Biost 518
Applied Biostatistics II

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

Lecture 11:
Regression Based Prediction

February 25, 2008

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

General Classification

• Clustering of observations
• Clustering of variables
• Quantification of distributions
• Comparing distributions
• Prediction of individual observations
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)

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

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

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)
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)

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)

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?)

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
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

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

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?)

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
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

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

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?)

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
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

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

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

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?)

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

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

Necessary Assumptions

- Independence
  - (between clusters for robust SE)
- Variance appropriate to the model
  - (relaxed for robust SE)
- Regression model accurately describes relationship of summary measures across groups
- Sufficient sample sizes for asymptotic distributions to be a good approximation
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

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
    - We generally find a confidence interval for the transformed summary measure, and then back transform to obtain the desired quantity

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
      - `xb` = linear prediction (works for all types)
      - `stdp` = SE of linear prediction (works for all types)
      - `p` = probability (works for logistic)

Computing CI for Predictions

- Just use the usual formula
  \[(\text{est}) \pm (\text{crit val}) \times (\text{std err})\]
  - In linear regression, we usually use the t distribution to obtain CI
    - Stata: (crit val) = `invttail(df, \alpha/2)`
    - degrees of freedom = \(n - \text{number of regression parameters}\)
  - In all other regressions, we would use the standard normal distribution
    - (crit val) = `invnorm(1-\alpha/2)` (1.96 for 95% CI)
Ex: Geom Mean FEV by ht, age

```
regress logfev height age
Number of obs = 654

logfev | Coef. Std. Err. t P>|t| [95% CI]
------- | -------- -------- -------- -------- --------
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
```

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
```

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
```

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( )
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

Stata Ex: Predicted Survival

• stcurve, survival at(lnadir=2 ps=70)

Prediction (Forecast) of Individual Measurements

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
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

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

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

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
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

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

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

Prediction Intervals: Normal Data

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

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)

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

Ex: Geom Mean FEV by ht, age

```
regress logfev height age
Number of obs = 654

Coef.   Std. Err.   t    P>|t|     [95% CI]
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 3.097021 2.320911 4.132662
```
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)

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

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
Better Approaches

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

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

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)

Plots of Coverage Probabilities
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

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

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)

Prediction Intervals

- Basic idea behind prediction intervals

Model: \( Y_i \mid X_i \sim N(\beta_0 + \beta_1 \times X_i, \sigma^2) \)

Estimated mean:
\[ \hat{\beta}_0 + \hat{\beta}_1 \times X_i \sim N(\beta_0 + \beta_1 \times X_i, \sigma^2 V) \]

Predicted observation:
\[ \hat{\beta}_0 + \hat{\beta}_1 \times X_i + \epsilon_i \sim N(\beta_0 + \beta_1 \times X_i, \sigma^2(1+V)) \]
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

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

\[
\hat{Y} + \hat{\beta} \times X_i + \epsilon - N(\hat{\beta}_0 + \hat{\beta}_1 \times X_i, \sigma^2(I + V))
\]

\[
\text{s.e.(Forecast)} = \sqrt{se^2(\hat{\beta}_0 + \hat{\beta}_1 \times X_i) + \hat{\sigma}^2}
\]

Prediction of Binary Measurements

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?
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

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 > 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

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”

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”
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

- 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

Stata Commands

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

  \texttt{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

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)
Example: Forward Stepwise

. stepwise, pr(0.10) pe(0.05) forward: logistic
deadIn4 age male smoker prevdis diab2 bmi systBP
cholesq 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

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

Example: Backward 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
Example: Backward Stepwise

|      | OR    | SE    | z    | P>|z| | [95% CI] |
|------|-------|-------|------|------|----------|
| age  | 1.111 | .0097 | 12.06| 0.000| 1.092    |
| male | 2.123 | .2232 | 7.16 | 0.000| 1.728    |
| smoker| 1.577 | .2414 | 2.97 | 0.003| 1.168    |
| prevdis| 2.023| .2154 | 6.61 | 0.000| 1.642    |
| diab2| 1.883 | .2300 | 5.18 | 0.000| 1.482    |
| bmi  | .979  | .0120 | -1.75| 0.079| .956     |
| systBP| 1.007 | .0023 | 2.88 | 0.004| 1.002    |
| logcrp| 1.553 | .1394 | 3.90 | 0.000| 1.302    |
| fib  | 1.002 | .0009 | 1.98 | 0.048| 1.000    |
| crp  | .980  | .0111 | -1.77| 0.077| .959     |

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

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

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)
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
```

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

```
g pfitHigh= pfit
recode pfitHigh 0/0.5=0 0.5/1=1
tabulate 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 | 94.87 | 90.22
| 1 |         185           10 |       195
| Sens: 94.87 | 5.13 | 100.00
| PV+: 94.87 | 94.87 | 94.87
Total |      1,977           17 |     1,994
| 99.15 | 84.67 | 99.15
| 100.00 | 100.00 | 100.00
```

Example: Other Thresholds

- Sensitivity, specificity will vary by threshold

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Specificity</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>43%</td>
<td>87%</td>
</tr>
<tr>
<td>0.10</td>
<td>68%</td>
<td>73%</td>
</tr>
<tr>
<td>0.15</td>
<td>84%</td>
<td>48%</td>
</tr>
<tr>
<td>0.20</td>
<td>91%</td>
<td>37%</td>
</tr>
<tr>
<td>0.50</td>
<td>99.6%</td>
<td>5%</td>
</tr>
</tbody>
</table>
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

Stata Commands

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

Area under ROC curve = 0.7556