International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org Volume 2 Issue 3 ‖ March. 2013 ‖ PP.42-51

Dummy Variable Multiple Regression Forecasting Model

Nwankwo, C. H[1], Oyeka, I.C.A.[2]

[1]*Department of Statistics, Nnamdi Azikiwe University, Awka. Nigeria*
[2]*Department of Statistics, Nnamdi Azikiwe University, Awka. Nigeria*

ABSTRACT: *A method of using multiple regression in making forecasts for data which are arranged in a sequence, including time-order, is presented here. This method uses dummy variables, which makes it robust. The design matrix is obtained by a cumulative coding procedure which enables it overcome the setback of equal spacing associated with dummy variable regression methods. The use of direct effects obtained by the method of path analysis makes the whole procedure unique.*

Keywords: *Cumulative Coding, Dummy, Path Analysis, Design Matrix, Robust, Direct Effect*

I. INTRODUCTION

Usually, multiple regression models are used for estimating the contributions or effects of independent variables on a dependent variable. A first order multiple regression model with two independent variables X_1 and X_2 is in the form [1].

$$Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\varepsilon_i \qquad (1)$$

This model is linear in the parameters and linear in the independent variables. Y_i denotes the response in the *i*th trial and X_{i1} and X_{i2} are the values of the two independent variables in the *i*th trial. The parameters of the model are β_0, β_1 and β_2, and the error term is ε_i.
Assuming that $E(\varepsilon_i) = 0$, the regression function for (1) is

$$E(Y)=\beta_0+\beta_1X_1+\beta_2X_2 \qquad (2)$$

Note that the regression function (2) is a plane and not a line.
The parameter β_0 is the y intercept of the regression plane.

The parameter β_1 indicates the change in the mean response per unit increase in X_1 when X_2 is held constant. Likewise, β_2 indicates the change in the mean response per unit increase in X_2 when X_1 is held constant.
When there are more than two independent variables, say p-1 independent variables $X_1, \ldots, X_{p-1}$, the first order model is

$$Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\ldots+\beta_{p-1}X_{i,p-1}+\varepsilon_i \qquad (3)$$

Assuming that $E(\varepsilon_i) = 0$, the response function for (3) is

$$E(Y)=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_{P-1}X_{p-1} \qquad (4)$$

The response function (4) is a hyper plane, which is a plane in more than two dimensions. It is no longer possible to picture this response surface as we were able to do with (2), which is a plane in 3 dimensions.
Multiple regression is one of the most widely used of all statistical tools. A regression model for which the response surface is a plane can be used either in its own right when it is appropriate, or as an approximation to a more complex response surface. Many complex response surfaces are often approximated well by a plane for limited ranges of the independent variables [1].

The use of multiple regression is largely limited to the problem of estimating the contributions, or estimation of the effects of the independent variables on the dependent variable. Forecasting values of the dependent variable using multiple regression models is often of interest to researchers, though forecasting is not a common feature of existing regression methods. The problem is that independent variables are not available for the period onto which forecasts are sought.

A method for carrying out forecasts of this sort is proposed here and will be applicable when data is in at least ordinal form.

II. PROPOSED METHOD

To achieve the objective of this paper, which is to develop a multiple regression forecasting model, we propose the use of dummy variable multiple regression modeling methods [2]. Specifically a 0,1 dummy variable coding system would be used in such a way that each category or level of a parent independent variable in a regression model is represented by a pattern of 1's and 0's, forming a dummy variable set. In order to avoid linear dependence among the dummy variables of a parent variable each parent variable is always represented by one dummy variable less than the number of its categories [3][1]. Thus if a given parent variable Z has z categories or levels, the corresponding design matrix X will be represented ordinally by z-1 column vectors of ordinally coded dummy variables, x_d, of 1's and 0's (for d=1,2,…,z-1). The 1's and 0's in each x_d are cumulative if the values of the level of the parent variable it represented are arranged together.

Specifically, the pth level (p = 1, 2, … z) of the z levels of a parent variable Z will be represented by d ordinally coded column vectors of 1's and 0's for d = 1, 2, … z-1 that is:

$$X_{id}=\begin{cases} 1, if\ yi\ is\ in\ level\ p\ of\ Z; p > d; d = 1,2,\dots,z \\ 0, otherwise \end{cases} \quad (5)$$

Then if, but without loss of generality, the observations in each level of Z are arranged all together, then the n *x* (z-1) design matrix X representing Z will consist of a set of z-1 cumulatively coded column vectors x_d of 1's and 0's of the form

Equation (6) is a prototype of ordinally coded design matrix X with z-1 cumulatively coded column vectors x_d

$$X = \begin{array}{c} \text{Level of } Z \\ 1 \\ \\ \\ \\ \\ \\ \\ \\ 2 \\ \\ \\ \\ \\ \\ \\ 3 \\ \\ \cdot \\ \cdot \\ \cdot \\ z \\ \\ \\ \end{array} \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdot & \cdot & \cdot & x_{1z-1} \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & 1 & \cdot & \cdot & \cdot & 1 \\ 1 & 1 & 1 & \cdot & \cdot & \cdot & 1 \\ 1 & 1 & 1 & \cdot & \cdot & \cdot & 1 \end{pmatrix} \quad . \quad . \quad . \quad (6)$$

of 1's and 0's representing the z levels of the parent variable Z. Note that the first n_1 elements of the first column x_1 of X representing the first level of Z are 0's while the remaining $n-n_1$ are all 1's. The firstn$_1$ + n_2 elements of

x_2 are 0's, while the remaining n-(n_1 + n_2) elements are all 1's and so on until finally all the elements of x_{z-1} are all 0's except the last n_z elements which are all 1's.
Note that all the observations in the first level (level 1) of Z are all coded 0 in all the columns of the design matrix X while observations in the last level (level z) of Z are all coded 1's in X.
Note also that Z may be any set of parent independent variables such as A, B, C etc with levels a, b, c, etc respectively. An ordinal dummy variable multiple regression model of y_i on the x_{ij}'s may be expressed as

$$y_i = \beta_0 + \beta_{1;A} X_{i1;A} + \beta_{2;A} X_{i2;A} + \ldots.. + \beta_{a-1;A} X_{i,a-1;A} + \ldots\ldots\ldots. + \beta_{c-1;C} X_{i,c-1;C} + e_i \qquad (7)$$

where β_j's are partial regression coefficients and e_i are error terms uncorrelated with x_{ij}'s, with $E(e_i) = 0$; A has 'a' levels, B has 'b' levels … C has 'c' levels, etc.
Note that the expected value of y_i is

$$E(y_i) = \beta_0 + \beta_{1;A} X_{i1;A} + \beta_{2;A} X_{i2;A} + \ldots\ldots\ldots \ldots + \beta_{a-1;A} X_{i,a-1;A} + \ldots.. + \beta_{c-1;C} X_{i,c-1;C} \qquad (8)$$

Equation 7 may alternatively be expressed in its matrix form as

$$\underline{y} = X \underline{\beta} + \underline{e} \qquad (9)$$

where $\underline{y}$ is an nx1 column vector of outcome values; X is an nxr cumulatively coded design matrix of 1's and 0's; $\underline{\beta}$ is an rx1 column vector of regression coefficient and $\underline{e}$ is an nx1 column vector of error terms uncorrelated with X with $E(\underline{e})=0$ where r is the rank of the design matrix X.
Use of the method of least squares with either equation (7) or (9) yields an unbiased estimator of $\underline{\beta}$ as

$$\underline{b} = (X'X)^{-1} \; X'\underline{y} \qquad (10)$$

where $(X'X)^{-1}$ is the matrix inverse of $(X'X)$, the resulting predicted regression model is

$$\hat{\underline{y}} = X \underline{b} \qquad (11)$$

The following analysis of variance (ANOVA) table (Table1), enables the testing of the adequacy of Equations (7) or (9) using the F test.

Table 1: Analysis of Variance (ANOVA) Table for Equation (9)

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (DF)	Mean Sum of Squares (MS)	F. Ratio
Regression	$SSR = \underline{b}'x'\underline{y} - n\bar{y}^2$	$r-1$	$MSR = \frac{SSR}{r-1}$	$F = \frac{MSR}{MSE}$
Error	$SSE = \underline{y}'\underline{y} - \underline{b}'x'\underline{y}$	$n-r$	$MSE = \frac{SSE}{n-r}$	
Total	$SST = \underline{y}'\underline{y} - n\bar{y}^2$	$n-1$		

The null hypothesis to be tested for the adequacy of the regression model (Equation 7 or 9) is

$$H_0 : \underline{\beta} = \underline{0} \; versus \; H_1 : \underline{\beta} \neq \underline{0} \qquad (12)$$

H_0 is tested using the test statistic F = $\frac{MSR}{MSE}$
This has an F distribution with r-1 and n-r degrees of freedom. H_0 is rejected at the α level of significance if

$$F \geq F_{(1-\alpha;\; r-1,\, n-r)} \qquad (13)$$

Otherwise we do not reject Ho, where $F_{(1-\square,\ r\text{-}1,\ n\text{-}r)}$ is the critical value of the F distribution with r-1 and n-r degrees of freedom for a specified α level.

If H_0 is rejected indicating that not all β_j's are equal to zero, then some other hypotheses concerning β_j's may be tested.

Note that β_k; is interpreted in ordinal dummy variable regression model as the amount by which the dependent variable y on the average changes for every unit increase in x_k compared with x_{k-1} or one unit decrease in x_k relative to x_{k+1} when all other independent variables in the model are held constant. That is, β_k measures the amount by which on the average the dependent variable y increases or decreases for every unit change in x_K compared with a corresponding unit change in either x_{k-1} or x_{k+1} respectively when all other independent variables in the model are held constant.

Research interest may be in comparing the differential effects of any two ordinal dummy variables of a parent independent variable on the dependent variable. For example one may be interested in testing the null hypothesis

$$H_0 : \beta_{l;A} = \beta_{j;A} \text{ versus } H_1 : \beta_{l;A} > \beta_{j;A} \tag{14}$$

Where the β_d's are estimated from Equation (10) as b_d's for $l = 1, 2 \ldots a\text{-}1$; $j = 1, 2 \ldots a\text{-}1$; $l \neq j$.
The null hypothesis of Equation (14) may be tested using the test statistic

$$t = \frac{b_{l;A} - b_{j;A}}{se\left(b_{l;A} - b_{j;A}\right)} = \frac{b_{l;A} - b_{j;A}}{\sqrt{\left(\underline{C}\left(X'X\right)^{-1}\underline{C}^1\right)MSE}} \tag{15}$$

where $\underline{C}$ is an r row vector of the form (0, 0,…,1,0,…-1,0,…0)
Where 1 and -1 correspond to the positions of $\beta_{l;A}$ and $\beta_{j;A}$ respectively in the rx1 column vector $\underline{b}$ and all other elements of $\underline{C}$ are 0. H_0 is rejected at the α level of significance if

$$t \geq t_{(1-\alpha;\ n-r)} \tag{16}$$

Otherwise we do not reject H_0, where $t_{(1-\alpha;\ n-r)}$ is the critical value of the t distribution with n-r degrees of freedom for a specified α level.

In general several other hypotheses may be tested. For example one may be interested in comparing the effects of the *i*th level of factor A, say and the *j*th level of factor C say or of some combinations of some levels of several factors. Thus interest may be in testing.

$$H_0 : \beta_{l;A} = \beta_{j;C} \text{ versus } H_1 : \beta_{l;A} \neq \beta_{j;C} \tag{17}$$

Using the test statistic

$$t = \frac{b_{l;A} - b_{j;C}}{se\left(b_{l;A} - b_{j;C}\right)} = \frac{\underline{C}\underline{b}}{\sqrt{\left(\underline{C}\left(X'X\right)^{-1}\underline{C}^1\right)MSE}} \tag{18}$$

where $\underline{C}$ is a row vector as in Equation (15) except that 1 and -1 now occurs at the positions corresponding to the *i*th level of factor A and *j*th level of factor C in $\underline{b}$. H_0 is rejected as in Equation (16).

Further interest may also be in estimating the total or overall effect of a given parent independent variable through the effects of its representative ordinal dummy variables on the dependent variable. To do this it should be noted that any parent variable is completely determined by its set of representative ordinal dummy variables.

III. ESTIMATION OF EFFECTS

For the purposes of parameter estimation in dummy variable regression, the dummy variables of a parent independent variable are treated as intermediate (independent) variables between the parent variable and the dependent variable of interest [4][5]. Each dummy variable of the parent independent variable is then treated as a separate variable determined by its parent variable and determining the specified dependent variable. Therefore in developing an expression for the regression effect of a parent independent variable on a specified

dependent variable use is made of the simple effects of the set of dummy variables representing that parent independent variable on the dependent variable.
Now an equation expressing the determination of a dependent variable by the d-th dummy variable, x_d (d=1, 2... s-1), of a parent independent variable V with s levels may be written in the form:

$$\underline{y} = \beta_d \underline{x}_d + x^{(x)} \underline{\beta}^{+} + \underline{e} \tag{19}$$

In equation (19) y is an n-column vector representing the dependent variable, x_d is an n-column vector denoting the d-th set of ordinal dummy variables representing the parent independent variable V, for d=1,2,……,s-1.
$x^{(x)}$ is an n x (s-2) matrix of full column rank, s-2, representing all the other s-2 ordinal dummy variables for V. β_d is the regression effects of x_d on y, and β^+ is an s-2 column vector of the regression effects of $x^{(x)}$ on y. e is an n-column vector of uncorrelated error terms. Note that equation (3.16) may be expressed more compactly in the form

$$\underline{y} = X \underline{\beta} + \underline{e} \tag{20}$$

Where X = (x_d, $x^{(x)}$) and

$$\underline{\beta} = (X'X)^{-1} X' \underline{y} \tag{21}$$

Without loss of generality, take β_d as the first component of β. In ordinal dummy variable regression, β_d is a regression coefficient measuring the net effect of the d-th level of a parent independent variable, V, on a dependent variable, y, relative to the effect of the immediately succeeding level ((d-1) level) of V, after adjustments have been made for the effects of other variables in the regression model.
Now to estimate the direct effect [6], 1960[2], B_v, of a given parent independent variable V on a dependent variable y, the dummies of the parent independent variable V are treated as intermediate variables between V and y. Then, following the method of path analysis, [6][4][2], B_v is obtained as a weighted sum of β_d given as

$$B_v = \sum_{d=1}^{s-1} \alpha_d \beta_d \tag{22}$$

where the weights α_d is the simple regression coefficient of x_d on V which is subject to the constraint

$$\sum_{d=1}^{s-1} \alpha_d = 1 \tag{23}$$

The difference between the total effect, b_v, namely the simple regression coefficient of y on the parent independent variable V, and B_v, the effect of V on y through the variables x_d, is the indirect effect of V on y [6][2].

IV. FORECASTING

Conventionally, forecasting using regression methods where the parent independent variables are categorical and represented by codes is carried out using the method of least squares. In this case, the value of the dependent variable to be predicted or forecasted, for given values of the independent variables, is obtained by inserting the values of these independent variables represented by codes in the fitted regression model, where the independent variable is coded from 1 to n (n being the number of observations in ascending order of time, or the orthogonal coding system where the earliest in time is coded $-(\frac{n-1}{2})$, increasing arithmetically by 1 unit until the latest in time is coded $(\frac{n-1}{2})$, if the number of observations are odd, example is t = -3,-2,-1,0,1,2,3 if there are 7 observations, alternatively the codes will be -(n-1), -(n-3), -(n-5), … starting with the earliest in time to the latest in time, if the number of observations is even.
These codes are used as independent variables against the real dependent variable, y. This method suffers from the restriction of equal spacing of levels in the codes using normal dummy variable regression models plus the additional difficulty of interpreting the regression coefficients generated.

In cases where the parent independent variables are each represented by a set of dummy variables of 1's and 0's, the use of direct effects as the forecast model coefficients is advocated. These uses of direct effects as coefficient has taken care of the requirement and constraints of equal spacing and are interpretable, hence more useful for practical purposes.

The multiple regression model expressing the dependence or relationship between the dependent variable 'Y' and the parent independent variables A, B, C…etc represented by their respective sets of ordinal dummy variables is

$$Y_i = \beta_0+\beta_{1;A}X_{i1;A} + \beta_{2;A}X_{i2;A} + \ldots + \beta_{a-1;A}X_{i,a-1;A} + \beta_{1;B}X_{i1;B} + \beta_{2;B}X_{i2;B} + \ldots + \beta_{b-1;B} X_{ib-1;B} + \beta_{1;C}X_{i1;C} + \beta_{2;C}X_{i2;C} + \ldots+\beta_{c-1;C}X_{ic-1;C} + \ldots + e_i \quad (24)$$

Where $X_{ij;A}$ is the ordinal dummy variable representing the jth level of factor A with regression effect $\beta_{j;A}$, j=1,2,…a-1; $X_{ij;B}$ is the ordinal dummy variable representing the jth level of factor B with regression effect $\beta_{j;B}$, j=1,2,…,b-1; $X_{ij;C}$ is the ordinal dummy variable representing the jth level of factor C with regression effect $B_{j;C}$; j=1,2,…c-1; etc, and e_i is the error term uncorrelated with the X_{ij}'s and $E(e_i) = 0$ for i=1,2,…n.

The estimated value of equation (24) is

$$E(y_i)=\beta_0+\sum_{j=1}^{a-1} \beta_{j;A} E(X_{ij;A})+\sum_{j=1}^{b-1} \beta_{j;B} E(X_{ij;B})+\sum_{j=1}^{c-1} \beta_{j;C} E(X_{ij;C})+\ldots \quad (25)$$

Now define the regression effect of any parent independent variable A say, on the dependent variable 'y' through the effects of the set of ordinal dummy variables representing that parent independent variable A. We take the partial derivative of equation (25), the expected value of 'y' with respect to A obtaining

$$\frac{dE(yi)}{dA} = \sum_{j=1}^{a-1} \beta_{j;A} \frac{dE(X_{ij;A})}{dA} + \sum_{j=1}^{b-1} \beta_{j;B} \frac{dE(X_{ij;B})}{dA} + \sum_{j=1}^{c-1} \beta_{j;C} \frac{dE(X_{ij;C})}{dA} + \ldots$$

Where $\frac{dE(X_{ij;A})}{dA} = \alpha_{j;A}$ is the simple regression coefficient of $X_{ij;A}$ regressing on the parent independent variable, A ,with $\sum_{j-1}^{a-1} \propto_{j;A} = 1$ (Oyeka,1993) and $\frac{dE(X_{ij;Z})}{dA} = 0$, for all parent independent variables $Z \neq A$; Z = B,C, . . . so that

$$\frac{dE(y_i)}{dA} = \sum_{j=1}^{a-1} \beta_{j;A} \propto_{J;A} + 0$$

Hence the partial regression effect of the parent independent variable of factor A through the effects of its representative set of ordinal dummy variables on the dependent variable 'y' is

$$\beta_{y;A}=\sum_{j=1}^{a-1} \propto_{j;A} \beta_{j;A} \quad (26)$$

Whose sample estimate is

$$\sum_{j=1}^{a-1} \propto_{j;A} b_{j;A} \quad (27)$$

where $b_{j;A}$ is the sample estimate of the partial regression coefficients $\beta_{j;A}$, j= 1, 2, … , a-1.
Estimates of the partial effects of other parent independent variables in the model are similarly obtained.

V. ILLUSTRATIVE EXAMPLE

Mean monthly maximum temperature data, in degree centigrade, was collected by a research centre for five consecutive years (2007-2011) as shown in table 2 below.

Table 2: Mean Monthly Maximum Temperature (°C)

YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
2007	**33**	**35**	**34**	**34**	**32**	**31**	**29**	**30**	**31**	**31**	**33**	**32**
2008	**33**	**33**	**34**	**33**	**31**	**31**	**30**	**29**	**29**	**31**	**31**	**32**
2009	**34**	**35**	**35**	**32**	**32**	**30**	**30**	**29**	**30**	**30**	**31**	**32**
2010	**31**	**35**	**34**	**32**	**32**	**30**	**29**	**29**	**30**	**31**	**32**	**33**
2011	**33**	**34**	**34**	**33**	**33**	**31**	**30**	**29**	**30**	**31**	**32**	**34**

Applying equation 5 to obtain the design matrix, note must be taken that two independent variables are involved here, they are ‘year’ represented in the dummy design matrix as X_{id}, i=1,2,3,4 and ‘months’ , represented in the dummy design matrix as M_{id} , i=1,2,…,11. Note also that the design matrix will be obtained using the cumulative coding system proposed in equation 5 for both the years and the months respectively.

Table 3 below shows the cumulatively coded design matrix for the year and month variables. Note that one year (2011) was dropped. Note also that one month (Dec) was dropped. P_m and P_x are the parent independent variables for month and year respectively.

Table3: Cummulatively coded design matrix for the maximum temperature data

P_m	P_x	X_{11}	X_{12}	X_{13}	X_{14}	M_{11}	M_{12}	M_{13}	M_{14}	M_{15}	M_{16}	M_{17}	M_{18}	M_{19}	$M_{1,10}$	$M_{1,11}$
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
5	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0
6	1	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0
7	1	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0
8	1	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0
9	1	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0
10	1	0	0	0	0	1	1	1	1	1	1	1	1	1	0	0
11	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0
12	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
3	2	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0
4	2	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0
5	2	1	0	0	0	1	1	1	1	0	0	0	0	0	0	0
6	2	1	0	0	0	1	1	1	1	1	0	0	0	0	0	0
7	2	1	0	0	0	1	1	1	1	1	1	0	0	0	0	0
8	2	1	0	0	0	1	1	1	1	1	1	1	0	0	0	0
9	2	1	0	0	0	1	1	1	1	1	1	1	1	0	0	0
10	2	1	0	0	0	1	1	1	1	1	1	1	1	1	0	0
11	2	1	0	0	0	1	1	1	1	1	1	1	1	1	1	0
12	3	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1
1	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	3	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
3	3	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0
4	3	1	1	0	0	1	1	1	0	0	0	0	0	0	0	0
5	3	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0
6	3	1	1	0	0	1	1	1	1	1	0	0	0	0	0	0
7	3	1	1	0	0	1	1	1	1	1	1	0	0	0	0	0
8	3	1	1	0	0	1	1	1	1	1	1	1	0	0	0	0
9	3	1	1	0	0	1	1	1	1	1	1	1	1	0	0	0
10	3	1	1	0	0	1	1	1	1	1	1	1	1	1	0	0
11	3	1	1	0	0	1	1	1	1	1	1	1	1	1	1	0
12	3	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1
1	4	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
2	4	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0
3	4	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0
4	4	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0
5	4	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0
6	4	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0
7	4	1	1	1	0	1	1	1	1	1	1	0	0	0	0	0

8	4	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0
9	4	1	1	1	0	1	1	1	1	1	1	1	1	0	0	0
10	4	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0
11	4	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0
12	4	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
1	5	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
2	5	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
3	5	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
4	5	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
5	5	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
6	5	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
7	5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
8	5	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
9	5	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
10	5	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
11	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
12	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

For the direct effects, the dummy variables (x_{id} ; d=1, 2, 3, 4) as independent variables, are each regressed on p_x, the parent independent variable for year, using the method of least squares. Similarly, the dummy variables (m_{id};d=1, . . . ,11) as independent variables, are each regressed on p_m, the parent independent variable for month, by the same method of least squares. Table 4 below show the outputs of these regressions.

Table 4: Simple Regression Coefficients

i	MODEL	COEFFICIENT(α_i)
1	X_{11} on p_x	0.200
2	X_{12} on p_x	0.300
3	X_{13} on p_x	0.300
4	X_{14} on p_x	0.200
1	M_{11} on p_m	0.038
2	M_{12} on p_m	0.070
3	M_{13} on p_m	0.094
4	M_{14} on p_m	0.112
5	M_{15} on p_m	0.122
6	M_{16} on p_m	0.126
7	M_{17} on p_m	0.122
8	M_{18} on p_m	0.114
9	M_{19} on p_m	0.098
10	$M_{1,10}$ on p_m	0.070
11	$M_{1,11}$ on p_m	0.038

Note that the sums of the regression coefficients for each of the variables (year and month) sum up to 1. Recall that the direct effects for the years, B_x , is obtained as $B_x = \sum_{i=1}^{4} \propto_i \beta_i$ where β_i are the regression coefficients obtained from regressing the maximum temperature data (as dependent variable) on the cumulatively coded dummy variable design matrix (table 3 without p_x and p_m) and α_i are the simple regression coefficients related to the years x_{11}, x_{12}, x_{13}, x_{14} respectively as shown in table 4 above. $B_m = \sum_{1}^{11} \propto_i \beta_i$ is obtained similarly for months where α_i are the simple regression coefficients related to the months m_{11}, m_{12}, . . . , $m_{1,11}$ respectively as shown in table 4.

Table 5 below show the coefficients (β) obtained from the regression.

Table 5: Regression Coefficients from the regression of maximum temperature data on the cumulatively coded dummy variable design matrix

i	Dummy Variable	Regression Coefficient (β_i)	p-value
	Constant	33.174	0.000
1	X_{11}	-0.785	.012
2	X_{12}	0.368	.223
3	X_{13}	-0.167	.567
4	X_{14}	0.500	.098
1	M_{11}	1.600	.001
2	M_{12}	-0.200	.664
3	M_{13}	-1.400	.004
4	M_{14}	-0.800	.088
5	M_{15}	-1.400	.004
6	M_{16}	-1.000	.034
7	M_{17}	-0.453	.358
8	M_{18}	0.853	.087
9	M_{19}	0.569	.202
10	$M_{1,10}$	1.231	.008
11	$M_{1,11}$	0.800	.088

For the regression above, we have R^2 =0.877 and MSE = 10.979 with p-value=0.000, hence a significant multiple regression exist.

Direct Effects and Test of Significance

The direct effect for year, $B_x=\sum_{i=1}^{4} \propto_i \beta_i$, is obtained as $3.3\text{x}10^{-3}$.

Similarly, the direct effect for month is

$B_m = \sum_{i=1}^{11} \propto_i \beta_i$, is -0.257.

Testing significance of the direct effect for the years B_x , the t- statistic where

$$t_X = \frac{Bx}{\sqrt{var\,(Bx)}}$$

follows the student t distribution with n-p degrees of freedom, p is the number of parameters in the model.

$Var(B_X) = Var(\alpha'\beta_i) = \underline{\alpha_i}'Var(\beta_i)\underline{\alpha_i}$ for i= 1,2,3,4

and $Var(\beta_i) = MSE(X_d'\,X_d)^{-1}$ [1]

so $Var(B_x) = \underline{\alpha_i}'\,MSE(X_d'\,X_d)^{-1}\underline{\alpha_i}$ i=1,2,3,4

this yields $Var(B_x) = 0.167$

Hence

$$t_x = \frac{0.0033}{\sqrt{0.167}} = 0.008$$

For 60 - 4 = 56 degrees of freedom, a p-value of more than 0.746 was observed; this supports the hypothesis of no significant direct effect, due to the years, on the regression equation.

Similarly, for a significance test of the direct effect of the months, B_m, on the regression equation
$Var(B_m)$ is 0.001
Hence

$t_m = \frac{Bm}{\sqrt{Var\,(Bm)}} = \frac{-0.257}{\sqrt{.001}} = -8.031$

For n-p which is 60-11 = 49 degrees of freedom, a p-value of less than 0.0005 was observed. This supports a significant direct effect due to the months.

Forecasting

The forecast equation for maximum temperature is therefore of the form

$$\check{T}_{ym}=B_O+B_x t_{x_i}+B_m t_{m_i} \quad i=1,2\ldots12 \qquad (28)$$

Where B_o is the overall mean effect estimated by the constant term obtained as in table 5 (the value here is 33.174). B_x is the direct effect of year on maximum temperature obtained by the method of path analysis (value here is $3.3x10^{-3}$), this value is not statistically significant hence it will be dropped. B_m is the direct effect of months on maximum temperature, also obtained by the method of path analysis (value here is -0.256892), this direct effect is statistically significant.

[7] interpreted the path coefficients (the direct effects) thus: given a path diagram as below

A ———— θ = 0.81————>B

If region A increases by 1 standard deviation from its mean, region B would be expected to increase by 0.81 its own standard deviations from its own mean while holding all other relevant regional connections constant.
Our forecast equation here, using the direct effects as coefficients is

$$\check{T} = 33.174 - 0.2568t_{m_i} \qquad i=1,2,...,12 \qquad (29)$$

t_{m_i} is the serial count of the months from January (ie t_{m_i}=1) to December (t_{m_i}= 12) of each year.
If forecast for December 2012, say, is required, and t_{m_i} = 12, then

$$\check{T} = 33.174 - .257(12) = 30.09^oC$$

Similarly, if forecast for February 2013 is required, t_{m_i} = 2, then

$$\check{T} = 33.174 - .257(2) = 32.66^oC$$

VI. CONCLUSION

So far it has been demonstrated how one can make forecasts on a variable which can be arranged in any sequence, including time-order, using the multiple regression approach and the robust method of dummy variables, with the design matrix obtained by a cumulative coding arrangement. The direct effects, obtained through the method of path analysis, are used as parameter estimates, where they are significant.

This procedure can be viewed as having the ability to break down a single sequence of data into its hitherto not visible components, thus creating a let-in into the bits and pieces that make up the variable, and the relevant pieces reassembled to obtain a forecasting model for the variable.

REFERENCES

[1]. Neter, J; Wasserman, W; Kutner, M.H. (1983): Applied Linear Regression Models. Richard D. Irwin Inc. Illinois.
[2]. Oyeka, I.C.A. (1993): "Estimating Effects in Ordinal Dummy Variable Regression". STATISTICA, anno LIII, n.2 PP 261-8.
[3]. Boyle, R. P. (1970): "Path Analysis and Ordinal Data". American Journal of Sociology, 47, 1970, 461-480.
[4]. Lyon, M (1971):"Techniques for Using Ordinal Measures in Regression and Path Analysis", in Herbert Costner (ed), Sociological Methods, Josey Bass Publishers, San Francisco.
[5]. Wikipedia, the free encyclopedia (2010), "Path Analysis (Statistics)". "http://en.wikipedia.org/wiki/path-analysis-(statistics)". Modified 20 November, 2010".
[6]. Wright, S. (1960); "Path Coefficients and Path Regression: Alternative to Complementary Concept": Biometrics, Volume 16, pp 189 – 202.
[7]. Brian, P .(2010): "Structural Equation Modelling (SEM) or Path Analysis". http//afni.nih.gov.