

Missing Data

This discussion borrows heavily from:

Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, by Jacob and Patricia Cohen (1975 edition). The 2003 edition of Cohen and Cohen's book is also used a little.

Paul Allison's Sage Monograph on Missing Data (Sage paper # 136, 2002).

Newman, Daniel A. 2003. Longitudinal Modeling with Randomly and Systematically Missing Data: A Simulation of Ad Hoc, Maximum Likelihood, and Multiple Imputation Techniques. Organizational Research Methods, Vol. 6 No. 3, July 2003 pp. 328-362.

Patrick Royston's series of articles in volumes 4 and 5 of The Stata Journal on multiple imputation. See especially Royston, Patrick. 2005. Multiple Imputation of Missing Values: Update. The Stata Journal Vol. 5 No. 2, pp. 188-201.

Often, part or all of the data are missing for a subject. This handout will describe the various types of missing data and common methods for handling it. We will not go over some of the more advanced methods, but the readings can help you with these if you are interested.

I. Types of missing data. There are several useful distinctions we can make.

- *Dependent versus independent variables.* Most methods involve missing values for IVs, although in recent years methods for dealing with missing data in the dependent variable have been developed.
- *Random versus selective loss of data.* A researcher must ask why the data are missing. In some cases the loss is completely at random (MCAR), i.e. the absence of values on an IV is unrelated to Y or other IVs. Unfortunately, in survey research, the loss often is not random. Refusal or inability to respond may be correlated with such things as education, income, interest in the subject, geographic location, etc. Selective loss of data is much more problematic than random loss.
- *Missing by design; or, not asked or not applicable.* These are special cases of random versus selective loss of data. Sometimes data are missing because the researcher deliberately did not ask the question of that particular respondent. For economic reasons, some questions might only be asked of a random subsample of the entire sample. For example there is a "short" version of the census (answered by everyone) and a "long" version that is only answered by 20%. This can be treated the same as a random loss of data, keeping in mind that the loss may be very high.

Other times, *skip patterns* are used to only ask questions of respondents with particular characteristics. For example, only married individuals might be asked questions about family life. With this selective loss of data, you must keep in mind that the subjects who were asked questions are probably quite different than those who were not (and that the question may not have been asked of others because it would not make any sense for them).

It can be quite frustrating to think you've found the perfect question, only to find that 3% of your sample answered it! However, keep in mind that, many times, most subjects actually

may be answering the same or similar questions, but at different points in the questionnaire. For example, married individuals may answer question 37 while unmarried individuals are asked the same thing in question 54 (perhaps with a slight change of wording to reflect the differences in marital status). Hence, it may be possible to construct a more or less complete set of data by combining responses from several questions. Often, the collectors or distributors of the data have already done this for you.

- *Many versus few missing data and their pattern.* Is only 1% of the data missing, or 40%? Is there much data missing from a few subjects or a little data missing from each of several subjects? Is the missing data concentrated on a few IVs or is it spread across several IVs?

II. Alternatives for handling missing data

We will discuss several different alternatives here. We caution in advance that, while many of these methods have been widely used, some are very problematic and their use is not encouraged (although you should be aware of them in case you encounter them in your reading.)

- *Compare the missing and non-missing cases on variables where information is not missing.* Whatever strategy you follow you may be able to add plausibility to your results (or detect potential biases) by comparing sample members on variables that are not missing. For example, in a panel study, some respondents will not be re-interviewed because they could not be found or else refused to participate. You can compare respondents and non-respondents in terms of demographic characteristics such as race, age, income, etc. If there are noteworthy differences, you can point them out, e.g. lower-income individuals appear to be underrepresented in the sample. Similarly, you can compare individuals who answered a question with those who failed to answer to see if there are any apparent differences between the groups.
- *Dropping variables.* When, for one or a few variables, a substantial proportion of cases lack data, the analyst may simply opt to drop the variables. This is no great loss if the variables had little effect on Y anyway. However, you presumably would not have asked the question if you did not think it was important, so just dropping variables may or may not be a good solution. Still, this is often the best or at least most practical approach. A great deal of missing data for an item might indicate there are fundamental problems with it and it is hence unusable. Perhaps a question was poorly worded, or perhaps there were problems with collecting the data.
- *Dropping subjects, i.e. listwise (also called casewise) deletion of missing data.* Particularly if the missing data is limited to a small number of the subjects, you may just opt to eliminate those cases from the analysis. That is, if a subject is missing data on *any* of the variables used in the analysis, it is dropped completely. The remaining cases, however, may not be representative of the population. Even if data is missing on a random basis, a listwise deletion of cases could result in a substantial reduction in sample size, if many cases were missing data on at least one variable. My guess is that listwise deletion is the most common approach for handling missing data, and it often works well, but you should be aware of its limitations if using it.

Another thing to be careful of, when using listwise deletion, is to make sure that your

selected samples remain comparable when you are doing a series of analyses. Suppose, for example, you do one regression where the IVs are X1, X2, and X3. You do a subsequent analysis with those same three variables plus X4. The inclusion of X4 (if it has missing data) could cause the sample size to decline. This could affect your tests of statistical significance. You might, for example, conclude that the effect of X3 becomes insignificant once X4 is controlled for – but this could be very misleading if the change in significance was the result of a decline in sample size, rather than because of any effect X4 has. Also, as we’ve noted before, many statistical tests that explicitly compare models assume that the same cases are being analyzed throughout.

Also, if the X4 cases are missing on a nonrandom basis, your understanding of how variable effects are interrelated could also get distorted. For example, suppose X1-X3 are asked of all respondents, but X4 is only asked of women. You might see huge changes in the estimated effects of X1-X3 once X4 was added. This might occur only because the samples analyzed are different, e.g. if you only analyzed women throughout the effects of X1-X3 might change little once X4 was added.

In SPSS, you might use Select If or Filter commands to consistently limit your data to the cases you want. You can also often avoid this problem by using a single Regression command with variables entered hierarchically, e.g. first you enter X1-X3, then you enter X4. If you use two separate regression commands, you may run into trouble.

| Good | Bad |
|--|--|
| <pre>REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT y /METHOD=ENTER x1 x2 x3 /METHOD=ENTER x4 .</pre> | <pre>REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT y /METHOD=ENTER x1 x2 x3 . REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT y /METHOD=ENTER x1 x2 x3 x4 .</pre> |

In Stata, you have to be a little more careful. Especially if your subsample selection is a little complicated, it may be best just to temporarily drop the cases you don’t want. Some other approaches:

```
. reg y x1 x2 x3 if !missing(x4)
. reg y x1 x2 x3 x4
```

The first regression will only get run on those cases which do not have missing values on x4. Another approach:

```
. gen touse = !missing(y, x1, x2, x3, x4)
. reg y x1 x2 x3 if touse
```

The variable `touse` will be coded 1 if there is no missing data in any of the variables specified; otherwise it will equal 0. The `if` statement on the `reg` command will limit the analysis to cases with nonzero values on `touse` (i.e. the cases with data on all 5 variables).

The `nestreg` prefix is another very good approach when you are estimating a series of nested models, e.g. first you estimate the model with `x1 x2 x3`, then you estimate a model with `x1 x2 x3 x4 x5`, etc. `nestreg` does listwise deletion on all the variables, and will also give you incremental F tests showing whether the variables added in each step are statistically significant, e.g.

```
. nestreg: reg y (x1 x2 x3) (x4 x5)
```

- *The “missing-data correlation matrix,” i.e. pairwise deletion of missing data.* Such a matrix is computed by using for each pair of variables (X_i, X_j) as many cases as have values for both variables. That is, when data is missing for either (or both) variables for a subject, the case is excluded from the computation of r_{ij} . In general, then, different correlation coefficients are not necessarily based on the same subjects or the same number of subjects.

This procedure is sensible if (and only if) the data are randomly missing. In this case, each correlation, mean, and standard deviation is an unbiased estimate of the corresponding population parameter. The varying N s on which the regression results are based is somewhat problematic. The results are clearly not as sturdy as if the maximum N had been obtained throughout, nor as frail as the minimum N would suggest. Still, if data are missing randomly, this approach is not too bad.

Unfortunately, the assumption that data are missing randomly is often not valid. If data are not missing at random, several problems develop:

- ✓ The pieces put together for the regression analysis refer to systematically different subsets of the population, e.g. the cases used in computing r_{12} may be very different than the cases used in computing r_{34} . Results cannot be interpreted coherently for the entire population or even some discernible subpopulation.
- ✓ One can obtain a missing-data correlation matrix whose values are mutually inconsistent, i.e. it would be mathematically impossible to obtain such a matrix with any complete population (e.g. such a matrix might produce a multiple R^2 of $-.3!$) It may be even worse, though, if you do get a consistent matrix. With an impossible matrix, you'll receive some sort of warning that the results are implausible, but with a consistent matrix the results might seem OK even though they are total nonsense.

Also, even if data are missing randomly, pairwise deletion is only practical for statistical analyses where a correlation matrix can be analyzed, e.g. OLS regression. It does not work for techniques like logistic regression. Even in the case of OLS, statistical software may or may not make it easy to use pairwise deletion, e.g. it is easy to do in SPSS but requires extra effort if you want to do it Stata.

For these and other reasons, pairwise deletion is not widely used or recommended. I would probably feel most comfortable with it in cases where only a random subset of the sample had been asked some questions while other questions had been answered by everyone, such as in the Census. (Of course, some of the same problems that make pairwise deletion problematic may also make other alternatives problematic as well; one of the worst things about pairwise deletion may be that it can cause you to miss the fact that problems are present.)

- *Nominal variables: Treat missing data as just another category.* Suppose the variable Religion is coded 1 = Catholic, 2 = Protestant, 3 = Other. Suppose some respondents fail to answer this question. Rather than just exclude these subjects, we could just set up a fourth category, 4 = Missing Data (or no response). We could then proceed as usual, constructing three dummy variables from the four category variable of religion. This method has been popular for years – but according to Allison & others, it produces biased estimates.
- *Substituted (plugged in) values, i.e. (Single) Imputation.* A common strategy, particularly if the missing data are not too numerous, is to substitute some sort of plausible guess [imputation] for the missing data. Common choices include:
 - ✓ The overall mean
 - ✓ An appropriate subgroup mean (e.g. the mean for blacks or for whites)
 - ✓ A regression estimate (i.e. for the non-MD cases, regress X on other variables. Use the resulting regression equation to compute X when X is missing)

In its documentation for the `impute` command (see examples below), the Stata 8 reference manual states that “[imputation] is not the only method for coping with missing data, but it is often much better than deleting cases with any missing data, which is the default.” However, many disagree with this. Note that these strategies tend to reduce variability and can artificially increase R^2 and decrease standard errors. Also, remember that cases may be missing precisely because they are atypical, hence plugging in “typical” values may not be wise. According to Allison, “All of these imputation methods suffer from a fundamental problem: Analyzing imputed data as though it were complete data produces standard errors that are underestimated and test statistics that are overestimated. Conventional analytic techniques simply do not adjust for the fact that the imputation process involves uncertainty about the missing values.”

- *Substituted (plugged in) value plus missing data indicator.* Allison calls this “Dummy variable adjustment”. As is explained below, this strategy is no longer recommended, except under certain conditions – in spite of what my old course notes and exams may say! This strategy proceeds as follows:
 - ✓ Plug in some arbitrary value for all MD cases (typically 0, or the variable's mean)
 - ✓ Include in the regression a dummy variable coded 1 if data in the original variable was missing (i.e. a value has been plugged in for MD), 0 otherwise.

Perhaps surprisingly, whatever value is substituted for MD, the metric coefficient will not be affected, and will be an unbiased estimate of the population b .

Although any arbitrary value can be used, it is particularly convenient to use the mean when “plugging in” values. The t-test of the coefficient for the missing data dichotomy then (supposedly) indicates whether or not data are missing at random.

According to Cohen and Cohen (1975), the advantages of this approach are

- ✓ avoid the risk of non-representativeness in dropping subjects if data are missing nonrandomly
- ✓ avoid loss of statistical power due to reduced N even if data are missing randomly
- ✓ capitalize on the information inherent in the absence-presence of values on the variable in question (e.g. are missing cases different from non-missing)
- ✓ capitalize on the information present on other variables although missing for some subjects on the variable in question

HOWEVER, while this technique has been used for many years (including in this class) Allison and others have recently been critical of it. Allison calls this technique “remarkably simple and intuitively appealing.” But unfortunately, “the method generally produces biased estimates of the coefficients.” See his book for examples. In the 2003 edition of their book, Cohen and Cohen no longer advocate missing data dummies and acknowledge that they have not been widely used.

NOTE!!! Buried in footnote 5 of Allison’s book is a very important footnote that is often overlooked (Thanks to Richard Campbell from Illinois-Chicago for pointing this out to me):

While the dummy variable adjustment method is clearly unacceptable when data are truly missing, it may still be appropriate in cases where the unobserved value simply does not exist. For example, married respondents may be asked to rate the quality of their marriage, but that question has no meaning for unmarried respondents. Suppose we assume that there is one linear equation for married couples and another equation for unmarried couples. The married equation is identical to the unmarried equation except that it has (a) a term corresponding to the effect of marital quality on the dependent variable and (b) a different intercept. It’s easy to show that the dummy variable adjustment method produces optimal estimates in this situation.

So, for example, you might have questions about mother’s education and father’s education, but the father is unknown or was never part of the family. Or, you might have spouse’s education, but there is no spouse. In such situations, the dummy variable adjustment method may be appropriate. Conversely, if there is a true value for father’s education but it is missing, Allison says the dummy variable adjustment method should not be used.

- **Advanced methods: Maximum Likelihood Estimation and Multiple Imputation.** Allison concludes that, of the conventional methods listed above, listwise deletion often works the best. However, he argues that, under certain conditions, Maximum Likelihood Methods and Multiple Imputation Methods can work better. As Newman (2003, p. 334) notes, “MI [multiple imputation] is a procedure by which missing data are imputed several times (e.g. using regression imputation) to produce several different complete-data estimates of the parameters. The parameter estimates from each imputation are then combined to give an overall estimate of the complete-data parameters as well as reasonable estimates of the standard errors.” Maximum Likelihood (ML) approaches “operate by estimating a set of parameters that maximize the probability of getting the data that was observed” (Newman, p.

332).

If time permits, I will give some simple examples showing how to use Patrick Royston's *ice* program for multiple imputation. You should read the program documentation carefully, as well as do additional reading on multiple imputation, if you want to use these techniques.

III. Simple Example. Here is an example from a previous exam:

1. A researcher collected the following data:

| Case # | Y | X1 | X2 | X3 |
|--------|----|---------|---------|---------|
| 1 | 30 | 2 | Missing | 12 |
| 2 | 37 | 2 | 1 | Missing |
| 3 | 41 | 3 | 1 | 20 |
| 4 | 42 | 1 | Missing | 16 |
| 5 | 45 | 3 | 2 | Missing |
| 6 | 49 | 1 | 2 | 27 |
| 7 | 51 | Missing | 1 | 30 |
| 8 | 55 | 3 | 2 | 33 |
| 9 | 58 | Missing | 2 | 19 |
| 10 | 60 | 2 | Missing | 24 |

a. Suppose the researcher believes that data are missing on a *random* basis, i.e. those who did not respond are no different than those who did. What would you recommend for her—pairwise deletion of missing data, or listwise deletion? Why?

Listwise deletion would result in 70% of the cases being deleted. Because data are missing randomly and MD is spread across several variables, pairwise deletion might be a reasonable option in this case. Still, I would probably do further examination to find out why so many cases were missing some data, i.e. I would want to be confident that the data really are missing randomly. This might occur in situations where, say, only random subsamples are asked some questions, such as in the short and long forms of the Census questionnaire.

b. Suppose the researcher believes that data may be missing on a *non-random* basis. What would you recommend for her—substitution of the mean for MD cases, or substitution of the mean plus including missing data dichotomies? Why?

In the past I recommended using the Cohen and Cohen method: substitute the mean for the MD cases, and then add a missing data dichotomy. A significant coefficient for the dichotomy supposedly indicated that data were missing on a non-random basis. That method has now been discredited however. The researcher probably needs to

better understand the reasons data are missing before deciding on a strategy. For example, are data missing because the question was not appropriate for the respondent (e.g. questions about marital satisfaction should not be asked of people who are not married)? Are they missing because some subjects refused to talk about sensitive topics? Were there problems with the questionnaire or with the data collection?

However, if the data are missing because they are non-existent (e.g. the question pertains to the spouse but there is no spouse) the dummy variable adjustment method may be appropriate.

[NOTE: In the interest of time, I may just let you go over most of sections IV & V on your own.]

IV. Using SPSS to handle missing data – traditional methods

First, a caution: While I am going to show you how to implement various methods using SPSS and Stata, in most cases you may be as well off or better off just using listwise deletion or dropping highly problematic variables that have a lot of MD. If you feel your missing data problems are extremely severe, you should consider using more advanced techniques than what we discuss here. In some cases, other software, such as LISREL, may better meet your needs.

A second caution: When using SPSS, Stata, or any program, be careful about permanently overwriting your original data. If you are going to “plug in” values for missing data, you may want to first create a copy of the original variable and then work on it.

A third caution: If you are only analyzing a subsample of the data (e.g. women only) you want to be careful that your “plugged in” values are not computed from the entire sample. In either SPSS or Stata, you may want to create an extract with only the cases you want first, or otherwise control the sample selection that is being used. In general, when manipulating your data, run checks to make sure things are coming out the way you wanted them too!!!

SPSS has an added-cost routine specifically designed to examine missing data. I haven't seen it, but it sounds interesting. Using more traditional SPSS features:

To assign the mean value to a variable:

- First, determine the mean of the variable, e.g. have something like
DESCRIPTIVES VARIABLES=VAR01
- Then, do something like

RECODE VAR01 (MISSING, SYSMIS = 32).

To assign a subgroup mean:

- First, determine the subgroup mean, e.g. have something like
MEANS TABLES=VAR01 BY RACE
- Then, do something like

IF (RACE = 1 AND MISSING(VAR01)) VAR01 = 29.

IF (RACE = 2 AND MISSING(VAR01)) VAR01 = 33.

To do mean substitution and create an MD indicator:

- Determine the mean
- Then, do something like

```
DO IF (MISSING(VAR01)).  
    COMPUTE MD01 = 1.  
    COMPUTE VAR01 = 32.  
ELSE.  
    COMPUTE MD01 = 0.  
END IF.
```

To substitute a regression estimate for the mean: Run a regression where your IV is the dependant variable. Then, using the beta coefficients, do something like

```
IF (MISSING(VAR01)) VAR01 = 2X1 + 3X2 + 7.
```

To control whether SPSS Regression uses listwise, pairwise, or mean substitution: On the regression card, use the MISSING subcommand. Here is the SPSS documentation.

MISSING Subcommand

MISSING controls the treatment of cases with missing values. By default, a case that has a user-missing or system-missing value for any variable named or implied on **VARIABLES** is omitted from the computation of the correlation matrix on which all analyses are based.

- The minimum specification is a keyword specifying a missing-value treatment.

LISTWISE *Delete cases with missing values listwise.* Only cases with valid values for all variables named on the current **VARIABLES** subcommand are used. If **INCLUDE** is also specified, only cases with system-missing values are deleted listwise. **LISTWISE** is the default if the **MISSING** subcommand is omitted.

PAIRWISE *Delete cases with missing values pairwise.* Each correlation coefficient is computed using cases with complete data for the pair of variables correlated. If **INCLUDE** is also specified, only cases with systemmissing values are deleted pairwise.

MEANSUBSTITUTION *Replace missing values with the variable mean.* All cases are included and the substitutions are treated as valid observations. If **INCLUDE** is also specified, user-missing values are treated as valid and are included in the computation of the means.

INCLUDE *Includes cases with user-missing values.* All user-missing values are treated as valid values. This keyword can be specified along with the methods **LISTWISE**, **PAIRWISE**, or **MEANSUBSTITUTION**.

Example

```
REGRESSION VARIABLES=POP15, POP75, INCOME, GROWTH, SAVINGS  
/DEPENDENT=SAVINGS  
/METHOD=STEP  
/MISSING=MEANSUBSTITUTION.
```

- System-missing and user-missing values are replaced with the means of the variables when the correlation matrix is calculated.

V. Using Stata to handle missing data – traditional methods

Again, I preface my comments by saying that you generally don't want to use most of these methods! As far as traditional methods go, listwise deletion tends to work as well or better as anything else.

Some things are easier to do in Stata than in Spss. While there are many ways to compute new variables with corrections for missing data, I find that the `impute` command is very handy. The basic syntax for `impute` is

```
impute depvar varlist [weight] [if exp] [in range], generate(newvar1)
```

The `generate` parameter creates a variable called `newvar1` (you can call it whatever you want). If the original variable (`depvar`) is not missing, `newvar1` = the original value. If `depvar` is missing, `newvar1` is set equal to a regression estimate computed using the vars in `varlist`. That is, `depvar` is regressed on `varlist`. If some of the vars in `varlist` themselves have missing data, the regression estimate will be based only on the nonmissing variables. If `depvar` and all the vars in `varlist` are missing, `newvar1` will also be missing, otherwise it will have a value.

First, here are some summary statistics for the data set I am using. As you can see, 95 cases are missing on `educ`, and the rest have complete data.

```
. use http://www.nd.edu/~rwilliam/stats2/statafiles/md.dta, clear
```

```
. sum
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-----|----------|-----------|-----|------|
| income | 500 | 27.79 | 8.973491 | 5 | 48.3 |
| educ | 405 | 13.01728 | 3.974821 | 2 | 21 |
| jobexp | 500 | 13.52 | 5.061703 | 1 | 21 |
| black | 500 | .2 | .4004006 | 0 | 1 |
| other | 500 | .1 | .3003005 | 0 | 1 |
| white | 500 | .7 | .4587165 | 0 | 1 |
| race | 500 | 1.4 | .6639893 | 1 | 3 |

To assign the mean value to a variable:

You could also do it pretty much the same way as you do with SPSS:

```
. gen xeduc = educ
(95 missing values generated)

. replace xeduc = 13.01728 if missing(educ)
(95 real changes made)

. sum educ xeduc
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-----|----------|-----------|-----|-----|
| educ | 405 | 13.01728 | 3.974821 | 2 | 21 |
| xeduc | 500 | 13.01728 | 3.576498 | 2 | 21 |

Here is how we can do it with the `impute` command:

```
. gen one = 1
. impute educ one, gen(xeduc)
19.00% (95) observations imputed
. sum educ xeduc
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-----|----------|-----------|-----|-----|
| educ | 405 | 13.01728 | 3.974821 | 2 | 21 |
| xeduc1 | 500 | 13.01728 | 3.576498 | 2 | 21 |

In this case, `educ` is regressed only on a constant, yielding a predicted value equal to the mean of `educ`. Hence, `xeduc = educ` when `educ` is not missing, `xeduc = the mean of educ` when `educ` is missing. In other words, the 95 missing cases all got assigned a value of 13.01728 on `xeduc`. As you see, however you do it, `educ` and `xeduc` have the same mean, but `xeduc1` has no missing cases. The standard deviation declines because there is no variability in the plugged-in values.

To assign a subgroup mean:

The `tab` command can show us what the subgroup means are, and we can fill in the subgroup means from there. This will work, but it is more tedious and probably more error prone than using `impute`:

```
. tab race, sum(educ)
```

| race | Summary of educ | | Freq. |
|-------|-----------------|-----------|-------|
| | Mean | Std. Dev. | |
| 1 | 14.072202 | 3.5997967 | 277 |
| 2 | 9.9302326 | 4.3865857 | 86 |
| 3 | 12.380952 | .79487324 | 42 |
| Total | 13.017284 | 3.9748214 | 405 |

```
. gen xeduc2 = educ
(95 missing values generated)

. replace xeduc2 = 14.072202 if race==1 & missing(xeduc2)
(73 real changes made)
```

```
. replace xeduc2 = 9.9302326 if race==2 & missing(xeduc2)
(14 real changes made)
```

```
. replace xeduc2 = 12.380952 if race==3 & missing(xeduc2)
(8 real changes made)
```

```
. tab race, sum(xeduc2)
```

| race | Summary of xeduc2 | | |
|-------|-------------------|-----------|-------|
| | Mean | Std. Dev. | Freq. |
| 1 | 14.072202 | 3.2012515 | 350 |
| 2 | 9.9302326 | 4.0646063 | 100 |
| 3 | 12.380952 | .72709601 | 50 |
| Total | 13.074683 | 3.6365787 | 500 |

Using the impute command is simpler and safer:

```
. impute educ black white other, gen(xeduc2)
```

```
19.00% (95) observations imputed
```

```
. tab race, sum(xeduc2)
```

| race | Summary of imputed educ | | |
|-------|-------------------------|-----------|-------|
| | Mean | Std. Dev. | Freq. |
| 1 | 14.072202 | 3.2012515 | 350 |
| 2 | 9.9302326 | 4.0646063 | 100 |
| 3 | 12.380952 | .72709601 | 50 |
| Total | 13.074683 | 3.6365787 | 500 |

As you see, with either approach, the subgroup means are identical to before, but there are no missing cases. Each missing case had the mean value for its racial subgroup plugged in.

Incidentally, if you didn't already have dichotomies computed for race, and just had a multi-category race var, the `xi` command could compute the dummies for you:

```
. xi: impute educ i.race, gen(xeduc2b)
```

```
i.race      _Irace_1-3      (naturally coded; _Irace_1 omitted)
```

```
19.00% (95) observations imputed
```

```
. tab race, sum(xeduc2b)
```

| race | Summary of imputed educ | | |
|-------|-------------------------|-----------|-------|
| | Mean | Std. Dev. | Freq. |
| 1 | 14.072202 | 3.2012515 | 350 |
| 2 | 9.9302326 | 4.0646063 | 100 |
| 3 | 12.380952 | .72709601 | 50 |
| Total | 13.074683 | 3.6365787 | 500 |

To do mean substitution and create an MD indicator:

After using one of the above methods to substitute the mean, you could do something like this:

```
. gen md = 0
```

```
. replace md = 1 if xeduc!=educ
```

If the original variable does not equal the imputed variable, that means a value was plugged in for missing cases. In such cases, $md = 1$. If educ does equal xeduc3, then no value was plugged in, and $md = 0$.

Again, if the data are missing because they are non-existent, rather than missing because values exist but are unknown, this could be a good method.

To substitute a regression estimate for the mean:

Just specify whatever vars you want to base your regression estimate (just be careful not to use the Y variable):

```
. impute educ jobexp black other white, gen(xeduc4)
19.00% (95) observations imputed
```

```
. sum xed*
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-----|----------|-----------|-----|-----|
| xeduc1 | 500 | 13.01728 | 3.576498 | 2 | 21 |
| xeduc2 | 500 | 13.07468 | 3.636579 | 2 | 21 |
| xeduc2b | 500 | 13.07468 | 3.636579 | 2 | 21 |
| xeduc3 | 500 | 13.01728 | 3.576498 | 2 | 21 |
| xeduc4 | 500 | 13.08214 | 3.659779 | 2 | 21 |

```
. tab race, sum(xeduc4)
```

| race | Summary of imputed educ | | |
|-------|-------------------------|-----------|-------|
| | Mean | Std. Dev. | Freq. |
| 1 | 14.071926 | 3.2328681 | 350 |
| 2 | 9.9475566 | 4.0879436 | 100 |
| 3 | 12.422792 | .8384331 | 50 |
| Total | 13.082139 | 3.6597795 | 500 |

In this particular example, jobexp is not that strongly related to education, hence including it as one of the predictors of education did not have much of an effect on the estimated values over and above what we got when we just used the subgroup differences in the means.

Again, keep in mind that the `impute` command will form a regression estimate based on all the nonmissing variables in varlist. So, for example, if a case was missing on both educ and jobexp, the imputation would be based on the regression of educ on black, other and white for all cases that were not missing on those variables. Unless a case is missing data on all the variables specified, `impute` will give a non-missing value for the imputed variable.

To control whether Stata Regression uses listwise, pairwise, or mean substitution:

Stata uses listwise deletion. As far as I know, there is no straightforward way to use pairwise deletion (if you desperately wanted it, I suppose you could compute the pairwise correlations and then use the `corr2data` command to create a data set with the desired correlations). If you want to do mean substitution, you'd have to compute the vars yourself, using methods like those described above.