Mid-term Assignment

Haosen Ge

March 17, 2015

1 Intepretation

Table 1: Coefficients of Original Data

	Estimate	Std. Error	T-Statistic	P-Value	Lower 95% CI	Upper 95% CI
intercept	-0.22	0.85	-0.26	0.80	-1.89	1.45
ELF_ethnic	2.01	0.63	3.18	0.00	0.77	3.25
polity2	-0.10	0.08	-1.20	0.24	-0.25	0.06

Table 2: Coefficients of Listwise Deletion Data

	Estimate	Std. Error	T-Statistic	P-Value	Lower 95% CI	Upper 95% CI
intercept	0.67	0.57	1.18	0.25	-0.45	1.79
ELF_ethnic	0.81	0.44	1.82	0.08	-0.06	1.67
polity2	-0.16	0.05	-3.16	0.00	-0.27	-0.06

Table 3: Coefficients of Imputed Data

	Estimate	Std. Error	T-Statistic	P-Value	Lower 95% CI	Upper 95% CI
intercept	1.94	0.46	4.22	0.00	1.04	2.84
ELF_ethnic	0.33	0.45	0.74	0.46	-0.55	1.22
polity2	-0.29	0.04	-7.49	0.00	-0.36	-0.21

I find that the coefficients of the imputed data are more biased than the listwise deletion data. From Figure 1, it can be observed that some obvious outliers are deleted. Because the number of variables in the dataset are very small, it is hard to justify the missing at random (MAR) assumption. The missing of polity score might be due to the score itself, which is nonignorable. Or, the missing of polity2 is due to some other variables which are not included in the model. The

application of mulitiple imputation might create more bias due to the potential violation of the assumption.

The 6 data points created in the imputed data are far smaller than the real data points. The variance of the variable polity2 and the covariance between polity2 and gini also increases significantly. The polity2 variable is truncated from its original distribution (ranges only from 4 to 10) which might violate the multivariate normal distribution assumption. The imputed data are drawn from the multivariate normal distribtuion while the real data are rescaled or truncated, which obviously will bias the result.

Table 4: Covariance of Original Data

	gini_net_std	polity2	ELF_ethnic
gini_net_std	1.00	-0.79	0.13
polity2	-0.79	3.41	-0.23
ELF_ethnic	0.13	-0.23	0.05

Table 5: Covariance of Imputed Data

	gini_net_std	polity2	ELF_ethnic
gini_net_std	1.00	-2.30	0.13
polity2	-2.30	7.55	-0.39
ELF_ethnic	0.13	-0.39	0.05

Table 6: Covariance of Listwise Deletion Data

	gini_net_std	polity2	ELF_ethnic
gini_net_std	0.45	-0.82	0.08
polity2	-0.82	3.76	-0.26
ELF_ethnic	0.08	-0.26	0.05

2 R code

```
loadPkg=function(toLoad){
  for(lib in toLoad){
    if(! lib %in% installed.packages()[,1])
    { install.packages(lib, repos='http://cran.rstudio.com/') }
```

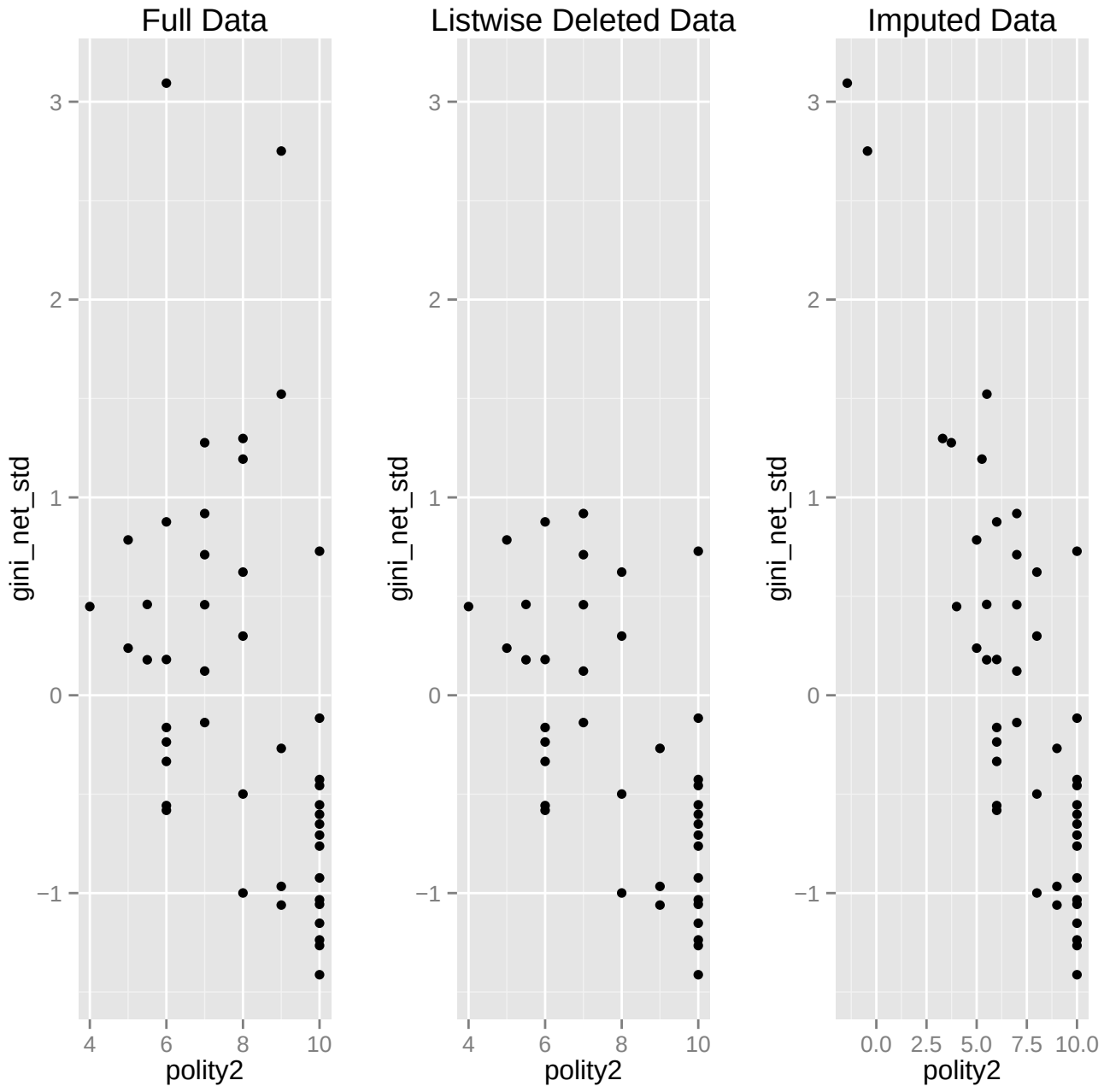

Figure 1: Data Distribution

```
    suppressMessages( library(lib, character.only=TRUE) ) }
}
setwd("C:/Users/Richard/Desktop/MLE/HW7")
load("midTermData.rda")
packs=c('foreign', 'lmtest', 'sandwich', 'Amelia',
        'sbgcop','stargazer','ggplot2','gridExtra','xtable')
loadPkg(packs)
# OLS function
  ols = function(formula, data, impute = FALSE){
  if(impute==TRUE){
  set.seed(6886)
  data=amelia(x=data,m=1)$imp$imp1}
  data=data[complete.cases(data),]
  dv = all.vars(formula)[1]
  ivs = all.vars(formula)[2:length(all.vars(formula))]
  y = data[,dv]
  x = data.matrix(cbind(1, data[,ivs]))
  n = nrow(x) # Number of observations
  p = length(ivs) # Number of parameters
  df = n-p-1 # degrees of freedom
  ## Get the coefficients
  xTx = t(x) %*% x
  xTy = t(x) %*% y
  beta = solve(xTx) %*% xTy
  ## Get the se
  yhat=x %*% beta
  e= y - yhat
  sigma2=sum(e^2)/df
  vcov=sigma2 * solve(xTx)
  se=sqrt(diag(vcov))
  ## Get the tstat, p-val and CI
  tstat = beta/se
  pval1=2*pt(abs(tstat),df,lower.tail=FALSE)
  up95=beta+qnorm(0.975)*se
  lo95=beta-qnorm(0.975)*se
  ## Get R-sq
  SSreg=sum((yhat-mean(y))^2)
  SStot=sum((y-mean(y))^2)
  R2=SSreg/SStot
  ## Get F stat
  Fstat=round((sum((yhat-mean(y))^2)/p)/(sum(e^2)/df),3)
  pval2=round(pf(Fstat,p,df,lower.tail=FALSE),3)
  ## Prepare the result
  coefficients = cbind(beta,se,tstat,pval1,lo95,up95)
  colnames(coefficients)=c("Estimate", "Std. Error",
                           "T-Statistic", "P-Value", "Lower 95% CI", "Upper 95% CI")
```

```
  rownames(coefficients)=c('intercept',ivs)
  varcov=vcov
  Rsq=R2
  Fstat=paste('F-statistic:',Fstat, 'on' ,p, 'and' ,df,
              'DF,',  'p-value:', pval2,sep=' ')
  result = list(coefficients=coefficients,varcov=varcov,Rsq=Rsq,Fstat=Fstat)
  return(result)
  }

set.seed(6886)
datalia = amelia(x=dataMiss,m=1)$imp$imp1
form = formula(gini_net_std ~ ELF_ethnic + polity2)

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

#lm1=lm(form,data=data)
#lm2=lm(form,data=dataMiss)
#lm3=lm(form,data=datalia)
xtable(model$coefficients)
xtable(modelListDel$coefficients)
xtable(modelAmelia$coefficients)


scat1=ggplot(data=data,aes(y=gini_net_std,x=polity2))+
  geom_point()+ggtitle("Full Data")
scat2=ggplot(data=dataMiss,aes(y=gini_net_std,x=polity2))+
  geom_point()+ggtitle("Listwise Deleted Data")
scat3=ggplot(data=datalia,aes(y=gini_net_std,x=polity2))+
  geom_point()+ggtitle("Imputed Data")
grid.arrange(scat1, scat2, scat3, ncol=3)
```