

**Methods for Synthesizing the Results of Regression:
An Empirical Example of Applying Multiple Imputation**

Meng-Jia Wu & Terri Pigott
Loyola University Chicago

The purpose of this paper is to test empirically the application of multiple imputation in synthesizing regression studies when the correlation matrices are provided. This work is a pilot investigation of a project funded by National Science Foundation (DRL-0723543) starting January 2008. The grant uses as a starting point the problems with synthesizing education production functions in order to examine the application of missing data methods to the synthesis of regression models. The current paper provides an overview of previous syntheses of education production functions, the major issues encountered in combining results from these studies, and a concrete example of adopting three multiple imputation techniques to deal with the synthesis of regression models.

Previous syntheses of education production function

The education production function, also termed the input-output model, is the most frequently used approach for studying the relationship between school inputs (predictors) such as per pupil expenditures and student outputs (outcomes) such as academic achievement. Not surprisingly, the various studies estimating education production functions have produced diverse and

conflicting findings. Hanushek (1981, 1986, 1989, 1991) was the first researcher to attempt the synthesis of education production functions and came to the conclusion that per-pupil expenditure was not highly related to student achievement. Hanushek summarized 187 regression models studying the impact of school inputs on student performance in 38 separately published articles or books. He characterized the regression coefficient for each school input according to its direction (positive or negative) and significance status (significant or non-significant), and counted the number of significant predictors to arrive at his conclusion. The tally procedure Hanushek used is referred to as the “vote-count” method (Light & Pillemer, 1984) in the meta-analysis literature. This method, as pointed out by Hedges & Olkin (1980), has poor statistical properties because it ignores the magnitudes of the relationship between each input variable and an output variable.

Hedges, Laine, and Greenwald (1994) re-analyzed the studies included in Hanushek’s review. The Hedges et al. synthesis focused on estimating the magnitude of the regression slopes of per pupil expenditure and teacher salary measured in dollars. While these two predictors were measured on the same scale, the output variables in the education production functions, student achievement, were not all measured on the same scale. Thus, the slopes of the per pupil expenditure and teacher salary could not be compared directly. To solve this problem, Hedges et al. calculated “half-standardized slopes”, the slope standardized by the standard deviation of the student achievement output

variable. The half-standardized slopes then represent the number of standard deviations of change in student achievement output associated with a one dollar change in the input variables, per-pupil expenditure and teacher salary. A few years later, Greenwald, Hedges & Laine (1996) expanded their search criteria and included more education production function studies in a new synthesis.

Greenwald, Hedges & Laine's method has one major drawback. The use of half-standardized slopes does not address the problem that the half-standardized slopes of any given predictor, say per-pupil expenditure, do not have the same meaning across studies due to the inclusion of additional and different predictors in the model. Such "unparallel" models make it inappropriate to combine the results across studies directly, because the effects of different predictors are held constant in different studies.

Linking unparallel models to missing data patterns

The problem of unparallel models can be re-conceptualized as the issue of missing data in the research. Table 1 provides a hypothetical example of four regression studies. Let us assume that each model estimates the same outcome Y and each includes a predictor X_1 . The studies differ in the number of additional predictors included in each model (X_2 , X_3 , and X_4). The column on the left side of the table shows the four statistical models in the four studies and the columns on the right side show the data structure for each variable in each model. When the original data from four studies are concatenated in the way shown in Table 1, a

variable that is not included in a study-specific model (e.g., X_3 in Study 2) can be seen as a missing predictor from the “full model”, defined as a model that contains all four independent variables. The data for the missing variable in studies constitute “missing blocks” (Little & Rubin, 2002) in the multivariate dataset after combining Studies 1, 2, 3, and 4.

Table 1. The monotone missing pattern

		Y	X_1	X_2	X_3	X_4
Study 1: $Y_{1l} = b_{10} + b_{11}X_{11l} + e_{1l}$	Study 1	Y_{11}	X_{111}	Missing	Missing	Missing
		Y_{12}	X_{112}			
		\vdots	\vdots			
		Y_{1n_1}	X_{11n_1}			
Study 2: $Y_{2l} = b_{20} + b_{21}X_{21l} + b_{22}X_{22l} + e_{2l}$	Study 2	Y_{21}	X_{211}	X_{221}	Missing	Missing
		Y_{22}	X_{212}	X_{222}		
		\vdots	\vdots	\vdots		
		Y_{2n_2}	X_{21n_2}	X_{22n_2}		
Study 3: $Y_{3l} = b_{30} + b_{31}X_{31l} + b_{32}X_{32l} + b_{33}X_{33l} + e_{3l}$	Study 3	Y_{31}	X_{311}	X_{321}	X_{331}	Missing
		Y_{32}	X_{312}	X_{322}	X_{332}	
		\vdots	\vdots	\vdots	\vdots	
		Y_{3n_3}	X_{31n_3}	X_{32n_3}	X_{33n_3}	
Study 4: $Y_{4l} = b_{40} + b_{41}X_{41l} + b_{42}X_{42l} + b_{43}X_{43l} + b_{44}X_{44l} + e_{4l}$	Study 4	Y_{41}	X_{411}	X_{421}	X_{431}	X_{441}
		Y_{42}	X_{412}	X_{422}	X_{432}	X_{442}
		\vdots	\vdots	\vdots	\vdots	\vdots
		Y_{4n_4}	X_{41n_4}	X_{42n_4}	X_{43n_4}	X_{44n_4}

Note. Y_{kl} : outcome score for person l in study k ; X_{kil} : score on predictor i for person l in study k ; b_{ki} : the slope for predictor i in study k ; n_k : sample size for study k .

It is important to distinguish the patterns of missing data in order to determine the methods for handling missing data. The example in Table 1 is said

to have a “monotone” missing data pattern (Little & Rubin, 2002), which indicates that when the value of a variable for an individual is missing, the values for the subsequent variables are all missing for that individual. Otherwise, the data set with missing data has an arbitrary missing pattern.

Application of multiple imputation techniques

Multiple imputation (Rubin, 1976) is a commonly used method to deal with missing data issues. Unlike single imputation which involves estimating and filling in only one value for each data point that is missing, multiple imputation replaces each missing value with a set of plausible values to capture the uncertainty of the data that are missing. The multiply imputed data sets are then analyzed by using the procedures that are designed for complete data and combining the results from these analyses. Specially, there are three distinct steps for conducting multiple imputation:

- 1) The missing data are filled in m times to generate m complete data sets.
- 2) The m complete data sets are analyzed using standard statistical analyses.
- 3) The results from the m complete data set are combined to obtain inferential results.

Three commonly used multiple imputation methods were applied in this paper to synthesize regression studies with correlation matrices of variables in the models provided: regression imputation, propensity scores, and Markov Chain Monte Carlo (MCMC). Regression imputation is a parametric method

while propensity score method is a nonparametric method. Both of them are suitable for monotone missing data patterns. MCMC method, on the other hand, works for an arbitrary missing data pattern. Those methods were introduced in more details using an example on education production functions.

Empirical example

An example is created here to demonstrate the application of multiple imputation to combine regression studies where the correlations among all variables in the model are reported. The general rationale of these three methods are illustrated using this example, followed by a discussion of the assumptions and related issues involved in adopting these methods.

We selected one of the frequently studied outcomes in the education production function literature, student's verbal ability (Y), as the dependent variable in this example. Per-pupil expenditure (X_1), studied in the previous syntheses, was used as the focal predictor in the regression model. Three other variables, student/teacher ratio (X_2), teacher's education level (X_3), and family size (X_4), which represent the influences from school, teacher and family respectively were used as other control variables in the regression models. Five studies were created with different control variables involved:

$$\text{Study 1: } Y_{1i} = B_{11}X_{11i} + e_{1i}$$

$$\text{Study 2: } Y_{2i} = B_{21}X_{21i} + B_{22}X_{22i} + e_{2i}$$

$$\text{Study 3: } Y_{3i} = B_{31}X_{31i} + B_{32}X_{32i} + B_{33}X_{33i} + e_{3i}$$

$$\text{Study 4: } Y_{4i} = B_{41}X_{41i} + B_{42}X_{42i} + B_{43}X_{43i} + B_{44}X_{44i} + e_{4i}$$

$$\text{Study 5: } Y_{5i} = B_{51}X_{41i} + B_{53}X_{53i} + e_{5i}$$

where

Y_{ki} is the i th student's achievement score in the study k (n_k is the sample size for study k);

B_{kj} is the standardized slope for the j th predictors in study k ;

X_{kji} is the value of the predictor X_j for the i th student in study k ; and

e_{ki} is the error of the i th student in study k .

Two sets of syntheses were conducted to test three methods. The first synthesis comprised studies one through four, which created the monotone pattern of missing data described earlier. The precision of the regression and propensity scores methods was investigated. The second synthesis comprised studies one through five, which allowed us to investigate the precision of the MCMC method in dealing with arbitrary missing pattern. In order to examine the precision, we created the example as if all the studies were from the same population. That is, the correlations among variables are all the same. The major difference between studies is the predictors that were involved. The population correlations used to create this example were from Dugan (1976), as shown in Table 2.

Based on the correlation matrix, the complete regression model with all the variables involved is:

$$Y_i = 0.278X_{1i} - 0.178X_{2i} + 0.173X_{3i} + 0.160X_{4i} + e_i$$

Table 2 Correlation matrix from Dugan (1976)

	Verbal ability (Y)	PPE ^a (X ₁)	S/T ratio ^b (X ₂)	Teacher ed (X ₃)	Family size (X ₄)
Y	1	.378	-.275	.263	.055
X ₁		1	-.323	.431	-.198
X ₂			1	-.067	.025
X ₃				1	-.262
X ₄					1

a. PPE= Per-pupil expenditure

b. S/T ratio=Student/teacher ratio

c. Teacher ed=Teacher's education level

The correlation matrix for each study 1 through 5 are based on the predictors assumed observed in the studies. For example, the correlation matrix for Study 2 (\mathbf{R}_2), which includes predictors for Per-pupil expenditure and student/teacher ratio, is

$$\mathbf{R}_2 = \begin{bmatrix} 1 & .378 & -.275 \\ & 1 & -.323 \\ & & 1 \end{bmatrix}$$

The correlation matrix for each study was used to generate the individual school level data that would result in that particular correlation matrix. Thus, the primary school-level data for each study, consisting of the per-pupil expenditure, student-teacher ratio and school achievement was generated to be consistent

with the resulting correlation matrix. In this example, we assumed equal sample size across studies. Specifically, the Cholesky decomposition was adopted to generate 1000 cases (schools) ($n_k=1000, k=1$ to 5) with the desired correlation among the variables for each study. Once the data were obtained for each study, multiple imputation was then conducted on the concatenate data (as shown in Table 1). Data from studies one through four were aggregated for investigating regression imputation and propensity methods for a monotone missing pattern; data from studies one through five were aggregated for investigating MCMC method for missing data with arbitrarily missing pattern.

After m imputations, m sets of complete data were created to estimate the slopes of the independent variables in the full model. Rubin (1987) provides the formulas for estimating the multiply-imputed point estimate for the slopes for the full model and the variance associated with the multiply-imputed slope estimates.

Monotone missing pattern: Combining studies one through four

Regression imputation. In regression imputation, a variable with missing values is imputed using the predicted value from the regression of the missing variable on the other observed variables in that case. Using the notation defined previously, for a missing variable X_j , a regression model is fitted using cases that have observed values for variables X_1 to X_5 . The general format of the fitted model looks like:

$$X_j = \beta_0 + \beta_1 X_1 + \dots + \beta_{j-1} X_{j-1} \quad (j = 1 \text{ to } 4 \text{ in this example})$$

The fitted model produces the estimated slopes and variance, and forms the posterior predictive distribution. In this example, five imputation ($m=5$) with five sets of simulated slopes and variances are drawn from that distribution. Five sets of filled-in data were generated from the five sets of simulated slopes and variances to create five complete sets of data for each of the studies one through four. More details for this method can be found in Rubin, 1987 (pp.166-167).

Table 3. The estimates of slope and standard error (SE) for each variable based on the regression method

	Slopes for the population model	Estimate	<u>Slopes</u>		SE
			Min	Max	
X_1	0.278	0.257	0.243	0.264	0.019
X_2	-0.178	-0.155	-0.157	-0.152	0.015
X_3	0.173	0.166	0.152	0.181	0.020
X_4	0.160	0.191	0.172	0.219	0.025

In Table 3, the estimates of the slope using regression method are shown in the right columns as compared to the population slopes from the complete model based on the correlation matrix from Dugan (1976). The largest difference between the population slopes and the estimated slopes based on the regression imputation method occurred in variable X_4 ($0.191-0.160=0.031$). None of the

differences between the population slopes and the estimated slopes using regression method is significant at .05 level.

Propensity score method. In the propensity score method, the conditional probability, or the propensity score, of each observation is estimated using logistic regression to represent the possibility of that observation being missing. The observations are then grouped based on the propensity scores. For each group, an approximate Bayesian bootstrap imputation is applied to fill in multiple sets of simulated data (Lavori, Dawson, and Shera, 1995). More details regarding this method can be found in Rosenbaum and Rubin (1983).

Table 4. The estimates of slope and standard error (SE) for each variable based on the propensity score method

	Slopes for the population model	Estimate	<u>Slopes</u>		SE
			Min	Max	
X ₁	0.278	0.288	0.277	0.297	0.018
X ₂	-0.178	-0.145*	-0.152	-0.138	0.016
X ₃	0.173	0.136	0.105	0.163	0.030
X ₄	0.160	0.173	0.151	0.211	0.030

* The difference between estimate slope and the population slope is significant at .05 level.

The results based on the propensity score methods for synthesizing studies one through four are shown in Table 4. The largest difference between the population slope and the estimated slopes based on the propensity score method occurred in variable X_3 ($0.136-0.173=-0.037$). After taking into account the variation in imputations, the difference between population and estimated slopes for X_2 is significant. The SEs based on the propensity score method tend to be larger than those based on the regression method.

General missing data pattern: Combining studies one through five

MCMC method. In statistics, the common usage of the MCMC method is to generate pseudo-random samples from multidimensional probability distributions via Markov chains, a sequence of random variables in which the distribution of each element depends only on the value of the previous one. The MCMC method constructs a Markov chain that is long enough to simulate stable estimates of interest. There are two repeated steps for applying the MCMC method to Bayesian inference with missing data. The imputation step simulates the missing values for a variable from a conditional distribution for missing values given the observed values. With the simulated values treated as if they were the true values, the posterior step simulates the posterior population mean vector and covariance matrix. The new estimates of the means and covariances are then used in the next imputation step. The two steps are iterated until the

results to be reliable. A thorough discussion of this method can be found in Schafer (1997).

Table 5 shows the estimates of the slope using MCMC method. This method produces estimates with smaller SEs for X_1 , X_2 , and X_3 , compared to the other two methods based on the studies with monotone missing pattern. However, the largest difference between the population slope and the estimated slopes based on the regression imputation method occurred in variable X_4 ($0.160 - 0.204 = -0.044$), which is the largest difference that have been observed among the three methods, though this difference is not significant at .05 level due to the large SE (0.034).

Table 5. The estimates of slope and standard error (SE) for each variable based on the MCMC method

	Slopes for the population model	<u>Slopes</u>			
		Estimate	Min	Max	SE
X_1	0.278	0.253	0.247	0.258	0.016
X_2	-0.178	-0.156	-0.169	-0.141	0.017
X_3	0.173	0.173	0.168	0.177	0.014
X_4	0.160	0.204	0.161	0.234	0.034

Conclusion

The example we created for this paper allowed us to examine the issues that might arise when applying multiple imputation methods to meta-analyze regression studies with correlations reported. It serves as a pilot investigation for our long term project focusing on the application of multiple imputation methods to synthesize education production functions. Generally speaking, the results based on regression imputation, propensity scores, and MCMC methods produced slope estimates similar to the population slopes. Though one important thing we have to keep in mind is that, by adopting these methods, we assume the missing variables from a study are missing at random (MAR). As Little and Rubin (2002) discuss, when a variable is MAR, the reason that this variable is missing is not due to the possible values for that missing variable but can be related to other completely observed variables in the data. Results from multiple imputation depend on the assumption of MAR data.

During the process of generating the data that reflect the desired correlation for each study, we realized that the data we simulate will never exactly reproduce the original, primary data. The correlation based on the simulated data approaches the desired correlation as the sample size for the study gets larger. We avoided this complication by creating 1000 cases for each study in this example as the trial to first applying the methods in the synthesis context. We will explore the impact of small and/or unequal sample sizes (therefore different rate of missing data for variables) in the near future.

Last, in the example, our imputation model is the same as the analysis model. Both of them contain the same four predictors. This may not be the case as we proceed in our project to synthesize the education production function literature, considering the diversity of the variables used in the literatures. Schafer (2003) discussed this issue in great detail and the suggestions from his work will be incorporated in our future study.

References

- Dugan, D. J. (1976). Scholastic achievement: Its determinants and effects in the education industry. *Education as an industry: a conference of the Universities-National Bureau Committee for Economic Research*. J. T. Froomkin, D. T. Jamison and R. Radner. New York, National Bureau of Economic Research: 53-83.
- Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student-achievement. *Review of Educational Research*, 66(3), 361-396.
- Hanushek, E. A. (1981). Throwing money at schools. *Journal of Policy Analysis and Management*, 1(1), 19-41.
- Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, 24, 1141-1177.
- Hanushek, E. A. (1989). The impact of differential expenditures on school performance. *Educational Researcher*, 18(4), 45-51.
- Hanushek, E. A. (1991). When school finance "reform" may not be good policy. *Harvard Journal on Legislation*, 28, 423-456.
- Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher*, 23(3), 5-14.
- Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, 93, 563-573.
- Light, R. J. & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge MA: Harvard University Press.
- Little, R. J and Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: John Wiley & Sons, Inc.
- Lavori, P. W., Dawsibm R., and Shera, D. (1995). A multiple imputation strategy for clinical trials with truncation of patient data. *Statistics in medicine*, 14, 1913-1925.

Rosenbaum, P. R. And Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons, Inc.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*, New York: Chapman and Hall.