

Chapter 6: Linear Model Selection & Regularization

In the regression setting, the standard linear model is commonly used to describe the relationship between a response Y and a set of variables X_1, \dots, X_p .

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$$

typically fit w/ least squares

Upcoming: more general models (non-linear)

The linear model has distinct advantages in terms of inference and is often surprisingly competitive for prediction. How can it be improved?

replace least squares w/ alternative fitting procedures.

We can yield both better prediction accuracy and model interpretability:

prediction accuracy: if true relationship is \approx linear \Rightarrow least squares will have low bias.

If $n \gg p \Rightarrow$ also have low variance \Rightarrow perform well on test data!

If n not much larger than $p \Rightarrow$ high variability \Rightarrow poor performance.

If $p > n \Rightarrow$ no longer have a unique solution \Rightarrow variance = $\infty \Rightarrow$ cannot be used at all!

goal: reduce variance without adding too much bias.

model interpretability: often many variables used in a regression are not associated w/ response.

By removing (setting $\hat{\beta}_i = 0$), we can obtain a more easily interpretable model.

Note: least squares will hardly ever result in $\hat{\beta}_i = 0$.

\Rightarrow need variable selection.

Same ideas apply to logistic regression.

1 Subset Selection

We consider methods for selecting subsets of predictors.

1.1 Best Subset Selection.

To perform *best subset selection*, we fit a separate least squares regression for each possible combination of the p predictors. $\binom{p}{2} = \frac{p(p-1)}{2}$ models with exactly 2 predictors, etc.

Algorithm:

1. Let M_0 denote the model with no predictors.
2. For $k=1, \dots, p$
 - (a) Fit all $\binom{p}{k}$ models that contain k predictors.
 - (b) Pick the best of those (call it M_k). "Best" is defined by $\downarrow \text{RSS}$ ($\uparrow R^2$).
3. Select a single best model from M_0, \dots, M_p using CV error, C_p , AIC/BIC, or adjusted R^2 traditional metrics, more later.

Why can't we use R^2 for step 3? as $p \uparrow$, $R^2 \uparrow$ always. Why might we not want to do this at all? computation
We can perform something similar with logistic regression. Fitting 2^p models! $p=10 \Rightarrow 1000$ models.

1.2 Stepwise Selection

For computational reasons, best subset selection cannot be performed for very large p . \rightarrow "impossible" with $p \geq 40$.

Best subset may also subset when p large because w/ a large search space can find good models in training that perform poorly on test data
 \Rightarrow high variability & overfitting can occur.

Stepwise selection is a computationally efficient procedure that considers a much smaller subset of models.

Forward Stepwise Selection: start with no predictors and add one predictor at a time until all predictors are in the model, choose the "best" from these.

1. Let M_0 denote the null model - no predictors
2. For $k=0, \dots, p-1$
 - (a) Consider $p-k$ models that augment the predictors in M_k w/ 1 additional predictor.
 - (b) Choose the best among these $p-k$ and call it M_{k+1} ($\uparrow R^2$).
3. Select a single best model from M_0, \dots, M_p using CV error, C_p , AIC/BIC, or adjusted R^2 . 2

Now we fit $1 + \sum_{k=0}^{p-1} \binom{p-k}{1} = 1 + \frac{p(p+1)}{2}$ models!

Backward Stepwise Selection: Begin w/ full model and take predictors away one at a time until you get to the null model.

1. Let M_p denote the full model, contains all predictors.

2. For $k = p, p-1, \dots, 1$:

(a) consider all models (k) that contain all but one of the predictors in M_k (k-1 predictors).

(b) Choose the best among them, call it M_{k-1} ($\uparrow R^2$).

3. Select the single best model from M_0, \dots, M_p using CV, C_p , AIC/BIC, or adjusted R^2 .

* Neither forward nor backwards stepwise selection are guaranteed to find the best model containing a subset of the p predictors.

Forward Selection can be used when $p > n$ (but only up to $n-1$ predictors (not up to p)).

1.3 Choosing the Optimal Model

Best subset, forward selection, backward selection all need a way to pick the "best" model - according to test error.

• $RSS + R^2$ are proxy for training error \Rightarrow not estimates of test error

$$\textcircled{2} C_p = \frac{1}{n} \left(\underbrace{RSS}_{\text{subset model}} + 2d \hat{\sigma}^2 \right)$$

\uparrow estimate of variance of ε (full model).
 \uparrow # predictors in subset model

\rightarrow ① estimate this directly (CV) or

② adjust training errors for model size.

adds a penalty to training error (RSS) to adjust for underestimation of test error.

(choose model w/ lowest value).

② AIC & BIC Low qit for models fit w/ MLE

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

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

choose model w/ lowest AIC or BIC. $\uparrow \log(n) \approx 2$ for $n > 7 \Rightarrow$ heavier penalty on models w/ many variables \Rightarrow results in smaller models.

② Adjusted R^2 (only for least squares).

$$R^2 = 1 - \frac{RSS}{TSS} \quad \text{Always } \uparrow \text{ as } d \uparrow$$

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

choose model w/ highest Adj R^2 .

① Validation and Cross-Validation

Directly estimate test error w/ Validation or CV and choose model w/ lowest estimated error.

Very general (can be used for any model) even when it's not clear how many "predictors" we have.

Now have fast computers \Rightarrow these are preferred.

2 Shrinkage Methods

The subset selection methods involve using least squares to fit a linear model that contains a subset of the predictors. As an alternative, we can fit a model with all p predictors using a technique that constrains (regularizes) the estimates.

↳ shrinks estimates towards zero.

Shrinking the coefficient estimates can significantly reduce their variance!

Help us to avoid overfitting!

2.1 Ridge Regression

Recall that the least squares fitting procedure estimates β_1, \dots, β_p using values that minimize

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

residual sum of squares.

Ridge Regression is similar to least squares, except that the coefficients are estimated by minimizing

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

$\hat{\beta}^R$

note we are not penalizing β_0
we want to penalize the relationships, not the intercept
(mean value of response when $x_{i1} = \dots = x_{ip} = 0$).

↳ $\lambda > 0$ tuning parameter (determine separately from fitting).

trades off 2 criteria: minimize RSS to fit data well

minimize $\lambda \sum_{j=1}^p \beta_j^2$ "shrinkage penalty" will be small when β_j close to zero \Rightarrow shrink estimates towards zero.

The tuning parameter λ serves to control the impact on the regression parameters.

When $\lambda = 0$, penalty has no effect \Rightarrow ridge regression = least squares.

As $\lambda \rightarrow \infty$, impact of penalty grows $\Rightarrow \hat{\beta}^R \rightarrow 0$.

Ridge regression will produce a different set of coefficients for each penalty ($\hat{\beta}_\lambda^R$).

4

Selecting a good λ is critical! How to choose? Cross validation!

The standard least squares coefficient estimates are scale invariant.

Multiply X_j by a constant c leads to a scaling of OLS best estimate by a factor of $\frac{1}{c}$.

\Rightarrow regardless of how the j^{th} predictor is scaled $X_j \hat{\beta}_j$ will remain the same.

In contrast, the ridge regression coefficients $\hat{\beta}_\lambda^R$ can change substantially when multiplying a given predictor by a constant.

e.g. say we have an income variable in ① dollars vs. ② thousands of dollars

due to the sum of squared coef. term, this change will not simply cause the coefficient to change by a factor of 1000.

$\Rightarrow X_j \hat{\beta}_{j,\lambda}^R$ depends not only on λ , but also on the scaling of X_j

(may even depend on scaling of other predictors)

Therefore, it is best to apply ridge regression after standardizing the predictors so that they are on the same scale:

i.e. have standard deviation of one.

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

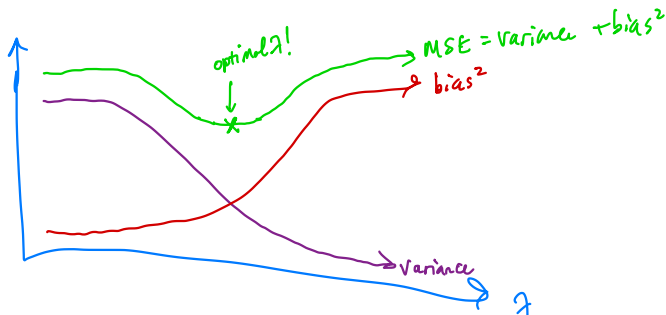
st. dev. of j^{th} predictor.

- ① standardize data
- ② tune model to choose λ (via cross validation).
- ③ fit ridge regression on standardized data w/ chosen λ .

Why does ridge regression work?

Because of the bias-variance trade-off!

As $\lambda \uparrow$, the flexibility of the ridge regression fit \downarrow
 \Rightarrow variance \downarrow and bias \uparrow



In situations where relationship between response + predictors is \approx linear
 OLS will have low bias.

when p almost as large as $n \Rightarrow$ OLS will have high variability!
 if $p > n$ least squares doesn't even have a unique solution.

ridge regression can still perform well in these scenarios by trading off a small amount of bias for a decrease in variance.

\hookrightarrow Ridge regression works best in high variance scenarios.

Also:

Cost advantage over subset selection

because for fixed λ , only fitting one model! (very fast model to fit).

Ridge regression improves predictive performance.

Does it also help us with interpretation?

2.2 The Lasso

Ridge regression does have one obvious disadvantage.

Unlike best subset, forward or backward selection, ridge regression will include all p variables in the final model.

penalty $\lambda \sum \beta_j^2$ will shrink all $\beta_j \rightarrow 0$ but $\beta_j \neq 0$ (unless $\lambda = \infty$)!

This may not be a problem for prediction accuracy, but it could be a challenge for model interpretation when p is very large.

We will always have all variables in the model, whether there is a relationship w/ Y or not.

"Least absolute shrinkage and selection operator"

The lasso is an alternative that overcomes this disadvantage. The lasso coefficients $\hat{\beta}_\lambda^L$ minimize

$$\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p |\beta_j| = \text{RSS} + \underbrace{\lambda \sum_{j=1}^p |\beta_j|}_{\text{L}_1 \text{ penalty}} \quad (\sum \beta_j^2 = \text{"L}_2 \text{ penalty"}).$$

As with ridge regression, the lasso shrinks the coefficient estimates towards zero.

L_1 penalty also has the effect of forcing some coefficients to be exactly zero when λ is sufficiently large.

\Rightarrow much like our selection methods, lasso performs variable selection!

As a result, lasso models are generally easier to interpret.

The lasso yields sparse models - models w/ only a subset of the variables.

Again selecting a good λ is critical. Use cross validation!

Why does the lasso result in estimates that are exactly equal to zero but ridge regression does not? One can show that the lasso and ridge regression coefficient estimates solve the following problems

↔
equivalent to
other formulations
w/ λ

lasso: minimize $\left\{ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 \right\}$ subject to $\sum_{j=1}^p |\beta_j| \leq S$

ridge: minimize $\left\{ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 \right\}$ subject to $\sum_{j=1}^p \beta_j^2 \leq S$

constraints

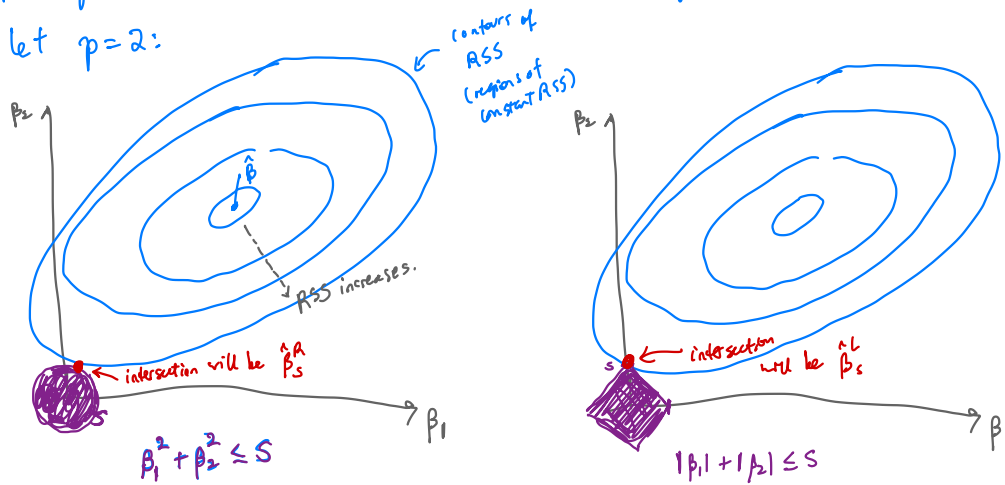
} constrained optimization problems.

In other words, when we perform the lasso we are trying to find the set of coefficient estimates that lead to the smallest RSS, subject to the constraint that there is a budget s for how large $\sum_{j=1}^p |\beta_j|$ can be.

When s is very large, this is not much of a constraint \Rightarrow coef estimates can be large (same for ridge).

But why does the lasso result in coefficient estimates exactly equal to 0?

let $p=2$:



Solution to lasso or ridge is the first point in the ellipses (RSS) contact the constraint region.

Ridge has a circular region \Rightarrow no sharp points, intersection won't generally occur on the axis.

Lasso corners on each axis \Rightarrow ellipse often intersects at the axis \Rightarrow one of the coefficients to equal zero.

If we believe there are predictors that do not have a relationship w/ Y (we just don't know which ones) lasso will perform better.

If not (everything is important), ridge will perform better.

Use CV to pick!

2.3 Tuning

We still need a mechanism by which we can determine which of the models under consideration is “best”.

For subset we have C_p , AIC/BIC, adjusted R^2 , CV error. ↙ equivalent

For both the lasso and ridge regression, we need to select λ (or the budget s).

How?

penalization
parameter

- ① scale data.
- ① choose a grid of λ values
- ② compute CV error (K-fold) for each λ .
- ③ select λ for which CV error is smallest
- ④ fit chosen model using all available observations and selected λ .

* Note * still important to scale variables x_1, \dots, x_p for lasso to have st. dev. = 1.

3 Dimension Reduction Methods

So far we have controlled variance in two ways:

- ① Using a subset of original variables
- best subset, forward/backward selection, lasso.
- ② shrinking coefficients towards zero.
- ridge regression, lasso.

These methods all defined using original predictor variables X_1, \dots, X_p .

We now explore a class of approaches that

- ① transform the predictors
- ② fit least squares using the transformed variables.

We refer to these techniques as dimension reduction methods.

- ① Let Z_1, \dots, Z_M represent $M < p$ linear combinations of our original predictors.

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

for constants $\phi_{1m}, \dots, \phi_{pm}$ $m=1, \dots, M$.

- ② fit the linear regression model using least squares

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

↑
regression coefficients

If choose ϕ_{jm} well, this can outperform least squares (w/ original data).

The term *dimension reduction* comes from the fact that this approach reduces the problem of estimating $p + 1$ coefficients to the problem of estimating $M + 1$ coefficients where $M < p$.

Dimension reduction serves to constrain β_j , since now they must take a particular form.

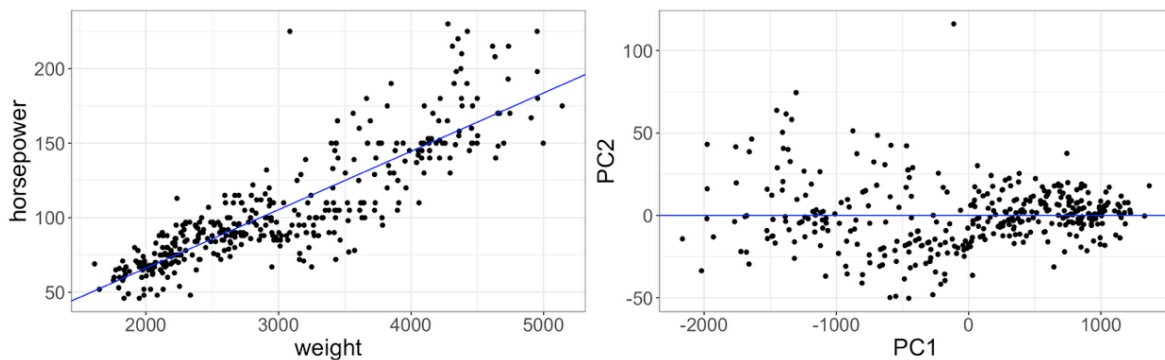
All dimension reduction methods work in two steps.

3.1 Principle Component Regression

Principal Components Analysis (PCA) is a popular approach for deriving a low-dimensional set of features from a large set of variables.

The *first principal component* directions of the data is that along which the observations vary the most.

We can construct up to p principal components, where the 2nd principal component is a linear combination of the variables that are uncorrelated to the first principal component and has the largest variance subject to this constraint.



The Principal Components Regression approach (PCR) involves

- 1.
- 2.

Key idea:

In other words, we assume that the directions in which X_1, \dots, X_p show the most variation are the directions that are associated with Y .

How to choose M , the number of components?

Note: PCR is not feature selection!

3.2 Partial Least Squares

The PCR approach involved identifying linear combinations that best represent the predictors X_1, \dots, X_p .

Consequently, PCR suffers from a drawback

Alternatively, *partial least squares (PLS)* is a supervised version.

Roughly speaking, the PLS approach attempts to find directions that help explain both the response and the predictors.

The first PLS direction is computed,

To identify the second PLS direction,

As with PCR, the number of partial least squares directions is chosen as a tuning parameter.

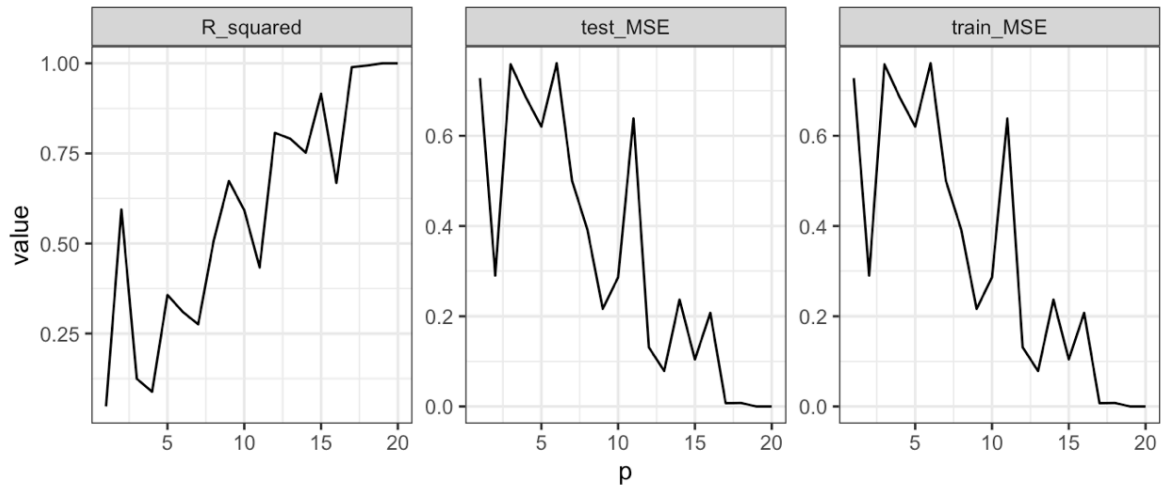
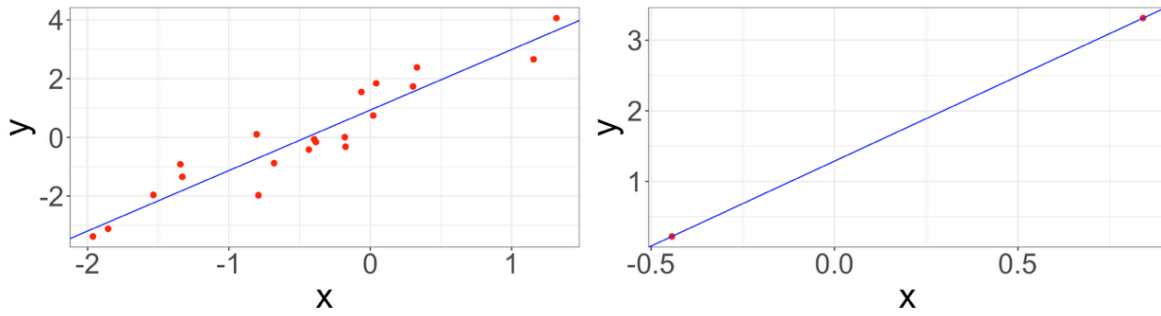
4 Considerations in High Dimensions

Most traditional statistical techniques for regression and classification are intended for the low-dimensional setting.

In the past 25 years, new technologies have changed the way that data are collected in many fields. It is not commonplace to collect an almost unlimited number of feature measurements.

Data sets containing more features than observations are often referred to as *high-dimensional*.

What can go wrong in high dimensions?



Many of the methods that we've seen for fitting *less flexible* models work well in the high-dimension setting.

1.

2.

3.

When we perform the lasso, ridge regression, or other regression procedures in the high-dimensional setting, we must be careful how we report our results.