

## Eco311 Optional Reading: Cluster Robust Standard Error (CRSE)

(Jing Li, Miami University)

1. Recall that for **i.i.d sample**, we can show the variance of sample means is

$$\begin{aligned} \text{var}(\bar{y}) &= \text{var}\left(\frac{y_1 + y_2 + \dots + y_n}{n}\right) = \\ &= \frac{\sum_i \text{var}(y_i) + \sum_i \sum_j \text{cov}(y_i, y_j)}{n^2} = \frac{\sum_i \sigma^2 + \sum_i \sum_j 0}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \quad (1) \end{aligned}$$

Notice that we let covariance  $\text{cov}(y_i, y_j) = 0$  since  $y_i$  and  $y_j$  are uncorrelated.

2. When the assumption of independence fails, formula (1) becomes invalid since it ignores those non-zero covariances.
3. In the presence of *positive* correlation  $\text{cov}(y_i, y_j) > 0$ , formula (1) is wrong by *underestimating* the variation of  $\bar{y}$ . In other words, if a correct new formula is used, standard error would become larger, t value would become smaller, and confidence interval would become wider.
4. When running regression, the i.i.d assumption fails when data are in **clusters** or groups, and generally speaking, we believe data within a cluster are correlated.
5. For instance, kids born in the same family can be seen as one cluster—their heights are correlated since they share the same parents. When we run regression, we need to adjust the traditional standard error in order to account for this within-family correlation. The new standard error is called **cluster robust standard error** (crse).
6. This situation is similar to heteroskedasticity, and recall that we use heteroskedasticity robust standard error to account for varying variances across observations. In short, heteroskedasticity and clustering data are two examples in which i.i.d assumption fails and traditional standard error is wrong.
7. As an illustration, we look at the Galton data in **mosaicData** package

```
> library(mosaicData)
> data(Galton)
> attach(Galton)
> head(Galton)
```

	family	father	mother	sex	height	nkids
1	1	78.5	67.0	M	73.2	4
2	1	78.5	67.0	F	69.2	4
3	1	78.5	67.0	F	69.0	4
4	1	78.5	67.0	F	69.0	4
5	2	75.5	66.5	M	73.5	4
6	2	75.5	66.5	M	72.5	4

The first four observations are in the first family (1st cluster). The three daughters are all tall (heights are 69.2, 69, 69) because their dad is tall. This notes show how to account for this within-family correlation when running regressions.

8. The `lm` function ignores that within-family correlation, and reports incorrect standard error, t value, and p value

```
> m = lm(height~father+sex)
> summary(m)$coef
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	34.4611308	2.13628139	16.13136	1.399998e-51
father	0.4278217	0.03078509	13.89704	6.693554e-40
sexM	5.1760424	0.15210598	34.02919	1.577629e-163

9. We can obtain the **correct** cluster robust standard error with `coeftest` function:

```
> library(sandwich)
> library(lmtest)
> coeftest(m, vcov = vcovCL, cluster = family)
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	34.461131	3.108462	11.0862	< 2.2e-16 ***
father	0.427822	0.044735	9.5634	< 2.2e-16 ***
sexM	5.176042	0.161969	31.9571	< 2.2e-16 ***

As expected, the cluster robust standard error is greater than the traditional standard error and t value is smaller than old one. For instance,  $0.044735 > 0.03078509$ ,  $9.5634 < 13.89704$ . This finding confirms that kids' heights within a family are indeed *positively* correlated.

10. Notice that the estimates for  $\beta_i$  remain unchanged.

11. To fully understand CRSE, we need to

(a) use matrix algebra to write the OLS estimate as

$$\hat{\beta} = \beta + (X'X)^{-1}X'U$$

where  $X$  is the  $n$  by  $k$  matrix for regressors, and  $U$  is the  $n$  by one vector of error terms.

(b) Conditional on  $X$ , the *variance-covariance matrix* of  $\hat{\beta}$  is

$$E((\hat{\beta} - \beta)(\hat{\beta} - \beta)'|X) = (X'X)^{-1}X'E(UU')X(X'X)^{-1}$$

(c) Next we partition  $X$  and  $U$  into  $c$  blocks, one block for each cluster (family):

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_c \end{pmatrix}, \quad U = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_c \end{pmatrix},$$

(d) The term in the middle of  $(X'X)^{-1}X'E(UU')X(X'X)^{-1}$  can be written as

$$\Omega \equiv X'E(UU')X = (x'_1, x'_2, \dots, x'_c) \begin{pmatrix} E(u_1u'_1) & 0 & \dots & 0 \\ 0 & E(u_2u'_2) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & E(u_cu'_c) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_c \end{pmatrix}$$

where we assume **independence across-clusters**:  $E(u_iu'_j) = 0, \forall i \neq j$ .

(e)  $\Omega$  is unknown since error term  $u$  is unobserved. We estimate  $\Omega$  using

$$\hat{\Omega} = (x'_1, x'_2, \dots, x'_c) \begin{pmatrix} \hat{u}_1\hat{u}'_1 & 0 & \dots & 0 \\ 0 & \hat{u}_2\hat{u}'_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \hat{u}_c\hat{u}'_c \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_c \end{pmatrix} = \sum_{i=1}^c x'_i\hat{u}_i\hat{u}'_ix_i$$

Basically we replace  $u$  with  $\hat{u}$ . Notice that  $\hat{u}_i\hat{u}'_i$  is the *outer product* of vector of

residuals in the  $i$ -th cluster.

(f) Finally, we multiply by a correction term

$$\text{cluster robust variance} = \frac{n-1}{n-p} \frac{c}{c-1} (X'X)^{-1} \widehat{\Omega} (X'X)^{-1} = \text{sandwich}$$

where  $n$  is sample size,  $p$  is the number of regressors (including constant term), and  $c$  is the number of clusters.

12. The R codes below compute the crse explicitly

```
> uhat = resid(m)
> id = as.numeric(family)
> x = cbind(rep(1,length(father)),father,as.numeric(sex=="M"))
> p = ncol(x)
> n = length(father)
> xpxinv = solve(t(x)%*%x)
> omega = matrix(rep(0,p^2), nrow=p)
> for (i in unique(id)) {
+ xi = x[id==i, , drop=FALSE]
+ omega = omega + t(xi)%*%tcrossprod(uhat[id==i])%*%xi
+ }
> sandw = xpxinv%*%omega%*%xpxinv
> c = length(unique(id))
> adj = (n-1)*c/(n-p)/(c-1)
> sqrt(diag(adj*sandw))
      father
3.10846241 0.04473515 0.16196856
```

The standard error 0.04473515 is the same as that reported by `lmtest` function