PH241 HW11

Kunal Mishra

4/12/2018

Basic Setup and Minimal EDA

```
library(dplyr)
library(DataExplorer)
library(lmtest)
library(dummies)

data = read.csv(file="HW11.csv", header=TRUE)
data %>% nrow
```

```
## [1] 975
```

```
data %>% head
```

```
##   X agegp alcgp tobgp casestatus
## 1 1     0     0     0          0
## 2 2     0     0     0          0
## 3 3     0     0     0          0
## 4 4     0     0     0          0
## 5 5     0     0     0          0
## 6 6     0     0     0          0
```

Question 1A

```
data.1A =
    data %>%
    mutate(lowAlc = ifelse(alcgp < 2, 1, 0)) %>%
    select(-(X))

data.1A %>% head
```

```
##   agegp alcgp tobgp casestatus lowAlc
## 1     0     0     0          0      1
## 2     0     0     0          0      1
## 3     0     0     0          0      1
## 4     0     0     0          0      1
## 5     0     0     0          0      1
## 6     0     0     0          0      1
```

Now that we've boiled our data down to a binary explanatory variable based on a threshold of 80g of alcohol per day, let's run logistic regression to examine whether there is an association. First, let's run it using just one explanatory variable – lowAlc.

```
fit1.A = glm(formula=casestatus~lowAlc, data=data.1A, family="binomial")
summary(fit1.A)
```

```
##
## Call:
```

```
## glm(formula = casestatus ~ lowAlc, family = "binomial", data = data.1A)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1240  -0.5387  -0.5387  -0.5387   2.0010
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.1270     0.1400  -0.907    0.364
## lowAlc       -1.7299     0.1752  -9.872   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 989.49  on 974  degrees of freedom
## Residual deviance: 893.06  on 973  degrees of freedom
## AIC: 897.06
##
## Number of Fisher Scoring iterations: 4
```

Examining CIs of e^coefficients

```r
exp(confint(fit1.A))
```

```
## Waiting for profiling to be done...
```

```
##                 2.5 %    97.5 %
## (Intercept) 0.6686045 1.1583156
## lowAlc      0.1255509 0.2496999
```

Now, using all the variables available to us, let's use multiple logistic regression

```r
fit2.A = glm(formula=casestatus~., data=data.1A, family="binomial")
summary(fit2.A)
```

```
##
## Call:
## glm(formula = casestatus ~ ., family = "binomial", data = data.1A)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0854  -0.5992  -0.3419  -0.1367   2.8028
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.89783    0.60205  -9.796  < 2e-16 ***
## agegp        0.76051    0.08271   9.195  < 2e-16 ***
## alcgp        1.46649    0.20603   7.118 1.10e-12 ***
## tobgp        0.42356    0.09422   4.495 6.95e-06 ***
## lowAlc       0.80591    0.38592   2.088   0.0368 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 989.49  on 974  degrees of freedom
```

```
## Residual deviance: 725.93  on 970  degrees of freedom
## AIC: 735.93
##
## Number of Fisher Scoring iterations: 5
```

Examining CIs of e^coefficients

```
exp(confint(fit2.A))
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %      97.5 %
## (Intercept) 0.0008161993 0.008672982
## agegp       1.8269388662 2.527847928
## alcgp       2.9153092216 6.546692913
## tobgp       1.2702755728 1.838997507
## lowAlc      1.0523284477 4.786066034
```

Now, lets run a likelihood ratio test to compare our resulting model from our first fit (casestatus = B0+B1*lowAlc) to the null, which assumes all Beta coefficients are 0 beyond the intercepts (casestatus = B0). We'll also examine our second fit (casestatus = B0+B1*lowAlc+B2*agegp*tobgp) against the null.

```
lrtest(fit1.A)
```

```
## Likelihood ratio test
##
## Model 1: casestatus ~ lowAlc
## Model 2: casestatus ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   2 -446.53
## 2   1 -494.74 -1 96.433  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(fit2.A)
```

```
## Likelihood ratio test
##
## Model 1: casestatus ~ agegp + alcgp + tobgp + lowAlc
## Model 2: casestatus ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -362.97
## 2   1 -494.74 -4 263.55  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpreting the LR test, we see that both models are significantly different from the null, and because they have lower log likelihoods, they are significantly better predictors of casestatus than the null.

Question 1B

In this subpart we'll need to do much of the same with slightly different data utilizing dummy variables for alcohol consumption as a categorical variable.

```
data.1B =
    dummy.data.frame(data=data, names=c("alcgp")) %>%
    select(-one_of("X", "alcgp0")) #Dropping alcgp0 as the reference group
```

```r
data.1B %>% head
```

```
##   agegp alcgp1 alcgp2 alcgp3 tobgp casestatus
## 1     0      0      0      0     0          0
## 2     0      0      0      0     0          0
## 3     0      0      0      0     0          0
## 4     0      0      0      0     0          0
## 5     0      0      0      0     0          0
## 6     0      0      0      0     0          0
```

Now, let's run logistic regression (univariate and multiple)

```r
fit1.B = glm(formula=casestatus~alcgp1+alcgp2+alcgp3, family="binomial", data=data.1B)
fit2.B = glm(formula=casestatus~., family="binomial", data=data.1B)

# Reporting Log Odds Ratios (Model fit)
summary(fit1.B)
```

```
##
## Call:
## glm(formula = casestatus ~ alcgp1 + alcgp2 + alcgp3, family = "binomial",
##     data = data.1B)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4924  -0.6889  -0.3806  -0.3806   2.3069
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.5885     0.1925 -13.444  < 2e-16 ***
## alcgp1        1.2712     0.2323   5.472 4.46e-08 ***
## alcgp2        2.0545     0.2611   7.868 3.59e-15 ***
## alcgp3        3.3042     0.3237  10.209  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 989.49  on 974  degrees of freedom
## Residual deviance: 842.99  on 971  degrees of freedom
## AIC: 850.99
##
## Number of Fisher Scoring iterations: 5
```

```r
# Reporting Odds Ratios (Coefficients)
exp(coef(fit1.B))
```

```
## (Intercept)      alcgp1      alcgp2      alcgp3
##  0.07512953  3.56527094  7.80261593 27.22570533
```

```r
#Reporting Confidence Intervals of Odds Ratios
exp(confint(fit1.B))
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %     97.5 %
## (Intercept)  0.05039729  0.1075009
```

```
## alcgp1       2.28528386  5.6990305
## alcgp2       4.71154372 13.1507976
## alcgp3      14.65657398 52.3110864
```

```r
# Reporting Log Odds Ratios (Model fit)
summary(fit2.B)
```

```
##
## Call:
## glm(formula = casestatus ~ ., family = "binomial", data = data.1B)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0694  -0.5940  -0.3387  -0.1354   2.8096
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.11205    0.37124 -13.770  < 2e-16 ***
## agegp        0.76073    0.08273   9.196  < 2e-16 ***
## alcgp1       1.49515    0.25176   5.939 2.87e-09 ***
## alcgp2       2.16355    0.28267   7.654 1.95e-14 ***
## alcgp3       3.57447    0.36247   9.861  < 2e-16 ***
## tobgp        0.42389    0.09420   4.500 6.80e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 989.49  on 974  degrees of freedom
## Residual deviance: 725.90  on 969  degrees of freedom
## AIC: 737.9
##
## Number of Fisher Scoring iterations: 5
```

```r
# Reporting Odds Ratios (Coefficients)
exp(coef(fit2.B))
```

```
##  (Intercept)        agegp       alcgp1       alcgp2       alcgp3
##  0.006023738  2.139845720  4.460000996  8.701936590 35.675544689
##        tobgp
##  1.527886396
```

```r
#Reporting Confidence Intervals of Odds Ratios
exp(confint(fit2.B))
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %      97.5 %
## (Intercept)  0.002817612  0.01209824
## agegp        1.827278688  2.52849920
## alcgp1       2.753714906  7.40827977
## alcgp2       5.044255550 15.31974148
## alcgp3      17.854914351 74.19286697
## tobgp        1.270745947  1.83952860
```

Now, lets run a likelihood ratio test to compare our resulting model from our first fit (casestatus = B0+B1*alcgp1+B2*alcgp2+B3*alcgp3) to the null, which assumes all Beta coefficients are 0 beyond the intercepts (casestatus = B0). We'll also examine our second fit (casestatus = B0+B1*agegp+B2*alcgp1+B3*alcgp2+B4*alcgp3+B5*tobgp) against the null.

```
lrtest(fit1.B)
```

```
## Likelihood ratio test
##
## Model 1: casestatus ~ alcgp1 + alcgp2 + alcgp3
## Model 2: casestatus ~ 1
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   4 -421.50
## 2   1 -494.74 -3 146.5  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(fit2.B)
```

```
## Likelihood ratio test
##
## Model 1: casestatus ~ agegp + alcgp1 + alcgp2 + alcgp3 + tobgp
## Model 2: casestatus ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -362.95
## 2   1 -494.74 -5 263.59  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpreting the LR test, we see that both models are significantly different from the null, and because they have lower log likelihoods, they are significantly better predictors of casestatus than the null.

Question 1C

This data requires the least cleaning of any question so far – we are using the given structure.

```
data.1C =
    data %>%
    select(-X)

data.1C %>% head
```

```
##   agegp alcgp tobgp casestatus
## 1     0     0     0          0
## 2     0     0     0          0
## 3     0     0     0          0
## 4     0     0     0          0
## 5     0     0     0          0
## 6     0     0     0          0
```

Now, let's run logistic regression (univariate and multiple)

```
fit1.C = glm(formula=casestatus~alcgp, family="binomial", data=data.1C)
fit2.C = glm(formula=casestatus~., family="binomial", data=data.1C)

# Reporting Log Odds Ratios (Model fit)
summary(fit1.C)
```

```
##
```

```
## Call:
## glm(formula = casestatus ~ alcgp, family = "binomial", data = data.1C)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4660  -0.6531  -0.4004  -0.4004   2.2643
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.4834     0.1459  -17.02   <2e-16 ***
## alcgp         1.0468     0.0935   11.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 989.49  on 974  degrees of freedom
## Residual deviance: 844.85  on 973  degrees of freedom
## AIC: 848.85
##
## Number of Fisher Scoring iterations: 4
```

```r
# Reporting Odds Ratios (Coefficients)
exp(coef(fit1.C))
```

```
## (Intercept)       alcgp
##  0.08346306  2.84844263
```

```r
#Reporting Confidence Intervals of Odds Ratios
exp(confint(fit1.C))
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %    97.5 %
## (Intercept) 0.06214785 0.1101768
## alcgp       2.37931880 3.4344229
```

```r
# Reporting Log Odds Ratios (Model fit)
summary(fit2.C)
```

```
##
## Call:
## glm(formula = casestatus ~ ., family = "binomial", data = data.1C)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0388  -0.6196  -0.3685  -0.1519   2.7336
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.88680    0.33604 -14.542  < 2e-16 ***
## agegp        0.74375    0.08178   9.094  < 2e-16 ***
## alcgp        1.10255    0.10317  10.687  < 2e-16 ***
## tobgp        0.43085    0.09393   4.587  4.5e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 989.49  on 974  degrees of freedom
## Residual deviance: 730.31  on 971  degrees of freedom
## AIC: 738.31
##
## Number of Fisher Scoring iterations: 5
```

```r
# Reporting Odds Ratios (Coefficients)
exp(coef(fit2.C))
```

```
## (Intercept)       agegp       alcgp       tobgp
## 0.007545561 2.103812880 3.011850604 1.538565912
```

```r
#Reporting Confidence Intervals of Odds Ratios
exp(confint(fit2.C))
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %     97.5 %
## (Intercept) 0.003793252 0.01418649
## agegp       1.799630604 2.48097955
## alcgp       2.470896204 3.70462950
## tobgp       1.280354429 1.85146937
```

Now, lets run a likelihood ratio test to compare our resulting model from our first fit (casestatus = B0+B1*alcgp) to the null, which assumes all Beta coefficients are 0 beyond the intercepts (casestatus = B0). We'll also examine our second fit (casestatus = B0+B1*agegp+B2*alcgp+B3*tobgp) against the null.

```r
lrtest(fit1.C)
```

```
## Likelihood ratio test
##
## Model 1: casestatus ~ alcgp
## Model 2: casestatus ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   2 -422.42
## 2   1 -494.74 -1 144.64  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
lrtest(fit2.C)
```

```
## Likelihood ratio test
##
## Model 1: casestatus ~ agegp + alcgp + tobgp
## Model 2: casestatus ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -365.16
## 2   1 -494.74 -3 259.17  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpreting the LR test, we see that both models are significantly different from the null, and because they have lower log likelihoods, they are significantly better predictors of casestatus than the null.

Question 1D

Now, lets run a likelihood ratio test to compare our models from part B and C

```r
lrtest(fit1.B, fit1.C)
```

```
## Likelihood ratio test
##
## Model 1: casestatus ~ alcgp1 + alcgp2 + alcgp3
## Model 2: casestatus ~ alcgp
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -421.50
## 2   2 -422.42 -2 1.8583     0.3949
```

```r
lrtest(fit2.B, fit2.C)
```

```
## Likelihood ratio test
##
## Model 1: casestatus ~ agegp + alcgp1 + alcgp2 + alcgp3 + tobgp
## Model 2: casestatus ~ agegp + alcgp + tobgp
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -362.95
## 2   4 -365.16 -2 4.4183     0.1098
```