Live Session - Week 3: Discrete Response Models Lecture 2

Devesh Tiwari

5/30/2017

Agenda

1. Brief review (10 mins)
2. Breakout Session: Odds ratio and interpreting coefficients (30 mins)
3. Breakout Session: Predicted values and confidence intervals (30 mins)
4. Breakout Session and conclusion: Interaction effects (remaining time)

Announcments

Lab 1 will be posted this week and it is due on June 18th by 11:59 PM PT.

This lecture begins the study of logistic regression models, the most important special case of the generalized linear models (GLMs). It begins with a discussion of why classical linear regression models is not appropriate, from both statistical sense and practical application sense, to model categorical respone variable.

This week

Topics covered

- Variable transformation: interactions among explanatory variables
- Variable transformation: quadratic term
- Categorical explanatory variables
- Odds ratio in the context of categorical explanatory variables
- Convergence criteria and complete separation

Please make sure that you are very familiar with the concepts and techniques coverd in this and last lecture, as they will be used again in the next two lectures in situations that are more general (from two categorical to $J > 2$ categories and from unordered cateogrical variables to ordinal variables). Especially in multinomial logistic regression models, the notions will be much heavier.

Required Readings:

BL2015: Christopher R. Bilder and Thomas M. Loughin. Analysis of Categorical Data with R. CRC Press. 2015.

- Ch. 2.2.5 – 2.2.7, 2.3

Breakout Session : Interpreting coefficients (20 minutes in breakout groups + 10 minutes group discussion)

Let's return to the data and models we ran last week (see below). Imagine that you were reporting these results in a report or paper. With that in mind:

(a) Interpret the coefficients for *k5* and *age* using Odds Ratios

(b) Calculate the 95 % Wald - interval for your interpretations above.

(c) Calculate the 95 % Profile LR intervals for your interpretations above. Are they the same? Why or why not?

```
rm(list = ls())
library(car)
require(dplyr)
library(Hmisc)
library(stargazer)

mroz.glm <- glm(lfp ~ k5 + k618 + age + wc + hc + lwg + inc,
                family = 'binomial', data = Mroz)

summary(mroz.glm)
```

```
##
## Call:
## glm(formula = lfp ~ k5 + k618 + age + wc + hc + lwg + inc, family = "binomial",
##     data = Mroz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1062  -1.0900   0.5978   0.9709   2.1893
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.182140   0.644375   4.938 7.88e-07 ***
## k5          -1.462913   0.197001  -7.426 1.12e-13 ***
## k618        -0.064571   0.068001  -0.950 0.342337
## age         -0.062871   0.012783  -4.918 8.73e-07 ***
## wcyes        0.807274   0.229980   3.510 0.000448 ***
## hcyes        0.111734   0.206040   0.542 0.587618
## lwg          0.604693   0.150818   4.009 6.09e-05 ***
## inc         -0.034446   0.008208  -4.196 2.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.75  on 752  degrees of freedom
## Residual deviance:  905.27  on 745  degrees of freedom
## AIC: 921.27
##
## Number of Fisher Scoring iterations: 4
```

Breakout Session 2: Predicted values and confidence intervals (20 minutes in breakout groups + 10 minutes group discussion)

This is an extension of of the examples from last week; we are going to generate some predicted values and their associated confience intervals. We can generate Wald-intervals and profile-LR intervals, though for this exercise we are going to generate Wald-intervals using the *predict.glm* function. The *predict.glm* function can return predicted values in terms of the log-odds (type = "link") and in terms of the predicted probability of an event occuring (type = "response"). *predict.glm* does not calculate confidence intervals, it calculates the predicted value's confidence interval instead (se.fit = TRUE). We are going to compare and contrast two ways to calculate predicted values and their confidence intervals: The wrong way and the right way.

(a) Calculate the predicted probability a woman participates in the labor force conditional on her having no kids, is 20 years old, did not go to college, her husband did not go to college, her logged wages is 1.0971, and family income excluding her own is 20. Calculate this the "wrong" way first by examining the predicted probability only (type = "response") and then calculate this again by transforming the log-odds (type = "link").

(b) Repeat this exercise, but for a woman with 4 kids under the age of 5 and who is 60 years old. Are there any women in the dataset who resemble this profile?

(c) Comment on your output. Why do you think the "wrong way" is actually wrong?

(d) If you have time, modify this code so you can graph the predicted probability a woman participates in the labor force by varying age between 20 and 60. Include the confidence intervals in your plot.

Breakout Session 3: Interaction effects (15 minutes in breakout groups, 10 minutes group discussion)

Suppose we are interested in understanding if the relationship between college education and labor force participation is conditional on age (alternatively, we could say that we are interested in understanding if the relationship between age and labor force participation is different for women who went to college versus those who did not). In order to test this hypothesis, we need to add an interaction term to the model (see below).

(a) Interpret the impact of college education on labor force participation for the null model, and then again for the model that includes the interaction term (do not worry about the confidence intervals). Based on these results of the newer model, do you think that women who attend college are more likely to participate in the labor force?

(b) Formally test the hypothesis that the interaction effect between age and college education is zero (that is, test the null hypothesis that the relationship between college education and labor force participation is NOT conditional on age).

(c) Comment on your findings.

```
mroz.interact.glm <- glm(lfp ~ k5 + k618 + age + wc + hc + lwg + inc + wc:age,
                         family = 'binomial', data = Mroz)

# Interpret impact of college
stargazer(mroz.glm, mroz.interact.glm, type = "text", summary = FALSE)
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                 lfp
```

```
##                         (1)            (2)
## ---------------------------------------------
## k5                   -1.463***      -1.437***
##                       (0.197)        (0.198)
##
## k618                   -0.065         -0.069
##                       (0.068)        (0.068)
##
## age                  -0.063***      -0.068***
##                       (0.013)        (0.014)
##
## wcyes                 0.807***       -0.141
##                       (0.230)        (1.010)
##
## hcyes                  0.112          0.096
##                       (0.206)        (0.207)
##
## lwg                   0.605***       0.611***
##                       (0.151)        (0.151)
##
## inc                  -0.034***      -0.034***
##                       (0.008)        (0.008)
##
## age:wcyes                             0.022
##                                      (0.023)
##
## Constant              3.182***       3.417***
##                       (0.644)        (0.690)
##
## ---------------------------------------------
## Observations            753            753
## Log Likelihood       -452.633       -452.171
## Akaike Inf. Crit.    921.266        922.342
## =============================================
## Note:              *p<0.1; **p<0.05; ***p<0.01
```