MINI PROJECT LOGISTIC REGRESSION

Project Summary

Humberto Carvalho – March 2017

EXERCISE – LOGISTIC REGRESSION

0. INTRODUCTION

The Exercise Logistic Regression requests conducting a logistic regression process to predict ever worked (**everwrk**) using age (**age_p**) and marital status (**r_maritl**) variables.

For this purpose, we create the subset **wrk.age.mar**, from the NH11 data set, with the required variables.

```
> wrk.age.mar <- subset(NH11, select = c("everwrk", "age_p", "r_maritl"))
> summary(wrk.age.mar)
              everwrk          age_p                                  r_maritl
 1 Yes            :12153   Min.   :18.00   1 Married - spouse in household:13943
 2 No             : 1887   1st Qu.:33.00   7 Never married                 : 7763
 7 Refused        :   17   Median :47.00   5 Divorced                      : 4511
 8 Not ascertained:    0   Mean   :48.11   4 Widowed                       : 3069
 9 Don't know     :    8   3rd Qu.:62.00   8 Living with partner           : 2002
 NA's             :18949   Max.   :85.00   6 Separated                     : 1121
                                          (Other)                         :  605
```

As the provided data is not ready to be modelled, taking into account that some data is missing, we previously clean up some variables using the MICE (Multiple Imputation by Chained Equations) tool.

Complementary, we store the data from the **everwrk** variable using a factor in order to ensure a proper treatment of this data when using the modelling functions. We also use the **droplevels** function to eliminate the unused levels from the **r_maritl** variable.

Besides this 1st goal, above mentioned, it is also required to predict the probability of working for each level of marital status.

1. Use glm to conduct a logistic regression to predict ever worked based on age and marital status

We apply the **glm** function on the **wrk.age.mar** getting the below results.

```
> wrk.age.mar.model <- glm(everwrk ~ age_p + r_maritl, data = wrk.age.mar, family = "binomial")
>
> summary(wrk.age.mar.model)

Call:
glm(formula = everwrk ~ age_p + r_maritl, family = "binomial",
    data = wrk.age.mar)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0388  -0.6172  -0.4882  -0.3647   2.6980

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                              -0.445835   0.057914  -7.698 1.38e-14 ***
age_p                                    -0.029358   0.001164 -25.213  < 2e-16 ***
r_maritl2 Married - spouse not in household  0.052649   0.125985   0.418   0.6760
r_maritl4 Widowed                         0.668590   0.066306  10.083  < 2e-16 ***
r_maritl5 Divorced                       -0.671813   0.065037 -10.330  < 2e-16 ***
r_maritl6 Separated                      -0.171445   0.093867  -1.826   0.0678 .
r_maritl7 Never married                   0.316595   0.039677   7.979 1.47e-15 ***
r_maritl8 Living with partner            -0.508524   0.074770  -6.801 1.04e-11 ***
r_maritl9 Unknown marital status          0.467248   0.301911   1.548   0.1217
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28231  on 33013  degrees of freedom
Residual deviance: 26664  on 33005  degrees of freedom
AIC: 26682

Number of Fisher Scoring iterations: 5
```

We also show the output from the **predict** and the **tapply** functions as well.

```
> predictEverwrk <- predict(wrk.age.mar.model, type = "response")
> summary(predictEverwrk)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.02626 0.09341 0.13530 0.15280 0.20020 0.41700
```

```
> tapply(predictEverwrk, wrk.age.mar$everwrk, mean)
          1 Yes           2 No       7 Refused 8 Not ascertained      9 Don't know
      0.1453099       0.1947971       0.1737882              NA         0.1483717
```

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

For this purpose, we use the **effects** package. See below the outcome from the **summary** after using **effects** on the model.

```
> wrk.age.mar.model_effect <- effect("r_maritl", wrk.age.mar.model)
> summary(wrk.age.mar.model_effect)

 r_maritl effect
r_maritl
    1 Married - spouse in household 2 Married - spouse not in household                    4 Widowed
                          0.13490824                          0.14117169                   0.23332140
                          5 Divorced                           6 Separated          7 Never married
                          0.07377796                          0.11612120                   0.17629620
              8 Living with partner           9 Unknown marital status
                          0.08574240                          0.19924931

 Lower 95 Percent Confidence Limits
r_maritl
    1 Married - spouse in household 2 Married - spouse not in household                    4 Widowed
                          0.12925910                          0.11426834                   0.21302623
                          5 Divorced                           6 Separated          7 Never married
                          0.06614909                          0.09908628                   0.16710663
              8 Living with partner           9 Unknown marital status
                          0.07540270                          0.12124841

 Upper 95 Percent Confidence Limits
r_maritl
    1 Married - spouse in household 2 Married - spouse not in household                    4 Widowed
                          0.14076435                          0.17317077                   0.25492377
                          5 Divorced                           6 Separated          7 Never married
                          0.08220921                          0.13564382                   0.18587833
              8 Living with partner           9 Unknown marital status
                          0.09735065                          0.30974221
```

Finally, the plot shows the probability of working for each level of marital status as well as the related confidence limits (lower and upper limits) on a clear graphically way.