6.1.3: Choosing the Optimal Model

We need to estimate the test error.

1) Adjust the training error
2) Directly estimate test error.

Cp, AIC, BIC, Adjusted R^2

- These can be used to select among different numbers of variables.

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

where $\hat{\sigma}^2$ is an estimate of $Var(\epsilon)$ and d is the number of predictors
(Note ϵ is often estimated with the full model)

- AIC is defined for models fit by maximum likelihood

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS - 2d\hat{\sigma}^2)$$

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

2/16/23 Adjusted $R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$

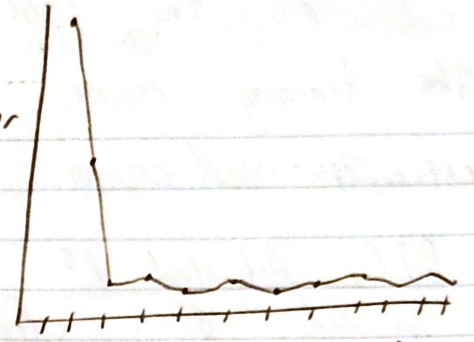
Adjusted R^2 pays a price for adding noise variables because they lead to only a small increase in RSS but make $(n-d-1)$ larger, shrinking adjusted R^2

Validation and Cross-Validation

- Direct estimate of test error w/out many assumptions of the model form.
- Computationally intense but not too bad nowadays

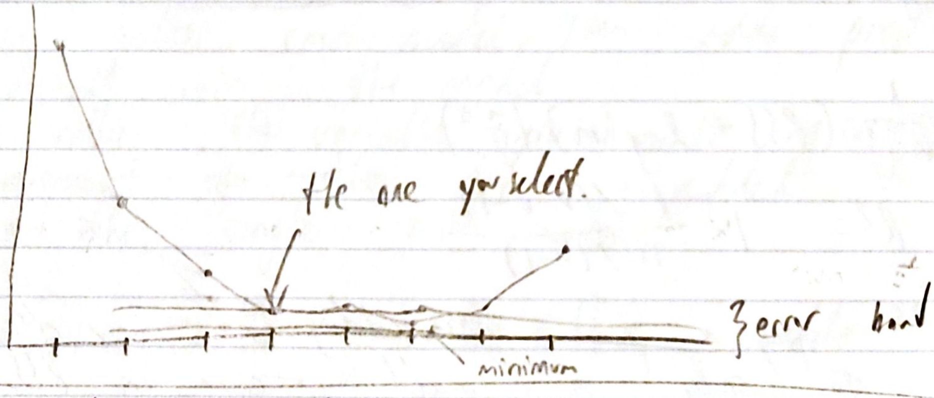
Here's a rule of thumb called the <u>one-standard error rule</u>

You use it in this setting, → where the CV error vs. # predictors curve is flat in the tail. This is because the <u>specific</u> ~~model~~ number of predictors depends on exactly how you split the training data into folds.



<u>One-standard error rule</u>

1) Calculate the standard error of the estimated test MSE for each model size.
2) Select the smallest model for which the estimated test error is within one standard error of the lowest point on the curve.



<u>6.2: Shrinkage Methods</u>

An alternative to subset methods

Instead: fit a model using all p predictors, but with a technique that <u>constrains</u> or <u>regularizes</u> the coefficient estimates

This works because shrinking the coefficient estimates can significantly reduce their variance.
The best techniques are ridge regression and the lasso.

6.2.1: Ridge Regression

Least squares minimizes

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

Ridge regression minimizes

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

where $\lambda \geq 0$ is a tuning parameter, which must be chosen well

$\lambda \sum_{j=1}^{p} \beta_j^2$ is a shrinkage penalty that makes the coefficients smaller.

- Note that we don't shrink β_0 as this is the mean of the response

Apply ridge regression after standardizing the predictors, meaning to use

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

estimated standard deviation of the j-th predictor.

Why does ridge regression improve over least squares?
As $\lambda \nearrow \infty$, flexibility decreases, bias increases

Ridge regression works best in situations where the least squares estimates have high variance.

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 variable selection, 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^*