Parametric Models for Covariance Structure: Examples

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$$

1. Model for the mean: $E[\boldsymbol{Y}_i] = X_i\boldsymbol{\beta}$

Here you want to take into account all the covariates that might have an effect on the response ... easy!

- Is the ozone suppresses growth ?
- How the growth change over time?
- Does AZT have an effect on CD4 cell depletion ?
- What is the average average time course of CD4+ cell depletion?
- Does diet affect protein content of milk?
- Does the protein content of milk change over time?

2. Model for the covariance matrix

Here you want to model the $Var(Y_{ij})$ and the the correlation between Y_{ij} and Y_{ik}... hard

- Is the protein content of milk measured perfectly or there is measurement error?
- Is the correlation between Y_{ij} and Y_{ik} a function of the time difference $t_{ij} - t_{ik}$?
- Are there unmeasured cows characteristics? Should we include in the analysis cows specific baseline contents of milk?

2. Model for the covariance matrix

$$V(\boldsymbol{Y}_i) = V(\boldsymbol{\epsilon}_i) = V_i(\boldsymbol{\alpha})$$

- $Var(\epsilon_{ij})$ can be explained by
  1. serial correlation $W_i(t_{ij})$
  2. random effects U_i
  3. measurement error Z_{ij}

$$\begin{aligned} \epsilon_{ij} &= U_i + W_i(t_{ij}) + Z_{ij} \\ Var(\epsilon_{ij}) &= var(U_i) + var(W_i(t_{ij}) + var(Z_{ij}) \\ &= v^2 + \sigma^2 + \tau^2 \\ \gamma(u) &= \sigma^2(1 - \rho(u)) + \tau^2 \end{aligned}$$

Which model should I pick?

1. Remove the effect of time, treatment and of others explanatory variables and estimate the residuals: $\hat{\epsilon}_{ij} = y_{ij} - X_i\hat{\boldsymbol{\beta}},$ where $\hat{\boldsymbol{\beta}}$ is the OLS estimate of $\boldsymbol{\beta}$
2. estimate the process variance by: $vtot = \sum \hat{\epsilon}_{ij}^2/(N-p)$
3. Estimate the empirical variogram $\hat{\gamma}(u)$
4. If $\hat{\gamma}(0) > 0$ then you need to include the measurement error component $\tau^2 = 0$
5. If t $\hat{\gamma}(u) < vtot$ then there is a random effect component $v^2 > 0$

CD4+ level

- HIV attack CD4+ cell which regulates the body's immunoreponse to infectious agent
- 2376 values of CD4+ cell number plotted against time since sieroconversion for 369 infected men enrolled in the MACS
- **Goals:**

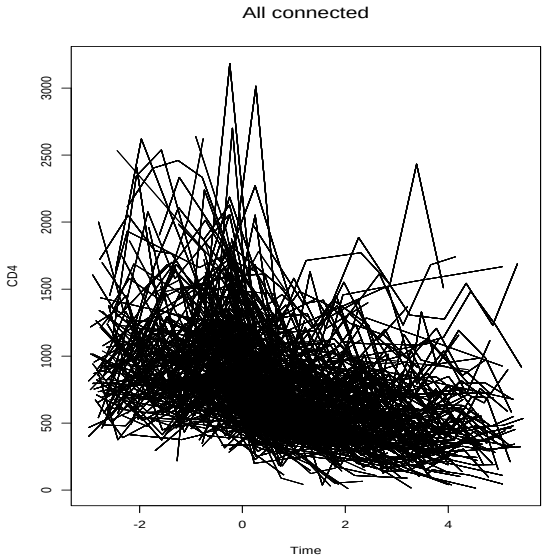
1. estimate the average time course of CD4+ cell depletion
2. identify factors which predict CD4+ cell changes
3. estimate the time course for an individual men taking into account of the measurement error in CD4+ cell determinations
4. characterize the degree of heterogeneity across men in the rate of progression

Aim 1: estimate the average time course of CD4+ cell depletion

- Y_{ij} = CD4+ level at time t_{ij} for subject i

$$
\begin{aligned}
Y_{ij} &= \beta_0 + \beta_1 t_{ij} + \epsilon_{ij} \\
E[Y_{ij}] &= \beta_0 + \beta_1 t_{ij} \\
\epsilon_{ij} &= W_i(t_{ij}) \\
Var(Y_{ij}) &= Var(\epsilon_{ij}) = \sigma^2 \\
Corr(\epsilon_{ij}, \epsilon_{ik}) &= \exp(-\phi \mid t_{ij} - t_{ik} \mid)
\end{aligned}
$$

The model for the covariance matrix is a model of serial correlation



Aim 2: identify factors which predict CD4 + cell changes

- $AZT_i = 1$ if subject i receive the antiviral drug
- $AZT_i = 0$ if subject i did not
- t_{ij} is the time after sieroconversion

$$
\begin{aligned}
Y_{ij} &= E[Y_{ij}] + \epsilon_{ij} \\
E[Y_{ij}] &= \beta_0 + \beta_1 t_{ij} + \beta_2 AZT_i + \beta_3 AZT_i t_{ij} \\
\epsilon_{ij} &= W_i(t_{ij}) \\
Var(Y_{ij}) &= Var(\epsilon_{ij}) = \sigma^2 \\
Corr(\epsilon_{ij}, \epsilon_{ik}) &= \exp(-\phi \mid t_{ij} - t_{ik} \mid)
\end{aligned}
$$

The model for the covariance matrix is a still model of serial correlation. We have changed the model for the mean

Parameter interpretation

- $\beta_0 + \beta_2$: CD4 level at sieroconversion for AZT group
- β_0 : CD4 level at sieroconversion for the non AZT group
- $\beta_1 + \beta_3$: Cells/year loss for the AZT group
- β_1: Cells/year loss for the non AZT group

Aim 3: estimate the time course for an individual men taking into account of the measurement error in CD4+ cell determinations

1. Model for the mean

$$\begin{aligned} Y_{ij} &= E[Y_{ij}] + \epsilon_{ij} \\ E[Y_{ij}] &= \beta_0 + \beta_1 t_{ij} + \beta_2 AZT_i + \beta_3 AZT_i t_{ij} \end{aligned}$$

2. Model for the Variance

$$\epsilon_{ij} = U_{i0} + U_{i1} t_{ij} + W_i(t_{ij}) + Z_{ij}$$

3. Model for the mean + Model for the Variance

$$\begin{aligned} Y_{ij} &= \beta_0 + U_{i0} + \beta_2 AZT_i + \\ &+ (\beta_1 + \beta_3 AZT_i + U_{i1}) t_{ij} + W_i(t_{ij}) + Z_{ij} \end{aligned}$$

This is model with Random intercept and slope + serial correlation + measurement error

Parameter interpretation

- $\beta_0 + U_{i0} + \beta_2$: CD4 level at sieroconversion for for subject i in the AZT group
- $\beta_0 + U_{i0}$: CD4 level at sieroconversion for for subject i not in the AZT group
- $\beta_1 + \beta_3 + U_{i1}$: Cells/year loss for subject i in the AZT group
- $\beta_1 + U_{i1}$: Cells/year loss for for subject i not in AZT group

Interpretation of the coefficients for a random effect models

- $\beta_{0i} =$ $\underbrace{\beta_0}_{\text{population average intercept}} + \underbrace{U_{i0}}_{\text{random effect}}$

$\beta_{0i} =$ *subject-specific intercept for the non AZT group*

- $\beta_{1i} =$ $\underbrace{\beta_1}_{\text{population average slope}} + \underbrace{U_{i1}}_{\text{random effect}}$

$\beta_{1i} =$ *subject-specific slope for the non AZT group*

- $(U_{i0}, U_{i1}) \sim MVN\left((0,0), \begin{bmatrix} v_1^2 & 0 \\ 0 & v_2^2 \end{bmatrix}\right)$

Aim 4: characterize the degree of heterogeneity across men in the rate of progression

We need to estimate v_1^2 and v_2^2!

Sitka Spruce tree

- data consist of measurements on 79 sitka spruce trees over two growing seasons
- the trees were grown in four controlled environment chambers, of which the first two containing 27 trees each, were treated with introduced ozone at 70 ppb whilst the remaining two, containing 12 and 13 trees, were controls
- response variable is the log-size measurement $y = \log(hd^2)$ where h denotes height and d denoted diameter
- question: is there a ozone effect on the growth pattern?

Remove the effects of explanatory variables

1. For example, you might want to obtain the residuals from a 2-way anova model (OLS) on day and treatment group (with interaction)

```
fit88 _ aov(logsize ~ as.factor(days) * as.factor(ozone),
            data = sitka[sitka$year == 88,])       #commands

           Df Sum of Sq  Mean Sq  F Value     Pr(F)
days        4   93.3623 23.34058 58.61906 0.0000000
ozone       1    3.8097  3.80967  9.56786 0.0021246
days:ozone  4    0.5629  0.14073  0.35345 0.8416089
Residuals 385  153.2969  0.39817
```

2. vtot $= 153.2969/\ 385 = 0.39817$
3. Estimate $\hat{\gamma}(u)$

```
# compute the pairwise differences of the residuals
# within each tree (pairdiff); there are (4+3+2+1 = 10)
#pairs for each tree.
vs _ pairdiffs^2 / 2
# compute the corresponding pairwise differences of the
# days (us)
# plot the sample variogram values
plot(rep(us, 79),vs,xlab="lag in days",ylab="Variogram")
```


Estimation of a Variogram: A toy Example

```
RES #commands in R
     [,1] [,2] [,3]
[1,]    2    1    3
[2,]    3    6    6
[3,]    1    7    9
[4,]    5    2   10
 TIME   = 1,4,5
 PAIRS _ apply(RES,1,function(x){xo _ outer(x,x,"-")
                        xo[col(xo) >row(xo)]} )
     [,1] [,2] [,3] [,4]
[1,]    1   -3   -6    3
[2,]   -1   -3   -8   -5
[3,]   -2    0   -2   -8
dim(PAIRS) = (2 + 1) x 4
VARIOGRAM _ PAIRS^2/2
US _ outer(TIME,TIME,'-')
US _ US[col(US) > row(US)]
US  _ -US
plot(rep(US,4),VARIOGRAM)
VARIOGRAM.mean _ apply(VARIOGRAM,1,mean)
lines(sort(US),VARIOGRAM.mean[order(US)])
```


Example: Protein contents of milk samples

- Y_{ij} = protein content
- i =cow, $i = 1, \ldots, 79$
- j = week, $j = 1, \ldots, 19$
- 25 cows received a barley diet
- 27 cows a mixed diet of barley and lupins
- 27 cows a diet of lupins only
- initial drop, settling-in period, gentle rise towards the end (see fig 1.4)
- the empirical variogram shows a smooth rise with increasing lag (fig 3.16)
- time is measured in weeks since calving, and the experiment was termined 19 weeks after the earliest calving
- About half of the 79 sequences of milk protein measurements are incomplete

Model for the Mean

$$\mu_g(t) = \begin{cases} \beta_{0g} + \beta_1 t & t \leq 3 \\ \beta_{0g} + 3\beta_1 + \beta_2(t-3) + \beta_3(t-3)^2 & t > 3 \end{cases}$$

- where g is the treatment, $g = 1, 2, 3$

Model for the Covariance Matrix

- *Serial correlation + random intercept + measu. error*

$$\begin{aligned} var(\epsilon_{ij}) &= \sigma^2 + \tau^2 + v^2 = \sigma^2(1 + \alpha_1 + \alpha_2) \\ \rho(u) &= e^{-\alpha_3 u} \\ \gamma(u) &= \tau^2 + \sigma^2(1 - e^{-\alpha_3 u}) = \sigma^2(1 + \alpha_1 - e^{-\alpha_3 u}) \end{aligned}$$

- unknown parameters:

  mean response: $\beta_{01},\ \beta_{02},\ \beta_{03},\ \beta_1,\ \beta_2,\ \beta_3$
- covariance structure: $\sigma^2,\ \alpha_1,\ \alpha_2,\ \alpha_3$

Does the diet affect the mean response profile?

- $H_0 : \beta_{01} = \beta_{02} = \beta_{03} = 0$
- $H_0 : \boldsymbol{\phi} = D\boldsymbol{\beta}_0 = 0$ where

$$D = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$$

- $T_0 = \hat{\boldsymbol{\phi}}'(D\hat{V}_1 D')^{-1}\boldsymbol{\phi}' = 15.98$
- $P(X_2^2 > 15.98) = 0.0003$
- we reject H_0 and conclude that diet affect mean response profile

Is there a rise in the mean response towards the end of the experiment?

- $H_0 : \beta_2 = \beta_3 = 0,\ \boldsymbol{\beta} = (\beta_2, \beta_2 3$
- $T_0 = \hat{\boldsymbol{\beta}}\hat{V}_2^{-1}\boldsymbol{\beta}$
- $P(\chi_2^2 > 1.29) = 0.525$
- we accept the null hypothesis in favor of a late rise in the mean response
- consider:

$$\mu_g(t) = \begin{cases} \beta_{0g} + \beta_1 t & t \leq 3 \\ \beta_{0g} + 3\beta_1 & t > 3 \end{cases}$$

- The fit of the variogram appears satisfactory, less satisfactory is the fit of the mean response (see fig 5.6)
- however the lack of fit is toward the end of the experiment, by which time the responses from almost half the animals are missing
- here, time is measured in weeks since calving and experiment was terminated 19 weeks before calving. Therefore almost half of the animals are missing. This increase the variability in the observed mean responses
- *is calving date independent of the measurement process?*

- if later-calving cows are also more likely to produce milk with a lower protein content, we would expect the observed mean responses to rises towards the end of the study.

Body weight of 26 cows

Data consists of body weights of 26 cows, measured at 23 unequally spaced times over a period of about 22 months.

- Y_{ij} = log weight of cow i at time j (10 days intervals)
- $i = 1, \ldots, 26$
- $j = 1, \ldots, 23$
- the treatments were allocated in a 2×2 factorial design
  - Control (4)
  - Iron dosing (4)
  - Infection (9)
  - Iron + infection (10)
- Look your data
- Estimate empirical variogram

What do you see?

- Measurement variance small
- Substantial between cows variability
- Gaussian correlation model appropriate

Model for the mean

$$
\begin{array}{lll}
EY_{ij} & = & \underbrace{\mu_j}_{\text{control}} + \\
\text{if iron :} & + & (\beta_{01} + \beta_{11}(t_j - 33) + \beta_{21}(t_j - 33)^2) \\
\text{if infection :} & + & (\beta_{02} + \beta_{12}(t_j - 33) + \beta_{22}(t_j - 33)^2) \\
\text{if iron + infection :} & + & (\beta_{03} + \beta_{13}(t_j - 33) + \beta_{23}(t_j - 33)^2)
\end{array}
$$

- Each treatment contrast is a quadratic function of time
- control mean is described by a separate parameter at each of the 23 data points μ_j

Model for the covariance matrix

$$
\begin{array}{rcl}
V(Y_{ij}) & = & \sigma^2 + v^2 + \tau^2 \\
\gamma(u) & = & \sigma^2(1 - \rho(u)) + \tau^2 \\
 & = & \sigma^2(1 - \exp(-\alpha_3 u^2) + \alpha_1)
\end{array}
$$

where $\tau^2 = \alpha_1 \sigma^2$ and $v^2 = \alpha_2 \sigma^2$

$$
\begin{array}{rcl}
\hat{\sigma}^2 & = & 0.0016 \\
\hat{\alpha}_1 & = & 0.353 \\
\hat{\alpha}_2 & = & 4.099 \\
\hat{\alpha}_3 & = & 0.0045
\end{array}
$$

- *Aim 1: can we use linear growth instead of quadratic?*

$$H_0 : \beta_{21} = \beta_{22} = \beta_{23} = 0$$

the quadratic curve is appropriate

- *Aim 2: Is there a main effects for iron?NO*

$$H_0 : \beta_{01} = \beta_{11} = \beta_{21} = 0$$

- *Aim 3: Is there a main effects for infection? YES*

$$H_0 : \beta_{02} = \beta_{12} = \beta_{22} = 0$$

- *Aim 4: Is there an interaction between iron and infection? NO*

$$
\begin{array}{rl}
H_0 : & \beta_{03} = \beta_{01} + \beta_{02} \\
 & \beta_{13} = \beta_{11} + \beta_{12} \\
 & \beta_{23} = \beta_{21} + \beta_{22}
\end{array}
$$

We refit the model with only the infection term

$$\mu(t) = -0.167 - 0.00134(t - 33) + 0.0000566(t - 33)^3$$

Conclusions

- Highly significant effect of infection
- No significant effect of iron
- No significant effect of interaction