Logistic Regression - National Health Interview Survey

Chinpei Tang

Exercise 1: Logistic Regression

Use the NH11 data set that we loaded earlier.

1. Use glm to conduct a logistic regression to predict ever worked (everwrk) using age (age$_p$) and marital status (r_{maritl}).

2. Predict the probability of working for each level of marital status.

Note that the data is not perfectly clean and ready to be modeled. You will need to clean up at least some of the variables before fitting the model.

Load the required library.

```
library(ggplot2)
library(effects)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

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

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

First, load the data into the workspace:

```
setwd("C:/Users/Chinpei/Documents/GitHub/Springboard_FDS/logistic_regression")
NH11 <- readRDS("dataSets/NatHealth2011.rds")
```

Then, examine the data:

```
summary(NH11)
```

```
##      fmx                fpx               wtia_sa           wtfa_sa
##  Length:33014       Length:33014       Min.   :  780.2   Min.   :  846
##  Class :character   Class :character   1st Qu.: 2933.3   1st Qu.: 3613
##  Mode  :character   Mode  :character   Median : 4494.4   Median : 5612
##                                        Mean   : 5607.1   Mean   : 7008
##                                        3rd Qu.: 7278.1   3rd Qu.: 9026
##                                        Max.   :65211.6   Max.   :71281
##
##      region         strat_p        psu_p            sex
##  Min.   :1.000   Min.   :  1   Min.   :1.00   1 Male  :14811
##  1st Qu.:2.000   1st Qu.: 82   1st Qu.:1.00   2 Female:18203
##  Median :3.000   Median :157   Median :1.00
##  Mean   :2.713   Mean   :155   Mean   :1.49
##  3rd Qu.:4.000   3rd Qu.:233   3rd Qu.:2.00
##  Max.   :4.000   Max.   :300   Max.   :2.00
##
##                            hispan_i
##  12 Not Hispanic/Spanish origin:27147
##  02 Mexican                    : 2181
##  03 Mexican-American           : 1348
##  06 Central or South American  :  955
##  01 Puerto Rico                :  567
##  04 Cuban/Cuban American       :  295
##  (Other)                       :  521
##                              mracrpi2         age_p
##  01 White                        :25074   Min.   :18.00
##  02 Black/African American       : 5193   1st Qu.:33.00
##  15 Other Asian (See file layout):  818   Median :47.00
##  10 Chinese                      :  477   Mean   :48.11
##  11 Filipino                     :  468   3rd Qu.:62.00
##  09 Asian Indian                 :  403   Max.   :85.00
##  (Other)                         :  581
##                             r_maritl                  everwrk
##  1 Married - spouse in household:13943   1 Yes            :12153
##  7 Never married                : 7763   2 No             : 1887
##  5 Divorced                     : 4511   7 Refused        :   17
##  4 Widowed                      : 3069   8 Not ascertained:    0
##  8 Living with partner          : 2002   9 Don't know     :    8
##  6 Separated                    : 1121   NA's             :18949
##  (Other)                        :  605
##                hypev                     aasmev
##  1 Yes            :10672   1 Yes            : 4100
##  2 No             :22296   2 No             :28882
##  7 Refused        :   20   7 Refused        :    9
##  8 Not ascertained:    0   8 Not ascertained:    0
##  9 Don't know     :   26   9 Don't know     :   23
##
##
##                aasmyr                    dibev           dibage
##  1 Yes            : 1335   1 Yes            : 3242   Min.   : 1.00
##  2 No             : 2749   2 No             :29260   1st Qu.:40.00
##  7 Refused        :    0   3 Borderline     :  485   Median :50.00
##  8 Not ascertained:    0   7 Refused        :   11   Mean   :49.72
##  9 Don't know     :   16   8 Not ascertained:    0   3rd Qu.:60.00
##  NA's             :28914   9 Don't know     :   16   Max.   :99.00
##                                                      NA's   :29772
##     difage2                   insln                    dibpill
##  Min.   : 0.00   1 Yes            :  945   1 Yes            : 2560
##  1st Qu.: 4.00   2 No             : 3765   2 No             : 2146
##  Median : 9.00   7 Refused        :    0   7 Refused        :    0
##  Mean   :14.96   8 Not ascertained:    0   8 Not ascertained:    0
##  3rd Qu.:18.00   9 Don't know     :    1   9 Don't know     :    5
##  Max.   :99.00   NA's             :28303   NA's             :28303
##  NA's   :29772
##                arth1                     arthlmt          wkdayr
##  1 Yes            : 8181   1 Yes            : 5058   Min.   :  0.000
##  2 No             :24788   2 No             : 8445   1st Qu.:  0.000
##  7 Refused        :    8   7 Refused        :    0   Median :  0.000
##  8 Not ascertained:    0   8 Not ascertained:    0   Mean   :  7.261
##  9 Don't know     :   37   9 Don't know     :    4   3rd Qu.:  2.000
##                            NA's             :19507   Max.   :999.000
##                                                      NA's   :11762
##     beddayr                   aflhca18        aldura10
##  Min.   :  0.00   1 Mentioned      :  683   Min.   : 0.00
##  1st Qu.:  0.00   2 Not mentioned  :11892   1st Qu.: 5.00
##  Median :  0.00   7 Refused        :   17   Median :10.00
##  Mean   : 11.25   8 Not ascertained:   20   Mean   :14.07
##  3rd Qu.:  2.00   9 Don't know     :  104   3rd Qu.:19.00
##  Max.   :999.00   NA's             :20298   Max.   :99.00
##                                             NA's   :32377
##     aldura17        aldura18                  smkev          cigsday
##  Min.   : 0.00   Min.   : 0.00   1 Yes            :13443   Min.   : 1.00
##  1st Qu.: 5.00   1st Qu.: 4.00   2 No             :19491   1st Qu.: 5.00
##  Median :12.00   Median :10.00   7 Refused        :   32   Median :10.00
##  Mean   :18.05   Mean   :18.19   8 Not ascertained:   28   Mean   :12.98
##  3rd Qu.:25.00   3rd Qu.:26.00   9 Don't know     :   20   3rd Qu.:20.00
##  Max.   :99.00   Max.   :99.00                             Max.   :99.00
##  NA's   :31905   NA's   :32331                             NA's   :26833
##      vigmin           modmin             bmi             sleep
##  Min.   : 10.00   Min.   : 10.00   Min.   :11.81   Min.   : 3.000
##  1st Qu.: 30.00   1st Qu.: 20.00   1st Qu.:23.57   1st Qu.: 6.000
##  Median : 45.00   Median : 30.00   Median :26.76   Median : 7.000
##  Mean   : 60.58   Mean   : 55.68   Mean   :29.90   Mean   : 7.862
##  3rd Qu.: 60.00   3rd Qu.: 60.00   3rd Qu.:31.31   3rd Qu.: 8.000
##  Max.   :999.00   Max.   :999.00   Max.   :99.99   Max.   :99.000
##  NA's   :19126    NA's   :14591
##                            ausualpl
##  1 Yes                         :27494
##  2 There is NO place           : 5061
##  3 There is MORE THAN ONE place:  348
##  7 Refused                     :   10
##  8 Not ascertained             :   92
##  9 Don't know                  :    9
##
```

```
str(NH11)
```

```
## 'data.frame':    33014 obs. of  36 variables:
##  $ fmx     : chr  "01" "01" "01" "01" ...
##  $ fpx     : chr  "03" "03" "01" "01" ...
##  $ wtia_sa : num  7521 5784 2512 3086 12530 ...
##  $ wtfa_sa : num  8814 10427 2791 3888 16609 ...
##  $ region  : num  3 3 1 3 3 1 3 3 3 3 ...
##  $ strat_p : num  223 201 3 166 125 31 190 190 217 173 ...
##  $ psu_p   : num  1 2 1 1 2 1 1 1 1 1 ...
##  $ sex     : Factor w/ 2 levels "1 Male","2 Female": 2 2 2 2 2 2 2 2 1 1 ...
##  $ hispan_i: Factor w/ 13 levels "00 Multiple Hispanic",..: 13 13 13 13 13 13 7 13 13 13 ...
##  $ mracrpi2: Factor w/ 9 levels "01 White","02 Black/African American",..: 1 2 2 2 1 1 1 1 2 1 ...
##  $ age_p   : num  47 18 79 51 43 41 21 20 33 56 ...
##  $ r_maritl: Factor w/ 10 levels "0 Under 14 years",..: 6 8 5 7 2 2 8 8 8 2 ...
##  $ everwrk : Factor w/ 5 levels "1 Yes","2 No",..: NA NA 1 NA NA NA NA NA 1 1 ...
##  $ hypev   : Factor w/ 5 levels "1 Yes","2 No",..: 2 2 1 2 2 1 2 2 1 2 ...
##  $ aasmev  : Factor w/ 5 levels "1 Yes","2 No",..: 1 2 2 2 2 2 2 2 2 2 ...
##  $ aasmyr  : Factor w/ 5 levels "1 Yes","2 No",..: 1 NA NA NA NA NA NA NA NA NA ...
##  $ dibev   : Factor w/ 6 levels "1 Yes","2 No",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ dibage  : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ difage2 : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ insln   : Factor w/ 5 levels "1 Yes","2 No",..: 2 NA NA NA NA NA NA NA NA NA ...
##  $ dibpill : Factor w/ 5 levels "1 Yes","2 No",..: 2 NA NA NA NA NA NA NA NA NA ...
##  $ arth1   : Factor w/ 5 levels "1 Yes","2 No",..: 1 2 1 2 2 1 2 2 1 2 ...
##  $ arthlmt : Factor w/ 5 levels "1 Yes","2 No",..: 2 NA 1 NA NA 2 NA 2 2 NA ...
##  $ wkdayr  : num  3 0 NA 0 1 0 0 1 NA 0 ...
##  $ beddayr : num  3 0 0 0 1 0 0 0 0 0 ...
##  $ aflhca18: Factor w/ 5 levels "1 Mentioned",..: 2 NA 2 NA NA 2 2 NA 2 NA ...
##  $ aldura10: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ aldura17: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ aldura18: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ smkev   : Factor w/ 5 levels "1 Yes","2 No",..: 2 2 2 1 3 2 2 2 2 1 ...
##  $ cigsday : num  NA NA NA 5 NA NA NA NA NA NA ...
##  $ vigmin  : num  NA NA NA NA NA 60 120 30 NA 120 ...
##  $ modmin  : num  15 NA 10 NA NA 30 30 120 NA 45 ...
##  $ bmi     : num  100 21.6 32.3 100 100 ...
##  $ sleep   : num  6 8 6 8 9 8 7 6 10 8 ...
##  $ ausualpl: Factor w/ 6 levels "1 Yes","2 There is NO place",..: 1 2 1 2 1 1 1 2 1 1 ...
##  - attr(*, "labels")='data.frame':   36 obs. of  2 variables:
##   ..$ name : Factor w/ 591 levels "aaseryr1","aasmev",..: 452 453 590 589 538 567 534 541 455 520 ..
##   ..$ label: Factor w/ 590 levels " AAU.050_01.010: Doesn't need doctor/haven't had problems",..: 359
```

We are looking at the following specific data:

- **Ever worked (everwrk):** 5-level factors with 18949 NA's. The 5 levels are: "1 Yes", "2 No", "7 Refused", "8 Not ascertained" and "9 Don't know".
- **Age (age_p):** continuous variable between 18 and 85. No NA.
- **Marital status (r_maritl):** 10-level factors with 6 major reported factors, and 605 observations grouped in "Other".

Now collect only the above 3 data into a single data frame:

```
NH11.ear <- NH11[c("everwrk", "age_p", "r_maritl")]
summary(NH11.ear)
```

```
##                everwrk          age_p      
##  1 Yes             :12153   Min.   :18.00  
##  2 No              : 1887   1st Qu.:33.00  
##  7 Refused         :   17   Median :47.00  
##  8 Not ascertained:     0   Mean   :48.11  
##  9 Don't know      :    8   3rd Qu.:62.00  
##  NA's              :18949   Max.   :85.00  
```

```
##
##                               r_maritl
##  1 Married - spouse in household:13943
##  7 Never married                : 7763
##  5 Divorced                     : 4511
##  4 Widowed                      : 3069
##  8 Living with partner          : 2002
##  6 Separated                    : 1121
##  (Other)                        :  605
```

Since there is a significant number of NA's in everwrk data, and logistic regression (and our interest) is the Yes and No prediction, we will omit the NA and "7 Refused", "8 Not ascertained" and "9 Don't know" data:

```
NH11.ear$everwrk <- factor(NH11.ear$everwrk, levels = c("2 No", "1 Yes"))
summary(NH11.ear)
```

```
##    everwrk          age_p                                r_maritl
##  2 No : 1887   Min.   :18.00   1 Married - spouse in household:13943
##  1 Yes:12153   1st Qu.:33.00   7 Never married                : 7763
##  NA's :18974   Median :47.00   5 Divorced                     : 4511
##                Mean   :48.11   4 Widowed                      : 3069
##                3rd Qu.:62.00   8 Living with partner          : 2002
##                Max.   :85.00   6 Separated                    : 1121
##                                (Other)                        :  605
```

Then do prediction using logistic regression:

```
everwrk.pred <- glm(everwrk ~ age_p + r_maritl, data = NH11.ear, family = binomial)
summary(everwrk.pred)
```

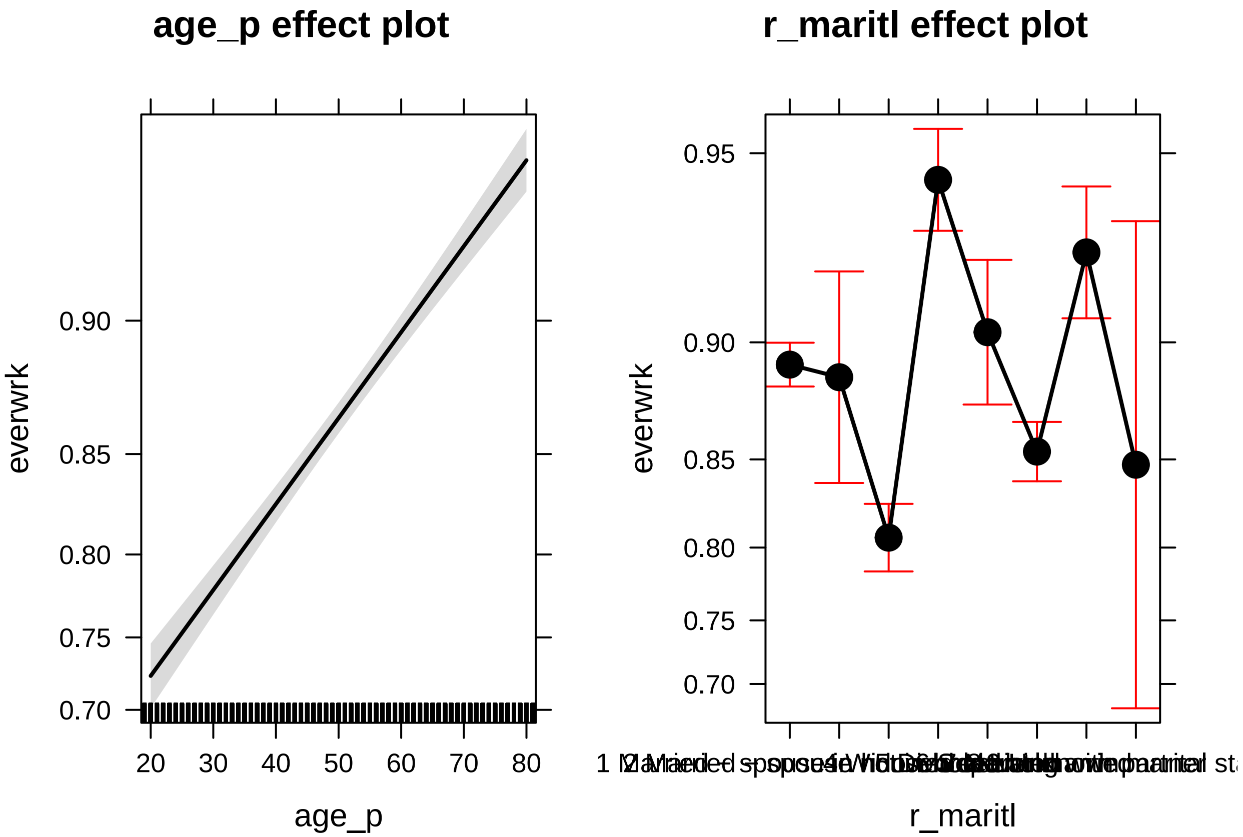
```
##
## Call:
## glm(formula = everwrk ~ age_p + r_maritl, family = binomial,
##     data = NH11.ear)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7308   0.3370   0.4391   0.5650   1.0436
##
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                  0.440248   0.093538   4.707
## age_p                                        0.029812   0.001645  18.118
## r_maritl2 Married - spouse not in household -0.049675   0.217310  -0.229
## r_maritl4 Widowed                           -0.683618   0.084335  -8.106
## r_maritl5 Divorced                           0.730115   0.111681   6.538
## r_maritl6 Separated                          0.128091   0.151366   0.846
## r_maritl7 Never married                     -0.343611   0.069222  -4.964
## r_maritl8 Living with partner                0.443583   0.137770   3.220
## r_maritl9 Unknown marital status            -0.395480   0.492967  -0.802
##                                             Pr(>|z|)
## (Intercept)                                 2.52e-06 ***
## age_p                                        < 2e-16 ***
## r_maritl2 Married - spouse not in household  0.81919    
## r_maritl4 Widowed                           5.23e-16 ***
## r_maritl5 Divorced                          6.25e-11 ***
## r_maritl6 Separated                          0.39742    
## r_maritl7 Never married                     6.91e-07 ***
## r_maritl8 Living with partner                0.00128 ** 
## r_maritl9 Unknown marital status             0.42241    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 11082  on 14039  degrees of freedom
## Residual deviance: 10309  on 14031  degrees of freedom
##   (18974 observations deleted due to missingness)
## AIC: 10327
## 
## Number of Fisher Scoring iterations: 5
```

Then, use the “effects” package to look at the probabilities of each of the cases:

```
summary(allEffects(everwrk.pred))
```

```
##  model: everwrk ~ age_p + r_maritl
## 
##  age_p effect
## age_p
##        20        30        40        50        60        70        80 
## 0.7241256 0.7795664 0.8265345 0.8652253 0.8963682 0.9209721 0.9401248 
## 
##  Lower 95 Percent Confidence Limits
## age_p
##        20        30        40        50        60        70        80 
## 0.7011322 0.7645495 0.8172380 0.8588624 0.8905221 0.9147761 0.9337009 
## 
##  Upper 95 Percent Confidence Limits
## age_p
##        20        30        40        50        60        70        80 
## 0.7459909 0.7938837 0.8354534 0.8713443 0.9019365 0.9267537 0.9459624 
## 
##  r_maritl effect
## r_maritl
##     1 Married - spouse in household 2 Married - spouse not in household 
##                           0.8917800                           0.8868918 
##                           4 Widowed                          5 Divorced 
##                           0.8061891                           0.9447561 
##                         6 Separated                     7 Never married 
##                           0.9035358                           0.8538900 
##               8 Living with partner            9 Unknown marital status 
##                           0.9277504                           0.8472992 
## 
##  Lower 95 Percent Confidence Limits
## r_maritl
##     1 Married - spouse in household 2 Married - spouse not in household
##                           0.8831439                           0.8377247
##                           4 Widowed                          5 Divorced
##                           0.7844913                           0.9332564
##                         6 Separated                     7 Never married
##                           0.8755978                           0.8386559
##                  8 Living with partner        9 Unknown marital status
##                           0.9082334                           0.6794427
##
##  Upper 95 Percent Confidence Limits
## r_maritl
##     1 Married - spouse in household 2 Married - spouse not in household
##                           0.8998502                           0.9225394
##                           4 Widowed                          5 Divorced
##                           0.8261864                           0.9543712
##                         6 Separated                     7 Never married
##                           0.9257318                           0.8679122
##                  8 Living with partner        9 Unknown marital status
##                           0.9433753                           0.9355916
```

```
plot(allEffects(everwrk.pred))
```



We can see that the higher the age is, the higher the probability the individual has ever worked, which makes sense.

For marital status, the divorced case has the highest probability of ever worked at 94.5%.