w271: Homework 3 (Due: 4pm Monday Week 4)

Grace Lin

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 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

```
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`.

```
df <- Mroz

df$totalKids <- df$k5 + df$k618

df$age_squared <- df$age * df$age

mod.fit <- glm(formula = lfp ~ age + inc + wc + hc + lwg + totalKids + I(age_squared), family =
binomial(link=logit), data = df)

mod.fit
```

```
##
## Call:  glm(formula = lfp ~ age + inc + wc + hc + lwg + totalKids + I(age_squared),
##     family = binomial(link = logit), data = df)
##
## Coefficients:
##    (Intercept)             age             inc           wcyes
##      -5.294073        0.318014       -0.034561        0.666013
##          hcyes             lwg       totalKids  I(age_squared)
##       0.098260        0.549976       -0.222490       -0.004114
##
## Degrees of Freedom: 752 Total (i.e. Null);  745 Residual
## Null Deviance:       1030
## Residual Deviance: 952   AIC: 968
```

Question 2:

Is the age effect statistically significant?

```
mod.fit_2 <- glm(formula = lfp ~ inc + wc + hc + lwg + totalKids, family = binomial(link=logit),
data = df)

anova(mod.fit, mod.fit_2, test = "Chisq")
```

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	745	952.0222	*NA*	*NA*	*NA*
2	747	972.0796	-2	-20.05737	4.411603e-05

2 rows

Model 2 (without age and age square) is statistically significant different from model 1(with age and age square). Thus, age effect is statistically significant.

```
stargazer(mod.fit, mod.fit_2, type = 'text')
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                 lfp
##                           (1)            (2)
## ------------------------------------------------
## age                     0.318***
##                         (0.109)
##
## inc                    -0.035***      -0.035***
##                         (0.008)        (0.008)
##
## wcyes                   0.666***      0.645***

##                         (0.218)        (0.215)
##
## hcyes                    0.098          0.117
##                         (0.199)        (0.194)
##
## lwg                     0.550***      0.583***
##                         (0.146)        (0.145)
##
## totalKids              -0.222***       -0.087*
##                         (0.064)        (0.053)
##
## I(age_squared)         -0.004***
##                         (0.001)
##
## Constant                -5.294**        0.263
##                         (2.282)        (0.226)
##
## ------------------------------------------------
## Observations              753            753
## Log Likelihood         -476.011      -486.040
## Akaike Inf. Crit.      968.022        984.080
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Because the p-value is much smaller than the chosen significance level of 0.05, age effect is statistically significant.

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.

$$log(\frac{\pi}{1-\pi}) = -5.294073 + 0.318014 * Age - 0.004114 * Age^2 + \dots$$

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

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

$$= exp(0.318014 * c + -0.004114 * c(2 \times age + c))$$

```
c = 5
age = 45
effect = exp(0.318014*c + -0.004114* c (2 * age + c))
cat('the effect is', effect)
```

```
## the effect is 3.317595
```

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.

```
# Compute 95% Wald Confidence Interval
ci.pi<-function(data, model, alpha){
  linear.pred = predict(object = mod.fit, newdata = data_q4, type = "link", se = TRUE)
  CI.lin.pred.lower = linear.pred$fit - qnorm(p = 1-alpha/2)*linear.pred$se
  CI.lin.pred.upper = linear.pred$fit + qnorm(p = 1-alpha/2)*linear.pred$se
  CI.pi.lower = exp(CI.lin.pred.lower) / (1 + exp(CI.lin.pred.lower))
  CI.pi.upper = exp(CI.lin.pred.upper) / (1 + exp(CI.lin.pred.upper))
  list(lower = CI.pi.lower, upper = CI.pi.upper)
}
```

```
# Estimate the confidence interval
# If the husband does not have college education:
data_q4 = data.frame(age = 40, inc = 20, wc = 'no', hc = 'no', lwg = 1, totalKids = 0, age_squar
ed = 40*40 )

ci_q4 = ci.pi(data=data_q4, model=mod.fit, alpha = 0.05)
cat("If the husbdand does not have college education, estimated 95% Wald CI for Probability", a
s.numeric(ci_q4), "\n")
```

```
## If the husbdand does not have college education, estimated 95% Wald CI for Probability 0.5861
286 0.7422584
```

```
#if husband has college education:
data_q4 = data.frame(age = 40, inc = 20, wc = 'no', hc = 'yes', lwg = 1, totalKids = 0, age_squa
red = 40*40 )


ci_q4 = ci.pi(data=data_q4, model=mod.fit, alpha = 0.05)
cat("If the husbdand has college education, Estimated 95% Wald CI for Probability", as.numeric(c
i_q4), "\n")
```

```
## If the husbdand has college education, Estimated 95% Wald CI for Probability 0.5849864 0.7788
481
```