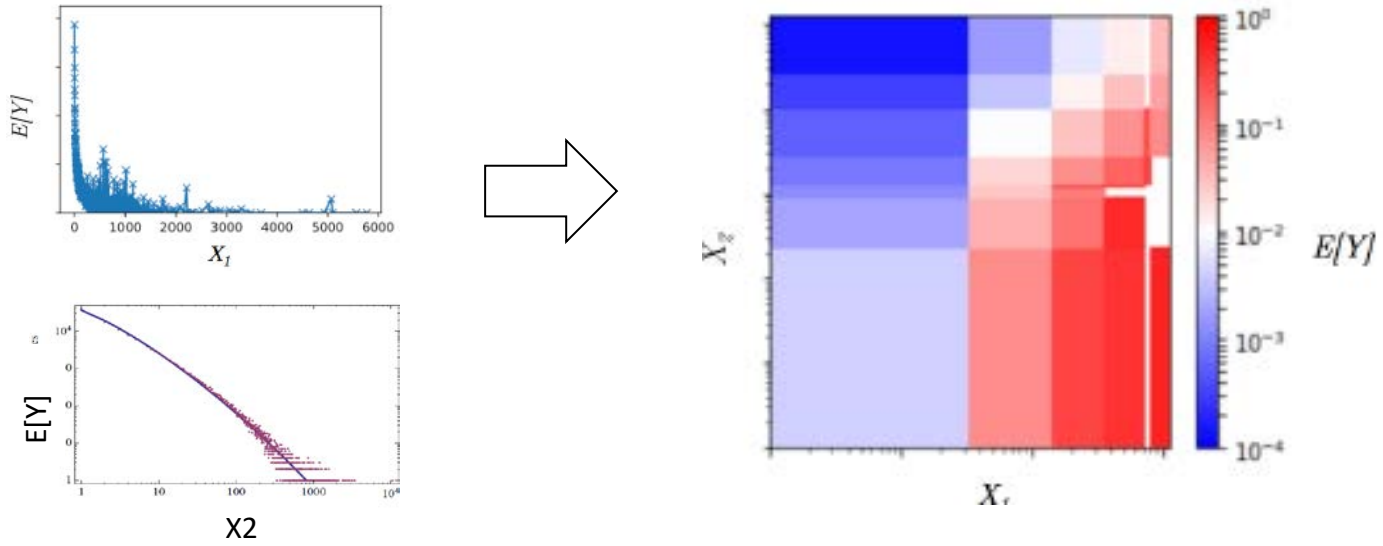
Structured Sum of Squares Decomposition (S3D)

https://github.com/peterfennell/S3D

- Successively picks features that collectively "best" explain an outcome variable
- Creates a non-linear model of data useful for visualization and prediction

$$R^2 = \frac{\sum_{g=1}^{G} N_g (\bar{y}_g - \hat{y})^2}{\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

$$\boldsymbol{Y = f(X_1, X_2)}$$

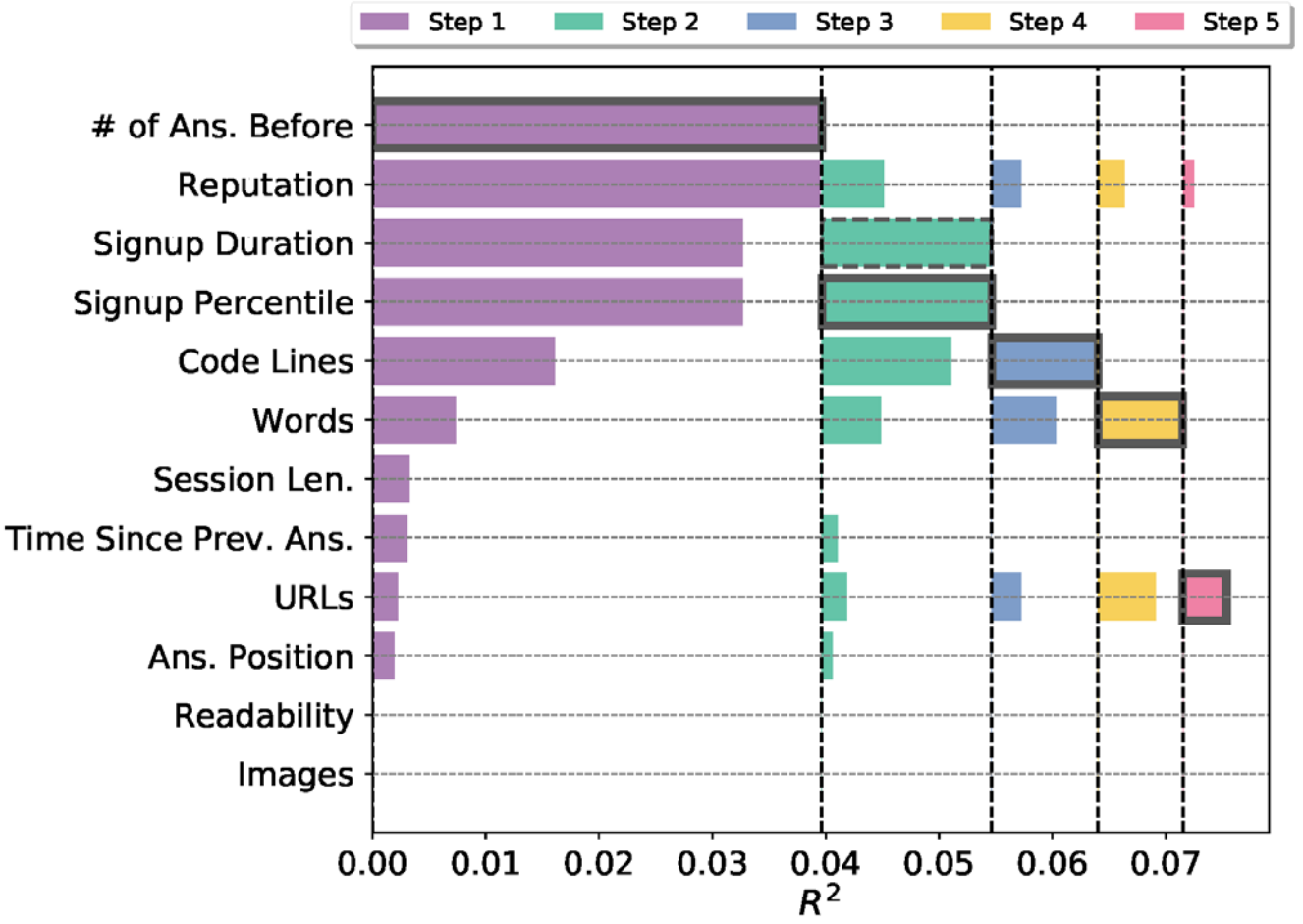


Color gives the average value of outcome in each bin.

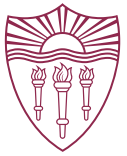
← Nonlinear approximation of data

S3D illustration on StackOverflow data

Identifying important features explaining whether user's answer is accepted as best answer to the question (Y=binary outcome)

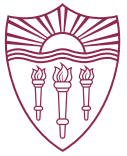


Considerations in High Dimensions



- Regularization or shrinkage plays a key role in high-dimensional problems.
- Appropriate tuning parameter selection is crucial for good predictive performance.
- The test error tends to increase as the dimensionality of the problem (i.e. the number of features or predictors) increases, unless the additional features are truly associated with the response.
  - Known as the *curse of dimensionality*

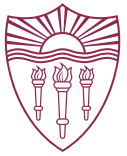
Considerations in High Dimensions



- *Curse of dimensionality*
  - Adding additional *signal* features that are truly associated with the response will improve the fitted model, in the sense of leading to a reduction in test set error.
  - Adding *noise* features that are not truly associated with the response will lead to a deterioration in the fitted model, and consequently an increased test set error.

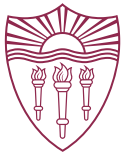
- Noise features increase the dimensionality of the problem, exacerbating the risk of overfitting without any potential upside in terms of improved test set error.

Considerations in High Dimensions



- In the high-dimensional setting, the multicollinearity problem is extreme: any variable in the model can be written as a linear combination of all of the other variables in the models.
- It is also important to be particularly careful in reporting errors and measures of model fit in the high-dimensional setting.
- One should *never* use sum of squared errors, p-values, R^2 statistics, or other traditional measures of model fit on the *training data* as evidence of good model fit in the high-dimensional setting.
- It is important to report results on an independent test set, or cross-validation errors.

Summary



- Best subset selection and stepwise selection methods.
- Estimate test error by adjusting training error to account for bias due to overfitting.
- Estimate test error using validation set approach and cross-validation approach.
- Ridge regression and the lasso as shrinkage (regularization) methods.
- Principal components regression and partial least squares.
- Considerations for high-dimensional settings.

Looking ahead

- Next week: Classification

- Virtual office hour
- https://usc.zoom.us/j/95136500603?pwd=VEJhblhWK25lT2N3RC9FNWk3eTJKQT09