Chapter 6: Linear Model Selection & Regularization

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

$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$
typically fit using least square

Upcoming: more general model (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 $n \gg p$ => also low variance => perform well on test data!
  [If n not much larger than p => high variability => poor performance on test data.
  If $n < p$ => least squares no longer has a unique solution => variance $= \infty$ => can't use this at all!]
  goal: reduce variance without adding too much bias.
- model interpretability: often many predictor variables in a regression model are not in fact associated w/ the response.

  By removing them (set $\hat{\beta}_i = 0$), we could obtain a more easily interpretable model.
  Note: least square will hardly ever $\hat{\beta}_i = 0$
  => 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. eg models w/ exactly 2 predictors $\Rightarrow \binom{p}{2} = \frac{p(p-1)}{2}$ models.

Algorithm:

1. Let $\mathcal{M}_0$ denote null model: no predictors
2. For $k = 1, \ldots, p$
   (a) Fit all $\binom{p}{k}$ models that contain k predictors
   (b) Pick best of those, call it $\mathcal{M}_k$. "Best" is defined by $\downarrow RSS$ ($\uparrow R^2$).
3. Select a single best model from $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ by using CV, C_p, AIC/BIC, or adjusted R^2 (more later.)

Why can't we use R^2 (RSS) to choose our model in step 3.?
adding predictors will always $\uparrow R^2$!

Why might we not want to do this procedure at all? Computation. Fitting 2^p models. $p=10 \approx 1000$ models.

We can perform something similar with logistic regression.

1.2 Stepwise Selection

For computational reasons, best subset selection cannot be performed for very large p. → impossible w/ $p \geq 40$.

Best subset may also suffer for p large etc w/ large search space
might happen upon a model that works well on training data that performs poorly on test data
$\Rightarrow$ high variability of coefs & overfitting can occur.

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

$\mathcal{M}_0$ ⇩ $\mathcal{M}_p$

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 $\mathcal{M}_0$ denote the null model — no predictors
2. For $k = 0, \ldots, p-1$
   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ w/ 1 additional predictor
   (b) Choose the best among these $p-k$ and call it $\mathcal{M}_{k+1}$ ($\uparrow R^2$, $\downarrow RSS$).
3. Select a single best model from $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using CV error, C_p, AIC/BIC, or adjusted R^2

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

$\mathcal{M}_p$ ⋮ ↓ $\mathcal{M}_0$

Backward Stepwise Selection: Begin w/ full model and take predictors away one at a time until we get to the null model. Choose the best one along that path.

1. Let $\mathcal{M}_p$ denote the full model – contains all p predictors
2. For $k = p, p-1, \ldots, 1$
   (a) consider all k models that contain all but 1 of the predictors in $\mathcal{M}_k$ ($k-1$ predictors).
   (b) Choose the best among them and call it $\mathcal{M}_{k-1}$ ($\uparrow R^2$, $\downarrow$ RSS).
3. Select the single best model from $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using CV error, AIC/BIC, or adjusted R^2.

greedy search. * Neither forward nor backwards stepwise selection are guaranteed to find the best model containing a subset of the p predictors. Seem to get decent results.

forward selection can be used when $p > n$ (but only up to $n-1$ predictors included – not p!).

1.3 Choosing the Optimal Model

Best subset, forward selection, backward selection all need a way to pick best model – according to test error

RSS & R^2 are proxies for training error ⇒ not good estimates of test error

↳ either (1) estimate this directly or (2) adjust training error for model size.

(2) $C_p = \frac{1}{n}\left(RSS + 2d\hat{\sigma}^2\right)$ ← estimate of variance of ε from full model.

↑ # predictors in submodel

add penalty to training error $\frac{RSS}{n}$ to adjust for under estimate of test error

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

proportional ⇒ same answer

(2) AIC & BIC maximum likelihood fit (linear model fit w/ least squares, this is the same).

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

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

choose model w/ low BIC. Since $\log(n) > 2$ for $n > 7$ ⇒ heavier penalty on models w/ many variables ⇒ result in smaller models.

(2) Adjusted R^2 (least squares models).

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

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

choose model w/ highest Adj. R^2.

(1) Validation and Cross-Validation

- Directly estimate test error w/ CV or validation method and choose the model w/ lowest estimated test error.
- Very general (can be used for any model) even when it's not clear how many "predictors" we're talking about.

Now have fast computers, CV is 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.

↳ 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 $\beta_1, \ldots, \beta_p$ using values that minimize

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

residual sum of squares.

$\hat{\beta}^R$

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

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \|\beta\|_2$$

ℓ_2 norm, $\|\beta\|_2$

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

$\lambda \geq 0$

penalization parameter

choose by tuning (separately from the fitting procedure).

trade-off 2 criteria: minimize RSS to fit data well

minimize $\lambda \sum_{j=1}^{p} \beta_j^2$ shrinkage penalty = small when β_j 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 $\lambda \to \infty$, impact of the penalty grows and $\hat{\beta}^R \to 0$.

Ridge regression will produce a different set of coefficients for each penalty λ $\left[\hat{\beta}^R_\lambda\right]$

Selecting a good λ is critical! How to choose?

The standard least squares coefficient estimates are scale invariant.

Multiply 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 jth predictor is scaled, $x_j \hat{\beta}_j$ will remain the same.

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

e.g. Say we have an income variable in ① dollars or ② thousands of dollars.

① = 1000 × ②

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

$\Rightarrow$ $x_j \hat{\beta}^R_{j,\lambda}$ depends not only on λ, but also on the scale of x_j

(may even depend on scaling of the 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 1.

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

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

★ ① standardize data

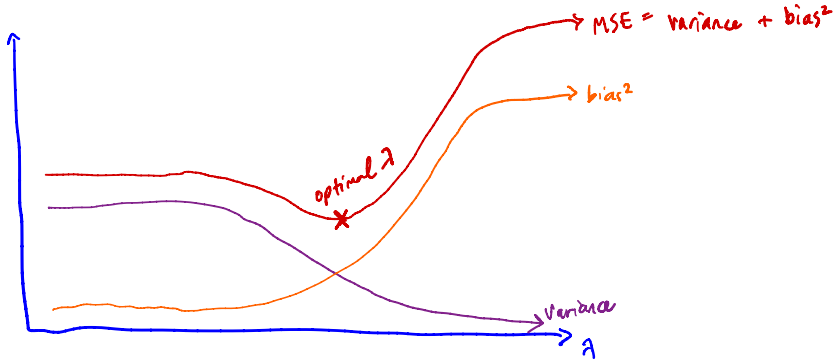
② tune model to choose λ

③ fit ridge regression w/ chosen λ.

Why does ridge regression work?

Because of the bias-variance trade-off!

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



In situations when the relationship between response and predictors is $\approx$ linear
least squares will have low bias.

When p almost as large as $n \Rightarrow$ least squares has high variability?
if $p > n$ least squares doesn't even have a solution

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

$\Rightarrow$ ridge regression works best in high variance scenarios.

Also
cost advantage over subset selection methods (sort of)
b/c for a fixed λ, only fit 1 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 or backward selection (generally a 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 $\beta_j \to 0$ 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. (Cons)

We will always have all variables in the model, whether there is a true 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

ℓ_1 norm, $\|\beta\|_1$

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

ℓ_1 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 subset selection methods, lasso also 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, choosing λ is critical.

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

lasso: minimize RSS subject to $\sum_{j=1}^{p} |\beta_j| \leq s$

Ridge: minimize RSS subject to $\sum_{j=1}^{p} \beta_j^2 \leq s$

constraints

$\Leftrightarrow$ equivalent to previous formulation

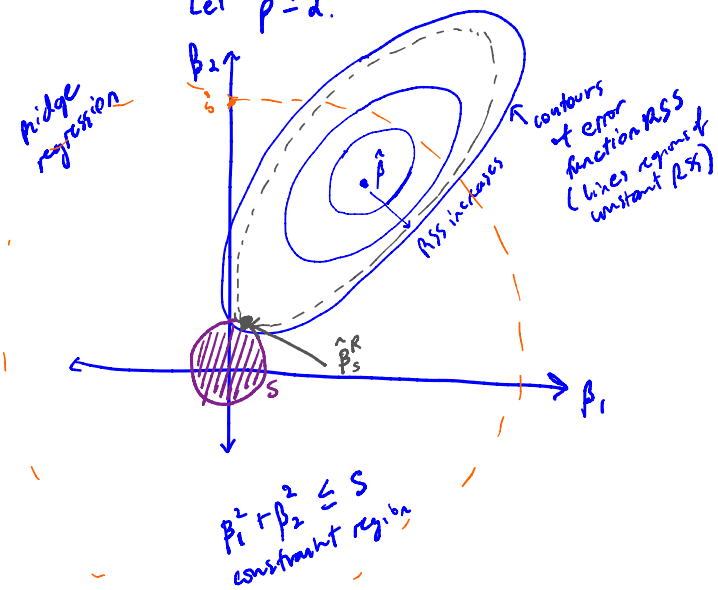
constrained optimization problems.

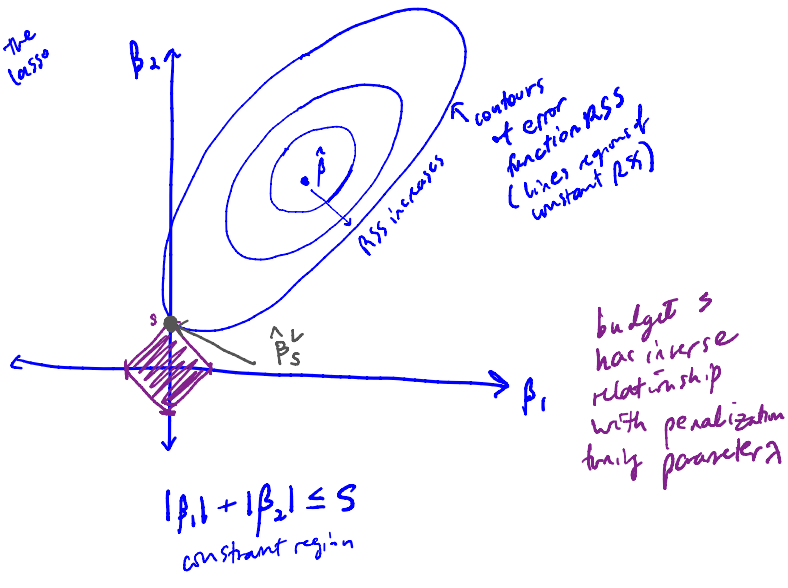
In other words, when we perform the lasso we are trying to find the set of coefficient estimates that lead to the smallest RSS, subject to the contraint that there is a budget s for how large $\sum_{j=1}^{p} |\beta_j|$ can be.

When s is large, this is not much of a constraint => coeff. estimates can be large

But why does lasso result in coefficient estimates exactly equal to zero (ridge does not)?

Let $p=2$.





Solution to either ridge or lasso is the first point in ellipses that contacts the constraint regions.

Since ridge regression results in a circular constraint, there are no sharp points, intersection doesn't usually occur on either axis.

Lasso has corners on axes! => the contours often intersect at the axis => one of the coefficients will equal to zero!

If you believe there are predictors that do not have a relationship w/ your response (Y) (you don't have to know which ones), lasso will perform better than ridge regression

↓ bias (enough) & ↓ 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

penalization parameter

How?

1. choose a grid of λ values.
2. Compute CV error for each λ (or validation method.) (LOOCV, k-fold)
3. Choose λ for which CV error is minimized.
4. Fit the model (lasso or ridge) on entire training data set using selected λ
5. ? make predictions, perform inference, etc.

NOTE: Still important to scale variables $x_1, \ldots, x_p$ for Lasso to all have st. dev. $= 1$.