BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 12

Empirical Standard Errors

Assessing Model Fit

Empirical Variance Estimation

We have focused on regression models for longitudinal data where the primary interest is in making inference about the regression parameters β.

For statistical inference about β we need

(i) an estimate, $\widehat{\beta}$

(ii) estimated standard error, $\text{SE}(\widehat{\beta})$

So far, we have made inferences about β using standard errors obtained under an assumed model for the covariance structure.

This approach is potentially problematic if the assumed covariance has been mis-specified.

How might the covariance be mis-specified?

For example, compound symmetry might be assumed but the correlations in fact decline over time.

Alternatively, an unstructured covariance might be assumed but the covariances also depend upon the treatment group.

If the assumed covariance has been mis-specified, we can correct the standard errors by using "empirical" or so-called "robust" variances.

Recall, the REML estimator of β is given by

$$\widehat{\beta} = \left[\sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}^{-1} X_i \right) \right]^{-1} \sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}^{-1} Y_i \right)$$

where $\widehat{\Sigma}$ is the REML estimate of Σ.

It has covariance matrix,

$$\text{Cov}(\widehat{\beta}) = \left[\sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}^{-1} X_i \right) \right]^{-1} \sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}^{-1} \text{Cov}\left(Y_i\right) \widehat{\Sigma}^{-1} X_i \right) \left[\sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}^{-1} X_i \right) \right]^{-1}$$

If $\text{Cov}(Y_i)$ is replaced by $\widehat{\Sigma}$, the REML estimate of Σ, $\text{Cov}(\widehat{\beta})$ can be estimated by

$$\left[\sum_{i=1}^{N}\left(X_i'\widehat{\Sigma}^{-1}X_i\right)\right]^{-1}$$

However, if the covariance has been mis-specified then an alternative estimator for $\text{Cov}\left(Y_i\right)$ is needed.

The empirical or so-called robust variance of $\widehat{\beta}$ is obtained by using

$$\widehat{V}_i = \left(Y_i - X_i\widehat{\beta}\right)\left(Y_i - X_i\widehat{\beta}\right)'$$

as an estimate of $\text{Cov}\left(Y_i\right)$.

Thus, the empirical variance of $\widehat{\beta}$ is estimated by

$$\left[\sum_{i=1}^{n}\left(X_i'\widehat{\Sigma}^{-1}X_i\right)\right]^{-1}\sum_{i=1}^{n}\left(X_i'\widehat{\Sigma}^{-1}\widehat{V}_i\widehat{\Sigma}^{-1}X_i\right)\left[\sum_{i=1}^{n}\left(X_i'\widehat{\Sigma}^{-1}X_i\right)\right]^{-1}$$

This empirical variance estimator is also known as the "sandwich estimator".

The remarkable thing about the empirical estimator of $\text{Cov}(\widehat{\beta})$ is that it provides a consistent estimator of the variance even when the model for the covariance matrix has been misspecified.

That is, in large samples the empirical variance estimator yields correct standard errors.

In general, its use should be confined to cases where N (number of individuals) is relatively large and n (number of measurements) is relatively small.

The empirical variance estimator may not be appropriate when there is severe imbalance in the data.

In summary, (with large samples) the following procedure will produce valid estimates of the regression coefficients and their standard errors:

(1) Choose a “working” covariance matrix of some convenient form.

(2) Estimate the regression coefficients under the assumed working covariance matrix.

(3) Estimate the standard errors using the empirical variance estimator.

Why not be a clever ostrich?

Why not simply ignore potential correlation among repeated measures (i.e., put head in sand) and assume an independence “working” covariance? Then, obtain correct standard errors using empirical variance estimator. Why should we bother to explicitly model the covariance?

Reasons:

(1) Efficiency: The optimal (most precise) estimator of β uses the true $\text{Cov}(Y_i)$. Given sufficient data, we can attempt to estimate $\text{Cov}(Y_i)$.

(2) When N (number of individuals) is not large relative to n (number of measurements) the empirical variance estimator is not recommended.

(3) Missing values: The empirical variance estimator uses the replications across individuals to estimate the covariance structure. This becomes problematic when there are missing data or when the times of measurement are not common.

In general, it is advantageous to model the covariance.

Table 1: Illustrative commands for an exponential model,
with empirical standard errors, using PROC MIXED in SAS.

```
PROC MIXED EMPIRICAL;
  CLASS id group time;
  MODEL y=group time group*time /S CHISQ;
  REPEATED time / TYPE=SP(EXP)(ctime) SUBJECT=id R RCORR;
```

Assessing Model Fit

Visual check of estimated and sample means over time.

Comparison of estimated and sample covariance structure.

Statistical comparison of selected model and higher-order (perhaps saturated) model.

Diagnostics based on analysis of residuals.

Residuals in Standard Linear Regression

Consider situation with a single measurement per subject.

Model: $Y_i = X_i\beta + e_i$, with $e_i \sim$ i.i.d. $N(0, \sigma^2)$.

True error: $e_i = Y_i - X_i\beta$.

Residual (estimated error): $r_i = Y_i - X_i\hat{\beta}_i$.

Residuals are uncorrelated with predicted values and with covariates.

Plots of residuals against predicted values, and against predictor variables should appear random, symmetric about 0, with constant variance, and no evidence of outliers.

Residuals in Regression Models for Longitudinal Data

Error and residual are now vectors.

True error: $e_i = Y_i - X_i\beta$.

Residual (estimated error): $r_i = Y_i - X_i\hat{\beta}$ with $r_{ij} = Y_{ij} - X_{ij}'\hat{\beta}$.

If mean model fits well, plot of r_{ij} against $\hat{\mu}_{ij} = X_{ij}'\hat{\beta}$ should not show any systematic patterns about the constant mean of zero.

Plot may also be useful for detecting outliers.

Correlation of Residuals

Problem: Components of the residual vector, r_i, are correlated and may not have constant variance.

Covariance of the residuals is not identical to the covariance of the (true) errors, but:

$$\text{Cov}(r_i) \approx \text{Cov}(e_i) = \Sigma_i.$$

Consequently, a plot of residuals versus predicted values may not show homogeneity of variance.

Can also show that residuals and covariates may be correlated, and so a plot of residuals versus a covariate may show a systematic trend.

Transformed Residuals

Idea: Transform the residuals to ones that have constant variance and zero correlation; then can interpret plots etc. in similar manner to those used for standard linear regression.

Use: $r_i^* = L_i^{-1} r_i = L_i^{-1}(Y_i - X_i\hat{\beta})$ where L_i is chosen to give uncorrelated residuals.

If we use the Cholesky decomposition such that L_i is a lower triangular matrix satisfying $\hat{\Sigma}_i = L_i L_i'$, then the residuals, r_i^*, are uncorrelated and have unit variance.

Similarly, can transform the predicted values, giving $\hat{\mu}_i^* = L_i^{-1} X_i \hat{\beta}$.

Plots of Transformed Residuals

Plot of r^*_{ij} against $\hat{\mu}^*_{ij}$ should show a random symmetric scatter around a constant mean of zero with constant variability and no outliers.

Similarly for plot of r^*_{ij} against transformed covariate values, $X^*_i = L_i^{-1} X_i$.

Plot of transformed residuals versus transformed time especially useful for assessing adequacy of model for change in mean response with time.

Often useful in plots to incorporate a lowess curve through the transformed residuals to help in evaluating systematic departures.

Normal plot (quantile or Q-Q plot) of the transformed residuals useful in assessing normality assumption.

Plot of $|r^*_{ij}|$ versus (transformed) time or versus μ^*_{ij} should also show no systematic pattern and should be centered approximately at 0.8, which is the mean of the absolute value from a $N(0,1)$ distribution.

SAS Code

```
* raw residual plots;
ods graphics on;
proc mixed data=lead plots=residualpanel(unpack) ;
class id trt;
model PbB = trt week trt*week / s outpm=fitted;
random intercept week / type=un subject=id g gcorr v vcorr;
run;   ods graphics off;

* standardized residual plots;
ods graphics on;
proc mixed data=lead plots=vcirypanel(unpack) ;
class id trt;
model PbB = trt week trt*week / s outpm=fitted vciry;
random intercept week / type=un subject=id g gcorr v vcorr;
run;  ods graphics off;
```

Evaluating the Covariance Structure: the Semi-Variogram

Semi-variogram is defined as:

$$\begin{aligned}\gamma(h_{ijk}) &= \frac{1}{2}E(r_{ij}-r_{ik})^2 \\ &= \frac{1}{2}E(r_{ij}^2+r_{ik}^2-2r_{ij}r_{ik}) \\ &= \frac{1}{2}\mathrm{Var}(r_{ij})+\frac{1}{2}\mathrm{Var}(r_{ik})-\mathrm{Cov}(r_{ij},r_{ik}).\end{aligned}$$

where h_{ijk} is the time elapsed between the j^{th} and k^{th} repeated measurements on the i^{th} individual.

Estimate of the semi-variogram is defined as one-half of the average squared difference between pairs of residuals on the same individual which are h units apart.

As a diagnostic for assessing the adequacy of the assumed covariance structure for a model, more useful to evaluate the semi-variogram for the transformed residuals with r_{ij} and r_{ik} replaced by r_{ij}^* and r_{ik}^*.

Then $\gamma(h_{ijk}) = \frac{1}{2}\mathrm{Var}(r_{ij}^*) + \frac{1}{2}\mathrm{Var}(r_{ik}^*) - \mathrm{Cov}(r_{ij}^*, r_{ik}^*) = \frac{1}{2}(1) + \frac{1}{2}(1) - 0 = 1.$

Thus, a plot of the transformed semi-variogram against transformed time should fluctuate randomly about 1.

In practice, fit a smooth (e.g. lowess) curve to the scatterplot of observed half squared differences between residuals obtained on the same individual and the corresponding time lags.

SAS Code

```
* run the model;
proc mixed data=lead;
class id trt;
model PbB = trt week trt*week / s outpm=fitted vciry;
random intercept week / type=un subject=id g gcorr v vcorr;
run;

* generate all pairs of residuals within a person;
proc variogram data=fitted outpair=pairs noprint;
by id;
compute lagd=2 maxlag=8 novariogram;
coordinates xc=week yc=week;
var scaledresid;
run;
```

```
* calculate semivariogram values for all pairs;
data semivar;
set pairs;
hijk=x2-x1;
sv = .5*(v1 - v2)**2;
run;

* plot smooth curve;
ods graphics on;
proc loess data=semivar;
model sv=hijk;
run;
ods graphics off;
```

Case Study: *Influence of Menarche on Changes in Body Fat*

- Prospective study on body fat accretion in a cohort of 162 girls from the MIT Growth and Development Study.

- At start of study, all the girls were pre-menarcheal and non-obese

- All girls were followed over time according to a schedule of annual measurements until four years after menarche.

- The final measurement was scheduled on the fourth anniversary of their reported date of menarche.

- At each examination, a measure of body fatness was obtained based on bioelectric impedance analysis.

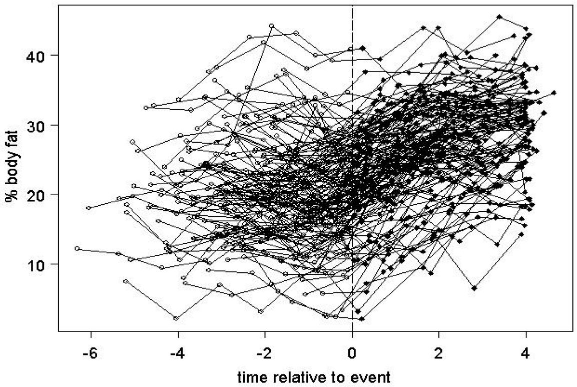

Figure 1: Timeplot of percent body fat against time, relative to age of menarche (in years).

Piecewise Linear Mixed Effects Model

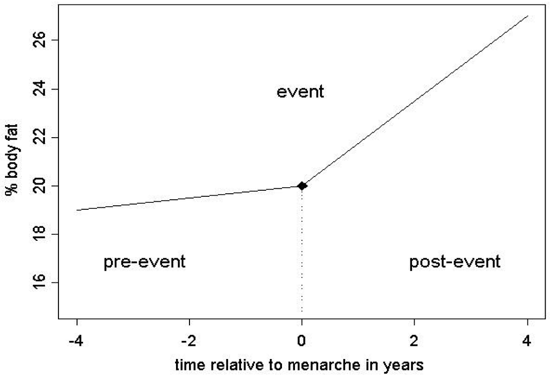

Figure 2: Graphical representation of piecewise linear trajectory.

We previously fitted the piecewise linear mixed effects model:

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij})_+ + b_{1i} + b_{2i} t_{ij} + b_{3i} (t_{ij})_+,$$

where $(t_{ij})_+ = t_{ij}$ if $t_{ij} > 0$ and $(t_{ij})_+ = 0$ if $t_{ij} \leq 0$.

The intercept β_1 is the average % body fat at menarche (when $t_{ij} = 0$).

The slope β_2 is the average rate of change in % body fat (per year) during the pre-menarcheal period.

The average rate of change in % body fat (per year) during the post-menarcheal period is given by $(\beta_2 + \beta_3)$.

Table 2: Estimated regression coefficients (fixed effects) and standard errors for the percent body fat data.

PARAMETER	ESTIMATE	SE	Z
INTERCEPT	21.3614	0.5646	37.84
time	0.4171	0.1572	2.65
$(\text{time})_+$	2.0471	0.2280	8.98

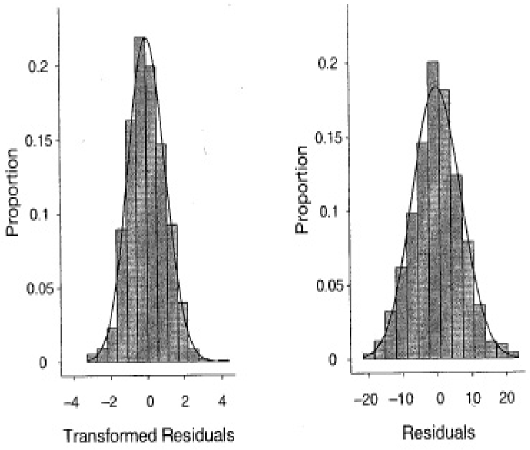

Figure 3: Frequency Plots of Transformed and Untransformed Residuals

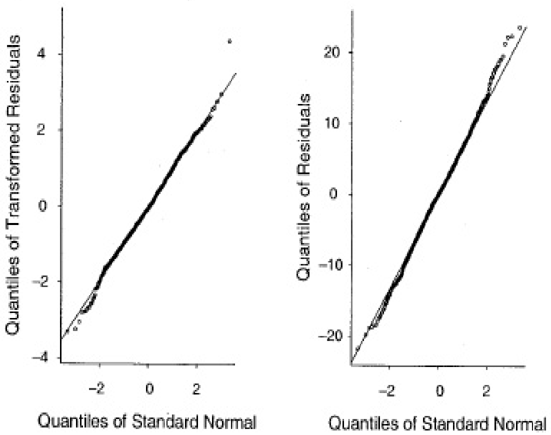

Figure 4: Normal Plots of Transformed and Untransformed Residuals

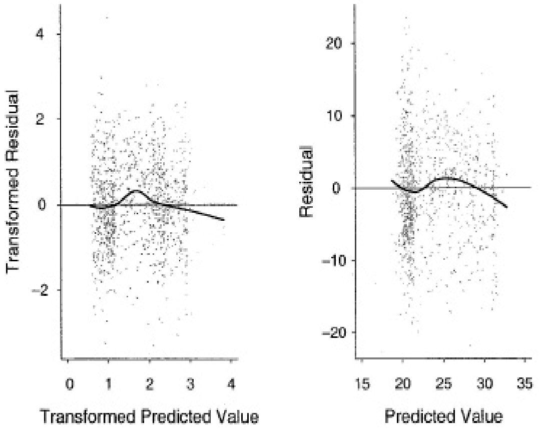

Figure 5: Plots of Residuals versus Predicted Values

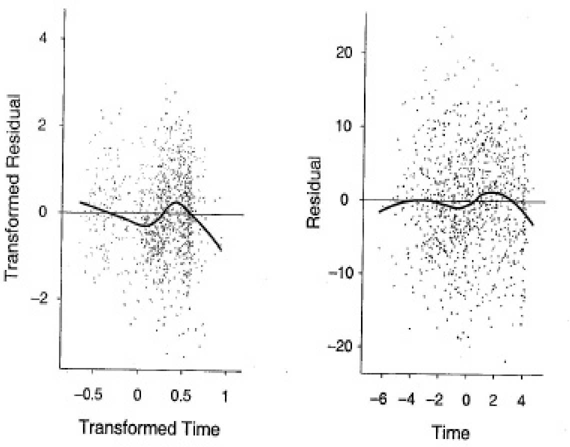

Figure 6: Plots of Residuals versus Time: Model with Linear Time Effect

Consider extension of the model to include a quadratic term in time after menarche:

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij})_+ + \beta_4 (t_{ij})_+^2 + b_{1i} + b_{2i} t_{ij} + b_{3i} (t_{ij})_+ + b_{4i} (t_{ij})_+^2,$$

where $(t_{ij})_+ = t_{ij}$ if $t_{ij} > 0$ and $(t_{ij})_+ = 0$ if $t_{ij} \leq 0$.

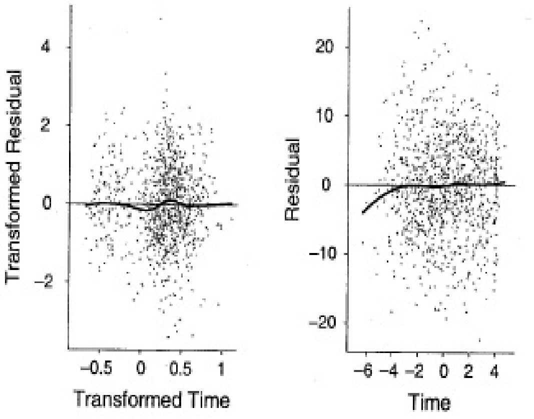

Figure 7: Plots of Residuals versus Time: Model with Quadratic Time Effect

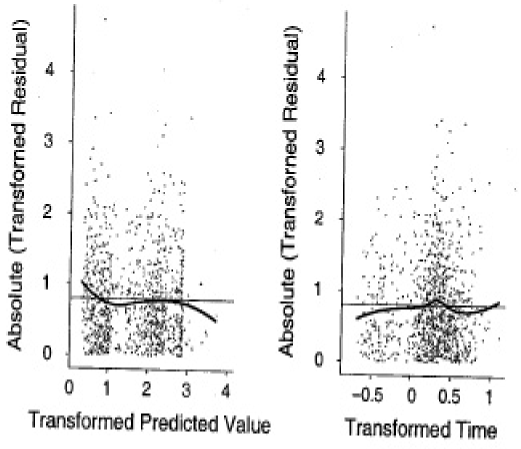

Figure 8: Plots of Absolute Residuals versus Time: Model with Quadratic Time Effect

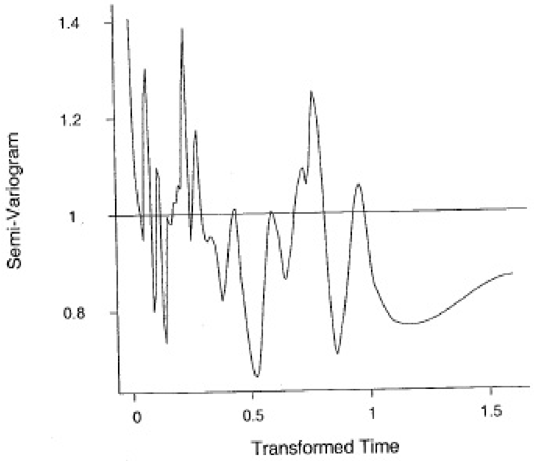

Figure 9: Empirical Semi-variogram: Model with Quadratic Time Effect

Table 3: Estimated regression coefficients (fixed effects) and standard errors for model with post-menarche quadratic term

PARAMETER	ESTIMATE	SE	Z
INTERCEPT	20.4201	0.5817	35.10
time	-0.0155	0.1612	-0.10
$(\text{time})_+$	4.8439	0.4055	11.94
$(\text{time})_+^2$	-0.6469	0.0772	-8.38

Strong evidence that rate of change in % body fat is highest just after menarche and then reduces with time.

No significant increase in % body fat prior to menarche.

In principle, could extend model to include higher order polynomial term(s) but biological rationale for the complexity may be difficult to justify, and model could be overly influenced by a few girls with longer follow-up post- (or pre-) menarche.

Example of Semi-variogram for Model with Incorrect Covariance Structure

Consider model with the intercept as the only random effect:

$$E(Y_{ij}|b_i) = \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij})_+ + \beta_4 (t_{ij})_+^2 + b_{1i}$$

Recall: Induced variance-covariance structure is then compound symmetry.

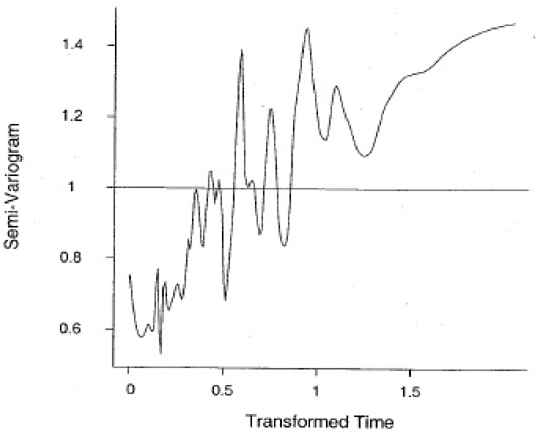

Figure 10: Empirical Semi-variogram: Model with Intercept as Only Random Effect