Chapter 6: Linear Model Selection & Regularization

In the regression setting (response numeric), 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 model using least squares.

↓

we will talk about other ways we could approach this fitting problem.

(later will go 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 true relationship is ≈ linear, least squares will have low bias.
  - If $n >> p$ ⟹ also low variance ⟹ perform well on test data!
  - But n not much larger than p ⟹ high variability ⟹ poor performance on test data.
  - If $n < p$ ⟹ no longer have unique solution ⟹ variance $= \infty$ ⟹ cannot use this at all!

  goal: reduce variance without adding too much bias.

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

  By removing them (setting $\hat\beta_i = 0$), we can obtain a more interpretable model.

  Note: Least squares will hardly ever result in $\hat\beta_i = 0$.

  goal: 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. $\binom{p}{k}$ models for each # of predictors in model (k).

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 the best of those call it $\mathcal{M}_k$. "Best" defined by $\downarrow$RSS, $\uparrow R^2$
3. Select a single best model from $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ using CV error, C_p, AIC/BIC, or adjusted R^2 more later.

We can't use R^2 for step 3. as $k\uparrow$ $R^2\uparrow$ always.

Why might we not want to do this procedure at all?

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. impossible for $p \geq 40$

Best subset also suffers when p is large because w/ large search space We can find good models on training data that perform poorly on test data.
high variability & overfitting of coeffs can occur.

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 $\mathcal{M}_0$ denote the null model – no predictors.
2. For $k = 0, \ldots, p-1$
    (a) consider all $p-k$ models that augment predictors in $\mathcal{M}_k$ w/ 1 additional predictor.
    (b). Choose the best among $p-k$ 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 are fitting $1 + \sum_{k=0}^{p-1} (p-k) = 1 + \frac{p(p+1)}{2}$ 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 – all p predictors

2. $K = p, p-1, \ldots, 1$:

   (a) Consider all k models that contain all but one of the 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 single best model using CV error, etc.

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

When $p>n$: forward selection can be used (but only up to $n-1$ predictors, not p).

1.3 Choosing the Optimal Model

Best Subset, forward, backward select all need to pick "best" model – according to test error.

RSS & R^2 are proxies for training error => not good estimates of test error. – ① estimate this directly ② adjust training for model size.

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

$\hat{\sigma}^2$ ← estimate of variance of ε full model

d ↑ # predictors in subset model

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

as $d\uparrow$, $C_p\uparrow$

② AIC & BIC

② Adjusted 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 $\beta_1, \dots, \beta_p$ 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?