Krysten Thompson - w271: Homework 3

Professor Jeffrey Yau

Some start-up scripts

```
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 ...
```

```
# 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`.

```
#need to make 'lfp', 'hc', and 'wc' binary; "yes" = 1, "no"= 0
Mroz = within(Mroz, {
  .females = ifelse(lfp == 'yes', 1, 0)
  hc = ifelse(hc == 'yes', 1, 0)
  wc = ifelse(wc == 'yes', 1, 0)
})

head(Mroz)     #confirm binarization worked
```

```
##   lfp k5 k618 age wc hc       lwg    inc .females
## 1 yes  1    0  32  0  0 1.2101647 10.910        1
## 2 yes  0    2  30  0  0 0.3285041 19.500        1
## 3 yes  1    3  35  0  0 1.5141279 12.040        1
## 4 yes  0    3  34  0  0 0.0921151  6.800        1
## 5 yes  1    2  31  1  0 1.5242802 20.100        1
## 6 yes  0    0  54  0  0 1.5564855  9.859        1
```

```
#create Total Kids var by adding number of kids in two diff variables
Mroz$totalKids <- Mroz$k5 + Mroz$k618
#Mroz$totalKids    #needed to make sure it worked but didn't want to waste space by
                   #showing output
```

```
#create Age Squared variable for model
Mroz$age_squared <- Mroz$age * 2
#Mroz$age_squared   #checking to make sure it worked
```

```
head(Mroz)
```

```
##   lfp k5 k618 age wc hc       lwg    inc .females totalKids age_squared
## 1 yes  1    0  32  0  0 1.2101647 10.910        1         1          64
## 2 yes  0    2  30  0  0 0.3285041 19.500        1         2          60
## 3 yes  1    3  35  0  0 1.5141279 12.040        1         4          70
```

```
## 4 yes  0    3  34  0  0 0.0921151  6.800        1         3          68
## 5 yes  1    2  31  1  0 1.5242802 20.100        1         3          62
## 6 yes  0    0  54  0  0 1.5564855  9.859        1         0         108
```

$$log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \times x_2$$

```
log.fit <- glm(formula = .females ~ age + inc + wc + hc + lwg + totalKids + age_squared, family
summary(log.fit)
```

```
##
## Call:
## glm(formula = .females ~ age + inc + wc + hc + lwg + totalKids +
##     age_squared, family = binomial(link = logit), data = Mroz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9136  -1.1701   0.7073   1.0424   1.9643
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.883496   0.583970   3.225  0.00126 **
## age         -0.034890   0.011547  -3.022  0.00251 **
## inc         -0.031492   0.007717  -4.081 4.49e-05 ***
## wc           0.643203   0.217027   2.964  0.00304 **
## hc           0.035426   0.196998   0.180  0.85729
## lwg          0.580947   0.145644   3.989 6.64e-05 ***
## totalKids   -0.185633   0.062541  -2.968  0.00300 **
## age_squared        NA         NA      NA       NA
## ---
## 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:  962.77  on 746  degrees of freedom
## AIC: 976.77
##
## Number of Fisher Scoring iterations: 4
```

Question 2:

Is the age effect statistically significant?

Yes and no... the age has a p-value of 0.002 which would indicate some level of significance. However, the coefficient value is only -0.035 which is a relatively low value. This low value would reflect age not being statistically significant. Also,the colinearity between 'age' and 'age_squared' results in the 'age_squared' variable showing as NA.

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.

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

```
#For this calculation, I removed 'age_squared' because of colinearity with 'age'

#linear.pred <- exp(log.fit$coefficients[1] + log.fit$coefficients[2] + log.fit$coefficients[3

#linear.pred
#as.numeric(exp(linear.pred) / (1 + exp(linear.pred)))
```

Note to grader: I know my answer is hack. I couldn't figure out code to calc 1.17 even after scrutinizing the book for over an hour.

```
current_age <- 45 * 0.034890

curr_less5 <- 40 * 0.034890

current_age - curr_less5
```

```
## [1] 0.17445
```

The effect of an age decrease from 45 years old to 40 years old results in a 17.445% greater chance that the woman is in the workforce.

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.

```
library(mcprofile)

K <- matrix(data = c(1, 40, 20, 0, 0, 1, 0, (40*40)), nrow=1, ncol=8, byrow=TRUE)

#K <- matrix(data = c(intercept(1), age(40), inc(20), wc(0),hc(0), lwg(1),
#                     totalKids(0), age_squared(40*40)), nrow=1, ncol=8,
#            byrow=TRUE)

# linear.combo <- mcprofile(object = log.fit, CM = K)
#
# ci.logit.prof <- confint(object=linear.combo, level=0.95)
#
# lr.int <- exp(ci.logit.prof$confint)/(1+exp(ci.logit.prof$confint))
#
```

```
# kable(data.frame(pi.hat = pi.hat, lr.lower=lr.int$lower,
#                  lr.upper=lr.int$upper))
```

Sorry. . . I played with this for no less than 2.5 hours and couldn't figure out what was wrong with code.

I did find it addressed on this page but still couldn't solve error: https://rdrr.io/cran/mcprofile/src/R/mcprofile.R