

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

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

Due Date: Friday, Oct 9, 2015, 12 AM (Beginning of Lab)

1. Merging data (8 points)

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.

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! - 4/4 Points" Also briefly point out which, if any, problems were not solved correctly and what the mistake was. See below for more examples.

a)

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 747 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] 747
```

The two indicators are 'country' and 'iso2c'

b)

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'

c)

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)

##
## Attaching package: 'countrycode'
```

```
##
## The following object is masked from 'package:psData':
##
##   countrycode_data

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

d)

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] 492

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       87276201
## 3  AFG 2009  AF Afghanistan      458.9558      213670260
## 4  AGO 2007  AO   Angola      3151.0207     -893342152
## 5  AGO 2008  AO   Angola      4242.3657     1678971010
## 6  AGO 2009  AO   Angola      3678.9492     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"
```

e)

(Challenging, but not 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] 255
```

2. Factors (2 points)

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"))
```

3. Regression with interaction (6 points)

a) 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)

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
## -4.445e+10 -1.554e+10 -9.421e+09 -5.007e+09  7.178e+11
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.868e+09  4.145e+09   1.657   0.0982 .
## polity2_binarydemocracy 9.321e+09  4.917e+09   1.896   0.0586 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.869e+10 on 475 degrees of freedom
## (15 observations deleted due to missingness)
## Multiple R-squared:  0.00751, Adjusted R-squared:  0.00542
## F-statistic: 3.594 on 1 and 475 DF, p-value: 0.05859
```

The average FDI for dictatorship is 6.8682351×10^9 USD.

The average FDI for democracy is 1.6189052×10^{10} USD.

b) Interaction

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

Imagine that we have a country with median `gdppc`. What would be its FDI if it were a 1) dictatorship and 2) democracy? Hint: You could either calculate using regression formula, or feed `newdata` to `predict`

Solution

Run regression

```
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
## -1.345e+11 -6.706e+09 -4.356e+09 -2.047e+09  6.804e+11
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.639e+09  4.598e+09   1.444  0.14942
## polity2_binarydemocracy -4.759e+09  5.536e+09  -0.860  0.39046
## gdppc           8.061e+04  2.932e+05   0.275  0.78352
## polity2_binarydemocracy:gdppc  9.372e+05  3.189e+05   2.939  0.00346 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.616e+10 on 462 degrees of freedom
## (26 observations deleted due to missingness)
## Multiple R-squared:  0.131, Adjusted R-squared:  0.1254
## F-statistic: 23.22 on 3 and 462 DF, p-value: 5.164e-14
```

FDI for democracy with median gdppc

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

##      1
## 5961997155
```

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
## 6962634500
```