STAT646 - Homework 6

Daniel Osorio - dcosorioh@tamu.edu
Department of Veterinary Integrative Biosciences
Texas A&M University

1. Context = “SNPs” data from package SNPassoc. These are data on 157 case / control individuals, for which we know gender, arterial blood pressure, “protein levels?, and genotype for 35 SNPs.

   (a) Use the association function to test for association between SNP 1 and case / control status, adjusted for gender and blood pressure. Assume a dominant pattern of inheritance, so that both CC and CT are phenotype one and only TT is phenotype two. Match the results by using the glm function to carry out logistic regression. What are the probabilities of being a case for females with arterial blood pressure = 13, comparing phenotype one and two? That is, use your logistic regression model to compute two estimated probabilities, both for females with b.p. = 13, one for phenotype one and one for phenotype two.

```
> library(SNPassoc)
> data(SNPs)
> # Using the package function
> Y <- setupSNP(SNPs, colSNPs = 6:40, info = SNPs.info.pos, sep = "")
> outAssociation <- association(casco ~ sex + blood.pre + snp10001,
+                              data = Y,
+                              model = "dominant")
> outAssociation
```

```
SNP: snp10001  adjusted by: sex blood.pre
          0    %  1    %   OR lower upper p-value   AIC
Dominant
T/T      24 51.1 68 61.8 1.00             0.2286 196.1
C/T-C/C  23 48.9 42 38.2 0.65  0.32  1.31
```

```
> # Using the glm model
> Y <- data.frame("casco" = Y$casco, "sex" = Y$sex,
+                 "blood.pre" = Y$blood.pre,
+                 "snp10001" = ifelse(Y$snp10001 == "T/T", 0, 1))
> glmOut <- glm(casco ~ sex + blood.pre + snp10001,
+               family = binomial, data = Y)
> # Comparing the result
> cf <- summary(glmOut)$coefficients
> exp(cf[4, 1])
```

```
[1] 0.6504373
```

```
> exp(cf[4, 1] + c(-1, 1) * 1.96 * cf[4, 2])
```

```
[1] 0.3230523 1.3095981
```

```
> 2 * (1 - pnorm(abs(cf[4, 1] / cf[4, 2])))

[1] 0.2283576

> # Female case
> newData <- data.frame("sex" = factor(rep("Female", 2)),
+                       "blood.pre" = rep(13, 2),
+                       "snp10001" = c(1, 0))
> predict(glmOut, newdata = newData, type = "response")

        1         2
0.6245024 0.7188599
```

(b) The tableHWE function uses an exact test to test whether Hardy Weinberg Equilibrium (HWE) holds with respect to each SNP. Manually implement the exact test and match the tableHWE p-value for SNP 1. The formula is given in Wigginton et al (2005). You don't need to use the recursion formulas given in equation (2). Note that you will need to compute lots of factorials to get the p-value. Factorials of big numbers can be problematic due to instability of the resulting numbers. I recommend that you use the lfactorial function in R to compute factorials on the log scale.

```
> Y <- setupSNP(SNPs, colSNPs = 6:40, info = SNPs.info.pos, sep = "")
> head(tableHWE(Y))

         HWE (p value)
snp10001   0.281639248
snp10002   0.004944837
snp10003            NA
snp10004            NA
snp10005   0.008019904
snp10006            NA

> testHWE <- function(snp) {
+   N <- length(snp)
+   snp <- as.character(snp)
+   snp_tbl <- table(snp)
+   if(length(snp_tbl) == 3) {
+     if(all(!is.na(snp))) {
+       al <- unlist(strsplit(snp, split = "/"))
+       al_uq <- unique(al)
+       al_tbl <- table(al)
+       oo <- order(al_tbl)
+       A <- names(al_tbl[oo[1]])
+       B <- names(al_tbl[oo[2]])
+       n_A <- al_tbl[oo[1]]
+       n_B <- al_tbl[oo[2]]
+       n_AB <- table(snp)[2]
+       n_AA <- (n_A - n_AB) / 2
+       n_BB = (n_B - n_AB) / 2
+       even <- floor(n_A / 2) == ceiling(n_A / 2)
+       if(even) {
+         n_AB_seq <- seq(from = 0, to = n_A, by = 2)
+       } else {
+         n_AB_seq <- seq(from = 1, to = n_A, by = 2)
+       }
+       k <- length(n_AB_seq)
+       PP <- rep(NA, k)
+       for(i in 1:k) {
+         n_ab <- n_AB_seq[i]
+         n_aa <- (n_A - n_ab) / 2
+         n_bb <- (n_B - n_ab) / 2
+         PP[i] <- exp(n_ab * log(2) + lfactorial(N) +
+                        lfactorial(n_A) + lfactorial(n_B) -
+                        lfactorial(n_aa) - lfactorial(n_ab) -
+                        lfactorial(n_bb) - lfactorial(2 * N))
+       }
+       p_obs <- PP[match(n_AB, n_AB_seq)]
+       p_val <- sum(PP[PP <= p_obs])
+       return(p_val)
+     } else {
+       return(NA) }
+   } else {
+     return(NA)
+   } }
> testHWE(Y$snp10001)
[1] 0.2816392
```

(c) In the notation of Wigginton et al (2005), nA and nB are the allele frequencies for the rarer and more common alleles, respectively; nA + nB = 2N = 2 × 157. Also, nAA is the frequency of individuals homozygous for the rarer allele, nAB is the frequency of heterozygous individuals, and nBB is the frequency of individuals homozygous for the more common allele; nAA + nAB + nBB = N = 157. What are the allele frequencies nA and nB and the genotype frequencies nAA, nAB, and nBB? Considering nA and nB as fixed, and assuming that HWE holds, how many of each genotype would you expect out of our 157 individuals?

```
> snp <- Y$snp10001
> N <- length(snp)
> snp <- as.character(snp)
> snp_tbl <- table(snp)
> al <- unlist(strsplit(snp, split = "/"))
> al_uq <- unique(al)
> al_tbl <- table(al)
> oo <- order(al_tbl)
> A <- names(al_tbl[oo[1]])
> B <- names(al_tbl[oo[2]])
> # Observed alleles
> print(n_A <- al_tbl[oo[1]])
 C
77
> print(n_B <- al_tbl[oo[2]])
  T
237
> # Observed genotypes
> print(n_AB <- table(snp)[2])
```

```
C/T
 53

> print(n_AA <- (n_A - n_AB) / 2)

 C
12

> print(n_BB <-  (n_B - n_AB) / 2)

 T
92

> p_A <- as.numeric(n_A / (2 * N))
> p_B <- as.numeric(1 - p_A)
> # Expected genotypes under HWE
> n_AA_hwe <- ceiling(N * p_A ^ 2)
> names(n_AA_hwe) <- "C/C"
> n_AA_hwe

C/C
 10

> n_AB_hwe <- ceiling(2 * N * p_A * p_B)
> names(n_AB_hwe) <- "C/T"
> n_AB_hwe

C/T
 59

> n_BB_hwe <- ceiling(N * p_B ^ 2)
> names(n_BB_hwe) <- "T/T"
> n_BB_hwe

T/T
 90
```