# Biost 518
# Applied Biostatistics II

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

## Lecture 9:
## Regression Based Prediction

March 3, 2006





## 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, count 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

  (est) +/- (crit val) * (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 minus 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
         gmfev    gmlofev    gmhifev
330.  3.097021   3.038578   3.156588
```



# 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
```



# 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
```



# 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( )`

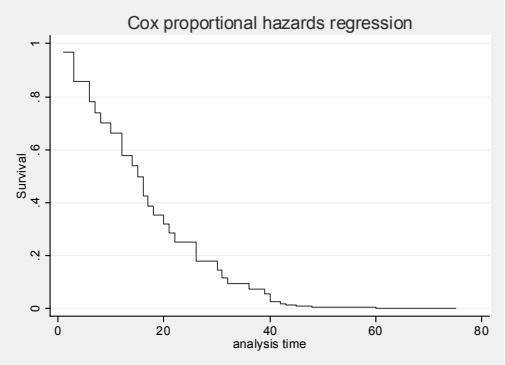
# Stata Ex: Relapse in PSA Data

```
. 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)`






# 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



# 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



## 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*`,`$\alpha/2$`)`
    - degrees of freedom = n minus number of regression parameters



## Ex: Geom Mean FEV by ht, age

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

## 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 <u>slightly</u> 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…

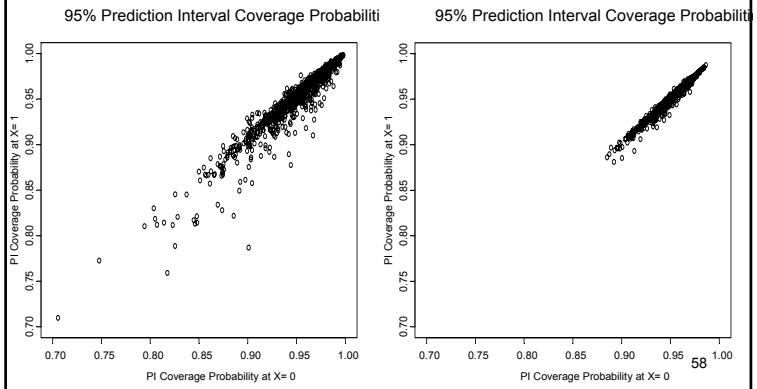
- All of these methods suffer from
  - Strong semiparametric assumptions
  - Multiple comparisons if more than one group
    - (But we do know how to get confidence bands)
  - Coverage probabilities defined across replicate experiments
    - On average (across experiments), 95% of observations will be 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 :

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

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

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 :

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



## 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 variance
- 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 :

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

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