Applied Multivariate Analysis Homework 4

110024516 邱繼賢

Problem 1.

Construct a binomial GLM with logit link function

$$y_x \sim Bin(n_x\ ,\ p_x)$$

$$logit(p_x) = \eta_x = X\beta$$

where X is a model matrix which contains main and interaction effects between all three predictors, *agegp, alcgp, tobgp*

Then, using the *step()* function which is a backward elimination by comparing AIC values and choose the smallest one.

Stop the algorithm when the AIC value by doing nothing is the smallest one.

```
data("esoph")
fit1 = glm(cbind(ncases, ncontrols) ~ agegp*alcgp*tobgp
           , esoph, family = binomial)
step(fit1)
```

```
## Start:  AIC=291.05
## cbind(ncases, ncontrols) ~ agegp * alcgp * tobgp
##
##                      Df Deviance    AIC
## - agegp:alcgp:tobgp 37   30.824 247.88
## <none>                    0.000 291.06
##
## Step:  AIC=247.88
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp + agegp:alcgp +
##     agegp:tobgp + alcgp:tobgp
##
##               Df Deviance    AIC
```

```
## - alcgp:tobgp  9   37.535 236.59
## - agegp:tobgp 15   50.309 237.36
## - agegp:alcgp 15   56.807 243.86
## <none>             30.824 247.88
##
## Step:  AIC=236.59
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp + agegp:alcgp +
##     agegp:tobgp
##
##               Df Deviance    AIC
## - agegp:tobgp 15   56.256 225.31
## - agegp:alcgp 15   62.776 231.83
## <none>             37.535 236.59
##
## Step:  AIC=225.31
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp + agegp:alcgp
##
##               Df Deviance    AIC
## - agegp:alcgp 15   82.337 221.39
## <none>             56.256 225.31
## - tobgp        3   80.300 243.35
##
## Step:  AIC=221.39
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp
##
##         Df Deviance    AIC
## <none>       82.337 221.39
## - tobgp  3  105.881 238.94
## - agegp  5  208.825 337.88
## - alcgp  3  210.270 343.32
```

```
##
## Call:  glm(formula = cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp,
##     family = binomial, data = esoph)
##
## Coefficients:
## (Intercept)      agegp.L      agegp.Q      agegp.C      agegp^4      agegp^5
```

```
##    -1.19039      3.99663     -1.65741      0.11094      0.07892     -0.26219
##     alcgp.L      alcgp.Q      alcgp.C      tobgp.L      tobgp.Q      tobgp.C
##     2.53899      0.09376      0.43930      1.11749      0.34516      0.31692
##
## Degrees of Freedom: 87 Total (i.e. Null);  76 Residual
## Null Deviance:       368
## Residual Deviance: 82.34     AIC: 221.4
```

By the result above, we can simplify our model into

$$
\begin{aligned}
y_x &\sim Bin(n_x \ , \ p_x) \\
logit(p_x) &= \eta_x = X\beta
\end{aligned}
$$

where model matrix X only contains the main effect of the predictors *agegp, alcgp, tobgp*

```
fit1.2 = glm(cbind(ncases, ncontrols) ~ agegp+alcgp+tobgp
           , esoph, family = binomial)
drop1(fit1.2, test = "Chi")
```

```
## Single term deletions
##
## Model:
## cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp
##        Df Deviance    AIC     LRT  Pr(>Chi)
## <none>       82.337 221.39
## agegp   5  208.825 337.88 126.488 < 2.2e-16 ***
## alcgp   3  210.270 343.32 127.933 < 2.2e-16 ***
## tobgp   3  105.881 238.94  23.544  3.11e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All the three predictors *agegp, alcgp, tobgp* are having significant contribution for our model.

```
summary(fit1.2)
```

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp,
```

```
##     family = binomial, data = esoph)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9507  -0.7376  -0.2438   0.6130   2.4127
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.19039    0.20737  -5.740 9.44e-09 ***
## agegp.L      3.99663    0.69389   5.760 8.42e-09 ***
## agegp.Q     -1.65741    0.62115  -2.668  0.00762 **
## agegp.C      0.11094    0.46815   0.237  0.81267
## agegp^4      0.07892    0.32463   0.243  0.80792
## agegp^5     -0.26219    0.21337  -1.229  0.21915
## alcgp.L      2.53899    0.26385   9.623  < 2e-16 ***
## alcgp.Q      0.09376    0.22419   0.418  0.67578
## alcgp.C      0.43930    0.18347   2.394  0.01665 *
## tobgp.L      1.11749    0.24014   4.653 3.26e-06 ***
## tobgp.Q      0.34516    0.22414   1.540  0.12358
## tobgp.C      0.31692    0.21091   1.503  0.13294
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 367.953  on 87  degrees of freedom
## Residual deviance:  82.337  on 76  degrees of freedom
## AIC: 221.39
##
## Number of Fisher Scoring iterations: 6
```

We can see that effect *agegp.L, agegp.Q, alcgp.L, alcgp.C, tobgp.L* are significant in Wald test, but there are many covariate classes with small n_i *'s*. By Hauck-Donner effect the standard errors can be over-estimated and so we need to be careful.

Problem 2.

Now, convert the three predictors *agegp, alcgp, tobgp* as numerical variable, so we do not have to represent them by dummy variables. Then the model can be simplified as

$$
\begin{aligned}
y_x \ &\sim \ Bin(n_x \ , \ p_x) \\
logit(p_x) \ &= \ \eta_x \ = \ \beta_0 \ + \ \beta_1 \times agegp \ + \ \beta_2 \times (agegp)^2 \ + \ \beta_3 \times alcgp \ + \ \beta_4 \times tobgp
\end{aligned}
$$

where

$$
agegp \ = \begin{cases} 1 \ , \ 25 \sim 34 \text{ years} \\ 2 \ , \ 35 \sim 44 \\ 3 \ , \ 45 \sim 54 \\ 4 \ , \ 55 \sim 64 \\ 5 \ , \ 65 \sim 74 \\ 6 \ , \ 75+ \end{cases} \quad , \quad alcgp \ = \begin{cases} 1 \ , \ 0 \sim 39 \text{ gm/day} \\ 2 \ , \ 40 \sim 79 \\ 3 \ , \ 80 \sim 119 \\ 4 \ , \ 120+ \end{cases} \quad , \quad tobgp \ = \begin{cases} 1 \ , \ 0 \sim 9 \text{ gm/day} \\ 2 \ , \ 10 \sim 19 \\ 3 \ , \ 20 \sim 29 \\ 4 \ , \ 30+ \end{cases}
$$

```r
fit2 = glm(cbind(ncases, ncontrols) ~ unclass(agegp) + I(unclass(agegp)^2) + unclass(alcgp) + unclass(tc
           , esoph, family = binomial)
drop1(fit2, test = "Chi")
```

```
## Single term deletions
##
## Model:
## cbind(ncases, ncontrols) ~ unclass(agegp) + I(unclass(agegp)^2) +
##     unclass(alcgp) + unclass(tobgp)
##                     Df Deviance    AIC     LRT  Pr(>Chi)
## <none>                   93.172 218.23
## unclass(agegp)       1  126.099 249.15  32.927 9.567e-09 ***
## I(unclass(agegp)^2)  1  108.779 231.83  15.607 7.796e-05 ***
## unclass(alcgp)       1  215.963 339.02 122.791 < 2.2e-16 ***
## unclass(tobgp)       1  114.342 237.40  21.170 4.203e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All the four variables *agegp, (agegp)^2, alcgp, tobgp* are having significant (in deviance-based test) contribution for our model.

Problem 3.

Test fot goodness-of-fit

$$\begin{cases} H_0 & : \text{ The model fits good enough} \\ H_1 & : \text{ The model does not fit well} \end{cases}$$

```
summary(fit2)
```

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ unclass(agegp) + I(unclass(agegp)^2) +
##     unclass(alcgp) + unclass(tobgp), family = binomial, data = esoph)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2757  -0.7828  -0.2313   0.5679   2.4646
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -10.10233    1.03074  -9.801  < 2e-16 ***
## unclass(agegp)        2.50576    0.50188   4.993 5.95e-07 ***
## I(unclass(agegp)^2)  -0.23417    0.06402  -3.658 0.000255 ***
## unclass(alcgp)        1.06511    0.10458  10.185  < 2e-16 ***
## unclass(tobgp)        0.43951    0.09559   4.598 4.27e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 367.953  on 87  degrees of freedom
## Residual deviance:  93.172  on 83  degrees of freedom
## AIC: 218.23
##
## Number of Fisher Scoring iterations: 5
```

We can see that the deviance = 93.172 on 83 degrees of freedom, and under $H_0 \, : \, D_S \overset{a}{\sim} \chi^2_{83} \Rightarrow$ p-value $= P\left(\chi^2_{83} \, > \, D_S\right) \, = \, 0.2087865 \, > \, 0.05$

$\therefore$ Do not reject H_0, the model fits the data well.

However, the chi-square (null distribution) is only an approximation that becomes more accurate as the n_i *'s* increase (often suggest $n_i \geq 5$). There are several covariate classes whose n_i *'s* are pretty small, so the test might not be accurate for this data.

Problem 4.

When moving to a category one higher in alcohol concumption, the log-odds of *ncases* increase by $\hat{\beta}_3 = 1.06511$, or the odds of *ncases* increase to $exp\left(\hat{\beta}_3\right) = 290.1158\%$

```
c(fit2$coef[4], exp(fit2$coef[4]))
```

```
## unclass(alcgp) unclass(alcgp)
##       1.065109       2.901154
```

And the 95% confidence intervals for this predicted effect (in log-odds and odds), which are computed using profile likelihood methods, are shown as below.

```
library(MASS)
confint(fit2)[4,]
```

```
##     2.5 %    97.5 %
## 0.8644407 1.2749782
```

```
exp(confint(fit2)[4,])
```

```
##    2.5 %   97.5 %
## 2.373678 3.578623
```

Problem 5.

Because this is a case-control study, namely retrospective study :

- β_1 , β_2 , β_3 , β_4 are estimable

- β_0 is inestimable $\Rightarrow$ cannot estimate probability

Therefore, we can only predict the effect of variable (such as **Problem 4.**), and can do nothing about predicting probability.