Supplementary Materials for "A Note on Posterior Predictive Checks to Assess Model Fit for Incomplete Data"

Dandan Xu[1], Arkendu Chatterjee[2], Michael J. Daniels[3,*]
[1]Department of Statistics, University of Florida, Gainesville, FL 32611
[2]Novartis, East Hanover, NJ 07936
[3]Department of Statistics and Data Science, The University of Texas, Austin, TX 78712
*email: mjdaniels@austin.utexas.edu

Derivation of the posterior predictive probability for the observed replication from Section 3

We derive the posterior predictive probability for the observed replication using the model in Section 3. We only consider a prior distribution for the regression parameter, α, $\alpha \sim N(0, \nu^2)$ for simplification and keep the other regression parameters fixed. First we derive the posterior distribution of α,

$$\begin{aligned} p(\alpha|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}_2) &\propto \exp\left[-\frac{1}{2\nu^2}\alpha^2\right] . \exp\left[-\frac{1}{2\tau^2}\sum_{i=1}^{n_1}(y_{2i} - \alpha - \phi y_{1i})^2\right] \\ &\propto \exp\left[-\frac{1}{2}\left(\frac{n_1}{\tau^2} + \frac{1}{\nu^2}\right)\left(\alpha - \frac{\frac{\sum_{i=1}^{n_1} y_{2i} - \phi_{21}\sum_{i=1}^{n_1} y_{1i}}{\tau^2}}{\frac{n_1}{\tau^2} + \frac{1}{\nu^2}}\right)^2\right]. \end{aligned}$$

Thus,

$$\alpha|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}_2 \sim N\left(\frac{n_1\nu^2(\bar{y}_{2,\text{obs}} - \phi\bar{y}_{1,\text{obs}})}{\tau^2 + n_1\nu^2}, \frac{\tau^2\nu^2}{\tau^2 + n_1\nu^2}\right).$$

Note that,

$$\begin{aligned} & y^{\text{rep}}_{2,\text{obs}}|y^{\text{rep}}_1, r^{\text{rep}}_{2i} = 1, \alpha \sim N(\alpha + \phi y^{\text{rep}}_1, \tau^2) \\ & y^{\text{rep}}_1|r^{\text{rep}}_{2i} = 1 \sim N(\mu + \xi, \sigma_2^2) \\ \Rightarrow\ & y^{\text{rep}}_{2,\text{obs}}|r^{\text{rep}}_{2i} = 1, \alpha \sim N(\alpha + \phi(\mu + \xi), \tau^2 + \phi^2\sigma_2^2). \end{aligned}$$

Now define $\overline{y}^{\text{rep}}_{2,\text{obs}} = \frac{1}{\sum I(r^{\text{rep}}_{2i}=1)}\sum_{i=1}^{n} y^{\text{rep}}_{2i} I(r^{\text{rep}}_{2i} = 1)$. Let $N = \sum I(r^{\text{rep}}_{2i} = 1) \sim Binomial(n, \eta)$ and assume the data are sorted so that the objects with missing values are at the end. So $\overline{y}^{\text{rep}}_{2,\text{obs}} = \frac{1}{N}\sum_{i=1}^{N} y^{\text{rep}}_{2\text{obs},i}$. Thus,

$$\begin{aligned} & \overline{y}^{\text{rep}}_{2,\text{obs}}|N, \alpha \sim N\left(\alpha + \phi(\mu + \xi), \frac{\tau^2 + \phi^2\sigma_2^2}{N}\right). \\ & \text{Since, } \alpha|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}_2 \sim N\left(\frac{n_1\nu^2(\overline{y}_{2,\text{obs}} - \phi\overline{y}_{1,\text{obs}})}{\tau^2 + n_1\nu^2}, \frac{\tau^2\nu^2}{\tau^2 + n_1\nu^2}\right) \\ \Rightarrow\ & \overline{y}^{\text{rep}}_{2,\text{obs}}|N, \boldsymbol{y}_{\text{obs}}, \boldsymbol{r}_2 \sim N\left(\frac{n_1\nu^2(\overline{y}_{2,\text{obs}} - \phi\overline{y}_{1,\text{obs}})}{\tau^2 + n_1\nu^2} + \phi(\mu + \xi), \frac{\tau^2 + \phi^2\sigma_2^2}{N} + \frac{\tau^2\nu^2}{\tau^2 + n_1\nu^2}\right). \end{aligned}$$

Therefore, in finite samples, $\overline{y}^{\text{rep}}_{2,\text{obs}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}_2$ is a mixture of normals, which is given by

$$\sum_{k=1}^{n}\binom{n}{k}\eta^k(1-\eta)^{n-k}N\left(\frac{n_1\nu^2(\overline{y}_{2,\text{obs}} - \phi\overline{y}_{1,\text{obs}})}{\tau^2 + n_1\nu^2} + \phi(\mu + \xi), \frac{\tau^2 + \phi^2\sigma_2^2}{k} + \frac{\tau^2\nu^2}{\tau^2 + n_1\nu^2}\right).$$

To explore the large sample properties, we note that

$$\frac{N}{n} \xrightarrow{P} \eta.$$

By Slutsky's theorem,

$$\overline{y}^{\text{rep}}_{2,\text{obs}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}_2 \sim AN\left(\frac{n_1\nu^2(\overline{y}_{2,\text{obs}} - \phi\overline{y}_{1,\text{obs}})}{\tau^2 + n_1\nu^2} + \phi(\mu+\xi), \frac{\tau^2+\phi^2\sigma_2^2}{n\eta} + \frac{\tau^2\nu^2}{\tau^2+n_1\nu^2}\right),$$

where AN represents an asymptotically normal distribution. Using the above results we obtain,

$$\begin{aligned}
P(\overline{y}^{\text{rep}}_{2,\text{obs}} > \overline{y}_{2,\text{obs}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}_2) &\approx \Phi\left(\frac{-\frac{\tau^2\overline{y}_{2,\text{obs}}+n_1\nu^2\phi\overline{y}_{1,\text{obs}}}{\tau^2+n_1\nu^2} + \phi(\mu+\xi)}{\sqrt{\frac{\tau^2+\phi^2\sigma_2^2}{n\eta} + \frac{\tau^2\nu^2}{\tau^2+n_1\nu^2}}}\right) \\
&= \Phi\left(\frac{-\frac{\tau^2}{\tau^2+n_1\nu^2}\sqrt{n_1}(\bar{y}_{2,\text{obs}} - \phi(\mu+\xi)) - \frac{n_1\nu^2}{\tau^2+n_1\nu^2}\phi\sqrt{n_1}(\bar{y}_{1,\text{obs}} - (\mu+\xi))}{\sqrt{\frac{n_1}{n\eta}(\tau^2+\phi^2\sigma_2^2) + \frac{n_1\tau^2\nu^2}{\tau^2+n_1\nu^2}}}\right) \\
&= \Phi\left(\frac{-o_p(1)A - O_p(1)\phi B}{\sqrt{O_p(1)(\tau^2+\phi^2\sigma_2^2) + O_p(1)\tau^2}}\right) \\
&= \Phi\left(\frac{m}{\sqrt{V}}\right),
\end{aligned}$$

where $A = \sqrt{n_1}(\bar{y}_{2,\text{obs}} - \phi(\mu+\xi))$, $B = \sqrt{n_1}(\bar{y}_{1,\text{obs}} - (\mu+\xi))$, m denotes the numerator, and V denotes the denominator.

Derivation of the posterior predictive probability for the complete replication from Section 3

Here we derive the posterior predictive probability for the complete replication using the model in Section 3. As for observed replication, the prior distribution on the regression parameter α is $\alpha \sim N(0, \nu^2)$.

Note that,

$$\begin{aligned}
& y^{\text{rep}}_{2,\text{mis}}|y^{\text{rep}}_1, r^{\text{rep}}_{2i} = 0, \alpha \sim N(\alpha + \Delta + \phi y_1, \tau^2) \\
& y^{\text{rep}}_1|r^{\text{rep}}_{2i} = 0 \sim N(\mu, \sigma_1^2) \\
\Rightarrow\ & y^{\text{rep}}_{2,\text{mis}}|r^{\text{rep}}_{2i} = 0, \alpha \sim N(\alpha + \Delta + \phi\mu, \tau^2 + \phi^2\sigma_1^2).
\end{aligned}$$

Now define $\overline{y}^{\text{rep}}_{2,\text{com}} = \frac{1}{n}\sum_{i=1}^n y^{\text{rep}}_{2i} = \frac{\sum_{i=1}^N y^{\text{rep}}_{2\text{obs},i} + \sum_{i=N+1}^n y^{\text{rep}}_{2\text{mis},i}}{n}$, where $N = \sum I(r^{\text{rep}}_{2i} = 1) \sim Binomial(n, \eta)$ and the data is sorted so that the objects with missing values are at the end. Thus,

$$\begin{aligned}
& \overline{y}^{\text{rep}}_{2,\text{com}}|N, \alpha \sim N\left(\alpha + \phi\mu + \frac{N}{n}\phi\xi + \frac{n-N}{n}\Delta, \frac{n\tau^2 + N\phi^2\sigma_2^2 + (n-N)\phi^2\sigma_1^2}{n^2}\right) \text{ and} \\
& \overline{y}_{2,\text{com}}|\boldsymbol{y}_{\text{obs}}, \alpha \sim N\left(\frac{n_1\overline{y}_{2,\text{obs}} + (n-n_1)(\alpha + \Delta + \phi\overline{y}_{1,\text{mis}})}{n}, \frac{(n-n_1)\tau^2}{n^2}\right) \\
\Rightarrow\ & \overline{y}^{\text{rep}}_{2,\text{com}} - \overline{y}_{2,\text{com}}|N, \boldsymbol{y}_{\text{obs}}, \alpha \sim N\left(\frac{n_1\alpha + n\phi\mu + N\phi\xi + (n_1-N)\Delta - n_1\overline{y}_{2,\text{obs}} - (n-n_1)\phi\overline{y}_{1,\text{mis}}}{n},\right. \\
& \left.\frac{(2n-n_1)\tau^2 + N\phi^2\sigma_2^2 + (n-N)\phi^2\sigma_1^2}{n^2}\right).
\end{aligned}$$

$$\text{Since, } \alpha|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}_2 \sim N\left(\frac{n_1\nu^2(\overline{y}_{2,\text{obs}} - \phi\overline{y}_{1,\text{obs}})}{\tau^2 + n_1\nu^2}, \frac{\tau^2\nu^2}{\tau^2 + n_1\nu^2}\right)$$

$$\Rightarrow \quad \overline{y}^{\text{rep}}_{2,\text{com}} - \overline{y}_{2,\text{com}}|N, \boldsymbol{y}_{\text{obs}}, \boldsymbol{r}_2 \sim N\left(\frac{n_1^2\nu^2(\overline{y}_{2,\text{obs}} - \phi\overline{y}_{1,\text{obs}})}{n(\tau^2 + n_1\nu^2)} + \frac{n\phi\mu + N\phi\xi + (n_1 - N)\Delta - n_1\overline{y}_{2,\text{obs}} - (n - n_1)\phi\overline{y}_{1,\text{mis}}}{n}, \frac{(2n - n_1)\tau^2 + N\phi^2\sigma_2^2 + (n - N)\phi^2\sigma_1^2}{n^2} + \frac{n_1^2\tau^2\nu^2}{n^2(\tau^2 + n_1\nu^2)}\right).$$

Therefore, in finite samples, $\overline{y}^{\text{rep}}_{2,\text{com}} - \overline{y}_{2,\text{com}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}_2$ is a mixture of normals. In the large sample case, since $\frac{N}{n} \xrightarrow{P} \eta$ and $\frac{N}{n} \sim AN(\eta, \frac{\eta(1-\eta)}{n})$, we can show that

$$\overline{y}^{\text{rep}}_{2,\text{com}} - \overline{y}_{2,\text{com}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}_2 \sim \quad AN\left(\frac{n_1^2\nu^2(\overline{y}_{2,\text{obs}} - \phi\overline{y}_{1,\text{obs}})}{n(\tau^2 + n_1\nu^2)} + \frac{n\phi\mu + n\eta\phi\xi + (n_1 - n\eta)\Delta - n_1\overline{y}_{2,\text{obs}} - (n - n_1)\phi\overline{y}_{1,\text{mis}}}{n}, \frac{(2n - n_1)\tau^2 + n\eta\phi^2\sigma_2^2 + n\eta(1-\eta)(\Delta - \phi\xi)^2 + (n - n\eta)\phi^2\sigma_1^2}{n^2} + \frac{n_1^2\tau^2\nu^2}{n^2(\tau^2 + n_1\nu^2)}\right).$$

Using the above results we obtain,

$$P(\overline{y}^{\text{rep}}_{2,\text{com}} > \overline{y}_{2,\text{com}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}_2)$$

$$\approx \Phi\left(\frac{-\frac{\tau^2}{\tau^2 + n_1\nu^2}\sqrt{n_1}(\bar{y}_{2,\text{obs}} - \phi(\mu + \xi)) - \frac{n_1\nu^2}{\tau^2 + n_1\nu^2}\phi\sqrt{n_1}(\bar{y}_{1,\text{obs}} - (\mu + \xi)) + \sqrt{\frac{n}{n_1}}\frac{n_1 - n\eta}{\sqrt{n}}(\Delta - \phi\xi) + \sqrt{\frac{n - n_1}{n_1}}\phi\sqrt{n - n_1}(\mu - \overline{y}_{1,\text{mis}})}{\sqrt{\frac{n_1}{n\eta}(\tau^2 + \phi^2\sigma_2^2) + \frac{n_1\tau^2\nu^2}{\tau^2 + n_1\nu^2} + \frac{n\eta}{n_1}(1 - \eta)(\Delta - \phi\xi)^2 + \tau^2(\frac{2n - n_1}{n_1} - \frac{n_1}{n\eta}) + \phi^2\sigma_2^2(\frac{n\eta}{n_1} - \frac{n_1}{n\eta}) + \phi^2\sigma_1^2\frac{n - n\eta}{n_1}}}\right)$$

$$= \Phi\left(\frac{-o_p(1)A - O_p(1)\phi B + O_p(1)\frac{1}{\sqrt{\eta}}(\Delta - \phi\xi)C - O_p(1)\sqrt{\frac{1-\eta}{\eta}}\phi D}{\sqrt{O_p(1)(\tau^2 + \phi^2\sigma_2^2) + O_p(1)\tau^2 + O_p(1)(1 - \eta)(\Delta - \phi\xi)^2 + O_p(1)\frac{2(1-\eta)}{\eta}\tau^2 + o_p(1)\phi^2\sigma_2^2 + O_p(1)\frac{1-\eta}{\eta}\phi^2\sigma_1^2}}\right),$$

where $C = \frac{n_1 - n\eta}{\sqrt{n}}$ and $D = \sqrt{n - n_1}(\overline{y}_{1,\text{mis}} - \mu)$.

WinBUGS code for the posterior predictive checks in Section 3

```
model{
##### likelihood (sigma1,sigma2 and tau are standard deviation)
for (i in 1:n)
{
  r2[i]~dbern(eta)
  y[i,1]~dnorm(mu+xi*r2[i], 1/sigma2^2*r2[i]+1/sigma1^2*(1-r2[i]))
  y[i,2]~dnorm(alpha+phi*y[i,1]+delta*(1-r2[i]), 1/tau^2)
}
##### prior
  alpha~dnorm(0,0.00001)
##### known (fixed) parameters
  eta<-param[1]
  mu<-param[2]
  xi<-param[3]
  sigma1<-param[4]
  sigma2<-param[5]
  phi<-param[6]
  delta<-param[7]
  tau<-param[8]
#### predictive replicates
for (i in 1:n)
{
  r2pred[i]~dbern(eta)
  ypred[i,1]~dnorm(mu+xi*r2pred[i], 1/sigma2^2*r2pred[i]+1/sigma1^2*(1-r2pred[i]))
  ypred[i,2]~dnorm(alpha+phi*ypred[i,1]+delta*(1-r2pred[i]), 1/tau^2)
}
#### posterior predictive check based on oberved data
check_obs<-step(sum(ypred[,2]*r2pred)-sum(y[,2]*r2)/sum(r2)*sum(r2pred))
#### posterior predictive check based on complete data
#check_com<-step(sum(ypred[,2])-sum(y[,2]))
}
```

WinBUGS code for the posterior predictive checks in Section 4

The observed data $(\boldsymbol{y}, \boldsymbol{r}_2)$ are sorted so that the subjects in the EG group are ordered before those in the EP group.

```
model{
  ##### likelihood (sigma1,sigma2 and tau are precision)
  for (i in 1:78)
  {
    r2[i]~dbern(eta)
    y[i,1]~dnorm(mu+xi*r2[i], sigma2*r2[i]+sigma1*(1-r2[i]))
    y[i,2]~dnorm(alpha+phi*y[i,1]+delta*(1-r2[i]), tau)
  }
  ##### prior
  alpha~dnorm(0,0.00001)
  eta~dbeta(1,1)
  mu~dnorm(0,0.00001)
  xi~dnorm(0,0.00001)
  sigma1~dgamma(.01,.01)
  sigma2~dgamma(.01,.01)
  tau~dgamma(.01,.01)
  phi~dnorm(0,0.00001)

  #### posterior predictive replicates
  for (i in 1:n)
  {
    r2pred[i]~dbern(eta)
    ypred[i,1]~dnorm(mu+xi*r2pred[i], sigma2*r2pred[i]+sigma1*(1-r2pred[i]))
    ypred[i,2]~dnorm(alpha+phi*ypred[i,1]+delta*(1-r2pred[i]),tau)
  }
  #### posterior predictive checks for EG based on oberved data and complete data replications

  check_obs_EG<-step(sum(ypred[1:38,2]*r2pred[1:38])-sum(y[1:38,2]*r2[1:38])/
                     sum(r2[1:38])*sum(r2pred[1:38]))

  check_com_EG<-step(sum(ypred[1:38,2])-sum(y[1:38,2]))


  #### posterior predictive checks for EP based on oberved data and complete data replications

  check_obs_EP<-step(sum(ypred[39:78,2]*r2pred[39:78])-sum(y[39:78,2]*r2[39:78])/
                     sum(r2[39:78])*sum(r2pred[39:78]))

  check_com_EP<-step(sum(ypred[39:78,2])-sum(y[39:78,2]))
```

```
  #### marginal means of Y_1 and Y_2
  mean_y1<-mu+eta*xi
  mean_y2<-alpha+phi*(mu+eta*xi)+(1-eta)*delta
}
```