THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 62
Building Models

Chris A. Mack
Adjunct Associate Professor

http://www.lithoguru.com/scientist/statistics/

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Model Building

- The general modeling building process:
  - Experimental Design
  - Data Collection
  - Model Building and Refinement
  - Model Validation
- We generally want $n > 10p$ for model building
- Usually the best way to build a model is to be guided by theory

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Automated Model Search

- Given several (or many) predictor variables, linear and/or higher order terms plus interactions, some unneeded, automated search for a model can be very efficient
- Full Search: Run all model subsets
  - For k parameters, there are 2^{k-1} subsets
- Forward Stepwise Regression
  - Add terms one at a time
- Backward Stepwise Regression
  - Start with full model, then remove terms

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

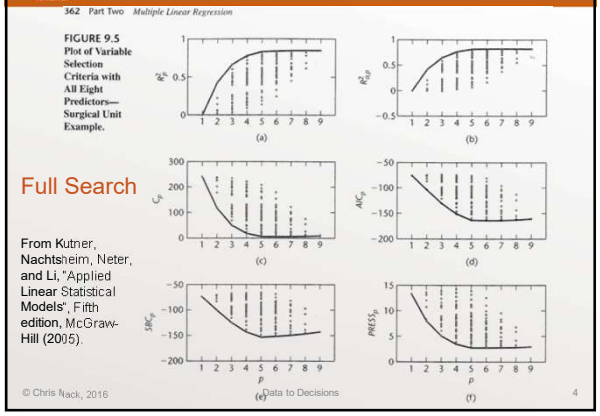
362 Part Two *Multiple Linear Regression*

FIGURE 9.5 Plot of Variable Selection Criteria with All Eight Predictors—Surgical Unit Example.

Full Search

From Kutner, Nachtsheim, Neter, and Li, "Applied Linear Statistical Models", Fifth edition, McGraw-Hill (2005).

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Forward Stepwise Regression

- Write down full possible model
  - Predictors, functions, interactions, etc.
- Pick two significance levels
  - α_{add} (typically 0.05 or 0.10)
  - α_{remove} (typically 0.01 or 0.05)
  - α_{remove} must be > α_{add}
- Step 1: regress y against each model term individually
  - Calculate $t = b_k/SE(b_k)$, and p-value for each term
  - Pick smallest p-value, if it is less than α_{add}, then add it

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Forward Stepwise Regression

- Step 2: repeat step 1, adding one new term to the existing model
  - Calculate $t = b_k/SE(b_k)$, and p-value for each term
  - Pick smallest p-value, if it is less than α_{add}, then add it to the model (now there are two terms in the model)
- Step 3: Check to see if a previously added term should be removed
  - For all previously added terms, find the one with the lowest t-score, If p-value > α_{remove}, remove it
- Repeat steps 2&3 until the model no longer changes
- Backward Stepwise Regression – run in reverse

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Omitted Variable Bias

- If a correct regressor (x_k) is missing from the model, then the remaining model parameters will be biased
  - The bias is proportional to the correlation between the missing x_k and the regressors used in the model
  - If x_k is orthogonal to the other variables, there is no bias

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Uncounted Degrees of Freedom

- Every time you test a regressor term for the model, consider it an added degree of freedom
  - When testing your final model, there is no clue about how many different models you went through to find this one!
  - As the total number of degrees of freedom (counted + uncounted) approaches the number of data points, the model becomes potentially useless (fitting noise)
  - Sometimes called "trolling for effects"

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Model Validation

- Just because a model can fit a data set does not mean it is good at making predictions
- Model validation: estimating how good the model will be at making predictions
- Three validation approaches
  - Model coherence – how well does this model fit within the framework of existing knowledge?
  - Compare model to new data
  - Data splitting: randomly split your data in two, use first group to calibrate (training sample), second group to validate (hold-out sample, test sample)

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Comparing Model to New Data

- Regress model to new data – how have the coefficients changed?
  - Use two-sample pooled t-test for each coefficient
- Compare existing model's predictions to new data
  - Mean squared predicted residuals (MSPR)
  - If MSPR is not too much greater than the original regression mean square error (MSE), the model is "valid" – we typically look for less than a factor of 2

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Data Splitting

- Typical data splitting uses about 50-80% of the data for training, 20-50% for validation
  - The data must be split randomly
  - E.g., break the data into five parts, then use five different 80/20 splits to train and test
    - Called 5-fold cross validation
    - Generalization: k-fold cross validation
    - Repeat many times with different random splits
  - Optimism principle: The validation data set will always have worse MSPR than the training MSE

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Lecture 62: What have we learned?

- What are the three automated model search approaches?
- Explain omitted variable bias
- How can model building result in uncounted degrees of freedom?
- What are the three model validation approaches?