MLE: Midterm

Jochen Rehmert

March 12, 2015

Bias through Missing Data Treatment

Table 2 presents three OLS regressions of Income inequality on the state level – measured through the GINI coefficient – on the level of democracy, measured by the Polity II index, and ethnic-linguistic fractionalization by looking at 46 countries. The Table presents results of three different OLS models that, by having the same regression equation, differ only in missingness of data and how this is treated.

The first model is run on the complete data set including all 46 countries. The results suggest that only ethnic fractionalization is significantly related to higher income inquality. The level of democracy, however, does not appear to be significantly related to the GINI coefficient.

The second model differs from the first one as 6 datapoints on the Polity II index variable were put to missing. Given the conventionality of list-wise deletion in virtually all regressions, the number of countries is decreased by 6 to 40. Moreover, the coefficients of the two covariates as well as the intercept are altered substantially. Now, the effect of the Polity II coefficient becomes negatively significant. The coefficient of ethnic fractionalization loses in significance. It becomes apparent, that list-wise deletion biases estimation of coefficients and standard errors if the missingness of data is not completely random – a very strong assumption.

The third model in Table 2 presents the same model ran on the missing data with the 6 missing data points being imputed (m=1). The Polity II coefficient's size and significance increases, whereas both decreases for ethnic fractionalization. Apparently, the imputed data model fares worse than the list-wise deletion model when comparing both with the complete data model (see Table 1).

	δ=| Estimate(Complete Data) - Estimate(Incomplete Data)|	
	Polity II	Ethnic Fractionalization
List-wise Deletion	0.0688	1.2052
Imputation	0.1911	1.6761

Table 1: Bias in Coefficients across Incomplete Data Models

When comparing the R^2 across the three models, the imputated data model appears to explain most of the variation of income inequality in the sample.

However, one should remember that usually several datasets are imputed with varying values for those cells empty, representing the uncertainty accompanied with multiple imputation. Thus, having set m=1, the imputed values are treated as if they were factual observed values. The actual uncertainty in these point estimations for the imputed data is thus not accounted for, which might explain the high R^2 as well as the rather far off coefficient estimations.

Figure 1 and 2 visually compares the coefficients sizes of all three models based on simulations with 10000 draws from a variance-covariance based multivariate normal distribution, taking in account only the systematic uncertainty of the model. In both figures, the coefficients values simulated for the imputed data models are the ones farest away from the complete data model's one, but those with the thinnest density distribution. This latter fact corresponds to the low uncertainty connected to the imputation with m=1. The biggest overlap in the distributions happens for the full data model and the list-wise deletion model for the very coefficient with the missing data. In contrast, for the ethnic frationalization coefficient, overlap is biggest between the list-wise deletion and the imputated data model.

```
## Loading required package:  Amelia
## Loading required package:  Rcpp
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.3, built:  2014-11-14)
## ## Copyright (C) 2005-2015 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

	Estimates	Std. Error	T-Statistic	P-Value	Lower 95% CI	Upper 95% CI
Complete data Model:						
Polity II	-0.0959	(0.0796)	-1.2044	0.2350	-0.2519	0.0601
Ethnic Fractionalization	2.0104	(0.6313)	3.1844	0.0027	0.7730	3.2477
Intercept	-0.2200	(0.8537)	-0.2577	0.7979	-1.8932	1.4532
N	46					
R^2	0.3396					
F-Stat	$F_{2,43}$= 11.057					
Prob > χ^2	0.000					
List-Wise Deletion Model:						
Polity II	-0.1647	(0.0522)	-3.1573	0.0032	-0.2669	-0.0624
Ethnic Fractionalization	0.8052	(0.4435)	1.8155	0.0776	-0.0641	1.6745
Intercept	0.6692	(0.5694)	1.1753	0.2474	-0.4468	1.7852
N	40					
R^2	0.4481					
F-Stat	$F_{2,37}$= 20.891					
Prob > χ^2	0.000					
Imputed Data Model:						
Polity II	-0.2870	(0.0383)	-7.4928	0.0000	-0.3621	-0.2119
Ethnic Fractionalization	0.3343	(0.4522)	0.7393	0.4638	-0.5519	1.2205
Intercept	1.9381	(0.4597)	4.2155	0.0001	1.0370	2.8392
N	46					
R^2	0.7039					
F-Stat	$F_{2,43}$= 51.113					
Prob > χ^2	0.000					

Table 2: OLS on Income Inequality: Bias through Missing Data Treatment

Figure 1: Simulated β for Polity II index

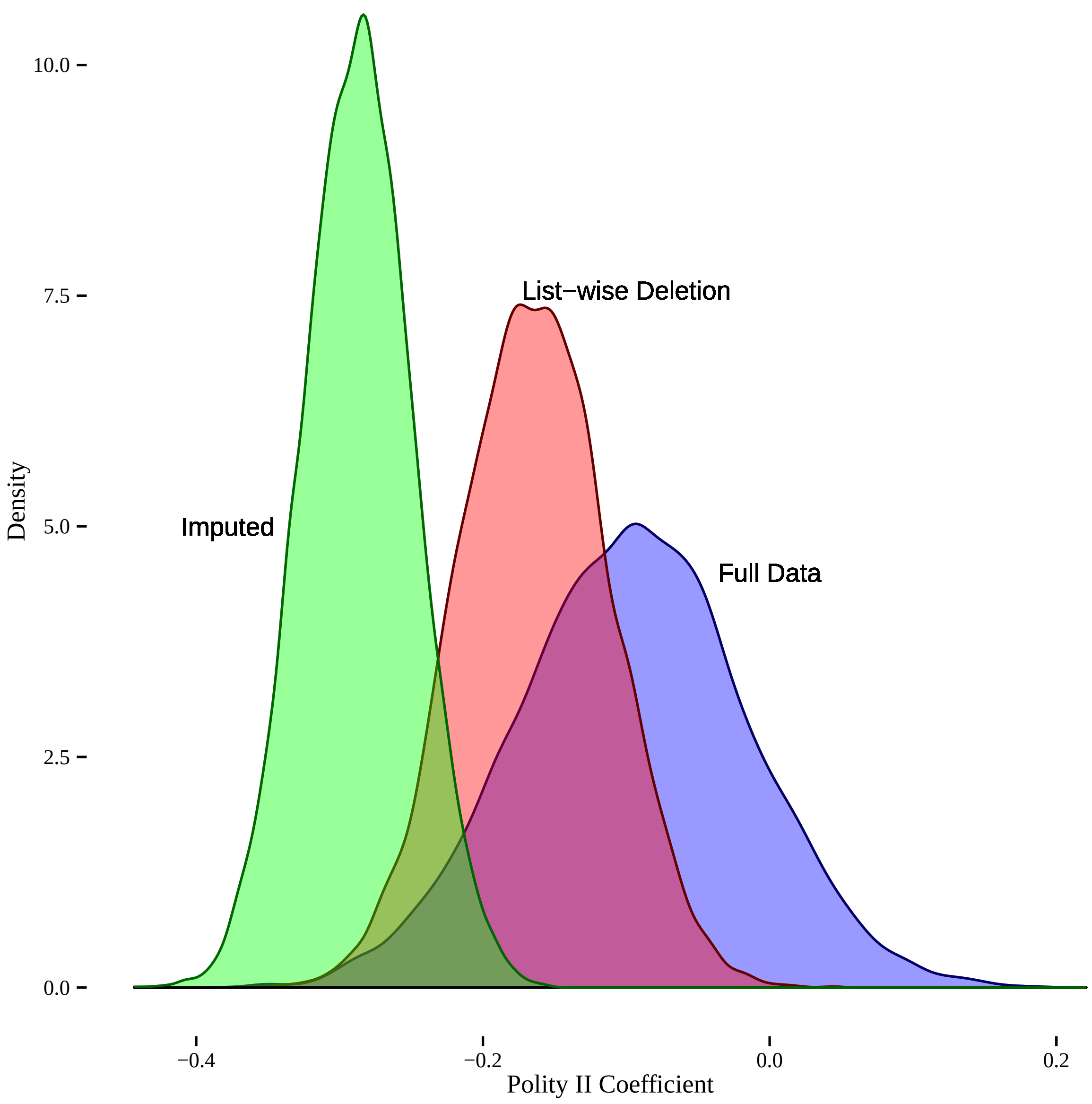
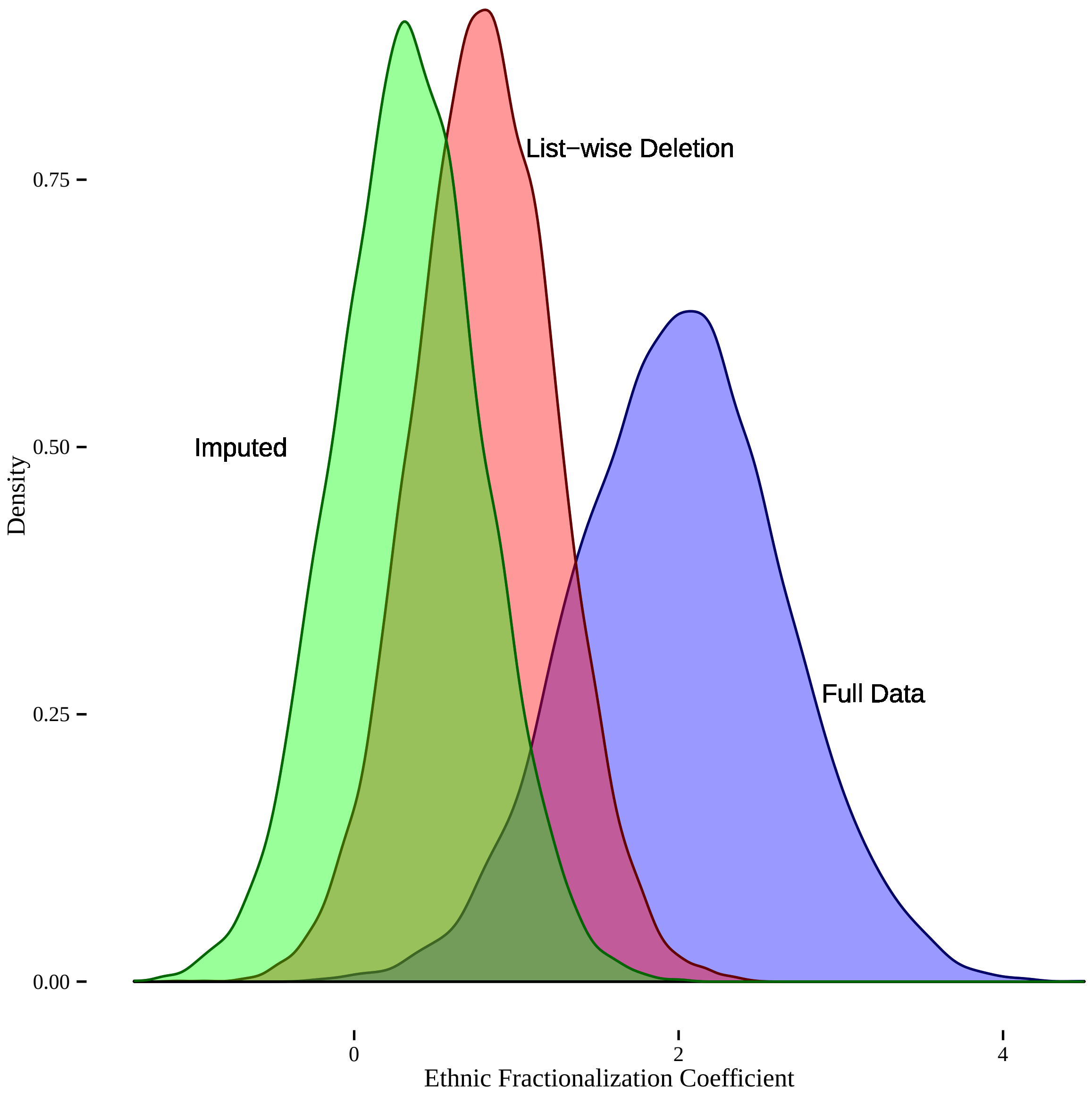


Figure 2: Simulated β for Ethnic Fractionalization



R Code

```
load("C:/Users/jochen/Dropbox/4_Fachsemester/MLE/midterm/midTermData.rda")
```

```
form = formula(gini_net_std ~ polity2 + ELF_ethnic)

#----- problems
# 1) output for p-val of Fstat does not round as wanted

set.seed(6886)

ols <- function(formula, data, impute=FALSE){
  #-------------------------------------------- handling missing data
  if((!require(Amelia)) && (impute==TRUE)){print("The 'Amelia' package is not installed. Please
  if(impute == TRUE){
    data <-  Amelia::amelia(data, m = 1)[['imputations']][['imp1']]
    data <- na.omit(data)
  } else {
    data <- na.omit(data)
  }
  mf <- model.frame(formula=formula, data=data)
  #-------------------------------------------- model matrix % parameter
  X  <- model.matrix(attr(mf, "terms"), data=mf)
  y  <- model.response(mf)
  n  <- nrow(mf)
  p  <- length(mf)
  df <- n-p
  #-------------------------------------------- estimation
  # beta: (X'X)^-1 %*% X'Y
  xTx = t(X) %*% X # X'X   dimensions: (p+1) * (p+1)
  xTy = t(X) %*% y # X'Y
  beta = solve(xTx) %*% xTy # getting beta
  # standard errors: sqrt(sigma^2 * (X'X)^-1)
  yhat = X %*% beta # predicted values
  e = y - yhat # residuals
  sigma2 = (sum(e^2))/(n-p) # e'e = sum of squared residuals
  varcov = sigma2 * solve(xTx) # variance-covariance matrix
  serror = sqrt(diag(varcov)) # standard error
  # t-stats
  tstat = beta/serror
  # p-value
  pval = 2 * (pt(abs(tstat), df, lower.tail=FALSE))
  #------------------------------------------- diagnostics & fit
  # R-squared
  SSreg = sum((yhat - mean(y))^2) # regression sum of squares
  SStot = sum((y - mean(y))^2) # total sum of squares
  rsq = SSreg/SStot
  # Adjusted R-squared
  adjr2 = 1-(1-rsq)*((n-1)/df)
  # F statistic
  MSReg = sum(yhat^2)/(p-1)
  MSRes = sum(e^2)/df
  fstat = MSReg/MSRes # f statistic
  fpval = pf(fstat, p, df, lower.tail=FALSE) # p-value for f statistic
  # root mean squared error
  rmse = sqrt(mean(e^2))
  # confidence intervals
  cilow95 = beta - qnorm(.975) * serror
  cihigh95 = beta + qnorm(.975) * serror
  #------------------------------------------ output essentials & cosmetics
  coefMat <- cbind(
    'Coefficient' = beta,
    'Std. Error' = serror,
    'T Stat' = tstat,
    'P-Value' = pval,
    'Low CI' = cilow95,
    'High CI' = cihigh95
  )
  colnames(coefMat) <- c("Estimates", "Std. Error", "T-Statistic", "P-Value", "Lower 95% CI", "

  ftest = paste0("F-statistic: ", round(fstat,3)," on ", (p-1), " and " ,df, " DF, p-value: ",

  return(list(
    coefficients= round(coefMat,4),
    varcov=varcov,
    Rsq=rsq,
    Fstat=ftest
      ))
}


model = ols(formula=form, data=data)

modelListDel = ols(formula = form, data=dataMiss)

modelAmelia = ols(formula = form, data=dataMiss, impute=TRUE)


## -- Imputation 1 --
##
##   1  2  3  4
```

```
# simulation
library(MASS);library(ggplot2);library(ggthemes)

sims = 10000 # number of simulations
coefModel=model$coefficients[,1]     # coefs of model 1
coefListW=modelListDel$coefficients[,1] # coefs of model 2
coefAmelia=modelAmelia$coefficients[,1] # coefs of model 3
varcovModel=model$varcov       # varcov of model 1
varcovListW=modelListDel$varcov # varcov of model 2
varcovAmelia=modelAmelia$varcov # varcov of model 3

set.seed(6886)
drawsModel = mvrnorm(sims, coefModel, varcovModel)
drawsListW = mvrnorm(sims, coefListW, varcovListW)
drawsAmelia = mvrnorm(sims, coefAmelia, varcovAmelia)


# transform to data frame
ggmodel <- data.frame(drawsModel)
gglistw <- data.frame(drawsListW)
ggamelia <- data.frame(drawsAmelia)

ggData <- cbind(ggmodel, gglistw, ggamelia)
colnames(ggData) <- c("intercept_model", "polity_model", "elf_model",
                       "intercept_listw", "polity_listw", "elf_listw",
                       "intercept_amelia", "polity_amelia", "elf_amelia")


m1 <- ggplot(ggData)
m1 <- m1 + geom_density(aes(x = polity_model), fill="blue", alpha = .4)
m1 <- m1 + geom_line(aes(x = polity_model),
           color = "blue", fill="blue", size = 0.5, alpha = .4, stat = "density")
m1 <- m1 + geom_density(aes(x = polity_listw), fill="red", alpha = .4)
m1 <- m1 + geom_line(aes(x = polity_listw),
           colour = "red", fill="red", size = 0.5, alpha = .4, stat = "density")
m1 <- m1 + geom_density(aes(x = polity_amelia), fill="green", alpha = .4)
m1 <- m1 + geom_line(aes(x = polity_amelia),
           colour = "green", fill="green", size = 0.5, alpha = .4, stat = "density")
m1 <- m1 + theme_tufte(ticks = TRUE, base_family = "serif", base_size = 11)
m1 <- m1 + scale_x_continuous(name="Polity II Coefficient") + scale_y_continuous(name="Density'
m1 <- m1 + geom_text(x=-0.10, y=7.56, label="List-wise Deletion", size=4, alpha=.6)
m1 <- m1 + geom_text(x=-0.378, y=5, label="Imputed", size=4)
m1 <- m1 + geom_text(x=0, y=4.5, label="Full Data", size=4)
```

```r
m2 <- ggplot(ggData)
m2 <- m2 + geom_density(aes(x = elf_model), fill="blue", alpha = .4)
m2 <- m2 + geom_line(aes(x = elf_model),
           colour = "blue", fill="blue", size = 0.5, alpha = .4, stat = "density")
m2 <- m2 + geom_density(aes(x = elf_listw), fill="red", alpha = .4)
m2 <- m2 + geom_line(aes(x = elf_listw),
           colour = "red", fill="red", size = 0.5, alpha = .4, stat = "density")
m2 <- m2 + geom_density(aes(x = elf_amelia), fill="green", alpha = .4)
m2 <- m2 + geom_line(aes(x = elf_amelia),
           colour = "green", fill="green", size = 0.5, alpha = .4, stat = "density")
m2 <- m2 + theme_tufte(ticks = TRUE, base_family = "serif", base_size = 11)
m2 <- m2 + scale_x_continuous(name="Ethnic Fractionalization Coefficient") + scale_y_continuous
m2 <- m2 + geom_text(x=1.7, y=0.78, label="List-wise Deletion", size=4, alpha=.6)
m2 <- m2 + geom_text(x=-0.7, y=.5, label="Imputed", size=4)
m2 <- m2 + geom_text(x=3.2, y=.27, label="Full Data", size=4)
```