Pol Sci 630: Problem Set 6 Solutions: Dummy Variables and Interactions

Prepared by: Anh Le (anh.le@duke.edu)

Due Date: Wednesday, Oct 12, 2016 (Beginning of class)

1 Merging data (8 points)

Insert your comments on the assignment that you are grading above the solution in bold and red text. For example write: "GRADER COMMENT: everything is correct! - 8/8 Points" Also briefly point out which, if any, problems were not solved correctly and what the mistake was. See below for more examples.

The most common merging task in political science is to merge datasets based on country-year. The biggest obstacle is that country codes can come in many forms (country name, World Bank code, COW code, ISO2, ISO3, etc.)

This exercise will let you dip your toes in the sea of pain that is merging real world data. You're expected to Google and read help files to figure out two packages: 1) `countrycode`, which converts between different types of country codes, and 2) `psData`, a package that automates the downloading of many common Political Science dataset.

This exercise is not technically hard, just requires you to figure out things on your own.

1.1 Download WDI data

Download GDP per capita ('NY.GDP.PCAP.CD') and FDI ('BX.KLT.DINV.CD.WD') from WDI, 2007-2009, `extra = FALSE`. What country indicators are there?

Note: There should be 792 rows

Solution

```
library(WDI)

## Loading required package:  RJSONIO

d_wdi <- WDI(indicator = c("NY.GDP.PCAP.CD", 'BX.KLT.DINV.CD.WD'),
             start = 2007, end = 2009, extra = FALSE)

names(d_wdi)
```

```
## [1] "iso2c"             "country"           "year"
## [4] "NY.GDP.PCAP.CD"    "BX.KLT.DINV.CD.WD"

nrow(d_wdi)

## [1] 792
```

The two indicators are 'country' and' 'iso2c'

1.2 Download Polity data

Use `PolityGet()` in package `psData` to download Polity data. Download the 'polity2' variable (*not* the entire dataset). Use 'iso3c' as the format for the country code.

What country indicators are there?

Note: There should be 15811 rows

Solution

```
library(psData)
d_polity <- PolityGet(vars = 'polity2', OutCountryID = 'iso3c')

## 577 duplicated values were created when standardising the country
ID with iso3c.
## 749 observations dropped based on missing values of the standardised
ID variable.

names(d_polity)

## [1] "iso3c"   "country" "year"    "polity2"

nrow(d_polity)

## [1] 15811
```

The country indicators are 'country' and 'iso3c'

1.3 Convert country code

To merge WDI and Polity data we must first create a common country ID. (We can't use country name, because there's no guarantee they will be the same). Use package `countrycode` to convert the country code in WDI data from 'iso2c' to 'iso3c'. Store this newly created country code in the WDI data frame.

Solution

```
library(countrycode)
d_wdi$iso3c <- countrycode(d_wdi$iso2c,
                           origin = 'iso2c', destination = 'iso3c')
```

1.4 Merge

Merge the WDI and the Polity data based on 'iso3c' and 'year' (Note: There should be 492 rows).

There are two variables showing country names in the merged dataset. Why? Clean them up so we only have 1 country name variable in the merged dataset.

Solution

```
d_merged <- merge(d_wdi, d_polity, by = c("iso3c", "year"))
nrow(d_merged)

## [1] 489

head(d_merged)

##   iso3c year iso2c   country.x NY.GDP.PCAP.CD BX.KLT.DINV.CD.WD
## 1   AFG 2007    AF Afghanistan       380.4010         188690000
## 2   AFG 2008    AF Afghanistan       384.1317          46033740
## 3   AFG 2009    AF Afghanistan       458.9558         197512727
## 4   AGO 2007    AO      Angola      3151.0224        -893342152
## 5   AGO 2008    AO      Angola      4242.3631        1678971010
## 6   AGO 2009    AO      Angola      3678.9477        2205298180
##     country.y polity2
## 1 Afghanistan      NA
## 2 Afghanistan      NA
## 3 Afghanistan      NA
## 4      Angola      -2
## 5      Angola      -2
## 6      Angola      -2

# Drop extra country variable and clean up
d_merged$country.y <- NULL
colnames(d_merged)[colnames(d_merged) == "country.x"] <- "country"
```

1.5 Check merged result

(Optional) Figure out which country years appear in WDI data but not in Polity data. Note: There should be 255 unmatched records.

In real research, this is useful to check that you are not throwing away data erroneously. There are more than one way to do this and should require some Googling.

Solution

```
# My favorite way
library(dplyr)
```

```
##
## Attaching package:  'dplyr'
## The following objects are masked from 'package:stats':
##
##    filter, lag
## The following objects are masked from 'package:base':
##
##    intersect, setdiff, setequal, union
```

```
d_unmatched <- anti_join(d_wdi, d_polity, by = c("iso3c", "year"))
nrow(d_unmatched)
```

```
## [1] 303
```

2 Factors and Regression with Factors (8 points)

"GRADER COMMENT: everything is correct! - 8/8 Points"

2.1 Dichotomize a continuous variable

Create a new factor variable in your merged dataset, called `polity2_binary` that is 1 (labeled 'democracy') when `polity2` ≥ 0, and 0 (labeled 'dictatorship') otherwise.

Solution

```
d_merged$polity2_binary <- ifelse(d_merged$polity2 >= 0, 1, 0)

d_merged$polity2_binary <- factor(d_merged$polity2_binary,
                                  levels = c(0, 1),
                                  labels = c("dictatorship", "democracy"))
```

2.2 Regression with one binary variable

Regress FDI on the binary variable `polity2_binary`. From the regression result, report the average amount of FDI that democracy and dictatorship gets.

Note: You should know this from the regression result, not from running `mean()`

Solution

```
# Fancy way (dplyr) to rename variables that you'll learn one day
d_merged <- d_merged %>%
  rename(gdppc = NY.GDP.PCAP.CD, fdi = BX.KLT.DINV.CD.WD)
```

```r
m_3a <- lm(fdi ~ polity2_binary, data = d_merged)
summary(m_3a)
```

```
##
## Call:
## lm(formula = fdi ~ polity2_binary, data = d_merged)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -4.583e+10 -1.551e+10 -9.231e+09 -4.631e+09  7.179e+11
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             6.387e+09  4.104e+09   1.556   0.1203
## polity2_binarydemocracy 9.767e+09  4.868e+09   2.007   0.0454 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.821e+10 on 475 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.008405,Adjusted R-squared:  0.006318
## F-statistic: 4.026 on 1 and 475 DF,  p-value: 0.04536
```

The average FDI for dictatorship is 6.386944×10^9 USD.
The average FDI for democracy is 1.6154234×10^{10} USD.

2.3 Regression with interaction and interpretation

Regress FDI against `polity2_binary`, gdppc, and their interaction term.

I want to plot FDI against gdppc with two lines, one representing democracy, the other representing dictatorship (similar to the last plot in the lab). What would be the intercept and slope of these two lines?

Solution

Run regression

```r
m_3b <- lm(fdi ~ polity2_binary + gdppc + polity2_binary * gdppc,
           data = d_merged)
summary(m_3b)
```

```
##
## Call:
## lm(formula = fdi ~ polity2_binary + gdppc + polity2_binary *
##     gdppc, data = d_merged)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
```

```
## -1.372e+11 -6.216e+09 -4.268e+09 -2.084e+09  6.804e+11
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    6.051e+09  4.555e+09   1.329  0.18465
## polity2_binarydemocracy       -4.076e+09  5.482e+09  -0.744  0.45755
## gdppc                          9.052e+04  2.899e+05   0.312  0.75502
## polity2_binarydemocracy:gdppc  9.166e+05  3.152e+05   2.908  0.00382 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.568e+10 on 462 degrees of freedom
##   (23 observations deleted due to missingness)
## Multiple R-squared:  0.1324,Adjusted R-squared:  0.1267
## F-statistic:  23.5 on 3 and 462 DF,  p-value: 3.607e-14
```

Intercept of dictatorship line: 6.0509365×10^9 Intercept of democracy line: 1.9751839×10^9

Slope of dictatorship line: 9.0520835×10^4 Slope of democracy line: 1.0070961×10^6

2.4 Demonstrating substative meaning of coefficients

In research, we usually have to demonstrate the substantive meaning of our regression result. A common way to do it is to give the estimated outcome for a "typical" country, varying one important factor.

For example, imagine that we have a country with median gdppc. What would be its FDI if it were a 1) dictatorship and 2) democracy, holding gdppc at the median value?

Hint: You could either calculate using regression formula, or feed `newdata` to `predict`

Solution

FDI for democracy with median gdppc

```
newdata <- data.frame(polity2_binary = factor("democracy"),
                      gdppc = median(d_merged$gdppc, na.rm=TRUE))

# Using regression formula
m_3b$coefficients['(Intercept)'] + m_3b$coefficients['polity2_binarydemocracy'] +
  m_3b$coefficients['gdppc'] * median(d_merged$gdppc, na.rm=TRUE) +
  m_3b$coefficients['polity2_binarydemocracy:gdppc'] * median(d_merged$gdppc, na.rm=TRUE)

## (Intercept)
##  6039452513

# Using predict
predict(m_3b, newdata = newdata)
```

```
##          1
## 6039452513
```

FDI for dictatorship with median gdppc

```
newdata <- data.frame(polity2_binary = factor("dictatorship"),
                      gdppc = median(d_merged$gdppc, na.rm=TRUE))
predict(m_3b, newdata = newdata)
```

```
##          1
## 6416245186
```