Separate Build Justification

Set-Up

Consider the simple two-part model where we model our response $y = (y_1, y_2, ..., y_n)$ as

$$\begin{aligned}
z_i &= \begin{cases} 1 & y_i > 0 \\ 0 & y_i = 0 \end{cases} \\
y_i^* &= x_{i,nz}\beta_1 + \beta_0 + \epsilon_i \qquad \text{where} \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \\
y_i &= z_i \cdot y_i^*
\end{aligned}$$

The subscript (nz) on our covariate for the y_i^* model represents the fact that that model is fit on only the data where the response is *non-zero*.

We model z_i on the full dataset as

$$p_i = P(z_i = 1) = \frac{1}{1 + e^{-(x_i\gamma_1 + \gamma_0)}}$$

Importantly, in this formulation we are assuming that the parameters from the model for z_i and the parameters from the model for y_i^* have priors that are independent.

Separate Fit

Now we might start out by fitting the models separately since this makes for an easier problem. Note, we use $f(\theta)$ to denote a probability distribution, and $L(\theta|y)$ to denote a likelihood function.

Under a bayesian frame we set up our posterior for each model as being proportional to the product of the likelihood and the joint priors.

In our model for y_i^* we have:

$$
\begin{aligned}
f(\beta_1, \beta_0, \sigma^2 \mid y) &\propto f(y \mid \beta_1, \beta_0, \sigma^2) f(\beta_1, \beta_0, \sigma^2) \\
&= L(\beta_1, \beta_0, \sigma^2 \mid y) f(\beta_1, \beta_0, \sigma^2) \\
&= \left[\prod_{i:y_i>0} f(y_i \mid \beta_1, \beta_0, \sigma^2) \right] f(\beta_1, \beta_0, \sigma^2)
\end{aligned}
$$

In our model for z_i we have:

$$
\begin{aligned}
f(\gamma_1, \gamma_0 \mid y) &\propto f(y \mid \gamma_1, \gamma_0) f(\gamma_1, \gamma_0) \\
&= L(\gamma_1, \gamma_0 \mid y) f(\gamma_1, \gamma_0) \\
&= \left[\prod_{i:y_i=0} (1 - p_i) \prod_{i:y_i>0} p_i \right] f(\gamma_1, \gamma_0)
\end{aligned}
$$

Because the proportionality constant is often too convoluted to derive exactly, we can use MCMC to simulate draws from each of these posteriors. Up until this point everything is fine, but when we want to actually build posterior predictive distributions we run into a problem. At this point we might have 2,000 MCMC parameter iterations for each posterior which look like this

$$
\begin{bmatrix}
\beta_1^{(1)} & \beta_0^{(1)} & (\sigma^2)^{(1)} \\
\vdots & \vdots & \vdots \\
\beta_1^{(2000)} & \beta_0^{(2000)} & (\sigma^2)^{(2000)}
\end{bmatrix}
\quad \text{and} \quad
\begin{bmatrix}
\gamma_1^{(1)} & \gamma_0^{(1)} \\
\vdots & \vdots \\
\gamma_1^{(2000)} & \gamma_0^{(2000)}
\end{bmatrix}
$$

and to get a posterior predictive distribution for a new data point we would take

$$
\begin{bmatrix}
y_{new}^{*(1)} \sim \mathcal{N}(\beta_1^{(1)} \cdot x_{new} + \beta_0^{(1)}, (\sigma^2)^{(1)}) \\
\vdots \\
\vdots \\
y_{new}^{*(2000)} \sim \mathcal{N}(\beta_1^{(2000)} \cdot x_{new} + \beta_0^{(2000)}, (\sigma^2)^{(2000)})
\end{bmatrix}
\quad \text{and} \quad
\begin{bmatrix}
p_{new}^{(1)} \sim Bern\left(\frac{1}{1+e^{-(\gamma_1^{(1)} \cdot x_{new} + \gamma_0^{(1)})}} \right) \\
\vdots \\
\vdots \\
p_{new}^{(2000)} \sim Bern\left(\frac{1}{1+e^{-(\gamma_1^{(2000)} \cdot x_{new} + \gamma_0^{(2000)})}} \right)
\end{bmatrix}
$$

And $(y_{new}^{*(1)}, ..., y_{new}^{*(2000)})$ would be our posterior predictive distribution under our model for y^*, while $(p_{new}^{(1)}, ..., p_{new}^{(2000)})$ would be our posterior predictive distribution under our model for p.

But remember that our response in this modeling schema is the product of the outputs from these two models. And so we want a the posterior predictive distribution of $y_{new} = y_{new}^* p_{new}$, but it's unclear how we should combine the predictive distributions from the individual models to get here. We might just match MCMC iteration i from each model together, but what makes this matching more correct than shuffling the iterations and then matching them up?

To understand what to do about this, we'll dive into some theory behind building the models simultaneously.

Simultaneous Fit

To get around our problem of how we combine the MCMC iterations for the models built separately, we could fit the models simultaneously. In this setting our posterior would be:

$$\begin{aligned} f(\beta_1, \beta_0, \gamma_1, \gamma_0, \sigma^2 \mid y) &\propto f(y \mid \beta_1, \beta_0, \gamma_1, \gamma_0, \sigma^2) f(\beta_1, \beta_0, \gamma_1, \gamma_0, \sigma^2) \\ &= L(\beta_1, \beta_0, \gamma_1, \gamma_0, \sigma^2 \mid y) f(\beta_1, \beta_0, \gamma_1, \gamma_0, \sigma^2) \end{aligned}$$

We can expand this by writing out the likelihood more fully based on whether y is zero or not:

$$f(\beta_1, \beta_0 \gamma_1, \gamma_0, \sigma^2 \mid y) \propto \left[\prod_{i:y_i=0} (1 - p_i) \prod_{i:y_i>0} p_i f(y_i \mid \beta_1, \beta_0, \sigma^2) \right] f(\beta_1, \beta_0, \gamma_1, \gamma_0, \sigma^2)$$

While it wasn't too difficult to write this out up to a proportionality constant, in practice it can be very difficult to figure out how to combine the two models in such a way that the MCMC algorithm still converges once you start using models that are more complicated than these very simple example ones.

But, there's important insight still to be found here. Let's group these terms based on the parameters that they use. In particular we'll group by which individual model the parameter belongs to:

$$= \left[\Big(\prod_{i:y_i=0} (1 - p_i) \prod_{i:y_i>0} p_i \Big) f(\gamma_1, \gamma_0) \right] \left[\Big(\prod_{i:y_i>0} f(y_i \mid \beta_1, \beta_0, \sigma^2) \Big) f(\beta_1, \beta_0, \sigma^2) \right]$$

We are able to split the joint prior in this way because we are assuming that there is no dependence in the priors *between* models.

But now, if we look at this closely we can see that what we really have here is a full separation into the posteriors for the individual models for p and y^* as seen in our derivation in the **separate fit** section. This means that we can write:

$$\begin{aligned} f(\beta_1, \beta_0, \gamma_1, \gamma_0, \sigma^2 \mid y) &\propto f(\gamma_1, \gamma_0 \mid y) f(\beta_1, \beta_0, \sigma^2 \mid y) \\ &= C \Big[f(\gamma_1, \gamma_0 \mid y) f(\beta_1, \beta_0, \sigma^2 \mid y) \Big] \end{aligned}$$

Finally, since these are all proper probability distributions we know that

$$\int_{-\infty}^{\infty} f(\beta_1, \beta_0, \gamma_1, \gamma_0, \sigma^2 \mid y)\ d\beta_1 d\beta_0 d\gamma_1 d\gamma_0 d\sigma^2 = C \int_{-\infty}^{\infty} f(\gamma_1, \gamma_0 \mid y)\ d\gamma_1 d\gamma_2 \int_{-\infty}^{\infty} f(\beta_1, \beta_0, \sigma^2 \mid y)\ d\beta_1 d\beta_0 d\sigma^2$$

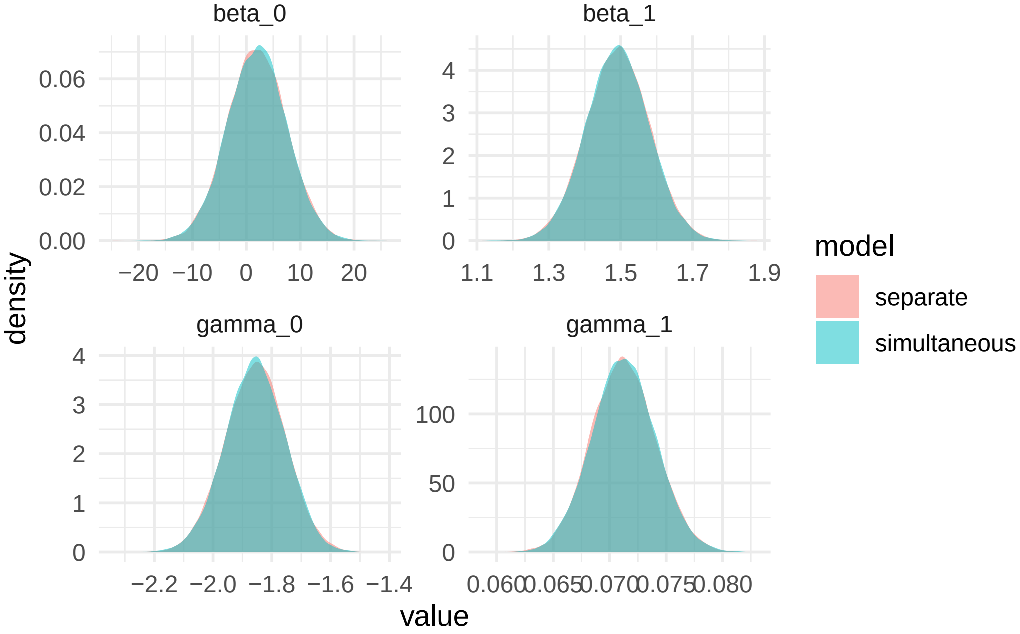
$$1 = C \cdot 1 \cdot 1$$

$$1 = C$$

And if $C = 1$ then the full posterior for the model built simultaneously is equal to the product of the posteriors for each model built separately. What this means for us is that we can fit the models separately, and then combine the results at the end to get our posterior predictive distributions.

Practical Backing

And indeed, when we fit the models simultaneously and also fit them separately making use of MCMC to generate samples from their posteriors, we find that they are nearly identical



Upshot

The major upshot here is that as long as we don't build in any correlations between the parameters in the two models, then we can build each model as complex as we might desire without having to worry about how we will eventually build the two models together. As we learned above, we can simply build them separately and combine the results at the end.