Model Selection

Chapter 10 in text: Linear Models with R

Model-Building Dilemma – want as many regressors as possible so that the

“information content” will predict adequately Y. However, want as few regressors as necessary because the variance of Yhat will increase as the number of regressors increases. In addition, more regressors in the model can increase costs in data collection and model implementation and model maintenance.

Principle of parsimony: The principle that the most acceptable explanation of an occurrence, phenomenon, or event is the simplest, involving the fewest entities, assumptions, or changes.

Parsimonious means the simplest model/theory with the least assumptions and variables but with greatest explanatory power.

One of the principles of reasoning used in science as well as philosophy is the principle of parsimony or Occam’s razor. The name comes from William of Ockham, a 14th century logician and Franciscan monk who used this principle in his philosophical reasoning. His principle proposed that entities should not be multiplied unnecessary i.e. unnecessary assumptions should be avoided for a theory/conclusion.

A balance between the two goals hopefully leads to the best regression model.

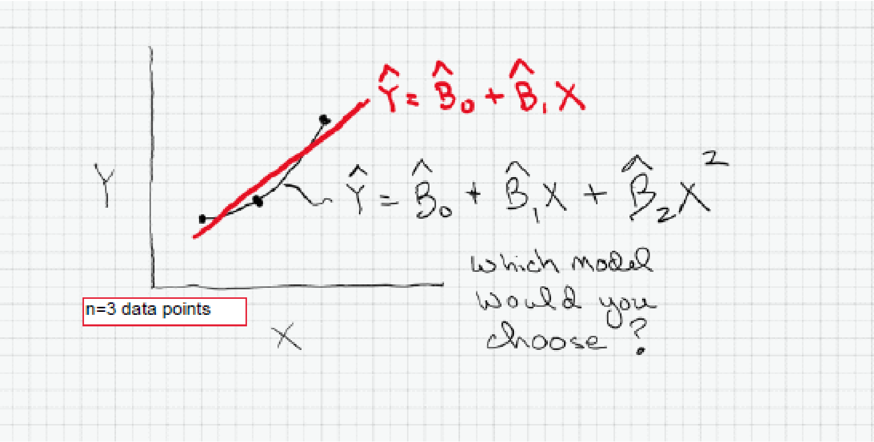
Considerations in model selection: All models are wrong, but some are useful. George Box (1919 - 2013)

No variable selection technique can guarantee the best regression model for the experimental situation of interest.

Model selection and implementation involves more than statistical considerations. For example, a model to predict the probability a patient who enters an emergency room will be admitted for a length of stay greater than or equal to 7 days may include variables that are not collected in all emergency rooms.

Complete reliance on statistical algorithms is not recommended. Need input from the experimenters who have knowledge of the data and the model objective.

The consequences of model misspecification include deleting variables that actually help predict Y, including variables that do not help to predict Y. Incorrect model selection can lead to increased variance in the predicted Y.


$\hat{Y} = \hat{B}_0 + \hat{B}_1 X$
$\hat{Y} = \hat{B}_0 + \hat{B}_1 X + \hat{B}_2 X^2$
Y
X
n=3 data points
which model would you choose?

The straight-line model follows the principle of parsimony. The quadratic model fits the data perfectly, but could be fitting noise - could be "overfitting".

Two approaches to model building to be discussed:

1. All possible regressions
2. Stepwise regression

Note: There are 2^k all possible regressions. For each subset of k regressors where k=1, 2, . . . , , number the regressors in the Full Model is equal to C(k,p) = p!/(k!(p-k)!

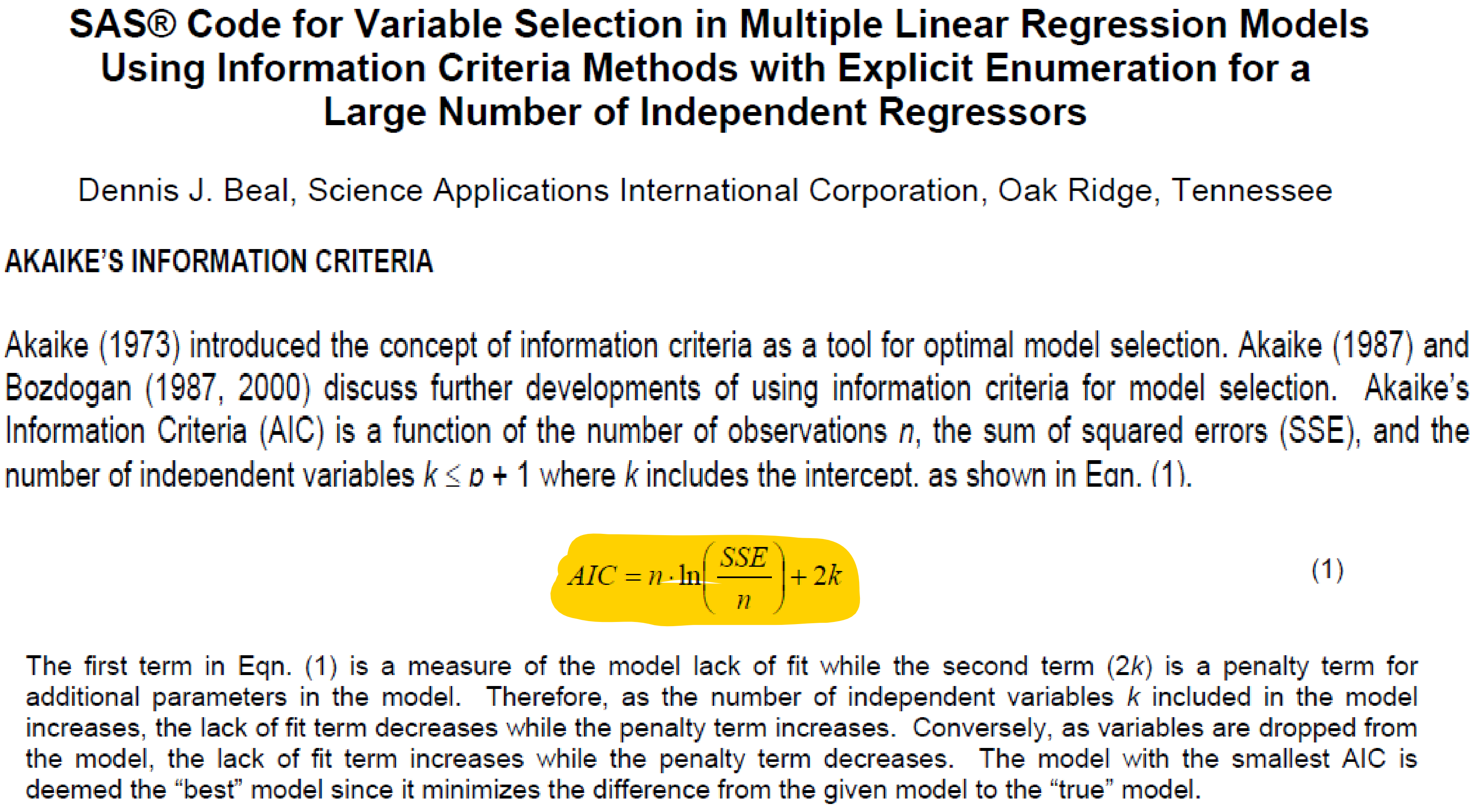
Criteria for Evaluating Regression Models Under Consideration

- Coefficient of Multiple Determination R^2 (want large values)
- Adjusted R^2 – computation considers the number of observations and the number of variables in the model under consideration

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

  Recall R_p^2 = SS(Regression) from the model with p parameters in consideration
  Want $R^2_{adj,p}$ to be large
- MSE_p – want MSE_p to be small. Note an increase in MSE_p can occur if the reduction in SSE_p when adding a variable is not sufficient to compensate for the loss of 1 degree of freedom for estimating σ^2
- $$C_p = \frac{SS_{\mathrm{Res}}(p)}{\hat{\sigma}^2} - n + 2p$$ which estimates the average mean square error of prediction. $(\sigma\text{hat})^2$ is MSE from the Full Model. Want Cp close to p. Note Cp for the Full Model is computed to be p.

Additional measures of model fit: AIC, BIC, and SBC estimate a measure of the difference between a given model and the "true" underlying model. The model with the smallest AIC, BIC, or SBC among all competing models is deemed the best model.



SAS® Code for Variable Selection in Multiple Linear Regression Models Using Information Criteria Methods with Explicit Enumeration for a Large Number of Independent Regressors

Dennis J. Beal, Science Applications International Corporation, Oak Ridge, Tennessee

AKAIKE'S INFORMATION CRITERIA

Akaike (1973) introduced the concept of information criteria as a tool for optimal model selection. Akaike (1987) and Bozdogan (1987, 2000) discuss further developments of using information criteria for model selection. Akaike's Information Criteria (AIC) is a function of the number of observations *n*, the sum of squared errors (SSE), and the number of independent variables $k \leq p + 1$ where *k* includes the intercept, as shown in Eqn. (1).

$$AIC = n \cdot \ln\left(\frac{SSE}{n}\right) + 2k \qquad (1)$$

The first term in Eqn. (1) is a measure of the model lack of fit while the second term (2*k*) is a penalty term for additional parameters in the model. Therefore, as the number of independent variables *k* included in the model increases, the lack of fit term decreases while the penalty term increases. Conversely, as variables are dropped from the model, the lack of fit term increases while the penalty term decreases. The model with the smallest AIC is deemed the "best" model since it minimizes the difference from the given model to the "true" model.

BAYESIAN INFORMATION CRITERIA

Sawa (1978) developed a model selection criterion that was derived from a Bayesian modification of the AIC criterion. Bayesian Information Criteria (BIC) is a function of the number of observations *n*, the SSE, the pure error variance fitting the full model (σ^2), and the number of independent variables $k \le p + 1$ where *k* includes the intercept, as shown in Eqn. (2).

$$BIC = n \ln\left(\frac{SSE}{n}\right) + \frac{2(k+2)n\sigma^2}{SSE} - \frac{2n^2\sigma^4}{SSE^2} \qquad (2)$$

The penalty term for BIC is more complex than the AIC penalty term and is a function of *n*, the SSE, and σ^2 in addition to *k*.

SCHWARZ BAYESIAN CRITERIA

Schwarz (1978) developed a model selection criterion that was derived from a Bayesian modification of the AIC criterion. Schwarz Bayesian Criteria (SBC) is a function of the number of observations *n*, the SSE, and the number of independent variables $k \le p + 1$ where *k* includes the intercept, as shown in Eqn. (3).

$$SBC = n \cdot \ln\left(\frac{SSE}{n}\right) + k \ln n \qquad (3)$$

The penalty term for SBC is similar to AIC in Eqn. (1), but uses a multiplier of ln *n* for *k* instead of a constant 2 by incorporating the sample size *n*.