Report of the Applied Statistics Curriculum Revision Committee

Updated July, 2020

This report has been prepared by a committee consisting of Professors Chuanshu Ji, Kai Zhang, Richard Smith (chair), Steve Marron and Yufeng Liu. Our main focus has been on STOR 664/665, though we also discussed other courses briefly, in particular the two machine learning courses, 565 (for MS students) and 767 (for PhD students).

Executive Summary:

STOR 664 and 665 should remain primarily "Linear Models" courses, with 664 focusing on normal-theory linear regression and ANOVA, and 665 on generalized linear models and random effects. Both courses should be streamlined to avoid overlap, as far as possible, though the choice of topics should also reflect that 664 is widely taken by MS and by OR PhD students, whereas 665 is primarily for PhD students in Statistics. Both courses should stress computation, primarily in R, and we recommend that they should include a "student project" component, where students would work independently on real-data projects with a written report due at end of semester. Consideration should also be given to changing the format of the CWE exam for these courses.

STOR 664:

This course has traditionally included a thorough treatment of the normal-theory linear model, including derivation of the least-squares equations, confidence intervals and hypothesis testing, variable selection, and diagnostics. Theoretical derivations (e.g. rigorous proofs of χ^2 and *F* distributions for some of the test statistics) may be of less interest to MS and OR students but we feel it should be retained in the course as it does not take up much time and makes this whole section of the course more complete. Some of the topics currently taught in the course feel out of date and there are others (e.g. LASSO) which have not been included though they are clearly relevant for modern applications.

We should strengthen the "analysis of variance" component by including other models beyond the simple one-way and two-way layouts. Examples of other models that could be included are latin squares, factorial designs and balanced incomplete block designs. These are useful techniques to know about for many statistical applications, and some OR (e.g. design of simulation experiments) – we do not proposal a formal "theory of design of experiments" but some basic design principles could be included.

Specific suggestions for course content:

Reduce time spent on diagnostics, e.g. separate course topics on DFFITS, DFBETAS, COVRATIO and the like are not very useful, but we could still spend time on the "influence diagram" diagnostic plot that is part of the standard R output.

Model selection techniques such as forward, backward and stepwise selection should still be included in the course, but the discussion could be shortened as these techniques are now readily available in R (the "step" function) so there is no need to discuss their detailed mechanics. LASSO is not currently taught in this course but it is a widely used variable selection technique and should be taught alongside the older methods. One can view LASSO as an alternative approach to ridge regression (which is currently taught) and the two can be combined in the elastic net method. This could possibly replace the methods of principal components regression and partial least squares regression that are currently taught in the course but arguably have been superseded by LASSO and related techniques (also, PCR and PLS are taught in STOR 565).

Computation should be primarily in R. In the future, we may consider extending this to also cover Python, but there would be significant overhead in trying to cover Python as well and it is better to spend the time on statistical rather than computing topics.

Other topics that could also be included as time permits: transformations, weighted and generalized least squares, nonlinear regression, random forests. There would be some merit in including a section on nonparametric regression (e.g. splines) but our overall feeling is that this would be hard to cover well in the time available and it is part of STOR 767.

We should consider including a "project" element where students would work individually on datasets of their own choosing, write a report and (possibly) make an in-class presentation. However, the more formal homework and midterm/final exam requirements are valuable preparation for the PhD qualifying exams so they should still be included.

STOR 665:

In recent years this course has started with a "review of linear models" but we recommend removing this component (essential theory should be included in 664) to get more quickly into new material.

The classical text on GLMs is *Generalized Linear Models* by McCullagh and Nelder (second edition, 1989 (https://www.routledge.com/Generalized-Linear-Models/McCullagh-Nelder/p/book/9780412317606) though several students also mentioned Agresti's book *Foundations of Linear and Generalized Linear Models* (Wiley, 2015, read online free through UNC libraries) as a valuable source for the basic theory and possibly a better course text than McCullagh and Nelder.

For the future, we recommend that the main focus of 665 should be on the following topics:

- Generalized Linear Models (including the quasi-likelihood fitting technique; practical aspects such as variable selection and diagnostics; and specific examples such as the binomial, Poisson ad multinomial distributions);
- Mixed effects (also known as random effects) models;
- Generalized mixed linear models (in effect, the intersection of the previous two techniques).

All three topics are covered in Faraway's *Extending the Linear Model with R* (Second Edition, 2016, https://www.routledge.com/Extending-the-Linear-Model-with-R-Generalized-Linear-Mixed-Effects-and/Faraway/p/book/9781498720960) though this doesn't go very far into the theory so we should possibly consider alternative texts for that.

Other topics that might be covered (time permitting):

- Some of the more applied multivariate analysis techniques, specifically principal components, factor analysis and some of the simpler methods of cluster analysis (note: PCA and cluster analysis are covered in both 565 and 767);
- Bayesian statistics, including the basic ideas of hierarchical models and MCMC.

At the moment, multivariate analysis is part of the "Time Series and Multivariate Analysis course" which was taught, most recently, in Spring 2019, but it is not clear whether there is any plan to continue this course (it is not on the schedule for 2020-21). Likewise, Bayesian statistics is also a separate course (Prof. Jan Hannig taught STOR 757 in Spring 2020) but we should clarify whether it is intended to teach that course on a regular basis and if so what the intended emphasis will be (we understand from Prof. Hannig that his course covered more the philosophical and decision theoretic aspects of Bayesian statistics than practical data analysis). One concern about Bayesian statistics is that if we are going to cover it at all, we should do so in enough detail that the students get a proper flavor of what the subject is really about – that seems unlikely given the number of topics currently proposed for 665.

CWE Exam

It has been pointed out that the equivalent exam in Biostatistics is a take-home exam and the suggestion has been made that we should move to the same system (for this sequence: presumably the theoretical statistics and probability sequences would still have closed-book exams). The Biostatistics department has had some honor code issues over the years but they are not regarded as a major issue. As a side note, I (RLS) have used take-home exams both in 664 in the past and for 556 in Springs 2019 and 2020; I have not so far had any occasion to accuse a student of an honor code violation.

Other courses in the Applied Statistics Curriculum

STOR 565 is a widely popular undergraduate Machine Learning course that is also recommended for our MS students. In addition, the department has decided that STOR 767 (the PhD-level Machine Learning course) will be taught every year in the future.

In view of these courses, our view is that there is no need to teach machine learning topics as a regular part of 664/665, but ML techniques that fit naturally into 664/665 could be included there. Examples are LASSO (as a variable selection technique in linear regression) and PCA; possibly random forest regression as part of 664, though this has not been included in the past and arguably is less important as a topic for 664.

The status of 754/755/756/757 should be clarified. 754 is "Time Series and Multivariate Analysis." If the above suggestions are adopted, the applied part of the multivariate analysis curriculum could become part of 665, but there is no space to include time series as well; this needs to be discussed. Similarly, 757 is officially "Bayesian Statistics and Generalized Linear Models" but there seems no need for a separate more advanced course on GLMs so it would be helpful to clarify both whether this course will be given in future and if so what the content is likely to be.

The consulting course STOR 765 is viewed as a vital part of the department's curriculum and we are not proposing any change to that course.