

# Regression

## Class Notes

Manuel Arellano

Revised: January 5, 2021

## 1 Means and predictors

Given some data  $\{y_1, \dots, y_n\}$  we could calculate a mean  $\bar{y} = (1/n) \sum_{i=1}^n y_i$  as a single quantity that summarizes the  $n$  data points.  $\bar{y}$  is an optimal predictor that minimizes mean squared error:

$$\bar{y} = \arg \min_a \sum_{i=1}^n (y_i - a)^2.$$

Now if we have data on two variables for the same units  $\{y_i, x_i\}_{i=1}^n$ , we can get a better predictor of  $y$  using the additional information in  $x$  calculating the regression line  $\hat{y}_i = \hat{a} + \hat{b}x_i$  where

$$(\hat{a}, \hat{b}) = \arg \min_{a,b} \sum_{i=1}^n (y_i - a - bx_i)^2.$$

More generally, if  $x_i$  is a vector  $x_i = (1, x_{2i}, \dots, x_{ki})'$ , we calculate the linear predictor  $\hat{y}_i = x_i' \hat{\beta}$  where

$$\hat{\beta} = \arg \min_b \sum_{i=1}^n (y_i - x_i' b)^2. \quad (1)$$

**The algebra of linear predictors** First order conditions of (1) are

$$\sum_{i=1}^n x_i (y_i - x_i' \hat{\beta}) = 0. \quad (2)$$

If  $\sum_{i=1}^n x_i x_i'$  is full rank (which requires  $n \geq k$ ) there is a unique solution:

$$\hat{\beta} = \left( \sum_{i=1}^n x_i x_i' \right)^{-1} \sum_{i=1}^n x_i y_i. \quad (3)$$

We may use the compact notation  $X'X = \sum_{i=1}^n x_i x_i'$  and  $X'y = \sum_{i=1}^n x_i y_i$  where  $y = (y_1, \dots, y_n)'$  and  $X = (x_1, \dots, x_n)'$ .

Denoting residuals as  $\hat{u}_i = y_i - x_i' \hat{\beta}$ , from the first order conditions (2) we can immediately say that as long as a constant term is included in  $x_i$ :

$$\frac{1}{n} \sum_{i=1}^n \hat{u}_i = 0, \quad \frac{1}{n} \sum_{i=1}^n x_{ji} \hat{u}_i = 0 \text{ for } j = 2, \dots, k.$$

Therefore, the mean of the residuals is zero and the covariance between the residuals and each of the  $x$  variables is also zero. Moreover, since  $\hat{y}_i$  is a linear combination of  $x_i$ , the covariance between  $\hat{u}_i$  and  $\hat{y}_i$  is also zero. We conclude that a linear regression decomposes  $y_i$  into two orthogonal components:

$$y_i = \hat{y}_i + \hat{u}_i,$$

so that  $\widehat{Var}(y_i) = \widehat{Var}(\hat{y}_i) + \widehat{Var}(\hat{u}_i)$ . An  $R^2$  measures the fraction of the variance of  $y_i$  that is accounted by  $\hat{y}_i$ :

$$R^2 = \frac{\widehat{Var}(\hat{y}_i)}{\widehat{Var}(y_i)}.$$

## 2 Consistency and asymptotic normality of linear predictors

If our data  $\{y_i, x_i\}_{i=1}^n$  are a random sample from some population we can study the properties of  $\hat{\beta}$  as an estimator of the corresponding population quantity:

$$\beta = [E(x_i x_i')]^{-1} E(x_i y_i), \quad (4)$$

where we require that  $E(x_i x_i')$  has full rank.

Letting the population linear predictor error be  $u_i = (y_i - x_i' \beta)$ , the estimation error is

$$\hat{\beta} - \beta = \left( \frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^n x_i u_i.$$

Clearly,  $E(x_i u_i) = 0$ , since  $\beta$  solves the first-order conditions  $E[x_i (y_i - x_i' \beta)] = 0$ . By Slutsky's theorem and the law of large numbers:

$$\text{plim}_{n \rightarrow \infty} (\hat{\beta} - \beta) = \left( \text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n x_i u_i = [E(x_i x_i')]^{-1} E(x_i u_i) = 0. \quad (5)$$

Therefore,  $\hat{\beta}$  is a consistent estimator of  $\beta$ .

Moreover, because of the central limit theorem

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i u_i \xrightarrow{d} \mathcal{N}(0, V)$$

where  $V = E(u_i^2 x_i x_i')$ . In addition, using Cramér's theorem we can assert that

$$\sqrt{n} (\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, W) \quad (6)$$

where

$$W = [E(x_i x_i')]^{-1} E(u_i^2 x_i x_i') [E(x_i x_i')]^{-1}, \quad (7)$$

and also for individual coefficients:

$$\sqrt{n} (\hat{\beta}_j - \beta_j) \xrightarrow{d} \mathcal{N}(0, w_{jj}) \quad (8)$$

where  $w_{jj}$  is the  $j$ -th diagonal element of  $W$ .

**Asymptotic standard errors and confidence intervals** A consistent estimator of  $W$  is:

$$\widehat{W} = \left( \frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2 x_i x_i' \right) \left( \frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1}. \quad (9)$$

The quantity  $\sqrt{\widehat{w}_{jj}/n}$  is called an asymptotic standard error of  $\hat{\beta}_j$ , or simply a standard error. It is an approximate standard deviation of  $\hat{\beta}_j$  in a large sample, and it is used as a measure of the precision of an estimate.

Due to Cramér's theorem:

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{w}_{jj}/n}} \xrightarrow{d} \mathcal{N}(0, 1). \quad (10)$$

The use of this statement is in calculating approximate confidence intervals. A 95% large sample confidence interval is:

$$\left( \widehat{\beta}_j - 1.96\sqrt{\widehat{w}_{jj}/n}, \widehat{\beta}_j + 1.96\sqrt{\widehat{w}_{jj}/n} \right). \quad (11)$$

### 3 Classical regression model

A linear predictor is the best linear approximation to the conditional mean of  $y$  given  $x$  in the sense:

$$\beta = \arg \min_b E \left\{ [E(y_i | x_i) - x_i' b]^2 \right\}. \quad (12)$$

That is,  $x_i' \beta$  minimizes the mean squared approximation errors where the mean is taken with respect to the distribution of  $x$ . Therefore, changing the distribution of  $x$  will change the linear predictor unless the conditional mean is linear, in which case  $E(y_i | x_i) = x_i' \beta$ .

If  $E \left\{ [E(y_i | x_i) - x_i' \beta]^2 \right\}$  is not zero or close to zero,  $x_i' \widehat{\beta}$  will not be a very informative summary of the dependence in mean between  $y$  and  $x$ . In general, the use of a linear predictor is hard to motivate if the conditional mean is notoriously nonlinear.

The classical regression model is a linear model that makes the following two assumptions:

$$E(y | X) = X\beta \quad (A1)$$

$$Var(y | X) = \sigma^2 I_n. \quad (A2)$$

The first assumption (A1) asserts that  $E(y_i | x_1, \dots, x_n) = x_i' \beta$  for all  $i$ . This assumption contains two parts. The first one is that  $E(y_i | x_1, \dots, x_n) = E(y_i | x_i)$ ; this part of the assumption will always hold if  $\{y_i, x_i\}_{i=1}^n$  is a random sample and is sometimes called strict exogeneity. The second part is the linearity assumption  $E(y_i | x_i) = x_i' \beta$ . Under A1  $\widehat{\beta}$  is an unbiased estimator:

$$E(\widehat{\beta} | X) = (X'X)^{-1} X'E(y | X) = \beta \quad (13)$$

and therefore also  $E(\widehat{\beta}) = \beta$  by the law of iterated expectations.

The second assumption (A2) says that  $Var(y_i | x_1, \dots, x_n) = \sigma^2$  and  $Cov(y_i, y_j | x_1, \dots, x_n) = 0$  for all  $i$  and  $j$ . Under random sampling  $Var(y_i | x_1, \dots, x_n) = Var(y_i | x_i)$  and  $Cov(y_i, y_j | x_1, \dots, x_n) = 0$  always hold. Assumption A2 also requires that  $Var(y_i | x_i)$  is constant for all  $x_i$  and this situation is called homoskedasticity. The alternative situation when  $Var(y_i | x_i)$  may vary with  $x_i$  is called heteroskedasticity. When the data are time series the zero covariance condition  $Cov(y_i, y_j | x_1, \dots, x_n) = 0$  is called lack of autocorrelation.

Under A2 the variance matrix of  $\widehat{\beta}$  given  $X$  is

$$\text{Var}(\widehat{\beta} | X) = \sigma^2 (X'X)^{-1}. \quad (14)$$

Moreover, under A2 since  $E(u_i^2 x_i x_i') = \sigma^2 E(x_i x_i')$  the sandwich formula (7) becomes

$$W = \sigma^2 [E(x_i x_i')]^{-1}. \quad (15)$$

To obtain an unbiased estimator of  $\sigma^2$  note that under A2, letting  $M = I_n - X(X'X)^{-1}X'$ , we have

$$E(\widehat{u}'\widehat{u}) = E[E(u'Mu | X)] = E(\text{tr}[ME(uu' | X)]) = \sigma^2 \text{tr}(M) = \sigma^2(n - k), \quad (16)$$

so that an unbiased estimator of  $\sigma^2$  is given by the degrees of freedom corrected residual variance:

$$\widehat{\sigma}^2 = \frac{\widehat{u}'\widehat{u}}{n - k}. \quad (17)$$

**Sampling distributions under conditional normality** Consider as a third assumption:

$$y | X \sim \mathcal{N}(X\beta, \sigma^2 I_n). \quad (A3)$$

Under A3:

$$\widehat{\beta} | X \sim \mathcal{N}(\beta, \sigma^2 (X'X)^{-1}), \quad (18)$$

so that also

$$\widehat{\beta}_j | X \sim \mathcal{N}(\beta_j, \sigma^2 a_{jj}) \quad (19)$$

where  $a_{jj}$  is the  $j$ -th diagonal element of  $(X'X)^{-1}$ . Moreover, conditionally and unconditionally we have

$$z_j \equiv \frac{\widehat{\beta}_j - \beta_j}{\sqrt{\sigma^2 a_{jj}}} \sim \mathcal{N}(0, 1). \quad (20)$$

This result, which holds exactly for the normal classical regression model, also holds under homoskedasticity as a large-sample approximation for linear predictors and non-normal populations, in light of (8), (15), and Cramér's theorem.

**Heteroskedasticity-consistent standard errors** Note that the validity of the large sample results in (9), (10) and (11) does not require homoskedasticity. This is why the asymptotic standard errors  $\sqrt{\widehat{w}_{jj}/n}$  calculated from (9) are usually called heteroskedasticity-consistent or White standard errors, after the work of Halbert White.

**Other distributional results** The other key exact distributional results in this context are

$$\frac{\widehat{u}'\widehat{u}}{\sigma^2} \sim \chi_{n-k}^2 \quad \text{independent of } z_j \quad (21)$$

and

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{\sigma}^2 a_{jj}}} \sim t_{n-k}. \quad (22)$$

In addition, letting now  $\widehat{\beta}_j$  denote a subset of  $r$  coefficients and  $A_{jj}$  the corresponding submatrix of  $(X'X)^{-1}$ , we have

$$\frac{(\widehat{\beta}_j - \beta_j)' A_{jj}^{-1} (\widehat{\beta}_j - \beta_j)}{\sigma^2} \sim \chi_r^2 \quad (23)$$

and

$$\frac{(\widehat{\beta}_j - \beta_j)' A_{jj}^{-1} (\widehat{\beta}_j - \beta_j) / r}{\widehat{\sigma}^2} \sim F_{r, (n-k)}. \quad (24)$$

## 4 Weighted least squares

The ordinary least squares (OLS) statistic  $\widehat{\beta}$  is a function of simple means of  $x_i x_i'$  and  $x_i y_i$ . Under heteroskedasticity it may make sense to consider weighted means in which observations with a smaller variance receive a larger weight. Let us consider estimators of the form

$$\widetilde{\beta} = \left( \sum_{i=1}^n w_i x_i x_i' \right)^{-1} \sum_{i=1}^n w_i x_i y_i \quad (25)$$

where  $w_i$  are some weights. OLS is the special case in which  $w_i = 1$  for all  $i$ .

Under appropriate regularity conditions

$$\text{plim} (\widetilde{\beta} - \beta) = [E(w_i x_i x_i')]^{-1} E(w_i x_i u_i). \quad (26)$$

Thus, in general to ensure consistency of  $\widetilde{\beta}$  we need that  $E(w_i x_i u_i) = 0$ . This result will hold if  $E(u_i | x_i) = 0$  and  $w_i = w(x_i)$  is a function of  $x_i$  only:

$$E(w_i x_i u_i) = E(w_i x_i E(u_i | x_i)) = 0,$$

but more generally  $\widetilde{\beta}$  is not a consistent estimator of the population linear projection coefficient  $\beta$  when  $E(y_i | x_i) \neq x_i' \beta$ .<sup>1</sup>

Subject to consistency, the asymptotic normality result is

$$\sqrt{n} (\widetilde{\beta} - \beta) \xrightarrow{d} \mathcal{N} \left( 0, [E(w_i x_i x_i')]^{-1} E(u_i^2 w_i^2 x_i x_i') [E(w_i x_i x_i')]^{-1} \right). \quad (27)$$

---

<sup>1</sup>Actually, if  $x_i$  has density  $f(x)$ ,  $\widetilde{\beta}$  is consistent for the optimal linear predictor under an alternative probability distribution of  $x_i$  given by  $g(x) \propto f(x) w(x)$ .

**Asymptotic efficiency** When weights are chosen to be proportional to the reciprocal of  $\sigma_i^2 = E(u_i^2 | x_i)$ , the asymptotic variance in (27) becomes

$$\left[ E \left( \frac{x_i x_i'}{\sigma_i^2} \right) \right]^{-1}. \quad (28)$$

Moreover, it can be shown that for any (conformable) vector  $q$ :

$$q' [E(w_i x_i x_i')]^{-1} E(\sigma_i^2 w_i^2 x_i x_i') [E(w_i x_i x_i')]^{-1} q \geq q' \left[ E \left( \frac{x_i x_i'}{\sigma_i^2} \right) \right]^{-1} q. \quad (29)$$

Statement (29) says that the asymptotic variance of any linear combination of weighted LS estimates  $q' \tilde{\beta}$  is the smallest when the weights are  $w_i \propto 1/\sigma_i^2$ . To prove (29) note that<sup>2</sup>

$$E \left( \frac{x_i x_i'}{\sigma_i^2} \right) - E(w_i x_i x_i') [E(\sigma_i^2 w_i^2 x_i x_i')]^{-1} E(w_i x_i x_i') = H' E(m_i m_i') H \quad (30)$$

where

$$H = \begin{pmatrix} I \\ -[E(\sigma_i^2 w_i^2 x_i x_i')]^{-1} E(w_i x_i x_i') \end{pmatrix}, \quad m_i = \begin{pmatrix} \frac{x_i}{\sigma_i} \\ \sigma_i w_i x_i \end{pmatrix}.$$

Also note that for any  $q$  we have  $q' [H' E(m_i m_i') H] q \geq 0$ .

**Generalized least squares** In view of (29) we can say that the estimator

$$\tilde{\beta}_{GLS} = \left( \sum_{i=1}^n \frac{x_i x_i'}{\sigma_i^2} \right)^{-1} \sum_{i=1}^n \frac{x_i y_i}{\sigma_i^2} \quad (31)$$

is asymptotically efficient in the sense of having the smallest asymptotic variance among the class of consistent weighted least squares estimators.  $\tilde{\beta}_{GLS}$  is a generalized least squares estimator (GLS).

In matrix notation:

$$\tilde{\beta}_{GLS} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y \quad (32)$$

where  $\Omega = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ .

In a generalized classical regression model we have  $E(y | X) = X\beta$  and  $\text{Var}(y | X) = \Omega$ .

The asymptotic normality result is

$$\sqrt{n} (\tilde{\beta}_{GLS} - \beta) \xrightarrow{d} \mathcal{N} \left( 0, \left[ E \left( \frac{x_i x_i'}{\sigma_i^2} \right) \right]^{-1} \right). \quad (33)$$

Usually  $\tilde{\beta}_{GLS}$  is an infeasible estimator because  $\sigma_i^2$  is an unknown function of  $x_i$ . In a feasible GLS estimation  $\sigma_i^2$  is replaced by a (parametric or nonparametric) estimated quantity. The large-sample properties of the resulting estimator may or may not coincide with those of the infeasible GLS.

---

<sup>2</sup>We are using the fact that if  $A$  and  $B$  are positive definite matrices, then  $A - B$  is positive definite if and only if  $B^{-1} - A^{-1}$  is positive definite.

## 5 Cluster-robust standard errors

Suppose the sample  $\{y_i, x_i\}_{i=1}^n$  consists of  $H$  groups or clusters of  $M_h$  observations each ( $n = M_1 + \dots + M_H$ ), such that observations are independent across groups but dependent within groups,  $H$  is large and  $M_h$  is small (fixed) for all  $h$ . For convenience let us order observations by groups and use a double-index notation  $(y_{hm}, x_{hm})$  for  $h = 1, \dots, H$  (group index) and  $m = 1, \dots, M_h$  (within group index).

The compact notation for linear regression was  $y = X\beta + u$ . A similar notation for the observations in cluster  $h$  is

$$y_h = X_h\beta + u_h \quad (34)$$

where  $y_h = (y_{h1}, \dots, y_{hM_h})'$ , etc. Using this notation the OLS estimator is

$$\hat{\beta} = (X'X)^{-1} X'y = \left( \sum_{h=1}^H X_h'X_h \right)^{-1} \sum_{h=1}^H X_h'y_h. \quad (35)$$

Note that in terms of individual observations we can write  $X'y = \sum_{h=1}^H \sum_{m=1}^{M_h} x_{hm}y_{hm}$ , etc.

The scaled estimation error is

$$\sqrt{H}(\hat{\beta} - \beta) = \left( \frac{X'X}{H} \right)^{-1} \frac{1}{\sqrt{H}} \sum_{h=1}^H X_h'u_h.$$

Applying the central limit theorem at cluster level, a consistent estimate of the variance of  $\sqrt{H}(\hat{\beta} - \beta)$  is given by

$$\left( \frac{X'X}{H} \right)^{-1} \frac{1}{H} \sum_{h=1}^H X_h'\hat{u}_h\hat{u}_h'X_h \left( \frac{X'X}{H} \right)^{-1}, \quad (36)$$

so that cluster-robust standard errors can be obtained as the square roots of the diagonal elements of the covariance matrix

$$\widehat{Var}(\hat{\beta}) = (X'X)^{-1} \left( \sum_{h=1}^H X_h'\hat{u}_h\hat{u}_h'X_h \right) (X'X)^{-1}. \quad (37)$$

This is the sandwich formula associated with clustering. Its rationale is as a large  $H$  approximation. There are many applications of this tool, both with actual cluster survey designs and with other data sets with potential group-level dependence.

## 6 Fixed effects

Data with a group structure are common in economics (e.g. schools and students, firms and workers, households and their members). Panel data is a prominent special case in which a group is the set of observations of an individual at different points in time. In such case we often use the notation  $(y_{it}, x_{it})$  ( $i = 1, \dots, N; t = 1, \dots, T_i$ ) instead of  $(y_{hm}, x_{hm})$  ( $h = 1, \dots, H; m = 1, \dots, M_h$ ).

In a regression with fixed effects we regress  $y_{hm}$  on  $x_{hm}$  and group dummy variables. Therefore, it is a regression with group-specific intercepts:<sup>3</sup>

$$y_{hm} = x'_{hm}\beta + \alpha_h + u_{hm} \quad (h = 1, \dots, H; m = 1, \dots, M_h), \quad (38)$$

or in compact notation:

$$y = X\beta + D\alpha + u \quad (39)$$

where  $\alpha = (\alpha_1, \dots, \alpha_H)'$  and  $D$  is an  $n \times H$  matrix of group dummy variables:

$$D = \begin{pmatrix} \iota_1 & 0 & \dots & 0 \\ 0 & \iota_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \iota_H \end{pmatrix} \quad (40)$$

where  $\iota_h$  is a vector of ones of order  $M_h$ .

A fixed effects regression coefficient  $\beta_j$  is the predictive effect of  $x_{jhm}$  on  $y_{hm}$  holding the other  $x_{hm}$ 's and the group-level effects constant. Any predictor that varies with  $h$  but not with  $m$  will be a linear combination of the group dummies and therefore redundant in the fixed effects regression.

**Formulas for partitioned regression** OLS estimates of  $\beta$  and  $\alpha$  in (39) solve the equations

$$\begin{pmatrix} X'X & X'D \\ D'X & D'D \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\alpha} \end{pmatrix} = \begin{pmatrix} X'y \\ D'y \end{pmatrix}. \quad (41)$$

To obtain separate expressions for  $\hat{\beta}$  and  $\hat{\alpha}$ , we solve for  $\hat{\alpha}$  in the second block of equations:

$$\hat{\alpha} = (D'D)^{-1} D' (y - X\hat{\beta}) \quad (42)$$

and insert the result in the first block to get:

$$\hat{\beta} = (X'QX)^{-1} X'Qy \quad (43)$$

where  $Q = I - D(D'D)^{-1}D'$ . According to (43),  $\hat{\beta}$  can be obtained as the OLS regression of  $\tilde{y} = Qy$  on  $\tilde{X} = QX$ , where  $\tilde{y}$  and  $\tilde{X}$  are regression residuals of  $y$  and  $X$  on  $D$ , respectively. Once we have  $\hat{\beta}$ ,  $\hat{\alpha}$  can be obtained as the OLS regression of the partial residual  $(y - X\hat{\beta})$  on  $D$ , in view of (42).

---

<sup>3</sup>Thus, in this section  $x_{hm}$  does not include an intercept term.



**Within-group estimation** When  $D$  is the matrix of group dummy variables (40),  $\tilde{y}$  and  $\tilde{X}$  are arrays of the original variables in deviations from group-specific means with elements  $\tilde{y}_{hm} = y_{hm} - \bar{y}_h$  and  $\tilde{x}_{hm} = x_{hm} - \bar{x}_h$  where  $\bar{y}_h = M_h^{-1} \sum_{m=1}^{M_h} y_{hm}$ , etc. Therefore, the fixed effects estimator  $\hat{\beta}$  is simply OLS of  $y$  on  $X$  in deviations from group means:

$$\hat{\beta} = \left[ \sum_{h=1}^H \sum_{m=1}^{M_h} (x_{hm} - \bar{x}_h) (x_{hm} - \bar{x}_h)' \right]^{-1} \sum_{h=1}^H \sum_{m=1}^{M_h} (x_{hm} - \bar{x}_h) (y_{hm} - \bar{y}_h). \quad (44)$$

As for the estimated fixed effects they can be obtained one by one as group averages of partial residuals:

$$\hat{\alpha}_h = \frac{1}{M_h} \sum_{m=1}^{M_h} (y_{hm} - x'_{hm} \hat{\beta}) = \bar{y}_h - \bar{x}'_h \hat{\beta} \quad (h = 1, \dots, H). \quad (45)$$

**Properties** Under the assumptions of the classical regression model, namely  $E(y | X, D) = X\beta + D\alpha$  and  $Var(y | X, D) = \sigma^2 I_n$ ,  $\hat{\beta}$  and  $\hat{\alpha}$  are unbiased with conditional variances given by:<sup>4</sup>

$$Var(\hat{\beta} | X, D) = \sigma^2 (X' Q X)^{-1} \quad (46)$$

$$Var(\hat{\alpha}_h | X, D) = \frac{\sigma^2}{M_h} + \sigma^2 \bar{x}'_h (X' Q X)^{-1} \bar{x}_h. \quad (47)$$

If the data are a clustered random sample, as  $H \rightarrow \infty$  for  $M_h$  fixed, we have  $Var(\hat{\beta} | X, D) \rightarrow 0$  and  $Var(\hat{\alpha}_h | X, D) \rightarrow \sigma^2 / M_h$ . Therefore,  $\hat{\beta}$  is consistent when  $H$  is large and  $M_h$  is small but not  $\hat{\alpha}_h$ . This is not surprising because  $\hat{\alpha}_h$  is an average of  $M_h$  observations, so its dispersion can only vanish as  $M_h$  increases. The lesson is that in data with a group structure some parameters may be more estimable than others.

**Cluster-robust standard errors in a regression with fixed effects** Including fixed effects is in general not a substitute for clustered standard errors. For example, a panel regression may include both individual fixed effects and errors that are correlated over time. Since a fixed effects regression is equivalent to a within-group regression, the clustering formula of the previous section can be applied to the data in deviations from group means:<sup>5</sup>

$$\widehat{Var}(\hat{\beta}) = (X' Q X)^{-1} \left( \sum_{h=1}^H \tilde{X}'_h \tilde{u}_h \tilde{u}'_h \tilde{X}_h \right) (X' Q X)^{-1} \quad (48)$$

where  $\tilde{u}_h = \tilde{y}_h - \tilde{X}_h \hat{\beta}$ , and  $\tilde{y}_h$  and  $\tilde{X}_h$  are the  $h$ -th blocks of  $\tilde{y} = Qy$  and  $\tilde{X} = QX$ , respectively.

---

<sup>4</sup>Note that  $Var\left[\begin{pmatrix} \hat{\beta} \\ \hat{\alpha} \end{pmatrix} | X, D\right] = \sigma^2 \begin{pmatrix} X'X & X'D \\ D'X & D'D \end{pmatrix}^{-1}$  and  $\hat{\alpha}_h - \alpha_h = \bar{u}_h - \bar{x}'_h (\hat{\beta} - \beta)$ .

<sup>5</sup>Arellano, M. (1987): "Computing robust standard errors for within-groups estimators," *Oxford Bulletin of Economics and Statistics*, 49, 431–434.