Live Session - Week 2: Discrete Response Models Lecture 2

Devesh Tiwari

5/23/2017

Agenda

1. Q&A (estimated time: 5 minutes)
2. An overview of this lecture and live session (estimated time: 15 minutes)
3. An extended example (estimated time: 65 minutes)
4. More take-home exercises (no need to turn them in, but we will ask volunteer to present their work in the next live session.)

1. Questions?

2. An Overivew of the Lecture (estimated time: 10 minutes)

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.

Topics covered in this lecture include

- An introduction to binary response models and linear probability model, covering the formulation of forme and its advantages limitations of the latter
- Binomial logistic regression model
- The logit transformation and the logistic curve
- Statistical assumption of binomial logistic regression model
- Maximum likelihood estimation of the parameters and an overview of a numerical procedure used in practice
- Variance-Covariance matrix of the estimators
- Hypothesis tests for the binomial logistic regression model parameters
- The notion of deviance and odds ratios in the context of logistic regression models
- Probability of success and the corresponding confidence intervals in the context of logistic regression models
- Common non-linear transformation used in the context of binary dependent variable
- Visual assessment of the logistic regression model
- R functions for *binomial distribution*

Recap some notations:

Recall that the probability mass function of the Binomial random variable is

$$P(W_j = w_j) = \binom{n_j}{w_j} \pi_j^{w_j} (1 - \pi_j)^{n_j - w_j}$$

where $w_j = 0, 1, \ldots, n_j$ where $j = 1, 2$

- the *link function* translates from the scale of mean response to the scale of linear predictor.

- The linear predicator can be expressed as

$$\eta(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- With $\mu(\mathbf{x}) = E(y|\mathbf{x})$ being the conditional mean of the response, we have in GLM

$$g(\mu(\mathbf{x})) = \eta(\mu(\mathbf{x}))$$

where $g()$ denotes some non-linear transformation. In the logit case, $g() = log_e(\frac{\mu}{1-\mu})$.

To estimate the parameters of a GLM model, MLE is used. Because there is generally no closed-form solution, numerical procedures are needed. In the case of GLM, the *iteratively weighted least squares* procedure is used.

3. An extended example (estimated time: 65 minutes)

Insert the function to *tidy up* the code when they are printed out

```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

Instructor's introduction to the example (estimated time: 5 minutes)

When solving data science problems, always begin with the understanding of the underlying question; our first step is typically **NOT** to jump right into the data. For the sake of this example, suppose the question is *"Do females who higher family income (excluding wife's income) have lower labor force participation rate?" If so, what is the magnitude of the effect?* Note that this was not Mroz (1987)'s objective of his paper. For the sake of learning to use logistic regression in answering a specific question, we stick with this question in this example.

Understanding the sample: Remember that this sample comes from *1976 Panel Data of Income Dynamics (PSID)*. PSID is one of the most popular dataset used by economists.

Breakout Session 1: EDA. Time: 10 mins in groups. 5 mins discussion

Take a look at the dataset called *Mroz*, which is located in the *car* package in R. You can find a description of the variables in this dataset by typing ?Mroz in the R-editor. Answer the following questions about the EDA portion of the modelling process. Wherever possible, conduct a brief EDA on this dataset when answering each question; but more importantly, think about the questions an effective EDA should answer and how you would modify your modeling strategy based on those answers. Remember, the dependent variable here is dichotomous!

(1) What questions about the data are you trying to answer when you examine univariate plots? What are you looking for?

(2) What questions about the data are you trying to answer when you examine bivariate plots (between the dependent variable of interest and the independent variable and also between independent variables of interest)? What are you looking for?

(3) What are interaction effects and how could you use EDA to explore whether they exist?

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

```
## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     combine, src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```r
`?`(Mroz)
describe(Mroz)
```

```
## Mroz
##
##  8  Variables      753  Observations
## ---------------------------------------------------------------------------
## lfp
##        n  missing distinct
##      753        0        2
##
## Value         no   yes
## Frequency    325   428
## Proportion 0.432 0.568
## ---------------------------------------------------------------------------
## k5
##        n  missing distinct     Info     Mean      Gmd
##      753        0        4    0.475   0.2377   0.3967
##
## Value          0     1     2     3
## Frequency    606   118    26     3
## Proportion 0.805 0.157 0.035 0.004
## ---------------------------------------------------------------------------
## k618
##        n  missing distinct     Info     Mean      Gmd
##      753        0        9    0.932    1.353     1.42
##
## Value          0     1     2     3     4     5     6     7     8
## Frequency    258   185   162   103    30    12     1     1     1
## Proportion 0.343 0.246 0.215 0.137 0.040 0.016 0.001 0.001 0.001
## ---------------------------------------------------------------------------
## age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      753        0       31    0.999    42.54    9.289     30.6     32.0
##      .25      .50      .75      .90      .95
##     36.0     43.0     49.0     54.0     56.0
```

```
##
## lowest : 30 31 32 33 34, highest: 56 57 58 59 60
## ---------------------------------------------------------------------------
## wc
##        n  missing distinct
##      753        0        2
##
## Value         no   yes
## Frequency    541   212
## Proportion 0.718 0.282
## ---------------------------------------------------------------------------
## hc
##        n  missing distinct
##      753        0        2
##
## Value         no   yes
## Frequency    458   295
## Proportion 0.608 0.392
## ---------------------------------------------------------------------------
## lwg
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      753        0      676        1    1.097   0.6151   0.2166   0.4984
##      .25      .50      .75      .90      .95
##   0.8181   1.0684   1.3997   1.7600   2.0753
##
## lowest : -2.054124 -1.822531 -1.766441 -1.543298 -1.029619
## highest:  2.905078  3.064725  3.113515  3.155581  3.218876
## ---------------------------------------------------------------------------
## inc
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      753        0      621        1    20.13    11.55    7.048    9.026
##      .25      .50      .75      .90      .95
##   13.025   17.700   24.466   32.697   40.920
##
## lowest : -0.029  1.200  1.500  2.134  2.200, highest: 77.000 79.800 88.000 91.000 96.000
## ---------------------------------------------------------------------------
```

```
# INSERT CODE HERE
```

Breakout Session 2: Comparing a linear model with a logit model. Time: 20 minutes (in groups) and 10 minutes discussion

In this exercise, we are going to examine the relationship between the dependent variable, *lfp*, and the remaining covariates via the CLM and logistic regression. Please follow the steps below as described:

(1) I build a linear model in the code below. Interpret the impact of the variable *k5* on *lpv*. Pay attention to the distribution of *k5*, what it stands for, and what the coefficient itself tells us.

```
mroz.lm <- lm(as.numeric(lfp) ~ k5 + k618 + age + wc + hc + lwg +
    inc, data = Mroz)
summary(mroz.lm)
```

```
##
## Call:
## lm(formula = as.numeric(lfp) ~ k5 + k618 + age + wc + hc + lwg +
```

```
##     inc, data = Mroz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9268 -0.4632  0.1684  0.3906  0.9602
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.143548   0.127053  16.871  < 2e-16 ***
## k5          -0.294836   0.035903  -8.212 9.58e-16 ***
## k618        -0.011215   0.013963  -0.803 0.422109
## age         -0.012741   0.002538  -5.021 6.45e-07 ***
## wcyes        0.163679   0.045828   3.572 0.000378 ***
## hcyes        0.018951   0.042533   0.446 0.656044
## lwg          0.122740   0.030191   4.065 5.31e-05 ***
## inc         -0.006760   0.001571  -4.304 1.90e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.459 on 745 degrees of freedom
## Multiple R-squared:  0.1503, Adjusted R-squared:  0.1423
## F-statistic: 18.83 on 7 and 745 DF,  p-value: < 2.2e-16
```

```
# INSERT CODE BELOW
```

(2) Using the GLM command, build a logistic model with the same covariates as above. Once again, interpret the impact of the variable *k5* (but don't spend too much time on it, as we will be discussing interpretation in the next breakout section!)

(3) Let's visually examine the relationsip between age and lfp for both the CLM and logistic model across two scenarios: One where *k5* equals zero and another when it equals three. In order to do this, we will need to use the predict.lm and the predict.glm functions in R. Take a minute to look at the documentations, but these two functions use our model results to generate predicted values on values specified by the user (see my code below on how to do that).

All told, you will generate 4 sets of predicted values, two for the clm model and two for the logit model. Plot all four of these predicted values against age (you don't have to do it all in a single plot, for now do what is easiset for you).

For this exercise, do not worry about the confidence intervals — we will tackle those next week.

Examine the plots and note anything that looks interesting or note-worthy. We will talk about this togther.

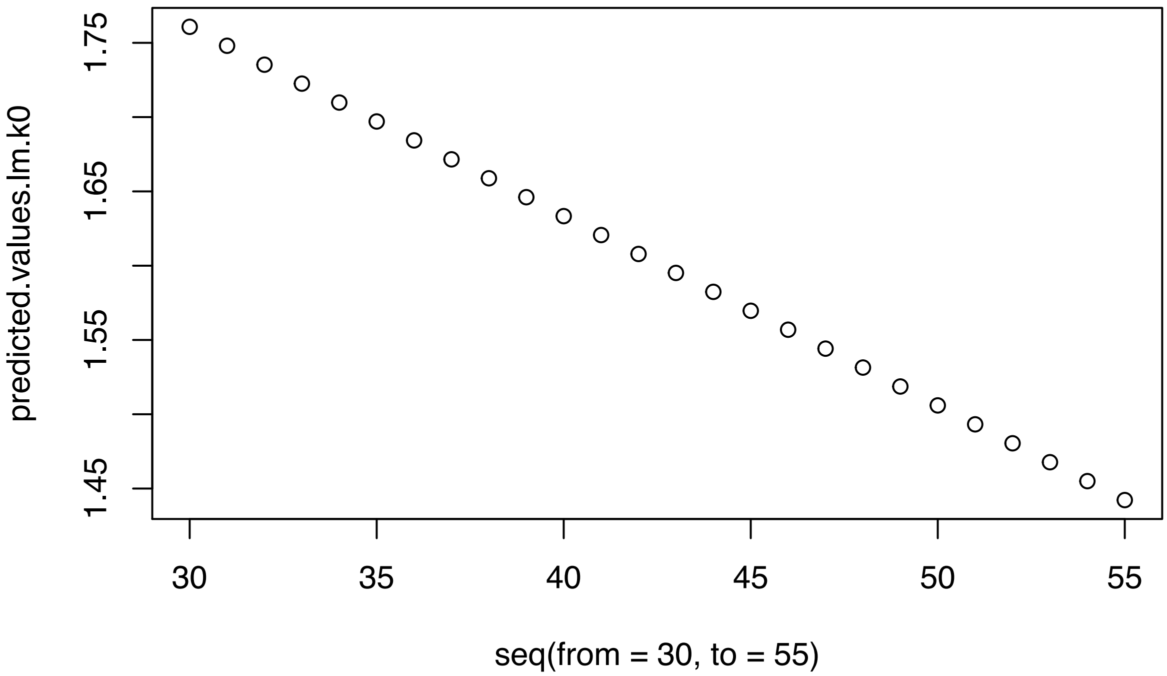
```
# Create the new df that will be used by the predict
# functions.  You will use this df for both the predict.lm
# and predict.glm functions

newdf <- data.frame(k5 = 0, k618 = 0, age = seq(from = 30, to = 55),
    wc = "no", hc = "no", lwg = 1.0971, inc = 20)
predicted.values.lm.k0 <- predict.lm(mroz.lm, newdata = newdf,
    se.fit = FALSE)
# predicted.values.glm.k0 <- predict.glm(FILL IN THE COMMAND
# HERE)

## Create two more predicted values charts (one for the clm
## and the other for the logit) but this time, set k5 to 3.
```

```r
# INSERT YOUR CODE

# Plots. Generate three more, one for each
plot(x = seq(from = 30, to = 55), predicted.values.lm.k0)
```




```r
# INSERT YOUR CODE
```

Breakout Session 3: Brief exercise on testing. Time: 20 minutes (in groups) and 10 minutes discussion.

Test the hypothesis that age makes no impact on *lfp* using both the Wald test and the Likelihood Ratio Test. In words, what is the point of each test and what do they tell you? HINT: For the LRT test, use the Anova function and use "LR" for the test option.

Breakout Session 4: Odds-ratio and interpretation. Time: Rest of class

Interpret the impact of *k5* on the dependent variable and the impact of *age* on the dependent variable. First, state your interpretation in terms of an odds-ratio (or log-odds ratio) and second in terms of predicted probability. What do you notice about stating your interpretation in terms of the predicted probability?

Take-home exercises

1. Use the model *mroz.glm* and test the hypothesis the hypothesis the wife's wage had no impact on her labor force participation. Set up the test. Write down the null hypothesis. Explain which test(s) you used. State the results. Explain the results.
2. Explain all of the deviance statistics in the model results (*summary(mroz.glm)*) and what do they tell us? (You answer may require you to perform further calculation using the deviance statistics.)
3. Expand the EDA and propose one additional specification based on your EDA.

4. Test this newly proposed model, call it mroz.glm2, and test the difference between the two models.
5. Study the model parameter estiamtion algorithm: Iterated Reweighted Least Square (IRLS) Reference: linked phrase