Model Selection and Multicollinearity

Bret Larget
Departments of Botany and of Statistics
University of Wisconsin—Madison
February 20, 2007

SAT scores

- Data analysis illustrates model selection and multicollinearity.
- Data set from 1982 on all fifty states.
- Variables:
  - sat: State average SAT score (verbal plus quantitative)
  - takers: Percentage of eligible students that take the exam
  - income: Median family income of test takers ($100)
  - years: Average total high school courses in English, science, history, mathematics
  - public: Percentage of test takers attending public high school
  - expend: Average state dollars spent per high school student ($100)
  - rank: Median percentile rank of test takers

The Big Picture

- When there are many possible explanatory variables, often times several models are nearly equally good at explaining variation in the response variable.
- $R^2$ and adjusted $R^2$ measure closeness of fit, but are poor criteria for variable selection.
- AIC and BIC are sometimes used as objective criteria for model selection.
- Stepwise regression searches for best models, but does not always find them.
- Models selected by AIC or BIC are often overfit.
- Tests after model selection are not valid, typically.
- Parameter interpretation is complex.

Geometry

- Consider a data set with $n$ individuals, each with a response variable $y$, $k$ explanatory variables $x_1, \ldots, x_k$, plus an intercept 1.
- This is an $n \times (k+2)$ matrix.
- Each row is a point in $k+1$ dimensional space (if we do not plot the intercept).
- We can also think of each column as a vector (ray from the origin) in $n$ dimensional space.
- The explanatory variables plus the intercept define a $k+1$ dimensional hyper-plane in this space. (This is called the column space of $X$.)
Geometry (cont.)

- The vector $y = \hat{y} + r$ where $r$ is the residual vector.
- In least squares regression, the fitted value $\hat{y}$ is the orthogonal projection of $y$ into the column space of $X$.
- The residual vector $r$ is orthogonal (perpendicular) to the column space of $X$.
- Two vectors are orthogonal if their dot product equals zero.
- The dot product of $w = (w_1, \ldots, w_n)$ and $z = (z_1, \ldots, z_n)$ is $\sum_{i=1}^{n} w_i z_i$.
- $r$ is orthogonal to every explanatory variable including the intercept.
- This explains why the sum of residuals is zero when there is an intercept.
- Understanding least squares regression as projection into a smaller space is helpful for developing intuition about linear models, degrees of freedom, and variable selection.

$R^2$

- The $R^2$ statistic is a generalization of the square of the correlation coefficient.
- $R^2$ can be interpreted as the proportion of the variance in $y$ explained by the regression.
- $$R^2 = \frac{SS_{Reg}}{SS_{Tot}} = 1 - \frac{SS_{Err}}{SS_{Tot}}$$
- Every time a new explanatory variable is added to a model, the $R^2$ increases.

Adjusted $R^2$

- Adjusted $R^2$ is an attempt to account for additional variables.
- $$\text{adj } R^2 = 1 - \frac{\text{MSErr}}{\text{MSTot}} = 1 - \frac{\text{SSErr}/(n - k - 1)}{\text{SSTot}/(n - 1)}$$
- The model with the best adjusted $R^2$ has the smallest $\hat{s}^2$.

Maximum Likelihood

- The probability of observable data is represented by a mathematical expression relating parameters and data values.
- For fixed parameter values, the total probability is one.
- Likelihood is the same expression for this probability of the observed data, but is considered as a function of the parameters with the data fixed.
- The principle of maximum likelihood is to estimate parameters by making the likelihood (probability of the observed data) as large as possible.
- In regression, least squares estimates $\hat{\beta}_i$ are also maximum likelihood estimates.
- Likelihood is only defined up to a constant, typically.
 Variable Selection

AIC

- Akaike’s Information Criterion (AIC) is based on maximum likelihood and a penalty for each parameter.
- The general form is
  \[ AIC = -2 \log L + 2p \]
  where \( L \) is the likelihood and \( p \) is the number of parameters.
- In multiple regression, this becomes
  \[ AIC = n \log \left( \frac{RSS}{n} \right) + 2p + C \]
  where \( RSS \) is the residual sum of squares and \( C \) is a constant.
- In R, the functions AIC and extractAIC define the constant differently.
- We only care about differences in AIC, so this does not matter (so long as we consistently use one or the other).
- The best model by this criterion minimizes AIC.

Statistics 572 (Spring 2007)  Multiple Linear Regression  February 20, 2007  9 / 14

 BIC

- Schwartz’s Bayesian Information Criterion (BIC) is similar to AIC but penalizes additional parameters more.
- The general form is
  \[ BIC = -2 \log L + (\log n)p \]
  where \( n \) is the number of observations, \( L \) is the likelihood, and \( p \) is the number of parameters.
- In multiple regression, this becomes
  \[ BIC = n \log \left( \frac{RSS}{n} \right) + (\log n)p + C \]
  where \( RSS \) is the residual sum of squares and \( C \) is a constant.
- In R, the functions AIC and extractAIC also find BIC setting with the extra argument \( k=\log(n) \) where \( n \) is the number of observations.
- The best model by this criterion minimizes BIC.

Statistics 572 (Spring 2007)  Multiple Linear Regression  February 20, 2007  10 / 14

 Stepwise Regression

- If there are \( p \) explanatory variables, we can in principle compute AIC (or BIC) for every possible combination of variables.
- There are \( 2^p \) such models.
- Instead, we typically begin with a model and attempt to add or remove variables that decrease AIC the most, continuing until no single variable change makes an improvement.
- This process need not find the global best model.
- It is wise to begin searches from models with both few and many variables to see if they finish in the same place.

Statistics 572 (Spring 2007)  Multiple Linear Regression  February 20, 2007  11 / 14

 R code

- The R function step searches for best models according to AIC or BIC.
- The first argument is a fitted lm model abject. This is the starting point of the search.
- An optional second argument provides a formula of the largest possible model to consider.
- Examples:
  
  \begin{verbatim}
  > form = formula(sat ~ takers + income + public + expend + years + rank)
  > fit.full = lm(form,data=SAT,subset=SAT$state != "Alaska")
  > aic1 = step(fit.full)
  > fit.0 = lm(sat ~ 1,data=SAT,subset=SAT$state != "Alaska")
  > aic2 = step(fit.0,scope=form)
  > bic1 = step(fit.full,k=log(49))
  \end{verbatim}

Statistics 572 (Spring 2007)  Multiple Linear Regression  February 20, 2007  12 / 14
Multicollinearity

- Multicollinearity is the situation where \( k = 2 \) or more explanatory variables lie very close to a hyper-plane of smaller dimension.
- In the most common case, two variables are highly correlated and their vectors are close to the same line.
- When multicollinearity is present, important variables can appear to be non-significant and standard errors can be large.
- Estimated coefficients can change substantially when parameters are added or dropped.
- Multicollinearity typically occurs when two or more variables measure essentially the same thing (possibly in different ways).
- It is best to remove excess variables to eliminate multicollinearity.
- Examinations of correlations is a first step.
- (Show with SAT data.)

3 variables

- It is possible for three variables to be multi-collinear without any pair-wise correlations being extreme.
- Picture points close to a plane or sheet held at an angle.
- The point would not look close to a line projected into any of the three pairs of dimensions.
- Demonstration!
- Principle components analysis is an alternative remedy for multicollinearity.