Logistic Regression

Jee Dong Jun

12.7.20

Parameters in Logistic Regression

- Y is response variable (0,1) and X is explanatory variable

$$\pi(x) = \exp(\alpha + \beta x)/(\exp(\alpha + \beta x) + 1)$$

- Equivalently, logit is linear

$$\log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x$$

Interpret β

- As x increase by one unit, odd increase by e^{β}
- $\frac{\partial \pi}{\partial x} = \beta \pi (1 - \pi)$
- Effect of α is not of interest.
- But when we use centred data, then it is value at the mean.

Looking at the data

- We use MLE in $\pi(x)$
- We need to check whether logistic regression is appropriate model.
- One way is to look at sample logit or adjusted sample logit and see whether they are linear.

$$log \frac{y_i + 0.5}{n_i - y_i + 0.5}$$

- We can also substitute quartile values of x into $\pi(x)$. Then, we can compare between different explanatory variables.

Logistic Regression with Retrospective Studies

- X rather than Y is random.
- For samples of subjects having Y=1 (case) and Y=0 (control), we observe X.
- Let $\rho_1 = P(Z = 1|y = 1)$, probability of sampling a case and $\rho_2 = P(Z = 1|y = 0)$
- Use Bayes' theorem to calculate $P(Y = 1, z = 1, x)$
- Also, suppose that $P(Z = 1|y, x) = P(Z = 1|y)$, then we can show that $P(Y = 1|z = 1, x)$ also follows logistic model if $P(Y = 1|x)$ follows logistics model.

Inference for logistics regression

- For single predictor

$$logit[\pi(x)] = \alpha + \beta x$$

- Significance test focus on $\beta = 0$
- We can use likelihood ratio test, wald test, score test.
- They all follow asymptotically chi-square 1

Confidence Interval

- From Wald approach, interval $\hat{\beta} + -z(SE)$
- For $\pi(x)$, we approximate by $\hat{\alpha} + \hat{\beta}x_o$
- Large sample SE is given by $var(\alpha) + x^2 var(\beta) + 2xcov(\alpha, \beta)$

Checking Goodness of fit

- Uses a likelihood-ratio test to compare the model to more complex ones.
- At each setting of x, we can calculate fitted value. Then we use Pearson test. (if x is categorical)
- It is important that table is grouped. IF ungrouped then it does not follow chi square

For continuous or ungrouped Data

- We group data or partition X into various spaces.
- The fitted value for 'yes' is sum of the estimated probabilities for all data having X in that category.
- Degree of freedom will be number of partition - number of parameter in logistic regression.
- Hosmer-Lemeshow test

Logistic model with categorical predictor

- Extend to include qualitative explanatory variables.
- $log \frac{\pi_i}{1-\pi_i} = \alpha + \beta_i$
- We can recode such that $\beta_I = 0$
- THe model has any many parametrs s observation.
- When factor has no effect, X and Y are independent.

Indicator Variables

- Use one hot encoding then this corresponds to the constraint $\beta_I = 0$
- Another use encoding suc that $\sum \beta_i = 0$
- Individual β_i is not important.
- Depending on coding individaul value might change, but model fit does not change.

For ordinal variable

- Above model does not take account of order.
- If ther is score $(x_1, x_2, .., x_I)$
- If there is order and we expect a monotone effect,

$$log \frac{\pi_i}{1 - \pi_i} = \alpha + \beta x_i$$

Cochran-Armitage Trend test

- Consider linear porbablity model $\pi_i = \alpha + \beta x_i$
- We fit this value using ordinary least square.
- Pearson statstic $X^2 I)$ can be decomposed into $z^2 + X^2(L)$
- $X^{@}(L)$ is asymptotically chi squared if linear model holds.
- z^2 is test for $\beta = 0$ when linear model holds. (Cochran-Armitage test)
- This test is equivalent to the score statistc in linear logit model.

Using directed models

- Partition G^2 into J-1 component (When 2 x j case)
- jth component where the first column combines column 1 j and second column is j+1
- In general, J-1 partition to IxJ table
- df for the subtable must sum to df for the full table.
- Each cell count in the full table must be in one and only one
- each marginal total of the full table must be marginal total for one and only one subtable

Multiple Logistic Regression

- Without ordinal assumption, we calculate ordinary G^2 and X^2
- $G^2(I|L) = G^2(I) - G^2(L)$ is likelihood-ratio statistic comparing the linear logit model and the independence model.
- Most of analysis can directly extend to multiple logistic regression.
- Instatantaneous rate of change for x_j is $\beta_j\pi(1-\pi)$ adjusting for other variable