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 subset model

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

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.

Shrinking the coefficient estimates can significantly reduce their variance!

2.1 Ridge Regression

Recall that the least squares fitting procedure estimates $\beta_1, \ldots, \beta_p$ using values that minimize

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

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

The standard least squares coefficient estimates are scale invariant.

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

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

Why does ridge regression work?

2.2 The Lasso

Ridge regression does have one obvious disadvantage.

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

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

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

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

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

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

2.3 Tuning

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

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

How?