# Diagnostics

Biostatistics 653

Applied Statistics III: Longitudinal Analysis

# Residuals

- We define a vector of residuals

$$\boldsymbol{r}_i = \boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}$$

for each individual. This vector has mean zero and provides an estimate of the errors

$$\boldsymbol{\epsilon}_i = \boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}$$

- We can plot the residuals $r_{ij}$ versus predicted means $\hat{\mu}_{ij} = X_{ij}\hat{\beta}$ and look for any systematic trends. The R x P plot should

  -- display no systematic pattern,

  -- have random scatter around a mean of 0, and

  -- *not* necessarily have homoscedastic variances.

# Residuals

- We should bear in mind that the components of $\boldsymbol{r}_i$ are correlated and will not necessarily have constant variance. Thus standard diagnostics for examining homogeneity of variance are not useful.
- Although the covariance of the residuals is not identical to the covariance of the errors, we can approximate the covariance of the residuals by

$$Cov(\boldsymbol{r}_i) \approx Cov(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_i$$

# Transformed Residuals

- It is desirable to facilitate model checking by transforming the residuals so that they have constant variance and zero correlation. We will do so using the Cholesky decomposition.
- Recall that the Cholesky decomposition may be used to represent a symmetric positive definite matrix **A** as $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$, where **L** is a lower triangular matrix (a lower triangular matrix has all 0's above the diagonal).
- We write $\widehat{\boldsymbol{\Sigma}}_i = \boldsymbol{L}_i \boldsymbol{L}_i^T$. Then, we use $\boldsymbol{L}_i^{-1}$ to create a set of transformed residuals

$$\boldsymbol{r}_i^* = \boldsymbol{L}_i^{-1}\boldsymbol{r}_i = \boldsymbol{L}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}})$$

that are approximately uncorrelated and have unit variance.

# Transformed Residuals

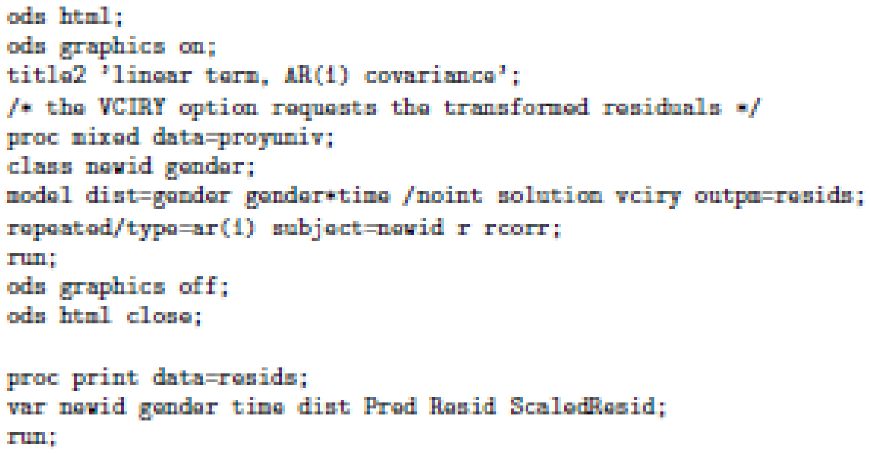
- These residuals have nice interpretations. The first element of $\boldsymbol{r}_i^*$ is the standardized residual for the first repeated measurement (often baseline). The second through last transformed residuals for each subject represent the standard deviations from the conditional mean of the response given all previous observations. That is, the *k'*th transformed residual is an estimate of

$$(Y_{ik} - E(Y_{ik}|Y_{i1}, \cdots, Y_{i,k-1}))/\sqrt{V\left(Y_{ik}|Y_{i1}, \cdots, Y_{i,k-1}\right)}$$

- Given a set of transformed residuals, the usual residual diagnostics from standard linear regression (e.g., R x P plot, Q-Q plot to detect departures from normality, skewness, etc.) may be applied.

# SAS Code

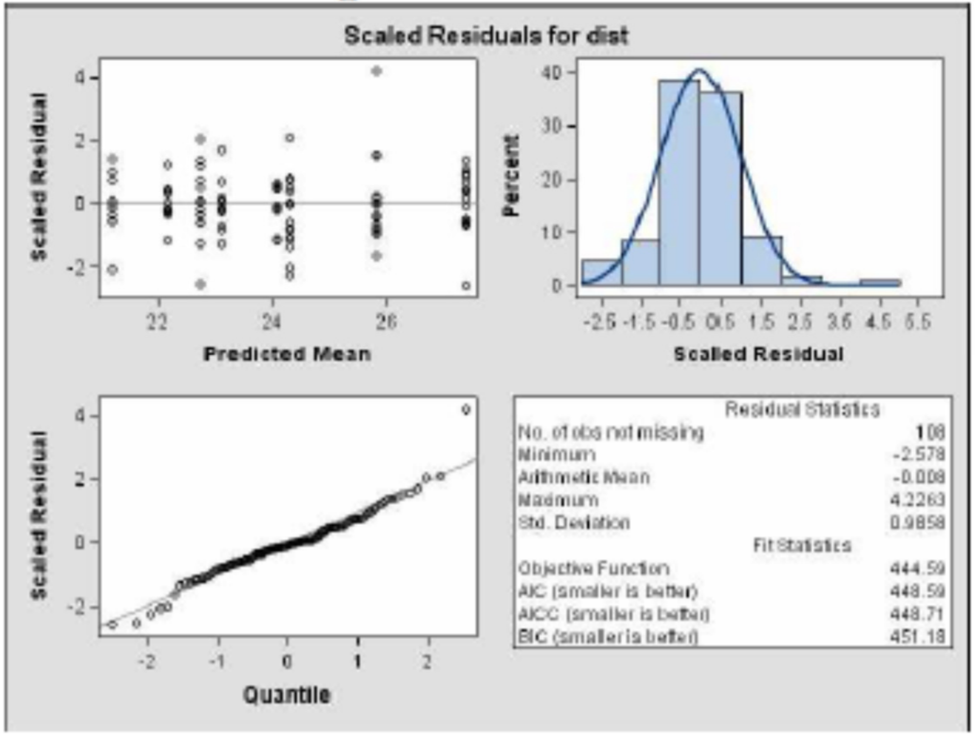
- Using the following SAS code, we request the transformed residuals (labeled by SAS as ScaledResid) and some plots for the dental data. Note that the predicted values for the plot are obtained as $\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}$.



```
ods html;
ods graphics on;
title2 'linear term, AR(1) covariance';
/* the VCIRY option requests the transformed residuals */
proc mixed data=proyuniv;
class newid gender;
model dist=gender gender*time /noint solution vciry outpm=resids;
repeated/type=ar(1) subject=newid r rcorr;
run;
ods graphics off;
ods html close;

proc print data=resids;
var newid gender time dist Pred Resid ScaledResid;
run;
```

# SAS Code

Figure 1: Residual



```
                  Residual Statistics
No. of obs not missing                   108
Minimum                               -2.578
Arithmetic Mean                       -0.008
Maximum                               4.2263
Std. Deviation                        0.9858
                    Fit Statistics
Objective Function                    444.59
AIC (smaller is better)               448.59
AICC (smaller is better)              448.71
BIC (smaller is better)               451.18
```

# Malhalanobis distance

- Based on the transformed residuals, we can calculate Malhalanobis distance

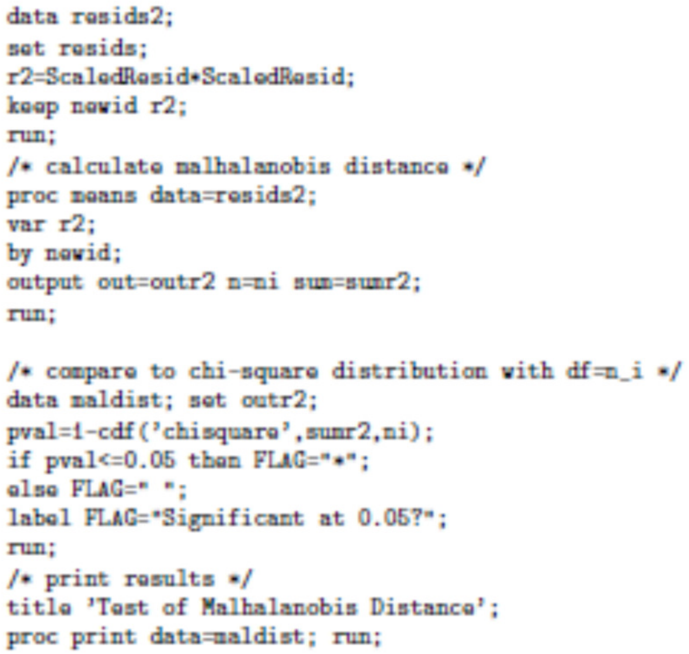
$$d_i = \boldsymbol{r}_i^{*T} \boldsymbol{r}_i^*$$

as a summary measure of the distance between observed and predicted responses. Under a correctly-specified model, these distances should be approximately chi-square with $df = n_i$ (the number of repeated measures on subject *i*).

- When we have a large dataset, it is good to bear in mind that we do expect to see some number of distances that are greater than the appropriate critical value.

# SAS Code

- The following SAS code is used to obtain the Malhalanobis distances and compare them to the $\chi^2_{n_i}$distribution.



```
data resids2;
set resids;
r2=ScaledResid*ScaledResid;
keep newid r2;
run;
/* calculate malhalanobis distance */
proc means data=resids2;
var r2;
by newid;
output out=outr2 n=ni sum=sumr2;
run;

/* compare to chi-square distribution with df=n_i */
data maldist; set outr2;
pval=1-cdf('chisquare',sumr2,ni);
if pval<=0.05 then FLAG="*";
else FLAG=" ";
label FLAG="Significant at 0.05?";
run;
/* print results */
title 'Test of Malhalanobis Distance';
proc print data=maldist; run;
```

# Semi-Variogram

- For longitudinal data, the semi-variogram is defined as half of the expected squared difference between residuals obtained on the same individual. We will use it as a diagnostic tool for assessing the adequacy of a selected covariance model. The semi-variogram, $\gamma(h_{ijk})$, is given by

$$\gamma\left(h_{ijk}\right) = \frac{1}{2} E\left(r_{ij} - r_{ik}\right)^2,$$

where $h_{ijk}$ is the time elapsed between the *j*'th and *k*'th repeated measurements on subject *i*.

# Semi-Variogram

- Because the residuals have mean zero, we can express the semi-variogram as

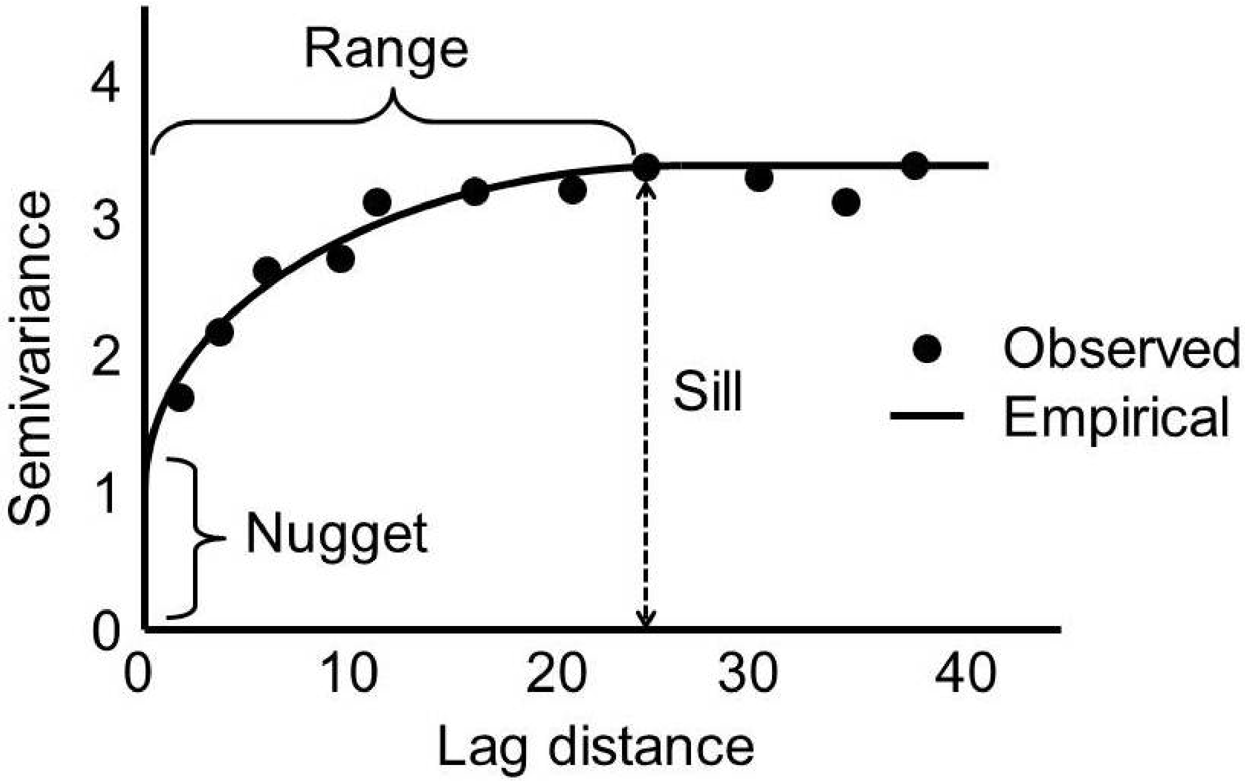
$$\gamma\left(h_{ijk}\right) = \frac{1}{2}E\left(r_{ij} - r_{ik}\right)^2 = \frac{1}{2}E\left(r_{ij}^2 + r_{ik}^2 - 2r_{ij}r_{ik}\right)$$
$$= \frac{1}{2}V\left(r_{ij}\right) + \frac{1}{2}V(r_{ik}) - Cov(r_{ij}, r_{ik})$$

- Applying the semi-variogram to the transformed residuals $r_{ij}^*$, we get

$$\gamma\left(h_{ijk}\right) = \frac{1}{2}V\left(r_{ij}^*\right) + \frac{1}{2}V(r_{ik}^*) - Cov\left(r_{ij}^*, r_{ik}^*\right) = \frac{1}{2} + \frac{1}{2} - 0 = 1$$

Thus in a correctly specified covariance model, the plot of the semi-variogram for the transformed residuals versus the time elapsed between the corresponding observations should fluctuate randomly around a horizontal line centered at 1.

# Semi-Variogram



# Semi-Variogram

- The empirical or sample semi-variogram, $\hat{\gamma}(h)$, is defined as half the average squared difference between pairs of residuals on the same individual whose corresponding observations are *h* units apart.
- If data are unbalanced, we can estimate this by fitting a smooth curve to the scatter-plot of the observed half squared differences between residuals obtained on the same subject and the corresponding time lags.
- We note that the empirical semi-variogram can be very sensitive to outliers; and that a single outlier can have a large amount of influence at several different time lags (this is because we are taking differences between pairs of residuals).

# Semi-Variogram

- To construct the sample semi-variogram, starting with the transformed residuals $r_{ij}^{*}$ and times $t_{ij}$, compute all possible

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

and

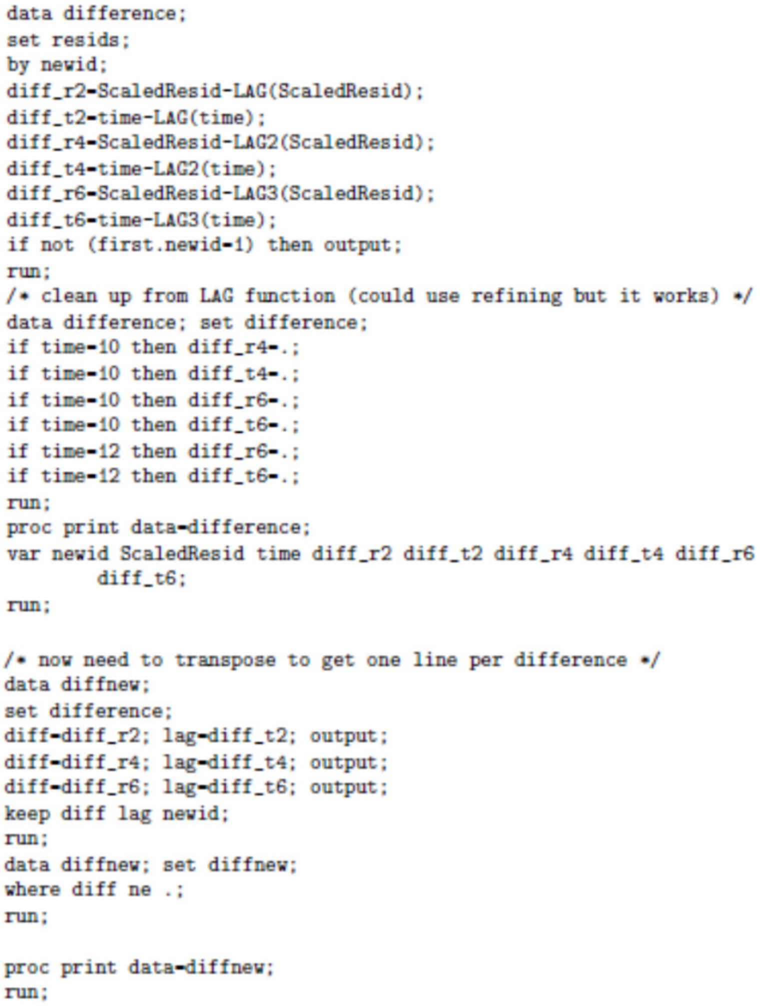
$$u_{ijk} = t_{ik} - t_{ij},$$

for j < k.

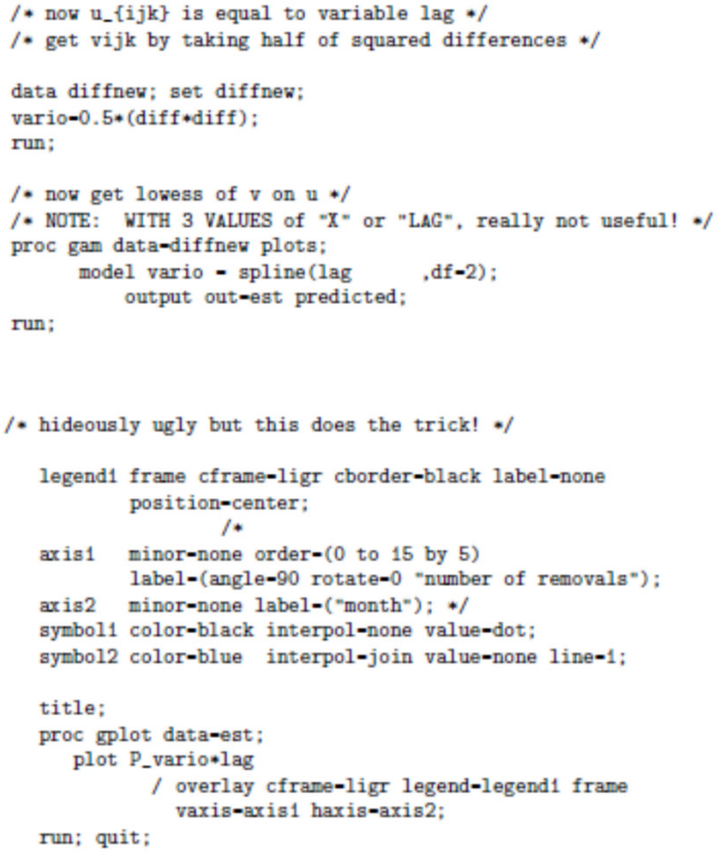
# SAS Code

- Note that while we have provided code for fitting a smooth curve to the scatter-plot of the sample values versus the time lags, this is really overkill for the dental data (for which we have only three unique lag values: 2 years, 4years, and 6 years). Really, we would just look at the means of v at each lag u, as there are only 3 unique lags.
- However, the code may be useful to you in the future when data may not be balanced and there may be a large number of unique lag values.

# SAS Code

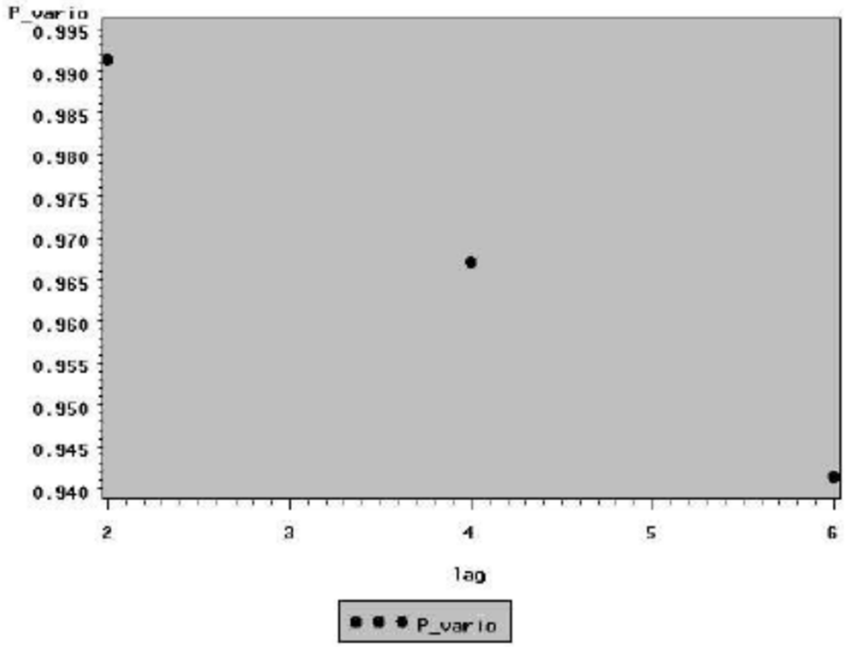


```
data difference;
set resids;
by newid;
diff_r2=ScaledResid-LAG(ScaledResid);
diff_t2=time-LAG(time);
diff_r4=ScaledResid-LAG2(ScaledResid);
diff_t4=time-LAG2(time);
diff_r6=ScaledResid-LAG3(ScaledResid);
diff_t6=time-LAG3(time);
if not (first.newid=1) then output;
run;
/* clean up from LAG function (could use refining but it works) */
data difference; set difference;
if time=10 then diff_r4=.;
if time=10 then diff_t4=.;
if time=10 then diff_r6=.;
if time=10 then diff_t6=.;
if time=12 then diff_r6=.;
if time=12 then diff_t6=.;
run;
proc print data=difference;
var newid ScaledResid time diff_r2 diff_t2 diff_r4 diff_t4 diff_r6
        diff_t6;
run;

/* now need to transpose to get one line per difference */
data diffnew;
set difference;
diff=diff_r2; lag=diff_t2; output;
diff=diff_r4; lag=diff_t4; output;
diff=diff_r6; lag=diff_t6; output;
keep diff lag newid;
run;
data diffnew; set diffnew;
where diff ne .;
run;

proc print data=diffnew;
run;
```



```
/* now u_{ijk} is equal to variable lag */
/* get vijk by taking half of squared differences */

data diffnew; set diffnew;
vario=0.5*(diff*diff);
run;

/* now get lowess of v on u */
/* NOTE:  WITH 3 VALUES of "X" or "LAG", really not useful! */
proc gam data=diffnew plots;
      model vario = spline(lag      ,df=2);
           output out=est predicted;
run;



/* hideously ugly but this does the trick! */

   legend1 frame cframe=ligr cborder=black label=none
           position=center;
                     /*
   axis1   minor=none order=(0 to 15 by 5)
           label=(angle=90 rotate=0 "number of removals");
   axis2   minor=none label=("month"); */
   symbol1 color=black interpol=none value=dot;
   symbol2 color=blue  interpol=join value=none line=1;

   title;
   proc gplot data=est;
      plot P_vario*lag
              / overlay cframe=ligr legend=legend1 frame
                vaxis=axis1 haxis=axis2;
   run; quit;
```

# Sample Semi-Variogram



# Influence Diagnostics

- We may also be interested in influence diagnostics. When we assess influence, we are interested on how much a single case effects the analysis results. We must bear in mind that we expect the data to impact analysis results (if not, something is wrong!); however, it is good to be aware if a single observation has a lot of influence on the fit of the model, even though that might not be a bad thing.
- Many influence diagnostic measures consider the impact on resulting results if an entire subject (with all repeated measures) were removed from the analysis. We will use the notation $Y_{-i}$ to denote the outcome data with the data from subject *i* removed. Similarly, we dene $X_{-i}$, which contains all the covariate data except for that for subject *i*, and $\hat{\beta}_{-i}$, which are the mean parameter estimates obtained from using the data ($Y_{-i}; X_{-i}$).

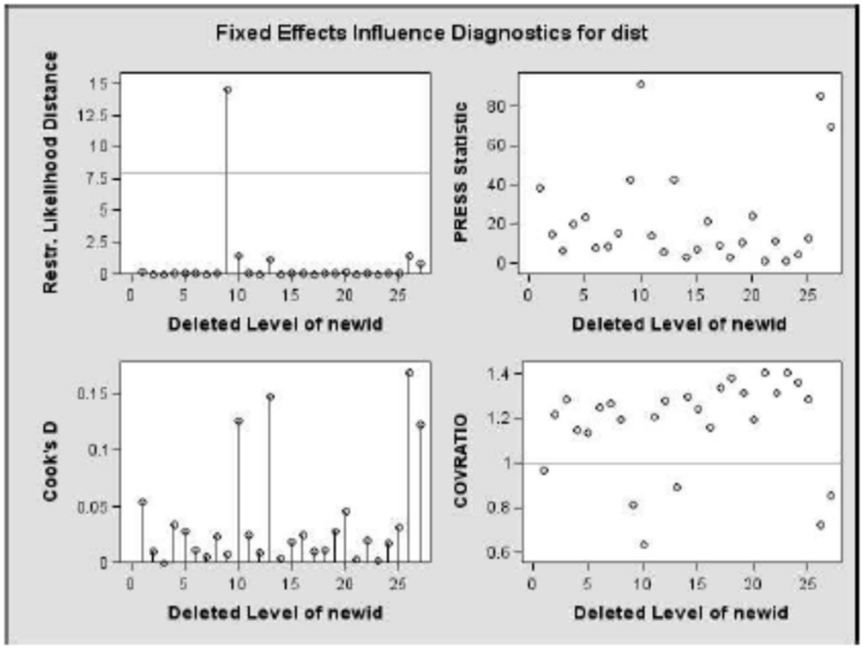
# Likelihood Distance

- In linear mixed models fit by maximum likelihood or REML, the likelihood distance (see Cook and Weisberg, 1982) can be used to measure overall influence on model fit. For this measure, you compute the full data parameter estimates $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}})$ and then re-compute the estimates based on the data without subject *i*, obtaining $\widehat{\boldsymbol{\theta}}_{-i}$. Then, we calculate the log-likelihood of the full data set (containing all N subjects) both at $\widehat{\boldsymbol{\theta}}$ and at $\widehat{\boldsymbol{\theta}}_{-i}$, taking twice the differences in these values.
- The likelihood distance for subject i is given by

$$LD_{-i} = 2(l(\widehat{\boldsymbol{\theta}}) - l(\widehat{\boldsymbol{\theta}}_{-i}))$$

- If this distance is large for one observation relative to all the others, then one should investigate its source of influence further (e.g., whether it has influence on the fixed effects estimates, their estimated variances, covariance parameter estimates, or estimates of their variances).

# Likelihood Distance



Subject 9, identified as influential by the likelihood distance, is a boy whose dental measures decreased at age 10, increased at 12, and decreased at 14 (definitely someone who should be examined more closely by the investigator!).

# Cook’s Distance

- A number of influence diagnostics have been developed to assess whether dropping one subject has undue influence on estimates of $\boldsymbol{\theta}$. Cook's distance $D_i$ for the mean parameters $\boldsymbol{\beta}$ measures the squared distance between the estimate based on all the data b and that based on the data with subject i excluded, $\widehat{\boldsymbol{\beta}}_{-\boldsymbol{i}}$ and is given by

$$D_i = \frac{\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{-i}\right)^T \hat{V}\left(\widehat{\boldsymbol{\beta}}\right)^{-1} (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{-i})}{rank(\boldsymbol{X})}$$

- Large values of $D_i$ indicate that subject i has a big impact on estimation of  relative to its variability.

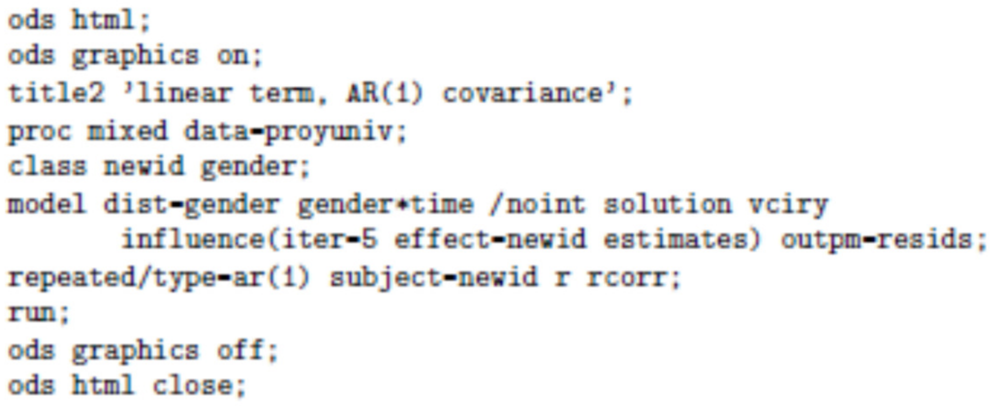
# MDFFITS

- A slight variation of Cook's distance is the multivariate DFFITS statistic, MDFFITS, which uses an *externalized* variance estimate (calculated by taking out subject i); Cook's distance does not use this externalized variance estimate.

$$S_i = \frac{\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{-i}\right)^T \hat{V}\left(\widehat{\boldsymbol{\beta}}_{-i}\right)^{-1} (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{-i})}{rank(\boldsymbol{X})}$$

- The difference means that if a subject affected both the estimate of  and made its variance large, then it might not be identified by Cook's distance, but would have a better chance of being identified by MDFFITS.
- Again, large values of MDFFITSi indicate that subject i has a big impact on estimation of  relative to its variability.

# MDFFITS Code

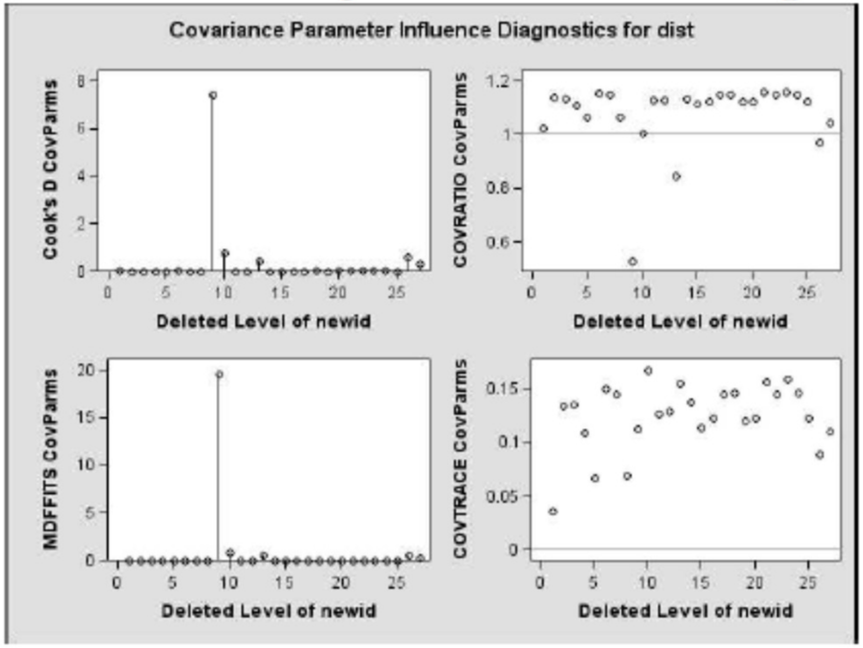


```
ods html;
ods graphics on;
title2 'linear term, AR(1) covariance';
proc mixed data=proyuniv;
class newid gender;
model dist=gender gender*time /noint solution vciry
       influence(iter=5 effect=newid estimates) outpm=resids;
repeated/type=ar(1) subject=newid r rcorr;
run;
ods graphics off;
ods html close;
```

# Covariance Parameter Diagnostics

- You can also define Cook's D and MDFFITS for covariance parameters. In this case, you would not divide the statistics by a matrix rank. When we consider influence on covariance parameter estimates, we see that subject 9 is identified as influential by both criteria.

Figure 2: Covariance parameter influence diagnostics



# Change in Precision of Estimates

- Other influence criteria are concerned with the effect of an individual on the precision of parameter estimates, as opposed to the effect on estimates themselves. This is important because an observation with a small Cook's D could still have a big effect on tests and confidence intervals if it has a big effect on precision.
- One single-number summary of a matrix is the trace (sum of diagonal elements), and another is the determinant. We can compare functions of the trace and determinant when one observation is left out of the analysis in order to characterize the influence of an individual observation.

# Covariance Trace and Covariance Ratio

- Two ways:

$$Cov\ Trace_i = |tr\left(\hat{V}(\widehat{\boldsymbol{\beta}}) - V(\widehat{\boldsymbol{\beta}}_{-i})\right) - rank(\boldsymbol{X})|$$

$$Cov\ Ratio_i = \frac{det_{ns}(V(\widehat{\boldsymbol{\beta}}_{-i}))}{det_{ns}(\hat{V}(\widehat{\boldsymbol{\beta}}))}$$

where $det_{ns}$ refers to the determinant of the nonsingular part of a given matrix.

- For these diagnostics, the benchmarks of “no influence" are 0 for the covariance trace and 1 for the covariance ratio. We can also define these quantities for the covariance matrix of the covariance parameters (in this case, we replace the rank of X in the covariance trace calculation with the rank of the covariance matrix of the covariance parameters).

# PRESS Residual

- The PRESS residual is the difference between the observed value and the predicted marginal mean, where the predicted value is obtained without including observation *i* in fitting. Formally,

$$\hat{\epsilon}_{ij,-i} = Y_{ij} - X_{ij}^{T}\hat{\beta}_{-i}$$

- The PRESS statistic is the sum of squared PRESS residuals in the deletion set and is given by

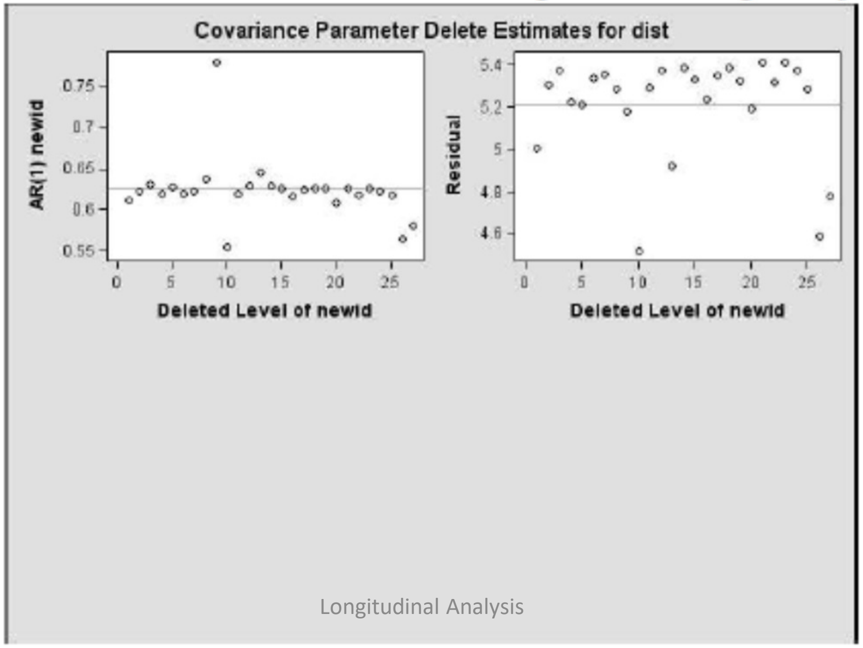
$$\sum_{j=1}^{n_i} \hat{\epsilon}_{ij,-i}^{2}$$

- Small values of this residual are desired.

# Other Deletion Statistics

- Similar deletion statistics can be obtained by examining the change in any individual parameter estimate when a subject is dropped from the dataset. In particular, the following figure shows the effects of deleting each subject in turn on the estimated gender-specific intercepts and slopes, plotting the estimates of obtained when each subject is deleted from the dataset.

Figure 4: Deletion effects on intercepts and slopes by gender



# Take Home Message

- Interpreting residual and influence diagnostics is an art more than a science.
- We may all reach different conclusions, but our next steps would be to inform the investigator of any troublesome observations.
- For example, while it is a bad idea to delete observations just to get better model fit, we would want to ensure that observation 9 in particular does not represent a data entry error.