Assessment of Assumptions

Each of the approaches to modeling that we have described depend upon assumptions concerning the structure of the data; to ensure that inference is appropriate we need to attempt to check that these assumptions are valid.

We first recap the assumptions:

GEE

Model:

$$\boldsymbol{Y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{e}_i,$$

with working covariance model $\text{var}(\boldsymbol{e}_i) = \boldsymbol{W}_i(\boldsymbol{\alpha})$, $i = 1, ..., m$.

G1 Marginal model $\text{E}[\boldsymbol{Y}_i] = \boldsymbol{x}_i\boldsymbol{\beta}$ is appropriate.

G2 m is sufficiently large for asymptotic inference to be appropriate.

G3 m is sufficiently large for robust estimation of standard errors.

G4 The working covariance $\boldsymbol{W}_i(\boldsymbol{\alpha})$ is not far from the "true" covariance structure; if this is the case then the analysis will be very inefficient (standard errors will be much bigger than they need to be).

LMEM

Model:

$$\boldsymbol{Y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

with $\boldsymbol{b}_i \sim \text{N}(\boldsymbol{0}, \boldsymbol{D})$, $\boldsymbol{\epsilon}_i \sim \text{N}(\boldsymbol{0}, \boldsymbol{E}_i)$, $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ independent, $i = 1, ..., m$.

L1 Mean model for fixed effects $\boldsymbol{x}_i\boldsymbol{\beta}$ is appropriate.

L2 Mean model for random effects $\boldsymbol{z}_i\boldsymbol{b}_i$ is appropriate.

L3 Variance model for $\boldsymbol{\epsilon}_i$ is correct.

L4 Variance model for $\boldsymbol{b}_i$ is correct.

L5 Normality of $\boldsymbol{\epsilon}_i$.

L6 Normality of $\boldsymbol{b}_i$.

L7 m is sufficiently large for asymptotic inference to be appropriate.

Bayesian Hierarchical Model

Model for LMEM, plus priors for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

Each of L1–L6 (asymptotic inference is not required if, for example, MCMC is used, though "appropriate" priors are needed).

Residual Analysis

Residuals may be defined with respect to different levels of the model.

A vector of unstandardized population-level residuals is given by

$$\boldsymbol{e}_i = \boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}.$$

A vector of unstandardized unit-level residuals is given by

$$\boldsymbol{\epsilon}_i = \boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta} - \boldsymbol{z}_i\boldsymbol{b}_i.$$

The vector of unstandardized random effects, $\boldsymbol{b}_i$, is also a form of residual.

Estimated versions of these residuals are given by

$$\widehat{\boldsymbol{e}}_i = \boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}},$$

$$\widehat{\boldsymbol{\epsilon}}_i = \boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}} - \boldsymbol{z}_i\widehat{\boldsymbol{b}}_i,$$

and $\widehat{\boldsymbol{b}}_i$.

Recall from consideration of the ordinary linear model that estimated residuals have dependencies induced by the estimation procedure; in the dependent data context the situation is much worse as the "true" residuals have dependencies due to the dependent nature of the data.

Hence standardization is essential.

Standardized Population Residuals

If $\boldsymbol{V}_i(\boldsymbol{\alpha})$ is the true error structure then

$$\text{var}(\boldsymbol{e}_i) = \boldsymbol{V}_i, \quad \text{and} \quad \text{var}(\widehat{\boldsymbol{e}}_i) \approx \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}}).$$

This dependence means that it is not possible to check whether the covariance model (both form of the correlation structure and mean-variance model) is correctly specified.

Plotting $\widehat{e}_{ij}$ versus x_{ij} may also be misleading due to the dependence within the residuals.

As an alternative, let $\widehat{\boldsymbol{V}}_i = \boldsymbol{L}_i\boldsymbol{L}_i^{\text{T}}$ denote the Cholesky decomposition of $\widehat{\boldsymbol{V}}_i = \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}})$, the estimated variance-covariance matrix.

We can use this decomposition to form

$$\widehat{\boldsymbol{e}}_i^{\star} = \boldsymbol{L}_i^{-1}\widehat{\boldsymbol{e}}_i = \boldsymbol{L}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}).$$

Note that: $\text{var}(\boldsymbol{e}_i^{\star}) = \boldsymbol{I}$.

We now discuss a tool for examining the correlation in a set of residuals.

The Variogram

For unequally-spaced data the ACF is not so convenient, unless we round the observations.

An alternative is provided by the *semi-variogram* which is defined, for a process Z_t and $d \geq 0$.

$$\gamma(d) = \frac{1}{2}\mathrm{E}\left[\{Z_t - Z_{t-d}\}^2\right].$$

For a second-order stationary process, $\mathrm{E}[Z_t] = \mu$ for all t and $\mathrm{cov}(Z_t, Z_{t-d})$ only depends on the distance d (which implies constant variance).

A smooth process is L_2-continuous, i.e.

$$\mathrm{E}\{(Z_t - Z_{t-d})^2\} \rightarrow 0$$

as $d \rightarrow 0$.

For a second-order stationary smooth process

$$\begin{aligned}\gamma(d) &= \frac{1}{2}\left\{\mathrm{E}[Z_t^2] + \mathrm{E}[Z_{t-d}^2] - 2\mathrm{E}[Z_t Z_{t-d}]\right\} \\ &= \sigma_z^2\{1 - \rho(d)\},\end{aligned}$$

where $\mathrm{var}(Z) = \sigma^2$.

The semi-variogram is also well-defined for an *intrinsically* stationary process for which $\mathrm{E}[Z_t] = \mu$ and for which

$$\mathrm{E}[(Z_t - Z_{t-d})^2] = 2\gamma(d).$$

As d increases then for observatons far apart in time

$$\gamma(d) \rightarrow \mathrm{var}(Z_t) = \sigma_z^2,$$

which (recall) is assumed constant.

Consider measurement error, ϵ_t with $\mathrm{E}[\epsilon_t] = 0$, $\mathrm{var}(\epsilon_t) = \sigma_\epsilon^2$, and

$$Y_t = Z_t + \epsilon_t,$$

then:

$$\gamma(d) = \frac{1}{2}\mathrm{E}\left[\{Y_t - Y_{t-d}\}^2\right] = \sigma_z^2\{1 - \rho(d)\} + \sigma_\epsilon^2,$$

and we have a "nugget" effect σ_ϵ^2.

The Variogram in Longitudinal Data Analysis

Define the semi-variogram of the population residuals, $e_{ij} = Y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta}$, as

$$\gamma_i(d_{ijk}) = \frac{1}{2}\mathrm{E}\left[\{e_{ij} - e_{ik}\}^2\right],$$

for $d_{ijk} = \mid t_{ij} - t_{ik} \mid \geq 0$.

Note: differences on the same individual.

The sample semi-variogram uses the empirical halved differences between pairs of population residuals

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

along with the spacings $u_{ijk} = t_{ij} - t_{ik}$.

With highly-irregular sampling times the variogram can be estimated from the pairs (u_{ijk}, v_{ijk}), $i = 1, ..., m$, $j < k = 1, ..., n_i$, with the resultant plot being smoothed.

The marginal distribution of each v_{ijk} is χ^2_1, and this large variability can make the variogram difficult to interpret.

The total variance is estimated as the average of $\frac{1}{2}(e_{ij} - e_{lk})^2$, for $i \neq l$, since

$$\frac{1}{2}\mathrm{E}\left[(e_{ij} - e_{lk})^2\right] = \frac{1}{2}\left\{\mathrm{E}[e_{ij}^2] + \mathrm{E}[e_{lk}^2]\right\} = \sigma^2,$$

assuming that observations on different individuals are independent (and the variance is constant over time, and for different individuals).

Consider the interpretation of the variogram for the model

$$Y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i + \delta_{ij} + \epsilon_{ij},$$

where $b_i \sim_{ind} N(0, \sigma_0^2)$ (note, univariate), $\epsilon_{ij} \sim_{ind} N(0, \sigma_\epsilon^2)$, and δ_{ij} represent error terms with serial dependence.

A simple and commonly-used form for serial dependence is the AR(1) model given by

$$\mathrm{cov}(\delta_{ij}, \delta_{ik}) = \sigma_\delta^2 \rho^{|t_{ij} - t_{ik}|}.$$

Under this model

$$\mathrm{var}(Y_{ij}|\boldsymbol{\beta}) = \sigma^2 = \sigma_0^2 + \sigma_\delta^2 + \sigma_\epsilon^2.$$

Consider the theoretical variogram for the residuals

$$e_{ij} = Y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta} = b_i + \delta_{ij} + \epsilon_{ij},$$

$i = 1, ..., m; j = 1, ...n_i$, with the AR(1) model.

For differences in residuals on the same individual

$$e_{ij} - e_{ik} = b_i + \delta_{ij} + \epsilon_{ij} - b_i - \delta_{ik} - \epsilon_{ik} = \delta_{ij} + \epsilon_{ij} - \delta_{ik} - \epsilon_{ik},$$

and so

$$\gamma_i(d_{ijk}) = \frac{1}{2}\mathrm{E}\left[(e_{ij} - e_{ik})^2\right] = \sigma_\delta^2(1 - \rho^{d_{ijk}}) + \sigma_\epsilon^2. \qquad (12)$$

As $d_{ijk} \to 0$, $\gamma_i(d_{ijk}) \to \sigma_\epsilon^2$ and b_i is the mean of e_{ij} and so its variance does not appear in (12).

The variogram is limited in its use for population residuals for the LMEM .

Consider, the mixed effects model with random intercepts and independent random slopes:

$$b_{i0} \sim N(0, \upsilon_{00}^2), \quad b_{i1} \sim N(0, \upsilon_{11}^2)$$

leads to non-constant marginal variance

$$\mathrm{var}(Y_{ij}|\boldsymbol{\beta}) = \upsilon_{00}^2 + 2\upsilon_{11}^2 t_{ij}^2,$$

so that we would not want to look at a variogram of population residuals because we do not have second-order stationarity.

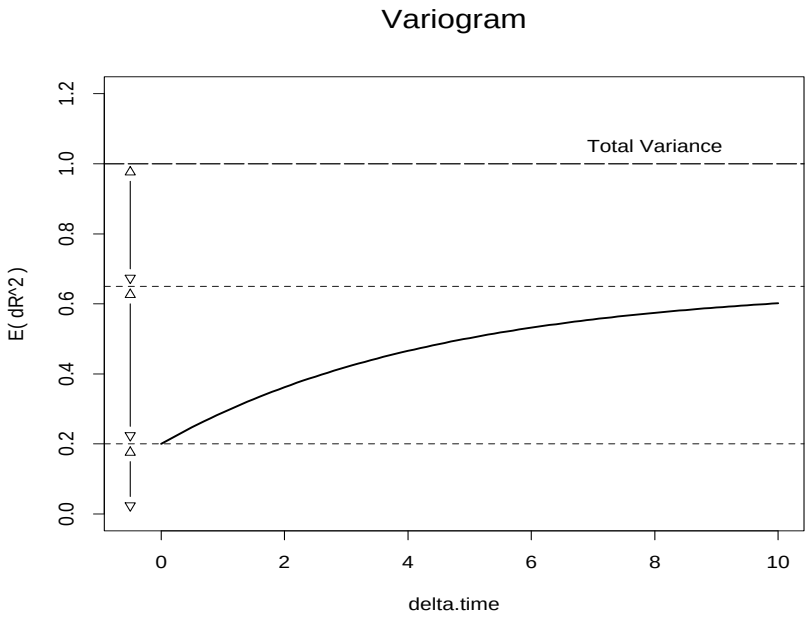

Figure 10: Theoretical variogram for a model with a random intercept, serial correlation, and measurement error.


```
lda.variogram <- function( id, y, x ){
# INPUT:  id = (nobs x 1) id vector
#          y = (nobs x 1) response (residual) vector
#          x = (nobs x 1) covariate (time) vector
# RETURN:  delta.y = vec( 0.5*(y_ij - y_ik)^2 )
#          delta.x = vec( abs( x_ij - x_ik ) )
uid <- unique( id )
m <- length( uid )
delta.y <- NULL
delta.x <- NULL
did <- NULL
for( i in 1:m ){
  yi <- y[ id==uid[i] ]
  xi <- x[ id==uid[i] ]
  n <- length(yi)
  expand.j <- rep( c(1:n), n )
  expand.k <- rep( c(1:n), rep(n,n) )
  keep <- expand.j > expand.k
  if( sum(keep)>0 ){
        expand.j <- expand.j[keep]
        expand.k <- expand.k[keep]
        delta.yi <- 0.5*( yi[expand.j] - yi[expand.k] )^2
        delta.xi <- abs( xi[expand.j] - xi[expand.k] )
        didi <- rep( uid[i], length(delta.yi) )
        delta.y <- c( delta.y, delta.yi )
        delta.x <- c( delta.x, delta.xi )
        did <- c( did, didi ) } }
out <- list( id = did, delta.y = delta.y, delta.x = delta.x )
out}
```

We illustrate the use of the `lda.variogram` function using the follicle data.

```
> mod0 <- lme( follicles ~ x1+x2,data=Ovary,random= ~1 )
# Obtain population residuals.
> res <- residuals(mod0,type="response")
# Obtain empirical variogram of these residuals.
> vario <- lda.variogram( Ovary$Mare, res, Ovary$Time)
> plot(vario$delta.x,vario$delta.y)
# Use a spline to reveal any trend in the variogram
# -- preferable to the use of a {\tt lowess()} smoother
# since the latter is not good with highly skewed data.
> lines(smooth.spline(vario$delta.x,vario$delta.y,df=10))
> var.est <- var(res)
> abline(h=var.est,lty=2)
# Now for GEE
> modgee <- geese(follicles ~ x1+x2,data=Ovary,id=Mare,
  corstr="independence")
> resgee <- Ovary$follicles-12.215-x1*(-3.339)-x2*(-0.869)
> vario <- lda.variogram( Ovary$Mare, resgee, Ovary$Time)
> plot(vario$delta.x,vario$delta.y,ylab="Empirical variogram",
  xlab="Time difference")
> lines(smooth.spline(vario$delta.x,vario$delta.y,df=10))
> var.est <- var(res)
> abline(h=var.est,lty=2)
```

We see what appears to be an increasing trend in the empirical semi-variogram plot in Figure 11 – suggests that the errors are correlated (or increasing variance?).

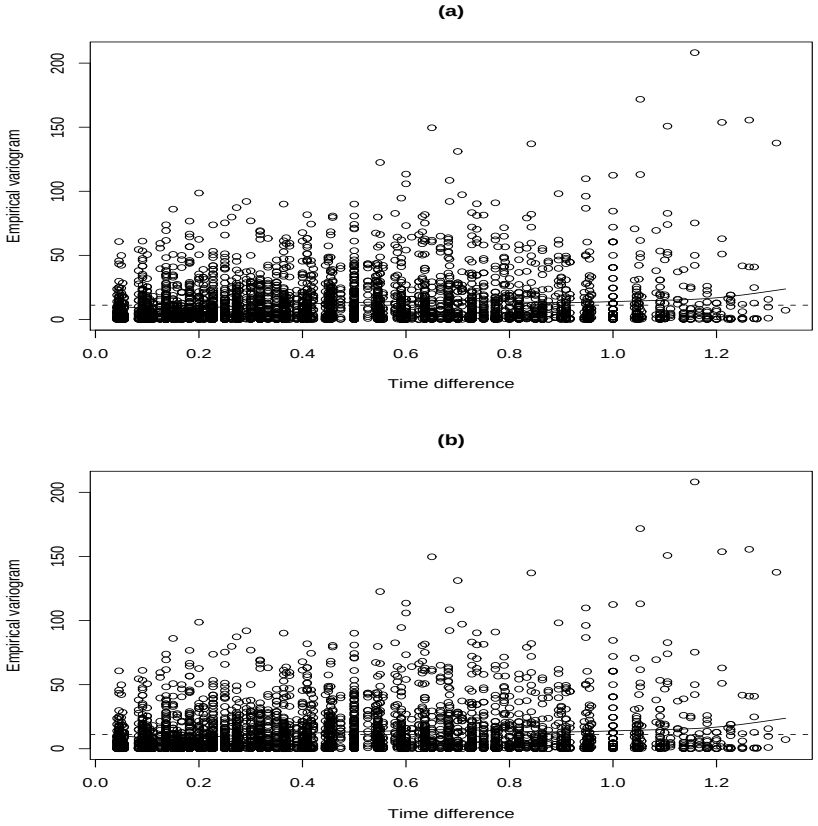

Figure 11: Empirical variogram of population residuals from (a) GEE, and (b) linear mixed effects model.

Variograms for Simulated Data

We now simulate data from various longitudinal models.

Each plot shows the raw data with individual fitted straight lines (top left), and then the variogram from the population residuals (top right), the subject-specific residuals (bottom left).

Consider the population residual process on subject i:

$$\begin{aligned} e_i(t) &= b_i + \delta_i(t) + \epsilon_i(t) \\ e_i(t-u) &= b_i + \delta_i(t-u) + \epsilon_i(t-u) \end{aligned}$$

where

- $b_i \sim N(0, \sigma_0^2)$,
- $\epsilon_i(s) \sim_{ind} N(0, \sigma_\epsilon^2)$, and
- δ_i are normal with $\text{cov}\{\delta_i(t), \delta_i(t-u)\} = \sigma_\delta^2 \rho^u$, expressing serial correlation on observations on the same individual.

We have

$$\gamma(u) = \text{E}[\{e_i(t) - e_i(t-u)\}^2] = \begin{cases} \sigma_\epsilon^2 + \sigma_\delta^2\{1 - \rho(u)\} & \text{for } u \neq 0 \\ \sigma_\epsilon^2 & \text{for } u = 0 \end{cases}$$

and $\text{var}\{e_i(t)\} = \sigma_\epsilon^2 + \sigma_0^2 + \sigma_\delta^2$.

If we define subject specific residuals as:

$$\begin{aligned} \epsilon_i^\star(t) &= \delta_i(t) + \epsilon_i(t) \\ \epsilon_i^\star(t-u) &= \delta_i(t-u) + \epsilon_i(t-u) \end{aligned}$$

then $\text{var}\{\epsilon_i^\star(t)\} = \sigma_\epsilon^2 + \sigma_\delta^2$.

Model for simulation:

$$Y_{ij} = 2 + 1 \times t_j + b_i + \delta_{ij} + e_{ij},$$

- For $i = 1, ..., m = 20$ individuals, with
- $j = 1, ..., n_i = 10$ observations on each,
- $b_i \sim N(0, \sigma_0^2)$,
- $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$,
- $\delta_{ij} \sim N(0, \sigma_\delta^2)$, $\text{cov}(\delta_{ij}, \delta_{ik}) = \sigma_\delta^2 \rho^{|t_j - t_k|}$.

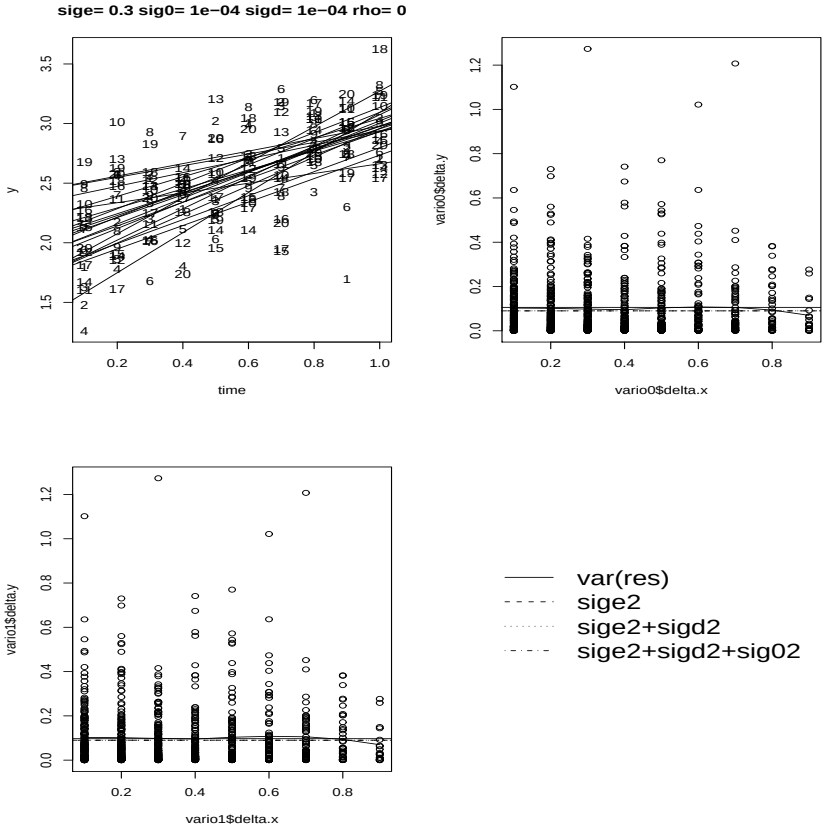

Figure 12: Measurement error only.

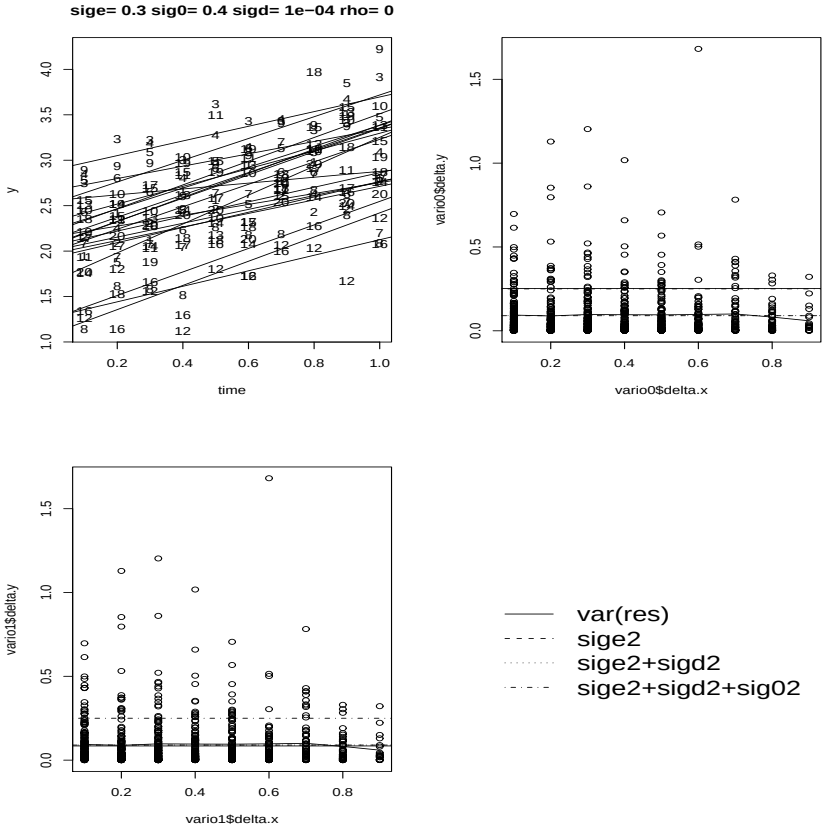

Figure 13: Measurement error and random effect for intercept.

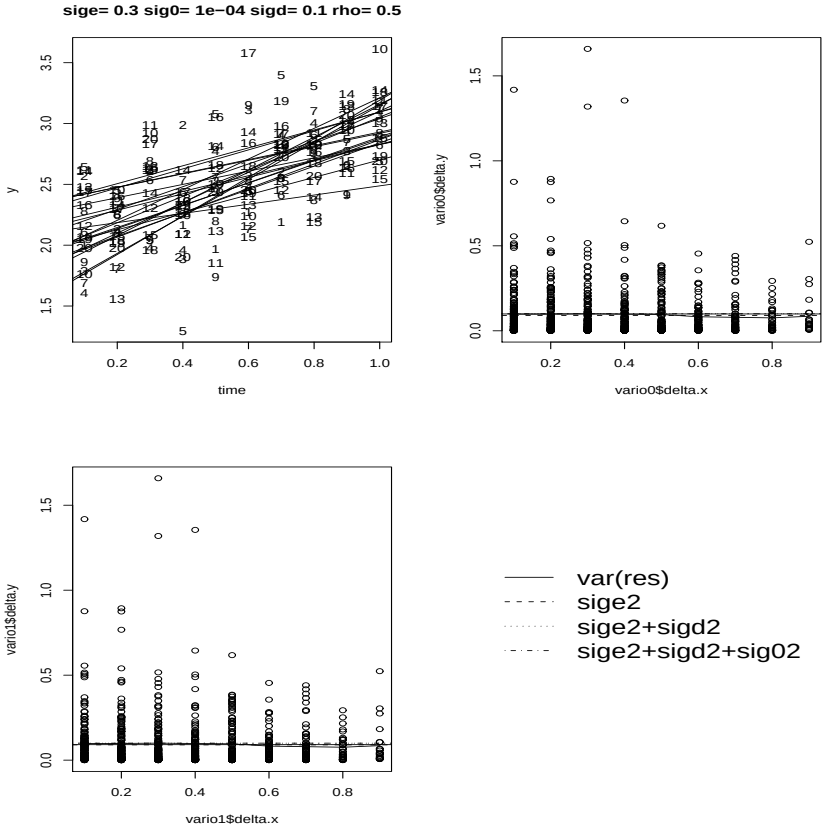

Figure 14: Measurement error and weak serial correlation.

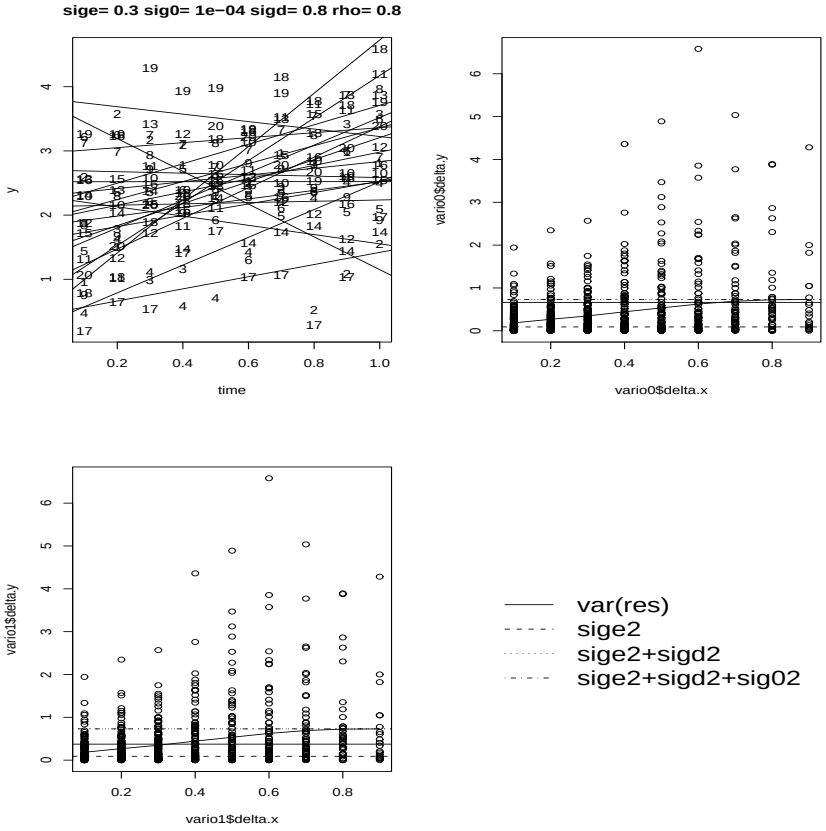

Figure 15: Measurement error and strong serial correlation.

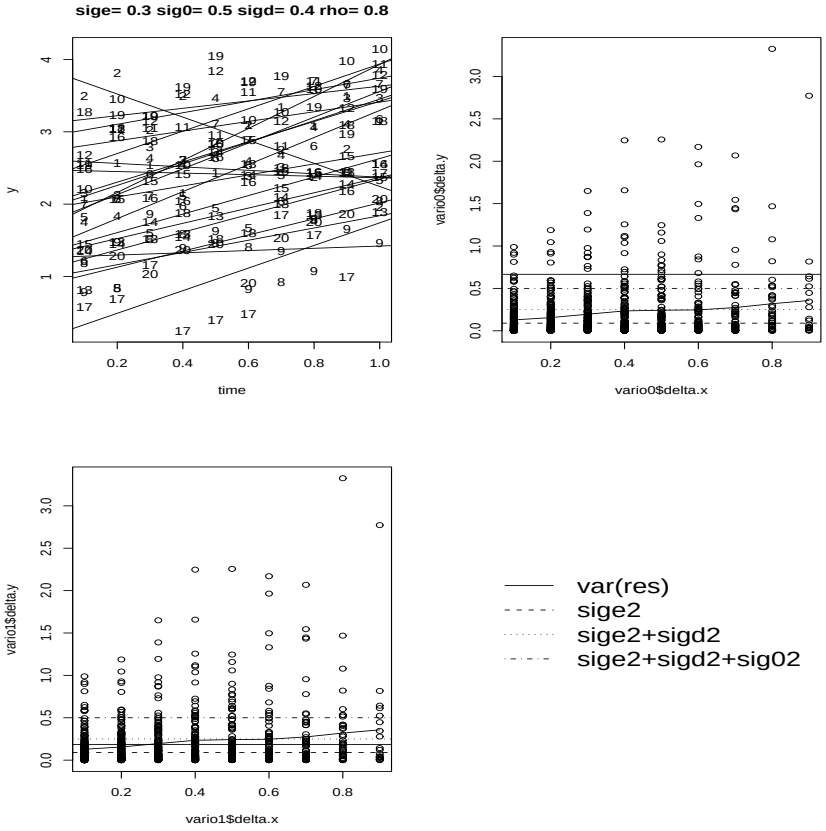

Figure 16: Measurement error and strong serial correlation and random effects.

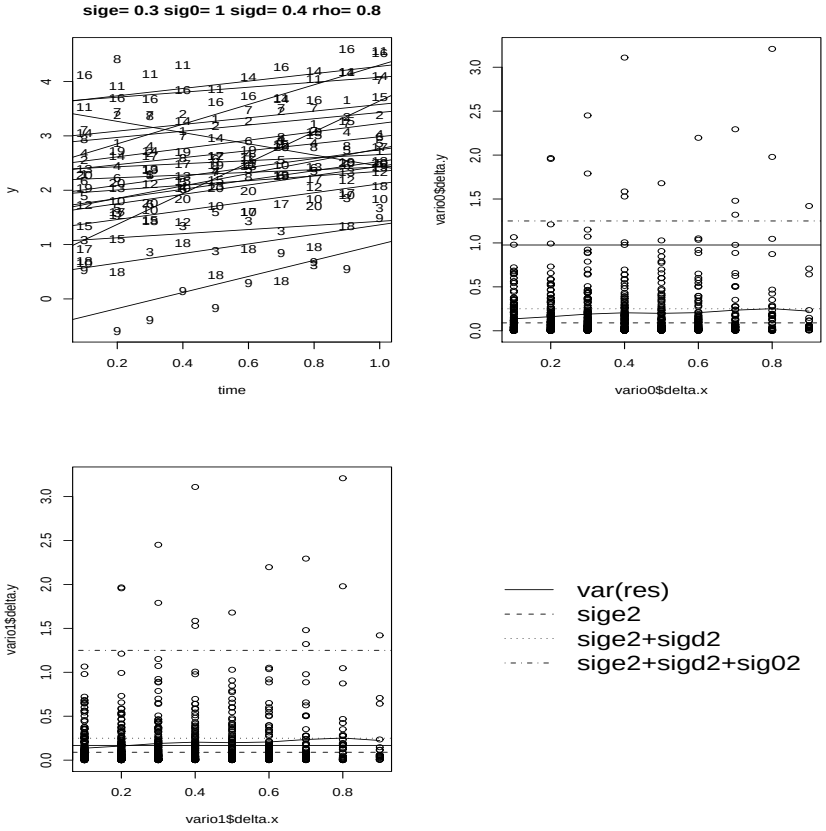

Figure 17: Measurement error and serial correlation and strong random effects.

The use of $\boldsymbol{b}_i$ in residual analysis

Predictions of the random effects $\widehat{\boldsymbol{b}}_i$ may be used to assess assumptions associated with the random effects distribution, in particular:

- Are the random effects normally distributed?
- If we have assumed independence between random effects, does this appear reasonable?
- Is the variance of the random effects independent of covariates $\boldsymbol{x}_i$?

It should be born in mind that interpretation of random effects predictions is more difficult since they are not direct functions of the data.

Recall that $\widehat{\boldsymbol{b}}_i$ are shrinkage estimators, and the amount of shrinkage is determined by the specific form assumed for the estimator (following from normality of random effects, or linearity of estimator, for example).

Hence assumptions about $\boldsymbol{b}_i$ may not be reflected in $\widehat{\boldsymbol{b}}_i$.

We may fit curves for particular individuals with n_i large, and then check the assumptions from these.

Example: Growth Curves

We now carry out a more comprehensive analysis of the dental growth data.

Suppose

- The primary aim is to answer the question of differences in average growth between boys and girls.
- A secondary aim is exploration/description of sources of variability in the data.

Figures 18–20 show various diagnostics that are based on individual fits to each boy and girl.

Figures 18 and 19 are useful for assessing the random effects distribution.

Figure 20 is useful for assessing the measurement error distribution.

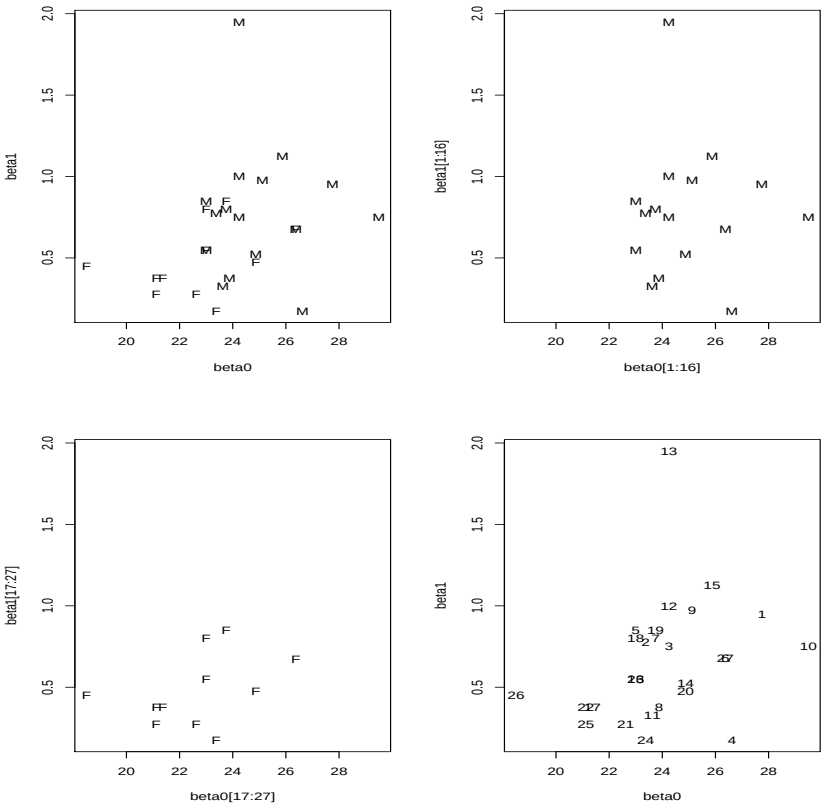

Figure 18: Bivariate $\widehat{\beta}_0, \widehat{\beta}_1$.

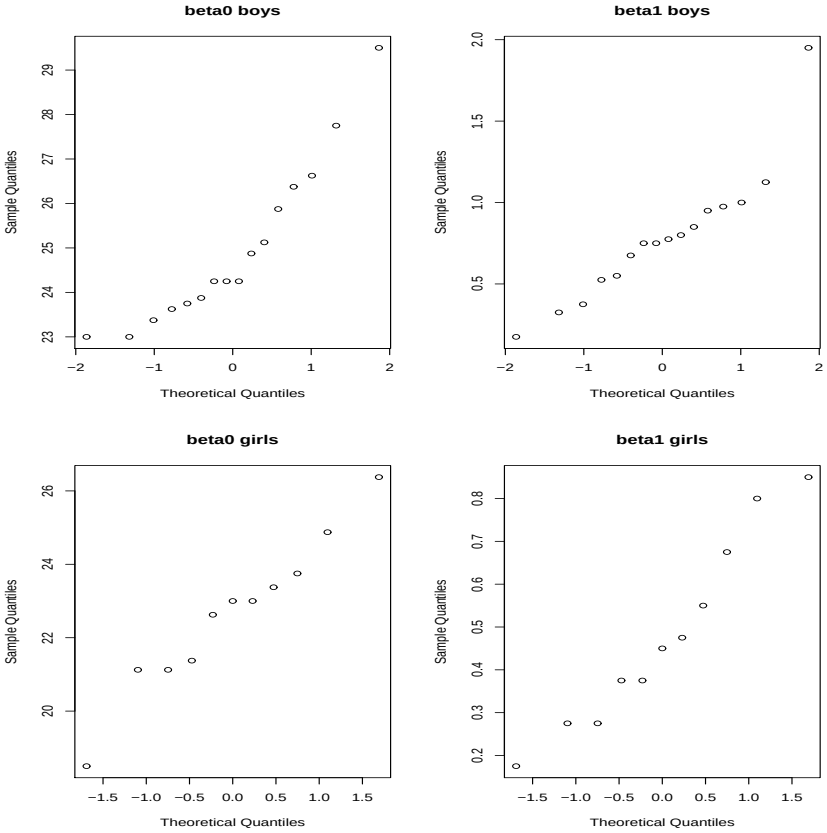

Figure 19: Univariate $\widehat{\beta}_0, \widehat{\beta}_1$.

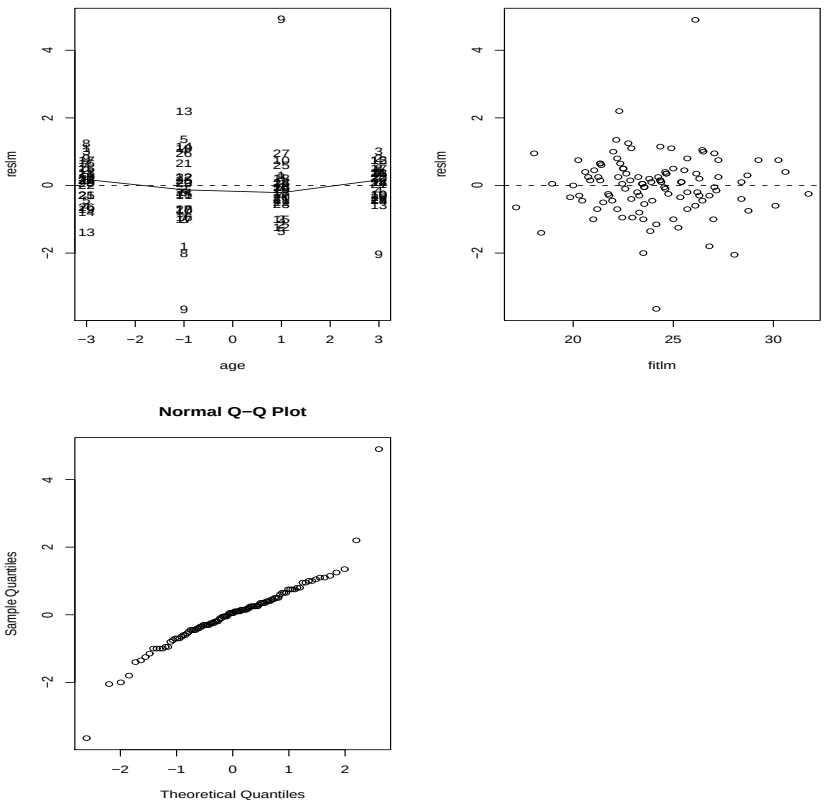

Figure 20: Residuals.

More diagnostics

Statistics:

IQR for $\widehat{\beta}_0$ for boys: 23.72–26.00, for girls: 21.25–23.56.

IQR for $\widehat{\beta}_1$ for boys: 0.54–0.96, for girls: 0.33–0.61.

$\text{sd}(\widehat{\beta}_0) = 1.83$ for boys and $\text{sd}(\widehat{\beta}_0) = 2.10$ for girls.

$\text{sd}(\widehat{\beta}_1) = 0.41$ for boys and $\text{sd}(\widehat{\beta}_1) = 0.22$ for girls.

Correlation $\widehat{\beta}_0, \widehat{\beta}_1$ is -0.00 for boys and 0.38 for girls.

Std dev of residuals for boys is 1.15, and for girls is 0.48.

Observations:

- Boy 10 is a slope outlier.
- Not clear that the random effects distriubtions are the same for boys and girls.
- Not clear we need random effects for slopes.
- Not clear measurement error is same for boys and girls.

We fit separate LMEM to each of the boys and girls data using REML.

Boys:

```
> modb <- lme( distance ~ I(age-11), data = Orthboy ,
+ random= ~I(age-11) | Subject )
> summary(modb)
Linear mixed-effects model fit by REML
Random effects:
 Formula: ~I(age - 11) | Subject
            StdDev   Corr
(Intercept) 1.642412 (Intr)
I(age - 11) 0.188629 -0.011
Residual    1.609114
Fixed effects: distance ~ I(age - 11)
               Value Std.Error DF  t-value p-value
(Intercept) 24.968750 0.4572219 47 54.60970       0
I(age - 11)  0.784375 0.1015638 47  7.72298       0
 Correlation:
            (Intr)
I(age - 11) -0.005
```


Girls:

```
> modg <- lme( distance ~ I(age-11), data = Orthgirl ,
+ random= ~I(age-11) | Subject )
> summary(modg)
Linear mixed-effects model fit by REML
Random effects:
 Formula: ~I(age - 11) | Subject
 Structure: General positive-definite, Log-Cholesky parametrizatio
            StdDev    Corr
(Intercept) 2.0775608 (Intr)
I(age - 11) 0.1612351 0.527
Residual    0.6678487
Fixed effects: distance ~ I(age - 11)
               Value Std.Error DF  t-value p-value
(Intercept) 22.647727 0.6344478 32 35.69675       0
I(age - 11)  0.479545 0.0662625 32  7.23706       0
 Correlation:
            (Intr)
I(age - 11) 0.381
```

CHAPTER 11: MODEL SELECTION/FORMULATION

Very context specific; we consider three distinct scenarios:

1. Confirmatory analyses in which an a priori hypothesis concerning a particular response/covariate relationship is of interest, and other variables have been measured and we wish to know which to adjust for.
2. Exploratory analyses where the aim is to gain clues as to structure in the data. For example, which covariates are causally related to a response, or characterizing sources of variability.
3. Prediction in which we are not concerned with causality, but merely with predicting a response given a set of variables.

We define the *null* model as that which contains an intercept only, the *minimal* model as the smallest model which is consistent with prior information.

So for example, in an epidemiological investigation we would almost always want to include terms for age and gender.

The minimal model may also be a function of the design so in matched case-control studies we include a term for each of the matching sets. Similarly in clinical trials in which treatments are randomized within a priori chosen strata, we again will include a term for strata.

With respect to the first two aims in particular, an understanding of causality and confounding is very useful in formulating models.

Causality and Confounding

We now discuss the choice of the mean model, via the concepts of causality and confounding.

Advantages of causal examination of models:

1. Graphical examination of variables can clarify assumptions/relationships between variables.
2. Understand which variables to control for.
3. In some situations the method of analysis is different from the conventional models we are used to fitting, e.g. time-varying exposures, adjustment for lack of compliance, missing data.
4. Interpretation of coefficients is aided.

Confounding

Conceptually, confounding is the confusion or mixing of effects due to variables other than those of primary interest.

In an observational study we cannot simply compare responses in exposed and unexposed populations because those populations will differ, in general, in other risk factors.

We say that the comparison of exposed and unexposed is *confounded* because the difference in responses results from a mixture of several effects, including but not limited to the exposure effect (Rothman and Greenland, 1998).

Confounding

Rothman and Greenland (1998) give the following criteria for a confounder:

1. A confounding factor must be a risk factor for the response.
2. A confounding factor must be associated with the exposure under study in the source population.
3. A confounding factor must not be affected by the exposure or the response. In particular it cannot be an intermediate step in the causal path between the exposure and the response.

Note that if a variable is assigned its value before the exposure is assigned, and before the response occurs, then it cannot be caused by either exposure or response.

Example: Confounding variable to adjust for

Suppose Y is the rate of lung cancer, X smoking rate, and Z diet and alcohol variables.

In this case Z is a confounder under the above definition since it satisfies 1.–3. The *causal diagram* in Figure 21 below illustrates one plausible mechanism for this situation, U could represent education level (or poverty) here.

If we obtain data on X, Z, Y then we will see an association between X and Y, but also between Z and Y, hence we must control for Z.

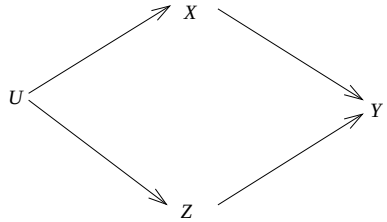

Figure 21: U denotes unmeasured variables.

Suppose $\mathrm{E}[Y|X,Z] = \alpha + \beta X + \gamma Z$ and $\mathrm{E}[Z|X] = a + bX$.

Then $\mathrm{E}[Y|X] = \alpha^\star + \beta^\star X$ where $\alpha^\star = \alpha + \gamma a$ and $\beta^\star = \beta + \gamma b$ showing overestimation of effect if b and γ of the same sign. Also shows that Z needs to be related to both X (through b) and Y (through γ) (much more complex with non-linear response model).

Example: Confounding variable to adjust for

Consider a cross-sectional study carried out to estimate the causal effect of smoking (X) on lung cancer (Y). A number of other variables would typically be measured in such a study, for simplicity we consider the additional variable gender (Z) only. We may hypothesize that Y is related to both X and Z and that men are more likely to smoke than women. This information may be summarized in the causal diagram in Figure 22. In this figure, U denotes an unobserved variable. For example U may include information on parental smoking. In this situation we need to adjust for Z in order to obtain an unbiased effect of exposure (smoking).

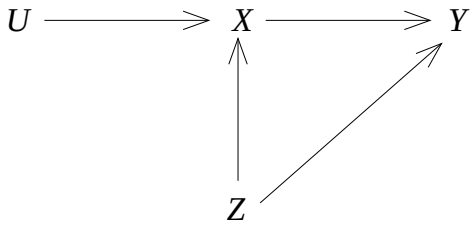

Figure 22: U denotes unmeasured variables.

We now consider further examples to illustrate some of the situations in which care must be taken in interpretation.

Example: Variables on the Causal Pathway

We first give an example of a variable that satisfies 1. and 2. but not 3.

In Figure 23, U denotes unobserved variables, X is smoking, Z is a variable representing tar deposits, and Y is lung cancer. If we looked at the marginal associations we would find relationships between X and Y but also between Z and Y. In this case we should not adjust for Z because this would dilute the causal effect of X on Y. In this example Z is not a confounder because it is on the causal pathway between X and Y (thus invalidating criteria 3.), Z is known as an *intermediary variable*.

Figure 23: Z is on the causal pathway.

Message: don't control for intermediary variables.

Example: Variables Affected by the Response

To further illustrate variables that satisfy 1. and 2. but contradict 3., we consider an example given by Greenland, Pearl and Robins (1999) in which Y represents endometrial cancer, X estrogen and Z uterine bleeding. The latter could be caused by X or Y and so, under this scenario, we have the causal diagram represented by Figure 24. Again we should not adjust for Z since the estimated causal effect of X on Y would be reduced. Note that we would observe marginal associations between X and Z and Y and Z and so 1. and 2. are satisfied.

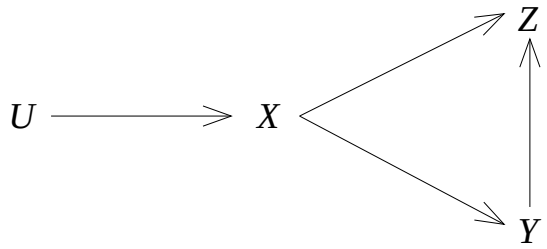

Figure 24: Variable Z afftected by the response.

References

Greenland, S., Pearl, J. and Robins, J.M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10, 37–48.

Lauritzen, S.L. (2000). Causal inference from graphical models. In *Complex Stochastic Systems*, eds. O.E. Barndorff-Nielsen, D.R. Cox and C. Klúppelberg, Chapman and Hall, London.

Little, R.J. and Rubin, D.B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes. *Annual Review of Public Health*, 21, 121–145.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference,* Cambridge University Press, Cambridge.

Robins, J.M., Greenland, S. and Hu, F.-C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association*, 94, 687–712.

Rosenbaum, P.R. and Rubin, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516–524.

Rothman, K.J. and Greenland, S. (1998). *Modern Epidemiology, Second Edition* . Lipincott-Raven.

Selection of Regressors

Recall our brief discussion of the mean-variance trade-off in the chapter on linear models.

Trade-off: as we include more covariates, bias is reduced, but variability may be increased, dependent on how strong a predictor the covariate is (and its association with other covariates).

We now describe some of the approaches to subset selection that have been proposed in the literature:

Forward selection. Begins with the simplest model. At each stage the 'best' unselected variable that satisfies the selection criterion is added. Best here is defined to be that variable whose deviance (or Wald or score statistic) is largest. This variable is added to the regression if its statistic is greater than a threshold of a specified significance level. This value, is contentious. Note that a maximum of p models will be considered in this procedure (out of 2^p).

Backward elimination. Begins with the full model. At each stage the covariate with the smallest deviance value that is less than a specified value is removed.

Stepwise regression (Efroymson's algorithm). Follows forward selection with backward elimination.

Difficulties

There are a number of problems with selection methods (Miller, 1990). In the conventional use of a hypothesis test, for the correct interpretation of significance levels the hypotheses must be specified before the data are examined.

Similarly for interval estimates to be valid the model must be specified *a priori*.

There now exists great potential for over-fitting in which models become too dataset-specific as they are refined on the basis of the examination of diagnostics.

In practice, if refinement is carried out through the fitting of alternative models (e.g. transformation of covariates, choice of distribution for the responses), then interval estimates will often be too narrow since they are produced by conditioning on the final model, and hence do not reflect the mechanism by which the model was selected.

Frequentist model selection difficulties

From a frequentist standpoint estimators and test statistics should be examined via their long-run behaviour *given* the model-fitting process, including refinement. To be more explicit, let P denote the procedure by which a final model M is decided upon. Then suppose it is of interest to examine the bias of a statistic T,

$$\mathrm{E}[T|P] = \mathrm{E}_{M|P}\{\mathrm{E}[T|M]\}. \tag{13}$$

In general it will be incorrect to report $\mathrm{E}[T \mid \widehat{M}]$ where $\widehat{M}$ is the final model chosen, since this does not reflect the procedure by which $\widehat{M}$ was chosen, but rather acts as if the final model is the "truth".

We know that

$$\mathrm{var}(T|P) = \mathrm{E}_{M|P}[\mathrm{var}(T|M)] + \mathrm{var}_{M|P}(\mathrm{E}[T|M]).$$

but $\mathrm{var}(T|\widehat{M})$ is reported (which approximates the first term only).

Under a frequentist approach inference follows from the behaviour of an estimator under repeated sampling from the true model, and if an initial model is clearly wrong on the basis of a residual plot (say), then it is very unlikely to be close to the "true" model and hence it is more appropriate to obtain properties of estimators under the assumed model.

Bayesian model selection difficulties

From a Bayesian standpoint the same problem of dredging exists because the posterior distribution should reflect all sources of uncertainty and *a priori* all possible models that may be entertained should be explicitly stated, with prior distributions being placed upon different likelihoods and the parameters of these likelihoods.

Model averaging (see later) should then be carried out across the different possibilities, a process which is fraught with difficulties not least in placing "comparable" priors over what may be fundamentally different objects.

(One solution is to place prior on "model-free" quantities.)

Model Averaging

Suppose the action concerns a parameter of interest ω (which for simplicity we assume is univariate) that is well-defined for all models.

We have

$$\mathrm{E}[\omega|\boldsymbol{y}] = \sum_{j=1}^{J} \mathrm{E}[\omega|\boldsymbol{y}, M_j] \times \Pr(M_j|\boldsymbol{y}),$$

and

$$\begin{aligned} \mathrm{var}(\omega|\boldsymbol{y}) &= \sum_{j=1}^{J} \mathrm{var}(\omega|\boldsymbol{y}, M_j) \times \Pr(M_j|\boldsymbol{y}) \\ &+ \sum_{j=1}^{J} \{\mathrm{E}[\omega|\boldsymbol{y}, M_j] - \mathrm{E}[\omega|\boldsymbol{y}]\}^2 \times \Pr(M_j|\boldsymbol{y}). \end{aligned}$$

This latter term shows how not only parameter uncertainty but *model* uncertainty is accounted for.

- Specification of priors is not trivial.
- Interpretation.
- Continuous model expansion.

A possible compromise

One solution is to never refine the model for a given data set. This approach is operationally pure but pragmatically dubious (unless one is in the context of a randomized experiment) since we may obtain appropriate inference for a model that is a very poor description of the phenomenon under study.

The philosophy suggested here is to think as carefully as possible about the initial model class before the analysis proceeds, but after fitting to carry out model checking and refine the model in the face of *clear* model misspecification, with refinement ideally being carried out within distinct *a priori* known classes.

With reference to (13), if a model is chosen because it is clearly superior to the alternatives, then it may be reasonable to assume that $\mathrm{E}[T \mid P] \approx \mathrm{E}[T \mid \widehat{M}]$, because $\widehat{M}$ would be consistently chosen in repeated sampling under these circumstances.

So, for example, examining quantile-quantile plots for different t distributions and picking the one that produces the straightest line would not be a good idea.

Inference then proceeds as if the final model were the one that were chosen initially. This is clearly a subjective procedure but can be informally justified via either philosophical approaches.

In a similar vein, under a Bayesian approach the above procedure is consistent with model-averaging but with the posterior model weight being concentated upon the chosen model (since alternative models are only rejected on the basis of clear inadequacy).

The aim is to provide probability statements, from either philosophical standpoints that are "honest" representations of uncertainty. The above approach is relevant to analyses that are more confirmatory in their outlook, as opposed to being used for prediction, or for more exploratory purposes (for example, to gain clues to models that may be appropriate for future data analyses).

We illustrate some of the difficulties of model selection with two simple examples.

Example 1: If we carry out a *single* hypothesis test and only report the estimate of β_1 in a simple linear regression *if* the null hypothesis of $\beta_1 = 0$ is rejected. Figure 25 results – the bias is clear. There are close links with publication bias in meta-analysis.

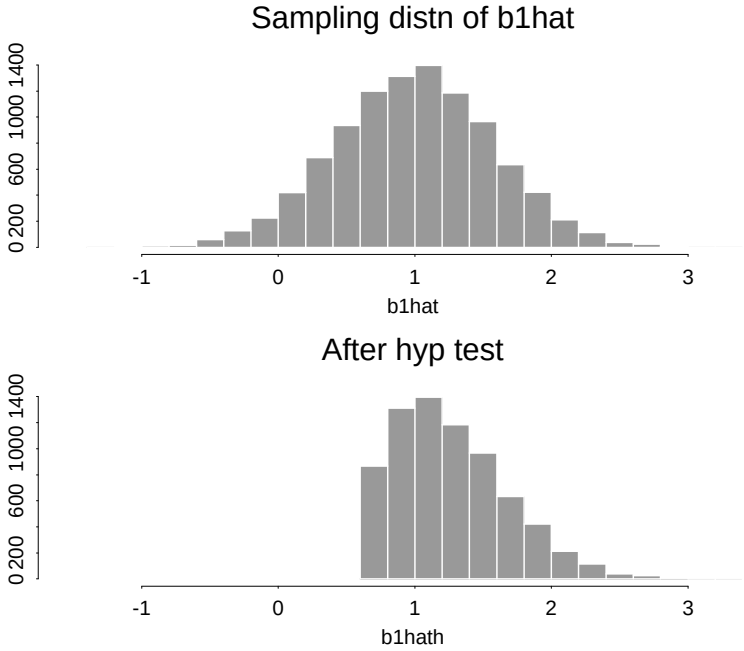

Figure 25: $\mathrm{E}[\widehat{\beta}_1] = 1.00$, while $\mathrm{E}[\widehat{\beta}_1|$ rejection of $H_0] = 1.27$.

Example 2: Suppose we are interested in β_1 but we wish to "control" for β_2 by testing whether the latter is significant. In the following simulation, a multiple linear regression in X_1 and X_2 was carried out. The true values were $\beta_1 = \beta_2 = 1$ and X_1, X_2 were simulated from a bivariate normal with means zero, variances one, and correlation 0.7. Figure 26 shows the results, in (a) we display the sampling distributions of $\widehat{\beta}_1$ from the adjusted model. The mean and standard deviation of the distribution of $\widehat{\beta}_1$ are 1.00 and 1.23 Panel (b) displays the sampling distribution of the *reported* estimator. The mean and standard deviation of the distribution of the reported estimate of β_1 are 1.23 and 1.01, respectively, showing positive bias and a reduced variance.

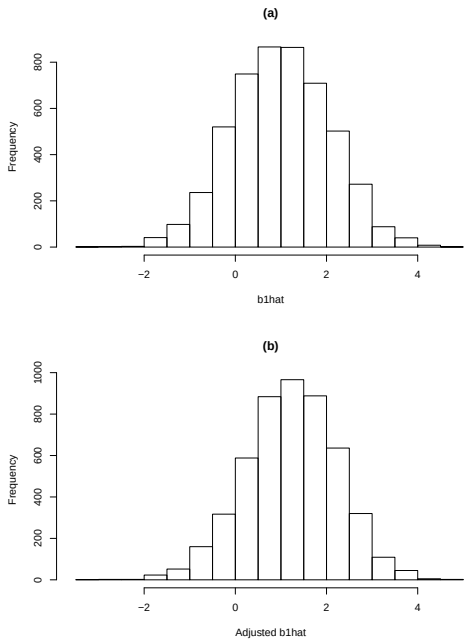

Figure 26: (a) Sampling distribution of $\widehat{\beta}_1$, (b) sampling distribution of $\widehat{\beta}_1$ given "control" for the possibility that $\beta_2 \neq 0$.

.

The above procedures assess the relative importance of regressors, an alternative is to examine:

All possible subsets: select that model that satisfies some criterion.

For nested models the $R^2(= 1 - RSS/CTSS)$ measure is non-decreasing in the number of variables and so will always suggest the most complex model. The adjusted R^2 measure defined by

$$R^2_a = 1 - \frac{RSS/(n-p-1)}{CTSS/(n-1)} = 1 - (1-R^2)\left(\frac{n-1}{n-p-1}\right),$$

is more useful and leads to the selection of the model that produces the smallest estimate of σ^2.

A widely-used statistic is that suggested by Mallows and defined by

$$C_p = \frac{RSS_p}{s^2_k} - (n-2p) = \frac{(n-p)s^2_p}{s^2_k} - (n-2p),$$

where RSS_p is the residual sum of squares from a model containing p parameters (including β_0), and s^2_k is the estimate from the maximal model. This criteria may be derived via consideration of the prediction error that results from choosing the model under consideration.

Full model: include all p regressors. Recall mean-variance trade-off. Increased standard errors if lots of regressors.

Bayesian shrinkage Remove regressors that are, a priori, thought to be of no use, for those that are definitely important put flat priors on the regressors, for those we're not sure about, put a prior centered on zero with a "small" variance. In this way these coefficients are shrunk towards zero and so to be important the association in the data must be really strong.

Closely related to ridge regression.

Penalized likelihood ratio statistics provide another set of criteria.

Penalized LR statistics

Consider nested models M_0 and M_1 with parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ of dimensionality d_0 and d_1, respectively,

The classical likelihood ratio procedure states that we should reject model M_0 if D_{10} is sufficiently large, as measured relative to a $\chi^2_{d_1-d_0}$ distribution, where $D_{10} = -2\log LR$ and

$$LR = \frac{L(\widehat{\boldsymbol{\theta}}_0)/L(\widehat{\boldsymbol{\theta}}_s)}{L(\widehat{\boldsymbol{\theta}}_1)/L(\widehat{\boldsymbol{\theta}}_s)}.$$

This procedure is inconsistent (in the sense of fixed type I error rate).

- The likelihood ratio approach is flawed if a fixed α (Type I error rate) is chosen, irrespective of n since as n increases β (Type II error rate) drops and so the alternative hypothesis (i.e. the more complex model) is favored.
- A more sensible procedure would decrease α as a function of n but it is not clear how to do this. The following illustrates how the Bayesian Information Criteria (BIC, or Schwarz criteria) achieves this automatically.

Laplace Approximation to the Bayes factor

To calculate the evidence in the data for models M_0 and M_1 we may calculate the Bayes factor $B_{01} = p(\boldsymbol{y}|M_0)/p(\boldsymbol{y}|M_1)$.

Recall that Laplace's approximation of

$$I = \int \exp\{ng(\boldsymbol{\theta})\}\, d\boldsymbol{\theta},$$

where $\dim(\boldsymbol{\theta}) = d$, is given by

$$\widetilde{I} = \left(\frac{2\pi}{n}\right)^{d/2} \exp\{ng(\widetilde{\boldsymbol{\theta}})\} \mid \widetilde{g}_2 \mid^{-1/2},$$

where $\widetilde{\boldsymbol{\theta}}$ is the value that maximizes $g(\boldsymbol{\theta})$, and

$$\widetilde{g}_2 = -\left.\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} g(\boldsymbol{\theta})\right|_{\widetilde{\theta}}.$$

Now suppose we wish to evaluate

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

and so let

$$g(\boldsymbol{\theta}) = \frac{1}{n}\log p(\boldsymbol{y} \mid \boldsymbol{\theta}) + \frac{1}{n}\log \pi(\boldsymbol{\theta}),$$

and in the Laplace approximation, $\widetilde{\boldsymbol{\theta}}$ corresponds to the posterior mode.

For large n, $\widetilde{\boldsymbol{\theta}} \rightarrow \widehat{\boldsymbol{\theta}}$, the MLE, and for iid data

$$\widetilde{g}_2(\widetilde{\boldsymbol{\theta}}) \rightarrow -\frac{1}{n} \left. \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) \right|_{\widehat{\theta}} \rightarrow \boldsymbol{I}_1(\widehat{\boldsymbol{\theta}}),$$

the expected information in a single observation, evaluated at the MLE.

Hence

$$\widehat{I} = \left(\frac{2\pi}{n} \right)^{d/2} p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}) \pi(\widehat{\boldsymbol{\theta}}) \mid \boldsymbol{I}_1(\widehat{\boldsymbol{\theta}}) \mid^{-1/2}, \qquad (14)$$

provides an approximation to $p(\boldsymbol{y})$.

We want an approximation to the Bayes factor

$$\mathrm{BF}_{01} = \frac{p(\boldsymbol{y} \mid M_0)}{p(\boldsymbol{y} \mid M_1)}.$$

Applying (14) to numerator and denominator gives

$$\widehat{\mathrm{BF}}_{01} = \frac{(2\pi/n)^{d_0/2}}{(2\pi/n)^{d_1/2}} \frac{p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}_0)}{p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}_1)} \frac{\pi(\widehat{\boldsymbol{\theta}}_0)}{\pi(\widehat{\boldsymbol{\theta}}_1)} \frac{\mid \boldsymbol{I}_0(\widehat{\boldsymbol{\theta}}_0) \mid^{1/2}}{\mid \boldsymbol{I}_1(\widehat{\boldsymbol{\theta}}_1) \mid^{1/2}}.$$

Note the dependence on the ratio of priors.

Bayesian Information Criteria

Now consider

$$\begin{aligned} -2 \log \widehat{BF}_{01} &= (d_1 - d_0) \log n + 2\{\log p(\boldsymbol{y} \mid \widehat{\theta}_1) - \log p(\boldsymbol{y} \mid \widehat{\theta}_0)\} \\ &= D_{10} - (d_1 - d_0) \log n + c \end{aligned}$$

where

$$D_{10} = 2 \log \frac{L(\widehat{\boldsymbol{\theta}}_1)}{L(\widehat{\boldsymbol{\theta}}_0)} > 0$$

is the deviance and

$$c = (d_1 - d_0) \log(2\pi) + 2 \log \left(\frac{\pi(\widehat{\theta}_1)}{\pi(\widehat{\theta}_0)} \right) + \log \frac{\mid \boldsymbol{I}_1 \mid}{\mid \boldsymbol{I}_0 \mid},$$

and is $O(1)$ (for fixed d_0, d_1).

Schwarz (1978) suggested using

$$S = D_{10} - (d_1 - d_0) \log n,$$

as a test criterion (i.e. setting $c = 0$). This is also known as the *Bayesian Information Criterion* (BIC).

Model M_0 is preferred if $S < 0$, model M_1 if $S > 0$.

Example: effect of sample size on BIC.

Suppose $d_1 - d_0 = 1$ so that the larger model has one more parameter than that nested within it.

A LR test with type I error $\alpha = 0.1$ would suggest model M_1 (i.e. reject M_0) if $D_{10} > 2.706$ (90% point of a χ^2_1).

Using BIC there is evidence for model M_1 if

$$S = D_{10} - \log n > 0,$$

i.e. if

$$D_{10} > \log n.$$

n	$\log n$	α
5	1.609	0.20
$e^{2.706}$	2.706	0.10
20	3.000	0.0833
100	4.605	0.032
1000	6.908	0.0086

Table 2: Comparison between a LR test and BIC.

Message is that BIC, form a frequentist perspective, has the effect of providing an α that decreases with n, and so as n increases favors more simple models when compared to the likelihood ratio approach.

This approximation may be used for model averaging also. If we have models M_j, $j = 0, ..., J$ and Bayes factors B_{j0} then

$$\Pr(M_j|\boldsymbol{y}) = \frac{\alpha_j B_{j0}}{\sum_{j=0}^{J} \alpha_j B_{j0}},$$

where $\alpha_j = \Pr(M_j)/\Pr(M_0)$ is the prior odds for M_j against M_0, $j = 0, ..., J$.

BIC

- BIC provides, from a frequentist perspective, a *consistent* model selection procedure.
- In general $\exp(S/2)$ provides an $O(1)$ approximation to B_{01}.
- Under a certain "unit information prior" the approximation is $O(n^{-1/2})$, however (Kass and Wasserman, 1995).
- In general, the choice of sample size is not trivial (e.g. survival analysis, hierarchical models).

General Criteria

The *Akaike Information Criteria* (AIC) for model j is given by

$$AIC_j = D(\widehat{\boldsymbol{\theta}}_j) + 2d_j,$$

where $D(\widehat{\boldsymbol{\theta}}_j) = 2\{L(\widehat{\boldsymbol{\theta}}_s) - L(\widehat{\boldsymbol{\theta}}_j)\}$, $j = 0, 1$ and $\boldsymbol{\theta}_s$ denote the parameters of the saturated model. For model comparison we examine

$$\begin{aligned} A_{10} &= AIC_0 - AIC_1 \\ &= \{-2\log L(\widehat{\boldsymbol{\theta}}_0) + 2d_0\} - \{-2\log L(\widehat{\boldsymbol{\theta}}_1) + 2d_1\} \\ &= D_{10} + 2(d_0 - d_1). \end{aligned} \tag{15}$$

Model M_0 (M_1) is preferred if $AIC_{10} < (>)0$, i.e. if $D_{10} < (>)2(d_1 - d_0)$.

Notes:

- AIC is, from a frequentist perspective, inconsistent.
- If $d_1 - d_0 = 1$ then AIC corresponds to a likelihood ratio test with $\alpha = 0.157$.

Prostate Cancer Example

For this example, forward selection, backwards elimination and stepwise regression all lead to the same model with for example lcavol being the most significant variable (F=111.3) followed by lweight (F=10.6), svi (F=10.1) and lbph (F=2.57). The latter may or not be included depending on the F-to-enter value. With the three most significant variables in the model, the coefficients listed in Table 3 were obtained. The three estimated coefficients all have smaller standard errors, though the uncertainty in the model search has not been acknowledged. We see that the estimated standard deviation is also smaller.

Figure 27 plots the C_p value versus number of parameters in the model. Here we pick out the model which gives C_p beneath the $C_p = p$ line, here occurring for the model with the five variables lcavol, lweight, age, lbph and svi.

	Full model			Stepwise/BIC model			C_p/AIC model		
Variable	Estimate	St. Er.	T score	Estimate	St. Er.	T score	Estimate	St. Er.	T score
lcavol	0.5870	0.0879	6.6768	0.5516	0.0747	7.3879	0.5656	0.0746	7.5829
lweight	0.4545	0.1700	2.6731	0.5085	0.1502	3.864	0.4237	0.1669	2.5390
age	-0.0196	0.0112	-1.7576	–	–	–	-0.0149	0.0108	-1.3847
lbph	0.1071	0.0584	1.8316	–	–	–	0.1118	0.0581	1.9265
svi	0.7662	0.2443	3.1360	0.6662	0.2098	3.1756	0.7210	0.2090	3.4492
lcp	-0.1055	0.0910	-1.1589	–	–	–	–	–	–
gleason	0.0451	0.1575	0.2866	–	–	–	–	–	–
pgg45	0.0045	0.0044	1.0236	–	–	–	–	–	–
σ	0.7804	–	–	0.7168	–	–	0.7073	–	–

Table 3: Parameter estimates from various models for the prostate cancer data.

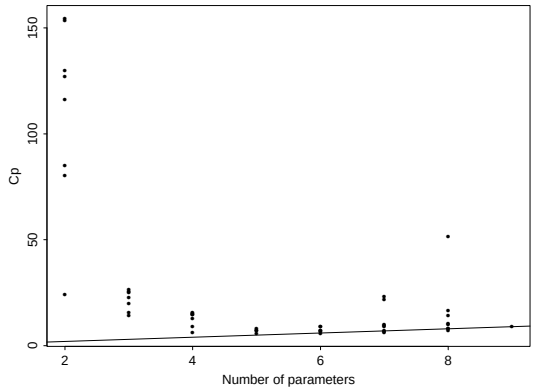

Figure 27: Mallows' C_p statistic plotted versus p, where $p-1$ is the number of covariates in the model. the line of equality is indicated since for a good model $\mathrm{E}[C_p] \approx p$.

Using the adjusted R^2, R^2_a to pick the best model (which recall is equivalent to picking that model with the smallest $\widehat{\sigma}^2$) gives a model with seven variables, gleason being that which is not included. This gave $\widehat{\sigma} = 0.7048$. The model giving the minimum AIC contained the variables lcavol, lweight, age, lbph and svi, i.e. the same as Mallows' C_p. The minimum BIC model was the same model as picked by the stepwise procedures.

We now describe a Bayesian analysis using an informative prior distribution. With the improper prior

$$\pi(\beta, \sigma^2) \propto \sigma^{-2},$$

inference is identical with the frequentist approach so that the values in columns 2 and 3 of Table 4 are also posterior means and posterior standard deviations.

Without more specific knowledge we here take the improper prior

$$\pi(\beta, \sigma^2) \propto \prod_{j=0}^{8} \pi(\beta_j) \times \sigma^{-2},$$

with $\pi(\beta_0) \propto 1$ and $\pi(\beta_j) \sim N(0, v_j)$. The standard deviations for the prior, $\sqrt{v}_j$, were chosen in the following way.

For the prostate data we believe that it is unlikely that any one covariate, over it's range (which we denote $x_{j\,\max}$), will change the log(PSA) by more than $Y_{\max} = 2$ units (on the log scale), giving a slope of $\beta_{\max} = |Y_{\max}/x_{\max}|$. The way we achieve this is by assuming that the approximate 95% point of the prior corresponds to the maximum value of β_j, $\beta_{\max}$, that we believe a priori is plausible. Formally we have

$$2\sqrt{v}_j = \beta_{\max} = \frac{Y_{\max}}{x_{j\,\max}} \quad \text{to give} \quad v_j = \frac{Y^2_{\max}}{2^2 X^2_{j\,\max}}.$$

	Prior model				Shrinkage model			
Variable	Median	St. Dev.	95% Interval		Median	St. Dev.	95% Interval	
lcavol	0	0.197	-0.387	0.387	0.4946	0.0806	0.3335	0.6495
lweight	0	0.273	-0.536	0.536	0.3472	0.1452	0.0592	0.6267
age	0	0.027	-0.053	0.053	-0.0133	0.0103	-0.0334	0.0072
lbph	0	0.275	-0.539	0.539	0.1116	0.0564	0.0006	0.2220
svi	0	1.020	-2.000	2.000	0.7660	0.2359	0.3039	1.2260
lcp	0	0.238	-0.466	0.466	-0.0386	0.0837	-0.2009	0.1278
gleason	0	0.340	-0.667	0.667	0.0577	0.1397	-0.0045	0.0110
pgg45	0	0.010	-0.020	0.020	0.0033	0.0039	-0.0045	0.0110
σ	–	–	–	–	0.7137	0.0554	0.6193	0.8370

Table 4: Prior and posterior summaries for the Bayesian shrinkage model for the prostate cancer data.

Conclusions

- For Confirmatory studies – try to avoid any model selection. Use background context to specify model.
- Exploratory studies – stepwise and all subsets may point to important variables, but attaching a p-value is difficult. Model averaging is another possibility.
- Prediction – some form of shrinkage should be used. Cross-validation may be used to choose a model. Bayesian model averaging also useful.

The Wishart Distribution

Consider the LMEM

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

with $\boldsymbol{b}_i \sim \mathrm{N}_{q+1}(\boldsymbol{0}, \boldsymbol{D})$, and $\boldsymbol{\epsilon}_i \sim \mathrm{N}_{n_i}(\boldsymbol{0}, \sigma^2_\epsilon \boldsymbol{I}_{n_i})$, $i = 1, ..., m$. A Bayesian analysis requires prior distributions on $\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2_\epsilon$; we assume independent priors

$$\pi(\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2_\epsilon) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{D})\pi(\sigma^2_\epsilon).$$

If $\boldsymbol{D}$ is a diagonal matrix with elements σ^2_k, $k = 0, 1, ..., q$, then a prior that leads to conjugate conditional distributions in a Gibbs sampling algorithm is

$$\pi(\sigma^2_0, ..., \sigma^2_q) = \prod_{k=0}^{q} \mathrm{IGa}(a_k, b_k),$$

where $\mathrm{IGa}(a_k, b_k)$ denotes the inverse gamma distribution with pre-specified parameters a_k, b_k, $k = 0, ..., q$.

For non-diagonal $\boldsymbol{D}$ we require a prior for the $(q+2)(q+1)/2$ elements, with the restriction that the resultant matrix is positive definite.

The conjugate choice is the so-called inverse Wishart distribution.

Suppose $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_r \sim_{iid} \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{S})$, where $\boldsymbol{S}$ is a non-singular variance-covariance matrix, and let

$$\boldsymbol{W} = \sum_{j=1}^{r} \boldsymbol{Z}_j \boldsymbol{Z}_j^{\mathrm{T}}. \tag{16}$$

Then $\boldsymbol{W}$ follows a Wishart distribution, denoted $\mathrm{W}_p(r, \boldsymbol{S})$, and

$$p(\boldsymbol{w}) = c^{-1} \mid \boldsymbol{w} \mid^{(r-p-1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{w}\boldsymbol{S}^{-1})\right\}$$

where

$$c = 2^{rp/2}\Gamma_p(r/2) \mid \boldsymbol{S} \mid^{r/2}, \tag{17}$$

with

$$\Gamma_p(r/2) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma((r+1-j)/2)$$

the generalized gamma function, and $n \geq p$ for a proper density. The mean is given by

$$\mathrm{E}[\boldsymbol{W}] = r\boldsymbol{S}.$$

The Wishart distribution is a multivariate version of the gamma distribution. Taking $p = 1$ yields

$$p(w) = \frac{(2S)^{-r/2}}{\Gamma(r/2)} w^{r/2-1} \exp(-w/2S),$$

for $w > 0$, the gamma distribution $\mathrm{Ga}(r/2, S/2)$. Further, taking $S = 1$ gives a χ^2_r random variable, which is clear from (16).

The Inverse Wishart Distribution

If $\boldsymbol{W} \sim \mathrm{W}_p(r, \boldsymbol{S})$, the distribution of $\boldsymbol{D} = \boldsymbol{W}^{-1}$ is known as the inverse Wishart distribution, and is given by

$$p(\boldsymbol{d}) = c^{-1} \mid \boldsymbol{d} \mid^{-(r+p+1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{d}^{-1}\boldsymbol{S})\right\}$$

where c is again given by (17). The mean is given by

$$\mathrm{E}[\boldsymbol{D}] = \frac{\boldsymbol{S}^{-1}}{r-p-1}$$

and is defined for $r > p+1$.

Note that if $p = 1$ we recover the inverse gamma distribution $\mathrm{IGa}(r/2, 1/2S)$.

We now consider a Gibbs sampling scheme and derive the conditional distribution for $\boldsymbol{W} = \boldsymbol{D}^{-1}$, with a $\mathrm{W}_{q+1}(r, \boldsymbol{R}^{-1})$ prior for $\boldsymbol{W}$.

Note that

$$\mathrm{E}[\boldsymbol{W}] = r\boldsymbol{R}^{-1},$$

and

$$\mathrm{E}[\boldsymbol{D}] = \boldsymbol{R}/(r-q-1-1),$$

so that $\boldsymbol{R}$, may be scaled to be a prior estimate of $\boldsymbol{D}$, with r acting as a strength of belief in the prior.

Conditional Conjugacy

Let $\boldsymbol{b} = (\boldsymbol{b}_1, ..., \boldsymbol{b}_m)$ and note that

$$\boldsymbol{b}^{\mathrm{T}}\boldsymbol{w}\boldsymbol{b} = \mathrm{tr}(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{w}\boldsymbol{b}) = \mathrm{tr}(\boldsymbol{w}\boldsymbol{b}\boldsymbol{b}^{\mathrm{T}}).$$

Then

$$\begin{aligned}
& p\ (\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{b}, \sigma^2_\epsilon) \propto \pi(\boldsymbol{b} \mid \boldsymbol{W}) \times \pi(\boldsymbol{W}) \\
\propto & \mid \boldsymbol{w} \mid^{m/2} \exp\left\{-\frac{1}{2}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{w}\boldsymbol{b}\right\} \mid \boldsymbol{w} \mid^{(r-q-1-1)/2} \exp\left\{\mathrm{tr}(\boldsymbol{w}\boldsymbol{R})\right\} \\
= & \mid \boldsymbol{w} \mid^{(m+r-q-1-1)/2} \exp\left\{-\frac{1}{2}\left[\boldsymbol{b}^{\mathrm{T}}\boldsymbol{w}\boldsymbol{b} + \mathrm{tr}(\boldsymbol{w}\boldsymbol{R})\right]\right\} \\
= & \mid \boldsymbol{w} \mid^{(m+r-q-1-1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{w}[\boldsymbol{b}\boldsymbol{b}^{\mathrm{T}} + \boldsymbol{R}])\right\}
\end{aligned}$$

Hence the conditional distribution is

$$\boldsymbol{W} \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{b}, \sigma^2_\epsilon) \sim \mathrm{W}_{q+1}\left\{r+m, \left(\boldsymbol{R} + \sum_{i=1}^m \boldsymbol{b}_i\boldsymbol{b}_i^{\mathrm{T}}\right)^{-1}\right\}.$$

So note that

$$\mathrm{E}[\boldsymbol{W} \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{b}, \sigma^2_\epsilon] = \frac{r+m}{\boldsymbol{R} + \sum_{i=1}^m \boldsymbol{b}_i\boldsymbol{b}_i^{\mathrm{T}}}.$$

Shows how r is acting like a sample size.

Issues with the Wishart Prior

- A problem with the Wishart distribution is that it is deficient in second moment parameters since there is only a single degrees of freedom parameter r. So, for example, it is not possible to have differing levels of certainty in the tightness of the prior distribution for different elements of $\boldsymbol{D}$. Note that with diagonal $\boldsymbol{D}$ and independent inverse gamma priors we have a precision parameter for each variance.
- The form of the conditional distribution suggests that it may be better to err on the side of picking $\boldsymbol{R}$ too small (though m is small, will always be influential).
- Intuition: as if our prior data for the precision consists of observing r normal random variables with variance-covariance matrices $\boldsymbol{R}$.
- We need to take $r \geq q+1$ for a proper prior, with the flattest prior corresponding to $r = q+1$. A proper prior is required to ensure propriety of the posterior distribution.
- Figure 28 displays samples from the Wishart distribution $W_2\{20, (20\boldsymbol{S})^{-1}\}$ where $\boldsymbol{S} = \begin{bmatrix} 0.4 & 0 \\ 0 & 1.0 \end{bmatrix}$. The mean is therefore $\mathrm{E}[\boldsymbol{W}] = \boldsymbol{S}^{-1} = \begin{bmatrix} 2.5 & 0 \\ 0 & 1.0 \end{bmatrix}$.

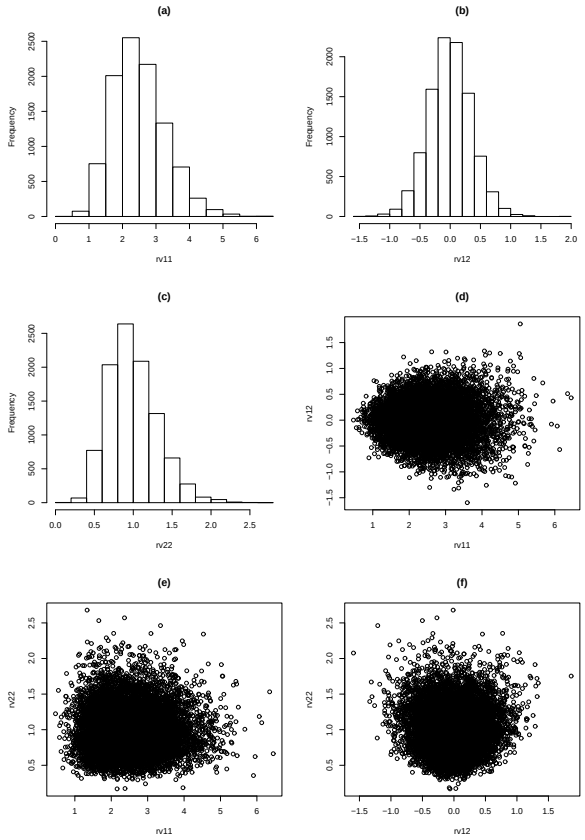

Figure 28: (a) Histogram of w_{11}, (b) Histogram of w_{12}, (c) Histogram of w_{22}, (d) Scatterplot of w_{11}, w_{12}, (e) Scatterplot of w_{11}, w_{22}, (f) Scatterplot of w_{12}, w_{22}

Example: Dental Data for Girls

Three-Stage Hierarchical Model:

First Stage:

$$y_{ij} = \beta_{0i} + \beta_{1i}(t_j - 11) + \epsilon_{ij},$$

with $\mathrm{N}(0, \sigma^2_\epsilon)$, $j = 1, ..., 4$, $i = 1, ..., 11$.

Second Stage: Let

$$\boldsymbol{\beta}_i = \begin{bmatrix} \beta_{0i} \\ \beta_{1i} \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \boldsymbol{D} = \begin{bmatrix} D_{00} & D_{01} \\ D_{10} & D_{11} \end{bmatrix},$$

and then

$$\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{D} \sim \mathrm{N}_2(\boldsymbol{\beta}, \boldsymbol{D}),$$

$i = 1, ..., m$.

Third Stage:

$$\pi(\sigma^2_\epsilon, \boldsymbol{\beta}, \boldsymbol{D}^{-1}) \propto \sigma_\epsilon^{-2} \times \mathrm{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 10^6 & 0 \\ 0 & 10^6 \end{bmatrix} \right) \times W_2(r, \boldsymbol{R}^{-1}).$$

Results below are for priors, with prior mean

$$\mathrm{E}[\boldsymbol{D}] = \begin{bmatrix} 1.0 & 0 \\ 0 & 0.1 \end{bmatrix}$$

and different degrees of freedom r.

We see sensitivity to the prior in inference for $\boldsymbol{D}$, but not for $\boldsymbol{\beta}$.

Note the greater shrinkage to the prior mean for the second and third priors, with $r = 28$ we have four times as much information in the prior, as in the data ($m = 11$).

r	$\boldsymbol{R}$	β_0	β_1
4	1.0 0 0 0.1	22.6 (21.4,23.8)	0.48 (0.33,0.63)
7	4.0 0 0 0.4	22.6 (21.5,23.7)	0.48 (0.31,0.65)
28	25 0 0 2.5	22.6 (21.8,23.5)	0.48 (0.28,0.67)

Table 5: Posterior medians and 95% intervals for population means, under three priors.

r	Diag $\boldsymbol{R}$	D_{00}	D_{01}	D_{11}
4	1.0 0.1	3.48 (1.66, 8.75)	0.13 (-0.10,0.54)	0.03 (0.01,0.10)
7	4.0 0.4	2.97 (1.51, 6.63)	0.10 (-0.14,0.46)	0.05 (0.02,0.12)
28	25 2.5	1.78 (1.14, 2.97)	0.04 (-0.10,0.20)	0.08 (0.05,0.14)

Table 6: Posterior medians and 95% intervals for population variances, under two priors.

The code below is for the analysis with $r = 4$, BUGS parametrizes the Wishart in terms of $\boldsymbol{R}^{-1}$ and r.

```
model
{
for( i in 1 : N ) {
for( j in 1 : T ) {
Y[i , j] ~ dnorm(mu[i , j],eps.tau)
mu[i , j] <- beta[i,1] + beta[i,2] * (x[j]-11)
}
beta[i,1:2] ~ dmnorm(beta.mu[1:2],iSigma[1:2,1:2])
}
beta.mu[1:2] ~ dmnorm(mean[1:2], prec[1:2, 1:2])
iSigma[1:2, 1:2] ~ dwish(R[1:2, 1:2], r)
Sigma[1:2, 1:2] <- inverse(iSigma[1:2, 1:2])
eps.tau <- exp(logtau)
logtau ~ dflat()
sigma <- 1 / sqrt(eps.tau)
}
list(x = c(8,10,12,14), N = 11, T = 4,
Y = structure(
.Data =   c(21,20,21.5,23,
........
24.5,25,28,28),
.Dim = c(11,4)),mean = c(0, 0),r=4,
R = structure(.Data = c(1, 0, 0,0.1),
.Dim = c(2, 2)),
prec = structure(.Data = c(1.0E-6, 0,0,1.0E-6),
.Dim = c(2, 2))))
```


Dental Data – GEE

We now examine the question of examining the population profiles for boys and girls, using GEE, LMEM and Bayesian hierarchical models.

A priori we would expect differences between boys and girls, and so we carry out separate analyses of boys and girls. As a secondary analysis we examine possible simplifications of the model.

For the GEE analyses, we used an exchangeable working correlation structure, since we expect correlation on observations on the same individual (and so an independence working model would be less efficient).

Letting α denote the common correlation parameter, for boys we found $\widehat{\alpha} = 0.47$ (0.20) and for girls $\widehat{\alpha} = 0.87$ (0.11).

Dental Data – LMEM

We assumed a non-diagonal $\boldsymbol{D}$ for both boys and girls, and a separate measurement error variance σ_ϵ, so that we have carried out separate analyses.

We obtained the following results using REML:

Error variances: $\widehat{\sigma}^B_\epsilon = 1.61$ and $\widehat{\sigma}^G_\epsilon = 0.668$.

Random effects matrices:

$$\widehat{\boldsymbol{D}}^B = \begin{bmatrix} 1.64^2 = 2.69 & -0.01 \times 1.64 \times 0.189 \\ -0.01 \times 1.64 \times 0.189 & 0.189^2 = 0.0357 \end{bmatrix}$$

for boys and

$$\widehat{\boldsymbol{D}}^G = \begin{bmatrix} 2.08^2 = 4.33 & 0.527 \times 2.08 \times 0.161 \\ 0.527 \times 2.08 \times 0.161 & 0.161^2 = 0.0259 \end{bmatrix}$$

for girls.

The larger $\boldsymbol{D}^G_{11}$ element for girls is consistent with greater α for girls with GEE.

Dental Data – Hierarchical Bayes

For the Bayesian analysis, as with the LMEM analyses we assumed separate measurement error variances and non-diagonal $\boldsymbol{D}$ for both boys and girls.

We specified the following priors (for both boys and girls):

$$\pi(\sigma^2_\epsilon, \boldsymbol{\beta}, \boldsymbol{D}^{-1}) \propto \sigma_\epsilon^{-2} \times \mathrm{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 10^6 & 0 \\ 0 & 10^6 \end{bmatrix}\right) \times W_2(r, \boldsymbol{R}^{-1}),$$

with $r = 4$ and $\boldsymbol{R}^{-1} = \begin{bmatrix} 4.0 & 0 \\ 0 & 0.4 \end{bmatrix}$, so that

$$\mathrm{E}[\boldsymbol{D}] = \boldsymbol{R} = \begin{bmatrix} 1.0 & 0 \\ 0 & 0.1 \end{bmatrix},$$

for both boys and girls.

We ran the Markov chain for 10,000 iterations, and discarded the initial 1,000 as burn-in.

Note that the intervals in Table 7 are not asymptotic but reflect the posterior uncertainty.

For inference on the differences $\beta^B_0 - \beta^G_0$ and $\beta^B_1 - \beta^G_1$, and the correlations and standard deviations $\sqrt{\boldsymbol{D}}_{11}$ and $\sqrt{\boldsymbol{D}}_{22}$ functions of interest were defined in WinBUGS.

For the Bayes analyses $\widehat{\sigma}_\epsilon^B = 1.68$ and $\widehat{\sigma}_\epsilon^G = 0.684$.

We assumed a non-diagonal $\boldsymbol{D}$ for both boys and girls, and obtained posterior medians of:

$$\widehat{\boldsymbol{D}}^B = \begin{bmatrix} 1.37^2 = 1.877 & 0.02 \times 1.37 \times 0.181 \\ 0.02 \times 1.37 \times 0.181 & 0.181^2 = 0.0328 \end{bmatrix}$$

for boys and

$$\widehat{\boldsymbol{D}}^G = \begin{bmatrix} 1.87^2 = 3.50 & 0.43 \times 2.08 \times 0.176 \\ 0.43 \times 1.87 \times 0.176 & 0.176^2 = 0.0310 \end{bmatrix}$$

for girls.

These estimates differ from the LMEM results in the directions expected due to the priors, the prior mean for $\boldsymbol{D}_{11}$ is smaller than the MLE, and the prior mean for $\boldsymbol{D}_{22}$ is larger than the MLE

Table 7 summarizes the analyses for the population parameters.

We see that very similar result from all analyses, which is reassuring (and not surprising in this very balanced situation).

Method	β_0^B	β_1^B	β_0^G	β_1^G	$\beta_0^B - \beta_0^G$	$\beta_1^B - \beta_1^G$
GEE	24.97	0.784	22.65	0.480	2.32	0.305
	(24.08,25.86)	(0.581,0.988)	(21.57,23.73)	(0.349,0.610)	(0.921,3.72)	(0.0632,0.546)
LMEM	24.97	0.784	22.65	0.480	2.32	0.305
	(24.07,25.87)	(0.585,0.984)	(21.40,23.89)	(0.350,0.609)	(0.789,3.85)	(0.0671,0.543)
Bayes	24.97	0.783	22.64	0.479	2.33	0.304
	(24.13,25.82)	(0.572,1.02)	(21.46,23.86)	(0.330,0.630)	(0.846,3.78)	(0.0482,0.575)

Table 7: Posterior medians and 95% intervals for population means, under GEE, LMEMs and Bayes hierarchical models.

Analysis of Reduced Dataset

Artificially create a reduced growth curve data set within which it is assumed that children randomly drop out of the study at some point after their first measurement. This yielded the data in Figure 29 – there are now 39 measurements on boys (previously 64), and 25 on girls (previously 44).

Again we analyze using GEE, LMEM and hierarchical Bayes, but in each case with simplified models (due to smaller sample size).

GEE: Single off-diagonal parameter in exchangeable structure, estimated as $\widehat{\alpha} = 0.84$.

Table 8 shows that the standard errors under exchangeable correlation structure appear too small (compared to LMEM and Bayes) – hence, also carried out with working independence, giving results more in line with other two analyses.

Reason? Negative bias in sandwich estimation due to small and unbalanced dataset?

LMEM: Single measurement error ($\widehat{\sigma}_\epsilon = 1.41$) and single variance-covariance matrix for boys and girls:

$$\widehat{\boldsymbol{D}} = \begin{bmatrix} 2.23^2 & 0.88 \times 2.23 \times 0.13 \\ 0.88 \times 2.23 \times 0.133 & 0.133^2 \end{bmatrix}.$$

Bayes: Single measurement error ($\widehat{\sigma}_\epsilon = 1.48$) and single variance-covariance matrix for boys and girls:

$$\widehat{\boldsymbol{D}} = \begin{bmatrix} 1.98^2 & 0.53 \times 1.98 \times 0.190 \\ 0.53 \times 1.98 \times 0.190 & 0.198^2 \end{bmatrix}.$$

Same Wishart prior as before – note shrinkage to prior mean. 95% interval estimate on correlation is (-0.53,0.93).

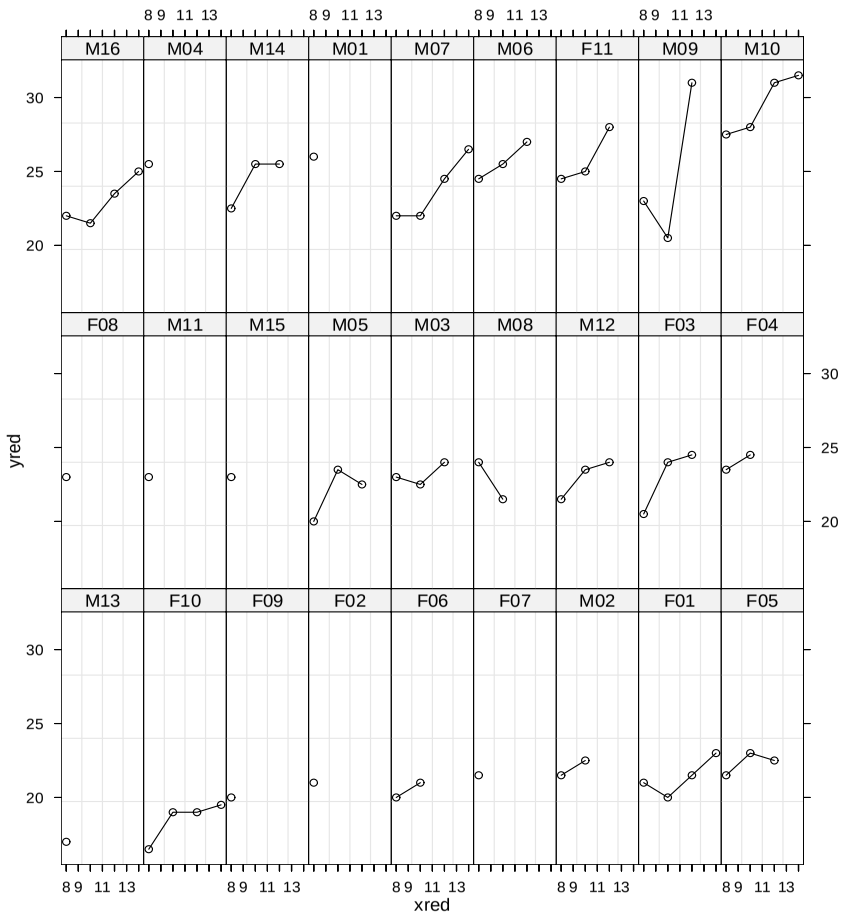

Figure 29: Distance versus age for reduced data.

Method	β_0^B	β_1^B	β_0^G	β_1^G	$\beta_0^B - \beta_0^G$	$\beta_1^B - \beta_1^G$
GEE E	24.78	0.671	22.77 (0.68)	0.502 (0.096)	2.011 (0.93)	0.210 (0.15)
	(,)	(,)	(21.43,24.10)	(0.179,0.888)	(0.190,3.83)	(-0.0760,0.496)
GEE I	24.88	0.772	22.17 (0.91)	0.535 (0.27)	2.700 (1.18)	0.237 (0.18)
	(,)	(,)				
LMEM	24.75	0.702	22.83 (0.81)	0.533 (0.18)	1.917 (1.04)	0.169 (0.23)
	(,)	(,)	(21.25,24.42)	(-0.120,3.953)	(0.179,0.888)	(-0.280,0.618)
Bayes	24.73	0.693	22.77 (0.79)	0.504 (0.21)	1.962 (1.00)	0.189 (0.26)
	(23.49,25.97)	(0.377,0.992)	(21.21,24.34)	(0.0928,0.917)	(-0.0336,3.895)	(-0.321,0.700)

Table 8: Posterior medians and 95% intervals for population means, under GEE (exchangeable and independence working covariance models), LMEMs and Bayes hierarchical models, for reduced data set.

CHAPTER 9: GENERAL REGRESSION MODELS

Motivating Examples:

Wakefield et al. (1994) model drug concentration data from 10 patients following administration of the drug Cadralazine. The regression model for individual i is non-linear in the parameters (and is not a GLM).

Yu et al. (2001) record the presence/absence of wheeze on asthmatic children each day for approximately 60 days, with daily pollution measures also being available. If we only had data from a single child then we might use a logistic or probit regression model.

Clayton and Kaldor (1987) examine the incidence of lip cancer in Scottish areas, with the proportion of individual's employed in agriculture being available for each area. Poisson counts, but with the potential for spatial dependence (exchangeability not reasonable here – areas close together more likely to be similar than those far apart).

We begin by considering the class of *Generalized Linear Mixed Models* (GLMMs), before turning to more general non-linear mixed effects models.

In this chapter we will consider both a *conditional* approach to modeling via the introduction of random effects, and a *marginal* approach using GEEs.

Example: Log-Linear Regression

For non-linear models marginal and conditional models have different interpretations – we illustrate with a simple example with non-dependent data, but overdispersion.

Conditional Models

Stage 1: $Y_i|\beta_0^c, \beta_1, b_i \sim_{ind} \text{Poisson}(\mu_i^c)$, where

$$\mu_i^c = \text{E}[Y_i|\beta_0^c, \beta_1, b_i] = N_i \exp(\beta_0^c + b_i + \beta_1 x_i),$$

is the conditional mean, $i = 1, ..., n$.

Stage 2: $b_i \sim_{iid} N(0, \sigma_0^2)$.

We now evaluate the marginal mean and variance – useful to recall the mean and variance of a lognormal random variable $Z \sim \text{LN}(\mu, \sigma^2)$:

$$\text{E}[Z] = \exp(\mu + \sigma^2/2),$$

and

$$\text{var}(Z) = \text{E}[Z]^2\{\exp(\sigma_0^2) - 1\}.$$

Note: $\exp(b_i) \sim \text{LN}(0, \sigma_0^2)$.

Marginally we have mean

$$\begin{aligned} \mathrm{E}[Y_i|\beta_0^c, \beta_1, \sigma_0^2] &= \mathrm{E}_{b_i|\sigma_0^2}\{\mathrm{E}[Y_i|\beta_0^c, \beta_1, b_i]\} \\ &= N_i \exp(\beta_0^c + \sigma_0^2/2 + \beta_1 x_i). \end{aligned}$$

Interpretation: $\exp(\beta_0^c)$ is not the expected response given that $x_i = 0$, rather it the median response given $x_i = 0$. The expected response at $x = 0$ is $\exp(\beta_0^c + \sigma_0^2/2)$. This is a marginal interpretation.

Alternatively, we might say that $\exp(\beta_0^c)$ is the average response for an individual with $x_i = 0$, and $b_i = 0$, a "typical" individual. This is a conditional interpretation.

We may also evaluate the marginal variance

$$\begin{aligned} \mathrm{var}(Y_i|\beta_0^c, \beta_1, \sigma_0^2) &= \mathrm{var}_{b_i|\sigma_0^2}\{\mathrm{E}[Y_i|\beta_0^c, \beta_1, b_i]\} \\ &+ \mathrm{E}_{b_i|\sigma_0^2}\{\mathrm{var}[Y_i|\beta_0, \beta_1, b_i]\} \\ &= \mathrm{E}[Y_i|\beta_0^c, \beta_1, \sigma_0^2](1 + \mathrm{E}[Y_i|\beta_0^c, \beta_1] \times c), \end{aligned}$$

where $c = \exp(\sigma_0^2) - 1$, and covariance

$$\mathrm{cov}(Y_i, Y_j|\beta_0^c, \beta_1, \sigma_0^2) = 0.$$

If we had considered the model:

$$Y_i|\beta_0^c, \beta_1, \delta_i \sim_{ind} \mathrm{Poisson}(N_i \exp(\beta_0^c + \beta_1 x_i)\delta_i),$$

with $\delta_i \sim_{iid} \mathrm{Ga}(\alpha, \alpha)$, then the marginal distribution of the data is obtained by integrating out the δ_i, and is negative binomial.

For the Poisson-normal model there is no closed form for the integrated marginal distribution:

$$p(y_i|\beta_0^c, \beta_1, \sigma_0^2) = \int_{b_i} p(y_i|\beta_0^c, \beta_1, \sigma_0^2) \times p(b_i|\sigma_0^2) db_i,$$

and so numerical/analytical approximations are needed.

In a likelihood context we may maximize

$$l(\beta_0^c, \beta_1, \sigma_0^2) = \sum_{i=1}^{n} \log p(y_i|\beta_0^c, \beta_1, \sigma_0^2),$$

and use standard theory. For estimates of the random effects, empirical Bayes methods may be used.

For a Bayesian analysis, we consider the posterior

$$p(\beta_0^c, \beta_1, \sigma_0^2|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\beta_0^c, \beta_1, \sigma_0^2)\pi(\beta_0^c, \beta_1, \sigma_0^2)}{p(\boldsymbol{y})},$$

where $\pi(\beta_0^c, \beta_1, \sigma_0^2)$ is the prior.

Computation is conveniently carried out by Markov chain Monte Carlo.

Marginal Model: We might specify

$$\mathrm{E}[Y_i|\beta_0^m, \beta_1] = \mu_i^m = N_i \exp(\beta_0^m + \beta_1 x_i),$$

so that now β_0^m may be interpreted marginally as the expected response at $x = 0$.

For the variance we may assume

$$\mathrm{var}(Y_i|\beta_0^m, \beta_1) = \mathrm{E}[Y_i|\beta_0^m, \beta_1] \times (1 + \alpha_q \times \mathrm{E}[Y_i|\beta_0^m, \beta_1]),$$

where α_q, the parameter in a *quadratic* variance model, needs to be estimated.

Alternatively:

$$\mathrm{var}(Y_i|\beta_0^m, \beta_1) = \alpha_l \mathrm{E}[Y_i|\beta_0^m, \beta_1],$$

where α_l, the parameter in a *linear* variance model, to be estimated.

In both cases we assume

$$\mathrm{cov}(Y_i, Y_j|\beta_0^m, \beta_1) = 0.$$

Inference for the marginal model may proceed via the estimating functions:

$$\boldsymbol{G}(\beta_0^m, \beta_1) = \boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}(\alpha)^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}),$$

where $\boldsymbol{V}(\alpha)$ is $n \times n$ variance-covariance matrix that is a function of the parameter α, and $\boldsymbol{D}$ is the $n \times 2$ matrix of derivatives with $i-$th row

$$\left[\frac{\partial \mu_i}{\partial \beta_0^m} \quad \frac{\partial \mu_i}{\partial \beta_1}\right].$$

For the linear variance model we have

$$\boldsymbol{G}(\beta_0^m, \beta_1) = \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{\mu}),$$

where $\boldsymbol{x}$ is the $n \times 2$ matrix with $i-$th row $[1 \;\; x_i]$ and so we do not need an estimate of α_l to obtain the estimate.

For standard errors etc, we need an estimate of α_l.

Non-Linear Mixed Effects Models

Likelihood Approach See Pinheiro and Bates (2000, Chapter 7).

Consider the model

$$y_{ij} = f(\boldsymbol{\phi}_{ij}, \boldsymbol{x}_{ij}) + \epsilon_{ij},$$

where $\boldsymbol{x}_{ij}$ are covariates (e.g. time), and

$$\boldsymbol{\phi}_{ij} = \boldsymbol{\phi}(\boldsymbol{\beta}, \boldsymbol{b}_i) = \boldsymbol{A}_{ij}\boldsymbol{\beta} + \boldsymbol{B}_{ij}\boldsymbol{b}_i,$$

with $\boldsymbol{A}_{ij}$ and $\boldsymbol{B}_{ij}$ functions of potentially time-varying covariates $\boldsymbol{x}_{ij}$,

- $\dim(\boldsymbol{\beta}) = p + 1$,
- $\dim(\boldsymbol{b}_i) = q + 1$
- $\epsilon_{ij} \sim \mathrm{N}(0, \sigma_\epsilon^2)$,
- $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D})$
- ϵ_{ij} and $\boldsymbol{b}_i$ independent, $i = 1, ..., m; j = 1, ..., n_i$.

Let $\boldsymbol{\alpha}$ represent σ_ϵ^2 and the parameters of $\boldsymbol{D}$ and $N = \sum_i n_i$.

The likelihood is, as usual, obtained by integrating out the random effects:

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= (2\pi\sigma^2)^{-N/2}(2\pi)^{-m/2}|\boldsymbol{D}|^{-m/2} \\ &\times \prod_{i=1}^{m} \int \exp\left[-\frac{(\boldsymbol{y}_i - \boldsymbol{f}_i)^{\mathrm{T}}(\boldsymbol{y}_i - \boldsymbol{f}_i)}{2\sigma_\epsilon^2} - \frac{\boldsymbol{b}_i \boldsymbol{D}^{-1} \boldsymbol{b}_i}{2} \right] d\boldsymbol{b}_i. \end{aligned}$$

where $\boldsymbol{f}_i = \boldsymbol{f}\{\boldsymbol{\phi}(\boldsymbol{\beta}, \boldsymbol{b}_i), \boldsymbol{x}_i\}$, $i = 1, ..., m$.

There are two difficulties here:

1. How to calculate the required integrals (which for non-linear models are analytically intractable, recall for linear models they were available in closed form).
2. How to maximize the resultant likelihood (for the linear model, we described Newton-Raphson and EM algorithms).

Example: Pharmacokinetics of Indomethacin

Six human volunteers received bolus intravenous doses (of the same size) of Indomethacine, and subsequently blood samples were taken, and the drug concentrations recorded.

Figure 30 shows the concentration-time data – the curves follow a similar pattern but there is clearly person to person variability.

The compartmental model that has previously been used for this drug is the two-compartment bi-exponential model:

$$\mathrm{E}[Y] = A_1 \exp\{-\alpha_1 t\} + A_2 \exp\{-\alpha_2 t\},$$

where Y is concentration, and t is time, and $A_1, A_2, \alpha_1, \alpha_2 > 0$.

Note: this model is unidentifiable since the parameter set $(A_1, \alpha_1, A_2, \alpha_2)$ gives the same fitted curve (and hence likeihood) as the set $(A_1, \alpha_1, A_2, \alpha_2)$. If this is a practical problem for a particular dataset (say $\alpha_1 \approx \alpha_2$) then we may parameterize in terms of α_1 and $\alpha_2 - \alpha_1$.

Figure 31 gives the log concentrations versus time – such a plot can be useful for picking the number of exponentials (and modeling the log concentration can provide initial estimates). Certainly not linear in time so more than a single exponential needed.

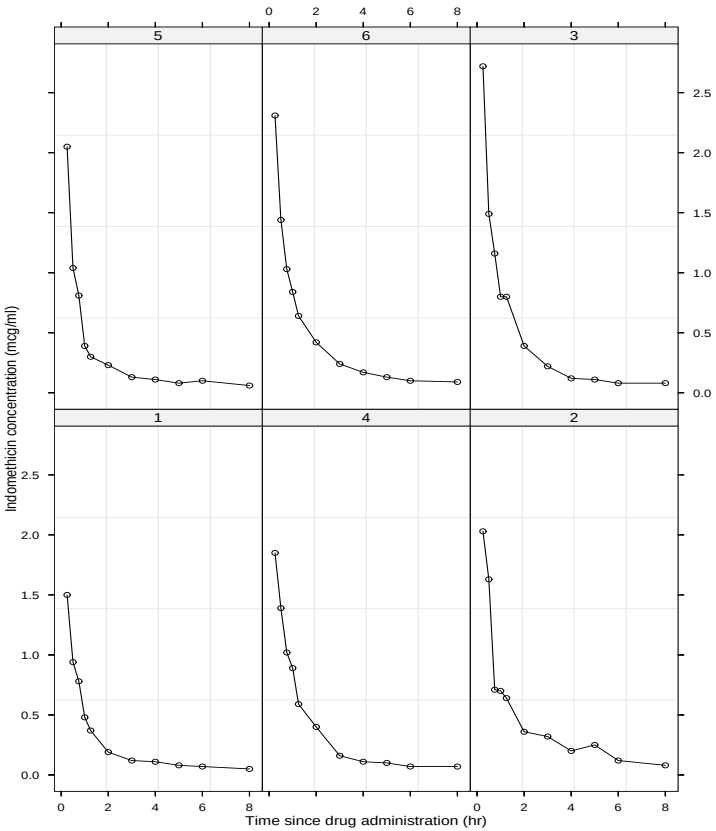

Figure 30: Concentration time data for Indomethacin.

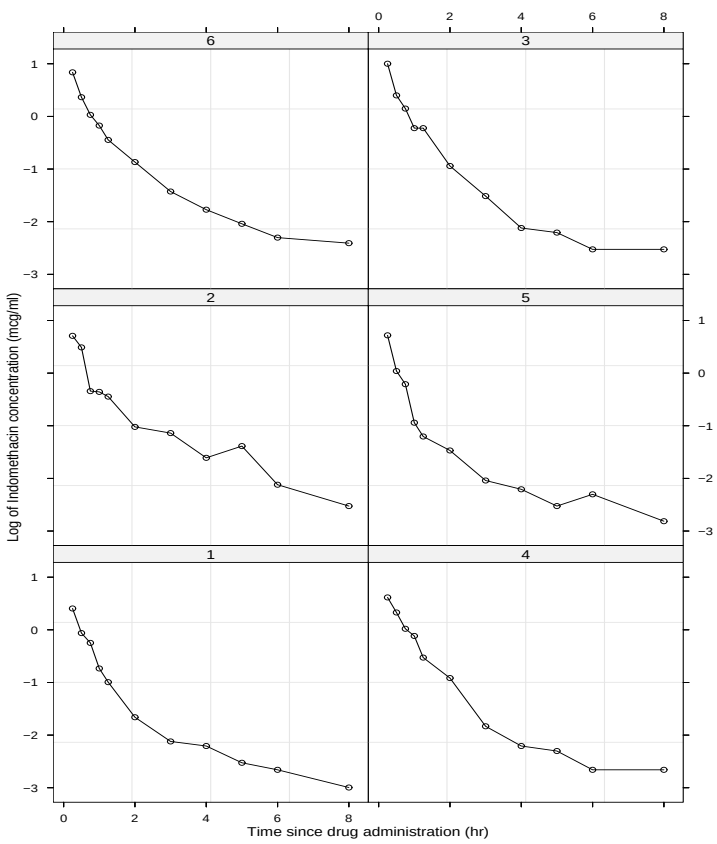

Figure 31: Log concentration time data for Indomethacin.

Individual fits

Let Y_{ij} be the drug concentration at time t_{ij} on indvidual i, $j = 1, ..., 11$, $i = 1, ..., 6$. We first fit bi-exponential models to each individual, using non-linear least squares.

We parameterize as

$$\mathrm{E}[Y_{ij} \mid \boldsymbol{\theta}_i] = \theta_{1i} \exp\{-\mathrm{e}^{\theta_{3i}} t_{ij}\} + \theta_{2i} \exp\{-\mathrm{e}^{\theta_{4i}} t_{ij}\},$$

for $i = 1, ..., 6$.

Even though the data are balanced, the standard errors are different for different individuals, as we see in Figure 32.

```
> indiv.lis <- nlsList( conc ~ SSbiexp(time,A1,lrc1,A2,lrc2),
  data=Indometh )
> indiv.lis
Call:
Model:conc~SSbiexp(time,A1,lrc1,A2,lrc2)|Subject
   Data: Indometh
Coefficients:
        A1      lrc1        A2       lrc2
1 2.029277 0.5793887 0.1915475 -1.7877849
4 2.198132 0.2423124 0.2545223 -1.6026859
2 2.827673 0.8013195 0.4989175 -1.6353512
5 3.566103 1.0407660 0.2914970 -1.5068522
6 3.002250 1.0882119 0.9685230 -0.8731358
3 5.468312 1.7497936 1.6757522 -0.4122004
Degrees of freedom: 66 total; 42 residual
Residual standard error: 0.0755502
```


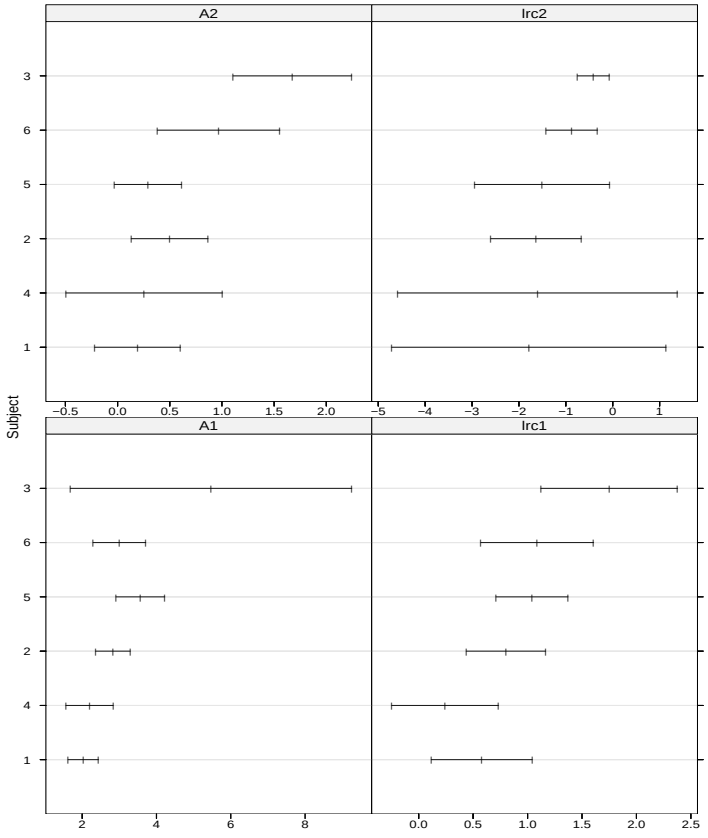

Figure 32: Asymptotic 95% CIs for elements of $\boldsymbol{\theta}_i$, $i = 1, ..., 6$.

The `nlme` algorithm

Within `nlme` an algorithm, introduced by Lindstrom and Bates (1990) is used.

The algorithm alternates between two steps:

Penalized Non-linear Least Squares (PNLS)

Condition on the current estimates of $\widehat{\boldsymbol{D}}$ and $\widehat{\sigma}^2_\epsilon$ and then minimize

$$\frac{1}{\widehat{\sigma}^2_\epsilon}\sum_{i=1}^{m}(\boldsymbol{y}_i-\boldsymbol{f}_i)^{\mathrm{T}}(\boldsymbol{y}_i-\boldsymbol{f}_i)+\boldsymbol{b}_i\widehat{\boldsymbol{D}}^{-1}\boldsymbol{b}_i,$$

to obtain estimates $\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{b}}_1,...\widehat{\boldsymbol{b}}_m$.

Note: may be viewed as finding the posterior mode for $\boldsymbol{\beta}$ and $\boldsymbol{b}_1,...,\boldsymbol{b}_m$.

Linear Mixed Effects (LME)

Carry out a first-order Taylor series of $\boldsymbol{f}_i$ about $\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{b}}_i$.

This results in a linear mixed effects model which can be maximized to obtain estimates of $\boldsymbol{D}$ and σ^2_ϵ.

We have likelihood

$$L(\boldsymbol{\beta},\boldsymbol{\alpha})=\mid\boldsymbol{D}\mid^{-m/2}\sigma_\epsilon^{-N}$$

$$\int\exp\left\{-\frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{y}_i-\boldsymbol{f}_i)^{\mathrm{T}}(\boldsymbol{y}_i-\boldsymbol{f}_i)-\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{b}_i\right\}d\boldsymbol{b}_i \quad (18)$$

where $\boldsymbol{f}_i=\boldsymbol{f}\{\boldsymbol{\phi}(\boldsymbol{\beta},\boldsymbol{b}_i),\boldsymbol{x}_i\}$, $i=1,...,m$.

Carry out a first-order Taylor series expansion of f_i about the estimates, obtained in the PNLS step at iteration k, of $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$, call these $\widehat{\boldsymbol{\beta}}^{(k)}$ and $\widehat{\boldsymbol{b}}_i^{(k)}$.

Specifically

$$\boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i) \approx \boldsymbol{f}_i\left(\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\boldsymbol{b}}_i^{(k)}\right) + \widehat{\boldsymbol{x}}_i^{(k)}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(k)}\right) + \widehat{\boldsymbol{z}}_i^{(k)}\left(\boldsymbol{b}_i - \widehat{\boldsymbol{b}_i}^{(k)}\right)$$

where

$$\widehat{\boldsymbol{x}}_i^{(k)} = \left.\frac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right|_{\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\boldsymbol{b}}_i^{(k)}}$$

$$\widehat{\boldsymbol{z}}_i^{(k)} = \left.\frac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{b}_i^{\mathrm{T}}}\right|_{\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\boldsymbol{b}}_i^{(k)}}$$

This gives

$$\boldsymbol{y}_i - \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i) \approx \boldsymbol{y}_i^{(k)} - \widehat{\boldsymbol{x}}_i^{(k)}\boldsymbol{\beta} - \widehat{\boldsymbol{z}}_i^{(k)}\boldsymbol{b}_i$$

where

$$\boldsymbol{y}_i^{(k)} = \boldsymbol{y}_i - \boldsymbol{f}_i\left(\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\boldsymbol{b}}_i^{(k)}\right) + \widehat{\boldsymbol{x}}_i^{(k)}\widehat{\boldsymbol{\beta}}^{(k)} + \widehat{\boldsymbol{z}}_i^{(k)}\widehat{\boldsymbol{b}}_i^{(k)}$$

Substitution in (18) allows the integral to be evaluated in closed-form to give log-likelihood

$$l(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{m}\log|\widehat{\boldsymbol{V}}_i| - \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{y}_i^{(k)} - \widehat{\boldsymbol{x}}_i^{(k)}\boldsymbol{\beta})^{\mathrm{T}}\widehat{\boldsymbol{V}}_i^{-1}(\boldsymbol{Y}_i - \widehat{\boldsymbol{x}}_i\boldsymbol{\beta})$$

where

$$\widehat{\boldsymbol{V}}_i = \widehat{\boldsymbol{z}}_i^{(k)}\boldsymbol{D}\widehat{\boldsymbol{z}}_i^{(k)\mathrm{T}} + \sigma_\epsilon^2\boldsymbol{I}_i,$$

which may be maximized to give ML estimates. REML estimates are obtained by adding the term

$$-\frac{1}{2}\sum_{i=1}^{m}\log|\,\widehat{\boldsymbol{x}}_i^{(k)\mathrm{T}}\widehat{\boldsymbol{V}}_i(\boldsymbol{\alpha})\widehat{\boldsymbol{x}}_i^{(k)}\,|$$

Alternatives: Laplace method (very close to the above algorithm), Gauss-Hermite quadrature, importance sampling.

Inference

Under the LB algorithm, the asymptotic distribution of the REML estimator $\widehat{\boldsymbol{\beta}}$ is

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathrm{N}_{p+1}\left(\boldsymbol{\beta}, \left[\sum_{i=1}^{m} \widehat{\boldsymbol{x}}_i^{\mathrm{T}} \widehat{\boldsymbol{V}}_i^{-1} \widehat{\boldsymbol{x}}_i\right]^{-1}\right),$$

where $\widehat{\boldsymbol{x}}_i = \widehat{\boldsymbol{x}}_i^{(k)}$ with k the final iteration, $i = 1, ..., m$

Similarly, the asymptotic distribution of $\boldsymbol{\alpha}$ is based on the information as calculated from the linear approximation to the likelihood.

Empirical Bayes estimates for the random effects are available, but caution should be given to using these for checking assumptions since they are strongly influenced by the assumption of normality being correct. If n_i is large then this will be less of a problem.

Indomethacin Example Revisited

First assume a diagonal $\boldsymbol{D}$ with random effects for first three elements only.

```
> nlme.indo <- nlme( indiv.lis,random=pdDiag(A1+lrc1+A2~1))
> summary(nlme.indo)
Nonlinear mixed-effects model fit by maximum likelihood
  Model: conc ~ SSbiexp(time, A1, lrc1, A2, lrc2)
 Data: Indometh
        AIC       BIC   logLik
  -93.18472 -75.66748 54.59236
Random effects:
 Formula: list(A1 ~ 1, lrc1 ~ 1, A2 ~ 1)
 Level: Subject
 Structure: Diagonal
             A1      lrc1        A2   Residual
StdDev: 0.57135 0.1581214 0.1115283 0.08149631
Fixed effects: list(A1 ~ 1, lrc1 ~ 1, A2 ~ 1, lrc2 ~ 1)
          Value Std.Error DF   t-value p-value
A1    2.8276029 0.2639744 57 10.711656   0e+00
lrc1  0.7732529 0.1100086 57  7.029021   0e+00
A2    0.4610197 0.1127560 57  4.088648   1e-04
lrc2 -1.3450041 0.2313139 57 -5.814627   0e+00
 Correlation:
     A1     lrc1   A2
lrc1  0.055
A2   -0.102  0.630
lrc2 -0.139  0.577  0.834
```

Now assume a non-diagonal $\boldsymbol{D}$ for all four parameters.

```
> nlme2.indo2 <- update( nlme.indo, random=A1+lrc1+A2+lrc2~1)
> summary(nlme.indo2)
Nonlinear mixed-effects model fit by maximum likelihood
  Model: conc ~ SSbiexp(time, A1, lrc1, A2, lrc2)
Random effects:
 Formula: list(A1 ~ 1, lrc1 ~ 1, A2 ~ 1, lrc2 ~ 1)
 Level: Subject
 Structure: General positive-definite, Log-Cholesky parametrizatic
         StdDev     Corr
A1       0.77583020 A1    lrc1  A2
lrc1     0.26863662 0.963
A2       0.38707000 0.459 0.682
lrc2     0.48253192 0.153 0.414 0.948
Residual 0.06962038
Fixed effects: list(A1 ~ 1, lrc1 ~ 1, A2 ~ 1, lrc2 ~ 1)
          Value Std.Error DF   t-value p-value
A1    2.8531611 0.3485825 57  8.185039   0e+00
lrc1  0.8755645 0.1253269 57  6.986245   0e+00
A2    0.6357872 0.1715520 57  3.706091   5e-04
lrc2 -1.2757709 0.2161119 57 -5.903288   0e+00
 Correlation:
     A1    lrc1  A2
lrc1 0.907
A2   0.411 0.676
lrc2 0.108 0.378 0.912
```


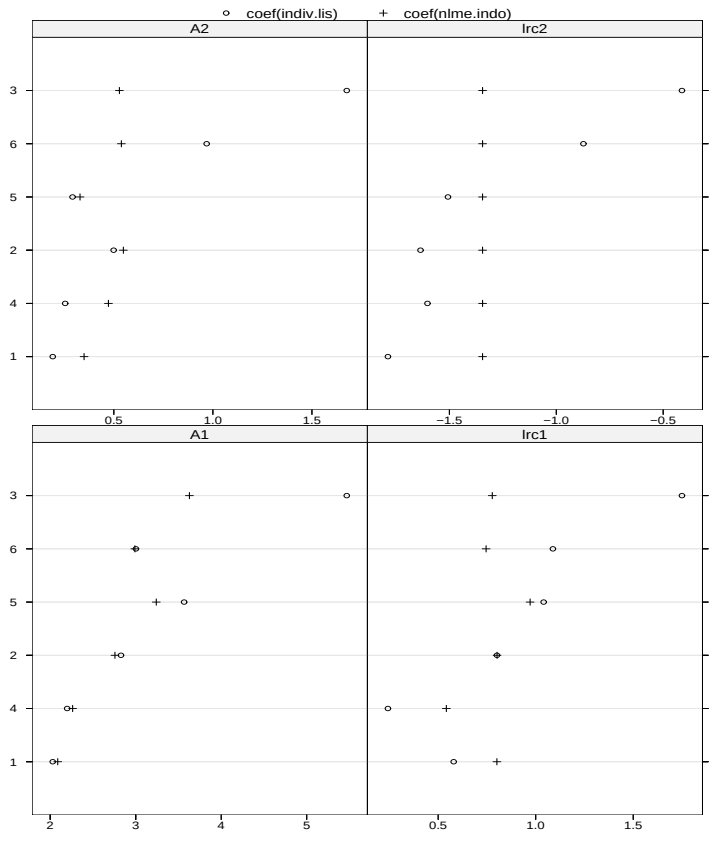

Figure 33: Comparison of non-linear LS and nlme estimates, with the latter from the model `nlme.indo` Created using the command `plot(compareFits(coef(indiv.is),coef(nlme.indo)))`.