MA477: Data Science

Lesson 8 Board Sheet — 27 January 2026

United States Military Academy, West Point
Instructor: MAJ Patrick Kuiper

1 Model Selection Lesson Objectives

- Understand the purpose and implementation of best-subsets and various stepwise approaches to regression modeling.
- Gain familiarity with various measures of model performance that are commonly used for model selection.

Discussion Questions: From Least Squares to Regularization

1. Discuss some of the advantages of using linear model (Least Squares). What are some of the limitations?
2. As the number of predictors approaches the number of observations, what changes do you expect in the behavior of the fitted coefficients, even if the training error continues to decrease?
3. In a setting where the number of predictors exceeds the number of observations, what happens?
4. If many different coefficient vectors produce zero training error, what criteria might you use to decide which solution is preferable for prediction or interpretation?
5. Instead of selecting a single subset of predictors, what else can we do?

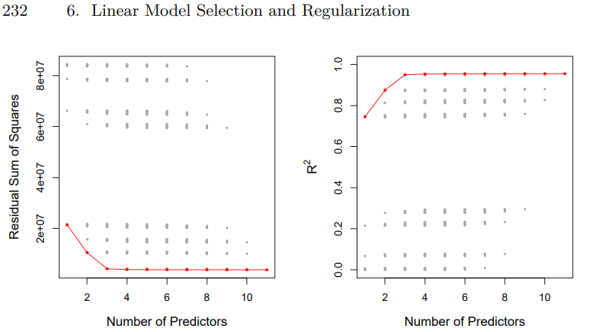

FIGURE 6.1. *For each possible model containing a subset of the ten predictors in the* `Credit` *data set, the RSS and* R^2 *are displayed. The red frontier tracks the* best *model for a given number of predictors, according to RSS and* R^2*. Though the data set contains only ten predictors, the* x*-axis ranges from* 1 *to* 11*, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables.*

Figure 1: Interaction Term

Adjusted Error Measures for Model Selection

Consider a linear regression model with:

- n observations,
- d predictors (excluding the intercept),
- Residual Sum of Squares RSS,
- Error variance estimate $\hat{\sigma}^2$ (typically from the full model),
- Total Sum of Squares $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

Mallows' C_p

$$C_p = \frac{1}{n}\left(RSS + 2d\,\hat{\sigma}^2\right)$$

Akaike Information Criterion (AIC) For least squares regression with Gaussian errors:

$$AIC = \frac{1}{n}\left(RSS + 2d\,\hat{\sigma}^2\right)$$

(up to an additive constant).

Bayesian Information Criterion (BIC)

$$BIC = \frac{1}{n}\left(RSS + \log(n)\,d\,\hat{\sigma}^2\right)$$

(up to an additive constant).

Adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$