Regression analysis with dependent data

Kerby Shedden

Department of Statistics, University of Michigan

December 16, 2019

Clustered data

Clustered data are sampled from a population that can be viewed as the union of a number of related subpopulations.

Write the data as

$$y_{ij} \in \mathcal{R},\ \mathbf{x}_{ij} \in \mathcal{R}^p,$$

$$i = 1, \ldots, m \qquad j = 1, \ldots, n_i$$

where i indexes the subpopulation and j indexes the individuals in the sample belonging to the i^{th} subpopulation.

There are m clusters (subpopulations), and there are n_i observations in cluster i. If the n_i are all the same, we have balanced clustering.

Clustered data

It may happen that units from the same subpopulation are more alike than units from different subpopulations, i.e. for $i \neq i'$, $j \neq j'$,

$$\text{cor}(y_{ij}, y_{ij'}) > \text{cor}(y_{ij}, y_{i'j'}).$$

Part of the within-cluster similarity may be explained by the covariates, i.e. units from the same subpopulation may have similar $\mathbf{x}_{ij}$ values, which leads to similar y values. In this case,

$$\text{cor}(y_{ij}, y_{ij'}) > \text{cor}(y_{ij}, y_{ij'}|\mathbf{x}_{ij}, \mathbf{x}_{ij'}).$$

Even after accounting for measured covariates, units in a cluster may still resemble each other more than units in different clusters:

$$\text{cor}(y_{ij}, y_{ij'}|\mathbf{x}_{ij}, \mathbf{x}_{ij'}) > \text{cor}(y_{ij}, y_{i'j'}|\mathbf{x}_{ij}, \mathbf{x}_{i'j'}).$$

Clustered data

A simple working model to account for this dependence is

$$\hat{y}_{ij} = \hat{\theta}_i + \hat{\beta}' \mathbf{x}_{ij}.$$

The idea here is that $\beta' \mathbf{x}_{ij}$ explains the variation in y that is related to the measured covariates, and the θ_i explain variation in y that is related to the clustering.

This working model would be correct if there were q omitted variables $\mathbf{z}_{ij\ell}$, $\ell = 1, \ldots, q$, that were constant for all units in the same cluster (i.e. $\mathbf{z}_{ij\ell}$ depends on ℓ and i, but not on j).

In that case, $\hat{\theta}_i$ would stand in for the value of $\sum_\ell \hat{\psi}_\ell \mathbf{z}_{ij\ell}$ that we would have obtained if the $\mathbf{z}_{ij\ell}$ were observed.

Clustered data

As an alternate notation, we can vectorize the data to express $y \in \mathcal{R}^n$ and $\mathbf{X} \in \mathcal{R}^{n \times p+1}$, then write

$$\hat{y} = \sum_i \hat{\theta}_i I_i + \mathbf{X}\hat{\beta},$$

where $I_i \in \mathcal{R}^n$ is the indicator of which subjects belong to cluster i.

If the observed covariates in $\mathbf{X}$ are related to the clustering (i.e. if the columns of $\mathbf{X}$ and the I_i are not orthogonal), then OLS apportions the overlapping variance between $\hat{\beta}$ and $\hat{\theta}$.

Test score example

Suppose y_{ij} is a reading test score for the j^{th} student in the i^{th} classroom, out of a large number of classrooms that are considered. Suppose $\mathbf{x}_{ij}$ is the income of a student's family.

We might postulate as a population model

$$y_{ij} = \theta_i + \beta \mathbf{x}_{ij} + \epsilon_{ij},$$

which can be fit as

$$\hat{y}_{ij} = \hat{\theta}_i + \hat{\beta} \mathbf{x}_{ij}.$$

Test score example

Ideally we would want the parameter estimates to reflect sources of variation as follows:

- "Direct effects" of parent income such as access to books, life experiences, good health care, etc. should go entirely to $\hat{\beta}$.
- Attributes of classrooms that are not related to parent income, for example, the effect of an exceptionally good or bad teacher, should go entirely to the $\hat{\theta}_i$.
- Attributes of classrooms that are correlated with parent income, such as teacher salary, training, and resources, will be apportioned by OLS between $\hat{\theta}_i$ and $\hat{\beta}$.
- Unique events affecting particular individuals, such as the severe illness of the student or a family member, should go entirely to ϵ.

Other examples of clustered data

- Treatment outcomes for patients treated in various hospitals.
- Crime rates in police precincts distributed over a number of large cities (the precincts are the units and the cities are the clusters).
- Prices of stocks belonging to various business sectors.
- Surveys in which the data are collected following a cluster sampling approach.

What if we ignore the θ_i?

The θ_i are usually not of primary interest, but we should be concerned that by failing to take account of the clustering, we may incorrectly assess the relationship between y and $\mathbf{X}$.

If the θ_i are nonzero, but we fail to include them in the model, the working model is misspecified.

Let $\mathbf{X}$ be the design matrix without intercept, and let Q be the matrix of cluster indicators (which includes the intercept):

$$y = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{23} \\ \cdots \end{pmatrix} \quad Q = \begin{pmatrix} 1 & 0 & \cdots \\ 1 & 0 & \cdots \\ 0 & 1 & \cdots \\ 0 & 1 & \cdots \\ 0 & 1 & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_{111} & \mathbf{X}_{112} & \cdots \\ \mathbf{X}_{121} & \mathbf{X}_{122} & \cdots \\ \mathbf{X}_{211} & \mathbf{X}_{212} & \cdots \\ \mathbf{X}_{221} & \mathbf{X}_{222} & \cdots \\ \mathbf{X}_{231} & \mathbf{X}_{232} & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix}$$

What if we ignore the θ_i?

Let $\tilde{\mathbf{X}} = [1_n|\mathbf{X}]$. The estimate that results from regressing y on $\tilde{\mathbf{X}}$ is

$$\begin{aligned} E\hat{\beta}^* &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'E[y] \\ &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(\mathbf{X}\beta + Q\theta). \end{aligned}$$

where $\hat{\beta}^* = (\hat{\beta}_0, \hat{\beta}')'$.

For the bias of $\hat{\beta}$ for β to be zero, we need

$$E[\hat{\beta}^*] - \tilde{\beta} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(\mathbf{X}\beta + Q\theta - \tilde{\mathbf{X}}\tilde{\beta}) = 0,$$

where $\beta_0 = E[\hat{\beta}_0]$ and $\tilde{\beta} = (\beta_0, \beta')'$. Thus we need

$$0 = \tilde{\mathbf{X}}'(\mathbf{X}\beta + Q\theta - \tilde{\mathbf{X}}\tilde{\beta}) = \tilde{\mathbf{X}}'(Q\theta - \beta_0).$$

What if we ignore the θ_i?

Let $S = Q\theta$ be the vector of cluster effects. Since the first column of $\tilde{\mathbf{X}}$ consists of 1's, we have that

$$\bar{S} = \beta_0.$$

For any other covariate $\mathbf{X_j}$, we have that

$$\mathbf{X}_j' S = \beta_0 \mathbf{X}_j' 1_n,$$

which implies that $\mathbf{X}_j$ and S have zero sample covariance.

This is unlikely in many studies, where people tend to cluster (in schools, hospitals, etc.) with other people having similar covariate levels.

Fixed effects analysis

In a fixed effects approach, we model the θ_i as regression parameters, by including additional columns in the design matrix whose covariate levels are the cluster indicators, i.e. we regress y on $[\mathbf{X}\ Q]$.

As the sample size grows, in most applications the cluster sizes n_i will remain bounded (e.g. a primary school classroom might have up to 30 students). Thus the number of clusters must grow, so the dimension of the parameter vector θ grows.

This puts us in a setting where the model dimension (p) and sample size (n) are growing together. This is not typical in standard regression modeling, and leads to the "Neyman Scott problem". Contemporary methods for "high dimensional" regression provide one way to work around this challenge.

Cluster effect examples

What does the inclusion of fixed effects do to the parameter estimates $\hat{\beta}$, which are often of primary interest?

The following slides show scatterplots of an outcome y against a scalar covariate $\mathbf{x}$, in a setting where there are three clusters (indicated by color).

Above each plot are the coefficient estimates, Z-scores, and r^2 values for fits in which the cluster effects are either included (top line) or excluded (second line). These estimates are obtained from the following two working models:

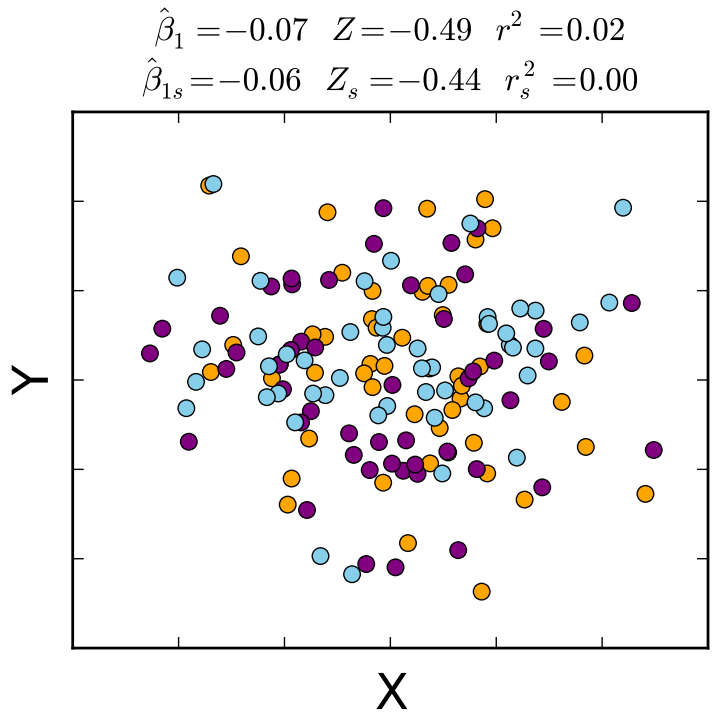
$$\hat{y} = \hat{\alpha} + \hat{\beta}\mathbf{x} + \sum \hat{\theta}_i I_i \qquad\qquad \hat{y} = \hat{\alpha}_s + \hat{\beta}_s \mathbf{x}.$$

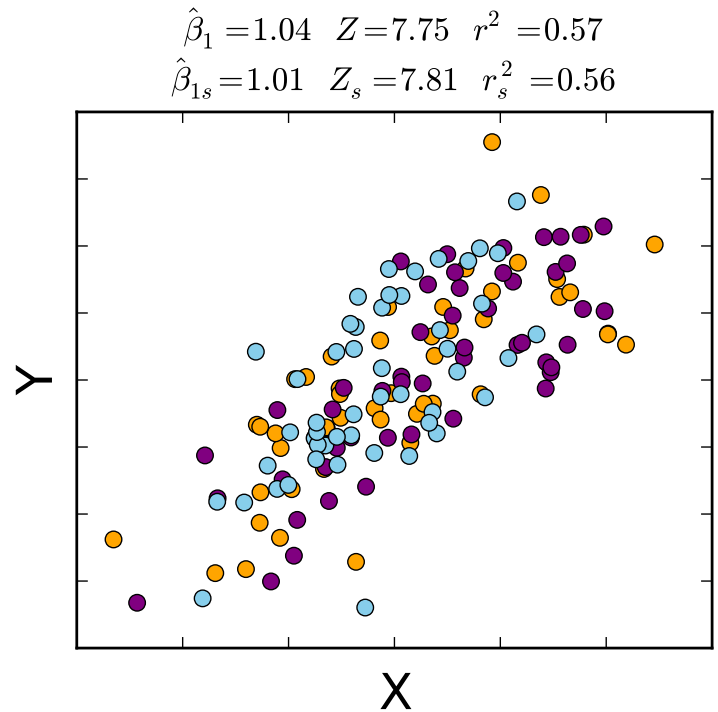
The Z-scores are $\hat{\beta}/\mathrm{SD}(\hat{\beta})$ and the r^2 values are $\mathrm{cor}(\hat{y}, y)^2$.

Cluster effect examples

Left: y is independent of both clusters and **x**; **x** and clusters are independent; $\hat{\beta} \approx 0$ with and without cluster terms in model.

Right: y relates to **x**, but not to the clusters; **x** and clusters are independent; $\hat{\beta}$, Z, and r^2 are similar with and without cluster effects in model.

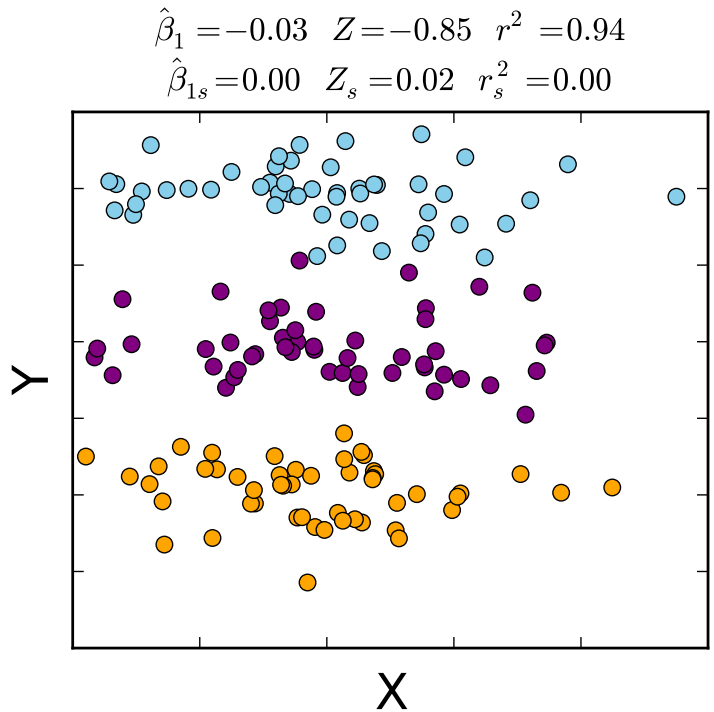


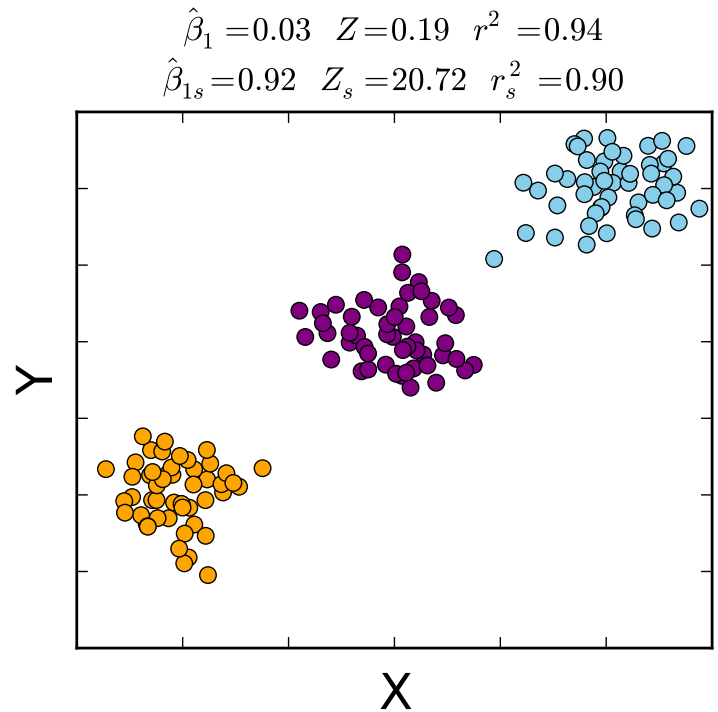


Cluster effect examples

Left: y relates to clusters, but not to **x**; **x** and clusters are independent; $\hat{\beta} \approx 0$ with and without cluster effects in model.

Right: y relates to clusters but not to **x**; **x** and clusters are dependent; when cluster effects are not modeled their effect is picked up by **x**.

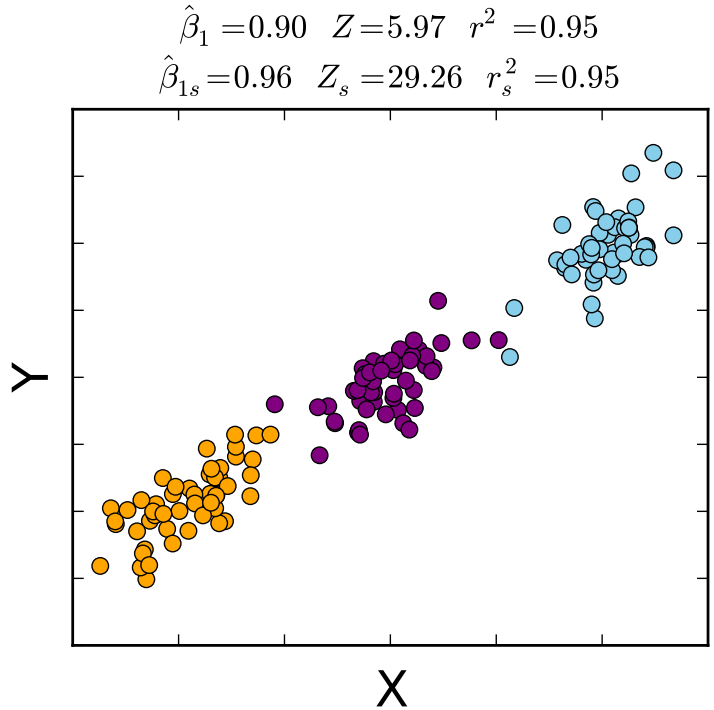


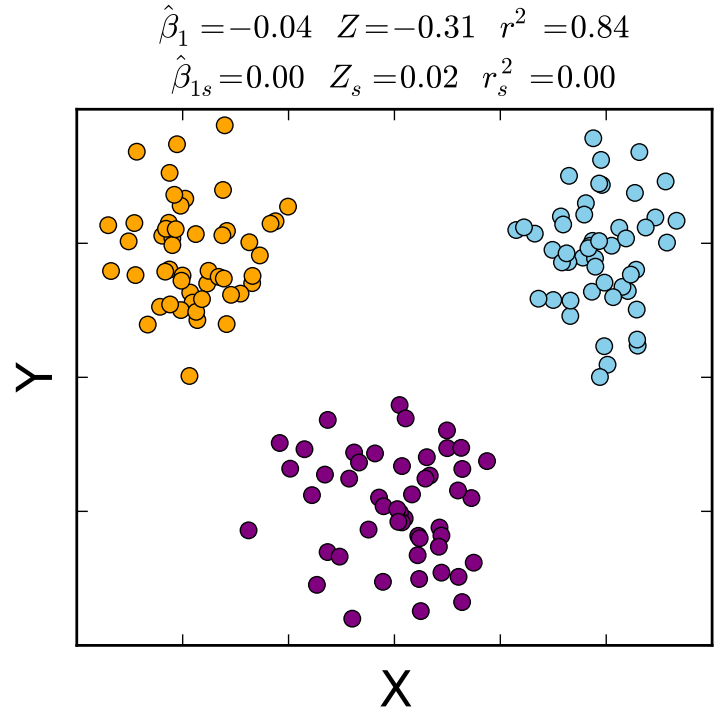


Cluster effect examples

Left: y relates to both clusters and **x**; **x** and clusters are dependent.

Right: y relates to clusters but not to **x**; **x** and clusters are dependent but the net cluster effect is not linearly related to **x**.

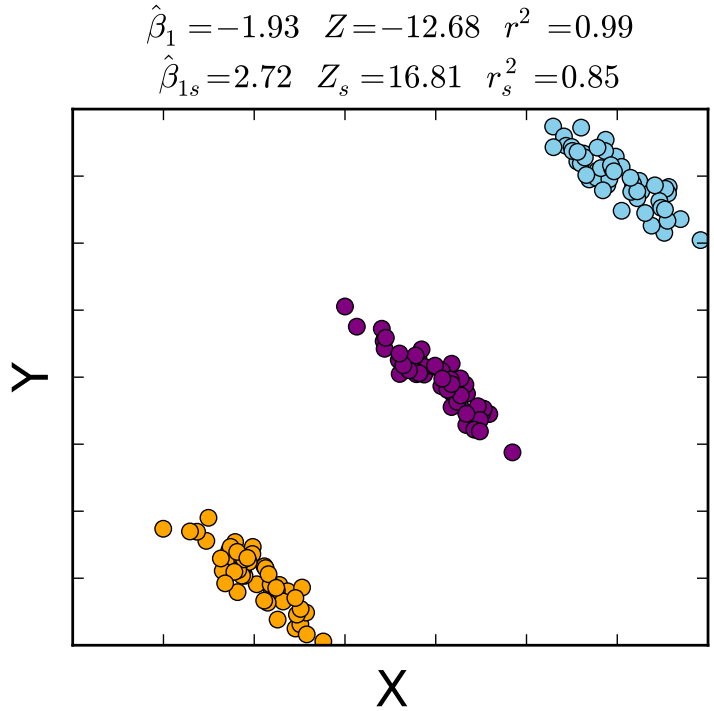


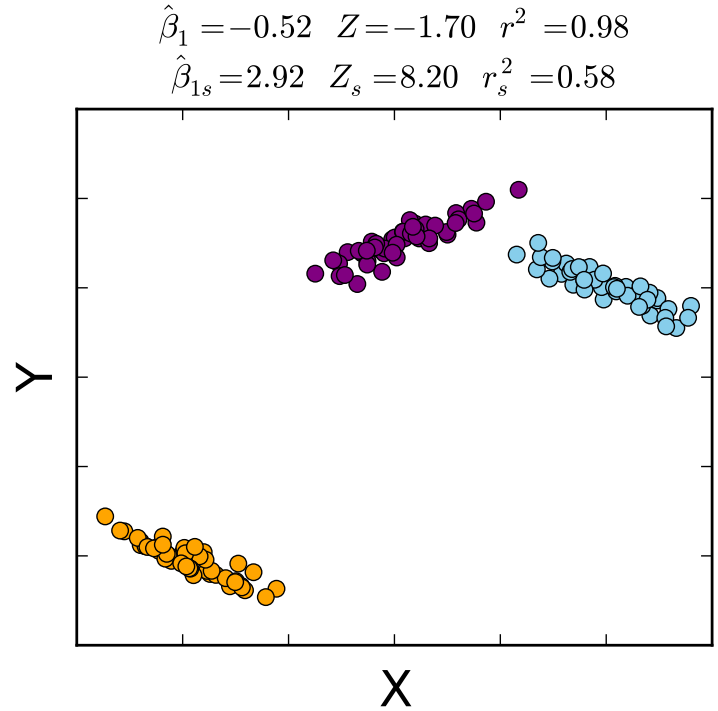


Cluster effect examples

Left: y relates to clusters and to **x**; **x** and clusters are dependent; the **x** effect has the opposite sign as the net cluster effect.

Right: y relates to clusters and to **x**; **x** and clusters are dependent; the signs and magnitudes of the **x** effects are cluster-specific.





Implications of fixed effects analysis for observational data

A stable confounder is a confounding factor that is approximately constant within clusters. A stable confounder will become part of the net cluster effect.

If a stable confounder is correlated with an observed covariate X, then this will create non-orthogonality between the cluster effects and the effects of the observed covariates.

Implications of fixed effects analysis for experimental data

Experiments are often carried out on "batches" of objects (specimens, parts, etc.) in which uncontrolled factors cause elements of the same batch to be more similar than elements of different batches.

If treatments are assigned randomly within each batch, there are no stable confounders (in general there are no confounders in experiments). Therefore the overall OLS estimate of β is unbiased as long as the standard linear model assumptions hold.

Implications of fixed effects analysis for experimental data

Suppose the design is balanced (e.g. exactly half of each batch is treated and half is untreated). This is an orthogonal design, so the estimate based on the working model

$$\hat{y}_{ij} = \hat{\beta}\mathbf{x}_{ij}$$

and the estimate based on the working model

$$\hat{y}_{ij} = \hat{\theta}_i + \hat{\beta}^*\mathbf{x}_{ij}$$

are identical ($\hat{\beta} = \hat{\beta}^*$). Thus they have the same variance. But the estimated variance of $\hat{\beta}$ will be greater than the estimated variance of $\hat{\beta}^*$ (since the corresponding estimate of σ^2 is greater), so it will have lower power and wider confidence intervals.

Random cluster effects

As we have seen, the cluster effects θ_i can be treated as unobserved constants, and estimated as we would estimate any other regression coefficient.

An alternative way to handle cluster effects is to view the θ_i as unobserved (latent) random variables.

In doing this, we now we have two random variables in the model: θ_i and ϵ_{ij}, which are taken to be independent of each other.

If the θ_i are independent and identically distributed, we can combine them with the error terms to get a single random error term per observation:

$$\epsilon_{ij}^{c} = \theta_i + \epsilon_{ij}.$$

Random cluster effects

Let $y_i \equiv (y_{i1}, \ldots, y_{in_i})'$ denote the vector of responses in the i^{th} cluster, let $\mathbf{x}_i \in \mathcal{R}^{n_i \times p}$ denote the matrix of predictor variables for the i^{th} cluster, and let $\epsilon_i^c \equiv (\epsilon_{i1}^c, \ldots, \epsilon_{in_i}^c)' \in \mathcal{R}^{n_i}$ denote the vector of random "errors" for the i^{th} cluster.

Thus we have the model

$$y_i = \mathbf{x}_i \beta + \epsilon_i^c,$$

for clusters $i = 1, \ldots, m$. The ϵ_i are taken to be uncorrelated between clusters, i.e.

$$\text{cov}(\epsilon_i^c, \epsilon_{i'}^c) = 0_{n_i \times n_{i'}}$$

for $i \neq i'$.

Random cluster effects

The structure of the covariance matrix

$$S_i \equiv \mathrm{cov}(\epsilon_i^c | \mathbf{x}_i)$$

is

$$S_i = \begin{pmatrix} \sigma^2 + \sigma_\theta^2 & \sigma_\theta^2 & \cdots & \cdots & \sigma_\theta^2 \\ \sigma_\theta^2 & \sigma^2 + \sigma_\theta^2 & \sigma_\theta^2 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma_\theta^2 & \sigma_\theta^2 & \cdots & \cdots & \sigma^2 + \sigma_\theta^2 \end{pmatrix} = \sigma^2 I + \sigma_\theta^2 \mathbf{1}\mathbf{1}^T,$$

where $\mathrm{var}(\epsilon_{ij}) = \sigma^2$ and $\mathrm{var}(\theta_i) = \sigma_\theta^2$.

Generalized Least Squares

Suppose we have a linear model with mean structure $E[y|\mathbf{X}] = \mathbf{X}\beta$ for $y \in \mathcal{R}^n$, $\mathbf{X} \in \mathcal{R}^{n \times p+1}$, and $\beta \in \mathcal{R}^{p+1}$, and variance structure $\mathrm{Cov}[y|\mathbf{X}] \propto \Sigma$, where Σ is a given $n \times n$ matrix.

We can write the model in generative form as $y = \mathbf{X}\beta + \epsilon$, where $\epsilon \in \mathcal{R}^n$ with $E[\epsilon|\mathbf{X}] = 0$, $\mathrm{Cov}[\epsilon|\mathbf{X}] = \Sigma$.

Factor the covariance matrix as $\Sigma = GG'$, and consider the transformed model

$$G^{-1}y = G^{-1}\mathbf{X}\beta + G^{-1}\epsilon.$$

Then letting $\eta \equiv G^{-1}\epsilon$, it follows that $\mathrm{Cov}(\eta) = I_{n \times n}$, and note that the population slope vector β of the transformed model is identical to the population slope vector of the original model.

Generalized Least Squares

The GLS estimator of β is defined to be the OLS estimator of β for the decorrelated response $G^{-1}y$ and the decorrelated predictors $G^{-1}\mathbf{X}$.

The GLS estimate of the regression slope can be expressed in terms of the original design matrix $\mathbf{X}$ and response vector y:

$$\begin{aligned}\hat{\beta}_{\text{GLS}} &= ((G^{-1}\mathbf{X})'G^{-1}\mathbf{X})^{-1}(G^{-1}\mathbf{X})'G^{-1}y \\ &= (\mathbf{X}'G^{-T}G^{-1}\mathbf{X})^{-1}\mathbf{X}'G^{-T}G^{-1}y \\ &= (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}y.\end{aligned}$$

Generalized Least Squares

We can use GLS to analyze clustered data with random cluster effects.

Let $y_i^* \equiv S_i^{-1/2} y_i$, $\epsilon_i^* \equiv S_i^{-1/2} \epsilon_i^c$, and $\mathbf{x}_i^* = S_i^{-1/2} \mathbf{x}_i$.

Let y^*, $\mathbf{X}^*$, and ϵ^* denote the result of stacking y_i^*, $\mathbf{x}_i^*$, and ϵ_i^* over i, respectively.

Generalized Least Squares

Since $\mathrm{cov}(\epsilon_i^*) \propto I$, the OLS estimate of β for the model

$$y^* = \mathbf{X}^* \beta + \epsilon^*$$

is the best estimate of β that is linear in y^* (by the Gauss-Markov theorem).

Since the set of linear estimates based on y^* is the same as the set of linear estimates based on y, it follows that the GLS estimate of β based on y is the best unbiased estimate of β that is linear in y.

Generalized Least Squares

When $\Sigma = \mathrm{Cov}[y|\mathbf{X}]$, the covariance matrix of $\hat{\beta}$ in GLS is

$$\mathrm{Cov}[\hat{\beta}|\mathbf{X}] = (\mathbf{X}^{*T}\mathbf{X}^{*})^{-1} = (\mathbf{X}^{T}\Sigma^{-1}\mathbf{X})^{-1}$$

Note that this is consistent with what we get in OLS, where $\Sigma = \sigma^2 I$ and $\mathrm{Cov}[\hat{\beta}|\mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

To apply GLS, it is only necessary to know Σ up to a multiplicative constant. The same estimated slopes $\hat{\beta}$ are obtained if we decorrelate with Σ or with $k\Sigma$ for $k > 0$.

However if $\Sigma = k \cdot \mathrm{Cov}[y|\mathbf{X}]$, then

$$\mathrm{Cov}[\hat{\beta}|\mathbf{X}] = k^{-1}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1},$$

and at the moment we have no way to estimate k.

Generalized Least Squares

When $\mathrm{Cov}[y|\mathbf{X}] \propto \Sigma$, the sampling covariance matrix for GLS is proportional to $(\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}$, and the parameter estimates are uncorrelated with each other if and only if $\mathbf{X}^T\Sigma^{-1}\mathbf{X}$ is diagonal – that is, if the columns of $\mathbf{X}$ are mutually orthogonal with respect to the inner product defined by Σ^{-1}: $\langle u, v\rangle \equiv u'\Sigma^{-1}v$.

This is related to the fact that $\hat{y}$ in GLS is the projection of y onto $\mathrm{col}(\mathbf{X})$ in the Mahalanobis metric defined by Σ^{-1}, $d(u, v) = (u - v)'\Sigma^{-1}(u - v)$. This generalizes the fact that $\hat{y}$ in OLS is the projection of y onto $\mathrm{col}(\mathbf{X})$ in the Euclidean metric.

Neither of these facts depend on the proportionality constant relating Σ to $\mathrm{Cov}[y|\mathbf{X}]$.

Generalized least squares with a "working" covariance

What if we perform GLS using a possibly miss-specified "working covariance" S?

For example, what happens if we use OLS when the actual covariance matrix of the errors is $\Sigma \neq I$?

Since

$$E[\epsilon^*|\mathbf{X}^*] = E[\epsilon^*|\mathbf{X}] = 0,$$

the estimate $\hat{\beta}$ remains unbiased. However it has two problems: it may not be the BLUE (i.e. it may not have the least variance among unbiased estimates), and the usual linear model inference procedures will be wrong.

Generalized least squares with "working" covariance

The sampling covariance when the error structure is mis-specified is given by the "sandwich expression:"

$$\mathrm{Cov}[\hat{\beta}] = (\mathbf{X}^{*\prime}\mathbf{X}^{*})^{-1}\mathbf{X}^{*\prime}\Sigma_{\epsilon^{*}}\mathbf{X}^{*}(\mathbf{X}^{*\prime}\mathbf{X}^{*})^{-1}$$

where

$$\Sigma_{\epsilon^{*}} = \mathrm{Cov}[\epsilon^{*}|\mathbf{X}^{*}].$$

This result covers two special situations: (i) we use OLS, so $\mathbf{X}^{*} = \mathbf{X}$ and $\Sigma_{\epsilon^{*}} = \Sigma_{\epsilon}$ is the covariance matrix of the errors, and (ii) we use a "working covariance" to define the decorrelating matrix G, and this working covariance is not equal to Σ^{*}.

Generalized least squares with "working" covariance

Another way to write the covariance of $\hat{\beta}$ is as follows:

$$\mathrm{Cov}[\hat{\beta}] = (\mathbf{X}'\Sigma_w^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma_w^{-1}\Sigma\Sigma_w^{-1}\mathbf{X}(\mathbf{X}'\Sigma_w^{-1}\mathbf{X})^{-1}$$

where Σ_w is the "working" (possibly incorrectly specifed) covariance matrix for ϵ.

From this expression, it is clear that if $\Sigma_w = \Sigma$ (the working model is correct), then

$$\mathrm{Cov}[\hat{\beta}] = (\mathbf{X}'\Sigma_w^{-1}\mathbf{X})^{-1} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}.$$

Iterated GLS

In practice, we will generally not know Σ, so it must be estimated from the data. Σ is the covariance matrix of the errors, so we can estimate Σ from the residuals.

Typically a low-dimensional parameteric model $\Sigma = \Sigma(\alpha)$ is used.

Given a model for the the unexplained variation (i.e. for Σ), then the fitting algorithm alternates between estimating α and estimating the regression slopes β.

Iterated GLS

For example, suppose our working covariance has the form

$$S_i(j,j) = \nu \qquad\qquad S_i(j,k) = r\nu \quad (j \neq k).$$

This is an exchangeable model with "intraclass correlation coefficient" (ICC) r between two observations in the same cluster.

There are several closely-related ICC estimates. One approach is based on the standardized residuals R^s_{ij}:

$$\sum_i \sum_{j<j'} R^s_{ij} R^s_{ij'} / (\sum_i n_i(n_i-1)/2 - p - 1)$$

where n_i is the size of the i^{th} cluster.

Iterated GLS

Another common model for the error covariance is the first order autoregressive (AR-1) model:

$$\Sigma_{ij} = \alpha^{|i-j|},$$

where $|\alpha| < 1$ is a parameter.

There are several possible estimates of α borrowing ideas from time series analysis. We will not provide more details here.

Generalized Least Squares and stable confounders

For GLS to give unbiased estimates of β, we must have

$$E[\epsilon^*_{ij}|\mathbf{X}] = E[\theta_i + \epsilon_{ij}|\mathbf{X}] = 0.$$

Since $E[\epsilon_{ij}|\mathbf{X}] = 0$, this is equivalent to requiring that $E[\theta_i|\mathbf{X}] = 0$.

Thus if the covariates have distinct within-cluster mean values, and the within-cluster mean values of the covariates are correlated with the θ_i, then the GLS estimate of β will be biased.

Likelihood inference for the random intercepts model

The random intercepts model

$$y_{ij} = \theta_i + \beta' \mathbf{x}_{ij} + \epsilon_{ij},$$

can be analyzed using a likelihood approach, using:

- A random intercepts density:

$$\phi(\theta|\mathbf{X})$$

- A density for the data given the random intercept:

$$f(y|\mathbf{x}, \theta)$$

Multilevel models and conditional independence

The models are hierarchical (multilevel), and encode conditional independence relationships.

At the base level of the hierarchy, the random effect θ_i are independent and identically distributed, and in particular are unrelated to $\mathbf{X}$:

$$p(\theta_1, \ldots, \theta_m | \mathbf{X}) = p(\theta_1, \ldots, \theta_m) = \prod_{i=1}^{m} \phi(\theta_i)$$

Conditionally on the random effects, the observed data $\{y_{ij}\}$ are independent, and follow distributions that depend on the (unobserved) random effects and the observed covariates:

$$p(\{y_{ij}\} | \{\theta_i\}, \mathbf{X}) = \prod_{ij} f(y_{ij} | \theta_i, \mathbf{X})$$

Likelihood inference for the random intercepts model

Since the random effects are not observed, we must estimate the parameters in terms of the marginal model for the observed data.

This is a model that depends on the structural parameters, which are $(\beta, \sigma^2_\theta, \sigma^2)$ in this case.

The marginal density is obtained by integrating out the random effects

$$p(y_{i1}, \ldots y_{in_i}|\mathbf{X}) = \int f(y_{i1}, \ldots, y_{in_i}|\mathbf{X}, \theta_i)\phi(\theta_i)d\theta_i.$$

Likelihood inference for the random intercepts model

For linear multilevel models, the marginal distribution $p(y|\mathbf{X})$ can be calculated explicitly. It is Gaussian, and therefore is characterized by its moments:

$$E[y_{ij}|\mathbf{x}_{ij}] = \beta'\mathbf{x}_{ij}$$

$$\mathrm{Var}[y_{ij}|\mathbf{x}_{ij}] = \sigma_\theta^2 + \sigma^2,$$

$$\mathrm{Cov}[y_{i_1 j_1}, y_{i_2 j_2}|\mathbf{x}_{i_1 j_1}, \mathbf{x}_{i_2 j_2}] = 0 \quad (\text{if } i_1 \neq i_2).$$

$$\mathrm{Cov}[y_{ij_1}, y_{ij_2}|\mathbf{x}_{ij_1}, \mathbf{x}_{ij_2}] = \sigma_\theta^2 \quad (\text{if } j_1 \neq j_2).$$

Marginal form of the generative model

Suppose

$$\epsilon|\mathbf{X} \sim N(0, \sigma^2) \qquad\qquad \theta|\mathbf{X} \sim N(0, \sigma_\theta^2).$$

In this case $y|\mathbf{X}$ is Gaussian, with mean and variance as given above. Thus the random intercept model can be equivalently written in marginal form as

$$y_{ij} = \beta' \mathbf{x}_{ij} + \epsilon^*_{ij}$$

where $\epsilon^*_{ij} = \theta_i + \epsilon_{ij}$.

It follows that $E[\epsilon^*_{ij}|\mathbf{X}] = 0$, $\mathrm{Var}[\epsilon^*_{ij}|\mathbf{X}] = \sigma_\theta^2 + \sigma^2$, and the ϵ^*_{ij} values have correlation coefficient $\sigma_\theta^2/(\sigma^2 + \sigma_\theta^2)$ within clusters.

Likelihood computation

Maximum likelihood estimates for the model can be calculated using a gradient-based optimization procedure applied to the marginal log-density, or using the EM algorithm.

Asymptotic standard errors can be obtained from the inverse of the Fisher information matrix. Likelihood ratio tests, AIC values, and other likelihood-based inference tools can be used.

For linear multilevel models, the parameter estimates from iterated GLS and the MLE for the random intercepts model will be similar but not identical. Both are consistent, and in general will be asymptotically equivalent.

Predicting the random intercepts

Since the model is fit by optimizing the marginal log-likelihood, we obtain an estimate of $\sigma^2_\theta \equiv \mathrm{Var}[\theta_i]$, but we don't automatically learn anything about the individual θ_i.

If there is an interest in the individual θ_i values, we can predict them using the best linear unbiased predictor (BLUP).

The population version of the BLUP is:

$$E_{\beta,\sigma^2,\sigma^2_\theta}[\theta_i|y_i, \mathbf{X}] = \mathrm{Cov}[\theta_i, y_i|\mathbf{X}] \cdot \mathrm{Cov}[y_i|\mathbf{X}]^{-1} \cdot (y_i - E[y_i|\mathbf{X}])$$

Since this depends on things we don't know, in practice we use the sample version of the BLUP (sometimes called the eBLUP):

$$E_{\hat{\beta},\hat{\sigma}^2,\hat{\sigma}^2_\theta}[\theta_i|y_i, \mathbf{X}] = \widehat{\mathrm{Cov}}[\theta_i, y_i|\mathbf{X}] \cdot \widehat{\mathrm{Cov}}[y_i|\mathbf{X}]^{-1} \cdot (y_i - \hat{E}[y_i|\mathbf{X}])$$

Predicting the random intercepts

The BLUP is truly a linear function of y, is unbiased, and is "best" in the sense of minimizing the expected squared prediction error.

However the eBLUP is none of of these things (it is not linear or unbiased, and it is unclear if it is "best").

Note also that due to the hierarchical structure of the model, to predict θ_i we only need to condition on y_i (the other $y_{i'}$, for $i' \neq i$, contain no information). Thus the BLUP for θ_i only depends on the data through y_i. But in the eBLUP, all the data are used indirectly, through the estimates of the structural parameters.

Predicting the random intercepts

The estimated second moments needed to calculate the BLUP are:

$$\widehat{\mathrm{Cov}}[\theta_i, y_i|\mathbf{X}] = \hat{\sigma}^2_\theta \cdot \mathbf{1}_{n_i}$$

and

$$\widehat{\mathrm{Cov}}[y_i|\mathbf{X}] = \hat{\sigma}^2 I + \hat{\sigma}^2_\theta \mathbf{1}\mathbf{1}^T.$$

where n_i is the size of the i^{th} group.

Predicting the random intercepts

For a given set of parameter values, the BLUP for the random intercepts model is a linear function of the data, with the following form:

$$E_{\hat{\beta},\hat{\sigma}^2,\hat{\sigma}^2_\theta}[\theta_i|y,\mathbf{X}] = n_i\sigma^2_\theta/(\hat{\sigma}^2 + n_i\hat{\sigma}^2_\theta) \cdot \mathbf{1}^T(y_i - \hat{E}[y_i|\mathbf{X}])/n_i,$$

Predicting the random intercepts

In the BLUP for θ_i, this term is the mean of the residuals:

$$\mathbf{1}^T(y_i - \hat{E}[y_i|\mathbf{X}])/n_i$$

and it is shrunk by this factor:

$$n_i\sigma_\theta^2/(\hat{\sigma}^2 + n_i\hat{\sigma}_\theta^2)$$

This shrinkage allows us to interpret the random intercepts model as a "partially pooled" model that is intermediate between the fixed effects model and the model that completely ignores the clusters.

Predicting the random intercepts

In the fixed effects model, the parameter estimates θ_i are unbiased, but they are "overdispersed", meaning that the sample variance of the $\hat{\theta}_i$ will generally be greater than σ^2_θ.

In the random intercepts model, the variance parameter σ^2_θ is estimated with low bias, and the BLUP's of the θ_i are shrunk toward zero.

Random slopes

Suppose we are interested in a measured covariate z whose effect may vary by cluster. We might start with the model

$$y_i = \mathbf{x_i}\beta + \gamma z_i + \epsilon,$$

For example, supose that $z_i \in \{0, 1\}^{n_i}$ is a vector of treatment indicators (1 for treated subjects, 0 for untreated subjects). Then γ is the average change associated with being treated (the population treatment effect).

In many cases, it is reasonable to consider the possibility that different clusters may have different treatment effects – that is, different clusters have different γ values. In this case we can let γ_i be the treatment response for cluster i.

Random slopes

Suppose we model the random slopes γ_i as being Gaussian (given X) with variance σ^2_γ. The marginal model is

$$E[y_i|\mathbf{x}_i, z_i] = \mathbf{x}_i'\beta$$

and

$$\mathrm{Cov}[y_i|\mathbf{x}_i, z_i] = \sigma^2_\gamma z_i z_i' + \sigma^2 I.$$

Again we can form a BLUP $E_{\hat{\beta},\hat{\sigma}^2_\gamma,\hat{\sigma}^2}[\gamma_i|y, \mathbf{X}, z]$, and the BLUP turns out to be a shrunken version of what would be obtained in a fixed effects model, where we regress y on $\mathbf{x}$ and z within every cluster.

Linear mixed effects models

The random intercept and random slope model are special cases of the linear mixed effects model:

$$y_i = \mathbf{X}_i\beta + Z_i\gamma_i + \epsilon_i$$

Here, $\mathbf{X}_i$ is a $n_i \times p$ design matrix for cluster i, Z is a $n_i \times q$ random effects design matrix for cluster i, $\gamma_i \in \mathcal{R}^q$ is a random vector with mean 0 and covariance matrix Ψ, and $\epsilon \in \mathcal{R}^{n_i}$ is a random vector with mean 0 and covariance matrix $\sigma^2 I$.