Published in final edited form as:

Multivariate Behav Res. 2017 ; 52(5): 551–561. doi:10.1080/00273171.2017.1333404.

Two-Part Predictors in Regression Models

John J. Dziak^a and Kimberly L. Henry^b

^aThe Methodology Center, The Pennsylvania State University

^bDepartment of Psychology & Colorado School of Public Health, Colorado State University

Abstract

Researchers often build regression models to relate a response to a set of predictor variables. In some cases there are predictors that apply to some participants, or to some measurement occasions, but not others. For example, a romantic partner's substance use may be a key predictor of one's own substance use. However, not all participants have a partner, and in a longitudinal study, participants may have a partner during only some occasions. This could be viewed as missing data, but of a very distinctive type: the values are not just unknown but also undefined. In this paper we present a simple method to accommodate this situation, along with a motivating example, the algebraic justification, a simulation study, and examples of how to carry out the technique.

Keywords

two-part predictors; two-part models; nested factors; relationship status; substance use

In many situations in which regression models are built to predict an outcome, there are important covariates that apply to only a subset of cases. For example, a romantic partner's illegal substance use may be a key predictor of one's own illegal substance use. However, in a cross-sectional study some participants may not have a partner; and in a longitudinal study, participants may have a partner only during some occasions. In these cases, the missing value on partner substance use is structurally, rather than randomly, missing. We refer to variables such as these as "conditionally relevant," (i.e., it is reasonable to include it for some participants/measurements but not for others). Many research questions that are salient to social and behavioral scientists involve conditionally relevant variables. For example, suppose marital satisfaction is of interest as a predictor variable. It is not reasonable to assume that lacking a spouse is the same as having the most unsatisfactory spouse possible, nor the most satisfactory, nor even an average one. As another example, what is the role of communication between mother and child about substance use among children who do not have a mother?

Clearly, assigning an arbitrary value a priori for the conditionally relevant covariate would be unreasonable; for example, one cannot assume that not having a partner is the same as

having a partner who does not use drugs. Another option might be to treat partner's substance use for singles as missing data and use multiple imputation (Rubin, 1987), essentially imputing different scenarios for what the partner's substance use might have been if they had been in a relationship. However, this is also somewhat arbitrary and ignores the fact that being single is fundamentally different than being in a relationship. Even if it were considered meaningful to impute (singly or multiply) partner's substance use, it would still be difficult. Any imputation model for partner's use would by necessity be based only on people in a relationship, and they may differ in systematic ways from single people. That is, the substance use of the partner the participant would have had, even if the construct is considered to be well-defined at all, is likely missing not at random, and there would be no way to diagnose this.

Another option would be to separate the data into groups and perform analyses separately, say for people in a relationship and single people. This would avoid the need to impute values for undefined constructs such as single people's exposure to partners' substance use. It would also allow all of the other covariates in the model to have different estimates for the two groups. However, there are some situations in which this approach becomes burdensome. For instance, consider a longitudinal study of emerging adults over a period of years. A participant might move in and out of a relationship several times. Furthermore, even in a cross-sectional study, dividing the sample into subgroups and doing all estimation and testing separately means that sample sizes will be much smaller and statistical power will be reduced. Such a strategy would likely be justified if the conditionally relevant covariate were the predictor of most interest in the study. Indeed, if the study hypothesis were specifically focused on partner influence (e.g., in Manne, Kashy, Weinberg, Boscarino, and Bowen, 2012), participants not in a relationship probably would not belong in the sample at all. However, if partner's influence were one of many covariates (e.g., perhaps being used to calculate a propensity score) or if more than one covariate is conditionally relevant, then the increase in sampling variance caused by having to analyze many different subgroups could outweigh the concomitant decrease in bias (a classic "bias-variance tradeoff"; see, e.g., Hastie, Tibshirani & Friedman, 2001).

Finally, for some models and in some software packages (perhaps especially those for structural equation modeling), it is possible to constrain some parameters to be equal across subsets, while allowing others to be unequal. Where feasible, this might improve the performance of a subsetting approach, because appropriate constraints can reduce sampling variance without causing a sizable increase in bias. Parke and co-authors (2004) and Widaman, Grimm, Early, Robins, and Conger (2013) utilized this approach in the context of models including an acculturation measure that is relevant for immigrant but not non-immigrant participants. However, in many commonly used regression software packages, this level of flexibility is not available.

In many cases, a simple alternative can help avoid these dilemmas. Our perspective is that covariates like these should not be treated as a form of missing data at all, but rather as a two-part variable that can be represented as a pair of interrelated covariates. The first variable in the pair is a dummy-coded indicator that denotes whether the covariate value is relevant (e.g., person is in a relationship or not) and the second denotes the covariate value

among people for whom the covariate value is relevant. This approach was considered by Cohen, Cohen, West, and Aiken (2003), who observed that “missing” could sometimes be appropriately considered a level on a dummy-coded categorical variable (p. 436), or that a dummy-coded variable could perhaps be used in order to treat “missingness as information” (p. 444). However, it was sharply criticized by Greenland and Finkle (1995) and Jones (1996). This paper will outline when an indicator-variable approach is appropriate and when it is inappropriate.

For simplicity, suppose first that the missing variable is dichotomous, although it is not limited to this case and we will provide examples of a continuous missing variable after outlining the simpler dichotomous case. As standard regression texts (e.g., Myers & Well, 2003; Cohen et al., 2003) explain, a k -level categorical variable can be expressed as a set of $k-1$ unidimensional indicator variables using dummy coding, and the resulting model can be fit very easily with simple regression models and standard software. Suppose race/ethnicity is considered as a covariate, and the sample participants can all be classified as either non-Hispanic Black, non-Hispanic White, or Hispanic. Let $B=1$ for non-Hispanic Blacks and 0 for others, $H=1$ for Hispanic participants and 0 for others, and $W=1$ for non-Hispanic White participants and 0 for others. Because of collinearity (the constraint that $B+H+W=1$), one of the three codes must be omitted as redundant, and this determines the reference group. For example, if White is the largest racial/ethnic group in the sample of interest, it might be reasonable to omit W and use only B and H . In this case, White individuals become the reference class and the k -level categorical variable is re-expressed as $k-1$ binary variables.

Returning to the partner substance use example, suppose now that instead of White, Black and Hispanic, we have a different categorical variable: single, in a relationship with a partner who uses drugs, and in a relationship with a partner who does not use drugs. Suppose that we decide to let the reference group be individuals who are in a relationship with a non-substance using partner. One could use the same setup as in the example above. Specifically, suppose that Relationship (REL) is coded as 1 for individuals in a relationship with a partner and 0 for other individuals. For individuals with $REL=1$, suppose Partner Use ($PUSE$) equals 1 if the partner uses drugs and $PUSE=0$ otherwise. The two variables REL and $PUSE$ together tell a complete story about the three-level categorical variable, but they tell the story in an awkward way because $PUSE$ is undefined for participants who are not in a relationship and therefore seems to be missing in their rows of the dataset. Therefore, one can define $PUSE^*$ as 1 for individuals in a relationship with a substance-using partner and 0 for all others. REL and $PUSE^*$ together carry the same information as REL and $PUSE$ together, but now do so more conveniently, because both variables are now well-defined for all participants.

Although everyone with $REL=0$ necessarily has $PUSE^*=0$, this does *not* imply that all singles would be with a partner who does not use drugs if they were in a relationship. This is because $PUSE^*$ is used only when REL is also included as a predictor in the model. Thus, partner drug use is no more being imputed as zero for unpartnered people, than Black versus White identification was being imputed as zero for Hispanics in the demographic example.

The use of dummy coding to represent whether a variable is observed has been controversial. It has previously been shown (Greenland & Finkle, 1995; Jones, 1996) that the technique is not appropriate for the usual situation of missing data—a situation in which a value of a covariate is missing, but the response has a possible value that would be meaningfully interpretable if known. For example, if all of the participants in a study had a partner, but some participants were unwilling to disclose whether their partners used drugs, then the missing covariate values are meaningful and the dummy-coding approach is not appropriate.

Schafer and Graham (2002) describe intuitively why an indicator variable approach for missingness is usually biased. In this case, the theoretical variable of interest has two possible values, “yes” and “no,” while the newly constructed variable for the dummy-coding approach would effectively have three possible values, “yes and willing to disclose use,” “no and willing to disclose nonuse,” and “unwilling to disclose.” If this three-level variable were of interest, then including a third category would change its meaning in a way that the investigator may not have desired and would likely bias the estimates and make interpretation difficult (Schafer & Graham, 2002, p. 149), both in terms of the covariate of interest and in terms of the other covariates in the model. That is, adding the indicator variable creates a different model, whose estimates are consequently biased for the corresponding quantities in the desired model. Schafer and Graham (2002) explain that a multiple imputation approach would be appropriate in that setting. However, the setting in which the use of a dummy-coded indicator variable is appropriate is clearly different. In our motivating example, being in a relationship with a romantic partner is part of the definition of the construct of partner’s substance use, so that specifying it does not obscure the original meaning of the construct. In fact, imputing a fictitious partner’s substance use for unpartnered individuals *would* obscure the meaning of the construct in this context, and could lead to very poor estimation performance, as we show later in a simulation study.

It is not essential to this two-part approach that the conditionally relevant predictor be binary when it is relevant, but only that it is unidimensional when it is relevant. Thus, in the current example, *PUSE* does not need to be binary, but could also be a count or a continuous index of the quantity or frequency of substance use. This intuition is the basis for our suggested method. This method allows all data to be included in the same model, with the indicator variable included in the model for all subjects (e.g., presence or absence of a relationship is known for all participants) and the conditionally relevant numerical variable included for only some subjects (e.g., *PUSE* is only relevant for people who are in a relationship). This is equivalent to estimating the model separately for each group but constraining all coefficients except those for the conditionally relevant covariate to be the same in both groups. Although this sounds complicated, it can be implemented easily with standard software. We propose the use of either of the following algebraically equivalent methods.

Method 1:

1. Create a dummy-coded variable (e.g., *REL*) that is 1 when the numerical covariate is relevant and 0 when it is not.

2. Create a recoded numerical variable (such as $PUSE^*$) to have its observed value for the observations in which it is relevant, and zero in the cases in which it is not relevant. (This imputed value must be zero, not another value such as the sample mean of nonmissing individuals).
3. Include the dummy-coded variable (e.g., REL) and the recoded numerical value (e.g., $PUSE^*$) as predictors in the model.

The following approach is algebraically equivalent:

Method 2:

1. Create a dummy-coded variable (e.g., REL) that is 1 when the numerical covariate is relevant and 0 when it is not.
2. Create a recoded numerical variable (such as $PUSE^*$) as before, to have its observed value for the observations in which it is relevant, and any value (e.g., perhaps the sample mean of nonmissing individuals) for cases in which it is not relevant.
3. Include the dummy-coded status variable and its *product* with the recoded numerical variable in the model (e.g., REL and $REL \times PUSE^*$, but not $PUSE^*$ itself).

Mathematically, Methods 1 and 2 are actually two ways of describing the same underlying method, in that they both give identical answers. Both approaches are presented here because they each give some insight into the idea being proposed. Also, there may be some situations in which a researcher might prefer one or the other because it might seem more palatable to readers. Method 1 might seem awkward if the covariate is measured on a scale which does not include zero, because it superficially appears to involve imputing a zero value, although as explained above this is not really what is being done. Method 2 might also seem awkward because it involves fitting an interaction without one of its parent main effects, which can confound regression results in some settings; however, in this setting it only reflects the natural confounding built into the meaning of the variable (partner's use is meaningless if there is no partner) and causes no difficulties in interpretation.

Algebraic Justification

To make the arguments above more concrete, consider a cross-sectional regression model to predict a response Y , representing participant's own substance use. For simplicity suppose here that Y is normally distributed conditional upon the covariates. In the data analysis presented later, Y is a dummy coded dichotomous variable, and therefore we will use a logistic regression. However, in this algebraic presentation we present a linear regression for convenience. Three covariates are of interest: AGE , which represents chronological age centered at 29; REL , which represents being in a relationship (1=yes, 0=no); and $PUSE$ which represents partner's use of drugs and is not recorded when $REL=0$. $PUSE$ could be normally distributed and perhaps on the same scale as Y , or it could be binary or count. Suppose that we wish to estimate a general effect of AGE and REL for everyone, but only

wish to consider *PUSE* when *REL*=1. Suppose an investigator believes that participants with partners can be reasonably described by the model

$$Y = \beta_{0, PARTNERED} + \beta_{AGE} AGE + \beta_{PUSE} PUSE + e, \quad (1)$$

and that participants without partners can be reasonably described by the model

$$Y = \beta_{0, SINGLE} + \beta_{AGE} AGE + e, \quad (2)$$

where the β_{AGE} values are assumed to be the same between subsets, but the β_0 values are not; and where for both subsets the residuals are independent and normally distributed with mean 0 and unknown constant variance σ^2 . Then fit the regression model (for Method 1 described above)

$$Y = \beta_0 + \beta_1 AGE + \beta_2 REL + \beta_3 PUSE^* + e, \quad (3)$$

or equivalently the regression model (for Method 2 described above)

$$Y = \beta_0 + \beta_1 AGE + \beta_2 REL + \beta_3 REL \times PUSE^* + e, \quad (4)$$

with the same assumption on the residuals as before (i.e., independent Normal(0, σ^2)).

First consider the special case in which *PUSE* is binary and coded as 0 or 1. For participants in a relationship with a partner who does not use drugs, *REL*=1 and *PUSE**=0, so for these participants (3) or (4) is written as $Y = \beta_0 + \beta_1 AGE + \beta_2 (1) + \beta_3 (0) + e$; and for participants in a relationship with a partner who does use drugs, *REL*=1 and *PUSE**=1, so for these participants (3) or (4) is written as $Y = \beta_0 + \beta_1 AGE + \beta_2 (1) + \beta_3 (1) + e$. Similarly, for participants not in a relationship, *REL*=0 and *PUSE**=0, so for these participants (3) or (4) is written as $Y = \beta_0 + \beta_1 AGE + \beta_2 (0) + \beta_3 (0) + e$. This is exactly dummy coding, with single individuals as the reference class. Even if *PUSE* is not binary, Equation (3) or (4) can similarly be written in a conditional or two-part form

$$Y = \begin{cases} \beta_{0, PARTNERED} + \beta_1 AGE + \beta_3 PUSE + e & \text{if } REL=1 \\ \beta_{0, SINGLE} + \beta_1 AGE + e & \text{otherwise} \end{cases} \quad (5)$$

where $\beta_{0, SINGLE} = \beta_0$ and $\beta_{0, PARTNERED} = \beta_0 + \beta_2$, and where the same assumptions about the residuals are being made as before (independent Normal(0, σ^2)). From (5), it can be seen that each parameter has an appropriate interpretation:

- $\beta_{0, SINGLE}$ is the fitted *Y* for a 29-year-old single person;

- $\beta_{0,PARTNERED}$ is the fitted Y for a 29-year-old person who is in a relationship with a person whose drug use is zero;
- β_1 is the expected difference in Y for a one-unit difference in age, holding constant relationship status (including partner drug use if relevant);
- $\beta_2 = \beta_{0,PARTNERED} - \beta_{0,SINGLE}$ is the expected Y for participants in a relationship with a non-user minus the expected Y for participants not in a relationship;
- β_3 is the expected difference in Y for a one-unit difference in $PUSE$, only for people in a relationship.

Thus, it is not assumed in (3), (4), or (5) that the effect of being single is the same as the effect of having a partner who does not use substances. That is, this approach is not imputing $PUSE=0$ for single people, even though it does involve coding $PUSE^*=0$ for them. Rather, the effect of being in a relationship with a partner who does not use, relative to being single, is freely estimated by the parameter β_2 . If REL had been omitted and only the artificial variable $PUSE^*$ had been used, then this faulty imputation would in fact be taking place. However, this would be inadvisable in the same way that using only one out of a set of dummy codes is inadvisable.

One might ask whether $PUSE^*$ has to be set to 0 for single participants. Could it not be set to some other value, such as the average value of $PUSE$ for participants in a relationship? This question would be especially salient if $PUSE$ were expressed on a scale in which zero was not meaningful (for example, on a 1 to 10 scale). In Equation (4), clearly any value can be used for $PUSE^*$ when $REL=0$, because it will simply be multiplied by zero regardless. In Equation (3), this is not the case, and using a different value for $PUSE^*$ for people not in a relationship will lead to a different estimate, at least of β_0 . Thus, Equation (4) might be a more convenient formulation. However, this is not really a question of model validity or substantive meaning, because it is never necessary to use a number other than zero in this context. Either way, if the steps above are followed correctly, the pseudo-imputed value never makes its way into the analysis.

Equations (3), (4), and (5) do assume that the residual error variance is the same between $REL=0$ and $REL=1$ individuals. However, this can be relaxed in some software (e.g., by using the LOCAL option in PROC MIXED in SAS software; see SAS Institute, 2011). As an additional caveat, we centered age at a common zero point for both single and partnered individuals (here, at age 29). In some situations (e.g., if unpartnered participants are on average much younger than partnered ones) it might be necessary to exercise some caution in interpretation to avoid something like Simpson's paradox (see, e.g., Baker & Kramer, 2001). However, hidden extrapolation such as this is a well-understood complication in multiple regression modeling (Ramsey & Schafer, 2002) and would have arisen even if the partner's use of drugs were not in the model. Note that the model represented by Equation (3), (4), or (5) does assume that the effect of age is the same for single people and people in a relationship. This was intentional, but could be easily relaxed either by considering the two groups separately or by including an interaction between relationship status and age. The models represented by Equations (3), (4), and (5) also assume the same error variance for

both single people and people in a relationship. This could be a problem, but it would have arisen even if *PUSE* had not been in the model at all. It could be avoided by fitting the two parts of Equation (5) separately for partnered and nonpartnered individuals. However, as mentioned earlier, subsetting the sample might become impractical if used for several covariates or for longitudinal data.

In summary, Equations (3) and (4) are equivalent in their assumptions to Equation (5). They will give the same fitted values. They are simply two ways of coercing standard regression software to fit Equation (5) without having to subset the data. In this section we used a cross-sectional linear model to make the notation as simple as possible. However, the simple algebraic arguments above apply equally well to longitudinal data and to logistic or other generalized linear models, after appropriate modifications.

Examples

We illustrate our proposed approach in cross-sectional and longitudinal settings using a real dataset from companion longitudinal studies (Thornberry, 2009; Thornberry, Freeman-Gallant, et al., 2003; Thornberry, Krohn, et al. 2003; Thornberry et al., 2014). The original study, the Rochester Youth Development Study (RYDS), began in 1988, and the intergenerational extension, the Rochester Intergenerational Study (RIGS), began in 1999. The original RYDS sample of 1,000 adolescents (referred to as G2) is representative of the 7th and 8th grade public school population of Rochester, New York at the start of the study. Youth at high risk for problem behaviors (males and adolescents residing in high crime neighborhoods) were oversampled. The RIGS selected G2's oldest biological child (G3; age two or older) and added new firstborns as they turned two in each subsequent year. In Year 1 of RIGS, the average age of G2 participants was 25.3 ($SD=.8$). As of Year 12 of the study (the last year considered for the analyses presented here), 518 G2 participants and their firstborn child (G3) had participated in the study. Annual interviews of G2 and G3 have occurred since Year 1, though beginning in Year 9, interviews with the G2 participants were discontinued once G3 turned 18. Procedures were approved by the University at Albany's Institutional Review Board. Thornberry et al. (2009) and Thornberry, Krohn et al. (2003) provide more details on the study.

Measures

Our outcome variable of interest is *G2's self-report illegal drug use* at each interview. For simplicity, we consider a binary variable, where 1=one or more instances of illegal drug use during the prior year and 0=no illegal drug use during the prior year.

At each annual interview, the participant's intimate partner relationship was ascertained: 0=no partner; 1=married; 2=not married, but living with a partner; 3=long-term partner (> 6 months), but not married and not living with that partner; 4=new partner (<6 months). We created a binary measure of *committed relationship status* (referred to as *REL* below): individuals who reported a 1, 2 or 3 were coded 1, individuals who reported a 0 or a 4 were coded 0.

We consider two predictors of interest—one dichotomous and one continuous. For the former, individuals in a committed relationship were asked how often their *partner used illegal drugs* (referred to as *PUSE* below) during the prior year. For simplicity, we consider a binary variable, where 1 implies one or more instances of partner illegal drug use during the prior year and 0 implies no partner illegal drug use during the prior year. For the latter, individuals in a committed relationship were asked how satisfied they were in the relationship using a 5-item scale. All item responses were provided using a 5-point scale, 1 (never) to 5 (always), with a higher score indicating more satisfaction. The average of the five items formed the relationship satisfaction scale considered here. To demonstrate that additional covariates can be added, we also consider *age* as a predictor in the model. We center this variable at the average age at Year 5 of the study (29 years).

Example with Cross-Sectional Data and a Binary Covariate

To begin, we consider a cross-sectional model and create a subset of just the Year 5 data. We fit a series of logistic regression models for demonstrative purposes. The results are presented in Table 2.

We first consider only partnered individuals and fit a model that is similar to (1) but uses logistic rather than linear regression. Specifically, we assume

$$\text{logit} (P (Y=1)) = \beta_0 + \beta_1 AGE + \beta_2 PUSE. \quad (6)$$

The estimate for β_1 is 0.02 ($SE=0.20$, $p=.92$), indicating essentially no relationship with age. The estimate for β_2 is 2.64 ($SE=0.42$, $p<.0001$), suggesting that participants whose partners use drugs have $\exp(2.64) \approx 14$ times higher odds of using drugs than those whose partners do not. This model says nothing about individuals without partners.

We now fit a model analogous to (2) on participants without partners, specifically,

$$\text{logit} (P (Y=1)) = \beta_0 + \beta_1 AGE. \quad (7)$$

The regression coefficient for age is $-.15$ ($SE=.23$, $p=.53$). If we believe that this is practically significantly different from the 0.02 observed in Model (6), then we might choose not to combine the partnered and unpartnered individuals into one analysis. The decision of whether to have equal or unequal coefficients could also be addressed using a likelihood ratio test in an appropriate larger model. However, for illustrative purposes we simply consider the combined model.

Thus, we consider all individuals and fit a model similar to (3), specifically,

$$\text{logit} (P (Y=1)) = \beta_0 + \beta_1 AGE + \beta_2 REL + \beta_3 PUSE^*. \quad (8)$$

The estimation for the effect of partner drug use is the same as in the subsetting analysis (i.e., 2.64). Additionally, we get an estimate of the coefficient for *REL*. Importantly, because *PUSE** is also in the model, the coefficient for *REL* does not compare all individuals having a partner with all individuals not having a partner. That is, it is not the general or average association between substance use and having a romantic partner. Instead, $\exp(\beta_2) \approx 0.38$ is the odds ratio of substance use for *individuals in a relationship with a nonusing partner* compared to *individuals not in a relationship*. To compare individuals in a relationship with a drug-using partner to individuals not in a relationship, one would use $\exp(\beta_2 + \beta_3) \approx 5.31$. This analysis suggests, intuitively, that having a partner who does not use drugs is associated with a much lower odds of using drugs relative to a person who does not have a partner, whereas having a partner who does use drugs is associated with a much higher odds of using drugs relative to a person who does not have a partner. Stated another way, relative to not being in a relationship, a non-drug-using partner is a possible protective factor and a drug-using partner is a possible risk factor for drug use; however, of course, this finding does not give information about the direction of causality. The relationship between substance use and being in a relationship in general (ignoring whether the partner uses drugs or not), if such a question were of interest, could be studied by fitting a model containing *REL* and excluding *PUSE* and *PUSE**.

We also fit a model similar to (4), specifically

$$\text{logit}(P(Y=1)) = \beta_0 + \beta_1 \text{AGE} + \beta_2 \text{REL} + \beta_3 \text{REL} \times \text{PUSE}^*. \quad (9)$$

Not surprisingly, the answers are the same as those of Model (8).

We also fit a version of (9) in which *PUSE** for unpartnered individuals was set to the mean for partnered individuals rather than to zero. Precisely the same estimates were obtained, for reasons described earlier. In contrast, if we had instead fit a naïve model that included *PUSE** but omitted *REL* (i.e., regressed *Y* only on *AGE* and *PUSE**), the same results would not have been obtained.

Example with Cross-Sectional Data and a Continuous Covariate

In order to illustrate the model with a continuous covariate, we used relationship satisfaction (*SAT*) instead of partner's use of drugs (*PUSE*). For participants in committed relationships, we centered relationship satisfaction at the mean; for those not in such relationships, we substituted a score of 0 for the centered score. Next, we regressed substance use on age (centered at age 29), an indicator of relationship status (1=in a relationship, 0=not in a relationship), and the product term of relationship status and the modified relationship satisfaction variable *SAT**. That is,

$$\text{logit}(P(Y=1)) = \beta_0 + \beta_1 \text{AGE} + \beta_2 \text{REL} + \beta_3 \text{REL} \times \text{SAT}^*. \quad (10)$$

The estimated coefficients for the model are presented in Table 3.

In Table 3, the negative β_3 represents the expected decrease in the log odds of substance use associated with a one-unit increase in relationship satisfaction, among people in a relationship. It is also possible to include both partner's drug use and relationship satisfaction in the model together, as in the following model.

$$\text{logit}(P(Y=1)) = \beta_0 + \beta_1 AGE + \beta_2 REL + \beta_3 REL \times SAT^* + \beta_4 REL \times PUSE^*. \quad (11)$$

Corresponding estimates are shown in the last column of Table 3, and generally agree in interpretation with previous results (among those in a relationship, relationship satisfaction predicts lower likelihood of drug use, and partner's use predicts higher likelihood). It would also be possible to include a $REL \times SAT^* \times PUSE^*$ interaction (e.g., perhaps individuals who are dissatisfied with living with a drug-using partner are less likely to use drugs themselves), but we do not explore this here, in order to save space.

Longitudinal Example

Now we return to the full dataset (all 12 years of data) and consider a longitudinal example. We fit a multilevel logistic regression model with individual participant as level 2 and measurement occasion as level 1. Here, the log odds of substance use for person i at measurement occasion j are expressed as

$$\text{logit}(P(Y_{ij}=1)) = \beta_0 + \beta_1 (AGE_{ij}) + \beta_2 (AGE_{ij})^2 + \beta_3 REL_{ij} + \beta_4 REL_{ij} \times PUSE_{ij}^* + a_i, \quad (12)$$

where the subject-level random intercepts a_i were assumed to be independent and normally distributed with unknown variance τ^2 . As was the case in the cross-sectional model, AGE_{ij} is centered at 29, and now we include a quadratic term for age as well. REL_{ij} is the dummy-coded indicator that individual i is in a relationship at measurement time j . $PUSE_{ij}^*$ is coded 1 for individuals who are in a relationship with a partner who uses substances, and 0 for all others (i.e., those in a relationship with a partner who does not use substances or those not in a relationship). Note that the values of both REL and $PUSE$ are free to change at each measurement occasion as individuals move in and out of relationships and partners' behaviors change; we assume that their regression coefficients are constant over time, although this could be relaxed by considering interactions with time. The results of the multilevel logistic regression model are presented in Table 4.

We observe a significant, negative regression coefficient for relationship status and a significant, positive regression coefficient for partner's drug use.

This longitudinal example shows the advantage of treating relationship status and partner's substance use as a single polytomous variable over the alternatives: multiply imputing the substance use of nonexistent people or dividing partnered and unpartnered occasions into separate data sets. If the multiple imputation approach were used, the resulting model would

be very difficult to meaningfully interpret, because the only information available for imputing *PUSE* is found in the partnered occasions, but *PUSE* needs to be imputed only for the unpartnered occasions. If the subsetting approach were used, it would be very difficult in standard software to enforce the assumption that the intercept $\beta_0 + a_i$ is the same across all occasions for a given participant i , or to make or test parsimonious assumptions such as equating β_1 and β_2 between the subsets.

Finally, we present the results of a similar model that includes relationship satisfaction rather than partner's use of drugs and thus demonstrates the model with a continuous covariate.

$$\text{logit}(P(Y_{ij}=1)) = \beta_0 + \beta_1(AGE_{ij}) + \beta_2(AGE_{ij})^2 + \beta_3REL_{ij} + \beta_4REL_{ij} \times SAT_{ij}^* + a_i,$$

(13)

The results are also presented in Table 4. As might be expected, we see a significant, negative association with relationship satisfaction.

Simulation Study

Although we have argued that considering undefined missing data as an additional dimension is logical, it is still reasonable to ask whether it really offers a noticeable empirical advantage over other approaches such as multiple imputation. Therefore, we performed a simulation experiment to compare the techniques. In each of 5000 datasets, we simulated 400 individuals assuming that Model (8) was correct in the population. *REL* was randomly generated under the assumption that 50% of individuals had a partner. For partnered individuals, *PUSE* was randomly generated under the assumption that 25% of partners used drugs. Independently from *PUSE*, *AGE* was randomly generated from a uniform distribution between 17 and 37 for unpartnered and 21 to 41 for partnered individuals, then centered at 25 for both groups. The estimates shown in Table 2 for Model (8) were used as population values in the simulation. We compared three estimation techniques: (1) fitting Model (8), which was the correct data-generating model; (2) instead of the correctly coded *PUSE**, use a singly mean-imputed value for *PUSE** which was 1 if *REL*=1 and *PUSE*=1, 0 if *REL*=1 and *PUSE*=0, and .25 if *REL*=0; or (3) instead of the correctly coded *PUSE**, use a multiply imputed value for *PUSE** when *REL*=0 (imputed using a logistic regression of *PUSE* on *AGE* within the partnered individuals). Of course, because Model (8) was the same model used to simulate the data, it was expected to perform best; however, we were interested in the extent to which other methods would underperform. The results are shown in Table 5.

As Table 5 shows, the only method that performed adequately was the use of the correct model. It had little or no bias, low noise (as measured by root mean squared error), and correct coverage (95% of the nominal 95% confidence intervals contained the data-generating values of the parameters).

Under the single imputation approach, the estimate for β_3 was actually unbiased. Intuitively this is because the procedure estimated β_3 while holding *REL* constant and assuming no interaction, which meant that it derived its information from the *REL*=1 subjects only (recall that the *REL*=0 subjects provided no information for that part of the model because *PUSE** was always constant for them). However, the inherent confounding between *REL* and *PUSE** in the model manifested itself anyway, in a severely biased estimate for β_2 . β_1 was unbiased, essentially because it was unrelated to the part of the model containing the misspecification.

The multiple imputation approach in this context had large bias and extremely poor coverage for all parameters except β_1 . Existing theory in multiple imputation (see, e.g., Schafer, 1999 and references therein) explains why it fails here: the imputation model is misspecified and incomplete because it ignores *REL*. However, the imputation model for *PUSE* cannot include *REL* in this context, because it is constant (*REL*=1) whenever *PUSE* is not missing. This dilemma is the reason why multiple imputation and other technical statistical solutions cannot overcome a conceptual shortcoming (namely, pretending that a value is missing at random, when it is actually structurally missing).

It is important to emphasize that we do not interpret these results as evidence that multiple imputation in general is flawed or should not be used. We simulated a situation in which the investigator had a choice between choosing a simple approach that appropriately fit the nature of the study and the question of interest, or else using an advanced approach blindly without considering its assumptions or meaning (i.e., applying a missing data technique to data that are really two-part rather than missing). Not surprisingly, the simpler approach performed much better.

Discussion

In this paper we describe a straightforward method for handling the common situation in which certain regression covariates are relevant only for certain participants. We argue that multiple imputation is not appropriate in this setting for two reasons. First, it is unnecessary, because it attempts to account for uncertainty that does not really exist. Second, it would likely introduce bias and noise. Therefore, we recommend instead treating the covariate as essentially two-dimensional, with a variable that is conditionally relevant nested within one level of a variable that is always relevant.

A positive side effect of this two-part representation is that it can sometimes give insights that would not have been obtained by treating a covariate as unidimensional. For example, Hedeker, Demirtas and Mermelstein (2009) and Hedeker, Mermelstein, and Demirtas (2012) modeled participants' self-reported positive and negative affect from covariates that included smoker status (yes versus no) and amount smoked (set at 0 for nonsmokers). Far from being redundant, the two covariates were found to have coefficients with different sign. That is, smokers reported feeling more positive than nonsmokers, but light smokers were more positive than heavy smokers; the authors were able to give a theoretical interpretation of these seemingly opposite findings.

There is some analogy between two-part covariates as described here and nested factors in ANOVA with designed experiments (see Maxwell & Delaney, 2004). When factors are nested, each level of one factor is observed with several levels of the second factor, but each level of the second factor is observed only within a single level of the first factor. It could be said that *PUSE* in our example is nested within one level of *REL*. A classic example of nested ANOVA is in animal breeding experiments in which each male is paired with several females (but each female with only one male), and the goal is to determine the degree to which a trait found in offspring depends on the mother, on the father, or both. In the social sciences, multilevel models are considered nested; for example, one school contains several students but each student attends only one school. However, these examples are unnecessarily complex for the issue of interest in this paper, especially because they generally involve random effects (quantification of random differences among populations of people or animals) rather than estimation of fixed effects (essentially, regression coefficients).

One remaining challenge that would be faced both by our proposed method and by competing methods (such as multiple imputation or subsetting) is the question of selection bias in a longitudinal setting. Whether an individual has a partner is not entirely random, but is related to events in the present or past. In some cases, more thought may be required to identify the target population at a given time. This issue has been discussed in the context of longitudinal biomedical studies, where it is especially salient because of missing data due to death. It is not sensible to impute or predict the health of people who have died, yet as time goes on, the exclusion of dead people means that naturally healthier people are likely to remain in the cohort longer. Depending on the exact research question, various approaches are available that do not require the imputation of uninterpretable data; some are described by Kurland, Johnson, Eggleston, and Diehr (2009). In such cases, stratifying by or conditioning on the time of death may be required. In addressing these issues, a rule of thumb or technical fix is less helpful than a precise statement of the question.

Furthermore, our recommendations apply only if the conditionally relevant variable is a covariate (i.e., treated as exogenous), rather than an outcome variable. If the conditionally defined variable is an outcome of the model, then a different technique is needed, such as two-part modeling as described in Olsen & Schafer (2001) and Flora (2011). Similarly, if the processes that determine substance use for single and partnered individuals are very different, then including them both in the same regression model may lead to results that do not describe either group well. This can be addressed by including and testing interaction terms (although in the case of linear models this does not remove the assumption of a common variance for both groups; see Jones 1996) or by comparing results for the different subsets. It cannot be addressed simply by using multiple imputation, because it is an issue of model misspecification, not merely of missing data.

Finally, we reiterate that we are *not* recommending the use of an indicator variable for missingness as a *general* solution to missing data. The flaws of such an approach are explored by Greenland and Finkle (1995), Jones (1996), and Schafer and Graham (2002). However, the proofs of Jones (1996) and the simulations of Greenland and Finkle (1995) assumed a true but unknown value for the missing covariate. For example, in the context of

Jones (1996), the magnitude of the bias introduced by using the missingness indicator method depended on the covariance between the missing covariate and the nonmissing covariates for the subset of subjects who are missing the covariate. In our example, this would seem to mean that our proposed method is biased if *the correlation between age and the true substance use of the partners of unpartnered subjects is nonzero*. In fact, this correlation is neither zero nor nonzero, but is meaningless. In response to this paradox, we note that Jones (1996) considers a conceptually non-nested situation, and we consider conceptually nested situation. In our situation, the “indicator and stratification methods” that are generally so unsatisfactory become natural and satisfactory approaches, while the usually preferred approaches such as multiple imputation become inappropriate. This underscores the importance of careful thought about the meaning and role of covariates in a model.

References

- Baker SG, Kramer BS. Good for women, good for men, bad for people: Simpson’s paradox and the importance of sex-specific analysis in observational studies. *Journal of Women’s Health & Gender-Based Medicine*. 2001; 10:867–872. DOI: 10.1089/152460901753285769
- Cohen, J., Cohen, P., West, SG., Aiken, LS. *Applied multiple regression/correlation analysis for the behavioral sciences*. 3rd. Mahway, New Jersey: Lawrence Erlbaum; 2003.
- Flora DB. Two-part modeling of semicontinuous longitudinal variables: A comparison of approaches. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*. 2011; 7:145–156. DOI: 10.1027/1614-2241/a000032
- Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*. 1995; 142:1255–64. DOI: 10.1093/oxfordjournals.aje.a117592 [PubMed: 7503045]
- Hastie, T., Tibshirani, R., Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. New York, NY: Springer; 2001.
- Hedeker D, Demirtas H, Mermelstein RJ. A mixed ordinal location scale model for analysis of Ecological Momentary Assessment (EMA) data. *Statistics and its Interface*. 2009; 2:391–401. DOI: 10.4310/SII.2009.v2.n4.a1 [PubMed: 20357914]
- Hedeker D, Mermelstein RJ, Demirtas H. Modeling between- and within-subject variance in ecological momentary assessment (EMA) data using mixed-effects location scale models. *Statistics in Medicine*. 2012; 31:3328–3336. DOI: 10.1002/sim.5338 [PubMed: 22419604]
- Jones MP. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*. 1996; 91:222–230. DOI: 10.2307/2291399
- Kurland BF, Johnson LL, Egleston BL, Diehr PH. Longitudinal data with follow-up truncated by death: Match the analysis method to research aims. *Statistical Science*. 2009; 24:211. doi: 10.1214/09-STS293 [PubMed: 20119502]
- Manne S, Kashy D, Weinberg DS, Boscarino JA, Bowen DJ. Using the interdependence model to understand spousal influence on colorectal cancer screening intentions: A structural equation model. *Annals of Behavioral Medicine*. 2012; 43:320–329. DOI: 10.1007/s12160-012-9344-y [PubMed: 22274725]
- Maxwell, SE., Delaney, HD. *Designing experiments and analyzing data: a model comparison perspective*. 2nd. Mahwah, NJ: Lawrence Erlbaum Associates; 2004.
- Myers, JL., Well, AD. *Research design and statistical analysis*. Mahwah, NJ: Lawrence Erlbaum Associates; 2003.
- Olsen MK, Schafer JL. A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*. 2001; 96:730–745. DOI: 10.1016/j.jhealeco.2009.11.010
- Parke RD, Coltrane S, Duffy S, Buriel R, Dennis J, Powers J, French S, Widaman KF. Economic stress, parenting, and child adjustment in Mexican American and European American families.

Child Development. 2004; 75:1632–1656. DOI: 10.1111/j.1467-8624.2004.00807.x [PubMed: 15566370]

Ramsey, FL., Schafer, DW. The statistical sleuth. 2nd. Pacific Grove, CA: Duxbury; 2002.

Rubin, DB. Multiple imputation for nonresponse in surveys. New York, NY: Wiley; 1987.

SAS Institute Inc. SAS/STAT® 9.3 User's Guide. Cary, NC: SAS Institute Inc; 2011.

Schafer JL. Multiple imputation: A primer. Statistical Methods in Medical Research. 1999; 8:3–15. DOI: 10.1177/096228029900800102 [PubMed: 10347857]

Schafer JL, Graham JW. Missing data: Our view of the state of the art. Psychological Methods. 2002; 7:147–177. DOI: 10.1037/1082-989X.7.2.147 [PubMed: 12090408]

Thornberry TP. The apple doesn't fall far from the tree (Or does it?): Intergenerational patterns of antisocial behavior. Criminology. 2009; 47:297–325. DOI: 10.1111/j.1745-9125.2009.00153.x [PubMed: 25308976]

Thornberry TP, Freeman-Gallant A, Lizotte AJ, Krohn MD, Smith CA. Linked lives: The intergenerational transmission of antisocial behavior. Journal of Abnormal Child Psychology. 2003; 31:171–184. DOI: 10.1023/A:1022574208366 [PubMed: 12735399]

Thornberry, TP., Krohn, MD., Lizotte, AJ., Smith, CA., Tobin, K. Gangs and Delinquency in Developmental Perspective. New York: Cambridge University Press; 2003.

Thornberry TP, Matsuda M, Greenman SJ, Bears Augustyn M, Henry KL, Smith CA, Ireland TO. Adolescent risk factors for child maltreatment. Child Abuse & Neglect. 2014; 38:706–722. DOI: 10.1016/j.chiabu.2013.08.009 [PubMed: 24075569]

Widaman KF, Grimm KJ, Early DR, Robins RW, Conger RD. Investigating factorial invariance of latent variables across populations when manifest variables are missing completely. Structural Equation Modeling. 2013; 20:384–408. DOI: 10.1080/10705511.2013.797819 [PubMed: 24019738]

Table 1

Coded values for each respective group for Equation 3 and 4

Group	<i>REL</i>	<i>PUSE</i>	<i>PUSE*</i>
In a relationship, partner doesn't use drugs	1	0	0
In a relationship, partner uses drugs	1	1	1
Single	0	Undefined	0

Table 2

Coefficient Estimates and Standard Errors for Logistic Regression Models From Cross-Sectional Data – Dichotomous Covariate

Coefficient	Model 6 N=282	Model 7 N=126	Model 8 N=408	Model 9 N=408
Intercept	−1.85 (0.19) **	−0.85 (0.20) **	−0.87 (0.20) **	−0.87 (0.20) **
Age minus 29	0.02 (0.20)	−0.15 (0.23)	−0.05 (0.15)	−0.05 (0.15)
In a relationship (<i>REL</i>)			−0.97 (0.27) **	−0.97 (0.27) **
Partner drug use (<i>PUSE</i>)	2.64 (0.42) **			
Partner drug use recoded (<i>PUSE*</i>)			2.64 (0.42) **	
Interaction (<i>REL</i> × <i>PUSE*</i>)				2.64 (0.42) **
−2 Log Likelihood	238.56	152.17	391.02	391.02

Notes. Coefficients shown are logistic regression coefficients (i.e., log odds ratios). Model numbers refer to the corresponding Equations in the text. Numbers in parentheses are standard errors. In Model (6), fit only to partnered individuals, the variable *PUSE* is defined as 1 if the individual's partner is using drugs, 0 if the partner is not, and undefined if the individual has no partner. Models (7) through (9) are fit to both partnered and unpartnered individuals, so their log-likelihoods can be compared to each other but not to that of Model (6). In Model (8), fit to partnered and unpartnered individuals, the variable *PUSE** is defined as 1 if the individual is in a relationship with a partner who uses drugs and defined as 0 otherwise. In Model (9), the variable *PUSE** is defined as 1 if an individual is in a relationship with a partner who uses drugs, and could be defined either as 0 or any other value, such as the sample mean, without changing any of the coefficient estimates, if the individual has no partner.

*
 $p < .01$,

**
 $p < .001$

Table 3

Results of Logistic Regression Model From Cross-Sectional Data – Continuous Covariate

Coefficient	Ignoring Partner Drug Use (Model 10)	Considering Partner Drug Use (Model 11)
Intercept	−0.87 (0.20) **	−0.87 (0.20) **
Age minus 29	−0.03 (0.14)	−0.05 (0.15)
In a relationship (<i>REL</i>)	−0.58 (0.25) *	−0.97 (0.27) **
In a relationship × Relationship Satisfaction (<i>REL</i> × <i>SAT</i> [®])	−0.74 (0.23) *	−0.45 (0.26)
In a relationship × Partner Drug Use (<i>REL</i> × <i>PUSE</i> [®])		2.48 (0.43) **

Notes. Coefficients shown are logistic regression coefficients (i.e., log odds ratios). Numbers in parentheses are standard errors.

N=408.

*
 $p < .01$,

**
 $p < .001$

Table 4

Results of Random-Intercept Logistic Regression Model for Longitudinal Data

Coefficient	Binary Covariate (Model 12)	Numerical Covariate (Model 13)
Intercept	-1.55 (0.14) **	-1.60 (0.15) **
(Age - 29)	-0.12 (0.02) **	-0.12 (0.02) **
(Age - 29) * (Age - 29)	0.01 (0.004) *	0.01 (0.004) *
In a relationship (<i>REL</i>)	-0.69 (0.13) **	-0.34 (0.12) *
In a relationship $\times$ Partner drug use (<i>REL</i> $\times$ <i>PUSE</i> *)	2.01 (0.19) **	
In a relationship $\times$ Relationship Satisfaction (<i>REL</i> $\times$ <i>SAT</i> *)		-0.43 (0.11) **

Notes. Coefficients shown are logistic regression coefficients (i.e., log odds ratios). Numbers in parentheses are standard errors. The variable *PUSE** is defined as 1 if an individual is in a relationship with a drug-using partner, 0 if in a relationship with a partner who doesn't use drugs, and could be defined either as 0 or any other value, such as the sample mean, without changing any of the coefficient estimates, if the individual has no partner. The variable *SAT**, represents the relationship satisfaction scale score for people in a relationship, and the mean score is substituted for people not in a relationship, but any score could be substituted without changing any of the coefficient estimates. N=518.

*
 $p < .01$,

**
 $p < .001$

Table 5

Comparative Performance of Methods in Simulation Study

	Correct Model	Incorrect Single Imputation	Incorrect Multiple Imputation
β_0 Bias	−0.007	−0.680	−0.356
β_1 Bias	−0.001	−0.001	0.003
β_2 Bias	−0.024	0.649	0.854
β_3 Bias	0.050	0.050	−1.378
β_0 RMSE	0.028	0.501	0.165
β_1 RMSE	0.000	0.000	0.000
β_2 RMSE	0.096	0.495	0.795
β_3 RMSE	0.171	0.171	1.934
β_0 Coverage	0.951	0.047	0.597
β_1 Coverage	0.948	0.948	0.962
β_2 Coverage	0.949	0.312	0.087
β_3 Coverage	0.950	0.950	0.001

Note: RMSE = square root of mean squared error of estimation. Coverage = obtained coverage for nominal 95%.