w271: Homework 3 (Due: Week 4)

Professor Jeffrey Yau

Due: 4pm Pacific Time on the Day of the Live Session of Week 4

Instructions (Please Read it Carefully!):

- **Page limit of the pdf report: None, but please be reasonable**
- Page setup:
  - Use the following font size, margin, and linespace:
    - fontsize=11pt
    - margin=1in
    - line_spacing=single
- Submission:
  - Each student submits his/her homework to the course github repo by the deadline; submission and revision made after the deadline will not be graded
  - Submit 2 files:
    1. A pdf file that details your answers. Include all the R codes used to produce the answers. *Please do not suppress the codes in your pdf file.*
    2. R markdown file used to produce the pdf file
  - Use the following file-naming convensation; fail to do so will receive 10% reduction in the grade:
    - StudentFirstNameLastName_HWNumber.fileExtension
    - For example, if the student's name is Kyle Cartman for homework 1, name your files as
      - KyleCartman_HW1.Rmd
      - KyleCartman_HW1.pdf
  - Although it sounds obvious, please print your name on page 1 of your pdf and Rmd files.
  - For statistical methods that we cover in this course, use only the R libraries and functions that are covered in this course. If you use libraries and functions for statistical modeling that we have not covered, you have to (1) provide an explanation of why such libraries and functions are used instead and (2) reference to the library documentation. **Lacking the explanation and reference to the documentation will result in a score of zero for the corresponding question.** For data wrangling and data visualization, you are free to use other libraries, such as dplyr, ggplot2, etc.
  - For mathematical formulae, type them in your R markdown file. **Do not write them on a piece of paper, take a photo, and either insert the image file or sumbit the image file separately. Doing so will receive a 0 for the whole question.**

– Students are expected to act with regards to UC Berkeley Academic Integrity.

In the live session of week 3, we discussed various ways of variable transformation. In this lab, you will practice using some of the variable transformation techniques and the concepts and techniques of applying a binary logistic regression covered in the first three weeks. This lab uses the `Mroz` data set that comes with the *car* library. We examine this dataset in one of our live sessions.

Some start-up scripts

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

# Describe the structure of the data, such as the number of
# observations, the number of variables, the variable names,
# and type of each of the variables, and a few observations of each of
# the variables
str(Mroz)
```

```
## 'data.frame':    753 obs. of  8 variables:
##  $ lfp : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ k5  : int  1 0 1 0 1 0 0 0 0 0 ...
##  $ k618: int  0 2 3 3 2 0 2 0 2 2 ...
##  $ age : int  32 30 35 34 31 54 37 54 48 39 ...
##  $ wc  : Factor w/ 2 levels "no","yes": 1 1 1 1 2 1 2 1 1 1 ...
##  $ hc  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ lwg : num  1.2102 0.3285 1.5141 0.0921 1.5243 ...
##  $ inc : num  10.9 19.5 12 6.8 20.1 ...
```

```r
# Provide summary statistics of each of the variables
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
## ---------------------------------------------------------------------------
```

```
# For datasets coming with a R library, we can put "?" in front of a
# dataset to display, under the help window, the description of the
# datasets
?Mroz
```

Question 1:

Estimate a binary logistic regression with `lfp`, which is a binary variable recoding the participation of the females in the sample, as the dependent variable. The set of explanatory variables includes `age`, `inc`, `wc`, `hc`, `lwg`, `totalKids`, and a quadratic term of `age`, called `age_squared`, where `totalKids` is the total number of children up to age 18 and is equal to the sum of `k5` and `k618`.

Answer: We first create a new varirables, such as the total number of kids and the quadratic term of age. Then, we estimate a binary logistic regression using the `glm()` function and display the estimation result.

```
# Create new explanatory variables

# Total number of kids
Mroz['totalKids'] <- Mroz$k5 + Mroz$k618
# Quadratic term of age (i.e. age squared)
Mroz['age_squared'] <- Mroz$age^2

# Estimate a bineary logistic regression with the variables specified in the questions
mroz.glm1 <- glm(lfp ~ age + age_squared + inc + wc + lwg + totalKids, family = 'binomial', da
# Note that another way to include a quadratic term is to include the transformation in the gl
#glm(lfp ~ age + I(age^2) + inc + wc + lwg + totalKids, family = 'binomial', data = Mroz)

# Display the estimation results
summary(mroz.glm1)
```

```
##
## Call:
## glm(formula = lfp ~ age + age_squared + inc + wc + lwg + totalKids,
##     family = "binomial", data = Mroz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8303  -1.1694   0.6764   1.0073   2.0829
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -5.150511   2.260965  -2.278 0.022726 *
## age          0.311895   0.108654   2.871 0.004098 **
## age_squared -0.004051   0.001265  -3.203 0.001359 **
## inc         -0.033435   0.007555  -4.425 9.63e-06 ***
## wcyes        0.713378   0.196114   3.638 0.000275 ***
## lwg          0.550747   0.145446   3.787 0.000153 ***
## totalKids   -0.221626   0.063799  -3.474 0.000513 ***
## ---
## 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:  952.27  on 746  degrees of freedom
## AIC: 966.27
##
## Number of Fisher Scoring iterations: 4
```

Question 2:

Is the age effect statistically significant?

Answer: To test the statistical significance of the age effect, we will apply LRT using R's `anova()` function, and to do so, we will estimate a "restricted" model with the age variables, which include both `age` and `age_squared` in the "full" model. We will call the restricted model `mroz.glm2`. Note also that because age is entered the logistic regression as a quadratic function, testing the statistical significance of the age effect include testing multiple hypotheses.

The model being estimated, surpressing the subscript for individuals, is

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 age + \beta_2 age\_squared + \beta_3 inc + \beta_4 wc + \beta_5 lwg + \beta_6 totalKids$$

where π denotes the probability that a female participating in the labor force. That is, $P(lfp_i = 1)$

$$H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0 H_1 : (\beta_1 \neq 0 \text{ and } \beta_2 = 0), \text{ or } (\beta_1 = 0 \text{ and } \beta_2 \neq 0), \text{ or } (\beta_1 \neq 0 \text{ and } \beta_2 \neq 0)$$

Note: I just explicitly write out all the alternative hypotheses. In most case, the following expression is being used

$$H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0 H_1 : H_0 \text{ is not true}$$

```r
mroz.glm2 <- glm(lfp ~ inc + wc + lwg + totalKids, family = 'binomial', data = Mroz)

# Display both Model 1 and Model 2
stargazer(mroz.glm1, mroz.glm2, type = 'text')
```

```
##
## ===============================================
##                        Dependent variable:
##                    ----------------------------
##                                lfp
##                         (1)           (2)
## -----------------------------------------------
## age                   0.312***
##                       (0.109)
##
## age_squared          -0.004***
##                       (0.001)
##
## inc                  -0.033***     -0.033***
##                       (0.008)       (0.007)
##
## wcyes                 0.713***      0.703***
##                       (0.196)       (0.193)
##
## lwg                   0.551***      0.584***
##                       (0.145)       (0.145)
##
## totalKids            -0.222***      -0.084
##                       (0.064)       (0.052)
##
## Constant              -5.151**       0.263
##                       (2.261)       (0.226)
##
## -----------------------------------------------
## Observations            753           753
## Log Likelihood        -476.133      -486.223
## Akaike Inf. Crit.     966.266       982.446
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```r
# Apply LRT
anova(mroz.glm1, mroz.glm2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: lfp ~ age + age_squared + inc + wc + lwg + totalKids
## Model 2: lfp ~ inc + wc + lwg + totalKids
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       746     952.27
## 2       748     972.45 -2   -20.18 4.149e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Questions 3:

What is the effect of a decrease in age by 5 years on the odds of labor force participation for a female who was 45 years of age.

Answer: Recall our model:

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 age + \beta_2 age\_squared + \beta_3 inc + \beta_4 wc + \beta_5 lwg + \beta_6 totalKids$$

The odds ratio for an increase in age by 5 is expressed in the following formula:

$$OR = exp(5\beta_1 + 5\beta_2(2 \times age + 5))$$

which depends on the level of age.

Let's compute the numerical change of the odds ratio by inserting the estimates to the formula above from the model stored in `mroz.glm1`, which is used here because we have tested that the age effect is significant.

```
c = -5
age = 45

OR.change = exp(c*(coefficients(mroz.glm1)[['age']] + coefficients(mroz.glm1)[['age_squared']]

OR.change
```

```
## [1] 1.176272
```

Therefore, the estimated odds of labor force participation (lfp) of females who are 45 **years of age increase by** 1.18 **times.**

Question 4:

Estimate the profile likelihood confidence interval of the probability of labor force participation for females who were 40 years old, had income equal to 20, did not attend college, had log wage equal to 1, and did not have children.

Answer:

```
library(mcprofile)

# Define the contrast matrix
K = matrix(data = c(1, 40, 40^2, 20, 0, 1, 0), nrow = 1, ncol = 7)

# Calculate -2log(Lambda)
linear.combo = mcprofile(object = mroz.glm1, CM = K)

# CI for the linear prredictor
ci.logit.profile <- confint(object = linear.combo, level = 0.95)
ci.logit.profile
```

```
##
##    mcprofile - Confidence Intervals
##
## level:        0.95
## adjustment:   single-step
##
##    Estimate lower upper
## C1    0.725 0.384  1.07
```

```r
names(ci.logit.profile)
```

```
## [1] "estimate"    "confint"     "CM"          "quant"       "alternative"
## [6] "level"       "adjust"
```

```r
# CI for probability
exp(ci.logit.profile$confint)/(1 + exp(ci.logit.profile$confint))
```

```
##       lower    upper
## 1 0.5948532 0.745044
```