Example 5

5.1 Evolution of lice on captive pigeons

5.1.1 Backround

Understanding the mechanisms responsible for the origin of new species is a fundamental topic in evolutionary biology that has been the focus of numerous experiments and much speculation dating back at least to Darwin, who argued that differential natural selection in a range of environments leads to reproductive isolation and thence, eventually, to the formation of new species. Chapter 3 is concerned with speciation induced by differential diets over approximately 40 generations of *Drosophila*. This chapter considers another experiment on the same theme, but with a different system and different environmental pressures.

The paper *Rapid experimental evolution of reproductive isolation from a single natural population* published by Villa *et al.* (PNAS 2019) is concerned with reproductive isolation developing in response to body-size evolution in isolated lineages of pigeon lice. Each lineage evolved over 60 generations on a different host pigeon. Half of the hosts were normal-sized captive feral pigeons, the other half were giant runts.

To establish their claim, the authors must show evidence of two phenomena: first that louse size evolves rapidly in giant runt hosts relative to that in captive feral hosts, and second that differential louse size induces sexual isolation. The evidence for both of these phenomena is essentially statistical. The mechanism by which size differences lead to reproductive isolation is important from an evolutionary standpoint, but this chapter deals only with size evolution, i.e., whether systematic louse-size changes are detectable in a 60-generation span. Our concern is not so much with the evolutionary implications of the authors' findings, but with the experimental design, the data analysis, and the inferences that follow. The goal is solely to examine the data for evidence of systematic body-size changes in response to host size.

5.1.2 Experimental design

The following synopsis of experimental procedure is taken directly from Villa *et al.* Before the start of the experiment, resident lice on all experimental pigeons were eradicated by housing the birds in low-humidity conditions for at least ten

weeks. According to the authors, this procedure kills both lice and eggs, while avoiding residues from insecticides. To begin the experiment, 800 lice taken from wild-caught feral pigeons were transferred to 32 lice-free experimental pigeons, 25 lice per host. Half of the experimental hosts were captive feral pigeons; the other half were giant runts, a domesticated breed that is threefold larger than feral pigeons. Pigeons were housed in eight aviaries, each aviary containing four birds of the same breed. Every six months, a random sample of lice from each bird was removed, photographed, and returned to the host. The sex, body length, metathorax width, and head width of each louse was recorded.

One aspect of this design is different from that in chapter 3. After measurements were made, the lice were returned to their host. This was done in order to minimize the effect of measurement on the host-parasite system. Otherwise, the act of measurement would reduce the resident population, and introduce instability in the lineage, which is not desirable. In the design in chapter 3, the flies removed for experimental purposes were reared separately for one generation on a standard diet, so it was not possible to return them to the main breeding line. However, the breeding lines were more tightly controlled, so plans could be made in advance to accommodate the numbers needed in any particular generation.

As always in situations of this sort, the phrase 'random sample of lice from each bird' must be treated with caution, particularly with regard to size measurements. Larger lice are more visible than smaller specimens, so it would be naive to expect the random sample to behave like a simple random sample of the resident lice on a given bird. Nonetheless, size-biased sampling need not be a serious concern for this experiment provided that it affects all birds equally.

5.1.3 Deconstruction of the experimental design

Since each measurement is made on one louse, it is evident that each observational unit is either one louse or one louse on one occasion, while the response Y_u is a point in the state space, which is $\{M, F\} \times \mathbb{R}^3$. Since a louse generation is approximately 24–25 days, and measurement occasions are six months apart, we can be sure that no louse was measured on more than one occasion. While there is no practical distinction between louse and louse-occasion as the observational unit, as a matter of principle the ordered pair is the correct choice.

The lice are arranged in 32 lineages, one lineage to each bird. Thus *lineage* and *bird* are equivalent as block factors, and *aviary* is a coarser partition or block factor with eight levels. With respect to birds, *host* or *breed* is a binary classification factor.

The baseline is the time at which de-lousing was complete, and the experiment was ready to commence with new lice lineages on captive birds. The paper mentions randomization only incidentally in the 'Materials and Methods' section, and the reference there is a little ambiguous, but two crucial choices appear to have been made at baseline. First, the 800 initial lice were partitioned into 32 lineages with 25 founders for each lineage. Second, each lineage was associated with a particular bird. Regardless of the biological and mechanical constraints in the laboratory, it seems reasonable and mathematically natural

to regard each of these steps as the outcome of an independent uniform randomization scheme. Since the objective is to study selective pressure, host size is the principal treatment. If the randomization was done in two steps as indicated, treatment is assigned to lineages in step two, in which case each lineage serves as one experimental unit.

By definition, a covariate is a pre-baseline variable, and it appears that there is only one. Measurement occasion or *time* is a function on the observational units, which is a quantitative factor. However, as indicated in the preceding paragraph, lineage could be regarded as a pre-baseline block factor, and it should certainly be used as the experimental unit to assess variability of the treatment effect estimate.

In addition to *time* and *lineage*, pre-baseline vital measurements including louse *sex* are available on the 800 founder lice. All pre-baseline variables are available for use as covariates as if the values were fixed and non-random, and initial response values are no different in that respect from any other pre-baseline measurements. Randomization ensures that the distribution of initial values is the same for all treatment groups, so the initial response values are uninformative for treatment effects. Generally speaking, when the response is a time series or temporal process, it is more convenient and mathematically more natural to treat initial response measurements as an integral part of the response process. A crucial point is that the probability model for the response at $t = 0$ must be consistent with the randomization: see sections 5.2.3, 5.2.5 and 11.4.5. The joint distribution implies a conditional distribution, which is available if needed for purposes of estimation or prediction.

The paper does not discuss how birds were assigned to aviaries, but it seems reasonable to regard that too as the outcome of a balanced randomization applied to birds, subject to restrictions mentioned earlier. We presume here that birds were quarantined in their aviaries during de-lousing, in which case *aviary* is a pre-baseline block factor. Since all birds in one aviary are of the same breed, a strong argument can be made that *aviary* is the experimental unit, not *lineage* as stated earlier. Both seem to be relevant. Whether they are pre-baseline or immediately post-baseline, *time*, *lineage*, *aviary*, and *treatment* are available for purposes of analysis and model construction.

Apart from the founders, louse sex is a post-baseline variable, and thus one of four components in the response. Genetic theory leads one to expect the sex ratio should steady at 50:50, and post-baseline counts in Table 5.3 confirm this. But the same table also shows that the baseline F:M ratio is 464:336, which is significantly in excess of 50:50.

Each lineage was associated with a particular pigeon at baseline, which means that *lineage* and *pigeon* are equivalent as block factors. A subsequent remark in the paper shows that this statement is not quite correct. When a bird died during the experiment, all lice from the dead bird were transferred to a new parasite-free bird of the same type. Thus, one lineage could span two or more birds. Unfortunately the data file does not indicate when deaths might have occurred, so we have no way to check the effect on lineages of host transfers.

5.2 Data analysis

5.2.1 Role of tables and graphs

However it is measured, the response of evolutionary interest is louse size. To keep matters as simple as reasonably possible, we focus on the single response, body length. Since we plan to use additive decompositions, the log transform is more or less automatic, so Y_u is the log body length for louse u. However, the range of variation in all size measurements is only a few percent of the average, so the log transformation has little effect on conclusions.

The purpose of a table or graph is to advance the narrative thread by drawing attention to the most important patterns or features in the data such as the nature and direction of various effects. It is natural enough to emphasize the effects of scientific interest—but not at the cost of misleading the reader. Every table or graph invites the question 'What is the point of this table?' or 'What feature does this graph illustrate?'. If the answer is not clearly apparent, the narrative is not advanced, and the reader is likely to be confused. Generally speaking, the data analyst examines numerous tables and graphs. Only the most useful of these are retained for presentation.

Table 5.1. Average log body length (in μm) of lice on two pigeon hosts

		Time in months								
Sex	Host	0	6	12	18	24	30	36	42	48
F	Feral	7.883	7.883	7.883	7.874	7.866	7.886	7.880	7.872	7.864
F	G.R.	7.885	7.894	7.882	7.882	7.882	7.895	7.894	7.899	7.886
M	Feral	7.720	7.716	7.705	7.700	7.702	7.712	7.709	7.713	7.700
M	G.R.	7.720	7.718	7.717	7.716	7.713	7.723	7.726	7.731	7.720
		Differences ×100: Giant runt − Feral								
F	G−F	0.2	0.1	−0.1	0.8	1.7	0.9	1.4	2.6	2.2
M	G−F	0.0	0.2	1.2	1.7	1.1	1.1	1.7	1.8	2.0

The first four rows of Table 5.1 show the average log body length of all lice measured on each occasion. Most impressive is the stability of body length for both louse sexes over 60 generations. If anything, there is a slight decrease in length for lice on both hosts, with a slightly greater decrease for captive feral pigeons.

The numbers in Table 5.1 are accurate to three decimal places or four decimal digits, but the first three digits are essentially constant at 7.88 for females and 7.72 for females, so we say that there are only 1–2 significant decimal digits. Usually, one is not enough to gauge accurately the statistical variation in the process. However, we have chosen to leave the table in its present form to emphasize how tiny are the size differences between lice on the two hosts.

The sex difference $7.88 - 7.72 = 0.16$ means that female lice are about 16% longer than males. The last two rows show that the mean difference for hosts tends to increase over time, reaching around 2% for both sexes after 48 months.

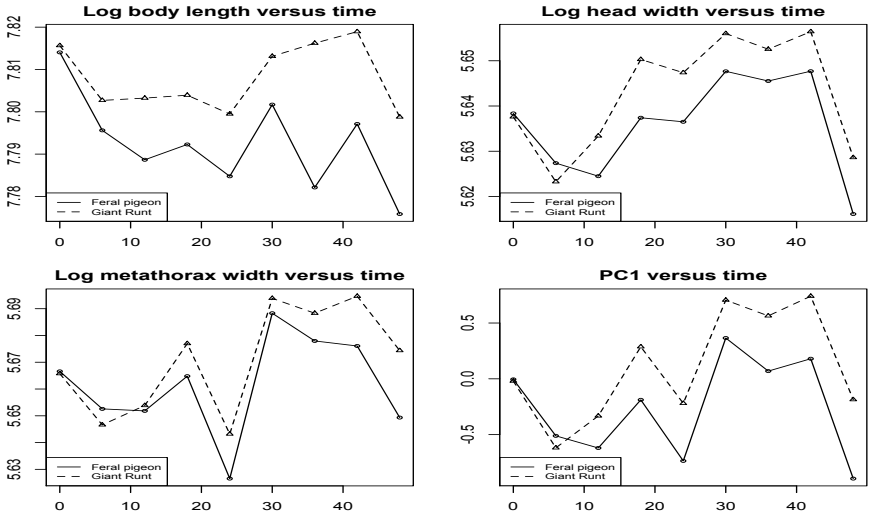

Figure 5.1: Average body sizes of lice for two hosts over time

It is remarkable that such a small size difference could have a detectable effect on sexual coupling.

The first panel of Figure 5.1 shows a plot of the same data with sexes combined. Automatic centering and re-scaling of the y-axis has the effect of exaggerating the variation and the magnitude of the divergence between the two groups. In other words, that which is emphasized by the table of averages is eliminated by the plot.

The remaining panels show similar plots for the head width, the metathorax width, and the first principal component, which is a roughly equally-weighted positive linear combination of the three standardized size variables. For all size variables, the temporal trajectory for louse size on giant runts is surprisingly similar to that for feral pigeons, and lice on giant runts are larger on average than those on feral pigeons. Apart from the uniform decrease in all size measurements in the initial and final intervals, no clear temporal trend is visible.

Ideally, it would be good to show error bars for every point. But size measurements for different lice on one pigeon are not independent, so honest error assessment is not straightforward. On balance, it is better to show no error bars than to show the naive default based on independence, which is misleading in this setting: contrast Fig. 5.2 with the table at the end of section 5.5.

5.2.2 Trends in mean squares

Table 5.1 and Fig. 5.1 illustrate temporal trends in average body size. To get a comparable impression of trends in variance, it is helpful to compute mean-squares associated with *louse sex*, *host size*, *aviary*, *lineage* and residuals at each of the nine time points.

Table 5.2. Trends in mean squares and variance components $\times 10^5$

	Time t in months								
MS	0	6	12	18	24	30	36	42	48
Host	12	340	190	996	1095	690	1024	3448	1506
Aviary	290	204	408	575	493	333	324	528	350
Lineage	83	85	85	94	87	80	100	79	152
Residual	111	52	60	68	56	52	56	57	64
	Variance-component estimates ($\times 10^5$)								
$\hat{\sigma}^2_{\text{aviary}}$	2.1	3.2	10.7	10.4	9.5	6.5	6.7	10.6	6.6
$\hat{\sigma}^2_{\text{lineage}}$	−1.1	2.8	3.3	3.8	2.8	2.2	5.5	0.3	12.1
$\hat{\sigma}^2_{\text{resid}}$	111.3	53.1	59.4	67.3	56.9	52.5	56.1	58.4	63.3

The dominant mean square is that for *louse sex* which starts off at 5.25 at baseline, drops to half that value at six months and decreases slowly to 1.64 at 48 months. For the other factors, the mean squares are shown in the top half of Table 5.2, together with the REML variance components for *aviary*, *lineage* and residual in the second half. For this fit, *host* and *sex* were eliminated as fixed effects, so the mean-squared residual does not coincide exactly with $\hat{\sigma}^2_0$.

Some of the following points are accommodated in subsequent analyses, but others are merely noted.

1. The residual variability at baseline is twice that on all subsequent occasions. One plausible explanation is that founder lice collected from wild pigeons are more variable in size than those resident on captive pigeons.

2. The lineage mean square is remarkably constant from baseline onwards. Relative to the residual mean square, it is below expectation at baseline, but not significantly so. After baseline, it is uniformly larger than the residual mean square, but not by a large factor.

3. The host mean square at baseline seems artificially low. There is strong evidence in the data, for example in the sex ratios, that the randomization scheme was more complicated that that depicted in the preceding section, so this may be a consequence of an effort to balance the randomization.

4. The between-aviary mean square at baseline is a little larger than expected from uniform random assignment: the F-ratio is 2.6, which is is at the upper 1.6% point of the reference null distribution.

5. Variance-component estimates on few degrees of freedom, such as those for aviary and lineage, have notoriously high variances.

The main issue to be addressed at this point is the size of the aviary mean square at baseline, and whether the mean square provides sufficient probative evidence to cast doubt on the randomization or to declare it inadequate or biased. The question is not whether the initial lice were labelled 1–800 and lots

drawn to determine which lice would be assigned to which birds, but whether the laboratory procedures actually employed are a reasonable facsimile of objective randomization. The only evidence before the court is shown in Table 5.1.

One traditional view is that the aviary mean square is selected for attention as the largest of three or four, so the p-value, or measure of extremity, is closer to 5%. That calculation tells us something, but it does not answer directly the question of interest to the court: 'Given the data, what is the probability that the allocation to aviaries was biased?' From another viewpoint in which sparseness prevails at odds level ρ, the odds against aviary bias given the mean-square ratio $F = 2.6$ on $6,367$ degrees of freedom are approximately $\rho\zeta_6(2.6)$, where $\zeta_6(2.6) = 3.81$. This calculation uses a modification for F-ratios of the sparsity argument in McCullagh and Polson (2018). The strength of the evidence is such that the initial presumption of innocence with probability $1/(1+\rho)$ is changed to $1/(1+\rho\zeta_6(2.6))$. For $\rho = 0.1$, which is not a strong prior presumption for this setting, the probability of a no aviary bias is changed by the evidence from 0.91 to $1/1.38 = 0.72$. So we take note and proceed with caution, giving the randomization a provisional pass. This point is revisited in section 5.4.2.

5.2.3 Initial values and factorial subspaces

If host size has an effect on louse size, it is an evolutionary development, so the effect is not immediate. Thus, *treatment* and *time* are the principal covariates whose effects are to be studied. In addition, the body length for *C. columbae* male lice is approximately 85% of that for females, so louse sex must also be taken into consideration. The effects of *lineage* and *aviary* are assumed to be additive random variables with independent and identically distributed components for each pigeon and each aviary respectively. Since their effects are additive zero-mean random variables, *lineage* and *aviary* do not contribute to the mean-value subspace.

Setting the two block factors aside temporarily, the factors *treatment* or *host size*, *time* and *sex* are to be taken into account. If we proceed to use factorial models in the naive manner, we may begin with all three main effects and check which interactions are needed. Or we may follow the authors' practice in their Tables S2–S5, which is to report the coefficients in the full three-factor interaction model. Both approaches are technically incorrect. Fitting either of the suggested models is a pointless exercise that serves only to confuse the narrative thread for this experiment.

The problem with the naive application of factorial models to this design lies in the role of time, and the fact that $t = 0$ corresponds to the experimental baseline. If Y_{ut} denotes the log body length of louse u at time t, the additive main-effects model for the conditional mean given sex and host has the form

$$E(Y_{ut} \mid s, h) = \beta_0 + \beta_1 t + \beta_2 s(u) + \beta_3 h(u),$$

in which $h(u)$ is a code for the host size, and $s(u)$ is the louse sex. At baseline,

the additive model implies

$$E(Y_{u0} \mid s, h) = \beta_0 + \beta_2 s(u) + \beta_3 h(u),$$

with three coefficients to be estimated. The presumption of randomization, which is that lice are assigned to hosts independently of their size, implies $\beta_3 = 0$. Thus, randomization contradicts both the additive model and any other factorial model that contains it as a subspace.

Whether or not randomization was explicitly employed in this experiment, it is reasonable to imagine or suppose that the initial assignment of lice was effectively randomized. Randomization has implications. The use of a model that contradicts those implications is a source for confusion; the use of a model that conforms with randomization is strongly advised.

The phenomenon described here—of time in relation to treatment and initial values—is not new. A simple example is given in Exercise 3.11 of McCullagh and Nelder (1987).

Only the most cynical reader would seriously consider the possibility that the researchers had deliberately assigned the lice differentially to hosts or to aviaries in an inappropriate manner. However, there might well be sound biological arguments for balancing the design in certain ways or for favouring females in the establishment of lineages. Deviations of this sort are normal practice, but they should be reported. Nonetheless, unintentional biased assignment can occur, so it is routine in many areas of application to check whether the baseline values are in conformity with randomization. That can be done here. While there is no indication of bias in Fig. 5.1, randomization implies that the mean squares for *aviary*, *lineage* and residual have the same expected value at baseline. However, the aviary-to-residual mean-square ratio in Table 5.2 is 2.61, which falls near the upper 98.5 percentile of the null distribution. This is not proof positive of bias, but it is a little troubling and calls for an explanation.

5.2.4 A simple variance-components model

The following linear models address directly the question that is of principal interest to an evolutionary biologist. Without straying from linearity in time, the null and alternative may be formulated as linear subspaces.

$$H_0: \quad E(Y_{ut}) = \beta_0 + \beta_1 t + \beta_2 s(u); \tag{5.1}$$

$$H_A: \quad E(Y_{ut}) = \beta_0 + \beta_{h(u)} t + \beta_2 s(u). \tag{5.2}$$

The model formulae `time+sex` and `host:time+sex` generate basis vectors for the two subspaces whose dimensions are three and four respectively. The alternative model has two linear trends in time, one for captive feral hosts $h(u) = 0$, and one for giant runts $h(u) = 1$.

For covariances, we start out following the authors' suggestion with three variance components

$$\operatorname{cov}(Y_u, Y_{u'}) = \sigma_0^2 \delta_{u,u'} + \sigma_1^2 \delta_{l,l'} + \sigma_2^2 \delta_{a,a'}, \tag{5.3}$$

where l, l' and a, a' are the lineages and aviaries respectively. This is a linear combination of three identity matrices, one on the lice, one on the lineages or pigeons with 32 blocks, and one on the aviaries with eight blocks. It is usually justified either by appeal to exchangeability based on recorded similarities of observational units, or, if that argument fails to convince, by appeal to randomization. Although neither argument carries weight in this instance, computation is cheap so we proceed.

For the log body length, the REML variance components in (5.3) paired with (5.2) are

lice	$\hat{\sigma}_0^2$	78.19×10^{-5},
lineages	$\hat{\sigma}_1^2$	1.84×10^{-5},
aviaries	$\hat{\sigma}_2^2$	1.40×10^{-5}.

Both the lineage and aviary variance components are small relative to the between-lice variance. Despite that, there is no compelling reason to declare them null simply because they are small. The fitted slope coefficients ($\times 10^4$) for the two pigeon breeds are

Parameter	Estimate	s.e.
Feral:time	−2.23	0.53
Giant:time	1.37	0.38
Difference	3.60	0.63

This analysis appears to provide reasonably strong evidence that lice transferred to captive feral pigeons decrease in size over time, and moderately strong evidence that lice transferred to giant runts increase in size over time. However, the analysis is based on linearity in time, which seems implausible given Fig. 5.1, and a covariance structure (5.3) that is both inadequate for the data and in conflict with randomization.

5.2.5 Conformity with randomization

Randomization implies that the body-size measurements at $t = 0$ are exchangeable with respect to some group of permutations, here assumed to be large enough that the responses for every pair of lice have the same joint distribution regardless of whether they are assigned to the same pigeon, to different pigeons in the same aviary or to different pigeons in different aviaries. Unfortunately, randomization implies $\sigma_1 = \sigma_2 = 0$ in (5.3).

Ever since the pioneering work of Edwards and Cavalli-Sforza (1963, 1964), Brownian motion has been the standard probabilistic model for the neutral evolution of a quantitative trait (Felsenstein, 2004, chapter 23). The conflict with randomization can be fixed only by introducing non-stationary temporal processes for the lineage and aviary effects, and the most natural way to incorporate Brownian motion is as follows:

$$\operatorname{cov}(Y_u, Y_{u'}) = \sigma_0^2 \delta_{u,u'} + \sigma_1^2 K(t,t')\, \delta_{l,l'} + \sigma_2^2 K(t,t') \delta_{a,a'} + \sigma_3^2 K(t,t'). \qquad (5.4)$$

The Brownian covariance function $K(t,t') = \min(t,t')$ is positive semi-definite, and $K(0,0) = 0$ ensures conformity with randomization. Environmental selective pressure exerts a genetic drift, and the mean model (5.2) contains one drift parameter for each host, so the differential drift is the treatment effect.

The rationale for (5.4) is as follows. The louse population as a whole evolves as a Brownian motion with volatility σ_3; each aviary evolves independently as a Brownian motion with volatility σ_2; and each lineage evolves independently as a Brownian motion with volatility σ_1. For the duration of this experiment, each louse belongs to the system, a lineage and an aviary, and the value for the louse is the sum of these three processes plus white noise. All three processes are neutral or drift-free. Drifts associated with host size occur in (5.2).

The REML log likelihood achieved by this Brownian modification exceeds that for (5.3) by approximately 57.9 units, and all four fitted coefficients are positive. Although these models are not nested, the difference is huge enough to leave no doubt that (5.3) is totally inadequate for these data.

The effect of these temporal correlations on the fitted regression coefficients is small but not negligible; their effect on standard errors is an eight-fold increase. The fitted slope coefficients ($\times 10^4$) for the two pigeon breeds are

Parameter	Estimate	s.e.
Feral:time	−4.22	4.1
Giant:time	0.33	4.1
Difference	4.55	3.3

The conclusion from this analysis is the essence of simplicity: the data are entirely consistent with neutral evolution of louse size on both hosts.

Apart from the Brownian contribution, Table 5.2 shows that the baseline variance is substantially larger than the residual variance on subsequent occasions. This observation suggests that (5.4) is not adequate on its own, and must be supplemented by an additional diagonal matrix for baseline observations. This differential baseline variance leads to a further 64.7-unit increase in the REML criterion. However its effect on conclusions is almost negligible; for comparison, the fitted coefficients ($\times 10^4$) are as follows:

Parameter	Estimate	s.e.
Feral:time	−4.39	4.3
Giant:time	0.30	4.1
Difference	4.69	3.6

The conclusion regarding neutrality of evolution is unaffected. The apparent evidence for a differential trend in the analysis at the end of section 5.2.4 is a consequence of a demonstrably inadequate variance assumption.

Brownian motion in (5.4) does a reasonable job of describing the temporal dependence, but the fit can be improved by using a low-index fractional Brownian motion. However, this and other modifications discussed in section 5.4 have only a small effect on drift estimates.

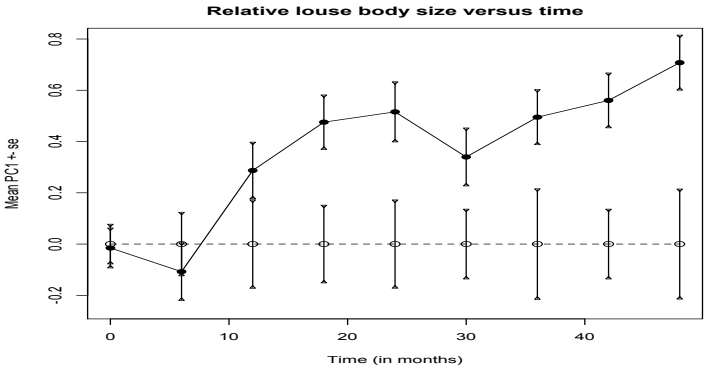

Figure 5.2: PC1 mean difference 'giant runt - feral' versus time

5.3 Critique of published claims

Villa *et al.* base their conclusions on the first principal component as a combined measure of overall louse size. Since the first principal component is essentially the standardized sum or average of the three size variables, this much is fine. The sample averages for each host are plotted in the fourth panel of Fig. 5.1, which shows that the divergence between the two mean trajectories is not appreciably greater than the temporal variability of any single trajectory. This is a disappointing conclusion for a four-year experiment, and not appealing as a headline story.

However, Villa *et al.* choose to emphasize the divergence over the variability by plotting the PC1 mean difference (giant runts minus controls) as a function of time in their Fig. 1C. A version of their plot is shown in Fig. 5.2, and is to be contrasted with the fourth panel of Fig. 5.1.

The plot symbol on the horizontal line in Fig. 1C or Fig. 5.2 is explicitly associated with controls. Error bars attached to to zero are not mentioned in captions or in text. The visual impression of remarkable temporal stability of louse size on feral pigeons contrasts starkly with the rapid increase for lineages on giant runts. The plot title and the scale on the y-axis confirm those impressions, which are in line with the authors' conclusion *Lineages of lice transferred to different sized pigeons rapidly evolved differences in size.* In my opinion, Fig. 1C or Fig. 5.2 gives a grossly misleading impression of stability for feral pigeons contrasted with a substantial trend for giant runts. In fact, Table 5.1 shows that louse body-size changes are no more than 2% over the entire period.

Taking correlations into account, the error bars for the non-zero line in Fig. 1C or Fig. 5.2 are too small by a factor increasing from about 1.0 to 15.0, and roughly proportional to time.

Tables S2–S5 in the Appendix to their paper report regression coefficients and their standard errors for the full factorial model with (5.3) as the covariance structure. These tables are cited in the *Results and Discussion* section to

support the chief claim: *Over the course of 4 y, lice on giant runts increased in size, relative to lice on feral pigeon controls (Fig. 1C and SI Appendix, Tables S2–S5).* It is unclear which coefficients are meant to justify this claim, but the coefficient of *host:time* in the PC1 analysis is reported with a t-ratio of 3.15. Overlooked in this computational blizzard is the fact that both the fitted mean and the fitted covariance contradict the randomization. In addition, the covariance assumption is non-standard for an evolutionary process, and is demonstrably inadequate for the task.

The formal analysis of the first principal component by linear Gaussian models follows the lines of section 5.2.5. Although the scale of the PC1-response is very different from that of the body length, the need for the Brownian-motion component is abundantly clear, as is the additional baseline variance. When these covariances are accommodated, the slope estimates and their standard errors are

Parameter	Estimate	s.e.
Feral:time	−0.0126	0.033
Giant:time	0.0016	0.033
Difference	0.0142	0.013

Nothing in this PC1 analysis points to a departure from neutral evolution of lice on either host. In conclusion, the evolutionary divergence described by Villa *et al.* may well exist on some time scale, but the evidence for it is not to be found in their data.

5.4 Further remarks

5.4.1 Role of louse sex

The variables *host* and *lineage* are treatment factors generated immediately post-baseline by randomization, and having a known distribution. For the 800 lineage founders, louse *sex* is a a pre-baseline variable; for the remaining lice, *sex* is a random variable not generated by randomization, and not recorded immediately post-baseline. One can speculate on the joint distribution, but in principle, the sex ratio for giant runts might not be the same as the sex ratio for controls. Thus, (5.1) and (5.2) are models for the conditional mean while (5.3) and (5.4) are models for the conditional covariance—given *host* and *lineage* plus the entire *sex*-configuration for all sampled lice.

Regardless of covariance assumptions, the interpretation in (5.2) of β_h as 'the effect of treatment' must be considered in the light of the fact that any additive effect possibly attributable to an effect of treatment on *sex* has been eliminated. Although not intermediate in the temporal sense, *sex* is not dissimilar mathematically to an intermediate response. It is possible that treatment could have an effect on the intermediate response, in which case the coefficients β_h in the conditional mean describe only one part of the treatment effect.

In the context of this experiment, no effect of treatment on *sex* is anticipated. Any effect that might be present is most likely to be a sampling artifact of little

Table 5.3. Louse counts by host, sex and time

		Time in months								
Host	Sex	0	6	12	18	24	30	36	42	48
Feral	F	231	67	39	57	38	56	23	55	19
	M	169	73	44	50	37	55	31	49	22
Giant runt	F	233	95	104	105	102	104	105	105	96
	M	167	102	95	92	98	97	91	95	104

or no evolutionary interest. Nonetheless, it is not difficult to examine the sex distribution at baseline and post-baseline for both treatment groups. Table 5.3 shows the louse counts by time, host and sex.

The post-baseline total count is quite constant at 200 for giant runts, but is much more variable for captive feral pigeons. The first is presumably a design target. We are left to wonder why the the control group does not have a similar target. Nevertheless, this is not a serious criticism. In both treatment groups, females account for 58% of lice at baseline, but close to 50% thereafter. As anticipated, there is little evidence of a difference in sex ratio between groups. If anything, the difference between the ratios is below expectation at nearly every time point.

The Poisson log-linear model *time:(host+sex)* is equivalent to the statement that *host* and *sex* are independent at each time point, or equivalently, that the sex ratio is the same for both pigeon breeds, but not necessarily 50:50. The residual deviance of 2.8 on nine degrees of freedom falls at the lower third percentile (0.03) of the null distribution, which shows that sample log odds ratios are uniformly closer to constant than the Poisson model predicts. Certainly, there is no suggestion of a treatment effect on sex ratios. Apart from the imbalance at baseline, the subsequent ratios are close to 50:50, so we can regard the sex indicator post-baseline as a Bernoulli process independent of treatment.

5.4.2 Persistence of initial patterns

One unintended consequence of the Brownian covariance model (5.4) is that baseline values are independent of all subsequent values. This is a strong assumption. It is not implied by randomization, and it is not necessarily a property that we could confidently expect to be supported by detailed examination of the data. Without contradicting the randomization, it is possible to introduce temporal correlations between baseline and non-baseline values by a simple modification such as replacing the last term in (5.4) with the shifted Brownian covariance $\min(t-\tau, t'-\tau)$ for some $\tau \leq 0$. For reasons that are explained in chapter 18, the REML criterion is independent of τ, so this particular modification has no effect on fitted values, on prediction or inference for contrasts. In fact, this covariance term could be replaced with the stationary version $-|t-t'|/2$.

The analysis of variance for baseline values already casts doubt on the fair-

ness of the randomization with respect to aviaries, so it is natural to check for correlations between initial and subsequent values associated with the same aviary. Does the pattern of louse size differences among aviaries at baseline persist in subsequent generations? The question is concerned with persistence of aviary patterns, so fairness of the randomization is not presumed.

One way to introduce persistent initial patterns is to replace the aviary term in (5.4) with independent shifted Brownian motions, one per aviary. The covariance contribution is then

$$\sigma_2^2 \delta_{a,a'} \min(t - \tau, t' - \tau),$$

with a single temporal shift $\tau \leq 0$ to be estimated from the data using the REML criterion. One boundary point $\tau = 0$ coincides with (5.4), and the other limit $\tau \to -\infty$ implies a constant aviary effect as in (5.3). For $\tau < 0$, this modification implies positive correlations within aviaries at baseline, which is a size pattern that contradicts our understanding of randomization. The interpretation is that, by accident or by design, some aviaries start out with larger lice than others, and the initial pattern leaves an imprint on the subsequent evolution.

For the PC1 variable, the profile REML log likelihood values for τ at zero, $\hat{\tau} = -2.6$ and $-\infty$ are 0.0, 11.5 and -18.5, showing that the constant aviary effect is decisively rejected by the data. It appears from this analysis that the initial aviary pattern for PC1 is non-zero and that it persists in the subsequent evolution. The particular temporal offset may be pure coincidence, but $\hat{\tau} = -2.6$ months is a very close approximation to the de-lousing quarantine period during which the pigeons had to be housed somewhere.

5.4.3 Observational units

Consider the statement near the beginning of section 5.1.3: 'since each measurement is made on one louse, it is evident that each observational unit is one louse...'. The premiss—that each measurement is made on one louse—is indisputable. Nevertheless, a conclusion that is obvious literally, is not necessarily true mathematically in the sense of the definition.

According to the definition, the observational units are the objects, or points in the domain, on which the response is defined as a stochastic process. Thus, each observational unit exists at baseline, not necessarily as a physical object, but as a non-random mathematical entity. For the models in section 5.2, with louse-time pairs as observational units, there is no birth or death, and no evolving finite population—only a fixed, arbitrarily large, set of lice in each lineage. In this mathematical framework, the lice are in 1–1 correspondence with the natural numbers, they live indefinitely in the product space, and their vital statistics are random variables recorderd in the state space. To each louse there corresponds a stochastic process, so the value for each louse evolves over time, but the population itself is fixed and arbitrarily large in every lineage.

It would be wrong to say that the Gaussian model is incorrect or that its flaws are fatal, but its shortcomings for this application are clear enough. If

the application calls for a finite randomly-evolving lineage, a more complicated mathematical structure is required. The remarkable thing is not that this Gaussian model is exquisitely tailored to this evolutionary process, but that a generic model that is missing the defining aspects of life, namely birth and death, should have anything useful to contribute at all.

Certainly, the lice do not exist in the physical sense at baseline. But lineages are established at baseline, and it is the lineages that evolve. They evolve randomly in two senses—in their composition as a finite set of lice, and in their values or features. If both aspects are important for a given application, a more complicated model is needed in which the observational units are lineage-time pairs. The state space for one measurement on one louse is $\mathcal{S} = \{M, F\} \times \mathbb{R}^3$; the state space for one observational unit is the set of finite subsets of $\mathcal{S}$. One finite subset of $\mathcal{S}$ is a complete description of the population size and the vital statistics of the residents at time t. The transitions from one finite subset to another are limited by birth, death and continuity in time.

A general process of the type described in the preceding paragraph is a complicated mathematical structure, and we make no effort to develop a general theory here. But there are simple versions that are essentially equivalent to imposing a pure birth-death process independently as a cohort restriction on the domain of a Gaussian process. The distribution of the values thus generated coincides with the Gaussian model in section 5.2, and none of the subsequent analyses are affected. For that setting, birth and death are immaterial.

The possibility that individual louse values or body-sizes might be related to the sample size or lineage size from which they come has not been considered up to this point, in part because such a dependence is not possible under the models in section 5.2. The notion that a sample can be extended indefinitely from a sub-sample such that the sub-sample values remain unchanged, is usually understood in applied work as an obvious fact. The possibility that the obvious fact might fail is "such an appalling vista that every sensible person would say 'It cannot be right..."'. But, just as it turned out a decade after Lord Denning's notorious judgement in 1980, from which this quote is taken, the appalling vista is not sufficient probative evidence to establish its alternative as fact. Failure also strikes at the heart of the most cherished notion in probability and applied statistics, which is the 'obvious fact' of distributional consistency for sub-samples as formulated by Kolmogorov (1934). If lice are the observational units for this process, consistency implies that the distribution for individuals is unrelated to the size of the sample from which they are taken. Fortunately, variability of sample sizes provides a weak check to test that implication.

Each of the 32×9 lineage-time pairs provides one sample, of which 15 are empty. The louse counts range from zero to 44, they are highly variable, and they tend to decrease over time. One lineage appears to go extinct at 30 months. The safest and the simplest way to test for a dependence on sample size is to include sample size as an additional 'covariate' in (5.2), retaining (5.4) for covariances. For both log body length and PC1, the fitted coefficient is negative and approximately one half of the standard error. This analysis offers no evidence of a sample-size dependence, which provides a little reassurance that the

earlier analysis with louse as the observational unit is reasonably sound.

If some birds preened more vigorously or more thoroughly than others, and larger or older lice were preferentially removed by preening, the more assiduous preeners would then host fewer and smaller lice. Differential preening could lead to a dependence of mean louse size on lineage size or on sample size, in which case the test in the preceding paragraph is a reasonable check.

5.5 Follow-up

5.5.1 New design information

Given the severity of the discrepancy between the conclusions presented above and those published by Villa et al. (2019), it seemed only appropriate to send a copy of sections 5.1–5.4 to the authors for comment. I contacted the lead author in early December 2020. Scott Villa, responded immediately, and later at the beginning of February 2021 offering further details about the experimental design, and challenging the conclusions on several points.

The randomization was carried out according to an elaborate protocol, which involved dislodging the CO_2-anesthetized lice over a custom-made 10×14 glass grid, generating a random grid number as the starting point for collection of specimens, and placing lice sequentially and cyclically in vials labelled 1–32 until each vial contained 25+ lice. It was designed to avoid unintentional biases, and it appeared to be adequate for the task.

The following summary of key design points that had previously been partially or totally misunderstood is taken from Villa's reply.

1. At time zero, 1600+ lice were collected from wild feral pigeons. No size measurements were made on the sub-sample of 800 founder lice that were transferred to captive birds. A second sample of 800 lice was photographed, measured, and frozen for subsequent genetic analysis.

2. The 800 founder lice were assigned to hosts at random, 25 per bird. Each founding population consisted of 13–14 females and 11–12 males with a deliberate female bias to ensure that a lineage would be established on every host.

3. The 800 lice measured at time zero did not contribute to the breeding population; their assignment to lineages was randomized, but purely virtual. The virtual sample had the same sex-ratio as the founders.

4. After baseline, the lice that were measured at 6-month intervals were frozen thereafter to use for genomic analyses of the populations over time. Throughout the experiment (months 6–48), the adult and immature lice that were removed but not photographed were immediately placed back on birds, thus ensuring stability of the lineages over time.

In light of the revised information, certain statements in the 'Materials and Methods' section of the published paper seem ambiguous or oddly phrased, for example,

> *We transferred 800 lice from wild caught feral pigeons to 16 giant runt pigeons and 16 feral pigeon controls (25 lice per bird). At this time (Time 0), we also randomly sampled 800 lice from the source population on wild caught feral pigeons and measured their body size.*

This remark suggests, correctly as it turns out, that the measured lice and the founder lice might be disjoint subsets. But that thought was dispelled by an earlier remark *Once photographed, the live lice were returned to their respective host*, which now turns out to be incorrect.

To learn about a natural host-parasite system, the scientist must manipulate the system to some extent. But as the degree of interference increases, the more is learned about the interference and the less about the system. The strong approving remark in the second paragraph of section 5.1.2 about the necessity of returning all lice to their host seems entirely correct as a matter of principle, if only to reduce interference and to minimize the possibility of lineage extinction. Regrettably, it seems now that photographed lice were not returned, perhaps because photography is damaging or destructive. Whether that degree of interference is acceptable or excessive is a matter of biological judgement best left to subject-matter experts, not a matter on which statistical expertise carries weight. As always, the over-riding concern is that the experiment be reported as it was conducted.

5.5.2 Modifications to analyses

At this point we accept the new design information, and ask what effect it has on the appropriateness of the analyses already performed, and what modifications are required.

Consider first the information that the association of time-zero measurements with lineages is virtual. This fact implies that the information content is unchanged if time-zero values are permuted in any manner that preserves sexes, while non-baseline values stay put. A baseline permutation that preserves sexes is one in which males are permuted with other males, females with other females, and non-baseline individuals are fixed. This set of permutations is a sub-group of size $464! \times 336!$ in the larger group of size $3105!$.

Any credible analysis that accommodates the virtual randomization must be invariant with respect to this group of permutations; similar remarks apply to numerical conclusions regarding temporal trends, variance components or other effects. The authors' block-factor assumption (5.3) applies to baseline and non-baseline values, so it contradicts baseline exchangeability, virtual or otherwise. The numerical values reported in their supplementary tables S1–S5 are also not invariant.

Non-virtual baseline exchangeability as discussed in section 5.2.5 implies that the marginal distribution of the initial 800 measurements is invariant with re-

spect to sex-preserving permutation. Virtual exchangeability is a much stronger condition because it implies also that the joint distribution of all 3105 measurements is invariant. Neither condition implies independence of initial and subsequent values, but virtual exchangeability implies that the dependence must be of a trivial type, which is ignorable in practice. The Brownian model (5.4) implies $\mathrm{cov}(Y_u, Y_{u'}) = 0$ for any pair $u \neq u'$ such that $t(u) = 0$ or $t(u') = 0$. Together with (5.2), it also satisfies the virtual exchangeability condition. By contrast, the standard random-effects model (5.3) having independent and identically distributed lineage effects that are constant in time, does not satisfy even the weaker exchangeability condition. It is also incompatible with the discussion in section 5.4.2.

The Brownian-motion model is in line with the standard genetic theory for trait evolution, and is compatible with virtual randomization as described above. Thus the conclusions as stated at the end of section 5.3 are confirmed. Average size differences between the two hosts shown in Table 5.1 are less than 2% and not close to statistical significance. The sex-adjusted PC1 mean differences GR−F at each non-zero time point are very similar to the unadjusted differences displayed in Fig. 5.2, but the correctly-computed standard errors tell a very different story.

Table 5.4. PC1 mean differences Giant Runt − Feral by time

	Time in months								
	0	6	12	18	24	30	36	42	48
Diff	0.000	−0.114	0.111	0.464	0.496	0.316	0.251	0.494	0.750
s.e.	0.00	0.24	0.33	0.40	0.46	0.51	0.56	0.60	0.65
Ratio	0.00	−0.48	0.34	1.16	1.09	0.62	0.45	0.82	1.16

Both the differences and the standard errors in this table are computed from a fitted Gaussian model, in which the temporal trend, previously modelled as a zero-mean random effect with covariance $\sigma_3^2(t \wedge t')$ in (5.4), is replaced with a non-random term in the mean. The moments are

$$E(Y_u) = \beta_0 + \beta_1 s(u) + \gamma_h(t), \tag{5.5}$$

$$\mathrm{cov}(Y_u, Y_{u'}) = \sigma_0^2 \delta_{u,u'} + \sigma_1^2 \delta_{l,l'}(t \wedge t') + \sigma_2^2 \delta_{a,a'}(t \wedge t') + \sigma_3^2 \delta_{uu'} I_{t=0}. \tag{5.6}$$

The mean subspace includes an additive constant for sex, and a host-dependent temporal trend $\gamma_h(t)$. The factorial model formula

```
sex + as.factor(time):host
```

generates a subspace of dimension $1 + 9 \times 2 = 19$, but the randomization constraint implies $\gamma_0(0) = \gamma_1(0)$, which reduces the dimension by one. The fitted differences $\hat{\gamma}_1(t) - \hat{\gamma}_0(t)$ are shown in the table, together with standard errors as estimated by REML and weighted least squares. They are automatically sex-adjusted, so they are not exactly the same as the sample differences shown in Fig. 5.2.

If the randomization constraint is ignored, the fitted difference is non-zero for $t = 0$. All estimates and standard errors throughout the table are altered, but only slightly.

At no time does the observed difference reach much above one standard error, so claims for rapid divergence are not supported by this analysis or by any modifications that include non-trivial temporal dependence: see Exercise 5.24. The same applies to the overall estimate of linear temporal trend, which is 0.0142 per month with standard error 0.013. Comparable analyses for body length and other size measurements point to similar conclusions.

It is possible to satisfy the randomization constraint by restricting the block factor terms in (5.3) to post-baseline times only. But Brownian-motion is a much better fit than the restricted block factor, which implies that the evidence for non-trivial temporal correlation is very strong (see Exercises 5.24 and 5.25). Any modified analysis that takes account of such correlations leads to very similar conclusions. It is unsafe to conclude that divergent evolution does not exist on some time scale, or even on a 48-month time scale, but it is safe to say that no evidence for it emerges from these data.

5.5.3 Further remarks

According to the reply by Scott Villa, the sex ratio of lice at baseline was intentionally biased towards females, with 13–14 females and 11–12 males as founders for each lineage. Following the initial seeding, male and female lice were sampled in approximately equal numbers, so information on the evolution of the sex ratio over time is not available. In light of this information, much of the speculation in section 5.4.1 is not relevant.

Villa also takes issue with a remark in section 5.2.1 that the overall change in body size is small, which was meant to sugest that changes of this magnitude ($< 2\%$) cannot be biologically significant. His counter-claim is that *body size changes on this scale are biologically relevant for this species, as the effect on mating behavior shows (Villa et al. 2019, Figs 2–5).*

The coefficient of variation of body length for female lice within aviaries is very stable at 2.4%–2.6% from six months onwards; the value for males is equally stable at 2.2%–2.4%. These numbers represent natural variability of body length within freely breeding populations, which is approximately 2.4%. The mean differences between hosts are shown in Table 5.1; they are almost uniformly less than 2%.

What are the implications for mating? The root mean square size discrepancy between a random pair from the same aviary is approximately $\sqrt{2 \times 2.4^2}$, or 3.4%, so the distribution of $F-M$-size differences is approximately $N(411, 83^2)$. A 2% increase in mean size for females implies that the distribution of size differences for mixed hosts is $N(411 + 50, 83^2)$. If size discrepancy is the chief determinant of sexual compatibility, and incompatibility is rare in each population, a mean difference of 0.6 standard deviations is not sufficient to make the incompatible fraction large in the mixed population. The movies provided by Villa et al (2019) illustrate size discrepancies of 1.8 and -2.6 standard de-

viations, so their relevance at the 0.6σ-scale is not immediately apparent. In the absence of a detailed morphological explanation, it is difficult to accept the authors' claim that body size changes on this scale ($\sim 0.6\sigma$) are biologically relevant for any species.

5.6 Exercises

5.1 According to the standard definition in section 13.2.1, two observational units u, u' belong to the same experimental unit if the treatment assignment probabilities given the baseline configuration satisfy $P(T_u = T_{u'}) = 1$. Section 5.1.3 makes the argument that each louse is one observational unit, and that each lineage is one experimental unit. But the author subsequently pivots to *aviary* as the experimental unit, hedging his bets by stating that 'both seem to be relevant'. Discuss the arguments pro and con of *louse-lineage* versus *louse-aviary* versus *lineage-aviary* as the observational-experimental units. In connection with the models in section 5.2, what are the substantive implications of one choice versus another?

5.2 According to Villa *et al.*,

> Pigeons combat feather lice by removing them with their beaks during regular bouts of preening. *Columbicola columbae*, a parasite of feral pigeons, avoids preening by hiding in spaces between adjacent feather barbs; preening selects for *C. columbae* small enough to fit between the barbs. Preening also exerts selection on traits critical for locomotion on the host.

In light of this information, comment on the remark in section 5.1.3 *... size-biased sampling need not be a serious concern for this experiment provided that it affects all birds equally.*

5.3 Download the data, compute the averages at each time point for the two pigeon breeds, and reconstruct the plots in Fig. 5.1 and Fig. 5.2.

5.4 The coefficient of variation is the standard-deviation-to-mean ratio, which is often reported as a percentage. For *body length* or other size variables, the coefficient of variation is essentially the same as the standard deviation of the log-transformed variable. Compute the coefficient of variation of *body length* separately for male and female lice on each occasion, and report this as a table of percentages. What patterns do you see in this table for males versus females or baseline versus non-baseline?

5.5 Use `anova(...)` to re-compute the mean squares in Table 5.2. Use Bartlett's statistic to test the hypothesis that the residual mean squares have the same expected value at all time points. What assumptions are needed to justify the null distribution?

5.6 For the model (5.3), what is the expected value of the within-lineage mean square at time t? For the Brownian-motion model (5.4), show that the variance of Y_u increases linearly with time. What is the expected value of the within-lineage mean square?

5.7 Use `lmer(...)` to fit the variance-components model (5.3) to the log body length with (5.2) as the mean-value subspace. Report the two slopes, the slope difference, and the three standard errors.

5.8 Explain why (5.3) is in conflict with randomization.

5.9 Compute the four covariance matrices $V_0, \ldots, V_3$ that occur in (5.4). Let Q be the ordinary least-squares projection with kernel (5.2). Compute the four quadratic forms $Y'Q'V_rQY$ and their expected values as a linear function of the four variance components. Hence or otherwise, obtain initial estimates.

5.10 Use `regress(...)` to compute the REML estimate of the variance components in (5.4). Hence obtain the estimated slopes, their difference, and the standard errors for all three.

5.11 For $n = 100$ points $t_1, \ldots, t_n$ equally spaced in the interval $(0, 48)$, compute the matrix

$$\Sigma_{ij} = \delta_{ij} + \theta(t_i \wedge t_j)$$

for small values of θ, say $0 \le \theta \le 0.02$. Find the maximum-likelihood estimate of β in the linear model $Y \sim N_n(\alpha + \beta t,\ \Sigma)$ with Σ known, and plot the variance of $\hat{\beta}$ as a function of θ. Comment on the effect of the Brownian-motion component.

5.12 Regress the 32×9 lineage-time averages (for PC1) against sample size using sample size as weights. You should find a statistically significant positive coefficient a little larger than 0.01. Explain why the conclusions from this exercise are so different from those at the end of section 5.4.2.

5.13 In Table S2 of their Appendix, Villa *et al.* fit the eight-dimensional factorial model *host:sex:time* to the first principal component values on 3096 lice. Show that this is equivalent to fitting four separate linear regressions $E(Y_u) = \alpha + \beta t_u$, with one intercept and one slope for each of the disjoint subgroups, Fer.F, Fer.M, Gr.F, Gr.M. Feral and female are the reference levels, so $sex_u = 1$ is the indicator vector for males. Deduce that the *host:time* coefficient is equal to the slope difference $\beta_{Gr.F} - \beta_{Fer.F}$ restricted to female lice. The fitted value is 0.009. What is the fitted slope difference for male lice?

5.14 The *sex* coefficient in Table S2 is -2.437. Which combination of the four α-values in the previous exercise does this correspond to?

5.15 The *host* coefficient in Table S2 is 0.449 with standard error 0.159. What does this imply about the average or expected baseline values for the four subgroups?

5.16 For the model with persistent aviary patterns described at the end of section 5.4.2, compute and plot the REML profile log likelihood for τ in the range $0.5 \le \tau \le 24$. Use PC1 as the response, and (5.2) for the mean-value subspace. The covariance should be a linear combination of five matrices, one each for the identity matrix and the identity restricted to baseline, two Brownian-motion product matrices as in (5.4), and one τ-shifted B-M product matrix. Ten to twelve points equally spaced on the log scale should suffice for plotting.

5.17 Use the profile log likelihood plot in the previous exercise to obtain a nominal 95% confidence interval for τ.

5.18 Distributional invariance. Consider a simplified version of the louse model in which there are 16 feral and 16 giant runt pigeons, no sex differences between lice, and no correlations among measurements. Two lice are associated with each bird at baseline, and two at each subsequent time $t = 1, \ldots, 7$ for a total of 512 observations. Each louse u is associated with a host type $h(u)$, feral or giant runt, and the joint distribution is Gaussian with moments

$$E(Y_u) = \beta_0 + \beta_{h(u)} t_u; \qquad \mathrm{cov}(Y_u, Y_{u'}) = \sigma^2 \delta_{u,u'}.$$

A baseline permutation is a 1–1 mapping $u \mapsto \tau(u)$ such that $t(u) > 0$ implies $\tau(u) = u$. Distributional invariance means that the permuted vector Y^τ with components $Y^\tau_u = Y_{\tau(u)}$ has the same distribution as Y. Show that the joint distribution is invariant with respect to baseline permutations. Note that $h(\tau(u))$ is not necessarily equal to $h(u)$.

5.19 Procedural invariance. Consider a sample of 512 observations generated according to the model in the previous exercise. The estimation procedure is invariant if $\hat\beta(Y) = \hat\beta(Y^\tau)$ and $\hat\sigma(Y) = \hat\sigma(Y^\tau)$ for every baseline permutation. Is it necessarily the case that distributional invariance implies procedural invariance? Explain why least-squares and maximum-likelihood are invariant procedures.

5.20 Consider the following statement taken from section 5.5. *Any credible analysis that accommodates the virtual randomization must be invariant with respect to the same group, and similar remarks apply to numerical conclusions regarding temporal trends, variance components or other effects.* Invariance in this setting means that each distribution in the model is exchangeable, or invariant with respect to sex-preserving baseline permutations. This is a demanding standard, and it is possible that subsequent statements in that same section may not live up to it. Show that the model-formula `Host:as.factor(Time)`, which is related to Table 5.4, corresponds to a set of vectors, some of which are not group-invariant. Investigate the implications, particularly for time zero.

5.21 According to the text in section 5.5, *Virtual randomization requires the time-zero average for feral hosts to be the same as that for giant runts, but the temporal trends are otherwise unconstrained.* It appears that the model matrix spanning this subspace is not constructible using factorial model formulae. Explain how to construct the desired matrix including a constant additive sex

effect. What is its rank? Fit the model as described in the text following Table 5.4. Include independent Brownian motions for aviaries and lineages, plus an additional baseline error term with independent and identically distributed components.

5.22 Use the fitted model from the previous exercise to compute the linear trend coefficient

$$\frac{\sum t(\hat{\gamma}_1(t) - \hat{\gamma}_0(t))}{\sum t^2}$$

and its standard error. You should find both numbers in the range 0.013–0.015 per month, similar to, but not exactly the same as those reported in the text.

5.23 The model in the previous two exercises has a baseline variance that is larger than the non-baseline residual variance. What is the ratio of fitted variances?

5.24 The fact that measured lice were not returned to their hosts is an interference in the system that may reduce or eliminate temporal correlations that would otherwise be expected. One mathematically viable covariance model that is in line with virtual randomization, replaces each occurrence of $t \wedge t'$ in (5.6) with the rank-one Boolean product matrix $(t > 0)(t' > 0)$, so that the only non-zero temporal correlations are those associated with lineage and aviary as strictly post-baseline block factors. Fit this modified block-factor model to the PC1 response with (5.5) for the mean subspace. Which model fits better? Is the log likelihood difference small or large? An informal comparison suffices at this point.

5.25 Construct two versions of Table 5.4, one based on the modified block-factor model, and one based on the combined variance model that includes both. Comment on any major discrepancy or difference in conclusions based on the various models.

5.26 What was the matter that Lord Denning refused to accept in his 1980 appeals-court judgement when he referred so melodramatically to the 'appalling vista that every sensible person would reject'? And why was this phenomenon so abhorrent to him?