Building a Regression Model I: Model Selection and Validation

Lecture 3
February 6, 2007
Psychology 791
Today’s Lecture

- A bit of Chapter 9 - the practical side of model fitting.
The Path to Finding a Model

- Up until now, we have discussed various regression models that can be applied to data.

- One thing that we have failed to discuss is, which model is the "best" model.

- How do you choose which model is the most appropriate for the data?

- We will discuss the idea of model selection and validation.
Process

- Model Building can be thought of as a 3 (or 4) step process:
  1. Data collection and preparation.
  2. Reduction of explanatory or predictor variables.
  4. Model validation.
Overview

Model Building/Selection Process
- Process
- Building a Regression Model

Data Collection

Data Preparation

Preliminary Model Investigation

Reduction of Explanatory Variables

Model Refinement and Selection

Criteria for Model Selection

Automatic Search Procedures

Model Validation

Wrapping Up

Building a Regression Model

FIGURE 9.1
Strategy for Building a Regression Model

Collect data → Preliminary check on data quality → Diagnostics for relationships and among interactions → Identify several potentially useful subsets of explanatory variables (include known essential variables) → Investigate curvatures and interaction effects more fully → Further model diagnostics needed? → Select tentative model → Model refinement and selection → Model validation → Final regression model

Lecture 3 Psychology 791
Step 1: Data Collection

- The data collection process really stems from the design of the research study.

- Four types of research designs are outlined in the book:
  - Controlled Experiments.
  - Controlled Experiments with Covariates.
  - Confirmatory Observational Studies.
  - Exploratory Observational Studies.
Controlled Experiments

- In a controlled experiment, the experimenter controls the levels of the explanatory variables and assigns treatments.

- With control over the experimental variables, data collection simply stems from collecting observations from the treatment conditions.

- Example: Subjects are randomly assigned to four treatment groups to determine which condition is best for weight loss.
Controlled Experiments with Covariates

- This type of design incorporates the idea above, in that the experimenter controls the levels of the explanatory variable.

- There are, however, other uncontrolled variables - or -

- Example: Subjects are randomly assigned to four treatment groups to determine which condition is best for weight loss.
  - However, they also believe that age and gender might also be a factor in weight loss.
Confirmatory Observational Studies

- Studies that are observational in nature that are designed to test specific hypotheses.

- The response variable cannot be controlled and are simply observed.

- Example: Previous research has shown that high stress levels lead to weight gain.
  - They are unable to control the stress levels of the subjects, so they simply measure it and see how it relates to weight gain.
## Exploratory Observational Studies

- These designs are when uncontrolled variables are measured in observational setting with no specific hypotheses.

- Basically a bunch of variables are collected and they want to see which ones have the most effect the response variable.

- Example: Researchers are interested in the stability of weight over time.
  - They are unsure of which variables are most important, so they gather a bunch of information they think might be predictive, for example, gender, amount of exercise, diet, etc.
Data Preparation

- Each of these four research designs outlines the way in which the data is collected.

- Once data is collected and is put into a data file, it should be checked for errors.

- Extreme outliers that may be input errors, measurement errors, etc.
Model Investigation

- Once you are comfortable that the data is correct, you can begin the analyses.

- During this time, you look for any clues that the data might give you as to the nature of the relationship.

- Check the scatter plots to determine the strength and nature of the relationship.

- Check residuals to determine what the functional form of the relationship is (linear, non-linear, etc).

- Check relationships between independent variables to determine if some interaction may exist.

- This step should involve a combination of prior research knowledge and brute force.
Reduction of Explanatory Variables

- The purpose of any research study is to determine which variables influence the response variable the most.

- We ultimately want to capture the most information with the fewest amount of variables.

- It doesn’t make sense to use three IV’s if two do just as well in prediction.

- The following is how variable selection works for each research design.
Controlled Experiments

- The number of IV’s is determined in the experimental design of the study.

- Any model alterations pertain to the functional form of the model and not the number of IV’s.

- Example: Still have the four treatment groups relating to weight loss.
Controlled Experiments with Covariates

- Reduction of variables relates to the Covariates.

- Again, the IV’s were determined in the design of the study.

- Model selection involves trying to identify which Covariates effect the relationship and which do not.

- Keep variables in that help, remove variable that do not contribute to the model.

- Example: Four treatment groups that relate to weight loss, but does gender have an effect? How about age?
Confirmatory Observational Studies

- There should be no reduction in variables for these designs.
- We are confirming something that is already known.
- Either it is confirmed or it is not.
- Example: Either stress relates to weight gain or it doesn’t.
Exploratory Observational Studies

- The most variable reduction is done in these types of experimental designs.

- Variable selection is a tricky process for these studies.

- You need to balance having too many variables with leaving out important variables.

- These search processes can be long and tedious.

- Luckily, computers will do this for you.

- This refers to model selection procedures which we will discuss a little later.

- Example: Identify which variables contribute most to long term weight stability.
Model Refinement and Selection

- The initial model is then run and analyzed.
- Things like residual plots are used at this stage to determine if the model is appropriate or needs to be changed.
- Fit of the model can also be determined.
- Once you find the model that gives the best fit and follows all assumptions, you can move to model validation.
- Model Validation is simply looking to see if your model makes 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.

- If you think about a model with 4 IV’s.

- There are 16 \(2^{4}\) different models that can be fit and analyzed.

- That is very tedious work.

- Computer programs will do this work for you, but you have to give it some kind of criteria of determining which model is better than another.
6 Different Criteria

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

1. $R^2_p$
2. $R^2_{a,p}$
3. $C_p$
4. $AIC_p$
5. $SBC_p$
6. $PRESS_p$
This criterion uses the coefficient of multiple determination, $R^2$, as a criteria for selecting variables ($p$ is the number of parameters).

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.

You can think of it as this: if you add another variable to the model and the $R^2$ does not increase a lot, then that variable is probably not important.
\[ R_{a,p}^2 \]

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

- \( R_{a,p}^2 \) is an adjusted \( R^2 \) that takes into account the number of parameters.

\[
R_{a,p}^2 = 1 - \left( \frac{n - 1}{n - p} \right) \frac{SSE_p}{SSTO}
\]

- We again do another inspection method.

- If the \( R_{a,p}^2 \) 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.

- You calculate the following:

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

- 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.
**Overview**

Model Building/Selection Process

Data Collection

Data Preparation

Preliminary Model Investigation

Reduction of Explanatory Variables

Model Refinement and Selection

Criteria for Model Selection

- Criteria for Model Selection
- 6 Different Criteria
- $R^2_p$
- $R^2_{a,p}$
- Mallows' $C_p$ Criterion
- $AIC_p$ and $SBC_p$
- $PRESS_p$

Automatic Search Procedures

Model Validation

Wrapping Up

Lecture 3

AIC<sub>p</sub> and SBC<sub>p</sub>

- Again, two more model selection criterion:

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

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

- You want to find models where both of these criteria are small.
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 $Y_i$.

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

Want to find models that have smaller PRESS values.
Automatic Search Procedures

- We have talked about ways to find good models by using values.
- But, what if we just want to leave it to the computer?
- These Automatic search procedures are built into every stat package to help you find the 'best' model.

- Here are a few methods which we will outline:
  - 'best' subsets algorithms.
  - Stepwise Regression.
  - Forward Stepwise Regression.
  - Forward Selection.
  - Backward Elimination.
’best’ subsets algorithms

- These procedures will find the "best" subset of models based on your search criteria (for example, one of the 6 mentioned above).

- You will receive information about the ’best’ model and several other ’good’ models based on your criteria.

- So, you essentially have your choice of models that fit the criteria.
## 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, an F test will be performed to determine if that variable is appropriate.

- This model differs from the previous algorithm in that it gives you the 'best' model possible given your variables and your criterion.
Forward Stepwise Regression

- Here is how this process works:
  - First, an individual model is fit for each IV.
  - 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.
  - Each stage 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.

- What will result is the best fitting model using this search criteria with significant parameters.
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 you have found the 'best' model, either by research design or brute force, next you need to determine if your 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

- Basically, you collect new data using the same method to see if it fits your data.

- You can measure the predictive ability of a model by using the mean squared prediction error:

  \[
  MS\text{PR} = \frac{\sum_{i=1}^{n^*} (Y_i - \hat{Y}_i)^2}{n^*}
  \]

- Want the MSPR to be close to the MSE.

- One note: Observational studies may be difficult to replicate.
Comparison with Theory

- May have some idea about how your results may relate in terms of theory.

- For example, if you find that eating more cake increases weight loss, then that might be a contradiction to theory.
Cross- Validation

- The most common way to validate the model is through cross-validation.

- In this analysis, you split your sample in half (randomly).

- You use half of the model to fit your model, then use the other half to determine the predictive ability of the model.

- Can then use the MSPR or the selection criteria for model fit.
Final Thought

• Today we talked about the practical side of regression.

• The way models are applied depend in large part on the research design under which the data were collected.

• As we will see, searching for best fitting models can result in trends that do not easily generalize beyond the specific study under investigation.

• As you work through the model building process, be aware of the data and what it is trying to tell you.
Next Time

- Model selection criteria example.
- Methods for searching for variables that have an effect on the response.
- More good clean fun.