

BLOST/STAT 570  
AUT 05  
Thomas Lumley

## Syllabus

1. Review of least squares regression: range of justifying assumptions. How to relate parameters and contrasts.
2. Where does randomness come from and what does it mean?
3. What models are for: prediction, description, inference about populations, inference about processes.
4. Describing uncertainty: relative to other plausible data (frequentist) or other plausible parameters (likelihood, especially Bayesian). Confidence, Conditionality, Sufficiency, and Likelihood principles.
5. Weighted least squares: precision weights and sampling weights. Robust standard errors. The jackknife and the bootstrap.
6. Causation, via counterfactuals and principal stratification. Holland causes. Confounding vs interaction. Causal graphs.
7. Stratification, adjustment, propensity scores, direct standardization. Model selection for prediction and for inference.
8. Generalized linear models: via exponential families, linear estimating functions, iteratively weighted least squares.
9. Logistic regression. Poisson regression.
10. Differences from linear regression: collapsibility vs confounding, model criticism is harder, link and variance can be chosen separately.
11. Retrospective sampling of binary data: logistic and conditional logistic regression and comparison to probability weights.
12. Survival models: exponential, parametric accelerated failure models, Cox model.
13.  $L_1$  regression and other resistant linear estimators

### In parallel

1. EDA: tables, scatterplots, scatterplot smoothers, coplots.
2. Identifying a substantive question and choosing a model.
3. Structure of data analysis reports and scientific papers

### Extra notes

1. How not to implement least squares estimates.

2. The Li–Duan theorem
3. A little estimating function theory.
4. Implementing regression methods: best practices.

*Statistical Models* by Anthony Davison is the recommended text. We will cover most of chapters 7–10 and parts of 1–5 and 11. Chapters 2–4, 7, and perhaps 12 will also be useful reference for the 580s sequence. However, the text gives more emphasis to likelihood than will I.

### Class project

As this is a statistical methods class the project involves developing, implementing, and applying statistical methodology. Much of the technical detail will appear in weekly assignments. You will then write up a report describing the method and using it to analyse some data that I will supply.

The statistical problem is logistic regression when the binary outcome variable is measured with error. Many medical diagnostic tests give some false positives, some false negatives, or both. You will examine the bias these errors produce in the logistic regression coefficients and in the area under the ROC curve, and will implement a modified version of logistic regression that fixes the bias.

The report will be due on Wednesday November 23, and worth 25% of the grade for the course. You will then review someone else's report. Based on comments from me and/or TAs and from the other reviewer you will have an opportunity to revise the written report and resubmit it at the end of the quarter. The review and the resubmission will each be worth 5% of your grade.

### Other assessment

The final exam will also be worth 35%, the weekly assignments 25%, and participation in assigned class discussions the remaining 5%. Note that some assignment questions may be repeated later in the course — being able to reconstruct your work is always a good habit.