Decompose Right Hand Side Variables from Linear Regression

Fan Wang

2020-04-01

Contents

Decompose RHS

> Go to the **RMD**, **R**, **PDF**, or **HTML** version of this file. Go back to fan's REconTools Package, R4Econ Repository (bookdown site), or Intro Stats with R Repository.

One runs a number of regressions. With different outcomes, and various right hand side variables.

What is the remaining variation in the left hand side variable if right hand side variable one by one is set to the average of the observed values.

- Dependency: *R4Econ/linreg/ivreg/ivregdfrow.R*

The code below does not work with categorical variables (except for dummies). Dummy variable inputs need to be converted to zero/one first. The examples are just to test the code with different types of variables.

```r
# Library
library(tidyverse)
library(AER)

# Load Sample Data
setwd('C:/Users/fan/R4Econ/_data/')
df <- read_csv('height_weight.csv')

# Source Dependency
source('C:/Users/fan/R4Econ/linreg/ivreg/ivregdfrow.R')
```

Data Cleaning.

```r
# Convert Variable for Sex which is categorical to Numeric
df <- df
df$male <- (as.numeric(factor(df$sex)) - 1)
summary(factor(df$sex))
```

```
## Female   Male
##  16446  18619
```

```r
summary(df$male)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   0.531   1.000   1.000
```

```r
df.use <- df %>% filter(S.country == 'Guatemala') %>%
  filter(svymthRound %in% c(12, 18, 24))
dim(df.use)
```

```
## [1] 2022   16
```

Setting Up Parameters.

```
# Define Left Hand Side Variab les
var.y1 <- c('hgt')
var.y2 <- c('wgt')
vars.y <- c(var.y1, var.y2)
# Define Right Hand Side Variables
vars.x <- c('prot')
vars.c <- c('male', 'wgt0', 'hgt0', 'svymthRound')
# vars.z <- c('p.A.prot')
vars.z <- c('vil.id')
# vars.z <- NULL
vars.xc <- c(vars.x, vars.c)

# Other variables to keep
vars.other.keep <- c('S.country', 'vil.id', 'indi.id', 'svymthRound')

# Decompose sequence
vars.tomean.first <- c('male', 'hgt0')
var.tomean.first.name.suffix <- '_mh02m'
vars.tomean.second <- c(vars.tomean.first, 'hgt0', 'wgt0')
var.tomean.second.name.suffix <- '_mh0me2m'
vars.tomean.third <- c(vars.tomean.second, 'prot')
var.tomean.third.name.suffix <- '_mh0mep2m'
vars.tomean.fourth <- c(vars.tomean.third, 'svymthRound')
var.tomean.fourth.name.suffix <- '_mh0mepm2m'
list.vars.tomean = list(
#                          vars.tomean.first,
                        vars.tomean.second,
                        vars.tomean.third,
                        vars.tomean.fourth
                        )
list.vars.tomean.name.suffix <- list(
#                                     var.tomean.first.name.suffix,
                                     var.tomean.second.name.suffix,
                                     var.tomean.third.name.suffix,
                                     var.tomean.fourth.name.suffix
                                   )
```

```
# Regressions
# regf.iv from C:\Users\fan\R4Econ\linreg\ivreg\ivregdfrow.R
df.reg.out <- as_tibble(
  bind_rows(lapply(vars.y, regf.iv,
                   vars.x=vars.x, vars.c=vars.c, vars.z=vars.z, df=df)))
# Regressions
# reg1 <- regf.iv(var.y = var.y1, vars.x, vars.c, vars.z, df.use)
# reg2 <- regf.iv(var.y = var.y2, vars.x, vars.c, vars.z, df.use)
# df.reg.out <- as_tibble(bind_rows(reg1, reg2))
```

```
# df.reg.out
```

```r
# Select Variables
str.esti.suffix <- '_Estimate'
arr.esti.name <- paste0(vars.xc, str.esti.suffix)
str.outcome.name <- 'vars_var.y'
arr.columns2select <- c(arr.esti.name, str.outcome.name)
arr.columns2select
```

Obtain Regression Coefficients from somewhere

```
## [1] "prot_Estimate"        "male_Estimate"        "wgt0_Estimate"        "hgt0_Estimate"          "svym
```

```r
# Generate dataframe for coefficients
df.coef <- df.reg.out[,c(arr.columns2select)] %>%
  mutate_at(vars(arr.esti.name), as.numeric) %>% column_to_rownames(str.outcome.name)
df.coef %>%
  kable() %>%
  kable_styling_fc()
```

	prot_Estimate	male_Estimate	wgt0_Estimate	hgt0_Estimate	svymthRound_Estimate
hgt	-0.2714772	1.244735	0.0004430	0.6834853	1.133919
wgt	-59.0727542	489.852902	0.7696158	75.4867897	250.778883

```r
str(df.coef)
```

```
## 'data.frame':    2 obs. of  5 variables:
##  $ prot_Estimate       : num  -0.271 -59.073
##  $ male_Estimate       : num  1.24 489.85
##  $ wgt0_Estimate       : num  0.000443 0.769616
##  $ hgt0_Estimate       : num  0.683 75.487
##  $ svymthRound_Estimate: num  1.13 250.78
```

```r
# Decomposition Step 1: gather
df.decompose_step1 <- df.use %>%
                        filter(svymthRound %in% c(12, 18, 24)) %>%
                        select(one_of(c(vars.other.keep, vars.xc, vars.y))) %>%
                        drop_na() %>%
                        gather(variable, value, -one_of(c(vars.other.keep, vars.xc)))
options(repr.matrix.max.rows=20, repr.matrix.max.cols=20)
dim(df.decompose_step1)
```

Decomposition Step 1

```
## [1] 1382   10
```

```r
head(df.decompose_step1, 10) %>%
  kable() %>%
  kable_styling_fc()
```

```r
# Decomposition Step 2: mutate_at(vars, funs(mean = mean(.)))
# the xc averaging could have taken place earlier, no difference in mean across variables
df.decompose_step2 <- df.decompose_step1 %>%
                        group_by(variable) %>%
                        mutate_at(vars(c(vars.xc, 'value')), funs(mean = mean(.))) %>%
```

S.country	vil.id	indi.id	svymthRound	prot	male	wgt0	hgt0	variable	value
Guatemala	3	1352	18	13.3	1	2545.2	47.4	hgt	70.2
Guatemala	3	1352	24	46.3	1	2545.2	47.4	hgt	75.8
Guatemala	3	1354	12	1.0	1	3634.3	51.2	hgt	66.3
Guatemala	3	1354	18	9.8	1	3634.3	51.2	hgt	69.2
Guatemala	3	1354	24	15.4	1	3634.3	51.2	hgt	75.3
Guatemala	3	1356	12	8.6	1	3911.8	51.9	hgt	68.1
Guatemala	3	1356	18	17.8	1	3911.8	51.9	hgt	74.1
Guatemala	3	1356	24	30.5	1	3911.8	51.9	hgt	77.1
Guatemala	3	1357	12	1.0	1	3791.4	52.6	hgt	71.5
Guatemala	3	1357	18	12.7	1	3791.4	52.6	hgt	77.8

```
                        ungroup()

options(repr.matrix.max.rows=20, repr.matrix.max.cols=20)
dim(df.decompose_step2)
```

Decomposition Step 2

```
## [1] 1382   16
```

```
head(df.decompose_step2,10) %>%
  kable() %>%
  kable_styling_fc_wide()
```

S.country	vil.id	indi.id	svymthRound	prot	male	wgt0	hgt0	variable	value	prot_mean	male_mean	wgt0_mean	hgt0_mean	svymthRound_mean	value_mean
Guatemala	3	1352	18	13.3	1	2545.2	47.4	hgt	70.2	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216
Guatemala	3	1352	24	46.3	1	2545.2	47.4	hgt	75.8	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216
Guatemala	3	1354	12	1.0	1	3634.3	51.2	hgt	66.3	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216
Guatemala	3	1354	18	9.8	1	3634.3	51.2	hgt	69.2	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216
Guatemala	3	1354	24	15.4	1	3634.3	51.2	hgt	75.3	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216
Guatemala	3	1356	12	8.6	1	3911.8	51.9	hgt	68.1	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216
Guatemala	3	1356	18	17.8	1	3911.8	51.9	hgt	74.1	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216
Guatemala	3	1356	24	30.5	1	3911.8	51.9	hgt	77.1	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216
Guatemala	3	1357	12	1.0	1	3791.4	52.6	hgt	71.5	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216
Guatemala	3	1357	18	12.7	1	3791.4	52.6	hgt	77.8	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216

```

ff_lr_decompose_valadj <- function(df, df.coef, vars.tomean, str.esti.suffix) {
  new_value <- (df$value +
                rowSums((df[paste0(vars.tomean, '_mean')] - df[vars.tomean])
                        *df.coef[df$variable, paste0(vars.tomean, str.esti.suffix)]))
  return(new_value)
}
```

Decomposition Step 3 Non-Loop

```

df.decompose_step3 <- df.decompose_step2
for (i in 1:length(list.vars.tomean)) {
    var.decomp.cur <- (paste0('value', list.vars.tomean.name.suffix[[i]]))
    vars.tomean <- list.vars.tomean[[i]]
    var.decomp.cur
    df.decompose_step3 <- df.decompose_step3 %>%
      mutate((!!var.decomp.cur) :=
               ff_lr_decompose_valadj(., df.coef, vars.tomean, str.esti.suffix))

}
```

```r
dim(df.decompose_step3)
```

Decomposition Step 3 With Loop

```
## [1] 1382   19
```

```r
head(df.decompose_step3, 10) %>%
  kable() %>%
  kable_styling_fc_wide()
```

S.country	vil.id	indi.id	svymthRound	prot	male	wgt0	hgt0	variable	value	prot_mean	male_mean	wgt0_mean	hgt0_mean	svymthRound_mean	value_mean	value_mh0me2m	value_mh0mep2m	value_mh0mepm2m
Guatemala	3	1352	18	13.3	1	2545.2	47.4	hgt	70.2	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216	73.19390	71.19903	71.68148
Guatemala	3	1352	24	46.3	1	2545.2	47.4	hgt	75.8	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216	78.79390	85.75778	79.43671
Guatemala	3	1354	12	1.0	1	3634.3	51.2	hgt	66.3	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216	63.61689	58.28285	65.56882
Guatemala	3	1354	18	9.8	1	3634.3	51.2	hgt	69.2	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216	66.51689	63.57185	64.05430
Guatemala	3	1354	24	15.4	1	3634.3	51.2	hgt	75.3	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216	72.61689	71.19213	64.87106
Guatemala	3	1356	12	8.6	1	3911.8	51.9	hgt	68.1	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216	64.33707	61.06626	68.35222
Guatemala	3	1356	18	17.8	1	3911.8	51.9	hgt	74.1	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216	70.33707	69.56385	70.04630
Guatemala	3	1356	24	30.5	1	3911.8	51.9	hgt	77.1	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216	73.33707	76.01161	69.69055
Guatemala	3	1357	12	1.0	1	3791.4	52.6	hgt	71.5	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216	66.83353	61.49949	68.78545
Guatemala	3	1357	18	12.7	1	3791.4	52.6	hgt	77.8	20.64819	0.5499276	3312.297	49.75137	18.42547	73.41216	73.13353	70.97578	71.45823

```r
df.decompose_step3 %>%
        select(variable, contains('value')) %>%
        group_by(variable) %>%
        summarize_all(funs(mean = mean, var = var)) %>%
        select(matches('value')) %>% select(ends_with("_var")) %>%
        mutate_if(is.numeric, funs( frac = (./value_var))) %>%
        mutate_if(is.numeric, round, 3) %>%
  kable() %>%
  kable_styling_fc_wide()
```

value_var	value_mean_var	value_mh0me2m_var	value_mh0mep2m_var	value_mh0mepm2m_var	value_var_frac	value_mean_var_frac	value_mh0me2m_var_frac	value_mh0mep2m_var_frac	value_mh0mepm2m_var_frac
21.864	NA	25.35	49.047	23.06	1	NA	1.159	2.243	1.055
2965693.245	NA	2949187.64	4192769.518	3147506.60	1	NA	0.994	1.414	1.061

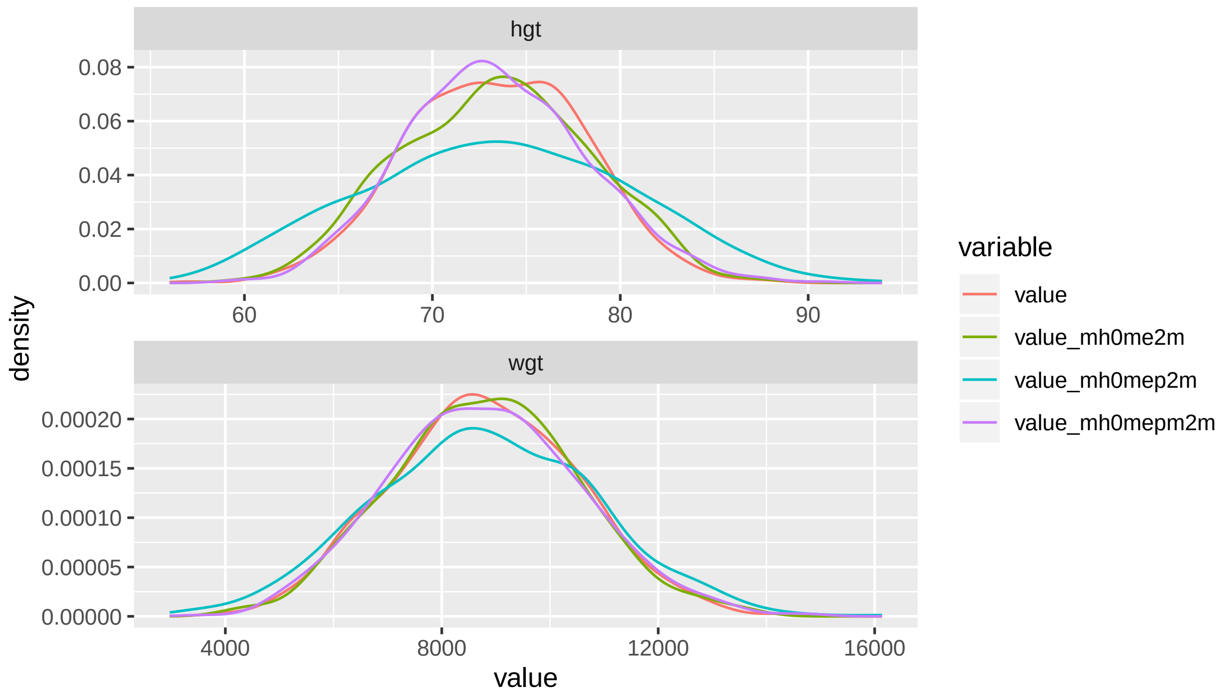
Decomposition Step 4 Variance

Graphical Results Graphically, difficult to pick up exact differences in variance, a 50 percent reduction in variance visually does not look like 50 percent. Intuitively, we are kind of seeing standard deviation, not variance on the graph if we think abou the x-scale.

```r
head(df.decompose_step3 %>%
    select(variable, contains('value'), -value_mean), 10) %>%
  kable() %>%
  kable_styling_fc()
```

variable	value	value_mh0me2m	value_mh0mep2m	value_mh0mepm2m
hgt	70.2	73.19390	71.19903	71.68148
hgt	75.8	78.79390	85.75778	79.43671
hgt	66.3	63.61689	58.28285	65.56882
hgt	69.2	66.51689	63.57185	64.05430
hgt	75.3	72.61689	71.19213	64.87106
hgt	68.1	64.33707	61.06626	68.35222
hgt	74.1	70.33707	69.56385	70.04630
hgt	77.1	73.33707	76.01161	69.69055
hgt	71.5	66.83353	61.49949	68.78545
hgt	77.8	73.13353	70.97578	71.45823

```
df.decompose_step3 %>%
    select(variable, contains('value'), -value_mean) %>%
    rename(outcome = variable) %>%
    gather(variable, value, -outcome) %>%
    ggplot(aes(x=value, color = variable, fill = variable)) +
        geom_line(stat = "density") +
        facet_wrap(~ outcome, scales='free', nrow=2)
```




```
head(df.decompose_step2[vars.tomean.first],3)
head(df.decompose_step2[paste0(vars.tomean.first, '_mean')], 3)
head(df.coef[df.decompose_step2$variable,
             paste0(vars.tomean.first, str.esti.suffix)], 3)
df.decompose.tomean.first <- df.decompose_step2 %>%
    mutate(pred_new = df.decompose_step2$value +
        rowSums((df.decompose_step2[paste0(vars.tomean.first, '_mean')]
                 - df.decompose_step2[vars.tomean.first])
            *df.coef[df.decompose_step2$variable,
                     paste0(vars.tomean.first, str.esti.suffix)])) %>%
        select(variable, value, pred_new)
head(df.decompose.tomean.first, 10)
df.decompose.tomean.first %>%
        group_by(variable) %>%
        summarize_all(funs(mean = mean, sd = sd))  %>%
  kable() %>%
  kable_styling_fc()
```

Additional Decomposition Testings Note the r-square from regression above matches up with the 1 - ratio below. This is the proper decomposition method that is equivalent to r2.

variable	value_mean	pred_new_mean	value_sd	pred_new_sd
hgt	73.41216	73.41216	4.675867	4.534947
wgt	8807.87656	8807.87656	1722.118824	1695.221845

```
df.decompose_step2 %>%
    mutate(pred_new = df.decompose_step2$value +
        rowSums((df.decompose_step2[paste0(vars.tomean.second, '_mean')]
                 - df.decompose_step2[vars.tomean.second])
            *df.coef[df.decompose_step2$variable,
                     paste0(vars.tomean.second, str.esti.suffix)])) %>%
        select(variable, value, pred_new) %>%
        group_by(variable) %>%
        summarize_all(funs(mean = mean, var = var)) %>%
        mutate(ratio = (pred_new_var/value_var))  %>%
  kable() %>%
  kable_styling_fc()
```

variable	value_mean	pred_new_mean	value_var	pred_new_var	ratio
hgt	73.41216	73.41216	2.186374e+01	25.3504	1.1594724
wgt	8807.87656	8807.87656	2.965693e+06	2949187.6357	0.9944345