6 `Cervix`: case-control study with errors in covariates

Carroll *et al.* (1993) consider the problem of estimating the odds ratio of a disease d in a case-control study where the binary exposure variable is measured with error. Their example concerns exposure to herpes simplex virus (HSV) in women with invasive cervical cancer ($d = 1$) and in controls ($d = 0$). Exposure to HSV is measured by a relatively inaccurate western blot procedure w for 1929 of the 2044 women, whilst for 115 women, it is also measured by a refined or "gold standard" method x. The data are given in the table below. They show a substantial amount of misclassification, as indicated by low sensitivity and specificity of w in the "complete" data, and Carroll *et al.* (1993) also found that the degree of misclassification was significantly higher for the controls than for the cases (p=0.049 by Fisher's exact test).

	d	x	w	Count
Complete data				
	1	0	0	13
	1	0	1	3
	1	1	0	5
	1	1	1	18
	0	0	0	33
	0	0	1	11
	0	1	0	16
	0	1	1	16
Incomplete data				
	1		0	318
	1		1	375
	1		0	701
	1		1	535

They fitted a prospective logistic model to the case-control data as follows

$$\begin{aligned} d_i &\sim \text{Bernoulli}(p_i) & i = 1, ..., 2044 \\ \text{logit}(p_i) &= \beta_{0C} + \beta x_i & i = 1, ..., 2044 \end{aligned}$$

where β is the log odds ratio of disease d. Since the relationship between d and x is only directly observable in the 115 women with "complete" data, and because there is evidence of differential measurement error, the following parameters are required in order to estimate the logistic model

$$\begin{aligned} \phi_{1,1} &= \text{P}(w = 1 \mid x = 0, d = 0) \\ \phi_{1,2} &= \text{P}(w = 1 \mid x = 0, d = 1) \\ \phi_{2,1} &= \text{P}(w = 1 \mid x = 1, d = 0) \\ \phi_{2,2} &= \text{P}(w = 1 \mid x = 1, d = 1) \\ q &= \text{P}(x = 1) \end{aligned}$$

The differential probability of being exposed to HSV ($x = 1$) for cases and controls is calculated as follows

$$
\begin{aligned}
\gamma_1 &= \mathrm{P}(x=1 \mid d=1) \\
&= \frac{\mathrm{P}(d=1 \mid x=1)\mathrm{P}(x=1)}{\mathrm{P}(d=1)} \\
&= \frac{1}{1+\frac{1+e^{\beta_{0c}+\beta}}{1+e^{\beta_{0c}}}\frac{1-q}{q}} \\
\gamma_2 &= \mathrm{P}(x=1 \mid d=0) \\
&= \frac{\mathrm{P}(d=0 \mid x=1)\mathrm{P}(x=1)}{\mathrm{P}(d=0)} \\
&= \frac{1}{1+\frac{1+e^{-\beta_{0c}-\beta}}{1+e^{-\beta_{0c}}}\frac{1-q}{q}}
\end{aligned}
$$

The graph for the above model is in Figure 6.

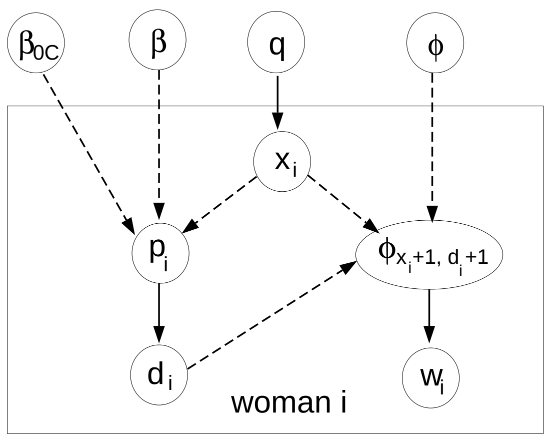

Figure 6: Graphical model for `cervix` example.

The role of the variables `x1` and `d1` is to pick the appropriate value of `phi` (the incidence of `w`) for any given true exposure status `x` and disease status `d`. Since `x` and `d` take the values 0 or 1, and the subscripts for `phi` take values 1 or 2, we must first add 1 to each `x[i]` and `d[i]` before using them as index values for `phi`. `BUGS` does not allow subscripts to be functions of variable quantities — hence the need to create `x1` and `d1` for use as subscripts. In addition, note that γ_1 and γ_2 were not simulated directly in `BUGS`, but were calculated as functions of other parameters. This is because the dependence of γ_1 and γ_2 on d would have led to a cycle in the graphical model which would no longer define a probability distribution.

Cervix: model specification in BUGS

```
model cervix;
const
   N = 2044;  # number of observations
var
     x[N], x1[N],  # ‘true’ HSV status (x[i] + 1)
     d[N], d1[N],  #  cancer status (d[i] + 1)
            p[N],  #  prob of case
               q,  #  incidence of HSV
  w[N], phi[2,2],  #  approx HSV status; rates for w being positive
    beta0C, beta,  #  intercept and log-odds ratio
  gamma1, gamma2;  #  prob HSV positive given control or case
data d, x, w in "cervix.dat";
inits in "cervix.in";
{
 for (i in 1:N) {
         x[i]   ~ dbern(q);           # incidence of HSV
   logit(p[i]) <- beta0C + beta*x[i]; # logistic model
          d[i]  ~ dbern(p[i]);        # incidence of cancer
         x1[i] <- x[i]+1; d1[i] <- d[i]+1;
          w[i]  ~ dbern(phi[x1[i],d1[i]]); # incidence of w
 }
 q      ~ dunif(0.0,1.0);           # prior distribution
 beta0C ~ dnorm(0.0,0.00001); beta   ~ dnorm(0.0,0.00001);
 for(j in 1:2) {
    for(k in 1:2){ phi[j,k] ~ dunif(0.0,1.0);  }
 }
# calculate gamma1 = P(x=1|d=0) and gamma2 = P(x=1|d=1)
  gamma1 <- 1/(1 + (1+exp(beta0C+beta))/(1+exp(beta0C)) * (1-q)/q);
  gamma2 <- 1/(1 + (1+exp(-beta0C-beta))/(1+exp(-beta0C)) * (1-q)/q);
}
```

Analysis

BUGS took 8 minutes to run for 1000 iterations, following a 500 iteration burn-in. The posterior means and standard errors are shown in the table below, and are compared to the pseudolikelihood (*PSL*) estimates obtained by Carroll *et al.* (1993).

		BUGS		*PSL*	
Parameter		mean	(S.E.)	estimate	(S.E.)
β_{0C}		-0.953	(0.240)	-0.981	(0.185)
β	(log odds ratio)	0.690	(0.416)	0.622	(0.355)
$\phi_{1,1}$	$\mathrm{P}(w=1 \mid x=0, d=0)$	0.307	(0.047)	0.317	(0.057)
$\phi_{1,2}$	$\mathrm{P}(w=1 \mid x=0, d=1)$	0.222	(0.084)	0.195	(0.089)
$\phi_{2,1}$	$\mathrm{P}(w=1 \mid x=1, d=0)$	0.586	(0.065)	0.577	(0.067)
$\phi_{2,2}$	$\mathrm{P}(w=1 \mid x=1, d=1)$	0.749	(0.067)	0.790	(0.067)
γ_1	$\mathrm{P}(x=1 \mid d=0)$	0.441	(0.053)	0.421	(0.057)
γ_2	$\mathrm{P}(x=1 \mid d=1)$	0.608	(0.076)	0.590	(0.079)