Nonparallel Regressions with Indicator Variables

Timothy G. Gregoire[1]

[1] School of Forestry & Environmental Studies, Yale University, New Haven, CT, USA

Correspondence: Timothy G. Gregoire, School of Forestry & Environmental Studies, Yale University, 360 Prospect St., New Haven, CT 06511, USA. Tel: 1-203-432-9398. E-mail: timothy.gregoire@yale.edu

Abstract

A multiple linear regression model which includes 0/1 variables to indicate membership in a group is a convenient way to model parallel regression surfaces. Building upon this, an extended model which includes predictor variables that are the product of the indicator and other variables will coincide with separate regressions for each group using only the other predictors. The algebraic basis for this concurrence is demonstrated. Least squares estimation is presumed. Examples with two groups and with four groups are presented.

Keywords: concurrent regressions, dummy variables, omitted group, partitioned matrices

1. Introduction

Most courses on linear regression include a discussion of using indicator, or dummy, variables to allow for parallel regression surfaces for different levels of a factor variable. Typical examples may be found in section 4.2.3 of Fox and Weisberg [2011, pp. 157-169]; section 2.3.1 of Harrell [2001, pp. 14-16]; Chapter 11 of Neter et al. [1996]. As a special case, the linear regression model with a single continuous predictor variable combined with a single binary indicator variable portrays a model comprising parallel lines. Weisberg [2005, § 6.2.2] and Christensen et al. [2011, § 7.4] both present a more general treatment within the linear model family.

When a third variable consisting of the product of the indicator with the continuous predictor is included, nonparallel regression surfaces are portrayed. It is not widely appreciated that in this case, the least squares coefficient estimates are identical to those obtained by fitting each of the two groups separately. This note explicates that result.

2. Method

Consider the simple linear regression (SLR) model

$$Y = X_s \beta_s + \epsilon \tag{1}$$

where both Y and ϵ are n-element vectors; the errors are iid with mean zero; $X_s = (X_0, X_1)$; and $\beta_s = (\beta_{s0}, \beta_{s1})'$. Here, X_s comprises the unit vector X_0 and an auxiliary variate vector X_1, evidently both of length n.

For this model the least squares estimator of β_s is

$$\widehat{\beta_s} = (X_s'X_s)^{-1}X_s'Y \tag{2}$$

Suppose that the n observations are split between two groups, with n_1 from Group I while the remaining $n_2 = n - n_1$ hail from Group II. Further suppose that we set up an indicator variable to distinguish group membership:

$$I_i = \begin{cases} 1, & \text{if the } i\text{th observation is a member of Group I;} \\ 0, & \text{else.} \end{cases}$$

Further suppose the data are arrayed such that $X_2 = (I_1, I_2, \ldots, I_n)'$ and $X_3 = (I_1 x_{11}, I_2 x_{12}, \ldots, I_n x_{1n})'$.

Consider the multiple linear regression (MLR) model

$$Y = X_m \beta_m + \epsilon \tag{3}$$
$$X_m = (X_0, X_1, X_2, X_3)$$
$$\beta_m = (\beta_{m0}, \beta_{m1}, \beta_{m2}, \beta_{m3})'.$$

The least squares estimator of β_m is

$$\widehat{\beta_m} = (X'_m X_m)^{-1} X'_m Y \tag{4}$$

It is commonly known that a regression with only X_0, X_1 and X_2 will portray parallel lines. The vertical distance between the fitted lines is given by $\widehat{\beta_{m2}}$.

For the MLR model given by (3), it is less well known that the least squares estimates of the coefficients are identical to those that obtain when SLR models for each group are fitted separately. The analytical basis for this assertion is demonstrated in the following section.

3. Nonparallel Regressions

Without loss of generality, assume that the data are arranged so that the first n_1 rows are the Group I observations. Therefore

$$X_2 = (1'_{n_1}, 0'_{n_2})' \tag{5}$$

where 1_{n_1} are the first n_1 elements of X_0, and 0_{n_2} is an n_2-element null vector. Likewise, let

$$X_3 = (x_{11}, x_{12}, \ldots, x_{1n_1}, 0'_{n_2})' \tag{6}$$

Evidently, X_s is a submatrix of X_m. Less evident is the relationship between $X'_s X_s$ and $X'_m X_m$:

$$X'_m X_m = \begin{pmatrix} X'_s X_s & X'_s X_I \\ X'_I X_s & X'_I X_I \end{pmatrix}$$

$$= \begin{pmatrix} X'_s X_s & X'_I X_I \\ X'_I X_I & X'_I X_I \end{pmatrix} \tag{7}$$

where

$$X_I = (X_2, X_3)$$

$$= \begin{pmatrix} 1'_{n_1} & , 0'_{n_2} \\ (x_{11}, \ldots, x_{1n_1})' & , 0'_{n_2} \end{pmatrix}'$$

Define $X_{II} = X_s - X_I$. From results on the inverse of a partitioned matrix, e.g. [Searle, 1966, p. 210], we know that when

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

then

$$M^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} - [-D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1}] \end{pmatrix}.$$

In the present context, the partitioned matrix of interest is $X'_m X_m$ as presented in (7) with

$$A = X'_s X_s$$
$$B = C = D = X'_I X_I.$$

By substitution, we get

$$(A - BD^{-1}C)^{-1} = (A - B)^{-1}$$

so that

$$(X_m' X_m)^{-1} = \begin{pmatrix} (A - B)^{-1} & -(A - B)^{-1} \\ -(A - B)^{-1} & B^{-1} + (A - B)^{-1} \end{pmatrix},$$

where

$$(A - B)^{-1} = (X_s' X_s - X_I' X_I)^{-1}$$
$$= (X_{II}' X_{II})^{-1}$$

and

$$B^{-1} + (A - B)^{-1} = (X_I' X_I)^{-1} + (X_{II}' X_{II})^{-1}.$$

Therefore,

$$(X_m' X_m)^{-1} = \begin{pmatrix} (X_{II}' X_{II})^{-1} & -(X_{II}' X_{II})^{-1} \\ -(X_{II}' X_{II})^{-1} & (X_I' X_I)^{-1} + (X_{II}' X_{II})^{-1} \end{pmatrix}.$$

Also,

$$X_m' Y = (X_0, X_1, X_2, X_3)' Y$$

$$= \begin{pmatrix} X_s' Y \\ X_I' Y \end{pmatrix}.$$

Therefore,

$$\widehat{\beta_m} = (X_m' X_m)^{-1} X_m' Y$$

$$= \begin{pmatrix} (X_{II}' X_{II})^{-1} (X_s' - X_I') Y \\ -(X_{II}' X_{II})^{-1} X_s' Y + [(X_I' X_I)^{-1} + (X_{II}' X_{II})^{-1}] X_I' Y \end{pmatrix}$$

$$= \begin{pmatrix} (X_{II}' X_{II})^{-1} X_{II}' Y \\ -(X_{II}' X_{II})^{-1} X_{II}' Y + (X_I' X_I)^{-1} X_I' Y \end{pmatrix}$$

$$= \begin{pmatrix} \widehat{\beta_{II}} \\ \widehat{\beta_I} - \widehat{\beta_{II}} \end{pmatrix} \tag{8}$$

where $\widehat{\beta_I}$ and $\widehat{\beta_{II}}$ are the coefficient estimators for the SLR fitted separately to the data from Group I and Group II, respectively.

In other words, $\widehat{\beta_{m0}}$ and $\widehat{\beta_{m1}}$ in (4) coincide with the intercept and slope estimates of β_s that one would obtain from the Group II data alone, whereas $\widehat{\beta_{m2}}$ and $\widehat{\beta_{m3}}$ are the difference between the estimates derived from the Group I data alone and those obtained from using Group II data alone.

The fitted values corresponding to an observation in Group I, say observation $i = 1$, is

$$\widehat{Y_1} = \widehat{\beta_{m0}} + \widehat{\beta_{m1}} x_{11} + \widehat{\beta_{m2}} x_{21} + \widehat{\beta_{m3}} x_{31}$$
$$= \widehat{\beta_{m0}} + \widehat{\beta_{m1}} x_{11} + \widehat{\beta_{m2}} + \widehat{\beta_{m3}} x_{11} \text{ (because } x_{31} \equiv x_{11})$$
$$= \widehat{\beta_{II0}} + \widehat{\beta_{II1}} x_{11} + (\widehat{\beta_{I0}} - \widehat{\beta_{II0}}) + (\widehat{\beta_{I1}} - \widehat{\beta_{II1}}) x_{11}$$
$$= \widehat{\beta_{I0}} + \widehat{\beta_{I1}} x_{11}.$$

Likewise, the fitted value corresponding to an observation in Group II, say observation $i = n$, is

$$
\begin{aligned}
\widehat{Y_1} &= \widehat{\beta}_{m0} + \widehat{\beta}_{m1} x_{1n} + \widehat{\beta}_{m2} x_{2n} + \widehat{\beta}_{m3} x_{3n} \\
&= \widehat{\beta}_{m0} + \widehat{\beta}_{m1} x_{11} \ (\text{ because } x_{2n} = x_{3n} = 0) \\
&= \widehat{\beta}_{II0} + \widehat{\beta}_{II1} x_{1n}.
\end{aligned}
$$

In the next section we exemplify these results using data presented by Robinson [2000] on $k = 2$ groups of rats subjected to a diet treatment. Identical results hold when there are more than two groups, as shown by our second example where group membership in one of $k = 4$ species groups is indicated.

4. Example 1

We present results from fitting separate, parallel, and nonparallel regressions to weight data on two groups of rats, one of which was subjected to a dietary treatment. The treated group included $n_1 = 14$ rats, while the control group included $n_2 = 12$ rats. The response variable of interest, Y, is the final weight of the rats, with implicit interest in whether the diet treatment successfully lowered final weight. Data were initially presented by Robinson [2000].

All models were fitted with packages and functions available in the R computing environment [R Core Team, 2012], particularly the gls function of the nlme package [Pinheiro et al., 2009].

Figure 1a displays a joint fit of these data using a simple linear regression with initial weight as the lone predictor. Even a cursory examination of this display reveals the inadequacy of the fitted model to capture the effectiveness of the treatment to lower final weight: an examination of the scatterplot itself strongly suggests that the treatment successfully lowered final weight on average. The joint SLR has failed to capture the noticeable effect of treatment, as might be expected *a priori*.

Figure 1b displays the separate SLR fits of these data. Clearly the effect of treatment is more adequately depicted. Yet it comes at the expense of reduced samples size, n_1 or n_2 for each group, rather than n for both together.

When the effect of initial weight on final weight is expected to be invariant to treatment, then a model for parallel regressions may be preferred. In the present context, such a model is given by

$$
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \tag{9}
$$

where Y is final weight, X_1 is initial weight, and X_2 is the vector of indicator values, with $X_{2i} = 1$ indicating that the ith rat is from the treated group (Group I). The fitted regressions are shown in Figure 1c. An advantage to fitting parallel regressions is that the full size of the sample is preserved for the fitting of the model. Another is that it allows for a common estimate of residual error for both groups, which may not only be apt but an important consideration for modeling.

Lastly we fitted model (3) to allow for differing intercepts and slopes yet avoiding the need to subset the data into separate groups for purposes of fitting the model. The graphical display of the fitted model is identical to that shown in Figure 1b.

Parameter estimates and the AIC summary measure of the adequacy of the fit of these competing models are displayed in Table 1. (The default REML method of estimation in the gls() function of the nlme package in R was used throughout). As demonstrated in the previous section, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for the jointly fitted, nonparallel model are identical to $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for the control group (Group II) fitted separately. Moreover, $\widehat{\beta}_2 = -57.2$ coincides with $\widehat{\beta}_0 = 14.9$ for the treated group fitted separately less $\widehat{\beta}_0 = 72.1$ for the control group fitted separately. Likewise, $\widehat{\beta}_3 = 0.173$ coincides with $\widehat{\beta}_1 = 0.700$ for the treated group fitted separately less $\widehat{\beta}_1 = 0.527$ for the control group fitted separately.

I often refer to $\widehat{\beta}_2$ as an "offset" to the intercept for the omitted group's regression, i.e., the "omitted group" (Christensen et al. [2011] uses the term reference category) denotes that group which was not assigned an indicator unit value. Similarly, $\widehat{\beta}_3$ can be regarded as an offset to the slope of the omitted group's regression. In the context of the present example, the offsets are a measure of treatment effect.

Despite the appeal of a common estimate of residual error for both groups, the data may suggest that it is not apt. The magnitude of difference in $\widehat{\sigma}$ between control and treated groups belies the presumption of a common error variance. Therefore we refitted both the parallel and non-parallel models to permit an estimate of the error
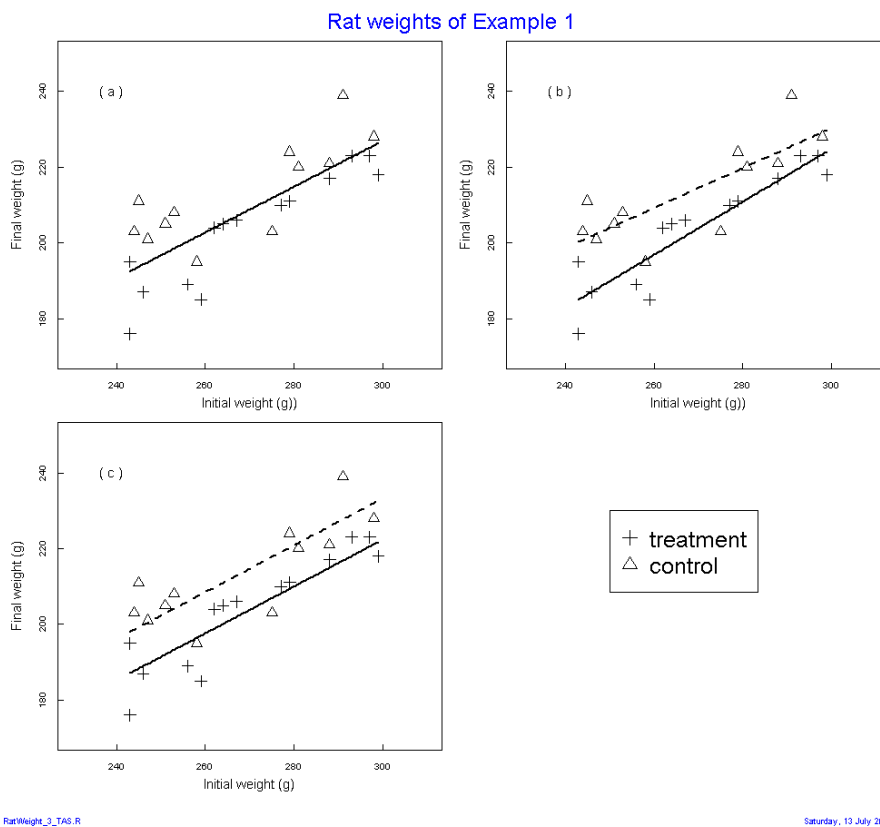
Figure 1. Depiction of the effect of diet treatment on rats; data from Robinson [2000]. a) joint SLR on all $n = 26$ observations; b) separate SLRs for treatment group (solid) and control group (dashed); c) parallel regressions for the two groups.

Table 1. Summary measures of model adequacy and parameter estimates for models fitted to the rat diet treatment data

Model	AIC	$\widehat{\sigma}$	$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$
Joint	193	9.1	45.5	0.605		
Treated (Group I) only	95	6.2	14.9	0.700		
Control (Group II) only	88	8.3	72.1	0.527		
Parallel[1]	179	7.3	47.4	0.620	-10.9	
Parallel[2]	180		40.8	0.644	-10.9	
Non-parallel[1]	182	7.2	72.1	0.527	-57.2	0.173
Non-parallel[2]	183		72.1	0.527	-57.2	0.173

[1] Common error variance for both groups.

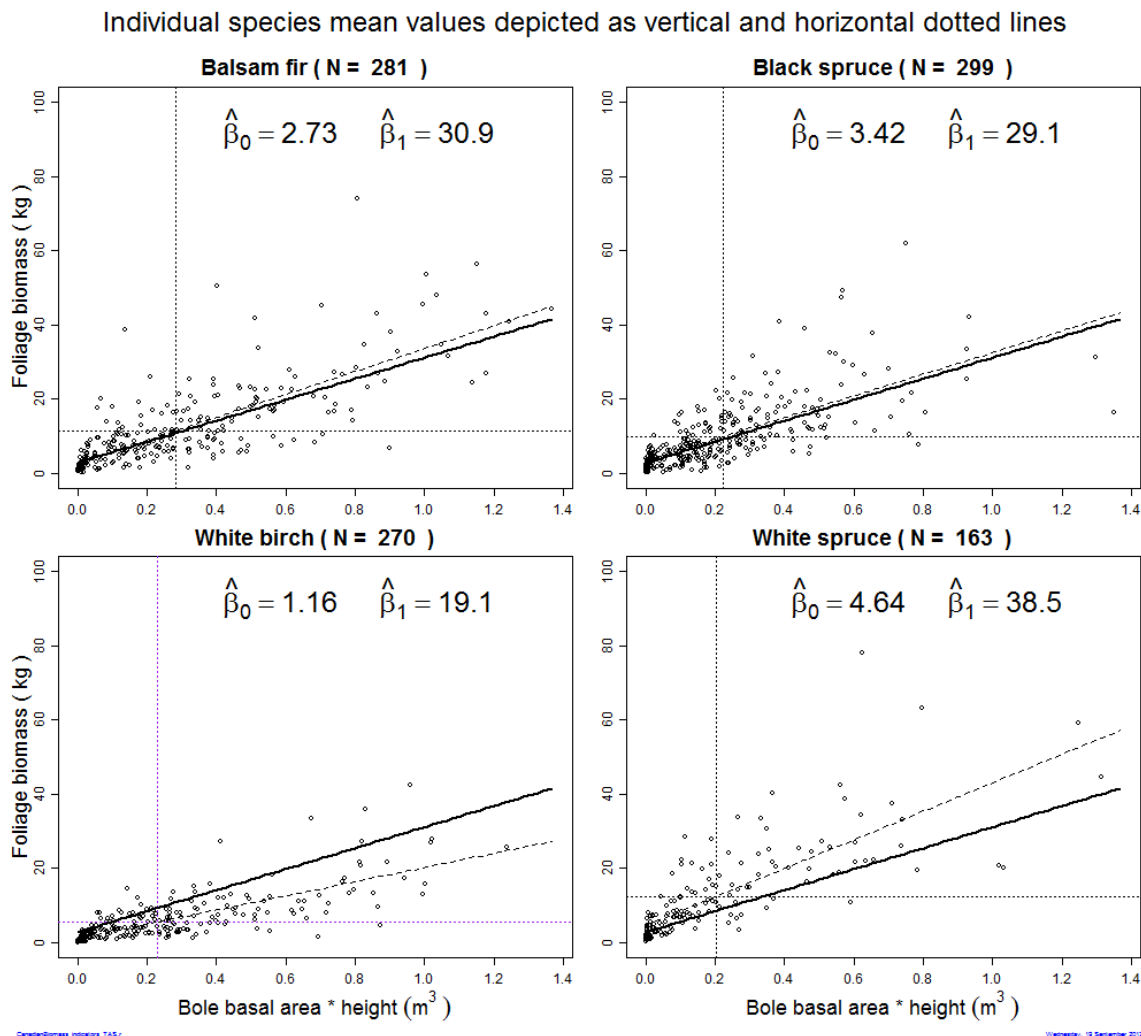[2] Separate error variance for each group.

Figure 2. Simple linear regression (SLR) of foliar biomass on the predictor (basal area × height) of the tree bole. In all four panels, the solid line shows the joint SLR on all $n = 1013$ observations of foliar biomass; in each separate panel, the superimposed dashed line shows the individual SLR for the species. The estimated coefficients for the individual SLRs are superimposed, also

variance separately for each group. [1] As may be surmised from the results displayed in Table 1, the latter model provides results, not merely coefficient estimates, that are identical in all respects to fitting models separately to both groups' data.

5. Example 5

The data for this example comprise 1013 measurements of foliar biomass among three coniferous and one deciduous species of forest trees. Foliar biomass of trees is modeled well by a simple linear regression using the cylindrical volume (the product of cross-sectional, aka basal, area and the height) of the tree's bole as the predictor, X_1. Figure 2 displays these data separately by species. The joint SLR fitted to all 1013 observations appears as the solid line in each panel, whereas the individual fit for each species alone appears as the superimposed dashed line. Summary statistics of these five fitted models appear in the uppermost five rows in Table 2.

We set up X_2, X_3, and X_4 as indicator variables for the three coniferous species (balsam fir, black spruce, and white spruce), respectively, and fitted parallel SLRs jointly, with the results as shown in Figure 3, and tabulated in Table 2. Judging by the lower AIC value for the parallel regressions model, it is superior to the joint SLR model fitted to

---

[1]In R we used the gls() function with weights as specified by the varIdent() variance function.
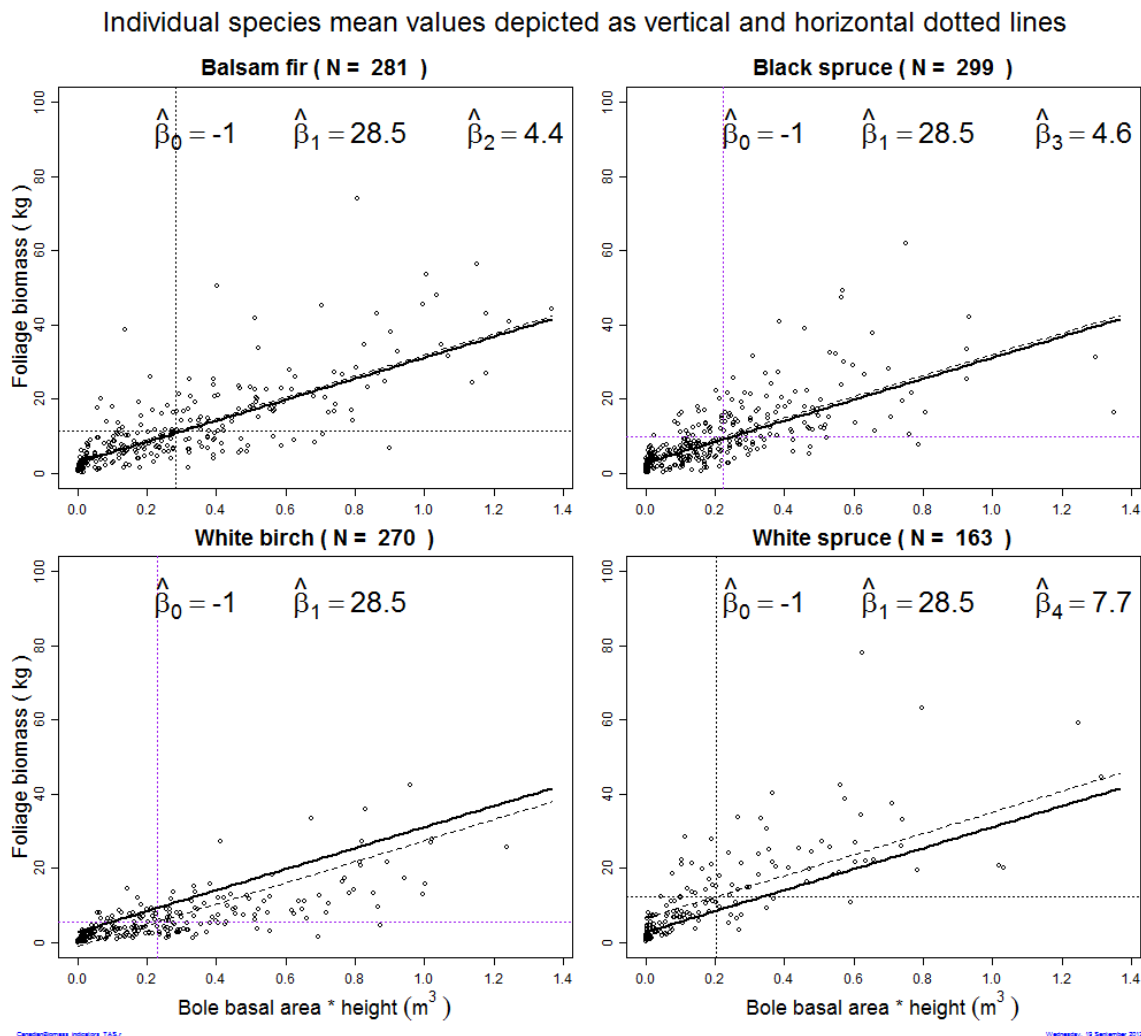
Figure 3. In each panel, the superimposed dashed line shows the individual but parallel SLR for the species. The estimated coefficients for the parallel SLRs are superimposed, also

all four species. Yet the much smaller error variance for the white birch model argued against the credibility of a common error variance in the model for parallel regression. When refitting the parallel regressions model to allow for distinct error variances for each species, the AIC showed a marked improvement — 6610 versus 6702.

By setting up X_5, X_6, and X_7 as the product of X_1 and each of the three indicator variables, we fitted a non-parallel regressions model, both with common and separate error variances. The coefficient estimates for both were identical. These results appear in the lowest two lines of Table 2. Because white birch constituted the omitted group in the indicator variable coding scheme, $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are identical to the coefficient estimates for white birch fitted alone. In contrast, $\widehat{\beta_0} + \widehat{\beta_2}$ yields a value of 2.73, which is the intercept estimate for the balsam fir regression fitted separately. Likewise, $\widehat{\beta_1} + \widehat{\beta_5}$ evaluates to 30.9, which coincides with the estimated slope of the balsam fir regression. In a similar fashion the coefficient estimates for white and black spruce may be recovered from the estimates for the non-parallel regressions model.

In Figure 2 and 3, there is a strong suggestion of increasing error variance with increasing tree size. For all the models summarized in Table 2 we re-fitted with analogous models that specified that $V[\epsilon] = \sigma^2 D^{2\delta}$, where D is the measured diameter of the tree bole. REML estimates of both σ and δ were obtained[2], which are shown in Table 3. In this specification, δ is the parameter that models the heteroscedasticity, whereas σ has units of measure

---

[2]We used the gls() function with weights as specified by the varPower() variance function to estimate δ.

Table 2. Summary measures of model adequacy and parameter estimates for models fitted to the foliar biomass data

Model	AIC	$\widehat{\sigma}$	$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_4}$	$\widehat{\beta_5}$	$\widehat{\beta_6}$	$\widehat{\beta_7}$
Joint	6839	7.1	2.84	28.3						
White birch	1510	3.9	1.16	19.1						
Balsam fir (I)	1903	7.1	2.73	30.9						
Black spruce (II)	1970	6.5	3.42	29.1						
White spruce (III)	1138	8.0	4.64	38.5						
Parallel[1]	6702	6.6	-1.00	28.5	4.40	4.56	7.65			
Parallel[2]	6610		-0.18	24.9	4.59	4.53	7.55			
Non-parallel[1]	6634	6.4	1.16	19.1	1.57	2.26	3.48	11.8	10.0	19.4
Non-parallel[2]	6521		1.16	19.1	1.57	2.26	3.48	11.8	10.0	19.4

[1] Common error variance for all species

[2] Separate error variance for each species

Table 3. Summary measures of model adequacy and parameter estimates for heterogeneous error variance models fitted to foliar biomass data.

Model	AIC	$\widehat{\delta}$	$\widehat{\sigma}$	$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_4}$	$\widehat{\beta_5}$	$\widehat{\beta_6}$	$\widehat{\beta_7}$
Joint	5988	1.22	0.22	1.37	32.1						
White birch	1190	1.53	0.05	0.49	21.6						
Balsam fir (I)	1685	1.42	0.13	1.12	35.0						
Black spruce(II)	1764	1.16	0.25	2.08	32.8						
White spruce (III)	987	1.20	0.28	1.97	47.3						
Parallel[1]	5876	1.29	0.18	0.35	32.3	0.92	1.68	1.76			
Parallel[2]	5717			0.38	30.2	0.86	1.80	1.87			
Non-parallel[1]	5781	1.26	0.19	0.58	20.6	0.66	1.45	1.32	13.6	12.4	27.1
Non-parallel[2]	5626			0.49	21.6	0.63	1.58	1.47	13.4	11.3	25.7

[1] Common δ and σ parameters for all species

[2] Separate δ and σ parameters for each species

kg/m^3 and acts as a scaling factor for the error variance. As is partly evident from the last line of results displayed in Table 3, not only the coefficient estimates but all other results (e.g. AIC, δ, σ) from fitting the non-parallel regression again coincide with fitting the models separately.

6. Discussion

The equivalence of the non-parallel regression models which incorporate indicator variables to the set of separate regression models for each group is evident in the results displayed long ago in Gujarati [1970]. Inasmuch as the emphasis in that article was on testing sets of coefficients, the equivalence went unremarked and unnoticed. The present note was prompted by a realization that this equivalence is not generally appreciated by instructors or their students.

The equivalence is retained even when modeling variance heterogeneity with a parametric function, as we did with the biomass data. This is an issue that is quite apart from the matter of whether the error variance is parameterized commonly for all groups, or separately for each, or a subset of each, group. Indeed, for the equivalence to hold, both variance parameters must be modeled separately by group.

References

Christensen, R., Johnson, W., Branscum, A., & Hanson, T. E. (2011). Bayesian Ideas and Data Analysis: an Introduction for Scientists and Statisticians. CRC Press, Boca Raton, 2 edition.

Fox, J., & Weisberg, S. (2011). An R Companion to Applied Regression. Sage, Los Angeles, 2 edition.

Gujarati, D. (1970). Use of dummy variables in testing for equality between sets of coeffcients in linear regressions: a generalization. *The American Statistician, 24,* 50C52. http://dx.doi.org/10.2307/2682446

Jr. Harrell, F. E. (2001). Regression Modeling Strategies: with Applications of Linear Models, Logistic Regression, and Survival Aalysis. Springer, New York.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied Linear Statistical Models. McGraw-Hill, Boston, 4 edition.

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2009). nlme: Linear and nonlinear mixed effects models. R package version 3:1-96. Technical report.

R Core Team. (2012). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org. ISBN 3-900051-07-0.

Robinson, G. K. (2000). *Practical Strategies for Experimenting*. Wiley, Chichester.

Searle, S. R. (1966). *Matrix Algebra for the Biological Sciences*. Wiley, New York.

Weisberg, S. (2005). *Applied Linear Regression*. Wiley, 3 edition.