Multi-response models

Jarrod Hadfield (j.hadfield@ed.ac.uk)

November 10, 2009

So far we have only fitted models to a single response variable. Multi-response models are not that widely used, except perhaps in quantitative genetics, and deserve wider use. They allow some of the assumptions of single response models to be relaxed and can be an effective way of dealing with missing data problems.

1 Relaxing the univariate assumptions of causality

Imagine we knew how much money 200 people had spent on their holiday and on their car in each of four years, and we want to know whether a relationship exists between the two. A simple correlation would be one possibility, but then how do we control for the repeated measures? An often used solution to this problem is to choose one variable as the response (lets say the amount spent on a car) and have the other variable as a fixed covariate (the amount spent on a holiday). The choice is essentially arbitrary, highlighting the belief that any relationship between the two types of spending maybe in part due to unmeasured variables, rather than being completely causal.

In practice does this matter? Lets imagine there was only one unmeasured variable: disposable income. There are repeatable differences between individuals in their disposable income, but also some variation within individuals across the four years. Likewise, people vary in what proportion of their disposable income they are willing to spend on a holiday versus a car, but this also changes from year to year. We can simulate some toy data to get a feel for the issues:

```
> id<-gl(200,4)                  # 200 people recorded four times
> av_wealth<-rlnorm(200, 0, 1)
> ac_wealth<-av_wealth[id]+rlnorm(800, 0, 1)
> # expected disposable incomes + some year to year variation
>
> av_ratio<-rbeta(200,10,10)
> ac_ratio<-rbeta(800, 2*(av_ratio[id]), 2*(1-av_ratio[id]))
> # expected proportion spent on car + some year to year variation
```

```
>
> y.car<-(ac_wealth*ac_ratio)^0.25     # disposable income * proportion spent on car
> y.hol<-(ac_wealth*(1-ac_ratio))^0.25 # disposable income * proportion spent on holiday
> Spending<-data.frame(y.hol=y.hol, y.car=y.car, id=id)
```

A simple regression suggests the two types of spending are negatively related but the association is weak with the $R^2 = 0.011$.

```
> summary(lm(y.car ~ y.hol, data = Spending))

Call:
lm(formula = y.car ~ y.hol, data = Spending)

Residuals:
      Min        1Q    Median        3Q       Max
-0.785418 -0.201515 -0.005233  0.178040  0.993145

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.10708    0.03831  28.898  < 2e-16 ***
y.hol       -0.10907    0.03663  -2.978  0.00299 **
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.2925 on 798 degrees of freedom
Multiple R-squared: 0.01099,        Adjusted R-squared: 0.009752
F-statistic: 8.869 on 1 and 798 DF,  p-value: 0.002989
```

With `id` added as a random term to deal with the the repeated measures, a similar conclusion is reached although the estimate is more negative:

```
> m5a.1 <- MCMCglmm(y.car ~ y.hol, random = ~id, data = Spending,
+     verbose = FALSE)
> summary(m5a.1$Sol[, "y.hol"])

Iterations = 3001:12991
Thinning interval = 10
Number of chains = 1
Sample size per chain = 1000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean             SD       Naive SE Time-series SE
     -0.202507       0.038393       0.001214       0.001184

2. Quantiles for each variable:
```

```
   2.5%     25%     50%     75%   97.5%
-0.2760 -0.2277 -0.2042 -0.1754 -0.1267
```

We may be inclined to stop there, but lets proceed with a multi-response model of the problem. The two responses are passed as a matrix using `cbind()`, and the rows of this matrix are indexed by the reserved variable `units`, and the columns by the reserved variable `trait`.

It is useful to think of a new data frame where the response variables have been stacked column-wise and the other predictors duplicated accordingly. Below is the original data frame on the left (`Spending`) and the stacked data frame on the right:

	y.hol	y.car	id
1	0.956558	1.421804	1
2	1.168930	0.899935	1
⋮	⋮	⋮	
800	0.968479	0.895636	200

$\Longrightarrow$

	y	trait	id	units
1	0.956558	y.hol	1	1
2	1.168930	y.hol	1	2
⋮	⋮	⋮	⋮	
800	0.968479	y.hol	200	800
801	1.421804	y.car	1	1
802	0.899935	y.car	1	2
⋮	⋮	⋮	⋮	
1600	0.895636	y.car	200	800

From this we can see that fitting a multi-response model is a direct extension to how we fitted models with categorical random interactions **??**:

```
> m5a.2 <- MCMCglmm(cbind(y.hol, y.car) ~ trait - 1, random = ~us(trait):id,
+     rcov = ~us(trait):units, data = Spending, family = c("gaussian",
+         "gaussian"), verbose = FALSE)
```

We have fitted the fixed effect `trait` so that the two types of spending can have different intercepts. I usually suppress the intercept (`-1`) for these types of models so the second coefficient is not the difference between the intercept for the first level of `trait` (`y.hol`) and the second level (`y.car`) but the actual trait specific intercepts. In other words the design matrix for the fixed effects has the form:

$$\begin{bmatrix} \texttt{trait[1]=="y.hol"} & \texttt{trait[1]=="y.car"} \\ \texttt{trait[2]=="y.hol"} & \texttt{trait[2]=="y.car"} \\ \vdots & \vdots \\ \texttt{trait[800]=="y.hol"} & \texttt{trait[800]=="y.car"} \\ \texttt{trait[801]=="y.hol"} & \texttt{trait[801]=="y.car"} \\ \texttt{trait[802]=="y.hol"} & \texttt{trait[802]=="y.car"} \\ \vdots & \vdots \\ \texttt{trait[1600]=="y.hol"} & \texttt{trait[1600]=="y.car"} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix}$$

A 2×2 covariance matrix is estimated for the random term where the diagonal elements are the variance in consistent individual effects for each type of spending. The off-diagonal is the covariance between these effects which if positive suggests that people that consistently spend more on their holidays consistently spend more on their cars. A 2×2 residual covariance matrix is also fitted. In Section **??** we fitted heterogeneous error models using `idh():units` which made sense in this case because each level of `unit` was specific to a particular datum and so any covariances could not be estimated. In multi-response models this is not the case because both traits have often been measured on the same observational unit and so the covariance can be measured. In the context of this example a positive covariance would indicate that in those years an individual spent a lot on their car they also spent a lot on their holiday.

A univariate regression is defined as the covariance between the response and the predictor divided by the variance in the predictor. We can therefore estimate a regression coefficient for these two levels of random variation, and compare them with the regression coefficient we obtained in the simpler model:

```
> id.regression <- m5a.2$VCV[, 2]/m5a.2$VCV[, 1]
> units.regression <- m5a.2$VCV[, 6]/m5a.2$VCV[, 5]
> plot(mcmc.list(m5a.1$Sol[, "y.hol"], id.regression, units.regression),
+     density = FALSE)
```

The regression coefficients (see Figure 1)differ substantially at the within individual (green) and between individual (red) levels, and neither is entirely consistent with the regression coefficient from the univariate model (black). The process by which we generated the data gives rise to this phenomenon - large variation between individuals in their disposable income means that people who are able to spend a lot on their holiday can also afford to spend a lot on their holidays (hence a positive covariation between `id` effects). However, a person that spent a large proportion of their disposable income in a particular year on a holiday, must have less to spend that year on a car (hence a negative residual (within year) covariation).

When fitting the simpler univariate model we make the assumption that the effect of spending money on a car directly effects how much you spend on a holiday. If this relationship was purely causal then all regression coefficients would have the same expectation, and the simpler model would be justified.

For example, we could set up a simpler model where two thirds of the variation in holiday expenditure is due to between individual differences, and holiday expenditure directly affects how much an individual will spend on their car (using a regression coefficient of -0.3). The variation in car expenditure not caused by holiday expenditure is also due to individual differences, but in this case they only explain a third of the variance.

```
> Spending$y.hol2 <- rnorm(200, 0, sqrt(2))[Spending$id] + rnorm(800,
```

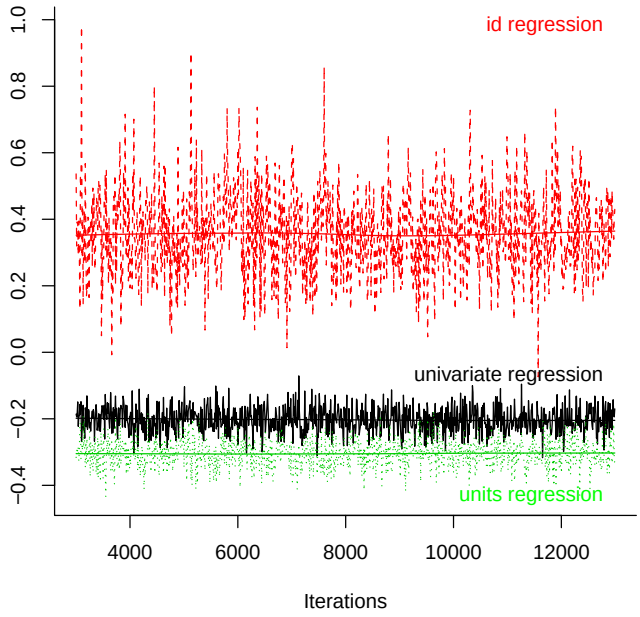

Figure 1: MCMC summary plot of the coefficient from a regression of car spending on holiday spending in black. The red and green traces are from a model where the regression coefficient is estimated at two levels: within an individual (green) and across individuals (red). The relationship between the two types of spending is in part mediating by a third unmeasured variable, disposable income.

```
+     0, sqrt(1))
> Spending$y.car2 <- Spending$y.hol2 * -0.3 + rnorm(200, 0, sqrt(1))[Spending$id] +
+     rnorm(800, 0, sqrt(2))
```

We can fit the univariate and multivariate models to these data, and compare the regression coefficients as we did before. Figure 2 shows that the regression coefficients are all very similar and a value of -0.3 has a reasonably high posterior probability. However, it should be noted that the posterior standard deviation is smaller in the simpler model because the more strict assumptions have allowed us to pool information across the two levels to get a more precise answer.

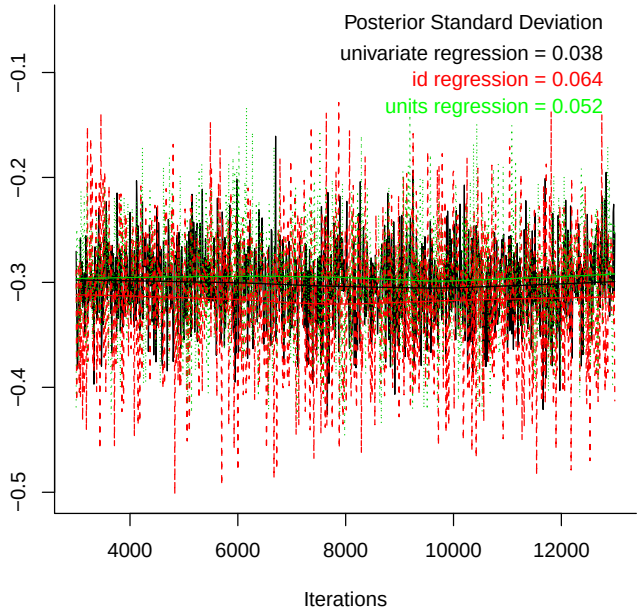

Figure 2: MCMC summary plot of the coefficient from a regression of car spending on holiday spending in black. The red and green traces are from a model where the regression coefficient is estimated at two levels: within an individual (green) and across individuals (red). In this model the relationship between the two types of spending is causal and the regression coefficients have the same expectation. However, the posterior standard deviation from the simple regression is smaller because information from the two different levels is pooled.

2 Multinomial

Multinomial models are difficult - both to fit and interpret. This is particularly true when each unit of observation only has a single realisation from the multinomial. In these instances the data can be expressed as a single vector of factors, and the family argument can be specified as `categorical`. To illustrate, using a very simple example, we'll use data collected on 666 Soay sheep from the island of Hirta in the St. Kilda archipelago (**?**, Table A2.5).

```
> data(SShorns)
> head(SShorns)
```

```
  id    horn    sex
1  1 scurred female
2  2 scurred female
3  3 scurred female
4  4 scurred female
5  5  polled female
6  6  polled female
```

The sex and horn morph were recorded for each individual, giving the contingency table:

```
> Ctable <- table(SShorns$horn, SShorns$sex)
> Ctable

         female male
  normal      83  352
  polled      65    0
  scurred     96   70
```

and we'll see if the frequencies of the three `horn` types differ, and if the trait is sex dependent. The usual way to do this would be to use a Chi square test, and to address the first question we could add the counts of the two sexes:

```
> chisq.test(rowSums(Ctable))

        Chi-squared test for given probabilities

data:  rowSums(Ctable)
X-squared = 329.5225, df = 2, p-value < 2.2e-16
```

which strongly suggests the three morphs differ in frequency. We could then ask whether the frequencies differ by sex:

```
> chisq.test(Ctable)

        Pearson's Chi-squared test

data:  Ctable
X-squared = 202.2962, df = 2, p-value < 2.2e-16
```

which again they do, which is not that surprising since the trait is partly sex limited, with males not expressing the polled phenotype.

If there were only two horn types, polled and normal for example, then we could have considered transforming the data into the binary variable *polled or not?* and analysing using a glm with sex as a predictor. In doing this we have reduced the dimension of the data from $k = 2$ categories to a single $(k - 1 = 1)$ contrast. The motivation for the dimension reduction is obvious; if being a

male increased the probability of expressing normal horns by 10%, it must by necessity reduce the probability of expressing polled horn type by 10%, because an individual cannot express both horn types simultaneously. The dimension reduction essentially constrains the probability of expressing either horn type to unity:

$$Pr(\texttt{horn[i]} = \text{normal}) + Pr(\texttt{horn[i]} = \text{polled}) = 1 \tag{1}$$

These concepts can be directly translated into situations with more than two categories where the unit sum constraint has the general form:

$$\sum_{j=1}^{k} Pr(y_i = j) = 1 \tag{2}$$

For binary data we designated one category to be the success (polled) and one category to be the failure (normal) which we will call the baseline category. The latent variable in this case was the log-odds ratio of succeeding versus failing:

$$l_i = \log\left(\frac{Pr(\texttt{horn[i]} = \text{polled})}{Pr(\texttt{horn[i]} = \text{normal})}\right) = \text{logit}\left(Pr(\texttt{horn[i]} = \text{polled})\right) \tag{3}$$

With more than two categories we need to have $k-1$ latent variables, which in the original horn type example are:

$$l_{i,\text{polled}} = \log\left(\frac{Pr(\texttt{horn[i]} = \text{polled})}{Pr(\texttt{horn[i]} = \text{normal})}\right) \tag{4}$$

and

$$l_{i,\text{scurred}} = \log\left(\frac{Pr(\texttt{horn[i]} = \text{scurred})}{Pr(\texttt{horn[i]} = \text{normal})}\right) \tag{5}$$

The two latent variables are indexed as `trait`, and the unit of observation (i) as `unit`, as in multi-response models. As with binary models the residual variance is not identified, and can be set to any arbitrary value. For reasons that will become clearer later I like to work with the residual covariance matrix $\frac{1}{k}(\mathbf{I}+\mathbf{J})$ where $\mathbf{I}$ and $\mathbf{J}$ are $k-1$ dimensional identity and unit matrices, respectively.

To start we will try a simple model with an intercept:

```
> IJ <- solve((1/3) * (diag(2) + matrix(1, 2, 2)))
> prior = list(R = list(V = IJ, fix = 1))
> m5c.1 <- MCMCglmm(horn ~ trait - 1, rcov = ~us(trait):units,
+     prior = prior, data = SShorns, family = "categorical", verbose = FALSE)
> plot(m5c.1$Sol)
```

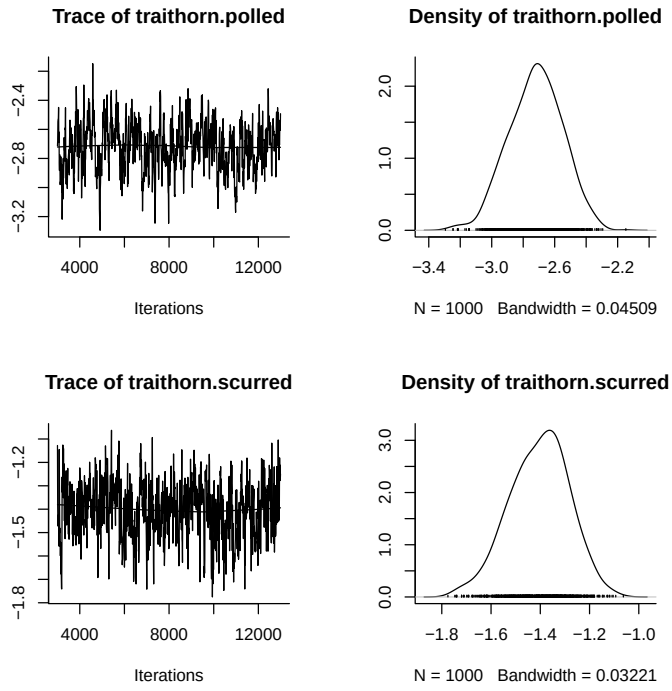

Figure 3: Phylogeny

The problem can be represented using the contrast matrix $\mathbf{\Delta}$ (**?**):

$$\mathbf{\Delta} = \begin{bmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{6}$$

where the rows correspond to the factor levels (`normal`, `polled` and `scurred`) and the columns to the two latent variables. For example column one corresponds to $l_{i,\text{polled}}$ which on the log scale is $Pr(\texttt{horn[i]} = \text{polled}) - Pr(\texttt{horn[i]} = \text{normal})$.

$$\exp\left((\mathbf{\Delta}\mathbf{\Delta}')^{-}\mathbf{\Delta}\mathbf{l}_i\right) \propto E \begin{bmatrix} Pr(\texttt{horn[i]} = \text{normal}) \\ Pr(\texttt{horn[i]} = \text{normal}) \\ Pr(\texttt{horn[i]} = \text{scurred}) \end{bmatrix} \tag{7}$$

The residual ($\mathbf{E}$) and any random effect (e.g. $\mathbf{A}$) covariance matrices are for estimability purposes estimated on the $J-1$ space: $\boldsymbol{C}_A = \mathbf{\Delta}'\mathbf{A}\mathbf{\Delta}$ and $\boldsymbol{C}_E = \mathbf{\Delta}'\mathbf{E}\mathbf{\Delta}$. Moreover, because there is only a single realization from the multinomial then all elements of $\boldsymbol{C}_E$ are non-estimable and must be fixed. As

with the binary model we fix $\mathbf{E} = \frac{1}{J}\mathbf{I}$ to give $\boldsymbol{C}_E = \frac{1}{J}(\mathbf{J} + \mathbf{I})$, where $\mathbf{J}$ is the unit matrix. We can visualize this unit sum constraint for three categories as a model parametrized on the simplex (Figure **??**).

3 Zero-inflated Poisson

Zero-inflation parameter is the probability that a zero comes from the extra-zero process as opposed to the Poisson process.