LTCC Advanced Course: Introduction to Semiparametric Modelling Lecture 3

Clifford Lam

Department of Statistics
London School of Economics and Political Science

Linear mixed model

- Longitudinal data of 48 pigs. Response is `weight`. They are measured over 9 successive weeks.

- Without taking repeated measures into account, a regression model can be

$$\texttt{weight}_{ij} = \beta_0 + \beta_1 \texttt{week}_j + \epsilon_{ij}, \quad 1 \leq i \leq 48,\ 1 \leq j \leq 9,$$

where $\epsilon_{ij} \sim i.i.d. N(0, \sigma^2)$. It leads to $\hat{\beta}_1 = 6.21$, $SD(\hat{\beta}_1) = 0.0818$.

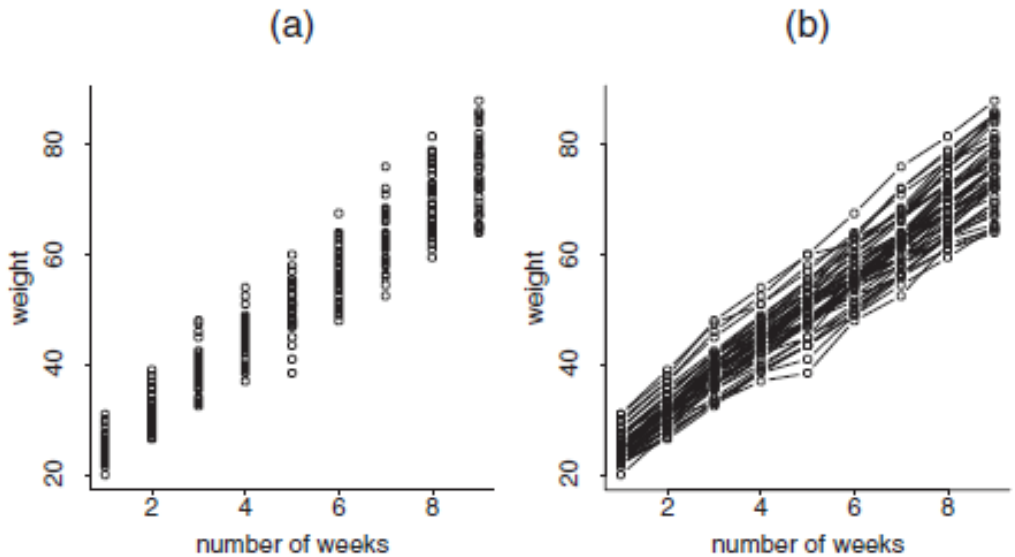

Figure: Representations of pig weight data. Panel (a) is a scatterplot of weight against week number. In (b), lines are used to connect those points pertaining to the same pig.

Linear mixed model

- A problem is that each individual pig is not that variable as seen in (b). **Within-pig** information should be helpful.
- Looking at the plots, a better model can be

$$\texttt{weight}_{ij} = \alpha_i + \beta_1 \texttt{week}_j + \epsilon_{ij}, \quad 1 \leq i \leq 48,\ 1 \leq j \leq 9,$$

  where α_i represents the intercept for the ith pig.
- There are then 48 intercepts and 1 slope parameter - too many for $n = 48$.
- A remedy is to use a random intercept:

$$\texttt{weight}_{ij} = \beta_0 + U_i + \beta_1 \texttt{week}_j + \epsilon_{ij},$$

  where the U_i's are treated as random sample from $N(0, \sigma_U^2)$, say.
- It is a **mixed model**, with **fixed component**

$$\beta_0 + \beta_1 \texttt{week}_j,$$

  and **random component**

$$U_i \sim N(0, \sigma_U^2).$$

- It leads to $\hat{\beta}_1 = 6.21$, $SD(\hat{\beta}_1) = 0.0391$. Later on how to fit.

Linear mixed model

- The mixed model allows for within-pig correlation. For two time points $j \neq j'$, for pig i,
$$\text{cov}(\texttt{weight}_{ij}, \texttt{weight}_{ij'}) = \text{var}(U_i) = \sigma_U^2.$$
- The correlation coefficient will be
$$\text{corr}(\texttt{weight}_{ij}, \texttt{weight}_{ij'}) = \frac{\sigma_U^2}{\sigma^2 + \sigma_U^2}.$$
- For the data, this within-pig correlation is estimated to be 0.775. So very high within-pig correlation indeed.

Linear mixed model - General formulation

- The general **linear mixed model** is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon},$$

where

$$E\begin{pmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \quad \text{and} \quad \text{cov}\begin{pmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{pmatrix} = \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix}.$$

- The pig weights example has

$$\mathbf{y} = \begin{pmatrix} \texttt{weight}_{1,1} \\ \vdots \\ \texttt{weight}_{1,9} \\ \vdots \\ \texttt{weight}_{48,1} \\ \vdots \\ \texttt{weight}_{48,9} \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & \texttt{week}_1 \\ \vdots & \vdots \\ 1 & \texttt{week}_9 \\ \vdots & \vdots \\ 1 & \texttt{week}_1 \\ \vdots & \vdots \\ 1 & \texttt{week}_9 \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},$$

$$\mathbf{Z} = \text{diag}(\mathbf{1}_{9\times 1}, \ldots, \mathbf{1}_{9\times 1}), \ \mathbf{u} = (U_1, \cdots, U_{48})^T, \ \mathbf{G} = \sigma_U^2\mathbf{I}, \ \mathbf{R} = \sigma^2\mathbf{I}.$$

Estimation and Prediction in LMM

- We can view the general LMM as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^*, \quad \text{where } \boldsymbol{\epsilon}^* = \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}.$$

- This is a linear model with correlated errors:

$$\text{cov}(\boldsymbol{\epsilon}^*) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}.$$

- Given $\mathbf{V}$, the **generalized least square** (GLS) estimator is

$$\tilde{\beta} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y},$$

  which is the best linear unbiased estimator (BLUE) for β. It is also MLE and uniformly minimum variance unbiased estimator (UMVUE) under normality.

- For predicting the random effects, can use the conditional expectation

$$\tilde{\mathbf{u}} = E(\mathbf{u}|\mathbf{y}),$$

  since it always minimises $E(\mathbf{u} - \tilde{\mathbf{u}})^2$ for an estimator $\tilde{\mathbf{u}}$ of $\mathbf{u}$. Note that $\tilde{\mathbf{u}}$ is not always linear in $\mathbf{y}$.

- If $\mathbf{u}$ and $\mathbf{y}$ forms a multivariate normal random vector, then best prediction (BP) is also the best linear prediction (BLP),

$$\text{BP}(\mathbf{u}) = \text{BLP}(\mathbf{u}) = E(\mathbf{u}) + \mathbf{C}\mathbf{V}^{-1}\{\mathbf{y} - E(\mathbf{y})\} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

  where $\mathbf{C} = \text{cov}(\mathbf{u}, \mathbf{y}) = \mathbf{G}\mathbf{Z}^T$.

Best linear unbiased prediction (BLUP)

- A systematic way to achieve best linear prediction is through BLUP. For $\mathbf{t}$ and $\mathbf{s}$ two arbitrary vectors, we want to find **linear** $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}$ such that we minimise the prediction error
$$E\{(\mathbf{s}^T\mathbf{X}\tilde{\boldsymbol{\beta}}+\mathbf{t}^T\mathbf{Z}\tilde{\mathbf{u}})-(\mathbf{s}^T\mathbf{X}\boldsymbol{\beta}+\mathbf{t}^T\mathbf{Z}\mathbf{u})\}^2,$$
subject to $E(\mathbf{s}^T\mathbf{X}\tilde{\boldsymbol{\beta}}+\mathbf{t}^T\mathbf{Z}\tilde{\mathbf{u}})=E(\mathbf{s}^T\mathbf{X}\boldsymbol{\beta}+\mathbf{t}^T\mathbf{Z}\mathbf{u})$.
- The solutions can be shown to be
$$\text{BLUP}(\boldsymbol{\beta})\equiv\tilde{\boldsymbol{\beta}}=(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y},$$
$$\text{BLUP}(\mathbf{u})\equiv\tilde{\mathbf{u}}=\mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}).$$
- These solutions coincide with the GLS solution for $\boldsymbol{\beta}$ and the BLP solution for $\mathbf{u}$ with $\boldsymbol{\beta}$ there substituted by $\tilde{\boldsymbol{\beta}}$.

Best linear unbiased prediction (BLUP)

- Another way to derive BLUP is by assuming

$$\mathbf{y}|\mathbf{u} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R}), \quad \mathbf{u} \sim N(\mathbf{0}, \mathbf{G}),$$

  and do MLE w.r.t. $\boldsymbol{\beta}$ and $\mathbf{u}$. It leads to minimising

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^T R^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \mathbf{u}^T \mathbf{G}^{-1} \mathbf{u},$$

  which is a generalized least square problem with a penalty term.

- This has solutions

$$\begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{pmatrix} = (\mathbf{C}^T \mathbf{R}^{-1} \mathbf{C} + \mathbf{B})^{-1} \mathbf{C}^T \mathbf{R} - 1\mathbf{y},$$

  where $\mathbf{C} = [\mathbf{X}, \mathbf{Z}]$ and $\mathbf{B} = \text{diag}(\mathbf{0}, \mathbf{G}^{-1})$. This is another formulation of BLUP($\boldsymbol{\beta}$) and BLUP($\mathbf{u}$).

- The fitted values are then

$$\text{BLUP}(\mathbf{y}) = \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{Z}\tilde{\mathbf{u}} = \mathbf{C}(\mathbf{C}^T \mathbf{R}^{-1} \mathbf{C} + \mathbf{B})^{-1} \mathbf{C}^T \mathbf{R}^{-1} \mathbf{y}.$$

Covariance Estimation

- To get the practical estimates of $\tilde{\beta}$ and $\tilde{\mathbf{u}}$, we need estimates of $\mathbf{G}$ and $\mathbf{R}$ (thus $\mathbf{V}$). This can be done though ML or REML (restricted ML).
- For ML, the model is $\mathbf{y} \sim N(\mathbf{X}\beta, \mathbf{V})$, so that the log-likelihood is

$$\ell(\beta, \mathbf{V}) = -\frac{1}{2}\{n\log(2\pi) + \log|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\beta)^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta)\}.$$

- In exercise 2 you will show that by keeping $\mathbf{V}$ fixed, this is maximised by

$$\tilde{\beta} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}.$$

- Substituting back into the log-likelihood, we obtain the **profile log-likelihood** for $\mathbf{V}$,

$$\ell_p(\mathbf{V}) = -\frac{1}{2}[\log|\mathbf{V}| + \mathbf{y}^T\mathbf{V}^{-1}\{\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\}\mathbf{y}] - \frac{n}{2}\log(2\pi).$$

- In the pig weights example, $\mathbf{V} = \sigma_U^2\mathbf{Z}\mathbf{Z}^T + \sigma^2\mathbf{I}$, so that $\ell_p(\mathbf{V})$ is a function of σ_U^2 and σ^2. Maximising, we get

$$\hat{\sigma}^2_{\text{ML}} = 4.38, \quad \hat{\sigma}^2_{U,\text{ML}} = 14.8$$

- The REML approach find linear combinations of $\mathbf{y}$ so that β is not entered into the likelihood of the various combinations (not covered in this course). With large samples ML and REML produce similar results. In the end we have

$$\ell_r(\mathbf{V}) = \ell_p(\mathbf{V}) - \frac{1}{2}\log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|.$$

Plug-in estimators

- With the covariance matrices estimated, we can finally have the **estimated BLUP** (EBLUP) as

$$\hat{\boldsymbol{\beta}} := \text{EBLUP}(\boldsymbol{\beta}) = (\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y},$$
$$\hat{\mathbf{u}} := \text{EBLUP}(\mathbf{u}) = \hat{\mathbf{G}}\mathbf{Z}^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

- The EBLUP($\mathbf{y}$) is then

$$\hat{\mathbf{y}} := \text{EBLUP}(\mathbf{y}) = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}}.$$

- There are two sources of variability: from estimating $\boldsymbol{\beta}$ and $\mathbf{u}$, and from estimating $\mathbf{G}$ and $\mathbf{R}$ (thus $\mathbf{V}$).
- From $\text{cov}(\tilde{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$, we estimate the standard error of $\hat{\beta}_i$ by

$$\widehat{\text{SD}}(\hat{\beta}_i) = \sqrt{[(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}]_{ii}}$$

Under large samples, the extra variability from estimating $\mathbf{G}$ and $\mathbf{V}$ is negligible, and thus this is a good estimate. But not for small samples. We need fully Bayesian approach for taking on variability due to estimating the variance components, and it is not covered in this course.

Likelihood ratio tests for LMM

- Normal theory tests for LMM are generally not applicable. Hence we use likelihood ratio tests instead.
- In general, the likelihood ratio for testing a null restricted model against an alternative unrestricted model is

$$\text{LR}(\mathbf{y}) = L(\hat{\boldsymbol{\theta}}_0; \mathbf{y}) / L(\hat{\boldsymbol{\theta}}; \mathbf{y}),$$

  where $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}$ are the maximum likelihood estimates of $\boldsymbol{\theta}$ under the null and the unrestricted model, respectively; $L(\boldsymbol{\theta}; \mathbf{y})$ is the likelihood function with parameter $\boldsymbol{\theta}$.
- Classical results state that

$$-2\log(\text{LR}(\mathbf{y})) \xrightarrow{\mathcal{D}} \chi^2_\nu,$$

  where $\nu = p_1 - p_0$, p_i is the number of independent parameters in model under hypothesis H_i.
- Main assumption is that the parameter of interest is NOT on the boundary of its parameter space.

Likelihood ratio tests for LMM - example

- In the pig weights example, we may be interested in testing if the intercepts of the individuals are significantly different from one another.
- This is equivalent to testing $H_0 : \sigma_U^2 = 0$ against $H_1 : \sigma_U^2 > 0$.
- The parameter space for σ_U^2 is $[0, \infty)$. Hence we are testing σ_U^2 on the boundary.
- For our pig weights model $y_i = \beta_0 + U_i + \beta_1 x_{ij} + \epsilon_{ij}$, **the subjects are independent of each other**.
- In fact, when the $\mathbf{y}$ vector can be partitioned into subvectors that are independent and the number of them goes to infinity, we have

$$-2\log\{\mathsf{LR}(\mathbf{y})\} \xrightarrow{\mathcal{D}} \frac{1}{2}\chi_s^2 + \frac{1}{2}\chi_{s+1}^2,$$

  where H_0 constraints one variance component and s regression coefficients to be 0. Note that the notation denotes it is a mixture of chi-square distributions, NOT the average of the chi-square random variables.
- In particular, under $H_0 : \sigma_U^2 = 0$, the asymptotic distribution is such that there is a 0.5 chance that $\hat{\sigma}_{U,\mathsf{ML}}^2 = 0$. And it leads to

$$-2\log\{\mathsf{LR}(\mathbf{y})\} \xrightarrow{\mathcal{D}} \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2.$$

Likelihood ratio tests for LMM

- REML can be used in carrying out hypothesis testing, but it requires that the models to be compared have the same mean structure, i.e. same fixed effects model.
- Irrespective of likelihood ratio tests using ML or REML, critical values of the tests can be obtained by simulations.
- The idea is to set the values of all fixed effect and variance component parameters equal to their estimates under the null distribution and then to simulate the distribution of the likelihood ratio test under the null model at the parameters and with the covariates equal to their observed values.
- To do this, we simulate N times a data set with fixed effect parameters at their estimated values and with the ϵ-values and random effects generated according to their estimated variances, both estimations under the null hypothesis.
- Then the likelihood ratio test statistic is calculated for each simulated data set.

Penalized splines as BLUPs in LMM

- We consider the model

$$y_i = f(x_i) + \epsilon_i, \;\; 1 \leq i \leq n,$$

where we assume $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$. We use the linear spline model

$$f(x_i) = \beta_0 + \beta_1 x_1 + \sum_{k=1}^{K} u_k (x_i - \kappa_k)_+.$$

- Let $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, $\mathbf{u} = (u_1, \ldots, u_K)^T$, with corresponding design matrices

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \text{and} \quad \mathbf{Z} = \begin{pmatrix} (x_1 - \kappa_1)_+ & \cdots & (x_1 - \kappa_K)_+ \\ \vdots & \ddots & \vdots \\ (x_n - \kappa_1)_+ & \cdots & (x_n - \kappa_K)_+ \end{pmatrix}.$$

- For the penalized spline we have a ridge regression-like fitting criterion

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^T \mathbf{D} \boldsymbol{\beta},$$

which is to be minimised. In this linear penalized spline case, we have, after division of σ^2,

$$\frac{1}{\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\|^2 + \frac{\lambda^2}{\sigma^2} \|\mathbf{u}\|^2.$$

Penalized splines as BLUPs in LMM

- Comparing to the BLUP fitting criterion of an LMM

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})^T R^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})+\mathbf{u}^T\mathbf{G}^{-1}\mathbf{u},$$

  the linear penalized spline criterion is actually in this form if we treat $\mathbf{u}$ as a set of random coefficients with

$$\text{cov}(\mathbf{u})=\sigma_U^2\mathbf{I}, \quad \text{where} \quad \sigma_U^2=\sigma^2/\lambda^2.$$

- The mixed model representation of the regression spline is then

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}+\boldsymbol{\epsilon}, \ \text{cov}\begin{pmatrix}\mathbf{u}\\ \boldsymbol{\epsilon}\end{pmatrix}=\begin{pmatrix}\sigma_U^2\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \sigma^2\mathbf{I}\end{pmatrix}.$$

- The fitted values can be rewritten as, according to the BLUP formulation,

$$\tilde{\mathbf{f}}=\mathbf{C}(\mathbf{C}^T\mathbf{C}+\lambda^2\mathbf{D})^{-1}\mathbf{C}^T\mathbf{y},$$

  where

$$\mathbf{C}=(\mathbf{X},\mathbf{Z}),\ \mathbf{D}=\text{diag}(0,0,1,1,\ldots,1),\ \lambda^2=\sigma^2/\sigma_U^2.$$

  This matches the form $\hat{\mathbf{y}}=\mathbf{X}(\mathbf{X}^T\mathbf{X}+\lambda^{2p}\mathbf{D})^{-1}\mathbf{X}^T\mathbf{y}$ for $p=1$ for formula of the fitted values of a p-th degree polynomial spline.

Penalized splines as BLUPs in LMM - example

- We simulate according to the model $y = \sin(3\pi x) + \epsilon$, $\sigma = 0.4$, $0 \leq x \leq 1$. Using linear spline model, we fit the data either
  (a) using ordinary least squares, or
  (b) using the mixed model formulation with $u_k \sim i.i.d. N(0, \sigma_U^2)$.
- In (a), we are basically treating $\sigma_U^2 = \infty$, so that the u_k's are free parameters. In (b), they are all restricted in size by $\sigma_U^2 < \infty$.

Penalized splines as BLUPs in LMM - example

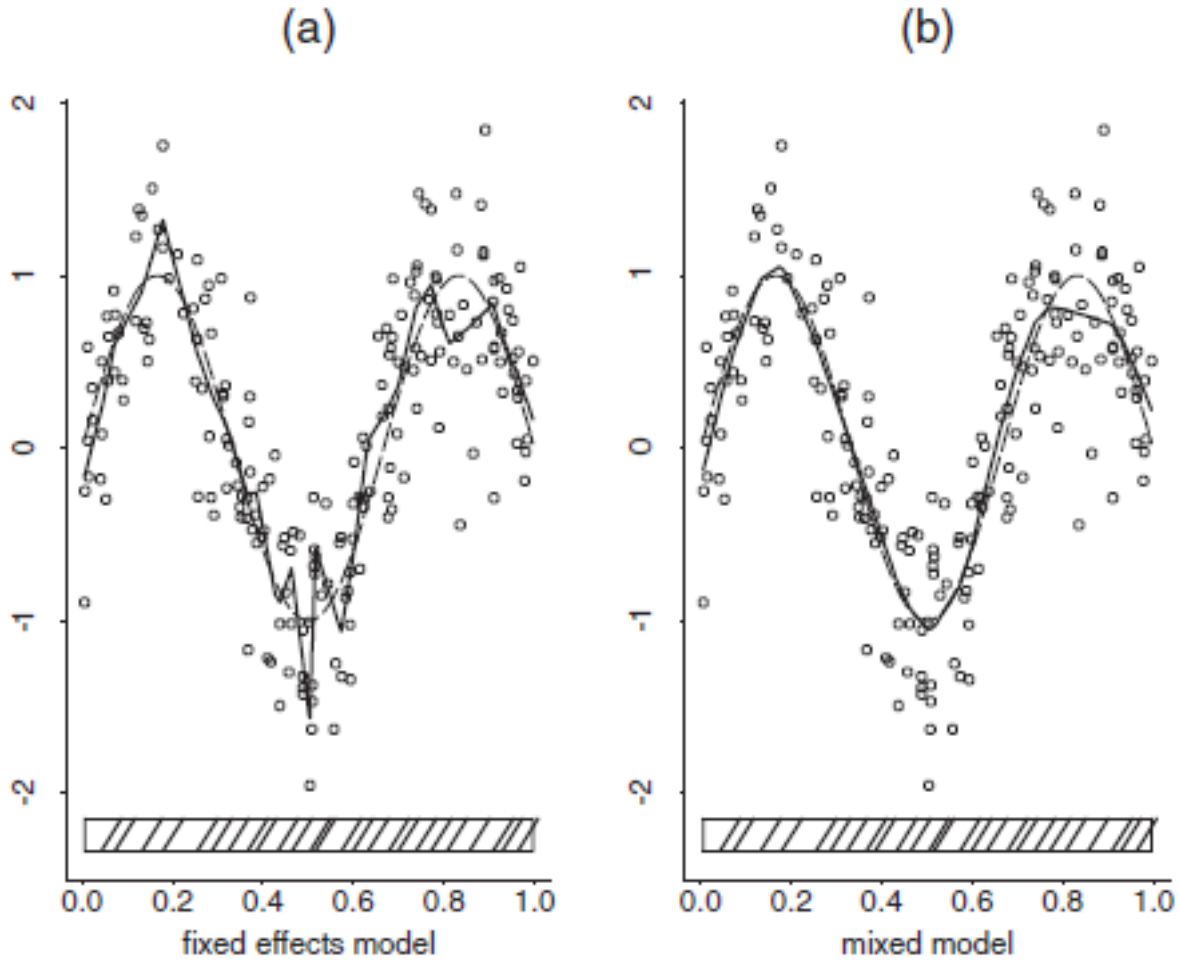

Figure: Comparison between treating the coefficients of the knots as fixed effects versus random effects. The solid curve is the estimated curve, while the dashed curve is the true function.

Degree of freedom in LMM

- In our model we have $\hat{\mathbf{y}} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{Z}\tilde{\mathbf{u}}$. We can define

$$\hat{\mathbf{y}} = \mathbf{H}_x\mathbf{y} + \mathbf{H}_z\mathbf{y}, \quad \text{where}$$
$$\mathbf{H}_x = \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}, \quad \mathbf{H}_z = \mathbf{Z}\mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{I} - \mathbf{H}_x).$$

- We define the degree of freedom for each component:

$$df_x = \text{tr}(\mathbf{H}_x), \quad df_z = \text{tr}(\mathbf{H}_z).$$

- Can show easily that df_x =Number of columns in $\mathbf{X}$ = dim$(\boldsymbol{\beta})$.
- For our pig weights data example, there are $m = 48$ pigs and $n = 9$ time points for each pig. We have $df_x = 2$, and we can show that

$$df_U = \frac{(m-1)n}{n + \sigma^2/\sigma_U^2} = 47\frac{9}{9 + \sigma^2/\sigma_U^2}.$$

- When $\sigma_U^2 = 0$, $df_U = 0$, since we basically removed the random effects completely. When $\sigma_U^2 = \infty$, $df_U = 47$, since we now have one intercept for each pig, but one parameter of intercept goes to β_0, hence 48-1=47.

Inference on $f(\cdot)$

- There are natural questions about $\hat{f}(x)$, such as standard deviation and confidence interval for $\hat{f}(x)$. We also want to see if f is in fact just linear or non-linear, or if it is monotonic.
- For the model $y = f(x) + \epsilon$, our estimator for all the design points is $\hat{\mathbf{y}} = \mathbf{L}\mathbf{y}$. For prediction with x not a design point x_i, we have $\hat{f}(x) = \ell_x^T \mathbf{y}$. Hence

$$\text{var}(\hat{f}(x)) = \ell_x^T \text{cov}(\mathbf{y}) \ell_x = \sigma^2 \|\ell_x\|^2.$$

- A pointwise variability band can then be

$$\hat{f}(x) \pm 2\hat{\sigma}\|\ell_x\|.$$

- We want adjustment to be made to the standard deviation of $\hat{f}(x)$:

  (1) variability in $\hat{\sigma}$ as an estimate of σ,
  (2) bias due to curvature,
  (3) variability in ℓ_x due to smoothing parameter estimation.

  The use of 2 before the standard deviation estimation is reasonable enough for (1). But (2) and (3) need more treatments.

Prediction and confidence intervals

- Consider the model $y_i = f(x_i) + \epsilon_i$, $\epsilon_i \sim i.i.d. N(0, \sigma^2)$, where we estimate by a general linear smoother $\hat{f}(x) = \ell_x^T \mathbf{y}$. For fixed value of smoothing parameter in ℓ_x, we have $\hat{f}(x) \sim N(E\{\hat{f}(x)\}, \sigma^2 \|\ell_x\|^2)$, and hence

$$\frac{\hat{f}(x) - E\{\hat{f}(x)\}}{\sigma \|\ell_x\|} \sim N(0,1).$$

- In linear regression, we estimate σ by $\hat{\sigma}$ and replace normal distribution by a t-distribution. In scatterplot smoothing, we have

$$\frac{\hat{f}(x) - E\{\hat{f}(x)\}}{\hat{\sigma} \|\ell_x\|} \overset{\text{approx}}{\sim} t_{[df_{res}]},$$

  where $[x]$ is the integer closest to x.
- If errors are not normal, then under large samples we still have approximate normality of the statistic on the left hand side of the above.
- Note that the intervals covered will then be centered at $E(\hat{f}(x))$ rather than $f(x)$. Hence we need to assume asymptotic unbiasedness, which is usually the case for large samples. Note that bias will be higher at peaks and valleys of the curve $f(x)$.
- For prediction intervals, we have

$$\hat{f}(x) \pm \begin{cases} t\left(1 - \frac{\alpha}{2}; df_{res}\right) \hat{\sigma} \sqrt{1 + \|\ell_x\|^2}, & \text{for small } n; \\ z\left(1 - \frac{\alpha}{2}\right) \hat{\sigma} \sqrt{1 + \|\ell_x\|^2}, & \text{for large } n. \end{cases}$$

Inference for penalized splines

- Without the mixed model assumption, the linear penalized spline has model

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} \beta_{1k}(x_i - \kappa_k) + \epsilon_i,$$

  where all parameters are constants and $\epsilon_i \sim i.i.d.N(0, \sigma^2)$.

- In this case, we have $\hat{f}(x) = \ell_x^T \mathbf{y}$, where

$$\ell_x = \mathbf{C}(\mathbf{C}^T\mathbf{C} + \lambda^2\mathbf{D})^{-1}\mathbf{C}_x^T,$$

  with $\mathbf{C}_x = (1, x, (x-\kappa_1)_+, \ldots, (x-\kappa_K)_+)^T$, $\mathbf{D} = \text{diag}(0, 0, 1, \ldots, 1)$, and $\mathbf{C} = (\mathbf{C}_1, ..., \mathbf{C}_n)^T$.

- With the mixed assumption, the model is then $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$, where for $1 \le i \le n$ and $1 \le k \le K$,

$$\mathbf{X} = (1, x_i)_{1 \le i \le n} \quad \mathbf{Z} = [(x_i - \kappa_k)_+]_{ij}, \quad \text{cov}\begin{pmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{pmatrix} = \begin{pmatrix} \sigma_U^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I} \end{pmatrix}.$$

Inference for penalized splines

- Define also
$$\mathbf{X}_x = (1, x), \mathbf{Z}_x = ((x-\kappa_1)_+, \ldots, (x-\kappa_K)_+),$$
and $\tilde{f}(x) = \mathbf{X}_x\tilde{\boldsymbol{\beta}} + \mathbf{Z}_x\tilde{\mathbf{u}}$, which is the BLUP of $f(x)$. Finally, let $\hat{f}(x) = \mathbf{X}_x\hat{\boldsymbol{\beta}} + \mathbf{Z}_x\hat{\mathbf{u}}$ which is the EBLUP of $f(x)$.
- Two kinds of arguments for modeling variability:
  (1) randomness of $\mathbf{u}$ is a device used to model curvature; only $\boldsymbol{\epsilon}$ accounts for variability about the curve. Variability of $\tilde{f}(x)$ should be assessed conditional on $\mathbf{u}$.
  (2) Randomness of $\mathbf{u}$ should be incorporated fully. Variability of $\tilde{f}(x)$ should be assessed unconditional on $\mathbf{u}$.
- With (1), we have
$$\text{var}\{\tilde{f}(x)|\mathbf{u}\} = \mathbf{C}_x \text{cov}\left(\left[\begin{array}{c}\tilde{\boldsymbol{\beta}}\\ \tilde{\mathbf{u}}\end{array}\right]\Big|\mathbf{u}\right)\mathbf{C}_x^T.$$
- In exercise 2, you will show that
$$\text{cov}\left(\left[\begin{array}{c}\tilde{\boldsymbol{\beta}}\\ \tilde{\mathbf{u}}\end{array}\right]\Big|\mathbf{u}\right) = \sigma^2\left(\mathbf{C}^T\mathbf{C} + \frac{\sigma^2}{\sigma_U^2}\mathbf{D}\right)^{-1}\mathbf{C}^T\mathbf{C}\left(\mathbf{C}^T\mathbf{C} + \frac{\sigma^2}{\sigma_U^2}\mathbf{D}\right)^{-1}.$$
- We estimate $\text{SD}\{\hat{f}(x)|\mathbf{u}\}$ by substituting σ^2 and σ_U^2 by $\hat{\sigma}^2$ and $\hat{\sigma}_U^2$ respectively.

Inference for penalized splines

- With the estimated SD for $\hat{f}(x)|\mathbf{u}$, if $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, then we use

$$\tilde{f}(x)|\mathbf{u} \sim N[E\{\tilde{f}(x)|\mathbf{u}\}, \mathsf{var}\{\tilde{f}(x)|\mathbf{u}\}],$$

  so that

$$\left.\frac{\tilde{f}(x) - E\{\tilde{f}(x)|\mathbf{u}\}}{\mathsf{SD}\{\tilde{f}(x)|\mathbf{u}\}}\right|\mathbf{u} \sim N(0,1),$$

  giving us an approximate $100(1-\alpha)\%$ confidence interval for $E\{\tilde{f}(x)|\mathbf{u}\}$ as

$$\hat{f}(x) \pm z\left(1 - \frac{\alpha}{2}\right)\widehat{\mathsf{SD}}\{\hat{f}(x)|\mathbf{u}\}.$$

- If $E\{\tilde{f}(x)|\mathbf{u}\} \approx f(x)$, then the above is similar to a confidence interval for $f(x)$.
- To adjust for bias is the same as using scheme (2), where variability is found unconditional to $\mathbf{u}$, since we have (to be shown in exercise 2)

$$E\{\tilde{f}(x) - f(x)\} = 0.$$

- We have

$$E[\{\tilde{f}(x) - f(x)\}^2] = \mathsf{var}\left\{\mathbf{C}_x \begin{bmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\mathbf{u}} - \mathbf{u} \end{bmatrix}\right\} = \mathbf{C}_x \mathsf{cov}\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} - \mathbf{u} \end{bmatrix}\mathbf{C}_x^T.$$

Inference for penalized splines

- In exercise 2, you will show that

$$\text{cov}\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} - \mathbf{u} \end{bmatrix} = \sigma^2 \left(\mathbf{C}^T\mathbf{C} + \frac{\sigma^2}{\sigma_U^2}\mathbf{D} \right)^{-1},$$

suggesting that we estimate

$$\widehat{\text{SD}}\{\hat{f}(x) - f(x)\} = \hat{\sigma}\sqrt{\mathbf{C}_x \left(\mathbf{C}^T\mathbf{C} + \frac{\hat{\sigma}^2}{\hat{\sigma}_U^2}\mathbf{D} \right)^{-1} \mathbf{C}_x^T}.$$

- Under certain assumptions,

$$\frac{\hat{f}(x) - f(x)}{\widehat{\text{SD}}\{\hat{f}(x) - f(x)\}} \overset{\text{approx}}{\sim} N(0,1),$$

and an approximate $100(1-\alpha)\%$ confidence interval for $f(x)$ is

$$\hat{f}(x) \pm \begin{cases} t\left(1-\frac{\alpha}{2}; df_{res}\right) \widehat{SD}\{\hat{f}(x) - f(x)\}, & \text{for small } n; \\ z\left(1-\frac{\alpha}{2}\right) \widehat{SD}\{\hat{f}(x) - f(x)\}, & \text{for large } n. \end{cases}$$

Inference for penalized splines

- Interval with bias adjusted is usually a bit wider, since it accounts for both components error (variance and squared bias) whereas the interval conditional on $\mathbf{u}$ accounts for only variance, and covers $E\{\tilde{f}(x)|\mathbf{u}\}$, not $f(x)$. But it is usually more preferred than the conditional version, though the latter is more commonly in use.

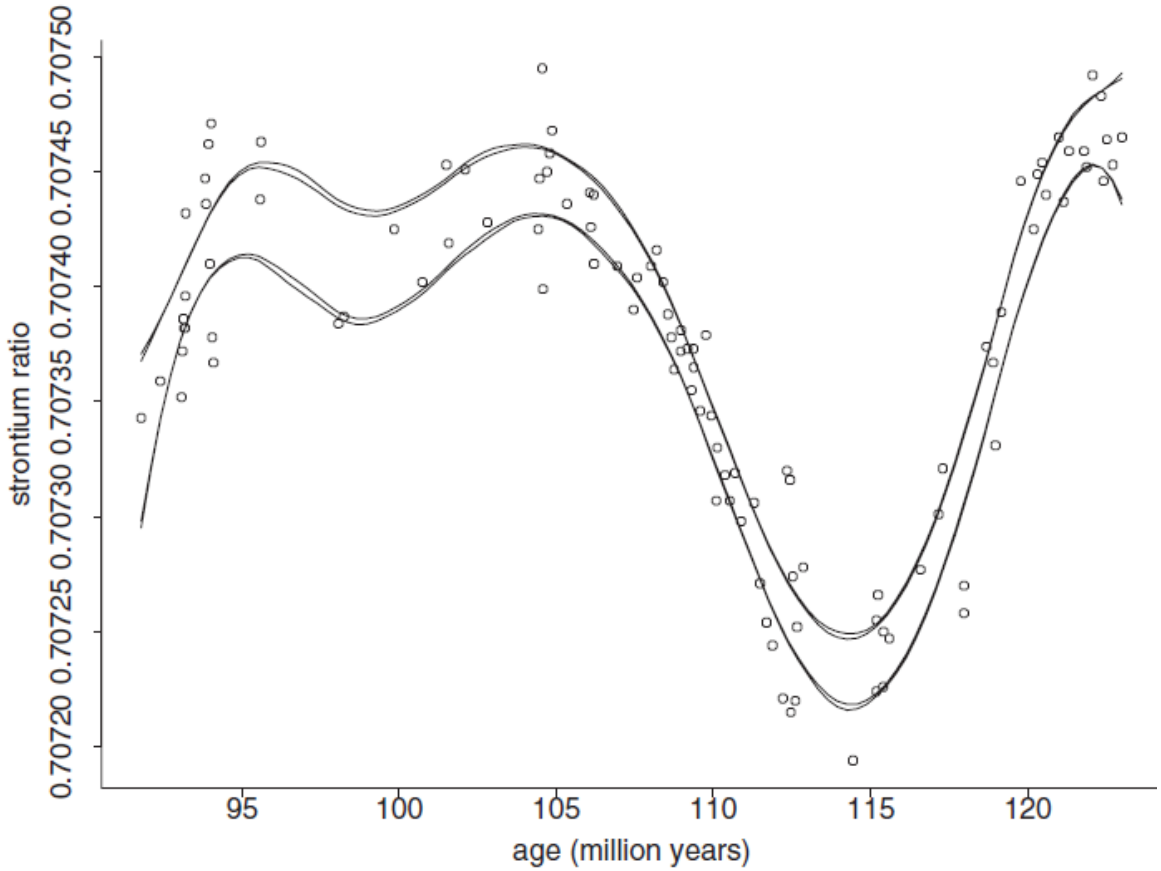

Figure: Fossil data with confidence bands that account for squared bias (outer) and do not account for squared bias (inner).

Simultaneous confidence band

- It is usually simulations based in practice. Suppose we want a simultaneous confidence band for f over a grid of M x-values $\mathbf{g} = (g_1, \ldots, g_M)$. Define

$$\mathbf{f}_g = (f(g_1), \ldots, f(g_M))^T$$

  to be the true function over $\mathbf{g}$ and let $\hat{f}_g$ be the corresponding EBLUP based on linear penalized splines in the mixed model framework.

- Then we have

$$\hat{\mathbf{f}}_g - \mathbf{f}_g = \mathbf{C}_g \begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\mathbf{u}} - \mathbf{u} \end{bmatrix} \overset{\text{approx}}{\sim} N\left\{\mathbf{0}, \hat{\sigma}^2 \mathbf{C}_g \left(\mathbf{C}^T\mathbf{C} + \frac{\hat{\sigma}^2}{\hat{\sigma}_U^2}\mathbf{D}\right)^{-1} \mathbf{C}_g^T\right\},$$

  where $\mathbf{C}_g = [1,\, \mathbf{g},\, (\mathbf{g} - \kappa_1 \mathbf{1})_+ \,, \ldots,\, (\mathbf{g} - \kappa_K \mathbf{1})_+]$.

Simultaneous confidence band

- A $100(1-\alpha)\%$ simultaneous confidence band for $\mathbf{f}_g$ is

$$\hat{\mathbf{f}}_g \pm m_{1-\alpha} \begin{bmatrix} \widehat{\mathrm{SD}}\{\hat{f}(g_1) - f(g_1)\} \\ \vdots \\ \widehat{\mathrm{SD}}\hat{f}(g_M) - f(g_M), \end{bmatrix}$$

  where $m_{1-\alpha}$ is the $(1-\alpha)$ quantile of the random variable

$$\sup_{x\in\chi} \left| \frac{\hat{f}(x) - f(x)}{\widehat{\mathrm{SD}}\{\hat{f}(x) - f(x)\}} \right| \approx \max_{1\le\ell\le M} \left| \frac{\left(\mathbf{C}_g \begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\mathbf{u}} - \mathbf{u} \end{bmatrix} \right)_\ell}{\widehat{\mathrm{SD}}\{\hat{f}(g_\ell) - f(g_\ell)\}} \right|.$$

- The idea is to simulate a realization of $\hat{f}(g_\ell) - f(g_\ell)$'s from the normal distribution described in the last page. Then the maximum statistic is calculated as above. This process is then repeated a large number of time, say $N = 10000$.
- The N simulated values of the maximum statistic are sorted from smallest to largest, and the one with rank $\lceil (1-\alpha)N \rceil$ is used as $m_{1-\alpha}$.
- Note that we cannot conclude the function is a straight line just by the fact that one can be drawn within the confidence band. The limits are basically about maximum and minimum values of the function with certain confidence level, but the structure within is not necessarily straight line.

Simultaneous confidence band

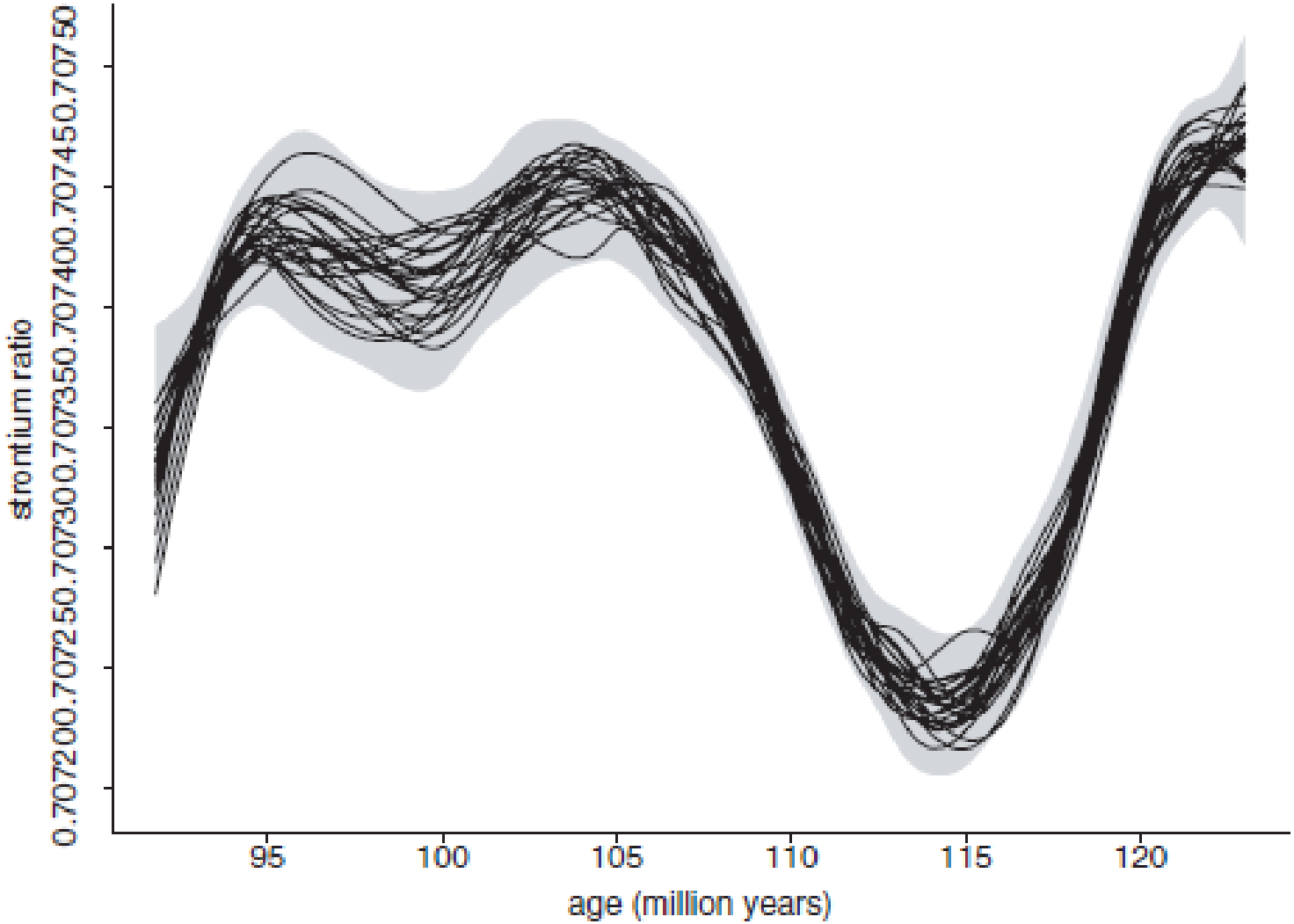

Figure: 25 simulated curves for the fossil data within the confidence band.

Restricted likelihood ratio tests

- For testing

$$H_0 : E(y|x) = \beta_0 + \beta_1 x \longleftrightarrow H_1 : E(y|x) = f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k (x - \kappa_k)_+,$$

  We can resort to the mixed model formulation, which is essentially testing

$$H_0 : \sigma_U^2 = 0 \longleftrightarrow H_1 : \sigma_U^2 > 0.$$

- If errors are $i.i.d. N(0, \sigma^2)$, then the restricted log-likelihood is

$$\begin{aligned} &-2\ell_r(\sigma_U^2, \sigma^2; \mathbf{y}) \\ &\quad = n \log(2\pi) + \log|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \log|\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}|. \end{aligned}$$

  The restricted likelihood ratio statistic is

$$-2 \log \mathsf{LR}_r(\mathbf{y}) = -2\{\ell_r(0, \hat{\sigma}_0^2; \mathbf{y}) - \ell_r(\hat{\sigma}_U^2, \hat{\sigma}^2; \mathbf{y})\},$$

  where $\hat{\sigma}_0^2$ minimises $-2\ell_r(0, \sigma^2; \mathbf{y})$.

- Null distribution is usually obtained by simulations.