Today
1. Ridge Regression & The Lasso (continued)
2. Dimension Reduction Methods
   - PCR
   - PLS

1. Ridge Regression & The Lasso (continued)

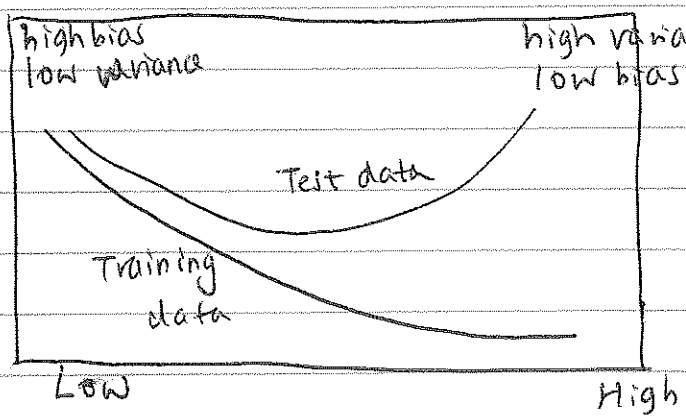
* Ridge Regression

$$\hat{\beta}_j^R \text{ minimize } RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

Note: It is better to apply ridge regression after standardizing the predictors $\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}$

Why can shrinking towards zero be a good thing to do?
- turns out that the LS estimates generally have low bias but can be high variance. In particular when n & p are of similar size or when $n < p$, then the LS estimate will be extremely variable.
- The penalty term makes the ridge regression estimates biased but can also substantially reduce variance
- There is a bias / variance trade-off

Cons & pros
- Ridge regression estimates will be more biased than the LS estimates, but have lower variance.
- Ridge regression will work best in situations where LS estimates have high variance
- Computation benefit: fit one model for any given λ
- Can be used when $p > n$



* Lasso

- Ridge Regression is not perfect. as the penalty term never forces any of the coefficients to be exactly zero. Thus, the final model will include all variables, which make it harder to interpret.
- A more modern approach is the Lasso.
- Similar way, use a different penalty term

Lasso estimates by minimizing

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

- Mathematically, prove that some coefficients end up being set to exactly zero.
- produce a model that has high predictive power, but simply to interpret.
- Lasso performs variable selection

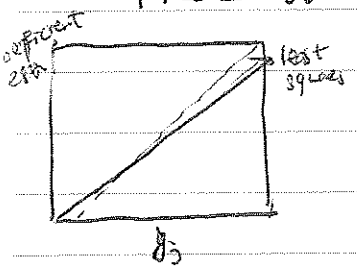
Ex 1: $n=p$ X a diagonal matrix with 1's on the diagonal and 0's in all off-diagonal elements

perform regression without an intercept.

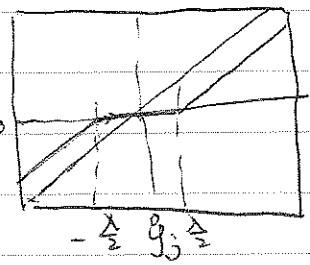
LS: minimize $\sum_{j=1}^{p}(y_j-\beta_j)^2 \Rightarrow \hat{\beta}_j=y_j$

Ridge Regression: minimize $\sum_{j=1}^{p}(y_j-\beta_j)^2+\lambda\sum_{j=1}^{p}(\beta_j)^2 \Rightarrow \hat{\beta}_j^R=\frac{y_j}{\lambda+1}$

The Lasso = minimize $\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$



$$\Rightarrow \hat{\beta}_j^L = \begin{cases} y_j - \frac{\lambda}{2} & \text{if } y_j > \frac{\lambda}{2} \\ y_j + \frac{\lambda}{2} & \text{if } y_j < -\frac{\lambda}{2} \\ 0 & \text{if } |y_j| \le \frac{\lambda}{2} \end{cases}$$



Note: ① Ridge Regression, each least squares coefficient estimate is shrunken by the same proportion. ② In contrast, the Lasso shrinks each least squares coefficient towards zero by a constant amount, $\frac{\lambda}{2}$; the least squares coefficients that are less than $\frac{\lambda}{2}$ in absolute value are shrunken entirely to zero.

* Selecting the tuning parameter λ

- cross-validation
- Choose a grid of λ values, and compute the cross-validation error for each value of λ, as described in chapter 5.
- Then choose the tuning parameter value for which the cross-validation error is smallest.
- Finally, the model is refit using all of the ~~available~~ available observations and the selected value of the tuning parameter.

2. Dimension Reduction Methods

- a class of approaches that transform the predictors and then fit a least squares model using the transformed variables. We will refer to these techniques as dimension reduction methods

* Basic idea

Let $Z_1, \cdots Z_M$ represent linear combinations of our original p predictors. That is, $Z_m = \sum_{j=1}^{p} \phi_{mj} X_j$ ①

for some constants $\phi_{m1}, \cdots \phi_{mp}$

Then fit a linear model with the ordinary least squares

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i \quad i = 1, \cdots n \quad ②$$

Note: ① $\sum_{m=1}^{M} \theta_m z_m = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{mj} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{mj} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}$

where $\beta_j = \sum_{m=1}^{M} \theta_m \phi_{mj}$ ③

⇒ ② can be thought of as a special case of the original linear regression model.

②' Dimension reduction serves to constrain the estimated β_j coefficients, since now they must take the form ③

③' can win in the bias-variance trade off

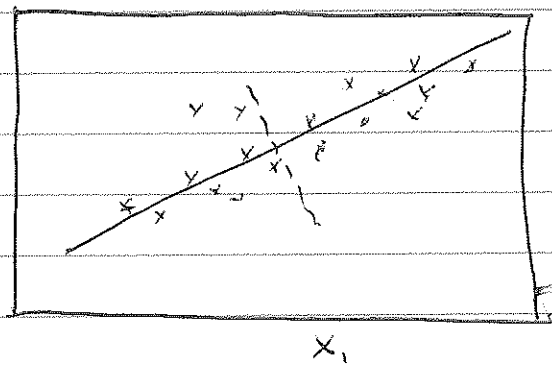
M=p. z_m's are linearly independent. ⇔ fit a least squares

I: PCR (principal Components Regression)

- the first pc is that linear combination of the variables with the largest variance
- the second principal component has largest variance, subject to being uncorrelated with the first one

and so on...

⇒ Hence with many correlated original variables, we replace them with a small set of pcs that capture their joint variance.



Example: PCR is applied to Credit dataset

Notes:

① PCR identifies linear combinations, or directions, that best represent the predictors $X_1, \cdots X_p$

These directions are identified in an unsupervised way, since the response variable Y is not used to help determine the PC directions.

i.e. the response does not supervise the identification of the principal components.

- Consequently, PCR suffers from a potentially serious drawback: there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.

Note: standardizing variables first

II PLS

- like PCR, PLS is a dimension reduction method, which first identifies a new set of features $Z_1, \dots Z_M$ that are linear combinations of the original features, and then fits a linear model via OLS using these M new features.
- Unlike PCR, PLS identifies these new features in a supervised way — i.e. it makes use of the response Y in order to identify new features that not only approximate the old features well, but also that are related to the response variable.
- In other words, the PLS approach attempts to find directions that help explain both the response and the predictors.

* Process:

- standardizing the p predictors
- PLS computes the first direction Z_1 by setting $\phi_{1j} = \beta_j$ ($Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$)
- One can show that this coefficient is proportional to the correlation between Y and X_j
- Hence, in computing $Z_1 = \sum_{j=1}^{p} \phi_{1j} X_j$, PLS places the highest weight on the variables that are most strongly related to the response.
- Subsequent directions are found by taking residuals and then repeating the above prescription