46-883 Section M3 (6 units)

Statistical Foundations of Business Analytics

Spring 2022

Carnegie Mellon University
Tepper School of Business

Professor Fallaw Sowell fs0v@andrew.cmu.edu

INTRODUCTION AND OVERVIEW

This course bridges traditional statistics and machine learning. The topics are in most statistics courses. However, our emphasis will be on taking data and making a decision versus theoretical statistical results.

The intuition and understanding of different statistical analyses require some theoretical development. This development provides tools and intuition to be a foundation to address future challenges.

Relative to a traditional statistics course, this course shows how statistical techniques and approaches address current machine learning problems. For example, we will address the method of maximum likelihood estimation. In addition to the theoretical issues of inference on the estimates, we will consider the practical problems of obtaining the estimates. How do we get the computer to evaluate a likelihood function and find the parameter estimates? This application demonstrates a general approach used to address most machine learning problems.

When you finish this course, given a data set, you should be able to select an appropriate R library and present a summary of the statistical results (parameter estimates, forecasts, and goodness of fit plots).

TEXTBOOKS

I believe online and/or pdf versions of these are available from the CMU library.

An Introduction to Statistical Learning: with Applications in R 2nd ed. 2021 edition (July 30, 2021) (ISBN: 978-1071614174) by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani published by Springer (Springer Texts in Statistics).

R for Everyone second edition (2017) (ISBN: 978-0-13-454692-6) by Jared P. Lander published by Pearson Addison-Wesley.

PROBLEM SETS

There will be three problem sets. These will be applied problems in R. You will take a data set, apply the appropriate R commands, and report the useful results. Your answers should include both the HTML file created when you knit your R markdown file and the R markdown file.

FINAL EXAM

The exam consists of applied problems. For each question, you will have a data set. You need to investigate the data, write appropriate R markdown programs to select a suitable model, and then present your results. Your answers should include both the HTML (or pdf) file created when you knit your R markdown file and the R markdown file.

GRADING

The final course grades will be determined by the following method:

Problem Set Average	65%
Cumulative Final Exam Grade	35%.

Topics

Wk 1: Linear Regression

Textbook Readings:

R for Everyone, Chapters 19 and 21. See Chapter 28 to learn about R markdown.

An Introduction to Statistical Learning, Chapter 3.

The basic problem, the terminology, and the model are presented. Restrictions on the error term are presented. The three basic questions of empirical work are presented. The least squares estimator is derived. Restrictions needed on the independent variables are noted. The sampling distributions are derived and the resulting confidence intervals for the parameters. Confidence interval can help select which variables should be in the model. Different model selection statistics are presented. Different plots help judge the overall fit of the selected model.

Wk 2: Ridge Estimation and Lasso Estimation

Textbook Readings:

R for Everyone, Chapter 22.

An Introduction to Statistical Learning, sections 5.1, 6.2

This section concerns multicollinearity in the linear regression model. Scatter plots and correlations help develop intuition. The problems associated with collinearity and near collinearity are highlighted with the simple example of solving two lines. Introduce the ridge estimator. Both the constrained optimization and the penalty function derivations are presented. Conditional on lambda, the sampling distribution is presented and used to explain the bias and variance trade-off. The high variance associated with a holdout sample approach to selecting lambda is noted. Cross-validation is presented as the best way to select lambda in the ridge estimation. The general problem of Regularization and tuning parameters is noted. The ability to used cross-validation in more general settings is noted. This section presents an introduction to Bayesian statistics. The linear model is then presented with a normal prior for the coefficients. The parameter value associated with the posterior mode is shown to be equivalent to the ridge estimator. The lasso estimation technique is introduced. The lasso's ability to perform variable selection is noted. The L1 and L2 norms are reviewed, and their relation to the penalty function objective functions that define ridge and lasso are noted. The elastic net estimator is presented as an approach to perhaps get the best of both estimators.

Wk 3: Maximum Likelihood Estimation

Textbook Readings:

The textbooks do not have a full presentation of maximum likelihood. However, when they present linear regression and logistic regression both sets of estimates are maximum likelihood estimators.

R for Everyone, Chapter 8 on writing functions.

An Introduction to Statistical Learning, Section 4.3

The likelihood function is derived from the probability density of a sample. The log-likelihood is a sum of terms and hence is more agreeable to statistical analysis. The asymptotic distribution for the maximum likelihood estimator is presented, and its relation to the information matrix is noted. The numerical optimization of multivariate functions is presented. The local gradient methods are presented as a generalization of Newton's Method. The R optim() function is introduced. The log-likelihood function for a sample from a normal distribution is presented. This is the foundation for linear regression. The log-likelihood function for a sample from a Bernoulli distribution is presented. This is the foundation for the logistic regression model. Replacing parameters in the

likelihood functions with linear combinations of covariates leads to the linear models in the lm() function of R. The derivatives needed for the linear models can be obtained by the chain rule.

Wk 4: Classification

Textbook Readings:

R for Everyone, Chapter 20

An Introduction to Statistical Learning, Chapter 4

Examples are given to make the distinction between regression and classification models. The discriminant function is defined. The classification problem can be viewed as the estimation of functions (boundaries) in the covariates to separate the different classes. Assume that the covariates are normally distributed to obtain quadratic discriminant analysis. When the covariances are the same this reduces to linear discriminant analysis. When there are two classes, the standard notation is, $y_i \in \{0, 1\}$. This suggests Bernoulli random variables with the probability of success equal to p. Covariates are introduced by modeling the probability p with the logistic function. Estimation is performed by maximum likelihood. The logistic regression model is extended to more than two cases. The need to have the probabilities sum to one is explained.

Wk 5: Trees, Bagging, Boosting and Random Forests

Textbook Readings:

R for Everyone, sections 23.4 - 23.7

An Introduction to Statistical Learning, Chapter 8

Trees are presented as a nonparametric approach to fitting a function between the covariates and the dependent variable. To avoid overfitting, regularization is performed by pruning the tree. Ensemble methods combine different models to produce the final model. Boosting creates simple models each directed at learning one area of the sample. The final model is a weighted average of the simple models. This results in a complex model with reduced bias. Bagging creates a large number of complicated models and then averages to create a final model. The complicated models are each unbiased and hence so is their average. The average has lower variance. Boosting and Bagging are general procedures that can apply to any model. We apply them to trees. The higher model variance associated with correlated trees is noted. Randomly selecting the covariates that can be included in the trees reduces the correlation and hence the model variance. This type of model is called a Random Forest.

Wk 6: Design of Experiments and A/B testing

There are no readings in the two textbooks on this topic.

The need for structural models to address control questions is presented. Models that are appropriate for prediction are contrasted with models that allow for causality and hence control. The Ice Cream Sales example is introduced. Building on the Ice Cream Sales example, the three factors with two levels Full Factorial design model is introduced. Main effects and interaction effects are presented. Fractional Factorial design models are introduced and the aliasing problem is demonstrated. With the Ice Cream Sales model as an example, the general field of Design of Experiments is reviewed. The basic terminology is presented and Randomization and Blocking are described. The Law of Large Numbers and the Central Limit theorem are presented. The CLT is used as the probability foundation of hypothesis testing about the mean of a random variable. Building on the idea of Randomization from the Design of Experiments and the CLT applied to Bernoulli random variables, A/B testing to improve management is presented as a hypothesis testing problem. The Size, Significance and Power requirements are used to establish the appropriate sample size and cut-off value. Common mistakes in A/B testing are noted.

Wk 7: Causality and Propensity Score Modeling

There are no readings in the two textbooks on this topic.

The distinction between controlled experiments and observational studies is noted. Examples of observational studies are introduced: the impact of Catholic School attendance on test scores and the impact of a new medical treatment on heart attack survivability. We note how confounding factors create bias in the estimated treatment effects. Covariates and randomization can help address this bias. Given an observational study where treatment effect are of interest, how might we proceed? The idea is to use observed covariates to estimate the probability of being treated. This probability is called the propensity score. The propensity score could be estimated with a logistic regression model. Individuals in the treatment group and the control group are matched if their propensity scores are close. If the covariates look to be balanced in the matched sample, the treatment effects can be estimated with a linear regression on the matched sample.

TAKE CARE OF YOURSELF

Do your best to maintain a healthy lifestyle this mini by eating well, exercising, avoiding drugs and alcohol, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress.

All of us benefit from support during times of struggle. You are not alone. There are many helpful resources available on campus and an important part of the university experience is learning how to ask for help. Asking for support sooner rather than later is often helpful.

If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit their website at http://www.cmu.edu/counseling/. Consider reaching out to a friend, faculty or family member you trust for help getting connected to the support that can help.