

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 with alternative fitting procedures.

We can yield both better prediction accuracy and model interpretability:

prediction accuracy: if the true relationship is \approx linear, least squares will have low bias
if $n \gg p \Rightarrow$ also have low variance \Rightarrow perform well on test data.

But if n is not much larger than $p \Rightarrow$ high variability \Rightarrow poor performance.

if $p > n$: no longer 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 in fact associated with the response.

By removing them (setting $\hat{\beta}_i = 0$) we could 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. $\leftarrow \binom{p}{2} = \frac{p(p-1)}{2}$ models w/ exactly 2 predictors, etc.

Algorithm:

1. let M_0 denote null model - 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 (M_k). Best is defined by \downarrow RSS (\uparrow R²).

3. Select a single best model from M_0, \dots, M_p using CV error, C_p , AIC/BIC, or adjusted R² more later.

Why can't we use R² for step 3? as $p \uparrow$, R² \uparrow always. Why might we not want to do this? Fitting 2^p models!
We can perform something similar with logistic regression. $p=10 \Rightarrow 1000$ 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 suffer when p large because w/ a large search space

We can find spurious good models that work on training data but perform poorly w/ test data.

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

Forward Stepwise Selection:

Start w/ no predictors and add predictors one 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 all $p-k$ models that augment the predictors in M_k with 1 additional predictor.

(b) choose the best among those $p-k$ and call it M_{k+1} (\uparrow R², \downarrow RSS). ₂

3. Select a single best model from M_0, \dots, M_p using CV error, C_p , AIC/BIC, adj R².

- 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 we get to null model.

1. Let M_p denote the full model — contains all p predictors.
2. For $k = p, p-1, \dots, 1$:
 - (a) consider all k models that contain all predictors except one of predictors in M_k ($k-1$ predictors).
 - (b) choose best among them and call it M_{k-1} ($\uparrow R^2$, \downarrow RSS).
3. Select the single best model from M_0, \dots, M_p using CV error, C_p , AIC/BIC, $\text{adj } R^2$.

* 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 $p!$).

1.3 Choosing the Optimal Model

Need a way way to pick "best" model that depends on test error (training error not a good estimate of this)
 ↳ either estimate this directly or adjust training errors for model size.

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

\uparrow estimate of variance for full model.
 \uparrow # of predictors in subset model

add penalty to training error (RSS) to adjust for underestimate of test error.

as $d \uparrow$, $C_p \uparrow$ (choose model w/ lowest value).

AIC & BIC can use for maximum likelihood fits

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

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

choose model w/ low value, since $\log(n) > 2$ for $n \geq 7 \Rightarrow$ heavier penalty on models w/ many variables \Rightarrow results in smaller models.

Adjusted R^2 (least squares models)

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

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

choose model w/ highest $\text{adj } R^2$.

Validation and Cross-Validation

- Directly estimate test error w/ validation or CV and choose model w/ lowest est. error.
- very general, can be used w/ any model, even when it's not clear how many "predictors" we have.

Now have fast computers \Rightarrow these are preferred.

give us some answer

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.

↳ shrink 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

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

note we are not penalizing β_0
we want to penalize the relationships not the intercept
(mean value of response when $x_{i1} = \dots = x_{ip} = 0$)
 $\lambda \geq 0$ tuning parameter (determined separately of the fitting procedure).

trades off 2 criteria: minimize RSS to fit data well
 $\lambda \sum_{j=1}^p \beta_j^2$ shrinkage penalty, small when β_j 's close to zero \Rightarrow shrinks estimates towards zero.
The tuning parameter λ serves to control the impact on the regression parameters.

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

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

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

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

The standard least squares coefficient estimates are scale invariant.

Multiplying X_j by a constant c leads to a scaling of least squares estimates by a factor of $\frac{1}{c}$.

\Rightarrow regardless of how 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.

$$\textcircled{1} = 1000 \times \textcircled{2}$$

due to the sum of squared coef. terms this change will not simply result in the coefficient estimate 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 the 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. standard deviation of one.

$$\tilde{x}_{ij} = \frac{x_{ij}}{\underbrace{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}}_{\text{s.t. dev. of } j^{\text{th}} \text{ predictor.}}}$$

*nice in.
workflow, recipe*

① standardize data

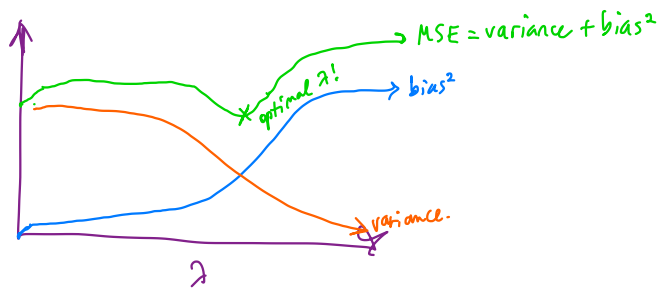
1.5 tune model to choose λ (using CV)

② fit ridge regression on training data using chosen λ .

Why does ridge regression work?

Because of the bias-variance trade-off!

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



In situations where relationship between response and predictors \approx linear
 least squares estimate will have low bias.

When p is almost as large as $n \Rightarrow$ least squares has 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 decrease in variance.

\Rightarrow ridge regression works very well in high variance scenarios.

Also

Cost advantage over subset selection

b/c for fixed λ , only fit one model! (very fast model to fit).

Ridge regression improves predictive performance.

Does it also help us w/ interpretation? No

2.2 The Lasso

Ridge regression does have one obvious disadvantage.

Unlike best subset, forward/backward selection (generally select model w/ a subset of variables) ridge regression will include all p variables in the final model.

penalty $\lambda \sum_{j=1}^p \beta_j^2$ will shrink $\beta_j \rightarrow 0$, 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 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

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

$\|\beta\|_1$, L_1 norm.

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 sufficiently large!

\Rightarrow much like best subset selection lasso perform 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.

variable selection

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

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

$$\sum_{j=1}^p |\beta_j| \leq s$$

$$\sum_{j=1}^p \beta_j^2 \leq s$$

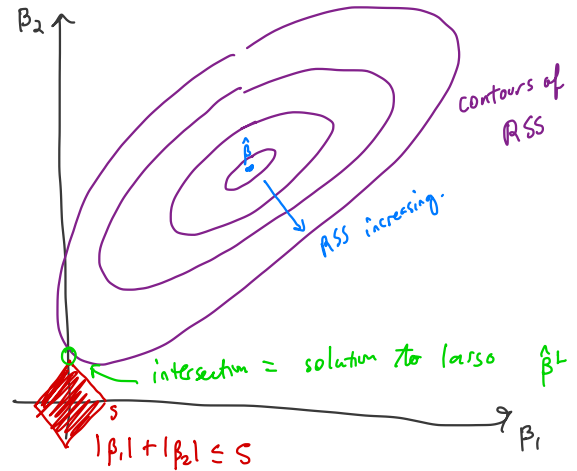
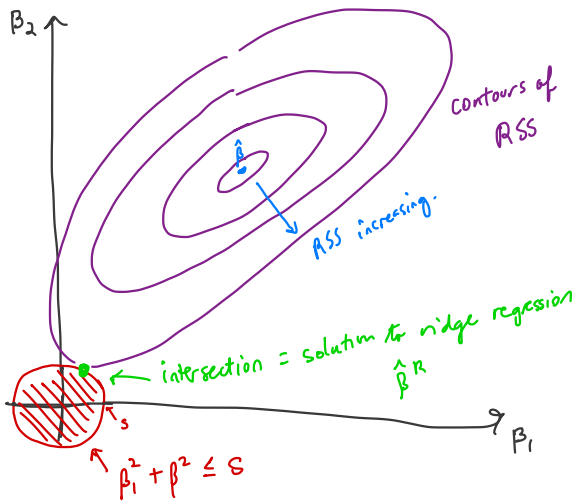
constraints

} 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 very large. Similar for Ridge as well.

But why does the lasso result in coef. estimates exactly = 0? let $p=2$.



Solution to both ridge and lasso is the first point the ellipses (RSS) contact the constraint regions.

Since ridge is a circle (no sharp points), intersection doesn't generally occur on the axis.

Lasso has corners on each axis \Rightarrow ellipse often will intersect at the axis \Rightarrow at least one of the coefficients equal to 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 than ridge (bias + variance)

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

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

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

How? CV!

penalization
parameter

- ① choose a grid of λ values.
- ② Compute CV error for each λ ← LOOCV or k -fold CV
- ③ select λ for which CV error is smallest
- ④ refit model using all available training data and selected λ .

NOTE: still very important to scale variables x_1, \dots, x_p for lasso to all 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, 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 then fit least squares on 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 ϕ_{jm} chosen well, this can outperform least squares.

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 $M < p$.

$$\begin{array}{ccc} & \uparrow & \\ & \beta_0, \beta_1, \beta_2, \dots, \beta_p & \\ & & \theta_0, \theta_1, \dots, \theta_M \end{array}$$

Note:

$$\sum_{m=1}^M \theta_m z_{im} = \sum_{m=1}^M \theta_m \sum_{j=1}^p \phi_{jm} x_{ij} = \sum_{j=1}^p \left[\sum_{m=1}^M \theta_m \phi_{jm} \right] x_{ij} = \sum_{j=1}^p \beta_j x_{ij}$$

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

$$\beta_j = \sum_{m=1}^M \theta_m \phi_{jm}$$

\Rightarrow special case of original linear regression model (with β_j constrained)

\hookrightarrow can bias coefficient estimates (trade-off lower variance, hopefully).

All dimension reduction methods work in two steps.

\hookrightarrow If $p > n$ ($p \approx n$) selecting $M \ll p$ can reduce variance.

① transformed predictors are obtained (get ϕ_{jm} 's).

② model is fit using M transformed predictors from ①.

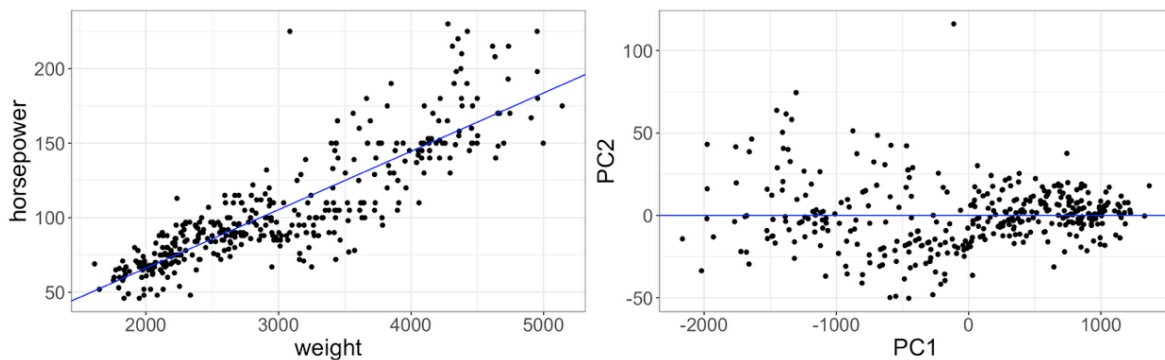
The selection of ϕ_{jm} 's can be done multiple ways, we will talk about 2.

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.

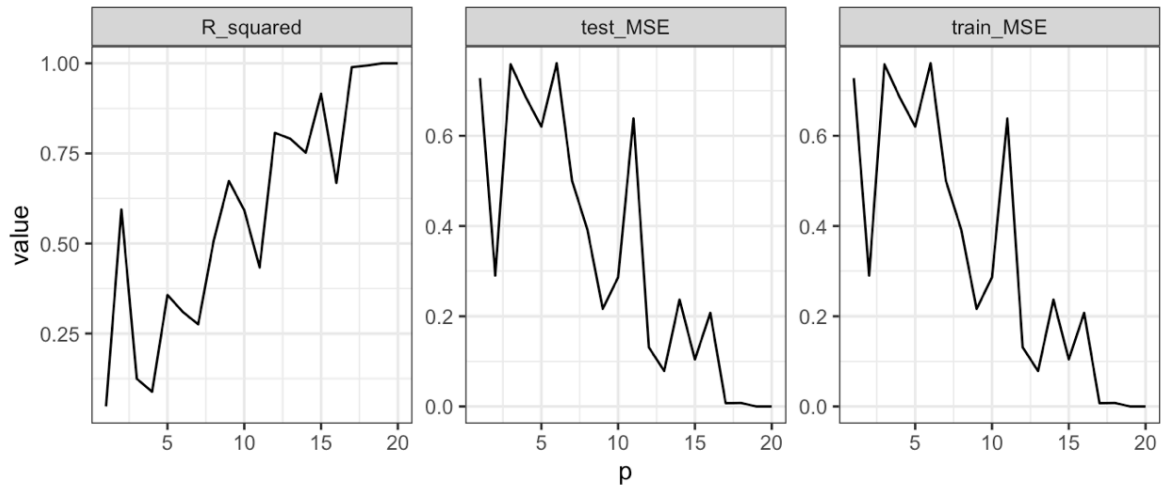
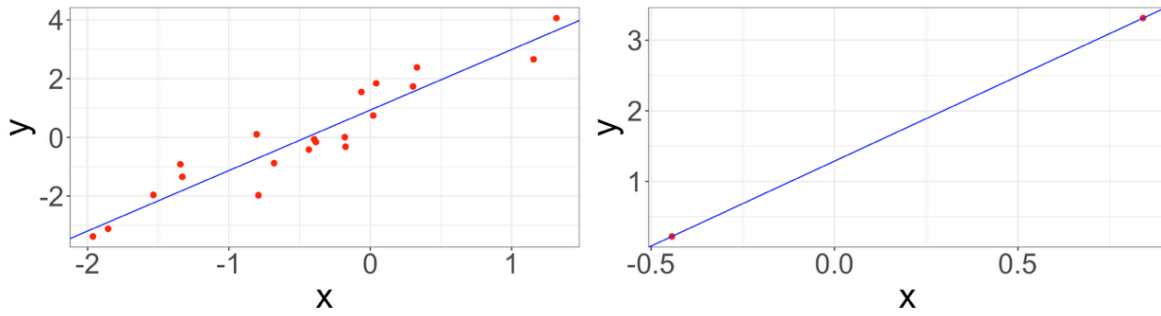
4 Considerations in High Dimensions

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.