Notes on the Leverage Regression Problem

**Example of the problem**

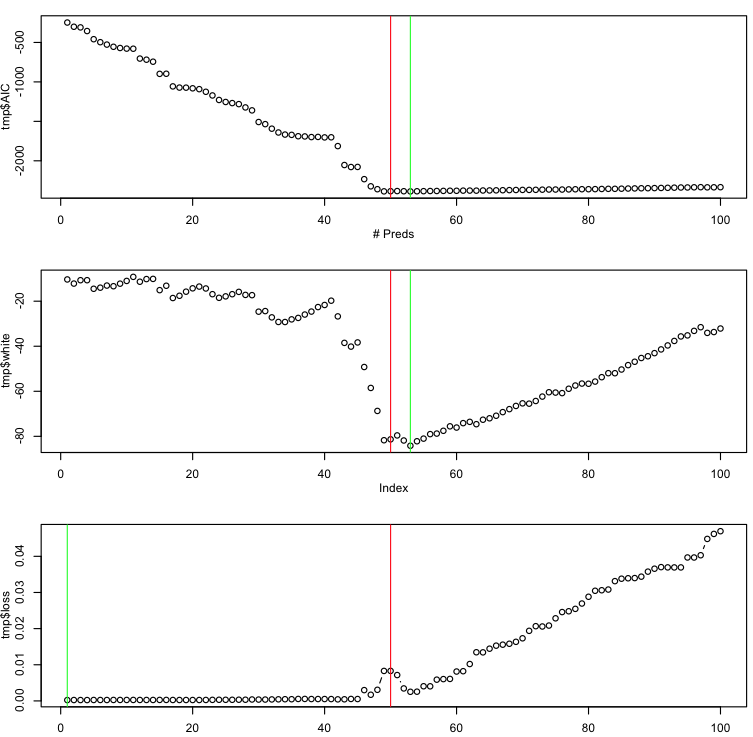
Consider a normal regression, with the usual setup,

*yi* = *β*’*xi* + *i* with *i* ~ N(0,) (1)

and with *n* cases and a vector *x* of *p* covariates. We will add predictors one-at-a time, keeping track of two model selection statistics and the estimator’s loss. The two statistics are the AIC and a version of AIC computed using a White (heteroscedastic consistent) variance estimate. The loss is the squared deviation between the LS estimator *b* and *β*, ||*b* - *β*||2

For the following figures, the data are generated to have *n* = 1,000 cases of model (1) with *p* = 100. The figures differ by having different signal levels obtained by varying the size of *β*. The wrinkle of the data generation is that the first  
 *nlarge* = 50 cases have standard deviation *s* = 1,000  
 (Note: this is the SD within a row of the design, not a column)  
and the remaining  
 *n* - *nlarge* have standard deviation 1.  
Notice that the model itself holds.

*t-stat for b ≈ 10* (*R*2 ≈ 0.91 for the model with all *p* = 100 regressors) The red line locates the number of leveraged points, and the green line locates the minimum of each series.

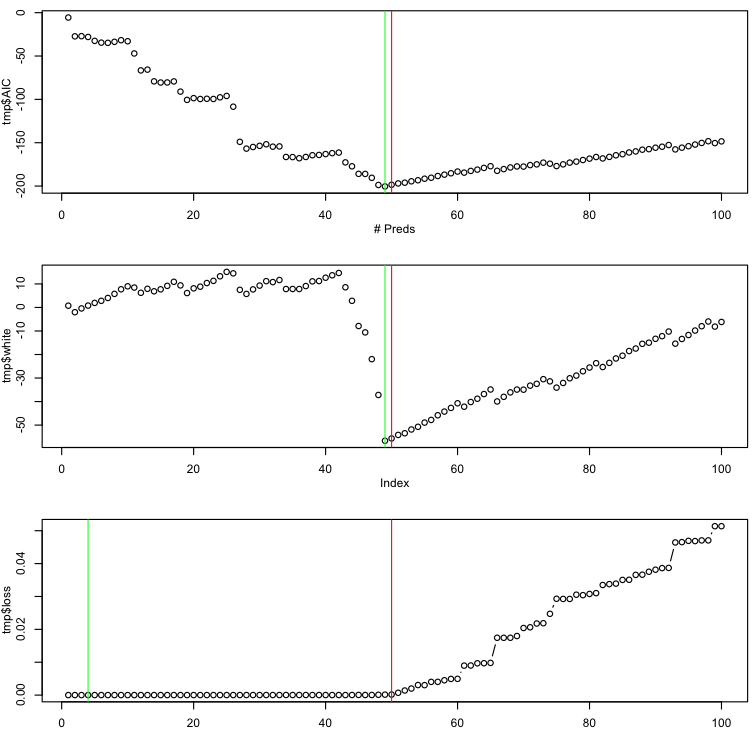


The model with lowest loss has no predictors, but both criteria prefer models with more than 50. Indeed, the loss is monotone increasing for fewer than 50, but both criteria are decreasing.

Also, the loss is not monotone increasing, with the local peak near 50, the number of leveraged cases.

If the level of signal decreases or decreases, the figure looks like these…

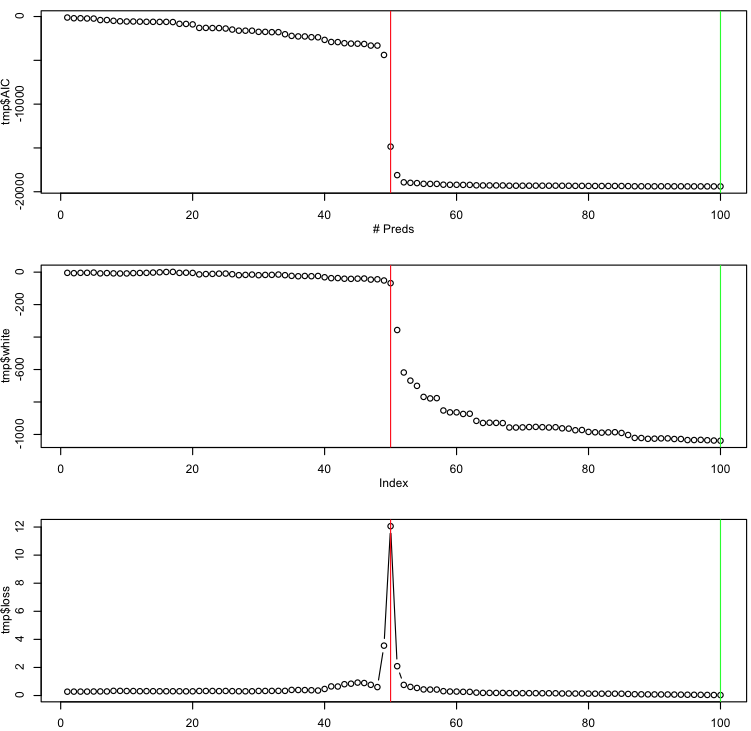
*t-stat for b ≈ 2* (*R*2 ≈ 0.30)



With less signal, the loss is monotone increasing. The loss is minimized with a very small model with just a few variables added.

Both criteria nonetheless pick the 50 predictor model.

*t-stat for b ≈ 350* (*R*2 ≈ 0.9999)



Finally, all *p* predictors are identified… the model with minimum loss uses all 100 features.  
  
though the peak in the loss at 50 has turned into a rather sharp spike.