Missing Data and Latent Variables

1 Case-Case Analysis for Gene-Environment Interactions

Define the following outcomes for an individual:

\[ D = 0, 1 \] disease status,
\[ G = 0, 1 \] genetic status,
\[ E = 0, 1 \] exposure status,

If \( G \) and \( E \) are independent (\([G][E]\)),

\[
Pr(D|GE) = \frac{Pr(GE | D) Pr(D)}{Pr(G) Pr(E)}.
\]

No \( G \times E \) interaction iff \( G \) and \( E \) are independent given \( D \) or \([GD][ED]\). This is equivalent to:

- Regress \( G \) on \( E \), controlling for \( D \), age and other other confounders, to look for \( E \) effect. No \( G \times E \) interaction if coefficients for \( E \) are 0.
- Regress \( E \) on \( G \), controlling for \( D \), age and other other confounders, to look for \( E \) effect. No \( G \times E \) interaction if coefficients for \( G \) are 0.

2 Example

“Sample” of individuals with high risk of breast cancer. In the high-risk clinic, we observe a test result \( T \) for genotype \( G \) given the proband’s age at diagnosis \( A \), disease status \( D \) and family history \( F \). In a gene by environmental interaction study, data on the individual’s environmental exposure, \( E \), is also recorded along with a variety of other covariates \( X \) thought to affect on risk. The appropriate likelihood for an individual is

\[ P(TEX|F,A,D).\]

Modelling the high-dimensional joint distribution of \( X \) may prove troublesome (continuous and categorical variables), and can be avoided by conditioning on it and using the likelihood

\[ P(TE|X,F,A,D),\]

however, what we are really interested in is

\[ P(GE|X,F,A,D) \]

but \( G \) is unobservable (currently)!

Because \( G \) is unobservable, we will focus attention on a model \( E|G,F,A,D,X \). In the example, the environmental risk factor is Alcohol Consumption. \( E = \{\text{drink, Iboozgrms}\} \), an indicator for whether an individual drinks alcohol, and the log of the amount (if a drinker) in grams/day.

Problems:

- genotype is unobserved; do observe test result for presence of BRCA1/2 mutations and have a probability model (BRCAPRO) for \( P(G|FAD) \)
- missing data; in \( E \) and \( X \)
3 Missing Data and Latent Variables

In a Bayesian analysis, missing data and latent variables are simply treated as parameters and require a prior distribution to quantify uncertainty regarding their values. Missing observations can be thought of as a special case of latent data.

3.1 Missing Data

Let $Y_o$ denote the observed responses, and $Y_m$ the missing or latent data, and $Y_c$ denote the “complete” or augmented data $Y_c = (Y_o', Y_m')$. Often the analysis is easier given the complete data, but plugging in any old set of values for the missing data leads to the wrong measures of uncertainty. In a Bayesian analysis we can avoid this problem.

The posterior distribution for $\beta$ given the observed data is

$$
\pi(\beta | Y_o, X) = \int \pi(\beta | Y_o, Y_m, X) \pi(Y_m | Y_o, X) dY_m
$$

where

$$
\pi(Y_m | Y_o, X) = \int f(Y_m | \beta, X) \pi(\beta | Y_o, X) d\beta.
$$

If the result is not tractable, this can be easily obtained by MCMC

$$
\pi(\beta | Y_o, X) = \sum_i \pi(\beta | Y_o, Y_m^{(i)}, X)
$$

where $Y_m$ is generated from the distribution $Y_m | \beta^{(i)}$ and $\beta | Y_o, Y_m$.

For missing covariates, one needs to explicitly specify a “prior” model for the data. The missing cases are then generated based on full conditional distributions.
4 BUGS MODEL

model LATENT;
const n=207;
var smokefbcc[n],othrace[n],vocotech[n],college[n],age1[n],age2[n],age3[n],
genotype[n],lpbrca12[n],pbrca12[n],drink[n],lboozgms[n],testres[n],
state.gen[n],state.dr[n],mu[n],mudrink[n],pi[n],
a[10],b[10],c[2],tau[2],pothr,psmoke;
data in "bugsnn";
isits in "cgenv-inits.txt";
{
for( i in 1 : n ) {
smokefbcc[i] ~ dbern(psmoke)
othrace[i] ~ dbern(pothr)
state.gen[i] <- genotype[i] + 1
testres[i] ~ dbern(c[state.gen[i]])
genotype[i] ~ dbern(pbrca12[i])
drink[i] ~ dbern(pi[i])
state.dr[i] <- drink[i] + 1
lboozgms[i] ~ dnorm(mudrink[i], tau[state.dr[i]])
mu[i] <- a[1] + a[2]*(smokefbcc[i]-mean(smokefbcc[]))
+ a[3]*(othrace[i]-mean(othrace[]))
+ a[4]*(vocotech[i]-mean(vocotech[]))
+ a[5]*(college[i]-mean(college[]))
+ a[6]*age1[i] + a[7]*age2[i] + a[8]*age3[i]
+ a[9]*(genotype[i]-mean(genotype[]))
+ a[10]*(lpbrca12[i]-mean(lpbrca12[]))
mudrink[i] <- mu[i]*drink[i]
logit(pii[i]) <- b[1] + b[2]*(smokefbcc[i]-mean(smokefbcc[]))
+ b[3]*(othrace[i]-mean(othrace[]))
+ b[4]*(vocotech[i]-mean(vocotech[]))
+ b[5]*(college[i]-mean(college[]))
+ (genotype[i]-mean(genotype[]))*b[9]
+ (lpbrca12[i]-mean(lpbrca12[]))*b[10]
}
for (j in 1:10){
b[j] ~ dnorm(0.0, .0001)
a[j] ~ dnorm(0.0,1.0E-04)
}
c[1] ~ dbeta(1.0, 999.00)
c[2] ~ dbeta(3.0, 1.0)
tau[2] ~ dgamma(.001, .001)
tau[1] ~ dgamma(1, 1)
psmoke ~ dbeta(5.7, 4.3)
pothr ~ dbeta(1, 4)
}

#list(b=c(1.9,1.6,-1.45, .22, .31, -8.9, 6.0,.55,-.030, 0.0),
  a=c(3.55,.05,-.7,-.5,-.04,-2.2,1.3,-.29,.01,-.02),
  tau=c(.001,1.09))
5 Bugs Output

Bugs> stats(c)

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>2.5%</th>
<th>97.5%</th>
<th>CI</th>
<th>median</th>
<th>sample</th>
</tr>
</thead>
</table>

Bugs> stats(a)

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>2.5%</th>
<th>97.5%</th>
<th>CI</th>
<th>median</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>3.550E+0</td>
<td>8.478E-2</td>
<td>3.382E+0</td>
<td>3.712E+0</td>
<td>3.548E+0</td>
<td></td>
<td>1100</td>
</tr>
<tr>
<td>[3]</td>
<td>-7.664E-1</td>
<td>5.188E-1</td>
<td>-1.744E+0</td>
<td>2.682E-1</td>
<td>-7.940E-1</td>
<td></td>
<td>1100</td>
</tr>
<tr>
<td>[6]</td>
<td>-1.954E+0</td>
<td>1.462E+0</td>
<td>-4.785E+0</td>
<td>8.569E-1</td>
<td>-1.891E+0</td>
<td></td>
<td>1100</td>
</tr>
<tr>
<td>[7]</td>
<td>4.418E-1</td>
<td>2.468E+0</td>
<td>-4.222E+0</td>
<td>5.172E+0</td>
<td>4.145E-1</td>
<td></td>
<td>1100</td>
</tr>
<tr>
<td>[8]</td>
<td>-2.404E0</td>
<td>2.270E0</td>
<td>-6.930E0</td>
<td>2.003E0</td>
<td>-2.390E0</td>
<td></td>
<td>1100</td>
</tr>
</tbody>
</table>

Bugs> stats(b)

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>2.5%</th>
<th>97.5%</th>
<th>CI</th>
<th>median</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>2.186E0</td>
<td>2.992E-1</td>
<td>1.612E+0</td>
<td>2.802E+0</td>
<td>2.177E+0</td>
<td></td>
<td>1100</td>
</tr>
<tr>
<td>[2]</td>
<td>1.812E0</td>
<td>5.197E-1</td>
<td>8.117E-1</td>
<td>2.887E+0</td>
<td>1.799E+0</td>
<td></td>
<td>1100</td>
</tr>
<tr>
<td>[3]</td>
<td>-1.320E0</td>
<td>9.690E-1</td>
<td>-3.240E+0</td>
<td>6.094E-1</td>
<td>-1.335E+0</td>
<td></td>
<td>1100</td>
</tr>
<tr>
<td>[5]</td>
<td>3.819E-1</td>
<td>6.009E-1</td>
<td>-7.413E-1</td>
<td>1.538E+0</td>
<td>3.831E-1</td>
<td></td>
<td>1100</td>
</tr>
<tr>
<td>[7]</td>
<td>1.375E+1</td>
<td>1.014E+1</td>
<td>-4.644E+1</td>
<td>3.694E+1</td>
<td>1.176E+1</td>
<td></td>
<td>1100</td>
</tr>
<tr>
<td>[8]</td>
<td>-5.120E0</td>
<td>7.748E+0</td>
<td>-2.220E+1</td>
<td>6.687E+0</td>
<td>-3.610E+0</td>
<td></td>
<td>1100</td>
</tr>
<tr>
<td>[9]</td>
<td>3.911E-1</td>
<td>1.051E+0</td>
<td>-1.746E+0</td>
<td>2.449E+0</td>
<td>3.941E-1</td>
<td></td>
<td>1100</td>
</tr>
<tr>
<td>[10]</td>
<td>-5.504E-2</td>
<td>1.150E-1</td>
<td>-2.611E-1</td>
<td>1.730E-1</td>
<td>-5.837E-2</td>
<td></td>
<td>1100</td>
</tr>
</tbody>
</table>

Bugs> stats(tau)

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>2.5%</th>
<th>97.5%</th>
<th>CI</th>
<th>median</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>1.805E+1</td>
<td>4.203E+0</td>
<td>1.066E+1</td>
<td>2.703E+1</td>
<td>1.784E+1</td>
<td></td>
<td>1100</td>
</tr>
<tr>
<td>[2]</td>
<td>1.364E+0</td>
<td>1.842E-1</td>
<td>1.028E+0</td>
<td>1.730E+0</td>
<td>1.359E+0</td>
<td></td>
<td>1100</td>
</tr>
</tbody>
</table>