EXPLORATORY / DESCRIPTIVE ANALYSIS OF LONGITUDINAL DATA

Exploration of Correlation Structure

- Recall that the **correlation among repeated measures** is what makes longitudinal data so powerful
- We will want to incorporate the **correlation structure** into our models for longitudinal data
  - the correlation may be of **direct scientific interest** in the study (more emphasis in earlier text by Diggle et al)

    or
  - knowing the correlation (or being able to model it) will result in **increased statistical efficiency** in modelling trends in the mean (by optimally weighting the data; given greater emphasis in current text by Fitzmaurice et al)
  - **reducing bias in SE and obtaining correct CI and p-value**
- Preliminary analyses and (later on) model checking will require **tools** for exploring the correlation structure:
  - scatterplot and correlation matrix
  - autocorrelation function and correlogram
  - variogram

- Consider a general (**marginal**) (**working**) model

$$Y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \epsilon_{ij} \tag{1}$$

where $\boldsymbol{x}_{ij}$ contains important covariate effects, generally including time and any fixed covariates.
– No subject-specific random variable b_i in model
– Instead ϵ_{ij} are correlated within subject i.

- We are interested in studying the correlation structure among the **error terms**,

$$\text{corr}(\epsilon_{ij}, \epsilon_{ik}), \quad \text{for } j \neq k$$

where ϵ_{ij} and ϵ_{ik} are two error terms on the **same subject (i) at different points (t_{ij} and t_{ik}) in time**

- For exploration purpose, define the (exploratory) **residuals**

$$r_{ij} = \hat{\epsilon}_{ij} = Y_{ij} - \boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}} = \text{resp} - \widehat{\text{pred}}$$

where $\widehat{\boldsymbol{\beta}}$ is some quick-and-dirty estimator for $\boldsymbol{\beta}$ (e.g., OLS, ignoring longitudinal structure)

Generating Residuals from Model (1) and Examining Dispersion

- **Example:** Protein content of cows milk:

  To explore the correlation structure among protein, first remove the time trends for each treatment group, by subtracting mean for each time in each treatment group:

- R program example:

```
> # read data
> data=read.csv("cows.csv")
>
> # Compute residuals, removing effects of diet and time
> dietgrp=levels(data$diet)
> dietgrp
[1] "barley" "lupins" "mixed"
>
> week=levels(as.factor(data$week))
> week
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14" "15"
[16] "16" "17" "18" "19"
>
> # subtract mean for each combination of diet and week
> data$protrs=data$prot
> for(d in 1:length(dietgrp))
+ for(w in 1:length(week))
```

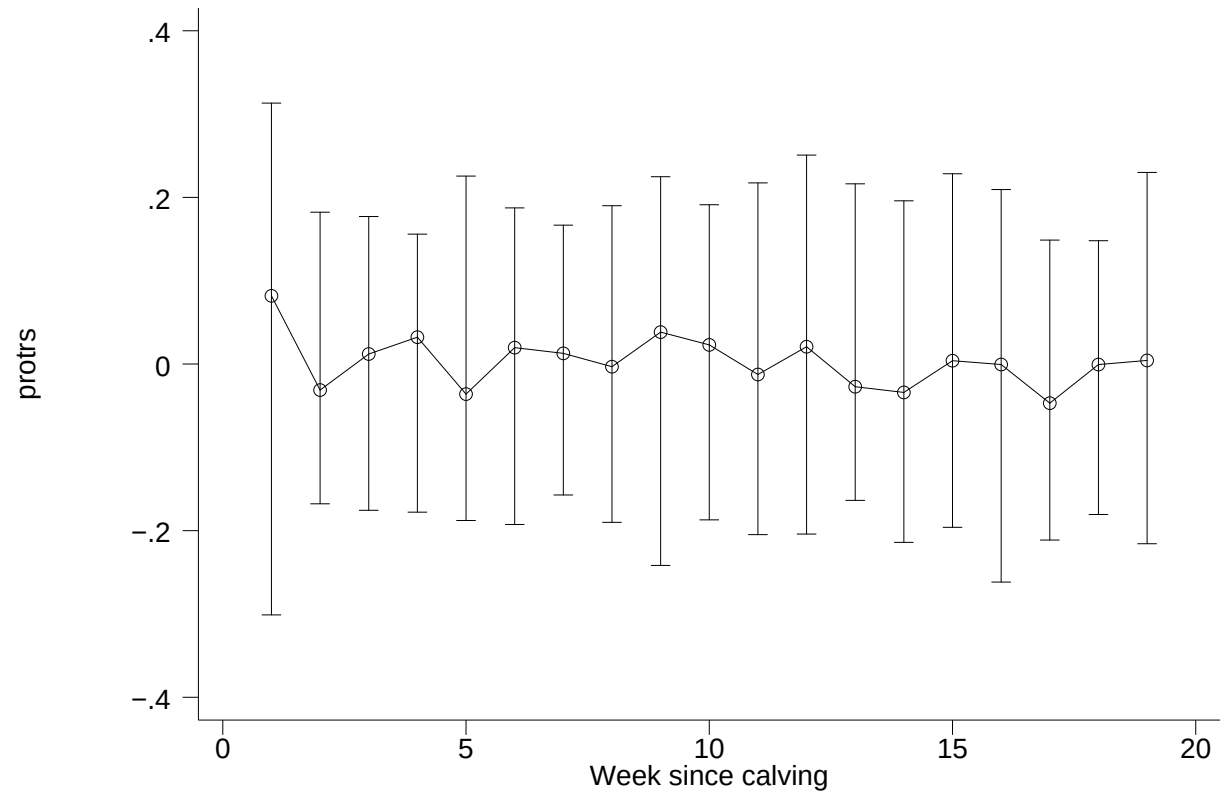
```
+ {
+ #select observations in this combination
+ index=which(data$diet==dietgrp[d] & as.factor(data$week)==week[w])
+ data$protrs[index]=data$prot[index]-mean(data[index,]$prot)#residual
+ }
```

- **Notes:**
  - Here the model (1) contains a dummy variable for each time point, each treatment group, and all two-way interaction terms
  - A simpler model could probably be used, but it is useful to be flexible during exploratory analyses
- As a preliminary step, we might examine the degree of dispersion (spread) among the residuals as a function of time

```
> ## create summary table for residual by week
>
> #Mean of residual by week
> mean.resid=aggregate(data$prot.resid, list(data$week), mean)
> #SD of residual by week
> sd.resid=aggregate(data$prot.resid, list(data$week), sd)
> #frequency of residual by week
> freq.resid=aggregate(rep(1,dim(data)[1]), list(data$week), sum)
>
> summ.resid=cbind(mean.resid,sd.resid[,2],freq.resid[,2])
> names(summ.resid)=c("week","mean","SD","freq")
> summ.resid
   week          mean        SD freq
1     1 -4.495939e-17 0.3974150   79
2     2 -6.263005e-17 0.2703849   78
3     3 -6.183589e-17 0.2547052   79
4     4 -4.496626e-17 0.2508063   79
5     5  3.416627e-17 0.3280103   78
6     6  5.058421e-17 0.2905210   79
7     7  1.845636e-16 0.2348227   77
8     8  6.344554e-17 0.3015270   77
9     9  5.189808e-17 0.2915378   77
10   10  0.000000e+00 0.2728563   78
11   11  4.555439e-17 0.3033748   78
12   12 -3.373777e-17 0.2805961   79
13   13 -7.401035e-17 0.2839899   78
14   14 -3.934427e-17 0.3160343   79
15   15 -1.505369e-16 0.3147629   59
16   16  3.553418e-17 0.3334893   50
```

```
17   17  2.896293e-17 0.2817822   46
18   18  0.000000e+00 0.2919066   46
19   19 -2.166338e-17 0.3046098   41
```

- Plot for median (and IQR) of residuals by week:



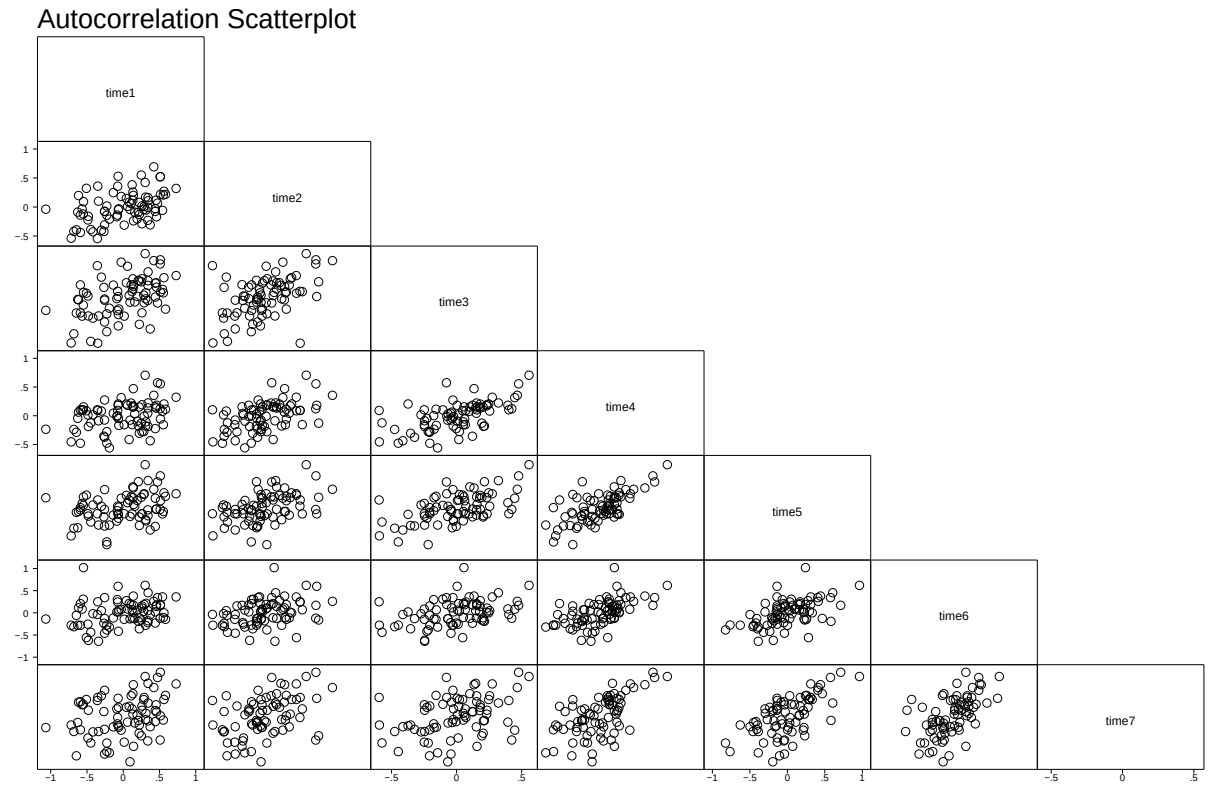
Notes:

- From the table and plot, appears as if the variance is fairly constant across time (except week 1)
- This somewhat unusual with longitudinal data . . . often we see variance increasing with time
- In most analyses, one would present **either** a table **or** a graphic, but not both
- In some data sets where time is more continuous, it might be necessary to group observation times in order to prepare such a table

  If data are grouped, it should only be for purposes of exploratory analysis
- Alternatively, the plot could be replaced with a scatterplot which shows the dispersion
- Also, if time is continuous, a scatterplot smoother might be required to remove time trends (here, we have just used a separate mean for each time point wihtin each treatment group)

Autocorrelation Scatterplot Matrix and Correlation Matrix

- Goal: examine graphically or otherwise the association across time between observations on the same subject
- For each **pair of** observation times, the residuals r_{ij} can be examined in a **scatterplot matrix**
- Here, we just examine the association among the first 7 time points for Cows data



Each point in this picture represents a pair of observations (`protrs` residuals) on the same subject in the original data set. E.g., the topmost cell plots week 2 vs. week 1

- We note several things:
  - the plots along the first diagonal, corresponding to a **time lag** of one week, show the strongest association
  - both association and information weaken as time lags get greater
  - Pearson's correlation coefficient would be a good summary measure of association for each plot

R program example:

```
> # keep week 1-7 and necessary variables only
> subset=data[data$week<=7,c(1,3,5)]
>
> # into wide format
> wide<-reshape(subset,v.names="protrs",idvar="id",timevar="week",direction="wide")
>
> # plot for Autocorrelation Scatterplot Matrix
> pairs(wide[,-1])
>
> # correlation matrix
> cor(wide[,-1],use="pairwise.complete.obs")
          protrs.1  protrs.2  protrs.3  protrs.4  protrs.5  protrs.6  protrs.7
protrs.1 1.0000000 0.4397052 0.4723762 0.3213910 0.3165455 0.2553003 0.2989087
protrs.2 0.4397052 1.0000000 0.4852995 0.5170111 0.4637478 0.3471221 0.4277654
protrs.3 0.4723762 0.4852995 1.0000000 0.6009765 0.5746046 0.3949390 0.3941880
protrs.4 0.3213910 0.5170111 0.6009765 1.0000000 0.6870871 0.5683348 0.6017598
protrs.5 0.3165455 0.4637478 0.5746046 0.6870871 1.0000000 0.5357291 0.6263782
protrs.6 0.2553003 0.3471221 0.3949390 0.5683348 0.5357291 1.0000000 0.5721362
protrs.7 0.2989087 0.4277654 0.3941880 0.6017598 0.6263782 0.5721362 1.0000000
```

- **Notes:**
  - As with the scatterplot matrix, we see that association is strongest with a **time lag** of one week, and weakens with greater lags (*across* diagonals)
  - As we read *along* a diagonal, lag is constant.Correlation tends to increase over time; this is not uncommon in longitudinal data analysis

Autocorrelation Function and Correlogram

- For now, let us ignore the differences in correlation along a diagonal, and make the working assumption that the residuals are **weakly stationary**, that is:
  - they have **constant mean** (over time and other factors).
    in this case: $\mathrm{E}(\epsilon_{ij}) = 0$.
    (easily satisfied since we have removed the time and treatment trends)
  - **constant variance**:

$$\mathrm{var}(\epsilon_{ij}) = \sigma^2$$

    (have checked this and does not appear unreasonable)
  - **correlation** that only depends on the **time lag**, $|t_{ij} - t_{ik}|$, between two residuals:

$$\mathrm{corr}(\epsilon_{ij}, \epsilon_{ik}) = \rho(u_{ijk}), \quad u_{ijk} = |t_{ij} - t_{ik}|$$

    (here we ignore differences in correlation later versus earlier in the study)

- Autocorrelation function (ACF) $\rho(u)$ is defined as

$$\rho(u) = \text{corr}(\epsilon_{ij}, \epsilon_{ik}), \quad u = |t_{ij} - t_{ik}|$$

- In this case, it makes sense to pool all of the **pairs of items** along each diagonal in the scatterplot matrix, and then compute the correlation for each pair. This results in in the (empirical) **autocorrelation function**:

- Computation of the empirical autocorrelation function

  For each time lag $u = 1, 2, \ldots$:

  1. Construct a data set of all pairs of observations (residuals) that are separated by lag u:

$$r_{ij} = \hat{\epsilon}_{ij}, r_{ik} = \hat{\epsilon}_{ik}, \quad \text{where} \quad |t_{ij} - t_{ik}| = u, \quad j < k$$

  2. Compute Pearson's sample correlation coefficient between the two observations for each observed lag u:

$$\widehat{\rho}(u) = \widehat{\text{corr}}(r_{ij}, r_{ik})$$

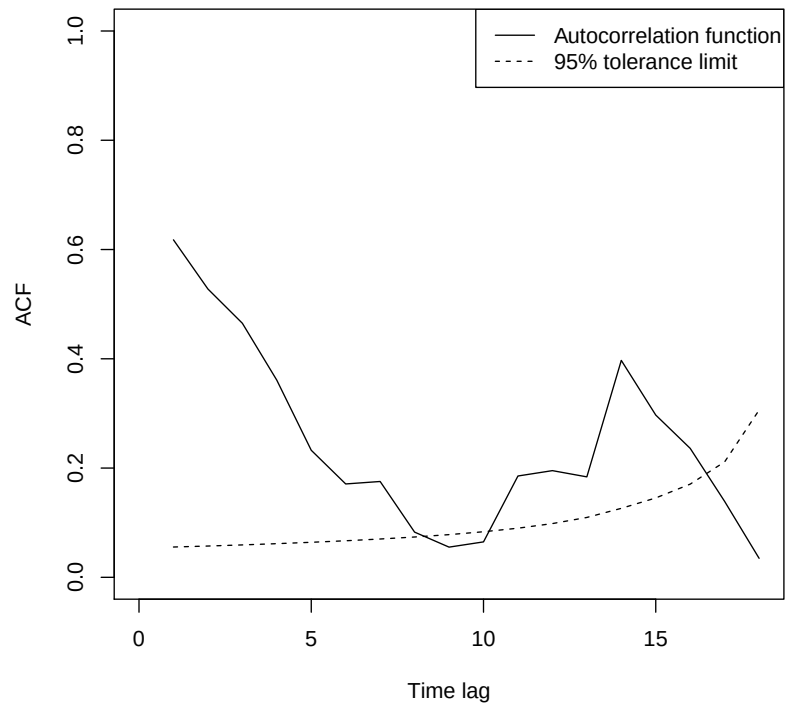
- A plot of the autocorrelation function versus the time lag is called a **correlogram**.

- **Tolerance limits** (TL) to complement the autocorrelation function:
  - Null hypothesis H_0: for a given lag u, the true correlation $\rho(u) = 0$
  - For a 95% TL: $\Pr(\widehat{\rho}(u)$ within $TL_{0.95}(u)|H_0) = 95\%$
  - 95% TL represents the bounds outside of which the empirical (observed) correlation will fall only 5% of chance, given null hypothesis of zero correlation is true
  - Correlations outside of the TL are significantly different than zero at the 5% level
- Calculation of **Tolerance limits**:
  - The simplest rule of thumb:
    A correlation coefficient r has standard error $\frac{1-r^2}{\sqrt{N}}$, where N is the number of independent pairs of observations in the calculation.
  - Under H_0 of no correlation, a correlation coefficient has standard error $\frac{1}{\sqrt{N}}$.
  - 95% Tolerance limits $= \frac{\pm 1.96}{\sqrt{N_u}}$, where N_u is the number of pairs of observations at lag u.

R program for autocorrelation function and correlogram:

```
> # extract all pairs within each subject
> data.pairs=list()
> for(i in 1:79)
+ {
+ subject.i=data[data$id==i,]
+ data.pairs.i=gtools::combinations(dim(subject.i)[1], 2,repeats=FALSE)#index of all possible pairs
+ data.pairs.add=data.frame(id=rep(i,dim(data.pairs.i)[1]),
+ week1=subject.i$week[data.pairs.i[,1]],week2=subject.i$week[data.pairs.i[,2]],
+ protrs1=subject.i$protrs[data.pairs.i[,1]],protrs2=subject.i$protrs[data.pairs.i[,2]]
+ )# all pairs within subject i
+ data.pairs=rbind(data.pairs,data.pairs.add)
+ }
> data.pairs$lag=abs(data.pairs$week2-data.pairs$week1)
>
> # Autocorrelation function and tolerance limit
>
> ACF=TL=rep(NA,18)
> for(lag in 1:18)
+ {
+ subset=data.pairs[which(data.pairs$lag==lag),4:5]
+ ACF[lag]=cor(subset)[1,2]
+ TL[lag]=1.96/sqrt(dim(subset)[1])
+ }
> summ.acf=cbind(lag=1:18,ACF,TL)
> summ.acf
      lag        ACF         TL
 [1,]   1 0.61775471 0.05548157
```

```
 [2,]   2 0.52777616 0.05727665
 [3,]   3 0.46503729 0.05933947
 [4,]   4 0.36084984 0.06161207
 [5,]   5 0.23253822 0.06413316
 [6,]   6 0.17091309 0.06691330
 [7,]   7 0.17543312 0.07013431
 [8,]   8 0.08273442 0.07381787
 [9,]   9 0.05528019 0.07815032
[10,]  10 0.06486405 0.08342314
[11,]  11 0.18548433 0.08993097
[12,]  12 0.19528698 0.09836958
[13,]  13 0.18387957 0.10956733
[14,]  14 0.39705065 0.12625470
[15,]  15 0.29701636 0.14528487
[16,]  16 0.23604609 0.17059610
[17,]  17 0.13893101 0.21135224
[18,]  18 0.03480314 0.30610057
>
> # correlogram
> plot(1:18,ACF,"l",xlab="Time lag",xlim=c(0,18),ylim=c(0,1))
> lines(1:18,TL,lty=2)
> legend("topright",c("Autocorrelation function","95% tolerance limit"),lty=1:2)
```


Autocorrelation function
95% tolerance limit
ACF
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
Time lag

- Interpretation of the correlogram:
  - As observations on the same subject are separated by greater lags (i.e., by greater amounts of time), two observations on the same individual are less alike, at least up to a lag of 10 weeks
  - There appears to be greater correlation around 14 week lag
    * some of this is probably noise
    * could represent unexplained factors operating on a 14-week cycle (e.g., delivery of fresh grain)

Variogram

- The sample autocorrelation function is useful when
  – time points are evenly-spaced, and
  – there are sufficient numbers of pairs (Y_{ij}, Y_{ik}) at each lag u
- When this is not the case (eg, CD4+ data)
  – need to group time points (eg, round to years)
  – **variogram** provides an alternative tool for exploring correlation structure
- The variogram function $\gamma(u)$ is defined as

$$\gamma(u) = \frac{1}{2}\mathrm{E}\left[\{\epsilon(t) - \epsilon(t-u)\}^2\right],$$

  where u is a positive lag, and t is any time
- given u, $\gamma(u)$ is constant for any t
- The variogram is also applicable to "weakly stationary" processes (like ACF)

- Recall, under **Weak stationarity**: Pick any u.
  For any t, the mean is constant:

$$\mathrm{E}\{\epsilon(t)\} = \mathrm{E}\{\epsilon(t-u)\} \, (= 0 \text{ in this case})$$

  variance is constant: $\mathrm{var}\{\epsilon(t)\} = \mathrm{var}\{\epsilon(t-u)\} = \sigma^2$

  within-subject correlation only depends on time lag u:

$$\mathrm{corr}\{\epsilon(t), \epsilon(t-u)\} = \rho(u)$$

- Under weak stationarity, relationship between variogram $\gamma(u)$ and ACF $\rho(u)$ is

$$\begin{aligned} \gamma(u) &= \frac{1}{2}\mathrm{E}\left[\{\epsilon(t) - \epsilon(t-u)\}^2\right] \\ &= \frac{1}{2}\left[2\sigma^2 - 2\sigma^2 \mathrm{corr}\{\epsilon(t), \epsilon(t-u)\}\right] \\ &= \sigma^2\{1 - \rho(u)\} \end{aligned}$$

- To satisfy the constant mean assumption, **variograms should be computed on** *residuals*, removing any effects of time and other important covariates

Computation of Sample Variogram

- Starting with the residuals $r_{ij} = \hat{\epsilon}_{ij}$ and the times t_{ij}, compute for all possible pairs within same subject:

$$v_{ijk} = \frac{1}{2}(r_{ik} - r_{ij})^2$$

and

$$u_{ijk} = t_{ik} - t_{ij}$$

for $j < k$

- Fit a model of v_{ijk} (dependent variable) on u_{ijk} (independent variable)

  Model examples:

  - ANOVA model with one parameter for each value of u (if number of possible u is small, eg, Cows data)
  - Smooth model fit (using, e.g., lowess) of v on u (eg, CD4+ data)

- Note: To estimate $\mathrm{var}\{\epsilon(t)\} = \sigma^2$, use the same way that we obtained $\widehat{\sigma}_t^2$ (half the average of between-subject squared differences)

$$\widehat{\sigma}^2 = \frac{1}{2}\text{average}\{(r_{ij} - r_{i'k})^2\}$$

over $i = 1, \ldots, m-1,\ i' = i+1, \ldots, m, j = 1, \ldots, n_i, k = 1, \ldots, n_{i'}$,
That is,

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{m-1}\sum_{i'=i+1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_{i'}}\{(r_{ij} - r_{i'k})^2\}}{2\sum_{i=1}^{m-1}\sum_{i'=i+1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_{i'}} 1}$$

Example: Protein content of cows' milk:

- (ANOVA model) There are 19 time points $\rightarrow$ 18 different lags. We can fit a discrete-lag variogram model for each of the 18 different lags:

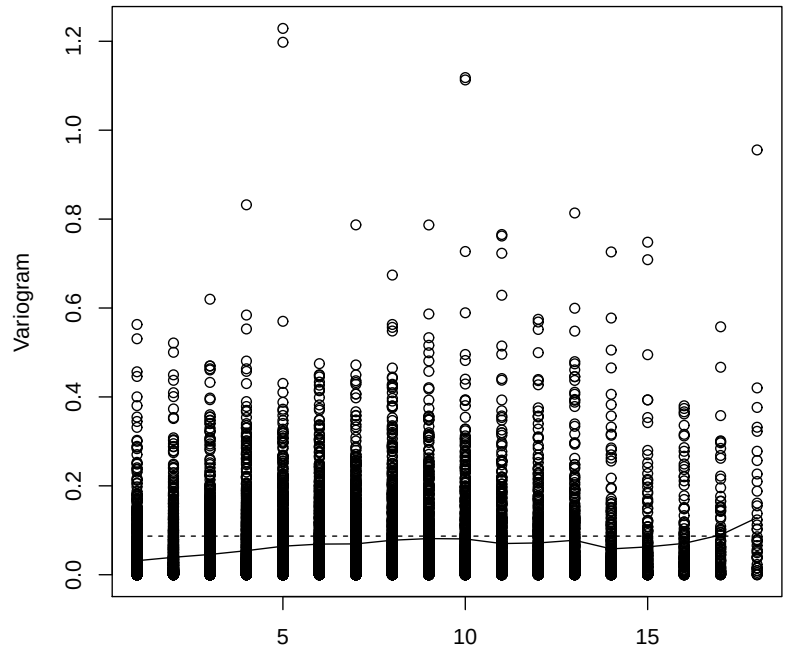
$$v_{ijk} = \alpha_{u_{ijk}} + e_{ijk}$$

and estimated variogram $\hat{\gamma}(u) = \hat{\alpha}_u$ is predicted value at each time lag u.

- R program example:

```
> data.pairs$v=(data.pairs$protrs1-data.pairs$protrs2)^2/2
>
> # ANOVA model
> model.anova=lm(v~as.factor(lag)-1,data=data.pairs)
> pred.anova=predict(model.anova,newdata=data.frame(lag=as.factor(1:18)))
>
> # variance
> numerator=0
> denominator=0
> for(i in 1:78)# sum up all possible pairs
+ {
+ subset.i=data[data$id==i,]
+ subset.i2=data[data$id>i,]
+ for(j in 1:dim(subset.i)[1])
+ numerator=numerator+sum((subset.i$protrs[j]-subset.i2$protrs)^2)
+ denominator=denominator+dim(subset.i)[1]*dim(subset.i2)[1]
+ }
```

```
> sigma2=numerator/denominator/2
> sigma2
[1] 0.08687206
>
> # Variogram
> plot(data.pairs$lag,data.pairs$v,xlab="Time lag",ylab="Variogram")
> lines(1:18,pred.anova)
> lines(1:18,rep(sigma2,18),lty=2)
```

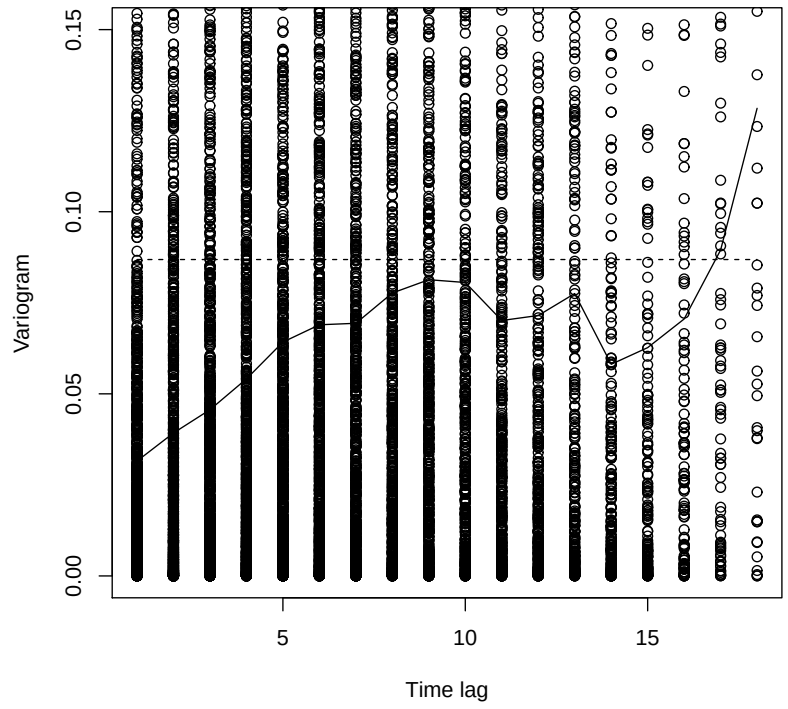


Note:

- The horizontal line is $\widehat{\sigma}^2 = 0.087$
- The points are the values of (v_{ijk}, u_{ijk}). Some of the v_{ijk}'s are very large, obscuring detail (this is normal)

- We can avoid plotting the large v_{ijk}'s

```
> # plot small v only
> plot(data.pairs$lag,data.pairs$v,ylim=c(0,0.15),xlab="Time lag",ylab="Variogram")
> lines(1:18,pred.anova)
> lines(1:18,rep(sigma2,18),lty=2)
```



- (Smooth fit) Instead of fitting a separate mean for each of the 18 lags, we can fit a smooth lowess function through the lags:

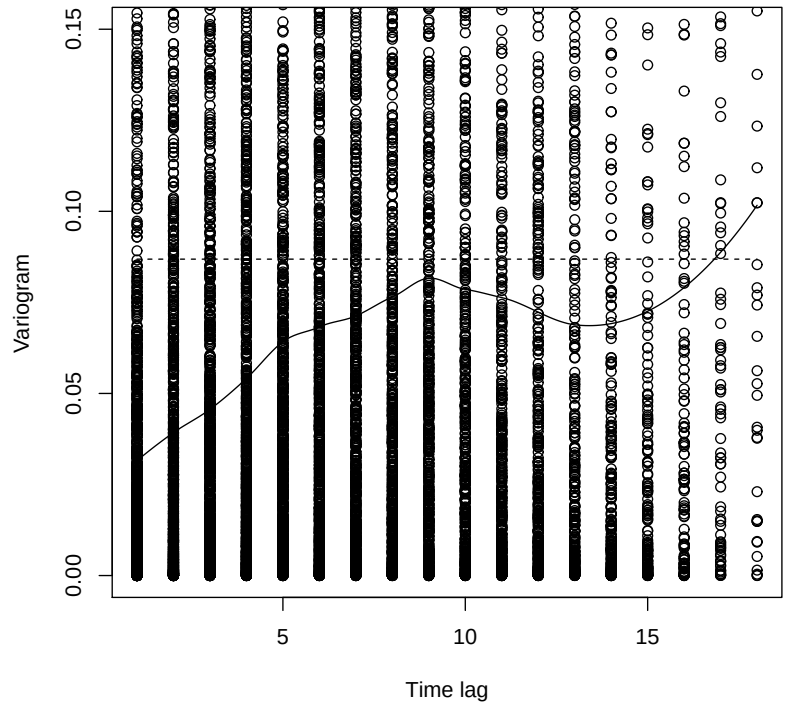
$$v_{ijk} = \mu(u_{ijk}) + e_{ijk},$$

  where $\mu()$ is a smooth function of time lag u obtained by lowess.

- R program:

```
# Lowess
model.lowess=loess(v~lag,data=data.pairs,span = 0.4)
pred.lowess=predict(model.lowess,newdata=data.frame(lag=(1:180)/10))

# Variogram
plot(data.pairs$lag,data.pairs$v,ylim=c(0,0.15),xlab="Time lag",ylab="Variogram")
lines((1:180)/10,pred.lowess)
lines(1:18,rep(sigma2,18),lty=2)
```


Variogram
0.00
0.05
0.10
0.15
5
10
15
Time lag

Interpretation of Variogram

- Recall that the variogram $\gamma(u)$ is related to the autocorrelation function $\rho(u)$ by

$$\gamma(u) = \sigma^2\{1 - \rho(u)\}$$

  which can be re-written as

$$\rho(u) = 1 - \gamma(u)/\sigma^2$$

- Therefore:
  - When $\widehat{\gamma}(u)$ is close to 0, autocorrelation is close to 1
  - When $\widehat{\gamma}(u)$ is increasing, autocorrelation is decreasing
  - When $\widehat{\gamma}(u)$ is close to $\widehat{\sigma}^2$, autocorrelation is close to 0
  - A gap between $\widehat{\gamma}(u)$ and $\widehat{\sigma}^2$ for large u indicates positive autocorrelation even at largest lag

Summary

- Three features of variance and correlation often seen in longitudinal data:
  - variance of observations increases with time ("fanning out")
  - correlation decreases with time lag separating observations
  - correlation increases (for a given lag) with mean time of two observations
- These principle deviations from standard OLS assumptions:
  We need to account for correlation and heteroscadisticity in analytical approaches to longitudinal data

Why we care about correlation and where does it come from?

- Empirical observations about correlation in many longitudinal studies:
  - correlations are positive
  - correlations decrease with time lag separating them
  - correlations do not appear to approach zero even for large time lags
  - correlations do not appear to approach one even for very small lags
  - correlations at a given lag increase as the study progresses

- Correlation among repeated measures on the **same individual** reflects underlying (unobserved) processes that are **not** accounted for in the (mean) model
  - all observations on the same subject share the same underlying processes
  - hence, observations on the same subject are more alike than observations on different subjects, yielding positive correlation between observations on the same subject
  - e.g., milk protein content partly reflects a characteristic (trait) of each individual cow
- Therefore, correlation reflects the **strength of contribution** of underlying processes to observed phenomena
  – Eg, higher correlation of weights within nepalese girls: reflects stronger contribution from underlying factor (eg, gene, diet)
- Also, the strength of correlation may vary with time lag separating observations
  - degree to which underlying **time-varying** processes contribute to observed phenomena
  - observations closer together on a given person are more correlated than are observations further apart in time
  - e.g., protein content may reflect amount and quality of grain eaten; two

observations close together in time (on same individual) are similar because they are taken under similar circumstances of nourishment

- Three contributing sources to observed patterns of correlation:
  - **between-individual heterogeneity**: some subjects are simply, on average, higher than others $\rightarrow$ positive correlation, even at long lags
  - **within-individual biological variability**: underlying biological processes change through time in smooth / continuous fashion $\rightarrow$ measurements close in time are more correlated than those far apart in time
  - **measurement error**: most measures in public health and bio-medical science are subject to error $\rightarrow$ even observations close together in time are not perfectly correlated

- **Statistical reasons** to study correlation:
  - e.g., suppose Y_1 is a measure before treatment and Y_2 a measure after treatment, $\mathrm{var}(Y_1) = \mathrm{var}(Y_2) = \sigma^2$
  - interest is on

$$\Delta = \mu_2 - \mu_1 = \mathrm{E}(Y_2) - \mathrm{E}(Y_1) = \text{treatment difference}$$

  - Estimated by difference of sample means $\hat{\Delta} = \bar{y}_2 - \bar{y}_1$
  - if data are **independent** (i.e., one set of subjects before treatment and a different set after treatment):

$$\begin{aligned} \mathrm{var}(\hat{\Delta}) &= \mathrm{var}(\bar{Y}_2) + \mathrm{var}(\bar{Y}_1) \\ &= \frac{2}{n}\sigma^2 \end{aligned}$$

    where n is the number of subjects at each time point
  - if data are **correlated** (i.e., as single set of individuals measured longitudinally both before and after treatment):

$$\mathrm{var}(\hat{\Delta}) = \mathrm{var}(\bar{Y}_2) + \mathrm{var}(\bar{Y}_1) - 2\mathrm{cov}(\bar{Y}_2, \bar{Y}_1)$$

$$
\begin{aligned}
&= \frac{2}{n}\sigma^2 - 2\mathrm{cov}(\frac{\sum_i Y_{i1}}{n}, \frac{\sum_i Y_{i2}}{n}) \\
&= \frac{2}{n}\sigma^2 - \frac{2}{n}\sigma^2 \mathrm{corr}(Y_{i1}, Y_{i2}) \\
&= \frac{2}{n}\sigma^2(1-\rho)
\end{aligned}
$$

where $\mathrm{corr}(Y_{i1}, Y_{i'2}) = 0$ for two different subjects,

within-subject correlation $\rho = \mathrm{corr}(Y_{i1}, Y_{i2})$

- The latter is smaller if $\rho > 0$!

- When studying **contrasts within individuals**, positive correlation helps
- More generally: Correlated observations provide more precise estimates of rates of change than would be obtained from independent observations from different individuals.