Curse(s) of dimensionality

more information is better than less, right?

There is such a thing as too much of a good thing.

High dimensionality data

- Has a lot of features/attributes *p*
  e.g. we may have *n* = 1,000 subjects and *p* = 200,000 single-nucleotide polymorphisms (SNPs).
- Quite often $n<p$ or $n<<p$

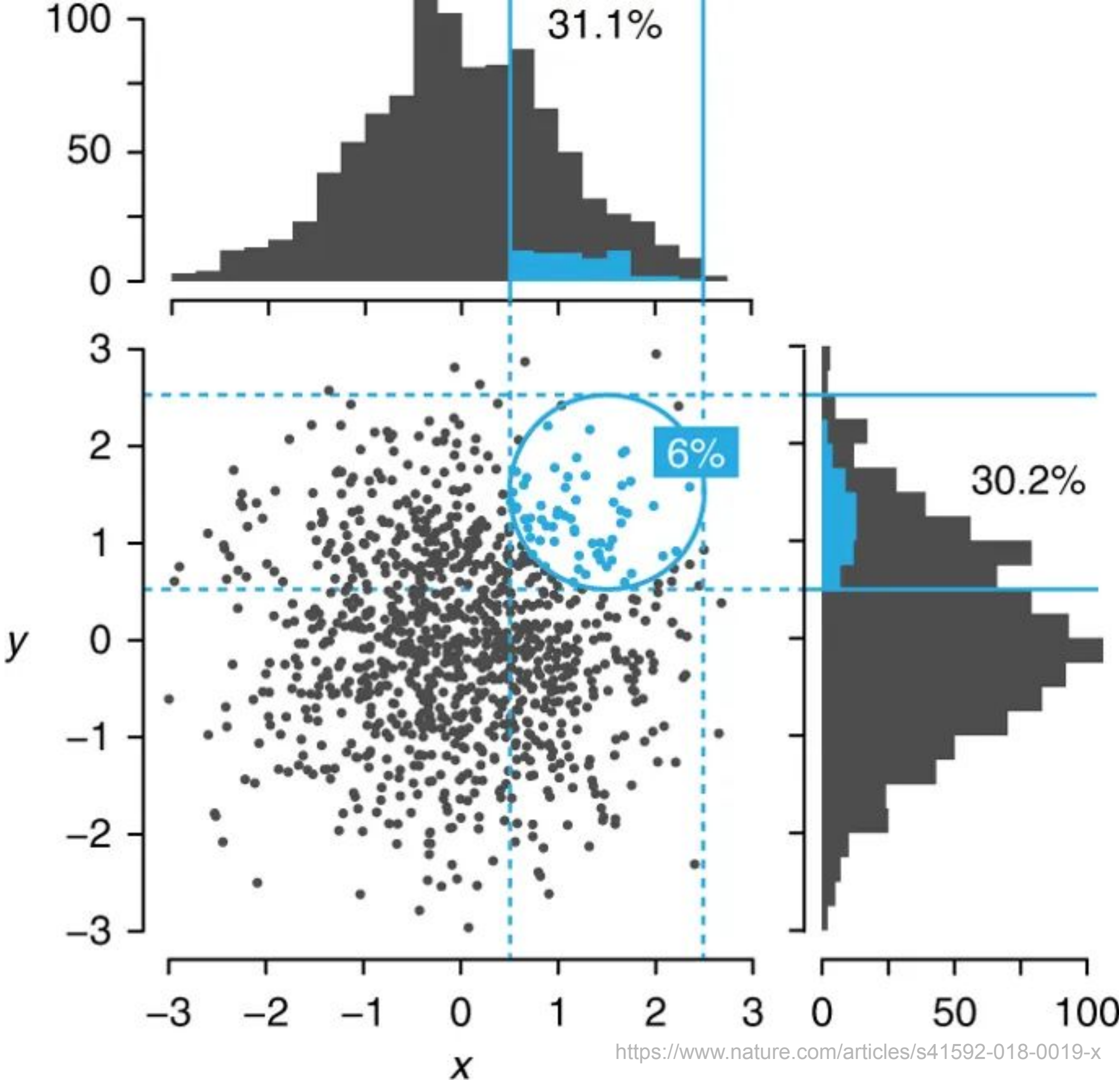
High dimensionality data

- Has a lot of features/attributes *p*
  e.g. we may have *n* = 1,000 subjects and *p* = 200,000 single-nucleotide polymorphisms (SNPs).
- Quite often n<p or n<<p

Main “curse” of dimensionality -- **data sparsity**

Data sparsity

As the dimensionality *p* increases, the ‘volume’ that the samples may occupy grows rapidly



https://www.nature.com/articles/s41592-018-0019-x

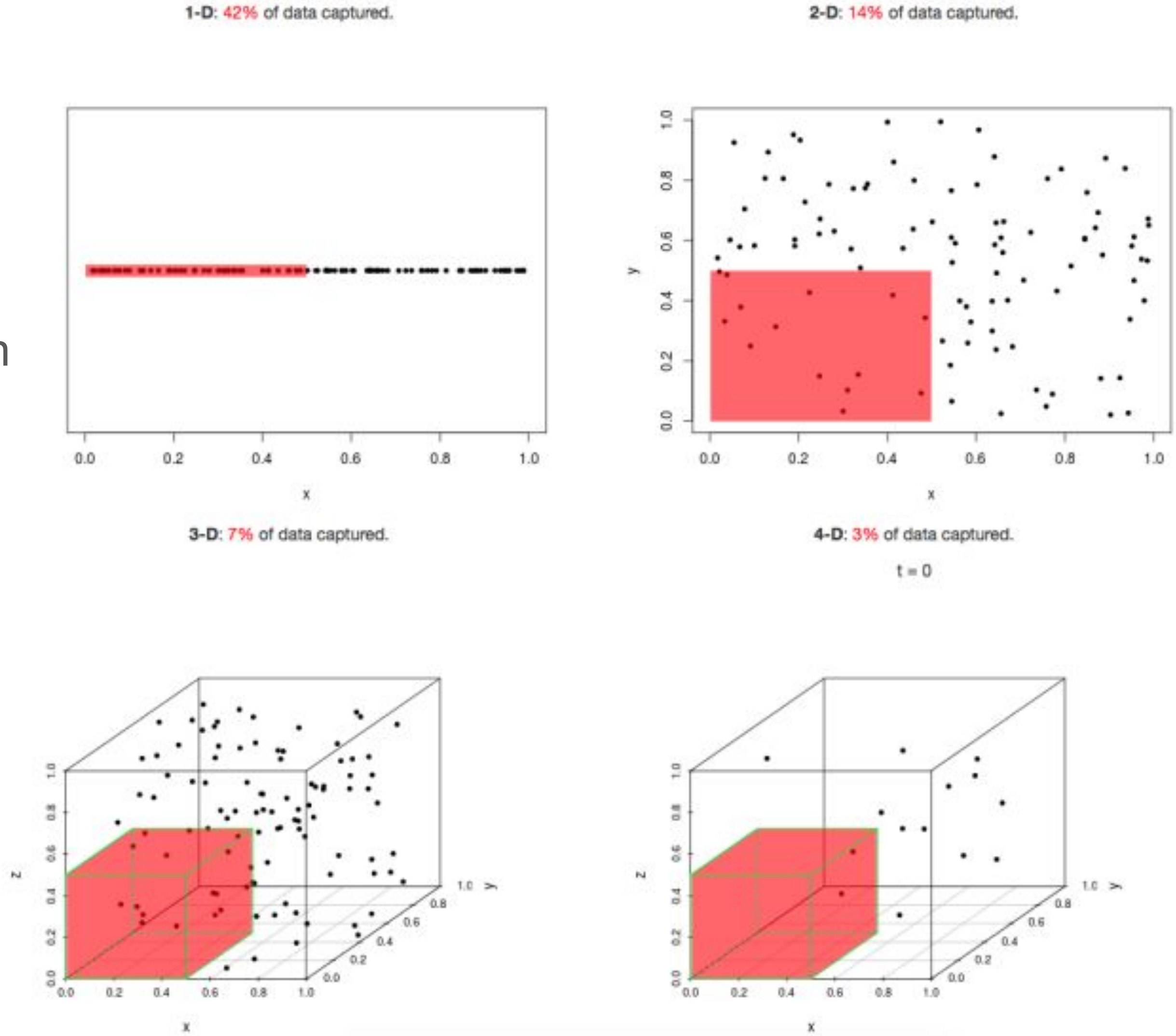
Distance measure

If

we treat the distance between points (e.g., Euclidean distance) as a measure of similarity,

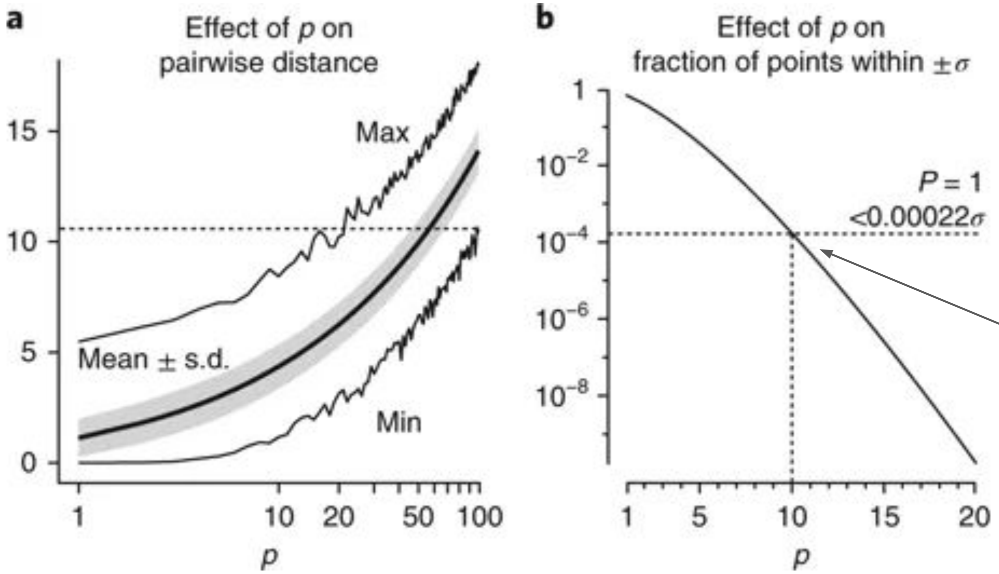
then:

greater distance --> greater dissimilarity



https://eranraviv.com/curse-of-dimensionality/

Distance measurement as *p* increases



Outlier detection becomes difficult
e.g. observation is further than 3 SD from mean

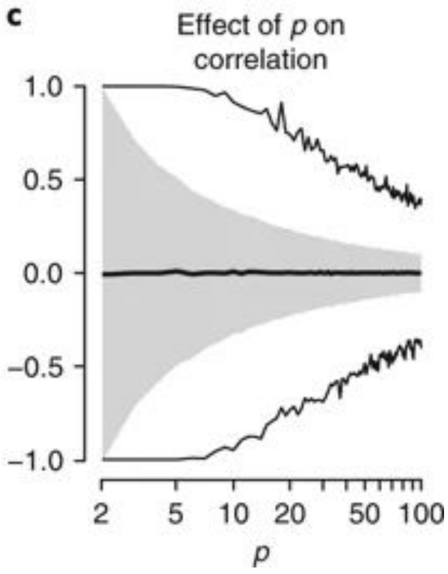
σ at *p* = 10 are as rare as points outside of 3.8*σ* at *p*=1

The average pairwise **distance** between two points **increases**

fraction of points within *σ* of the mean **drops** rapidly with increasing *p*

https://www.nature.com/articles/s41592-018-0019-x

correlation as *p* increases



correlation between two random vectors **decreases in range**

As the number of variables increases, the number of subjects in each set of categories decreases

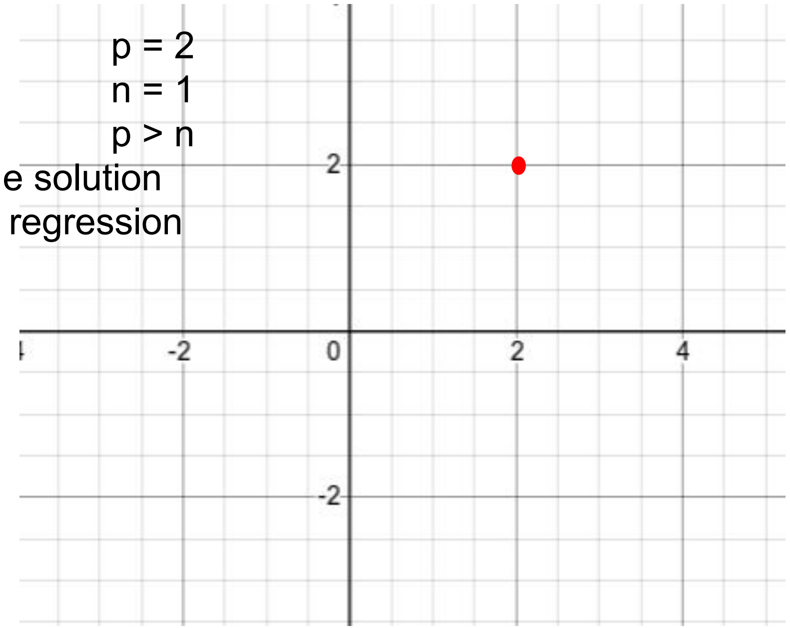
https://www.nature.com/articles/s41592-018-0019-x

When $p>n$ → overfitting

When p > n, there is no longer a unique least squares coefficient estimate.

The variance is infinite so the method cannot be used at all

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

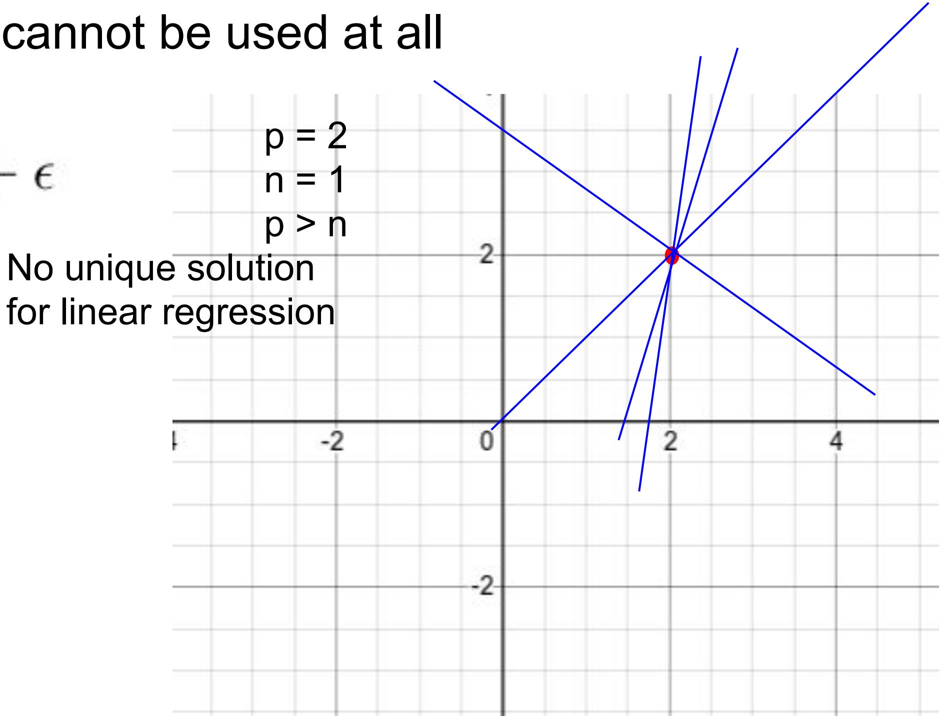


No unique solution
for linear regression

When $p>n$ → overfitting

When p > n, there is no longer a unique least squares coefficient estimate.
The variance is infinite so the method cannot be used at all

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



High dimensionality data

“Curses” of dimensionality:

- **data sparsity,**
- overfitting,
- multicollinearity,
- multiple testing

Effects are amplified by poor data quality.

Quality may increase with the number of variables.

Linear Model Selection and Regularization

(p > n case)

Several strategies what to do when p > n

- Subset Selection
- Shrinkage (also known as regularization)
- Dimension Reduction

Several strategies what to do when p > n

- Subset Selection -- we select a subset of the p predictors that we believe to be related to the response

- Shrinkage (also known as regularization) -- fit model using all predictors, but shrink some of the estimates to zero (or near zero)

- Dimension Reduction -- project p predictors into a M-dimensional subspace, where M < p and then use M projections as predictors to fit the model

Subset selection

1. Best Subset Selection
2. Forward Stepwise Selection
3. Backward Stepwise Selection

Best subset selection

Try all possible models (2^p) and select 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.

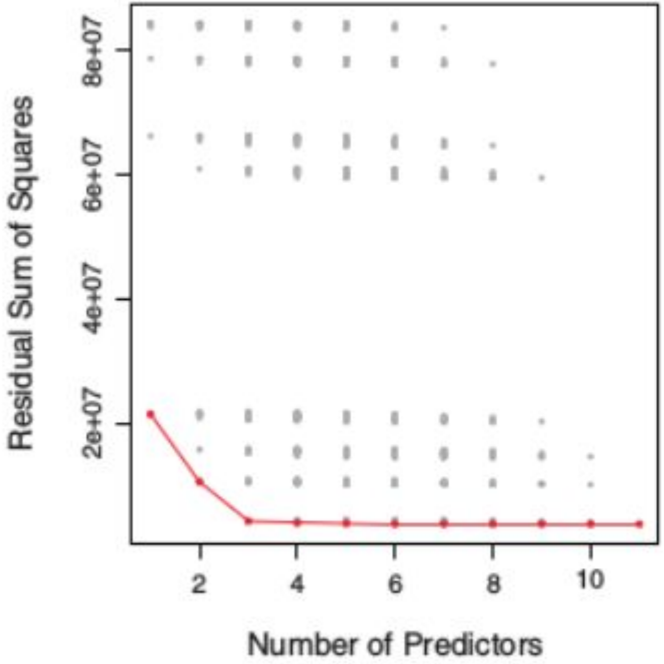
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.

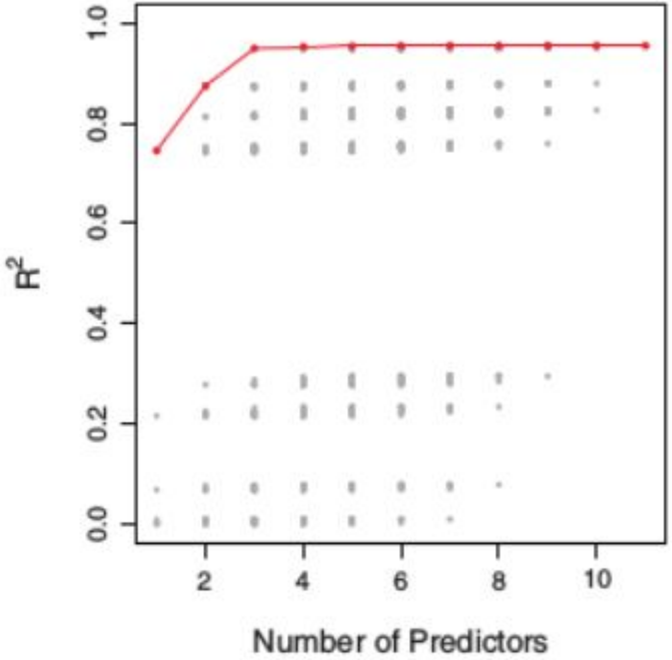
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Best subset selection

10 variables, 1 of which is categorical, so 2 dummy variables are created





Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Comparing models with different numbers of predictors

Problem: Training error decreases as more variables are added to the model and does not represent test error well

Best model selection with respect to test error

indirect

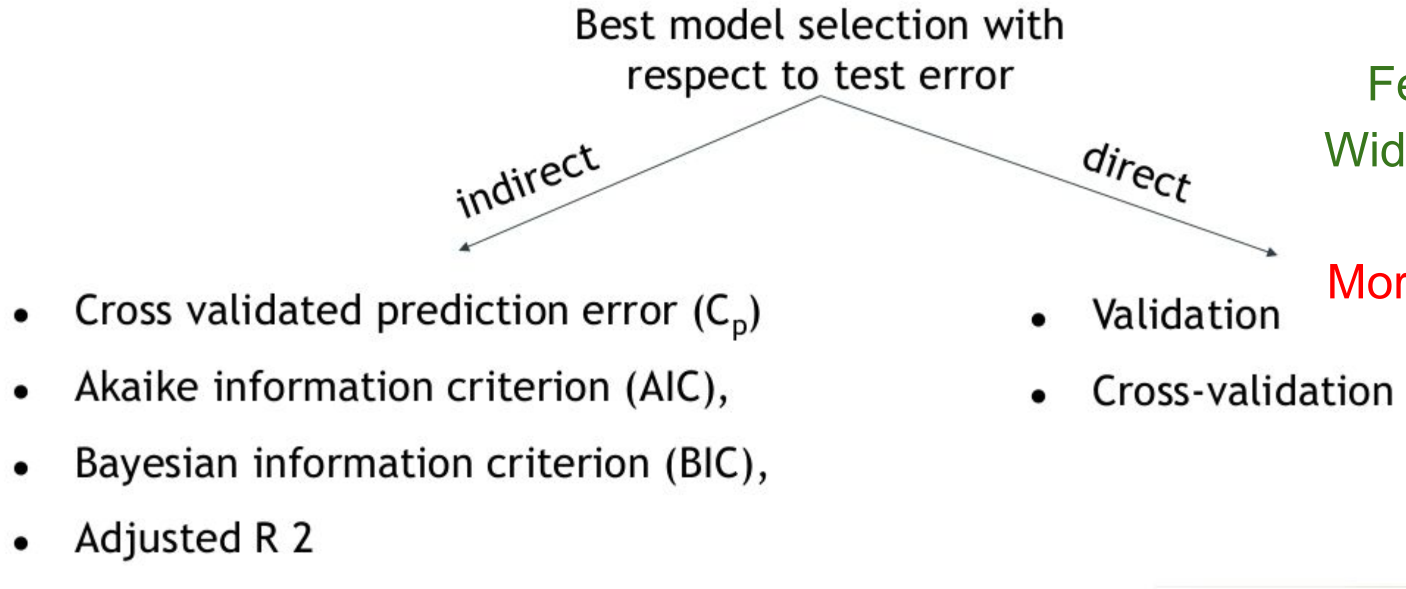
- Cross validated prediction error (C_p)
- Akaike information criterion (AIC),
- Bayesian information criterion (BIC),
- Adjusted R 2

direct

- Validation
- Cross-validation

Comparing models with different numbers of predictors

Problem: Training error decreases as more variables are added to the model and does not represent test error well



Direct estimate

Fewer assumptions

Wider range of model selection tasks

More computationally intensive

Forward subset selection

Adds predictors to the model one at a time. The variable that gives the greatest additional improvement to the fit is added to the model

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.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Forward subset selection

Adds predictors to the model one at a time. The variable that gives the greatest additional improvement to the fit is added to the model

# 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

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Backward subset selection

Begins with full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time

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.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Backward subset selection

Begins with full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time

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.

Needs n > p to fit full model :(

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Hybrid Stepwise Selection

1. Variables are added to model sequentially (like forward selection)
2. After adding each new variable, any variables that no longer provides an improvement in the model fit may also be removed (like backward selection)

Shrinkage (regularization)

1. Ridge Regression
2. Lasso Regression
3. Elastic nets

Ridge regression

To fit linear regression model -- minimize:

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

To fit ridge regression model -- minimize:

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

Ridge regression

To fit linear regression model -- minimize:

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

λ ≥ 0 is a tuning parameter

To fit ridge regression model -- minimize:

Shrinkage penalty

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

Ridge regression

$\lambda \geq 0$ is a tuning parameter

To fit ridge regression model -- minimize:

Shrinkage penalty

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

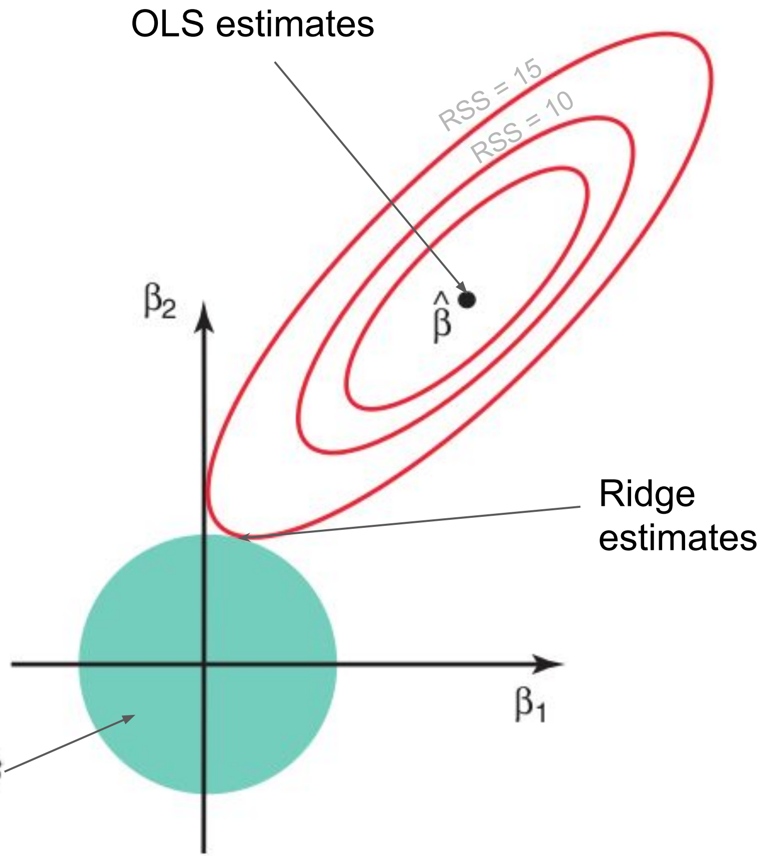
Shrinkage penalty is small, when $\beta_1,\ldots,\beta_p$ are close to zero

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

Ridge regression

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

$$\mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$



Let’s say we have only 2 predictors β_1 and β_1

$$\lambda(\beta_1^2 + \beta_2^2)$$

$$\beta_1^2 + \beta_2^2 \leq s$$

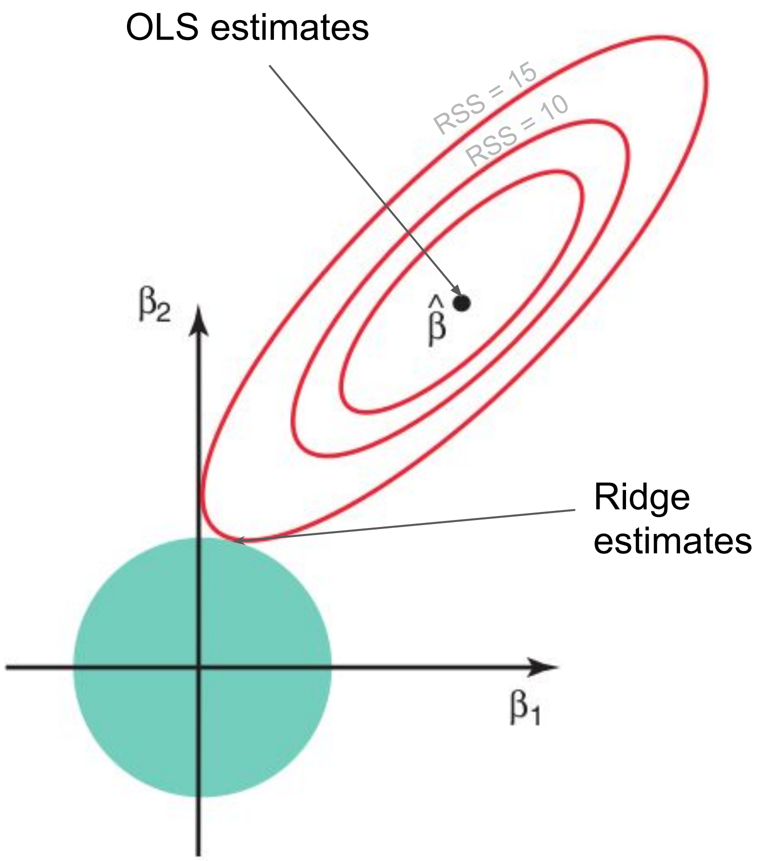
large value of s corresponds to λ = 0

Ridge regression

$$\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Introduces bias,
- may significantly decrease the variance of the estimates.

If variance effect is larger, this would decrease the test error

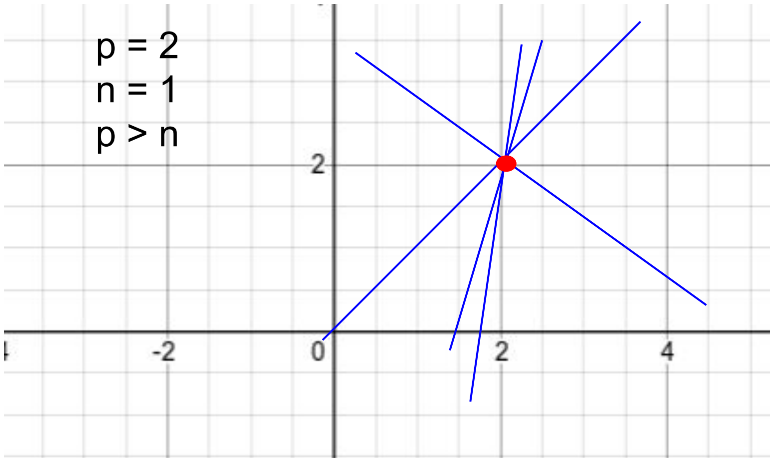


Ridge regression

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Introduces bias,
- may significantly decrease the variance of the estimates.

If variance effect is larger, this would decrease the test error



Ridge and Lasso regression

To fit ridge regression model -- minimize:

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

To fit lasso regression model -- minimize:

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

Ridge and Lasso regression

To fit ridge regression model -- minimize:

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

ℓ_2 penalty

To fit lasso regression model -- minimize:

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

ℓ_1 penalty

Lasso regression

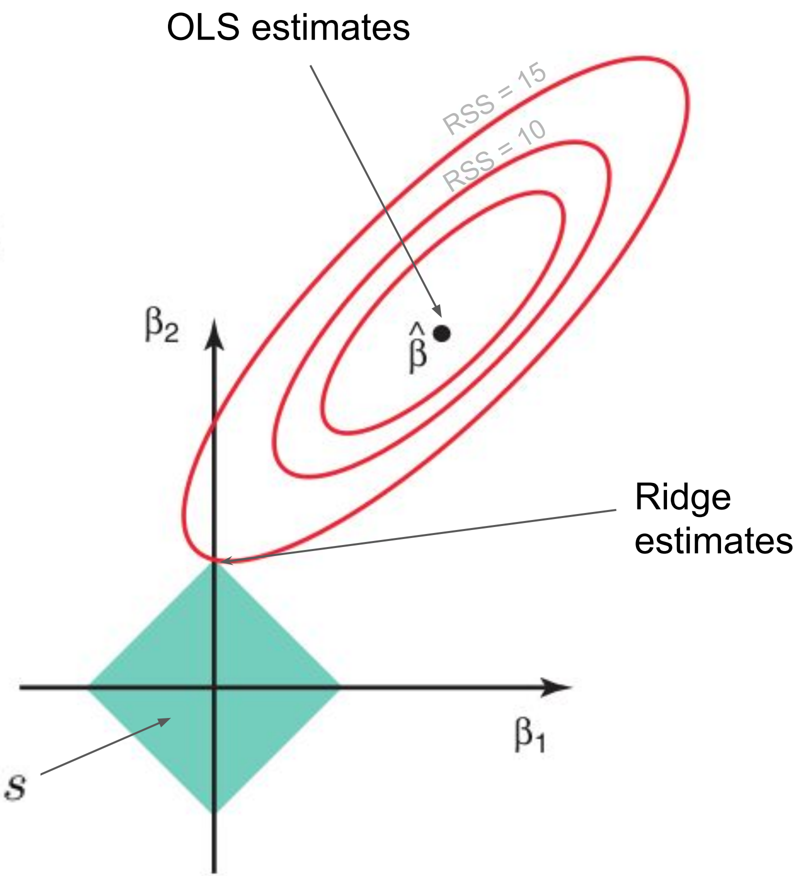
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \ =$$

$$\text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

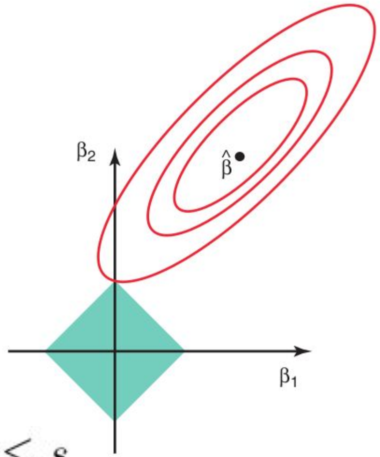
Let's say we have only 2 predictors β_1 and β_1

$$\lambda(|\beta_1| + |\beta_2|)$$

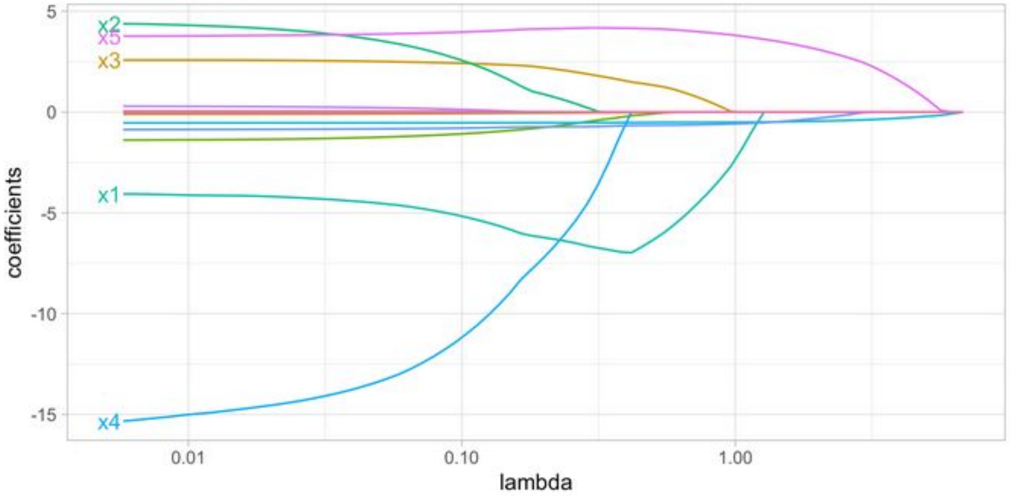
$$|\beta_1| + |\beta_2| \leq s$$

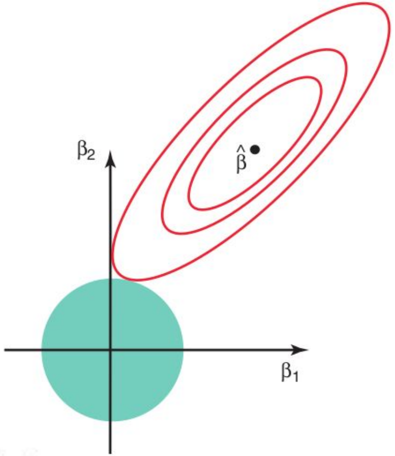


large value of s corresponds to λ = 0

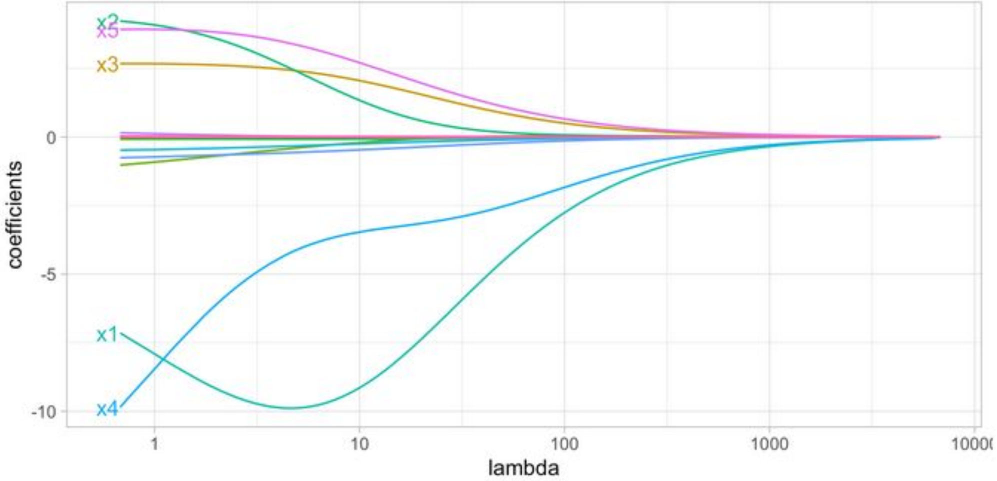


$$|\beta_1| + |\beta_2| \leq s$$





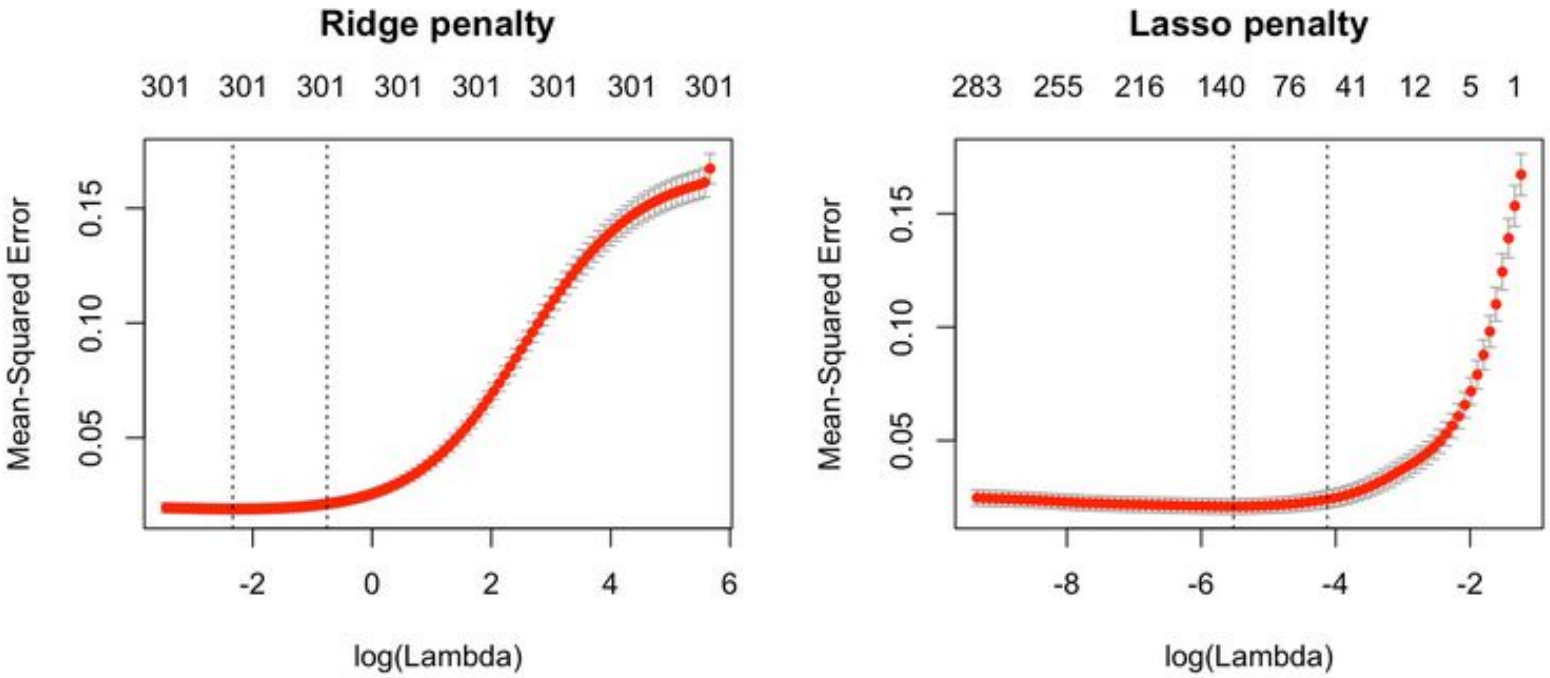
$$\beta_1^2 + \beta_2^2 \leq s$$



https://bradleyboehmke.github.io/HOML/regularized-regression.html

How to choose λ ?

How to choose λ ?

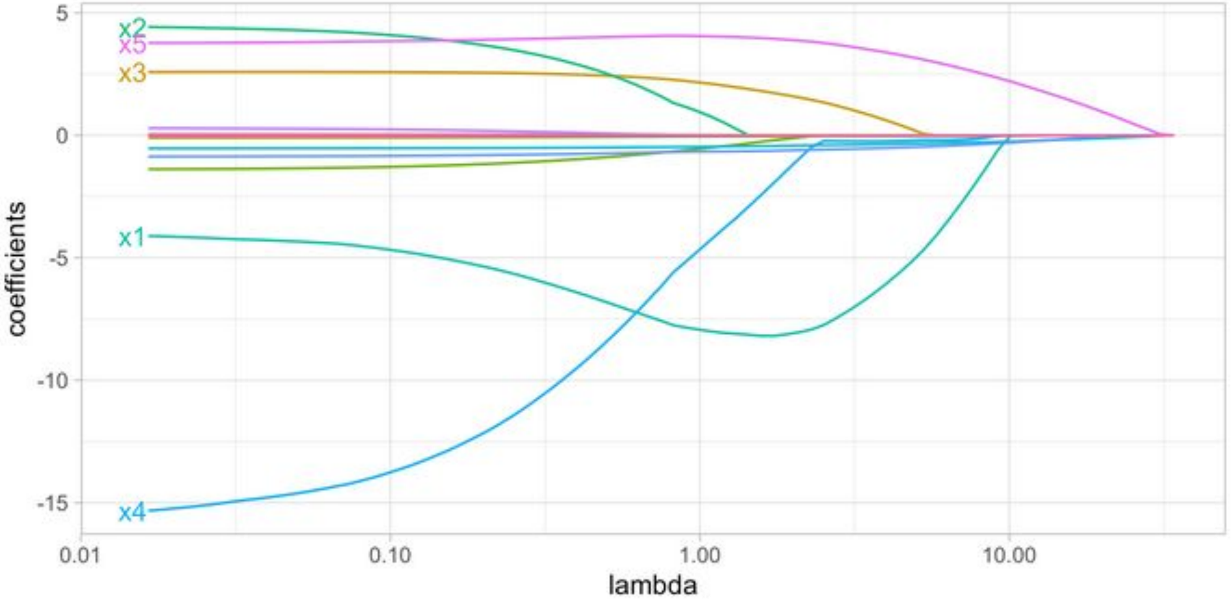


https://bradleyboehmke.github.io/HOML/regularized-regression.html

Elastic net

Combines Lasso and Ridge at the same time

$$\text{minimize}\left(\text{RSS} + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|\right)$$



https://bradleyboehmke.github.io/HOML/regularized-regression.html

https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html

https://bradleyboehmke.github.io/HOML/regularized-regression.html

Dimension reduction -- PCA

Next day...