Optional Reading: Weighted Least Squares (WLS)

(Jing Li, Miami University)

1. This note provides more discussion about WLS, whose key idea is that each observation in the data is assigned a weight based on its variance. Observations with higher variance (i.e., more uncertainty, or noisier) are given less weight, and observations with lower variance (i.e., more precision, or more signal) are given more weight. In particular, this note is concerned with using aggregate data, for which the weight is related to population.

2. Assume individual-level data is homoskedasitic, i.e., $var(y_i) = \sigma_y^2 = constant, \ \forall i$. The regression at individual level is

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

   OLS applied to regression (1) is BLUE.

3. Sometime we need to work with date aggregated at, say, state level. Suppose state j has n_j individuals (population is n_j). Then it follows that the aggregate data of sum is not homoskedastic

$$var\left(\sum_{i=1}^{n_j} y_i\right) = n_j \sigma_y^2 = h(n_j) \neq constant \tag{2}$$

   where h makes it clear that the variance of aggregate data is a function of population. If population varies across states, the issue of heteroskedasticity arises! This is an example of the function form of heteroskedasticity being known.

4. The regression with aggregate state level data is

$$\sum_{i=1}^{n_j} y_i = \beta_0 + \beta_1 \sum_{i=1}^{n_j} x_i + u_j \tag{3}$$

   where $u_j = \sum_{i=1}^{n_j} u_i$ is the state-level error term. The heteroskedasticity shown in (2) implies that it is necessary to employ heterskedasticity-robust standard error, t value, and p value for the OLS estimate of state-level regression (3). OLS now is not BLUE.

5. Furthermore, the heteroskedasticity shown in (2) motivates a transformed regression

that is obtained by dividing the square root of n_j

$$\frac{\sum_{i=1}^{n_j} y_i}{\sqrt{n_j}} = \beta_0 \frac{1}{\sqrt{n_j}} + \beta_1 \frac{\sum_{i=1}^{n_j} x_i}{\sqrt{n_j}} + \frac{u_j}{\sqrt{n_j}} \tag{4}$$

The benefit is that the new error term in the transformed regression is homoskedastic

$$var\left(\frac{u_j}{\sqrt{n_j}}\right) = \sigma_u^2 = constant$$

Thus, the conditions of Gauss-Markov theorem are satisfied by the transformed regression.

6. WLS is the OLS applied to the transformed regression (4). For this example, some people say that the weight is

$$weight = \frac{1}{\sqrt{n_j}} = \frac{1}{\sqrt{population_j}}$$

but R uses the label for the squared weight, or

$$R\ weight = \frac{1}{population_j}$$

Notice that there is no constant term (intercept) in the transformed regression since $\frac{1}{\sqrt{n_j}} \neq 1$.

7. Next we use simulated or toy data as an illustration.

```
> # individual level data
> set.seed(11111)
> pop = c(10,20,40,80)
> state = rep(1:4,times=pop)
> n = length(state)
> x = rnorm(n)
> beta1 = 1
> y = beta1*x+rnorm(n)
> d = data.frame(y,x,state)
> head(d)
           y            x state
```

```
1 -0.2490008  0.003630422     1
2  1.4481986  0.798122717     1
3  0.2542727  0.893397342     1
4  0.3756962 -1.195893897     1
5 -2.3783292 -1.269878611     1
6 -0.8365046  0.181969053     1
```

There are four states $j = 1, 2, 3, 4$, with populations n_j being 10, 20, 40, and 80 for each state. The true values of coefficients are $\beta_1 = 1, \beta_0 = 0$.

8. The result of OLS applied to individual level data, or regression (1), is

```
> # OLS with individual-level data
> summary(lm(y~x,data=d))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02948    0.08532  -0.346     0.73
x            0.92972    0.08596  10.816   <2e-16 ***
Residual standard error: 1.031 on 148 degrees of freedom
Multiple R-squared:  0.4415,        Adjusted R-squared:  0.4377
F-statistic:   117 on 1 and 148 DF,  p-value: < 2.2e-16
```

(a) $\hat{\beta}_1 = 0.92972$, close to true value 1

(b) standard error 0.08596 is correct, because data are homoskedastic at individual level

(c) the residual standard error 1.031 is close to true value 1.

(d) In short, every result is correct, and OLS with individual level data is BLUE

9. Next, we consider state-level aggregate data.

```
> # state-level data
> da = aggregate(cbind(y,x)~state,data=d,FUN=sum)
> da
  state           y            x
1     1   0.4121972   0.01454022
2     2 -10.6683936 -12.12418429
```

```
3      3  -1.3616950  -2.14848310
4      4 -15.4928033 -10.14507910
> sum(d$y[1:10])
[1] 0.4121972
> sum(d$x[1:10])
[1] 0.01454022
```

For instance, state 1 has 10 individuals, and its aggregate y and x values are 0.4121972 and 0.01454022.

10. The estimated state-level regression (3) is

```
> # OLS with state-level data
> summary(lm(y~x,data=da))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3856     2.8226   0.137   0.9038
x             1.1742     0.3538   3.318   0.0801 .
Residual standard error: 3.637 on 2 degrees of freedom
Multiple R-squared:  0.8463,        Adjusted R-squared:  0.7694
F-statistic: 11.01 on 1 and 2 DF,  p-value: 0.08006
```

The residual standard error 3.637 is greater than the individual-level RSE 1.031, consistent with that $u_j = \sum_{i=1}^{n_j} u_i$. The reported standard error, t value, and p value (in blue color) are wrong since they assume homoskedasticity.

11. The heteroskedasticity-robust results are below (in green color)

```
> library("lmtest")
> library("sandwich")
> mols = lm(y~x,data=da)
> coeftest(mols, vcov = vcovHC(mols, type = "HC0"))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.38561    0.48848  0.7894  0.51260
x            1.17415    0.23832  4.9269  0.03881
```

12. The slope coefficient 1.1742 estimated by OLS can be improved because OLS treats the four states equally by assigning equal weights to them. Nevertheless, state 1 should

receive higher weight thanks to its smaller variance. Next, we report the WLS results

```
> # WLS with state-level data
> summary(lm(y~x, weight=1/pop,data=da))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3700     1.4197   0.261   0.8188
x             1.0051     0.2078   4.837   0.0402 *
Residual standard error: 0.4857 on 2 degrees of freedom
Multiple R-squared:  0.9212,      Adjusted R-squared:  0.8819
F-statistic: 23.39 on 1 and 2 DF,  p-value: 0.04019
```

The WLS estimate 1.0051 is much closer to the true value 1 than the OLS estimate 1.17415. The se of WLS 0.2078 is less than the heteroskedasticity-robust se 0.23832. Both findings highlight the superiority of WLS over OLS. We can explicitly fit the transformed regression using codes below

```
> yw = da$y/sqrt(pop)
> xw = da$x/sqrt(pop)
> cw = rep(1,4)/sqrt(pop)
> coef(lm(yw~cw+xw-1))
       cw        xw
0.3699852 1.0050859
```

13. Another possibility is that the state-level data is average (eg., GDP per capita) rather than sum (eg., total GDP). The regression with average data at the state-level is

$$\frac{\sum_{i=1}^{n_j} y_i}{n_j} = \beta_0 + \beta_1 \frac{\sum_{i=1}^{n_j} x_i}{n_j} + \bar{u}_j \tag{5}$$

where $\bar{u}_j = \frac{\sum_{i=1}^{n_j} u_i}{n_j}$. In this case, the heteroskedasticity is present as well, but has a different math form

$$var\left(\frac{\sum_{i=1}^{n_j} y_i}{n_j}\right) = \frac{\sigma_y^2}{n_j} = h(n_j) \neq constant$$

We can show the proper weight is $\sqrt{n_j}$,

$$var\left(\sqrt{n_j}\bar{u}_j\right) = \sigma_u^2 = constant$$

In terms of R weight, we need to use

$$R\ weight = population_j$$

So states with big population receive big weights.

14. Results of applying WLS to state average data are

```
> # WLS with state-level average data
> da = aggregate(cbind(y,x)~state,data=d,FUN=mean)
> summary(lm(y~x, weight=pop,data=da))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.04141    0.04603  -0.900   0.4633
x            0.85643    0.19057   4.494   0.0461 *
Residual standard error: 0.4167 on 2 degrees of freedom
Multiple R-squared:  0.9099,        Adjusted R-squared:  0.8648
F-statistic:  20.2 on 1 and 2 DF,  p-value: 0.04611
```

15. We can use matrix algebra to run the transformed regression, and duplicate the coefficients above

```
> YW = da$y*sqrt(pop)
> XW = cbind(da$x,rep(1,4))*sqrt(pop)
> bwls = solve(t(XW)%*%XW)%*%(t(XW)%*%YW)
> bwls
            [,1]
[1,]  0.85642822
[2,] -0.04140733
```

16. To sum up, WLS is preferred over OLS when aggregate data are used. The form of weights depends on whether data is average or sum. OLS can still be used, but robust standard error, t value, and p value should be reported.