HOW TO IMPUTE INTERACTIONS, SQUARES, AND OTHER TRANSFORMED VARIABLES

*Paul T. von Hippel**

Researchers often carry out regression analysis using data that have missing values. Missing values can be filled in using multiple imputation, but imputation is tricky if the regression includes interactions, squares, or other transformations of the regressors. In this paper, we examine different approaches to imputing transformed variables; and we find one simple method that works well across a variety of circumstances. Our recommendation is to transform, then impute*—i.e., calculate the interactions or squares in the incomplete data and then impute these transformations like any other variable. The transform-then-impute method yields good regression estimates, even though the imputed values are often inconsistent with one another. It is tempting to try and "fix" the inconsistencies in the imputed values, but methods that do so lead to biased regression estimates. Such biased methods include the* passive imputation *strategy implemented by the popular* ice *command for Stata.*

1. TRANSFORMING VARIABLES IN MULTIPLE IMPUTATION

Multiple imputation (MI) is an increasingly popular method for repairing data with missing values (Rubin 1987). In MI, the researcher makes

I thank Paul Allison, Doug Downey, Jerry Reiter, and Donald Rubin for helpful feedback on an earlier draft. Direct correspondence to Paul T. von Hippel, Department of Sociology, Ohio State University, Columbus OH 43210; e-mail: von-hippel.1@osu.edu.

*Ohio State University

several copies of the incomplete data set and replaces missing values with random imputations that are consistent with the data's multivariate structure. The researcher then analyzes each imputed data set separately and combines the separate analyses to get a single set of final estimates.

MI requires two statistical models: an *imputation model* and an *analysis model*. The *imputation model* is used to fill in missing values. The *analysis model* is used to analyze the imputed data. For example, the analysis may be a regression of Y on X.

It is generally recognized that the imputation model and analysis model should be *compatible*. That is, any relationship in the analysis model should also be part of the imputation model. For example, if a complete variable Y is to be regressed on an incomplete variable X, then missing X values should be imputed conditionally on Y. Failing to condition on Y—that is, omitting Y from the imputation model—will result in imputed X values that have no relationship to Y. When the data are analyzed, the slope of Y on X will be biased toward zero, since no association between X and Y was allowed for in the imputation model.

The compatibility requirement also applies to squared term, interactions, and other transformed variables. For example, if Y is to be regressed on X_1, X_2, the squared term X_2^2, and the interaction $I_{12} = X_1X_2$, then the square and the interaction should play some role in the imputation model. If these transformed variables are omitted from the imputation model, then their slopes in the regression analysis will be biased toward zero.

Although it is generally agreed that MI should account for squares and interactions, it is not agreed how best to do this. Aside from a two-page discussion of interactions in Allison (2002), there is little research that addresses the topic. In this note, we review several approaches to imputing squares and interactions. We find that certain methods—including methods that are implemented in popular software—have biases, some of which have not been remarked before.

When using imputation software that relies on a parametric imputation model—for example, the MI procedure in SAS or the *ice* command in Stata—there are two principal approaches to imputing transformations such as squares or interactions. One method—which we call *transform, then impute*—is to calculate the squares and interactions in the incomplete data, and then impute these transformations like any other variable. The other method—which we call *impute, then*

transform—is to impute variables in their raw form, and then calculate transformations such as squares and interactions in the imputed data.

We find that, when the regression model includes interactions, the transform-then-impute method yields good regression estimates, while the impute-then-transform method is biased. Allison's (2002) results suggest this, but we provide a detailed mathematical justification for the result. We also go further than Allison, demonstrating that the same result holds for squared terms as well as interactions, and for logit and probit analysis as well as linear regression. In addition, we demonstrate that several more sophisticated approaches to imputing squares and interactions—including the *passive imputation* method implemented in Royston's (2005) *ice* command for Stata—are just disguised variants of the impute-then-transform method, and share that method's biases.

In all of these situations, the source of bias is the same: the analysis model is inconsistent with the imputation model. The analysis model includes the transformed variables and specifies their relationship to Y. The imputation model, by contrast, either ignores these relationships or undoes them by editing the imputed data.

2. IMPUTING LINEAR TERMS

To review the basics of imputation, let's temporarily leave transformation out of the picture and suppose that we just want to carry out a normal linear regression of Y on X. Let's also suppose, without losing generality, that X has been standardized, and that Y has been scaled so that the regression intercept is zero and the regression residual $e_{Y \cdot X}$ is standard normal. So the regression model is

$$\begin{aligned} &Y = \alpha_{Y \cdot X} + \beta_{Y \cdot X} X + e_{Y \cdot X}, \quad \text{where} \\ &\alpha_{Y \cdot X} = 0,\ E(X) = 0,\ Var(X) = 1, \quad \text{and} \quad e_{Y \cdot X} \sim Normal(0, 1), \end{aligned} \tag{1}$$

which implies that (X, Y) have a mean and covariance matrix of

$$\mu_{XY} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \Sigma_{XY} = \begin{bmatrix} 1 & \\ 1 + \beta_{Y \cdot X}^2 & \beta_{Y \cdot X} \end{bmatrix}. \tag{2}$$

This mean and covariance matrix are not hard to calculate by hand, but as this paper progresses the calculations will get more complicated and

tedious. For help with these calculations, we use the MathStatica package, version 1.5 (Rose and Smith 2002), running under Mathematica software, version 5.2 (Wolfram Software 2005).

If both X and Y were complete, we would simply fit model (1) to the complete data. But suppose that only Y is complete, while X has missing values. If X is normal (and even if it is not), missing X values are commonly imputed by normal linear regression on Y. Using such a model, we would replace missing values of the incomplete variable X with the imputed variable $X^{(m)}$:

$$X^{(m)} = \alpha_{X \cdot Y} + \beta_{X \cdot Y} Y + e^{(m)}_{X \cdot Y} \quad \text{with} \quad e^{(m)}_{X \cdot Y} \sim Normal\left(0, \sigma^2_{X \cdot Y}\right), \qquad (3)$$

where the parameters for the regression of X on Y

$$\alpha_{X \cdot Y} = 0, \quad \beta_{X \cdot Y} = \frac{\beta_{Y X}}{1 + \beta^2_{Y X}}, \quad \text{and} \quad \sigma^2_{X \cdot Y} = \frac{1}{1 + \beta^2_{Y X}} \qquad (4)$$

can be derived, using standard formulas, from μ_{XY} and Σ_{XY} (e.g., see Johnson and Wichern 2002). In some imputation software (such as *ice* for Stata or IVEware for SAS) the regression model used to impute $X^{(m)}$ is specified explicitly, while in other imputation software (such as the MI procedure in SAS) the regression model is implicit in the assumption that (X, Y) are multivariate normal with mean μ_{XY} and covariance matrix Σ_{XY}. Notice that it does not really matter whether (X, Y) fit a bivariate normal distribution. What matters is that the imputed data reproduce the mean μ_{XY} and covariance matrix Σ_{XY} of the complete data. If they do, then a regression of Y on $X^{(m)}$ in the imputed data will have the same parameters as a regression of Y on X in the complete data, so that valid conclusions can be drawn from an analysis based on imputed values.

It is worth pointing out that the imputed values of $X^{(m)}$ are not the same as the missing values of X. In fact, when X is normal the correlation between X and $X^{(m)}$ is only $\beta^2_{Y \cdot X}/(1 + \beta^2_{Y \cdot X})$, according to MathStatica. This correlation is always less than one and can be as small as zero. In short, although X and $X^{(m)}$ fit the same regression equation, we should not expect them to have much more than that in common. This will be important to remember when we look at more complicated regressions involving squares and interactions.

Our conclusions so far are based on the idea that $X^{(m)}$ can be imputed using the population parameters μ_{XY} and Σ_{XY}. In a real data

analysis, these parameter values are not known and have to be estimated from a sample of data in which some values of X are missing. Consistent estimates $\hat{\mu}_{XY}^{(m)}$ and $\hat{\Sigma}_{XY}^{(m)}$ can be obtained from such incomplete data if values are missing *ignorably*, or *missing at random (MAR)*—that is, if the probability that X is missing depends only on observed values of Y (Little and Rubin 2002). Typically the estimates $\hat{\mu}_{XY}^{(m)}$ and $\hat{\Sigma}_{XY}^{(m)}$ that we use for imputation are drawn at random from the Bayesian posterior density of $(\mu_{12Y},\ \Sigma_{12Y})$.

Using parameter estimates instead of the true parameter values does not change the fact that consistent regression estimates can be obtained from imputed data. Since $\hat{\mu}_{12Y}^{(m)}$ and $\hat{\Sigma}_{12Y}^{(m)}$ are consistent estimates of $\mu_{12\mathrm{Y}}$ and $\Sigma_{12\mathrm{Y}}$, it follows that data which are imputed using $\hat{\mu}_{12Y}^{(m)}$ and $\hat{\Sigma}_{12Y}^{(m)}$ can be analyzed to get consistent estimates of $\mu_{12\mathrm{Y}}$ and $\Sigma_{12\mathrm{Y}}$—and therefore to get consistent estimates for the regression of Y on X.

Since imputed values have a random component, the results obtained by analyzing imputed data will be at least a little different if the data are imputed again. The variation from one imputed data set to another is called *imputation variation*, just as variation from one sample to another is called *sampling variation*. To estimate this imputation variation and incorporate it into our estimates, we repeat the cycle of imputation and analysis multiple times. More specifically, in multiple imputation we repeat M times the process of drawing estimates $\hat{\mu}_{12Y}^{(m)}$ and $\hat{\Sigma}_{12Y}^{(m)}$, $m = 1, \ldots, M$, from the posterior density of $(\mu_{12\mathrm{Y}},\ \Sigma_{12\mathrm{Y}})$. For each set of estimates we impute values $X^{(m)}$ by regression on Y, and then we regress Y on the observed and imputed values of X. The result is M sets of point estimates $\hat{\alpha}_{Y.X}^{(m)}$, $\hat{\beta}_{Y.X}^{(m)}$, and $\hat{\sigma}_{Y.X}^{(m)}$—one set for each imputed data set—along with M sets of standard-error estimates $S(\hat{\alpha}_{Y.X}^{(m)})$, $S(\hat{\beta}_{Y.X}^{(m)})$, and $S(\hat{\sigma}_{Y.X}^{(m)})$.

The M point estimates are averaged to obtain an MI point estimate. For example, an MI estimate of the intercept is

$$\hat{\alpha}_{Y.X}^{MI} = \frac{1}{M}\sum_{m=1}^{M}\hat{\alpha}_{Y.X}^{(m)}. \tag{5}$$

Again, this does not change the fundamental result: if the estimates from a single imputed data set are consistent estimates of the regression parameters, then their average will be consistent as well.

The standard error of an MI point estimate is a combination of two components. The first component is sampling variation, which

we estimate by averaging the squared standard errors from the individual imputed data sets. The second component is imputation variation, which we estimate by calculating the variance of the point estimates from one imputed data set to another. To get an MI standard error, we add these two sources of variation and take the square root of the sum. For example, an MI standard error for the regression intercept is

$$\hat{S}_{MI}(\hat{\alpha})_{Y.X}^{MI} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left[S\left(\hat{\alpha}_{Y.X}^{(m)}\right)\right]^2 + \frac{1+1/M}{M-1}\sum_{m=1}^{M}\left(\hat{\alpha}_{Y.X}^{(m)} - \hat{\alpha}_{Y.X}^{MI}\right)^2}. \quad (6)$$

3. IMPUTING SQUARES

Imputation is straightforward if you just want to regress Y on X. But what if the regression includes the square of X or an interaction between X and another regressor? To answer this question, let's begin by working out how to impute a square. Then we'll extend our approach to interactions, which are a little more complicated.

Suppose that the intended analysis is a normal linear regression of Y on X and X^2. We can assume without losing generality that X has been standardized and that Y has been scaled so that the residual $e_{Y.X}$ is standard normal. For certain calculations it will also be useful to assume that X has a normal distribution. So the regression model is

$$\begin{aligned} &Y = \alpha_{Y.X2} + \beta_{Y.X}X + \beta_{Y.2}X^2 + e_{Y.X2}, \quad \text{where} \\ &X \sim \textit{Normal}(0, 1) \quad \text{and} \quad e_{Y.X2} \sim \textit{Normal}\,(0, 1) \end{aligned} \quad (7)$$

Note that X^2, as the square of a standard normal variable, has a chi-square distribution with 1 degree of freedom. Therefore Y, as the sum of two normal variables and a chi-square variable, is not normal either, but has some skew.

The mean and covariance matrix of (X, X^2, Y) are

$$\mu_{X2Y} = \begin{bmatrix} 0 \\ 1 \\ \beta_{Y.2} \end{bmatrix} \quad \text{and} \quad \Sigma_{X2Y} \begin{bmatrix} 1 & & \\ 0 & 2 & \\ \beta_{Y.1} & 2\beta_{Y.2} & 1+\beta_{Y.1}^2+2\beta_{Y.2}^2 \end{bmatrix}. \quad (8)$$

Notice that X is uncorrelated with its square. This result is well-known for standard normal variables (Rice 1994), but it is also true for any symmetric variable with a finite variance and a mean of zero.

Having established the properties of the complete data, let's suppose that the data are incomplete. In fact, as a thought experiment, let's consider the artificial and extreme situation where every value of X and X^2 is missing and has to be imputed. This thought experiment offers a stringent test for any imputation method. When Y is regressed exclusively on imputed values, any biases depend entirely on the quality of the imputations. In a realistic data set where the regressors have observed as well as imputed values, the biases will be smaller but in the same direction.

If a real data set were missing all its values of X and X^2, we would have no basis for imputation. But let's imagine that, despite missing all the X and X^2 values, we nevertheless know the mean μ_{X2Y} and covariance matrix Σ_{X2Y} of the complete data, and we can use μ_{X2Y} and Σ_{X2Y} to produce imputations. Any biases revealed under these circumstances will be worth worrying about. If an imputation method is biased when it is based on the true parameter values, it must surely be biased when the parameters have to be estimated.

3.1. *Transform, Then Impute*

Our recommendation for imputing a squared term is a method that we call *transform, then impute*. This method ignores the fact that X^2 is derived from X and simply imputes $X^{(m)}$ and $X^{2(m)}$ like any other pair of variables. A straightforward way to do this is to impute $X^{(m)}$ and $X^{2(m)}$ simultaneously, using an equation where both dependent variables are regressed simultaneously on Y:

$$\begin{bmatrix} X^{(m)} \\ X^{2(m)} \end{bmatrix} = \begin{bmatrix} \alpha_{X\cdot Y} \\ \alpha_{2\cdot Y} \end{bmatrix} + \begin{bmatrix} \beta_{X\cdot Y} \\ \beta_{2\cdot Y} \end{bmatrix} Y + \begin{bmatrix} e^{(m)}_{X\cdot Y} \\ e^{(m)}_{2\cdot Y} \end{bmatrix}, \quad \text{with}$$

$$\begin{bmatrix} e^{(m)}_{X\cdot Y} \\ e^{(m)}_{2\cdot Y} \end{bmatrix} \sim \textit{Bivariate normal}(0, \Sigma_{X2\cdot Y}). \tag{9}$$

Equivalently, we could impute the variables successively instead of simultaneously—first imputing $X^{(m)}$ by regression on Y, then imputing $X^{2(m)}$ by regression on $X^{(m)}$ and Y. Or vice versa: first impute $X^{2(m)}$

and then impute $X^{(m)}$. Or the regression-based imputations could be carried out implicitly, under a multivariate normal model for (X, X^2, Y). However the regressions are carried out, the results will be the same. Since the regression parameters that are used for imputation can be derived from μ_{X2Y} and Σ_{X2}, the imputed data will have the same mean and covariance matrix as the complete data. And that ensures that a regression of Y on X in the imputed data will have the same parameters as in the complete data.

Notice that it does not really matter whether (X, X^2, Y) fit a multivariate normal distribution. They do not. But normal imputations will yield the right regression estimates as long as they preserve the mean and covariance matrix of the complete data.

It bears repeating that although the imputed variables fit the same regression as the complete variables, the imputed and complete variables are not the same. The most striking difference is that the imputed value for $X^{2(m)}$ is not the square of the imputed value for $X^{(m)}$. Another difference is that $X^{(m)}$, unlike X, is not normal but skewed. ($X^{(m)}$ is the weighted sum of the normal variable $e^{(m)}_{X\cdot Y}$ and the skewed variable Y.)

In short, under the transform-then-impute method, the imputed values differ in striking ways from the observed values. Nevertheless, regression estimates based on the imputed data will be unbiased, at least when the parameters of the complete data are known in advance.

3.2. *Impute, Then Transform*

As noted earlier, a disquieting result of the transform-then-impute method is that the imputed value of $X^{2(m)}$ is not the square of the imputed value for $X^{(m)}$. It is quite possible, for example, for $X^{2(m)}$ to have a negative value, whereas the true value of X^2, being a square, could never be negative.

In order to avoid such discrepancies between the imputed values, some analysts simply impute $X^{(m)}$ and then square the imputed value to get $(X^{(m)})^2$. We call this approach *impute, then transform*.

Unfortunately, the impute-then-transform method yields biased regression estimates. Although the imputed values look plausible, they will not lead to the right estimates when Y is regressed on $X^{(m)}$ and $(X^{(m)})^2$. More specifically, suppose that $X^{(m)}$ is imputed by normal linear regression on Y:

$$X^{(m)} = \alpha_{X\cdot Y} + \beta_{X\cdot Y} Y + e^{(m)}_{X\cdot Y}, \quad \text{with} \quad e^{(m)}_{X\cdot Y} \sim \mathit{Normal}\big(0, \sigma^2_{X\cdot Y}\big). \quad (10)$$

The imputation parameters $\alpha_{X\cdot Y}$, $\beta_{X\cdot Y}$, $\sigma^2_{X\cdot Y}$ can be derived from μ_{XY} and Σ_{XY}. If we square the imputed values to get $(X^{(m)})^2$, then the mean $\mu^{(m)}_{X2Y}$ of the imputed data $(X^{(m)}, (X^{(m)})^2, Y)$ is the same as the mean μ_{X2Y} of the complete data. But the covariance matrix of the imputed data is

$$\Sigma^{(m)}_{X2Y} = \begin{bmatrix} 1 & & \\ \big(6\beta^5_{Y.1}\beta_{Y.2} + 8\beta^3_{Y.1}\beta^3_{Y.2}\big)/P^3 & 2 + 48\beta^4_{Y.1}\beta^2_{Y.2}\big(\beta^2_{Y.1} + \beta^2_{Y.2}\big)/P^4 & \\ \beta_{Y.1} & \big(6\beta^4_{Y.1}\beta_{Y.2} + 8\beta^2_{Y.2}\beta^3_{Y.2}\big)/P^2 & 1 + \beta^2_{Y.1} + 2\beta^2_{Y.2} \end{bmatrix},$$

$$\text{where} \quad P = 1 + \beta^2_{Y.1} + 2\beta^2_{Y.2}, \quad (11)$$

which is not the same as the covariance Σ_{X2Y} of the complete data. The differences between Σ_{X2Y} and $\Sigma^{(m)}_{X2Y}$ relate to the squared term. The complete-data square X^2 has a variance of 2, but the variance of the squared imputation $(X^{(m)})^2$ is larger than that. The covariance between X and X^2 is zero in the complete data, but the corresponding covariance is not zero in the imputed data. (This is possible because $X^{(m)}$, unlike X, is skewed.) And, most importantly, the covariance between and $(X^{(m)})^2$ and Y in the imputed data is smaller than the corresponding covariance between X^2 and Y in the complete data. This is because $(X^{(m)})^2$ was calculated from $X^{(m)}$ alone, without direct input from Y.

To see what the impute-then-transform method does to regression estimates, we used Mathstatica software to derive, from $\mu^{(m)}_{X2Y}$ and $\Sigma^{(m)}_{X2Y}$, the parameters for a regression of Y on the imputed variables $X^{(m)}$ and $X^{2(m)}$. The resulting expressions are too complicated to display, but we can get a feel for their biases by calculating a few example values. To calculate these examples, we fix the complete data intercept $\alpha_{Y.X2}$ at 0, fix the complete-data residual standard deviation $\sigma_{Y.X2}$ at 1, and set the complete data slopes $\beta_{Y.X}$ and $\beta_{Y.2}$ to various values between 0 and 1.[1] Given these settings of the complete-data regression parameters, we calculate the slope, intercept, and residual variance that

[1] Slopes greater than 1 would be too strong for most social science regressions; given the covariance matrix of (X, X^2, Y), slopes exceeding 1 would imply that X and X^2 explained more than $R^2 = .75$ of the variance in Y. Slopes less than 0 would be realistic, but would not affect the results except by changing their sign.

would be obtained when Y is regressed on the imputed variable $X^{(m)}$ and its square $(X^{(m)})^2$. The results are shown in Table 1.

As Table 1 shows, when values are imputed using the impute-then-transform method, the regression of Y on the imputed data is biased. The most serious bias is in the slope $\beta_{Y2}^{(m)}$ of the imputed square $(X^{(m)})^2$; the value of this slope in the imputed data is typically less than a third of the corresponding slope $\beta_{Y.2}$ in the complete data. The bias in $\beta_{Y2}^{(m)}$ is greatest when the complete-data slope $\beta_{Y.2}$ of X^2 is large or the complete-data slope $\beta_{Y.X}$ of X is small. This makes sense since the slope of X^2 reflects the relationship between X^2 and Y, and that relationship is neglected by the impute-then-transform method. The more important the X^2–Y relationship, the more severe the consequences of ignoring it.

The residual variance in the imputed data is typically larger than it should be, especially when the slope of X^2 in the complete data is very strong. Again, this makes sense because the impute-then-transform method effectively ignores the X^2–Y relationship. The more important this relationship, the more residual variance is left unexplained by ignoring it.

The slope $\beta_{Y.X}^{(m)}$ of the imputed variable $X^{(m)}$ is also biased toward zero, but the bias is much milder than the bias in the slope $\beta_{Y2}^{(m)}$ of $(X^{(m)})^2$. Under all of the settings that we tested, the value of $\beta_{Y.X}^{(m)}$ is never more than 6% less than the corresponding slope $\beta_{Y.X}$ in the complete data, and even that 6% bias occurs only when the slope β_{Y2} of X^2 in the complete data is very strong.

The intercept $\alpha_{Y.X2}^{(m)}$ in the imputed data is larger than the intercept $\alpha_{Y.X2}$ in the complete data. This is not surprising. It is well-known that estimation error in the intercept is negatively correlated with estimation error in the slopes (e.g., see Rice 1994). When the estimated slopes are too small, as they are here, the estimated intercept will be too large.

To sum up, when the complete data parameters are known, the transform-then-impute method gives unbiased regression estimates but also gives imputed values that are inconsistent with one another. By contrast, the impute-then-transform method yields plausible-looking imputed values but badly biased regression estimates.

3.3. *Passive Imputation, and Transform, Impute, then Transform Again*

An attempt to combine the competing imputation methods is a strategy that we call *transform, impute, then transform again*. Under this strategy,

TABLE 1
Regression Estimates Obtained by the Impute-Then-Transform Method

Slope of X in Complete Data ($\beta_{Y.X}$)		Slope of X^2 in Complete Data ($\beta_{Y.2}$)				
		0	0.25	0.5	0.75	1
0	Intercept in imputed data ($\alpha^{(m)}_{Y.X2}$)	0	0.25	0.5	0.75	1
	Slope of X in imputed data ($\beta^{(m)}_{Y.X}$)	0	0	0	0	0
	Slope of X^2 in imputed data ($\beta^{(m)}_{Y.2}$)	0	0	0	0	0
	Residual S.D. in imputed data ($\sigma^{(m)}_{Y.X2}$)	1	1.1	1.2	1.5	1.7
0.25	Intercept in imputed data ($\alpha^{(m)}_{Y.X2}$)	0	0.25	0.49	0.73	0.97
	Slope of X in imputed data ($\beta^{(m)}_{Y.X}$)	0.25	0.25	0.25	0.25	0.25
	Slope of X^2 in imputed data ($\beta^{(m)}_{Y.2}$)	0	0.0046	0.015	0.023	0.027
	Residual S.D. in imputed data ($\sigma^{(m)}_{Y.X2}$)	1	1.1	1.2	1.5	1.7
0.5	Intercept in imputed data ($\alpha^{(m)}_{Y.X2}$)	0	0.22	0.44	0.66	0.9
	Slope of X in imputed data ($\beta^{(m)}_{Y.X}$)	0.5	0.5	0.5	0.5	0.5
	Slope of X^2 in imputed data ($\beta^{(m)}_{Y.2}$)	0	0.027	0.06	0.087	0.1
	Residual S.D. in imputed data ($\sigma^{(m)}_{Y.X2}$)	1	1.1	1.2	1.5	1.7
0.75	Intercept in imputed data ($\alpha^{(m)}_{Y.X2}$)	0	0.19	0.38	0.58	0.8
	Slope of X in imputed data ($\beta^{(m)}_{Y.X}$)	0.75	0.74	0.73	0.73	0.73
	Slope of X^2 in imputed data ($\beta^{(m)}_{Y.2}$)	0	0.062	0.12	0.17	0.2
	Residual S.D. in imputed data ($\sigma^{(m)}_{Y.X2}$)	1	1.1	1.2	1.4	1.7
1	Intercept in imputed data ($\alpha^{(m)}_{Y.X2}$)	0	0.16	0.33	0.52	0.72
	Slope of X in imputed data ($\beta^{(m)}_{Y.X}$)	1	0.98	0.96	0.94	0.94
	Slope of X^2 in imputed data ($\beta^{(m)}_{Y.2}$)	0	0.09	0.17	0.23	0.28
	Residual S.D. in imputed data ($\sigma^{(m)}_{Y.X2}$)	1	1.1	1.2	1.4	1.7

$X^{(m)}$ and $X^{2(m)}$ are imputed as in the transform-then-impute strategy, but then $X^{2(m)}$ is deleted and replaced with $(X^{(m)})^2$.

Unfortunately, this compromise strategy yields exactly the same biased regression estimates as the impute-then-transform method. Once $X^{2(m)}$ is deleted, it is as though it was never imputed in the first place. The imputed variable $X^{(m)}$ is the same as it is under the impute-then-transform strategy, so the square $(X^{(m)})^2$ is the same, and so are the estimates that result from regressing Y on $X^{(m)}$ and $(X^{(m)})^2$.

An iterative variant of this strategy is implemented in software such as IVEware for SAS, *ice* for Stata, and MICE for R and S-Plus. In these packages, squares are used as regressors but are not themselves imputed as dependent variables. After each variable is imputed, its square is recalculated—a process that Royston (2005) calls *passive imputation*. The process is repeated until convergence.

Although it sounds sophisticated, passive imputation yields the same biased estimates as the basic impute-then-transform method. If passive imputation were applied to our simple example, $X^{(m)}$ would be imputed by regression on Y, and then $X^{(m)}$ and its square $(X^{(m)})^2$ would be used as regressors to impute missing values of Y, if there were any. The imputed Y values would be used to impute $X^{(m)}$ again, and so on. After a few rounds of imputation, we would reach the same result that we get from the simple impute-then-transform strategy: we would have $X^{(m)}$ values imputed by linear regression on Y, along with the square $(X^{(m)})^2$ of those imputed values. When Y is regressed on $X^{(m)}$ and $(X^{(m)})^2$, the regression will have the same biases as it does under the impute-then-transform method.

3.4. *Binary Variables: Probit and Logit Regression*

Our discussion has focused on normal variables, because the mathematics of normal variables is tractable and leads to closed-form solutions. But our general point is not limited to normal variables. The variables in a linear regression can have any distribution, and it will still be the case that good regression estimates require an imputation method that preserves the mean and covariance matrix of the complete data.

Our results also extend to binary dependent variables that are modeled by logit and probit regression. A crude way to see this is to remember that in some binary regressions—specifically, in analyses

where the conditional probabilities being modeled are not too close to 0% or 100%—the logit and probit curves are approximately linear, and the results of a probit or logit regression are close to those of a linear regression whose slopes have been multiplied by 2.5 (to get probit slopes) or multiplied by 4 (to get logit slopes) (Wooldridge 2002). An imputation method that works for the linear regression should work for the logistic or probit regression as well.

A more elegant justification relies on the latent relationship between linear and and binary regression (Long 1997; Wooldridge 2002). A binary dependent variable Y^* can be viewed as an indicator that some underlying continuous variable Y has exceeded a threshold value. When Y exceeds the threshold, $Y^* = 1$; otherwise $Y^* = 0$. If Y is normal and appropriately scaled, then the parameters for a linear regression of Y on X are the same as the parameters for a probit regression of Y^* on X. A logit regression, in turn, has approximately the same slopes as the probit regression, multiplied by 1.65 (Long 1997; Wooldridge 2002). An imputation method that works for the underlying linear regression of Y on X, therefore, should work for the derived logit or probit regression of Y^* on X as well. In the next section we look at this relationship concretely, by explicitly deriving the indicator Y^* in a probit regression from the continuous Y^* in a linear regression.

3.5. *An Applied Example*

To illustrate the imputation of squared terms in practice, we now apply the competing imputation methods to a data set analyzed in Allison (2002).[2] The data come from *U.S. News and World Report*, and provide statistics on 1302 U.S. colleges and universities, including the annual cost of room and board X (in thousands of dollars) and the graduation rate Y. The X-Y relationship is curved: In general the graduation rate Y increases with the cost of room and board X; but as the graduation rate approaches 100%, the effect of additional room and board costs is necessarily diminished. We modeled these diminishing returns by regressing Y on X and X^2, expecting a positive slope for X and a negative slope for X^2. Before imputing the data, we centered X around

[2] The data were downloaded from http://www.ssc.upenn.edu/~allison/#Data in September 2007.

its mean of \$4.1 thousand. Mean-centering reduces the collinearity between X and X^2, and ensures that the values of X^2 will be less extreme than they would be in noncentered data. In addition, mean-centering eases interpretation in the sense that the intercept can be interpreted as the average graduation rate, and a one-unit change from the mean value of X is also a one-unit change in the value of X^2.

The choice of imputation method is important in this analysis, since fully 40% of cases are missing the value of X. By contrast, only 7.5% of cases are missing the value of Y (including one case where Y was recorded but had to be deleted because, at $Y = 125\%$, it was impossibly high). Notice that the *U.S. News* data differ in several ways from the idealized data that we used in our earlier calculations. Unlike the idealized data, the *U.S. News* data have missing values on Y as well as X; X is partly rather than completely missing, and neither X nor Y is normal (X is skewed right and Y is skewed left).

We applied all of the imputation methods that we described earlier. The passive imputation method was carried out using IVEware for SAS (Raghunathan, Solenberger, and Van Hoewyk 2002), while the other imputation methods used the MI procedure in SAS software, version 9. For each imputation method, we used $M = 40$ imputations. That may seem like a lot of imputations since the usual rule of thumb is that $M = 3$ to 10 imputations are enough (Rubin 1987). But the usual rule guarantees only that point estimates are reliable, and that standard errors and p values are correct on average. If we also want the standard errors and p values to be *reliable*—in the sense that they would not change much if the data were imputed again—then the number of imputations should be similar to the percentage of cases that are incomplete, which in this example is about 40 (Bodner 2008). To make the estimates even more reliable, before analyzing the imputed data we deleted cases with imputed Y values; such cases add only noise to the estimates (Little 1992; von Hippel 2007).

The linear regression estimates produced by the different imputation methods are displayed in Table 2(a). The biases are as we expected. Compared with the transform-then-impute method, which in our idealized calculations gave consistent estimates, all of the other methods—impute-then-transform, transform-then-impute-then-transform again, and passive imputation—give estimates that are severely biased and practically indistinguishable from one another. As our calculations

TABLE 2(a)
Linear Regression Predicting the Graduation Rate

	Transform, Then Impute	Impute, Then Transform	Transform, Impute, Then Transform Again	Passive Imputation
Intercept (average of Y)	61.0***	60.7***	60.6***	60.6***
	(0.7)	(0.6)	(0.6)	(0.6)
X (room and board, in thousands)	8.7***	8.2***	8.3***	8.3***
	(0.5)	(0.5)	(0.5)	(0.5)
X^2	−0.7*	−0.4	−0.4	−0.4
	(0.3)	(0.3)	(0.3)	(0.3)
Residual variance	264.8	266.1	266.0	264.5

TABLE 2(b)
Probit Regression Predicting Whether the Graduation Rate Exceeds 60%

	Transform, Then Impute	Impute, Then Transform	Transform, Impute, Then Transform Again	Passive Imputation
Intercept	0.03	−0.01	0.00	−0.01
	(0.05)	(0.05)	(0.05)	(0.05)
X (room and board, in thousands)	0.59***	0.54***	0.55***	0.55***
	(0.05)	(0.05)	(0.05)	(0.05)
X^2	−0.07*	−0.04	−0.04	−0.04
	(0.03)	(0.03)	(0.03)	(0.03)

*$p < .05$, **$p < .01$, ***$p < .001$. Standard errors appear in parentheses.

suggested, the alternatives to the transform-then-impute method are simply different ways to get the same biased results.

The biases of the alternative methods are in the directions predicted by our mathematical calculations. Compared with the transform-then-impute method, the other methods give X^2 a slope that is strongly biased toward zero. Specifically, under the transform-then-impute method, the slope of X^2 is -0.7 and statistically significant, but under the other methods the slope of X^2 is nonsignificant and only -0.4. The slope of X is also biased toward zero, but only a little bit; under the transform-then-impute method the slope of X is 8.7, while under the other methods it is 8.2 or 8.3.

Bias in the intercept and in the residual variance is negligible. Bias in the intercept is negatively correlated with bias in the slopes, and since the two slopes have opposite biases—one positive, the other negative—the net effect on the intercept is close to nil. Bias in the residual variance is also trivial; since the squared term did not explain much of the variance to begin with, little excess variance is left unexplained if we impute the squared term incorrectly.

It is worth taking a moment to interpret the results. According to the best imputation method—transform-then-impute—the graduation rate for a college with average room-and-board expenses is 61%. Increasing room and board to \$1000 above the mean predicts an 8% rise in the graduation rate ($8.7 \times 1 - 0.7 \times 1^2$), but an additional \$1000 increase predicts a smaller rise of just 6.6% ($8.7 \times (2 - 1) - 0.7 \times (2^2 - 1^2)$). The diminishing effects of additional expenditures are modeled by the squared term. Under inferior imputation methods, the squared term has a smaller coefficient and is nonsignificant, which might prevent researchers from realizing that room-and-board expenses have diminishing returns. And if researchers do not report that expenses have diminishing returns, then parents reading research summaries might spend more than they really need to.

We also carried out a probit regression in which X and X^2 were used to predict a dummy variable Y^* that turns from 0 to 1 when the graduation rate Y exceeds 60% (the median graduation rate). Given the relationship between a probit regression of Y^* and a linear regression of Y (see Section 3.4), it is not surprising that the probit results, in Table 2(b), display the same pattern as the linear results in Table 2(a). Under the biased imputation methods, the probit slopes of X and X^2, especially X^2, are smaller than they are under the transform-then-impute method. A logistic regression would display the same biases, since logistic slopes are approximately equal to probit slopes multiplied by 1.65 (Long 1997).

4. IMPUTING INTERACTIONS

Let's look now at the problem of imputing interactions. The mathematics for interactions are similar to those for squares. After all, a square is just a variable interacting with itself.

Suppose that the target analysis is a linear regression where the inputs X_1 and X_2 are bivariate standard normal variables with a correlation of ρ, and the outcome is

$$Y = \alpha_{Y\cdot 12I} + \beta_{Y\cdot 1} X_1 + \beta_{Y\cdot 2} X_2 + \beta_{Y\cdot I} I_{12} + e_{Y\cdot 12I}, \tag{12}$$

where $I_{12} = X_1 X_2$ is the interaction term and $e_{Y\cdot 12I} \sim \mathrm{N}(0, \sigma^2_{Y\cdot 12I})$ is normal error. To keep things simple, we assume that all the parameters have a value of 1, i.e. $(\alpha_{Y\cdot 12I}, \beta_{Y\cdot 1}, \beta_{Y\cdot 2}, \beta_{Y\cdot I}, \sigma^2_{Y\cdot 12I}) = (1, 1, 1, 1, 1)$. The assumption of normal regressors and parameters of 1 are not really necessary, but they make the calculations simpler and more transparent.

The mean and covariance matrix of (X_1, X_2, I_{12}, Y) are, according to Mathstatica,

$$\mu_{12IY} = \begin{bmatrix} 0 \\ 0 \\ \rho \\ 1+\rho \end{bmatrix} \quad \text{and}$$

$$\Sigma_{12IY} = \begin{bmatrix} 1 & & & \\ \rho & 1 & & \\ 0 & 0 & 1+\rho^2 & \\ 1+\rho & 1+\rho & 1+\rho^2 & 4+2\rho+\rho^2 \end{bmatrix}. \tag{13}$$

Notice that the interaction I_{12} has no correlation with X_1 or X_2. This is a bivariate analog to our earlier and more familiar finding that a standard normal variable is uncorrelated with its square.

Now suppose that all values of X_2 are deleted; then the interaction I_{12} must be deleted as well, since in real data I_{12} could not be calculated without X_2. Again, we imagine that, despite missing all values of X_2 and I_{12}, we nevertheless know μ_{12IY} and Σ_{12IY} and we can use μ_{12IY} and Σ_{12IY} to replace the missing variables with imputed variables $X_2^{(m)}$ and $I_{12}^{(m)}$. This is a rigorous thought experiment: If an imputation method fails even when we have perfect knowledge of the complete-data parameters, then that method must be fundamentally unsound.

4.1. *Transform, Then Impute*

As noted earlier, normal linear regression, whether implicit or explicit in the imputation model, is the most widely implemented method for imputing continuous variables. Using a normal linear regression model

with two dependent variables, we can impute $X_2^{(m)}$ and $I_{12}^{(m)}$ by regression on the complete variables X_1 and Y:

$$\begin{bmatrix} X_2^{(m)} \\ I_{12}^{(m)} \end{bmatrix} = \begin{bmatrix} \alpha_{2\cdot 1Y} \\ \alpha_{I\cdot 1Y} \end{bmatrix} + \begin{bmatrix} \beta_{2\cdot 1} & \beta_{2\cdot Y} \\ \beta_{I\cdot 1} & \beta_{I\cdot 1} \end{bmatrix} \begin{bmatrix} X_1 \\ Y \end{bmatrix} + \begin{bmatrix} e_{2\cdot 1Y}^{(m)} \\ e_{I\cdot 1Y}^{(m)} \end{bmatrix}, \quad \text{where}$$

$$\begin{bmatrix} e_{2\cdot 1Y}^{(m)} \\ e_{I\cdot 1Y}^{(m)} \end{bmatrix} \sim \text{Bivariate normal } (0, \Sigma_{2I\cdot 1Y}). \tag{14}$$

Alternatively, we can impute the variables one at a time, first imputing $X_2^{(m)}$ by regression on X_1 and Y, and then imputing $I_{12}^{(m)}$ by regression on X_1, Y, and $X_2^{(m)}$. Or vice versa: first impute $I_{12}^{(m)}$ and then impute $X_2^{(m)}$. Or we can regress impheitly under multivariate normal model for (X_1, X_2, I_{12}, Y). As long as the regression parameters used for imputation are derived from μ_{12IY} and Σ_{12IY}, the result will be the same: the mix of complete and imputed variables (X_1, $X_2^{(m)}$, $I_{12}^{(m)}$, Y) will have the same mean μ_{12IY} and covariance matrix Σ_{12IY} as the complete data. So when Y is regressed on X_1, $X_2^{(m)}$, and $I_{12}^{(m)}$, the regression parameters will be the same as if all the variables were complete.

In short, in an idealized situation where the mean and covariance matrix of the complete data are known, the transform-then-impute method yields unbiased regression estimates.

4.2. *Impute, Then Transform, and Its Variants: Passive Imputation and Transform, Impute, Then Transform Again*

It bears repeating that the imputed variables $X_2^{(m)}$ and $I_{12}^{(m)}$ are not the same as the complete variables X_2 and I_{12}. Unlike the complete variable X_2, the imputed variable $X_2^{(m)}$ is not normal. More disturbingly, while the complete data interaction I_{12} is the product of the complete variables X_1 and X_2, the imputed interaction $I_{12}^{(m)}$ is not the product of the observed X_1 and the imputed $X_2^{(m)}$. It is quite possible, for example, for $I_{12}^{(m)}$ to have a negative value in a case where both X_1 and $X_2^{(m)}$ are positive.

Some researchers find this last anomaly so troubling that they replace the imputed interaction $I_{12}^{(m)}$ with the product $X_1 X_2^{(m)}$. We call this procedure *transform, impute, then transform again*. Alternatively, a researcher may calculate the product $X_1 X_2^{(m)}$ without having previously

imputed $I_{12}^{(m)}$. This approach is called *transform then impute*. An iterative version of the transform-then-impute method is called *passive imputation*, implemented in Stata's *ice* command and, by a different name, in IVEware for SAS and MICE for R.

All these techniques are variants of the basic impute-then-transform method, and all of them lead to the same biased estimates. The problem is that the product $X_1 X_2^{(m)}$, although its individual values look plausible, does not vary in the right way with the other variables, especially with Y. According to Mathstatica, the imputed-then-transformed data $(X_1,\ X_2^{(m)},\ X_1 X_2^{(m)},\ Y)$ has the same mean μ_{12IY} as the complete data, but the covariance matrix of the imputed-then-transformed data is

$$\Sigma_{12IY}^{(m)} = \begin{bmatrix} 1 & & & \\ \rho & 1 & & \\ \frac{2}{3}\rho(1-\rho^2) & \frac{2}{9}(\rho^4+4\rho^2+1)(1-\rho^2) & \frac{1}{9}(6\rho^6-10\rho^4+11\rho^2+11) & \\ 1+\rho & 1+\rho & \frac{1}{3}(2-2\rho(\rho^2-2\rho-1)) & 4+2\rho+\rho^2 \end{bmatrix}, \tag{15}$$

which is clearly not the same as the covariance matrix Σ_{12IY} of the complete data. All of the variances and covariances involving the recalculated interaction $X_1 X_2^{(m)}$ differ from those involving the original complete-data interaction I_{12}. Whereas the complete-data interaction I_{12} had zero covariance with X_1 and X_2, the recalculated interaction $X_1 X_2^{(m)}$ has a small covariance with both X_1 and $X_2^{(m)}$. (This is possible because $X_2^{(m)}$, unlike X_2, has skew.) Most importantly, the recalculated interaction $X_1 X_2^{(m)}$ has a much smaller covariance with Y than did the original interaction I_{12}. This is because the product $X_1 X_2^{(m)}$ was calculated from the observed and imputed X values alone, without additional input from Y.

To see what the impute-then-transform method does to regression estimates, we can use Mathstatica software to derive, from $\mu_{12\mathrm{IY}}$ and $\Sigma_{12IY}^{(m)}$, the parameters for a regression of Y on the observed variable X_1, the imputed variable $X_2^{(m)}$, and the composite interaction $X_1 X_2^{(m)}$. Figure 1 plots the results as a function of the correlation ρ between X_1 and X_2. A horizontal reference line shows the value 1, which in the

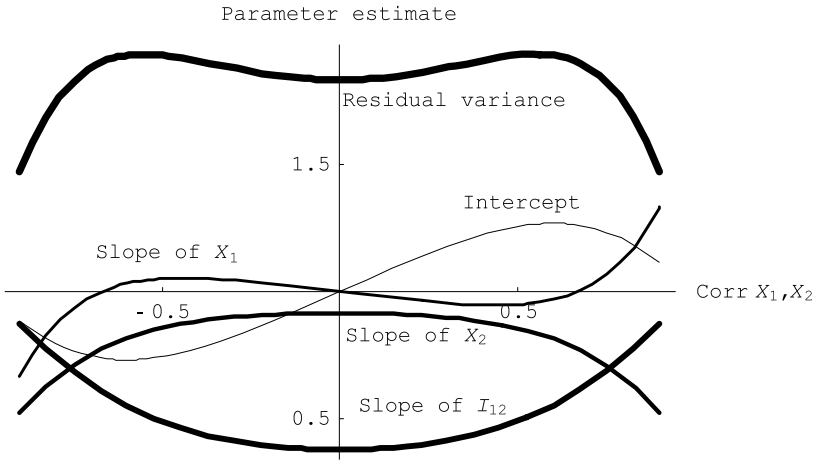

FIGURE 1. Values of the regression estimates when interactions are imputed using the impute-then-transform method.

complete data is the correct value for all of the parameters. If the estimates obtained by the impute-then-transform method were unbiased, they would all be 1. But they are not.

Instead, as an estimate of the complete-data regression, a regression using the imputed-then-transformed data has several biases. The residual variance is too large, and the slopes of X_2 and I_{12} are biased toward zero, with the size of the bias depending on ρ. The intercept and the slope of X_1 are also biased, with both the size and direction of the bias depending on ρ. The exact shape of these biases would have been hard to guess, but the broadest patterns are not surprising. Under most conditions the slope of the interaction is much more biased than the other slopes; this makes sense since the interaction is the term that was neglected in imputation. The residual variance is too large under all circumstances; this too makes sense. When the regressors are not imputed well, they leave a lot of the variation in Y unexplained.

Would the impute-then-transform method yield better results if the imputation model for X_2 were better specified? In equation (14), $X_2^{(m)}$ was imputed by linear regression on X_1 and Y. But clearly $X_2^{(m)}$ does not depend on X_1 and Y in a purely linear way. Instead, since the regression equation (12) for Y contains an interaction between X_1 and X_2, we know that the $X_2 - Y$ relationship varies with X_1. It follows that

the imputation equation for X_2 should include an interaction between X_1 and Y.

Suppose that we add the interaction $I_{1Y} = X_1 Y$ to the regression equation that is used to impute missing values of X_1

$$X_2^{(m)^*} = \alpha^*_{21YI} + \beta^*_{2\cdot 1} X_1 + \beta^*_{2\cdot Y} Y + \beta^*_{2\cdot I} I_{1Y} + e^{(m)}_{2\cdot 1YI}, \quad \text{with}$$
$$e^{(m)}_{2\cdot 1Y} \sim \mathit{Normal}\big(0, \sigma^2_{2\cdot 1YI}\big), \tag{16}$$

where the regression parameters are derived from the mean and covariance matrix of $(X_1,\ Y, I_{1Y}, X_2)$:

$$\mu_{1YI2} = \begin{bmatrix} 0 \\ 1+\rho \\ 1+\rho \\ 0 \end{bmatrix} \quad \text{and}$$

$$\Sigma_{1YI2} = \begin{bmatrix} 1 & & & \\ 1+\rho & 4+2\rho+\rho^2 & & \\ 1+3\rho & 3+6\rho+3\rho^2 & 8+10\rho+13\rho^2 & \\ \rho & 1+\rho & 1+\rho+2\rho^2 & 1 \end{bmatrix}. \tag{17}$$

And suppose that, after imputing $X_2^{(m)*}$, we calculate interactions $X_1 X_2^{(m)*}$.

Do these imputed interactions come close to replicating the covariance structure of the complete data? Unfortunately, they do not. Although $(X_1,\ X_2^{(m)^*},\ X_1 X_2^{(m)^*}, Y)$ has the same mean μ_{12IY} as the complete data, the covariance matrix of the imputed data is not the covariance matrix Σ_{12IY} of the complete data. Instead, the covariance matrix of $(X_1,\ X_2^{(m)^*},\ X_1 X_2^{(m)^*}, Y)$ is

$$\Sigma^{(m)^*}_{12IY} = \begin{bmatrix} 1 & & & \\ \rho & 1 & & \\ P_{31}/P & P_{32}/P^2 & P_{33}/P^2 & \\ 1+\rho & 1+\rho & 1+\rho^2 & 4+2\rho+\rho^2 \end{bmatrix}, \tag{18}$$

where the Ps represent polynomials in ρ:

$$
\begin{aligned}
P &= 17 + 4\rho + \rho^2 \\
P_{31} &= 2 + 8\rho - 2\rho^2 - 8\rho^5 \\
P_{32} &= 106 + 16\rho + 116\rho^2 + 36\rho^3 + 106\rho^4 + 44\rho^5 \\
&\quad - 200\rho^6 - 32\rho^7 - 96\rho^8 - 64\rho^9 - 32\rho^{10} \\
P_{33} &= 403 + 112\rho + 765\rho^2 - 144\rho^3 + 126\rho^4 + 304\rho^5 \\
&\quad - 204\rho^6 + 96\rho^7 + 96\rho^8 + 32\rho^9 + 96\rho^{10}.
\end{aligned}
$$

If we derive regression estimates from this mean and covariance matrix, the result will still be biased when compared with the complete-data regression. In fact, when plotted, the biases in these estimates are almost indistinguishable from the biases of the transform-then-impute method plotted in Figure 1. It appears that, although adding an extra interaction when imputing $X_2^{(m)^*}$ may improve the marginal distribution of the imputed variable, this refinement does not substantially improve the accuracy of the regression estimates that are derived from the imputed data.

4.3. *Stratify, Then Impute*

An alternative imputation strategy—which we call *stratify, then impute*—presents itself when one of the interacting variables is discrete and has no missing values. For example, X_1 may be a complete dummy variable that takes values of 1 and 0. In this situation, we can divide the data into two strata, a stratum with $X_1 = 0$ and a stratum with $X_1 = 1$. Within each stratum, we impute missing values of X_2 by linear regression on Y, and if Y has missing values, we can impute them by linear regression on X_2. This is an elegant solution since it allows for the interaction without having to incorporate it into the imputation model. Within each stratum no interaction is required, and simple linear regression can produce imputations with the same mean and covariance matrix as the complete data. The regression of Y on the imputed Xs will be the same, on average, as if Y were regressed on the complete Xs.

The stratify-then-impute method is an ideal solution, but unfortunately it is not always available. Defining strata is straightforward when the stratifying variable X_1 is complete and takes just a few discrete

values. But strata are harder to define when X_1 is continuous or has missing values itself.

4.4. *An Applied Example*

To illustrate the imputation of interactions in practice, we again apply the competing imputation methods to Allison's (2002) *U.S. News* data on colleges and universities. Again we predict each college's graduation rate Y, which is missing for 7.5% of cases, but now our predictors are the college's average combined SAT scores X_2, which is missing for 40% of cases, and a complete variable X_1 indicating whether the college is public ($X_1 = 0$) or private ($X_1 = 1$). Allison's (2002) own analysis of these data includes other regressors, but to illustrate the properties of competing imputation methods it suffices to regress Y on X_1, X_2, and $I_{12} = X_1 X_2$. Despite the omission of supplementary variables, our results are very similar to Allison's, but more comprehensive because more methods are tested.

To ease interpretation, we rescaled the combined SAT score by subtracting the public-college mean of 942 and dividing by 100. Under this scaling, the intercept represents the average graduation rate at a public college, and a single unit on the rescaled score represents 100 points on the SAT. Note that the *U.S. News* data differ in several ways from the idealized data that we used in our earlier calculations. Unlike the idealized data, the *U.S. News* data have missing values on Y as well as X_2; X_2 is partly rather than completely missing; and the complete variable X_1 is a dummy rather than a normal variable.

We applied all of the imputation methods that we described earlier, using $M = 40$ imputations (Bodner 2008). The passive imputation method was carried out using IVEware for SAS (Raghunathan et al. 2002), while the other imputation methods used the MI procedure in SAS software, version 9. Table 3(a) gives the regression estimates that are obtained after the six different imputation strategies are applied.

Although we do not know the true parameter values, we expect to get our best estimates from the stratify-then-impute method and the transform-then-impute method. Reassuringly, these two methods give very similar results. By contrast, the other methods—impute-then-transform (with and without an extra interaction to impute X_2), transform-impute-then-transform-again, and passive imputation—return estimates that are very similar to each other, and quite biased.

TABLE 3(a)
Linear Regression Predicting the Graduation Rate

	Stratify, Then Impute	Transform, Then Impute	Impute, Then Transform	Impute, Then Transform, with an Extra Interaction to Impute X_2	Transform, Impute, Then Transform Again	Passive Imputation
Intercept	50.5***	50.5***	50.4***	50.6***	50.5***	50.5***
	(0.7)	(0.8)	(0.8)	(0.8)	(0.8)	(0.8)
X_1. Private	12.9***	12.9***	12.8***	12.9***	12.8***	12.9***
	(0.9)	(0.9)	(0.9)	(0.9)	(0.9)	(0.9)
X_2. Combined SAT (in hundreds)	9.7***	10.0***	8.5***	8.6***	8.5***	8.5***
	(0.7)	(0.9)	(0.7)	(0.7)	(0.7)	(0.7)
I_{12}. Private * SAT	−2.0*	−2.3*	−0.4	−0.5	−0.4	−0.4
	(0.9)	(1.0)	(0.8)	(0.8)	(0.9)	(0.8)
Residual variance	194.3	194.7	197.4	196.0	196.2	198.2

TABLE 3(b)
Probit Regression Predicting Whether the Graduation Rate Exceeds 60%

	Stratify, Then Impute	Transform, Then Impute	Impute, Then Transform	Impute, Then Transform, with an Extra Interaction to Impute X_2	Transform, Impute, Then Transform Again	Passive Imputation
Intercept	−0.98***	−0.95***	−0.92***	−0.88***	−0.91***	−0.90***
	(0.10)	(0.09)	(0.09)	(0.09)	(0.10)	(0.09)
X_1. Private	1.18***	1.16***	1.12***	1.08***	1.11***	1.09***
	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)
X_2. Combined SAT (in hundreds)	0.96***	0.93***	0.76***	0.65***	0.75***	0.74***
	(0.12)	(0.12)	(0.10)	(0.13)	(0.11)	(0.11)
I_{12}. Private * SAT	−0.38**	−0.34*	−0.13	0.02	−0.11	−0.12
	(0.14)	(0.14)	(0.12)	(0.14)	(0.12)	(0.12)

*$p < .05$, **$p < .01$, ***$p < .001$. Standard errors appear in parentheses.

Compared with the sound methods, the biased methods give too large a residual variance, too small a coefficient for X_2, and much too small a coefficient for the interaction. All these biases are consistent with those displayed in Figure 1. The intercept and the slope of X_1 appear to be approximately unbiased; again, this is consistent with Figure 1, where the bias for these estimates is small if the correlation between X_1 and

X_2 is small. (In these data, the correlation between X_1 and X_2 is just 0.14.)

It is worth taking a minute to interpret the results. According to either of the sound imputation methods—transform-then-impute or stratify-then-impute—the intercept is 50.5, implying that public colleges have an average graduation rate of 50.5%. The slope for X_1, the private college indicator, is 12.9, implying that a private college with the same combined SAT score as an average public college would be expected to have a graduation rate that is 12.9% higher. The slope for X_2, the combined SAT score, is about 10 and the interaction between X_1 and X_2 is about -2. So in public colleges a 100-point boost in the combined SAT score predicts a 10% increase in the graduation rate, but in private colleges the increase in the graduation rate is just 8% (10 $-$ 2). In other words, the graduation gap between public and private colleges is smaller at institutions where the students have high SAT scores. This diminishing return to private education is modeled by the interaction, which under the best imputation methods is decent-sized and significant. Under the inferior methods, the same interaction is nonsignificant and trivial in size. This bias might lead researchers to overlook the diminishing returns of private education for high-scoring students. Parents reading a research summary might mistakenly believe that their high-scoring children would benefit from private schooling much more than is actually the case.

We also carried out a probit regression in which the interacting variables were used to predict a dummy variable Y^* that turns from 0 to 1 when the graduation rate Y exceeds the median value of 60%. Given the relationship between a probit regression of Y^* and a linear regression of Y (see Section 3.4), it is not surprising that the probit results in Table 3(b) display the same pattern as the linear results in Table 3(a). Under the biased imputation methods, the probit slopes of both Xs and especially of their interaction are smaller than they are under the sound methods. A logistic regression would display the same biases as well, since logistic slopes are approximately equal to probit slopes multiplied by 1.65 (Long 1997).

5. CONCLUSION

When using a parametric model to prepare data for linear, probit, or logistic regression, there is one widely applicable and reasonably accurate

way to impute transformed variables such as squares and interactions. That method is to *transform, then impute*—that is, to transform the variable in the incomplete data and then to impute the transformation like any other variable. The transform-then-impute method yields good regression estimates, even though the imputed values are often inconsistent with one another. (For example, the imputed value for the square of X is not the square of the imputed value for X.) It is tempting to try and "fix" the inconsistencies in the imputed values, but methods that do so lead to biased regression estimates. Such biased methods include the *passive* imputation strategy implemented by the popular *ice* command for Stata.

Our calculations and analyses have been limited to two types of transformed variable: the square and the interaction. These two transformations cover the vast majority of analyses where a raw variable and its transformation are used in the same regression. However, there is no reason to think that the correct approach for imputing other transformations—such as cubed terms or splines—would be any different. The general principle—that an imputation method should preserve the covariances among the raw and transformed variables—is not limited to a particular kind of transformation.

It is a little disappointing to find that none of the methods we looked at can give us everything we want. If we want consistent regression estimates, we have to accept imputed values that do not make a lot of sense—for example, imputed squares that are negative. On the other hand, if we want imputed values that make sense, we have to accept biased regression estimates. We have to wonder if we have missed something. Isn't there a method that gives good regression estimates and sensible imputed values as well? Our analyses demonstrated one such method—stratify, then impute—but unfortunately that method has limited application. It does not apply to squared terms, and it can only be used for interactions if one of the interacting variables is both complete and discrete.

The general idea of imputing within strata seems promising, though, and stratification is actually the basis for nonparametric methods that impute by resampling. Such resampling methods include hot-deck imputation and the approximate Bayesian bootstrap (Little and Rubin 2002), which fill in missing values by resampling observed values from strata of similar cases. Although resampling methods have a number of practical difficulties, future research may find that resampling

provides a sound general approach to imputing transformations. Perhaps resampling can produce accurate regression estimates and sensible imputed values at the same time.

REFERENCES

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

———. 2005. "Imputation of Categorical Variables with PROC MI," Presented at the 30th Meeting of SAS Users Group International, April 10–13, Philadephia, PA. Retrieved May 29, 2007 (www2.sas.com/proceedings/sugi30/113-30.pdf).

Bodner, T. E. 2008. "What Improves with Increased Missing Data Imputations?" *Structural Equation Modeling* 15(4):651–75.

Johnson, Richard A., and Dean W. Wichern. 2003. *Applied Multivariate Statistical Analysis*. 5th ed. New York: Prentice Hall.

Little, Roderick J. A. 1992. "Regression with Missing X's: A Review." *Journal of the American Statistical Association* 87(420):1227–37.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*, 2nd ed. New York: Wiley.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

Raghunathan, T. E., Peter Solenberger, and John Van Hoewyk. 2002. "IVEware: Imputation and Variance Estimation Software." User manual. Survey Methodology Program, University of Michigan. Retrieved June 19, 2008 (http://www.isr.umich.edu/src/smp/ive/).

Rice, John A. 1994. *Mathematical Statistics and Data Analysis*. Belmont, CA: Duxbury.

Rose, Colin, and Murray Smith. 2002. *Mathematical Statistics with Mathematica*. New York: Springer.

Royston, Patrick. 2005. "Multiple Imputation of Missing Values: Update." *Stata Journal* 5(2):1–14.

Rubin, Donald B. 1987. *Multiple Imputation for Survey Nonresponse*. New York: Wiley.

von Hippel, Paul T. 2007. "Regression with Missing Ys: An Improved Strategy for Analyzing Multiply-Imputed Data." Pp. 83–117 in *Sociological Methodology*, vol. 37, edited by Yu Xie. Boston, MA: Blackwell Publishing.

———. 2009. "Imputing Skewed Variables." Ohio State University. Unpublished manuscript.

Wolfram Software. 2005. *Mathematica* Version 5.2.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross-Section and Panel Data*. Cambridge, MA: MIT Press.