# Appendix: Bias for estimating $γ\_{u}$ from imputed data when $R\_{A}=U$

Here we give an informal justification for the bias result. In this section we use the superscript $\*$ to denote true values of parameters.

First consider the imputation model

$$a\_{i}=ϕ\_{0}^{\*}+ϕ\_{Y}^{\*}y\_{i}+δ\_{i},$$

for $J=\{i:R\_{A,i}=0\}$, i.e. non-missing $A$s, with $δ\_{i}∼N(0,τ^{2})$. Note that $u\_{i}$ does not appear in this model because $u\_{i}=0$ for all $i\in J$.

Now $y\_{i}∼N(μ\_{Y,i}=γ\_{0}^{\*}+γ\_{A}^{\*}a\_{i},σ\_{Y}^{2})$ for $i\in J$. In analogy with p175 of [[11](#ref-Frost2000)], consider a hypothetical imputation model based on $μ\_{Y,i}$:

$$a\_{i}=ψ\_{0}^{\*}+ψ\_{Y}^{\*}μ\_{Y,i}+ξ\_{i},$$

from which it is apparent that $ψ\_{Y}^{\*}=1/γ\_{A}^{\*}$. Additionally, across all observations, $Var(μ\_{Y})=γ\_{A}^{\*}​^{2}σ\_{A}^{\*2}$

In analogy with [[11](#ref-Frost2000)] (referencing [[17](#ref-Snedecor1968)]) we have that

$$ϕ\_{Y}^{\*}=\frac{ψ\_{Y}^{\*}γ\_{A}^{\*2}σ\_{A}^{\*2}}{γ\_{A}^{\*2}σ\_{A}^{\*2}+σ\_{Y}^{\*2}}=\frac{γ\_{A}^{\*}σ\_{A}^{\*2}}{γ\_{A}^{\*2}σ\_{A}^{\*2}+σ\_{Y}^{\*2}}.$$

Moreover, $ϕ\_{0}^{\*}=0$ because $A$ and $Y$ are both centred.

The imputation model is then used to impute values for the missing $a\_{i}$s i.e. for $i\notin J$, if we knew the true imputation model,

$$a\_{i,imp}=ϕ\_{0}^{\*}+ϕ\_{Y}^{\*}y\_{i}+δ\_{i}=\frac{γ\_{A}^{\*}σ\_{A}^{\*2}}{γ\_{A}^{\*2}σ\_{A}^{\*2}+σ\_{Y}^{\*2}}(γ\_{A}^{\*}a\_{i}+γ\_{U}^{\*}u\_{i}+ϵ\_{i})+δ\_{i}.$$

Returning to the outcome model,

$$y\_{i}=γ\_{0}^{\*}+γ\_{A}^{\*}a\_{i}+γ\_{U}^{\*}u\_{i}+ϵ\_{i}.$$

In the absence of missing data, we would of course simply solve using least squares, and if $γ=(γ\_{0},γ\_{A},γ\_{U})$ and $\hat{y}\_{i}(γ)=γ\_{0}+γ\_{A}a\_{i}+γ\_{U}u\_{i}$, then $\tilde{γ}=argmin(\sum\_{i=1}^{n}(y\_{i}-\hat{y}\_{i})^{2})$, then of course $E\_{Y}[\tilde{γ}\_{U}]=γ\_{U}^{\*}$.

As we have missing data, rewriting the outcome model to replace the missing $a\_{i}$s with their imputed versions, for substitution into the least squares formula we have:

$$\hat{y}\_{i}(γ)=γ\_{0}+γ\_{A}((1-u\_{i})a\_{i}+u\_{i}a\_{i,imp})+γ\_{U}u\_{i}.$$

The residual sum of squares can then be written as

$$\hat{γ}=argmin\sum\_{i=1}^{n}\left\{(γ\_{0}^{\*}-γ\_{0})+(γ\_{A}^{\*}-γ\_{A})a\_{i}+\left(γ\_{U}^{\*}-γ\_{A}γ\_{U}^{\*}κ-γ\_{U})u\_{i}+(-γ\_{A}κϵ\_{i}-γ\_{A}δ\_{i}\right)u\_{i}+(γ\_{A}-γ\_{A}κγ\_{A}^{\*})u\_{i}a\_{i}\right\}^{2},$$

where $κ=\frac{γ\_{A}^{\*}σ\_{A}^{\*2}}{γ\_{A}^{\*2}σ\_{A}^{\*2}+σ\_{Y}^{\*2}}$.

To consider minimising this expression, consider each bracket in turn. To minimise the first bracket, it is clear that $E\_{Y}[\hat{γ}\_{0}]=γ\_{0}^{\*}$. It is also apparent that $E\_{Y}[\hat{γ}\_{A}]=γ\_{A}^{\*}$, since we must minimise the second bracket, and the fourth and fifth brackets are additional error contributed by the imputed data, which cannot be reduced. This leaves the third bracket, which is minimised with

$$E\_{Y}[γ\_{U}^{\*}-γ\_{A}γ\_{U}^{\*}κ-γ\_{U}]=0.$$

Rearranging yields,

$$E\_{Y}[γ\_{U}]=γ\_{U}^{\*}\frac{σ\_{Y}^{\*2}}{γ\_{A}^{\*2}σ\_{A}^{\*2}+σ\_{Y}^{\*2}},$$

as claimed.

# Supplementary figures



Figure S1: Results for scenarios (i) and (ii). Average model based and empirical standard errors. Columns are different parameter estimates, rows are different values of $σ\_{Y}$. Within each graph, the y-axis varies $β\_{U}$.



Figure S2: Results for scenarios (ii) and (iii) with $R\_{A}=U$. Average model based and empirical standard errors. Columns are different parameter estimates, rows are different values of $σ\_{Y}$. Within each graph, the y-axis varies $α\_{U}$.



Figure S3: Results for scenario (iii). Average model based and empirical standard errors. Columns are different parameter estimates, rows are different values of $β\_{U}$. Within each graph, the y-axis varies $α\_{U}$.



Figure S4: Results for scenarios (iv). Average model based and empirical standard errors. Columns are different parameter estimates, rows are different values of $σ\_{Y}$. Within each graph, the y-axis varies $β\_{A}$.



Figure S5: Results for scenarios (v). Average model based and empirical standard errors. Columns are different parameter estimates, rows are different values of $β\_{U}$. Within each graph, the y-axis varies $β\_{A}$.



Figure S6: Results for scenario (vi). Average model based and empirical standard errors. Columns are different parameter estimates, rows are different values of $β\_{U}$. Within each graph, the y-axis varies $β\_{A}$.