The why, the how, and the when to impute: a practitioners' guide to survey-to-survey imputation of poverty

January 9, 2025

Paul Corral, Andres Ham, Leonardo Lucchetti, and Henry Stemmler

The below is a short excerpt from a larger study that will be published later this Fiscal Year.

0.0.1 Omitted Variables

Omitted variables and endogeneity have traditionally not been major concerns in S2S or small area estimation. This is because, in the current period, any omitted variables are accounted for by the constant term, which adjusts to ensure that OLS predictions of the dependent variable remain unbiased. However, over time, the effects of omitted variables can influence predictions considerably. For a more detailed discussion, see the annex (Annex 1.1).

The simulations presented here take as an example the S2S application of Afghanistan (Barriga-Cabanillas et al., 2023). The authors presume that the decrease in poverty predicted in rural areas is driven by a reduction in conflict. The models do not control for conflict. Additionally, the authors follow recommendations of the SWIFT program and include "fast-changing consumption variables to better capture welfare changes during shocks" (Barriga-Cabanillas et al. 2023, p6). Among the variables the authors include to the model are a list of food consumption dummies that include meat, eggs, and chocolate. However, it is quite likely that the model used by the authors suffers from omitted variable bias. Mainly, conflict is likely negatively correlated to expenditure and is also likely to be negatively correlated to these consumption dummies as it may affect the availability of these goods. The discussion in the annex (Annex 1.1) illustrates that in this case, the omission of conflict from the model is likely to lead to biased estimates of poverty due to omitted variable bias (OVB).

To simulate the potential impact of omitted variable bias on poverty predictions over time the data simulated in section **??** is expanded to include 2 new covariates:

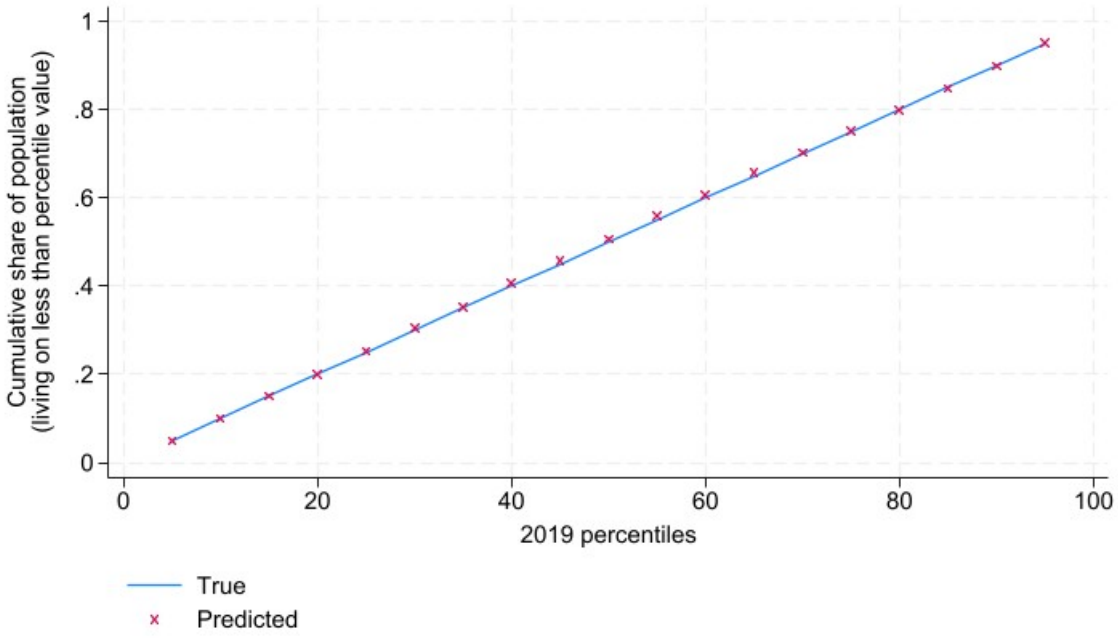
1. Conflict. Conflict is assumed to be negatively correlated with welfare with a coefficient equal to -0.3.
   (a) It is simulated as a binary variable taking value 1 when a random uniform number between 0 and 1 is less than $c = 0.4$
2. Eggs purchased. The variable is assumed to be positively correlated to welfare with a coefficient equal to 0.4.
   (a) It is simulated as a binary variable taking value 1 when a random uniform number between 0 and 1 is less than $(0.5x + 0.5Conflict)$, with $x = 1$ in the baseline. The DGP for the covariate illustrates the negative correlation of conflict and the likelihood of purchasing eggs.

Under the baseline simulation, the value for $x = 1$ and $c = 0.4$. Welfare is simulated assuming the true DGP, and thus includes conflict and eggs with their true coefficients. The prediction model is produced omitting conflict and is executed using the complete data to avoid other potential sources of bias (Table 1). As presented in Annex 1.1, the coefficient on eggs is upward biased. Due to the omission of conflict, the constant term of the regression is downward biased. The adjustment in the constant term ensures that the model's prediction of the dependent variable is unbiased. As can be seen, omitting conflict from the model also leads to a larger RMSE, which will likely affect poverty predictions over time. Nevertheless, for predictions to the same period OVB has little if any impact (1).[1]

Table 1: Omitted variable model comparison

	OVB Model	True Model
x1	0.1022	0.1017
x2	0.4965	0.4970
x3	-0.2492	-0.2481
x4	-0.1997	-0.1995
x5	-0.1423	-0.1462
eggs	0.5745	0.3927
conflict		-0.3182
Intercept	2.8264	3.0092
R2	0.5327	0.5634
RMSE	0.5120	0.4949
Observations	20,000	20,000
$\hat{y}$	2.9826	2.9826

Figure 1: Resulting poverty prediction under OVB in the same period



Source: Based on simulated data illustrated in Section **??** with added covariates. Predictions are assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5.

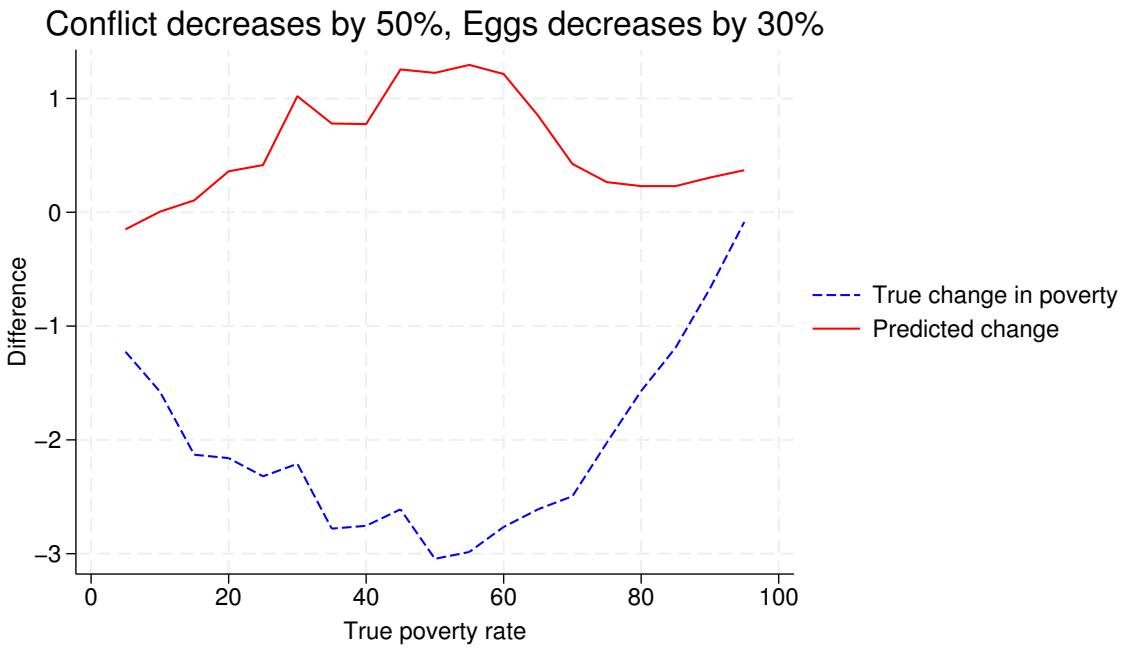
For the simulation, different values of c and x are determined, which yield a change in conflict likelihood

[1]The reason for this is that even in the presence of OVB var $\left[x\hat{\beta}\right] + \hat{\sigma}_e^2$ will be the same to the one where conflict is present and since $\hat{y}$ is also equal poverty predictions in the same period are unaffected.

and egg purchases. No other covariate is changed. Using the adjusted covariates, the true resulting welfare is calculated – including conflict and the correct coefficients. Then the model parameters obtained from the original data, excluding conflict, is used to obtain predictions of poverty across the welfare distribution. As can be seen in the discussion in the Annex 1.1, the direction of the prediction bias is not immediately clear. In many instances the predictions will suggest an increase in poverty whereas in reality poverty has dropped. In other instances it may underestimate the total change.

Under the assumed data-generating process (DGP), imagine that conflict decreases by 50% ($c = 0.2$), and egg purchases decline by 30% (x =0.7) due to factors unrelated to conflict.[2] In such a case, S2S predictions over time would likely show an increase in poverty based on most thresholds. However, in reality, poverty would have actually decreased (Figure 2). This is just an illustrative example where the purpose is to illustrate the potential problems encountered when imputing over time, particularly in instances that a major shock has occurred in between the year of the actual welfare data and the target data. The inclusion of "fast-changing" consumption variables as covariates is likely to introduce OVB to prediction models for imputations to other periods unless the shock only affects welfare but not individual components of welfare.

Figure 2: Resulting poverty prediction under OVB in a different period



Source: Based on simulated data illustrated in Section **??** with added covariates. Predictions are assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5.

The impact of omitted variable bias extends beyond "fast-changing" consumption variables. Consider a model that includes an indicator for subsistence farmers, who are more likely to be poor and thus negatively correlated with welfare. Now imagine a massive drought occurs between the year the model was trained and the target year. This drought could reduce the number of people farming. If those former farmers are unable to meet their needs, the model might underestimate poverty because it omitted the effect of the drought. Another example involves the omission of remittances which are positively related to welfare and likely positively related to "fast-changing" consumption variables, such as eggs or

[2]Note that because of the relationship to conflict egg purchases do not actually drop by 30% since it is offset by the decrease in conflict.

meat. Under this instance the coefficient on eggs would be upward biased, and since the impact of the omitted variable will be absorbed by the intercept it will likely also be upward biased. Thus, changes in remittances that are not captured by the model are likely to lead to biased predictions.

The magnitude and direction of the bias that arises due to OVB is often not clear and could compound other biases just as easily as it could be offset by other biases. Nevertheless, unless actual welfare data are available for the imputed year it is impossible to truly know how the model is performing.

References

Barriga-Cabanillas, O., Chawla, P., Redaelli, S., & Yoshida, N. (2023). *Updating poverty in afghanistan using the swift-plus methodology* (tech. rep.). The World Bank.

Yoshida, N., Takamatsu, S., Yoshimura, K., Aron, D. V., Chen, X., Malgioglio, S., Shivakumaran, S., & Zhang, K. (2022). *The concept and empirical evidence of swift methodology*. World Bank.

1 Annex

1.1 The Case of Afghanistan - Omitted Variable Bias

The Afghanistan team's discussion inadvertently demonstrates why survey-to-survey imputation across time periods requires careful consideration. While Barriga-Cabanillas et al. (2023) report overall poverty reduction driven by rural improvements—despite GDP contraction, reduced aid, and a locust outbreak—they attribute this to decreased conflict.[3] However, since conflict is not explicitly included in their model, its effects would only manifest through changes in other variables. This potential omitted variable bias raises concerns about the model's reliability for temporal predictions. When imputing over time, it is possible for multiple issues to arise. Since true expenditure is not available it is impossible to know if the model's biases cancel eachother out or compound.

A key aspect of the team's imputation is the inclusion of fast changing consumption variables, such as an indicator of whether or not the household consumed eggs over the past week. Barriga-Cabanillas et al. (2023, p2) follow the proposed approach from Yoshida et al. (2022) and argue that these dummies are included "to better capture welfare changes in a context where large economic shocks have occurred". This is what the authors refer to as the "SWIFT Plus methodology". Nevertheless, these variables are likely subject to omitted variable bias which can lead to biased estimates of the coefficients leading to biased poverty measures. In the case of the Afghanistan imputation the omitted variable is conflict. In others it may be the introduction of a cash transfer, for example.

If conflict is negatively related to expenditure and negatively related to consumption of certain goods, such as meat and eggs, this would lead to coefficients that are upward biased. Assume the following, simplified, model:

$$Y = \beta_0 + \beta_1 Eggs + \beta_2 Conflict + \varepsilon \tag{1}$$

By omitting conflict the model is now:

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 Eggs + u \tag{2}$$

where $u = \beta_2 Conflict + \varepsilon$. In a model where conflict is not included $\tilde{\beta}_1$would be greater than β_1. The OLS estimator for $\tilde{\beta}_1$ is given by:

$$\tilde{\beta}_1 = \frac{\text{Cov}\left(Eggs, Y\right)}{\text{Var}\left[Eggs\right]}$$

$$\tilde{\beta}_1 = \frac{\text{Cov}\left(Eggs, \beta_0 + \beta_1 Eggs + \beta_2 Conflict + \varepsilon\right)}{\text{Var}\left[Eggs\right]}$$

and relying on the linearity of covariance:

$$\tilde{\beta}_1 = \beta_1 + \frac{\beta_2 \text{Cov}\left(Eggs, Conflict\right)}{\text{Var}\left[Eggs\right]}$$

[3] Note that nothing to control for those shocks was included in the original model.

Since we assume that conflict and welfare are negatively correlated, and negatively correlated with the likelihood of buying eggs, then we know that $\beta_2 < 0$ and that $\text{Cov}\left(Eggs, Conflict\right) < 0$. This would lead to $\frac{\beta_2 \text{Cov}(Eggs, Conflict)}{\text{Var}[Eggs]} > 0$ which means that $\tilde{\beta}_1$ is upward biased.

In addition, the intercept will be underestimated. The intercept is equal to:

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \overline{Eggs}$$

$$\tilde{\beta}_0 = \beta_0 + \beta_1 \overline{Eggs} + \beta_2 \overline{Conflict} - \tilde{\beta}_1 \overline{Eggs}$$

$$\tilde{\beta}_0 = \beta_0 + \beta_2 \overline{Conflict} + (\beta_1 - \tilde{\beta}_1) \overline{Eggs} \tag{3}$$

we know that $\beta_2 \overset{`}{<} 0$, and that $(\beta_1 - \tilde{\beta}_1) < 0$. This suggest that the intercept term is downward biased.

When predicting over time, the direction of the prediction bias is dictated by changes in conflict and the share of households who purchased eggs. Any change in conflict, will lead to a change in egg purchases and expenditure. The direction of the bias in prediction is undetermined:

$$E\left[\bar{\tilde{y}} - \bar{y}\right] = E\left[\tilde{\beta}_0 + \tilde{\beta}_1 \overline{Eggs}' - \beta_0 - \beta_1 \overline{Eggs}' - \beta_2 \overline{Conflict'}\right]$$

replace $\tilde{\beta}_0$ with the value of Eq. 3, and re-arrange:

$$E\left[\bar{\tilde{y}} - \bar{y}\right] = E\left[\beta_2 \left(\overline{Conflict} - \overline{Conflict'}\right) + \left(\beta_1 - \tilde{\beta}_1\right) \left(\overline{Eggs} - \overline{Eggs}'\right)\right]$$

then a prediction will be **upward biased** if:

$$\beta_2 \left(\overline{Conflict} - \overline{Conflict'}\right) > \left(\tilde{\beta}_1 - \beta_1\right) \left(\overline{Eggs} - \overline{Eggs}'\right)$$

and downward biased if:

$$\beta_2 \left(\overline{Conflict} - \overline{Conflict'}\right) < \left(\tilde{\beta}_1 - \beta_1\right) \left(\overline{Eggs} - \overline{Eggs}'\right)$$

and while they could cancel eachother out, it is unlikely.

Note that since poverty also depends on the model's predicted root mean squared error ($RMSE$), which would change over time since it is a function of the omitted variable – conflict – then there are 2 sources of bias. The first is due to the egg coefficient and the intercept, and the second is due to the difference in the distribution of errors which changes due to the omitted variable. A decrease in conflict would potentially lead to a shrinking of the $RMSE$, lowering the poverty likelihood of households but would be ignored in the Afghanistan exercise.le populations.