MAST90139 Assignment 1

Haonan Zhong 867492 (Wed 14:15 Qiuyi Li)

```
domviolence <- read.csv(file="domviolence.csv")

# Factor all the categorical predictors before fitting an initial model
domviolence$age=factor(domviolence$age)
domviolence$ms=factor(domviolence$ms)
domviolence$mmo=factor(domviolence$mmo)
domviolence$smok=factor(domviolence$smok)
domviolence$alc=factor(domviolence$alc)
domviolence$falc=factor(domviolence$falc)
domviolence$educ=factor(domviolence$educ)
domviolence$reg=factor(domviolence$reg)
domviolence$dv=factor(domviolence$dv)

# Obtain a simple summary of the domviolence dataframe
summary(domviolence)
```

```
##  age     ms      mmo      smok    alc      falc     educ    reg     dv
##  0:398   1:875   0: 258   0:983   0:1209   0:1032   0: 58   1:275   0:947
##  1:532   2: 98   1:1058   1:333   1: 107   1: 284   1:632   2:316   1:369
##  2:242   3: 51                                     2:626   3:378
##  3:144   4: 47                                             4:347
##          5: 28
##          6:217
```

Question 1a

```
# Fit an initial model that contains all the predictors
model0 <- glm(dv ~ ., family = binomial, data = domviolence)

# Compute hypothesis testing to remove non-significant predictors
anova(model0, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: dv
##
## Terms added sequentially (first to last)
##
##
```

```
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                  1315     1561.6
## age   3  26.1373      1312     1535.5 8.926e-06 ***
## ms    5  31.3925      1307     1504.1 7.835e-06 ***
## mmo   1   4.0785      1306     1500.0  0.043431 *
## smok  1  17.9658      1305     1482.1 2.249e-05 ***
## alc   1   2.4734      1304     1479.6  0.115787
## falc  1   9.7522      1303     1469.8  0.001791 **
## educ  2  23.4457      1301     1446.4 8.106e-06 ***
## reg   3  28.7213      1298     1417.7 2.563e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above table returns a p-value of 0.115787 for `alcohol`, which indicates alcohol does not have significant effect, hence we will drop this variable.

```
model0 <- glm(dv ~ age + ms + mmo + smok + falc + educ + reg,
              family = binomial, data = domviolence)
anova(model0, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: dv
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                  1315     1561.6
## age   3  26.1373      1312     1535.5 8.926e-06 ***
## ms    5  31.3925      1307     1504.1 7.835e-06 ***
## mmo   1   4.0785      1306     1500.0  0.043431 *
## smok  1  17.9658      1305     1482.1 2.249e-05 ***
## falc  1  10.5232      1304     1471.5  0.001179 **
## educ  2  22.6593      1302     1448.9 1.201e-05 ***
## reg   3  28.7468      1299     1420.1 2.531e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As depicted by the result, all the remaining terms are significant. Next, we will perform a stepwise selection using AIC to further simplify our model.

```
model1 <- step(model0, trace = 0)
summary(model1)
```

```
##
## Call:
## glm(formula = dv ~ age + ms + smok + falc + educ + reg, family = binomial,
##     data = domviolence)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7848  -0.8195  -0.5870   1.0787   2.3143
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.52027    0.35749   1.455 0.145567
## age1        -0.56976    0.17674  -3.224 0.001265 **
## age2        -0.88831    0.23941  -3.710 0.000207 ***
## age3        -0.92992    0.29064  -3.200 0.001376 **
## ms2          0.30849    0.24311   1.269 0.204467
## ms3          0.57716    0.30935   1.866 0.062078 .
## ms4          1.38101    0.32680   4.226 2.38e-05 ***
## ms5          0.42905    0.45987   0.933 0.350828
## ms6          0.03606    0.21089   0.171 0.864219
## smok1        0.53873    0.14646   3.678 0.000235 ***
## falc1        0.44970    0.15069   2.984 0.002842 **
## educ1       -0.97425    0.29679  -3.283 0.001029 **
## educ2       -1.32296    0.30928  -4.278 1.89e-05 ***
## reg2        -0.91585    0.20992  -4.363 1.28e-05 ***
## reg3         0.01977    0.17628   0.112 0.910686
## reg4        -0.43371    0.18632  -2.328 0.019929 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1561.6  on 1315  degrees of freedom
## Residual deviance: 1421.4  on 1300  degrees of freedom
## AIC: 1453.4
##
## Number of Fisher Scoring iterations: 4
```

As the summary table suggests, stepwise selection has removed `mmo`.

Question 1b

```
# Convert age and educ back to numerical form
domviolence$age <- as.integer(domviolence$age)
domviolence$educ <- as.integer(domviolence$educ)
model2 <- glm(dv ~ age + ms + smok + falc + educ + reg, family = binomial, domviolence)
summary(model2)
```

```
##
## Call:
## glm(formula = dv ~ age + ms + smok + falc + educ + reg, family = binomial,
##     data = domviolence)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8416  -0.8351  -0.6002   1.0940   2.3086
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.79432    0.42005   1.891 0.058623 .
## age         -0.33551    0.09131  -3.674 0.000239 ***
## ms2          0.33731    0.24119   1.398 0.161963
## ms3          0.56094    0.30786   1.822 0.068446 .
## ms4          1.36151    0.32560   4.181 2.90e-05 ***
## ms5          0.56923    0.44937   1.267 0.205250
## ms6          0.18601    0.19852   0.937 0.348776
## smok1        0.51611    0.14554   3.546 0.000391 ***
## falc1        0.42240    0.14963   2.823 0.004759 **
## educ        -0.48947    0.12230  -4.002 6.27e-05 ***
## reg2        -0.91205    0.20919  -4.360 1.30e-05 ***
## reg3         0.04867    0.17486   0.278 0.780745
## reg4        -0.41621    0.18466  -2.254 0.024199 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1561.6  on 1315  degrees of freedom
## Residual deviance: 1427.9  on 1303  degrees of freedom
## AIC: 1453.9
##
## Number of Fisher Scoring iterations: 4
```

```
anova(model2, model1, test='Chi')
```

```
## Analysis of Deviance Table
##
## Model 1: dv ~ age + ms + smok + falc + educ + reg
## Model 2: dv ~ age + ms + smok + falc + educ + reg
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1303     1427.9
## 2      1300     1421.4  3   6.4509  0.09162 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the test above, the p-value 0.09162 is greater than the significance level of 0.05. Thus, we will accept `model2`, since it's not significantly different from `model1`, and it's simpler than `model1` in terms of model complexity.

Question 1c

```
model3 <- glm(dv ~ (age + ms + smok + falc + educ + reg)^2, family = binomial, domviolence)

# Perform stepwise selection to simplify model 3
suppressWarnings(model4 <- step(model3, trace = FALSE))
summary(model4)
```

```
##
## Call:
## glm(formula = dv ~ age + ms + smok + falc + educ + reg + ms:falc +
##     smok:falc + educ:reg, family = binomial, data = domviolence)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9001  -0.8120  -0.6126   1.0027   2.3227
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.13349    0.65665   1.726 0.084320 .
## age         -0.37147    0.09332  -3.980 6.88e-05 ***
## ms2          0.76595    0.27660   2.769 0.005620 **
## ms3          0.46111    0.38588   1.195 0.232102
## ms4          1.31717    0.36560   3.603 0.000315 ***
## ms5          0.53650    0.50969   1.053 0.292523
## ms6          0.07792    0.22467   0.347 0.728714
## smok1        0.67308    0.17019   3.955 7.66e-05 ***
## falc1        0.63780    0.20800   3.066 0.002167 **
## educ        -0.62665    0.25004  -2.506 0.012204 *
## reg2        -1.33985    0.91057  -1.471 0.141172
## reg3         0.41698    0.73350   0.568 0.569716
## reg4        -1.90241    0.81954  -2.321 0.020270 *
## ms2:falc1   -1.64084    0.57575  -2.850 0.004373 **
## ms3:falc1    0.50303    0.66004   0.762 0.445985
## ms4:falc1    0.23587    0.82867   0.285 0.775926
## ms5:falc1    0.48348    1.15166   0.420 0.674621
## ms6:falc1    0.24844    0.40581   0.612 0.540400
## smok1:falc1 -0.47591    0.33210  -1.433 0.151841
## educ:reg2    0.19103    0.37798   0.505 0.613283
## educ:reg3   -0.15840    0.30731  -0.515 0.606259
## educ:reg4    0.60461    0.33414   1.809 0.070379 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1561.6  on 1315  degrees of freedom
## Residual deviance: 1406.5  on 1294  degrees of freedom
## AIC: 1450.5
##
## Number of Fisher Scoring iterations: 4
```

```
# Further simplify the model using Chi-square test
anova(model4, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: dv
##
## Terms added sequentially (first to last)
```

```
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       1315     1561.6
## age        1  23.3206      1314     1538.3 1.371e-06 ***
## ms         5  31.2069      1309     1507.1 8.526e-06 ***
## smok       1  19.6686      1308     1487.4 9.210e-06 ***
## falc       1  10.2754      1307     1477.2  0.001348 **
## educ       1  18.2615      1306     1458.9 1.926e-05 ***
## reg        3  31.0159      1303     1427.9 8.435e-07 ***
## ms:falc    5  12.7910      1298     1415.1  0.025418 *
## smok:falc  1   1.5394      1297     1413.5  0.214704
## educ:reg   3   7.0232      1294     1406.5  0.071163 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We will first remove the intetaction term between `smok` and `falc`, given it's p-value is larger than the significance level of 0.05.

```
model4 <- glm(dv ~ age + ms + smok + falc + educ + reg + ms:falc + educ:reg,
              family = binomial, data = domviolence)
anova(model4, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: dv
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       1315     1561.6
## age        1   23.321      1314     1538.3 1.371e-06 ***
## ms         5   31.207      1309     1507.1 8.526e-06 ***
## smok       1   19.669      1308     1487.4 9.210e-06 ***
## falc       1   10.275      1307     1477.2  0.001348 **
## educ       1   18.262      1306     1458.9 1.926e-05 ***
## reg        3   31.016      1303     1427.9 8.435e-07 ***
## ms:falc    5   12.791      1298     1415.1  0.025418 *
## educ:reg   3    6.508      1295     1408.6  0.089347 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again, we will remove interaction between `educ` and `reg`, given it's high p-value.

```
model4 <- glm(dv ~  ms + age + smok + falc + educ + reg + ms:falc,
              family = binomial, data = domviolence)
anova(model4, test = "Chi")
```

```
## Analysis of Deviance Table
```

```
##
## Model: binomial, link: logit
##
## Response: dv
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     1315     1561.6
## ms       5   42.425      1310     1519.2 4.833e-08 ***
## age      1   12.103      1309     1507.1 0.0005034 ***
## smok     1   19.669      1308     1487.4 9.210e-06 ***
## falc     1   10.275      1307     1477.2 0.0013482 **
## educ     1   18.262      1306     1458.9 1.926e-05 ***
## reg      3   31.016      1303     1427.9 8.435e-07 ***
## ms:falc  5   12.791      1298     1415.1 0.0254178 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, as the table suggests, all the remaining terms are relevant. A summary table of the final model is printed below.

```
summary(model4)
```

```
##
## Call:
## glm(formula = dv ~ ms + age + smok + falc + educ + reg + ms:falc,
##     family = binomial, data = domviolence)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9645  -0.8312  -0.5834   1.0333   2.3199
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.79898    0.42542   1.878 0.060369 .
## ms2          0.79651    0.27412   2.906 0.003665 **
## ms3          0.43946    0.38210   1.150 0.250094
## ms4          1.31189    0.36282   3.616 0.000299 ***
## ms5          0.48817    0.50342   0.970 0.332192
## ms6          0.14320    0.22283   0.643 0.520462
## age         -0.34707    0.09181  -3.780 0.000157 ***
## smok1        0.53324    0.14649   3.640 0.000273 ***
## falc1        0.52629    0.19063   2.761 0.005766 **
## educ        -0.49007    0.12337  -3.972 7.12e-05 ***
## reg2        -0.90821    0.21067  -4.311 1.63e-05 ***
## reg3         0.02792    0.17609   0.159 0.874038
## reg4        -0.42353    0.18623  -2.274 0.022953 *
## ms2:falc1   -1.78134    0.57027  -3.124 0.001786 **
## ms3:falc1    0.32013    0.65570   0.488 0.625388
## ms4:falc1    0.24874    0.83284   0.299 0.765197
## ms5:falc1    0.59486    1.13590   0.524 0.600494
```

```
## ms6:falc1    0.11761    0.40123   0.293 0.769425
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1561.6  on 1315  degrees of freedom
## Residual deviance: 1415.1  on 1298  degrees of freedom
## AIC: 1451.1
##
## Number of Fisher Scoring iterations: 4
```

Therefore, the "best" model is of the form:

$$
\begin{aligned}
logit(\theta) = {} & 0.79898 + 0.79651 \times \texttt{ms.2} + 0.43946 \times \texttt{ms.3} + 1.31189 \times \texttt{ms.4} + 0.48817 \times \texttt{ms.5} \\
& + 0.14320 \times \texttt{ms.6} + 0.53324 \times \texttt{smok.1} + 0.52629 \times \texttt{falc.1} - 0.90821 \times \texttt{reg.2} + 0.02792 \times \texttt{reg.3} \\
& - 0.42353 \times \texttt{reg.4} - 0.34707 \times \texttt{age} - 0.49007 \times \texttt{educ} - 1.78134 \times \texttt{ms.2} : \texttt{falc.1} + 0.32013 \times \texttt{ms.3} : \texttt{falc.1} \\
& + 0.24874 \times \texttt{ms.4} : \texttt{falc.1} + 0.59486 \times \texttt{ms.5} : \texttt{falc.1} + 0.11761 \times \texttt{ms.6} : \texttt{falc.1}
\end{aligned}
$$

Question 2 Odds ratio calculation and interpretation

Marital Status:

OR at various levels of `falc`	`falc` = 0	`falc` = 1
`ms` = 1 vs. `ms` = 1	1	1
`ms` = 2 vs. `ms` = 1	$e^{0.79651} = 2.2178$	$e^{0.79651-1.78134} = 0.3735$
`ms` = 3 vs. `ms` = 1	$e^{0.43946} = 1.5519$	$e^{0.43946+0.32013} = 2.1374$
`ms` = 4 vs. `ms` = 1	$e^{1.31189} = 3.7132$	$e^{1.31189+0.24874} = 4.7618$
`ms` = 5 vs. `ms` = 1	$e^{0.48817} = 1.6293$	$e^{0.48817+0.59486} = 2.9536$
`ms` = 6 vs. `ms` = 1	$e^{0.14320} = 1.1540$	$e^{0.14320+0.11761} = 1.2980$

(`falc` = 0, `ms` = 6 vs. `ms` = 1) When there were no concern caused by family member's use of alcohol during childhood. The estimated odds of a never married woman to respond positively about domestic violence or mental abuse experienced during the previous 12 months is 15.40% more than a married woman. (OR = 1.1540)

(`falc` = 1, `ms` = 6 vs. `ms` = 1) On the other hand, if there were concern caused by family member's use of alcohol when grow up. The estimated odds ratio of responding positively for a never married woman is 29.80% higher compared to a married woman. (OR = 1.2980)

Smoking: Odds ratio = $e^{\beta_6} = e^{0.53324} = 1.7044$
The estimated odds of a non-smoking woman (`smok` = 1) are 70.44% more likely to respond positively about physical domestic violence or mental abuse experienced during the previous 12 month compared to a smoking woman (`smok` = 0).

Family alcohol:

OR at various `ms` levels	`falc` = 0 vs. `falc` = 0	`falc` = 1 vs. `falc` = 0
`ms` = 1	1	$e^{0.52629} = 1.6926$
`ms` = 2	1	$e^{0.52629-1.78134} = 0.2851$
`ms` = 3	1	$e^{0.52629+0.32013} = 2.3313$
`ms` = 4	1	$e^{0.52629+0.24874} = 2.1707$
`ms` = 5	1	$e^{0.52629+0.59486} = 3.0684$
`ms` = 6	1	$e^{0.52629+0.11761} = 1.9040$

(`ms` = 5, `falc` = 1 vs. `falc` = 0) The odds of a widowed woman to respond positively about physical domestic violence or mental abuse experienced during the previous 12 months, who has concern over family member's alcohol abuse when grow up, is three times (OR = 3.0684) higher than a widowed woman who has no concern.

(`ms` = 1, `falc` = 1 vs. `falc` = 0) Whilst a married woman who has concern over family's alcohol abuse during childhood is 69.26% more likely to respond positively to questions about physical domestic violence or mental abuse experienced during the previous 12 months compared to a married woman who has no concern caused. (OR = 1.6926)

Region:

	`north`	`east`	`south`	`west`
OR for each region vs. `north`	1	$e^{-0.90821} = 0.4032$	$e^{0.02792} = 1.0283$	$e^{-0.42353} = 0.6547$

(`south` vs. `north`) The estimated odds of a woman from the south region to respond positively about physical domestic violence or mental abuse experienced during the previous 12 months is almost the same as a woman from the north region. (OR = 1.0283)

(`east vs. north`) Contrarily, the estimated odds of a woman from the east region to respond positively about physical domestic violence or mental abuse experienced during the previous 12 months is 59.67% less than a woman from the north region. (OR = 0.4032)

Age: $e^{\beta_{11}} = e^{-0.34707} = 0.7068$
For every increase in age, the estimated odds of a woman responding positively about physical domestic violence or mental abuse experienced during the previous 12 month decreased by 29.32%.

Education: $e^{\beta_{12}} = e^{-0.49007} = 0.6126$
For every increase in the number of years in formal education, the estimated odds of a woman responding positively about physical domestic violence or mental abuse experienced during the previous 12 month decreased by 38.74%.