Supplementary Material to *Bias analysis for misclassification errors in both the response variable and covariate*

Juxin Liu*
liu@math.usask.ca
Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, S7N 5E6, Canada

Annshirley Afful
aaa877@math.usask.ca
Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, S7N 5E6, Canada

Holly Mansell
holly.mansell@usask.ca
College of Pharmacy and Nutrition, University of Saskatchewan, Saskatoon, S7N 5E5, Canada

Yanyuan Ma
yzm63@psu.edu
Department of Statistics, Pennsylvania State University, University Park, PA 16802

April 10, 2022

*Corresponding author

1. Supplementary notes

Supplementary Note 1: Proof of the relations among the dependence parameters when Y/Y^* and X/X^* are binary variables.

$$
\begin{aligned}
& P(Y^*=y, X^*=1-x|Y=y, X=x) - P(Y^*=y|Y=y, X=x)P(X^*=1-x|Y=y, X=x) \\
= \; & P(Y^*=y|Y=y, X=x) - P(Y^*=y, X^*=x|Y=y, X=x) \\
& -P(Y^*=y|Y=y, X=x) + P(Y^*=y|Y=y, X=x)P(X^*=x|Y=y, X=x) \\
= \; & -\Big(P(Y^*=y, X^*=x|Y=y, X=x) - P(Y^*=y|Y=y, X=x)P(X^*=x|Y=y, X=x)\Big) \\
= \; & -D_{yx},
\end{aligned}
$$

$$
\begin{aligned}
& P(Y^*=1-y, X^*=x|Y=y, X=x) - P(Y^*=1-y|Y=y, X=x)P(X^*=x|Y=y, X=x) \\
= \; & P(X^*=x|Y=y, X=x) - P(Y^*=y, X^*=x|Y=y, X=x) \\
& -P(X^*=x|Y=y, X=x) + P(Y^*=y|Y=y, X=x)P(X^*=x|Y=y, X=x) \\
= \; & -\{P(Y^*=y, X^*=x|Y=y, X=x) - P(Y^*=y|Y=y, X=x)P(X^*=x|Y=y, X=x)\} \\
= \; & -D_{yx},
\end{aligned}
$$

and

$$
\begin{aligned}
& P(Y^*=1-y, X^*=1-x|Y=y, X=x) - \\
& P(Y^*=1-y|Y=y, X=x)P(X^*=1-x|Y=y, X=x) \\
= \; & P(X^*=1-x|Y=y, X=x) - P(Y^*=y, X^*=1-x|Y=y, X=x) - \\
& P(X^*=1-x|Y=y, X=x) + P(Y^*=y|Y=y, X=x)P(X^*=1-x|Y=y, X=x) - \\
& P(Y^*=y|Y=y, X=x)P(X^*=x|Y=y, X=x) \\
= \; & -P(Y^*=y, X^*=1-x|Y=y, X=x) + P(Y^*=y|Y=y, X=x)P(X^*=1-x|Y=y, X=x) \\
= \; & D_{yx}.
\end{aligned}
$$

Supplementary Note 2: Relationship between $\mathbf{p}$ and $\mathbf{p}^*$

Recall $\mathbf{p}^*$ and $\mathbf{p}$ represent the joint distribution for (Y^*, X^*) and (Y, X), respectively when both variables are binary.

$$
\begin{aligned}
\mathbf{p}^* &= M\mathbf{p} \qquad \text{(S.1)} \\
\mathbf{p}^* &= \{(M_Y \otimes \mathbf{1}) \circ (\mathbf{1} \otimes M_X) + \mathbf{D}\}\mathbf{p},
\end{aligned}
$$

where $\otimes$ is the Kronecker product and $\circ$ is the Hadamard product (i.e., elementwise multiplication of matrices) and

$$\begin{aligned}
\mathbf{1} &= (1,1)', \\
M_Y &= \begin{pmatrix} SN_{Y1} & SN_{Y0} & 1-SP_{Y1} & 1-SP_{Y0} \\ 1-SN_{Y1} & 1-SN_{Y0} & SP_{Y1} & SP_{Y0} \end{pmatrix}, \\
M_X &= \begin{pmatrix} SN_{X1} & SN_{X0} & 1-SP_{X1} & 1-SP_{X0} \\ 1-SN_{X1} & 1-SN_{X0} & SP_{X1} & SP_{X0} \end{pmatrix}, \\
\mathbf{D} &= \begin{pmatrix} D_{11} & -D_{10} & -D_{01} & D_{00} \\ -D_{11} & D_{10} & D_{01} & -D_{00} \\ -D_{11} & D_{10} & D_{01} & -D_{00} \\ D_{11} & -D_{10} & -D_{01} & D_{00} \end{pmatrix}.
\end{aligned}$$

For *nondifferential* misclassification, (S.2) holds and can be further simplified to

$$\mathbf{p}^* = (M_Y \otimes M_X + \mathbf{D})\mathbf{p}, \tag{S.2}$$

where $\otimes$ is the Kronecker product and

$$M_Y = \begin{pmatrix} SN_Y & 1-SP_Y \\ 1-SN_Y & SP_Y \end{pmatrix}, \quad M_X = \begin{pmatrix} SN_X & 1-SP_X \\ 1-SN_X & SP_X \end{pmatrix}.$$

Supplementary Note 3: Derivations of L_v and L_m for binary Y and X

In the most general setting with differential and dependent misclassification errors for binary Y and X, the likelihood function based on validation data is

$$\begin{aligned}
L_v(\boldsymbol{\theta}) &= \prod_{i=1}^{n_v} p_{y_i x_i} \Big[\Big\{ SN_{Yx_i}^{y_i^*} (1-SN_{Yx_i})^{1-y_i^*} \Big\}^{y_i} \Big\{ 1 - SP_{Yx_i}^{y_i^*} (1-SP_{Yx_i})^{1-y_i^*} \Big\}^{1-y_i} \\
&\qquad \Big\{ SN_{Xy_i}^{x_i^*} (1-SN_{Xy_i})^{1-x_i^*} \Big\}^{x_i} \Big\{ SN_{Xy_i}^{x_i^*} (1-SN_{Xy_i})^{1-x_i^*} \Big\}^{1-x_i} \\
&\qquad + (-1)^{I(y_i^*=y_i)+I(x_i^*=x_i)} D_{y_i x_i} \Big].
\end{aligned}$$

Using the rule of total probability and definition of the dependence parameters, we have

$$\begin{aligned}
L_m(\boldsymbol{\theta}) &= \prod_{i=n_v+1}^{n} \sum_{yx} p_{yx} \Big\{ P(Y_i^* = y_i^* | Y_i = y) P(X_i^* = x_i^* | X_i = x) + D_{yx} (-1)^{\{I(y_i^*=y)+I(x_i^*=x)\}} \Big\} \\
&= \prod_{i=n_v+1}^{n} \sum_{yx} p_{yx} \Big[\Big\{ SN_{Yx}^{y_i^*} (1-SN_{Yx})^{1-y_i^*} \Big\}^{y} \Big\{ (1-SP_{Yx})^{y_i^*} SP_{Yx}^{1-y_i^*} \Big\}^{1-y} \Big\{ SN_{Xy}^{x_i^*} (1-SN_{Xy})^{1-x_i^*} \Big\}^{x} \\
&\qquad \Big\{ (1-SP_{Xy})^{x_i^*} SP_{Xy}^{1-x_i^*} \Big\}^{1-x} + (-1)^{I(y_i^*=y)+I(x_i^*=x)} D_{yx} \Big], \\
&= \prod_{i=n_v+1}^{n} \Big(L_{m,ind}(\boldsymbol{\theta} | y_i^*, x_i^*) + (-1)^{(y_i^*+x_i^*)} \delta \Big).
\end{aligned} \tag{S.3}$$

Supplementary Note 4: Boundaries of D parameters.

The proof is given in a general setting, that is, categorical Y and X. Let $D_{ij,st} = P(Y^* = i, X^* = j|Y = s, X = t) - P(Y^* = i|Y = s, X = t)P(X^* = j|Y = s, X = t)$. Because

$$P(Y^* = i, X^* = j|Y = s, X = t) \;\; \leq \;\; P(Y^* = i|Y = s, X = t),$$

we have

$$\begin{aligned} D_{ij,st} &\leq P(Y^* = i|Y = s, X = t) - P(Y^* = i|Y = s, X = t)P(X^* = j|Y = s, X = t) \\ &= P(Y^* = i|Y = s, X = t)\Big(1 - P(X^* = j|Y = s, X = t)\Big). \end{aligned}$$

Similarly

$$\begin{aligned} D_{ij,st} &\leq P(X^* = j|Y = s, X = t) - P(Y^* = i|Y = s, X = t)P(X^* = j|Y = s, X = t) \\ &= P(X^* = j|Y = s, X = t)\Big(1 - P(Y^* = i|Y = s, X = t)\Big). \end{aligned}$$

On the other hand, due to the fact $P(A \cap B) = P(A) + P(B) - P(A \cup B) \geq P(A) + P(B) - 1$, we have

$$\begin{aligned} D_{ij,st} &\geq P(Y^* = i|Y = s, X = t) + P(X^* = j|Y = s, X = t) - 1 - \\ &\quad\; P(Y^* = i|Y = s, X = t)P(X^* = j|Y = s, X = t) \\ &= -\Big(1 - P(Y^* = i|Y = s, X = t)\Big)\Big(1 - P(X^* = j|Y = s, X = t)\Big). \end{aligned}$$

Moreover, simply dropping the first item in the definition of $D_{ij,st}$,

$$D_{ij,st} \;\; \geq \;\; -P(Y^* = i|Y = s, X = t)P(X^* = j|Y = s, X = t).$$

Note that for binary case under the nondifferential misclassification assumption,

$$\begin{aligned} P(Y^* = i|Y = s, X = t) &= P(Y^* = i|Y = s) \\ &= SN_Y^{is}(1 - SN_Y)^{(1-i)s} SP_Y^{(1-i)(1-s)}(1 - SP_Y)^{i(1-s)}, \\ P(X^* = j|Y = s, X = t) &= P(X^* = j|X = t) \\ &= SN_X^{jt}(1 - SN_X)^{(1-j)t} SP_X^{(1-j)(1-t)}(1 - SP_X)^{j(1-t)}. \end{aligned}$$

Supplementary Note 5: Proof of $\delta = E(\mathbf{cov}(Y^*, X^*|Y, X))$.

The proof is for a general setting with differential misclassification errors in both Y and X, which includes nondifferential misclassification errors as a special case. By definition of covariance, we have

$$\begin{aligned} \mathrm{cov}(Y^*, X^*|Y, X) &= E(Y^* X^*|Y, X) - E(Y^*|Y, X)E(X^*|Y, X) \\ &= P(Y^* = X^* = 1|Y, X) - P(Y^* = 1|Y, X)P(X^* = 1|Y, X). \end{aligned}$$

Also we have

$$
\begin{aligned}
& E\Big(P(Y^* = X^* = 1|Y, X)\Big) = \sum_{y,x} P(Y^* = X^* = 1|Y = y, X = x)p_{yx} \\
= \quad & \Big(D_{11} + SN_{Y1}SN_{X1}\Big)p_{11} + \Big(-D_{10} + SN_{Y0}(1 - SP_{X1})\Big)p_{10} + \\
& \Big(-D_{01} + (1 - SP_{Y1})SN_{X0}\Big)p_{01} + \Big(D_{00} + (1 - SP_{Y0})(1 - SP_{X0})\Big)p_{00} \\
= \quad & \delta + SN_{Y1}SN_{X1}p_{11} + SN_{Y0}(1 - SP_{X1})p_{10} + (1 - SP_{Y1})SN_{X0}p_{01} + (1 - SP_{Y0})(1 - SP_{X0})p_{00}.
\end{aligned}
$$

While

$$
\begin{aligned}
& E(P(Y^* = 1|Y, X)P(X^* = 1|Y, X)) = \sum_{y,x} P(Y^* = 1|Y = y, X = x)P(X^* = 1|Y = y, X = x)p_{y,x} \\
& = SN_{Y1}SN_{X1}p_{11} + SN_{Y0}(1 - SP_{X1})p_{10} + (1 - SP_{Y1})SN_{X0}p_{01} + (1 - SP_{Y0})(1 - SP_{X0})p_{00}.
\end{aligned}
$$

Therefore, combining the above three equalities together, we have $\delta = E(\mathrm{cov}(Y^*, X^*|Y, X))$.

2. Supplementary plots

2.1 Binary variables

Here are the plots for the simulation scenario in Section 4.1 when X and/or Y is subject to differential errors.

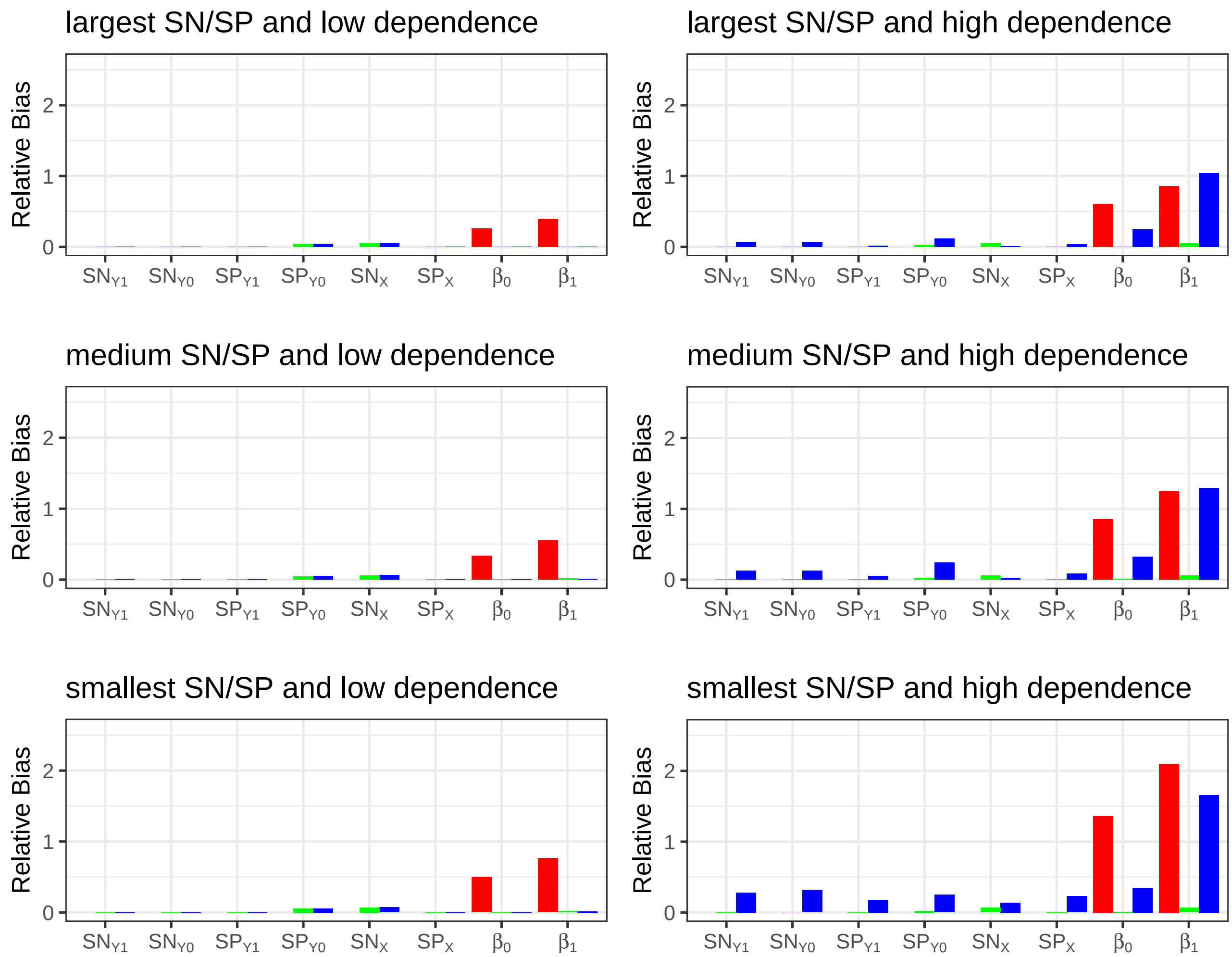

Figure S.1: Relative bias of model parameters when Y is subject to differential misclassification error and X is subject to nondifferential misclassification error with $n_v/n = 10\%$: red for naïve model, blue for independent misclassification error model, green for dependent misclassification model.

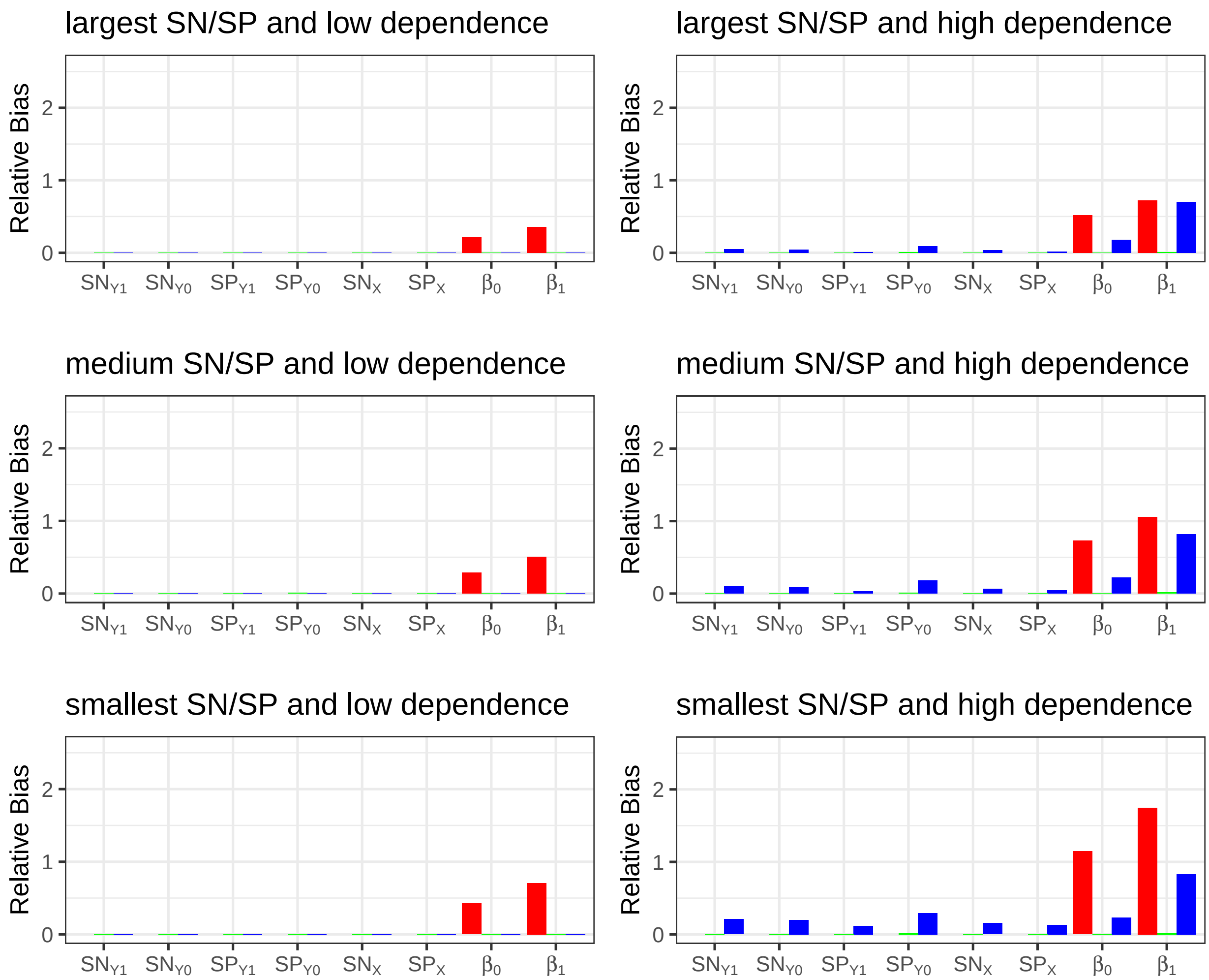

Figure S.2: Relative bias of model parameters when Y is subject to differential misclassification error and X is subject to nondifferential misclassification error with $n_v/n = 30\%$: red for naïve model, blue for independent misclassification error model, green for dependent misclassification model.

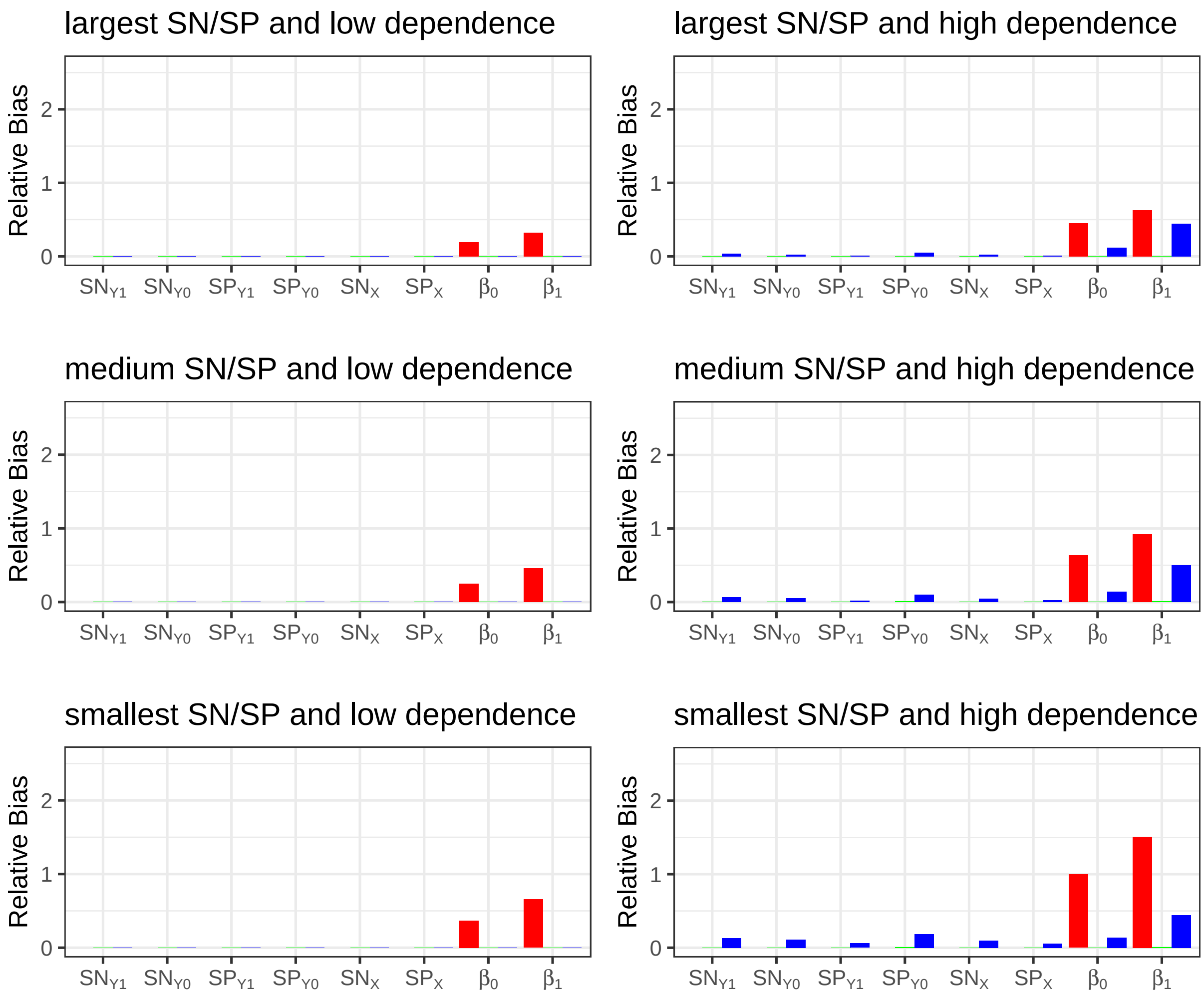

Figure S.3: Relative bias of model parameters when Y is subject to differential misclassification error and X is subject to nondifferential misclassification error with $n_v/n = 50\%$: red for naïve model, blue for independent misclassification error model, green for dependent misclassification model.

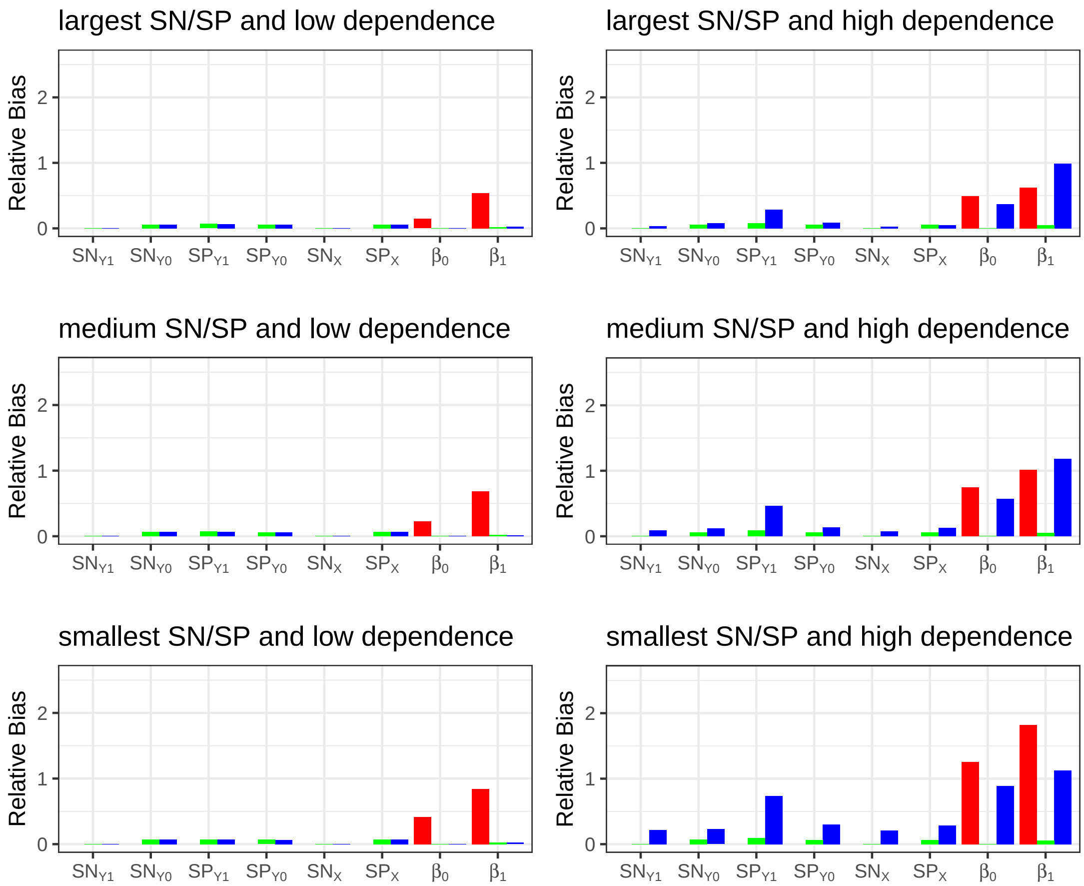

Figure S.4: Relative bias of model parameters when Y is subject to nondifferential misclassification error and X is subject to differential misclassification error with $n_v/n = 10\%$: red for naïve model, blue for independent misclassification error model, green for dependent misclassification model.

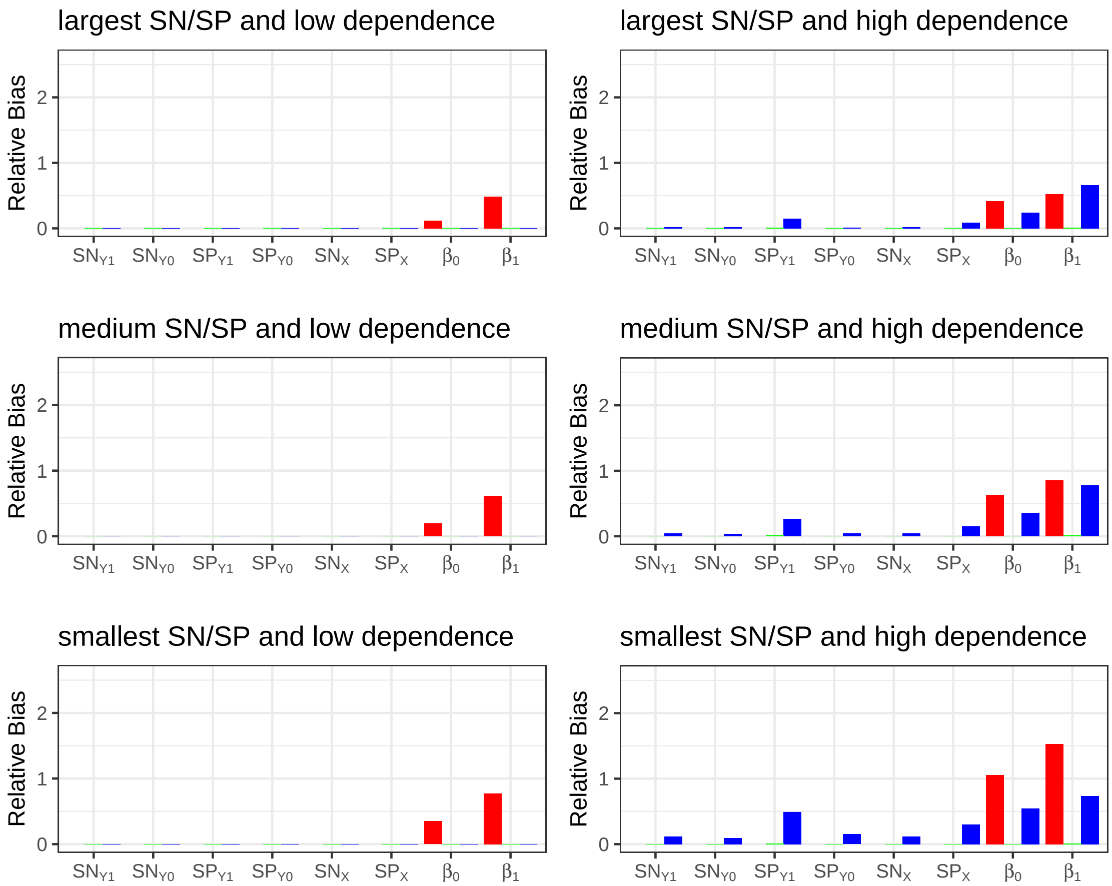

Figure S.5: Relative bias of model parameters when Y is subject to nondifferential misclassification error and X is subject to differential misclassification error with $n_v/n = 30\%$: red for naïve model, blue for independent misclassification error model, green for dependent misclassification model.

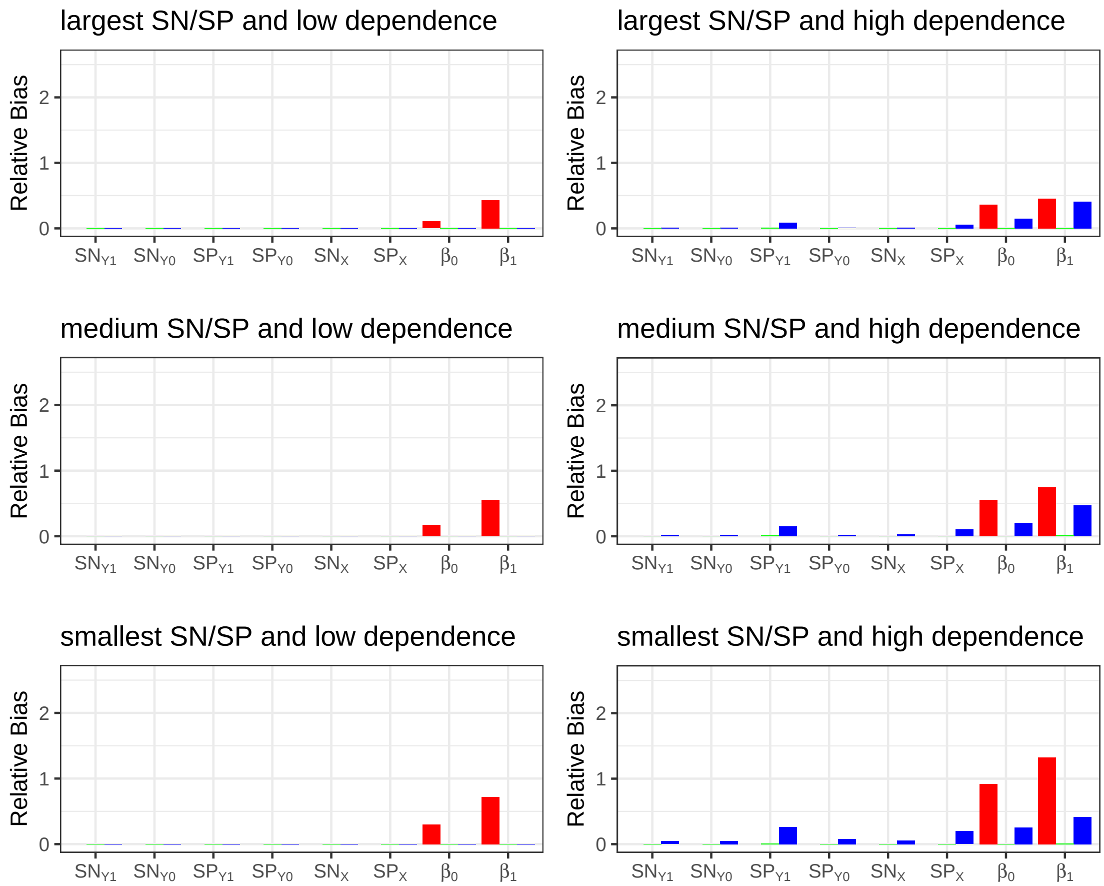

Figure S.6: Relative bias of estimated parameters when Y is subject to nondifferential misclassification error and X is subject to differential misclassification error with $n_v/n = 50\%$: red for naïve model, blue for independent misclassification error model, green for dependent misclassification model.

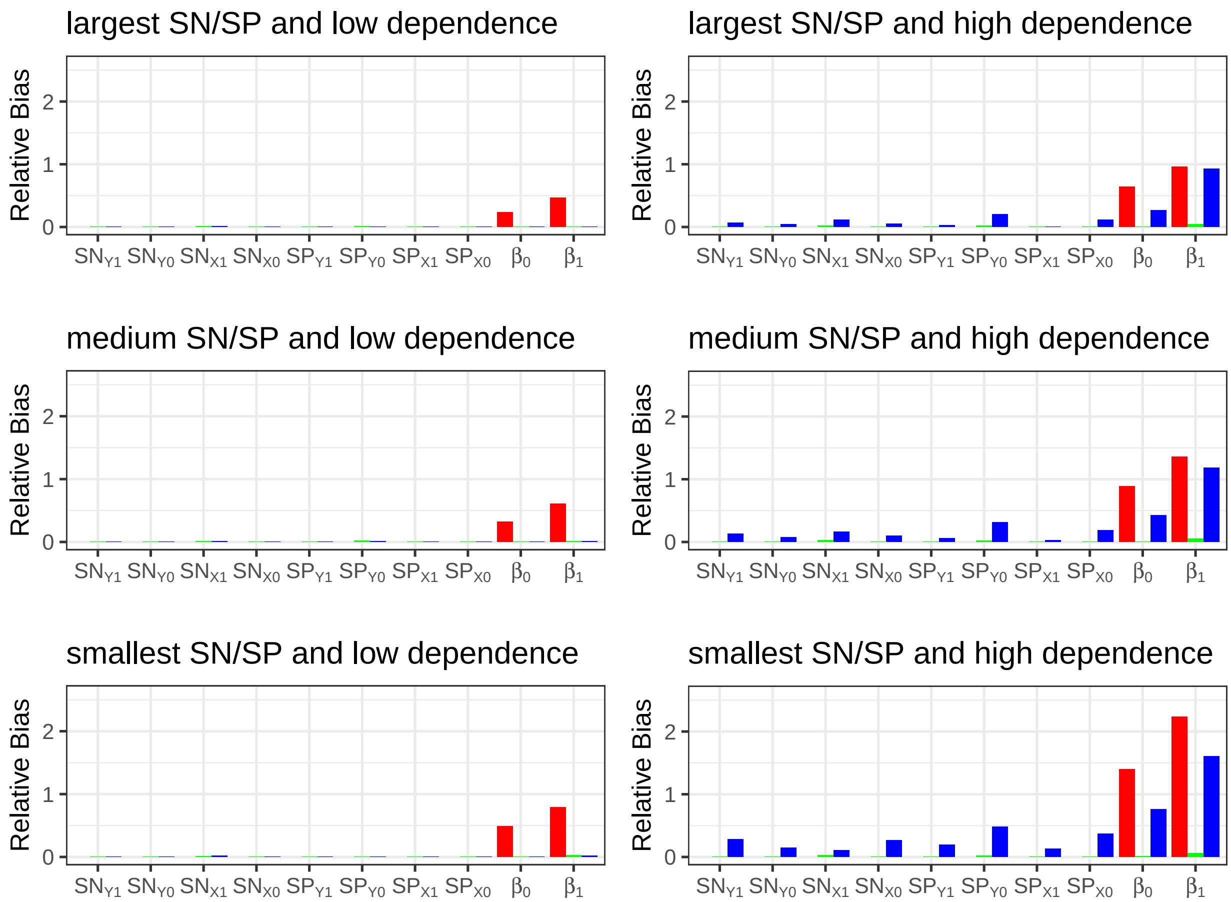

Figure S.7: Relative bias of model parameters when both Y and X are subject to differential misclassification errors with $n_v/n = 10\%$: red for naïve model, blue for independent misclassification error model, green for dependent misclassification model.

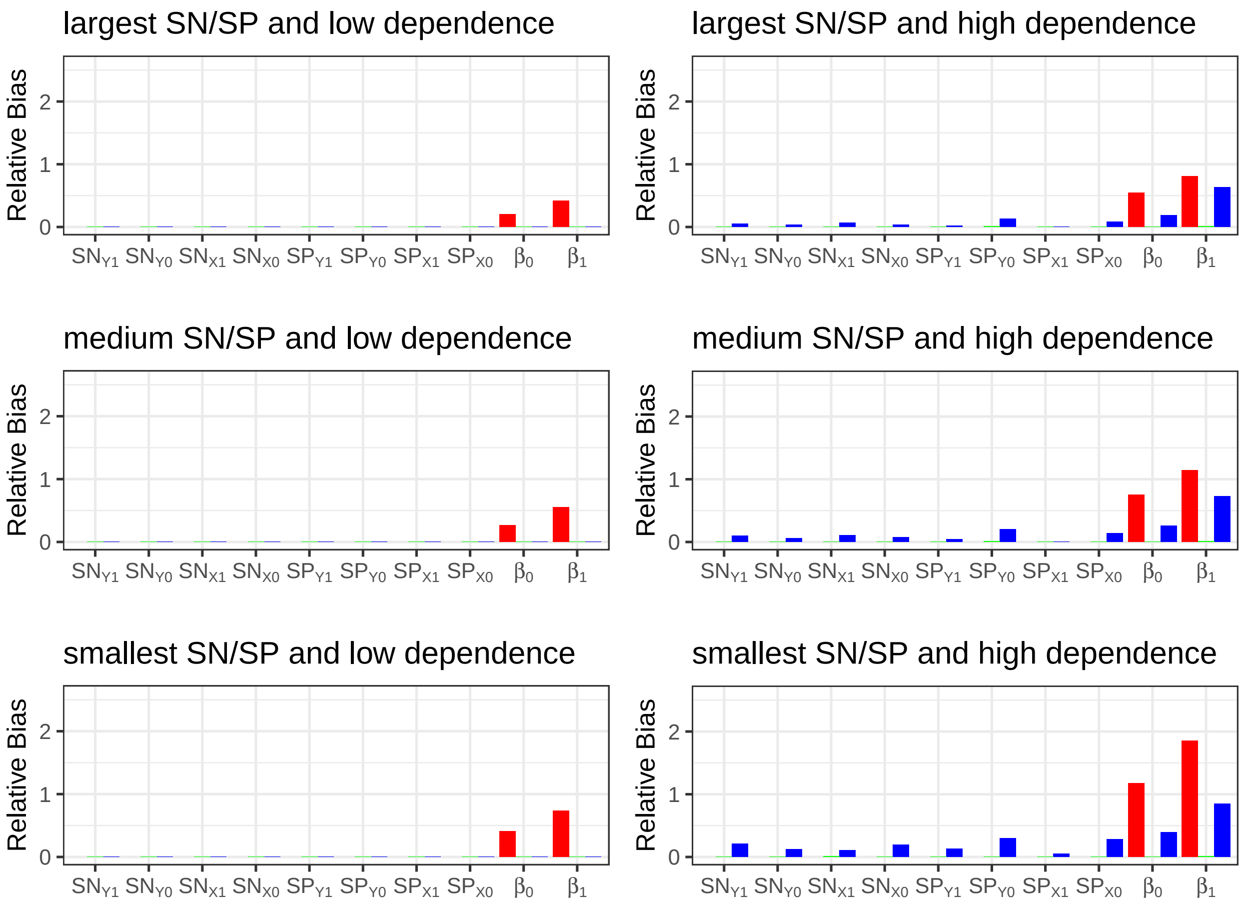

Figure S.8: Relative bias of model parameters when both Y and X are subject to differential misclassification errors with $n_v/n = 30\%$: red for naïve model, blue for independent misclassification error model, green for dependent misclassification model.

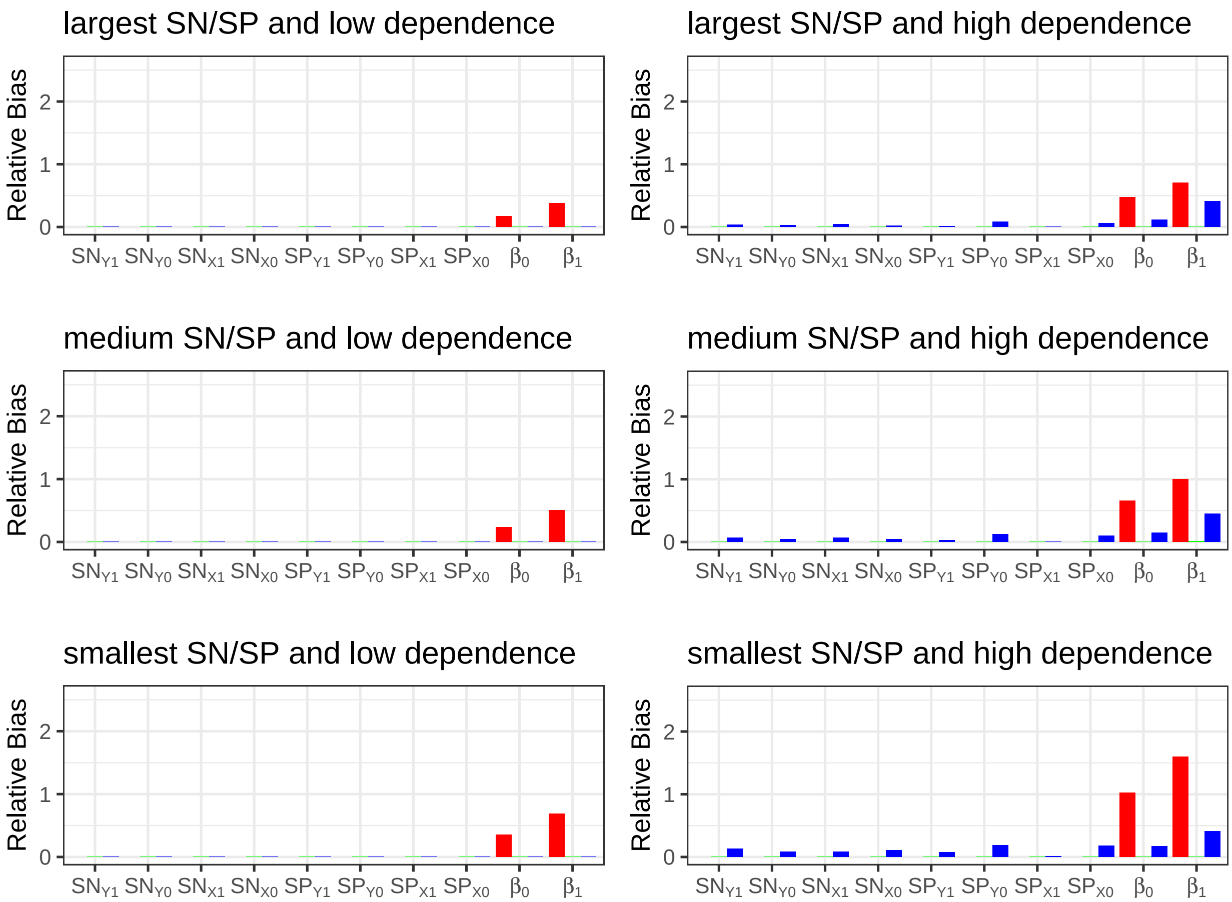

Figure S.9: Relative bias of model parameters when both Y and X are subject to differential misclassification errors with $n_v/n = 50\%$: red for naïve model, blue for independent misclassification error model, green for dependent misclassification model.

2.2 Categorical variables

Here are the plots for the simulation scenario in Section 4.2 when Y, X, Y^*, and X^* are all trinary variables.

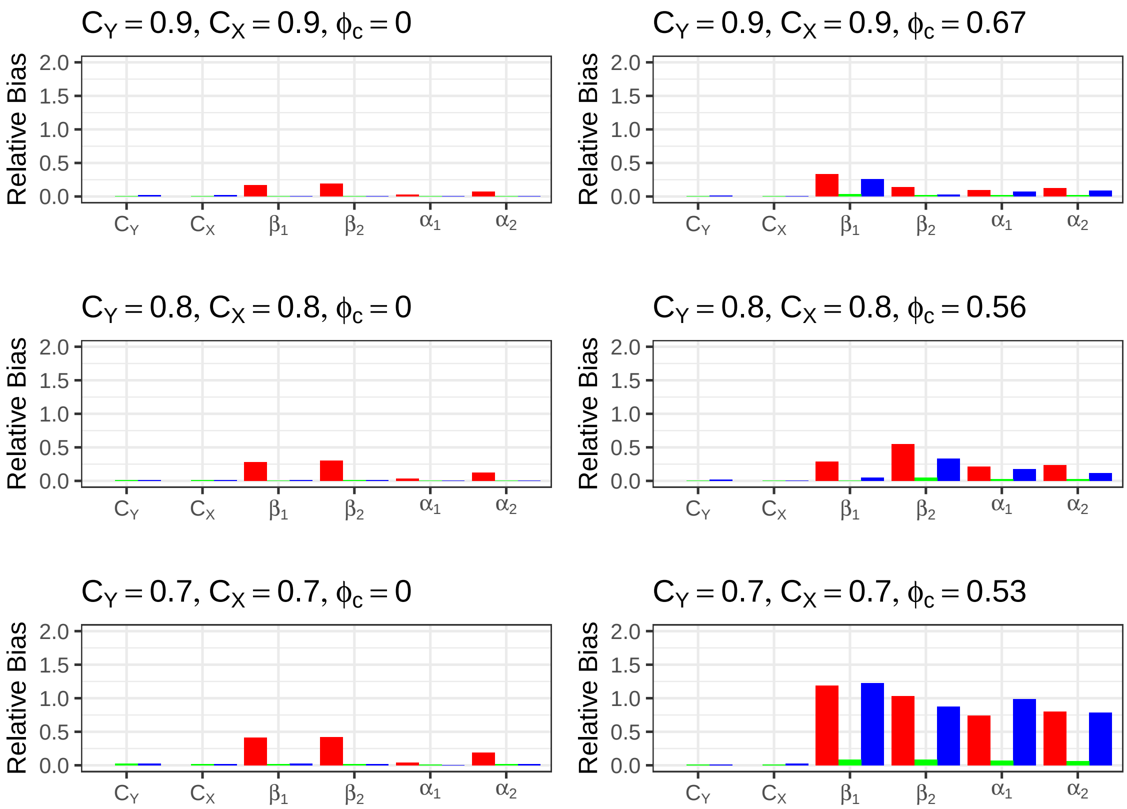

Figure S.10: Relative bias of estimated parameters when both Y and X are trinary variables with $n_v/n = 10\%$.

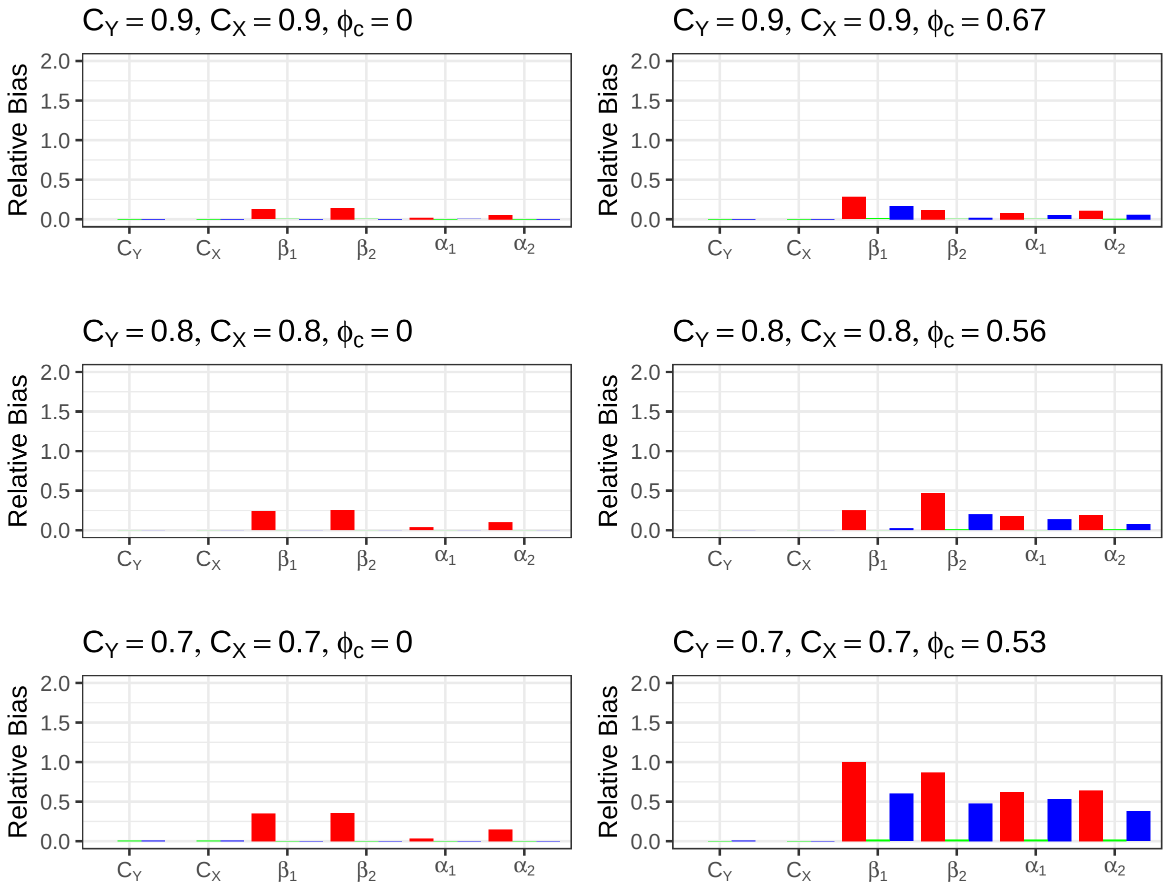

Figure S.11: Relative bias of estimated parameters when both Y and X are trinary variables with $n_v/n = 30\%$.

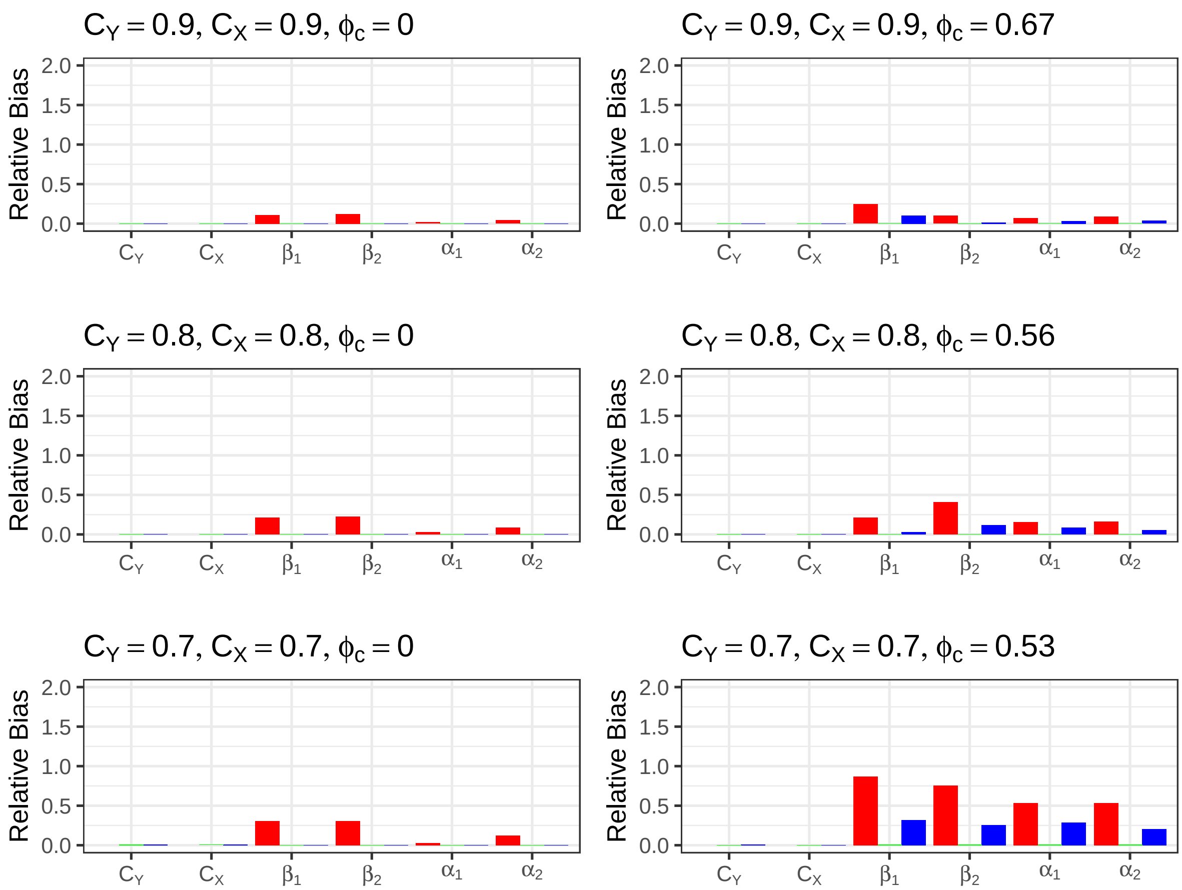

Figure S.12: Relative bias of estimated parameters when both Y and X are trinary variables with $n_v/n = 50\%$.