Cluster-Robust Standard Errors

Lucas Szwarcberg, Eric Shen, and Charles Hua

Stat 244 Project

1 Introduction

Traditional regression approaches, encompassed by much of the Stat 244 material, use the assumption that the error terms of observations are homoskedastic and uncorrelated. In particular, the covariance matrix of the observed data $\boldsymbol{y}$ is typically modeled as $\sigma^2\mathbf{I}$, where σ^2 is an estimable scalar parameter. This gives a relatively straightforward model for the regression techniques and inference that follow.

This assumption is a strong one and may be unrealistic or invalid in certain settings. This is particularly the case with cross-sectional data that are implicitly grouped into different clusters. In such cases of "clustered" data, the errors within clusters may exhibit local and differing covariance structures. For example, suppose a survey of undergraduates at a university is taken. The errors of observations for students within the same majors or classes may be correlated in unique ways, rendering the assumption of homoskedastic and uncorrelated errors invalid. Similar phenomena may occur when considering experiments that sample across different countries, cities, industries, households, hospitals, or other aggregate units (MacKinnon, Nielsen, and Webb, forthcoming, 2).

Given an estimator of the regression parameters $\hat{\boldsymbol{\beta}}$ in regression-based approaches, applying traditional techniques used to characterize uncertainty, such as estimators of the $\text{Var}(\hat{\boldsymbol{\beta}})$ and standard errors derived from these estimators, may no longer be accurate under this

clustering phenomenon. For example, it has been shown that applying the assumption of homoskedastic and uncorrelated errors to clustered data can lead to estimates of $\text{Var}(\hat{\boldsymbol{\beta}})$ that systematically underestimate the true variance. The same can happen even when heteroskedasticity-robust variance estimators are used (which relax the homoskedasticity assumption, but still assume errors are uncorrelated) (MacKinnon and Webb 2020, 2). Thus, performing regression without accounting for clustering can produce misleadingly narrow confidence intervals, large t-statistics, and small p-values (Cameron and Miller 2015, 318). Properly modelling and controlling for clustered data leads to more accurate estimates and improves the validity of regression-based inference.

Adjusting for clustering is reasonable in various settings, such as when the sampling or treatment assignment for experiments occurs at the cluster, as opposed to the unit, level. Much work in statistical inference and econometrics has increasingly addressed ways to account for clustering, i.e. relax the assumption of homoskedastic and uncorrelated errors to derive more accurate standard errors under clustered data settings. There are several different ways to do so. Some include explicitly modeling intra-cluster error correlations into a linear model used in regression itself. This lead to techniques such as fixed or random effects models, which are examples of hierarchical modelling (Agresti 2015, 296).

Another strategy is to obtain $\hat{\boldsymbol{\beta}}$ through a standard linear model with ordinary least squares (OLS), but use different estimators for the variance of $\hat{\boldsymbol{\beta}}$ which still perform well under clustered settings. This report focuses on these "cluster-robust variance estimators", which translate into "cluster-robust standard errors". This technique has a range of applications, particularly in econometrics, such as with difference-in-differences methods, and applied political science. The method is widely used, as recent literature demonstrates (Angrist and Pischke 2009; Cameron, Gelbach, and Miller 2011; Abadie et al. 2022). We start by describing how clustered data can be modelled. We then introduce common cluster-robust variance estimators, how to use them in inference, and an extension of this technique to feasible GLS. The report concludes with data examples and a brief discussion of broader topics.

2 Theoretical Setup

We start from the setup for a linear model with a general covariance structure:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad E(\boldsymbol{\varepsilon}) = 0 \qquad \text{Var}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\mathsf{T}) = \boldsymbol{\Sigma} \tag{1}$$

where $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ are $n \times 1$ vectors, $\boldsymbol{X}$ is a known $n \times p$ matrix with rows $\boldsymbol{x}_i$, $\boldsymbol{\beta}$ is an unknown $p \times 1$ parameter vector, and $\boldsymbol{\Sigma}$ is a $n \times n$ matrix. From the lecture notes, when we assume $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ for some constant σ^2, which corresponds to homoskedastic and uncorrelated data, the OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$.

More general models assume a more general structure for $\boldsymbol{\Sigma}$, and may treat $\boldsymbol{\Sigma}$ as an unknown quantity to be estimated. In class, we saw that when the residuals are Normally distributed and when $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{V}$ for some unknown σ^2 and known, positive-definite $\boldsymbol{V}$, then $\boldsymbol{\beta}$ can be estimated by generalized least squares. In this case, $\boldsymbol{V}$ being diagonal corresponds with weighted least-squares regression, and a model where errors are heteroskedastic but uncorrelated (as cross-terms in the covariance are 0, but the diagonal elements differ).

We will consider a context in which $\boldsymbol{\Sigma}$ is not proportional to a known matrix such as $\boldsymbol{V}$, but where there is a known structure in terms of clusters: there will be correlations within but not between clusters (i.e. there is intra-cluster but no inter-cluster correlation). Suppose that there are a fixed number, G, of clusters and that $\boldsymbol{\Sigma}$ takes the following block-diagonal form (MacKinnon, Nielsen, and Webb, forthcoming, 1):

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_G \end{bmatrix} \tag{2}$$

where each $\boldsymbol{\Sigma}_g$ is a positive-definite square matrix associated with each cluster g, $1 \leq g \leq G$. Each cluster need not have the same number of associated observations; if we let n_g be the number of observations in cluster g, $\boldsymbol{\Sigma}_g \in \mathbb{R}^{n_g \times n_g}$. We do not necessarily know $\boldsymbol{\Sigma}$, and will assume that we need to estimate it as well.

3 The Method of Cluster-Robust Standard Errors

3.1 Cluster-Robust Variance Estimators and Standard Errors

First consider the general model in (1). As discussed in the introduction, one strategy to estimate $\boldsymbol{\beta}$ as well as its variance in this model is to use the standard OLS estimator $\hat{\boldsymbol{\beta}}$, but to determine standard errors for $\hat{\boldsymbol{\beta}}$ through special methods, such as cluster-robust variance estimators, that account for the true covariance structure. This is our focus, and will use this assumption that $\hat{\boldsymbol{\beta}}$ itself is the OLS estimator throughout this section. Taking the variance of $\hat{\boldsymbol{\beta}}$ in the context of the general model gives:

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}) &= \text{Var}((\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}) \\
&= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\text{Var}(\boldsymbol{y})((\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T})^\mathsf{T} \\
&\qquad \text{since } \text{Var}(\boldsymbol{AB}) = \boldsymbol{A}\text{Var}(\boldsymbol{B})\boldsymbol{A}^\mathsf{T} \text{ for constant matrix } \boldsymbol{A} \\
&= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}. \qquad (3)
\end{aligned}$$

In this context, a standard error for $\hat{\beta}_j$ (the jth element of $\hat{\boldsymbol{\beta}}$) is an estimator of $\sqrt{\text{Var}(\hat{\boldsymbol{\beta}})_{jj}}$. Overall, $\text{Var}(\hat{\boldsymbol{\beta}})$ is sometimes called a sandwich covariance matrix, as it takes the form of some element "sandwiched" between two instances of a different element (here, $\boldsymbol{X}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{X}$ is sandwiched by $(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}$) (Davidson and MacKinnon 2004, 198).

As in the lecture notes, if the errors are homoskedastic and uncorrelated (and thus non-clustered), or equivalently if $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, then (3) simplifies to $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}$, which implies that $\sqrt{\text{Var}(\hat{\beta}_j)} = \sigma\sqrt{(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}_{jj}}$. To estimate the standard deviation of $\hat{\beta}_j$ without knowing σ, we can use the square root of the estimator $\widehat{\text{Var}}(\hat{\beta}_j) = s\sqrt{(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}_{jj}}$. This estimator replaces σ with $s = \sqrt{s^2}$, where $s^2 = \frac{1}{n-p}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ is an unbiased estimator for σ^2. s^2 is a finite-sample correction for unbiasedness to the maximum likelihood estimate for σ^2, $\hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$. Observe that $\hat{\sigma}^2$ is an average of estimated residuals across many observations, encoded in the vector $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$:

$$\hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}})^2.$$

Now consider the clustered data setting. We show how to obtain standard errors that account for clustering, i.e. are cluster-robust, starting with the simplest form and then discussing analogous finite-sample corrections. Under the clustering model, suppose that the gth cluster, $1 \leq g \leq G$, corresponds with n_g data observations, which are contained in the $n_g \times p$ matrix $\boldsymbol{X}_g$; this matrix contains the rows of $\boldsymbol{X}$ belonging to the gth cluster. Then (3) can be rewritten (MacKinnon and Webb 2020, 2) as:

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G} \boldsymbol{X}_g^\mathsf{T}\boldsymbol{\Sigma}_g\boldsymbol{X}_g\right)(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}. \tag{4}$$

Cluster-robust variance estimators (sometimes called CRVEs) estimate the above by using a sample-based analogue to the middle factor in the right-hand side, $\sum_{g=1}^{G} \boldsymbol{X}_g^\mathsf{T}\boldsymbol{\Sigma}_g\boldsymbol{X}_g$. A simple cluster-robust variance estimator has the form (Cameron and Miller 2015, 324):

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G} \boldsymbol{X}_g^\mathsf{T}(\boldsymbol{y}_g - \boldsymbol{X}_g\hat{\boldsymbol{\beta}})(\boldsymbol{y}_g - \boldsymbol{X}_g\hat{\boldsymbol{\beta}})^\mathsf{T}\boldsymbol{X}_g\right)(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} \tag{5}$$

where we write $\boldsymbol{y}_g$ for the observations belonging to the gth cluster, analogously to the construction of $\boldsymbol{X}_g$. Unlike in the homoskedastic, uncorrelated and non-clustered case from class, where the unknown component of $\boldsymbol{\Sigma}$ was estimated by replacing it with an average across all estimated residuals as shown above, in this case we replace each cluster's variance matrix with a single sampling term for that cluster. Each of the sampling terms uses the vector of estimated residuals in an individual cluster, $\boldsymbol{y}_g - \boldsymbol{X}_g\hat{\boldsymbol{\beta}}$ (MacKinnon and Webb 2020, 4). As with sandwich covariance matrices, $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$ is a "sandwich estimator", as it takes the same form of a "filling" surrounded by two instances of "bread"; sandwich estimators describe a class of estimators that can be characterized by using similar ideas to achieve consistency.

Now, suppose that $G \to \infty$ and that the n_g do not vary as G changes (so the number of data points also goes to infinity). It can be shown that each $\boldsymbol{\Sigma}_g$ cannot possibly be consistently estimated by its sample counterpart since it is analogous to a single observation. However, the matrix $\left(\sum_{g=1}^{G} \boldsymbol{X}_g^\mathsf{T}\boldsymbol{\Sigma}_g\boldsymbol{X}_g\right)$ has a fixed size $p \times p$, and it turns out that it can be consistently estimated by the expression we replaced it with above (Davidson and MacKinnon 2004, 197).

Initial derivations of the estimator (5) showed it to be consistent as $G \to \infty$ for a finite number of observations per cluster (not necessarily the same for all clusters). More recently, researchers have shown it to be consistent even when the number of observations per cluster grows, in addition to the number of clusters going to infinity (Cameron and Miller 2015, 324). In all cases, though, the number of clusters needs to go to infinity for the estimator to be consistent, rather than just the number of observations.

The estimator (5) has been shown to be downward-biased. A number of finite-sample adjustment factors are used in practice for scaling the estimator (5) to mitigate bias, analogously to the scaling factor $\frac{n}{n-p}$ used to convert between the sample analogue of the variance and the unbiased estimator of the variance in the uncorrelated, homoskedastic case. The most common adjustment, which is used by Stata, is $\frac{G(n-1)}{(G-1)(n-p)}$, giving the alternative estimator

$$\widehat{\mathrm{Var}}_{\mathrm{R1}}(\hat{\boldsymbol{\beta}}) = \frac{G(n-1)}{(G-1)(n-p)}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g^\mathsf{T}(\boldsymbol{y}_g - \boldsymbol{X}_g\hat{\boldsymbol{\beta}})(\boldsymbol{y}_g - \boldsymbol{X}_g\hat{\boldsymbol{\beta}})^\mathsf{T}\boldsymbol{X}_g\right)(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} \quad (6)$$

In the trivial case where each cluster contains only one observation, so $G = n$, the clustered data reduces to heteroskedastic, uncorrelated data (the covariance matrix is diagonal) and the adjustment reduces to $\frac{n}{n-p}$, as intuitively expected; indeed, it can be shown that the estimator (5) reduces to a heteroskedasticity-robust variance estimator. An alternative correction is $\frac{G}{G-1}$, which is similar to the first when n is large (MacKinnon and Webb 2020, 4). However, neither of these corrections fully eliminates the downward bias of (5) in the general case; in fact, nor does any other known correction (Cameron and Miller 2015, 325).

Yet another variance estimator that has somewhat better finite-sample properties is

$$\widehat{\mathrm{Var}}_{\mathrm{R2}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g^\mathsf{T}\boldsymbol{M}_g^{-1/2}(\boldsymbol{y}_g - \boldsymbol{X}_g\hat{\boldsymbol{\beta}})(\boldsymbol{y}_g - \boldsymbol{X}_g\hat{\boldsymbol{\beta}})^\mathsf{T}\boldsymbol{M}_g^{-1/2}\boldsymbol{X}_g\right)(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}$$

where $\boldsymbol{M}_g^{-1/2}$ is some inverse square root of $\mathbf{I}_{n_g} - \boldsymbol{X}_g(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}_g^\mathsf{T}$. $\boldsymbol{M}_g^{-1/2}$ intuitively rescales the estimated residuals in each cluster. In practice, $\widehat{\mathrm{Var}}_{\mathrm{R1}}(\hat{\boldsymbol{\beta}})$ is the most used variance estimator (MacKinnon and Webb 2020, 4).

3.2 Confidence Intervals and Hypothesis Tests

We now turn to how inference can be performed by using these standard errors to obtain a test statistic with a known distribution, which can be used in inference. We have seen from lecture that if we assume in model (1) that $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, we can use $\text{s.e.}(\hat{\beta}_j) = s\sqrt{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{jj}}$ as an estimate for the standard deviation of $\hat{\beta}_j$. Moreover, under the assumption that residuals have a Normal distribution, the statistic $\frac{\hat{\beta}_j - \beta_j}{\text{s.e.}(\hat{\beta}_j)}$ is distributed as t_{n-p}.

In the general model with clustering, we can do the same as above to obtain an analogous estimator, referred to here as the Wald t-statistic, where $\text{s.e.}(\hat{\beta}_j)$ is instead given by the square root of the appropriate diagonal entry of (5) or some other cluster-robust variance estimator, that is, $\text{s.e.}(\hat{\beta}_j) = \sqrt{\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})}$. Under some technical conditions including a fixed number of clusters G and increasing n, this statistic has an asymptotic t_{G-1} distribution (MacKinnon and Webb 2020, 5):

$$\frac{\hat{\beta}_j - \beta_j}{\text{s.e.}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})_{jj}}} \dot{\sim} t_{G-1}.$$

For this reason, the t_{G-1} distribution is used in Stata as an asymptotic approximation for this statistic to compute confidence intervals and perform hypothesis tests. For example, a $(1-\alpha)\%$ confidence interval for β_j is

$$[\hat{\beta}_j - \text{s.e.}(\hat{\beta}_j)t_{1-\alpha/2,G-1}, \hat{\beta}_j + \text{s.e.}(\hat{\beta}_j)t_{1-\alpha/2,G-1}]$$

where $t_{1-\alpha/2,G-1}$ denotes the $1-\alpha/2$ quantile of the t_{G-1} distribution (4).

It has been formally shown that under some conditions where G increases with n to infinity, the Wald statistic tends to a $\mathcal{N}(0,1)$ distribution asymptotically, agreeing with the intuition that since $G \to \infty$, we get the limiting distribution $t_{G-1} \to \mathcal{N}(0,1)$ (Cameron and Miller 2015, 340). These conditions allow the cluster sizes n_g to increase but not too fast, and limit how much the cluster sizes can vary (MacKinnon and Webb 2020, 4). Assuming large G, this means the statistic can also be approximated as being distributed $\mathcal{N}(0,1)$, which is what some R packages do by default (Heiss 2021).

However, no exact distribution or properties of the statistic are known in the finite-sample case, even if the residuals have a Normal distribution; additionally, both these t- and Normal approximations do poorly when G is small (MacKinnon and Webb 2020, 14). This is unlike the unclustered case $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ where we had an exact distribution in small samples.

Cluster-robust standard errors can also be applied in hypothesis tests for general linear hypotheses. Specifically, consider testing the hypotheses H_0 : $(\mathbf{\Lambda}\boldsymbol{\beta} = \mathbf{0})$ and H_1 : ($\boldsymbol{\beta}$ is unrestricted) for some fixed matrix $\mathbf{\Lambda} \in \mathbb{R}^{\ell \times p}$. It can be shown that the statistic

$$(\mathbf{\Lambda}\hat{\boldsymbol{\beta}})^{\mathsf{T}}(\mathbf{\Lambda}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\mathbf{\Lambda}^{\mathsf{T}})^{-1}(\mathbf{\Lambda}\hat{\boldsymbol{\beta}})$$

where $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$ is a cluster-robust variance estimator as presented above, can be used as a statistic for this test, and can be approximated using the χ^2_ℓ distribution or ℓ times the $F_{\ell,G-1}$ distribution, as long as $\ell \ll G$; other bootstrapping methods exist (5).

3.3 Comparison to Non-Clustered Variance in a Special Case

In this section, we discuss an important special case of the model in (1) and (2), in which the variance given in (4) for the clustered model is an interpretable multiple of the variance expression for the unclustered (homoskedastic, uncorrelated) model. This provides additional insight into how the magnitude of cluster-robust standard errors will tend to differ from that of regular ones. Consider the special case where regression is univariate (with an intercept), where errors are homoskedastic, and the within-cluster correlation ρ is the same for all observations within and across clusters. We index observations by gi, where g is the cluster number and i is the number of the unit within the cluster. Then, we have:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{Gn_G} \end{bmatrix}, \qquad \mathbf{\Sigma}_g = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \text{ for all } g \tag{7}$$

where σ^2 and ρ are scalars.

Now, define $\text{Var}(\hat{\boldsymbol{\beta}})$ as in (4), and define $\widetilde{\text{Var}}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}$, which is the variance for the usual non-clustered model with homoskedastic, uncorrelated errors, as reviewed in the previous section. We can see $\widetilde{\text{Var}}(\hat{\boldsymbol{\beta}})$ as the variance we might calculate if we knew σ^2 but misspecified $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$. In the model above, it can be shown (Moulton 1986, 387) that the ratio between the two is:

$$\frac{\text{Var}(\hat{\boldsymbol{\beta}})}{\widetilde{\text{Var}}(\hat{\boldsymbol{\beta}})} = 1 + \left(\frac{\text{Var}(n_g)}{\bar{n}} + \bar{n} - 1\right)\rho_x\rho, \qquad \text{with } \rho_x = \frac{\sum_g \sum_j \sum_{i\neq j}(x_{gi} - \bar{x})(x_{gi} - \bar{x})}{\text{Var}(x_{gi})\sum_g n_g(n_g - 1)} \tag{8}$$

where $\bar{n}$ is the average of the group sizes n_g, and where ρ_x is a measure of the correlation of the regressor within clusters. This is often called the "Moulton factor", or sometimes a "variance inflation factor" (Cameron and Miller 2015, 322), which does not however bear any close relationship to the one seen in Stat 244.

This expression relates variances depending on the unknown σ^2, rather than standard errors that can be computed from the data and that are square roots of estimates of those variances, like that in (5). Nevertheless, it provides intuition about how cluster-robust standard errors will tend to differ from regular ones:

- The ratio is increasing in the within-cluster correlations of the regressor and of the error. If either is zero, or if all clusters have size 1, then the ratio is 1 and the two variance expressions are equal.
- The ratio is increasing in the variance of the cluster size.
- If all clusters have the same size (equivalently, $\text{Var}(n_g) = 0$), then the ratio is increasing in the group size.

The last point can be reformulated as saying that the fewer clusters there are (for a fixed total n), the more important it is that the model be correctly specified when calculating the variance of $\hat{\boldsymbol{\beta}}$.

4 Feasible Generalized Least Squares (FGLS)

While most of the paper has focused on deriving correct standard errors for OLS estimates in a setting with clustering, there are other estimators of $\hat{\beta}$ itself that can be used. In particular, the feasible generalized least squares (FGLS) estimator often leads to improvements in efficiency. In this section, which serves as a discussion of one extension of the cluster-robust standard error method, we develop this estimator, explain its relationship to the GLS development from the lecture notes, and describe how the idea of cluster-robust standard errors can be extended to guard against the extra assumptions that are made in deriving the FGLS estimator.

The FGLS model arises from some components of the generalized least squares (GLS) model being unknown. In the lecture notes, the GLS model was developed in a setting where our model (1) had $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{V}$ with σ^2 being unknown and $\boldsymbol{V}$ being a known $n \times n$ positive-definite matrix. In that context, the following estimator was derived: $\hat{\boldsymbol{\beta}}_{\text{GLS}} = \left(\boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{y}$. This is only a computable estimator when $\boldsymbol{V}$ is known. The FGLS method extends GLS by using a consistent estimator of $\boldsymbol{\Sigma}$ instead of requiring knowing the entire structure of $\boldsymbol{V}$ aside from σ^2. This is only possible if a model is specified for intracluster error correlation, allowing estimates of the intracluster error parameters to be generated. This approach is suitable under the assumption that the model for intracluster error correlation is correctly specified, and leads to valid parameter estimates that are more efficient than ordinary least squares under statistical inference.

We consider the GLS expression $(\boldsymbol{X}^\intercal \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{\Sigma}^{-1} \boldsymbol{y}$ with $\boldsymbol{\Sigma}$ being block-diagonal and containing the $\boldsymbol{\Sigma}_g$ as in (3). Suppose that we have available a consistent estimator $\hat{\boldsymbol{\Sigma}}_g$ of $\boldsymbol{\Sigma}_g$, derived from a model for the intracluster error correlations. The FGLS estimator of $\boldsymbol{\beta}$ is then (Cameron and Miller 2015, 325):

$$\hat{\boldsymbol{\beta}}_{\text{FGLS}} = \left(\sum_{g=1}^{G} \boldsymbol{X}_g^\intercal \hat{\boldsymbol{\Sigma}}_g^{-1} \boldsymbol{X}_g \right)^{-1} \sum_{g=1}^{G} \boldsymbol{X}_g^\intercal \hat{\boldsymbol{\Sigma}}_g^{-1} \boldsymbol{y}_g$$

with $\boldsymbol{X}_g$ and $\boldsymbol{y}_g$ defined as in (5).

Interestingly, the method of cluster-robust standard errors from Section 3 can be employed

to guard against the possibility that the model for the intracluster correlations was incorrectly specified, such that our $\hat{\boldsymbol{\Sigma}}_g$ was not actually consistent. The cluster-robust estimate of the above variance matrix of the FGLS estimator is represented as follows, paralleling the OLS counterpart (5) (Cameron and Miller 2015, 325):

$$\widehat{\mathrm{Var}}_{\mathrm{R}}(\hat{\boldsymbol{\beta}}_{\mathrm{FGLS}}) = (\boldsymbol{X}^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}_g^{-1}\hat{\boldsymbol{\varepsilon}}_g\hat{\boldsymbol{\varepsilon}}_g^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}_g^{-1}\boldsymbol{X}_g\right)(\boldsymbol{X}^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X})^{-1} \tag{9}$$

where $\hat{\boldsymbol{\varepsilon}}_g = \boldsymbol{y}_g - \boldsymbol{X}_g\hat{\boldsymbol{\beta}}_{\mathrm{FGLS}}$.

To briefly illustrate what is meant by a model of the intracluster correlations, recall model (7), in which each $\boldsymbol{\Sigma}_g$ was assumed to be identical, and in which all the intracluster correlations between different observations were assumed to be ρ. In such a case, we can easily obtain method-of-moments estimators of $\hat{\sigma}^2$ and $\hat{\rho}$ and use those to calculate $\hat{\boldsymbol{\Sigma}}_g$ (Trivedi and Cameron 2005, 835, 837). Then, the use of cluster-robust standard errors for the FGLS estimator as in (9) will allow the standard errors to remain accurate in the eventuality that the model for the $\boldsymbol{\Sigma}_g$'s is incorrect, while allowing greater efficiency if it is correct.

FGLS is considered to be asymptotically more efficient than other methods. Generally, there is no obvious method to determine when the efficiency improvements under the FGLS method are large. For finite samples, FGLS may lead to minimal improvements in efficiency, although these minimal efficiency improvements can still be valuable (Cameron and Miller 2015, 318).

5 Coding Examples

5.1 Synthetic Demonstration

In this first example, we construct a one-dimensional synthetic model with clustered errors, and use it to compare different measures of the standard deviation of an estimated coefficient. The advantage of this approach is that our model assumptions are guaranteed to be satisfied (e.g., residuals are uncorrelated across clusters). Our code is given in Appendix A and

implements all calculations from scratch.

We set $n = 500$ and $G = 40$, and generate clusters by assigning each observation equal probability $\frac{1}{G}$ of being in each cluster. For each cluster, we generate a random symmetric, positive-definite matrix $\boldsymbol{\Sigma}_g$, combining them to create the block-diagonal $\boldsymbol{\Sigma}$. We let $p = 2$, modelling the data matrix $\boldsymbol{X}$ with a column of 1s and a covariate column of n independent samples from $\mathcal{N}(0, 1)$. We set the true parameters $\beta_0 = 5, \beta_1 = 10$, giving a model specification obeying our setup from (1) and (2). Knowing $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, the standard deviation in this clustered setting (the square root of $\text{Var}(\hat{\beta}_1)$, found from (4)) is 0.0576.

In each of 500 runs, we treat $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}$ as unknown, but assume $\boldsymbol{X}$ and the cluster assignments are given (and fixed across runs). We generate noise $\boldsymbol{\varepsilon}$ with covariance given by $\boldsymbol{\Sigma}$, yielding $\boldsymbol{y}$. We obtain the OLS estimate $\hat{\boldsymbol{\beta}}$. For the slope $\hat{\beta}_1$, we compute the non-robust standard error under the homoskedastic, uncorrelated assumption, and the cluster-robust standard error $\widehat{\text{Var}}_{\text{R1}}(\hat{\beta}_1)$ from (6). These two standard errors took average values of 0.0675 and 0.0565 respectively. The sample standard deviation of $\hat{\beta}_1$ across all runs was 0.0589.

We can make a few observations. First, the sample standard deviation of $\hat{\beta}_1$ is close to the theoretical standard deviation, aligning with our expectations of convergence. Second, the sample mean of the cluster-robust standard errors are close to the theoretical quantity they seek to estimate. In comparison, the average non-robust standard error differs significantly, illustrating the importance of using the robust methods we described. Here, the non-robust error was higher than the robust error; in practice, the opposite is usually the case. For intuition for how this can happen, note that in our special case from earlier, this could have been indicated to happen if $\rho < 0$, according to (8).

5.2 Analyzing Datasets in R

We consider two real-world datasets in which clustering can be used to model the data, and where cluster-robust standard errors are substantially different from non-robust standard errors, i.e., standard errors in the homoskedastic, uncorrelated case. Code and diagrams

are in Appendix B. Here we rely upon the `sandwich` (Zeileis, Köll, and Graham 2020) and `lmtest` (Zeileis and Hothorn 2002) R packages, configuring them to use the finite-sample adjustment from (6).

The first example considers data collected on the physical qualities of penguins (Horst, Hill, and Gorman 2020). Adapting a discussion given by Heiss (2021), we regress weight on bill length and flipper length (with intercept) and obtain an OLS estimate $\hat{\boldsymbol{\beta}}$. Considering a plot of the estimated residuals against estimates for each observation, however, shows that the residuals can be grouped into species-specific clusters (see Figure 1), suggesting the presence of species-specific error dependence. Using non-robust and heteroskedasticity-robust standard errors for the coefficients on bill length and flipper length give smaller estimated values, being approximately 7 and 3, approximately. Clustering by species and using a cluster-robust standard error, by comparison, gives a significantly higher value corresponding to a wider 95% confidence interval for the coefficient values. Note that cluster-robust inference tells us that there is much higher uncertainty in the OLS estimates; for instance, the confidence interval for the coefficient of bill length changes from $[47.2, 72.9]$ under the non-robust setting to $[5.3, 115.0]$ under the cluster-robust setting with 3 clusters (using the t distribution with $G - 1 = 2$ degrees of freedom). See Figure 2.

The second dataset is experimental data collected by Giné and Mansuri (2018) investigating the effect of several treatments (including instructing individuals about the importance of voting, instructing them about vote privacy, etc.) on the voter turnout of women in Pakistan, across 67 neighborhoods in 9 villages. This example is also used to study cluster robustness in MacKinnon and Webb (2020, 22). The outcomes $\boldsymbol{y}$ are binary, indicating whether or not a woman voted. We consider a simplification of one experiment run by Giné and Mansuri, regressing the outcomes on indicators of two treatments T_1 and T_2, using OLS (Giné and Mansuri use weighted least-squares and also add regressors for demographics) (Giné and Mansuri 2018, 221). Again, a non-robust error gives the smallest standard error and smallest CI for the coefficients of T_1 and T_2, as compared to cluster-robust standard errors when

clustering by village or neighborhood. The standard errors clustered by neighborhood (which are the finest clustering) produce the widest CIs for the coefficient estimates, compared to the standard errors clustered by village and to the non-robust standard errors, as seen in Figure 3. Here too, the confidence intervals with clustering were computed using the appropriate t_{G-1} distribution.

6 Discussion and Conclusion

This report has looked at cluster-robust standard errors and inference to address the limitations of traditional regression approaches that assume homoskedastic and uncorrelated error terms. This strong assumption may lead to systematically inaccurate estimates of parameter variances and, therefore, inference results. We primarily discussed the use of different estimators for the parameter variances, which work well when the data can be modelled as belonging to clusters, with intra-cluster error correlation. These methods translate into inferential procedures, and are being used with increasing frequency, particularly in the economics and econometrics literature. As the data examples suggest, the prevalence of clustered data in real life demonstrates the wide range of applications for these clustered standard errors. We also described how a model for clustering structure can be used to obtain more efficient parameter estimates using FGLS, and how cluster-robust standard errors can be used to protect against misspecification of such a model.

There is a growing body of literature to examine additional methods for dealing with clustered standard errors beyond the ones considered in this paper. Many of these focus on the problem of "few clusters"; this is heavily tied to the fact that the t_{G-1} distribution mentioned earlier only holds asymptotically, and that inferential procedures may not be very accurate with low G. These alternative methods rely on bootstrapping: for example, the wild cluster bootstrap method and the pairs cluster bootstrap procedure offer advancements useful for cases with few clusters (Esarey and Menger 2019). These bootstrapping procedures can also

be applied to extensions of cluster-robust methods, such as "multi-way" clustering, in which data is modeled as being clustered in two or more dimensions, and performing hypothesis tests. Cluster-robust standard errors have been applied to some generalized linear model settings, including probit and logit models (Cameron and Miller 2015, 354). Additionally, we have assumed throughout that our initial parameter estimates were derived by analyzing the entire dataset at once; other extensions for clustered standard errors consider situations where clusters are individually estimated (MacKinnon and Webb 2020, 5, 17).

The topic of cluster-robust standard errors can broadly be framed as one specific set of methods that can be used to address the overall issue of applying regression to clustered data. As some argue, clustered standard errors also offer more flexibility and robustness over other common ways to deal with clustering, including hierarchical models (6). At the same time, while there exist guarantees for the cluster-robust variance estimators consistently estimating the true variance of the data, there are still other issues common to clustered data which motivate further areas of inquiry. These include inferential problems that arise when the number of clusters G is too small and the sizes of clusters severely vary, as well as overall methodological considerations as to how to determine the number of clusters and how to organize data into clusters when such information is not known *a priori*—in addition to the overall question of when modelling data as clustered is even appropriate. While these concepts fall out of scope here, they reflect additional important areas of work in statistical inference and econometrics, reflecting ongoing efforts to improve accurate modelling and inference on real-world data.

Bibliography

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2022. "When Should You Adjust Standard Errors for Clustering?" *The Quarterly Journal of Economics* (October). https://doi.org/10.1093/qje/qjac038.

Agresti, Alan. 2015. *Foundations of Linear and Generalized Linear Models.* Wiley. ISBN: 978-1-118-73003-4.

Angrist, Joshua David, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton: Princeton University Press. ISBN: 978-0-691-12034-8.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2011. "Robust Inference with Multiway Clustering." *Journal of Business & Economic Statistics* 29, no. 2 (April): 238–249. https://doi.org/10.1198/jbes.2010.07136.

Cameron, A. Colin, and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50 (2): 317–372. https://doi.org/10.3368/jhr.50.2.317.

Davidson, Russell, and James G. MacKinnon. 2004. *Econometric Theory and Methods.* New York: Oxford University Press. ISBN: 978-0-19-512372-2. http://qed.econ.queensu.ca/ETM/ETM-davidson-mackinnon-2021.pdf.

Esarey, Justin, and Andrew Menger. 2019. "Practical and Effective Approaches to Dealing With Clustered Data." *Political Science Research and Methods* 7 (3): 541–559. https://doi.org/https://doi.org/10.1017/psrm.2017.42. https://pubs.aeaweb.org/doi/pdfplus/10.1257/app.20130480.

Giné, Xavier, and Ghazala Mansuri. 2018. "Together We Will: Experimental Evidence on Female Voting Behavior in Packistan." *American Economic Journal: Applied Economics* 10 (1): 207–235. https://doi.org/https://doi.org/10.1257/app.20130480. https://pubs.aeaweb.org/doi/pdfplus/10.1257/app.20130480.

Heiss, Andrew. 2021. *Robust and clustered standard errors with R.* Technical report. https://evalf21.classes.andrewheiss.com/example/standard-errors/.

Horst, Allison M, Alison P Hill, and Kristen B Gorman. 2020. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data.* R package version 0.1.0. https://doi.org/10.5281/zenodo.3960218. https://allisonhorst.github.io/palmerpenguins/.

MacKinnon, James G., Morten Ørregaard Nielsen, and Matthew D. Webb. Forthcoming. "Cluster-Robust Inference: A Guide To Empirical Practice." *Journal of Econometrics,* https://doi.org/10.1016/j.jeconom.2022.04.001.

MacKinnon, James G., and Matthew D. Webb. 2020. *When and How to Deal with Clustered Errors in Regression Models.* Queen's Economics Department Working Paper, 1421, May. Accessed November 21, 2022. http://qed.econ.queensu.ca/pub/faculty/mackinnon/working-papers/qed_wp_1421.pdf.

Moulton, Brent R. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics* 32, no. 3 (August): 385–397. https://doi.org/10.1016/0304-4076(86)90021-7.

Trivedi, Pravin K., and A. Colin Cameron. 2005. *Microeconometrics.* Cambridge University Press. ISBN: 978-0-511-12495-2.

Zeileis, Achim, and Torsten Hothorn. 2002. "Diagnostic Checking in Regression Relationships." *R News* 2 (3): 7–10. https://CRAN.R-project.org/doc/Rnews/.

Zeileis, Achim, Susanne Köll, and Nathaniel Graham. 2020. "Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in *R*." *Journal of Statistical Software* 95 (1). https://doi.org/10.18637/jss.v095.i01.

Appendix A Code for synthetic example

```
import numpy as np
import sklearn.datasets as skd
import scipy.linalg as spl

np.random.seed(244)

num_sims = 500

n = 500
G = 40

betas = [5, 10]
p = len(betas)

# Values of the CR and classical standard errors and of slope estimates to
↪ be collected
crse_vals = []
classical_se_vals = []
betahat_vals = []

# Generate cluster sizes
cluster_sizes = np.random.multinomial(n, [1/G]*G)

# Mapping of observations to cluster indices
cluster_indices = []
for i in range(G):
    cluster_indices.append(np.repeat(i, cluster_sizes[i]))
cluster_indices = np.concatenate(cluster_indices)

# Generate matrices £\bdSigma_g£ that are symmetric, positive definite
# and diagonal
cluster_cov_matrices = [skd.make_spd_matrix(cluster_sizes[i]) for i in
↪ range(G)]
cov_matrix = spl.block_diag(*cluster_cov_matrices)

# Generate covariates
x = np.random.randn(n)

# Calculate the theoretical cluster-robust standard deviation of the
↪ coefficient estimators
# (Formula 4 from Section 3.1, using the most common finite-sample
↪ adjustment)
```

```
x_matrix = np.column_stack((np.ones(n), x))  # Create the design matrix
xtx_inv = np.linalg.inv(x_matrix.T @ x_matrix)  # Calculate (X'X)^(-1)
xtsx = x_matrix.T @ cov_matrix @ x_matrix  # calculate X'ΣX
theoretical_crsd = np.sqrt((xtx_inv @ xtsx @ xtx_inv).diagonal())

# Run the simulations
for i in range(num_sims):
    errors = np.random.multivariate_normal(np.zeros(n), cov_matrix)
    y = (betas[0] + betas[1] * x + errors)[:, np.newaxis]

    betahat = xtx_inv @ x_matrix.T @ y

    # Add the estimated slope to our list
    betahat_vals.append(betahat[1])

    residuals = y - x_matrix @ betahat
    # "Lowercase s", classical estimate of the standard deviation
    classical_s = 1 / (n - p) * residuals.T @ residuals

    # Calculate the classical standard error of the slope
    classical_se = classical_s * np.sqrt(xtx_inv[1, 1])
    classical_se_vals.append(classical_se)

    # Calculate the cluster-robust standard error of the slope
    # (Formula 4 from Section 3.1, using the most common finite-sample adjustment)
    filling_matrix = np.zeros((2, 2))
    for i in range(G):
        x_g_matrix = x_matrix[cluster_indices == i, :]
        assert(x_g_matrix.shape == (cluster_sizes[i], p))
        y_g = y[cluster_indices == i]
        assert(y_g.shape == (cluster_sizes[i], 1))

        residuals_g = y_g - x_g_matrix @ betahat
        filling_matrix += x_g_matrix.T @ residuals_g @ residuals_g.T @ x_g_matrix
    crvar = (G * (n - 1)) / ((G - 1) * (n - p)) * xtx_inv @ filling_matrix @ xtx_inv
    assert(crvar.shape == (p, p))

    crse_vals.append(np.sqrt(crvar[1, 1]))

print(f"Theoretical CRSD of betahat_1: {theoretical_crsd[1]:.4f}")

```

```python
# Calculate the average classical standard error of the slope across simulations
mean_classical_se = np.mean(classical_se_vals)
print(f'Sample mean of classical SE of betahat_1: {mean_classical_se:.4f}')

# Calculate the average cluster-robust standard error of the slope across simulations
mean_crse = np.mean(crse_vals)
print(f'Sample mean of CRSE of betahat_1: {mean_crse:.4f}')

# Calculate the standard deviation of the estimated slopes across simulations
sd_betahat = np.std(betahat_vals)
print(f'Sample standard deviation of betahat_1: {sd_betahat:.4f}')

from scipy import stats
from matplotlib import pyplot as plt
import statsmodels.api as sm

wald_stats = (np.array(betahat_vals).flatten() - betas[1]) / np.array(crse_vals)
sm.qqplot(wald_stats, dist=stats.t, distargs=(G-1,), line="s")

plt.title("Q-Q Plot of CR Wald Stat vs. t(G-1)")
plt.xlabel("Theoretical quantiles")
plt.ylabel("Sample quantiles")
plt.show()

sm.qqplot(wald_stats, line="s")

plt.title("Q-Q Plot of CR Wald Stat vs. Standard Normal")
plt.xlabel("Theoretical quantiles")
plt.ylabel("Sample quantiles")
plt.show()
```

Appendix B Code and diagrams for dataset example

```r
# For reproducibility of results
set.seed(244)

# Import libraries
library(haven)
```

```r
library(tidyverse)
library(broom)
library(palmerpenguins)
library(modelsummary)
library(sandwich)
library(lmtest)

#####

# PENGUIN DATA
# Use palmerpenguins' dataset and drop missing values
penguin_data <- penguins
penguin_data <- penguin_data %>% drop_na(sex)

# Regress body mass on bill length, flipper length, and species
# Assume OLS estimation of the parameters (beta) throughout
# Non-robust: assuming homoskedastic and uncorrelated errors
modelp_simple <- lm(body_mass_g ~ flipper_length_mm + bill_length_mm
                                  + species,
                    data = penguin_data)
tidy(modelp_simple, conf.int=T) %>%
  filter(term %in% c("bill_length_mm", "flipper_length_mm"))

# Plot the residuals against fitted values to see that there is clustering
residuals <- augment(modelp_simple, data = penguin_data)
ggplot(residuals, aes(x=.fitted, y=.resid)) +
  geom_smooth(method="lm") +
  geom_point(aes(color=species))

# As it turns out, using a model that is only heteroskedasticity-robust
# also doesn't give great standard error estimates
modelp_hetero_robust <- coeftest(modelp_simple,
                                 vcov = vcovHC)
tidy(modelp_hetero_robust, conf.int=T) %>%
  filter(term %in% c("bill_length_mm", "flipper_length_mm"))

# Cluster-robust standard errors, clustering by species
# and specifying a t(2) distribution
# Given G clusters, the df for t distribution is G-1
modelp_robust_clustered <- coeftest(modelp_simple,
                                    vcov = vcovCL,
                                    type = "HC1",
                                    df = 2,
                                    cluster = ~species)
tidy(modelp_robust_clustered, conf.int=T) %>%
```

```r
  filter(term %in% c("bill_length_mm", "flipper_length_mm"))

# Compare all CIs
modelplot(
  list("LM, cluster-robust (3 clusters)" = modelp_robust_clustered,
       "LM, heteroskedasticity-robust" = modelp_hetero_robust,
       "LM, non-robust" = modelp_simple),
  coef_omit = "species|Intercept") +
  guides(color=guide_legend(reverse=T))

#####

# VOTING DATA
# Dataset used is available at
# https://www.openicpsr.org/openicpsr/project/113595/version/V1/view
# Read Stata file
votingdata <- read_dta('female_voting_data.dta')
# Drop empty values
votingdata <- votingdata %>% drop_na(t1)
votingdata <- votingdata %>% drop_na(t2)
votingdata <- votingdata %>% drop_na(voted)

# Regress whether or not a woman voted (binary) on two treatments t1, t2
# Use weighted least-squares estimation of the parameters (beta)
# Non-robust: assuming homoskedastic and uncorrelated errors
modelv_simple <- lm(voted ~ t1 + t2,
                    data = votingdata)
tidy(modelv_simple, conf.int=T) %>% filter(term %in% c("t1", "t2"))

# Cluster-robust standard errors, clustering by village (9)
modelv_robust_clustered_village <- coeftest(modelv_simple,
                                            vcov = vcovCL,
                                            type = "HC1",
                                            cluster = ~village_code,
                                            df = 8)

tidy(modelv_robust_clustered_village, conf.int=T) %>%
  filter(term %in% c("t1", "t2"))

# Cluster-robust standard errors, clustering by neighborhood (67)
modelv_robust_clustered_neighborhood <- coeftest(modelv_simple,
                                                 vcov = vcovCL,
                                                 type = "HC1",
                                                 cluster =
                                                 ↪  ~neighborhood_code,
```

```
                                                  df = 66)
tidy(modelv_robust_clustered_neighborhood, conf.int=T) %>%
  filter(term %in% c("t1", "t2"))

# Compare all CIs
modelplot(
  list("LM, cluster-robust, cluster by neighborhood (67 clusters)" =
         modelv_robust_clustered_neighborhood,
       "LM, cluster-robust, cluster by village (9 clusters)" =
       ↪  modelv_robust_clustered_village,
       "LM, non-robust" = modelv_simple),
  coef_omit = "Intercept") +
  guides(color=guide_legend(reverse=T))
```

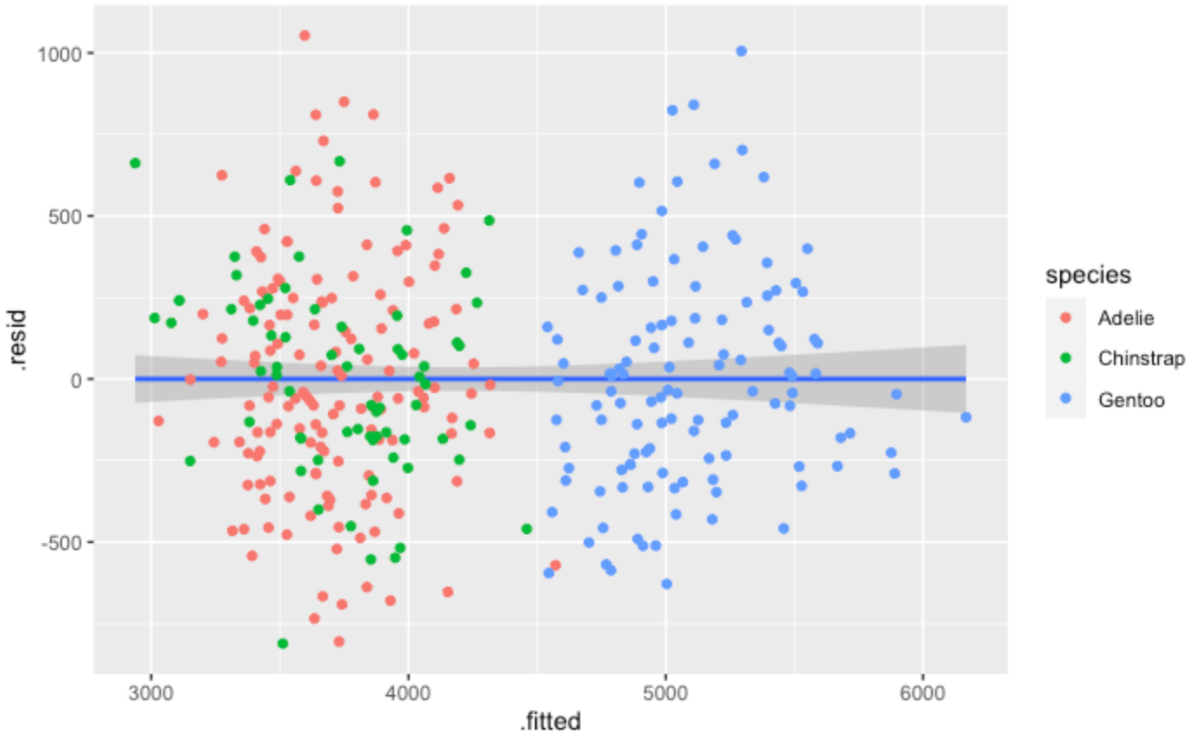

Figure 1: Plots of residuals against fitted values for predicting penguin body mass, color-coded by species. This visually depicts the presence of species-specific clusters.

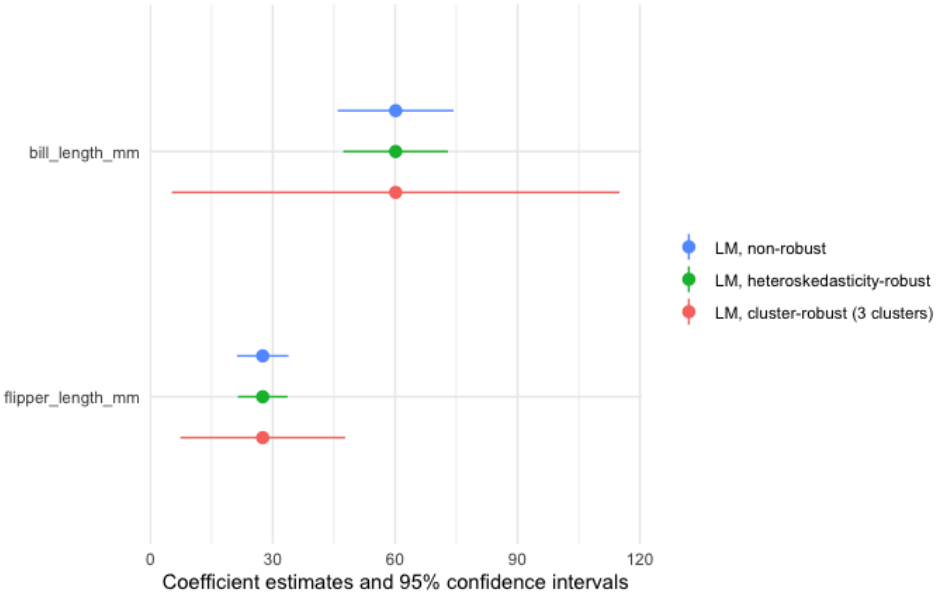

Figure 2: 95% confidence interval estimates for the regression coefficients of bill length and flipper length on predicting penguin body mass under different unrobust and cluster-robust standard errors.

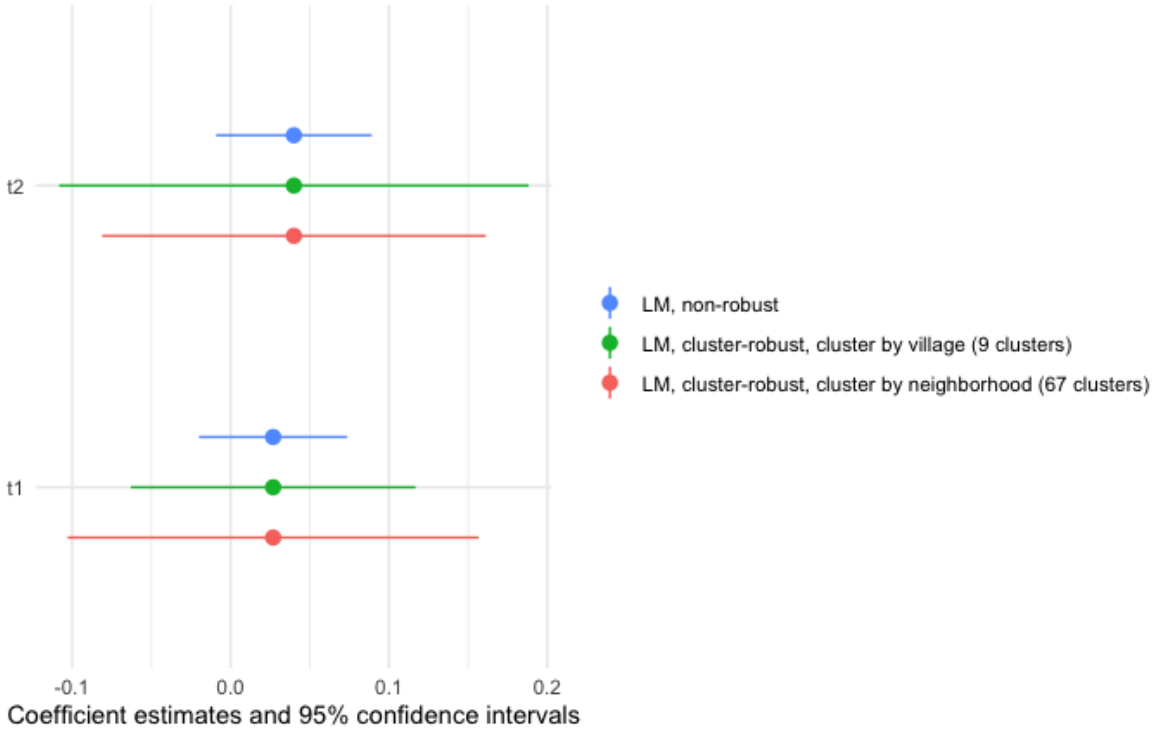

Figure 3: 95% confidence interval estimates for the regression coefficients of two treatments on predicting individual voter turnout among Pakistani women under different unrobust and cluster-robust standard errors.