MLE midterm

Jason Douglas Todd

Friday, March 06, 2015

Discussion

Table 1 summarizes the results of an ordinary least squares (OLS) regression on data, which regresses an indicator of inequality (Gini: gini_net_std) on measures of ethnolinguistic fractionalization (ELF: ELF_ethnic) and democracy (Polity: polity2) at the country level. A scatterplot of the data can be seen in the top panel of Figure 2. Results for the full data include a substantively small and statistically insignificant intercept and coefficient for Polity scores, and a coefficient on ELF that is both substantively meaningful and statistically significant. Interpreting this result, *ceteris paribus*, an ethnolinguistically homogenous state would score around two points lower on this particular measure of income inequality than one with extreme ethnolinguistical diversity; this indicates a negative correlation between ethnolinguistic fractionalization and egalitarian income distributions when controlling for Polity scores.

	Estimate	Std. Error	T-Statistic	P-Value	Lower 95% CI	Upper 95% CI
(Intercept)	-0.22	0.85	-0.26	0.80	-1.94	1.50
ELF_ethnic	2.01	0.63	3.18	0.00	0.74	3.28
polity2	-0.10	0.08	-1.20	0.24	-0.26	0.06

Table 1: Results of OLS on data (model)

A second dataset, dataMiss, is identical to data save induced missingness on six observations of polity2. There are many approaches to missing data, including list-wise deletion, mean imputation, multiple imputation via expectation maximization (EM), multiple imputation by chained equations (MICE), and copula methods. This discussion will focus upon list-wise deletion (also known as complete case analysis) and EM-based multiple imputation. List-wise deletion makes no attempt to understand the pattern of missingness, nor does it make any attempt to ameliorate or impute missing values. Instead, any observation for which data is missing on at least one variable is removed from the dataset, at times appreciably reducing the sample size. Unless data is missing completely at random (MCAR), meaning that "missingness does not depend on the values of the data Y, missing or observed," list-wise deletion will typically introduce bias into the coefficient estimates, as well as decrease statistical power through the reduction in n (Little and Rubin 2002, 12).

Unfortunately in the case of dataMiss, MCAR does not correctly characterize the pattern of missingness. The two remaining potential patterns are missing at random (MAR) and missing not at random (MNAR). The former pattern occurs when "missingness depends only on the components Y_{obs} of Y that are observed, and not on the components that are missing", or, alternatively, when other variables in the dataset predict missingness (ibid.). Multiple imputation is appropriate for data MCAR or MAR and, once the cause of missingness is accounted for, parameter estimates should be unbiased. MNAR, however, describes data in which the distribution of missingness depends upon the unobserved values Y_{mis} of Y. With MNAR, even multiple imputation may produce biased parameter estimates, although the extent of the bias is not necessarily of substantive importance.

As Graham puts it, "The best way to think of all missing data is as a continuum between MAR and MNAR", and such is the case here (Graham 2009, 167). As can clearly be seen in the middle panel of Figure 2, the pattern of missingness in dataMiss displays a strong element of MAR. Specifically, an observation's Polity score is missing whenever the value of income inequality exceeds a threshold of 1. Not as obvious from the scatterplot is the relationship between the missingness and the value of polity2 itself, a relationship best described as MNAR. Table 2 reveals a clear trend in the probability of missingness conditional upon the value of polity2, yet this could only be calculated with the entirety of data in hand. The probability of missingness increases as Polity scores increase, peaking at Polity scores of nine, with full democracies producing no missing data.

	polity2	Pr(NA|polity2)
1	4	0.000
2	5	0.000
3	5.5	0.000
4	6	0.125
5	7	0.167
6	8	0.333
7	9	0.400
8	10	0.000

Table 2: Probability of missingness conditional on value of polity2

Having discussed the patterns of missingness, their implications, and their relationships to various approaches to missing data, as well as the particular patterns of missingness in the data at hand, it should be clear that both list-wise deletion and imputation via EM will, in the present case, produce biased parameter estimates. First, inspecting the bottom panel of Figure 2 reveals that the imputed Polity scores contain a consistent and, at times, substantive downward bias. This bias undoubtedly affects the results of any statistical models utilizing the imputed data. Thus Tables 3 and 4 present OLS results produced with these two approaches, with EM-based imputation implemented by the Amelia II package in R (Honaker et al. 2011). As compared to those of Table 1, the parameter estimates produced with list-wise deletion and imputation are qualitatively different. The estimated intercept, negative and insignificant in the original specification, is positive with list-wise deletion and Amelia II, and statistically significant and substantively large when imputed. The coefficient estimate on ELF–large, positive, and highly significant using data–decreases in size and significance as one moves from the full dataset to list-wise deletion to Amelia II. Interestingly, the coefficient estimate on Polity scores, originally negative and insignificant, is increasingly negative and increasingly significant with list-wise deletion and EM-based imputation. Compared to the original model, the standard errors for all parameter estimates decrease with list-wise deletion and imputation. Figure 1 permits visual inspection of the parameter estimates and standard errors across all three models.

	Estimate	Std. Error	T-Statistic	P-Value	Lower 95% CI	Upper 95% CI
(Intercept)	0.67	0.57	1.18	0.25	-0.48	1.82
ELF_ethnic	0.81	0.44	1.82	0.08	-0.09	1.70
polity2	-0.16	0.05	-3.16	0.00	-0.27	-0.06

Table 3: Results of OLS on dataMiss via list-wise deletion (modelListDel)

	Estimate	Std. Error	T-Statistic	P-Value	Lower 95% CI	Upper 95% CI
(Intercept)	1.94	0.46	4.22	0.00	1.01	2.87
ELF_ethnic	0.33	0.45	0.74	0.46	-0.58	1.25
polity2	-0.29	0.04	-7.49	0.00	-0.36	-0.21

Table 4: Results of OLS on dataMiss with imputation via Amelia II (modelAmelia)

In summary, then, the pattern of missingness in dataMiss is neither MCAR nor purely MAR, but falls along the continuum between MAR and MNAR. Because missingness is not MCAR and the proportion of missing data is not insubstantial, list-wise deletion is an inappropriate remedy and produces biased parameter estimates. This bias is visible in Figure 1. Additionally, because missingness is not exclusively MAR, multiple (or in this case, single) imputation also fails to deliver unbiased parameter estimates. This failure is illustrated in Figure 1, as well as in the bottom panel of Figure 2.

Figure 1: Coefficient plot illustrating results from Tables 1, 3, and 4

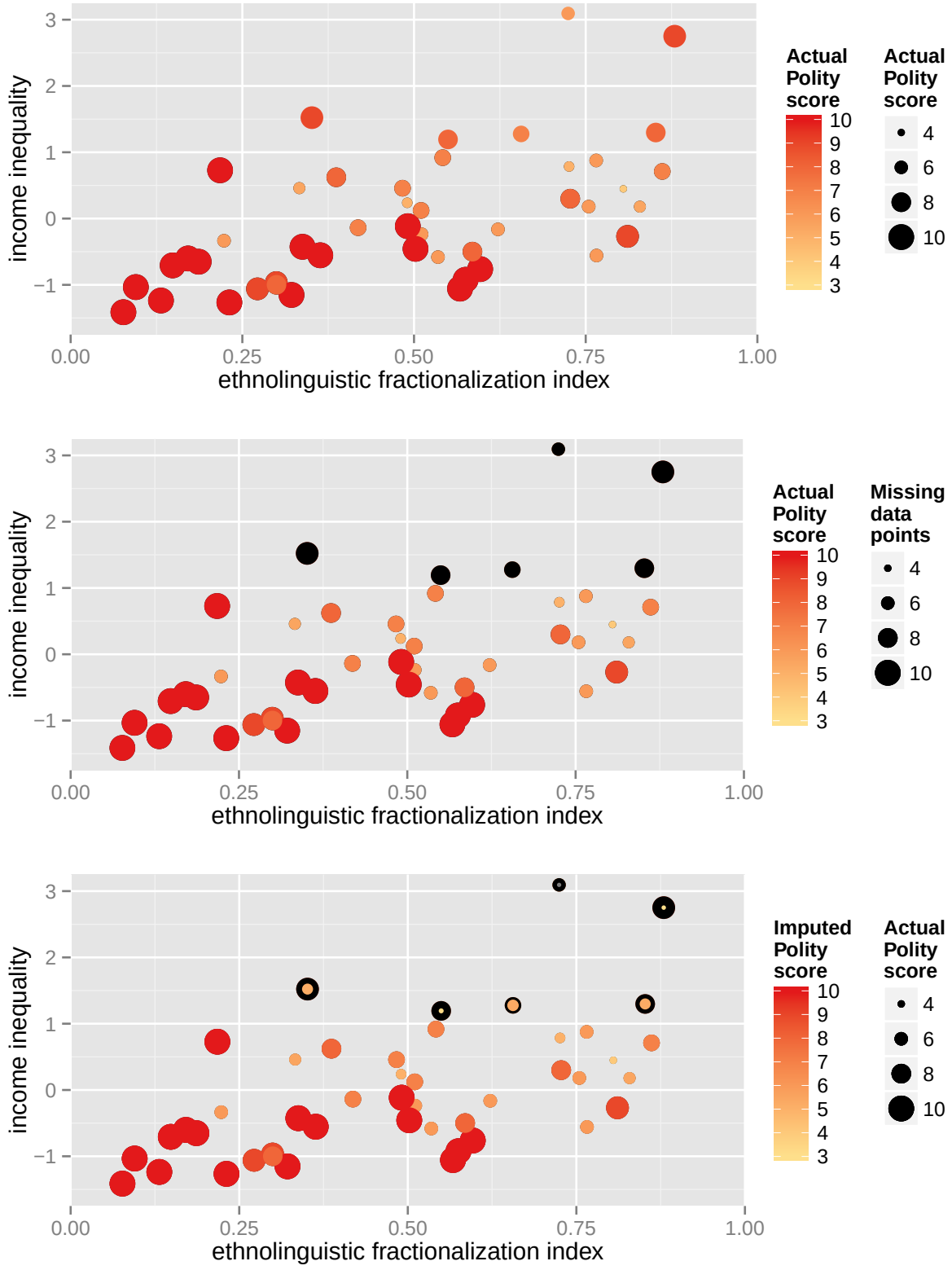

Figure 2: Scatterplots comparing (t) full data, (m) missing data, and (b) imputed data

References

Graham, John W. 2009. "Missing Data Analysis: Making It Work in the Real World." *Annual Review of Psychology*. 60: 549-576.

Honaker, James, Gary King, and Matthew Blackwell. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software*. 45 (7): 1-47.

Rubin, Donald B., and Roderick JA Little. 2002. *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley & Sons.

Code Appendix

```
# clears work environment, sets working directory, loads dataframes, and defines formula
rm(list=ls())
setwd('C:/Users/Jason/Box Sync/Home Folder jdt34/733 - Maximum Likelihood Estimation/week 7 midterm')
load('midTermData.rda')
form = formula(gini_net_std ~ ELF_ethnic + polity2)

library(ggplot2)
library(Amelia)
library(gridExtra)
library(xtable)
options(xtable.comment=F)

freqtable = as.data.frame((table(data$polity2) - table(dataMiss$polity2)) / table(data$polity2))
freqtable$Freq = round(freqtable$Freq, 3)
colnames(freqtable) = c('polity2', 'Pr(NA|polity2)')

# loads OLS function

# INPUT:
# 1) formula = formula object specifying model to be run
# 2) data    = dataframe object containing data (DV and IVs)
# 3) impute  = logical parameter indicating whether missing data should be
#              imputed via Amelia (m=1)
#      ^ default  = FALSE
# OUTPUT: list object
# 1) coefficients = (p+1)x6 matrix object summarizing coefficient estimates
#      ^ rownames = '(Intercept)' and original IV names
#      ^ colnames = 'Estimate', 'Std. Error', 'T-Statistic', 'P-Value',
#                   'Lower 95% CI', 'Upper 95% CI'
# 2) varcov       = matrix object containing variance-covariance matrix of
#                   coefficient estimates
#      ^ rownames = rownames(coefficients)
#      ^ colnames = rownames(coefficients)
# 3) Rsq          = numeric object containing value of model's R^2 statistic
# 4) Fstat        = character object providing F-statistic, relevant degrees
#                   of freedom, and p-value (rounded to 3 decimal places)
#      ^ format   = 'F-statistic: 0.204 on 1 and 8 DF,  p-value: 0.663'

ols = function(formula, data, impute=FALSE) {
  dv     = all.vars(formula)[1]
  ivs    = all.vars(formula)[2:length(all.vars(formula))]
  if(impute==T) {
    if(! 'Amelia' %in% installed.packages()[,1]) {
      install.packages('Amelia', repos='http://cran.rstudio.com/')
    }
    suppressMessages(library('Amelia', character.only=T))
    # set.seed(6886)
    found = amelia(data, m=1, p2s=0)
    Adata = found$imputations$imp1
    y     = Adata[,dv]
    x     = data.matrix(cbind(1, Adata[,ivs]))
```

```
  } else {
    data  = na.omit(data[,c(dv, ivs)])
    y     = data[,dv]
    x     = data.matrix(cbind(1, data[,ivs]))
  }
  n      = nrow(x)
  p      = length(ivs)
  df     = n - p - 1
  xTx    = t(x) %*% x
  xTy    = t(x) %*% y
  betas  = solve(xTx) %*% xTy
  yHats  = x %*% betas
  resids = y - yHats
  msRes  = sum(resids^2) / df
  msReg  = sum(yHats^2) / p
  fstat  = round((msReg / msRes), 3)
  fpval  = round(pf(fstat, p, df, lower.tail=F), 3)
  Fstat  = paste('F-statistic:', fstat, 'on', p, 'and', df, 'DF, p-value:', fpval)
  ssRes  = sum(t(resids) %*% resids)
  ssReg  = sum((mean(y) - yHats)^2)
  ssTot  = ssRes + ssReg
  Rsq    = ssReg / ssTot
  sigma2 = ssRes / df
  varcov = solve(xTx) * sigma2
  ses    = sqrt(diag(varcov))
  tstats = betas / ses
  pvals  = 2 * pt(abs(tstats), df, lower.tail=F)
  band   = qt(.975, df) * ses
  lower  = betas - band
  upper  = betas + band
  coefficients              = cbind(betas, ses, tstats, pvals, lower, upper)
  colnames(coefficients)    = c('Estimate', 'Std. Error', 'T-Statistic',
                                'P-Value', 'Lower 95% CI', 'Upper 95% CI')
  rownames(coefficients)[1] = '(Intercept)'
  rownames(varcov) = colnames(varcov) = rownames(coefficients)
  output = list(coefficients, varcov, Rsq, Fstat)
  names(output) = c('coefficients', 'varcov', 'Rsq', 'Fstat')
  return(output)
}

# sets seed and runs models
set.seed(6886)
model        = ols(formula=form, data=data)
modelListDel = ols(formula=form, data=dataMiss)
modelAmelia  = ols(formula=form, data=dataMiss, impute=T)

cp.coefs = rownames(model$coefficients)
cp.model = as.data.frame(model$coefficients)
cp.model$model = 'no missing data'
cp.modelListDel = as.data.frame(modelListDel$coefficients)
cp.modelListDel$model = 'list-wise deletion'
cp.modelAmelia = as.data.frame(modelAmelia$coefficients)
cp.modelAmelia$model = 'EM-based imputation'
```

```
cp.data = rbind(cp.model, cp.modelListDel, cp.modelAmelia)
rownames(cp.data) = NULL
colnames(cp.data) = c('est', 'se', 'tstat', 'pval', 'lo95', 'up95', 'Model')
cp.data$coef = cp.coefs

# Figure 1: coefficient plot
cp = ggplot(data=cp.data,
            aes(color=Model)
            ) +
  geom_linerange(aes(x=coef,
                     ymin=lo95,
                     ymax=up95),
                 position=position_dodge(width=.5)
                 ) +
  geom_point(aes(x=coef,
                 y=est),
             position=position_dodge(width=.5),
             shape=21,
             fill='white'
             ) +
  geom_hline(yintercept=0,
             linetype=2,
             color='black') +
  coord_flip() +
  xlab('') +
  ylab('') +
  labs(title='Coefficient plot by model') +
  scale_x_discrete(limits=c('polity2', 'ELF_ethnic', '(Intercept)')) +
  scale_y_continuous(breaks=seq(-4, 4, 1)) +
  scale_color_discrete(breaks=c('no missing data',
                                'list-wise deletion',
                                'EM-based imputation'),
                       labels=c('no missing\ndata',
                                'list-wise\ndeletion',
                                'EM-based\nimputation')
                       ) +
  theme(axis.ticks=element_blank(),
        legend.position='bottom')

MIamelia = amelia(x=dataMiss, m=1, p2s=0)
MIdata = MIamelia$imputations$imp1
NAdata = data[which(is.na(dataMiss$polity2)),]

# Figure 2: top panel
full = ggplot(data=MIdata,
              aes(x=ELF_ethnic,
                  y=gini_net_std,
                  size=polity2)
              ) +
  geom_point() +
  geom_point(data=data,
             aes(color=polity2)
             ) +
```

```
  scale_color_continuous(low='#fee08b',
                         high='#e31a1c',
                         name='Actual\nPolity\nscore',
                         limits=c(3, 10),
                         breaks=seq(3, 10, 1)
                         ) +
  scale_size_continuous(name='Actual\nPolity\nscore'
                        ) +
  scale_x_continuous(limits=c(0, 1), expand=c(0, 0)) +
  scale_y_continuous(limits=c(-1.75, 3.25), expand=c(0, 0)) +
  labs(x='ethnolinguistic fractionalization index',
       y='income inequality'
       ) +
  guides(color=guide_colorbar(order=1),
         size=guide_legend(order=2)) +
  theme(legend.title=element_text(size=10),
        legend.position='right',
        legend.box='horizontal')

# Figure 2: middle panel
missing = ggplot(data=MIdata,
                 aes(x=ELF_ethnic,
                     y=gini_net_std,
                     size=polity2)
                 ) +
  geom_point() +
  geom_point(data=data,
             aes(color=polity2)
             ) +
  geom_point(data=NAdata,
             color='black'
             ) +
  scale_color_continuous(low='#fee08b',
                         high='#e31a1c',
                         name='Actual\nPolity\nscore',
                         limits=c(3, 10),
                         breaks=seq(3, 10, 1)
                         ) +
  scale_size_continuous(name='Missing\ndata\npoints'
                        ) +
  scale_x_continuous(limits=c(0, 1), expand=c(0, 0)) +
  scale_y_continuous(limits=c(-1.75, 3.25), expand=c(0, 0)) +
  labs(x='ethnolinguistic fractionalization index',
       y='income inequality'
       ) +
  guides(color=guide_colorbar(order=1),
         size=guide_legend(order=2)) +
  theme(legend.title=element_text(size=10),
        legend.position='right',
        legend.box='horizontal')

# Figure 2: bottom panel
impute = ggplot(data=MIdata,
```

```
                aes(x=ELF_ethnic,
                    y=gini_net_std,
                    size=polity2)
                ) +
  geom_point() +
  geom_point(data=data,
             aes(color=polity2)
             ) +
  geom_point(data=NAdata,
             color='black'
             ) +
  geom_point(data=MIdata,
             aes(color=polity2)
             ) +
  scale_color_continuous(low='#fee08b',
                         high='#e31a1c',
                         name='Imputed\nPolity\nscore',
                         limits=c(3, 10),
                         breaks=seq(3, 10, 1)
                         ) +
  scale_size_continuous(name='Actual\nPolity\nscore'
                        ) +
  scale_x_continuous(limits=c(0, 1), expand=c(0, 0)) +
  scale_y_continuous(limits=c(-1.75, 3.25), expand=c(0, 0)) +
  labs(x='ethnolinguistic fractionalization index',
       y='income inequality'
       ) +
  guides(color=guide_colorbar(order=1),
         size=guide_legend(order=2)) +
  theme(legend.title=element_text(size=10),
        legend.position='right',
        legend.box='horizontal')

# produces tables and figures
xtable(model$coefficients,
       caption='Results of OLS on data (model)\\label{tab:fulltable}')
xtable(freqtable,
       caption='Probability of missingness conditional on value of polity2\\label{tab:freqtable}',
       digits=c(1,0,3))
xtable(modelListDel$coefficients,
       caption='Results of OLS on dataMiss via list-wise deletion (modelListDel)\\label{tab:missingtable
xtable(modelAmelia$coefficients,
       caption='Results of OLS on dataMiss with imputation via Amelia II (modelAmelia)\\label{tab:imput
cp
grid.arrange(full, missing, impute, ncol=1)
```