Lecture 14.

Last class:

Model Selection (Two main ingredients):

1. Model Selection Criteria
2. Model selection strategy

Selection criteria including:

1. Adjusted R^2: $R^2_{adj} = 1 - \frac{SS(res)/n-k-1}{SS(tot)/n-1}$

   k: number of predictors in a model $(k \le p)$

2. AIC: $AIC = 2q - 2\ln L(\hat{\theta})$

   q: number of parameters in a model

   $L(\hat{\theta})$: likelihood function evaluated at $\hat{\theta}$

3. BIC: $BIC = q\ln(n) - 2\ln L(\hat{\theta})$

- R^2_{adj}, AIC and BIC explicitly penalizes too many parameters in unnecessarily complex models)
- R^2_{adj}, AIC and BIC all try to prevent overfitting, which can lead to poor predictions

  modeling relationships b/t some covariates and response that are due to chance.
- All three methods can be used to compare fitted models.

4. MSPE

Model Selection Strategy:

- Used with some selection criteria
- Suppose we have p predictors, and we want to find a subset $(k \le p)$ that gives the "best" model.

(1) All possible subset regression.

With p predictors, how many models to fit in total?

- $\binom{p}{0} = 1$ (intercept only model)
- $\binom{p}{1} = p$ (models with one covariate)

$\vdots$

- $\binom{p}{p} = 1$ (model w/ all covariates)

$$\sum_{j=0}^{p} \binom{p}{j} = \sum_{j=0}^{p} \binom{p}{j} 1^{j} 1^{p-j} = (1+1)^p = 2^p$$

$\therefore$ There is a total of 2^p models we must fit.

- In theory, we can fit all 2^p models, and choose the "best" one according to our criteria. Thus, we can find the optimal model (based on criteria).
- Not feasible when p is very large.
  e.g. $p=10$, $2^p = 1024$, $p=35$, $2^p > 30$ billion

Idea: To find a good/useful model with reasonable computational time (not necessarily optimal). Following strategies focus on adding/removing variables one at a time.

(2) Forward Selection (FS)

Idea: Start w/ no covariates and add one variable at a time.

1. Start w/ model with just an intercept.
2. Fit p models with 1 covariate, i.e.
   $$Y_i = \beta_0 + \beta_1 x_{ij} + \varepsilon_i, \quad i=1,\dots,n \text{ and } j=1,\dots,p.$$
3. Pick the best of the p models according to the selection criteria, and add that covariate (say x_a) to our model.
4. Fit p-1 models w/ x_a and another covariate, i.e
   $$Y_i = \beta_0 + \beta_1 x_{ia} + \beta_2 x_{ij} + \varepsilon_i, \quad i=1,\dots,n \text{ and } j=1,\dots,p \underbrace{\setminus a}_{\text{excluding}}.$$
   (i). If none of p-1 models improve criteria, STOP; o/w
   (ii). Pick best of p-1 models according to criteria, so we end up with 2 covariates in our model.

5. Repeat the process by adding variables one at a time, until no more variables improve the selection criteria.

- Final model is one w/ the best criteria when we stop.
- Compared w/ strategy (1), this is less computationally intensive. The max # of models is $p + (p-1) + (p-2) + \cdots + 2 + 1 = p(p+1)/2$ models. (compare w/ 2^p)
- However, the final model might not be optimal (out of 2^p), but might be good enough.

→ e.g. $p=3$, if the best one-variable model is one that includes x_1, but the optimal model is one that includes x_2 and x_3, FS will never find this optimal solution.

(3) Backwards Elimination (BE)

Idea: Start w/ p predictors and remove 1 var at a time.

1. Start w/ full model (w/ all p predictors)
2. fit p models resulting from removing one predictor from the regression. (each model has p-1 covariates)
3. Choose the best of the p models based on criteria and remove the variable (say x_b) from our model (if no improvement then STOP).
4. Fit p-1 models without x_b and another variable (2 variables removed)
   (i) If none of p-1 models improve the criteria, STOP, o/w
   (ii) Pick best of p-1 models according to criteria, and we have p-2 variables in our model.
5. Repeat the process by removing 1 var. at a time, until no more improvements.

- Same computational complexity as FS.
- Why would we prefer FS over BE? If $p >> n$, then (X^TX) is not invertible
- Once a variable is removed, it can't re-enter the model.

(4) Forward and Backward stepwise regression

- Compromise between FS and BE.

1. Start w/ forward selection and find the best one-var model (say x_a).
2. FS : Select the best two covariate model using FS (w/ x_a included, say we add x_b). If two-var model doesn't improve selection criteria, STOP., o/w.
3. BE : With x_b added to our model, determine if any other x variables in our model should be dropped (at this stage, we only have x_a, but at later stages, we will have much more).
4. Repeat the process (steps 2 and 3) until no more improvements on model can be made.

- This method allows us to add/remove a variable more than once.

Note :

- Basic methods described here can be used to get a "good" (useful) model.
- However, they are primitive, some more sophisticated methods:
  LAsso, ridge, elastic net, etc.
- Variable/model selection is a hard problem.