Biost 518 / Biost 515
Applied Biostatistics II / Biostatistics II

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

Lecture 14:

Prediction

(with acknowledgements to Thomas Lumley)

March 10, 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

2

Setting for Predictions

General Classification

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

4

1. Cluster Analysis

- Focus is on identifying similar groups of observations
- Divide a population into subgroups based on patterns of similar measurements
  - Univariate, multivariate
  - Known or unknown number of clusters
- (All variables treated symmetrically: No delineation between outcomes and groups)

5

2. Clustering Variables

- Identifying hidden variables indicating groups that tend to have similar measurements of some outcome
- Interest in some particular outcome measurement
- Predictors that imprecisely measure some abstract quality
- Desire to find patterns in predictors that more precisely reflect the abstract quality

6

3. Quantifying Distributions

- Focus is on distributions of measurements within a population
- Scientific questions about tendencies for specific measurements within a population
  - Point estimates of summary measures
  - Interval estimates of summary measures
    - Quantifying uncertainty
  - Decisions about hypothesized values
- May desire estimates within subgroups
  - E.g., estimates by sex, age, race

7

Example: Estimation of Median

- Statistical Tasks
  - Sample of patients newly diagnosed with stage II breast cancer
    - Follow for survival time (may be censored)
  - Statistical analysis
    - Best estimate of the median survival (K-M?)
    - Quantify uncertainty in that estimate
    - Compare to some clinically important time range (e.g., 10 years)

8

4. Comparing Distributions

- Comparing distributions of measurements across populations
- 4a. Identifying groups that have different distributions of some measurement
- 4b. Quantifying differences in the distribution of some measurement across predefined groups (effects or associations)
- 4c. Quantifying differences in effects across subgroups (interactions or effect modification)

9

4a. Identifying Groups

- Identifying groups that have different distributions of some measurement
- Focus is on some particular outcome measurement
- Identify groups based on other measurements
  - E.g., quantifying distributions within subgroups
  - E.g, stepwise regression models
- (cf: Cluster analysis where all measurements are treated symmetrically)

10

Example: Identifying Groups

- Statistical Tasks
  - Sample subjects to measure risk factors and disease prevalence
    - Cohort study
    - Case-control study
  - Statistical analysis
    - Stepwise model building
      - (Rank most interesting variables by p value?)

11

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

12

Example: Continuous Prediction

- Creatinine clearance
  - Creatinine
    - Breakdown product of creatine
    - Removed by the kidneys by filtration
      - Little secretion, reabsorption
  - Measure of renal function
    - Amount of creatinine cleared by the kidneys in 24 hours

13

Example: Continuous Prediction

- Problem:
  - Need to collect urine output (and blood creatinine) for 24 hours
- Goal:
  - Find blood, urine measures that can be obtained instantly, yet still provide an accurate estimate of a patient's creatinine clearance

14

Example: Continuous Prediction

- Statistical Tasks:
  - Training sample
    - Measure true creatinine clearance
    - Measure sex, age, weight, height, creatinine
  - Statistical analysis
    - Regression model that uses other variables to predict creatinine clearance
    - Quantify accuracy of predictive model
      - (Mean squared error?)

15

Example: Discrimination

- Diagnosis of prostate cancer
  - Use other measurements to predict whether a particular patient might have prostate cancer
    - Demographic: Age, race, (sex)
    - Clinical: Symptoms
    - Biological: Prostate specific antigen (PSA)
  - Goal is a diagnosis for each patient

16

Example: Discrimination

- Statistical Tasks:
  - Training sample
    - "Gold standard" diagnosis
    - Measure age, race, PSA
  - Statistical analysis
    - Regression model that uses other variables to predict prostate cancer diagnosis
    - Quantify accuracy of predictive model
      - ROC curve analysis
        - » Sensitivity vs 1 – Specificity
        - » True Positives vs False Positives

17

Example: Interval Prediction

- Determining normal range for PSA
  - Identify the range of PSA values that would be expected in the 95% most typical healthy males
  - Age, race specific values

18

Example: Interval Prediction

- Statistical Tasks:
  - Training sample
    - Measure age, race, PSA
  - Statistical analysis
    - Regression model that uses other variables to define prediction interval
      - (Mean plus/minus 2 SD?)
      - (Confidence interval for quantiles?)
    - Quantify accuracy of predictive model
      - (Coverage probabilities?)

19

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

20

So far: Inference for Associations

- Necessary assumptions for classical regressions (no robust SE)
  - Independence of response measurements
  - Appropriate within group variance
    - Linear regression: Equal variance across groups
    - Other regressions: Appropriate mean-variance relationship
      - Hence, some dependence on model fit
  - Sufficiently large sample size for asymptotic normal distribution of estimates to be a good approximation

21

So far: Inference for Associations

- Necessary assumptions for first order trends using robust SE
  - Independence of response measurements across identified clusters
    - May have correlated response within identified clusters
  - (Robust SE accounts for heteroscedasticity in large samples)
    - Lack of "model fit" leads to conservative inference due to mixing systematic and random error
  - Sufficiently large sample size for asymptotic normal distribution of estimates to be a good approximation

22

Now: Inference for Predictions

- Additional assumptions for predictions
  - Estimation of summary measures within groups
    - We need to know that our regression model accurately describes the relationship between summary measures across groups
  - Prediction of individual observations
    - We need to know the shape of the distribution within each group

23

Optimality Criteria

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

25
25

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.

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

27
27

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?

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

29
29

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)

30
30

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

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

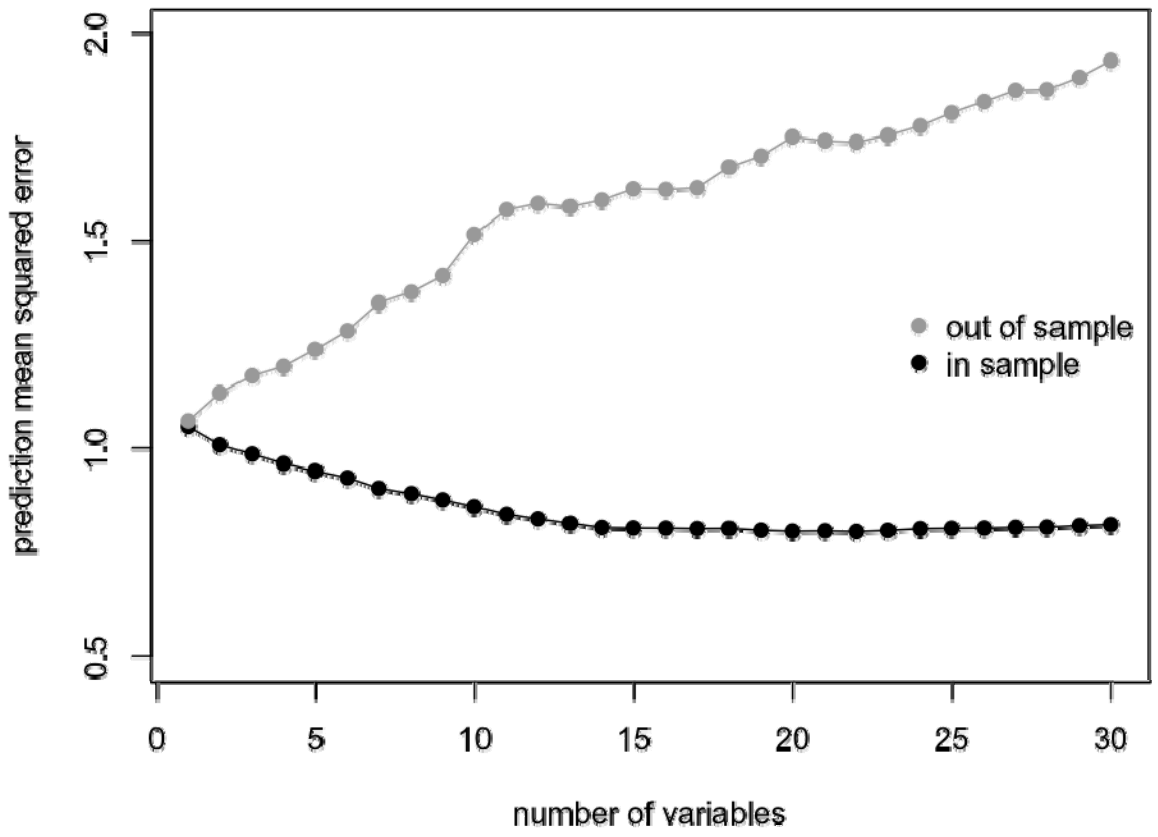
32
32

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

33
33

Simulated Example



34
34

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

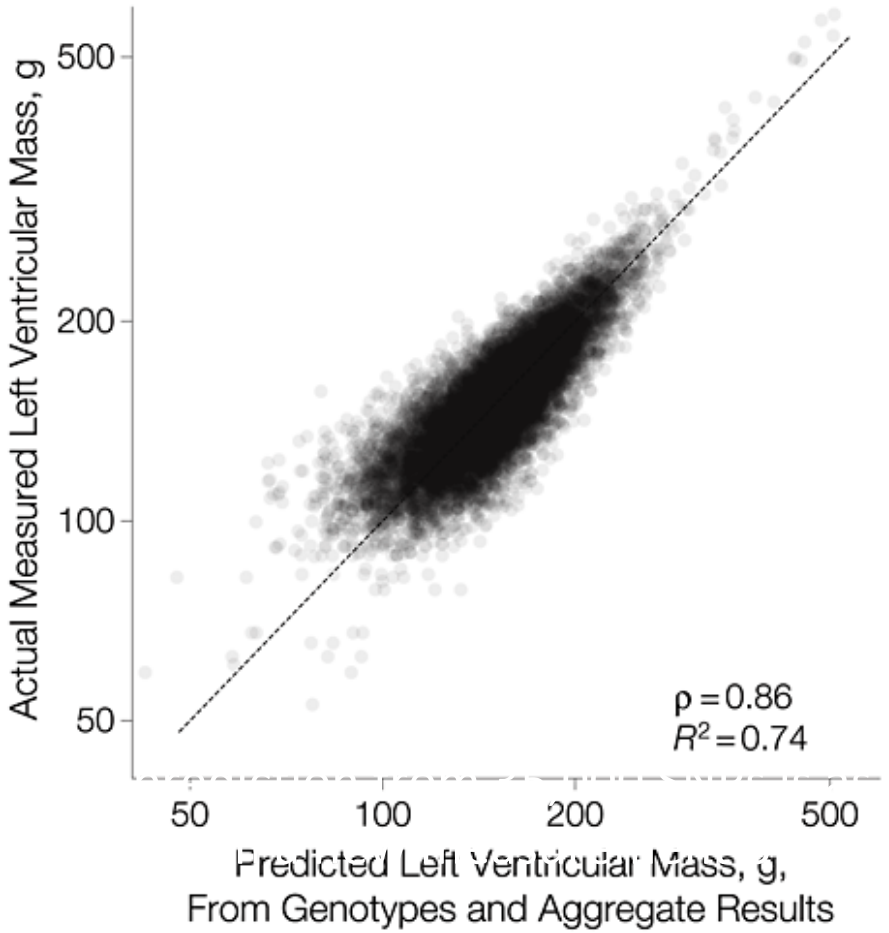
35
35

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]

GWAS Predictions



37
37

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

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.

39
39

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

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

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.

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

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.

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.

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).

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

47
47

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.

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

Automated fitting of predictive models

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.

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

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

53
53

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

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

55
55

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

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

57
57

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

58
58

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.

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.

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.

61
61

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.

62
62

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

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.

Summary

- Prediction can be
  - prediction of a summary statistic, with confidence interval
  - point prediction of a best guess
  - interval prediction
- Model accuracy is important
  - regression model for fitted mean must be accurate
  - for interval prediction, assumptions about distribution of outcome must be accuate

65
65

Summary

- Prediction intervals are wider than confidence intervals, and do not shrink to zero width with increasing sample size

- Prediction intervals depend on the distribution of the data: only the Normal distribution is widely available in software

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

Estimation (Prediction) of Summary Measures

Examples

- Estimate age, height, and sex specific mean (or geometric mean) FEV
  - Linear regression to obtain estimates and CI
- Estimate probability (or odds) of remaining in remission for 24 months by age, PSA
  - Logistic regression to obtain estimates and CI
- Estimate median time to liver failure in PBC patients by age, bilirubin, etc.
  - Proportional hazards regression for estimates (and CI?)

69

Issues

- Which statistic provides the best estimate?
  - Definition of best?
    - Consistent (correct with infinite sample size)
    - Precise (minimal variability, minimal squared error)
  - Answer: Common regression models provide the best estimate in a wide variety of settings
- Is best good enough in particular setting?
  - Answer: CI for the value of true summary measure for each group

70

General Methods

- Estimated summary measure involves a linear function of regression parameters
- Linear, logistic, Poisson regression this is all that is needed
- Proportional hazards regression also needs an estimate of the survival distribution in the reference group
  - We are not yet very good at putting confidence bounds on this part of the estimates

71

Estimating θ Within Groups

- Relevant sampling for scientific question
  - (or appropriate weighting of observations)
- Independence
  - (or correct modeling of dependent data within clusters)
- Sufficient sample sizes to approximate asymptotic distributions
- Variance appropriate to the model
  - (relaxed when using robust Huber-White SE)
- Regression model accurately describes summary measures across all groups
  - Linearity on appropriate scale; any effect modification of interest
  - Proportional hazards assumption holds with PH regression
- Shape of distribution same in each group
  - (or can be correctly derived using parameter estimates)

72

Inference About θ Within Groups

- Relevant sampling for scientific question
  - (or appropriate weighting of observations)
- Independence
  - (or correct modeling of dependent data within clusters)
- Sufficient sample sizes to approximate asymptotic distributions
- Variance appropriate to the model
  - (relaxed when using robust Huber-White SE)
- Regression model accurately describes summary measures across all groups
  - Linearity on appropriate scale; any effect modification of interest
  - Proportional hazards assumption holds with PH regression
- Shape of distribution same in each group
  - (or can be correctly derived using parameter estimates)

73

Obtaining Point Estimates

- Substitution of predictor values provides the estimate of the modeled transformation of the summary measure
  - Linear regression: mean
  - Linear regression on logs: log geometric mean
  - Logistic regression: log odds
  - Poisson regression: log rate
  - Proportional hazards: log hazard ratio applied to baseline survival estimate

74

Obtaining Interval Estimates

- Under the appropriate assumptions, we can obtain standard errors for each such estimate
  - Notable exception: Proportional hazards
    - More work to be done to get interval estimates
- Standard errors for the point estimates must take the covariance of the regression parameter estimates
  - The covariance matrix is typically stored but not printed
- In logistic, Poisson, PH regression, we generally find a confidence interval for the transformed summary measure, and then back transform to obtain the desired quantity

75

Statistical Software

- Traditionally: provide commands to
  - Estimate predicted (fitted) value
    - Linear predictor, exponentiated, or transformed to mean (logistic)
  - Estimate standard error of predicted value

- Better: provide commands to obtain arbitrary estimates and CI of linear predictors
  - Possibly back-transformed to relevant quantity

76

Stata Commands: `predict`

- Get predictions for covariate combinations in dataset
- After performing any regression command, Stata command
  - `predict` *`varname`*`, [`*`what`*`]`
    - *`varname`* is the name of the variable to store predictions
    - *`what`* is an option specifying what you want computed
      - xb = linear prediction (works for all types)
      - stdp = SE of linear prediction (works for all types)
      - p = probability (works for logistic)
- To get CI, just use the usual formula: (est) +/- (crit val) * (std err)
  - In linear regression: we usually use the t distribution to obtain CI
    - Stata: (crit val) = `invttail(`*`df`*`,`*`α/2`*`)`
    - degrees of freedom = n minus number of regression parameters
  - In all other regressions: we use the standard normal distribution
    - (crit val) = `invnorm(1-`*`α/2`*`)` (1.96 for 95% CI)

77

Stata Commands: `lincom`

- Get predictions and CI for specified covariate combinations
- After performing any regression command, Stata command
  - `lincom` *`expr`*`, [`*`eform`*`]`
    - *`expr`* is an expression involving regression parameters
      - Cannot include additive constants with logistic, Poisson, PH regression (a silly restriction – perhaps revised in later versions?)
    - *`eform`* is an option specifying that estimates should be exponentiated

78

Ex: Geom Mean FEV by ht, age

```
. regress logfev height age
Number of obs =     654
------------------------------------------------------------------------------
      logfev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |        .044   .002       26.71   0.000      .041       .047
         age |        .020   .003        6.23   0.000      .014       .026
       _cons |       -1.97   .078      -25.16   0.000     -2.12      -1.82
------------------------------------------------------------------------------

. lincom 66*height + 10*age + _cons
 ( 1)  66*height + 10*age + _cons = 0

------------------------------------------------------------------------------
      logfev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |     1.13044    .009702    116.52   0.000     1.1114     1.14949
------------------------------------------------------------------------------


. lincom 66*height + 10*age + _cons, eform
 ( 1)  66*height + 10*age + _cons = 0

------------------------------------------------------------------------------
      logfev |      exp(b)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |     3.09702   .0300474    116.52   0.000     3.03858    3.15659
------------------------------------------------------------------------------
```

Ex: Geom Mean FEV by ht, age

```
regress logfev height age
Number of obs =     654
------------------------------------------------------------------------------
      logfev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |    .044    .002    26.71   0.000      .041    .047
         age |    .020    .003     6.23   0.000      .014    .026
       _cons |   -1.97    .078   -25.16   0.000    -2.12   -1.82
------------------------------------------------------------------------------


predict flogfev
predict sefit, stdp
g gmfev= exp(flogfev)
g gmlofev = exp(flogfev – invttail(651, .025) * sefit)
g gmhifev = exp(flogfev + invttail(651, .025) * sefit)
list gmfev gmlofev gmhifev if age==10 & height==66
          gmfev    gmlofev    gmhifev
330.  3.097021   3.038578   3.156588
```

80

Ex: Odds Relapse by NadirPSA

```
. logit relapse24 lognadir, robust

. predict lorel, xb
. predict selo, stdp

. g odds= exp(lorel)
. g oddslo= exp(lorel - 1.96 * selo)
. g oddshi= exp(lorel + 1.96 * selo)

. list odds oddslo oddshi if nadir==1
          odds     oddslo     oddshi
 10.  .4911836   .2388794   1.009971
```

81

Ex: Prob Relapse by NadirPSA

```
. logit relapse24 lognadir, robust

. predict prel

. g prob = odds / (1+odds)
. g problo= oddslo / (1 + oddslo)
. g probhi= oddshi / (1 + oddshi)

. list prel prob problo probhi if nadir==1
          prel       prob     problo     probhi
 10.  .3293918   .3293918    .192819   .5024805
```

82

Prediction in PH Regression

- Recall that there is no intercept in PH models
  - Instead there is a “baseline hazard function” which is related to the survival function in the reference group
  - Stata will allow prediction of baseline survival function in their “stcox” command
    - Specify option `basesurv(`*`newvar`*`)` in `stcox`
    - Then use `stcurve, survival at( )`

83

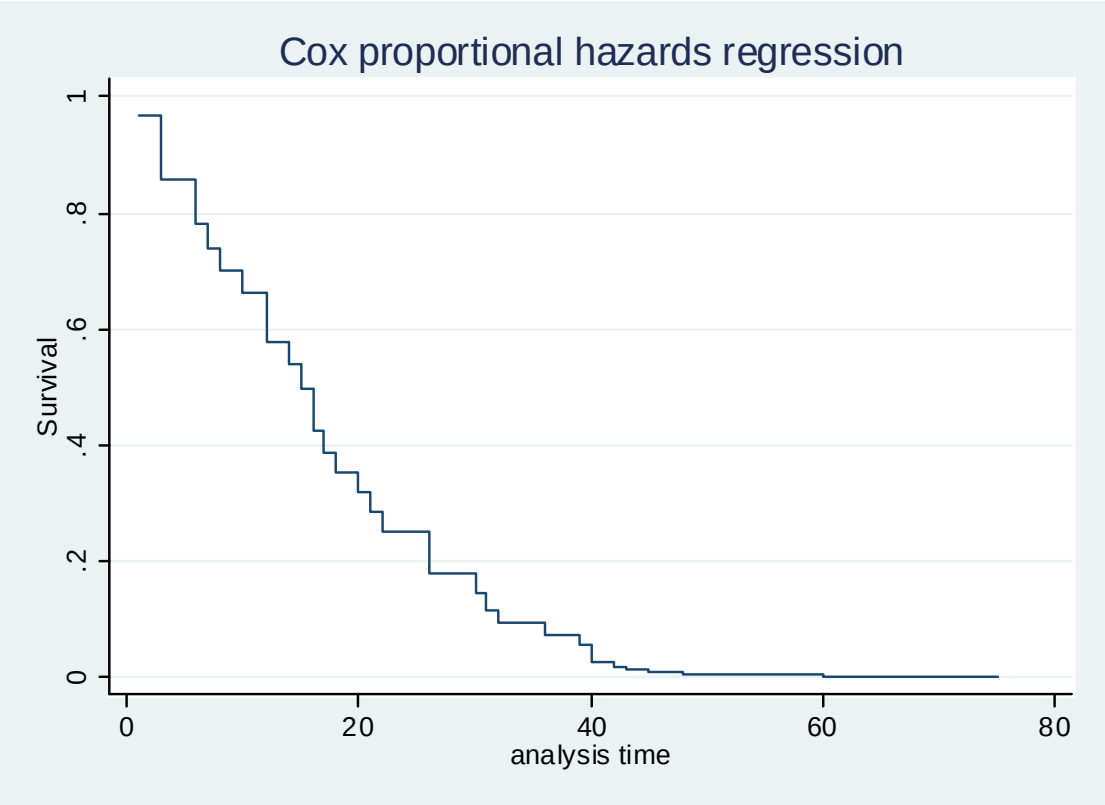
Stata Ex: Relapse in PSA Data

```
. g relapse=0
. replace relapse=1 if inrem=="no"
. stset obstime relapse
. g lnadir= log(nadir)
. stcox lnadir ps, robust basesurv(bslnS)
No. of subjects = 48                 Number of obs =     48
No. of failures = 34                 Time at risk  =   1408
                                     Wald chi2(2)  =  33.18
Log pseudolklhd = -97.1              Prob > chi2   = 0.0000
       |      Robust
    _t |   HR   SE      z    P>|z|     [95% C I]
-------+------------------------------------------------
lnadir | 1.56  .124   5.66   0.000   1.34    1.83
    ps | .960  .0162 -2.41   0.016   .929    .992
```

84

Stata Ex: Predicted Survival

- `stcurve, survival at(lnadir=2 ps=70)`



85

Comments on PH Regression

- We can thus easily obtain estimated summary measures for any group based on semi-parametric PH assumption
  - Survival probabilities
  - Quantiles (median, etc.)
  - (Restricted mean (area under survival curve))
- We do not yet provide SE for those estimates

86

Prediction (Forecast) of Individual Measurements

Examples

- Estimate “normal range” for FEV by age, height, and sex groups
  - Linear regression
- Estimate probability (or odds) of remaining in remission for 24 months by age, PSA
  - Logistic regression
- Estimate range of times to liver failure in PBC patients by age, bilirubin, etc.
  - Proportional hazards regression

88

Issues

- Which statistic provides the best estimate?
  - Definition of best?
    - Consistent (correct with infinite sample size)
    - Precise (minimal variability, minimal squared error)
  - Answer: Common regression models provide the best estimate in a wide variety of settings
- How variable is "best" in particular setting?
  - Answer: Prediction (Stata: Forecast) interval for the value of individual observation in each group

89

Necessary Assumptions

- Independence
  - (between identified clusters for robust SE)
- Variance appropriate to the model
  - (NOT relaxed for robust SE)
- Regression model accurately describes relationship of summary measures across groups
- **Shape of distribution same in each group**
- Sufficient sample sizes for asymptotic distributions to be a good approximation

90

Comments

- These are strong assumptions
  - Consequently, we do not have many methods that provide robust inference
    - Robust SE will only work here for correlated response, not for heteroscedasticity
  - For the most part, precise methods have only been well developed for
    - Binary or Poisson variables
      - All we need is an estimate of the probability or rate
    - Normally distributed data

91

Obtaining Point Estimates

- Substitution of predictor values provides the estimate of the modeled transformation of the summary measure
  - Linear regression: mean
  - Linear regression on logs: log geometric mean
  - Logistic regression: log odds
  - Poisson regression: log rate
  - Proportional hazards: log hazard ratio applied to baseline survival estimate

92

Obtaining Interval Estimates

- Under the appropriate assumptions, we can obtain standard errors for each such estimated summary measure
  - Notable exception: Proportional hazards
    - More work to be done to get interval estimates
  - We generally find a confidence interval for the transformed summary measure, and then back transform to obtain the desired quantity
- THEN: Add in variability within group

93

Statistical Software

- No statistical package that I know of will provide prediction intervals except for normally distributed data
  - Even then, I do not think that they are behaving the way we want them to
    - Frequentist intervals describe behavior across repeated experiments, not within one experiment

94

Prediction Intervals: Normal Data

- Obtaining point estimates
  - The point prediction is typically the mean (or log geometric mean) from the regression model

95

Obtaining Interval Estimates

- Under the appropriate assumptions, we can obtain standard errors for each such prediction
  - The standard error accounts for
    - Uncertainty in estimating the regression parameters
    - The within group standard deviation
      - Spread of data about the group specific means

96

Stata Commands: Predict

- After performing any regression command, the Stata command "`predict`" will compute estimates and standard errors
  - `predict` *`varname`*`, [`*`what`*`]`
    - *`varname`* is the name of the variable where you want the predictions stored
    - *`what`* is an option specifying what you want computed
      - stdf = standard error of forecast (works for linear regression)

97

Computing Prediction Intervals

- Just use the usual formula

  (est) +/- (crit val) * (std err)

  – In linear regression, we usually use the t distribution to obtain CI
    - Stata: (crit val) = `invttail(`*df*`,`*α/2*`)`
    - degrees of freedom = n minus number of regression parameters

98

Ex: Geom Mean FEV by ht, age

```
regress logfev height age
Number of obs =     654
------------------------------------------------------------------------------
      logfev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |    .044    .002    26.71   0.000     .041    .047
         age |    .020    .003     6.23   0.000     .014    .026
       _cons |   -1.97    .078   -25.16   0.000    -2.12   -1.82
------------------------------------------------------------------------------

predict flogfev
predict sefore, stdf
g predfev= exp(flogfev)
g predlofev = exp(flogfev – invttail(651, .025) * sefore)
g predhifev = exp(flogfev + invttail(651, .025) * sefore)
list predfev predlofev predhifev if age==10 & height==66
       predfev  predlofev  predhifev
330.  3.097021   2.320911   4.132662
```

99

Compare: CI for Parameter

- Using the "standard error of the prediction"
  - 95% CI for geometric mean of 66" tall 10 yo
    - From slide 33: (3.039, 3.157)
  - Tells us how precisely we know the geometric mean, which is a single number
  - As n becomes infinite, the width of the CI goes to 0
    - We will know the geometric mean for that group exactly
      - (if our model is correct)

100

Compare: Prediction Interval

- Using the "standard error of the forecast"
  - 95% PI for FEV measurements of 66" tall 10 year olds
    - From slide 52: (2.321, 4.133)
  - Tries to predict the range containing 95% of measurements in the population of 66" tall 10 year olds
  - As n becomes infinite, the width of the PI (on the log scale) would be +/- 1.96 SD

101

Caveat

- This “forecast” or “prediction interval” assumes that the log FEV measurements are normally distributed
  - This is a pretty strong assumption

Extensions

- I know how to get approximate intervals based on some slightly weaker semi-parametric assumptions
    - Uses nonparametric estimates of the error distribution
    - This would work for censored data as well
  – Most software packages will not do this

103

Better Approaches

- It would be better to find nonparametric confidence intervals for
  - the 2.5th percentile
  - the 97.5th percentile

104

But Still…

- All of these methods suffer from
  - Strong semiparametric assumptions
  - Multiple comparisons if more than one group
    - (But we do know how to get confidence bands)
  - Coverage probabilities defined across replicate experiments
    - On average (across experiments), 95% of observations will be within an interval
    - But in any given experiment, the intervals might truly cover less or more of the population

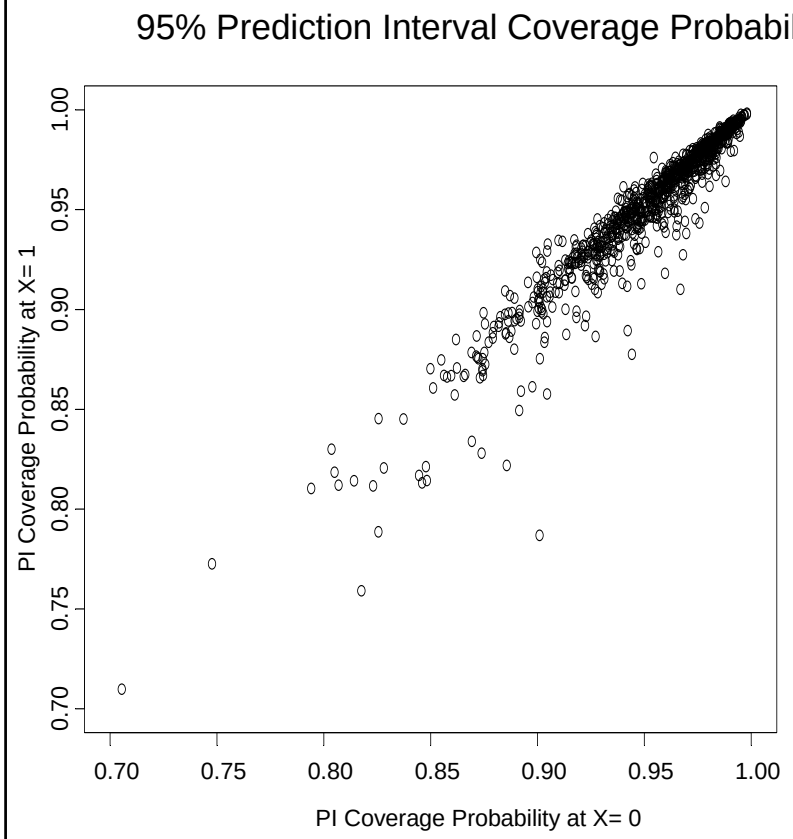
105

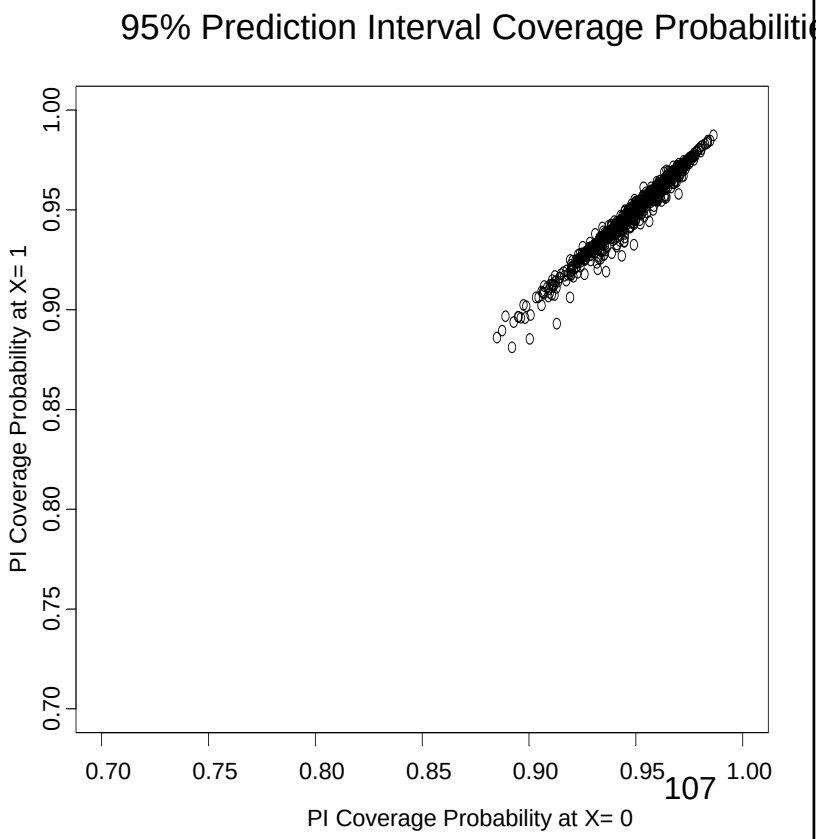
Simulation Study

- Perform 1000 simulated regressions
  - X is normally distributed, mean 0, sd 1
  - N= 25 or 100
  - Generate 95% prediction intervals for
    - X = 0 (mean)
    - X = 1 (1 sd from the mean)
  - Calculate true coverage probability of each prediction interval
    - (I know the truth in this case)

106

Plots of Coverage Probabilities





107

Coverage Probabilities

- Sample size N= 25
  - Mean coverage probability: 0.950
  - Interquartile range: 0.935 – 0.978
  - Range: 0.706 – 0.998

- Sample size N= 100
  - Mean coverage probability: 0.950
  - Interquartile range: 0.941 – 0.962
  - Range: 0.885 – 0.986

108

Joint Coverage of 2 Pred Intvl

- Sample size N= 25
  - Mean coverage probability: 0.906
  - Interquartile range: 0.874 – 0.956
  - Range: 0.501 – 0.996

- Sample size N= 100
  - Mean coverage probability: 0.903
  - Interquartile range: 0.884 – 0.926
  - Range: 0.784 – 0.974

109

Correlated Response

- Prediction Intervals can be computed for correlated response
  - Stata, however, does not provide the obvious approximation
  - Thus for the SEP dataset we would have options of
    - Using mean p60 and adjusting the PI “by hand”
    - Identifying clusters and computing PI “by hand”
    - (More advanced models
      - mixed effects, repeated measures)

110

Prediction Intervals

- Basic idea behind prediction intervals

Model :

$$Y_i \mid X_i \sim N\left(\beta_0 + \beta_1 \times X_i, \sigma^2\right)$$

$$Y_i \mid X_i = \beta_0 + \beta_1 \times X_i + \varepsilon_i$$

Estimated mean :

$$\hat{\beta}_0 + \hat{\beta}_1 \times X_i \sim N\left(\beta_0 + \beta_1 \times X_i, \sigma^2 V\right)$$

Predicted observation :

$$\hat{\beta}_0 + \hat{\beta}_1 \times X_i + \varepsilon_i \sim N\left(\beta_0 + \beta_1 \times X_i, \sigma^2 (1+V)\right)$$

111

Computing Prediction Intervals

- We use an estimate for the within group variance
  - So we usually use the t distribution instead of the normal distribution
- With correlated response data, the degrees of freedom can be more complicated
  - But if n is large, it makes little difference

112

With Correlated Response

- With a balanced design the "Root MSE" is still consistent for the within group standard deviation
- Hence, we can approximate the standard error of the forecast as

Estimated mean :

$$\hat{\beta}_0 + \hat{\beta}_1 \times X_i \sim N\left(\beta_0 + \beta_1 \times X_i, \sigma^2 V\right)$$

Predicted observation :

$$\hat{\beta}_0 + \hat{\beta}_1 \times X_i + \varepsilon_i \sim N\left(\beta_0 + \beta_1 \times X_i, \sigma^2 (1+V)\right)$$

$$\hat{se}(Forecast) = \sqrt{se^2\left(\hat{\beta}_0 + \hat{\beta}_1 \times X_i\right) + \hat{\sigma}^2}$$

113

Prediction of Binary Measurements

114

Classification (Discrimination)

- Sometimes the scientific question is one of deriving a rule to classify subjects
  - Diagnosis of prostate cancer
    - Based on age, race, and PSA, should we make a diagnosis of prostate cancer?
  - Prognosis of patients with primary biliary cirrhosis
    - Based on age, bilirubin, albumin, edema, protime, is the patient likely to die within the next year?

115

Prediction of a Binary Variable

- Classification can be regarded as trying to predict the value of a binary variable
  - Before (slides 34-35) we were estimating the probability and odds of relapse within a particular group: A summary measure
  - Now we want to decide whether a particular individual will relapse: An individual measurement
  - Obvious connection:
    - The probability or odds tells us everything about the distribution of values
    - The only possible values are 0 or 1

116

Typical Approach

- Use regression model to estimate probability of the event in each group
- Form a decision rule based on estimated probability of the event
  - If estimate $\geq$ c, predict measurement is 1
  - If estimate < c, predict measurement is 0
- Quantify accuracy of decision rule
  - Sens, Spec, Pred Val Pos, Pred Val Neg

117

Often: Stepwise Model Building

- Consider a large number of covariates that might possibly be predictive
  - Starting model
    - No covariates: “Forward stepwise regression”
    - All covariates: “Backward stepwise regression”
  - Add or remove covariates based on the corresponding partial t or partial Z test
    - “P to enter” and “P to remove”
    - Avoid infinite loops: “P to enter” < “P to remove”

118

Caveats

- Stepwise model building “overfits” your data
  - “P values” are not true p values—instead they are anti-conservative
- You will quite often obtain different models depending upon whether you go “forward” or “backward”

119

Use of Stepwise Model Building

- Exploratory data analyses
  - Statistical question 4a: Which covariates should we rigorously investigate first, because they seem to have the strongest association with response?
    - Provides an order that we might consider the covariates
    - Does not tell us whether any of the covariates are truly associated
      - Many false positives

120

Use of Stepwise Model Building

- Predictive models
  - Statistical question 5: What is our best estimate for an individual's measurement?
    - We are not interested in the association between the covariates in the model and the response
    - We do not mind confounding or surrogate variables
    - We will judge accuracy of our predictive model by evaluating sens, spec, PV+, PV- in an independent sample

121

Stata Commands

- Stata has prefix command “stepwise” that works with most regression commands

```
stepwise, pe(#) pr(#) [forward]:
```

  - “P to enter”: a number between 0 and 1
  - “P to remove”: a number between 0 and 1
  - forward or backward: backward is default

122

Example

- Stepwise model building in inflammatory markers data set to predict who will die within 4 years
  - No subjects were censored before 4 years
  - Use logistic regression
  - Consider variables
    - age, male, smoker, prevdis, diab2, bmi ,systBP, cholest, cholsqr, crp, logcrp, fib
      - (Note that I am allowing cholesterol to have a U shaped trend, and I consider a transformation of CRP as well)

123

Example: Forward Stepwise

```
. stepwise, pr(0.10) pe(0.05) forward: logistic deadIn4 age male smoker
   prevdis diab2 bmi systBP cholest cholsqr crp logcrp fib


begin with empty model
p = 0.0000 <  0.0500  adding   age
p = 0.0000 <  0.0500  adding   logcrp
p = 0.0000 <  0.0500  adding   male
p = 0.0000 <  0.0500  adding   prevdis
p = 0.0000 <  0.0500  adding   diab2
p = 0.0005 <  0.0500  adding   smoker
p = 0.0032 <  0.0500  adding   systBP
```

124

Example: Forward Stepwise

```
Logistic regression     Number of obs   =     4861
                        LR chi2(7)      =   412.54
                        Prob > chi2     =   0.0000
Log likelihood = -1345  Pseudo R2       =   0.1330
------------------------------------------------------------------------------
deadIn4 |  OR      SE      z    P>|z|     [95% CI]
--------+---------------------------------------------------------------------
    age | 1.115  .0095  12.81  0.000  1.097  1.134
 logcrp | 1.444  .0731   7.26  0.000  1.308  1.595
   male | 2.122  .2216   7.20  0.000  1.729  2.604
prevdis | 2.056  .2181   6.80  0.000  1.670  2.531
  diab2 | 1.824  .2193   5.00  0.000  1.441  2.309
 smoker | 1.698  .2555   3.52  0.000  1.264  2.281
 systBP | 1.007  .0023   2.94  0.003  1.002  1.011
------------------------------------------------------------------------------
```

125

Example: Forward Stepwise

- Interpretation
  - Provides an ordering of the variables with respect to observed strength of association
    - In the case of forward stepwise, Stata lists variables according to "P value"
  - We cannot trust the P values due to the data driven analyses
  - It is possible that confounding relationships kept some variables out of the model

126

Example: Backward Stepwise

```
. stepwise, pr(0.10) pe(0.05): logistic deadIn4 age male smoker prevdis
   diab2 bmi systBP cholest cholsqr crp logcrp fib
                      begin with full model
p = 0.2157 >= 0.1000  removing cholsqr
p = 0.3768 >= 0.1000  removing cholest


Logistic regression     Number of obs   =    4861
                        LR chi2(10)     =  421.22
                        Prob > chi2     =  0.0000
Log likelihood = -1341  Pseudo R2       =  0.1358
```

127

Example: Backward Stepwise

deadIn4	OR	SE	z	P>|z|	[95% CI]	
age	1.111	.0097	12.06	0.000	1.092	1.130
male	2.123	.2232	7.16	0.000	1.728	2.609
smoker	1.577	.2414	2.97	0.003	1.168	2.129
prevdis	2.023	.2154	6.61	0.000	1.642	2.492
diab2	1.883	.2300	5.18	0.000	1.482	2.393
bmi	.979	.0120	-1.75	0.079	.956	1.003
systBP	1.007	.0023	2.88	0.004	1.002	1.011
logcrp	1.553	.1394	4.90	0.000	1.302	1.851
fib	1.002	.0009	1.98	0.048	1.000	1.003
crp	.980	.0111	-1.77	0.077	.959	1.002

128

Example: Backward Stepwise

- Interpretation
  - Provides an ordering of the variables with respect to observed strength of association
    - In the case of backward stepwise, Stata lists variables according to original order
  - We cannot trust the P values due to the data driven analyses
  - Compare to forward
    - Some additional variables with P > 0.05
    - But also some additional with P < 0.05

129

Stepwise for Classification

- We sometimes use stepwise model building to derive a classification rule
  - To ensure valid estimates of classification rates, we usually divide a sample into
    - Training sample used to build a regression model, and
    - Validation sample used to compute the classification rates
      - Sensitivity, specificity, predictive value of the positive, predictive value of the negative

130

Example

- Prognostic model for death in 4 years
  - Training sample containing about 60% of data
  - Backward stepwise variable selection
  - Estimated probability of death used to classify
    - Some arbitrary threshold
  - Use all other cases (validation set) to compute
    - Sensitivity, specificity (condition on survival status)
    - PV+, PV- (condition on estimated p > threshold)

131

Example: Model Building

```
. g training= uniform()
. stepwise, pr(0.10) pe(0.05): logistic deadIn4 age male smoker prevdis
   diab2 bmi systBP cholest cholsqr crp logcrp fib if training <= 0.60
begin with full model
p = 0.9919 >= 0.1000  removing cholsqr
p = 0.4914 >= 0.1000  removing cholest
p = 0.4475 >= 0.1000  removing fib
p = 0.1908 >= 0.1000  removing smoker
Logistic regression     Number of obs   =       2875
(output deleted – we do not care about it)
. predict pfit
```

132

Example: Sens, Spec, PV+, PV-

- Consider a rule that predicts death if the estimated *pfit* is greater than 0.5
  - Create a variable indicating *pfit* > 0.5
  - Cross tabulate deadIn4 and pfit
    - Sensitivity and specificity from row percentages
    - PV+ and PV- from column percentages

133

Example: Sens, Spec, PV+, PV-

```
. g pfitHigh= pfit
. recode pfitHigh 0/0.5=0 0.5/1=1
. tabula deadIn4 pfitHigh if training > 0.6, row col
           |        pfitHigh
   deadIn4 |          0              1 |     Total
-----------+--------------------------+----------
         0 |      1,792              7 |     1,799
           |Spec: 99.61          0.39 |    100.00
           |PV-:  90.64         41.18 |     90.22
-----------+--------------------------+----------
         1 |        185             10 |       195
           |      94.87  Sens:    5.13 |    100.00
           |       9.36  PV+:    58.82 |      9.78
-----------+--------------------------+----------
     Total |      1,977             17 |     1,994
           |      99.15          0.85 |    100.00
           |     100.00        100.00 |    100.00
```

134

Example: Other Thresholds

- Sensitivity, specificity will vary by threshold

Threshold	Specificity	Sensitivity
0.05	43%	87%
0.10	68%	73%
0.15	84%	48%
0.20	91%	37%
0.50	99.6%	5%

135

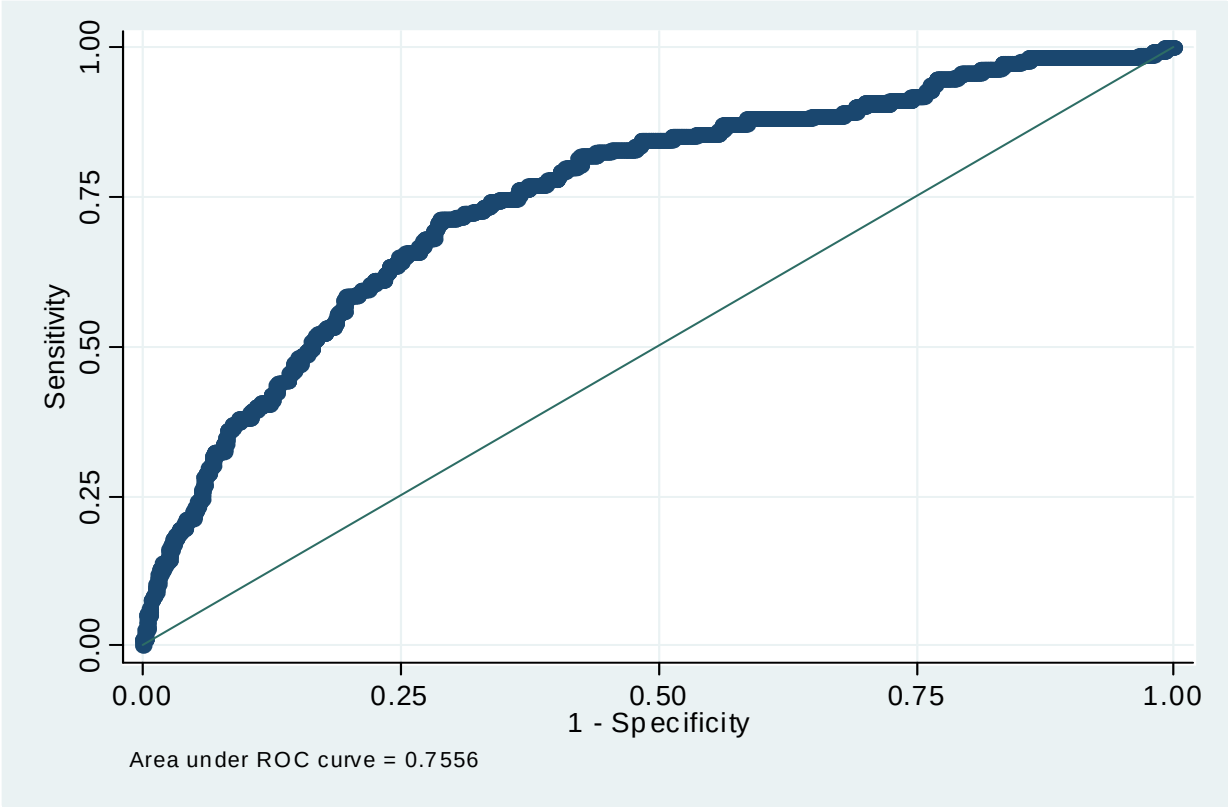
ROC Curve Analysis

- Receiver Operating Curves (from Engr)
  - Compare sens and spec as threshold varies
  - Y axis: Sensitivity (True Positive rate)
  - X axis: 1 – Specificity (False Positive rate)
  - Interpretation
    - Sometimes summarize area under curve (AUC)
    - A diagonal line: Like flipping a coin (AUC = 0.5)
    - ROC curve in upper left: Ideal (AUC = 1.0)
    - Comparing two rules:
      - If one ROC curve always above the other, that rule will always have better PV+ and PV- for all prevalences

136

Stata Commands

```
. roctab deadIn4 pfit if training > 0.60, graph
```



137