STAA 552: Final Exam Part 2

Your Name Here

Honor Code from Part 1 applies here, too.

DMD (Q5 - Q13)

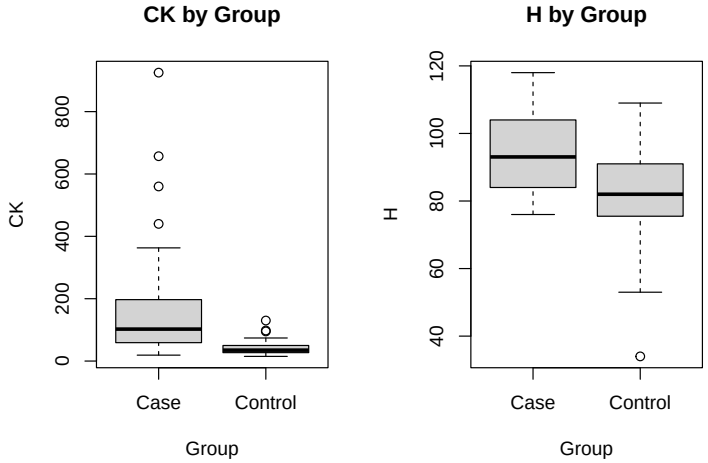
Duchene Muscular Dystrophy (DMD) is a genetically transmitted disease, passed from a mother to her children. Boys with the disease show obvious symptoms. But affected girls usually do not suffer symptoms and hence may unknowingly carry the disease. Doctors must rely on some kind of test to detect the presence of the disease. This data includes n = 120 women, of whom 38 are known DMD carriers and 82 who are not carriers. **DMD carrier is the event of interest.** A blood sample was obtained from each woman and the levels of two enzymes were recorded: creatine kinase (CK) and hemopoxin (H). The data is available from Canvas as DMD.csv.

The data includes the following variables:

- Carrier: 1 if DMD carrier; 0 otherwise.
- Group: Case if DMD carrier; Control otherwise. **This variable is redundant to Carrier, but MAY be helpful for graphing.**
- CK: creatine kinase in units/L.
- H: hemopoxin in units/L.

Q5

Create a summary plots of (a) carrier (or group) and **CK** (b) carrier (or group) and **H**.



Q6

For **Model 1**, fit an appropriate model including only additive effects (no interaction). Show the coefficients table (including coefficient estimates and Wald test p-values). Note: This model will generate a warning about "fitted probabilities numerically 0 or 1". For now, use the model "as is". We will consider alternatives in later questions.

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##
## Call:
## glm(formula = Carrier ~ CK + H, family = binomial, data = DMD)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -16.16695    3.65473  -4.424 9.71e-06 ***
## CK            0.06838    0.01510   4.530 5.91e-06 ***
## H             0.12732    0.03460   3.680 0.000234 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 149.840  on 119  degrees of freedom
## Residual deviance:  62.224  on 117  degrees of freedom
## AIC: 68.224
##
## Number of Fisher Scoring iterations: 8
```

Q7 (0 pts)

For **Model 2**, refit the model using log(CK) and H (no transformation). Show the coefficients table (including parameter estimates and Wald test p-values).
Note: You should no longer get a warning with this model.

```
##
## Call:
## glm(formula = Carrier ~ log(CK) + H, family = binomial, data = DMD)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -28.91340    5.80017  -4.985 6.20e-07 ***
## log(CK)       4.02043    0.82910   4.849 1.24e-06 ***
## H             0.13652    0.03654   3.736 0.000187 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 149.840  on 119  degrees of freedom
## Residual deviance:  61.992  on 117  degrees of freedom
## AIC: 67.992
##
## Number of Fisher Scoring iterations: 7
```

Q8

8a. (2 pts) Use your summary plot from Q5 to briefly justify why the log transformation for **CK** is reasonable. CK had a distribution with a long right tail. This means that most values are lower, but some extreme outliers exist. Doing a log transform may make this more "log normal" as it brings in the long tail ***** Response

---

8b. (2 pts) Using AIC criteria, is Model 1 or Model 2 preferred? Model 2 has a lower AIC and as such is preferred ***** Response

---

Q9 (6 pts)

Using **Model 2**, provide a detailed one sentence **interpretation** of the coefficient corresponding to **H** in context. Also provide a relevant confidence interval.

Interpretation:
For each 1-unit increase in H, the odds of being a DMD carrier are multiplied by .131.

95%CI: 0.07311998 0.21832719

```
##      2.5 %     97.5 %
## 0.07311998 0.21832719
```

Q10

Working with **Model 2**, perform a likelihood ratio test (LRT) corresponding to **H**. Calculate both the test statistic and p-value "by hand" using summary() output.

```r
#Q10
Model2_reduced <- glm(Carrier ~ log(CK), data = DMD, family = binomial)
summary(Model2_reduced)
```

```
##
## Call:
## glm(formula = Carrier ~ log(CK), family = binomial, data = DMD)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -13.2051     2.4132  -5.472 4.45e-08 ***
## log(CK)       3.0759     0.5924   5.192 2.08e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 149.840  on 119  degrees of freedom
## Residual deviance:  86.984  on 118  degrees of freedom
## AIC: 90.984
##
## Number of Fisher Scoring iterations: 6
```

```r
Deviance_full <- deviance(Model2)
Deviance_reduced <- deviance(Model2_reduced)

test_stat <- Deviance_reduced - Deviance_full
p_value <- pchisq(test_stat, df = 1, lower.tail = FALSE)
```

Q11

Using **Model 2**, run an appropriate goodness of fit test and show the results. Make a conclusion in context.

```
## Warning: package 'ResourceSelection' was built under R version 4.4.2
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  as.numeric(DMD$Carrier), fitted(Model2)
## X-squared = 5.9088, df = 8, p-value = 0.6575
```

Conclusion: there is no evidence of lack of fit

Q12 (2 pts)

Using **Model 2**, calculate McFadden's (pseudo) R^2.

```
## [1] 0.5862766
```

Q13

Given that we may want to use this model to make predictions about whether a woman is a DMD carrier, calculate the model **accuracy** using a "standard" cutoff of 0.5.

```
##           Actual
## Predicted  0  1
##         0 78  8
##         1  4 30
```

```
## [1] 0.9
```

Accuracy = .9

Appendix

```
#Retain this code chunk!!!
library(knitr)
library(tidyverse)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)

DMD <- read.csv("DMD.csv")

#Q5

par(mfrow = c(1,2))
boxplot(CK ~ Group, data = DMD, main = "CK by Group",
        xlab = "Group", ylab = "CK")
boxplot(H ~ Group, data = DMD, main = "H by Group",
        xlab = "Group", ylab = "H")
#Q6

Model1 <- glm(Carrier ~ CK + H, data = DMD, family = binomial)
```

```r
summary(Model1)

#Q7
Model2 <- glm(Carrier ~ log(CK) + H, data = DMD, family = binomial)
summary(Model2)

#Q9
confint(Model2, "H")

#Q10
Model2_reduced <- glm(Carrier ~ log(CK), data = DMD, family = binomial)
summary(Model2_reduced)

Deviance_full <- deviance(Model2)
Deviance_reduced <- deviance(Model2_reduced)

test_stat <- Deviance_reduced - Deviance_full
p_value <- pchisq(test_stat, df = 1, lower.tail = FALSE)
#Q11
library(ResourceSelection)
hl_test <- hoslem.test(as.numeric(DMD$Carrier), fitted(Model2), g=10)
hl_test
#Q12
Model_null <- glm(Carrier ~ 1, data = DMD, family = binomial)
logL_full <- logLik(Model2)
logL_null <- logLik(Model_null)

R2_McFadden <- 1 - as.numeric(logL_full / logL_null)
print(R2_McFadden)

#Q13
pred_prob <- predict(Model2, type = "response")

pred_class <- ifelse(pred_prob > 0.5, 1, 0)

confusion_matrix <- table(Predicted = pred_class, Actual = DMD$Carrier)
print(confusion_matrix)

accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
accuracy
```