Part II:

General linear models for dependent data

The general linear model

Definition

- By a *general linear model for dependent data*, we mean a statistical model with the following assumptions/components:

  ⋆ for the k^{th} cluster, given $\boldsymbol{X}_k$, we have that:

$$
\begin{aligned}
\mathsf{E}[\boldsymbol{Y}_k |\ \boldsymbol{X}_k] &= \boldsymbol{\mu}_k = \boldsymbol{X}_k\boldsymbol{\beta} \\
\mathsf{Cov}[\boldsymbol{Y}_k] &= \boldsymbol{\Sigma}_k
\end{aligned}
$$

  - $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{kn_k})^T$ is an $n_k \times 1$ mean vector
  - $\boldsymbol{\beta}$ is a p-vector of regression coefficients
  - $\boldsymbol{\Sigma}_k$ is an $n_k \times n_k$ covariance matrix

  ⋆ responses across clusters are independent of each other

- Sometimes it will be useful to use a representation of problem that encompasses all K clusters into a single matrix notation

- Towards this, let $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_K)^T$ denote the $N \times 1$ vector of responses and $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K)^T$ the $N \times p$ matrix of covariates for all study units across all clusters

- Finally, let

$$\boldsymbol{\mu} = \mathsf{E}[\boldsymbol{Y} \mid \boldsymbol{X}] = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)^T$$

denote the $N \times 1$ vector of response means and

$$\mathsf{Cov}[\boldsymbol{Y}] \equiv \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\Sigma}_K \end{bmatrix}$$

the $N \times N$ variance-covariance matrix for $\boldsymbol{Y}$

- While independence across clusters provides a clear simplification of the form of $\mathbf{\Sigma}$, we may also want/need to put some structure on the component $\mathbf{\Sigma}_k$ sub-matrixes

- The form of $\mathbf{\Sigma}_k$ can depend on many things:
  - ⋆ the value(s) of certain covariates, including time
  - ⋆ design considerations
    - ∗ e.g. whether the data are balanced or unbalanced

- While substantive knowledge and exploratory data analyses can help guide how one approaches this, it is worth considering a few examples

Specification of $\mathbf{\Sigma}_k$: Example #1

- For some clustered data settings, one option is that the correlation is common to all pairs of observations:

$$\mathbf{\Sigma}_k = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \ldots & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ \rho & \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots & \\ \rho & \rho & \rho & \ldots & 1 \end{bmatrix}$$

  - ⋆ same value of ρ across the K clusters

- Referred to as an *exchangeable* or *compound symmetric* structure

- May be reasonable for the CMS data which consists of patients within hospitals

Specification of $\boldsymbol{\Sigma}_k$: Example #2

- For longitudinal settings, one might adopt a correlation matrix that is a function of the distance between two observations:

$$\boldsymbol{\Sigma}_k = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n_k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n_k-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n_k-3} \\ \vdots & \vdots & \ddots & \vdots & \\ \rho_{n_k-1} & \rho_{n_k-2} & \rho_{n_k-3} & \cdots & 1 \end{bmatrix}$$

  ⋆ take the same set of values of $(\rho_1, \ldots, \rho_{n_k-1})$ across the K clusters

- Referred to as a *banded* correlation structure

- May be reasonable for equally spaced observations such as the dental growth data

Specification of $\boldsymbol{\Sigma}_k$: Example #3

- Building on Example # 2, one might adopt a correlation matrix that decays as a function of time between observations:

$$\boldsymbol{\Sigma}_k = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{n_k-2} \\ \rho & 1 & \rho & \ldots & \rho^{n_k-2} \\ \rho^2 & \rho & 1 & \ldots & \rho^{n_k-3} \\ \vdots & \vdots & \ddots & \vdots & \\ \rho^{n_k-1} & \rho^{n_k-2} & \rho^{n_k-3} & \ldots & 1 \end{bmatrix}$$

  ⋆ the same value of ρ across the K clusters

- Referred to as an *auto-regressive* correlation structure

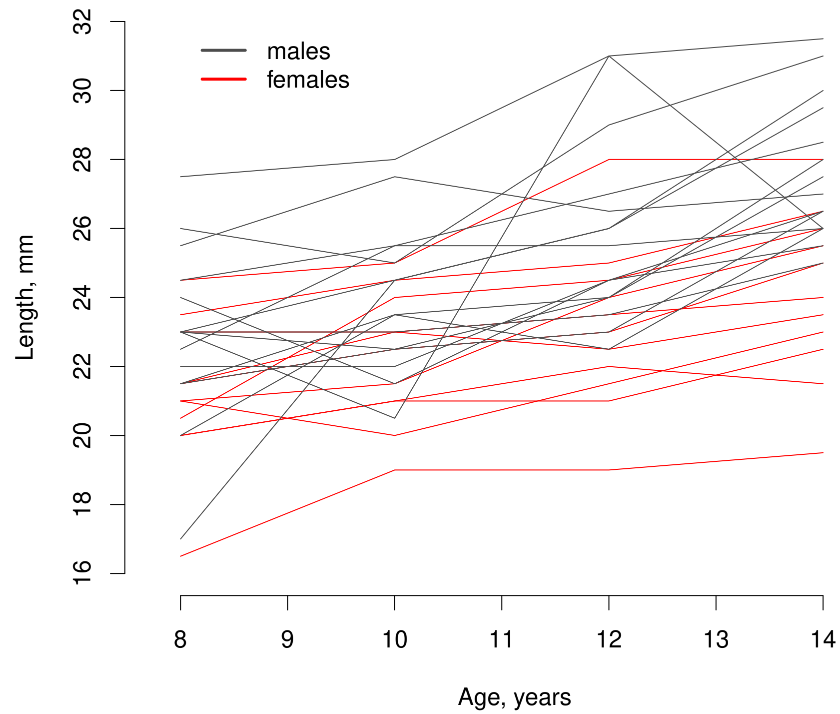
Estimation/inference

- For any given specification of the mean and covariance models, we would like
  - ⋆ consistent estimation
  - ⋆ valid inference
- While primary interest often lies with the mean model, estimation/inference for $\boldsymbol{\beta}$ is generally intertwined with $\boldsymbol{\Sigma}$
- Consider two broad set of tools:
  - ⋆ least squares estimation/inference
  - ⋆ likelihood-based estimation/inference

Two-stage least squares

- Recall the dental growth data

- Suppose the goal is to estimate and formally compare the average growth trajectory between males and females



- One strategy for analyzing these data could be to:

(1) estimate the growth trajectory for each child

(2) characterize the variation in the child-specific coefficients between the males and females

Stage 1

- Estimate subject-specific growth trajectories

- One simple model is to assume that the k^{th} subject's response vector varies randomly around a linear growth curve:

$$\boldsymbol{Y}_k \;=\; \boldsymbol{Z}_k\boldsymbol{\beta}_k \;+\; \boldsymbol{\epsilon}_k$$

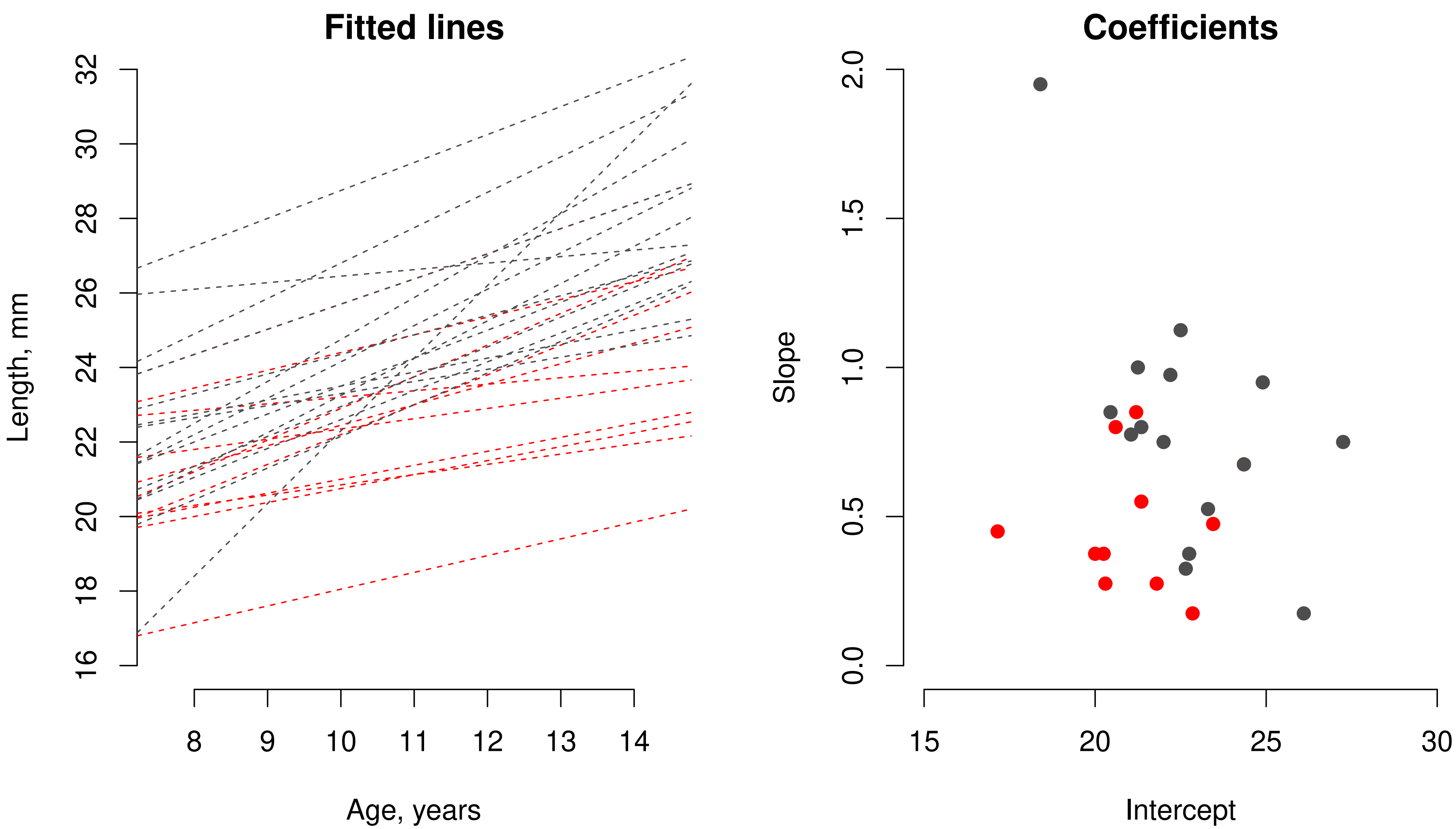
  ⋆ $\boldsymbol{Z}_k \subset \boldsymbol{X}_k$ are restricted to be within-subject or time-dependent

- $\boldsymbol{\epsilon}_k$ represent observation-specific random variations around each subjects' underlying growth curve

  ⋆ assumed to be i.i.d with mean 0 and variance σ_k^2

- Use subject-specific OLS of $\boldsymbol{Y}_k$ on $\boldsymbol{Z}_k$ to obtain estimates, $\widehat{\boldsymbol{\beta}}_k$

- Results for the dental growth data:
  - ⋆ after standardizing age to ensure the intercepts are interpretable



Stage 2

- Explain variation across the subject-specific coefficient estimates

- For example, one could assume that the $\boldsymbol{\beta}_k$ are a random sample from some population for which

$$\boldsymbol{\beta}_k = \boldsymbol{W}_k \boldsymbol{\beta} + \boldsymbol{\gamma}_k$$

  ⋆ $\boldsymbol{W}_k \subset \boldsymbol{X}_k$ are restricted to be subject-specific or time-invariant

- $\boldsymbol{\gamma}_k$ represent cluster-specific random variation around the population growth curve

  ⋆ assumed to be i.i.d with mean 0 and variance-covariance matrix $\boldsymbol{G}$

- Use OLS of the 'observed' $\widehat{\boldsymbol{\beta}}_k$ on $\boldsymbol{W}_k$ to obtain estimates, $\widehat{\boldsymbol{\beta}}$

```
> ## Stage 1
> ##
> betaMat <- data.frame(gender=rep(NA, K), beta0=rep(NA, K), beta1=rep(NA, K))
> for(k in 1:K)
+ {
+   temp.k <- growth[growth$id == k,]
+   fit.k  <- lm(length ~ ageStar, data=temp.k)
+   betaMat[k,2:4] <- c(temp.k$gender[1], fit.k$coef)
+ }
>
> ## Stage 2
> ##
> summary(lm(beta0 ~ gender, data=betaMat))
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.7182     1.3991  14.094 4.18e-13 ***
gender        1.4909     0.8466   1.761    0.091 .
...
```

- Marginal evidence of a difference in average length between males and females at age 8

```
>
> summary(lm(beta1 ~ gender, data=betaMat))
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1591     0.2280   0.698    0.492
gender        0.3205     0.1380   2.322    0.029 *
...
```

- Suggestive of a significant difference in the slope of the growth trajectories between males and females

Issues

- The design matrix is constrained at each stage of the analysis
  - ⋆ stage 1 is restricted to within-subject covariates
  - ⋆ stage 2 is restricted to between-subject covariates

- Information is lost by having summarized the response vector for subject k at stage 1

- Noting that the $\widehat{\boldsymbol{\beta}}_k$'s are statistics (i.e. just summaries of the data), the fact that they may arise on the basis of a different number of observations across the K clusters is ignored

- The fact that observations are correlated is ignored

- All of these are key motivators for combining stages 1 and 2 into a single model formulation
  - ⋆ linear mixed effects models (Part III)

Weighted least squares

- Recall in Methods I, in the (standard) linear regression setting with independent data, we considered the class of *weighted least squares* estimators:

$$\widehat{\boldsymbol{\beta}}_{\text{WLS}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y}$$

  ⋆ $\boldsymbol{W}$ is an $N \times N$ matrix

- Can show that for any (non-trivial) $\boldsymbol{W}$:

$$\begin{aligned} \mathsf{E}[\widehat{\boldsymbol{\beta}}_{\text{WLS}}] &= \boldsymbol{\beta} \\ \mathsf{Cov}[\widehat{\boldsymbol{\beta}}_{\text{WLS}}] &= \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1} \end{aligned}$$

  where $\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ and $\boldsymbol{B} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T\boldsymbol{X}$

- ‘Robust’ in the sense that inference is valid regardless of the choice of $\boldsymbol{W}$

- We also noted that one can obtain efficiency gains by being wise (and confident!) when choosing $\boldsymbol{W}$

- Specifically, the *generalized least squares* estimator:

$$\widehat{\boldsymbol{\beta}}_{\mathsf{GLS}} \;=\; (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$$

  is the *best linear unbiased estimator* of $\boldsymbol{\beta}$, with

$$\mathsf{Cov}[\widehat{\boldsymbol{\beta}}_{\mathsf{GLS}}] \;=\; (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$$

  ⋆ obtained by setting $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$

  ⋆ optimality via the Gauss-Markov Theorem

- Operationally, one needs an estimate of $\boldsymbol{\Sigma} = \mathsf{diag}(\sigma_1^2,\, \ldots,\, \sigma_n^2)$

  ⋆ numerous options that make use of residuals from some fitted model

Q: Can we translate these ideas into the dependent data setting? **Yes!**

The WLS estimator

- Notationally, recall that
  - ⋆ $\boldsymbol{X}_k$ is a cluster-specific $n_k \times p$ matrix of covariates
  - ⋆ $\boldsymbol{Y}_k$ is a cluster-specific $n_k \times 1$ response vector
- Let $\boldsymbol{W}_k$ denote a cluster-specific symmetric $n_k \times n_k$ matrix of weights
- Consider estimating $\boldsymbol{\beta}$ via minimization of objective function:

$$\mathcal{Q}_{\boldsymbol{W}}(\boldsymbol{\beta}) = \sum_{k=1}^{K} (\boldsymbol{Y}_k - \boldsymbol{X}_k\boldsymbol{\beta})^T \boldsymbol{W}_k (\boldsymbol{Y}_k - \boldsymbol{X}_k\boldsymbol{\beta})$$

- Notice that the summation is over k
  - ⋆ form is motivated by the assumption that clusters are independent of each other

- It is relatively straightforward to show that the solution to

$$\frac{\partial}{\partial \boldsymbol{\beta}} \mathcal{Q}_{\boldsymbol{W}}(\boldsymbol{\beta}) \;=\; \sum_{k=1}^{K} \boldsymbol{U}_{\boldsymbol{W}}(\boldsymbol{\beta}; \boldsymbol{Y}_k) \;=\; \sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{W}_k (\boldsymbol{Y}_k - \boldsymbol{X}_k \boldsymbol{\beta}) \;=\; \mathbf{0}$$

  is

$$\widehat{\boldsymbol{\beta}}_{\mathsf{WLS}} \;=\; \left(\sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{W}_k \boldsymbol{X}_k \right)^{-1} \left(\sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{W}_k \boldsymbol{Y}_k \right).$$

- It is also relatively straightforward to show that $\mathsf{E}[\widehat{\boldsymbol{\beta}}_{\mathsf{WLS}}] \;=\; \boldsymbol{\beta}$, regardless of the choice of $\boldsymbol{W}$

- Performing a Taylor series expansion and appealing to the central limit theorem, we have that

$$\sqrt{K}(\widehat{\boldsymbol{\beta}}_{\mathsf{WLS}} - \boldsymbol{\beta}) \;\longrightarrow\; \mathsf{MVN}_p(\mathbf{0},\; \boldsymbol{C}_{\boldsymbol{W}})$$

  as $K \longrightarrow \infty$

- The asymptotic variance-covariance matrix is

$$\boldsymbol{C}_{\boldsymbol{W}} \;=\; \boldsymbol{F}_{\boldsymbol{W}}^{-1}\boldsymbol{I}_{\boldsymbol{W}}\boldsymbol{F}_{\boldsymbol{W}}^{-1}$$

where

$$\boldsymbol{F}_{\boldsymbol{W}} \;=\; \mathsf{E}\left[\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{U}_{\boldsymbol{W}}(\boldsymbol{\beta};\boldsymbol{Y})\right] \quad \text{and} \quad \boldsymbol{I}_{\boldsymbol{W}} \;=\; \mathsf{E}\left[\boldsymbol{U}_{\boldsymbol{W}}(\boldsymbol{\beta};\boldsymbol{Y})\boldsymbol{U}_{\boldsymbol{W}}(\boldsymbol{\beta};\boldsymbol{Y})^T\right]$$

- Given the structure of $\boldsymbol{U}_{\boldsymbol{W}}(\boldsymbol{\beta};\boldsymbol{Y}_k)$, these expectations have analytically tractable forms and one can show that

$$\mathsf{Cov}[\widehat{\boldsymbol{\beta}}_{\mathsf{WLS}}] \;=\; \boldsymbol{A}_{\boldsymbol{W}}^{-1}\boldsymbol{B}_{\boldsymbol{W}}\boldsymbol{A}_{\boldsymbol{W}}^{-1}$$

where

$$\boldsymbol{A}_{\boldsymbol{W}} \;=\; \sum_{k=1}^{K}\boldsymbol{X}_k^T\boldsymbol{W}_k\boldsymbol{X}_k \quad \text{and} \quad \boldsymbol{B}_{\boldsymbol{W}} \;=\; \sum_{k=1}^{K}\boldsymbol{X}_k^T\boldsymbol{W}_k\boldsymbol{\Sigma}_k\boldsymbol{W}_k\boldsymbol{X}_k$$

- In practice, since we don't know the 'true' $\boldsymbol{\Sigma}_k$, we need to estimate $\boldsymbol{B}_{\boldsymbol{W}}$

- One way to forward would be to empirically estimate it as the expectation of the square of the 'scores':

$$\begin{aligned}\widehat{\boldsymbol{B}}_{\boldsymbol{W}} &= \sum_{k=1}^{K} \boldsymbol{U}_{\boldsymbol{W}}(\widehat{\boldsymbol{\beta}};\boldsymbol{Y}_k)\boldsymbol{U}_{\boldsymbol{W}}(\widehat{\boldsymbol{\beta}};\boldsymbol{Y}_k)^T \\ &= \sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{W}_k(\boldsymbol{Y}_k - \boldsymbol{X}_k\widehat{\boldsymbol{\beta}})(\boldsymbol{Y}_k - \boldsymbol{X}_k\widehat{\boldsymbol{\beta}})^T \boldsymbol{W}_k \boldsymbol{X}_k,\end{aligned}$$

and base inference on

$$\widehat{\mathsf{Cov}}[\widehat{\boldsymbol{\beta}}_{\mathsf{WLS}}] = \boldsymbol{A}_{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{B}}_{\boldsymbol{W}}\boldsymbol{A}_{\boldsymbol{W}}^{-1}.$$

Example #1

- Suppose $\boldsymbol{X}_k = \boldsymbol{X}_0$ for all k
  - ⋆ balanced and complete data
  - ⋆ e.g. the dental growth curve data restricted to the 11 females
- Furthermore, suppose we take $\boldsymbol{W}_k = \boldsymbol{W}_0$ for all k
  - ⋆ each cluster is assigned the <u>same</u> weighting structure
- We then have that

$$\widehat{\boldsymbol{\beta}}_{\text{WLS}} = (\boldsymbol{X}_0^T \boldsymbol{W}_0 \boldsymbol{X}_0)^{-1} \boldsymbol{X}_0^T \boldsymbol{W}_0 \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{Y}_k$$

  - ⋆ can be viewed as the regression of the study unit-specific averages

Example #2

- For the special case in which we take $\boldsymbol{W}_k = \boldsymbol{I}_k \; \forall \; k$, we obtain the OLS estimator:

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} \;=\; \left(\sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{X}_k\right)^{-1} \left(\sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{Y}_k\right).$$

- It's interesting to note that $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ minimizes:

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\beta}) \;&=\; \sum_{k=1}^{K} (\boldsymbol{Y}_k \;-\; \boldsymbol{X}_k \boldsymbol{\beta})^T (\boldsymbol{Y}_k - \; \boldsymbol{X}_k \boldsymbol{\beta}) \\
&=\; \sum_{k=1}^{K} \sum_{i=1}^{n_k} (Y_{ki} \;-\; \boldsymbol{X}_{ki} \boldsymbol{\beta})^2
\end{aligned}$$

  ⋆ each of the $N = \sum_k n_k$ study units is assigned <u>equal</u> weight in the objective function

- The variance-covariance matrix for $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ is:

$$\text{Cov}[\widehat{\boldsymbol{\beta}}_{\text{OLS}}] \;=\; \left(\sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{X}_k\right)^{-1} \left(\sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{\Sigma}_k \boldsymbol{X}_k\right) \left(\sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{X}_k\right)^{-1}$$

which can be estimated by:

$$\widehat{\text{Cov}}[\widehat{\boldsymbol{\beta}}_{\text{OLS}}] \;=\; \boldsymbol{A}_{\boldsymbol{W}}^{-1} \widehat{\boldsymbol{B}}_{\boldsymbol{W}} \boldsymbol{A}_{\boldsymbol{W}}^{-1}$$

where

$$\boldsymbol{A}_{\boldsymbol{W}} \;=\; \sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{X}_k$$

$$\widehat{\boldsymbol{B}}_{\boldsymbol{W}} \;=\; \sum_{k=1}^{K} \boldsymbol{X}_k^T (\boldsymbol{Y}_k - \boldsymbol{X}_k \widehat{\boldsymbol{\beta}}_{\text{OLS}})(\boldsymbol{Y}_k - \boldsymbol{X}_k \widehat{\boldsymbol{\beta}}_{\text{OLS}})^T \boldsymbol{X}_k$$

Example #3

- By the Gauss-Markov Theorem, the most efficient WLS estimator of $\boldsymbol{\beta}$ is the one where one sets $\boldsymbol{W}_k = \boldsymbol{\Sigma}_k^{-1}$:

$$\widehat{\boldsymbol{\beta}}_{\text{GLS}} = \left(\sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{X}_k \right)^{-1} \left(\sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{Y}_k \right)$$

- For this estimator we have $\boldsymbol{A}_{\boldsymbol{W}} = \boldsymbol{B}_{\boldsymbol{W}}$ so that

$$\text{Cov}[\widehat{\boldsymbol{\beta}}_{\text{GLS}}] = \left(\sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{X}_k \right)^{-1}$$

  ⋆ We are going refer to this as the *generalized least squares* (GLS) estimator

- Note, the optimal weighting uses information about variation and co-variation within the cluster

- In practice, use of the GLS estimator requires solving the practical challenge of not knowing the true $\mathbf{\Sigma}_k$

- In principle, we could plug in the following:

$$\widehat{\mathbf{\Sigma}}_k \;=\; (\boldsymbol{Y}_k - \boldsymbol{X}_k\widehat{\boldsymbol{\beta}})(\boldsymbol{Y}_k - \boldsymbol{X}_k\widehat{\boldsymbol{\beta}})^T$$

  although this, in general, is going to be an unreliable estimate of Cov$[\boldsymbol{Y}_k]$

  ⋆ estimating a variance-covariance matrix on the basis of a single observation, the resulting matrix is not invertible.

- Forging ahead with this estimate can often result in highly variable weights and instability in the estimation process

- In practice, therefore, we typically place some structure on $\mathbf{\Sigma}_k$ prior to using it as an inverse-weight

  ⋆ structure across clusters

  ⋆ structure within clusters

- Before we consider how we might do this, however, it is important to note that we may no longer be using the optimal weighting strategy

- We are therefore faced with a trade-off in that choosing increasingly parsimonious structures for $\mathbf{\Sigma}_k$ will likely result in:
  - ⋆ increasingly stable estimation and weights that are not highly variable
  - ⋆ increasing losses in efficiency

Structuring $\mathbf{\Sigma}_k$

- Suppose we have balanced and complete data
  - ⋆ $n_k = n \; \forall \; k$
  - ⋆ e.g. a longitudinal study in which the timing of observations is the same $\forall \; k$ such as the dental growth data

- In this instance, it may be reasonable (for the purposes of weighting) to take

$$\mathbf{\Sigma}_k \;=\; \mathbf{\Sigma}_0 \quad \forall \; k$$

  - ⋆ interpret $\mathbf{\Sigma}_0$ as either a common variance-covariance structure or as some average of the cluster-specific $\mathbf{\Sigma}_k$, once covariates have in the mean model have been taken into account
  - ⋆ 'reasonable' in so far as it is a *working approximation* to the true $\mathbf{\Sigma}_k$

- Given a consistent estimate of $\boldsymbol{\beta}$, say $\widehat{\boldsymbol{\beta}}$, a consistent estimate of $\boldsymbol{\Sigma}_0$ is:

$$\widehat{\boldsymbol{\Sigma}}_0 \;=\; \frac{1}{K}\sum_{k=1}^{K}(\boldsymbol{Y}_k - \boldsymbol{X}_k\widehat{\boldsymbol{\beta}})(\boldsymbol{Y}_k - \boldsymbol{X}_k\widehat{\boldsymbol{\beta}})^T$$

  - ⋆ average of the empirical covariance of the cluster-specific residuals
  - ⋆ the clusters are the ‘independent’ replications

- One can then use this estimate to inform a new weighting scheme to give:

$$\widehat{\boldsymbol{\beta}}_{\text{WLS}} \;=\; \left(\sum_{k=1}^{K}\boldsymbol{X}_k^T\widehat{\boldsymbol{\Sigma}}_0^{-1}\boldsymbol{X}_k\right)^{-1}\left(\sum_{k=1}^{K}\boldsymbol{X}_k^T\widehat{\boldsymbol{\Sigma}}_0^{-1}\boldsymbol{Y}_k\right)$$

- Note, if it truly is the case that $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_0 \;\forall\; k$ then this estimator will be asymptotically equivalent to $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$
  - ⋆ in this case, inference would be based on

$$\widehat{\text{Cov}}[\widehat{\boldsymbol{\beta}}_{\text{WLS}}] \;=\; \left(\sum_{k=1}^{K}\boldsymbol{X}_k^T\widehat{\boldsymbol{\Sigma}}_0^{-1}\boldsymbol{X}_k\right)^{-1}$$

- If it is not the case, however, that $\mathbf{\Sigma}_k = \mathbf{\Sigma}_0 \; \forall \; k$, then $\widehat{\boldsymbol{\beta}}_{\text{WLS}} \neq \widehat{\boldsymbol{\beta}}_{\text{GLS}}$
    - ⋆ hence, it will not be optimal (in the Guass-Markov sense)
    - ⋆ if $\mathbf{\Sigma}_0$ is 'reasonable', however, then we might expect it to be more efficient than $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$

- Either way, inference would be based on

$$\widehat{\text{Cov}}[\widehat{\boldsymbol{\beta}}_{\text{WLS}}] \;=\; \boldsymbol{A}_{\boldsymbol{W}}^{-1} \widehat{\boldsymbol{B}}_{\boldsymbol{W}} \boldsymbol{A}_{\boldsymbol{W}}^{-1}$$

where

$$\boldsymbol{A}_{\boldsymbol{W}} \;=\; \sum_{k=1}^{K} \boldsymbol{X}_k^T \widehat{\mathbf{\Sigma}}_0^{-1} \boldsymbol{X}_k$$

and

$$\widehat{\boldsymbol{B}}_{\boldsymbol{W}} \;=\; \sum_{k=1}^{K} \boldsymbol{X}_k^T \widehat{\mathbf{\Sigma}}_0^{-1} (\boldsymbol{Y}_k - \boldsymbol{X}_k \widehat{\boldsymbol{\beta}}_{\text{WLS}})(\boldsymbol{Y}_k - \boldsymbol{X}_k \widehat{\boldsymbol{\beta}}_{\text{WLS}})^T \widehat{\mathbf{\Sigma}}_0^{-1} \boldsymbol{X}_k$$

- Note, inference based on this estimate of $\text{Cov}[\widehat{\boldsymbol{\beta}}_{\text{WLS}}]$ will be *robust* in the sense that it will be valid (in large samples) regardless of whether $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_0$ $\forall\ k$
  - ⋆ $\widehat{\boldsymbol{\Sigma}}_0^{-1}$ is simply one choice of weighting scheme
  - ⋆ as long as $\widehat{\boldsymbol{B}}_{\boldsymbol{W}} \rightarrow \boldsymbol{B}_{\boldsymbol{W}}$, inference will be valid

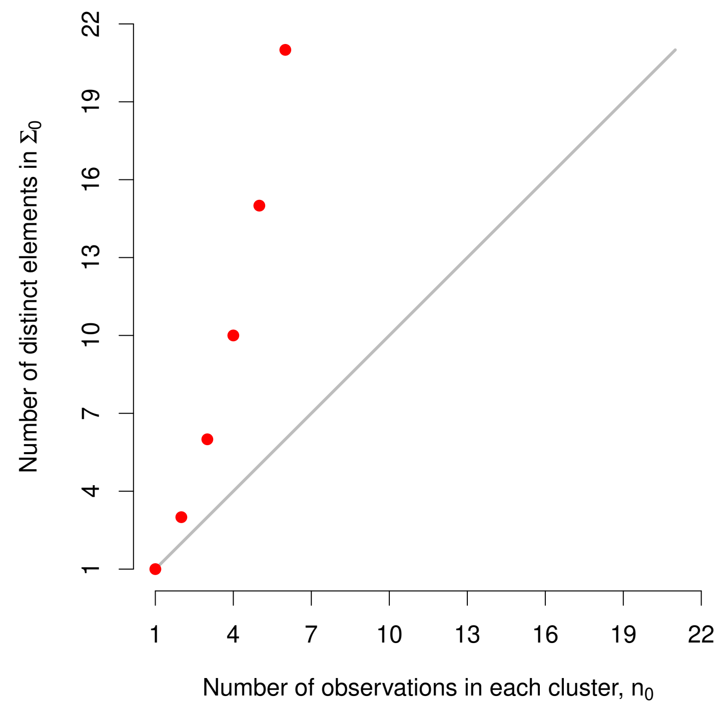
Modeling the within-cluster dependence

- Even if we are willing to take $\mathbf{\Sigma}_k = \mathbf{\Sigma}_0 \; \forall \; k$ in the weighting scheme, we should be aware that consistency of $\widehat{\mathbf{\Sigma}}_0$ hinges on K

- Interestingly, as n gets large the number of distinct parameters in $\mathbf{\Sigma}_0$ increases in a non-linear fashion:
  - ⋆ for fixed n, the number of distinct parameters is:

$$n + n(n-1)/2$$

  - ⋆ implies that the extent to which K is 'large enough' for valid inference depends, in part, on n



- For small to moderate K, we may wish to adopt some simplifying structure for Σ_0
  - i.e. structure the internal elements of $\boldsymbol{\Sigma}_0$
  - as a function of some small('ish) number of parameters, $\boldsymbol{\alpha}$

$$\{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\} \quad \Rightarrow \quad \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_0 \ \forall\ k \quad \Rightarrow \quad \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_0(\boldsymbol{\alpha}) \ \forall\ k$$

- There is substantial scope for flexibility in how we specify dependence between study units within a cluster
  - several examples on slides 87-89

- In practice, we can use substantive knowledge about $\boldsymbol{Y}_k$ and exploratory data analysis to guide the selection of a simpler covariance model
  - see Part I of the notes

- For example, we might believe that a reasonable model for dependence is the exchangeable or compound symmetric covariance model:

$$\boldsymbol{\Sigma}_0(\boldsymbol{\alpha}) \;=\; \begin{bmatrix} \alpha_1 & \alpha_2 & \ldots & \alpha_2 \\ \alpha_2 & \alpha_1 & \ldots & \alpha_2 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_2 & \alpha_2 & \ldots & \alpha_1 \end{bmatrix}$$

- Use of this structure as a weighting scheme corresponds to the working assumptions that:
  - ⋆ $\mathsf{V}[Y_{ki}] = \alpha_1 \;\forall\; i$
  - ⋆ $\mathsf{Cov}[Y_{ki}, Y_{kj}] = \alpha_2 \;\forall\; i \neq j$

Q: Can we think of settings where this might be ‘reasonable’?

- Given an initial estimate $\widehat{\boldsymbol{\beta}}$, simple moment-based estimators of α_1 and α_2 are:

$$\hat{\alpha}_1 = \frac{1}{K}\sum_{k=1}^{K}\left\{\frac{1}{n}\sum_{i=1}^{n}(Y_{ki} - \boldsymbol{X}_{ki}\widehat{\boldsymbol{\beta}})^2\right\}$$

$$\hat{\alpha}_2 = \frac{1}{K}\sum_{k=1}^{K}\left\{\frac{1}{n(n-1)}\sum_{i\neq j}(Y_{ki} - \boldsymbol{X}_{ki}\widehat{\boldsymbol{\beta}})(Y_{kj} - \boldsymbol{X}_{kj}\widehat{\boldsymbol{\beta}})\right\}$$

- We can then obtain a new WLS estimator as:

$$\widehat{\boldsymbol{\beta}}_{\mathsf{WLS}} = \left(\sum_{k=1}^{K}\boldsymbol{X}_k^T\boldsymbol{\Sigma}_0(\widehat{\boldsymbol{\alpha}})^{-1}\boldsymbol{X}_k\right)^{-1}\left(\sum_{k=1}^{K}\boldsymbol{X}_k^T\boldsymbol{\Sigma}_0(\widehat{\boldsymbol{\alpha}})^{-1}\boldsymbol{Y}_k\right).$$

- Inference for this estimator can be based on:

$$\widehat{\text{Cov}}[\widehat{\boldsymbol{\beta}}_{\text{WLS}}] \;=\; \boldsymbol{A}(\widehat{\boldsymbol{\alpha}})^{-1}\widehat{\boldsymbol{B}}(\widehat{\boldsymbol{\alpha}})\boldsymbol{A}(\widehat{\boldsymbol{\alpha}})^{-1}$$

where

$$\boldsymbol{A}(\widehat{\boldsymbol{\alpha}}) \;=\; \sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{\Sigma}_0(\widehat{\boldsymbol{\alpha}})^{-1}\boldsymbol{X}_k$$

and

$$\widehat{\boldsymbol{B}}(\widehat{\boldsymbol{\alpha}}) \;=\; \sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{\Sigma}_0(\widehat{\boldsymbol{\alpha}})^{-1}(\boldsymbol{Y}_k - \boldsymbol{X}_k\widehat{\boldsymbol{\beta}}_{\text{WLS}})(\boldsymbol{Y}_k - \boldsymbol{X}_k\widehat{\boldsymbol{\beta}}_{\text{WLS}})^T\boldsymbol{\Sigma}_0(\widehat{\boldsymbol{\alpha}})^{-1}\boldsymbol{X}_k$$

- Note, inference based on this estimate of Cov$[\widehat{\boldsymbol{\beta}}_{\text{WLS}}]$ will be *robust* in the sense that it will be valid (in large samples) regardless of whether $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_0(\boldsymbol{\alpha}) \;\forall\; k$
  - ⋆ $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\alpha}})_0^{-1}$ is simply one choice of weighting scheme
  - ⋆ as long as $\widehat{\boldsymbol{B}}_{\boldsymbol{w}} \rightarrow \boldsymbol{B}_{\boldsymbol{w}}$, inference will be valid

Simulation

- We can investigate the interplay between K and n with a simulation study

- Generate outcomes according to the mean model:

$$\mathsf{E}[Y_{ki}|\ X_{1,ki}, X_{2,ki}] \ = \ \beta_0 \ + \ \beta_1 X_{1,ki} \ + \ \beta_2 X_{2,ki}$$

  - ⋆ $X_{1,ki} \in \{1, \ldots, n\}$ is a study unit specific ‘time’ variable
  - ⋆ $X_{2,ki}$ is a cluster-specific binary variable such that half of the clusters have $X_{2,ki}$=0 and the other half $X_{2,ki}$=1
  - ⋆ $\boldsymbol{\beta}$ = (0, 1, 1)

- Compound symmetric dependence structure $\forall\ k$, with $\mathsf{V}[Y_{ki}]$=1 and ρ=0.5

- Sample sizes:
  - ⋆ K = 30, 60
  - ⋆ n = 4, 6

- Consider three WLS estimators that differ in the 'working' correlation structure that informs the weighting scheme:

(1) *working independence:* $\text{Cor}[\boldsymbol{Y}_k] = \boldsymbol{\Sigma}_0 = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$

(2) *working exchangeable:* $\text{Cor}[\boldsymbol{Y}_k] = \boldsymbol{\Sigma}_0 = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix}$

(3) *working unstructured:* $\text{Cor}[\boldsymbol{Y}_k] = \boldsymbol{\Sigma}_0 = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{12} & 1 & \dots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1n} & \rho_{2n} & \dots & 1 \end{bmatrix}$

```
> ##
> library(mvtnorm)
>
> ##
> genData <- function(K, nT, qX2, betaV, sigSq, tauSq)
+ {
+   ##
+   X1.ki  <- rep(1:nT, K)
+   X2.ki  <- rep(rep(c(0,1), c(K-round(K*qX2), round(K*qX2))), rep(nT, K))
+   eta.ki <- matrix(cbind(1, X1.ki, X2.ki) %*% betaV,
+                    nrow=K, ncol=nT, byrow=TRUE)
+   ##
+   Sigma0 <- matrix(tauSq, nrow=nT, ncol=nT)
+   diag(Sigma0) <- tauSq + sigSq
+   ##
+   Y.ki <- matrix(NA, nrow=K, ncol=nT)
+   for(k in 1:K) Y.ki[k,] <- rmvnorm(1, mean=eta.ki[k,], sigma=Sigma0)
+   ##
+   return(data.frame(id=rep(1:K, rep(nT, K)),
+                     X1=X1.ki,
+                     X2=X2.ki,
+                     Y=c(t(Y.ki))))
+ }
```

```
> ##
> qX2   <- 0.5
> betaV <- c(0, 1, 1)
> sigSq <- 0.5
> tauSq <- 0.5
>
> ##
> simData <- genData(K=30, nT=4, qX2, betaV, sigSq, tauSq)
>
> ##
> library(gee)
>
> ##
> fit.WI <- gee(Y ~ X1 + X2, data=simData, id=id, corstr="independence")
> fit.WE <- gee(Y ~ X1 + X2, data=simData, id=id, corstr="exchangeable")
> fit.WU <- gee(Y ~ X1 + X2, data=simData, id=id, corstr="unstructured")
```

- Simulated R=10,000 datasets for each (K, n) combination

- For each estimator consider bias associated with $\widehat{\boldsymbol{\beta}}_{\text{WLS}}$ as an estimate of $\boldsymbol{\beta}$

- Also consider the performance of two estimators of $\text{Cov}[\widehat{\boldsymbol{\beta}}_{\text{WLS}}]$:

(1) naïve estimator:

$$\widehat{\text{Cov}}_n[\widehat{\boldsymbol{\beta}}_{\text{WLS}}] \;=\; \boldsymbol{A}_{\boldsymbol{W}}^{-1}$$

(2) robust estimator:

$$\widehat{\text{Cov}}[\widehat{\boldsymbol{\beta}}_{\text{WLS}}] \;=\; \boldsymbol{A}_{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{B}}_{\boldsymbol{W}}\boldsymbol{A}_{\boldsymbol{W}}^{-1}$$

$\star$ for both estimators report the ratio of the mean of the estimated standard errors to the standard deviation of the point estimates $\times$ 100

- Instances out of R=10,000 that `gee()` converged:

	$n = 4$	$n = 6$
$K = 30$		
Working independence	10,000	10,000
Working exchangeable	10,000	10,000
Working unstructured	9,864	8,869
$K = 60$		
Working independence	10,000	10,000
Working exchangeable	10,000	10,000
Working unstructured	10,000	9,962

⋆ when K=30 and n=6, a little over 10% of the working unstructured estimators fail to converge

- Mean of the point estimates:

	$n = 4$			$n = 6$		
	β_0	β_1	β_2	β_0	β_1	β_2
$K = 30$						
Working independence	0	1	1	0	1	1
Working exchangeable	0	1	1	0	1	1
Working unstructured	0	1	1	0	1	1
$K = 60$						
Working independence	0	1	1	0	1	1
Working exchangeable	0	1	1	0	1	1
Working unstructured	0	1	1	0	1	1

$\star$ no bias across the board

- Ratio for the naïve standard errors:

	$n = 4$			$n = 6$		
	β_0	β_1	β_2	β_0	β_1	β_2
$K = 30$						
Working independence	95	140	62	82	139	52
Working exchangeable	98	101	97	98	100	96
Working unstructured	92	88	93	85	76	84
$K = 60$						
Working independence	97	140	64	83	141	53
Working exchangeable	100	100	100	100	101	98
Working unstructured	98	95	98	94	89	93

⋆ working independence does very poorly

⋆ working exchangeable does well since it is the 'correct' structure

⋆ unstructured performs quite poorly when K=30

- Ratio for the robust standard errors:

	$n = 4$			$n = 6$		
	β_0	β_1	β_2	β_0	β_1	β_2
$K = 30$						
Independence	96	98	96	96	97	96
Exchangeable	96	98	96	96	97	96
Unstructured	92	92	90	87	87	85
$K = 60$						
Independence	99	99	99	99	99	98
Exchangeable	99	99	99	99	99	98
Unstructured	97	96	96	93	93	92

⋆ working independence and exchangeable seem to be equivalent

  ∗ can show this analytically in this setting

⋆ for fixed K, unstructured gets worse as n increases

Hypothesis testing

- Using the fact that the asymptotic sampling distribution of $\widehat{\boldsymbol{\beta}}_{\text{WLS}}$ is a multivariate Normal, one can perform hypothesis testing using the usual Wald test

- Specifically, consider testing the linear null hypotheses:

$$\mathrm{H}_0 : \boldsymbol{Q}\boldsymbol{\beta} \ = \ \mathbf{0}$$

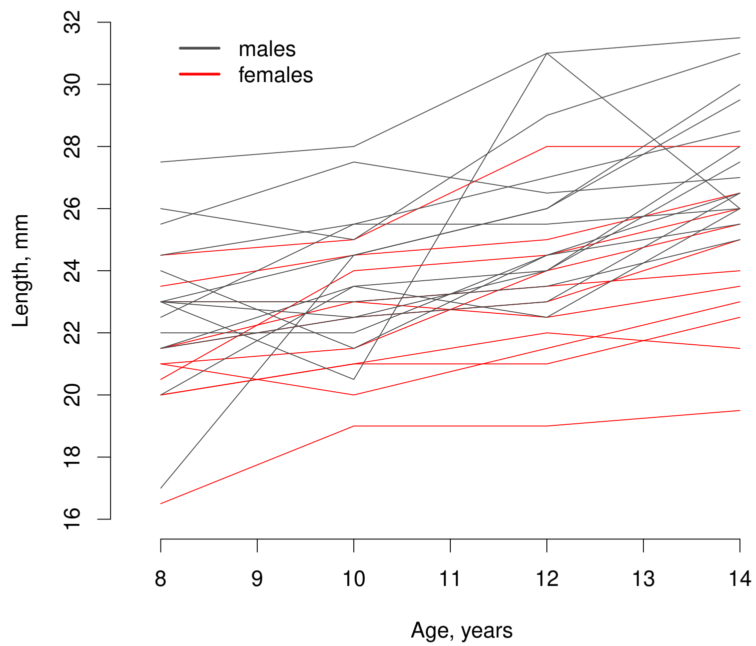
where $\boldsymbol{Q}$ is a matrix of full rank with $\dim(\boldsymbol{Q}) = r \times p$ with $r < p$

- Evaluate evidence regarding the null on the basis of the multivariate Wald statistic:

$$(\boldsymbol{Q}\widehat{\boldsymbol{\beta}}_{\text{WLS}})^T(\boldsymbol{Q}\widehat{\text{Cov}}[\widehat{\boldsymbol{\beta}}_{\text{WLS}}]\boldsymbol{Q}^T)^{-1}(\boldsymbol{Q}\widehat{\boldsymbol{\beta}}_{\text{WLS}}) \ \sim \ \chi^2_r$$

- Note, the likelihood ratio test is not available

Dental growth data



- Formally characterize dental growth among males and females aged 8 to 14 years using the model:

$$\mathsf{E}[Y_{ki}] \;=\; \beta_0 \;+\; \beta_1 A^*_{ki} \;+\; \beta_1 G_k \;+\; \beta_3 A^*_{ki} G_k$$

⋆ use $A^*_{ki} = A_{ki} - 8$ to ensure that β_0 is interpretable

- We can fit this model using the gee() function in R:

```
> ##
> growth$ageStar <- growth$age - 8
>
> ##
> library(gee)
>
> ## Weighted least squares
> ##
> fit0.GEE <- gee(length ~ ageStar * gender, id=id, data=growth,
                  corstr="independence")
Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate
       (Intercept)            ageStar         gendermale ageStar:gendermale
        21.2090909          0.4795455          1.4909091          0.3204545
>
> fit1.GEE <- gee(length ~ ageStar * gender, id=id, data=growth,
                  corstr="exchangeable")
...
> fit2.GEE <- gee(length ~ ageStar * gender, id=id, data=growth,
                  corstr="unstructured")
...
```

```
> ##
> summary(fit0.GEE)
...
Model:
 Link:                      Identity
 Variance to Mean Relation: Gaussian
 Correlation Structure:     Independent
...
Coefficients:
                     Estimate Naive S.E.   Naive z Robust S.E.  Robust z
(Intercept)        21.2090909  0.5700227 37.207453   0.5604314 37.844221
ageStar             0.4795455  0.1523450  3.147760   0.0631326  7.595845
gendermale          1.4909091  0.7504697  1.986635   0.7939739  1.877781
ageStar:gendermale  0.3204545  0.2005715  1.597708   0.1213679  2.640356

Estimated Scale Parameter:  5.105977
...
Working Correlation
     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1
```

```
> ##
> summary(fit1.GEE)
...
Model:
 Link:                      Identity
 Variance to Mean Relation: Gaussian
 Correlation Structure:     Exchangeable
...
Coefficients:
                     Estimate Naive S.E.   Naive z Robust S.E.  Robust z
(Intercept)        21.2090909 0.63937427 33.171637   0.5604314 37.844221
ageStar             0.4795455 0.09607315  4.991462   0.0631326  7.595845
gendermale          1.4909091 0.84177534  1.771148   0.7939739  1.877781
ageStar:gendermale  0.3204545 0.12648618  2.533514   0.1213679  2.640356

Estimated Scale Parameter:  5.105977
...
Working Correlation
          [,1]      [,2]      [,3]      [,4]
[1,] 1.0000000 0.6023071 0.6023071 0.6023071
[2,] 0.6023071 1.0000000 0.6023071 0.6023071
[3,] 0.6023071 0.6023071 1.0000000 0.6023071
[4,] 0.6023071 0.6023071 0.6023071 1.0000000
```

```
> ##
> summary(fit2.GEE)
...
Model:
 Link:                      Identity
 Variance to Mean Relation: Gaussian
 Correlation Structure:     Unstructured
...
Coefficients:
                     Estimate Naive S.E.   Naive z Robust S.E.  Robust z
(Intercept)        21.2212826  0.6234615 34.037840  0.55506777 38.231877
ageStar             0.4784452  0.1026902  4.659112  0.06475658  7.388365
gendermale          1.5043743  0.8208252  1.832758  0.78469921  1.917135
ageStar:gendermale  0.3167658  0.1351980  2.342978  0.12435591  2.547252

Estimated Scale Parameter:  5.10616
...
Working Correlation
          [,1]      [,2]      [,3]      [,4]
[1,] 1.0000000 0.5064509 0.7487428 0.5160647
[2,] 0.5064509 1.0000000 0.5318310 0.5963414
[3,] 0.7487428 0.5318310 1.0000000 0.7625703
[4,] 0.5160647 0.5963414 0.7625703 1.0000000
```

Comments

- Theory sketched out here may be considered *semi-parametric* in the sense that estimation/inference regarding $\boldsymbol{\beta}$ relies solely on specification of a model for the mean of $\boldsymbol{Y}_k$ and (possibly) the variance-covariance matrix
  - ⋆ will form the basis for *generalized estimating equations*

- Hypothesis testing regarding $\boldsymbol{\beta}$ is straightforward but investigating questions regarding the variance-covariance matrix is not
  - ⋆ $\boldsymbol{\Sigma}_k$ is viewed, primarily, as a nuisance rather than a quantity of intrinsic interest
  - ⋆ no clear means of obtaining estimates of uncertainty regarding the components of $\boldsymbol{\Sigma}_k$
  - ⋆ motivates the use of likelihood-based methods

- While efficiency considerations motivated careful consideration of the dependence model, they also motivate the use of likelihood-based methods
  - ⋆ parametric modeling of the <u>entire</u> distribution of $\boldsymbol{Y}_k$
  - ⋆ especially useful in small-sample settings

- In estimating the asymptotic covariance matrix, we exploited 'independent' replication across clusters
  - ⋆ may become problematic in some missing data settings
  - ⋆ another motivation for likelihood-base inference

Maximum likelihood

- Suppose we assume that

$$\boldsymbol{Y}_k \;\sim\; \mathsf{MVN}_n(\boldsymbol{X}_k\boldsymbol{\beta},\; \sigma^2\boldsymbol{V}_0)$$

  - ⋆ assume $n_k = n$ and $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_0 \;\forall\; k$
  - ⋆ decompose $\boldsymbol{\Sigma}_0 = \sigma^2\boldsymbol{V}_0$
    - ∗ σ^2 is a common variance component
    - ∗ $\boldsymbol{V}_0$ is a common correlation matrix

- The log-likelihood, as a function of the unknown parameters, is proportional to:

$$\ell(\boldsymbol{\beta},\sigma^2,\boldsymbol{V}_0) \;\propto\; -\,Kn\log(\sigma^2) \;-\; K\log|\boldsymbol{V}_0| \\ \qquad -\;\frac{1}{\sigma^2}\sum_{k=1}^{K}(\boldsymbol{Y}_k - \boldsymbol{X}_k\boldsymbol{\beta})^T\boldsymbol{V}_0^{-1}(\boldsymbol{Y}_k - \boldsymbol{X}_k\boldsymbol{\beta})$$

- For a given $\boldsymbol{V}_0$, the MLE for $\boldsymbol{\beta}$ is the WLS estimator:

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{V}_0) = \left(\sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{V}_0^{-1} \boldsymbol{X}_k\right)^{-1} \left(\sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{V}_0^{-1} \boldsymbol{Y}_k\right)$$

- Substitution of this estimator into $\ell(\boldsymbol{\beta}, \sigma^2, \boldsymbol{V}_0)$ yields the log-profile likelihood:

$$\ell(\widehat{\boldsymbol{\beta}}(\boldsymbol{V}_0), \sigma^2, \boldsymbol{V}_0) \propto - Kn \log(\sigma^2) - K \log |\boldsymbol{V}_0| - \frac{\mathsf{RSS}(\boldsymbol{V}_0)}{\sigma^2}$$

where

$$\mathsf{RSS}(\boldsymbol{V}_0) = \sum_{k=1}^{K} (\boldsymbol{Y}_k - \boldsymbol{X}_k \widehat{\boldsymbol{\beta}}(\boldsymbol{V}_0))^T \boldsymbol{V}_0^{-1} (\boldsymbol{Y}_k - \boldsymbol{X}_k \widehat{\boldsymbol{\beta}}(\boldsymbol{V}_0))$$

- Again for a given $\boldsymbol{V}_0$, differentiating with respect to σ^2 yields the MLE:

$$\hat{\sigma}^2(\boldsymbol{V}_0) = \frac{\mathsf{RSS}(\boldsymbol{V}_0)}{Kn}$$

- Finally, substitution of $\widehat{\boldsymbol{\beta}}(\boldsymbol{V}_0)$ and $\hat{\sigma}^2(\boldsymbol{V}_0)$ into $\ell(\boldsymbol{\beta}, \sigma^2, \boldsymbol{V}_0)$ yields the reduced log-profile likelihood corresponding to $\boldsymbol{V}_0$:

$$\ell_r(\boldsymbol{V}_0) \ \propto \ - Kn \log \mathsf{RSS}(\boldsymbol{V}_0) \ - \ K \log |\boldsymbol{V}_0|$$

- Maximization of this function yields the MLE, $\widehat{\boldsymbol{V}}_0$
  - ⋆ obtaining $\widehat{\boldsymbol{V}}_0$ will generally require numerical optimization routines
  - ⋆ dimensionality of the optimization problem is $n(n-1)/2$

- Finally, $\widehat{\boldsymbol{V}}_0$ can then be substituted into the expressions for $\widehat{\boldsymbol{\beta}}(\boldsymbol{V}_0)$ and $\hat{\sigma}^2(\boldsymbol{V}_0)$ to give the corresponding MLEs

- Note, inference for $\boldsymbol{\beta}$ could be based on:
  - ⋆ a Wald test
  - ⋆ a score test
  - ⋆ a likelihood ratio test

Issues

- The MLEs for σ^2 and $\boldsymbol{V}_0$ generally exhibit small-sample bias

- In the independent data setting, for example, it is well known that the MLE

$$\hat{\sigma}^2 = \frac{\text{RSS}}{N},$$

  where RSS is the residual sum of squares based on $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ exhibits small-sample bias but that:

$$\tilde{\sigma}^2 = \frac{\text{RSS}}{N - p}$$

  is unbiased

- In general, bias in the estimates of σ^2 and $\boldsymbol{V}_0$ has implications for likelihood-based inference based on the inverse-information matrix

- It turns out that $\tilde{\sigma}^2$ can be obtained via *restricted or residual maximum likelihood*

Restricted maximum likelihood

- Consider the general linear model for dependent data:

$$\boldsymbol{Y} \sim \text{MVN}_N(\boldsymbol{X\beta}, \boldsymbol{\Sigma})$$

- Suppose that $\boldsymbol{\Sigma}$ can be represented as a function of $\boldsymbol{\alpha}$, a set of (unknown) variance-covariance parameters

- The REML estimator of $\boldsymbol{\alpha}$ is defined as the maximum likelihood estimator based on a transformed outcome $\boldsymbol{Y}^* = A\boldsymbol{Y}$ such that the distribution of $\boldsymbol{Y}^*$ does not depend on $\boldsymbol{\beta}$
  - ⋆ note, the matrix A is a linear operator

- Put another way, REML considers a linear transformation of the response such that the resulting distribution (or part of it) depends solely on $\boldsymbol{\alpha}$
  - ⋆ estimation of $\boldsymbol{\beta}$ does not impact estimation of $\boldsymbol{\Sigma}$

Operationalization

- Consider the $N \times N$ matrix:

$$A = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

  - ⋆ $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is the so-called *hat matrix*

- For this choice of A, $\boldsymbol{Y}^* = A\boldsymbol{Y}$ is the vector of OLS residuals
  - ⋆ singular multivariate Normal distribution
  - ⋆ mean zero regardless of the value of $\boldsymbol{\beta}$

- To obtain a non-singular distribution, on which we can base estimation for $\boldsymbol{\Sigma}$, we could use only $N - p$ rows of A
  - ⋆ intuitively, remove a certain number of rows to 'account' for the fact that p regression parameters were estimated
  - ⋆ it turns out that it doesn't matter which rows we use

- Strategy we are going to use is to consider the transformation

$$\boldsymbol{Y} \;\Rightarrow\; (\boldsymbol{Z}, \widehat{\boldsymbol{\beta}})$$

where $\boldsymbol{Z} = B^T\boldsymbol{Y}$ with B the $N \times (N-p)$ matrix defined by the requirements that:

$$\begin{aligned} BB^T &= A \\ B^TB &= \boldsymbol{I} \end{aligned}$$

and $\widehat{\boldsymbol{\beta}}$ is the MLE for $\boldsymbol{\beta}$ for fixed $\boldsymbol{\alpha}$:

$$\widehat{\boldsymbol{\beta}} \;=\; (\boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$$

⋆ notice that the transformation is a linear one

- Standard results for the distribution of a transformation imply that:

$$f(\boldsymbol{Z}, \widehat{\boldsymbol{\beta}}) = f_{\boldsymbol{Y}}(g_1(\boldsymbol{Z}, \widehat{\boldsymbol{\beta}}), g_2(\boldsymbol{Z}, \widehat{\boldsymbol{\beta}}))|J|$$

  - ⋆ $g_1(\cdot)$ and $g_2(\cdot)$ are the inverse transformation functions
  - ⋆ J is Jacobian of the transformation

- It turns out, however, that

$$\begin{aligned} \mathsf{E}[\boldsymbol{Z}] &= \mathbf{0} \\ \mathsf{Cov}[\boldsymbol{Z}, \widehat{\boldsymbol{\beta}}] &= \mathbf{0} \end{aligned}$$

  regardless of the true value of $\boldsymbol{\beta}$

- Since zero covariance in the multivariate Normal setting is equivalent to independence, it follows that

$$f(\boldsymbol{Z}, \widehat{\boldsymbol{\beta}}) = f(\boldsymbol{Z})f(\widehat{\boldsymbol{\beta}}) = f_{\boldsymbol{Y}}(g_1(\boldsymbol{Z}, \widehat{\boldsymbol{\beta}}), g_2(\boldsymbol{Z}, \widehat{\boldsymbol{\beta}}))|J|$$

- Since

$$
\begin{aligned}
f_{\boldsymbol{Y}}(\boldsymbol{Y}) &= (2\pi)^{-N/2}|\boldsymbol{\Sigma}|^{-1/2} \\
&\quad \times \exp\left\{-\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\right\}
\end{aligned}
$$

and

$$
\begin{aligned}
f(\widehat{\boldsymbol{\beta}}) &= (2\pi)^{-p/2}|\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}|^{1/2} \\
&\quad \times \exp\left\{-\frac{1}{2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right\}
\end{aligned}
$$

and that the Jacobian doesn't depend on $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$, we have that the pdf of $\boldsymbol{Z}$, expressed as a function of $\boldsymbol{Y}$, is proportional to:

$$
\begin{aligned}
\frac{f_{\boldsymbol{Y}}(\boldsymbol{Y})}{f(\widehat{\boldsymbol{\beta}})} &= (2\pi)^{-(N-p)/2}|\boldsymbol{\Sigma}|^{-1/2}|\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}|^{-1/2} \\
&\quad \times \exp\left\{-\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})\right\}
\end{aligned}
$$

- The REML estimator of $\boldsymbol{\Sigma}$, therefore, maximizes the so-called *restricted log-likelihood*:

$$\ell^*(\boldsymbol{\alpha}) \ \propto \ - \ \log|\boldsymbol{\Sigma}| \ - \ \log|\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}| \ - \ (\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})$$

- Contrast this with the log-profile likelihood for the MLE:

$$\ell(\boldsymbol{\alpha}) \ \propto \ - \ \log|\boldsymbol{\Sigma}| \ - \ (\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})$$

- Returning to the notation where $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_0 = \sigma^2\boldsymbol{V}_0 \ \forall \ k$, recall that the MLE of σ^2 for a given $\boldsymbol{V}_0$ is:

$$\hat{\sigma}^2(\boldsymbol{V}_0) \ = \ \frac{\mathsf{RSS}(\boldsymbol{V}_0)}{Kn}$$

- In contrast, differentiating $\ell^*(\sigma^2, \boldsymbol{V}_0)$ with respect to σ^2 yields the REML estimator:

$$\tilde{\sigma}^2 \ = \ \frac{\mathsf{RSS}(\boldsymbol{V}_0)}{Kn-p}$$

- Substitution of $\tilde{\sigma}^2(\boldsymbol{V}_0)$ into $\ell^*(\sigma^2, \boldsymbol{V}_0)$ yields the reduced log-likelihood corresponding to $\boldsymbol{V}_0$:

$$\ell_r^*(\boldsymbol{V}_0) \;\propto\; -\, Kn \log \mathsf{RSS}(\boldsymbol{V}_0) \;-\; K \log |\boldsymbol{V}_0| \;-\; \log |\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}|$$

  - ⋆ $\boldsymbol{V}$ is a block-diagonal $N \times N$ matrix with common non-zero blocks $\boldsymbol{V}_0$

- Maximization of this function yields the MLE, $\widetilde{\boldsymbol{V}}_0$
  - ⋆ only a simple modification from the maximum likelihood algorithm

- This can also be contrasted with the ML analogue:

$$\ell_r(\boldsymbol{V}_0) \;\propto\; -\, Kn \log \mathsf{RSS}(\boldsymbol{V}_0) \;-\; K \log |\boldsymbol{V}_0|$$

- One can then use $\widetilde{\boldsymbol{V}}_0$ to obtain the REML estimator of σ^2:

$$\tilde{\sigma}^2 \;=\; \frac{\mathsf{RSS}(\widetilde{\boldsymbol{V}}_0)}{Kn - p}$$

- Finally, $\widetilde{\boldsymbol{V}}_0$ and $\tilde{\sigma}^2$ can be combined to form an estimate of $\boldsymbol{\Sigma}$ which can then be used to inform the weighting scheme for the estimation of $\boldsymbol{\beta}$:

$$\widetilde{\boldsymbol{\beta}}_{\mathsf{WLS}} \;=\; (\boldsymbol{X}\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Y}$$

- Inference for this estimator could then be based on:

$$\widehat{\mathsf{Cov}}[\widetilde{\boldsymbol{\beta}}_{\mathsf{WLS}}] \;=\; \boldsymbol{X}\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}$$

- Note that, by Slutsky's Theorem, the fact that we have estimated $\boldsymbol{\Sigma}$ does not impact the asymptotic distribution
  - ⋆ since $\widetilde{\boldsymbol{\Sigma}} \;\longrightarrow\; \boldsymbol{\Sigma}$

$$\sqrt{K}(\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{\Sigma}}) - \boldsymbol{\beta}) \;\equiv_d\; \sqrt{K}(\widehat{\boldsymbol{\beta}}(\boldsymbol{\Sigma}) - \boldsymbol{\beta})$$

Comments

- One cannot perform likelihood-based testing for $\boldsymbol{\beta}$ when you use REML
  - ⋆ key additional term in the restricted likelihood, $\log|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}|$, depends on the design matrix $\boldsymbol{X}$
  - ⋆ consequently, the additional term does not cancel out when you form the usual likelihood ratio test statistic and the asymptotic sampling distribution isn't necessarily a χ^2

- One can, however, still perform a Wald test or compare models using other measures such as the Akaike Information Criterion (AIC)

- Interestingly, REML is the default for many of the implementations in `R`

- Asymptotically, use of the MLE or the REMLE for $\boldsymbol{\Sigma}$ is equivalent so the distinction is only important in 'small samples'
  - ⋆ since the order of $\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}$ is p, the distinction rests on the relative size of p and $N = Kn$

Dental growth data

- Consider again the model:

$$\mathsf{E}[Y_{ki}] \;=\; \beta_0 \;+\; \beta_1 A^*_{ki} \;+\; \beta_1 G_k \;+\; \beta_3 A^*_{ki} G_k$$

- Consider the following correlation structures:

  Model 1. exchangeable (compound symmetric)

  Model 2. auto-regressive

  Model 3. unstructured (symmetric)

- Perform estimation/inference via ML and REML using the `gls()` function in the `nlme` library

```
> ##
> library(nlme)
>
> fit10.ML   <- gls(length ~ ageStar * gender, method="ML", data=growth,
+                   corr=corCompSymm(form = ~ 1 | id))
> fit10.REML <- gls(length ~ ageStar * gender, method="REML", data=growth,
+                   corr=corCompSymm(form = ~ 1 | id))
>
> ##
> fit20.ML   <- gls(length ~ ageStar * gender, method="ML", data=growth,
+                   corr=corAR1(form = ~ 1 | id))
> fit20.REML <- gls(length ~ ageStar * gender, method="REML", data=growth,
+                   corr=corAR1(form = ~ 1 | id))
>
> ##
> fit30.ML   <- gls(length ~ ageStar * gender, method="ML", data=growth,
+                   corr=corSymm(form = ~ 1 | id))
> fit30.REML <- gls(length ~ ageStar * gender, method="REML", data=growth,
+                   corr=corSymm(form = ~ 1 | id))
```

```
> ##
> summary(fit10.ML)
Generalized least squares fit by maximum likelihood
  Model: length ~ ageStar * gender
  Data: growth
       AIC      BIC    logLik
  426.1665 442.0329 -207.0833

Correlation Structure: Compound symmetry
 Formula: ~1 | id
 Parameter estimate(s):
      Rho
0.6103379

Coefficients:
                       Value Std.Error  t-value p-value
(Intercept)        21.209091 0.6402482 33.12636   0.000
ageStar             0.479545 0.0950982  5.04264   0.000
gendermale          1.490909 0.8429260  1.76873   0.080
ageStar:gendermale  0.320455 0.1252026  2.55949   0.012
...
Residual standard error: 2.21576
...
```

```
> ##
> summary(fit10.REML)
Generalized least squares fit by REML
  Model: length ~ ageStar * gender
  Data: growth
      AIC      BIC    logLik
  431.125 446.7561 -209.5625

Correlation Structure: Compound symmetry
 Formula: ~1 | id
 Parameter estimate(s):
      Rho
0.6250796

Coefficients:
                       Value Std.Error  t-value p-value
(Intercept)        21.209091 0.6500272 32.62801  0.0000
ageStar             0.479545 0.0944705  5.07614  0.0000
gendermale          1.490909 0.8558006  1.74212  0.0846
ageStar:gendermale  0.320455 0.1243761  2.57650  0.0114
...
Residual standard error: 2.288431
...
```

- Compare models for dependence using AIC
  - ⋆ $2 \times$ (Number of parameters $-$ value of the maximized log-likelihood)
  - ⋆ lower values indicate superior fit

	Number of	ML		REML	
	parameters	log-like	AIC	log-like	AIC
Exchangeable	6	-207.1	426.2	-209.6	431.1
Auto-regressive	6	-213.0	437.9	-214.8	441.7
Unstructured	11	-203.6	429.2	-206.0	433.9

  - ⋆ suggests that the exchangeable and unstructured models are superior to the auto-regressive structure
  - ⋆ not much difference between the exchangeable and unstructured models
  - ⋆ in practice, we might base final conclusions on the more parsimonious model

- Recall from the EDA that there was some evidence of heteroskedasticity
  - ⋆ empirical standard deviations increase with age:

$$\widehat{\boldsymbol{S}} = \begin{bmatrix} \mathbf{2.12} & & & \\ 0.83 & \mathbf{1.90} & & \\ 0.86 & 0.90 & \mathbf{2.36} & \\ 0.84 & 0.88 & 0.95 & \mathbf{2.44} \end{bmatrix}$$

- We can formally characterize and investigate this by using ML to fit a model that permits the age-specific variances to vary

```
>
> fit11.ML   <- gls(length ~ ageStar * gender, method="ML", data=growth,
+                  corr=corCompSymm(form = ~ 1 | id),
+                  weights=varIdent(form = ~ 1 | age))
>
> fit11.REML <- gls(length ~ ageStar * gender, method="REML", data=growth,
+                  corr=corCompSymm(form = ~ 1 | id),
+                  weights=varIdent(form = ~ 1 | age))
```

```
> summary(fit11.ML)
...
Correlation Structure: Compound symmetry
 Formula: ~1 | id
 Parameter estimate(s):
      Rho
0.6156417
Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | age
 Parameter estimates:
        8        10        12        14
1.0000000 0.8456271 1.0373947 0.9005826

Coefficients:
                       Value Std.Error  t-value p-value
(Intercept)        21.236378 0.6319446 33.60481  0.0000
ageStar             0.478601 0.0940939  5.08642  0.0000
gendermale          1.338854 0.8319937  1.60921  0.1107
ageStar:gendermale  0.332153 0.1238803  2.68124  0.0086
...
Residual standard error: 2.335967
...
```

- Again compare models with AIC:

	Number of	ML		REML	
	parameters	log-like	AIC	log-like	AIC
Homoskedastic					
Exchangeable	6	-207.1	426.2	-209.6	431.1
Auto-regressive	6	-213.0	437.9	-214.8	441.7
Unstructured	11	-203.6	429.2	-206.0	433.9
Heteroskedastic					
Exchangeable	9	-206.0	430.0	-208.5	435.1
Auto-regressive	9	-211.8	441.6	-213.8	445.6
Unstructured	14	-202.6	433.3	-205.1	438.2

  - ⋆ suggests that the homoskedastic model is superior
  - ⋆ exchangeable and unstructured models remain superior to the auto-regressive structure

- Formally evaluate heterskedasticity with the ML fits:

```
> ##
> lrt.gls <- function(fit.F, fit.R, digits=3)
+ {
+     nP.F    <- nrow(fit.F$apVar) + length(fit.F$coef)
+     nP.R    <- nrow(fit.R$apVar) + length(fit.R$coef)
+     test.df <- nP.F - nP.R
+     test.stat <- as.numeric(2 * abs(fit.F$logLik - fit.R$logLik))
+     p.value   <- 1 - pchisq(test.stat, test.df)
+     return(round(c(test.stat, test.df, p.value), digits=digits))
+ }
>
> ##
> lrt.gls(fit11.ML, fit10.ML)
[1] 2.190 3.000 0.534
> lrt.gls(fit21.ML, fit20.ML)
[1] 2.290 3.000 0.515
> lrt.gls(fit31.ML, fit30.ML)
[1] 1.898 3.000 0.594
```

- Find that there is insufficient evidence that the age-specific variances differ

Summary

- Goals:
  - ⋆ perform estimation/inference for regression parameters from a model for a continuous response while acknowledging within-cluster dependence
  - ⋆ learn about the structure of the correlation towards improving efficiency or because it is of intrinsic scientific interest
- Approach:
  - ⋆ specify a linear regression model for the mean structure
  - ⋆ use methods that are robust to misspecification of the dependence structure
  - ⋆ build and validate explicit models for the dependence structure
- Estimation/inference:
  - ⋆ two-stage and weighted least squares
  - ⋆ maximum and restricted maximum likelihood

$$
\begin{aligned}
\mathsf{Cov}[Z,\ \hat{\boldsymbol{\beta}}] &= \mathsf{E}[Z(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\mathsf{T}] \\
&= \mathsf{E}[B^\mathsf{T}\boldsymbol{Y}(\boldsymbol{Y}^\mathsf{T}G^\mathsf{T} - \boldsymbol{\beta}^\mathsf{T})] \\
&= B^\mathsf{T}\mathsf{E}[\boldsymbol{Y}\boldsymbol{Y}^\mathsf{T}]G^\mathsf{T} - B^\mathsf{T}\mathsf{E}[\boldsymbol{Y}]\boldsymbol{\beta}^\mathsf{T} \\
&= B^\mathsf{T}(\mathsf{Cov}[\boldsymbol{Y}] + \mathsf{E}[\boldsymbol{Y}]\mathsf{E}[\boldsymbol{Y}]^\mathsf{T})G^\mathsf{T} - B^\mathsf{T}\mathsf{E}[\boldsymbol{Y}]\boldsymbol{\beta}^\mathsf{T} \\
&= B^\mathsf{T}(\boldsymbol{\Sigma} + \boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}^\mathsf{T})G^\mathsf{T} - B^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^\mathsf{T} \\
&= B^\mathsf{T}\boldsymbol{\Sigma}G^\mathsf{T} + B^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}^\mathsf{T}G^\mathsf{T} - B^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^\mathsf{T} \\
&= B^\mathsf{T}\boldsymbol{\Sigma}G^\mathsf{T} && B^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta} = 0 \\
&= B^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1} \\
&= B^\mathsf{T}\boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1} \\
&= 0 && B^\mathsf{T}\boldsymbol{X} = 0
\end{aligned}
$$