ORIGIN := 0

Adding Autocorrelation Structures to Linear and Linear Mixed Models

Linear Models require the fundamental assumption of independence between each instance of the response variable, or equivalently, of all residuals. As part of the construction of a model, therefore, validation plots should be constructed to look for patterns in residuals versus each independent covariate or factor as well as with the model's fitted values. If patterns occur, then remedial measures are necessary to ensure proper P-values of associated the tests. The concept of independence typically enters into formal descriptions of Linear Fixed Models as: $\varepsilon_i \sim N(0,\sigma^2)$ = "residuals (ε) are Normally distributed with mean 0 and variance σ^2 for each observation i. In terms of matrix algebra the above may be expanded into: $\varepsilon_i \sim N(0,\Sigma)$ where Σ is the i by i square covariance matrix, with $\Sigma=\sigma^2 I$, and I being the identity matrix. Equal variances (σ^2) occur along the main diagonal with pairwise covariances, equal to zero, everywhere else.

For instance, for i = 3 observations, *covariance matrix* Σ = $\begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix}$ **$= \sigma^2 I$, where I =** $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

Dividing all entries of the matrix by σ^2, yields the corresponding correlation matrix P = $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

As one can see from the above definitions, this fundamental assumption of Linear Models is a stringent requirement not often met by real data. Mixed Linear Models therefore seek to relax this requirement by allowing observations from Groups to be correlated. In matrix terms, now using i = 3 observations *per group j*, this translates into:

$\Sigma_j =$ $\begin{pmatrix} \sigma_d^2 + \sigma^2 & \sigma_d^2 & \sigma_d^2 \\ \sigma_d^2 & \sigma_d^2 + \sigma^2 & \sigma_d^2 \\ \sigma_d^2 & \sigma_d^2 & \sigma_d^2 + \sigma^2 \end{pmatrix}$ **where all covariances have the same value σ_d^2 variance of all observations = $\sigma_d^2 + \sigma^2$, and** $P_j =$ $\begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$

The term: $\rho_j = \dfrac{\sigma_d^2}{\sigma_d^2 + \sigma^2}$ **is called the *Intraclass correlation coefficient* and refers to the correlation between observations with each group. In addition, variance within each group can be seen to be comprised to two components: σ_d^2 and σ^2. In Mixed Linear Models, σ_d^2 refers to "between"variance attributed to the random differences between the groups, whereas σ^2 refers to (pooled) variance within groups".**

This partial relaxation of the requirement of independence within Mixed Linear Model is called the *Compound Symmetric* Correlation Structure, and is but one of several such structures in current use. Correlation structures are often employed to correct for patterns of non-independence observed in validation plots from models of messy real data.

Shown in this Worksheet are several other Correlation Structures implemented by gls(), lme() and other functions of the {nlme} package in R. Background mathematics can be found in Chapter 5 of Pinheiro & Bates 2004 (PB), *Mixed-Effects Models in S and S-Plus.* Simplified presentation with the following example comes from Ch. 6 of Zuur et al. 2009 (Za), *Mixed Effects Models and Extensions in Ecology with R.* In the latter, emphasis is placed on "Temporal Autocorrelation". The idea here is that successive observations in a time series (or analogous series repeated measures in a spatial context) mostly occur at regularly spaced intervals exhibiting "stationarity" (p. 145). "Stationarity" implies that correlations are modeled only in terms of relative positions t, t+1, t+2 etc, in a time series, but are unrelated to overall position in the series. From this, multiple ways have been developed to convert ρ in P above into different ρ's depending on position within P.

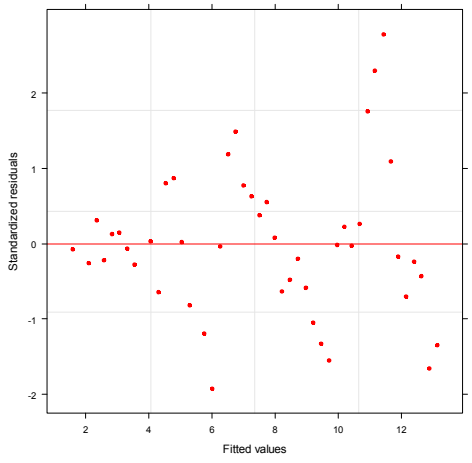
```
library(nlme)
setwd("c:/DATA/Models")
H=read.table("Hawaii.txt",header=T)
H$Birds=sqrt(H$Moorhen.Kauai)
H
```

Following Za, the square root of Moorhen.Kauai counts are placed in "Birds" and used as the response variable. "Rainfall" and "Year" (both covariates) are the independent variables.

```
> H
   Year Stilt.Oahu Stilt.Maui Coot.Oahu Coot.Maui Moorhen.Kauai Rainfall     Birds
1  1956        163        169       528       177             2    15.16  1.414214
2  1957        272        190       338       273            NA    15.48        NA
3  1958        549        159       449       256             2    16.26  1.414214
4  1959        533        211       822       170            10    21.25  3.162278
5  1960         NA        232        NA       188             4    10.94  2.000000
6  1961        134        155       717       149            10    19.93  3.162278
7  1962        175        282        12       205            12    12.63  3.464102
8  1963        356        170       169       108            10    20.09  3.162278
9  1964        485        164        98        79             8    10.02  2.828427
10 1965        184        162       112        53            NA    30.91        NA
11 1966        242        253        77        75            17    12.49  4.123106
12 1967        209        188       106        80             7    32.70  2.645751
13 1968        175        226        64       104            44    18.40  6.633250
14 1969        162        171        15       122            50    25.21  7.071068
15 1970        322        207       130        78            26    18.61  5.099020
16 1971        362        189        84        76            10    20.13  3.162278
17 1972        342        274       132       161            NA    15.71        NA
18 1973        509        352       122       114             7    10.27  2.645751
19 1974        107        245        41        63             1    18.68  1.000000
20 1975        110        340        83       151            38    13.74  6.164414
21 1976        249        335       148       177            92    12.83  9.591663
22 1977        738        302       148       154           113    11.50 10.630146
...
33 1988        478        296       250        57            36    26.79  6.000000
34 1989        740        215       542        92            32    40.63  5.656854
35 1990        515        172       415        41            98    35.20  9.899495
36 1991        371        197       261        61           116    16.09 10.770330
37 1992        562        366       450       175           107    16.98 10.344080
38 1993        393        284       339        90           129    12.69 11.357817
39 1994         NA        212        NA        68           240    13.93 15.491933
40 1995        598        191       303        57           294    13.45 17.146428
41 1996        530        297       214        87           348    31.00 18.654758
42 1997        558        262       317       113           210    23.08 14.491377
43 1998        502        424       899       396           131     6.76 11.445523
44 1999        442        364      1004       384           106     9.66 10.295630
45 2000        499        526       634       422           138     9.72 11.747340
46 2001        620        337       668       538           132    10.53 11.489125
47 2002        692        297       589       234            73    15.07  8.544004
48 2003        240        258       885        85            92    13.83  9.591663
```

Constructing and Validating a Linear Fixed Model:

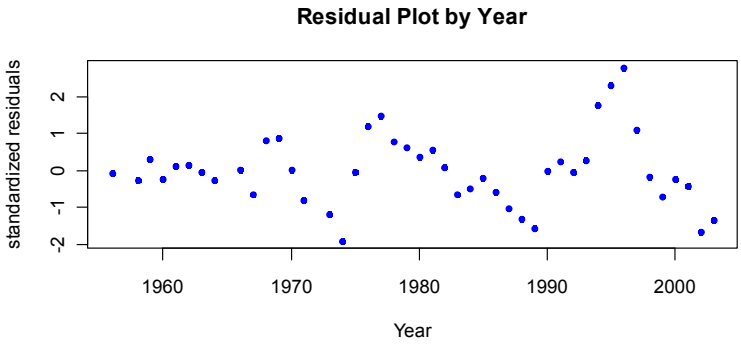
```
#LINEAR FIXED MODEL
LMgls=gls(Birds~Rainfall+Year,na.action=na.omit,data=H)
#VALIDATION PLOT
plot(LMgls,col='red',pch=20)
```

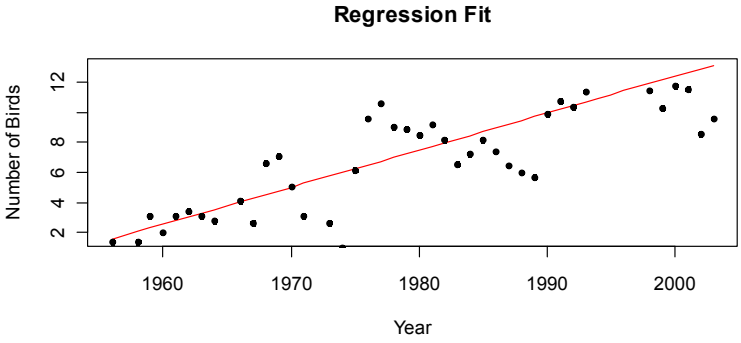



```
#HOMEMADE VALIDATION PLOTS
op=par(mfrow = c(2,1))
plot(H$Year[!is.na(H$Birds)],resid(LMgls,type="pearson"),
        col="blue",pch=20,xlab="Year",ylab="standardized residuals",
        main="Residual Plot by Mean-Centered Year")
plot(H$Year[!is.na(H$Birds)],fitted(LMgls),
        col="red",type='l',xlab="Year",ylab="Number of Birds",
        main="Regression Fit")
points(H$Year[!is.na(H$Birds)],H$Birds[!is.na(H$Birds)],
        col="black",pch=20)
par(op)
```

The plots show clear non-randomness and successive correlation by Year & fit indicative of autocorrelation.

A regular pattern of the residuals around the regression fit (shown in red) suggest that additional unaccounted for independent variable(s) may be involved. At the very least, observations in successive years appear to be correlated.

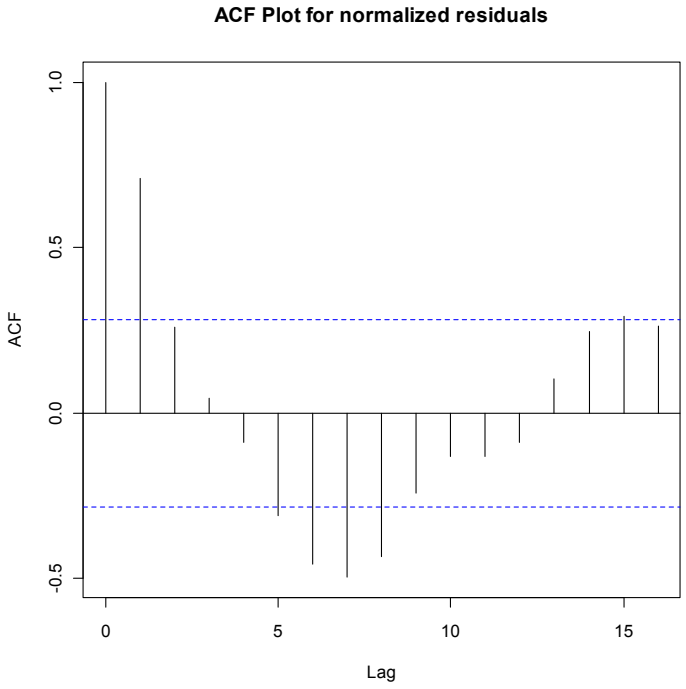




ACF Plot:

Auto Correlation Function (ACF) plots provide another graphical tool for judging autocorrelation. The "Lag" indicates number of constant intervals (i.e., offset) in the time series over which correlation is calculated. Because of the assumption of "stationarity", all Lags (gaps) between measurements are considered equivalent regardless of where the gaps occur in the time series.

```
#AUTOCORRELATION (ACF) PLOT:
E=residuals(LMgls,type="normalized")
I=!is.na(H$Birds)
Ea=vector(length=length(H$Birds))
Ea=NA
Ea[I]=E
acf(Ea,na.action=na.pass,
   main="ACF Plot for normalized residuals")
```



Strong positive correlation is observed over Lags 1-3, with negative correlation over Lags 4-12, followed by positive correlation in larger Lags. Perhaps there's evidence for longer-term peridicity in addition to short-term association here, but this will not be considered in what follows.

Adding Correlation Structures in gls():

Zuur et al.'s results match the anova() and AIC/BIC findings here. However, Za do not show paremeter estimates from summary(). Displaying the summary results shows a high negative (-1) correlation between Intercept and Rainfall and engenders similar strange results in the models that follow. Therefore, the initial Linear Fixed model will be refit using mean-centered independent variables. As a result, values of the Intercept estimate changes (with minor differences in slopes), but results of the marginal t-tests remain the same.

```
#OBSERVED CORRELATION IN LINEAR FIXED MODEL:
summary(LMgls)
#MEAN-CENTERING INDEPENDENT VARIABLES:
H$Year=H$Year-mean(H$Year)
H$Rainfall=H$Rainfall-mean(H$Rainfall)
#REFITTING LINEAR FIXED MODEL:
LMgls=gls(Birds~Rainfall+Year,na.action=na.omit,data=H)
summary(LMgls)
```

before mean centering data:

```
> summary(LMgls)
Generalized least squares fit by REML
  Model: Birds ~ Rainfall + Year
  Data: H
       AIC      BIC    logLik
  228.4798 235.4305 -110.2399

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) -477.6634  56.41907 -8.466346  0.0000
Rainfall       0.0009   0.04989  0.017245  0.9863
Year           0.2450   0.02847  8.604858  0.0000

 Correlation:
         (Intr) Ranfll
Rainfall -0.036
Year     -1.000  0.020

Standardized residuals:
        Min          Q1         Med          Q3
   Max
-1.91985793 -0.58712230 -0.03223775  0.38320859
2.77801077

Residual standard error: 2.608391
Degrees of freedom: 45 total; 42 residual
```

after mean-centering data:

```
> summary(LMgls)
Generalized least squares fit by REML
  Model: Birds ~ Rainfall + Year
  Data: H
       AIC      BIC    logLik
  228.4798 235.4305 -110.2399

Coefficients:
               Value Std.Error   t-value p-value
(Intercept) 7.355073 0.3899178 18.863136  0.0000
Rainfall    0.000860 0.0498915  0.017245  0.9863
Year        0.245013 0.0284738  8.604858  0.0000

 Correlation:
         (Intr) Ranfll
Rainfall  0.018
Year     -0.072  0.020

Standardized residuals:
        Min          Q1         Med          Q3
     Max
-1.91985793 -0.58712230 -0.03223775  0.38320859
2.77801077

Residual standard error: 2.608391
Degrees of freedom: 45 total; 42 residual
```

Compound Symmetry:

```
#CORRELATION STRUCTURES:
#COMPOUND SYMMETRY = NESTED lme():
LMcgls1=gls(Birds~Rainfall+Year,na.action=na.omit,data=H,
  correlation=corCompSymm(form =~Year))
summary(LMcgls1)
LMe1=lme(Birds~Rainfall+Year,na.action=na.omit,data=H,
        random=~1|Year)
summary(LMe1)
anova(LMgls,LMcgls1)
anova(LMgls,LMe1)
```

< compound symmetric correlation in specified by corCompSymm() in gls()

< compound symmetric correlation implicit in lme()

"Compound Symmetric" correlation structure is the term used to connote a single ρ for all off diagonal correlations in P. This correlation structure is also imposed by Linear Mixed Models using lme().

> summary(LMcgls1)

```
Generalized least squares fit by REML
  Model: Birds ~ Rainfall + Year
  Data: H
       AIC      BIC    logLik
  230.4798 239.1682 -110.2399

Correlation Structure: Compound symmetry
 Formula: ~Year
 Parameter estimate(s):
          Rho
3.392348e-18

Coefficients:
               Value Std.Error   t-value p-value
(Intercept) 7.355073 0.3899178 18.863136  0.0000
Rainfall    0.000860 0.0498915  0.017245  0.9863
Year        0.245013 0.0284738  8.604858  0.0000

 Correlation:
         (Intr) Ranfll
Rainfall  0.018
Year     -0.072  0.020

Standardized residuals:
        Min          Q1         Med          Q3
Max
-1.91985793 -0.58712230 -0.03223775  0.38320859
2.77801077

Residual standard error: 2.608391
Degrees of freedom: 45 total; 42 residual
```

> summary(LMe1)

```
Linear mixed-effects model fit by REML
 Data: H
       AIC      BIC    logLik
  230.4798 239.1682 -110.2399

Random effects:
 Formula: ~1 | Year
        (Intercept)  Residual
StdDev:    2.442313 0.9158673

Fixed effects: Birds ~ Rainfall + Year
               Value Std.Error DF   t-value p-value
(Intercept) 7.355073 0.3899178 42 18.863136  0.0000
Rainfall    0.000860 0.0498915 42  0.017245  0.9863
Year        0.245013 0.0284738 42  8.604858  0.0000
 Correlation:
         (Intr) Ranfll
Rainfall  0.018
Year     -0.072  0.020

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3
   Max
-0.67410712 -0.20615240 -0.01131943  0.13455352
0.97542470

Number of Observations: 45
Number of Groups: 45
```

^ Similar results are obtained, although note that ρ = 0 in the summary from gls().

Variance structure:

$$\Sigma_j = \begin{pmatrix} \sigma_d^2 + \sigma^2 & \sigma_d^2 & \sigma_d^2 \\ \sigma_d^2 & \sigma_d^2 + \sigma^2 & \sigma_d^2 \\ \sigma_d^2 & \sigma_d^2 & \sigma_d^2 + \sigma^2 \end{pmatrix}$$

where all covariances have the same value σ_d^2 variance of all observations = $\sigma_d^2 + \sigma^2$, and

$$P_j = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$

In gls(), ρ_j is estimated to be 0 and fixed estimates are the same as the Linear Fixed Model (LMgls) indicating little has been gained by adding Compound Symmetry correlations.

In lme(), estimated variance components and ρ can be calculated from the Random Factor Standard Deviations:

$\sigma_d^2 = (2.442313)^2$

$\sigma^2 = (0.9158673)^2$

and intraclass correlation ρ = $\dfrac{2.442313^2}{\left(2.442313^2 + 0.9158673^2\right)} = 0.877$

These two estimates of ρ obviously do not match, and I haven't yet been able to determine why. My guess is the difference may in part be due to reporting, but may also relate to subtle differences in the methods employed by the two function that blow up when little of value can be estimated from the data. Za only conclude that this correlation structure makes "no improvements in the model" (p. 149.)

Anova results are identical and match Za p. 149.

>

Only slight AIC support is seen for a model containing compound symmetry.

> anova(LMgls,LMcgls1)

```
        Model df      AIC      BIC    logLik   Test L.Ratio p-value
LMgls       1  4 228.4798 235.4305 -110.2399
LMcgls1     2  5 230.4798 239.1682 -110.2399 1 vs 2       0       1
```

> anova(LMgls,LMe1)

```
      Model df      AIC      BIC    logLik   Test L.Ratio p-value
LMgls     1  4 228.4798 235.4305 -110.2399
LMe1      2  5 230.4798 239.1682 -110.2399 1 vs 2       0       1
```

AR-1 Autocorrelation:

AR-1 Autocorrelation weights the off main-diagonal correlations according to the integer difference between elements (i.e., the Lag) in the time series $\rho^{|i-j|}$, where i & j are indices marking different locations in the sequence.

For a sequence of 4 elements (maximum difference of 3), the correlation matrix is the following:

$$P_j = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

Compared with Compound Symmetry, with AR-1 we now we have different correlations between elements depending on how big the Lag between observations happens to be. Because $|\rho|$ is typically < 1, $\rho^3 < \rho^3 < \rho$ etc., and correlation becomes vanishingly small with higher powers.

```
#AR-1 CORRELATION:
LMcgls2<-gls(Birds~Rainfall+Year,na.action=na.omit,data=H,
  correlation=corAR1(form =~ Year))
summary(LMcgls2)
anova(LMgls,LMcgls2)
```

Here we see major improvement to the model as judged by AIC/BIC and the formal Liklihood Ratio test.

Values for AIC & BIC are similar, but not identical, to that reported by Zuur et al. presumably due to estimations on mean-centered data here.

Values of estimates for the Fixed factors show subtle differences from Linear Fixed Model and from the results of lme().

Autocorrelations are reported by:

Lag 1 ϕ = 0.6698197

Lag 2 ϕ^2 = (0.6698197)^2

Lag 3 ϕ^3 = (0.6698197)^3

…

```
> summary(LMcgls2)
Generalized least squares fit by REML
  Model: Birds ~ Rainfall + Year
  Data: H
       AIC      BIC    logLik
  200.4019 209.0903 -95.20097

Correlation Structure: ARMA(1,0)
 Formula: ~Year
 Parameter estimate(s):
     Phi1
0.6698197

Coefficients:
               Value Std.Error   t-value p-value
(Intercept)  7.169698 0.7891140  9.085757  0.0000
Rainfall    -0.018731 0.0345885 -0.541550  0.5910
Year         0.236668 0.0524666  4.510844  0.0001

 Correlation:
         (Intr) Ranfll
Rainfall  0.021
Year      0.249 -0.015

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.26298182 -0.60129562 -0.01286182  0.63966754  3.63725014
```

```
> anova(LMgls,LMcgls2)
        Model df      AIC      BIC     logLik   Test  L.Ratio p-value
LMgls       1  4 228.4798 235.4305 -110.23991
LMcgls2     2  5 200.4019 209.0903  -95.20097 1 vs 2 30.07787  <.0001
```

In general, Za repeat advice from several sources that suggests little value in spending much time finding a "best fit" correlation structure for real problems. Many will suffice with only minor differences seen in AIC/BIC levels and in model estimates. Therefore, it seems that AR-1 correlation is probably sufficient for most purposes here. However, a popular alternative correlation structure is given below.

ARMA Correlation Structures:

Auto-Regressive Moving Average (ARMA) correlation structures are briefly described in Za along with an example of their use. ARMA is clearly more sophisticated than AR-1, involving both p "autoregressive" and q "moving average" parameters. Each must be specified and, due to convergence problems in software, should probably not exceed 3 for either p or q. Initial estimates between (-1,1) for all p & q also need to be supplied and these may need to be played with to find software convergence. Keeping in mind that exact models of correlation often matter little, Za describe ARMA as a "black box" from which a reasonable correlation structure might be found that will suffice with little further justification. Because alternate ARMA models do not internest, relative value of model fits should be judged by AIC or BIC.

```
#ARMA CORRELATION:
LMform=formula(Birds~Rainfall+Year)
LMcgls3=gls(LMform,na.action=na.omit,data=H,
  correlation=corARMA(c(0.2),p=1,q=0))
LMcgls4=gls(LMform,na.action=na.omit,data=H,
  correlation=corARMA(c(0.3,-0.3),p=2,q=0))
LMcgls5=gls(LMform,na.action=na.omit,data=H,
  correlation=corARMA(c(0.3,-0.3,0.2),p=2,q=1))
LMcgls6=gls(LMform,na.action=na.omit,data=H,
  correlation=corARMA(c(0.3,-0.3,0.3,-0.3),p=2,q=2))
```

```
AIC(LMcgls3,LMcgls4,LMcgls5,LMcgls6)
summary(LMcgls4)
anova(LMgls,LMcgls4)
```

```
> AIC(LMcgls3,LMcgls4,LMcgls5,LMcgls6)
         df      AIC
LMcgls3   5 197.7546
LMcgls4   6 194.5268
LMcgls5   7 196.5096
LMcgls6   8 198.0618
```

< slight preference is seen for model LMcgls4 based on AIC

```
> summary(LMcgls4)
Generalized least squares fit by REML
  Model: LMform
  Data: H
       AIC      BIC    logLik
  194.5268 204.9528 -91.2634

Correlation Structure: ARMA(2,0)
 Formula: ~1
 Parameter estimate(s):
      Phi1       Phi2
 0.9923617 -0.3567427

Coefficients:
               Value Std.Error   t-value p-value
(Intercept)  7.319921 0.6893037 10.619297  0.0000
Rainfall    -0.017918 0.0268054 -0.668465  0.5075
Year         0.240154 0.0484653  4.955172  0.0000

 Correlation:
         (Intr) Ranfll
Rainfall  0.012
Year     -0.066  0.013

Standardized residuals:
        Min          Q1         Med          Q3         Max
-1.86109211 -0.52771157 -0.05902324  0.45273990  2.83144005

Residual standard error: 2.68344
Degrees of freedom: 45 total; 42 residual
```

Slight differences in values of the estimates are noted between the preferred ARMA model here and AR-1

- or for that matter, results from lme().

Strong preference is shown for the preferred ARMA autocorrelation model over the Linear Fixed Model > without autocorrelation.

```
> anova(LMgls,LMcgls4)
        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
LMgls       1  4 228.4798 235.4305 -110.2399
LMcgls4     2  6 194.5268 204.9528  -91.2634 1 vs 2 37.95303  <.0001
```