Logistic Regression

Exercise: logistic regression

Using the NH11 data set:

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.

Part One

Load data, create a new data frame only with variables of interest, examine the data

```
NH11 <- readRDS("NatHealth2011.rds")
NH11.subset <- NH11[c("everwrk", "age_p", "r_maritl")]
summary(NH11.subset)
```

```
##                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
```

The everwrk variable has many NA's, and other values besides Yes and No.

Create a 'training' data set to build our model by removing all other values besides Yes and No. Remove unused factors to prevent errors in modelling.

```
train <- subset(NH11.subset, everwrk == "1 Yes" | everwrk == "2 No")
train$everwrk <- factor(train$everwrk)
train$r_maritl <- factor(train$r_maritl)
summary(train)
```

```
##   everwrk          age_p                                   r_maritl
##  1 Yes:12153   Min.   :18.00   1 Married - spouse in household:5458
##  2 No : 1887   1st Qu.:39.00   7 Never married                :2843
##                Median :60.00   4 Widowed                      :2518
##                Mean   :55.98   5 Divorced                     :1907
##                3rd Qu.:73.00   8 Living with partner          : 601
##                Max.   :85.00   6 Separated                    : 467
##                                (Other)                        : 246
```

Use glm to conduct a logistic regression to predict ever worked (everwrk) using age (age_p) and marital status (r_maritl)

```
trainLog <- glm(everwrk ~ age_p + r_maritl, data = train, family = binomial)
summary(trainLog)
```

```
##
## Call:
## glm(formula = everwrk ~ age_p + r_maritl, family = binomial,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0436  -0.5650  -0.4391  -0.3370   2.7308
##
## 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
## AIC: 10327
```

```
##
## Number of Fisher Scoring iterations: 5
```

Age, Widowed, Divorced, Never Married and Living with Partner are all statistically significant predictors of ever worked. However, only Widowed and Never Married show a postive correlation.

Part Two

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

Create a dataset with predictors set at each level of marital status, predict probability of working at each level.

```
predictData <- with(train, expand.grid(r_maritl = levels(train$r_maritl), age_p = mean(age_p)))

cbind(predictData, predict(trainLog, type = "response", se.fit = TRUE, interval = "confidence",
                           newdata = predictData))
```

```
##                               r_maritl    age_p        fit      se.fit
## 1     1 Married - spouse in household 55.97728 0.10822000 0.004259644
## 2 2 Married - spouse not in household 55.97728 0.11310823 0.021393167
## 3                            4 Widowed 55.97728 0.19381087 0.010634762
## 4                           5 Divorced 55.97728 0.05524394 0.005361664
## 5                          6 Separated 55.97728 0.09646417 0.012707502
## 6                        7 Never married 55.97728 0.14611000 0.007459212
## 7                 8 Living with partner 55.97728 0.07224958 0.008904955
## 8              9 Unknown marital status 55.97728 0.15270076 0.063528455
##   residual.scale
## 1              1
## 2              1
## 3              1
## 4              1
## 5              1
## 6              1
## 7              1
## 8              1
```

Based on this model, the probability of working for each level of marital status ranges between 7.2 % and 19.4%