Biostatistics 201a - Lecture 26 - 11/28/11

Contents:

- Final Exam announcements
- Model Selection Wrap-up
- Logistic Regression Basics / GLMs

pages = 10

11/28/11

Announcements:

* Final Tu. 3-6 on Dec. 6th

- Study guide, final practice, etc. posted
- Logistic regression handout in handout/course material section of website
- Friday: 8-9 wrap-up lecture
  9-10 review session
- Extra office hours next M, Tu; extra review session next M; watch e-mail for exact times + locations

---

Wrap-up of Model Selection:

- There are automated procedures; most common

  forward stepwise: add variables one at a time until nothing sufficiently improves model

  backwards stepwise: delete variables one at a time until everything left is useful.

Problems w/ these procedures

- "Greedy" - do what is locally best at each step, not what will necessarily produce the best overall model

- Backwards stepwise is affected by multicollinearity and related issues
- forward selection tends to miss things that might be useful in combination, but don't look great individually
- Both have problem that once a variable is in (or out) you can't get it back
  - ⇒ Can be fixed by allowing "mixed stepwise" - have entry steps and exit steps need to be careful not to get stuck in endless loops
- None of these approaches respect the hierarchical principal or contextual clues (e.g. which of two multiple variables makes more sense)

Generally a better idea to do this all manually - that way you can think about what happens at each step

These sorts of procedures are not guaranteed to find the "best" model or even to

all pick the same model, though they usually pick fairly good ones.

The only way to get the "best" fitting model is to do what is called "all subsets" selection - try all possible models. This is a pain.

If you have $K = 10$ possible predictors, there are $2^{10} = 1024$ possible models. This also has a massive multiple testing or overfitting problem - often doesn't produce best model for new data points.

Logistic Regression

So far we've been working with numerical / roughly continuous outcome variables, Y.

What do we do if Y is qualitative? e.g. Yes/no, gets a disease or not, etc.

Logistic regression specifically deals w/ the case when Y is binary - only 2 outcome values. There are variants - e.g. ordinal logistic, multinomial logistic, etc. - for multicategory outcomes.

Let's imagine that

$Y = \begin{cases} 1 & \text{if person gets a disease} \\ 0 & \text{if don't} \end{cases}$

X's are risk factors for disease

What happens if we just try to use MLR?

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

Y: 1 or 0

$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$: need not be 1 or 0

won't give good predictions!

Step 1: Instead of predicting the individual Yes/No or 1/0 we predict $P(Y=1)$ – i.e. how likely the person is to get the disease. This is better in that it's continuous but no guarantee $\beta_0 + \beta_1 X_1 + \cdots$ will be between 0 and 1

Step 2: Transform to the odds scale:

$$\text{Odds}(\text{getting disease}) = \frac{P(Y=1)}{1-P(Y=1)} = \frac{P(Y=1)}{P(Y=0)}$$

This can take on any value between 0 (event can't happen) and ∞ (event always happens)

If $P(Y=1) = 1/2$ – equally likely to happen or not

$$\text{odds} = \frac{1/2}{1-1/2} = 1 \rightarrow \text{equal probability}$$

This is a key cut off: odds = 1

odds > 1 event is more likely than not to happen

odds < 1 event is less likely to happen than not

If $P(Y=1) = 2/3$, $\text{odds} = \frac{2/3}{1/3} = 2$

or 2:1

$P(Y=1) = 1/3$, $\text{odds} = \frac{1/3}{2/3} = 1/2$

We've still got a problem.... which is that $\beta_0 + \beta_1 X_1 + \cdots$ could be negative

Step 3: Take logs

$$\text{logit} = \ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \beta_0 + \beta_1 X_1 + \cdots$$

"called log odds"

modeled as a linear function of X's

the logit or log odds can range from $-\infty$ to ∞

log odds >0 iff odds >1
log odds <0 iff odds <1
log odds $=0$ iff odds $=1$

This implies that if X has a positive ($\beta>0$) coefficient odds and log odds increase as X increases. Similarly if $\beta<0$, as X increases the odds or log odds decrease. If $\beta=0$, X has no effect on the log odds, i.e. the variable is not a significant predictor.

This is just like MLR except

(1) No good form of equation for the estimated regression coefficients, b's. Computer does something called iteratively reweighted least squares to approximate the optimal b's.

(2) Interpretation of the regression coefficients (β, b) must be in terms of log odds, or if we transform back, in terms of odds ratios. Interpretations on probability scale are hard.

(3) Tests, both overall and for individual variables, use different distributions:

Overall test: χ^2 = "chi-squared" test instead of f

Individual tests: Z instead of t.

Logistic regression is a special case of something called the <u>generalized linear model</u>. The idea is that the average value of Y is related to a linear expression in the X's via a <u>link function</u>, g:

$$E(Y) = g^{-1}(\beta_0 + \beta_1 X_1 + \cdots)$$

- Regular MLR uses the "identity link" – basically you don't need a
- Logistic regression uses the logit link...
- Other useful links, e.g. probit, log, etc. for other kinds of data (e.g. counts)