MA 542 2018 SPRING

Applied Regression Analysis

Chapter 9

Model Selection and Validation

Model Building Process

In this chapter we discuss the idea of model selection and validation. (i.e., How to choose the best model for the data (selection) and the application (validation))

- Model building can be thought as a four step process.

  1. Data collection and preparation.
  2. Reduction of explanatory (or predictor) Variables.
  3. Model refinement and selection.
  4. Model validation.

.

Data Collection and Preparation

The data collection process totally based on the design of the research study.

- There are basically four types of research designs.

  1. Controlled Experiments.
  2. Controlled Experiments with Co-variates.
  3. Confirmatory Observational Studies.
  4. Exploratory Observational Studies.

.

Data Collection : *Controlled Experiments*

In a controlled experiment, the experimenter controls the levels of explanatory variables (factors) and assign treatments (combination of levels) to experimental units.

Then the response is observed from those units.

E.g:- Effect of the size of a graphic presentation and the time allowed for junior executives on a measure of the accuracy of the analysis.

- **Factors :** Size of a graphic, Time allowed.
- **Treatments :** Combination of size levels and time levels.
- **Response (Y):** Accuracy of the analysis.
- **Experimental Units :** Junior executives.

Data Collection : *Controlled Experiments with Co-variates*

This method is used when we need to include some other variables in the regression model which are not in the design study.

Uncontrolled variables which can not be incorporated in the design study are called co-variates.

E.g:- In our previous example : Suppose they also believe that gender and number of years of education of junior executives also effect the accuracy.

- **Controlled Variables :** Size (X_1), Time(X_2).
- **Co-variates :** Gender (X_3), # of years (X_4).

Data Collection : *Confirmatory Observational Studies*

Observational studies are used when the level of variables cannot be controlled.

Confirmatory Observational Studies are used when knowledge from previous studies are available. We also can include new variables to the study.

Variables involved in the hypothesis are called the primary variables.

New variables are called control variables

E.g:- Study : Effect of smoking on Lung cancer (unable to control smoking).

- **Primary Variable :** Smoking.
- **New Variables :** Age, Gender, Race.

Data Collection : *Exploratory Observational Studies*

This method is used when the prior knowledge is not available and level of variables cannot be controlled.

Basically a bunch of variables are collected and the experimenter want to see which ones have the most effect on the response.

E.g:- Researchers are interested in the stability of the weight over time.

- **Possible Predictors :** Gender, amount of exercise, diet, etc,....

Data Preparation:
Once data is collected it should be checked and plots prepared to identify gross data errors, extreme outliers, etc,...

Model Investigation :

Once we are conformable that data is correct, the analysis can be begun.

Here we try to identify the function form of each predictor variable and interaction terms should be in the model.

Experimenters previous knowledge about the study can be used to decide appropriate transformations and interactions to include.

- Check the scatter plot to determine the strength and the nature of the relationship.
- Check the residuals what the function form of the relationship is (linear, non-linear etc).
- Check the relationship between interaction terms to determine the interaction terms.

Reduction of Explanatory Variables :

The purpose of any research study is to determine which variable influence the response variable the most and to capture the most information with the fewest amount of variables. The most variable reduction is done in Exploratory Observational Studies.

Here we need to balance of having too many variables with leaving out important variables.

E.g:- Identifying which variables contribute the most to long term weight stability .

Model Refinement and Selection :

The initial model or several "GOOD" regression models need to be run and analyzed.

The residual plots and other diagnostics method can be used to determine if the model is appropriate or needs to be changed.

Once we find the model that gives the best fit and follows all the assumptions, we can move to the model validation.

Model validation is simply looking to see if our model make sense.

Criteria for Model Selection

• At the model selection stage, it would be helpful to have some information about which models are better than others and which variables are important or not.

• Consider a model with 4 explanatory variables. Then there are $16(2^4)$ different models that can be fit and analyzed.

• That is very tedious work.

• Computer programs can be used to do this , but we have to give it some kind of **criteria of determining** which model is better than another.

.

Criteria for Model Selection

- The following 6 criteria are options for use in model selection:

  1. R_p^2 : Coefficient of multiple determination.
  2. $R_{a,p}^2$: Adjusted coefficient of multiple determination.
  3. C_p : Mellow's C_p criterion.
  4. AIC_p : Akaike information criterion.
  5. SBC_p : Schwarz Bayesian criterion.
  6. $PRESS_p$: Prediction sum of squares.

Here p is the number of parameters in the model.

.

Coefficient of multiple determination (R_p^2)

- This is more of an abstract method of model fit.
- Sets of variables are compared and the one with the highest R^2 is assumed to be better.

$$R_p^2 = 1 - \frac{SSE_p}{SSTO}$$

Key point: if you add another variable to the model and the R^2 does not increase a lot, then that variable is probably not important.

.

Adjusted coefficient of multiple determination ($R^2_{a,p}$)

- R^2 itself does not take into account the number of variable in the model, plus there is that sticky problem of the ever increasing R^2.

- $R^2_{a,p}$ is an adjusted R^2 that takes into account the number of parameters.

$$R^2_{a,p} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE_p}{SSTO}$$

Key point: If the $R^2_{a,p}$ does not increase too much, then the last variable(s) do not contribute to the fit.

.

Mallows C_p Criterion

• This criterion is concerned with the total mean square error of the n fitted values for each regression model.

• Here we calculate:

$$C_p = \frac{SSE_p}{MSE(X_1, X-2, \cdots, X_{p-1})} - (n-2p),$$

where
SSE_p : SSE for the fitted subset with p parameters.
$MSE(X_1, X-2, \cdots, X_{p-1})$: MSE for the model with P parameters (considering all $P-1$ predictors).

• **Key point:** Want to find models where this C_p value is small and the value is near p.

When C_p is small: the mean squared error is small.

When C_p is close to p: bias of the regression model is small.

Note: We can show that $E(C_p) = p$, when $E[\hat{Y}] = E[Y]$,
(i.e. the model is unbiased).

AIC_p and SBC_p

- AIC_p : Akaike information criterion.

$$AIC_p = n \ln(SSE_p) - n \ln(n) + 2p$$

- SBC_p : Schwarz bayesian criterion.

$$SBC_p = n \ln(SSE_p) - n \ln(n) + \ln(n)p$$

- Both depend on SSE and number of parameters in the model(p).

Key point: We want to find models where both of these criteria are small. .

Prediction sums of squares ($PRESS_p$)

• The prediction sums of squares is a measure of how well the use of the fitted values for the subset model can predict the observed responses Y_i.

$$PRESS_p = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i(i)})^2,$$

where

$Y_{i(i)}$: predicted value for the i^{th} case for the regression function fitted without the i^{th} case.

Key point: Want to find models that have smaller PRESS values.

.

Automatic Search Procedures

Here we discuss how to use the computer to find the best model.

Here are few methods:

1. 'Best' subsets algorithms
2. Stepwise Regression.
   - Forward Stepwise Regression.
   - Forward Selection.
   - Backward Elimination.

.

'Best' subsets algorithms

• These procedures consider all the possible subsets and find **the 'best' subset of models** based on our search criteria (for example, one of the 6 mentioned above).

• We will receive information about the best model and several other good models based on our criteria.

• So, we essentially have our choice of models that fit the criteria.

• This method may not be helpful when there are many (30 or 40) explanatory variables.

Stepwise Regression

• This procedures goes through a step-by-step process of adding variables until the best model is produced based on your search criteria.

• At each step, t-test (or partial F test) will be performed to determine if that variable is appropriate.

• This model differs from the previous algorithm in that it give us the 'best' model possible (*a single model*) given our variables and criterion. On the other hand, the best subset algorithm gives several 'good' regression models.

Forward Stepwise Regression

- Here is how this process works:
  - First, a SLR model is fit for each X variable.
  - The variable with the largest significant t statistic is kept.
  - Next, that first variable is run in the model with each of the remaining variables.
  - The next variable with the largest significant t is then kept.
  - At each stage with more than one variable, there is also another test that is done to determine if a variable in the model should be dropped.
  - This process is repeated until no significant t values are found and all variables should be kept.
- Variables are tested for addition to model and then removal from the model.
- The result is the best fitting model using this search criteria with significant parameters.

Other Stepwise Procedures

Forward Selection:

- This is a simplified version of Forward Stepwise procedure.
- Each variable is added in the same way, one by one, but there is no test as to whether a variable should be dropped.

Backward Elimination:

- The idea is the same as the forward selection, except all variables are put in the model at first..
- Then working backwards, each variable is tested for it to be removed.
- So, each variable is removed one at a time, instead of added one at a time.

Model Validation

- Once we have found the 'best' model, next we need to determine if our model is valid.

- There are three basic ways to validate a regression model:

  1. Collection of new data to check model and its predictive ability.
  2. Comparison of results with theoretical expectations, earlier empirical results, and simulation results.
  3. Use of a holdout sample to check the model and its predictive ability.

.

Collection of New Data

- Here basically, we collect new data using the same method to see if it fits our data.
- Predictive ability of a model can be measured by using **the mean squared prediction error:**

$$MSPR = \frac{\sum_{i=1}^{n^*}(Y_i - \hat{Y}_i)^2}{n^*}$$

- Key point: If $MSPR \simeq MSE$, then the model is valid.

Note: Collection of new data may not be easy (possible) for some studies.

E.g: Observational studies may be difficult to replicate.

.

Comparison with Theory

- We may have some idea about how our results may relate in terms of theory or early results.

E.g: If we find that eating more cake increases weight loss, then that might be a contradiction to theory (or to previous results).

- This method is not that common in regression.

.

Cross- Validation

- The most common way to validate the model is through cross-validation.
- Here the data set is randomly splitted into two sets (training set and validation set)
- A set (training set) is used to fit the regression model and the other set (validation set) is used to determine the predictive ability of the model.
- We can then use the *MSPR* or a selection criteria for model fit to test the predictive ability.

.