On-line supplementary material for "Analysing Mark-recapture-recovery Data in the Presence of Missing Covariate Data via Multiple Imputation"

Hannah Worthington, Ruth King and Stephen T. Buckland

A Derivation of Predictive Distributions

Here we detail the derivation of the covariate prediction distributions for unobserved covariate values for model 2 described in Section 3.1.1 relating to the first order Markov model with additive time and age effects.

Case (ii) $b_i < t < c_i$

Consider the predictive distribution of $z_{i,t}$ conditional on $w_{i,t-1}$ and $y_{i,t+k}$ such that all covariate values in the interval $[t, t+k-1]$ are unknown ($k \geq 1$). We have that

$$\begin{aligned} f(z_{i,t}|w_{i,t-1}, y_{i,t+k}, \widehat{\boldsymbol{\eta}}) &\propto f(z_{i,t}|w_{i,t-1}, \widehat{\boldsymbol{\eta}}) f(y_{i,t+k}|z_{i,t}, w_{i,t-1}, \widehat{\boldsymbol{\eta}}) \\ &= f(z_{i,t}|w_{i,t-1}, \widehat{\boldsymbol{\eta}}) f(y_{i,t+k}|z_{i,t}, \widehat{\boldsymbol{\eta}}). \end{aligned}$$

Given that

$$z_{i,t}|w_{i,t-1}, \widehat{\boldsymbol{\eta}} \sim N(w_{i,t-1} + \widehat{\kappa}_t + \widehat{\gamma}_j, \widehat{\sigma}^2)$$

and

$$y_{i,t+k}|z_{i,t}, \widehat{\boldsymbol{\eta}} \sim N\left(z_{i,t} + \sum_{g=1}^{k} (\widehat{\kappa}_{t+g} + \widehat{\gamma}_{j+g}), k\widehat{\sigma}^2\right),$$

we have that

$$
\begin{aligned}
f(z_{i,t}|w_{i,t-1}, y_{i,t+k}, \widehat{\boldsymbol{\eta}}) &\propto \exp\left(\frac{-(z_{i,t}-(w_{i,t-1}+\widehat{\kappa}_t+\widehat{\gamma}_j))^2}{2\widehat{\sigma}^2}\right) \\
&\quad \times \exp\left(\frac{-\left(y_{i,t+k}-\left(z_{i,t}+\sum_{g=1}^{k}(\widehat{\kappa}_{t+g}+\widehat{\gamma}_{j+g})\right)\right)^2}{2k\widehat{\sigma}^2}\right) \\
&\propto \exp\left(\frac{-1}{2k\widehat{\sigma}^2}\left(kz_{i,t}^2-2kz_{i,t}(w_{i,t-1}+\widehat{\kappa}_t+\widehat{\gamma}_j)\right.\right. \\
&\quad \left.\left. -2y_{i,t+k}z_{i,t}+z_{i,t}^2+2z_{i,t}\left(\sum_{g=1}^{k}(\widehat{\gamma}_{t+g}+\widehat{\gamma}_{j+g})\right)\right)\right) \\
&\propto \exp\left(\frac{-(k+1)}{2k\widehat{\sigma}^2}\right. \\
&\quad \left. \times\left(z_{i,t}-\left(\frac{k(w_{i,t-1}+\widehat{\kappa}_t+\widehat{\gamma}_j)+y_{i,t+k}-\sum_{g=1}^{k}(\widehat{\kappa}_{t+g}+\widehat{\gamma}_{j+g})}{k+1}\right)\right)^2\right).
\end{aligned}
$$

Thus the result follows that

$$
z_{i,t}|w_{i,t-1}, y_{i,t+k}, \widehat{\boldsymbol{\eta}} \sim N\left(\frac{k(w_{i,t-1}+\widehat{\kappa}_t+\widehat{\gamma}_j)+y_{i,t+k}-\sum_{g=1}^{k}(\widehat{\kappa}_{t+g}+\widehat{\gamma}_{j+g})}{k+1}, \frac{k\widehat{\sigma}^2}{k+1}\right).
$$

Case (iii) $t < b_i$

Consider the predictive distribution of z_{i,f_i} conditional on y_{i,b_i} such that all covariates values in the interval $[f_i, b_i - 1]$ are unknown. We have that

$$
f(z_{i,f_i}|y_{i,b_i}, \widehat{\boldsymbol{\eta}}) \propto f(y_{i,b_i}|z_{i,f_i}, \widehat{\boldsymbol{\eta}})f(z_{i,f_i}|\widehat{\boldsymbol{\eta}}).
$$

Given that

$$
y_{i,b_i}|z_{i,f_i}, \widehat{\boldsymbol{\eta}} \sim N\left(z_{i,f_i}+\sum_{g=1}^{k}(\widehat{\kappa}_{f_i+g}+\widehat{\gamma}_{1+g}), k\widehat{\sigma}^2\right)
$$

where $k = b_i - f_i$ and

$$
z_{i,f_i}|\widehat{\boldsymbol{\eta}} \sim N(\widehat{\nu}_{f_i}, \widehat{\tau}^2),
$$

we have that

$$f(z_{i,f_i}|y_{i,b_i}\widehat{\boldsymbol{\eta}}) \propto \exp\left(\frac{-\left(y_{i,b_i}-\left(z_{i,f_i}+\sum_{g=1}^{k}\left(\widehat{\kappa}_{f_i+g}+\widehat{\gamma}_{1+g}\right)\right)\right)^2}{2k\widehat{\sigma}^2}\right)\times\exp\left(\frac{-(z_{i,f_i}-\widehat{\nu}_{f_i})^2}{2\widehat{\tau}^2}\right)$$

$$\propto \exp\left(\frac{-1}{2k\widehat{\sigma}^2\widehat{\tau}^2}\left(-2\widehat{\tau}^2 y_{i,b_i}z_{i,f_i}+\widehat{\tau}^2 z_{i,f_i}^2+2\widehat{\tau}^2 z_{i,f_i}\left(\sum_{g=1}^{k}\left(\widehat{\kappa}_{f_i+g}+\widehat{\gamma}_{1+g}\right)\right)\right.\right.$$

$$\left.\left.+k\widehat{\sigma}^2 z_{i,f_i}^2-2k\widehat{\sigma}^2 z_{i,f_i}\widehat{\nu}_{f_i}\right)\right)$$

$$\propto \exp\left(\frac{-(\widehat{\tau}^2+k\widehat{\sigma}^2)}{2k\widehat{\sigma}^2\widehat{\tau}^2}\left(z_{i,f_i}-\left(\frac{\widehat{\tau}^2\left(y_{i,b_i}-\sum_{g=1}^{k}\left(\widehat{\kappa}_{f_i+g}+\widehat{\gamma}_{1+g}\right)\right)+k\widehat{\sigma}^2\widehat{\nu}_{f_i}}{\widehat{\tau}^2+k\widehat{\sigma}^2}\right)\right)^2\right).$$

Thus the result follows that

$$z_{i,f_i}|y_{i,b_i}\widehat{\boldsymbol{\eta}} \sim N\left(\frac{\widehat{\tau}^2\left(y_{i,b_i}-\sum_{g=1}^{k}\left(\widehat{\kappa}_{f_i+g}+\widehat{\gamma}_{1+g}\right)\right)+k\widehat{\sigma}^2\widehat{\nu}_{f_i}}{\widehat{\tau}^2+k\widehat{\sigma}^2}, \frac{k\widehat{\sigma}^2\widehat{\tau}^2}{\widehat{\tau}^2+k\widehat{\sigma}^2}\right).$$

B Simulation Study - Convergence of Regression Parameters

Here we provide estimates of the survival regression parameters against the number of multiple imputations used within the two-step algorithm for the simulation study conducted in Section 4 for a typical dataset for each possible scenario considered. Figure 1 considers the case $p_w = 1$ for each combination of the recapture and recovery parameter values and Figure 2 the analogous plots for $p_w = 0.6$.

[Figure 1 about here.]

[Figure 2 about here.]

C Simulation Study - Recapture and Recovery Probabilities

Figure 3 provides boxplots of the recapture and recovery probabilities for the simulation study conducted in Section 4 for each possible scenario considered.

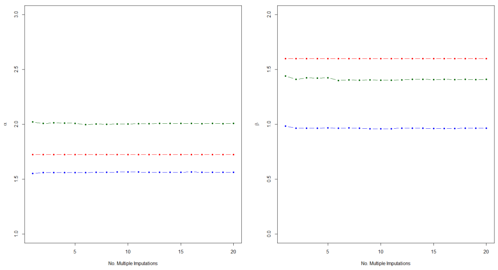
[Figure 3 about here.]

D Bayesian Analysis of Soay Sheep

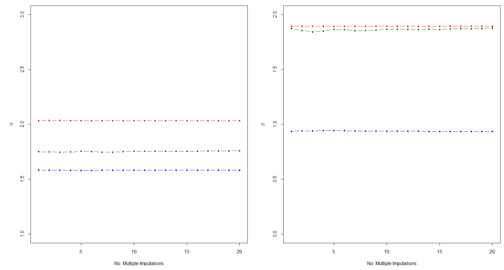
We consider a Bayesian analysis of the Soay sheep dataset, with the corresponding results provided in Section 5 of the paper. The following vague priors are specified:

$$
\begin{aligned}
\nu_t &\sim N(0, 0.001) \quad t = 1, \dots, 19 \\
\tau &\sim \Gamma(0.01, 0.01) \\
\kappa_t &\sim N(0, \tau_\kappa) \quad t = 2, \dots, 20 \\
\tau_\kappa &\sim \Gamma(0.01, 0.01) \\
\gamma_j &\sim N(0, \tau_\gamma) \quad j = 2, \dots, 14 \\
\tau_\gamma &\sim \Gamma(0.01, 0.01) \\
\sigma &\sim \Gamma(0.01, 0.01) \\
\alpha_k &\sim N(0, 0.001) \quad \text{for all age groups } k \\
\beta_k &\sim N(0, 0.001) \quad \text{for all age groups } k \\
p_t &\sim Beta(1, 1) \quad t = 2, \dots, 20 \\
\lambda_t &\sim Beta(1, 1) \quad t = 2, \dots, 20.
\end{aligned}
$$

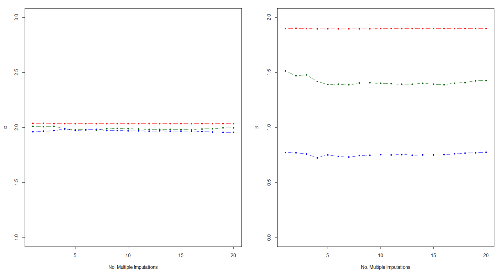
The Markov chain Monte Carlo (MCMC) simulations are conducted in `rjags` (Plummer, 2003). Two chains of 100000 iterations are run, with the first 25000 iterations discarded as burn-in. The Brooks-Gelman-Rubin statistic suggested that this was a conservative burn-in with $\widehat{R} < 1.01$ for all model parameters.



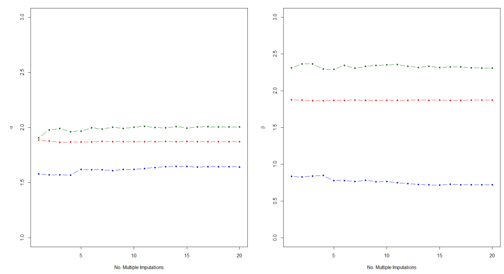
(a)



(b)



(c)



(d)

Figure 1: MLEs of the survival regression parameters for each age group plotted against the number of imputed datasets used for the simulation study with $p_w = 1$ for scenarios (a) $p = 0.9$, $\lambda = 0.9$; (b) $p = 0.9$, $\lambda = 0.3$; (c) $p = 0.3$, $\lambda = 0.9$ and; (d) $p = 0.3$, $\lambda = 0.3$. Red corresponds to lambs (year 1); green to yearlings (year 2); and blue to adults (years 3+).

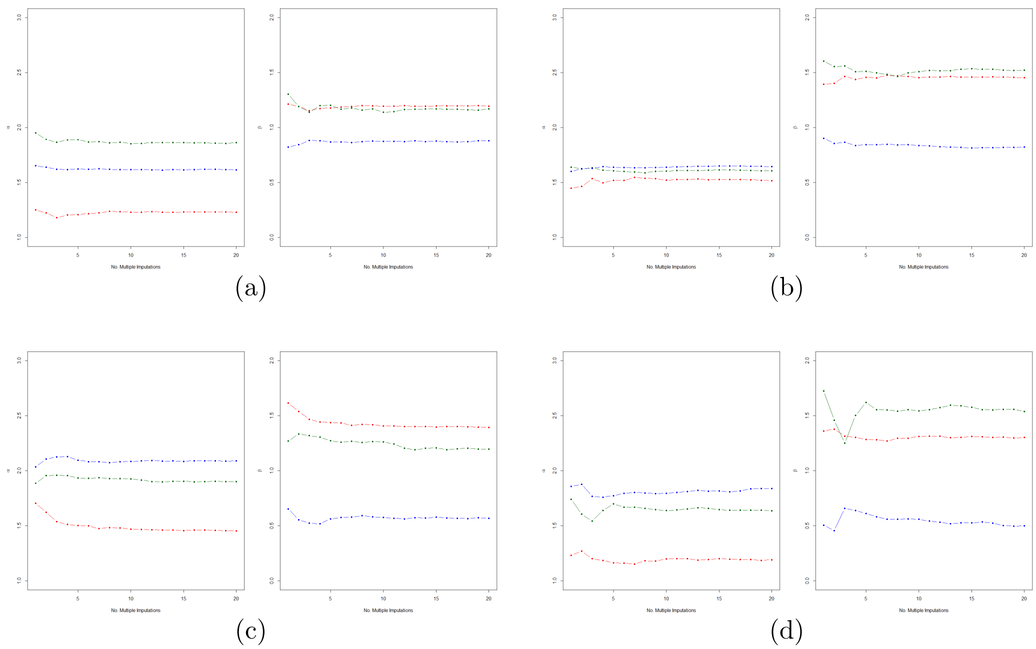


Figure 2: MLEs of the survival regression parameters for each age group plotted against the number of imputed datasets used for the simulation study with $p_w = 0.6$ for scenarios (a) $p = 0.9$, $\lambda = 0.9$; (b) $p = 0.9$, $\lambda = 0.3$; (c) $p = 0.3$, $\lambda = 0.9$; and (d) $p = 0.3$, $\lambda = 0.3$. Red corresponds to lambs (year 1); green to yearlings (year 2); and blue to adults (years 3+).

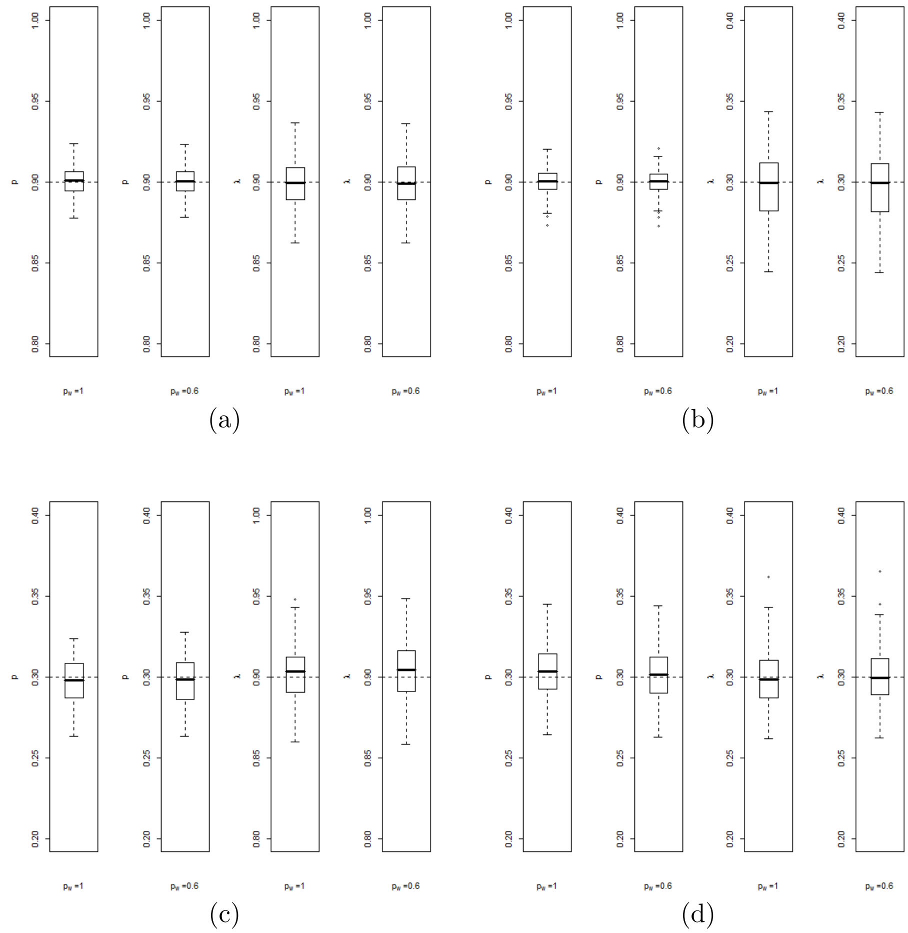


Figure 3: Boxplots of the capture and recovery probabilities (horizontal line is the true value) for the simulation study for scenarios (a) $p = 0.9$, $\lambda = 0.9$; (b) $p = 0.9$, $\lambda = 0.3$; (c) $p = 0.3$, $\lambda = 0.9$; and (d) $p = 0.3$, $\lambda = 0.3$.