BIOMETRICS INFORMATION
(You're 95% likely to need this information)

PAMPHLET NO. # 56 DATE: April 10, 1997

SUBJECT: The Use of Indicator Variables in Non-linear Regression

Indicator variables[1] are one of the chief concepts behind statistical techniques such as simple linear regression and analysis of variance (ANOVA). However, their implementation often goes unnoticed because most statistical packages create them automatically "behind the scenes". Indicator variables are used to identify the different levels of class variables such as treatments or blocks. This allows the proposed statistical model to recognize the different groups. This pamphlet will discuss how to use indicator variables in more elaborate models such as nonlinear curve fitting.

To begin with, let us consider an example of simple linear regression where a straight line is fitted through observed y_i's and x_i's. The statistical model is

$$y_i = \alpha_0 + \beta_0 x_i + \varepsilon_i$$

where, α_0 is an unknown parameter[2] representing the intercept,
β_0 is an unknown parameter representing the slope, and
as usual, $i = 1, 2, \ldots, n$ and the errors $\varepsilon_i \sim N(0, \sigma^2)$.

A measure of model fit (or lack-of-fit, depending on your politics) is provided by the residual sum of squares (also called sum of squared deviations or sum of squared errors), denoted here as *SSE*. It is calculated by adding the squared difference between the observed and fitted values obtained from the model over all observations. For this example, it is simply

$$SSE = \sum_{i=1}^{n} \left[y_i - \left(\hat{\alpha}_0 + \hat{\beta}_0 x_i \right) \right]^2 .$$

Now suppose that a more complicated model is proposed where a treatment is expected to change the slope when compared to a control. This results in a model with one intercept α_0 and two different slope parameters β_0 and β_1. Such a model requires adding a single[3] indicator variable d_1 to the data, whose purpose is to distinguish data from the treatment from those in the control group. For example, d_1 could be assigned a value of one for data from the treatment and 0 for data from the control. The statistical model would now look something like

$$y_i = \alpha_0 + \beta_0 x_i + \beta_1 d_{1i} x_i + \varepsilon_i \ .$$

After fitting such a model, the sum of squared errors *SSE* could again be calculated to measure model fit:

[1] These are sometimes referred to as dummy variables, and are usually confined to a few integer values such as 1, 0, or -1. For further discussion, the reader is referred to chapter 6 of Bergerud (1996), chapter 10 of Neter et.al. (1990) or pp.33-38 of Lesperance (1995).

[2] Parameters are the unknown quantities that make up the model, such as α_0, β_0, etc. and are estimated using observed data.

[3] If a classification factor has b levels, then b-1 indicator variables are needed.

$$SSE = \sum_{i=1}^{n} \left[y_i - \left(\hat{\alpha}_0 + \hat{\beta}_0 x_i + \hat{\beta}_1 d_{1i} x_i \right) \right]^2 .$$

If this model was more appropriate than the simple regression, its *SSE* would be significantly smaller (i. e. the observed data would deviate less from the fitted model). Testing whether the change in slope between the two groups, which is represented by β_1, is statistically different from zero is equivalent to testing if the fit of the more complicated model has improved enough to justify the extra parameter β_1.

Methods such as this can be applied to almost any model, including non-linear regression and even generalized linear models. For models that assume independent and identically distributed normal errors, an exact formal procedure is always available. It is called the *extra sums of squares principle* (Draper and Smith 1981, Wetherhill 1981, Bergerud 1991). The idea is that the improved model fit caused by fitting a complicated model (after trying a simpler model) can be measured by the reduction in *SSE* (scaled by the additional number of parameters needed for the complicated model). This scaled value, when compared to the best measure of background variation (usually the *MSE* from the more complicated model), has an *F* distribution. In other words,

$$\frac{\dfrac{SSE_s - SSE_c}{\nu_c - \nu_s}}{\dfrac{SSE_c}{n - \nu_c}} \sim F_{\nu_c - \nu_s, n - \nu_c}$$

where, SSE_s is the residual sum of squares for the simpler model,
SSE_c is the residual sum of squares for the more complicated model (the one with more parameters),
ν_s and ν_c are the number of unknown parameters in each of the two models, and
n is the number of observations.

The following example will discuss how the above method can be used for non-linear regression. Consider an experiment where data has been collected on hardwood seedlings to measure the effect of light on radial growth. For each seedling, a section of the stem is removed and the width of the most recent growth ring is measured along with the complete stem radius and a measure of light availability. The experimenter believes that a Michaelis Menton[4] curve is appropriate to model this particular data. Figure 1 shows the general shape of this curve. Notice that it is characterized by a steep increase in growth as light level increases, which quickly levels off. The form of this model is

$$r_i = \frac{\alpha_0 l_i}{\dfrac{\alpha_0}{\beta_0} + l_i} + \varepsilon_i$$

where, r_i is radial growth rate (width of most recent growth ring per stem radius)
l_i is light level.
α_0 is an unknown parameter representing the asymptotic (maximum) relative growth.
β_0 is an unknown parameter representing the slope at zero light.
As usual, $i = 1,2,\ldots,n$ and the errors $\varepsilon_i \sim N(0, \sigma^2)$.

[4] The motivation for this model is given in Pacala et. al. (1994)

Seedlings were collected from three different biogeoclimatic subzones and the researcher is most interested in knowing whether the different subzones affect the asymptotic radial growth rate α of the seedlings. To answer this question, the first step will be to create two indicator variables to discriminate between the three different subzones. For example, the first indicator variable d_1 could be 1 for all data in the first subzone and 0 otherwise, while the second indicator variable d_2 could be 1 for all data in the second subzone but 0 otherwise. Using this parameterization the new, more complicated model is

$$r_i = \frac{\alpha_0 l_i + \alpha_1 d_{1i} l_i + \alpha_2 d_{2i} l_i}{\frac{\alpha_0 + \alpha_1 d_{1i} + \alpha_2 d_{2i}}{\beta_0} + l_i} + \varepsilon_i$$

Let us examine the above equation more carefully. Notice that this is actually three separate curves, with different asymptotic relative growths and a common β_0 . Data from the first subzone are fitted by

$$r_i = \frac{\alpha_0 l_i + \alpha_1 (1) l_i + \alpha_2 (0) l_i}{\frac{\alpha_0 + \alpha_1 (1) + \alpha_2 (0)}{\beta_0} + l_i} + \varepsilon_i = \frac{(\alpha_0 + \alpha_1) l_i}{\frac{(\alpha_0 + \alpha_1)}{\beta_0} + l_i} + \varepsilon_i \ .$$

Data from the second subzone are fitted by

$$r_i = \frac{\alpha_0 l_i + \alpha_1 (0) l_i + \alpha_2 (1) l_i}{\frac{\alpha_0 + \alpha_1 (0) + \alpha_2 (1)}{\beta_0} + l_i} + \varepsilon_i = \frac{(\alpha_0 + \alpha_2) l_i}{\frac{(\alpha_0 + \alpha_2)}{\beta_0} + l_i} + \varepsilon_i \ ,$$

and data from the third subzone are fitted by

$$r_i = \frac{\alpha_0 l_i + \alpha_1 (0) l_i + \alpha_2 (0) l_i}{\frac{\alpha_0 + \alpha_1 (0) + \alpha_2 (0)}{\beta_0} + l_i} + \varepsilon_i = \frac{\alpha_0 l_i}{\frac{\alpha_0}{\beta_0} + l_i} + \varepsilon_i \ .$$

In other words, α_1 and α_2 represent the difference between asymptotic relative growth for the first and second subzones respectively, when compared to the third subzone.

After adding the two indicator variables into the raw dataset, the next step is to fit the two different models: a complicated model that includes the indicator variables, and the simpler model without them. The SAS program to perform this task using proc NLIN is in the Appendix (note that the derivative statements in this program can be omitted if the DUD method is specified). Recall that the objective is to test the null hypothesis:

H_0: $\alpha_1 = \alpha_2 = 0$ or H_0: the asymptote is the same for the three different subzones.

The edited SAS results are given below.

Non-Linear Least Squares Summary Statistics

Model 1 (No indicator variables)

Source	DF	Sum of Squares	Mean Square
Regression	2	1. 0143723744	0. 5071861872
Residual	28	0. 0339809292	0. 0012136046
Uncorrected Total	30	1. 0483533035	

Parameter	Estimate	Std. Error
A0	0. 2137887362	0. 01287405908
B0	0. 0591584834	0. 03101092395

Model 2 (Two indicator variables)

Source	DF	Sum of Squares	Mean Square
Regression	4	1. 0392379476	0. 2598094869
Residual	26	0. 0091153559	0. 0003505906
Uncorrected Total	30	1. 0483533035	

Parameter	Estimate	Std. Error
A0	0. 1707111115	0. 00788637335
A1	0. 0873390258	0. 01108455410
A2	0. 0350858098	0. 01032020544
B0	0. 0683310452	0. 02039468305

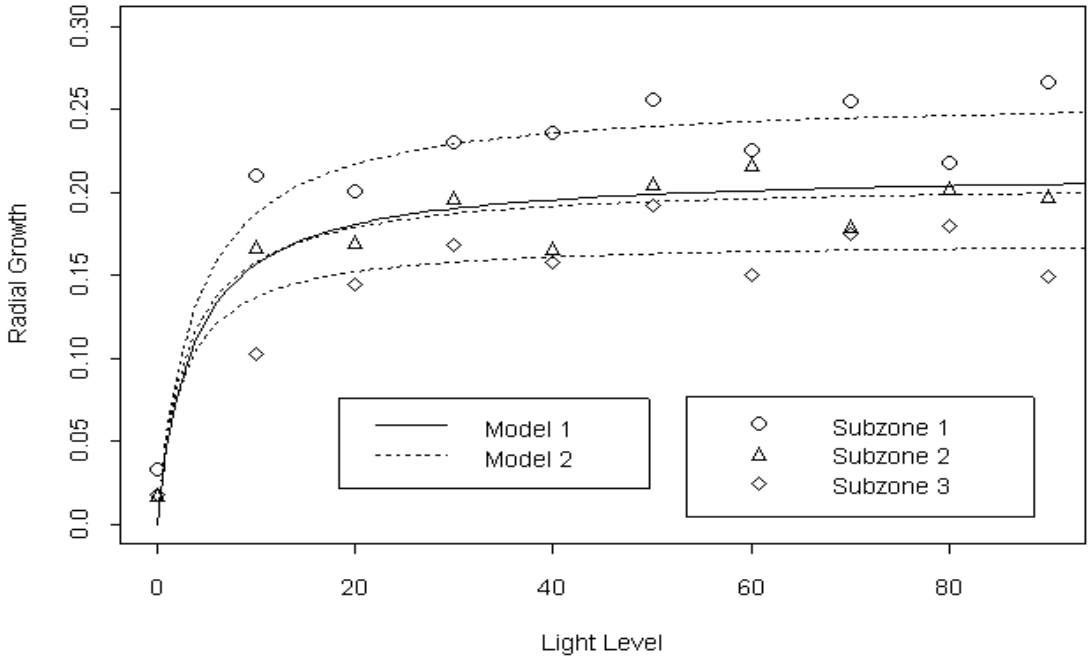

Figure 1. Graph of radial growth (r) against light level (l) from three different subzones.

A graph of the fitted models is given in Figure 1. To formally test the null hypothesis, we will use the extra sums of squares principle. Extracting the relevant information from the SAS output and doing the calculations, we see that

$$F = \frac{\frac{SSE_s - SSE_c}{\nu_c - \nu_s}}{\frac{SSE_c}{n - \nu_c}} = \frac{\frac{0.03398 - 0.00912}{4 - 2}}{\frac{0.00912}{26}} = 35.51 \text{ with } df = 2, 26.$$

This gives p ≤ 0. 0001 so we reject the null hypothesis and conclude that the asymptotic relative growth is different for the three subzones.

Interested readers should note that the significance of indicator variables in practically any model, including generalized linear models such as logistic regression, can invariably be tested using the *approximate likelihood ratio test*[5] (Kalbfleisch, 1985). This test is more general than the extra sums of squares principle since it does not require that the model has independent normal errors, but it is only valid for large sample sizes.

Contact:

Peter Ott, Research Branch, B.C. Ministry of Forests, Victoria, B.C. V8W 9C2. 387-7982.

References:

Bergerud, W. A. 1991. *Pictures of Linear Models*. Biom. Info. Hand. 1. Res. Br. B.C. Min. For. Victoria, B.C. ISSN 1183-9759. 43pp.

Bergerud, W. A. 1996. *Introduction to Logistic Regression Models: with worked forestry examples*. Biom. Info. Hand. 7. Res. Br. B.C. Min. For. Victoria, B.C. Work. Pap. 26/1996. 147pp.

Draper, N. R. and Smith, H. 1981. *Applied Regression Analysis, 2nd Edition*. John Wiley & Sons. NY. 709pp.

Kalbfleisch, J. G. 1985. *Probability and Statistical Inference. Volume 2: Statistical Inference, 2nd Edition*. Springer-Verlag. NY. 360 pp.

Lesperance, M. L. 1995. Categorical data analysis workshop notes. Unpublished. B.C. Ministry of Forests Research Branch. 122pp.

Neter, J. , Wasserman, W. , and Kutner, M. H. 1990. *Applied Linear Statistical Models, 3rd Edition*. Irwin. Boston. 1181pp.

Pacala, S. W. , Canham, C. D. , Silander, J. A. Jr. , and Kobe, R. K. 1994. Sapling growth as a function of resources in a north temperate forest. Can. J. For. Res. 24:2172-2183.

Wetherhill, G. B. 1981. *Intermediate Statistical Methods*. Chapman and Hall. NY. 390pp.

[5] Under the null hypothesis that the simpler model is correct, this test claims that for large sample sizes, the distribution of twice the difference between the maximum value of the log-likelihood for the more complicated model and the maximum value of the simpler log-likelihood, is approximately distributed as a Chi-square variable having degrees of freedom equal to the difference in the number of unknown parameters. This test will be available under most conditions, but of course deriving an explicit log-likelihood function requires an in-depth understanding of mathematical statistics.

Appendix:

```
data fake;
 input group l r d1 d2;
* group=subzone, l=light-level, r=radial-growth, d1&d2=indicator-vars;
 cards;
    1  0 0.03260295  1  0
    1 10 0.21016232  1  0
    1 20 0.20013950  1  0
    1 30 0.23010207  1  0
    1 40 0.23623511  1  0
    1 50 0.25552097  1  0
    1 60 0.22553702  1  0
    1 70 0.25496982  1  0
    1 80 0.21782379  1  0
    1 90 0.26674286  1  0
    2  0 0.01722150  0  1
    2 10 0.16693288  0  1
    2 20 0.16963460  0  1
    2 30 0.19683460  0  1
    2 40 0.16607336  0  1
    2 50 0.20516229  0  1
    2 60 0.21686777  0  1
    2 70 0.17975210  0  1
    2 80 0.20268201  0  1
    2 90 0.19793763  0  1
    3  0 0.01786722  0  0
    3 10 0.10198515  0  0
    3 20 0.14471916  0  0
    3 30 0.16844120  0  0
    3 40 0.15750866  0  0
    3 50 0.19213788  0  0
    3 60 0.15011088  0  0
    3 70 0.17530425  0  0
    3 80 0.17945065  0  0
    3 90 0.14953142  0  0
;

proc nlin data=fake;
 title 'Fit of the Simpler Model';
 parameters a0=0.25 b0=0.1;
 model r=a0*l/(a0/b0+l);
 der. a0=(l**2)/((a0/b0+l)**2);
 der. b0=(l*a0**2)/(((a0/b0+l)**2)*(b0**2));
 output out=new1 p=pred1; run;

proc nlin data=fake;
 title 'Fit of the More Complicated Model';
 parameters a0=0.35 a1=-0.20 a2=-0.10 b0=0.1;
 model r=(a0*l+a1*l*d1+a2*l*d2)/((a0+a1*d1+a2*d2)/b0+l);
 der. a0=(l**2)/(((a0+a1*d1+a2*d2)/b0+l)**2);
 der. a1=(d1*(l**2))/(((a0+a1*d1+a2*d2)/b0+l)**2);
 der. a2=(d2*(l**2))/(((a0+a1*d1+a2*d2)/b0+l)**2);
 der. b0=((a0*l+a1*l*d1+a2*l*d2)*(a0+a1*d1+a2*d2))/((b0**2)
        *((a0+a1*d1+a2*d2)/b0l)**2);
 output out=new2 p=pred2; run;

quit;
```