MSIS 5503 – Statistics for Data Science – Fall 2021-Assignment 13

1) For the stature.dat data set:

   a. Build a Binary Logistic regression model that uses height, hand, foot, to predict whether a person is a female or male.

   Answer:

```
df <- read.table('stature.csv',
                 header = TRUE, sep = ',')
print(head(df))
gender <-df$gender
height <- df$height
hand <- df$hand
foot <- df$foot

# Logistic Regression
#
logimod1 <- glm(gender ~ height+hand+foot, data = df, family =
"binomial")
summary(logimod1)
```

```
> summary(logimod1)

Call:
glm(formula = gender ~ height + hand + foot, family = "binomial",
    data = df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.10364  -0.23614   0.01218   0.18162   1.91500

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -76.158078  13.272097  -5.738 9.57e-09 ***
height        0.033505   0.009335   3.589 0.000332 ***
hand          0.014573   0.040623   0.359 0.719799
foot          0.070514   0.038038   1.854 0.063771 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 214.714  on 154  degrees of freedom
Residual deviance:  66.629  on 151  degrees of freedom
AIC: 74.629

Number of Fisher Scoring iterations: 7
```

   b. Write out the predicting equations for gender and interpret each coefficient in terms of log-odds. In each case, indicate what this means in terms of the probability of being a male (increases, stays the same, or decreases).

   Answer:

The equation for log-odds is
-76.158078 + 0.033505*height + 0.014573*hand + 0.070514*foot

Interpretation:
The change in the log-odds ratio of male, for unit change in height is 0.0335, controlling for hand and foot. This also means that the probability of being male increases, with increase in height. Similarly for hand size and foot size.

c. Calculate the Predicted Probability that a person is Male given height = 1680, hand=200, foot=250.
Confirm the calculation by hand.
Answer:

```
#c prediction
df_a <- data.frame(height=1680, hand=200, foot=250)
p_log_odds <- predict(logimod1, df_a)
p_odds <- exp(p_log_odds)
p_prob <- (p_odds/(1+p_odds))

print(paste("Predicted Log Odds = ", round(p_log_odds,4),
            "Predicted Odds Ratio = ", round(p_odds,4),
            "Predicted Probability of male = ",round(p_prob,4)))
```

```
[1] "Predicted Log Odds =  0.6734 Predicted Odds Ratio =  1.961 Predicted Probability of male =  0.6623"
>
```

Calculation by hand
log odds = -76.158078 + 0.033505*1680 + 0.014573*200 + 0.070514*250 =0.6735
odds ratio = $e^{0.6735}$ = 1.961
predicted probability = 1.961/(1+1.961) =0.6623

d. Produce the classification table for the full model (Model 1) and interpret it. Assume that if predicted probability >= 0.5, the person is a male.
Answer:

```
glm.probs <- predict(logimod1,type = "response")

pred_gender <- ifelse(glm.probs >= 0.5,1,0)
df_class <- cbind(gender, pred_gender)
class_tbl <- xtabs(~ gender + pred_gender, data = df_class)
class_pct <- class_tbl/length(gender)
print(class_tbl)
print(class_pct)
print(paste("Correct classification Rate (percent): ",
            (class_pct[1,1] + class_pct[2,2])*100))
```

```
> class_pct <- class_tbl/length(gender)
> print(class_tbl)
      pred_gender
gender  0  1
     0 66  9
     1  6 74
> print(class_pct)
      pred_gender
gender          0          1
     0 0.42580645 0.05806452
     1 0.03870968 0.47741935
> print(paste("Correct classification Rate (percent): ",
+             (class_pct[1,1] + class_pct[2,2])*100))
[1] "Correct classification Rate (percent):  90.3225806451613"
> 
```

66 out of 75 females were identified correctly, and 74 out of 80 males were identified correctly.

e. Produce the classification table for the model (Model 2) **without** non-significant predictors (at $\alpha = 0.05$) (Model 2) and interpret it. Assume that if predicted probability >= 0.5, the person is a male.
Answer:

```
> print(class_tbl)
      pred_gender
gender  0  1
     0 66  9
     1  9 71
> print(class_pct)
      pred_gender
gender          0          1
     0 0.42580645 0.05806452
     1 0.05806452 0.45806452
> print(paste("Correct classification Rate (percent): ",
+             (class_pct[1,1] + class_pct[2,2])*100))
[1] "Correct classification Rate (percent):  88.3870967741935"
> 
```

66 out of 75 females were identified correctly, and 71 out of 80 males were identified correctly.

f. Which model would you prefer based on the results of the classification tables? Why?
Answer: Based on the classification rate Model 1 would be preferred.

2) (**4 points**) Dataset: hsdemo.csv

For the hsdemo.csv data:

a. Build a Multinomial Logistic Regression model that predicts prog (Program type) using predictor ses, math, science and write. Set the reference category to *General* and for ses, set the reference category to "low".
Answer:

```r
dfm <- read.table('hsbdemo.csv',
                  header = TRUE, sep = ',')

prog <- dfm$prog
ses <- dfm$ses
math <- dfm$math
science <- dfm$science
write <- dfm$write

library(nnet)
#
# Set general as reference category for multinomial dependent
variable prog
#
prog <-factor(prog)
prog <- relevel(prog, ref = "general")
#
# Set low as reference category for predictor variable ses
#
ses <-factor(ses)
ses <- relevel(ses, ref = "low")
#
# Specify the polytomous logistic regression model
#
mmod1 <- multinom(prog ~ ses + math+science+write)
#
# Model summary
summary(mmod1)
```

```
> # Model summary
> summary(mmod1)
Call:
multinom(formula = prog ~ ses + math + science + write)

Coefficients:
         (Intercept)   seshigh sesmiddle        math     science       write
academic   -3.741253 1.1706592 0.4998732  0.11456924 -0.08349463  0.04341104
vocation    4.227472 0.5132541 1.0500691 -0.02626292 -0.04371232 -0.02713430

Std. Errors:
         (Intercept)   seshigh sesmiddle       math    science      write
academic    1.425803 0.5558803 0.4786722 0.03212528 0.02805457 0.02728789
vocation    1.571957 0.6755999 0.5117155 0.03530094 0.02893982 0.02869655

Residual Deviance: 332.0517
AIC: 356.0517
```

b. Write out the predicting equations for each type of program.

$$\ln\left(\frac{\mathbf{P}(\boldsymbol{academic})}{\mathbf{P}(\mathbf{general})}\right) = \mathbf{-3.7412 + 0.1147math - 0.0835science + 0.04341write + 1.1707\ (ses = high) - 0.4999(ses = middle)}$$

$$\ln\left(\frac{\mathbf{P}(\boldsymbol{vocation})}{\mathbf{P}(\mathbf{general})}\right) = \mathbf{4.2275 - 0.0263math - 0.0437science - 0.02713write + 0.5132\ (ses = high) - 1.05007(ses = middle)}$$

c. Interpret the *science score* coefficient and the [*sesmiddle*] coefficient for the academic category for the *academic* program in terms of log-odds. Also interpret in words what this means for the probability of being in the academic program vs the general program.

Answer:

Controlling for math and write scores and for ses, one unit increase in science score results in 0.0835 **decrease** in the log odds of being in academic versus general program.
The probability will decrease by a factor of $e^{0.0835}$, that is it will change from p to p/ $e^{0.0835}$ with one unit increase in science score while controlling for other variables.

Controlling for math , science and write scores relative log odds of being in academic versus general program will **decrease** by 0.4999, if moving from (ses = low) to (ses = medium)
The probability will decrease by a factor of $e^{0.4999}$, that is it will change from p to p/ $e^{0.4999}$ if moving from (ses = low) to (ses = medium) while controlling for other variables

d. Predict the probability of being admitted to the vocation program when ses is "high" and the scores for write, math and science are all 37. Show calculations by hand, and then verify using R.
Answer:

```
#  Probability of being in the three different programs when
math=science=write = 37, ses = high;
#
dses <- data.frame( math = 37,science = 37,write = 37,ses = "high")
print(dses)
predict(mmod1, newdata = dses, "probs")
```

```
> print(dses)
  math science write  ses
1   37      37    37 high
> predict(mmod1, newdata = dses, "probs")
 general academic vocation
0.186758 0.224795 0.588447
>
```

Probability for vocation is 0.588447

- **<u>Calculating by Hand:</u>**
  - $$\ln\left(\frac{\mathrm{P}(academic)}{\mathrm{P(general)}}\right) = -3.7412 + 0.1147\mathrm{math} - 0.0835\mathrm{science} + 0.04341\mathrm{write} + 1.1707\ (\mathrm{ses} = \mathrm{high}) - 0.4999(\mathrm{ses} = \mathrm{middle})$$

$= -3.7412 + 0.1147 * 37 - 0.0835 * 37 + 0.04341 * 37 + 1.1707$
=0.19007

- $\frac{\mathrm{P}(academic)}{\mathrm{P(general)}}$ = exp(0.19007) = 1.20933
- $P(academic)$ = 1.20933*$P(general)$

- Similarly, using
  - $\ln\left(\frac{\mathrm{P}(vocation)}{\mathrm{P(general)}}\right) = 4.2275 - 0.0263\mathrm{math} - 0.0437\mathrm{science} - 0.02713\mathrm{write} + 0.5132\ (\mathrm{ses = high}) - 1.05007\ (\mathrm{ses = middle})$
  - $= 4.2275 - 0.0263 * 37 - 0.0437 * 37 - 0.02713 * 37 + 0.5132$ =1.14689
  - $\frac{\mathrm{P}(vocation)}{\mathrm{P(general)}}$ = exp(1.14689) = 3.14838
  - P(vocation) = 3.14838*$P(general)$
- But $P(academic)$ + P(vocation) + $P(general)$ =1
  - 5.35771*$P(general)$ = 1
  - So, $P(general)$ = 1/5.35771= 0.186646

    So P(vocation) = 3.14838 * 0.186646 = 0.5876