Introduction to R for Epidemiologists

Jenna Krall, PhD

Thursday, March 5, 2015

Outline

1. Interactions in regression
2. Splines
3. Logistic regression
4. Generalized linear models
5. Survival analysis

Interactions

Interaction terms allow the association between the covariate and outcome to differ by a third variable

- Does the association between air pollution and birthweight differ by temperature?
- Does the association between population and murder rate differ by robbery rate?
- Does the association between birthweight and gestational age differ by survival status?

Interactions

Analysis of Covariance (ANCOVA) is the same as linear regression with one categorical covariate and one continuous covariate

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_P x_P$

where

- x_1, x_2, ..., x_{P-1} are indicator variables for whether an observation belongs in that group
  - Indicator variable is 1 if condition is met, 0 otherwise
  - When $x_1 = x_2 = ... = x_P = 0$, indicates reference group
- x_P is the continuous predictor
- β_0 is the intercept, or the $E(y)$ when $x_1 = x_2 = ... = x_P = 0$. In other words, the mean y for the reference group when $x_P = 0$
- β_1 is the slope for x_1, or the difference in $E(y)$ for comparing group 1 to the reference group, controlling for x_P
- $\beta_0 + \beta_1$ is the mean y for group 1, when $x_P = 0$
- β_P is the mean change in y for a one unit change in x_p, controlling for the categorical variable

Logistic regression

Applied when the outcome of interest is binary

- What is the association between smoking and lung cancer?
- Is gestational age associated with survival in very low birthweight infants?

$\text{logit}(E(y|x)) = \beta_0 + \beta_1 x$

- logit is the (natural) log odds, $log(p/(1-p))$
- $E(y)$ is the mean y. Recall that for binary y, the mean of y is simply the proportion of 1's
- β_0 is the log odds of y when $x = 0$
- β_1 is the difference in log odds of y for a one unit change in x
  - The difference in log odds is the same as the log odds ratio
  - β_1 quantifies the association between x and y

Survival analysis in R

For time-to-event data with censoring

- What is the time until AIDS for HIV positive individuals?
- What is the time-to-relapse for smokers?
- Is a new drug associated with time until lung cancer?

Survival data in R

Hunger games survival analysis: Do “career” tributes survive longer?

“which covariates are associated with the odds (or hazard ratios) being ever in your favor?”

- `http://www.bdkeller.com/writing/hunger-games-survival-analysis/`

(*source: Brett Keller*)

Other models

R can also be used to fit more complex models including

- Ordinal logistic regression
- Mixed and random effect models
- Time series models
- Bayesian modeling