Lecture 21
Monday, July 22, 2019 2:42 PM

[1] All Possible regressions

- Here we fit all possible regression models and choose the one that is 'best'.
- This involves fitting the intercept only model, the $q = \binom{q}{1}$ models with one explanatory variable, the $\binom{q}{2}$ models with two explanatory variables ...., the $\binom{q}{q} = 1$ model with all q explanatory variables.

  ↳ $\sum_{i=0}^{q} \binom{q}{i} = 2^q$ is the total number of models we need to fit

- We identify the "best" model by comparing all of them on the basis of some "decision criteria". This criterion is a metric that is chosen to match our modeling objective:
  - Explanatory ⟶ R^2_{adj}, AIC / AICC, BIC
  - Predictive ⟶ Predictive MSE (this is different from the MSE we're so far familiar with).

- $R^2_{adj} = 1 - \left(\frac{n-1}{n-p-1}\right)(1-R^2)$

  We use R^2_{adj} to compare models because it accounts for the number of explanatory variables in the model, and it doesn't necessarily increase when additional variables are included.

- Akaike Information Criteria (AIC)

$$AIC = 2k - 2 \ln L(\hat{\theta})$$

  where θ are the parameters and k is equal to the number of parameters.
  * Smaller values are better
  * the $2k$ term is a penalty term that protects us from overfitting.

- Corrected AIC (AICC or AICc)

$$AICC = \left(\frac{2n}{n-k-1}\right)k - 2 \ln L(\hat{\theta})$$

  * this "corrected" version adjust penalty based on the sample size (harsher penalty when n is small).

- Bayesian Information Criteria (BIC)

$$BIC = \ln(n)k - 2 \ln L(\hat{\theta})$$

  * an alternative version of a samplesize-dependent penalty.

The preceding 4 quantities are referred to as "goodness-of-fit" metrics. They allow us to compare the adequacy of multiple models with potentially different numbers of explanatory variables.

Each metric contains a penalty term which exists to protect us from over fitting which is a problem that arises when by adding many explanatory variables into the model, we can make the fitted values arbitrarily close to the observed values. This is sometimes referred to as "modeling the noise".

* The All-possible-regressions approach is nice because we're guaranteed to find the optimal model. However, 2^q can be a computational intensive / impossible number of models to fit. As such, interest lies is using model selection techniques that are less computationally expensive, and that still find us a "good" model

May not "optimal"

Our solution to this problem are the stepwise selection techniques: