

Applied Regression Analysis

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

Session 9

Applied Regression Analysis

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

Part 4: Extension to Other Regression Models

Lecture Outline

-
- Topics:
 - General regression model
 - Logistic regression
 - Proportional hazards regression
 - Example: Prognostic value of PSA

General Regression Model

.....

Types of Variables

.....

- Statistical classification of scientific data
 - Binary data
 - E.g., sex, death
 - Nominal data: unordered, categorical data
 - E.g., race, marital status
 - Ordinal categorical data
 - E.g., stage of disease
 - Quantitative data
 - E.g., age, blood pressure
 - Right censored data
 - E.g., time to death (when not everyone has died)

Summary Measures

.....

- The measures commonly used to summarize and compare distributions vary according to the types of data
 - Means: binary; quantitative
 - Medians: ordered; quantitative; censored
 - Proportions: binary; nominal
 - Odds: binary; nominal
 - Hazards: censored
 - hazard = instantaneous rate of failure

Regression Models

- Regression methods differ primarily according to the parameter compared across groups
 - Means → Linear regression
 - Geom Means → Linear regression on logs
 - Odds → Logistic regression
 - Rates → Poisson regression
 - Hazards → Proportional Hazards regression
 - Quantiles → Parametric survival regression

General Regression

- General notation for variables and parameter
 - Y_i Response measured on the i th subject
 - X_i Value of the predictor for the i th subject
 - θ_i Parameter of distribution of Y_i
 - The parameter might be the mean, geometric mean, odds, rate, instantaneous risk of an event (hazard), etc.

Simple Regression

- General notation for simple regression model
 - $g(\theta_i) = \beta_0 + \beta_1 \times X_i$
 - $g(\)$ "link" function used for modeling
 - β_0 "Intercept"
 - β_1 "Slope (for predictor X)"
 - The link function is usually either none (means) or log (geom mean, odds, hazard)

Regression Analysis

- The major difference between the various regression models is then the interpretation of the parameters
 - Issues related to inclusion of covariates remains the same
 - Address the scientific question
 - Predictor of interest
 - Effect modifiers
 - Address confounding
 - Increase precision
 - (There are some additional issues specific to the use of each model, but these will not be addressed here)

Example

.....

Example: Prognosis in Prostate Cancer

- The use of PSA prognostically in hormonally treated prostate cancer
 - 50 men who received some type of hormonal treatment for advanced prostate cancer were followed for disease progression
 - All men were followed at least 24 months
 - But some men had not experienced progression of prostate cancer at the time of data analysis
 - Scientific question:
 - Does the lowest post-treatment value of PSA predict length of time in remission?
 - Is its prognostic value independent of other prognostic variables such as bone scan or Karnofsky score?

Linear Regression Approach

- Compare Nadir PSA between men who did and did not progress within 24 months
 - Response: Nadir PSA
 - Summary measure: Geometric Mean
 - Predictor of interest: Relapse within 24 months
 - Potential confounders:
 - Karnofsky performance status
 - Bone scan score 3 or greater

Linear Regression: Stata Output

```
. regress lnadir relapse24 ps bss3, robust

Number of obs =      48
Root MSE      = 1.6697

lnadir |          Robust
relapse24 | Coef  St Err  t    P>|t|  [95% CI]
ps | -.010  .027  -0.37  0.713  -.065  .045
bss3 | .463  .479  0.97  0.339  -.503  1.429
_cons | -.037  2.323  -0.02  0.987  -4.719  4.645
```

Linear Regression Interpretation

- Comparison of those who relapsed early and those who remained in remission for at least 24 months
 - Geometric mean of nadir PSA is 14.7 times higher in men who relapsed early than men with the same performance status and bone scan score who remained in remission for at least 24 months (95% CI: 4.4 to 48.3 times higher)
 - These results are atypical of what we might expect with no true difference between relapse groups of the same performance status and bone scan score: $P < 0.0005$
 - (Calculations: $e^{2.684} = 14.7$; $e^{1.491} = 4.40$; $e^{3.878} = 48.3$)

Logistic Regression Approach

- Compare odds of relapse within 24 months across groups defined by Nadir PSA
 - Response: Relapse24
 - Summary measure: Odds
 - Predictor of interest: log (NadirPSA)
 - Potential confounders:
 - Karnofsky performance status
 - Bone scan score 3 or greater

Applied Regression Analysis,
June, 2003

16

Logistic Regression: Stata Output

```
. logit relapse24 lnadir ps bss3, robust

Number of obs   =          48
                = Robust
relapse24      |      Coef   St Err   z    P>|z|    [95% CI]
lnadir         |      .876   .316   2.77  0.006   .256  1.495
ps             |     -.054   .038  -1.40  0.160  -1.129 .021
bss3           |      .842   .786   1.07  0.284  -1.699  2.383
_cons         |     2.921   3.188   0.92  0.360  -3.328  9.170
```

- (Note that with the "logistic" command, Stata suppresses the intercept and performs the exponentiation to get the OR per 1 unit difference in the predictor)

Applied Regression Analysis,
June, 2003

17

Logistic Regression Interpretation

- Comparison of odds of relapse within 24 months across groups defined by Nadir PSA
 - Odds of relapse within 24 months is 1.835 times higher for every doubling of the nadir PSA when comparing groups with the same performance status and bone scan score (95% CI: 1.194 to 2.819 times higher)
 - These results are atypical of what we might expect with no true difference between relapse groups of the same performance status and bone scan score: P = 0.006
 - (Calculations: $2^{0.876} = 1.835$; $2^{0.256} = 1.194$; $2^{1.495} = 2.819$)

Applied Regression Analysis,
June, 2003

18

Survival Regression Approach

- Compare instantaneous risk of relapse across groups defined by Nadir PSA
 - Response: Observation time and relapse status
 - Summary measure: Hazard (instantaneous risk)
 - Predictor of interest: log (NadirPSA)
 - Potential confounders:
 - Karnofsky performance status
 - Bone scan score 3 or greater

Survival Regression: Stata Output

```
. cox obstime lnadir ps bss3, dead(relapse) robust

Number of obs = 48
obstime | Robust
relapse | Coef StErr z P>|z| [95% CI]
lnadir | .402 .085 4.70 0.000 .234 .569
ps | -.037 .018 -2.05 0.040 -.072 -.002
bss3 | .738 .411 1.80 0.072 -.067 1.543
```

- (Note that with the "stcox" command, Stata performs the exponentiation to get the HR per 1 unit difference of predictors)

Survival Regression Interpretation

- Comparison of risk of relapse across groups defined by Nadir PSA
 - Risk of relapse is 1.321 times higher for every doubling of the nadir PSA when comparing groups with the same performance status and bone scan score (95% CI: 1.176 to 1.484 times higher)
 - These results are atypical of what we might expect with no true difference between relapse groups of the same performance status and bone scan score: $P < 0.0005$
 - (Calculations: $2^{0.402} = 1.321$; $2^{0.234} = 1.176$; $2^{0.569} = 1.484$)

Comments

- All regression methods are more alike than they are different
 - In general, they all compare the distribution of some response variable across groups defined by the predictor of interest holding adjustment variables constant
 - There are technical differences worthy of greater scrutiny
 - This is particularly true of the proportional hazards model used here for the survival analysis

Comments

- The hard part of every data analysis is deciding which scientific question to answer
 - As a rule, if you state the question precisely enough, the statistical model for analysis is also specified
 - But, as with the PSA example, our scientific goals are often only vaguely specified
 - In this case, we could examine the association between nadir PSA and relapse by several different methods
 - Mean nadir PSA across relapse groups
 - Odds of relapse in a fixed amount of time across PSA groups
 - Risk of relapse at each time across PSA groups
 - The basic idea behind covariate adjustment is the same for all of these models, though the exact interpretation of that adjustment varies