Lecture 22

Wednesday, July 24, 2019 3:06 PM

$\boxed{2}$ Forward Selection

1. Fit the intercept-only model (i.e., the one with no explanatory variables):

$$y = \beta_0 + \varepsilon$$

2. Fit the q simple linear regressions:

$$y = \beta_0 + \beta_1 v_k + \varepsilon, \quad k = 1, 2, \ldots, q$$

3. By some <u>decision criteria</u> compare the q models in Step 2 to the baseline model from step 1. Identify the explanatory variable v_ℓ that improves the model to <u>most</u>.

4. Lock in the best explanatory variable from Step 3 as x_1:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

5. Fit $q-1$ two-variable models:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 v_k + \varepsilon \quad k = 1, 2, \ldots, q, \; k \neq \ell$$

   And by the selected decision criteria compare these to the model from step 4. Whichever explanatory improves the model the most is locked in as x_2.

6. Continue this process until no remaining explanatory variable improves the model. The model that can't be improved upon is the final model.

<u>Remark:</u> The decision criteria could be:

- R^2_{adj}
- AIC, AICc, BIC
- p-values corresponding to the formal comparison of full vs. reduced models

In terms of computational efficiency, this is much faster (even in the worst case scenario) that "all possible" models. In the worst case we fit

$$1 + q + (q-1) + (q-2) + (q-3) + \cdots + 2 + 1 = 1 + \frac{q(q+1)}{2} \text{ models}$$

* this is $O(q^2)$ vs. $O(2^q)$

$\boxed{3}$ Backward Elimination

1. Fit the full model that includes all q explanatory variables:

$$y = \beta_0 + \beta_1 v_1 + \beta_2 v_2 + \cdots + \beta_q v_q + \varepsilon$$

2. Fit the q models that result from eliminating one of the v_k's, $k = 1, 2, \ldots, q$. Each of these will have $q-1$ explanatory variables

3. By some decision criteria compare each of those models to the one from Step 1. Whichever model fits the best, adopt it as the new optimal model (and hence remove the associated v_k from consideration as it is least important).

4. Fit the $q-1$ models that result from (one at a time) eliminating each of the explanatory variables from the model from Step 3. By the selected decision criteria, compare each of these models to the on from Step 3 and determine which one provides the biggest improvement. Choose this model as the new optimal one (and hence eliminate the associated v_k from consideration since it is the least important).

5. Continue this process until no remaining terms in the model can be removed without worsening how well the model fits. The model that can't be improved upon is the final model.

<u>Remarks:</u>

- The decision criteria here are the same as the ones listed above.
- The number of models fit in the worst case scenario (when none of the q explanatory variables are important) is also

$$1 + \frac{q(q+1)}{2}$$

In forward selection and backward elimination, once the explanatory variables enter/leave the model, this decision is final and cannot be reversed. An alternative approach that more closely mimics the "all possible regressions" approach, but that is still more computationally efficient is a <u>hybrid</u> approach that combines the decision processes of both forward selection and backward elimination.

$\boxed{4}$ Hybrid Selection

1. Start as in Forward Selection.

2. At each stage you have three decisions
   - Add an important variable into the model
   - Eliminate an unimportant variable from the model
   - Stop

   Take the action that improves the model the most (by some decision criteria) relative to what the current optimal model is.

   * Note that to make this decision, we must fit q models at every step. If ℓ explanatory variables are in the current best model, then we must fit the ℓ models with $\ell-1$ variables that result from backward elimination and the $q-\ell$ models with $\ell+1$ variables that result from forward selection.

3. Continue until the model cannot be improved either by adding a new explanatory variable or eliminating an existing explanatory variable.

<u>Remarks:</u>

- Here an explanatory variable could enter/leave the model several times.

* <u>Note</u>: These stepwise selection algorithms may not find the "optimal" model. And the approaches may not even choose the same model. But in general, they will find you a "pretty good" model.

<u>Example</u>: Credit Card Data.

This dataset contained $q=11$ possible explanatory variables. By "all-possible regressions" and using R^2_{adj} as the decision criteria, we found the optimal model to be the one with the following explanatory variables:

- Income
- Limit
- Rating
- Cards
- Age
- Gender
- Student

Using AIC as the decision criteria, all three of the stepwise algorithms identified the optimal model as the one containing the following explanatory variables:

- Income
- Limit
- Rating
- Cards
- Age
- Student