Biost 518 / Biost 515
Applied Biostatistics II / Biostatistics II

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Lecture 13:
Prediction
(with acknowledgements to Thomas Lumley)

March 13, 2014

1

Lecture Outline

- General Setting
- Prediction of Summary Measures
  - Necessary Assumptions for Inference
  - Special cases
    - Means, Geometric Means, Odds, Probabilities, Rates, Hazard Ratios, Survival probabilities
- Prediction of Individual Observations
  - Necessary Assumptions for Inferences
  - Special cases
    - Continuous measurements, binary measurements

Setting for Predictions

3

General Classification

- Clustering of observations
- Clustering of variables
- Quantification of distributions
- Comparing distributions
- Prediction of individual observations

5. Prediction

- Focus is on individual measurements
- Point prediction:
  - Best single estimate for the measurement that would be obtained on a future individual
    - Continuous measurements
    - Binary measurements (discrimination)
- Interval prediction:
  - Range of measurements that might reasonably be observed for a future individual

Regression Based Inference

- Estimation of summary measures
  - Point, interval estimates within groups
  - Tests hypotheses about absolute measurements
- Inference about associations
  - First order trends in summary measures across groups
    - Point, interval estimates of contrasts across groups
    - Tests hypotheses about relative measurements
- Inference about individual predictions
  - Point, interval estimates

Optimality Criteria

7

Prediction and Classification

- Training sample of covariates X and outcome Y used to develop a model
- The model is used on observations where X is known and Y is not, to estimate Y
- 'Prediction' is the general term
  - sometimes 'prediction' means specifically that Y will occur in the future
- 'Classification' or 'discrimination' is used for binary outcomes

8

Scientific and Statistical Question

- What is the best estimate of the outcome for this new person?
  - point estimation of a summary, point prediction
- What is the uncertainty in the best estimate?
  - confidence interval around the summary
- What is the uncertainty in the outcome?
  - prediction interval for new observation.

9

Goals for a Prediction Model

- Accurate prediction
  - the predicted value should be as close as possible to the new outcome
- Honest estimate of prediction error
  - we need to know how good the prediction is
- Cost of variables
  - if possible, we don't want to measure too many difficult or expensive things to compute the prediction

10

More Controversial

- Face validity
  - for people to use a prediction model it helps if it makes sense to them (more true for physicians than financial analysts)
- Causal grounding
  - Even if we don't care why the model predicts well, a model that predicts well for good reasons is likely to extrapolate better to new settings.
- Usefulness of information
  - what will be done with the prediction model that wouldn't be done just as well without it?

11

Prediction Accuracy

- In order to choose the most accurate prediction, need a way to measure prediction accuracy, a loss function
- For continuous variables, we might use
  - squared error: $E[(\text{outcome} - \text{prediction})^2]$
  - absolute error: $E[|\text{outcome} - \text{prediction}|]$
  - the expected values are averages over the possible covariate values at which we are prediction and the distribution of outcomes at those covariate values

12

Loss Functions: Continuous

- Minimizing squared error implies the best possible prediction is the mean of the outcome at the new covariate values
- Minimizing absolute error implies the best possible prediction is the median of the outcome at the new covariate values
- We are familiar with regression models for the mean, so squared error loss is convenient.
  - note: using a transformation of outcome implies minimizing squared error loss on the transformed scale
- We sometimes "penalize" the loss function by
  - The number of covariates included, or
  - The magnitude of the regression parameters (shrinkage)
    - "LASSO"

13

Loss Functions: Binary

- For a binary outcome there are only two errors
  - predict 1 when outcome is 0
  - predict 0 when outcome is 1
- We can assign an appropriate cost to each one

14

Honest Estimates of Prediction Error

- "Prediction is hard, especially about the future"
  (variously and unreliably attributed)
- Choosing a prediction model will often involve considering many possible models
- Estimating prediction error on the same sample used for model selection will give an over-optimistic estimate.
- In most situations when model selection is done the bias is unacceptably large

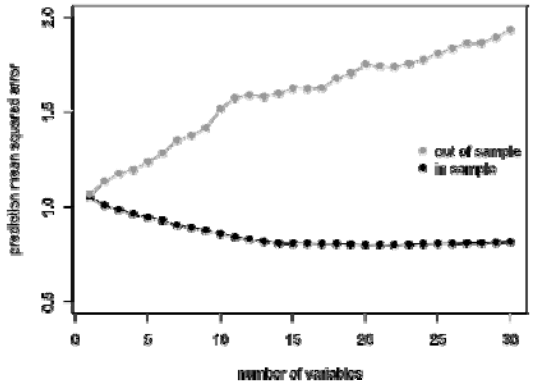
15

Simulated Example

- 100 observations of 50 independent Normal(0,1) predictors and a Normal(0,1) outcome
  - no predictors have any relationship to outcome
  - adding variables will improve in-sample prediction, worsen out of sample prediction
- Model chosen by minimizing AIC, a popular criterion designed for prediction (corresponds roughly to $p<0.15$)
  - in-sample prediction error 0.85
  - out of sample prediction error 1.57

16

Simulated Example



17

Example: GWAS Disclosure

- Genome-wide association studies estimate the association between an outcome variable and hundreds of thousands of genetic predictors taken one at a time
- Prediction in new samples is usually very poor – an R^2 of 0.05 would be regarded as good.
- Because of the very large number of predictors, prediction in the original sample is nearly perfect

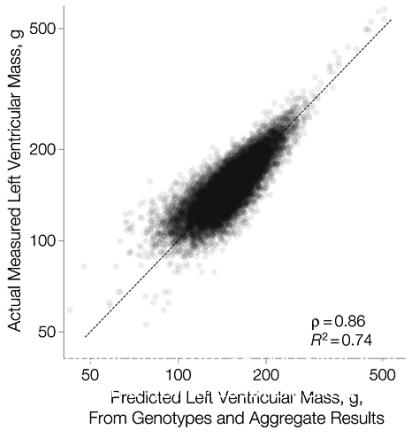
18

Example: GWAS Disclosure

- Since prediction in the original sample is nearly perfect
  - someone who can obtain a complete or partial genotype for a study participant, and the corresponding association estimates can estimate their previously observed outcome accurately
  - publishing all the association estimates leaks information about individual participant outcome values
- [PLoS Genetics October 2009; JAMA commentary Feb 17,2010]

19

GWAS Predictions



20

Out-of-sample Error

- True estimates of prediction error require independent data
- Can fake this by sample splitting
  - use part of the data to choose the model, part to estimate the error [more later]
- Sample splitting captures the model-selection component of prediction error
  - does not capture error in generalizing to new population
  - distributions and associations in genuine new data will be slightly different

21

Cost of Variables

- A prediction model is only useful if the benefit of the information is greater than the cost of using the model
  - monetary cost of obtaining variables
  - risk or discomfort from measuring variables, eg biopsy, radiation dose from x-ray imaging.
- Ideally use a small number of variables that would already be available for other reasons.

22

Example: Framingham Risk Score

- Predicts 10-year risk of coronary heart disease, uses age, sex, blood pressure, smoking, HDL and total cholesterol
  - age, sex, smoking, blood pressure are measured for everyone already
  - cholesterol would probably be measured for people whose CHD risk is being estimated.
  - using total and HDL cholesterol rather than LDL cholesterol means fasting before the blood sample is not needed
  - Carotid artery ultrasound gives slightly more accurate predictions, but is not routinely available

23

Example: Mayo Model for PBC

- Predictive model for time to death in the rare liver disease primary biliary cirrhosis
- Disease stage measured by liver biopsy is strongly predictive, but biopsy is unpleasant and carries some risk
- One goal of the model was to obtain good prediction from blood sample and clinical examination, and not require liver biopsy

24

Face Validity

- Willingness to use a predictive model can depend on whether the model looks plausible.
- If there are many models with equally good prediction (often true), picking one that looks plausible can be helpful it getting it accepted.

25

Causal Grounding

- For pure prediction, it doesn't matter whether the predictors cause the differences in outcome as long as the prediction is accurate
  - C-reactive protein levels in the blood predict heart attack, quite likely just a symptom of atherosclerosis
  - Good credit ratings predict low risk of car accidents, are used by insurance companies, but do not have a direct effect

26

Causal Grounding

- If an association between predictor and outcome is not due to a stable causal mechanism, it is more likely to change in future data
  - recession lowers many people's credit scores, does not increase car crashes.
  - treatments could affect C-reactive protein without affecting risk of heart attack.

27

Usefulness of Information

- Screening
  - screening is done on the general population and the result is that some of them are diagnosed as sick or at risk
  - "screening takes healthy people and makes them sick"
  - screening is useful only if something can usefully be done with the result
  - the cost of making the prediction and the cost of a false positive result are important, especially if there are very few true positives.

28

Example: Mammography

- Mammograms clearly reduce breast cancer mortality in women over 50 (community randomized trials)
- Less clear in younger women
  - outcome is much rarer, so more false positives and fewer true positives
  - accuracy of test is lower
  - tumors may be more likely to have metastasized before detection
- US Preventive Services Taskforce changed its recommendation in recent years (controversially).

29

Usefulness of Information

- Diagnosis, prognosis
  - people are self-selected because they have a complaint, so more likely to have disease, less risk of making healthy people sick
  - predictive model may be useful because it affects treatment
  - predictive model may be useful to give information about likely future, even if it can't be modified
  - may also be useful just in explanation

30

Example: Mayo Clinic PBC Model

- Mayo model for primary biliary cirrhosis is used in the scheduling of liver transplants
  - affects treatment
  - doesn't predict survival, because availability of liver transplant is a big change from when the model was developed.

31

Example: Factor V Leiden

- Factor V Leiden is a genetic variant that leads to higher risk of blood clots, especially in leg veins
  - One of the most common genetic tests in adults
- Does not predict prognosis or affect treatment in people who have had a clot
- Predicts future risk but does not affect treatment in relatives of people who have had a clot
- Main motivation appears to be explanation of why the clot happened

32

Automated fitting of predictive models

33

Fitting predictive models

- Given unlimited amounts of data:
  - Step 1: fit a very large number of models to some of the data
  - Step 2: evaluate the out-of-sample prediction error of each fitted model on new data and choose the best one
  - Step 3: evaluate the out-of-sample prediction error of the best model on another set of new data, to get an honest estimate.

34

In practice

- We don't have infinite amounts of data or computing
- Need to fake having independent data by cross-validation
- Need a search strategy for models rather than fitting all of them
- Lots of modern statistical research in this area
  - expert advice is useful if you have to do prediction
  - we will look at one simple but respectable approach

35

Traditional forward selection

- Try all models with a single predictor, pick the one with the smallest p-value (if <0.05)
- Now try all models with that predictor plus one more, and pick the additional predictor with the smallest p-value (if <0.05)
- Repeat until no additional variable has p<0.05
- Stata, like most statistics packages, automates this for you with the stepwise prefix

36

Traditional forward selection

- Doesn't work very well, partly because p<0.05 is probably the wrong threshold
- For a single test, p<0.05 might be too stringent
  - not much loss from having one extra unnecessary variable
- The fitting algorithm does many tests
  - not obvious whether this implies higher or lower p-value threshold is better
- If we had independent data we could run forward selection for a range of thresholds and pick the best one

37

Cross-validation

- Divide the data into 10 parts
- Fit the model to 9 parts and make predictions on the 10th part
- Repeat, leaving each tenth of the data out in turn
- For every observation in the sample, we now have a prediction from independent data and an observed outcome
  - calculate the out-of-sample prediction error

38

Cross-validation

- Cross-validation gives an approximately unbiased (but imprecise) estimate of prediction error
- The number of parts to split into is not critical, but 10 is popular and works reasonably well
  - with large data sets, could use 20 or 50 parts for more precise estimates

39

Using cross-validation to choose p

- Split the data into 10 parts
- For 9/10ths of the data
  - run forward selection with several thresholds (eg p=0.001, 0.005,0.01,0.05, 0.1, 0.15)
  - using the resulting several models, compute predictions for the left-out 1/10 of the data and store them
- Repeat, leaving out each 1/10 of the data in turn
- Compute the out-of-sample prediction error for each p-value threshold

40

Using cross-validation to choose p

- Pick the p-value threshold with the lowest out-of-sample prediction error
- Run forwards selection on the whole data set with that p-value threshold to get a prediction model

41

Cross-validation and forward selection

- The models fitted to each 9/10 of the data may not be the same
  - we're not evaluating the models, just the threshold
- This approach, for different model selection procedures, is part of most modern approaches to predictive model building
  - many methods also average over multiple models or 'shrink' coefficients towards zero, to reduce bias.

42

Cross-validation and forward selection

- There isn't a completely honest estimate of the prediction error of the final model
  - the out-of-sample error from cross-validation for the best threshold is not very biased, because it is only chosen from a small set of alternatives.

43

Simulated example

- Same simulated example: 100 observations of 50 Normal(0,1) predictors, all independent of outcome
- Cross-validation with a range of p-values from 0.5 to 0.005
- 'Best' p-value threshold 0.02
- Resulting model has two predictors
  - in-sample prediction error 1.009
  - cross-validation error estimate 1.16
  - true out-of-sample prediction error 1.13
- Not perfect, but not too bad.

44

What variables to start with?

- Intelligent choice of variables to put into automated model selection will give better results
  - variables that are likely to be related to outcome
  - appropriate transformations of the variables
  - correlation is not a problem
  - multiple versions of the same variable are ok.
- Looking at the data can help choose good transformations, but makes assessment of prediction error less reliable.

45

Predicting a binary variable

- Procedure is essentially the same for binary data
- For logistic regression, use the out-of-sample predictions from cross-validation to estimate the total loss for each p-value threshold
- Choose the p-value threshold that minimizes the this loss, then refit the model with all the data, using this threshold

46

Survival predictions

- In censored data the mean is often not estimable
- Prediction error for a Cox model can't be defined in terms of error from the predicted mean
  - cross-validation to choose p-value threshold is more complicated.
  - automated predictive model fitting is beyond scope of this course, but methods do exist.

47

Summary

- Prediction can be
  - prediction of a summary statistic, with confidence interval
  - point prediction of a best guess
  - interval prediction
- Importance of model "accuracy" depends on the use you are going to make of the predictions (and what you consider optimal)
- If you want unbiased (or consistent) estimates and CI for a particular summary measure
  - regression model for fitted mean must be accurate
  - for interval prediction, assumptions about distribution of outcome must be accuate
- If you want good average performance across a population
  - Interpretation of the regression model is unimportant

48

Summary

- The biases caused by model selection for prediction are serious, but there are ways to avoid them
- Cross-validation is a practical way to get an honest estimate of prediction error
- Ask an expert about modern statistical methods

49