6.2.2: The Lasso

- A disadvantage of ridge regression is that it will include all p predictors in the final model.

- The lasso overcomes this. It minimizes

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

This forces some coefficients to be exactly 0 when λ is large enough. So: the lasso performs <u>variable selection</u>, yielding sparse models

The Variable Selection Property of the Lasso

Has to do with Lagrange multipliers

Use Lasso if you expect that some coefficients are 0.

6.2.3: Selecting the Tuning Parameter

Need to select λ.

Cross-validation approach:

1) Choose a grid of λ values
2) Compute CV error for each λ
3) Select λ where CV error is smallest, call it λ^*
4) Refit the model on all the data using λ^*

2/19/23

Day 14

6.3: Dimension Reduction Methods

This is a class of approaches that <u>transforms</u> the predictors and then fits least squares using the transformed variables.

Let $Z_1, \ldots, Z_M$ represent $M<p$ linear combinations of our original p predictors,

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \,, \quad \text{for constants } \phi_{1m}, \ldots, \phi_{pm},\ m \in [M].$$

Then we can fit the linear regression model (w/ least squares)

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i \,, \quad i = 1, \ldots, n$$

If $\phi_{1m}, \ldots, \phi_{pm}$ are chosen wisely, then dimension reduction approaches can outperform least squares regression.

This works well when ~~$M<<p$~~ $p>n$, especially if you choose $M<<p$.

All dimension reduction methods follow two steps:
1) Get $Z_1, \ldots, Z_m$
2) Estimate $\theta_1, \ldots, \theta_m$

We review two approaches: Principal Components and Partial Least Squares

6.3.1: Principal Components Regression (PCA)
An Overview of PCA
- PCA is a technique for reducing the dimension of an $n \times p$ data matrix X.
- The first principal component direction of the data is that along which the observations vary the most.
  - i.e. Projecting the 100 observations onto this line gives the largest variance possible.

So, (1) center the data, (2) fit ~~orthogon~~ orthonormal components to the data that maximize variance.

This is about simplifying the predictors.

Principal Component Regression

- Find the first M components, then use these components in a linear regression.
- We assume that the component directions are associated with Y.
- Can do better than least squares linear regression by using most of the information while avoiding overfitting.

PCR does better if fewer components are required.

PCR is not a feature extraction method

"One can think of ridge regression as continuous PCR." Cool

In PCR, M is often chosen by cross validation.

~~In PCR,~~ Standardize the predictors prior to doing PCR, otherwise a variable having high variance can have an outsized impact on the result.

6.3.2: Partial Least Squares (PLS)

- In PCR, the principal components are identified in an unsupervised way - as Y is not used to find the principal components.
- Hence, there's no guarantee that the principal components are associated with the response.

PLS is a supervised alternative to PCR.

How do you compute the PLS direction?

1) The first direction Z_1 is computed by setting each ϕ_{j1} equal to β_j from a simple linear regression of Y onto X_j, so

$$Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$$

is ~~mo~~ strongly correlated with the response.

Okay, I don't understand this