ASSET(Association analysis for SubSETs) Package

October 15, 2013

Introduction

The ASSET package consists of two main functions: (1) h.traits and (2) h.types. The function h.traits is suitable for conducting meta-analysis of possibly different traits when summary level data are available from individual studies. The function allows for correlation among different studies/traits, which, for example, may arise due to shared subjects across studies. This function can also be used to conduct "meta-analysis" across multiple correlated traits on the same individuals by appropriately specifying the correlation matrix for the multivariate trait. Input arguments to this function are vectors/matrices of the estimated log-odds ratios, standard errors and number of cases and controls for each SNP and study. The function h.types is suitable for analysis of case-control studies when cases consist of distinct disease subtypes. This function assumes individual level data are available. The main input argument for h.types is a data frame containing the SNP variables, response variable and covariates for all subjects.

```
> library(ASSET)
```

Examples of h.traits

Get the path to the data.

```
> datafile <- system.file("sampleData", "vdata.rda", package="ASSET")
```

Load the data frames. There are 4 data frames, data1 - data4 for the 4 independent studies. Each study has the SNPs SNP1-SNP3 genotyped, and information on each subject's age and case-control status. Each SNP is coded as the number of copies of the minor allele or NA for missing genotypes.

```
> load(datafile)
> data1[1:5, ]
```

```
    CC AGE SNP1 SNP2 SNP3
456  1  70    1    0    0
457  1  55    1    0    0
458  1  48    0    0    1
459  1  72    1    0    2
460  1  74    2    0    0
```

```
> SNPs    <- paste("SNP", 1:3, sep="")
> nSNP    <- length(SNPs)
> studies <- paste("STUDY", 1:4, sep="")
> nStudy  <- length(studies)
```

Let us determine the number of non-missing cases and controls for each SNP and study.

```
> case    <- matrix(data=NA, nrow=nSNP, ncol=nStudy)
> control <- matrix(data=NA, nrow=nSNP, ncol=nStudy)
> for (i in 1:nStudy) {
+   data <- eval(parse(text=paste("data", i, sep="")))
+   caseVec <- data[, "CC"] == 1
+   controlVec <- !caseVec
+   for (j in 1:nSNP) {
+     temp <- !is.na(data[, SNPs[j]])
+     case[j, i] <- sum(caseVec & temp, na.rm=TRUE)
+     control[j, i] <- sum(controlVec & temp, na.rm=TRUE)
+   }
+ }
> case
```

```
     [,1] [,2] [,3] [,4]
[1,] 1897 1363 1714  686
[2,] 1909 1369 1726  691
[3,] 1875 1341 1732  696
```

```
> control
```

```
     [,1] [,2] [,3] [,4]
[1,] 1955 1802 1262  667
[2,] 1955 1773 1268  670
[3,] 1925  749 1269  674
```

Run a logistic regression for each SNP and study

```
> beta  <- matrix(data=NA, nrow=nSNP, ncol=nStudy)
> sigma <- matrix(data=NA, nrow=nSNP, ncol=nStudy)
> for (i in 1:nStudy) {
+   data <- eval(parse(text=paste("data", i, sep="")))
+   for (j in 1:nSNP) {
+     data[, "SNP"] <- data[, SNPs[j]]
+     fit <- glm(CC ~ AGE + SNP, data=data, family=binomial())
+     coef <- summary(fit)$coefficients
+     beta[j, i] <- coef["SNP", 1]
+     sigma[j, i] <- coef["SNP", 2]
+   }
+ }
> beta
```

```
            [,1]       [,2]       [,3]       [,4]
[1,]  0.30837615 0.09041508  0.1799979 0.13116360
[2,]  0.09311754 0.20472698  0.1465665 0.05729745
[3,] -0.08212701 0.08909210 -0.0621090 0.01181724
```

```
> sigma

           [,1]       [,2]       [,3]       [,4]
[1,] 0.04637132 0.05410822 0.05931264 0.07970842
[2,] 0.10214703 0.08211686 0.09299885 0.11889584
[3,] 0.04954003 0.07202424 0.05736282 0.08468467

>
```

Call the h.traits function. Since the studies are independent, we do not need to specify the cor option.

```
> res <- h.traits(SNPs, studies, beta, sigma, case, control, meta=TRUE)
```

Compute a summary table. Notice that in the Subset.2sided results, the first 2 SNPs have missing values for OR.2, CI.low.2, and CI.high.2 since the estimated betas were all positive for these SNPs.

```
> h.summary(res)

$Meta
   SNP       Pvalue    OR CI.low CI.high
1 SNP1 3.268265e-12 1.218  1.216   1.220
2 SNP2 3.666743e-03 1.150  1.145   1.156
3 SNP3 2.994911e-01 0.968  0.967   0.970

$Subset.1sided
   SNP       Pvalue    OR CI.low CI.high                Pheno
1 SNP1 2.202379e-11 1.268  1.183   1.359 STUDY1,STUDY3,STUDY4
2 SNP2 3.389452e-02 1.196  1.014   1.412        STUDY2,STUDY3
3 SNP3 3.190294e-01 0.929  0.804   1.074        STUDY1,STUDY3

$Subset.2sided
   SNP       Pvalue     Pvalue.1  Pvalue.2  OR.1 CI.low.1 CI.high.1  OR.2
1 SNP1 5.154931e-11 5.154931e-11 1.0000000 1.268    1.181     1.361    NA
2 SNP2 5.527301e-02 5.527301e-02 1.0000000 1.196    0.996     1.437    NA
3 SNP3 1.582707e-01 3.604528e-01 0.1020498 1.093    0.903     1.323 0.929
  CI.low.2 CI.high.2              Pheno.1       Pheno.2
1       NA        NA STUDY1,STUDY3,STUDY4
2       NA        NA        STUDY2,STUDY3
3    0.851     1.015               STUDY2 STUDY1,STUDY3
```

Intead of searching over all possible subsets, let us define our own subset function to determine which nsubsets to search over. We will only consider subsets where the first m traits are in the subset (m = 1, 2, ...). The DLM p-value will also be computed using only these subsets.

```
> sub.def <- function(logicalVec) {
+   sum <- sum(logicalVec)
+   ret <- all(logicalVec[1:sum])
+   ret
+ }
```

Call the h.traits function with the zmax.args pval.args options defined

```
> res <- h.traits(SNPs, studies, beta, sigma, case, control, meta=TRUE,
+          zmax.args=list(sub.def=sub.def), pval.args=list(sub.def=sub.def))

> h.summary(res)

$Meta
   SNP       Pvalue    OR CI.low CI.high
1 SNP1 3.268265e-12 1.218  1.216   1.220
2 SNP2 3.666743e-03 1.150  1.145   1.156
3 SNP3 2.994911e-01 0.968  0.967   0.970

$Subset.1sided
   SNP       Pvalue    OR CI.low CI.high                       Pheno
1 SNP1 2.096558e-11 1.218  1.150   1.290 STUDY1,STUDY2,STUDY3,STUDY4
2 SNP2 1.430355e-02 1.169  1.032   1.325        STUDY1,STUDY2,STUDY3
3 SNP3 2.477190e-01 0.921  0.801   1.059                      STUDY1

$Subset.2sided
   SNP       Pvalue     Pvalue.1   Pvalue.2  OR.1 CI.low.1 CI.high.1  OR.2
1 SNP1 1.713498e-11 1.713498e-11 1.00000000 1.218    1.150     1.290    NA
2 SNP2 9.590357e-03 9.590357e-03 1.00000000 1.169    1.039     1.316    NA
3 SNP3 6.352379e-02 2.085630e-01 0.05586139 1.093    0.951     1.256 0.929
  CI.low.2 CI.high.2                    Pheno.1       Pheno.2
1       NA        NA STUDY1,STUDY2,STUDY3,STUDY4
2       NA        NA        STUDY1,STUDY2,STUDY3
3    0.862     1.002                      STUDY2 STUDY1,STUDY3
```

Session Information

```
> sessionInfo()

R version 3.0.2 (2013-09-25)
Platform: i386-w64-mingw32/i386 (32-bit)

locale:
[1] LC_COLLATE=C
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:
[1] grid      stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
[1] ASSET_1.0.0 rmeta_2.16  msm_1.2     MASS_7.3-29
```

```
loaded via a namespace (and not attached):
[1] mvtnorm_0.9-9996 splines_3.0.2    survival_2.37-4  tools_3.0.2
```