

Lab#6 Discussion

Biostatistics 210

0. BACKGROUND

This lab is meant to give you practice in predictor selection in the context of trying to build a model which is likely to give good prediction in the future. .

2. PREDICTOR SELECTION

The dataset has six predictors and it is easy to explore a large number of possible models. Note that the total number of possible models is $2^6 = 64$ which is not feasible to explore. Instead, do a series of backward selection models based on which variable has the largest p-value (equivalents the z-statistic which the least extreme).

The six model sequence is

1. *All predictors*
2. *Remove age_don*
3. *Remove prev_tx*
4. *Remove hlamat*
5. *Remove age*
6. *Remove txtype*

When we delete age_don, prev-tx and hlamat we have fairly straightforward choices since all these variables are conventionally statistically significant. However, we find that the BIC suggests we delete txtype and move to a two variable model (age and year). However, both AIC and the cross-validation suggest we stick with a three variable model. I would be torn. First, the AIC is known to be lenient and so is the cross-validation. It is never quite as tough as external validation.

# of Variables	AIC	BIC	Area under ROC (CV)
6	2190.0	2270.9	0.6695
5	2188.1	2262.2	0.6706
4	2186.3	2253.7	0.6721
3	2180.4	2207.4	0.6763
2	2184.1	2204.3	0.6738
1	2191.5	2205.0	0.6678

Discuss which model appears optimal based on the minimizing the BIC and (CV) area under the ROC curve. Which model would you choose as like? If you have time, examine the effect of categorizing the continuous variables on the BIC and CV area under the ROC curve.

I would be torn. First, the AIC is known to be lenient and so is the cross-validation. It is never quite as tough as external validation. The BIC exacts a strong penalty for complexity. I would tend to feel that txtype might add an individual level to the prediction which might help in it's adoption and would be inclined to include it but that is my personal preference.