## Parametric Models for covariance structure

We consider the General Linear Model for correlated data, but assume that the covariance structure of the sequence of measurements on each unit is to be specified by the values of unknown parameters

- Parametric modelling approach very useful for data in which the measurements on different units are not made at a common set of times
- $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$
- $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{in_i})$
- $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m)$
- $\boldsymbol{t} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_m),\ N = \sum_{i=1}^m n_i$



## Parametric Models for covariance matrices

$$\begin{aligned} E[\boldsymbol{Y}] &= X\boldsymbol{\beta} \\ Var(\boldsymbol{Y}) &= V(\boldsymbol{\alpha}) \end{aligned}$$

We considered two examples:

1. Uniform correlation model

$$\epsilon_{ij} = U_i + Z_{ij},\ Z_{ij} \sim N(0, \sigma^2)$$

2. Exponential correlation model

$$\begin{aligned} \epsilon_{ij} &= W_{ij} \\ W_{ij} &= \rho W_{ij-1} + Z_{ij},\ Z_{ij} \sim N(0, \sigma^2) \end{aligned}$$

In these cases $\boldsymbol{\alpha} = c(\sigma^2, \rho)$

We now consider more general models for the covariance matrix $V(\boldsymbol{\alpha})$ which can be specified by looking at the variogram.



## Interpretation of the Variogram

Diggle (1988) proposed the following model

$$Y_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + U_i + W_i(t_{ij}) + Z_{ij}$$

This model contained three sources of variation:

| | |
|---|---|
| random intercept | $U_i$ |
| serial process | $W_i(t_{ij})$ |
| measurement error | $Z_{ij}$ |

If we further assume

$$\begin{aligned} \text{var}(U_i) &= \nu^2 \\ \text{cov}(W(s), W(t)) &= \rho(\mid s - t \mid) \\ \text{var}(Z_{ij}) &= \tau^2 \end{aligned}$$

Then we can use the Variogram to characterize these variance components



## Interpretation of the Variogram

- the subject-specific intercept $U_i$ expresses the degree to which all observations on the same subject are similar, $U_i$ is a subject "trait" variable, and $\nu^2$ is the between-subject variance
- $W_i(t_{ij})$ is the serial process an within-subject variance
- $Z_{ij}$ is the measurement error (noise)

## Example on the Protein Content of Milk

Consider the mean model

$$E(Y_{ij}) = \beta_0 + \beta_1 \texttt{mixed}_i + \beta_2 \texttt{barley}_i + \alpha_j$$

where $\alpha_j$ is a factor for time $j$

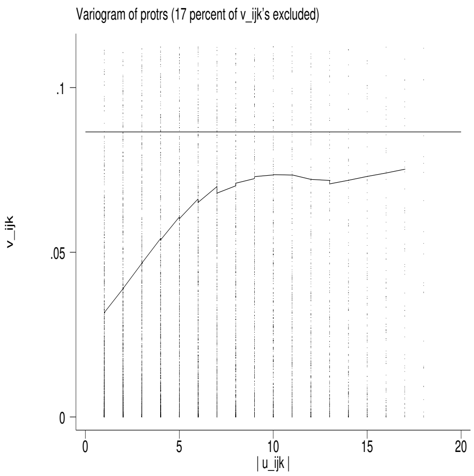
- First, compute residuals, allowing for a different mean for each time and diet
- The overall variance of the residuals is $0.2942^2 = 0.087$
- There are 19 time points, so there are 18 lags
- **Note:** the horizontal line is $\hat{\sigma}^2$
- Looking at the variogram there is evidence of:

  serial correlation (the variogramis increasing with lag)

  random intercept (the variogram does not start at zero)

  measurement error (the variogram does not reach the total variance $\hat{\sigma}^2$






Recall that the sample variogram with a bandwidth of 0.7 was This was generated by stata code



```
. use cows

 . * Compute residuals, removing effects of diet and time
 . sort diet week

. by diet week : egen protmn=mean(prot)

. gen protrs = prot - protmn

. sort id week

. * Compure smooth lowess variogram with
    bandwidth 0.7
. variogram protrs , bw(0.7) incl(1.3)
```



### Parametric Models for covariance structure

1. **Serial correlation:**
$$\epsilon_{ij} = W_i(t_{ij})$$
2. **Serial correlation + measurement error:**
$$\epsilon_{ij} = W_i(t_{ij}) + Z_{ij}$$
3. **Random intercept + measurement error:**
$$\epsilon_{ij} = U_{i0} + Z_{ij}$$
4. **Random intercept + random slope + meas. error:**
$$\epsilon_{ij} = U_{i0} + U_{i1}t_{ij} + Z_{ij}$$
5. **Random intercept + serial correlation + meas. error:**
$$\epsilon_{ij} = U_{i0} + W_i(t_{ij}) + Z_{ij}$$

## Models

To develop a model we need to understand the sources of variations

- **Random effects**: stochastic variation between units
- **Serial correlation**: time-varying stochastic process within an unit
- **Measurement error**: measurement process introduces a component of random variation (sampling within units)



## How do we incorporate these qualitative features into specific models?

1. Make explicit separation between mean and covariance structures as

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

It follows that

$$\boldsymbol{\epsilon} \sim MVN(0, V(\boldsymbol{t}, \boldsymbol{\alpha}))$$

$$\epsilon_{ij} = \underbrace{\boldsymbol{d}_{ij}^{'}\boldsymbol{U}_i}_{\text{random effects}} + \underbrace{W_i(t_{ij})}_{\text{serial corr}} + \underbrace{Z_{ij}}_{\text{meas error}}$$

For example: $\boldsymbol{d}_{ij}^{'}\boldsymbol{U}_i = U_{i0} + U_{i1}t_{ij}$



## Serial Correlation: Example

- $corr(W_i(t_1), W_i(t_2)) = \rho|\ t_{i1} - t_{i2}\ |$

  $W_i(t_{ij}) \sim N(0, \sigma^2 H_i)$

$$\begin{aligned} H_i &= \begin{pmatrix} 1 & e^{-\theta|t_{i1}-t_{i2}|} & \ldots & e^{-\theta|t_{i1}-t_{in}|} \\ \vdots & \vdots & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} \\ \boldsymbol{t} &= (-1, 0, 2, 20),\ \theta = .2 \\ H &= \begin{pmatrix} 1 & .82 & .55 & .015 \\ & 1 & .67 & .018 \\ & & 1 & .027 \\ & & & 1 \end{pmatrix} \end{aligned}$$



## Serial Correlation

$\boxed{\epsilon_{ij} = W(t_{ij}); Var(\boldsymbol{\epsilon}_i) = \sigma^2 H_i}$

- $Var(\epsilon_{ij}) = \sigma^2$
- $Cov(\epsilon_{ij}, \epsilon_{ik}) = \sigma^2\rho(|\ t_{ij} - t_{ik}\ |)$
- variogram $\gamma(u) = \sigma^2(1 - \rho(u)),\ \gamma(0) = 0$
- three popular choices of $\rho(u)$ are:

1. the exponential correlation model $\rho(u) = e^{-\phi u}$
2. the Gaussian correlation model $\rho(u) = e^{-\phi u^2}$
3. first order autoregressive model $\rho(u) = e^{-\phi|t_j - t_{j-1}|}$

## Model-fitting

1. **Formulation** – choosing the general form of the model;

a) Mean

b) Association

2. **Estimation** – fit the model

a) Weighted least squares for $\boldsymbol{\beta}$

b) ML for covariance parameters $\boldsymbol{\alpha}$ or subset

c) Iterate a) and b) to converge

3. **Inference** – calculating confidence intervals or testing hypotheses about parameters of interest

4. **Diagnostic** – checking that the model fits the data

   Examine residuals for lack of fit, correlation



## Formulation

- Formulation of the model is a continuation of exploratory data analysis
- Focus on the mean and covariance structure

1. Look at the residuals
2. Do time plots, scatterplot matrices and empirical variogram plots
3. Do you have stationarity? If not.. you need to transform the data or use inherently non-stationary model as the random effects model.
4. Once the stationarity has been achieved, use empirical variogram to estimate the underlying covariance structure.



## Estimation

1. Unknown parameters are: $\boldsymbol{\beta},\ \boldsymbol{\alpha}$ and $\sigma^2$
2. Given $\boldsymbol{\alpha}$, find REML estimates of $\boldsymbol{\beta}$ and $\sigma^2$

   $\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}),\ \hat{\sigma}^2(\boldsymbol{\alpha})$
3. $L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\alpha}) = L(\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}), \hat{\sigma}^2(\boldsymbol{\alpha}), \boldsymbol{\alpha}) = L^*(\boldsymbol{\alpha})$
4. $\hat{\boldsymbol{\alpha}} = \text{argmax} L^*(\boldsymbol{\alpha})$
5. $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\alpha}}),\ \hat{\sigma}^2 = \hat{\sigma}^2(\hat{\boldsymbol{\alpha}})$



## Diagnostic

AIM: compare the data with the fitted model

1. Superimpose the fitted mean response profiles on a time plot of the average observed responses within each combination of treatment and times
2. Superimpose the fitted variogram on a plot of the empirical variogram

## Examples and Summary

Nepal Data set

This data contains anthropologic measurements on Nepalease children. The study design called for collecting measurements on 2258 kids at 5 time points, spaced approximately 4 months a part.

- Scientific question: estimate association between arm circumference and child's weight taking into account of the correlation
- **Model for the mean:**

$$E[\text{arm}_{ij}] = \beta_0 + \beta_1 \text{wt}_{ij} + \beta_2 \text{age}_j + \beta_3 \text{sex}_i + \epsilon_{ij}$$



- **Models for the covariance matrix:**

Indipendence model:

$$\epsilon_{ij} = Z_{ij},\ \text{corr}(Z_{ij}, Z_{ij'}) = 0$$

Uniform:

$$\epsilon_{ij} = U_i + Z_{ij}$$

Exponential:

$$\begin{aligned} \epsilon_{ij} &= W_{ij} + Z_{ij} \\ W_{ij} &= \rho W_{ij-1} + Z_{ij} \end{aligned}$$

Uniform + Exponential:

$$\begin{aligned} \epsilon_{ij} &= U_i + W_{ij} + Z_{ij} \\ W_{ij} &= \rho W_{ij-1} + Z_{ij} \end{aligned}$$



## Indipendence

- One very simple model is to set $\rho_{jk} = 0$ and calculate OLS

```
. regress arm wt age sex
<snip>
--------------------------------------------------------------------
  arm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+------------------------------------------------------------
   wt |   .4978829    .016785    29.66   0.000     .4649392    .5308266
  age |  -.0395428   .0026665   -14.83   0.000    -.0447763   -.0343092
  sex |   .2410595   .0473892     5.09   0.000     .1480494    .3340697
_cons |   9.545213   .1392856    68.53   0.000      9.27184    9.818587
--------------------------------------------------------------------------
```

- Interpretation of $\hat{\beta}_1 = 0.498$, estimated coefficient of weight:
- While the independence correlation model is most likely wrong, the $\hat{\beta}$'s are **still valid estimates** (but not very good ones!)
- The standard errors, however, are **wrong** (and so, therefore, is the rest of the table!) (the actual s.e. of $\hat{\beta}_1$ is more like $0.037$)



## Exchangeable Correlation Model

A better model is to assume $\rho_{jk} = \rho$ (same for all pairs of observations). This is called the **uniform**, **exchangeable** or **compound symmetry** correlation model. In stata:

```
. xtreg arm wt age sex , re
<snip>
------------------------------------------------------------------------------
  arm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   wt |   .6572958   .0231161    28.43   0.000     .6119891    .7026024
  age |  -.0588112   .0036029   -16.32   0.000    -.0658728   -.0517496
  sex |   .2900758   .0928129     3.13   0.002     .1081658    .4719857
_cons |   8.419315   .2176118    38.69   0.000     7.992804    8.845826
-------------+----------------------------------------------------------------
sigma_u |  .60715091
sigma_e |  .35223707
    rho |  .74818317   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

- Interpretation of $\hat{\beta}_1 = 0.657$ is the same as in the independence model
- This is a **much better** estimate because we have made some attempt to account for the correlation in the repeated measures

- The estimated **within-subject** correlation coefficient is $\hat{\rho} = 0.748$ which denotes the correlation between two arm circumference measures on the same child.
- This estimate excludes any similarities in arm circumference due to weight, age or sex. i.e., $\hat{\rho}$ is **adjusted** for weight, age and sex.
- Furthermore, if the correlation model is (approximately) correct, the standard errors are also (approximately) correct
- How does this model arise? Suppose that

$$\epsilon_{ij} = U_i + Z_{ij}$$

where $U_i \sim N(0, \nu^2)$, $Z_{ij} \sim N(0, \tau^2)$ and the $U_i$'s and $Z_{ij}$'s are all independent of one another

Then:

$$\begin{aligned} \text{var}(\epsilon_{ij}) &= \nu^2 + \tau^2 \\ \text{cov}(\epsilon_{ij}, \epsilon_{ik}) &= \rho\nu^2 \\ \rho &= \frac{\nu^2}{\nu^2+\tau^2} \end{aligned}$$

- The $U_i$ allows each child to have his/her own intercept $\beta_0 + U_i$, and the term $U_i$ is called a **random effect**



- In this model, the estimates are

$$\widehat{var}(U_i) = \hat{\nu}^2 = 0.60715^2 = 0.369$$

and

$$\widehat{var}(Z_{ij}) = \hat{\tau}^2 = 0.35224^2 = 0.124$$



## Exponential Correlation Model

A different model is to assume that the correlation of observations closer together in time is larger than that of observations farther apart

- One model for this is the **exponential** or **AR (1)** correlation model

$$\rho_{jk} = \rho^{|t_j - t_k|}$$

- A fit of this model is available in stata:

```
. prais arm wt age sex
<snip>
------------------------------------------------------------------------------
  arm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   wt |   .6724654    .024466    27.49   0.000     .6244464    .7204844
  age |  -.0615307   .0041314   -14.89   0.000    -.0696393   -.0534221
  sex |    .303858   .0939725     3.23   0.001     .1194196    .4882965
_cons |   8.311044   .2259521    36.78   0.000     7.867571    8.754517
-------------+----------------------------------------------------------------
  rho |   .8287446
------------------------------------------------------------------------------
```

- Again, the interpretation of $\hat{\beta}_1 = 0.672$ is the same as in the independence model, and it is a valid estimate



- If the exponential correlation model is correct, standard errors will also be correct
- The estimated **within-subject** correlation coefficient for two observations separated by 4 months is

$$\widehat{corr}(\epsilon_{ij}, \epsilon_{i,j-1}) = \hat{\rho} = 0.829$$

(again, adjusted for weight, age and sex)

- How does this model arise? Suppose that $\epsilon_{ij} = W_{ij}$, and

$$W_{ij} = \rho W_{i,j-1} + Z_{ij}$$

where $Z_{ij} \sim N\{(0, \sigma^2(1-\rho^2)\}$ are all independent of one another

- This is sometimes called a **first order autoregressive** (AR(1)) model and it allows each subject's error term $\epsilon_{ij}$ at a given time to be a function of his error term $\epsilon_{i.j-1}$ at the **previous** time

## Exchangeable plus Exponential

- Suppose that:

$$\begin{aligned}\epsilon_{ij} &= U_i + W_{ij} + Z_{ij}\\ W_{ij} &= \rho W_{i,j-1} + Z_{ij}\end{aligned}$$

- The $U_i$ provides for a subject-specific intercept
- The $W_{ij}$ provide for an autoregressive error structure
- This will induce a new correlation structure
- If we fit this model, we get:

```
. xtregar arm wt age sex
<snip>
------------------------------------------------------------------------------
         arm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          wt |   .6446496   .0229793    28.05   0.000     .5996109    .6896882
         age |  -.0574427   .0036364   -15.80   0.000    -.0645698   -.0503155
         sex |   .2832056    .087249     3.25   0.001     .1122007    .4542105
       _cons |   8.514987   .2113676    40.29   0.000     8.100714     8.92926
-------------+----------------------------------------------------------------
      rho_ar |  .18543604   (estimated autocorrelation coefficient)
     sigma_u |  .57772352
     sigma_e |  .38673782
     rho_fov |  .69055107   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```



## In summary

- Indipendence model: $\hat{\beta}_1 = 0.497$ and $se = 0.0167$
- Uniform model: $\hat{\beta}_1 = 0.657$ and $se = 0.0231$ and $\hat{\rho} = 0.748$
- Exponential model: $\hat{\beta}_1 = 0.672$ and $se = 0.0244$ and $\hat{\rho}_{ar} = 0.8287$
- Uniform + Exponential model: $\hat{\beta}_1 = 0.644$ and $se = 0.0229$ and $\hat{\rho} = 0.69$ and $\hat{\rho}_{ar} = 0.185$
- We see that the exchangeable correlation (0.691) is similar to the model without the exponential correlation (0.748), and the exponential correlation (0.185) is now much smaller
- The regression parameter estimates are also similar to the exchangeable case
- This suggests that the exchangeable correlation model may be capturing the main correlation pattern here



## In summary

- Modelling the correlation in longitudinal data is important to be able to obtain correct inferences on regression coefficients ($\beta$)
  - statistical efficiency
  - correct standard errors
- These are **marginal models** because the interpretation of the regression coefficients is the **same** as that in cross-sectional data
  - Exchangeable correlation model: subject-specific formulation
  - Exponential correlation model: transition model formulation
- Three basic elements of correlation structure
  - random effects
  - autocorrelation or serial dependence
  - observation-level noise, measurement error



## Evaluating Covariance Models

- Once you have chosen a (set of) covariance model(s), how do you evaluate whether it fits the data well, or how do you compare several of them?
- Several tools, and each work with either ML or ReML

  LRTs for comparing **nested** models

  Akaike's Information Criterion

  Examining fitted model variograms

## Comparing Covariance Models with Akaike's Information Criterion (AIC)

- Useful for comparing several models when some of them may not be nested within others
- In the study of arm circumference of Nepalese children, we considered three covariance models, and obtained three values of $2L = 2\log(l)$:
  - Exponential plus exchangeable: $2L = -1208.6$
  - Exponential only: $2L = -1301.6$
  - Exchangeable only: $2L = -1213.7$
- We can compare the simpler models to exponential plus exchangeable via LRT, but what about choosing among the three?
- To compare all three, use

$$AIC = -2(L - q)$$

where $q =$ the number of parameters in the covariance model

Then, pick the one with the **smallest** AIC



- The idea is to penalize for using more parameters, hence the $-q$
- For the Nepalese children, the three AICs based on ML are:
  - Exponential plus exchangeable: AIC =
  - Exponential only: AIC =
  - Exchangeable only: AIC =

from which we would conclude that the exchangeable only is definitely better than the exponential only, but that the exchangeable plus exponential is the best