STAT 410 - Linear Regression

Lecture 14

Meng Li

Department of Statistics

Nov 7, 2017

Criteria for model selection

For a linear regression model with p regressors including the intercept, we will have $2^p - 1$ regression models to consider.

- R^2: Coefficient of determination
  - $R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$
  - R^2 increases as the p increases, regardless of that the added predictors are linearly associated with y or not.
  - For example, if $p = n$, we will have $R^2 = 1$.
  - Therefore, R^2 is not a good measurement to compare models with different number of parameters.
- R^2_{adj}: Adjusted coefficient of determination
  - $$R^2_{adj} = 1 - \frac{SS_{Res}/(n-p-1)}{SS_T/(n-1)} = 1 - \frac{MS_{Res}}{MS_T}$$
  - The inclusion of new parameters is penalized by $n-p-1$ while R^2 increases (or equivalently, SS_{Res} decreases).
  - Therefore, R^2_{adj} is (more) suitable to compare models with different number of parameters as it introduces a penalty on model complexity.

- AIC: Akaike's Information Criterion
  - The Akaike's Information Criterion (AIC) is defined as

$$\text{AIC} = -2\log L(\hat{\theta}) + 2K,$$

    where $\hat{\theta}$ is the MLE of the parameter θ (which is possibly a vector), $L(\hat{\theta})$ is the likelihood evaluated at the MLE, and K is the number of parameters in the model.
  - Only differences in AIC are meaningful, thus any common additive constant can be ignored.
  - For MLR, we have

$$\text{AIC} = -2\log L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) + 2(p+1),$$

    where p is the number of β's in MLR including the intercept.
  - This further reduces to

$$\text{AIC} = n\log\frac{SS_{Res}}{n} + 2p + \text{ constant.}$$

- AIC_c: Corrected AIC:

$$\text{AIC}_c = \text{AIC} + \frac{2K(K+1)}{n-K+1}.$$

  - AIC is information-based criterion (specifically, it is an estimate of the Kullback-Leibler information disregarding additive constants), and AIC_c tends to have smaller bias.
  - AIC_c induces a heavier penalty on the number of parameters.

- BIC:Bayesian information criterion
  - BIC is defined by

$$\text{BIC} = -2\log L(\hat{\theta}) + K\log n.$$

  - It penalizes the complexity of the model where complexity refers to the number of parameters in the model.
- General comments:
  - Model selection is an “unsolved" problem. There are no magic procedures that gives us the “best" model.
  - There are other model selection criteria such as prediction-based approaches and Bayesian model selections.
  - A popular data analysis strategy is to calculate R^2_{adj}, AIC, AIC_c and BIC and compare the models which minimize AIC, AIC_c and BIC with the model that maximizes R^2_{adj}.

Example: bridge construction

We calculate the discussed criteria for all subsets of regressors in the bridge construction example:

Subset size	Predictors	R^2_{adj}	AIC	AIC_C	BIC
1	log(Dwgs)	0.702	–94.90	–94.31	–91.28
2	log(Dwgs), log(Spans)	0.753	–102.37	**–101.37**	**–96.95**
3	log(Dwgs), log(Spans), log(CCost)	**0.758**	**–102.41**	–100.87	–95.19
4	log(Dwgs), log(Spans), log(CCost), log(DArea)	0.753	–100.64	–98.43	–91.61
5	log(Dwgs), log(Spans), log(CCost), log(DArea), log(Length)	0.748	–98.71	–95.68	–87.87

Therefore, we may choose the size 2 model including log(Dwgs) and log(Spans) as the regressors.

- The approach of all subsets cannot scale up if p is large.
- Next we shall introduce model selection techniques other than considering all subsets of regressors.