CPH 931 — Fall 2008 — Dr. Charnigo

Lecture 5

Handling missing data

Introduction. In Lectures 1 through 4 we have discussed how to handle various problems preventing successful application of ordinary least squares in fitting a linear regression model. Some of our solutions have entailed using an estimation framework other than ordinary least squares (weighted least squares for heteroscedasticity, ridge regression for multicollinearity, robust regression for nonnormal errors or outliers), while some have entailed using a more general statistical model (nonparametric regression model when the mean response is not a linear function, linear mixed model when the errors are correlated due to repeated measurements).

Today we consider another problem, that of missing data. This problem is not one of statistical assumptions being violated. Rather, missing data can arise for one of the following reasons.

1. Some subjects may refuse to provide sensitive information.

2. Some subjects may drop out of a study.

3. A researcher may find a gross error in a database, but correcting the gross error may not be possible because the researcher no longer has contact with the subject. Thus, the gross error is effectively replaced by a missing value.

4. A researcher may not be able to acquire data on all variables for some subjects. This can happen if the researcher is acquiring data from preexisting medical records rather than contemporaneously as part of his/her study.

Types of missing data. Consider a data set with a variable V that has some missing values and variables $U_1, \ldots, U_k$ that do not have any missing values.

Data are missing completely at random (MCAR) if the probability of observing V does not depend on the true value of V — assumed to exist whether or not we actually observe it — or on the values of $U_1, \ldots, U_k$:

$$\mathbb{P}(\text{observing } V \mid V = v, U_1 = u_1, \ldots, U_k = u_k) = \text{constant}.$$

Data are missing at random (MAR) if the probability of observing V does not depend on the true value of V after we control for $U_1, \ldots, U_k$:

$$\mathbb{P}(\text{observing } V \mid V = v, U_1 = u_1, \ldots, U_k = u_k) = f(u_1, \ldots, u_k).$$

Data are missing not at random (MNAR) if the probability of observing V depends on the true value of V even after we control for $U_1, \ldots, U_k$:

$$\mathbb{P}(\text{observing } V \mid V = v, U_1 = u_1, \ldots, U_k = u_k) = f(v, u_1, \ldots, u_k).$$

Illustrating types of missing data. Suppose that we survey patients at a medical practice as part of an investigation on factors relating to patients' exercise habits. Not ruling out that socioeconomic factors may relate to patients' exercise habits, one question we may ask of patients is how much money they make.

However, some patients may regard this as sensitive information. In particular, suppose that 30% of "wealthy" patients decline to respond, 20% of "middle class" patients decline to respond, and 40% of "poor" patients decline to respond. (The exact definitions of wealthy, middle class, and poor are unimportant here, except that we are using these terms to characterize income rather than assets accumulated over time.)

Under these suppositions, the income data are not MCAR: the probability of observing a patient's income depends on the patient's true income because this probability is greater for a middle class patient than for a wealthy patient or a poor patient.

However, if there are other items on the survey — such as occupation, educational attainment, and manner of payment for medical services — that are strongly related to income, then the income data may be MAR: the probability of observing a patient's income may no longer depend on the patient's true income once we take into account the patient's occupation, educational attainment, and manner of payment for medical services.

On the other hand, if there are no other items on the survey that are strongly related to income, then the income data are MNAR.

Implications for statistical inference. By now you have encountered many statistical models, but there is almost always a common theme to the statistical inference: we want to quantify population-level relationships between a response variable and one or more explanatory variables.

When the data are MNAR, the relationships between V and $U_1, \ldots, U_k$ may differ between the subpopulation for which V is observable and the subpopulation for which V is not observable. Since we can only observe V in members of the first subpopulation, quantifying population-level relationships is problematic. If we fit a statistical model using only those records without missing values, then we are quantifying relationships not in the full population but in the subpopulation for which V is observable.

One approach to this problem if V is categorical, or if we are willing to discretize a continuous V, is to imagine that V has another category — "I refuse to specify" — and treat the missing values as if the subjects had responded "I refuse to specify". Then we can fit a statistical model using all of the records. This approach makes sense when missing values occur due to subjects' refusals to provide sensitive information.

More sophisticated approaches for analyzing MNAR data exist, but they are cumbersome. In particular, they may entail explicit estimation of

$\mathbb{P}(\text{observing } V \mid V = v, U_1 = u_1, \ldots, U_k = u_k)$ as a function of $v, u_1, \ldots, u_k$. Thus, we may prefer trying to avoid MNAR data altogether by collecting information on many variables. Even though having MCAR data is often too much to hope for, the data may be MAR if we are sufficiently broad in our collection of information.

When the data are MAR, the relationships between V and $U_1, \ldots, U_k$ cannot differ between the subpopulation for which V is observable and the subpopulation for which V is not observable. The marginal distributions of V and $U_1, \ldots, U_k$ most certainly can differ between such subpopulations, but the relationships between V and $U_1, \ldots, U_k$ cannot. Thus, quantifying population-level relationships is possible without explicitly estimating $\mathbb{P}(\text{observing } V \mid V = v, U_1 = u_1, \ldots, U_k = u_k)$, although care is required.

For the rest of this lecture, I assume that the data are MAR.

Motivating example. In {Mercury.xls} I have a modified version of the data from {http://lib.stat.cmu.edu/DASL/Datafiles/MercuryinBass.html}. The website's description of the original data set is as follows.

"Largemouth bass were studied in 53 different Florida lakes to examine the factors that influence the level of mercury contamination. Water samples were collected from the surface of the middle of each lake in August 1990 and then again in March 1991. The pH level, the amount of chlorophyll, calcium, and alkalinity were measured in each sample. The average of the August and March values were used in the analysis. Next, a sample of fish was taken from each lake with sample sizes ranging from 4 to 44 fish. The age of each fish and mercury concentration in the muscle tissue was measured. (Note: Since fish absorb mercury over time, older fish will tend to have higher concentrations). Thus, to make a fair comparison of the fish in different lakes, the investigators used a regression estimate of the expected mercury concentration in a three year old fish as the standardized value for each lake. Finally, in 10 of the 53 lakes, the age of the individual fish could not be determined and the average mercury concentration of the sampled fish was used instead of the standardized value."

The modification I made to {Mercury.xls} was to set the 3-year standard mercury values to missing for the 10 lakes in which the ages of the individual fish could not be determined. Thus, there are missing values on 18.9% of the records, which is slightly greater than the $10\% - 15\%$ at which I would begin to feel uncomfortable confining attention to records without missing values.

On page 1 of {Mercury.rtf} are results for a linear regression model fit by ordinary least squares, using only the 43 observations with non-missing 3-year standard mercury values. The response variable is a log-transformed version of 3-year standard mercury, hereafter denoted Y. The sole explanatory variable is the lake's pH level, hereafter denoted X. Page 2 shows a scatterplot of Y values against X values along with the fitted regression line $\widehat{y} = 1.53885 - 0.35773 \cdot x$. Note that $R^2 = 0.3818$.

Single imputation. Suppose that I ask you to guess the value of Y for Lake Annie in our motivating example. One possibility, based on "imputation by mean substitution", is the sample mean of Y, which is -0.82713.

However, such a guess ignores the potentially relevant information that the value of X for Lake Annie is 5.1. A better guess, based on "imputation by regression", is $1.53885 - 0.35773 \cdot 5.1 = -0.28557$.

We can similarly impute a value of Y for each of the other nine lakes in which the ages of the individual fish could not be determined. This is referred to as "single imputation" because we are making a single number best guess for the value of Y wherever it is missing.

Note that, in this example, the variable on which missing values occur is the response variable (i.e., $V = Y$, $k = 1$, and $U_1 = X$). If missing values had occurred on the explanatory variable, then to perform single imputation we could have employed an auxiliary linear regression model in which the roles of X and Y were reversed.

Consequences of single imputation. On page 3 of {Mercury.rtf} are results for a linear regression model fit by ordinary least squares, using all 53 observations and treating the singly imputed values as if they had been observed in the first place. Page 4 provides the corresponding scatterplot.

Discussion questions. As was the case last week, we can benefit more from the discussion questions if they are in the middle of the lecture than if they are at the end.

1. What has happened to the estimates of the intercept and slope? How can we explain this?
2. What has happened to R^2? How can we explain this?
3. If published, why would the results on page 3 be misleading?

Another difficulty with single imputation. Aside from the problem identified in the third discussion question, another difficulty with single imputation is that many data sets will have more than one variable with missing values.

Consider a data set with variables V_1 and V_2 that have some missing values — not necessarily on the same subjects — and variables $U_1, \ldots, U_k$

that do not have any missing values. (Although I do not state them explicitly, the definitions of MCAR, MAR, and MNAR can be generalized to this situation; we continue to assume that the data are MAR.) What if we wish to impute values for V_1 and V_2?

Mean substitution is possible, but this may be undesirable since $U_1, \ldots, U_k$ may provide clues on the missing values of V_1 and V_2.

Imputation by regressing V_1 on $U_1, \ldots, U_k$ and V_2 on $U_1, \ldots, U_k$ is possible, but what if V_1 and V_2 are correlated? Then not using V_2 to impute values for V_1 is failing to use clues provided by V_2 on the missing values of V_1 (and vice versa). On the other hand, we cannot use V_2 to impute values for V_1 on subjects for whom the values of V_2 are also missing. Consequently, imputation by regression seems piecemeal and haphazard.

Expectation maximization. The difficulty noted above can be overcome by using an "expectation maximization" (EM) algorithm to estimate the means, variances, and covariances of $V_1, V_2, U_1, \ldots, U_k$. Once these means, variances, and covariances are estimated, we can make our best guesses for the missing values using all available clues in each instance.

A technical assumption for the EM algorithm is that $V_1, V_2, U_1, \ldots, U_k$ have a multivariate normal distribution. Since that is only possible if $V_1, V_2, U_1, \ldots, U_k$ are individually normally distributed, technically the EM algorithm is not valid if one or more of the variables in the data set is dichotomous. Still, people often use the EM algorithm even if some of the variables in the data set are dichotomous. However, if V_1 (or V_2) is dichotomous and coded as (say) 0 or 1, then the best guesses for the missing values of V_1 must be rounded.

Multiple imputation. Returning to page 4 of {Mercury.rtf}, the situation would seem less unsatisfactory if the imputed values did not all fall on the fitted regression line. Indeed, let us imagine randomly perturbing the imputed values, moving some above the fitted regression line and some below.

In general, rather than working with best guesses for the missing values of V (or of V_1 and V_2), we can work with random perturbations of best guesses. Then we are not artificially diminishing the uncertainty associated with V. A method called "Markov Chain Monte Carlo" (MCMC) uses the output from the EM algorithm to generate random perturbations of best guesses. Again, if V is dichotomous, then the random perturbations of best guesses must be rounded.

Page 5 of {Mercury.rtf} shows ("EM (Posterior Mode) Estimates") the estimated means, variances, and covariance for Y and X based on the EM algorithm.

Pages 7 through 14 show five "complete" data sets generated via MCMC. Values of Y that were not missing have not been altered. Values of Y that were missing have been filled in with random perturbations of best guesses. In the first complete data set, the filled-in value of Y for Lake Annie (record 2) is 0.23114. In the second complete data set, the filled-in value of Y for Lake Annie (record 55) is -0.23191. The corresponding filled-in values from the other three complete data sets are -0.28748, -0.61005, and 0.66434.

The process of creating multiple complete data sets in this manner is called multiple imputation.

Combining results from complete data sets. Once we have performed multiple imputation, we can fit our statistical model using each of the complete data sets. Pages 15 through 19 of {Mercury.rtf} show the results from the five complete data sets for fitting a linear regression model with Y as the response variable and X as the sole explanatory variable. Each set of parameter estimates is slightly different, since the random perturbations used to create each complete data set were different.

Of course, we do not want to report five different estimates for the intercept and five different estimates for the slope. We need to somehow combine the five sets of parameter estimates into "overall" estimates for the intercept and slope. The intuitive choice — to simply average over the complete data sets — is in fact correct. So, we average 1.31402, 1.86004, 2.23131, 1.61610, 1.50964 to get 1.706221. And, we average -0.32436, -0.40643, -0.46546, -0.37073, -0.35121 to get -0.383640.

The overall estimates 1.706221 and -0.383640 are shown on page 20 of {Mercury.rtf}, along with standard errors calculated by

$$\sqrt{W + (m+1)B/m},$$

where

$$W := \sum_{j=1}^{m} \left\{ se(\widehat{\theta}_j) \right\}^2 / m,$$

$$B := \sum_{j=1}^{m} \left(\widehat{\theta}_j - \sum_{k=1}^{m} \widehat{\theta}_k / m \right)^2 / (m-1),$$

$\widehat{\theta}_j$ is the estimate of θ (generic notation for a parameter such as an intercept or slope) based on the j^{th} complete data set, $se(\widehat{\theta}_j)$ is the corresponding standard error, and m is the number of complete data sets. Confidence intervals and p-values are also reported, using a T reference distribution on degrees of freedom calculated by

$$(m-1)\left(1 + \frac{1}{(1+1/m)(B/W)}\right)^2.$$

Closing remarks on multiple imputation. A few remarks are in order.

1. Although the choice to create $m = 5$ complete data sets may seem arbitrary, this is accepted practice.

2. Page 20 of {Mercury.rtf} shows that we did not gain much from multiple imputation in this example. The coefficient estimates and standard errors are not much different from what they were in the original analysis based on records without missing values. Thus, the original analysis appears to have been adequate in this example — at least to the extent that the expected value of log 3-year standard mercury is believed to be linear in the lake's pH level.

3. Multiple imputation can also be used when there are missing values on one or more explanatory variables and when there are missing values on both the response variable and one or more explanatory variables. My past experiences suggest that multiple imputation is more likely to be helpful in these situations than when the missing values occur solely on the response variable.

4. While today's lecture sought to provide a tool for handling missing data in a linear regression seting, multiple imputation can be used with other (parametric) statistical models, such as logistic regression and linear mixed models.