Applied GLM Final Project

Lauren Bergam

April 21, 2019

```r
library(ggplot2)

load("alzheimers_data_filtered.Rdata")
nacc <- alzheimers_data_filtered
```

```r
# Clean data, errant value for apoe_01
nacc <- nacc[(nacc$apoe_01 == 0 | nacc$apoe_01 == 1), ]

# Contingency table for apoe and dementia diagnosis
table_apoe01 <- table("APOE Carrier" = factor(nacc$apoe_01, levels = c("1", "0")),
                      "Dementia" = factor(nacc$DEMENTED, levels = c("1", "0")))
table_apoe01
```

```
##              Dementia
## APOE Carrier    1    0
##            1  125 2273
##            0  124 5162
```

```r
# Fisher's test
fisher.test(table_apoe01)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table_apoe01
## p-value = 2.344e-10
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.762173 2.973824
## sample estimates:
## odds ratio
##   2.289073
```

Interpretation: The odds of having dementia is 2.3 times more likely for individuals who are a carrier of the apolipoprotein E than those who are non-carriers. The chi-squared test gives a very small p-value and therefore, dementia and apoe carriers are not independent from each other.

```r
# Contingency table for age
nacc["age_threshold"] = ifelse(nacc$age_visit > 70, 1, 0)
age_table <- table("Age > 70" = factor(nacc$age_threshold, levels = c("1", "0")),
                   "Dementia" =  factor(nacc$DEMENTED, levels = c("1", "0")))
age_table
```

```
##         Dementia
## Age > 70    1    0
##        1  242 4577
##        0    7 2858
```

```r
# Fisher's test
fisher.test(age_table)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  age_table
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  10.29485 54.42148
## sample estimates:
## odds ratio
##    21.5832
```

Since age is a discrete, non-binary variable, we choose a split value (in this case 70) to classify the data into two groups for a contingency table. Interpretation: The odds of having dementia is 21.6 times more likely for individuals over the age of 70 than those who are equal to or under the age of 70. The chi-squared test gives a very small p-value and therefore, dementia and age are not independent from each other.

```r
# Contingency table for animal test
nacc["animal_threshold"] = ifelse(nacc$animal < 10, 1, 0)
animal_table <- table("Animal < 10 " = factor(nacc$animal_threshold, levels = c("1", "0")),
                      "Dementia" =  factor(nacc$DEMENTED, levels = c("1", "0")))
animal_table
```

```
##             Dementia
## Animal < 10     1    0
##           1   99  161
##           0  150 7274
```

```r
# Fisher's Test
fisher.test(animal_table)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  animal_table
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  21.8292 40.5110
## sample estimates:
## odds ratio
##   29.76251
```

Since number of animals recalled is a discrete, non-binary variable, we choose a split value (in this case 10) to classify the data into two groups. Interpretation: The odds of having dementia is 29.8 times more likely for individuals who can recall fewer than 10 animals within one minute than those who can recall 10 or more. The chi-squared test gives a very small p-value and therefore, dementia and the results from the animal test are not independent from each other.

```r
# Contingency table for years of education
nacc["educ_threshold"] = ifelse(nacc$educ < 12, 1, 0)
educ_threshold <- table("# of years of education < 12 " = factor(nacc$educ_threshold, levels = c("1", "0
                        "Dementia" =  factor(nacc$DEMENTED, levels = c("1", "0")))
educ_threshold
```

```
##                               Dementia
## # of years of education < 12     1    0
##                              1   19  275
##                              0  230 7160
```

```r
# Fisher's Test
fisher.test(educ_threshold)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  educ_threshold
## p-value = 0.003705
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.251961 3.501128
## sample estimates:
## odds ratio
##   2.150489
```

```r
# Chi-squared test
chisq.test(educ_threshold)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  educ_threshold
## X-squared = 9.0815, df = 1, p-value = 0.002582
```

Interpretation: The odds of having dementia is 2.2 times more likely for individuals who have fewer than 12 years of education than those who have 8 years or more. The chi-squared test gives a very small p-value and therefore, dementia and years of education are not independent from each other.

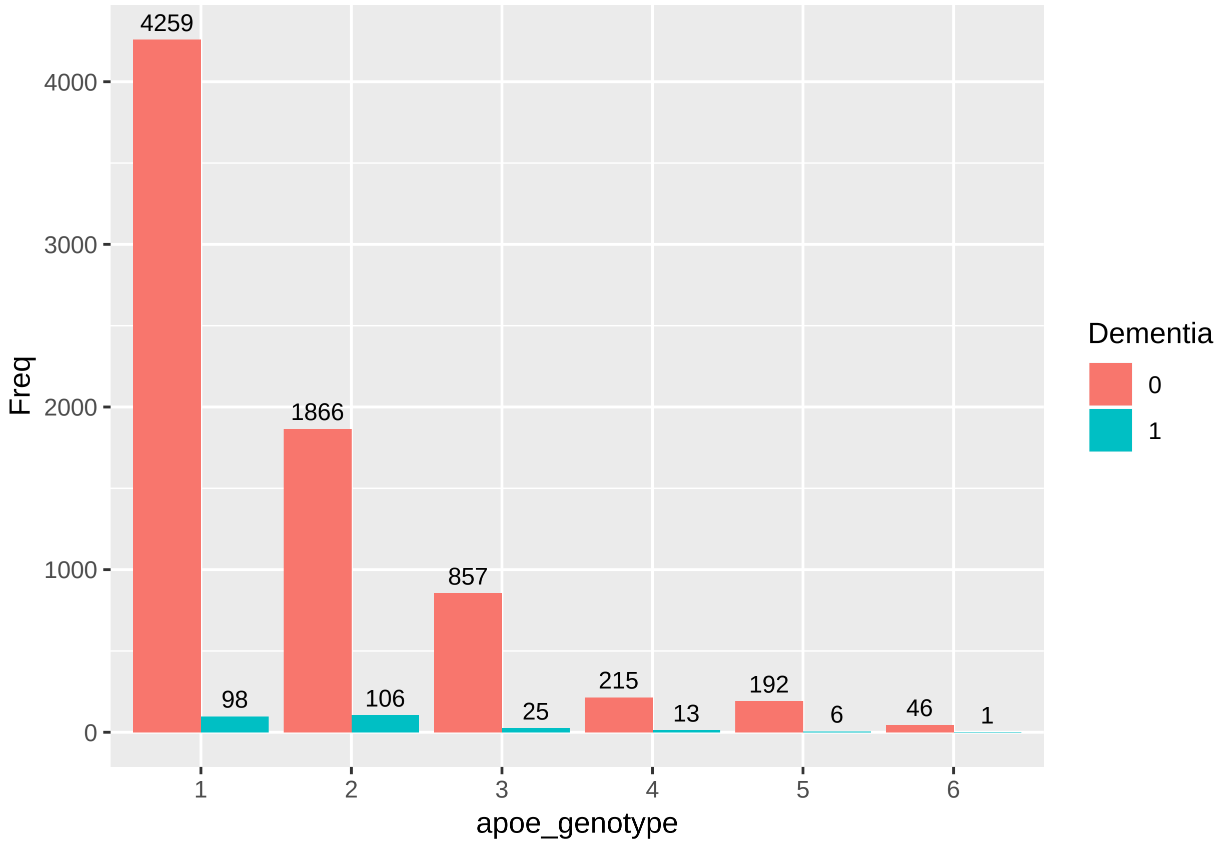
```r
# 6x2 contingency table for APOE genotype
apoe_table <- table(apoe_genotype = factor(nacc$apoe), Dementia = factor (nacc$DEMENTED))
apoe_table
```

```
##              Dementia
## apoe_genotype    0    1
##             1 4259   98
##             2 1866  106
##             3  857   25
##             4  215   13
##             5  192    6
##             6   46    1
```

```r
# Convert to dataframe with frequencies for each permutation
apoe_df <- as.data.frame(apoe_table)
apoe_df
```

```
##    apoe_genotype Dementia Freq
## 1              1        0 4259
## 2              2        0 1866
## 3              3        0  857
## 4              4        0  215
## 5              5        0  192
## 6              6        0   46
## 7              1        1   98
## 8              2        1  106
## 9              3        1   25
## 10             4        1   13
## 11             5        1    6
## 12             6        1    1
```

```r
# Plot frequency table for each genotype
ggplot(apoe_df, aes(x = apoe_genotype, y = Freq, fill = Dementia)) +
  geom_bar(position = "dodge", stat = "identity") +
  geom_text(aes(label = Freq), vjust = -0.5, size = 3, position = position_dodge(0.9))
```



```r
# Model the log odds of having dementia for each genotype compared to the baseline of genotype 1
apoe_model <- glm(Dementia ~ apoe_genotype, weights = Freq, data = apoe_df, family = binomial)
summary(apoe_model)
```

```
##
## Call:
## glm(formula = Dementia ~ apoe_genotype, family = binomial, data = apoe_df,
##     weights = Freq)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -14.3596   -5.5234    0.6842    9.8094   27.2715
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.77182    0.10217 -36.917  < 2e-16 ***
## apoe_genotype2  0.90371    0.14286   6.326 2.52e-10 ***
## apoe_genotype3  0.23726    0.22717   1.044  0.29629
## apoe_genotype4  0.96613    0.30334   3.185  0.00145 **
## apoe_genotype5  0.30609    0.42698   0.717  0.47346
## apoe_genotype6 -0.05682    1.01596  -0.056  0.95540
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2197.7  on 11  degrees of freedom
## Residual deviance: 2154.1  on  6  degrees of freedom
## AIC: 2166.1
##
## Number of Fisher Scoring iterations: 7
```

```r
exp(-3.77182)*100
```

```
## [1] 2.301015
```

```r
exp(-3.77182+0.96613)*100
```

```
## [1] 6.046504
```

```r
exp(-3.77182-0.05682)*100
```

```
## [1] 2.173916
```

Interpretation: The intercept is the log odds of having dementia given the patient has genotype 1. Exponentiating shows that the odds of having dementia given genotype 1 are 2.3%. The odds of the other genotypes are calculated relative to this baseline. The highest odds of having dementia given any of the genotypes is 6.0% for genotype 4 and the lowest odds is 2.1% for genotype 6.