Statistics 622
Advanced Quantitative Modeling

Professor Robert Stine
444 Huntsman Hall
stine@Wharton.upenn.edu

Overview

This half-semester course extends regression modeling beyond the scope of Statistics 621. These extensions include methods for

- *visualizing* regression models,
- *building* models from large data sets,
- *modeling* categorical data for classification, and
- *capturing* richer structures using novel methods from data mining (trees and nets).

The course emphasizes the practical use of methodologies that are often associated with data mining and machine learning. The treatment of these methods in Statistics 622 adds both breadth (e.g., logistic regression) and depth (e.g., model selection) to the topics covered in Statistics 621.

The course begins with regression, with an emphasis on the use of regression to find optimal predictions for business forecasting. This topic allows review of the foundations of regression, but the emphasis lies in new areas such as the trade-off between over-predicting versus under-predicting. Better minimization of costs motivates finding more predictive models. For example, the bank that predicts who will default on a loan more accurately than rivals reaps higher profits. A company that is able to identify those most interested in its products can launch a new product with a finely tuned campaign.

The pursuit of more accurate predictions for these applications leads to methods for finding predictors concealed in large databases. What's large? Research methods handle more than a million predictors. We won't go quite so far, but we'll work with large datasets having 1000's of predictors. The trick is to avoid over-fitting, building a model that fits the data at hand but predicts new data poorly. (For an example of a large problem, see Foster and Stine, 2004, "Variable selection in data mining: Building a predictive model for bankruptcy," available from my web page at www-stat.wharton.upenn.edu/~stine).

The course next turns to methodologies that extend the regression model. Some of these methodologies enrich the form of the predictive model. We'll use these extensions as diagnostics to see whether the regression model has missed useful structure. For example, neural networks blend several regression models into a complex equation that can capture features that ordinary regression misses.

Generalized linear models, in particular logistic regression, extend regression modeling to the analysis of categorical responses. These models are commonly used to develop models of choice, as in identifying which factors affect the retail purchase decision. Classification trees (CART) models present a more radical departure from regression. These models organize the data into subsets that behave similarly.

Audience and Prerequisites

This course presumes that students are familiar with inferential statistics as covered in Statistics 603 (including hypothesis tests, confidence intervals, p-values) and Statistics 621 (use and interpretation of least squares regression). The course also presumes familiarity with most recent version of the statistics package JMP, which is used in Statistics 603 and 621.

Materials

Readings will be assigned from the casebook

> *Business Analysis Using Regression* (abbreviated BAR in outline)
> (Foster, Stine, and Waterman, Springer, 1998),

supplemented by

> Lecture notes and related papers available via WebCafe.

The software used in the course is

> JMP-IN, version 5.1. Readings from the accompanying text *JMP Start Statistics*, Third Edition (Sall, Creighton, and Lehman, Thomson, 2004) are an important supplement to the class notes.

Familiarity with this software is *fundamental*. Assignments require students to analyze data using this package. The software/manual package is available from the bookstore and on the Wharton network.

Office hours and TA

The class has a TA, and I will post office hours as soon as possible. My office hours are Tuesday and Thursday from 4:30-5:30 and also on Wednesday from 4–5 pm. Otherwise contact me by e-mail or post your question in Web Café.

Requirements

Total grade = 45% assignments + 55% final

Assignments

These consist of small, weekly data analyses tasks intended to reinforce classroom discussion and develop familiarity with the methods. The exercises are similar to those used in Statistics 621.

Final Exam

Closed-book, short answer and multiple-choice format. The exam occurs in the final class period. Sample final exams from prior years – with solutions – will be posted prior to the exam.

For the final, you are allowed one 8.5 × 11 sheet for any notes that you want to bring. (You probably will not need them, but the process of figuring out what to bring is a useful way to prepare for the final.)

Course Outline and Readings

Class 1. The Cost of Uncertainty **Tuesday, 1/11**

Two key uses of *regression* in business are resource allocation and demand prediction. Statistics 621 covers resource allocation and the foundations of prediction. We'll focus on the use of predictions.

In particular, once one has a model for demand, how should the prediction be used? The answer depends on the cost of errors. Over-predict demand and you end up with *visible* storage costs and depreciation. Under-predict demand and you miss *hidden* opportunities to sell. The successful use of predictions from regression depends on knowing costs of these errors. At last, you'll see how summaries like RMSE turn into dollars.

BAR, pages 12-22, Class 3 (inference in regression).

Class 2. Getting better predictions **Thursday, 1/13**

An analysis of costs shows that models with more accurate predictions lower costs. The addition of more predictors to a regression offers the chance to understand the underlying structure better as well as the opportunity for more accurate predictions. More accurate predictions in turn offer the opportunity for higher profits.

The trick to building a *multiple regression*, however, is making sure that the added predictors really do give more accurate predictions. After all, more complicated models generally fit the observed data better than simpler models. Often more complicated models predict new data poorly, having been over-customized to the data in hand. Before complicating a model, we need to be sure the added features help.

BAR, Class 4, particularly 105-108 and the following car example.
Review software for multiple regression, pages 307-325 JMP-IN.

Class 3. Calibration **Tuesday, 1/18**

Making decisions from regression requires your model to be calibrated. If you predict demand to be $100 million next month, you need to be comfortable that you are right on average. That's what *calibration* is all about: being right on average. It's not the stiffest test, but if you use a model that is not calibrated, you've left money on the table.

BAR, particularly pages 12-22.
JMP, Chapter 10 (particularly splines and polynomials).

Class 4. Finding hidden structure **Thursday, 1/20**

Models that fail to account for lurking factors lead to expensive mistakes. Automatic methods that patiently search for useful predictors can help – if you have the data and know how to use them. *Stepwise regression* is the classic method for doing just this. When used carefully, it can find features, particularly combinations known as interactions that are hard to identify otherwise but offer dramatic improvements in model accuracy.

BAR, Class 8, particularly pages 199-201.
JMP, pages 325-329.

Class 5. Visualizing models **Tuesday, 1/25**

Regression becomes more useful once you learn how to use interactions to find and describe complex patterns in data. These can be hard to understand, but recent visualization tools make it possible to see things clearly. We'll use *interactive surface plots* to understand how interactions contribute to a model.

BAR pages 39-52, 105-132

Class 6. Avoiding over-confidence **Thurday, 1/27**

Automated modeling tools like stepwise regression are greedy. They resemble the "EverReady bunny": the model grows bigger and bigger and the process never stops. Unfortunately, at the same time, its predictions get worse and worse. This phenomenon is known as overfitting. The solution is figuring out how to unplug stepwise before it goes too far.

Saving some data for later can be pretty useful. But, do you really have enough to save just to see how well your model is doing? Often, setting aside data is the only way to know. That's the essence of *cross-validation*: build a model on one part of the data, and test it on another. Cross-validation is sometimes too expensive in its need for data, but the *Bonferroni* approach is thrifty.

BAR pages 199-201, 220-227
JMP pages 325-329

Class 7. Diagnostics: Have we missed something? **Tuesday, 2/1**

Regression models with interactions are flexible, but are limited when compared to the scope of *neural networks*. The problem is to find (or "train") a network to capture real patterns, not noise. We'll use these as a diagnostic to search for structure that our regression model has missed.

BAR pages 62-72
JMP: See on-line Statistics Manual for discussion of neural networks.

Class 8. From regression to decisions **Thursday, 2/3**

What do you do when the actions that you can take are discrete, but the model that you have built predicts a range of values? How should you use the predicted value to choose your action?

Class 9. Models for classification **Tuesday, 2/8**

Sometimes your data is not up to regression. The response could be an action (buy/not buy) rather than an amount. When the response that you want to model is categorical, you're in the domain of *logistic regression*. We start with models having one predictor. This setting gives us a chance to figure out how to fit a logistic regression. No more least squares.

BAR pages 273-298 BAR
JMP-IN 289-294, 529-531, 535-543

Class 10. Improving classification **Thursday, 2/10**

Logistic regression can use more than one predictor, just like ordinary regression. But once we allow several predictors, how are we supposed to decide which features belong in a *multiple logistic regression*? And once we have identified them, what do they mean?

BAR pages 273-298
JMP pages 289-294, 529-531, 535-543

Class 11. Tree models for classification **Tuesday, 2/15**

Regression predicts the response with an equation. A neural net combines several regression equations. Trees are, well, trees! Decision trees, that is. How do you pick the "best" tree? We start at the beginning, comparing trees to models that use equations with just one predictor.[1]

Class 12. Growing trees **Thursday, 2/17**

As we gain experience, we'll move on to trees that pick predictive features from a wider assortment as well as consider trees that compete directly with regression.

JMP pages 445-457.

Class 13. Review and comparisons **Tuesday, 2/22**

Which method is the right one to use for a new problem that confronts you? Now you have enough tools to be able to use something other than regression, if you can figure out which one to use.

Class 14. In-class final **Thursday, 2/24**

[1] If you like trees and want more details, then have a look at *Classification and Regression Trees* by Breiman, Friedman, and Olshen. It's a lucid classic. It's not required, but you can find it at Amazon and elsewhere.