Live Session - Week 3: Discrete Response Models Lecture 2

Devesh Tiwari and Jeffrey Yau

Sept 19, 2017

Agenda

Announcments

Lab 1 has been posted!!

It is due on Sunday October 1, by 11:59 PM PT.

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)

I printed output from two models based on the data we examined last week. In your breakout sessions, answer the following questions

(a) Describe, in words, the difference between the two models. What is the second model testing?

(b) Interpret the coefficeints for *k5* and *age* using odds ratios, in both models.

(c) Calculate the 95 % Wald - interval for your interpretations above [Take home].

(d) Calculate the 95 % Profile LR intervals for your interpretations above. Are they the same? Why or why not [Take Home]?

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

Mroz$totalKids <- Mroz$k5 + Mroz$k618

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

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


stargazer(mroz.glm, mroz.interact.glm, type = 'text')
```

```
##
## ===============================================
##                       Dependent variable:
##                   -----------------------------
##                               lfp
##                        (1)           (2)
## -----------------------------------------------
## totalKids            -0.186***     -0.189***
##                      (0.063)       (0.063)
##
## age                  -0.035***     -0.047***
##                      (0.012)       (0.013)
##
## wcyes                 0.643***     -1.236
##                      (0.217)       (0.953)
##
## hcyes                 0.035         0.005
##                      (0.197)       (0.199)
##
## lwg                   0.581***      0.589***
##                      (0.146)       (0.146)
##
## inc                  -0.031***     -0.031***
```

```
##                        (0.008)        (0.008)
##
## age:wcyes                             0.045**
##                                       (0.022)
##
## Constant               1.883***      2.406***
##                        (0.584)        (0.642)
##
## -----------------------------------------------
## Observations             753            753
## Log Likelihood         -481.385      -479.341
## Akaike Inf. Crit.      976.771        974.681
## ===============================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Group Discussion 1: Choosing among models

(1) Based on the discussion thus far, which model do you prefer? Why? What does it mean for one model to be "better" than another?

(2) What is the residual devieance of a model? Could you use that information from a model to decide which model is "better" than another?

Demo: Assessing the explanatory power of different models

```r
# If the models are nested, we can use the Anova function
anova(mroz.glm, mroz.interact.glm, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: lfp ~ totalKids + age + wc + hc + lwg + inc
## Model 2: lfp ~ totalKids + age + wc + hc + lwg + inc + wc:age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       746     962.77
## 2       745     958.68  1   4.0898  0.04314 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# If the models are not nested, we can compare the AIC values
AIC(mroz.glm, mroz.interact.glm)
```

```
##                   df      AIC
## mroz.glm           7 976.7709
## mroz.interact.glm  8 974.6811
```

Group Disucssion 2: Predicted probabilities

It is really important to be able to graphically present the relationship between changes in covariates of interest and the predicted probability that a given event occurs (the dependent variable). A graphical presentation will help you assess the practical significance of your model results, mainly because you are forced to show how practically significant changes in X impact your dependent variable.

I am going to generate confidence intervals using the Wald-standard errors as generated from 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.

Take a look at the plots and code below, what is wrong with the "wrong" way of producing CI using the predict.glm function? After taking a look at the predicted probability chart, do you think that this is a real problem or are the instructors just being unneccsarily picky?

Now, re-run the following code in order to generate predicted probability charts for women who have 4 children under the age of 5. What do you notice?

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

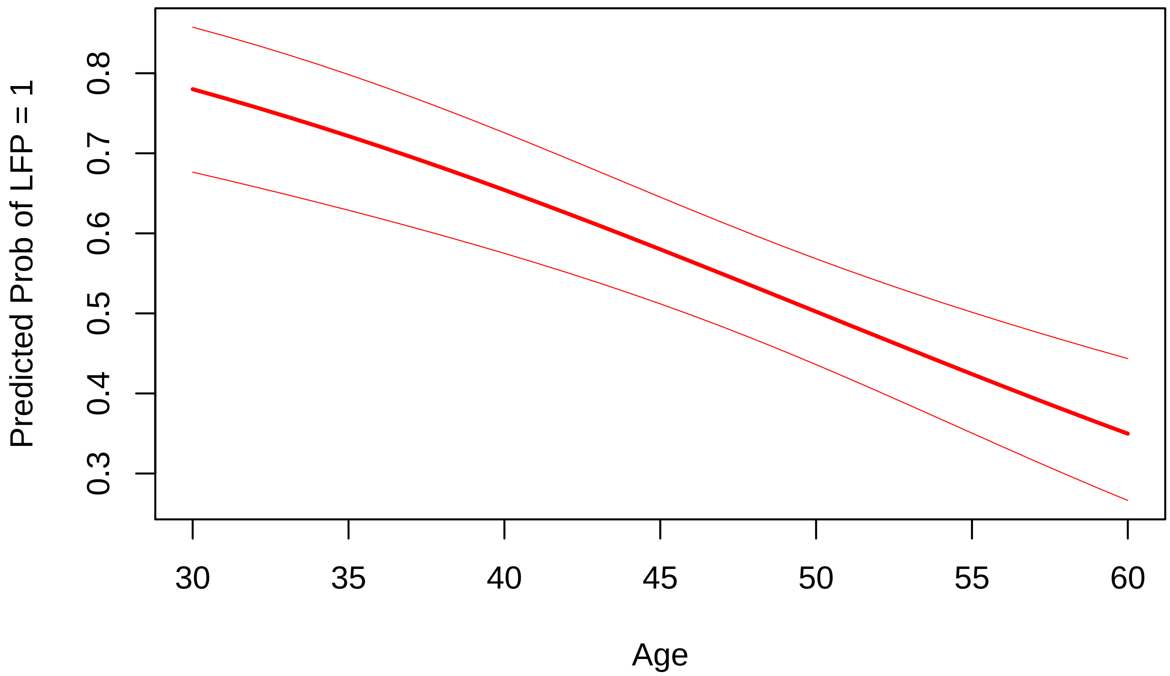
```
##
## Call:
## glm(formula = lfp ~ k5 + k618 + age + wc + hc + lwg + inc, family = "binomial",
##     data = Mroz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1062  -1.0900   0.5978   0.9709   2.1893
##
## Coefficients:
##              Estimate Std. Error z value 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
```

```
## The right way to do it

newdf <- data.frame(k5 = 0,
                    k618 = 0,
                    age = seq(from = 30, to = 60, by = 1),
                    wc = 'no',
                    hc = 'no',
                    lwg = mean(Mroz$lwg),
```

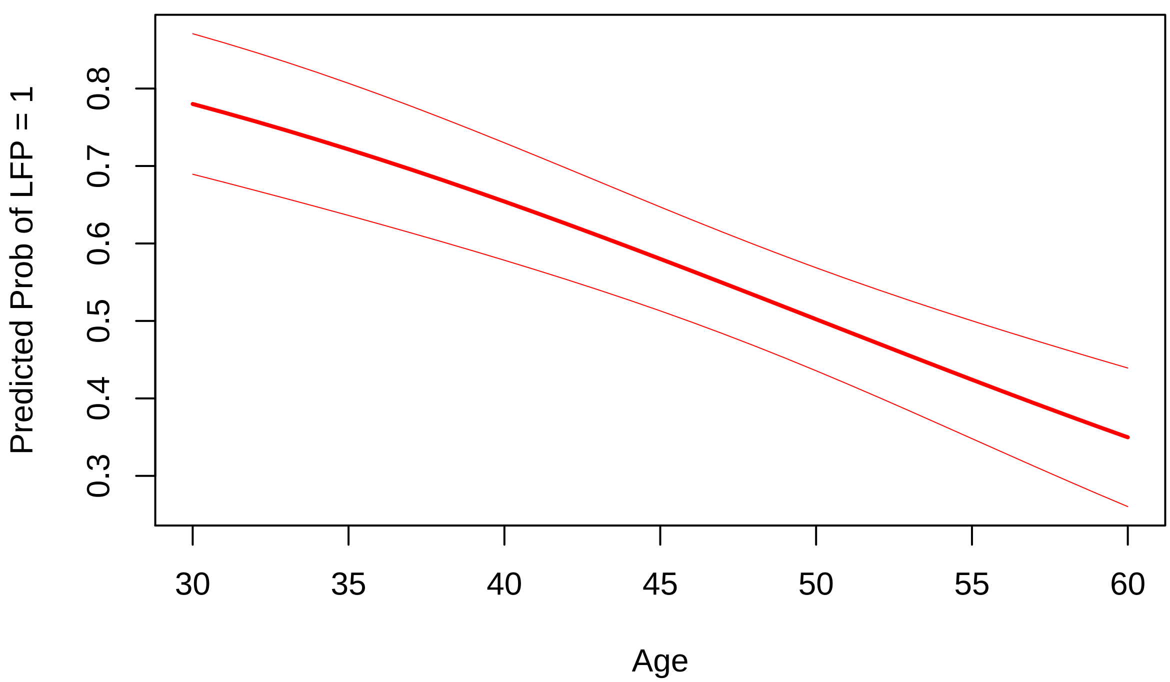
```
                    inc = mean(Mroz$inc))

lp.hat <- predict.glm(mroz.old.glm, newdata = newdf, type = "link", se.fit = TRUE)
head(lp.hat)
```

```
## $fit
##            1            2            3            4            5
##  1.266070708  1.203200157  1.140329606  1.077459055  1.014588503
##            6            7            8            9           10
##  0.951717952  0.888847401  0.825976850  0.763106299  0.700235748
##           11           12           13           14           15
##  0.637365196  0.574494645  0.511624094  0.448753543  0.385882992
##           16           17           18           19           20
##  0.323012441  0.260141889  0.197271338  0.134400787  0.071530236
##           21           22           23           24           25
##  0.008659685 -0.054210867 -0.117081418 -0.179951969 -0.242822520
##           26           27           28           29           30
## -0.305693071 -0.368563622 -0.431434174 -0.494304725 -0.557175276
##           31
## -0.620045827
##
## $se.fit
##         1         2         3         4         5         6         7
## 0.2696810 0.2586632 0.2478150 0.2371596 0.2267244 0.2165410 0.2066468
##         8         9        10        11        12        13        14
## 0.1970852 0.1879072 0.1791715 0.1709461 0.1633081 0.1563435 0.1501462
##        15        16        17        18        19        20        21
## 0.1448146 0.1404475 0.1371369 0.1349606 0.1339740 0.1342032 0.1356421
##        22        23        24        25        26        27        28
## 0.1382530 0.1419712 0.1467124 0.1523814 0.1588787 0.1661073 0.1739759
##        29        30        31
## 0.1824018 0.1913114 0.2006402
##
## $residual.scale
## [1] 1
```

```
lp.hat.mean <- lp.hat$fit
lp.hat.lci <- lp.hat$fit - 1.96 * lp.hat$se.fit
lp.hat.uci <- lp.hat$fit + 1.96 * lp.hat$se.fit

pi.hat <- exp(lp.hat.mean) / (1 + exp(lp.hat.mean))
pi.hat.lci <- exp(lp.hat.lci) / (1 + exp(lp.hat.lci))
pi.hat.uci <- exp(lp.hat.uci) / (1 + exp(lp.hat.uci))

### Plot predicted probabilities
age <- newdf$age
plot(age, pi.hat, ylim = range(c(pi.hat.lci, pi.hat.uci)),
     xlab = "Age", ylab = "Predicted Prob of LFP = 1", type = 'l', col = 'red', lwd = 2)
lines(age, pi.hat.lci, col = 'red', lwd = 0.5)
lines(age, pi.hat.uci, col = 'red', lwd = 0.5)
```




```r
#### The wrong way
pi.hat.response <- predict.glm(mroz.old.glm, newdf, type = "response", se.fit = TRUE)
pi.hat.response.lci <- pi.hat.response$fit - 1.96*pi.hat.response$se.fit
pi.hat.response.uci <- pi.hat.response$fit + 1.96*pi.hat.response$se.fit

plot(age, pi.hat.response$fit, ylim = range(c(pi.hat.response.lci, pi.hat.response.uci)),
     xlab = "Age", ylab = "Predicted Prob of LFP = 1", type = 'l', col = 'red', lwd = 2)
lines(age, pi.hat.response.lci, col = 'red', lwd = 0.5)
lines(age, pi.hat.response.uci, col = 'red', lwd = 0.5)
```



Take home exercise

Create predicted probability charts for the following models *mroz.glm* and *mroz.interact.glm* in order to determine whether or not you think that the interaction term is neccessary.