

# Handling Missing Data for Regression Models

**JEFF GILL**

Center for Applied Statistics  
Department of Political Science  
Washington University, St. Louis

## Background

▶ Goal:  $f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta})$  unbiased.

▶ Define:

$$\mathbf{Z}_{mis} = (X_{mis}, Y_{mis})$$

$$\mathbf{Z}_{obs} = (X_{obs}, Y_{obs})$$

▶ We stipulate a  $n \times k$  matrix,  $\mathbf{R}$ , corresponding to  $\mathbf{X}$  that contains 0 when the  $\mathbf{X}$  matrix data value is *not* missing, and 1 when it is missing.

▶  $\phi$  is a parameter that “causes” missingness.

▶ Standard Terms from Rubin:

<b>MCAR</b>	$p(\mathbf{R} \mathbf{Z}_{obs}, \mathbf{Z}_{mis}) = p(\mathbf{R} \phi)$	missingness not related to observed or unobserved
<b>MAR</b>	$p(\mathbf{R} \mathbf{Z}_{obs}, \mathbf{Z}_{mis}) = p(\mathbf{R} \mathbf{Z}_{obs}, \phi)$	missingness depends only on observed data
<b>Non-Ignorable</b>	$p(\mathbf{R} \mathbf{Z}_{obs}, \mathbf{Z}_{mis}) = p(\mathbf{R} \mathbf{Z}_{obs}, \mathbf{Z}_{mis}, \phi)$	missingness depends on unobserved data

## Background

- ▶ For the data matrix  $\mathbf{D}$ , if  $\mathbf{M}$  is a dichotomous indicator matrix valued 1 if a datum is missing and 0 if it is not, the missing data generating mechanism is described by  $p(\mathbf{M}|\mathbf{D})$  (Little and Rubin, 2002, p.12).
- ▶ So we can summarize the definitions according to this structure:

Assumption	Definition
MCAR	$p(\mathbf{M} \mathbf{D}) = p(\mathbf{M})$
MAR	$p(\mathbf{M} \mathbf{D}) = p(\mathbf{M} \mathbf{D}_R)$
NI	$p(\mathbf{M} \mathbf{D}) = p(\mathbf{M} \mathbf{D})$

## Why Case-wise Deletion is Evil

- ▶ Consider the computation of a mean  $\mu$  from data  $\mathbf{y}$  where some data are non-randomly missing.
- ▶ When  $\mu_R$  is the mean of respondents and  $\mu_M$  is the mean of missing data, we write the overall mean as:

$$\mu = \pi_R \mu_R + (1 - \pi_M) \mu_M.$$

where  $\pi_R$  is the *proportion* of observed responses.

- ▶ The bias produced by casewise deletion is the expected fraction of missing data times the difference in means for observed and missing data (Little and Rubin, 2002, p.43):

$$\mu_R - \mu = (1 - \pi_R)(\mu_R - \mu_M).$$

- ▶ In the special case of MCAR missingness,  $\mu_R = \mu_M$  and the statistic is unbiased, but this is commonly violated in the social sciences.

## Why Case-wise Deletion is Evil

- Preliminaries: we are interested in obtaining the posterior mode of an unknown  $k$ -dimensional  $\boldsymbol{\theta}$  coefficient vector, given an outcome variable vector  $\mathbf{y}$ , and an observed  $\mathbf{X}$  matrix of explanatory data values assumed to be distributed iid according to  $f(\mathbf{X}|\boldsymbol{\theta})$ .
- Normally we would then:
  - as a **Likelihoodist**: find  $\boldsymbol{\theta}$  that maximizes  $\ell(\boldsymbol{\theta}|\mathbf{X})$ .
  - as a **Bayesian**: produce a posterior distribution from  $\pi(\boldsymbol{\theta}|\mathbf{X}) \propto p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})$ .

Neither of these approaches are directly possible if there happen to be missing data in  $\mathbf{X}$  unless the data are considered “missing completely at random,”

## Why Case-wise Deletion is Evil

- First, segment the  $\mathbf{X}$ -matrix into two constituent parts:  $\mathbf{X} = [\mathbf{X}_{obs}, \mathbf{X}_{mis}]$ , and restate the distribution function:

$$f(\mathbf{X}|\boldsymbol{\theta}) = f(\mathbf{X}_{obs}, \mathbf{X}_{mis}|\boldsymbol{\theta}) = f(\mathbf{X}_{obs}|\boldsymbol{\theta})f(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta}).$$

- Now segment the log likelihood function into two distinct components:

$$\begin{aligned}\ell(\boldsymbol{\theta}|\mathbf{X}) &= \ell(\boldsymbol{\theta}|\mathbf{X}_{obs}, \mathbf{X}_{mis}) \\ &= \ell(\boldsymbol{\theta}|\mathbf{X}_{obs}) + \log f(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta})\end{aligned}$$

- Rearrange this form to create a statement with both unknowns collected on the right-hand side:

$$\ell(\boldsymbol{\theta}|\mathbf{X}_{obs}) = \ell(\boldsymbol{\theta}|\mathbf{X}_{obs}, \mathbf{X}_{mis}) - \log f(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta}).$$

## Why Case-wise Deletion is Evil

- We can average over this uncertainty by taking expectations with respect to  $\mathbf{X}_{mis}$  on both sides:

$$\begin{aligned} & \int \ell(\boldsymbol{\theta}|\mathbf{X}_{obs})f(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta})d\mathbf{X}_{mis} \\ &= \int \ell(\boldsymbol{\theta}|\mathbf{X}_{obs}, \mathbf{X}_{mis})f(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta})d\mathbf{X}_{mis} \\ & \quad - \int \log f(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta})f(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta})d\mathbf{X}_{mis}. \end{aligned}$$

- LHS simplifies back to  $\ell(\boldsymbol{\theta}|\mathbf{X}_{obs})$  because the integral ends up operating over just the isolated complete PDF for  $\mathbf{X}_{mis}$ :

$$\begin{aligned} & \int \ell(\boldsymbol{\theta}|\mathbf{X}_{obs})f(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta})d\mathbf{X}_{mis} \\ &= \ell(\boldsymbol{\theta}|\mathbf{X}_{obs}) \int f(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta})d\mathbf{X}_{mis} \\ &= \ell(\boldsymbol{\theta}|\mathbf{X}_{obs}). \end{aligned}$$

## Why Case-wise Deletion is Evil

- Now we have an expression based only the observed data that relates the obtainable likelihood to two quantities that can be manipulated.

$$\begin{aligned}\ell(\boldsymbol{\theta}|\mathbf{X}_{obs}) &= \int \ell(\boldsymbol{\theta}|\mathbf{X}_{obs}, \mathbf{X}_{mis})f(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta})d\mathbf{X}_{mis} \\ &\quad - \int \log f(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta})f(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta})d\mathbf{X}_{mis}.\end{aligned}$$

## Multiple Imputation

- ▶ 3 steps:
  - ▷ generate reasonable missing the data ( $Z_{mis}$ )  $m$  times to get  $m$  replicate datasets,
  - ▷ analyze/regress each dataset separately,
  - ▷ combine results with summary process.
- ▶ Imputation step assumes a conditional distribution for the missing data conditioning on observed values.
- ▶ Oddly enough  $m = 5$  to  $10$  is sufficient.
- ▶ Combining process uses means for coefficients and an intuitive ANOVA approach for standard errors.

## Multiple Imputation in R

- ▶ Data From Greene, "Econometrics". National figures scaled by CPI already, 1968-1982.
- ▶ Use the `mice` package in R.
- ▶ `Amelia` has gotten better, but still has some suspect assumptions.

## Multiple Imputation in R

```
invest.df <- read.table("http://jgill.wustl.edu/data/investment.dat",header=TRUE)
invest.df
```

```
   real.investment trend real.gnp real.interest  inflation
1          0.161     1    1.058         5.16      4.40
2          0.172     2    1.088         5.87      5.15
3          0.158     3    1.086         5.95      5.37
4          0.173     4    1.122         4.88      4.99
5          0.195     5    1.186         4.50      4.16
6          0.217     6    1.254         6.44      5.75
7          0.199     7    1.246         7.83      8.82
8          0.163     8    1.232         6.25      9.31
9          0.195     9    1.298         5.50      5.21
10         0.231    10    1.370         5.46      5.83
11         0.257    11    1.439         7.46      7.40
12         0.259    12    1.479        10.28      8.64
13         0.225    13    1.474        11.77      9.31
14         0.241    14    1.503        13.42      9.44
15         0.204    15    1.475        11.02      5.99
```

## Multiple Imputation in R

```
##### FIRST DO A STANDARD LINEAR MODEL ON THE FULL DATA #####
```

```
invest.lm <- lm(real.investment ~ trend + real.gnp + real.interest + inflation,
               data=invest.df); summary(invest.lm)
```

```
Residuals:
```

	Min	1Q	Median	3Q	Max
	-0.0102779	-0.0022946	0.0004119	0.0029377	0.0080418

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-5.091e-01	5.513e-02	-9.234	3.28e-06
trend	-1.658e-02	1.972e-03	-8.409	7.59e-06
real.gnp	6.704e-01	5.500e-02	12.189	2.52e-07
real.interest	-2.326e-03	1.219e-03	-1.908	0.0854
inflation	-9.401e-05	1.347e-03	-0.070	0.9458

```
Residual standard error: 0.006714 on 10 degrees of freedom
```

```
Multiple R-Squared: 0.9724, Adjusted R-squared: 0.9614
```

```
F-statistic: 88.19 on 4 and 10 DF, p-value: 9.333e-08
```

## Multiple Imputation in R

```
##### NOW DELIBERATELY DELETE SOME DATA TO SIMULATE MISSINGNESS #####
```

```
invest.missing.df <- invest.df  
invest.missing.df[1,1] <- invest.missing.df[8,5] <- invest.missing.df [9,3] <-  
invest.missing.df[13,4] <- NA
```

```
sum(is.na(invest.missing.df))/prod(dim(invest.missing.df))  
[1] 0.05333333
```

```
# FROM THE lm HELP FILE:
```

```
# na.action: a function which indicates what should happen when the data  
# contain 'NA's. The default is set by the 'na.action' setting  
# of 'options', and is 'na.fail' if that is unset. The  
# 'factory-fresh' default is 'na.omit'.
```

```
#
```

```
# STRONGLY ADVISED: options(na.action="na.fail")
```

## Multiple Imputation in R

```
##### SECOND MODEL IS THE SAME SPECIFICATION BUT ON DATASET WITH MISSING #####
invest.missing.lm <- lm(real.investment ~ trend + real.gnp + real.interest
                        + inflation, data=invest.missing.df,na.action=na.omit)
summary.lm(invest.missing.lm)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.002759	-0.001341	-0.001181	0.001311	0.003809

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.4459603	0.0324644	-13.737	9.25e-06
trend	-0.0144363	0.0012779	-11.297	2.88e-05
real.gnp	0.6032710	0.0332504	18.143	1.80e-06
real.interest	-0.0031976	0.0007746	-4.128	0.00616
inflation	0.0023307	0.0009748	2.391	0.05395

Residual standard error: 0.002866 on 6 degrees of freedom

Multiple R-Squared: 0.9959, Adjusted R-squared: 0.9931

F-statistic: 362.5 on 4 and 6 DF, p-value: 2.79e-07

## Multiple Imputation in R

```
##### USE MULTIPLE IMPUTATION FROM THE MICE PROGRAM #####
```

```
library(mice)
```

```
# m=10 SPECIFIES 10 IMPUTATIONS (DEFAULT=5)
```

```
m <- 10
```

```
imp.invest <- mice(invest.missing.df,m)
```

```
# WE CAN NOW USE THE complete FUNCTION TEN TIMES FOLLOWED BY 10 lm CALLS
```

```
imp.invest.array <- array(NA,c(dim(invest.df),m,7))
```

```
for (i in 1:m) imp.invest.array[,,i] <- as.matrix(complete(imp.invest,i))
```

## Multiple Imputation in R

```
##### THE lm.mids FUNCTION DOES THIS IN ONE COMMAND #####
```

```
invest.mids <- lm.mids(real.investment ~ trend + real.gnp + real.interest  
  + inflation,data=imp.invest)  
pool(invest.mids)
```

Pooled coefficients:

(Intercept)	trend	real.gnp	real.interest	inflation
-0.387215446	-0.011858096	0.546190553	-0.005146479	0.003709159

Fraction of information about the coefficients missing due to nonresponse:

(Intercept)	trend	real.gnp	real.interest	inflation
0.1102147	0.1597954	0.1218865	0.1205782	0.1535351

## Multiple Imputation in R

```
# A FUNCTION TO CONVENIENTLY GET COEFFICIENTS AND STANDARD ERRORS FROM A lm.mids  
# OBJECT, param=1 GIVES THE COEFFICIENT ESTIMATES, param=2 GIVES THE STANDARD ERRORS
```

```
lm.mids.vals <- function(obj,param) {  
  out.mat <- NULL  
  for (i in 1:obj$call1$m)  
    out.mat <- rbind(out.mat,summary.lm(obj$analyses[[i]])$coef[,param])  
  out.mat  
}
```

```
# USE THIS FUNCTION TO GET THREE REQUIRED VECTOR:
```

```
impute.coef.vec <- apply(lm.mids.vals(invest.mids,1),2,mean)
```

(Intercept)	trend	real.gnp	real.interest	inflation
-0.387215446	-0.011858096	0.546190553	-0.005146479	0.003709159

## Multiple Imputation in R

```
# CALCULATE THE STANDARD TWO ERROR COMPONENTS BY HAND
```

```
( between.var <- apply(lm.mids.vals(invest.mids,1),2,var) )  
  (Intercept)          trend          real.gnp real.interest    inflation  
3.982399e-04  7.837802e-07  4.919964e-04  2.755307e-07  8.322247e-07
```

```
( within.var <- apply(lm.mids.vals(invest.mids,2)^2,2,mean) )  
  (Intercept)          trend          real.gnp real.interest    inflation  
3.536579e-03  4.533230e-06  3.898968e-03  2.210504e-06  5.047015e-06
```

```
# USE THESE TO COMPUTE THE FINAL STANDARD ERROR VECTOR
```

```
m <- 10
```

```
( impute.se.vec <- sqrt(within.var + ((m+1)/m)*between.var) )  
  (Intercept)          trend          real.gnp real.interest    inflation  
0.063044769  0.002322798  0.066634554  0.001585430  0.002441815
```

## Multiple Imputation in R

```
# THE DEGREES OF FREEDOM FOR THE T-STATISTIC NEEDS TO BE ADJUSTED.  
# SEE LITTLE AND RUBIN (1987), PAGE 257
```

```
( impute.df <- (m-1)*(1 + (1/(m+1))) * within.var/between.var)^2 )  
  (Intercept)          trend      real.gnp real.interest inflation  
    29.39766      20.95260      26.63908      26.91547  21.65926
```

```
# CREATE A REGRESSION TABLE:
```

```
out.table <- round( cbind( impute.coef.vec,impute.se.vec,  
  impute.coef.vec/impute.se.vec,  
  1-pt(abs(impute.coef.vec/impute.se.vec),impute.df) ),6 )  
dimnames(out.table) <- list( c("(intercept)","trend","real.gnp","real.interest",  
  "inflation"), c("Estimate","Std. Error","t value","Pr(>|t|)") )
```

## Multiple Imputation in R

```
out.table
```

	Estimate	Std. Error	t value	Pr(> t )
(intercept)	-0.390189	0.062981	-6.195350	0.000001
trend	-0.011984	0.002325	-5.153349	0.000025
real.gnp	0.549227	0.066607	8.245852	0.000000
real.interest	-0.005107	0.001576	-3.241617	0.001627
inflation	0.003673	0.002421	1.516784	0.071848

```
# NOT QUITE THE IMPROVEMENT WE HOPED FOR.  WHY?  TWO REASONS: (1) I PARTICULARLY  
# CHOSE THE FOUR MISSING VALUES TO "DO THE MOST DAMAGE", AND (2) THIS IS A  
# RELATIVELY SMALL DATASET SO MI DOES NOT HAVE A LOT OF INFORMATION WITH WHICH  
# TO BUILD THE POSTERIOR DISTRIBUTION OF THE MISSING DATA.
```

## Multiple Imputation in R

```
# OR PUT THE PIECES TOGETHER WITH ONE SUMMARY FUNCTION
summary(pool(invest.mids))
```

	est	se	t	df
(Intercept)	-0.387215446	0.063044769	-6.141912	8.880183
trend	-0.011858096	0.002322798	-5.105092	8.324535
real.gnp	0.546190553	0.066634554	8.196807	8.751254
real.interest	-0.005146479	0.001585430	-3.246110	8.765773
inflation	0.003709159	0.002441815	1.519017	8.395695

  

	Pr(> t )	lo 95	hi 95
(Intercept)	1.801275e-04	-0.530126490	-0.244304402
trend	8.169365e-04	-0.017178345	-0.006537846
real.gnp	2.157312e-05	0.394797063	0.697584043
real.interest	1.041540e-02	-0.008747632	-0.001545326
inflation	1.654781e-01	-0.001875817	0.009294134

## Some Text From a Recent Paper

- ▶ Missing data values are addressed here with *multiple imputation* (Little and Rubin 1983, 1987; Rubin 1987) using the `mice` (multiple imputation by chained equations) package in the `R` statistical environment. The commonly used methods of listwise deletion and mean imputation lead to biased and misleading results. Essentially, multiple imputation creates a posterior distribution for the missing data conditional on the observed data, and draws randomly from this distribution to create multiple replications (5-10) of the original dataset. The model analysis is performed on each of these replicates and then averaged (with a standard error adjustment). See King, *et al.* (2001) for a review of missing data issues in political science. All data, `R` code, `WinBUGS` code, and diagnostics to implement the statistical model described here, as well as our data imputations used in this analysis are available at the website:  
for replication purposes.

## Book References to Note

- ▶ Little, Roderick J. A., and Donald B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.
- ▶ Rubin, Donald. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- ▶ Schafer, Joseph L. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, 1997.