Dissertation Study 1 - Cross National Multilevel Model

Dylan Wiwad

June 21, 2018

This is the supplemental code document for Study 1 in my dissertation. This contains all the code and analysis regarding my handling of the missing data, the linear regressions, and then ultimately the multilevel model of World Values Survey data.

First just loading the data and all required packages.

Handling of Missing Data

The missing data in the WVS dataset in this particular wave of the WVS comes in three distinct flavors: (1) not asked in survey, (2) not answered, and (3) don't know. I suspected that all the missing data that was simply not asked in the survey is clustered under country. For instance, certain questions simply were not asked in certain countries. To explore this, I created new datasets for each of the key variables and looked at whether all observations from certain countries were missing.

```
ineq_mis <- wvs[ which(wvs$ineq==-4), ]
attrib_mis <- wvs[ which(wvs$attrib==-4),]
id_mis <- wvs[ which(wvs$ideology==-4),]
inc_mis <- wvs[ which(wvs$inc.lad==-4),]
relig_mis <- wvs[ which(wvs$relig.import==-4),]
educ_mis <- wvs[ which(wvs$educ==-4),]

# just get the counts
count(ineq_mis$country)
```

```
##     x freq
## 1 170 3029
## 2 586  733
```

```
count(attrib_mis$country)
```

```
##     x freq
## 1 170 3029
## 2 756 1212
## 3 826 1093
```

```
count(id_mis$country)
```

```
##     x freq
## 1 156 1500
## 2 170 3029
## 3 586  733
## 4 826 1093
```

```
count(inc_mis$country)
```

```
##     x freq
## 1 348  650
## 2 608 1200
## 3 705 1007
```

```r
count(relig_mis$country)
```

```
##     x freq
## 1 156 1093
```

```r
count(educ_mis$country)
```

```
##     x freq
## 1 191 1196
## 2 392 1054
```

According to the codebook for the WVS and the above country codes, all the missing data comes from China, Colombia, Pakistan, Switzerland, Great Britain, Croatia, Japan, Hungary, Slovenia, and the Phillipines.

I'm going to print the country counts in the full data set here to compare:

```r
count(wvs$country)
```

```
##      x freq
## 1    8  999
## 2   31 2002
## 3   32 1079
## 4   36 2048
## 5   50 1525
## 6   51 2000
## 7  100 1072
## 8  112 2092
## 9  152 1000
## 10 156 1500
## 11 158  780
## 12 170 6025
## 13 191 1196
## 14 203 1147
## 15 214  417
## 16 222 1254
## 17 233 1021
## 18 246  987
## 19 268 2008
## 20 276 2026
## 21 348  650
## 22 356 2040
## 23 392 1054
## 24 410 1249
## 25 428 1200
## 26 440 1009
## 27 484 2364
## 28 498  984
## 29 499  240
## 30 554 1201
## 31 566 1996
## 32 578 1127
## 33 586  733
## 34 604 1211
## 35 608 1200
## 36 616 1153
## 37 630 1164
```

```
## 38 642 1239
## 39 643 2040
## 40 688 1280
## 41 703 1095
## 42 705 1007
## 43 710 2935
## 44 724 1211
## 45 752 1009
## 46 756 1212
## 47 792 1907
## 48 804 2811
## 49 807  995
## 50 826 1093
## 51 840 1542
## 52 858 1000
## 53 862 1200
## 54 914  800
```

Notice how, for instance, country 586 (Pakistan) was missing 733 observations support for inequality and political ideology but there were only 733 observations in the complete dataset for Pakistan. This suggests that these questions simply were not asked in Pakistan at all. So here I list wise delete all the rows that have any missing values that simply were not asked (despite these values likely not being MCAR due to country clustering).

```
# ideology
wvs <- wvs[ which(wvs$ideology>=-3), ]
# Equality
wvs <- wvs[ which(wvs$ineq>=-3),]
# Attributions
wvs <- wvs[ which(wvs$attrib>=-3),]
wvs <- wvs[ which(wvs$attrib <= 2),] # this also removes people who said neither
# Sex
wvs <- wvs[ which(wvs$sex>=-3),]
# Age
wvs <- wvs[ which(wvs$age>=-3),]
# Educ
wvs <- wvs[ which(wvs$educ>=-3),]
# Income Ladder
wvs <- wvs[ which(wvs$inc.lad>=-3),]
# Religiosity
wvs <- wvs[ which(wvs$relig.import>=-3),]
```

Now, with these all removed I need to convert all the missing values (coded as -2 and -1) intro straight NAs.

```
wvs$ineq[wvs$ineq<=0] <- NA
wvs$attrib[wvs$attrib<=0] <- NA
wvs$ideology[wvs$ideology<=0] <- NA
wvs$sex[wvs$sex<=0] <- NA
wvs$age[wvs$age<=0] <- NA
wvs$educ[wvs$educ<=0] <- NA
wvs$inc.lad[wvs$inc.lad<=0] <- NA
wvs$relig.import[wvs$relig.import<=0] <- NA
```

Now with this all dealt with I can move towards testing for MCAR using Little's (1981) protocol. I'll do this in a smaller trimmed dataset containing only the eight variables I actually care about.

```r
key_cols <- c("ineq", "attrib", "ideology", "sex", "age", "educ", "inc.lad", "relig.import")
key_Vars <- wvs[key_cols]

# Run the MCAR test, as described in Little, 1988
mcar_test <- LittleMCAR(key_Vars)
```

```
## this could take a while
```

```r
# If I try to print the whole thing it prints out all the data too and hides stuff, so lets just get th
mcar_test$chi.square
```

```
## [1] 3500.663
```

```r
mcar_test$df
```

```
## [1] 425
```

```r
mcar_test$p.value
```

```
## [1] 0
```

```r
mcar_test$missing.patterns
```

```
## [1] 86
```

```r
mcar_test$amount.missing
```

```
##                         ineq       attrib     ideology          sex
## Number Missing  2.333000e+03 7862.0000000 1.317100e+04 68.000000000
## Percent Missing 3.641216e-02    0.1227057 2.055656e-01  0.001061306
##                          age         educ      inc.lad relig.import
## Number Missing  1.420000e+02 3.880000e+02 6509.0000000 1.335000e+03
## Percent Missing 2.216257e-03 6.055687e-03    0.1015888 2.083593e-02
```

Thus, we reject the null hypothesis that the data are MCAR. However, the huge sample makes for nearly certain rejection of the null regardless of the truth of the null hypothesis. However, as a logic check, I will also run the regression analysis with imputed data in the full sample after running the analysis with list wise deletion.

Given the huge data set (and reasons discussed in the manuscript) I opted for list wise deletion because it is likely to not introduce any more bias into the analysis than imputing in excess of 20,000 data points.

```r
# Ideology
wvs <- wvs[ which(wvs$ideology>=1), ]
# Equality
wvs <- wvs[ which(wvs$ineq>=1),]
# Attributions
wvs <- wvs[ which(wvs$attrib==1 | wvs$attrib==2),]
# Sex
wvs <- wvs[ which(wvs$sex>=1),]
# Age
wvs <- wvs[ which(wvs$age>=1),]
# Educ
wvs <- wvs[ which(wvs$educ>=1),]
# Income Ladder
wvs <- wvs[ which(wvs$inc.lad>=1),]
# Religiosity
wvs <- wvs[ which(wvs$relig.import>=1),]
```

Thus, I'm not left with a complete cases data set of 40,031 observations.

Initial Linear Regression

Now getting right to it and converting everything to z-scores and running a simple linear regression of support for economic inequality on attributions for poverty, controlling for political ideology, education, income, religiosity, age, and gender.

```r
# Turning everything into z-scores for the regression
wvs$zineq <- scale(wvs$ineq, center=TRUE,scale=TRUE)
wvs$zideol <- scale(wvs$ideology, center=TRUE,scale=TRUE)
wvs$zattrib <- scale(wvs$attrib, center=TRUE,scale=TRUE)
wvs$zsex <- scale(wvs$sex, center=TRUE,scale=TRUE)
wvs$zage <- scale(wvs$age, center=TRUE,scale=TRUE)
wvs$zeduc <- scale(wvs$educ, center=TRUE,scale=TRUE)
wvs$zinclad <- scale(wvs$inc.lad, center=TRUE,scale=TRUE)
wvs$zrelig.import <- scale(wvs$relig.import, center=TRUE,scale=TRUE)

summary(lm(zineq~zattrib+zideol+zsex+zage+zeduc+zinclad+zrelig.import, data=wvs))
```

```
##
## Call:
## lm(formula = zineq ~ zattrib + zideol + zsex + zage + zeduc +
##     zinclad + zrelig.import, data = wvs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2688 -0.7566  0.0598  0.7993  1.9800
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.752e-15  4.887e-03   0.000   1.0000
## zattrib       -5.727e-02  4.945e-03 -11.581  < 2e-16 ***
## zideol         1.195e-01  4.924e-03  24.275  < 2e-16 ***
## zsex          -9.058e-03  4.896e-03  -1.850   0.0643 .
## zage          -1.236e-02  5.021e-03  -2.462   0.0138 *
## zeduc          1.337e-01  5.281e-03  25.313  < 2e-16 ***
## zinclad        3.819e-02  5.181e-03   7.372 1.71e-13 ***
## zrelig.import -2.406e-02  4.901e-03  -4.909 9.17e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9777 on 40023 degrees of freedom
## Multiple R-squared:  0.04429,    Adjusted R-squared:  0.04412
## F-statistic:   265 on 7 and 40023 DF,  p-value: < 2.2e-16
```

Multilevel Model

Procedural steps

First things first, I just need to merge the two datasets. I already brought in the country level data at the start of this markdown document, so I'll merge GDP and Gini into the WVS data here.

```
wvs$ID <- seq.int(nrow(wvs))

# Inserting inequality
wvs$gini = 0
for(i in 1:length(wvs$ID))
{wvs$gini[i]=country$Gini[which(country$Code == wvs$country[i])]
}
# Inserting GDP
wvs$gdpcap = 0
for(i in 1:length(wvs$ID))
{wvs$gdpcap[i]=country$GDPpercap[which(country$Code == wvs$country[i])]
}

# Converting the new country variables to z scores.
wvs$zgini <- scale(wvs$gini, center=TRUE,scale=TRUE)
wvs$zgdpcap <- scale(wvs$gdpcap, center=TRUE,scale=TRUE)
```

Modeling

First step in running an MLM is determining whether or not it's actually necessary. Here is the null model:

```
summary(lme(zineq~1, data = wvs, random = ~ 1|S003A, method = "ML", na.action = "na.omit"))
```

```
## Linear mixed-effects model fit by maximum likelihood
##  Data: wvs
##       AIC      BIC   logLik
##    110489 110514.8 -55241.5
##
## Random effects:
##  Formula: ~1 | S003A
##         (Intercept)  Residual
## StdDev:   0.2879902 0.9594195
##
## Fixed effects: zineq ~ 1
##                   Value  Std.Error    DF    t-value p-value
## (Intercept) -0.01615016 0.04279758 39985 -0.3773616  0.7059
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.39125493 -0.77014612  0.08513048  0.83397588  2.14890285
##
## Number of Observations: 40031
## Number of Groups: 46
```

Calculating the ICC based on the output of the null model:

```r
ICC <- (.2843273*.2843273)/((.2843273*.2843273)+(.9598677*.9598677))
ICC
```

```
## [1] 0.08066552
```

While the effect of the clustering is small (8%), the data set is large so this is enough to bias the model output. So, moving forward with the MLM. First the predictor only model.

```r
summary(lme(zineq~zattrib, data = wvs, random = ~ 1|S003A, method = "ML", na.action = "na.omit"))
```

```
## Linear mixed-effects model fit by maximum likelihood
##  Data: wvs
##        AIC      BIC    logLik
##   110154.1 110188.5 -55073.03
##
## Random effects:
##  Formula: ~1 | S003A
##         (Intercept)  Residual
## StdDev:   0.2886435 0.9553832
##
## Fixed effects: zineq ~ zattrib
##                   Value  Std.Error    DF    t-value p-value
## (Intercept) -0.01734578 0.04289097 39984  -0.404416  0.6859
## zattrib     -0.09255196 0.00503155 39984 -18.394307  0.0000
##  Correlation:
##         (Intr)
## zattrib 0.002
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.55714260 -0.75418583  0.09084066  0.81893626  2.24508347
##
## Number of Observations: 40031
## Number of Groups: 46
```

And now the full model with all Level 1 and LEvel 2 covariates.

```r
summary(lme(zineq~zattrib+zideol+zsex+zage+zeduc+zinclad+zrelig.import+zgini+zgdpcap, data = wvs, rando
```

```
## Linear mixed-effects model fit by maximum likelihood
##  Data: wvs
##        AIC      BIC    logLik
##   88248.09 88348.59 -44112.04
##
## Random effects:
##  Formula: ~1 | country
##         (Intercept)  Residual
## StdDev:   0.2793627 0.9555429
##
## Fixed effects: zineq ~ zattrib + zideol + zsex + zage + zeduc + zinclad + zrelig.import +      zgini
##                     Value  Std.Error    DF    t-value p-value
## (Intercept)   -0.01057870 0.05578415 32023  -0.189636  0.8496
## zattrib       -0.07119684 0.00565790 32023 -12.583618  0.0000
## zideol         0.10462003 0.00543172 32023  19.260953  0.0000
## zsex          -0.01165118 0.00537530 32023  -2.167541  0.0302
## zage          -0.01477523 0.00573148 32023  -2.577908  0.0099
```

```
## zeduc          0.08976762 0.00618525 32023  14.513171  0.0000
## zinclad        0.08205534 0.00636245 32023  12.896817  0.0000
## zrelig.import  0.00296278 0.00593695 32023   0.499041  0.6178
## zgini          0.02600693 0.05209044    31   0.499265  0.6211
## zgdpcap       -0.14483175 0.08184387    31  -1.769610  0.0866
##  Correlation:
##               (Intr) zattrb zideol zsex   zage   zeduc  zincld zrlg.m
## zattrib        0.008
## zideol        -0.001  0.095
## zsex          -0.001 -0.024  0.000
## zage          -0.003 -0.010  0.005  0.036
## zeduc          0.004 -0.006  0.036  0.014  0.185
## zinclad       -0.008  0.066 -0.014  0.047  0.056 -0.361
## zrelig.import  0.000 -0.001  0.002 -0.016  0.010  0.026  0.024
## zgini          0.166  0.000 -0.004 -0.001  0.016 -0.003  0.006  0.002
## zgdpcap        0.468  0.015  0.004 -0.003 -0.014  0.003 -0.019 -0.009
##               zgini
## zattrib
## zideol
## zsex
## zage
## zeduc
## zinclad
## zrelig.import
## zgini
## zgdpcap       -0.026
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.78056329 -0.75835625  0.08756768  0.80308354  2.39068942
##
## Number of Observations: 32064
## Number of Groups: 34
```

There is the final output - attributions for poverty are related to support for inequality, controlling for both individual and country level covariates.

Missing Data Logic Check

In order to check if list wise deletion resulted in roughly similar outcomes to data with imputed values, I ran a multiple imputation using mice, just to compare the coefficients. First, here is the original model with non-standardized scores.

```r
summary(lm(ineq~attrib+ideology+sex+age+educ+inc.lad+relig.import, data=wvs))
```

```
##
## Call:
## lm(formula = ineq ~ attrib + ideology + sex + age + educ + inc.lad +
##     relig.import, data = wvs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6845 -2.2291  0.1762  2.3550  5.8337
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.961162   0.104763  47.356  < 2e-16 ***
## attrib       -0.370641   0.032003 -11.581  < 2e-16 ***
## ideology      0.150872   0.006215  24.275  < 2e-16 ***
## sex          -0.053388   0.028855  -1.850   0.0643 .
## age          -0.023824   0.009677  -2.462   0.0138 *
## educ          0.177437   0.007010  25.313  < 2e-16 ***
## inc.lad       0.044696   0.006063   7.372 1.71e-13 ***
## relig.import -0.066906   0.013628  -4.909 9.17e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.881 on 40023 degrees of freedom
## Multiple R-squared:  0.04429,    Adjusted R-squared:  0.04412
## F-statistic:   265 on 7 and 40023 DF,  p-value: < 2.2e-16
```

Now, I will run the imputation. This allows me to retain an extra (roughly) 15,000 rows by imputing upwards of 20,000 missing data points. The imputation creates five different imputed datasets. This allows us to make sure we aren't just getting one biased imputation; we do it five times and then pool across them.

```
imp <- mice(key_Vars)
```

```
##
##  iter imp variable
##   1   1  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   1   2  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   1   3  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   1   4  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   1   5  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   2   1  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   2   2  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   2   3  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   2   4  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   2   5  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   3   1  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   3   2  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   3   3  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   3   4  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   3   5  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   4   1  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   4   2  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   4   3  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   4   4  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   4   5  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   5   1  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   5   2  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   5   3  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   5   4  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
##   5   5  ineq  attrib  ideology  sex  age  educ  inc.lad  relig.import
```

```
## Warning in as.POSIXlt.POSIXct(Sys.time()): unknown timezone 'zone/tz/2018c.
## 1.0/zoneinfo/America/Detroit'
```

Now lets re-run this regression with the newly imputed dataset.

```r
fit <- with(imp, lm(ineq~attrib+ideology+sex+age+educ+inc.lad+relig.import))
summary(fit)
```

```
##              term    estimate   std.error  statistic       p.value
## 1   (Intercept)  4.88746678 0.083655238  58.423918  0.000000e+00
## 2        attrib -0.35825201 0.025679652 -13.950812  3.610294e-44
## 3      ideology  0.15116791 0.004962786  30.460291 2.456487e-202
## 4           sex -0.07263329 0.023075782  -3.147598  1.646934e-03
## 5           age -0.03951136 0.007620509  -5.184872  2.168187e-07
## 6          educ  0.18662429 0.005567006  33.523278 2.906714e-244
## 7       inc.lad  0.05199861 0.004875909  10.664392  1.572349e-26
## 8  relig.import -0.05416888 0.010887038  -4.975539  6.523443e-07
## 9   (Intercept)  4.85659290 0.083778514  57.969433  0.000000e+00
## 10       attrib -0.34018202 0.025741092 -13.215524  8.050313e-40
## 11     ideology  0.15144402 0.004981297  30.402529 1.391997e-201
## 12          sex -0.07587161 0.023081844  -3.287069  1.012905e-03
## 13          age -0.03364854 0.007623426  -4.413835  1.017212e-05
## 14         educ  0.18413534 0.005567467  33.073448 7.304850e-238
## 15      inc.lad  0.05214209 0.004877058  10.691302  1.177359e-26
## 16 relig.import -0.05243502 0.010889150  -4.815345  1.472805e-06
## 17  (Intercept)  4.82973956 0.083674587  57.720507  0.000000e+00
## 18       attrib -0.32635651 0.025750269 -12.673907  9.133902e-37
## 19     ideology  0.15047629 0.004994770  30.126773 5.259119e-198
## 20          sex -0.07102475 0.023079321  -3.077419  2.088894e-03
## 21          age -0.03367118 0.007620803  -4.418325  9.963167e-06
## 22         educ  0.18490478 0.005567980  33.208595 8.902053e-240
## 23      inc.lad  0.05138373 0.004876130  10.537808  6.073081e-26
## 24 relig.import -0.05148234 0.010886853  -4.728854  2.262700e-06
## 25  (Intercept)  4.82989299 0.083711483  57.696899  0.000000e+00
## 26       attrib -0.33517979 0.025756901 -13.013203  1.152527e-38
## 27     ideology  0.15688001 0.004970265  31.563710 5.417102e-217
## 28          sex -0.06481414 0.023081380  -2.808070  4.985443e-03
## 29          age -0.03707829 0.007623711  -4.863548  1.155728e-06
## 30         educ  0.18453714 0.005565091  33.159770 4.384100e-239
## 31      inc.lad  0.05145908 0.004875985  10.553577  5.136644e-26
## 32 relig.import -0.05943983 0.010887056  -5.459679  4.787619e-08
## 33  (Intercept)  4.81459089 0.083729577  57.501674  0.000000e+00
## 34       attrib -0.33809187 0.025745671 -13.131989  2.427529e-39
## 35     ideology  0.15745500 0.004973467  31.659001 2.781986e-218
## 36          sex -0.06762249 0.023053900  -2.933234  3.355695e-03
## 37          age -0.03644970 0.007613943  -4.787231  1.694747e-06
## 38         educ  0.18569910 0.005561304  33.391287 2.238848e-242
## 39      inc.lad  0.05325887 0.004868441  10.939614  7.886789e-28
## 40 relig.import -0.05899757 0.010873941  -5.425592  5.797140e-08
```

```r
round(summary(pool(fit)), 2)
```

```
##               estimate std.error statistic      df p.value
## (Intercept)       4.84      0.09     54.15  257.82       0
## attrib           -0.34      0.03    -11.82  101.70       0
## ideology          0.15      0.01     24.73   31.33       0
## sex              -0.07      0.02     -2.99 2385.40       0
## age              -0.04      0.01     -4.46  312.41       0
## educ              0.19      0.01     32.65 2831.10       0
```

```
## inc.lad          0.05       0.00      10.53 4776.37        0
## relig.import    -0.06       0.01      -4.76  267.50        0
```

So above are each of the five regressions with imputed data, and then just one final regression with the pooled imputations. What we see is that when I impute the data using mice, the outcome is not much different than when I used listwise deletion. The key predictor, attributions for poverty, still has a beta of -.34 (as opposed to the -.37 we see with listiwse deletion).

As such, I am comfortable using listwise deletion in this dataset as I then don't have to impute and analyze tens of thousands of observations that don't actually exist.