Article

Obtaining Predictions from Models Fit to Multiply Imputed Data

Sociological Methods & Research
1-11

DOI: 10.1177/0049124115610345
smr.sagepub.com

Andrew Miles[1]

Abstract

Obtaining predictions from regression models fit to multiply imputed data can be challenging because treatments of multiple imputation seldom give clear guidance on how predictions can be calculated, and because available software often does not have built-in routines for performing the necessary calculations. This research note reviews how predictions can be obtained using Rubin's rules, that is, by being estimated separately in each imputed data set and then combined. It then demonstrates that predictions can also be calculated directly from the final analysis model. Both approaches yield identical results when predictions rely solely on linear transformations of the coefficients and calculate standard errors using the delta method and diverge only slightly when using nonlinear transformations. However, calculation from the final model is faster, easier to implement, and generates predictions with a clearer relationship to model coefficients. These principles are illustrated using data from the General Social Survey and with a simulation.

Keywords

multiple imputation, prediction, missing data, linear transformation, nonlinear transformation

[1] University of Toronto Mississauga, ON, Canada

Corresponding Author:
Andrew Miles, William G. Davis Building, Room DV-3217, University of Toronto Mississauga, 3359 Mississauga Road, Mississauga, Ontario L5L 1C6, Canada.
Email: andrew.miles@utoronto.ca

Multiple imputation (MI) is an approach to handle missing data that have been growing in popularity among social scientists. MI is attractive from a statistical standpoint because it allows analysts to obtain accurate parameter estimates and standard errors under weaker assumptions about patterns of missingness than is the case with the traditional methods such as listwise deletion or mean imputation (Enders 2010). The adoption of MI technique has been facilitated by the availability of nontechnical, application-oriented treatments of the subject, and the introduction of easy to use MI routines in major statistical packages such as SAS Version 9.4 (2013), Stata Release 13 (2013), SPSS Version 22.0 (2013), and R Version 3.1 (2014). Available software makes it relatively straightforward to generate imputations, fit regression models to each of the imputed data sets, and then combine estimates and standard errors to obtain a final model.

Unfortunately, procedures for working with these final models (hereafter called ''combined models'') are much less developed, and analysts may be disappointed to find that the available software does not allow them to perform many of the postestimation tasks they are used to, such as calculating fit statistics, residuals, or predicted values. At times this omission is intentional, as not all tasks are easily adapted to MI analyses (e.g., obtaining likelihood-based fit statistics, see Table VIII in White, Royston, and Wood 2011), but in other instances it reflects a simple lack of implementation.

One area where implementation lags is in using fitted models to make predictions. Here, I use ''predictions'' as a blanket term for any value $\hat{p}$ that can be calculated by applying some type of transformation $t()$ to the vector of coefficients from a fitted model ($\hat{\beta}$).

$$\hat{p} = t(\hat{\beta}). \tag{1}$$

Predictions in this sense include marginal effects (on the scale of the linear predictor), predicted values, and out-of-sample predictions (i.e., predicted values using hypothetical covariate values). Many software packages do not have built-in prediction capabilities for models estimated using MI data, and many of the most accessible texts about MI do not offer explicit guidance on how predictions can be obtained (Allison 2002; Enders 2010; Schafer 1999; Schafer and Graham 2002). This makes it difficult for analysts who lack a deep knowledge of MI to write their own prediction routines.

White et al. (2011) are a notable exception. They note that, as with coefficient estimates, predictions can be obtained by calculating them separately

from models fit to each of the m imputed data sets and then combining them using Rubin's rules. The point estimate of a prediction is the average of the m estimates:

$$\bar{p} = \frac{1}{m}\sum_{j=1}^{m}\hat{p}_j,$$

where $\hat{p}_j$ is the prediction from imputed data set $j, j = 1, \ldots, m$. The standard error is computed using both the within and between imputation variance of the prediction. The within imputation variance is:

$$V_W = \frac{1}{m}\sum_{j=1}^{m}\widehat{SE}_j^2,$$

where $\widehat{SE}_j$ is the estimated standard error of the prediction from imputed data set $j, j = 1, \ldots, m$. Between imputation variance is the variance of the prediction across the m imputed data sets:

$$V_B = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{p}_j - \bar{p})^2.$$

With these quantities, the total variance of the prediction can be calculated as:

$$V_{\bar{p}} = V_W + V_B\left(1 + \frac{1}{m}\right), \tag{2}$$

and the standard error is $\sqrt{V_{\bar{p}}}$.

This "predict then combine" (PC) approach to calculating model-based predictions is straightforward but requires calculating predictions and their standard errors m times (and the recommended m has been increasing, see Graham, Olchowski, and Gilreath 2007). This can be tedious if the MI software being used does not automate the process and analysts must calculate these quantities themselves, and time consuming if the analysis model is complex and predictions require more than a few seconds to estimate.

A Simpler Method for Calculating Linear Predictions

Fortunately, accurate linear predictions and their standard errors can be obtained directly from the combined model. Linear predictions are calculated using equation (1), with the added stipulation that the transformation $t()$ be

restricted to the subset of linear transformations, $g()$. Linear transformations yield the same result whether they are applied to separate vectors and then combined, or if the vectors are first combined and then the transformation applied (Strang 1980:348). This implies that transforming m vectors of coefficients and then taking their mean is equivalent to first averaging the vectors and transforming the result. That is,

$$\bar{p}_l = \frac{1}{m}\sum_{j=1}^{m} g(\hat{\beta}_j) = g\left(\frac{1}{m}\sum_{j=1}^{m}\hat{\beta}_j\right), \tag{3}$$

where $\bar{p}_l$ is the combined linear prediction and $\hat{\beta}_j$ is the $k \times 1$ vector of coefficients from the model estimated using imputed data set j, $j = 1, \ldots, m$.[1]

The variance of a linear transformation can be calculated using the delta method (Greene 2008:1056):

$$V_g = \boldsymbol{s}\hat{\mathbf{V}}\boldsymbol{s}' \tag{4}$$

where $\boldsymbol{s}$ is a $1 \times k$ row vector of partial first derivatives of $g()$ (with respect to each parameter), and $\hat{\mathbf{V}}$ is the $k \times k$ covariance matrix from a fitted model.[2] As before, calculating the variance of the transformation in each imputed data set and then combining is equivalent to calculating the variance from the combined model (proof given in Online Appendix A1):

$$V_{\bar{p}} = \frac{1}{m}\sum_{j=1}^{m}\boldsymbol{s}\hat{\mathbf{V}}_j\boldsymbol{s}' + V_B\left(1+\frac{1}{m}\right) = \boldsymbol{s}\hat{\mathbf{V}}_C\boldsymbol{s}', \tag{5}$$

where $\hat{\mathbf{V}}_j$ is the $k \times k$ covariance matrix from the model fit to imputed data set j, $j = 1, \ldots, m$, and $\hat{\mathbf{V}}_C$ is the $k \times k$ covariance matrix from the combined model.

Given equations (3) and (5), it is evident that the major difference between the PC approach and the "combine then predict" approach (CP) is the number of times the predictions and standard errors are computed. Because the predictions from a combined model need only be estimated once rather than m times, the amount of time needed to estimate these parameters will be approximately,

$$t_{CP} \approx \frac{1}{m}\sum_{j=1}^{m} t_j, \tag{6}$$

where t_j is the time needed to calculate predictions from the model fit to imputed data set j, $j = 1, \ldots, m$.[3]

Table 1. Predicted Prestige Scores Obtained Using the Predict Then Combine (PC) and Combine Then Predict (CP) Methods.

	Predict Then Combine (PC)		Combine Then Predict (CP)	
	Predicted Prestige Score	*S.E.*	Predicted Prestige Score	*S.E.*
Percentile				
1%	39.10	0.37	39.10	0.37
5%	39.24	0.36	39.24	0.36
10%	39.67	0.35	39.67	0.35
25%	40.02	0.34	40.02	0.34
50%	40.38	0.34	40.38	0.34
75%	41.17	0.34	41.17	0.34
90%	42.09	0.37	42.09	0.37
95%	42.45	0.39	42.45	0.39
99%	42.88	0.41	42.88	0.41
Estimation time (in minutes)				
PC: 1.93				
CP: 0.10 (anticipated)				

Empirical Illustration

I illustrate the equivalence of the PC and CP approaches using data from the 1988 to 2010 waves of the pooled General Social Survey. Data were imputed 20 times using the expectation maximization bootstrap approach recently outlined by Honaker and King (2010). To increase efficiency, cases originally missing data on the outcome were used during imputation but excluded prior to analyses (Von Hippel 2007).[4]

The model is adapted from Blau and Duncan's (1967) status attainment model and linearly regresses respondents' occupational prestige scores on a series of dummy variables representing the highest degree earned by respondents and their fathers, on fathers' occupational prestige scores, and on indicators for each cross section of the survey. Neither coding details nor the estimated coefficients from the model are essential to the discussion and so are not presented here, but are included in Online Appendices A2 and A3 for reference.

Table 1 shows the model-predicted prestige scores obtained using both PC and CP along with the estimation times for each. These scores represent the predicted occupational prestige of several hypothetical persons (in 1988), all with only a high school education and whose fathers also have a high school education, but whose fathers also hold jobs with various levels of prestige. In

Table 2. Marginal Effects of Father's Occupational Prestige on Respondent's Occupational Prestige.

	Predict Then Combine (PC)		Combine Then Predict (CP)	
	Marginal Effect	*S.E.*	Marginal Effect	*S.E.*
Year				
1988	0.064	0.029	0.064	0.029
1989	0.080	0.029	0.080	0.029
1990	0.063	0.031	0.063	0.031
1991	0.070	0.030	0.070	0.030
1993	0.065	0.025	0.065	0.025
1994	0.092	0.021	0.092	0.021
1996	0.076	0.021	0.076	0.021
1998	0.050	0.022	0.050	0.022
2000	0.071	0.023	0.071	0.023
2002	0.059	0.022	0.059	0.022
2004	0.096	0.020	0.096	0.020
2006	0.064	0.019	0.064	0.019
2008	0.064	0.026	0.064	0.026
2010	0.082	0.027	0.082	0.027
Estimation time (in minutes)				
PC: 8.66				
CP: 0.43 (anticipated)				

all cases, the predicted values and standard errors are identical. PC took 1.93 minutes to complete using Stata's margins command,[5] compared to an anticipated 0.10 minutes for CP.[6]

Estimation directly from the combined model works equally well for marginal effects. Table 2 presents marginal effects of father's occupational prestige from a model that includes interactions between father's prestige and survey year (model included in Online Appendix A3). Once again, predictions and standard errors from the combined model exactly match those calculated using PC. However, the anticipated computational time needed for CP is only 0.43 compared to nearly 9 minutes for PC.

Nonlinear Transformations

At times, analysts may wish to perform nonlinear transformations of model coefficients such as estimating predicted probabilities from a logit model or predicted counts from a Poisson model. In such cases, PC and CP no longer give identical results. But which method should be preferred? White and

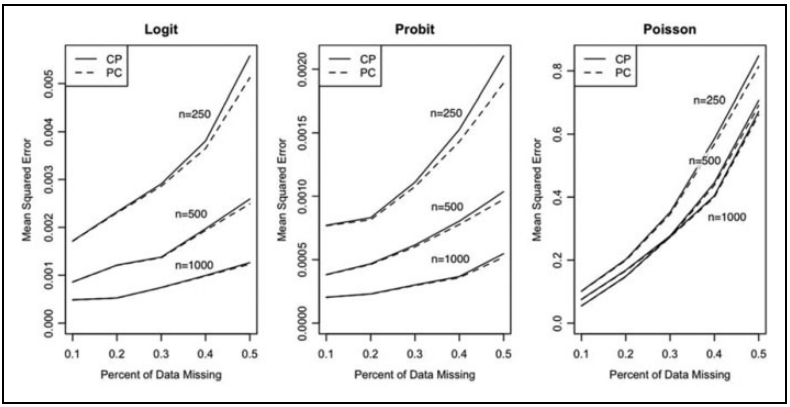

Figure 1. Mean squared error in nonlinear prediction using combine then predict (CP) and predict then combine (PC) approaches.

colleagues (2011:390) note that combining predictions either before or after a nonlinear transformation often give similar results but give no guidance on whether one method is theoretically more appropriate.

Rubin (1996) notes that MI can be used to estimate any population quantity from a survey that has a normal sampling distribution (see also Schafer and Graham 2002). Both PC and CP approaches combine quantities that can be reasonably thought of as estimates of population parameters—for PC, these are the nonlinear predictions themselves (e.g., the probability of marriage in a given culture) and for CP, they are the model coefficients used to generate these predictions. The normality condition is usually met for both approaches provided that the sample is large enough (Enders 2010:220-21). These considerations suggest that both PC and CP are appropriate for estimating nonlinear predictions from models fit to MI data.

In light of this fact, the key consideration becomes the relative advantages of each method. Here, CP stands out: it is faster, easier to implement, and generates predictions whose relationship to the final model's coefficients is direct and easy to see. But these advantages mean little if, in practice, it does a worse job of recovering population parameters. The next section, therefore, presents a simulation designed to test the relative accuracy of CP and PC.

Simulation

Figure 1 displays results from a simulation designed to compare the adequacy of CP and PC for making nonlinear predictions. Each panel shows the

mean squared error (MSE) for nonlinear predictions obtained using CP and PC for three common models—the logit, the probit, and the Poisson. These predictions are probabilities for the logit and probit models and counts for the Poisson model. The relative performance of CP and PC is compared for proportions of missing data ranging from 10 percent to 50 percent in each variable (far more than occurs in most real data situations) and with sample sizes set to 250, 500, and 1,000. In keeping with White et al.'s (2011) proposed rule of thumb, the number of imputations used was matched to the proportion of missing data (e.g., $m = 30$ when 30 percent of the data were missing). Further details on the simulation can be found in Online Appendix A4.

Results show similar patterns for all three models. MSE increases in step with the proportion of missing data, as we would expect, but decreases as sample sizes become larger. At $n = 250$, PC and CP perform similarly until the proportion of missing data reaches about 0.30, at which point PC begins to have a slightly lower MSE than CP.[7] However, these differences in MSE are minor and diminish as the sample size increases. By $n = 1{,}000$, the two methods have virtually identical MSE's at all levels of missingness. Practically, these results suggest that CP and PC perform equally well in recovering population parameters that require a nonlinear transformation of model coefficients and that this remains true even in (highly unlikely) "worst case" scenarios—small samples with high rates of missing data.

Conclusion

Predictions are useful for understanding and presenting regression results, but at present analysts wishing to obtain these quantities from models estimated with multiply imputed data have little guidance about how to do so and must use software with limited capabilities. This research note helps alleviate the conceptual side of the problem by reviewing how Rubin's rules can be used to combine predictions and their standard errors (PC method) and then demonstrating the utility of calculating predictions directly from combined models (CP method). Predictions from combined models are identical to those obtained using Rubin's rules when they involve linear transformations of the coefficients and use the delta method to obtain standard errors and are very similar when applying nonlinear transformations.

However, the CP approach also has practical advantages: It simplifies calculations, reduces computational time, and makes a clearer connection between predictions and the coefficients from the final (combined) model. It is particularly useful for analysts working in software packages that lack prediction routines for models fit to MI data. In these cases, researchers using

CP can perform one set of calculations using standard prediction formulas rather than multiple calculations in different imputed data sets, which then must be combined using Rubin's rules. These considerations make CP particularly attractive in applied MI analyses.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The linear transformation applied determines the type of prediction that is produced. For example, premultiplying a $k \times 1$ coefficient vector $\hat{\beta}$ by a $1 \times k$ vector of fixed covariate values will produce a single (hypothetical) out-of-sample prediction or (if the covariates are chosen appropriately) a marginal effect on the scale of the linear predictor. Predicted values require premultiplication by $\mathbf{X}$, an $n \times k$ model matrix of observed covariate values (including a leading column of 1's), which is not defined for combined models. $\mathbf{X}$ can be approximated by averaging the model matrices from models fit to each of the m imputed data sets. In the linear case, multiplying $\hat{\beta}$ by this averaged model matrix is equivalent to calculating predicted values in each of the m data sets and then combining using Rubin's rules.
2. As an example, suppose you have the model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ and wish to calculate the predicted value of y when $x_1 = 3$ and $x_2 = 4$. By inserting these values, you obtain the prediction equation $E(y) = 1\beta_0 + 3\beta_1 + 4\beta_2$, which describes a liner transformation of the coefficients. Calculating the standard error of the prediction requires you to know $\boldsymbol{s}$, which contains the partial derivatives of this equation with respect to each term. These are $\frac{\partial E(y)}{\partial \beta_0} = 1$, $\frac{\partial E(y)}{\partial \beta_1} = 3$, $\frac{\partial E(y)}{\partial \beta_2} = 4$, so $\boldsymbol{s} = [1 \quad 3 \quad 4]$.
3. See Online Appendix B for sample code for both methods in Stata and R.
4. Removing imputed outcome values may not be necessary with low levels of missingness, many imputations (e.g., $m > 5$), or when auxiliary variables that are highly correlated with the outcome are used in imputation (Allison 2009; Young and Johnson 2010).
5. See http://www.ats.ucla.edu/stat/stata/faq/ologit_mi_marginsplot.htm for guidance on how to use margins with multiply imputed data and the user-written margins command that implements these suggestions.

6. The pooled model estimation times in Tables 1 and 2 were approximated using equation (6). The reason is that Stata 13 currently cannot make predictions from the pooled model, and so predictions and standard errors were calculated directly using matrix operations. This makes it difficult to directly compare computation times. The approximation gives a sense for how long it would take the margins command to perform the calculations just once, for the pooled model, rather than m times (the actual computation time using matrix operations was 0.008 seconds).
7. Perhaps a more readily interpretable metric is bias in estimates. Here, again PC performs better but practically these differences are minor enough not to matter. With 50 percent missing data and a sample size of 250 (the worst case scenario), differences in bias are 0.006 and 0.007 on the probability scale for the logit and probit models, respectively, and 0.02 on the count scale for the Poisson model.

Supplemental Material

The online appendices are available at http://smr.sagepub.com/supplemental

References

Allison, Paul D. 2002. *Missing Data*. Thousand Oaks, CA: Sage.

Allison, Paul D. 2009. "Missing Data." Pp. 72-89 in *The Sage Handbook of Quantitative Methods in Psychology*, edited by Roger E. Millsap and Alberto Maydeu-Olivares. Sage.

Blau, Peter M. and Otis Dudley Duncan. 1967. *The American Occupational Structure*. New York: The Free Press.

Enders, Craig K. 2010. *Applied Missing Data Analysis*. New York: The Guilford Press.

Graham, John W., Allison E. Olchowski, and Tamika D. Gilreath. 2007. "How Many Imputations Are Really Needed? Some Practical Clarifications of Multiple Imputation Theory." *Prevention Science* 8:206-13.

Greene, William H. 2008. *Econometric Analysis*. 6th ed. Upper Saddle River, NJ: Pearson/Prentice Hall.

Honaker, James and Gary King. 2010. "What to Do about Missing Values in Time-series Cross-section Data." *American Journal of Political Science* 54:561-81.

IBM Corp. Released 2013. *IBM SPSS Statistics for Windows, Version 22.0*. Armonk, NY: IBM Corp.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rubin, Donald B. 1996. "Multiple Imputation after 18+ Years." *Journal of the American Statistical Association* 91:473-89.

SAS Institute, Inc. 2013. *SAS/STAT, Version 9.4*. Cary, NC: SAS Institute, Inc.

Schafer, J. L. 1999. "Multiple Imputation: A Primer." *Statistical Methods in Medical Research* 8:3-15.

Schafer, Joseph L. and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7:147-77.

StataCorp. 2013. *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP.

Strang, Gilbert. 1980. *Linear Algebra and Its Applications*. 2nd ed. New York: Academic Press.

Von Hippel, Paul T. 2007. "Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data." *Sociological Methodology* 37:83-117.

White, Ian R., Patrick Royston, and Angela M. Wood. 2011. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice." *Statistics in Medicine* 30:377-99.

Young, Rebekah and David R. Johnson. 2010. "Imputing the Missing Y's: Implications for Survey Producers and Survey Users." Paper presented at 64th Annual Conference of the American Association for Public Opinion Research, May 13-16, Chicago.

Author Biography

Andrew Miles is Assistant Professor of Sociology at the University of Toronto. The goal of his research is to understand the processes underlying human behavior, with an emphasis on identities, morality, and dual-process models of cognition.