Biostatistics 201A - Lecture 27 - 11/30/11

Contents:

- Logistic regression example

pages: 9

Logistic Regression Example

11/30/11

Outcome $Y = \begin{cases} 1 & \text{mother transmits HIV to baby} \\ 0 & \text{doesn't transmit} \end{cases}$

Predictors: treatment regimen, viral load, age, illness duration, delivery method (C-section or not)

Model:

$$\ln\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_7 X_7 + \varepsilon$$

$p_x = P(Y=1)$ for given X_s = probability of transmission

(1) Overall, is the model useful?

(a) Test:

H_0: $\beta_1 = \beta_2 = \cdots \beta_7 = 0$ - none of tx, viral load, etc., help predict risk of transmission

H_A: At least one $\beta_j \neq 0$ - at least one of the predictors is useful

Instead of f_{obs} our test statistic is χ^2_{obs} - basically same idea.

Large values of χ^2_{obs} tell us the model has explained a lot.

$$p\text{-value} = P(\chi^2_m \geq \chi^2_{obs})$$

m = # predictors

As usual small p-values make us reject.

Example: $m=7$ predictors, $\chi^2_{obs} = 32.47$ ("likelihood ratio chi-squared test" = LR chi2(m)); p-value = 0.000

So we reject H_0 and conclude that one of more of our predictors is useful (surprise!)

(b) Goodness of fit:

- pseudo R^2 - like R^2 but not really a % - more than one way to define it - key is that it's between 0 and 1 w/ higher values being better
- ROC = receiver operating characteristic curve measures how accurately we predict outcome in terms

of sensitivity and specificity
Idea is for any person in data set you can get a predicted probability, p, of (in this case) HIV transmission
Set a thresh-hold for guessing the event will happen - usually

$p > .5$ guess yes
$p < .5$ guess no

but it depends on how important the different types of mistakes are.

Then count how many points in your data set are correctly predicted using this thresh-hold = "error rate". Basically ROC cure looks at error rate as a function of your thresh-hold

---

② Interpreting regression coefficients

b_0 = intercept = log odds (of transmission) when all the X's are 0

b_j = "slope" for variable X_j

= change in log odds associated with a 1 unit change in X_j assuming all else = 1 held fixed. Similar variants for indicators, interactions and transformations

Example: $b_7 = -.4$ - delivery variable

We assume two equivalent moms (same tx, same age, same illness profile) but one has a C-section and one doesn't

P_A = P(mom A transmits virus) ← C-section

P_B = P(mom B " ") ← natural delivery

model: delivery.

$$\ln\left(\frac{P_A}{1-P_A}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_6 X_6 + \beta_7 \cdot 1$$

all equal

$$\ln\left(\frac{P_B}{1-P_B}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_6 X_6 + \beta_7 \cdot 0$$

$$\underbrace{\ln\left(\frac{p_A}{1-p_A}\right) - \ln\left(\frac{p_B}{1-p_B}\right)}_{\text{difference in log odds}} = \beta_7$$

everything else drops out

$$\beta_7 = \ln\left(\frac{\frac{p_A}{1-p_A}}{\frac{p_B}{1-p_B}}\right)$$ properties of log

$$= \ln(\text{"odds ratio"})$$

for mom A vs B

Get the odds ratio by exponentiating:

$$e^{\beta_7} = \frac{\frac{p_A}{1-p_A}}{\frac{p_B}{1-p_B}} = OR_{A \text{ vs } B}$$

Odds ratios $= 1$ → same likelihood of event (i.e. variable had no effect)

Odds ratios > 1 mean A has a higher likelihood of event than B

Odds ratios < 1 mean A has a lower likelihood of the event than B

Example: For delivery $b_7 = -.4$

log odds $< 0 \Rightarrow$ mom w/ C-section has a lower chance of transmitting HIV than the otherwise = mom who has a natural delivery

$$OR_{\substack{\text{C-section} \\ \text{vs not}}} = e^{-.4} = .67 < 1$$

which all implies C-section safer

Specifically mom w/ the C-section is only "2/3rds as likely" or "has odds 2/3rds as high" as the mom w/ natural delivery. Or you can say her odds are 33% (1-.67) lower.

What about a continuous predictor?

b_i = change in log odds

e^{b_i} = change in odds or odds ratio comparing two

people who differ by 1 unit in X_j.

e.g. for Age $b_4 = .1$, $e^{b_4} = 1.10$

1.10 means odds are 1.1 times as big or 10% higher of transmitting HIV for each additional year of age

Y: Whether or not mother transmits HIV to baby (1 = Yes, 0 = No)
X1 = Indicator for whether mother is taking HAART A
X2 = Indicator for whether mother is taking HAART B
X3 = Mother's viral load (log scale)
X4 = Mother's age
X5 = Number of years the mother has been HIV positive
X6 = Number of weeks during the pregnancy the mother was on HAART
X7 = Delivery method (1 = C-section, 0 = normal delivery)

```
. logit HIVplus HAART_A HAART_B VLoad Age YrsHIV WksHAART Delivery
```

Logistic regression

Number of obs	=	300	
LR chi2(7)	=	32.47	
Prob > chi2	=	0.000	
Pseudo R2	=	0.500	

Log likelihood = -26.51722

HIVplus	Coef (= b)	Std. Err. (= s_b)	z	P>|z|	[95% Conf.	Interval]
HAART_A	-0.70	0.250	2.80	0.005	[-1.19,	-0.21]
HAART_B	-1.80	0.300	6.00	0.000	[-2.39,	-1.21]
VLoad	0.00001	0.0000025	4.00	0.000	[.000005,	.000015]
Age	0.10	0.050	2.00	0.046	[0.00,	0.20]
YrsHIV	0.10	0.080	1.25	0.211	[-0.06,	0.26]
WksHAART	-0.05	0.010	-5.00	0.000	[-0.07,	-0.03]
Delivery	-0.40	0.150	-2.67	0.004	[-0.69,	-0.11]
_cons	-5.00	0.500	-10.00	0.000	[-5.98,	-4.02]

```
. logistic HIVplus HAART_A HAART_B VLoad Age YrsHIV WksHAART Delivery
```

Logistic regression

Number of obs	=	300
LR chi2(7)	=	32.47
Prob > chi2	=	0.000
Pseudo R2	=	0.500

Log likelihood = -26.51722

HIVplus	OddsRatio (= e^b)	z	P>|z|	[95% Conf.	Interval] (= $e^{\text{CI for } \beta}$)
HAART_A	0.4966	2.80	0.005	[0.3042,	0.8106]
HAART_B	0.1652	6.00	0.000	[0.0916,	0.2982]
VLoad	1.00001	4.00	0.000	[1.000005,	1.000015]
Age	1.1052	2.00	0.046	[1.0020,	1.2190]
YrsHIV	1.1052	1.25	0.211	[0.9448,	1.2928]
WksHAART	0.9512	-5.00	0.000	[0.9328,	0.9700]
Delivery	0.6703	-2.67	0.004	[0.4996,	0.8994]