





LIBRARY
OF THE
MASSACHUSETTS INSTITUTE
OF TECHNOLOGY

D28
M414
.151-65

151-65

WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

DUMMY VARIABLES AND THE
ANALYSIS OF COVARIANCE

151-65

William L. White
Chi-Yuan Lin

MASSACHUSETTS INST. OF TECH.
MAR 13 1966
DEWEY LIBRARY

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139



MASS. INST. TECH.
JAN 25 1966
DEWEY LIBRARY

DUMMY VARIABLES AND THE
ANALYSIS OF COVARIANCE

151-65

William L. White
Chi-Yuan Lin

MAH 16 1966
DEWEY LIBRARY

The authors are respectively Assistant Professor and Graduate Student at the Sloan School of Management at M.I.T.

The authors wish to acknowledge that it was the generous support of the Ford Foundation's grant to M.I.T.'s School of Industrial Management for Research in Business Finance which made this work possible.

DUMMY VARIABLES AND THE ANALYSIS OF COVARIANCE

Introduction

Until the last few years, most regression studies have been constructed around either time series data or cross section data. As the art of model building developed, considerable interest arose in the testing of hypotheses which required the pooling together of data from several cross sections, e.g., the data on several firms in each of many years. A well known problem in time series analysis had been that of autocorrelated errors. It was known that this phenomenon had its counterpart in intertemporal or pooled cross section analysis. Regressions run on pooled cross section data assume that error terms are independent drawings while in fact the elements in each cross section are often sampled in each of the cross section years. If the errors are not independent, that is, if there are consistent components to the error term of each element every time it is sampled, it has been shown that the unexplained variance and the estimates of the slope coefficients which arise from the usual regression techniques are in error. Carter^[2] showed that the inclusion in the pooled regression equation of a discrete (mutually exclusive and exhaustive) dummy variable (zero-one) for each cell would result in unbiased estimates of unexplained variance and slope coefficients and thus retain the desired properties of classical regression analysis.

Until recently, these techniques have required such large computational capacity that they were of little practical interest. With the advent of more exotic computation facilities, however, some earlier statistical devices which require the pooling of data from cross sections and time series have come to be of real use.

During this same time, the dummy variable technique suggested by Carter has been shown to have further interest in a paper by Suits. Suits [6] suggests that the coefficient of these dummy variables can be of significant interest in and of themselves. That is, the coefficients associated with the dummy variables specifically account for those peculiarities associated with each cell that are not "explained" by the included independent variables.

While the inclusion of dummy variables may be a device to improve the quality of estimate in a pooled cross section analysis, it should be noted that the validity of conclusions drawn from such analyses depends upon a typically untested assumption. Dummy variable regressions assume that the intra cell slopes are identical and tests the hypothesis that the cells have homogeneous intercepts. If that hypothesis is rejected, the coefficients of the dummy variables are taken as estimates of the inter-cell difference. In the language of the analysis of covariance, dummy variables assume that there is no difference in within cell relations (slopes) and that the only source of variation is due to differences between the cells -- differences reflected in their intercepts. Put in this way, it becomes clear that dummy variable regressions are a kind of analysis of covariance -- a partitioning of explained variance into a within cell and a between cell component.

It is the purpose of this article to suggest that the technique of the analysis of covariance dominates, in a statistical sense, the use of dummy variables. We shall first show how dummy variable regressions and the analysis of covariance are related. This comparison will show why we think the analysis of covariance is a more useful technique than the use of dummy variables in a regression analysis. Then in an attempt to make our prescription of the analysis of covariance somewhat easier to apply we shall develop a set of computational short cuts. In our view these result in computational difficulties of such small extra cost that they seem clearly to suggest that the analysis of covariance is a technique much more useful than simply inserting dummy variables into regressions.

In what follows we will present the test for the significance of the inclusion of a set of dummy variables in a regression equation; the tests required in the analysis of covariance; and then contrast these two procedures. Finally, we shall suggest a set of problems for which the analysis of covariance framework, which partitions variance into between and within components, can be used to separate long and short run relations between variables.

The Test for the Significance of a Set of Dummy Variables

The test for the significance of a set of dummy variables is exactly the same as a test for the significance of the set of any regular (continuous) included variables.* It results from a partitioning of the total variance of the dependent variable into that due to the regression of the regular independent variables, that due to the inclusion of the dummy variables and a residual.

Let y_{ij} and x_{ij} be the j th pair of observations in the i th cell (group) ($i=1,2,3,\dots,I, j=1,2,3,\dots,J$) of the dependent variable and independent variable, respectively.** The classical regression model, without dummy variables is

$$(1) \quad y_{ij} = \alpha + \beta x_{ij} + \epsilon_{ij}.$$

If f_i is the effect of the i th cell, then the model is

$$(2) \quad y_{ij} = \alpha_0 + f_i + \beta_w x_{ij} + \epsilon_{ij}, \quad ***$$

or

$$(2)' \quad y_{ij} = \alpha_i + \beta_w x_{ij} + \epsilon_{ij},$$

where α_i is the sum of α_0 and f_i .

In fact, when dummy variables are used in a regression equation, they are added as $I-1$ **** additional independent variables as shown in equation (2)".

$$(2)'' \quad y_{ij} = \beta_0 + \beta_w x_{ij} + \sum_{k=1}^{I-1} f_i \delta_{ik} + \epsilon_{ij},$$

where $\delta_{ik} = 0$ if $i \neq k$ and $\delta_{ik} = 1$ if $i = k$.

The test for the statistical significance of this set of dummy variables is summarized in Table 1.

* Analysis-of-variance in regression contexts is discussed in Statistics and Econometrics texts, such as Bryant, [1], and Fraser, [3].

** In order to keep the notation clear, we shall include only one x in the equations. The analysis is easily extended to include more than one x and, in fact, all the tables refer to equations with $n-1$ x 's.

*** As there will be need for several β 's, this one has a subscript w to signify it is the slope within the individual cells.

**** Although there are I groups, one dummy variable has to be omitted in a regression equation, in order to obtain determinate estimates of parameters, as suggested by Suits, [6].

TABLE 1

Source	Computational Short Cut of SS	df	MS
Reduction due to x's	$SS(y) - SSR(A)$	$n-1^*$	(1)
Additional reduction due to dummy variables	$[SS(y) - SSR(D)] - [SS(y) - SSR(A)] = SSR(A) - SSR(D)$	$I-1$	(2)
Residual	$SSR(D)$	$IJ-n - (I-1)$	(6a)
Total	$SS(y)$	$IJ-1$	

The direct computation of the sum of squares for Column 1 of Table 1 is tedious, ^{**} especially in a multiple regression, and hence a computational short-cut is quite desirable. Column 2 suggests such a short-cut. In addition, this representation displays the contribution of dummy variables more clearly and will be of considerable interest later.

Define:

$$SS(y) = \sum_{ij} (y_{ij} - \bar{y})^2, \text{ i.e., the total sum of squares of the dependent variable, with } IJ-1 \text{ degrees of freedom.}$$

$$SSR(A) = \sum_{ij} (y_{ij} - a - bx_{ij})^2, \text{ i.e., the sum of squared residuals from a classical regression (through All the data), with } IJ-n \text{ degrees of freedom.}$$

* n is the number of independent variables (constant included).

** An algebraic expression for these sums of squares can be found in most statistics textbooks as a part of their discussion of regression analysis. See Bryant [1], pp. 214-216 or Frazer [3], pp. 296-304.

$$SSR(D) = \sum_{ij} (y_{ij} - a_i - b_w x_{ij})^2, \text{ i.e., the sum of squared residuals from a regression with Dummy variables included, with } IJ - n - (I - 1) \text{ degrees of freedom.}$$

It should be clear that $SSR(D)$ measures the unexplained or residual portion of the variation. Also, the combined reduction due to the x 's and Z 's is $SS(y) - SSR(D)$. Since the reduction due to the inclusion of the x 's is $SS(y) - SSR(A)$, the incremental reduction due to the inclusion of the Z 's, the dummy variables, is $[SS(y) - SSR(D)] - [SS(y) - SSR(A)]$. That is, it is the reduction due to the combined effects of the x 's and the Z 's minus the effect of the x 's. Simplifying, this becomes $SSR(A) - SSR(D)$. In other words, the contribution of the dummy variables is measured by the difference between the residual sum of squares after the inclusion of the x 's, but before the inclusion of the dummy variables, and the residual sum of squares after the inclusion of both the x 's and the dummy variables.

Mean squares (Column 4) are obtained by dividing the SS by its corresponding degrees of freedom. In the dummy variable framework, the significance of the different intercepts is tested by the F ratio $\frac{MS(2)}{MS(6a)}$. If the computed F ratio is found to be larger than the critical F values, then the cell effects are thought to be non zero and contribute significantly to an understanding of the variation of the dependent variable.

The Analysis of Covariance

The analysis of covariance extends the regression equation with dummy variables to allow for the inclusion of differing cell slopes. It adds a third step in the progression of models; namely,

$$(3) \quad y_{ij} = \alpha_i + \beta_i x_{ij} + \epsilon_{ij}$$

It results in a partitioning of the total sum of squares into the following components:

TABLE 2

Source	Computational Short Cut for SS	df	MS
Due to x's.	$SS(y) - SSR(A)$	$n-1$	(1)
Due to dummy variables -- different cell intercepts.	$SSR(A) - SSR(D)$	$I-1$	(2)
Due to different cell slopes.	$SSR(D) - SSR(I)$	$(I-1)(n-1)$	(5)
Residual	$SSR(I)$	$I(J-n)$	(6b)
Total	$SS(y)$	$IJ-1$	

where

$$SSR(I) = \sum_{ij} (y_{ij} - \alpha_i - b_i x_{ij})^2, \text{ i.e., the sum of squared residuals from } I \text{ regressions through each of the individual cells, with } I(J-n) \text{ degrees of freedom in total.}$$

In this presentation, the residual sum of squares in the dummy variable regression analysis has been split into two components. One is the reduction in explained variance due to allowing different within cell slopes and the other becomes the new estimate of residual or unexplained variance.

Within the analysis of covariance framework, the hypotheses to be tested are: first, are the within cell slopes equal; and second, are the cell intercepts different. The test procedure is first to determine if there is any evidence that the cell slopes differ. This test is performed by dividing the mean square contribution due to different cell slopes by the residual mean square, that is $\frac{MS(5)}{MS(6b)}$. If there is no evidence that all slopes differ, i.e., no evidence that $SSR(D)$ is significantly different from $SSR(I)$, one next tests whether the introduction of many intercepts

rather than a single intercept significantly reduces the sum of squares of residuals by forming the ratio $\frac{MS(2)}{MS(6b)}$.*

The Relationship Between the Use of Dummy Variables and the Analysis of Covariance

Having developed the tests implied in both regression analysis with dummy variables and the analysis of covariance, it is possible to see how these methods relate to each other. In both methods, the significance of the cell intercepts is tested by comparing the mean square reduction due to many as opposed to one intercept with an estimate of residual mean square. The tests differ only in what is considered the residual mean square. The dummy variable regression procedure assumes cell slopes are the same; that is, it assumes $SSR(D)$ is equal to $SSR(I)$. Thus it tests $SSR(A) - SSR(D)$ against $SSR(D)$. The analysis of covariance procedure does not assume $SSR(D)$ equals $SSR(I)$. In fact, it explicitly tests this hypothesis. If the hypothesis of equal cell slopes is rejected, the analysis of covariance procedure does not allow one to test the homogeneity of cell intercepts.

Within the analysis of covariance framework, the test for homogeneous intercepts is conditional on the cell slopes have been shown not to differ significantly from each other. Thus the analysis of covariance procedure does all that the dummy variable regression technique can do and in addition it explicitly tests the untested assumption of equal slopes which dummy variable regressions require. In this sense the analysis of covariance dominates the dummy variable regression techniques.**

In addition to the partitioning of explained variance as shown in Table 2, there are more informative uses to which one can put the analysis of covariance. The table 2 partitioning was employed to ease the comparison with regression analysis encompassing dummy variables. Table 3 shows a further possibility.*** In it the difference between $SSR(A)$ and $SSR(D)$

*Some statisticians recommend "pooling". If it is concluded that $SSR(D) = SSR(I)$, then $MS(5)$ and $MS(6b)$ would be combined to form an "error" MS based on $(I-1)(n-1)$ more degrees of freedom, namely $IJ-n-(I-1)$ degrees of freedom.

**The only additional computation required is a set of regressions run through the data for each cell. This would seem to come at little cost with modern computers and regression programs.

***This method of partitioning the variance can be found in Mood [5], pp.350-356.

TABLE 3

Source	Computational Short Cut for SS	df	MS
Due to x's	$SS(y) - SSR(A)$	$n-1$	(1)
Due to cell mean relation	$SSR(M)$	$I-n$	(3)
Due to the difference of within and between cell slopes	$SSR(A) - SSR(D) - SSR(M)$	$n-1$	(4)
Due to different cell slopes	$SSR(D) - SSR(I)$	$(I-1)(n-1)$	(5)
Residuals	$SSR(I)$	$I(J-n)$	(6b)
Total	$SS(y)$	$IJ-1$	

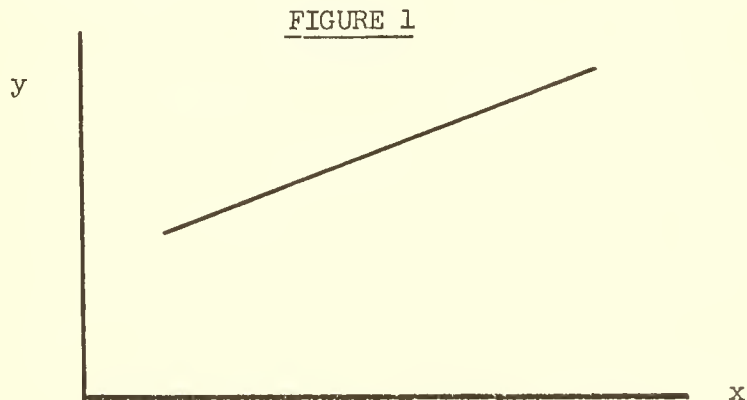
is further partitioned into two parts -- $SSR(M)$, and $SSR(A)$ minus $SSR(D)$ minus $SSR(M)$. The $SSR(M)$ term is the sum of squares of residuals from a regression through the cell means. Under this partitioning, as before, one tests for homogeneous within cell slopes by comparing $MS(5)$ with $MS(6b)$. Given that these intra-cell slopes are homogeneous, one tests for the significance of different intra-cell intercepts in two steps. The first step is a test for the existence of a between cell relation. This is the purpose of the regression through the cell mean data. Its existence is measured by comparing $MS(3)$ with $MS(6b)$. If the regression through the cell means "fits" as well as the individual within cell regressions the between cell relation is said to exist.

Given that this between cell relation exists, and that the intra-cell slopes are homogeneous -- that an intra-cell slope exists, one compares the two. If the slope of the regression through the cell means differs from the slope of the regression within the cells, it is said that there exist different intra-cell intercepts.

This test is an extremely useful one to test the hypothesis that the within cell regression relation differs from the between cell regression relation within a body of pooled cross section data. This last specification of the analysis of covariance is more subtle than that in Table 2, as it specifically tests first for the existence of a homogeneous within cell relation and for the existence of a between cell regression relation and only then compares the two to see if they are equal. That is, it will not offer evidence that the between cell relation differs from the within cell relation unless it is satisfied that both relationships actually exist.

At this point, a graphical presentation of what the analysis of covariance procedure enables one to test may be useful.

Suppose one ran a regression between a dependent variable y and an independent variable x where y and x are data from several firms in each of several years. Let the diagram in Figure 1 stand for the estimated regression relation between y and x .



This regression relation could arise from any one of the possibilities shown in Figures 2 through 5, where the oval represents the cluster of data for each firm.

FIGURE 2

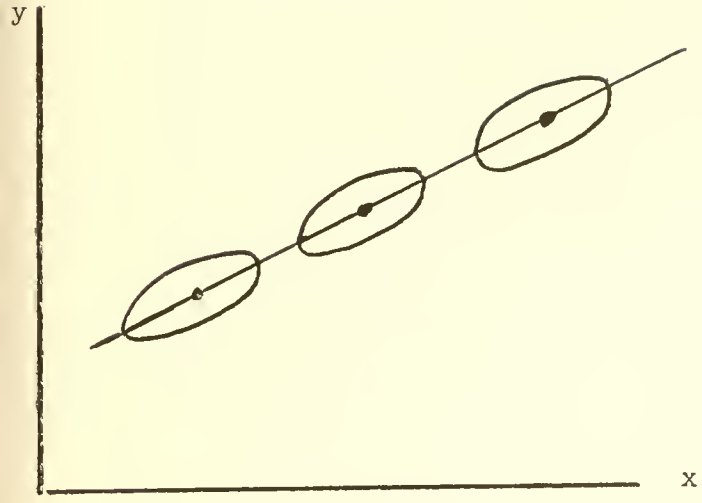


FIGURE 3

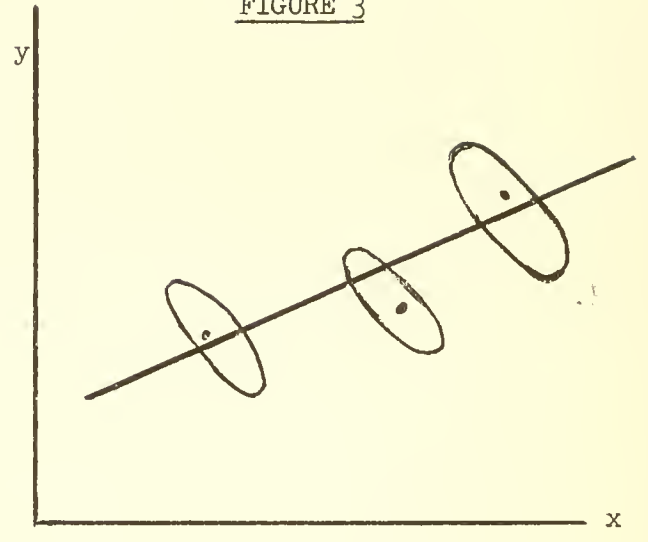


FIGURE 4

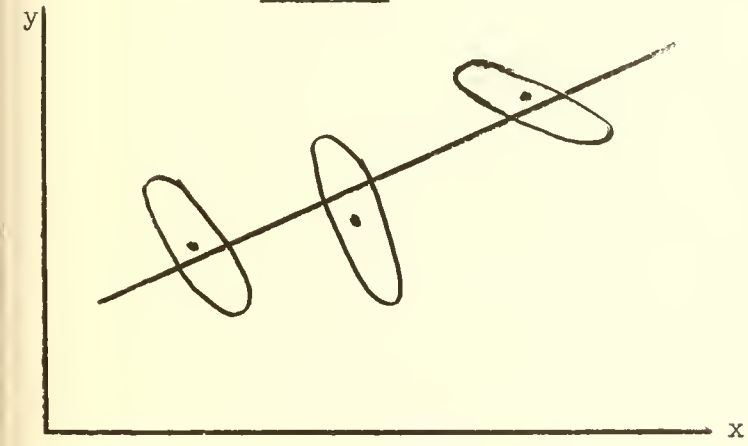
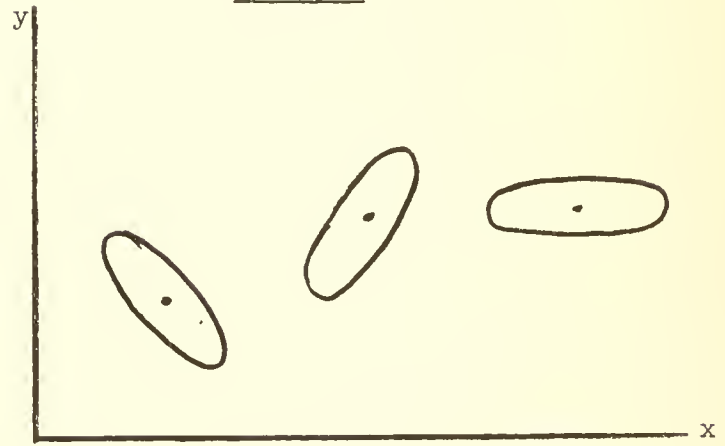


FIGURE 5



In Figure 2 the within firm regression relation between y and x is the same as the between firm relation. In Figure 3 the within firm regression relation is quite different from the between firm regression relation. This is a situation in which regression analysis using dummy variables would be acceptable. Figure 4, on the other hand, presents a situation where there is some evidence of a between firm regression relation but non homogeneous relations of y to x among firms. Figure 5 presents the most pervase of the possibilities. It is meant to suggest the possibility of a generally upward sloping relation between y and x without a significant between firm regression relation and heterogeneous within firm relations. Given a regression relation as in Figure 1, it is the purpose of the analysis of covariance technique to distinguish whether the underlying process is that in Figure 2 or one of those in Figures 3 through 5.

The next section of this paper presents an example of how this specification of the analysis of covariance procedure can exploit the data generated by an actual process for considerable information about its nature.

The Analysis of Covariance and the Framing of Hypotheses

The kinds of problems discussed in this paper all concerned pooled cross section data. The dummy variable regression technique concentrates interest on the within cell slope and the within cell intercepts. The analysis of covariance framework, however, can be used to concentrate attention on a different aspect of the problem. It can be used to partition the variation explained by the regression equation into two parts -- that variation arising from relations between the cell means and that variation which arises from relations within the cells. For a kind of problem of significant interest to the authors such a difference in focus is quite important. Suppose the pooled cross sections are data on stock prices of a set of firms in each of several years and variables associated with the firms which are thought to affect these stock prices.

It seems sensible to hypothesize that variations in price may occur because one firm differs (on the average) from another, or because in a specific year one firm pursues policies different from its average policy. Thus the total of the variation being explained can be thought of as arising from variations between the firms and variations within the firms over the period studied. To be specific, it was hypothesized that stock prices responded to debt and dividends payout policies and that these responses were composed of two types of influences:

- a. The influence of debt and dividend policies, which were said to be described by the average of the variables, and
- b. the influence of short run variation in debt and dividends around these desired or policy levels.

Thus in any specific year, the stock price of, say, Standard Oil of New Jersey, is thought to differ from that of Texaco not only because Standard pursues different financial policies but also because, in that year, Standard and/or Texaco may have debt ratios or dividend payout ratios which differ from their target or average ratios due to the peculiarities of that year. Stated in even another way, variations in stock prices are thought to arise from variations in established financial policies between companies, and from within company year-to-year aberrations around these financial policies.

For such a problem, the analysis of covariance as presented in Table 3 seems a most appropriate statistical tool. First, within its framework it is possible to test for the existence of differences in stock prices which arise from between company differences. These are the differences thought to arise from the fact that the different companies pursue different financial policies which will be measured by the average of the variable in question over the time period studied. Second, it is possible to test for the existence of a response of stock prices to short run fluctuations in financial variables around their average or target level. Finally, it is possible to test to see if the two relations -- the long run and the short run -- of debt and dividends on stock prices are equal. If, in fact, the evidence is such that one has reason to believe the data are generated

by two sets of forces -- short run and long run -- there is considerable reason to question the validity of some other studies of the relation between stock prices and financial variables which made no attempt to isolate these two different types of influences.

The financial variables which affect stock prices are said to be the ratio of debt to total capitalization and the ratio of dividends to net profits. The process by which these two variables affect stock prices is said to be the following. In determining the price they are willing to pay for a particular security, investors formulate a return they require from the stock in order to justify their holding this stock rather than some other. The return they expect is composed of dividends and price appreciation. The rate of return they realize is the sum of the dividend price ratio -- the dividend yield -- plus the percentage growth in price. Thus, given their required rate of return, stockholders estimate a growth rate and the prospective dividend per share and pay a price such that the prospective dividend yield and capital gain is at least as high as their required rate of return. That is, given their expectation about the growth rate, g , and the dividend per share, DIV , stockholders set prices so that

$$\frac{DIV}{PRICE} + g = r$$

where r is the return they require in order to hold this particular security.

The required rate of return is taken to be higher if the company employs more debt to finance itself. The more of the money that finances the company which requires fixed contractual payments -- the higher the debt ratio, the higher the risk and thus the higher the required return is said to be. The role of the dividend payout ratio is more difficult to state a priori. Since the capital gains tax rate is lower than the income tax rate for most stockholders, low payout ratios would seem to cause stockholders to require lower before tax returns as this would allow most of the return to come in the form of capital gains which would arise from the retention of earnings. On the other hand, the feeling that dividends now might be preferred to possible capital gains later might make higher payout ratios induce stockholders to accept lower returns because they thought of them as more secure returns. Thus, we have no strong a priori views on the sign of the slope coefficient of the dividend payout ratio term.

The model as developed can be presented as

$$\left(\frac{\text{DIV}}{p} + g\right)_{ij} = \overline{\left(\frac{\text{DIV}}{p} + g\right)_j} + a_0 + b_1 \left(\frac{D}{D+E}\right)_{ij} + b_2 \left(\frac{\text{DIV}}{\text{PRO}}\right)_{ij} + \epsilon_{ij}$$

where

$\left(\frac{\text{DIV}}{p} + g\right)_{ij}$ is the dividend yield plus growth rate -- the total rate of return for the *i*th firm in the *j*th year.

$\overline{\left(\frac{\text{DIV}}{p} + g\right)_j}$ is the average rate of return for all *I* firms in year *j*. It is a "market" rate of return for that year.

$\left(\frac{D}{D+E}\right)_{ij}$ is the ratio of debt to total capitalization or of debt to debt plus equity of the *i*th firm in the *j*th year.

$\left(\frac{\text{DIV}}{\text{PRO}}\right)_{ij}$ is the ratio of dividends to profits for the *i*th firm in the *j*th year.

That is, the required return for the *i*th firm in year *j* depends first on the state of the stock market in year *j* as measured by the average required return for all stocks covered. In addition to this effect of the year, however, there are two effects which are peculiar to each company -- that arising from its debt policy and that arising from its dividend policy.

To allow the pooling of the annual cross sections, the model will be written for testing purposes as

$$\left(\frac{\text{DIV}}{p} + g\right)_{ij} - \overline{\left(\frac{\text{DIV}}{p} + g\right)_j} = a_0 + b_1 \left(\frac{D}{D+E}\right)_{ij} + b_2 \left(\frac{\text{DIV}}{\text{PRO}}\right)_{ij} + \epsilon_{ij}$$

The sample used to test this relation is drawn from 20 firms in the Food industry with the data on each firm drawn from the years 1949 to 1960.

A regression through all the data yielded the following results:

$$\left(\frac{\text{DIV}}{p} + g\right)_{ij} - \overline{\left(\frac{\text{DIV}}{p} + g\right)_j} = .0933 - .0131\left(\frac{D}{D+E}\right)_{ij} - .1611\left(\frac{\text{DIV}}{\text{PRO}}\right)_{ij} + \epsilon_{ij}$$

(4.1796)* (.3554) (4.3735)

$$R^2 = .0768$$

$$F(2, 237) = 9.8645^{**}$$

In this regression, the correlation, while significant, is quite low. There seems no relation between debt ratios and required returns. Finally, the significant and negative coefficient of dividends suggests that prices rise as payout ratios rise.***

A regression which employed dummy variables, however, yielded somewhat different results. These were

$$\left(\frac{\text{DIV}}{p} + g\right)_{ij} - \overline{\left(\frac{\text{DIV}}{p} + g\right)_j} = .2196 - .1275\left(\frac{D}{D+E}\right)_{ij} - .3356\left(\frac{\text{DIV}}{\text{PRO}}\right)_{ij} + f_i + \epsilon_{ij}$$

(7.0614) (3.0913) (7.1064)

$$R^2 = .1957$$

$$F(21, 218) = 2.5264^{****}$$

In this equation, the correlation, while small, is significant. Moreover, there seem to be significant relations within each firm (each cell) between debt, dividends, and stock prices -- the t ratios are quite significant. Surprisingly, higher debt ratios mean higher stock prices and, as before, higher dividend payout ratios mean higher stock prices.

* The numbers in parentheses are the "t" ratios associated with each slope coefficient.

** This is significantly different from zero at the .01 level.

*** This result is like that found by others who have studied stock prices. See for instance Myron Gordon, [4].

**** Significantly different from zero at the .01 level.

If we now apply the analysis of covariance technique to this problem and specifically test first for the existence of a relation through the cell means and second for the existence of homogeneous within cell (within firm) relations between debt and dividends and stock prices, these first results are shown to be quite misleading. First, as can be seen in Table 4, there is evidence of the existence of a regression relation through the cell means. This follows from the fact that the cell mean regression fits as well as the individual cell regression -- from the fact that MS(3) does not significantly differ from MS(6b). However, from the estimated regression equation through the cell mean data,

$$\overline{\left[\left(\frac{DIV}{p} + g \right)_{ij} - \left(\frac{DIV}{p} + g \right)_j \right]_i} = .0179 + .0115 \left(\frac{\bar{D}}{D+E} \right)_i - .0371 \left(\frac{\overline{DIV}}{PRO} \right)_i + \epsilon_i$$

(.7848) (.3443) (.9596)

we see that the slope coefficients for the debt and dividend terms from this cell mean regression are not like those obtained from the dummy variable regression. These slope coefficients are different in sign, in magnitude, and in their statistical significance. They suggest that while debt and dividend explain part of the between firm variation in prices, they do not explain very much.

Second, the within cell relations or what we call the short run relation of prices to debt and dividend fluctuations around their means are not homogeneous among firms. That this is so can be seen from the comparison of MS(5) and MS(6b). For some firms, stock prices rise as dividend payout rise and for other firms stock prices fall.

Thus, while both the regression through all the data and the dummy variable regressions gave evidence of a relation between dividends and debt ratios and stock prices, with dividends seeming to play a quite significant role, a closer examination shows little evidence of such a relation between firms. Furthermore, what within firm relation that does exist is not homogeneous between firms. An application of the analysis of covariance procedure to this model of stock prices suggests great care must be used in the interpretation of simple cross section studies and any pooled

TABLE 4

Source	Computation Short Cut for SS	df	MS	Computed F	Critical F .01
Due to x's	$SS(y) - SSR(A) = .1168949$	$3-1=2$	⁽¹⁾ .0584475	12.7509	4.78
Due to cell mean relation	$SSR(M) = .0654840$	$20-3=17$	⁽³⁾ .0038520	.8404	2.10
Due to the difference of within and between cell slope	$SSR(A) - SSR(D) - SSR(M) = .1154781$	$3-1=2$	⁽⁴⁾ .0577391	12.5963	4.78
Due to different cell slopes	$SSR(D) - SSR(I) = .3991303$	$(20-1)(3-1) = 38$	⁽⁵⁾ .0105034	2.2914	1.72
Residuals	$SSR(I) = .8250828$	$20(12-3) = 180$	^(6b) .0045838		
Total	$SS(y) = 1.5220701$	$20 \times 12-1 = 239$			

where $SSR(M) = .0654840$
 $SSR(A) = 1.4051752$
 $SSR(D) = 1.2242131$
 $SSR(I) = .8250828$

cross section studies such as those employing dummy variables which do not explicitly test for homogeneous within cell behavior and the existence of a well defined between cell behavior. The results from the regressions which allowed each firm its own set of slope coefficients showed the results from the dummy variable regression to be most misleading.

Conclusion

The purpose of this article has been threefold. It has attempted to draw the link between regression analyses with dummy variables and the analysis of covariance. In addition it has advanced the view that the analysis of covariance dominates dummy variable regressions in the examination of pooled cross section data -- it generates more informative tests at what seem to be low cost. Finally, we have presented an application of the analysis of covariance to a kind of problem where its use results in the ability to separate long run and short run relationships between variables. This was an example which attempted to isolate the long run and the short run relation of financial policies to stock prices. This example was meant to serve as evidence that indiscriminant use of regression analysis using dummy variables can be most misleading, and careful use of the analysis of covariance can be much more informative.

REFERENCES

- [1] Bryant, E. C., Statistical Analysis, New York, McGraw-Hill, 1960.
- [2] Carter, A. H., "The Estimation and Comparison of Residuals Regressions Where There are Two or More Related Sets of Observations," Biometrika, Vol. 36, 1949, pp. 26-46.
- [3] Fraser, D. A. S., Statistics: An Introduction, New York, Wiley, 1958.
- [4] Gordon, M., The Investment, Financing and Valuation of the Corporation, Homewood, Illinois, Irwin, 1962.
- [5] Mood, A. M., Introduction to the Theory of Statistics, New York, McGraw Hill, 1950.
- [6] Suits, D. B., "Use of Dummy Variables in Regression Equations," Journal of the American Statistical Association, Vol. 52, 1957, pp. 548 - 551.



LIBRARY
Date Due

NOV. 09 1984

~~FE~~

~~RE~~

~~RE~~

JUL 17 '81

AG 21 '84

AG 11 '87

AG 8 '89

Lib-26-67

MIT LIBRARIES
3 9080 003 899 587

142-65

MIT LIBRARIES
3 9080 003 899 595

143-65

MIT LIBRARIES
3 9080 003 899 603

144-65

MIT LIBRARIES
3 9080 003 899 611

145-65

MIT LIBRARIES
3 9080 003 868 624

146-65

MIT LIBRARIES
3 9080 003 899 629

147-65

MIT LIBRARIES
3 9080 003 900 153

148-65

BASEMENT

MIT LIBRARIES
3 9080 003 869 135

149-65

SIGN this card and present it with book
at the CIRCULATION DESK.

MIT LIBRARIES
3 9080 003 900 138

150-65

MASSACHUSETTS INSTITUTE OF
TECHNOLOGY LIBRARIES

MIT LIBRARIES
3 9080 003 869 168

151-65

1751893

HD28 M.I.T. Alfred P. Sloan
.M414 School of Management
Nos.142-65 Working Papers.
Nos.151-65

DEWEY LIBRARY

