Building the Regression Model I: Model Selection and Validation

Paweł Polak

November 8, 2017

Linear Regression Models - Lecture 11

Content:

- Variable Selection
- Model Selection
- R^2, R_a^2, and Information Criteria: AIC, BIC
- Predicted Sum of Squares (PRESS)
- Variable selection techniques: Forward Selection or Backward Elimination
- Stepwise Regression

General Linear Model

- *Independent responses* of the form $Y_i \sim N(\mu_i, \sigma^2)$, where

$$\mu_i = \mathbf{X}_i^\top \boldsymbol{\beta}$$

for some known vector of *explanatory* variables $\mathbf{X}_i^\top = (X_{i1}, \ldots, X_{ip})$.

- Unknown *parameter* vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{P-1})^\top$, where $P < N$.

- This is the *linear model* and is usually written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

(in vector notation) where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{P-1} \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix},$$

where $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, for $i = 1, 2, \ldots, N$.

Variable Selection

- We often have data on a *large* number of *explanatory* variables and wish to build a regression model using some *subset* of them.
- Using fewer variables makes the resulting model more *manageable*, especially if more data is to be collected at a later time.
- Including unnecessary variables yield *less precise* inferences and *complicate interpretation*.

The principle of parsimony

- The principle of parsimony says that when two competing models have the *same predictive* power, the model with the *lower* number of parameters should be used.
- *Occam's Razor* – simple models are preferred over complicated ones.

Model Selection

- All possible regressions
  - Consider *all possible subsets* of the pool of explanatory variables and find the "*best*" model according to some *criteria*.
- Automatic methods
  - When the number of explanatory variables is *large* it is more efficient to use a *search algorithm* to find the "best" model.
- Different criteria may be used to select the best model, e.g., Adjusted R^2, C_p, *AIC and BIC*.
- These criteria assign scores to each model and allow us to choose the model with the *best score*.

R^2

- The *coefficient of multiple determination* is

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- The *proportion* of the *variability in Y explained* by regression model.

- Since R^2 *increases* with the size of the model, it is *not* a good criterion for variable selection – always chooses to include all variables.

Adjusted R^2

- The *adjusted coefficient of multiple determination*, uses the *mean squares* instead of the sums of square, i.e.,

$$R_a^2 = 1 - \frac{MSE}{MST} = 1 - \left(\frac{N-1}{N-P}\right)\frac{SSE}{SST}$$

- Since the term includes the number of model parameters, P, it *penalizes* for *model complexity*.

AIC

- *Akaike's Information Criterion* (*AIC*) tries to balance the *conflicting* demands of model *accuracy* and *parsimony*.

- It can be expressed as:

$$AIC_P = N\log(SSE/N) + 2P$$

- $N\log(SSE/N)$ measures the *model fit*

- $2P$ is the penalty for using P *parameters*

- *Low values* indicate a better model.

SBC (aka BIC)

- Several modifications of AIC have been suggested.
- *Schwarz's Bayesian Information Criterion* (BIC) is defined as:

$$BIC_P = N \log(SSE/N) + (\log N)P$$

- $N \log(SSE/N)$ measures the model fit
- $(\log N)P$ is the penalty for using more parameters
- The difference between AIC and BIC lies in the *severity* of the penalty.
- The penalty is *larger* for BIC when $N > 8$.
- Hence, BIC tends to favor *more* parsimonious models compared to AIC which has a tendency to *overfit* (i.e., include too many explanatory variables).

PRESS

- The *prediction sum of squares* (*PRESS*) criterion measures how well the fitted values for a model can predict the observed response.

- Procedure:
  - *Remove* the i-th observation and fit the model with the remaining $N-1$ observations to obtain $\hat{Y}_{i(i)}$
  - Use this model to calculate the *prediction error* for the left-out observation $Y_i - \hat{Y}_{i(i)}$
  - Repeat this process for each observation

- The *PRESS* statistic is then defined

$$PRESS_P = \sum_{i=1}^{N} (Y_i - \hat{Y}_{i(i)})^2$$

- The model with the *smallest PRESS* statistic is considered "best".

- Leaving one item out at a time is known as *leave-one-out cross-validation.*

Variable selection techniques

- When the number of explanatory variables is *large* it is not feasible to fit all possible models.
- It is more efficient to use a *search algorithm* to find the best model.
- A number of such algorithms exist, including *forward selection*, *backward elimination* and *stepwise regression*.
- Assume we are choosing from a set of P possible explanatory variables $v_k, k = 1, \ldots, P$.
- In each algorithm our goal is to find the *subset* of v_k, $k = 1, \ldots, P$, that best balances model fit and parsimony.
- We discuss each algorithm in detail.

Forward Selection

(1) Fit the P simple linear regression models:

$$Y_i = \beta_0 + \beta_1 v_{ki} + \varepsilon_i, \qquad k = 1, \ldots P - 1$$

- Set $X_1 = v_k$, where v_k is the variable that has the *most significant* regression coefficient (i.e., the smallest p-value)

- If no variable is significant (e.g., none of the p-values are smaller than a preset significance level α) the algorithm stops.

(2) *Lock* in the variable found in (1), and repeat the procedure with models that include *two* explanatory variables:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 v_{ki} + \varepsilon_i, \qquad k = 1, \dots P-1, v_k \neq X_1$$

- Set $X_2 = v_k$, where v_k is the variable that has the *most significant* regression coefficient.
- If no variable is significant stop the algorithm.

(3) Continue until *no* remaining v_k generate a p-value that is smaller than the preset significance level α.

Comments

- The criteria for choosing whether to include a new variable can vary.
- As an alternative to using p-values one can instead use a criteria such as the *AIC*.
    - In each step choose the variable whose inclusion *lowers* the *AIC* the most.
    - If no variables lower the *AIC* than stop the algorithm.
- The R function *step()* can be used to perform variable selection.
    - It chooses to include variables using *AIC*.
    - To perform forward selection we need to specify a *starting model* and the *range of models* to be examined in the search.

Backward Elimination

(1) Start by fitting a model that includes *all* possible variables:

$$Y_i = \beta_0 + \beta_1 v_{1i} + \ldots + \beta_{P-1} v_{Pi} + \varepsilon_i, \qquad k = 1, \ldots, P-1$$

- Find the variable v_k which has the *least significant* regression coefficient (i.e., the largest p-value).
- If its p-value is smaller than some preset significance level, stop the algorithm, otherwise drop the variable.

(2) Fit the *largest* model excluding the dropped v_k.

- Find the variable which has the least significant regression coefficient.
- If its p-value is smaller than some preset significance level, stop the algorithm, otherwise drop the variable.

(3) Continue until the algorithm stops.

Comments

- Alternatively, *AIC* or *BIC* can be used as a criteria for determining whether to drop variables.
    - Start with a *full* model.
    - In each step choose the variable whose *exclusion* lowers the *AIC* the most.
    - If the exclusion of any variable does not lower the AIC than stop the algorithm.

Stepwise regression

(1) Start in the *same* manner as in *forward* selection and add the *most significant* variable from a series of P simple linear regressions.

(2) Once a new variable has been included in the model, check other variables already included in the model for their *partial* significance.

  - Remove the least significant explanatory variable whose p-value is greater than the preset significance level.

(3) Continue until no variables can be added and none removed, according to the specified criteria.

- Note AIC can be used instead of p-values.

Least Angle Regression (LAR)

The LARS algorithm, introduced by Efron et al. (2004) is a general estimation algorithm which can be used for computation of the LASSO paths, or the LAR paths. The steps of the algorithm are:

- as in classic Forward Selection, we start with all coefficients equal to zero;
- we find the predictor most correlated with the response, say x_{j_1};
- we take the largest step possible (importantly it has a closed form expression) in the direction of this predictor until some other predictor, say x_{j_2}, has as much correlation with the current residual; (Forward Selection would continue along this direction)
- when we reach equal correlation, LARS continues along a direction which is *equiangular* between the two predictors until a third variable x_{j_3} becomes most correlated with residuals;
- LARS then proceeds *equiangularly* between x_{j_1}, x_{j_2}, and x_{j_3}, until fourth variable enters, and so on.