1 Arabidopsis growth curves

1.1 Data description

The file `PlantGrowth.dat` contains the heights in mm. of 70 *Arabidopsis* plants measured every four days from day 29 to day 69 following the planting of seeds. The ultimate heights range from 19mm to 40mm, and most heights are recorded to the nearest millimetre.

The cumulative number of plants brearded was zero up to and including day 29, eight by day 33, 40 by day 37, and all 70 by day 41. Thus, sprouted plants were first recorded on day 33, and all plants had appeared by day 41. By day 65 or earlier, the growth was complete; for each plant, the height recorded on day 69 was the same as the height on day 65.

Plant age is most naturally measured from its birth at brearding rather than the date on which seed was planted. In this experiment, all seeds were planted on the same date, but the date of brearding varies from plant to plant. The brearding date is deemed to be the last date on which the height was recorded as zero rather than the first date on which the height was positive. In other words, eight plants were deemed to be born on day 29, 32 on day 33, and so on.

The typical growth trajectory for *Arabidopsis* begins at zero on day 0, reaching a plateau whose height varies from plant to plant. Regardless of the ultimate height, the semi-maximum height is attained in about 13 days, which is fairly constant from plant to plant. By inspection of the graphs in Fig 1, it appears that the standard *Arabidopsis* growth trajectory is roughly $h(t) \propto t^2/(\tau^2 + t^2)$, where $\tau \simeq 13$ is such that $h(\tau) = \frac{1}{2}h(\infty)$.

The graphs in Fig. 1 give the impression that the number of plants is no more than 30, but there are in fact 69 distinct growth trajectories for 70 plants. The illusion is caused in part by heights being rounded to 1mm, so that, at any fixed time, there are usually fewer than 20 distinct heights.

Two strains of plant are included in the study, the first 40 called 'cis' and the remaining 30 labelled '108'. One goal of this project was to compare the two strains and to assess the significance of the observed differences. The time series plot for all plants in Fig. 1 reveals that both types have similar growth trajectories, but that the ultimate height of the '108'strain is about 40% greater than the 'cis' strain. The age-specific ratio of sample mean heights '108'/'cis' for plants aged 4–32 days is

day	4	8	12	16	20	24	28	32
ratio	1.06	1.39	1.37	1.36	1.42	1.44	1.43	1.42

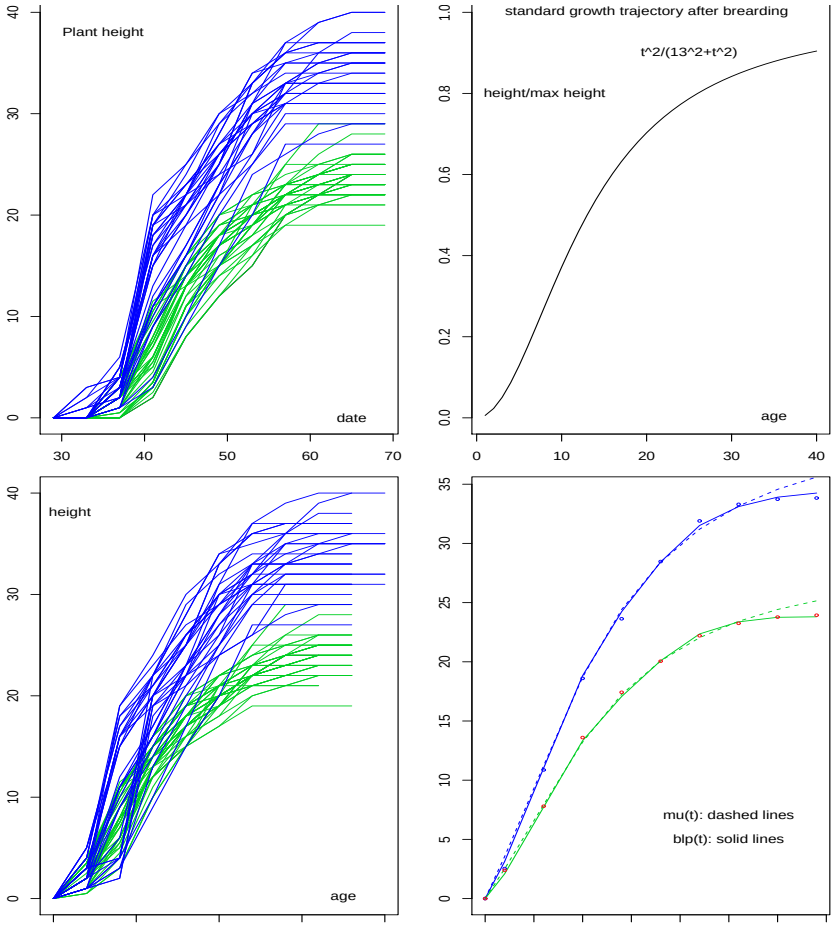

Figure 1: Heights in mm of 70 *Arabidopsis* plants of two strains, plotted against calendar time in panel 1, and against age in panel 3 (lower left). Lower right panel shows the fitted mean functions (dashed lines) together with the best linear predictor (solid lines) of plant height for each strain.

which is remarkably constant from day 8 onwards. Note that the x-axis in the first panel of Fig. 1 is calendar time, not plant age.

1.2 Growth curve models

The growth curve for plant i is modelled as a random function $\eta_i(t)$ whose value at time zero is, in principle at least, exactly zero, and whose temporal trajectory is continuous. In the analyses that follow, the mean trajectory is $\beta_{s(i)}h(t)$ with no intercept. Here $s(i)$ is the strain of plant i, the plateau levels β_0, β_1 depend on the strain, and the ratio of means is constant over time. The observation $Y(t) = \eta(t) + \epsilon(t)$ is contaminated by measurement error, which is assumed to have mean zero with constant variance σ_0^2, and to be independent for all times $t > 0$ and all plants. The error distribution for values reported as zero need not be the same as the error distribution for positive measured values.

Brownian motion (BM) starting from zero at time $t = 0$ is a continuous random function with covariance function $\mathrm{cov}(B(t), B(t')) = \min(t, t')$. We are thus led initially to consider the additive Gaussian model with moments

$$\begin{aligned} E(Y_i(t)) &= \beta_{s(i)}h(t), \\ \mathrm{cov}(Y_i(t), Y_j(t')) &= \sigma_0^2\delta_{ij}\delta_{t,t'} + \sigma_1^2 K(t,t') + \sigma_2^2\delta_{ij}K(t,t') \end{aligned} \tag{1}$$

where $K(t, t') = \min(t, t')$ for the Brownian-motion model.

One objection sometimes raised to Brownian motion as a model for a growth curve is that it is not suficiently smooth, in fact nowhere differentiable. If a compelling argument could be made that physical growth is a differentiable function, one would have to reconsider the Brownian-motion model, perhaps replacing it with a smoother random function or a family of random functions having varying degrees of smoothness. But in the absence of a compelling demonstration of differentiability, the lack of differentiability of BM is not a strong argument against its use for growth curves. The Brownian motion component of the model can be replaced by any continuous random function deemed suitable, such as fractional Brownian motion (FBM), and the data can be permitted to discriminate among these. Despite the perception that physical growth curves are smooth in time, trajectories smoother than BM are firmly rejected by the data in favour of rougher trajectories. See variation (iii) below.

Leaving aside the measurement error component, the growth-curve model (1) has two additional variance components, one Brownian motion with volatility coefficient σ_1 that is common to all plants regardless of strain, and

another with coefficient σ_2 that is plant specific and independent for each plant. In other words, for $t > 0$ the measured value on plant i is a sum of one non-random term and three independent random processes

$$Y_i(t) = \beta_{s(i)}h(t) + \epsilon_{it} + \sigma_1\eta_0(t) + \sigma_2\eta_i(t), \tag{2}$$

where $\eta_0, \eta_1, \ldots, \eta_{70}$ are independent and identically distributed Brownian trajectories starting from zero at time zero. In this model, the variances

$$\begin{aligned} \mathrm{var}(Y_i(t)) &= \sigma_0^2 + (\sigma_1^2 + \sigma_2^2)t \\ \mathrm{var}(Y_i(t) - Y_j(t)) &= 2\sigma_0^2 + 2\sigma_2^2 t \end{aligned}$$

are both increasing linear functions of plant age.

The average height at age t of a large number of plants of strain s is $\beta_s h(t) + \sigma_1\eta_0(t)$, which is a Gaussian variable with mean $\beta_s h(t)$ and covariance $\sigma_1^2 K(t, t')$. That is to say, $\sigma_1\eta_0(t)$ is the deviation of $\beta_s h(t)$ from the mean trajectory averaged over infinitely many plants of strain s. From the fitted model, the estimated, or predicted, plant height trajectory $E(Y_{i^*}(t) \,|\, \mathrm{data})$ for a new plant i^* is shown for both strains in the fourth panel of Fig. 1. Each fitted trajectory is the sum of the fitted mean $\hat{\beta}_s h(t)$ plus the conditional expected value of $\sigma_1\eta_0(t)$ given the data. The latter term $E(\eta_0(t) \,|\, \mathrm{data})$ is piecewise linear, a linear spline with knots at the observation times.

Only the 628 response values at strictly positive plant ages are included in the likelihood computations, the heights at $t \leq 0$ being exactly zero by construction. For the mean model, $\hat{\tau} = 12.842$ days is used throughout. The three variance-components estimated by maximum residual likelihood are

parameter	estimate	$S.E.$
σ_0^2	1.040	0.151
σ_1^2	0.066	0.042
σ_2^2	0.432	0.052

with asymptotic standard errors as indicated. Asymptotic standard errors of variance components are worth reporting, but are often less reliable as indicators of significance than standard errors of regression coefficients. The first coefficient implies that the standard deviation of the measurement error is around 1mm, which is about right for laboratory measurements of plant height. The small value of σ_1^2 implies that $h(t)$ is a close approximation to the mean trajectory averaged over plants, and the relatively large standard error suggests that this term may be unnecessary. Nevertheless, the reduced

model with only two variance components is demonstrably inferior: the increase in residual log likelihood is 13.78, i.e. the likelihood ratio chi-squared statistic is 27.56 on one degree of freedom. In this instance, the comparison of $\hat{\sigma}_1^2$ with its asymptotic standard error gives a misleading impression of the significance of that term.

The regression parameters governing the mean, τ included if necessary, are estimated by weighted least squares. For the 70 plants in this study, the plateau estimates in mm for the two strains are as follows:

strain	coefficient	*S.E.*
cis	28.35	1.76
108	40.13	1.92
diff	11.86	0.99

The Box-Tidwell method has been used here in the calculation of standard errors to make allowance for the estimation of τ. (The unadjusted standard errors are 1.54, 1.58 and 0.90 respectively.) This analysis makes it plain that the difference between the two strains is highly significant.

1.3 Technical points

Non-linear mean model with variance components
The inverse quadratic model for the mean growth curve

$$\mu_{it} = E(Y_{it}) = \beta_{s(i)} t^2/(\tau^2 + t^2)$$

has three parameters to be estimated, the two asymptote heights β_0, β_1 and the temporal scale parameter τ. Two options for the estimation of parameters are available as follows.

The most straightforward option is to use ordinary maximum likelihood (not REML) for the estimation of all parameters jointly. Since the model for fixed τ is linear in β, this can be done by computing the profile likelihood for a range of τ values, say $12 \leq \tau \leq 14$ in suitably small steps, and using the `kernel=0` option in `regress()` as follows.

```
h <- age^2/(tau^2 + age^2)
fit0 <- regress(y~h:strain-1, ~BM+BMP, kernel=0)
fit0$llik
```

Although all ages used in the computation are strictly positive, the model formula is such that the mean height at age zero is exactly zero. We find that

the log likelihood is maximized at $\hat{\tau} \simeq 12.782$. A plot of the profile log likelihood values against τ can be used to generate an approximate confidence interval if needed: the 95% limits are approximately $(11.7, 14.2)$ days.

A follow-up step is needed in order for the standard errors of the β-coefficients to be computed correctly from the Fisher information. To compensate for the estimation of τ, the derivative of the mean vector with respect to τ at $\hat{\tau}$ must be included as an additional covariate, as decribed by Box and Tidwell (197?)

```
deriv <- -2*tau * fit$fitted * h / age^2
fit0a <- regress(y~deriv+h:strain-1, ~BM+BMP, kernel=0)
```

It is a property of maximum likelihood estimators for exponential-family models that the residual vector $y - \hat{\mu}$ is orthogonal to the tangent space of the mean model (with respect to the natural inner product $\hat{\Sigma}^{-1}$). Consequently, the coefficient of `deriv` at $\hat{\tau}$ is exactly zero by construction, and all other coefficients β, σ^2 are unaffected. The ordinary maximum-likelihood estimates of the variance components are $(1.0467,\ 0.0497, 0.4283)$, the plateau coefficients are $(28.293, 40.042)$ mm, and the standard error of the difference is 0.975. In this instance, the unadjusted standard error is 0.941, so the effect of the adjustment is not great.

The second method is closer in spirit to REML, where the variance components are estimated from the residual likelihood, i.e. the marginal likelihood based on the residuals. The mean-value model has a three-dimensional tangent space at τ,

$$\mathcal{X}_\tau = \text{span}\{\partial\mu/\partial\beta_0, \partial\mu/\partial\beta_1, \partial\mu/\partial\tau\} = \text{span}\{hS_0, hS_1, deriv\}$$

where S_r is the indicator vector for strain r. The aim is to find $\hat{\tau}$ such that $\mathcal{X}_\tau$, is orthogonal to the residual vector. The only difference between this procedure and maximum likelihood is that the variance components, which determine the inner product matrix, are estimated by maximizing the residual likelihood rather than the full likelihood. If we fix τ, h and `deriv` as before, the command

```
fit0b <- regress(y~deriv+h:strain-1, ~BM+BMP)
```

uses the default kernel $\mathcal{K} = \mathcal{X}_\tau$ in the estimation of the variance components using REML, and $\tau = 12.842$ is such that the coefficient of *deriv* is zero. The estimated variance components are $(1.0400, 0.0659, 0.4319)$, the plateau coefficients are $(28.353, 40.138)$, and the standard error of the difference is 0.988.

When an iterative function such as `regress()` is used repeatedly in a loop as above, the overall efficiency can be substantially improved by supplying the previously-computed vector of variance components

```
fit0 <- regress(y~..., ~BM+BMP, start=fit0$sigma)
```

Estimated variance components may be negative provided that the combined matrix is positive definite. The argument `pos=c(0,1,1)` can be used to force positivity on selected coefficients.

Fitted and predicted values

The mean functions for the two strains are $\beta_0 h(t)$ and $\beta_1 h(t)$, and the fitted curves with β_s replaced by $\hat{\beta}_s$ are shown as dashed lines in the lower right panel of Fig. 1. The fitted mean is not to be confused with the predicted growth curve for a new plant i^* of strain s, which is deemed to have a response

$$Y_{i^*}(t) = \beta_s h(t) + \sigma_1 \eta_0(t) + \sigma_2 \eta_{i^*}(t) + \epsilon_{i^* t}.$$

Although this new plant is not one of those observed in the experiment, its growth trajectory is not independent of the sample responses because the covariances

$$\rho_t(i, t') = \mathrm{cov}(Y_{i^*}(t), Y_i(t')) = \sigma_1^2 \,\mathrm{cov}(\eta_0(t), \eta_0(t')) = \sigma_1^2 K(t, t')$$

are not all zero. The conditional distribution given the data is Gaussian with conditional mean

$$E(Y_{i^*}(t) \,|\, \mathrm{data}) = \beta_s h(t) + \rho_t' W(y - \mu) \tag{3}$$

where ρ_t is the n-component vector of covariances, μ is the n-component vector of means, and $W = \Sigma^{-1}$ is the inverse covariance matrix of the response values for the sampled plants. The fitted conditional distribution, or the fitted predictive distribution, has a mean $\hat{\beta}_s h(t) + \hat{\rho}_t' \hat{W}(y - \hat{\mu})$, called the best linear predictor (BLUP). This is shown as a pair of solid curves in the lower right panel in Fig. 1, one curve for each strain.

If we had taken the plant age to be two days rather than four at the time of the first positive measurement, and $h(t) = t/(\tau + t)$ to be linear at the origin rather than quadratic, the graphs of fitted means and best linear predictors would look rather different: see Fig. 2. Even with the reduced two-day offset for the origin, the inverse linear function is less satisfactory as a description of the growth curve than the inverse quadratic, so the variance coefficient σ_1^2 needs to be increased by a factor of roughly 7.3 to compensate

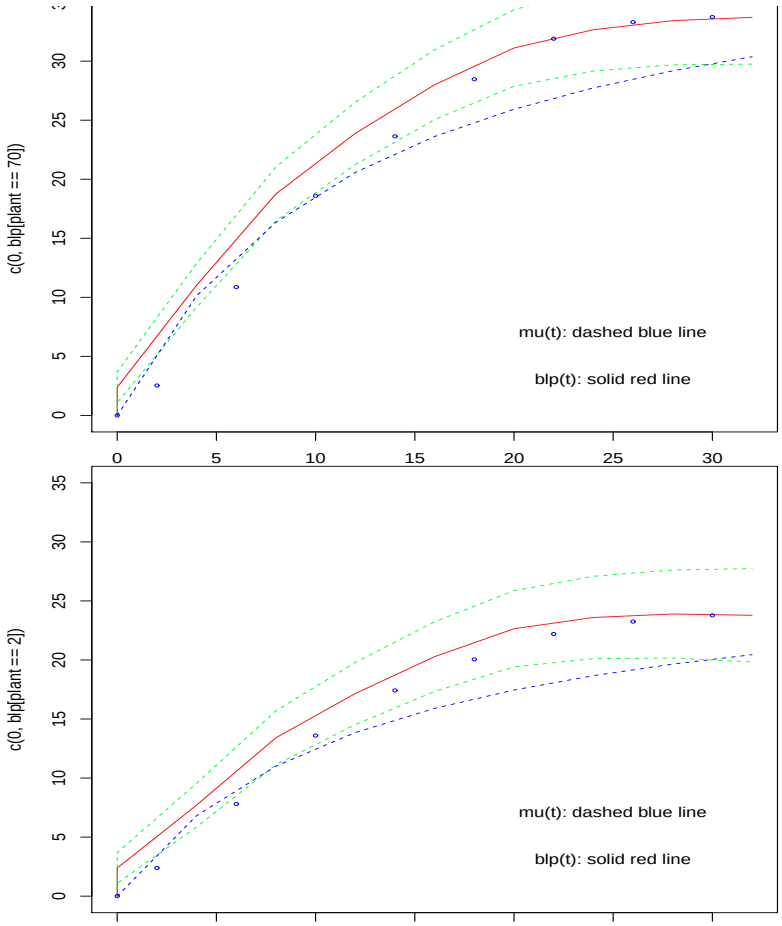

Figure 2: Fitted mean growth curves (dashed lines) and best linear predictors (solid lines) of plant height for two strains, using an inverse linear model for the mean trajectory and Brownian motion for the deviations. Sample average heights at each age are indicated by dots.

for the larger deviations. Using log likelihood for model comparison, the inverse linear model is decisively rejected. Despite the less satisfactory fit, the best linear predictor for the inverse linear model (shown as the pair of solid lines in Fig. 2) is not appreciably different from the best linear predictor for the inverse quadratic model in Fig. 1. The maximum difference is approximately one millimetre (or 3%) at age 40 days. The difference between fitted means is much larger.

In certain areas of application such as animal breeding, the chief goal is to make predictions about the meat or milk production of the future progeny of a specific individual bull. This bull is not an extra-sample individual, but one of those experimental animals whose previous progeny have been observed and measured. Such predictions are seldom called for in plant studies. Nevertheless, from a probabilistic viewpoint, the procedure is no different. If i^* is one of the sampled plants and t is an arbitrary time point, the covariance of $Y_{i^*}(t)$ and $Y_i(t')$ is

$$\rho_{i^*t} = \sigma_1^2 K(t,t') + \sigma_2^2 \delta_{i,i^*} K(t,t'),$$

which involves two of the three variance components. The conditional expected value (3) then yields an interpolated curve for each plant.

Variations on the theme

(i) In the inverse quadratic model, the height of plant i at age t is Gaussian with mean $\beta_{s(i)}h(t)$ whose limit as $t \to \infty$ is $\beta_{s(i)}$. What is the variance of the ultimate height?

(ii) For the inverse linear model in which brearding is deemed to have occurred two days prior to the first positive measurement, estimate τ together with the plateau coefficients. Obtain the standard error for the estimated limiting difference of mean heights for the two strains.

(iii) The Brownian motion component of the model can be replaced with fractional Brownian motion with parameter $0 < \nu < 1$, whose covariance function is

$$\operatorname{cov}(Y(s), Y(t)) = s^{2\nu} + t^{2\nu} - |s-t|^{2\nu},$$

where $s, t \geq 0$. The index ν is called the Hurst coefficient, and $\nu = 1/2$ is ordinary Brownian motion. Show that the fit of the plant growth model can be appreciably improved by taking $\nu \simeq 1/4$.

(iv) Bearing in mind that the heights are measured to the nearest millimetre, comment briefly on the magnitude of the estimated variance components for the FBM model.

(v) In the fractional Brownian model with $\nu < 1/2$, the temporal increments for non-overlapping intervals are negatively correlated. Suggest a plausible mechanism that could lead to negative correlation.

(vi) For 1000 equally spaced t-values in $(0, 10]$ compute the FBM covariance matrix K and its Choleski factorization $K = L'L$. (If $t = 0$ is included, K is rank deficient, and the factorization may fail.) Thence compute $Y = L'Z$, where the components of Z are iid standard Gaussian, and plot the FBM sample path, Y_t against t. Repeat this exercise for various values of ν in $(0, 1)$ and comment on the nature of FBM as a function of the Hurst coefficient.

(viii) Several plants reach their plateau well before the end of the observation period. How is the analysis affected if repeated values are removed from the end of each series?

(ix) Explain the Box-Tidwell method.

(x) Investigate the relation between brearding date and ultimate plant height. Is it the case that early-sprouting plants tend to be taller than late-sprouting plants?

1.4 Modelling strategies

1. Choice of temporal origin. The distinction between calendar time and plant age is fundamental. The decision to measure plant age relative to the time of brearding is crucial, and has a greater effect on conclusions than any subsequent choice.

2. Selection of a characteristic mean curve. The mean curve must pass through the origin at age zero, so a logistic function $e^t/(1+e^t)$ cannot be used. The graphs in Fig. 1 suggest an inverse quadratic curve, which may or may not be appropriate for other plants.

3. Use of a non-stationary covariance model. Plant growth curves are intrinsically non-stationary because they are tied to the origin at age zero. Animal growth curves using weight in place of height are not similarly constrained.

4. Brownian motion. It seems reasonable that every growth curve should be continuous. It seems reasonable also to model the response as a sum of the actual height plus measurement error, thereby making a distinction between plant height and the measurements made at a finite set of selected times. The particular choice (BM) is not crucial, and can be improved by FBM. It is also possible to mix these by

using FBM for the plant-specific deviation, and BM for the common deviation, or vice-versa.

5. Positivity. Plant heights are necessarily positive at all positive ages, whereas any Gaussian model puts positive probability on negative heights. This is one of those minor compromises that is frequently needed in applied work.

6. Response transformation, usually $y \mapsto \log(y)$, is an option that must always be considered. The log transformation might be reasonable for animal growth curves, but it was rejected here because of the role of zero height in determining the age origin.

7. Limiting behaviour. Plants do not grow indefinitely or live for ever, so the capacity of the growth model for prediction is limited to the life span of a typical plant.

8. Other differences. The emphasis on growth curves overlooks the possibility that the two strains may differ in other ways. In fact, the average brearding time for strain '108' is two days less than the time for strain 'cis', with a standard deviation of 0.43 days. No single summary tells the whole story.

1.5 Miscellaneous R functions

The following is a list of various R functions used in the construction of covariance matrices, and in the fitting of variance-components models.

```
BM <- outer(age, age, "pmin")     (BM covariance matrix)
FBM <- outer(age^p, age^p, "+") - abs(outer(age, age, "-"))^p
Plant <- outer(plant, plant, "==")     (plant block factor)
BMP <- BM * Plant     (component-wise matrix multiplication)
FBMP <- FBM * Plant     (iid FBM for each plant)
mht0 <- tapply(height[strain==0], age[strain==0], mean)
mht1 <- tapply(height[strain==1], age[strain==1], mean)
tapply(brearded, strain, mean)
L <- t(chol(FBM))  (Choleski factorization)
fit <- regress(y~h:strain-1, ~BM+FBMP, kernel=0)
```

Computer files

```
PlantGrowth.dat  PlantGrowth.R
```