Model Validation

Stat 230: Applied Regression Analysis

First, other model comparison metrics

Metrics that avoid testing

- Adjusted R^2
- Akaike's Information Criteria (AIC)
- Bayesian Information Criteria (BIC)
- Mallow's C_p

These do not rely on p-values, and they balance both complexity and fit

Mallow's C_p

Let p be # of coefficients for the model in question

$$C_p = \underbrace{\frac{SSE_p}{MSE_{full}}}_{\text{error part}} - \underbrace{(n-2p)}_{\text{penalty for complexity}}$$

- Want small C_p, but also want $C_p \approx p$
- $C_m = m$ for the biggest model (m = # of coefficients)
- C_p focuses on prediction
- Or, C_p minimizes SSE + penalty

AIC

$$AIC = \underbrace{n \log(SSE/n)}_{\text{error part}} + \underset{\text{penalty for complexity}}{2p}$$

- choose model with smaller AIC
- models do not need to be nested to be compared
- based on asymptotic theory
- generally favors models that are bigger than the “true” model

BIC

$$BIC = \underbrace{n \log(SSE/n)}_{\text{error part}} + \underbrace{p \log(n)}_{\text{penalty for complexity}}$$

- choose model with smaller BIC
- models do not need to be nested to be compared
- based on asymptotic theory
- usually larger penalty than AIC $\Longrightarrow$ leads to smaller models

Model validation

Example: Candy rankings

Selected model

Starting with a model that contained all predictor variables, dropped variables that were unimportant:

term	estimate	std.error	statistic	p.value
(Intercept)	51.804	3.749	13.817	<0.001
caramelTRUE	−14.510	7.609	−1.907	0.060
peanutyalmondyTRUE	−18.192	7.505	−2.424	0.018
hardTRUE	16.201	7.313	2.215	0.030
barTRUE	20.275	6.898	2.939	0.004

But wait!

- I replaced the real win percentage with randomly generated values between 0 and 100!
- We still got “significant” results…

Beware of inference after model selection!

If we use the same data set to conduct inference as we used to select our model we run into (related) issues

- overfitting
- selection bias
- under-estimation of MSE

Validation

If you want to run inference on variables used in model selection…

1. Collect new data for inference -> might be possible in controlled experiments
2. Compare results with past results
3. Use a holdout sample -> most practical

Data splitting

- Split your original data set into two pieces: a **training set** and a **validation set**
- Training set is only used for model selection, should have at least 6 to 10 times as many rows as there are slope coefficients in the biggest model considered
- What split?
    - Random 50/50 split is a starting point
    - Increase size of training set until you get 6-10 x rows as slopes
- Sometimes you don’t have enough data to do this!

Interpreting results

- Fit the model selected to the validation set and check
  - coefficients - are they similar?
  - significance tests - similar results?
  - MSE - similar prediction error?
- If the results are similar between the sets, no big issues with bias from model selection
  - Customary to refit the model to the entire (training + validation) data set
- If they are quite different, trust the results from the validation set