Assessing Model Assumptions

- More on serial correlation
- Comments on residuals in general
- Normality assumptions for random effects

Note: Different groups sometimes propose different (but related) strategies.

Serial Correlation

Correlation within a cluster from successive measurements over time.

Further decompose $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i)$:

$$\boldsymbol{\epsilon}_i = \boldsymbol{\epsilon}_{(1)i} + \boldsymbol{\epsilon}_{(2)i}$$

$$\begin{aligned} \boldsymbol{\epsilon}_{(1)i} &= \text{serial correlation} \\ \boldsymbol{\epsilon}_{(2)i} &= \text{measurement error} \end{aligned}$$

Marginal covariance:

$$var(\mathbf{Y}_i) = \mathbf{V}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \tau^2 \mathbf{H}_i + \sigma^2 \mathbf{I}_{n_i}$$

$H_{ik} = g(|t_{ij} - t_{ik}|)$ for some function $g(\cdot)$ with $g(0) = 1$.

Informal Check for Serial Correlation

Do we need to model serial correlation? $\rightarrow$ hard because residual variability dominated by $\mathbf{Z}\widehat{\mathbf{b}}$

Idea: Orthogonalize $\mathbf{r}_i$ (OLS residuals – remove systematic effects) from $\mathbf{Z}_i$

- Lets us study variability not explained by random effects

Set $\mathbf{A}_i = n_i \times (n_i - q)$ matrix with $\mathbf{A}_i^T \mathbf{Z}_i = 0$ and $\mathbf{A}_i^T \mathbf{A}_i = \mathbf{I}$.

$$\Rightarrow \widetilde{\mathbf{r}}_i = \mathbf{A}_i^T \mathbf{r}_i \sim N(0, \mathbf{A}_i^T \mathbf{V}_i \mathbf{A}_i)$$

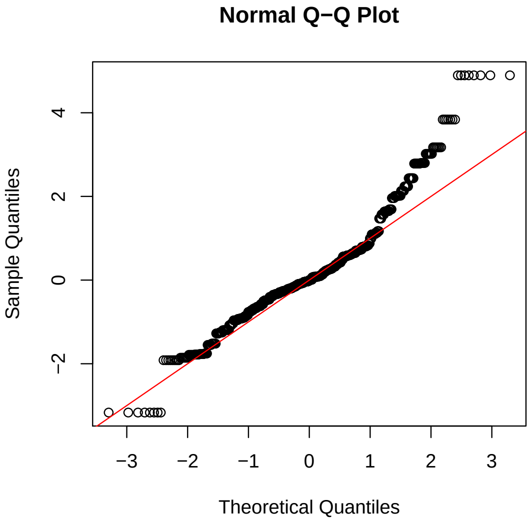
$$\mathbf{A}_i^T \mathbf{V}_i \mathbf{A}_i = \tau^2 \mathbf{A}_i^T \mathbf{H}_i \mathbf{A}_i + \sigma^2 \mathbf{I}_{n_i - q}.$$

No serial correlation means that the $\widetilde{\mathbf{r}}$'s are all $N(0, \sigma^2)$

Deviation from normality implies model is off: **possibly** need serial correlation component.

Informal Check for Serial Correlation

Random intercept + slope



Semi-Variograms

Empirical, nonparametric approach for studying serial correlation (Diggle, 1998)

Semi-Variogram for Random Intercept(Diggle, 1998)

$$\mathbf{V}_i = \nu^2 \mathbf{J}_{n_i} + \tau^2 \mathbf{H}_i + \sigma^2 \mathbf{I}$$

where $\mathbf{J}$ is matrix of 1's and ν^2 is variance of random intercept.

$$var(r_{ij}) = \nu^2 + \tau^2 + \sigma^2.$$

$$cor(r_{ij}, r_{ik}) = \rho(|t_{ij} - t_{ik}|) = \frac{\nu^2 + \tau^2 g(|t_{ij} - t_{ik}|)}{\nu^2 + \tau^2 + \sigma^2}$$

Then the semivariogram is defined as

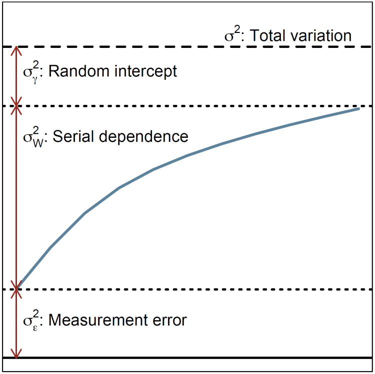
$$v(|t_{ij} - t_{ik}|) = \frac{1}{2} E(r_{ij} - r_{ik})^2 = \sigma^2 + \tau^2 (1 - g(|t_{ij} - t_{ik}|))$$

Semi-Variograms

Essentially: looking at similarity between pairs of observations

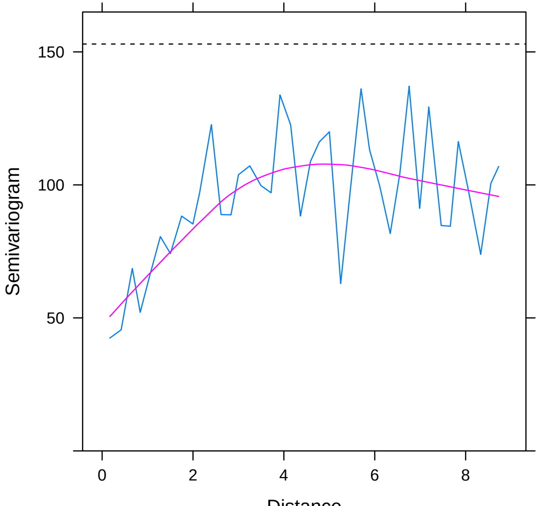
x-axis: Lag $= |t_k - t_j|)$

y-axis: Semivariance $= v(|t_{ij} - t_{ik}|)$



Variograms for CF Data

Sample variogram uses the empirical squared differences between pairs of residuals from the same subject.



Remarks

- Can compare the shape of variogram to theoretical correlation structures
- Can de-correlate the residuals from the fitted model: then should get a horizontal line if the correlation correctly specified.
- Can still be hard to tease apart a "best" structure: different strategies can give different answers
- As long as you're not strictly interested in the serial correlation, probably good enough just to include it (even if $g(\cdot)$ not optimal)

Residuals in Linear Mixed Models

Residual analysis is useful for checking model assumptions and looking for outliers. What is a "residual" for LMM?

- Marginal: $\mathbf{Y}_i - \mathbf{X}_i\widehat{\beta}$
  Deviation of individual curve from population mean
- Subject specific: $\mathbf{Y}_i - \mathbf{X}_i - \mathbf{Z}_i\widehat{\mathbf{b}}_i$
  Deviation of observations from subject specific predicted line
- Random effect: $\widehat{\mathbf{b}}_i$
  Deviation from population profile

Decorrelated Residuals

$$\widehat{\mathbf{V}}_i = \mathbf{L}_i \mathbf{L}_i^T$$

$$\mathbf{r}_i^* = \mathbf{L}_i^{-1} \mathbf{r}_i = \mathbf{L}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \widehat{\beta})$$

which are uncorrelated with variance 1.

Note that $\mathbf{r}_{i_1}^*$ is just standardized residual, but $\mathbf{r}_{i,k}^*$ is an estimate for

$$\frac{Y_{ik} - E[Y_{ik} | Y_{i1}, \ldots, Y_{i(k-1)}]}{sd(Y_{ik} | Y_{i1}, \ldots, Y_{i(k-1)})}$$

We can do all of the usual thinks with de-correlated residuals (e.g. normality, outlying observations, outlying individuals)

Outlying Individuals

Calculate Mahalanobis distance

$$d_i = {\mathbf{r}_i^*}^T \mathbf{r}_i^*$$

Then $d_i \sim \chi^2_{n_i}$ if model correctly specified $\rightarrow$ p-value

More formal notions of local influence for observations and subjects are in Verbeke and Molenberghs.

Normality Assumption of Random Effects

Recall: assume $\mathbf{b} \sim N(0, \mathbf{D}(\boldsymbol{\theta}))$

What is the impact of violations of normality?

How do we assess normality?

Impact of Normality Assumption

In General:

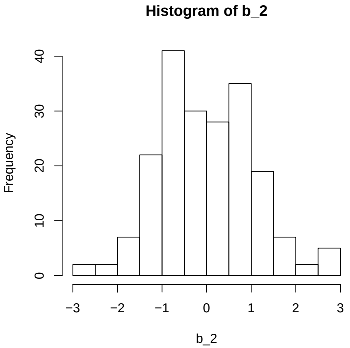
- Normality assumption significantly affects $\widehat{\mathbf{b}}$
- Normality assumption has little effect on β and $\boldsymbol{\theta}$ estimation
- Normality assumption affects the SEs and consequently β and $\boldsymbol{\theta}$ inference

Assessing Normality

Can we just look at the empirical estimates of $\widehat{b}_i$? $\rightarrow$ Only sometimes.

(1) The $\widehat{b}_i$ all have different individual distributions

(2) Shrinkage effect makes them look pretty normal anyway



Estimated $\widehat{b}_{i_2}$ from CF data.

Assessing Normality - Use More Complex Model!

Need to compare results under normality to results from relaxed model.

Heterogeneity Model:

$$\mathbf{b}_i \sim \sum_{j=1}^{g} \pi_j N(\boldsymbol{\mu}_j, \mathbf{D})$$

$$\sum_{j=1}^{g} \pi_j = 1 \quad \text{and} \quad \sum_{j=1}^{g} \pi_j \boldsymbol{\mu}_j = 0$$

Essentially: unobserved heterogeneity in the model.

Would like to test $H_0 : g = 1$ vs $H_A : g = 2$ (hard! $\leftarrow$ boundary problem)

Could also test: $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ OR $H_0 : \pi_1 = 0$ OR $H_0 : \pi_2 = 0$ (also hard!)

Heterogeneity Model

Conditional Model:

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i, \quad \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i)$$

$$\mathbf{b}_i \sim \sum_{j=1}^{g} \pi_j N(\boldsymbol{\mu}_j, \mathbf{D})$$

$$\sum_{j=1}^{g} \pi_j = 1 \quad \text{and} \quad \sum_{j=1}^{g} \pi_j \boldsymbol{\mu}_j = 0$$

Marginal Model:

$$\mathbf{Y}_i \sim \sum_{j=1}^{g} \pi_j N(\mathbf{X}_i\beta + \mathbf{Z}_i\boldsymbol{\mu}, \mathbf{V}_i)$$

Heterogeneity Model (2)

Estimation Usual EM-algorithm

Goodness of Fit Assess Need for Mixture

If $F_i(\cdot)$ is CDF, then $F_i(\mathbf{Y}_i) \sim \mathit{Unif} \rightarrow$ Hard due to multidimensionality

Instead: consider $\boldsymbol{a}_i^T \mathbf{Y}_i$

Heterogeneity Model Goodness of Fit

$\boldsymbol{a}_i^T \mathbf{Y}_i$ is univariate such that

$$\mathcal{U}_i = F_i(\boldsymbol{a}_i^T \mathbf{Y}_i) = \sum_{j=1}^{g} \pi_j \Phi \left(\frac{\boldsymbol{a}_i^T (\mathbf{Y}_i - \mathbf{X}_i \beta - \mathbf{Z}_i \boldsymbol{\mu}_j)}{\sqrt{\boldsymbol{a}_i^T \mathbf{V}_i \boldsymbol{a}}} \right) \sim Unif$$

KS-test assesses uniformity with estimates plugged in.

Any choice of $\boldsymbol{a}_i$ leads to valid test, but affects power: set equal to largest eigenvector of $\mathbf{R}_i^{-1} \mathbf{Z}_i \mathbf{D}^* \mathbf{Z}_i^T$ with $\mathbf{D}^* = \sum (\pi_j \boldsymbol{\mu}_j \boldsymbol{\mu}_j^T + \mathbf{D}_j)$, the overall covariance of $\mathbf{b}_i$.

Can test range of g to evaluate number of components.