Longitudinal data

Further Statistical Methods, Lecture 10
Hilary Term 2004

Steffen Lauritzen, University of Oxford; April 13, 2005

Longitudinal data

Longitudinal data can be seen as a specific type of multi-level data, where the level one units refer to observations over *time* of the value of specific quantities, taken on the same level two unit.

Typically level two units are here *individuals* $j = 1, \ldots, N$. For each of them we have observations $Y_{ij}, i = 1, \ldots, n_j$ taken at *times* $t_1, \ldots, t_{n_j}$.

Models for longitudinal data differ from general multilevel data partly by almost always using *time as a covariate*, but specifically by using *time in the dependence structure* between measurements taken on the same units.

Covariates for longitudinal data

As in the multilevel data we may have covariates $x_{ij} = (x_{ij1}, \ldots x_{ijk})^\top$ and $z_j = (z_{j1} \ldots, z_{jl})^\top$ at both levels.

But for longitudinal data x_{ij} typically include time or functions of time, such as e.g.

$$x_{ij1} = 1, \quad x_{ij2} = t_{ij}, \quad x_{ij3} = t_{ij}^2$$

corresponding to a quadratic trend, or

$$x_{ij1} = 1, \quad x_{ij2} = \cos(2\pi f t_{ij}), \quad x_{ij3} = \sin(2\pi f t_{ij})$$

corresponding to a periodic trend with period $\lambda = 1/f$, etc.

A general linear model

The general linear model for longitudinal data is then given as

$$Y_{ij} = \alpha^\top z_j + \beta^\top x_{ij} + \epsilon_{ij},$$

where the errors ϵ_{ij} are multivariate Gaussian and *correlated* as

$$\mathrm{Cov}(\epsilon_{ij}, \epsilon_{i'j'}) = v_{ii'jj'}$$

where

$$v_{ii'jj'} = \begin{cases} c(t_{ij}, t_{i'j}) & \text{if } j = j' \\ 0 & \text{otherwise,} \end{cases}$$

for some *covariance model* determined by the function c. The models thus allow for correlation between observations from the same individual but assume independence between individuals.

Correlation models

A flexible class of covariance models has three components:

$$c(t_{ij}, t_{i'j}) = \nu^2 + \sigma^2 \rho(t_{ij} - t_{i'j}) + \tau^2 \delta_{ii'},$$

where $\delta_{ii'}$ is 1 for $i = i'$ and 0 otherwise.

The first component ν^2 reflects the intrinsic correlation between measurements taken on the same individual, as in the multilevel case.

The second component describes a (stationary) serial correlation as known from time series analysis.

The final component τ^2 corresponds to an instantaneous noise term.

The variogram

The *variogram* for a stochastic process $X(t)$ is the function

$$\gamma(u) = \frac{1}{2}\mathbf{E}\left[\{X(t) - X(t-u)\}^2\right], \quad u \geq 0.$$

For the error process with three components just defined we get

$$\gamma(u) = \nu^2 + \sigma^2\{1 - \rho(u)\}, \text{ for } u > 0.$$

Choosing ρ so that $\rho(0) = 1, \lim_{t\to\infty} \rho(t) = 0$ yields

$$\gamma(0) = \nu^2, \quad \lim_{t\to\infty} \gamma(u) = \sigma^2 + \tau^2 \tag{1}$$

whereas the process variance is

$$\mathbf{V}\{Y(t_{ij})\} = c(t_{ij}, t_{ij}) = \nu^2 + \sigma^2 + \tau^2, \tag{2}$$

as reflected in the following diagram, taken from Diggle et al. (2002).

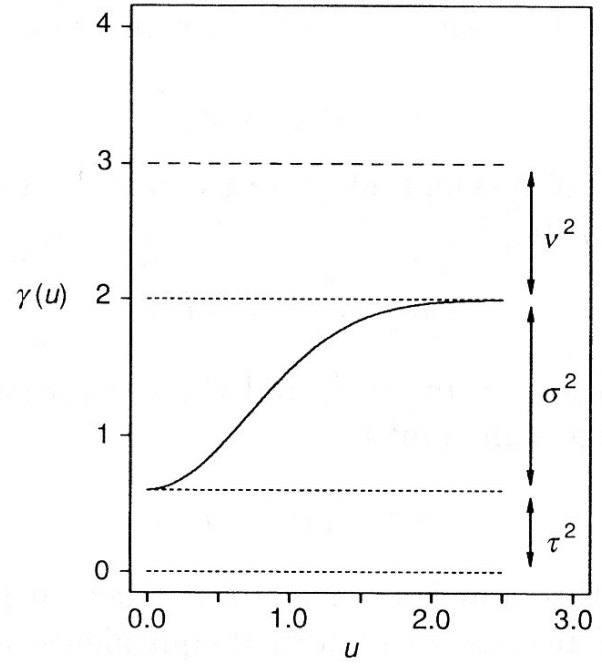

Fig. 5.4. The variogram for a model with a random intercept, serial correlation, and measurement error.

Sample variogram

To identify reasonable suggestions for the covariance structure, residuals r_{ij} from a least squares fit of the parameters are calculated and the *sample variogram* is based on a curve through points (u_{ijk}, v_{ijk}), where

$$u_{ijk} = t_{ij} - t_{kj}, \quad v_{ijk} = \frac{1}{2}(r_{ij} - r_{kj})^2$$

or rather averages of v_{ijk} for indices corresponding to identical time differences u.

Such a sample variogram gives a first idea of the importance of the three components of variance using (1) and (2) and some idea of the shape of the serial correlation function ρ.

An example of a sample variogram, taken from Diggle et al. (2002) is seen below. Note that there are few large time differences, so the variogram becomes noisy for large lags, here around lag 10.

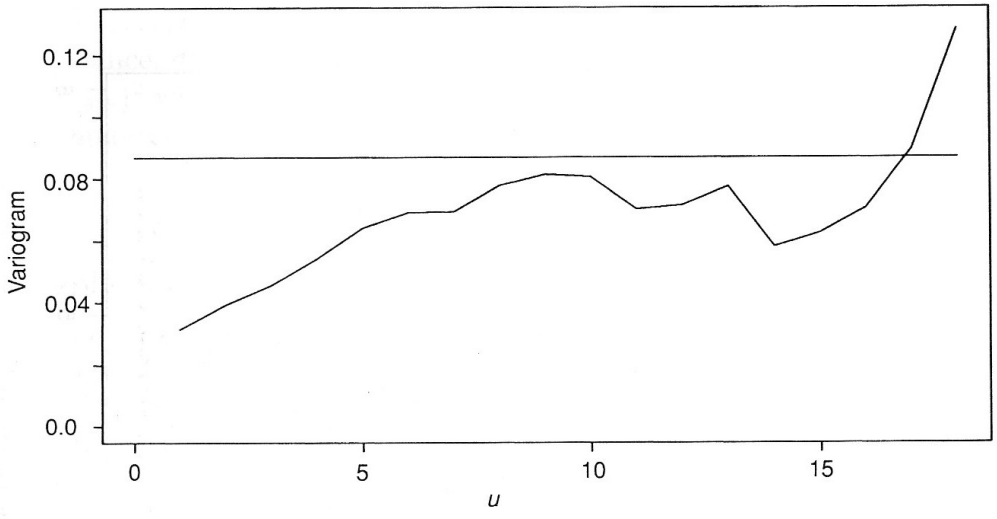

Fig. 3.16. Sample variogram of milk protein residuals. Horizontal line estimates process variance.

In this case there is essentially no within pig correlation.

Choice of correlation function

Generally the time series are often many but short, so there is little information about the shape of the serial correlation function and one is forced to rather ad hoc choices.

The serial correlation function must be positive definite to ensure matrices of the form $m_{rs} = \rho(t_r - t_s)$ are positive definite, for all choices of k and $t_1, \ldots, t_k$.

Typical choices which satisfy these restrictions are

$$\rho_1(t) = e^{-\theta|t|}, \quad \rho_2(t) = e^{-\theta t^2/2},$$

known as the *exponential* and *Gaussian* correlation model.

It can be difficult to distinguish these from the sample variogram.

Estimation of parameters

In principle this is done in the same way as in other multi-level models, using *residual maximum likelihood* (REML).

Straight maximum likelihood yields strongly biased estimates of the variance parameters and should be avoided.

Routines for calculating the REML estimates are available in many forms of software.

They can be calculated using the following steps:

1. Calculate estimates $(\tilde{\alpha}, \tilde{\beta})$ of the linear parameters by ordinary least squares (OLS), ignoring the correlation;

2. Calculate the residuals

$$r_{ij} = y_{ij} - \tilde{\alpha}^\top z_j - \tilde{\beta}^\top x_{ij}$$

from the OLS analysis;

3. The vector R of residuals is $\mathcal{N}(0, W)$ where the covariance matrix W has the form

$$W = \nu^2 A + \sigma^2 B(\theta) + \tau^2 C$$

where A, B, C are known matrices, B possibly depending on θ;

4. Calculate the MLE of $(\nu^2, \sigma^2, \tau^2, \theta)$ *based on the likelihood for the residuals*;

5. Calculate the final estimates $(\hat{\alpha}, \hat{\beta})$ using *weighted least squares* (WLS) with weights determined by the given covariance model and its estimated parameters.