Introduction to contextual analysis

Iversen, Gudmund R.

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Introduction to Contextual Analysis

Gudmund R. Iversen

ZUMA-Arbeitsbericht Nr. 87/08

Zentrum für Umfragen, Methoden und Analysen e.V. (ZUMA)
Postfach 5969
D-6800 Mannheim 1
Tel. (0621) 180040

Introduction

to

Contextual Analysis

Gudmund R. Iversen

Swarthmore College

This is a slightly expanded version of lectures given at ZUMA, Mannheim in September 1986. The lectures benefitted from the inputs of Hartmut Esser and Michael Wiedenbeck as well as from the many comments by the people who listened to the lectures.

CONTENTS

Introduction

The extent to which we are affected by our individual and group characteristics lies at the very heart of social science inquiry. The classical theorists were concerned with this question, and the interplay of individual and group concerns all the social sciences.

But we have lost track of the group in much of empirical research. In particular, survey research examines the isolated individual and does not take the context into account. The implication of the very word 'social' is that group membership is important, but this is more in theory and less in practice.

The reasons for the lack of prominence of the group in social research are both conceptual and practical. The concept of group itself is difficult. For example, I am a man. Do I do certain things because I am affected by that characteristic of me as an individual, or is it because I belong to the group of all men? After solving the question of what is meant by a group, we need to identify the relevant groups. As individuals we are members of an ever-changing set of overlapping groups through family, friends, work, residence, etc. These groups do not have well defined borders, and within a group we are more affected by those close to us than those farther away.

After identifying the group we need to measure the characteristic of the group. We may be interested in the level of a variable, like the mean test score or the percentage of people in a group that belong to a certain type. Alternatively, we may be interested in the homogeneity of a group as measured by the variance. More homogeneous groups have less tension, and this may be a relevant concept. It may also be that we are only interested in differentiating between different groups without measuring their characteristics, and that can be done through the use of dummy variables.

After deciding on the proper measurements, we have to collect data on the groups as well as on the individuals within the groups. Individual data are often collected through surveys, and these data must be merged with data on the groups, often census data.

Finally, after these hurdles have been cleared, the individual and group data must be analyzed by taking both levels of data into account. Such contextual analysis is discussed here. Only the most simple cases are discussed, and there is room for generalizations in many different directions. In the end we are only limited by our imagination and scope of the study.

The statistical equations we use for any analysis generally and contextual analysis in particular should reflect the real world process that generated the data in the first place. This means that we must understand this substantive process before we can write down any equations that are to be used as model for the process. In contextual analysis we work with concepts of effects of variables on the level of the individual as well as on the level of the group. It is important that we understand what is meant by effects of this kind, and there are here discussions of two different types of individual and group effects. They lead to different equations for the analysis of the data, and the choice between the two must be made on the substantive level, not on the statistical level.

This text gives an introduction to the main issues that come up in contextual analysis. One such issue is in what ways one variable can affect another variable, a second issue is how we can measure these effects, and a third issue deals with the possibility of recovering data on the level of the individual when we observe data on the level of the group. The text builds on the earlier work by Boyd and Iversen (1979). For other discussions of contextual analysis see, among others, Blalock (1984) and van den Eeden and Hüttner (1982).

Contingency tables

Contextual analysis began with the study of contingency tables. This is the situation we are in when we study the relationship between two categorical variables X and Y, and we have data on X and Y for observations belong to several different groups. Early treatments of contextual analysis with contingency tables are found in Kendall and Lazarsfeld (1950) and in Blau (1960).

Data on two categorical variables X and Y can be arranged in a contingency table, and when both variables have two categories 0 or 1, we get a table with frequencies n_{ij}:

		X = 0	X = 1	Sum
Y	1	n_{11}	n_{12}	$n_{1\cdot}$
	0	n_{21}	n_{22}	$n_{2\cdot}$
Sum		$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

From these frequencies it is possible to compute various proportions. Dividing the cell frequencies by the column totals gives the column proportions p_{ij} and the marginal row proportions $p_{i\cdot}$ as shown in the table:

		X = 0	X = 1	
Y	1	p_{11}	p_{12}	$p_{1\cdot}$
	0	p_{21}	p_{22}	$p_{2\cdot}$
Sum		1.00	1.00	1.00

In addition, let $p_{\cdot 1} = n_{\cdot 1}/n$ and $p_{\cdot 2} = n_{\cdot 2}/n$ be the two marginal column proportions.

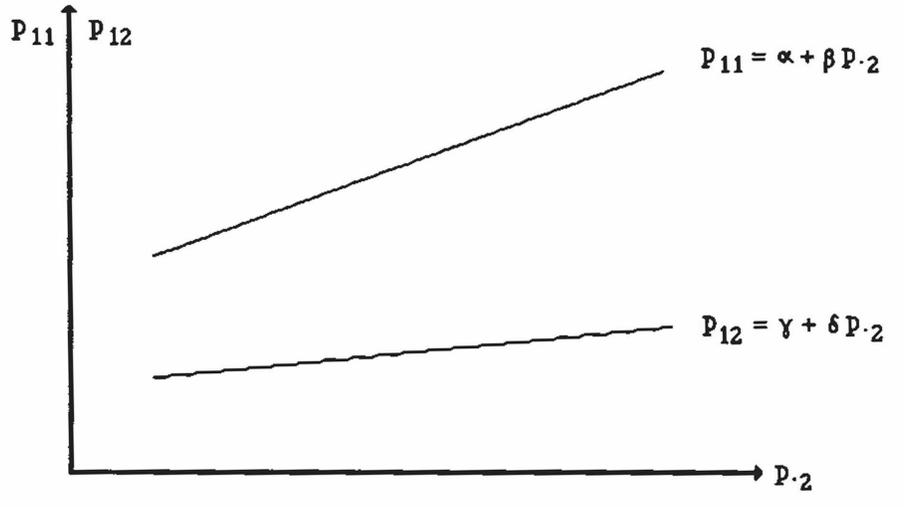
If p_{11} is different from p_{12}, then there is a relationship between X and Y in this table. That means that an individual's category of the independent variable X determines, in part, in what category the individual will fall on the dependent variable Y. We express this by saying that there is an effect of X on Y on the level of the individual. This can also be expressed by saying that there is an individual level effect of X on Y.

Next, suppose we have two tables, representing data from two different groups. Furthermore, suppose the two column p's are equal to each other in each of the tables, but that the values differ in the two tables. In that case, it makes no difference for Y which category of X an individual belongs to, but it does make a difference what group an individual belongs to. We express this by saying that there is a group effect on Y.

When we have tables for several groups, it is possible to plot the two column proportions p_{11} and p_{12} as functions of the marginal column proportion $p_{\cdot 2}$, as done by Davis et al. (1961). The marginal proportion is a group characteristic, telling us the proportion of individuals in the group with X equal to 1. The two column proportions tell us the

propensity of the individuals in the two columns to have the characteristic that Y is equal to 1. The purpose of this plot is to see if the propensity of having Y equal to 1 in any way is related to the composition of the group on the X variable.

Suppose both column proportions are related linearly to the marginal column proportion. These linear relationships can be expressed mathematically, as done by Iversen (1973), and shown in a figure like the one below:



Now there are several possibilities. If the two lines are both horizontal, then it does not matter for Y which group an individual belongs to. In that case group composition does not affect Y. If the lines are horizontal and the intercepts are different, then it matters for Y whether an individual is in the first or the second column, but not what group the individual is in. This is because p_{11} is the same in every group, and p_{12} is the same in every group. Thus, two horizontal lines are a sign that there is only an individual level effect of X on Y, and there is no group effect. On the other hand, if the two lines are the same and have a nonzero slope, then it only matters for Y what group an individual is in, not what column the individual is in. Since the lines overlap, the two p's are equal in any given table, and there is no individual effect present. But with a nonzero slope the common values of the p's differ from one table to the next, and it matters for Y what group a person belongs to. Thus, there is a group effect present. Finally, if the the two slopes are different, then there is an individual-group interaction effect in addition to the possible individual and group effects.

When the two column proportions are linearly related to the marginal proportion, it is possible to express these relationships as lines in a graph or equivalently as two linear equations. Let us consider the following equations:

$$p_{11} = \alpha + \beta p_{\cdot 2} \qquad\qquad p_{12} = \gamma + \delta p_{\cdot 2}$$

The Greek letters are the intercepts and slopes of the lines and are known as the parameters for the lines. With actual data we would not get proportions that lie exactly on two straight lines, and the analysis would require residual terms. The meaning of the parameters are discussed below, but first we want to rewrite the equations.

It is possible to analyze the relationship between X and Y in a contingency table with

two rows and two columns using dummy variables. The tables above show that the observations in the first column are assigned an X-value of 0 and the observations in the second column are assigned a 1. Similarly, the observations in the first row are assigned a Y-value of 1 and the observations in the second row are assigned a 0. Regressing Y on X within a group with these values of Y and X gives the equation

$$Y = p_{11} + (p_{12} - p_{11})\,X + e$$

Thus, the intercept equals p_{11} and the slope equals $p_{12} - p_{11}$.

Using the equations above we can express the slope and intercept in terms of the four parameters and $p_{\cdot 2}$ in the equations

$$p_{11} = \alpha + \beta p_{\cdot 2} \qquad\qquad p_{12} - p_{11} = (\gamma - \alpha) + (\delta - \beta)p_{\cdot 2}$$

When X is represented by a dummy variable with values 0 and 1 as we have done here, the mean of X equals the marginal proportion $p_{\cdot 2}$. Thus, the equations above show that when the column proportions are linearly related to the marginal column proportion, then the intercept and slope can be written as linear functions of the group mean.

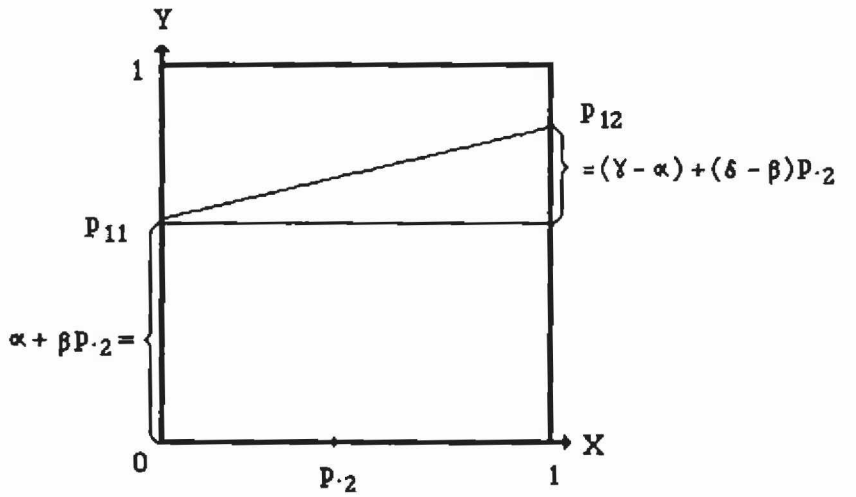
The values of the four parameters α, β, γ and δ determine the existence of the various effects. Only the individual effect of X on Y is present when $\beta = \delta = 0$ and $\alpha \neq \gamma$. In that case the intercept (p_{11}) is equal to the constant α and the slope ($p_{12} - p_{11}$) is equal to the constant $\gamma - \alpha$. The column proportions are the same in all the tables, but they are different from each other. By definition, this is the case of the individual effect of X.

There is only a group effect present when $\alpha = \gamma$ and $\beta = \delta$ ($\neq 0$). In that case the intercept is a linear function of the marginal proportion $p_{\cdot 2}$ and the slope within each group equals zero. Because this effect is produced by the marginal proportion $p_{\cdot 2}$, which is the mean of X, we say that this group effect is an effect of the variable X. Group effects can be due to other variables, not necessarily X, but here we discuss the case when the group effect is due to X.

Both individual and group effects are present when $\alpha \neq \gamma$ and $\beta = \delta$ ($\neq 0$). In that case the intercept vary from one group to the next, but the slopes are the same, meaning that the regression lines within the groups are parallel. Finally, there is also an individual-group interaction effect present when $\beta \neq \delta$. In that case the regression lines are no longer parallel.

It is possible to illustrate these equations, as shown in the graph below. The graphshows the scatterplot of the frequencies in one of the tables using X and Y as dummy variables. The scatterplot of X and Y has n_{11} observations at the point (0,1), n_{12} observations at the point (1,1), n_{21} observations at the point (0,0) and n_{22} observations at the point (1,0). Regressing Y on X gives the regression line which goes through the point p_{11} on the west side of the square and through the point p_{12} on the east side of the square. Thus, the regression line has intercept p_{11} and slope $p_{12} - p_{11}$.

Finally, the model specifies that both intercept and slope are linear functions of the marginal proportion $p_{\cdot 2}$. When the data have been generated by this model it is possible to study the relationship between X and Y in a set of 2x2 contingency tables in terms of individual, group and interaction effects. The analysis is performed by first computing the column proportions p_{11} and p_{12} as well as the marginal proportion $p_{\cdot 2}$ in each table. The next step consists of plotting both p_{11} and $p_{12} - p_{11}$ against $p_{\cdot 2}$. If the graphs reveal linear relationships, simple regressions can be used to estimate the parameters for the two lines and thereby establish the presence of the various effects.



The same results are obtained by regressing p_{11} and p_{12} against $p_{\cdot 2}$. But using the slope $p_{12} - p_{11}$ instead of p_{12} makes the analysis in this section more consistent with the analysis for two interval (metric) variables, as seen below.

The parameters can also be estimated in a different way. If the model is true and the column proportions are related linearly to the marginal proportion $p_{\cdot 2}$, then we can substitute the model equations for p_{11} and $p_{12} - p_{11}$ into the regression equation for the relationship between X and Y. This results in the equation

$$Y = \alpha + (\gamma - \alpha)X + \beta p_{\cdot 2} + (\delta - \beta)Xp_{\cdot 2} + \varepsilon$$

This equation suggests estimating the effect parameters using a multiple regression analysis on the data for all the individuals in all the groups. The dependent variable Y equals 0 or 1, depending upon whether an individual is in the first or the second row, and X is another dummy variable equalling 0 or 1 depending upon whether an individual is in the first or second column. The second explanatory variable is $p_{\cdot 2}$, the proportion of observations in the second column in a group, and every individual in a group has the same value of this variable. The last explanatory variable is the product $Xp_{\cdot 2}$, and it equals 0 or $p_{\cdot 2}$ depending upon whether an individual has $X = 0$ or $X = 1$.

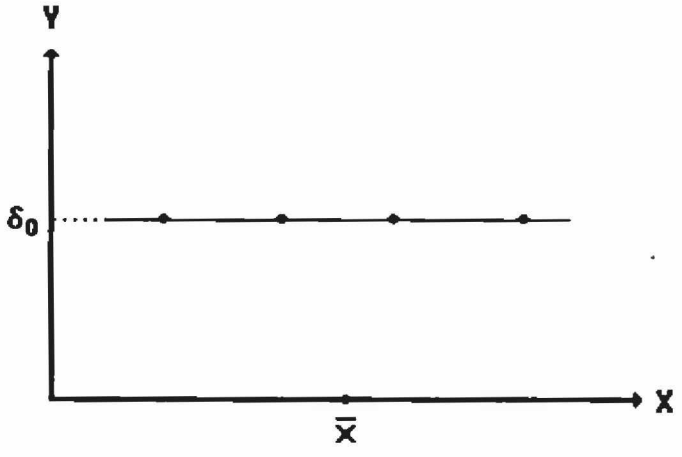
The model discussed here specifies that the intercepts and slopes within the groups are functions of the marginal proportion of X. But it is also possible to imagine other models for the intercept and slope where they are functions of other variables.

Meaning of effects

Let us examine what happens when a variable X causally determines another variable Y and that X is the only variable affecting Y. The variable X can then affect the other variable Y in different ways. One way is on the level of the individual and another is on the level of the group.

There is an individual level effect present when two individuals in the same group, but with different values of X, end up with different values of Y. Similarly, there is a group effect present when two individuals have the same value of X but belong to different groups, and they have different values of Y. In addition, there may be an individual-group interaction effect of X on Y.

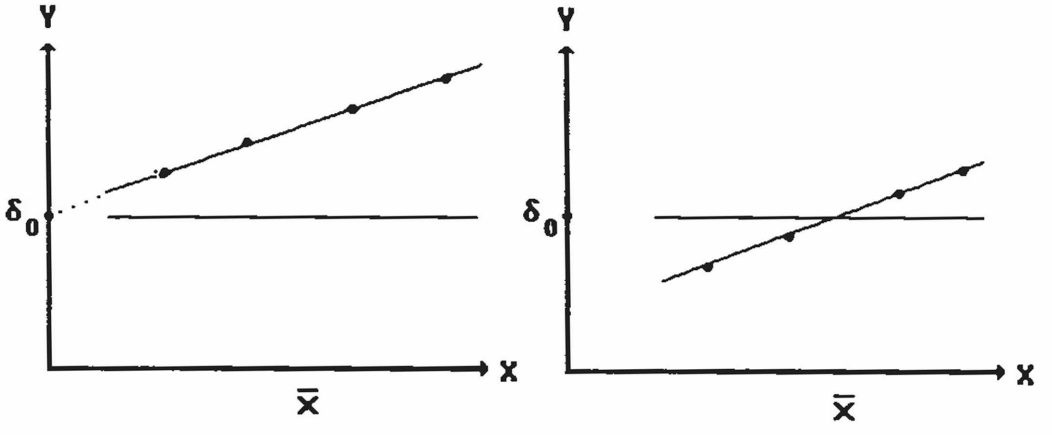
Let us first consider the individual effect of X on Y. When this effect is not present, then all individuals will have the same value of Y irrespective of what values they have on X. This can be illustrated in the graph below. The graph shows the X and Y values as pointsfor four individuals and the mean of X. When there is no individual level effect of X on Y, then the values of Y are the same for all individuals, no matter what their X-values are. This means that the data points lie on a horizontal line. Let this line have intercept denoted by δ_0.



When we introduce the individual level effect of X on Y, the data points no longer lie on a horizontal line. Let us assume that the relationship between X and Y is linear. The data points then lie on a line with nonzero slope. A major question becomes how the points moved when the individual effect was introduced. One possibility is that the points moved in such a way that the intercept for the new line is the same as the intercept for the old line. Another possibility is that the points moved in such a way that the new line pivoted around the mean point. These two possibilities are shown in the two graphs below.

In the graph on the left the new line has the same intercept as the horizontal no-effect line. In this case we say that the line is anchored at the intercept. Thus, the effect of X is such that all the points have moved up from where they were when X had no effect. In the graph on the right, the new line goes through the same mean point as the horizontal no-effect line. In that case we say that the line is pivoted around the balance point of the line. Here, the effect of X is such that the points to the left of the mean of X have moved down and the

points to the right of the mean have moved up.

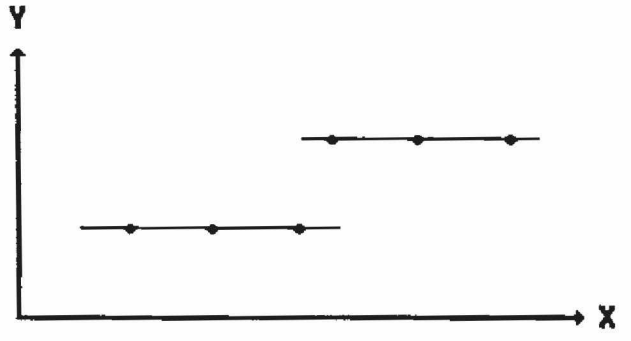


Thus, the individual level effect of X can be of the anchored type or the balanced type. It is possible to imagine other types as well, but we limit the dicussion to these two. The essential difference between the two have to do with how we look at the value of X for an individual. When we study in greater detail what are called the anchored and the balanced models, we find that the essential difference between them is how we regard the X values in the various groups. In the anchored model the magnitude of the effect of X is measured according to the actual, observed value. In the balanced model the effect of X depends upon the X-value relative to the group mean.

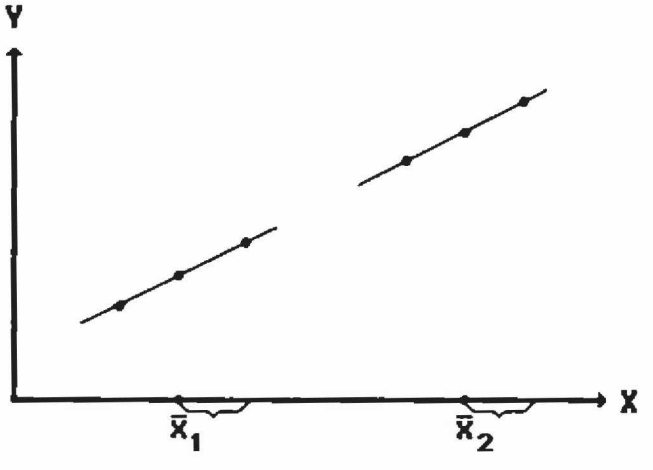
In the two hypothetical examples below there is an X variable with values from 1 to 9. The anchored model of individual level effects takes these values of X as they are. For example, a person with a score of 5 has a value which is 3 more than a person with a score of 2, and so on for any two people. What matters for the effect of X on Y is the actual value of X. If there is no group effect present, two people with the same score on X would be expected to have the same score on Y even if they are in different groups.

On the other hand, in the balanced model it is not the actual value of X that matters but the value of X relative to the group mean. In the balanced model an individual is measured against the group mean, and a value of 5 for a person in a group with mean 3 is the same as having a value of 9 in a group with mean 7. In both cases the individuals are 2 units above the mean, and that is what is important. The individual level effect of X will be the same for those two individuals since they are the same distance above their group mean, in spite of the fact that their original observed values of X are different.

The presence of a group effect can be illustrated with data from two groups. When there is only a group effect present the lines within each group are horizontal, but the lines are at different levels of Y. The lines are horizontal because there is no individual level effect of X. The group effect can be seen in the graph above. The graph shows three observations in each group, and within each group the values of Y are the same because of the lack of an individual level effect. The group effect is present because two individuals in different groups will have different values of Y, even if they have the same value of X.



It is now possible to combine both individual and group effects and look at them at the same time. Consider the graph below. The graph contains data from two groups, with three data points in each group. The data show that there is an individual level effect present, because the lines for the two groups have nonzero slopes. Within each group we find that the higher the value of X, the higher is the value of Y.

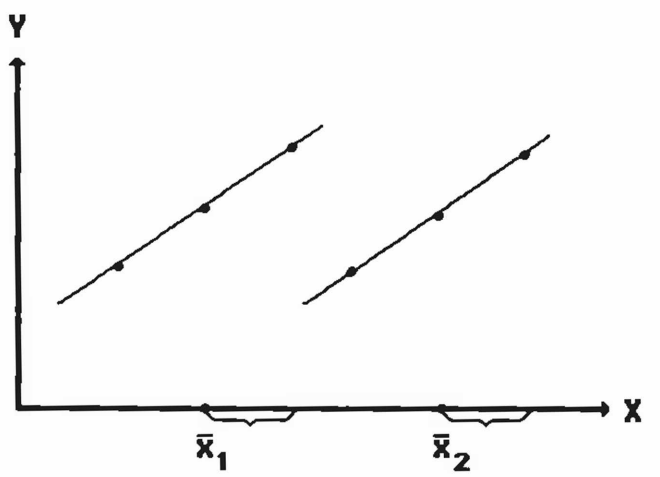


But is there a group effect present? One way to look at this graph is to say that we started with the graph containing the group effect only, and within each group the individual level effect was introduced by pivoting the line around the balance point. This is the way the individual effect is introduced in the balanced model. Thus, according to the balanced model the graph displays a group effect in addition to the individual effect. Another way to see that is in terms of prediction. If we need on the individual value of X in order to predict Y, then we only have an individual effect of X. But if we need the individual value of X and also what group an observation belongs to, then there is a group effect present in addtion to the individual effect. In the balanced model we use deviations from the group mean as values for the X-variable, and if we are told a value of say 2, we know that this individual has a value 2 larger than the mean of the group. In the graph above that may the observation to the right of the mean of the first group or the observation to the right of the mean of the second

group. In order to predict Y, we also need to know which group the observation belongs to.

Another way to look at this graph is to conclude that there is no group effect, only an individual effect. This is because the only thing we need to know in order to predict the Y-value is a person's X-value. There is no new knowledge contained in knowing whether the person is in the first or the second group, and this means there is no group effect. According to the anchored model, where we use the actual values of X, there is no group effect present. Thus, depending upon whether we measure X as deviation from the group mean or as the observed value, the data in the graph may or may not contain a group effect in addition to the individual effect.

The reverse situation occurs in the graph below. Because the lines through the data points have nonzero slopes, there is an individual level effect present in this graph. The more interesting question is whether the graph shows the presence of a group effect. One answer is yes, because if we control for the individual value of X by choosing two observations from different groups but with the same X-value, then they will have different Y-values. Thus, it does make a difference what group an observation belongs to. This is according to the anchored model where we use the actual X-values.



It is also possible to answer the question by a no. This is because it may have been that these lines first were horizontal and with the same intercept, in other words containing neither individual nor group effects. Then the individual effect was introduced by pivoting the lines around their mean points, and that left the lines where they are shown in the graph. Two individuals equally far above the means in their respective groups will have the same value of Y. Thus, for the balanced model there is no group effect present since it is only the position of X relative to the group mean that matters.

Anchored model

We have two variables X and Y, and we want to study the effect of X on Y. In this section we work with the individual level effect being of the anchored type. This is the model which forms the basis for the work by Boyd and Iversen (1979).

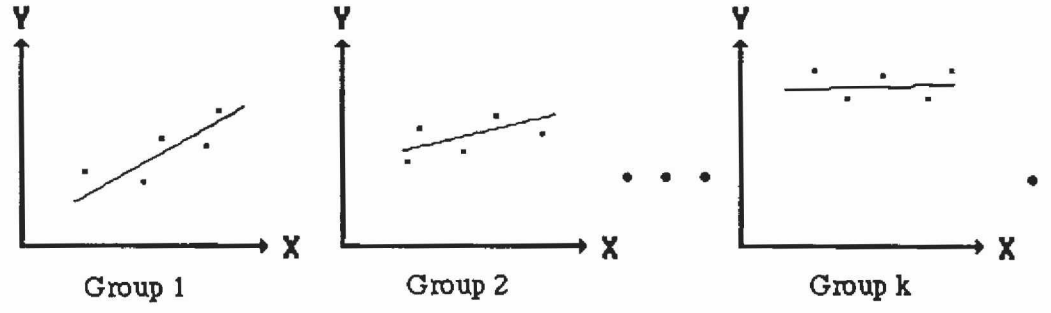
In addition to data on the two variables for a set of individuals, we also know that each individual belongs to one of several groups. Thus, the data matrix looks like

Y	X	Group
y_{11}	x_{11}	1
y_{21}	x_{21}	1
.	.	.
.	.	.
y_{ik}	x_{ik}	k
.	.	.
.	.	.

The second subscript on X and Y refers to the group number, and the first subscript refers to the individual within the group.

It is possible to make a scatterplot of X and Y for all the data, and if the relationship looks linear it can be analyzed using simple regression analysis. But it may be that the relationship between X and Y is different from group to group, and the overall relationship between the variables may not give a very good representation of how X and Y are related. To examine this possibility we control for group membership and make a separate scatterplot of X and Y for each group. The only way to find out if the groups are relevant is to actually break the data up into groups and find out if the relationships between X and Y are different in the various groups.

Suppose we get different scatterplots and relationships within the groups, as shown in the these plots:



The graphs tell us that the variables are not related in same way in each group. In this case the overall scatterplot does not give a good representation of the way in which X and Y are related. There seems to be something about these groups that influences the way in which the two variables are related. Thus, we should take the groups into account in the analysis.

When the relationships between X and Y are linear in the different groups, we can do a separate simple regression analysis for each group. The equations for these analyses can be expressed as

$$y_{ik} = \delta_{0k} + \delta_{1k}x_{ik} + \varepsilon_{ik}$$

Thus, we have a set of intercepts $\delta_{01}, \delta_{02}, \ldots, \delta_{0k}, \ldots$ and a set of slopes $\delta_{11}, \delta_{12}, \ldots, \delta_{1k}, \ldots$, one intercept and one slope for each group.

The next task is to study these different intercepts and slopes. There must be some reason why the intercepts differ and why the slopes differ, and we want to know what determines these intercepts and slopes. This question can be expressed in the <u>model</u> equations

$$\text{intercept } \delta_0 = \text{function of something}$$
$$\text{slope } \delta_1 = \text{function of something}$$

This raises two issues; 1) what kinds of functions do we have and 2) functions of what variables. The choices of functions and variables depend on the substantive problem.

We want to express these functions mathematically, and they could be of any kind. We usually start with the most simple mathematical functions, and this means linear functions. But it is important to realize that it may well be that more complicated functions will work better and that linear functions are only one of many possibilities.

The next question is what variables go into these equations. There is nothing in the data which tells what these variables should be, the choice depends entirely on the substantive nature of the study. The variables represent characteristics of one kind or another of the groups, and we can use either one or several variables. The variables may be categorical (nominal) variables represented by dummy variables, and in that case we get into analysis of covariance. For such an approach see Schuessler (1969). Perhaps more often the variables are means or proportions of some kind. As a special, but important case, we consider here the case where it is the group mean of X which determines the group intercepts and slopes. But, while this is an important case, it should be stressed that it represents only one of many possibilities.

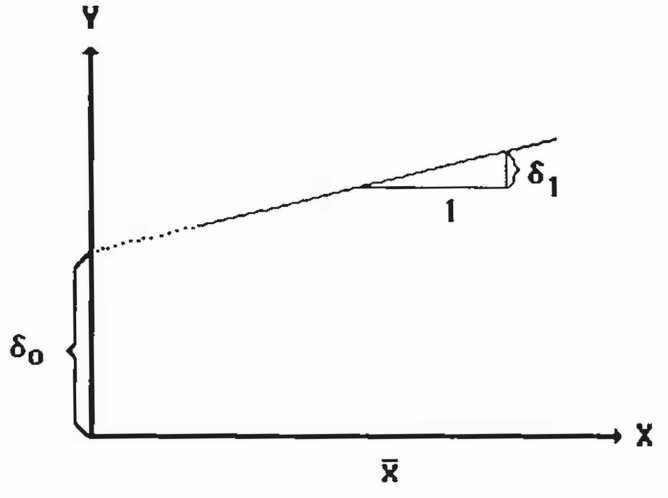
With the choice of linear functions and the group mean as the variable, the model equations can be written

$$\delta_{0k} = \alpha_0 + \alpha_2 x_k$$
$$\delta_{1k} = \alpha_1 + \alpha_3 x_k$$

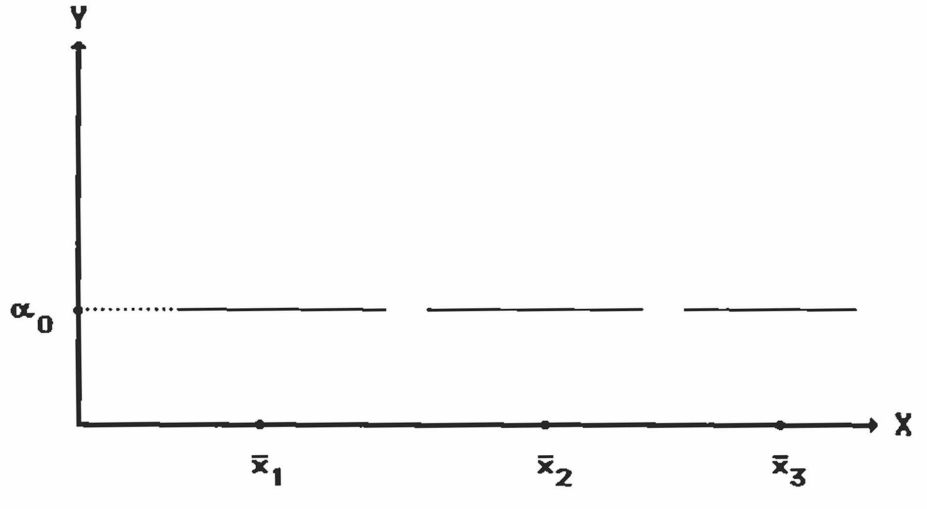
In the text and equations the mean of X in the k-th group is denoted x_k without a bar above the x while the bar is included in most of the graphs. This model states that the mean of X in group number k linearly determines the way in which X is related to Y in that group. The model is expressed using the four alphas as parameters. It is worth noting that this is a deterministic model in the sense that the group mean is the only variable which determines the intercept and slope. There are no residual terms in this model. It would be possible to introduce residual variables in the two model equations, but that would make the model more complicated and this should not be done unless there are substantive reasons for the inclusion of such residuals. As the model stands, the deltas are completely determined by the alpha parameters and the group means. If we replace the deltas in the model equations by their

estimates, then we have to introduce residuals in the model equations. This point is discussed further in the section on parameter estimation. This model is a direct generalization of the contextual model for contingency tables.

In this model the intercept of the within group lines are specified by the model, and we express this by saying that the lines are anchored on the Y axis by the model. The anchored model can be illustrated in the figure below. The figure shows how it is the intercept and slope in each group that are determined by the model equations.



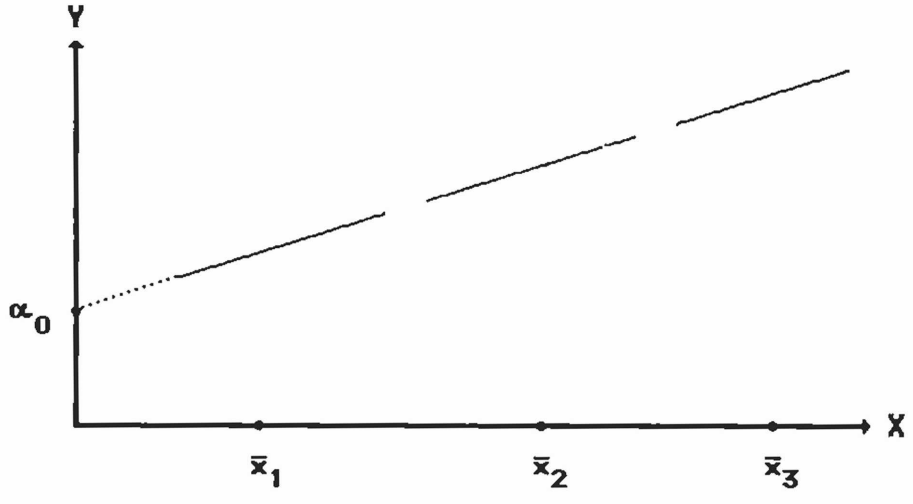
The four parameters α_0, α_1, α_2 and α_3 in the model equations determine the presence of various types of effects. In order to get a better understanding of the model let us examine certain combinations of parameter values.



$\alpha_0 \neq 0$, $\alpha_1 = \alpha_2 = \alpha_3 = 0$. In that case all the group intercepts are equal to the common value α_0 and all the slopes are equal to zero. This means that the lines for all the groups are horizontal and have the same intercept, as shown in the figure above. There is no

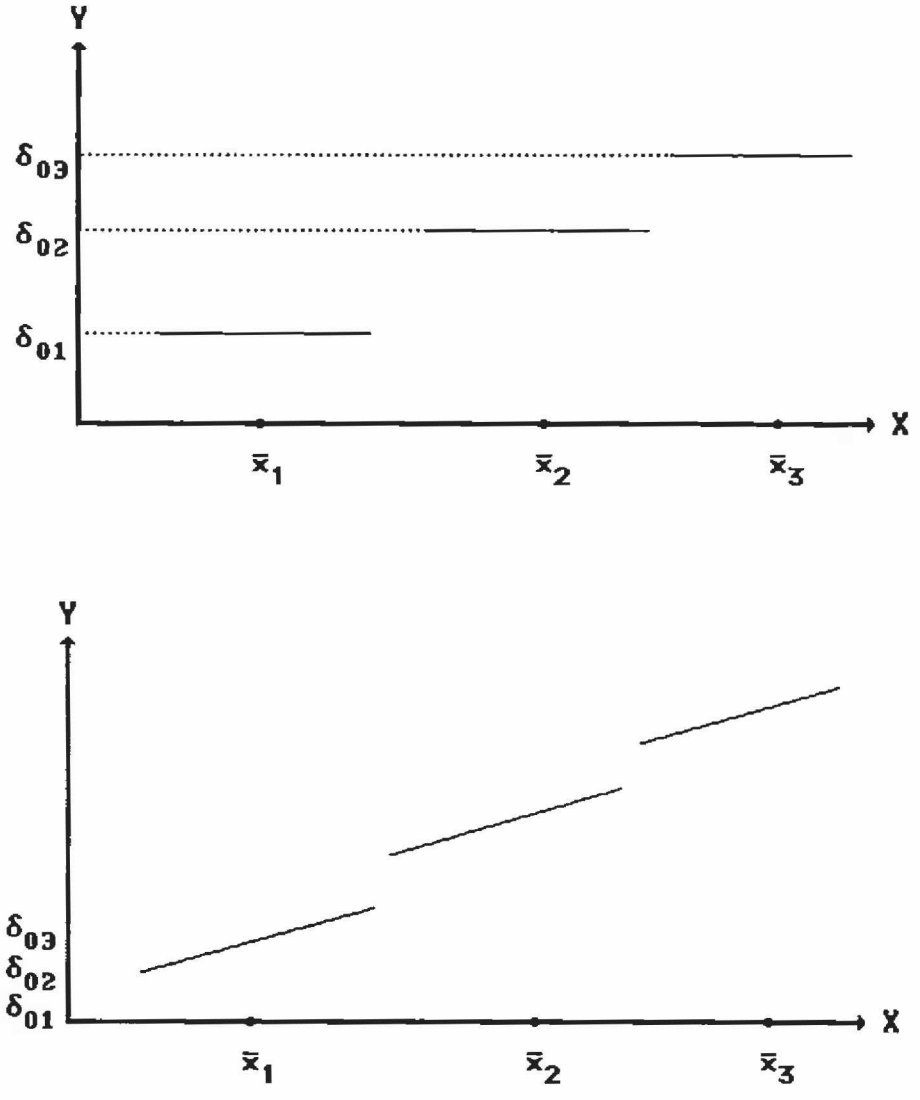
relationship between X and Y in any of the groups, since the regression line is horizontal for each group. The parameter α_0 sets the level of Y in the groups and serves no other purpose. In this case there is no individual level effect because of the horizontal lines, and there is no group effect because the lines have the same intercept.

$\alpha_0 \neq 0$, $\alpha_1 \neq 0$, $\alpha_2 = \alpha_3 = 0$. In this case all lines have the same intercept $\delta_{0k} = \alpha_0$ and the same slope $\delta_{1k} = \alpha_1$. The graph of the regression lines are shown in the figure below. The lines in the various groups may or may not overlap, depending upon the range of the X-values in each group. In this graph the lines do not overlap since it is easier to draw the graph using lines which do not overlap. Compared to the figure above with no effects, the three lines have swung upward from their anchoring point at α_0. The difference now is that the parameter α_1 is no longer equal to zero, and this means that the lines have nonzero slopes. Because the lines have slopes that are different from zero, there is an individual level effect present of X on Y. Thus, the parameter α_1 is a measure of the individual effect.

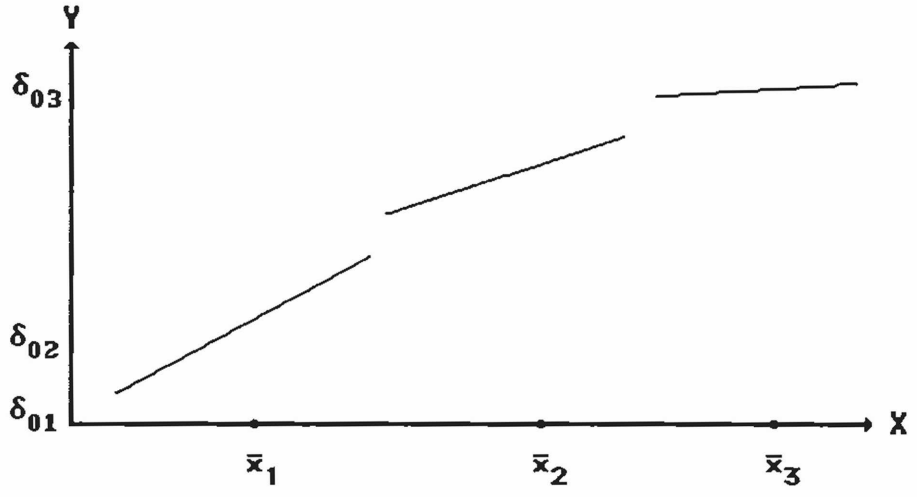


$\alpha_0 \neq 0$, $\alpha_1 = 0$, $\alpha_2 \neq 0$, $\alpha_3 = 0$. With these values of the parameters, the intercepts become linearly related to the group means while the slopes are all equal to zero. In this case the three regression lines for the different groups can be drawn as shown in the graph below. Within a group all the values of Y are the same, and the line is horizontal for each group. This means that there is no individual effect of X present here. But the intercepts are different, and that makes the level of Y different in the various groups. Thus, group membership affects Y, and this means there is a group effect present. According to our model, the intercepts are determined by the group mean of X, and the group effect here is therefore a group effect of the X variable. The intercepts are different because the parameter α_2 is different from zero, and this parameter becomes the measure of the group effect.

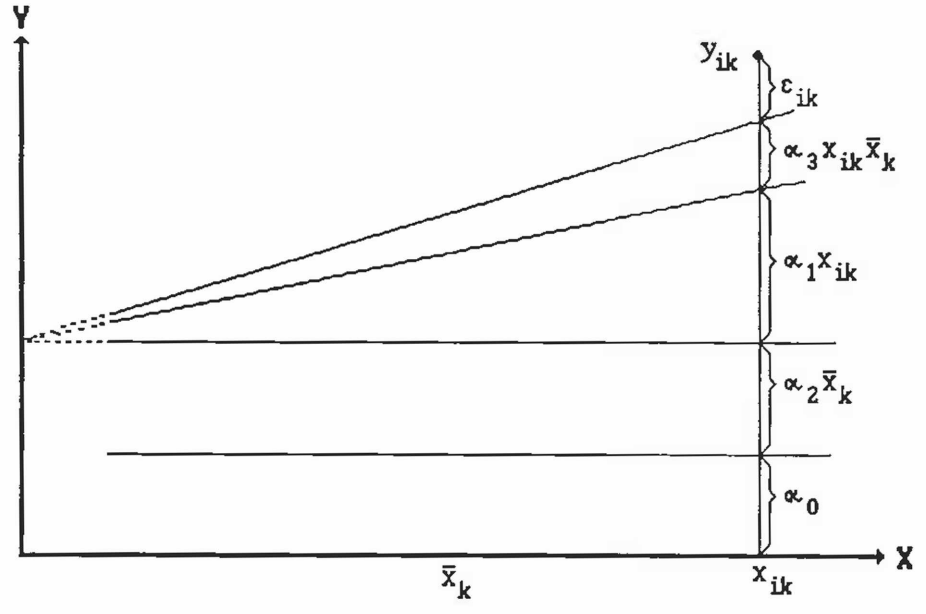
$\alpha_0 \neq 0$, $\alpha_1 \neq 0$, $\alpha_2 \neq 0$, $\alpha_3 = 0$. In this case the intercept is again a linear function of the group mean, but the slopes are nonzero and equal to the same value α_1. From this configuration of the coefficient we find that there is both an individual and a group effect present. The lines are parallel but with different intercepts, and they are shown in the graph below. The individual level effect is present because the lines are nonhorizontal, while the group effect is present because of the differing intercepts.



All α's different from zero. In this case both the intercepts and the slopes vary with the group means. Thus, the lines have different intercepts and they are no longer parallel since the slopes are different. Since the lines have nonzero slopes, there is an individual level effect of X present. Also, there is a group effect of X present, since the intercepts are different. What is new here is that the lines are no longer parallel. Such nonparallel lines indicates the presence of an individual-group interaction effect of X in addition to the other two effects. This effect is produced by the interaction parameter α_3. With the presence of the interaction variable a graph of three lines would look like the graph below. In this particular example the different intercepts are a positive function of the group mean since they increase with increasing group means. The slopes are such that in this case the larger the group mean the smaller the slope.



The complete anchored model is illustrated in the graph below. This graph reflects the process that generated the data under this model. An individual started with a Y-value of α_0, the group effect added an amount $\alpha_2 x_k$, the individual effect added an amount $\alpha_1 x_{ik}$, and the interaction effect added an amount $\alpha_3 x_{ik} x_k$. Finally, the residual variable added an amount ε_{ik}. These effects are added together to produce the observed value y_{ik}. The anchored nature of this model can be seen from the way the three top lines in the graph are anchored on the Y axis.



For usages of the anchored model see, for example, Hero and Durand (1985) and Knoke (1981).

Estimation of anchored parameters

There are two sets of parameters in the models discussed here. First, there are the deltas used to characterize the relationship between X and Y within each of the groups. Second, there are the alphas used to model how the deltas are determined, in the model discussed here as linear functions of the group means. In order to study the presence of individual, group and interaction effects in a set of data, we need to estimate both sets of parameters.

The deltas can be estimated by regressing Y on X within each group. Within the k-th group we have the estimated relationship between X and Y expressed in the equation

$$y_{ik} = d_{0k} + d_{1k}x_{ik} + f_{ik}$$

where the f's are the residuals. By regressing Y on X within each group the analyses result in the estimated intercepts $d_{01}, d_{02}, \ldots, d_{0k}, \ldots$ and slopes $d_{11}, d_{12}, \ldots, d_{1k}, \ldots$, one intercept and one slope for each group. These estimated deltas do not tell us anything directly about the existence of the various effects, but to the extent that intercepts and slopes differ from group to group we know that there are possible individual, group and interaction effects present in the data. It is therefore always important, as a first step, to examine the relationships between X and Y within the groups and see if they differ from group to group.

There are two main ways in which the effect parameters (alphas) can be estimated. The first method is known as the separate equations method and uses the group as the unit in the analysis. The second method uses the individual as the unit and is known as the single equation method.

Separate equations. The model equations specify that the group intercepts and slopes are linearly related to the group means, as expressed in the equations

$$\delta_{0k} = \alpha_0 + \alpha_2 x_k$$
$$\delta_{1k} = \alpha_1 + \alpha_3 x_k$$

We do not have the actual deltas for the group intercepts and slopes on the left sides of these equations. We only have the estimated deltas from the regressions within the group, but we can use these estimated deltas instead of the true deltas. The observed estimated intercept d_{0k} differs from the true parameter value δ_{0k} by some amount ζ_k; that is, $d_{0k} = \delta_{0k} + \zeta_k$. The same holds true for the slope, and we can write $d_{1k} = \delta_{1k} + \xi_k$. If we now substitute for the two deltas in the model equations, we get two equations where the observed intercepts and slopes depend on the alphas, the group means and the residuals. These equations can be used to estimate the alphas.

By regressing the observed intercepts and slopes on the group means we get estimates of the alphas. These two simple regression analyses can be expressed in the equations

$$d_{0k} = a_0 + a_2 x_k + u_k$$

$$d_{1k} = a_1 + a_3 x_k + v_k$$

where u and v are the estimated residuals. The regression coefficients in these two analyses

are estimates of the alphas.

The unit in this analysis is the group. This means that the separate equation method may not be a very good way to estimate the alphas if we only have a few group, since each of these simple regression analyses will be based only on a small number of observed data points. When the number of individuals vary a great deal across the groups, it may be better to use a weighted regression analysis in order for the larger groups to count more.

Single equation. By substituting for δ_{0k} and δ_{1k} from the model equations into the equation for the relationship between X and Y we get the equation

$$y_{ik} = (\alpha_0 + \alpha_3 x_k) + (\alpha_1 + \alpha_3 x_k)x_{ik} + \varepsilon_{ik}$$

Rearranging the equation it can be written

$$y_{ik} = \alpha_0 + \alpha_1 x_{ik} + \alpha_2 x_k + \alpha_3 x_{ik} x_k + \varepsilon_{ik}$$

This equation shows that when we have a linear relationship between X and Y within the groups and the group intercepts and slopes are modelled to be linear functions of the group means, then the value of Y for the i-th individual in the k-th group is a function of the individual value of X, the group mean of X and the product of the individual value of X and the group mean.

This equation can be used for estimation of the parameters through a multiple regression analysis with three explanatory variables. In order to perform this regression analysis two new columns must be constructed from the original data matrix. We already have individual values for Y and X, and we now need one column for the group variable and another column for the interaction variable. The column for the group variable is constructed by assigning the group mean x_k to every individual in the group. The column for the interaction variable is constructed by multiplying the individual and group columns. That way the data matrix looks like

Y	Ind.	Group	Int.
y_{11}	x_{11}	x_1	$x_{11}x_1$
y_{21}	x_{21}	x_1	$x_{21}x_1$
.	.	.	.
.	.	.	.
y_{ik}	x_{ik}	x_k	$x_{ik}x_k$
.	.	.	.
.	.	.	.

Regression Y on these three variables results in the estimated regression equation

$$y_{ik} = A_0 + A_1 x_{ik} + A_2 x_k + A_3 x_{ik} x_k + e_{ik}$$

where the A's are the estimated coefficients and the e's are the estimated residuals. The coefficients from this single equation analysis are denoted by capital A's in order to distinguish these estimates from the a's obtained from the separate equations estimation.

Ordinarily the two sets of estimated coefficients are different. Limited experiences from Monte Carlo studies indicate that the coefficients from the single equation are usually

closer to the true parameter values. Also, most often the A's have smaller standard deviations than the a's. The main reason for the smaller standard deviations is that the individual is the unit of this analysis, and the number of individuals is usually much larger than the number of groups. But at the same time, the explanatory variables in the multiple regression are constructed in such a way that they are correlated among themselves, and that tends to increase the standard deviations of the coefficients.

The various residuals are related in the following way. If we substitute for d_{0k} and d_{1k} into the relationship between X and Y in the k-th group, we get

$$y_{ik} = (a_0 + a_2x_k + u_k) + (a_1 + a_3x_k + v_k)x_{ik} + f_{ik}$$

or rearranged,

$$y_{ik} = a_0 + a_1x_{ik} + a_2x_k + a_3x_{ik}x_k + (f_{ik} + u_k + v_kx_{ik})$$

We know that the A's are the coefficients which give the best fit with the smallest residual sum of squares, and if we use any other set of coefficients we get a larger residual sum of squares. This means that we have the inequality

$$\Sigma\Sigma\, e_{ik}^2 \leq \Sigma\Sigma\, (f_{ik} + u_k + v_kx_{ik})^2.$$

Equality only occurs if the a's are equal to the A's.

Group means

The single equation with the relationship between Y and the individual, group and interaction variable represents one equation for each individual. It is possible to add the equations for all the individuals in a particular group and divide by the group size. On the left side we get the mean of Y in that group, and we get the following relationship between the means for X and Y,

$$y_k = \alpha_0 + \alpha_1 x_k + \alpha_2 x_k + \alpha_3 x_k^2 + \varepsilon_k$$

$$y_k = \alpha_0 + (\alpha_1 + \alpha_2) x_k + \alpha_3 x_k^2 + \varepsilon_k$$

This shows that when all three effects are present, the group means are related in a nonlinear way since the equation above contains the square of the group mean for X. If there is no interaction effect present, there is a linear relationship between the group means.

The last equation also shows that if we have only group data (group means), we cannot hope to obtain separate estimates for α_1 and α_2. The coefficient for the group mean can only estimate the sum of the two coefficients α_1 and α_2, and without additional data there is no way of untangling this estimate and get separate estimates for these two coefficients. The equation also shows that when we are using group data and perform a so-called ecological regression analysis, the coefficient for the group mean of x contains both the coefficient for the individual effect (α_1) and the coefficient for the group effect (α_2).

Within group relationships

It may be that we are particularly interested in the relationship between X and Y within each of the groups. One way to study this relationship within a particular group is to use the data in that group and regress Y on X. This will give the intercept d_{0k} and slope d_{1k} for the k-th group. With the contextual model it is possible to get better estimates of this intercept and slope. We can use the estimated alphas in the model equations and thereby estimate the intercepts and slopes in the groups.

If we use the A's from the single equation, the estimated slopes and intercepts can be found from the equations

$$\delta_{0k} = A_0 + A_2 x_k \qquad\qquad \delta_{1k} = A_1 + A_3 x_k$$

In these estimates we use information from all the groups to estimate the line in a particular group instead of just the information in that group. Computations like these are illustrated in the example below of the anchored model.

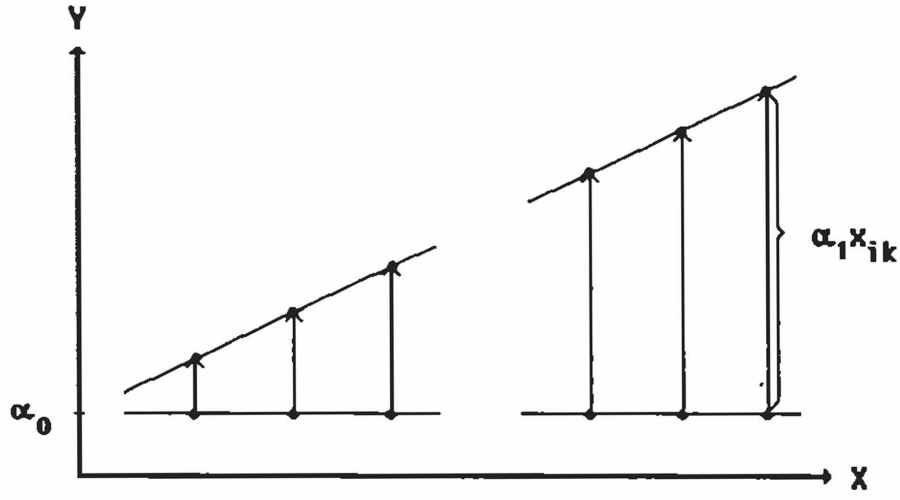
Measuring effects

Usually there is considerable interest in the question of how large the effects are on the dependent variable from the individual, group and interaction variables. The form of the effects can be seen in the regression coefficients, but these coefficients do not tell us how large the effects are of the corresponding variables. The main reason for this is that the three variables are measured in different units. In multiple regression, a common way around that difficulty is to change the coefficients to standardized coefficients in order to compare how large the effects are of the different variables. It is also common in multiple regression to look at the sums of squares for the different explanatory variables in the analysis as the effects of the variables. The main difficulty with this approach is that when the variables themselves are correlated, there is no way of getting a unique sum of squares for each variable. But it is not clear why squared deviations, as they are computed in multiple regression necessarily measure the effects of the variables the way we may want to measure effects.

Another way to look at the problem of measuring effects is to take a closer look at the process we believe generated the observed data in the first place. This process is mirrored in the model we are working with. This model is such that when there are no effects present, then the observations are equal to the common term α_0 plus a residual term. When there is only an individual level effect present, the observed values of Y are thought to be equal to

$$y_{ik} = \alpha_0 + \alpha_1 x_{ik} + \varepsilon_{ik}$$

For the i-th individual in the k-th group we see that the individual level effect adds a term $\alpha_1 x_{ik}$ to otherwise would have been the value of the dependent variable.

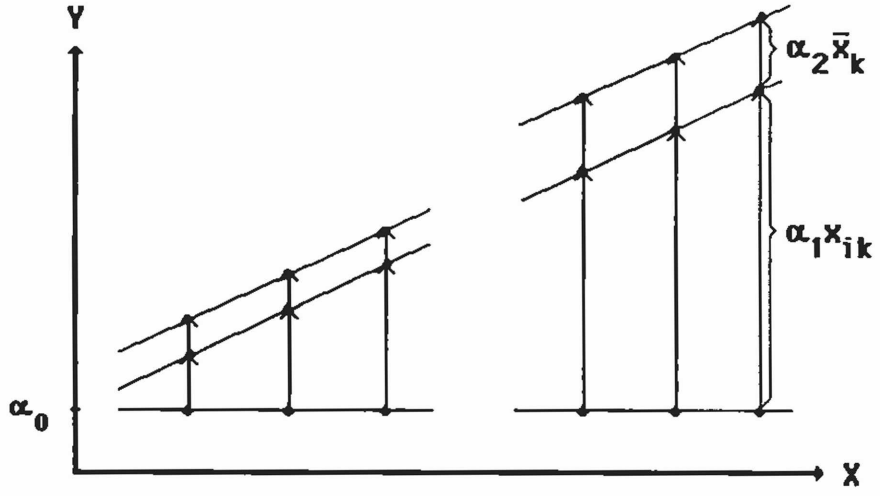


The workings of the individual level variable can be seen in the picture above. When there were no effects present, aside from the residual, the observations were located on the horizontal line with intercept α_0. When the observations were exposed to the individual level variable, the effect was to move the observations as marked by the arrows in the picture. The points in the scatterplot moved from the horizontal line up to a line with the same intercept

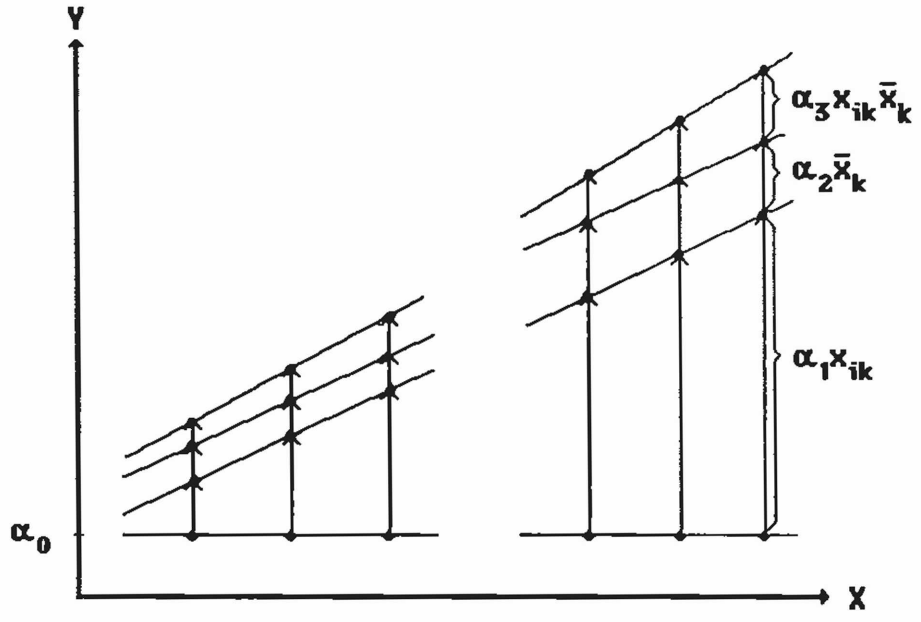
α_0 but with slope α_1. This means that an individual with X-value of x_{ik} moved a distance of $\alpha_1 x_{ik}$. That distance can be taken as the effect of the individual level variable on that individual. The effects are shown as vertical arrows on the graph.

One way to think of the overall effect of the individual level variable is to add up the total distance the points were moved by the individual level variable. The sum of the arrows in the graphs becomes $|\alpha_1|\Sigma\Sigma|x_{ik}|$. The absolute values signs are needed since we are only interested in how far the points moved, not whether they moved up or down. When the X-values are all positive, the sum of the absolute values becomes equal to the number of observations times the mean.

The presence of the group effect adds another $\alpha_2 x_k$ to the value of Y. This amount can be thought of as the effect of the group variable, and the effect is shown in the next graph. Within a group the line has been lifted a distance of $\alpha_2 x_k$. For each individual in the group that distance can be taken as the effect of the group variable. By adding up these distances we get the overall effect of the group variable, and this sum can be written $|\alpha_2|\Sigma n_k|x_k|$. Again, we need the absolute values since we are only interested in how far the points were moved by the group variable, not the direction in which they were moved. When the X-values are all positive, the sums involving the X's are the same for both the individual and the group effect. This means that the coefficients themselves for these two variables can be compared directly.



The interaction variable adds the value $\alpha_3 x_{ik} x_k$ to Y for the i-th individual in the k-th group. This is the distance an individual is moved on the scatterplot by the interaction variable, and the effect can be seen in the graph below. The observations are now located on nonparallel lines, and it was the interaction variable that made the lines have different slopes. The effect of the interaction variable can be taken as the sum of the distances travelled by the points due to the interaction variable. This sum can be written $|\alpha_3|\Sigma\Sigma|x_{ik}x_k|$. As before, the absolute values are needed, since we are only concerned with how far the points move, not the direction in which they move. Often, the X-variable has positive values only, and in that case we can disregard the absolute value signs for the X's.



Finally, the residual variable moves the points off the lines to where we actually find the observations when we plot the observed data. In the spirit of thinking of effects as distances moved, the effect of the residual variable is taken as the sum of distances from the lines to the points. That means the effect of the residual variable can be written $\Sigma\Sigma\,|\varepsilon_{ik}|$.

In the analysis of actual data these effects are found by replacing the regression coefficients and the residuals by the corresponding estimates obtained from the data.

Centering

The anchored model is based on the regression of Y on the individual, group and interaction variables as seen in the equation

$$y_{ik} = A_0 + A_1 x_{ik} + A_2 x_k + A_3 x_{ik} x_k + e_{ik}$$

The three estimated coefficients tell us about the presence of the individual, group and interaction effects, but they do not tell us much about the magnitudes of the effects.

One approach to the question of measuring effects in multiple regression is to partition the total sum of squares of the dependent variable into a regression sum of squares and a residual sum of squares. The regression sum of squares tells us how large the combined effect on Y is of the explanatory variables, and we would like to have this combined effect partitioned into separate effects for each of the individual, group and interaction variables. But in this anchored model the three explanatory variables are correlated among themselves, and with such correlated variables it is not possible to partition the regression sum of squares into unique components for each of the variables.

What we can do is analyze the effects of the variables sequentially. Let the regression sum of squares be denoted RegrSS(ind,gr,int) when we regress Y on all three variables. Next, let us regress Y on only the individual and group variables. This results in a new, and smaller, regression sum of squares which we can denote RegrSS(ind,gr). This sum is smaller because we are now using only the two explanatory variables (individual and group) instead of all three. Since the interaction variable is not included in this analysis, the reduction in the regression sum of squares must be due to the absence of the interaction variable. Thus, we can take this reduction, that is, the difference between the two sums of squares, as a measure of the effect of the interaction variable. That way the effect of the interaction variable is found as the difference RegrSS(ind,gr,int) - RegrSS(ind,gr).

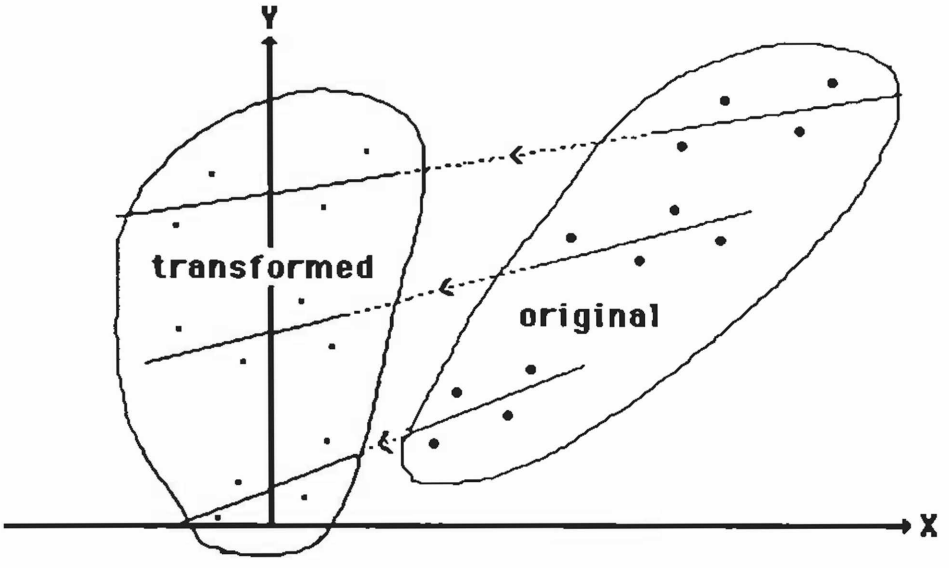
Similarly, if we regress Y on the individual level variable only, we get a regression sum of squares which can be denoted RegrSS(ind). The difference between the regression sum of squares obtained when we regress Y on both the individual and group variables and the one we get for the individual variable only must be due to the group variable. That is, the effect of the group variable can be found as the difference RegrSS(ind,gr) - RegrSS(ind). Finally, the effect of the individual level variable can taken simply as RegrSS(ind).

The major problem with this approach is the sequential nature of the analysis. The interaction effect is measured as the effect of the interaction variable, after the individual and group variables have been allowed to account for their effects. Similarly, the group effect is measured as the effect of the group variable after the individual variable has been allowed to account for its effect. But it would not be necessary to take the variables in this particular order with the individual variable first, then the group variable and finally the interaction variable. It may make sense to do the interaction variable last, since it is formed by the other two variables. But perhaps we should have taken the group variable first, and then measured the individual variable after the group variable. That would have given different effects for the individual and group variables from what we have above.

The lack of unique sums of squares for the three variables comes from the fact that they are correlated among themselves. In particular, since the interaction variable is constructed as a product of the individual and group variables, it is usually strongly correlated with the other two variables.

This raises the question whether it is possible to transform the data in some way and get unique measures of the effects of the three variables. The essential information about individual, group and interaction effects is contained in the intercepts and slopes of the lines for the various groups, and the transformation should be such that those intercepts and slopes are not changed. But it is possible to slide the data within a group along the direction of the group line and center the data around the Y-axis. This will change both X and Y coordinates, but the transformed X and Y values will give the same intercept and slope as the original data, thus preserving the information about the various effects. The advantage of the transformed data points is that the effect variables are no longer as correlated.

The transformation is illustrated in the graph showing original and transformed data points. The original data are shown as points on the right side, resulting in three regression lines, one for each group. These groups have different intercepts and different slopes, indicating that all three effects are present in these data. The transformation moves the original points within a group in such a way that the relative positions of the points in the group are maintained, and the intercept and slope of the new line is the same as for the old line. In addition, the transformation centers all the groups on the Y-axis in such a way that the transformed mean of X equals zero in each group.



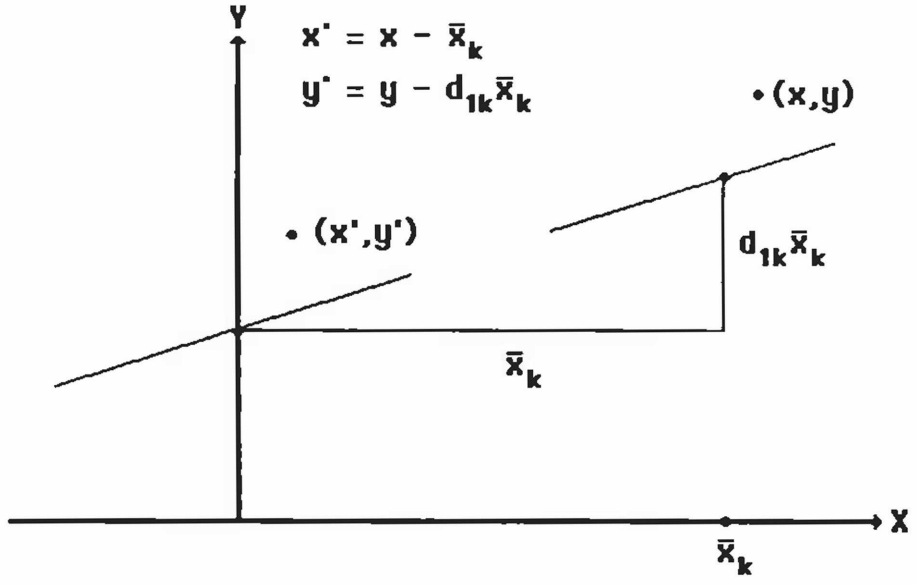
The coordinates of the transformed data points become

$$\text{new } x = \text{old } x - x_k$$

$$\text{new } y = \text{old } y - d_{1k}x_k$$

This transformation is illustrated in the graph below. The graph shows the regression line for the k-th group together with one observation with coordinates (x,y). This point is transformed to the new point with coordinates (x',y'). To see how the transformation works, we can consider the mean point in the middle of the original line. This point gets transformed to a point on the Y-axis in the middle of the new line. The transformation consisted of moving

a distance x_k to the left, meaning that we subtract an amount x_k from the X-value. The slope of the lines is d_{1k}, which implies that the mean point is moved down a distance $d_{1k}x_k$, meaning that we subtract this amount from the Y-value. All the points in the scatterplot are then moved the same way, a distance x_k to the left and a distance $d_{1k}x_k$ down. Thus, we subtract x_k from all the original X-values and $d_{1k}x_k$ from the original Y-



values. Let an original point be denoted (x_{ik},y_{ik}) and the transformed point (x'_{ik},y'_{ik}). These coordinates are then related according to the equations

$$x'_{ik} = x_{ik} - x_k$$
$$y'_{ik} = y_{ik} - d_{1k}x_k$$

After all the data points have been transformed this way, the new regression lines for all the groups are located in such a way that their mean points lie on the Y-axis, and the new lines have the same intercepts and slopes as the old lines.

It is now possible to examine how X and Y are related after the transformations. The original relationship was expressed in the equation

$$y_{ik} = \delta_{0k} + \delta_{1k}x_{ik} + \varepsilon_{ik}$$

When we substitute for the old coordinates of Y the equation can be written

$$y'_{ik} = \delta_{0k} + \delta_{1k}(x_{ik} - x_k) + \varepsilon_{ik}$$

This equation shows that the intercepts and slopes have not changed, and the estimated intercept is equal to the mean of the Y'-values.

We still model that the intercepts and slopes are functions of the group means. But it works better to express the model equations as deviations from the overall mean, denoted by

x, and the model is written

$$\delta_{0k} = \alpha'_0 + \alpha'_2(x_k - x)$$
$$\delta_{1k} = \alpha'_1 + \alpha'_3(x_k - x)$$

Compared to the earlier model we see that the parameters are related according to the equations

$$\alpha_0 = \alpha'_0 - \alpha'_2 x \qquad \alpha'_0 = \alpha_0 + \alpha_2 x$$
$$\alpha_1 = \alpha'_0 - \alpha'_3 x \qquad \alpha'_1 = \alpha_1 + \alpha_3 x$$
$$\alpha_2 = \alpha'_2 \qquad \alpha'_2 = \alpha_2$$
$$\alpha_3 = \alpha'_3 \qquad \alpha'_3 = \alpha_3$$

Finally, when we substitute for the model equation into the relationship between the transformed variables we get the single equation

$$y'_{ik} = \alpha'_0 + \alpha'_1(x_{ik} - x_k) + \alpha'_2(x_k - x) + \alpha'_3(x_{ik} - x_k)(x_k - x) + \epsilon_{ik}$$

The advantage of this single equation is that these explanatory have much less collinearity. The individual and group variables are always uncorrelated, as are the group and interaction variables. The group and interaction variables are uncorrelated when the variance of X is the same in all the groups.

When the three variables are uncorrelated, the regression sum of squares can be broken into three unique components and the effects of the three variables become

$$\text{Individual effect} = (\alpha'_1)^2 \Sigma\Sigma(x_{ik} - x_k)^2$$
$$\text{Group effect} = (\alpha'_2)^2 \Sigma n_k(x_k - x)^2$$
$$\text{Interaction effect} = (\alpha'_3)^2 \Sigma\Sigma[(x_{ik} - x_k)(x_k - x)]^2$$

For a single individual the effects are shown in the graph below. First, if there are no effects, then all the observations lie on the horizontal line with intercept α'_0. The group effect moves the k-th line a distance $\alpha'_2(x_k - x)$, and this distance is the group effect for each observation in that group. This determines the intercept of the line, and the individual effect pivots the line around this intercept from the horizontal position to a line with slope α'_1. For the i-th individual the individual effect moves the observation a distance of $\alpha'_1(x_{ik} - x_k)$. Finally, the interaction effect changes the line such that the slope becomes $\alpha'_1 + \alpha'_3(x_k - x)$, and that moves the observation as shown on the graph. After these effects the effect of the residual variable is added in, and we get the observed value y'_{ik}.

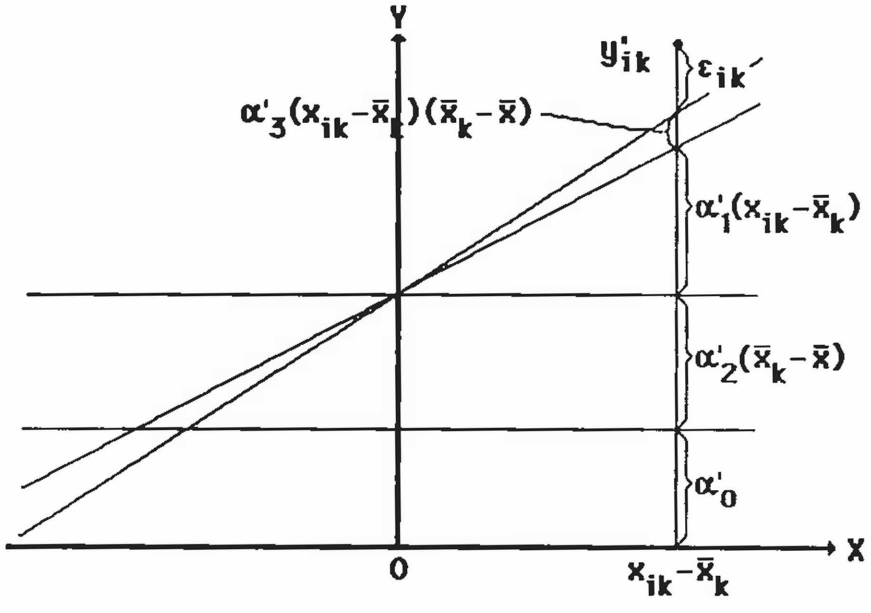
Even though we are accustomed to summing the squares of the distances in order to get the effect of a variable, it is also possible to think of the effect of a variable as the sum of the actual distances instead of their squares. In that case we get the following effects:

$$\text{Individual effect} = |\alpha'_1| \Sigma\Sigma|x_{ik} - x_k|$$

Group effect = $|\alpha'_2|\Sigma n_k|x_k - x|$

Individual effect = $|\alpha'_3|\Sigma\Sigma|(x_{ik} - x_k)(x_k - x)|$

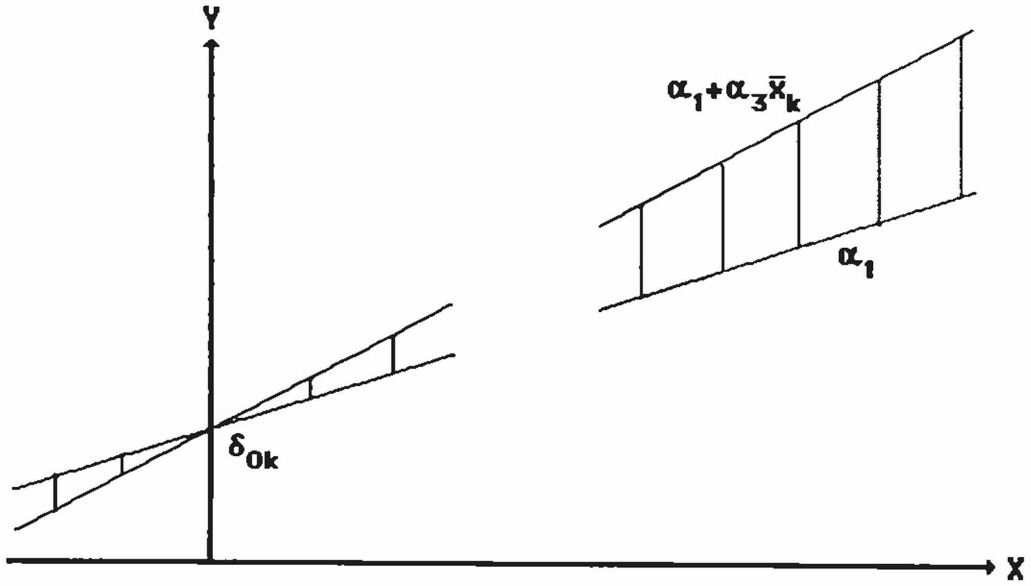
The absolute values are needed since we are only interested in how far a certain effect moved an observation, not the direction in which the observation was moved. In order to compute the estimated effects we replace the parameters above by their estimates.



The essential feature of the centering method is that it measures the effects by pivoting on the Y-axis, while the original lines are located with anchoring on the Y-axis. This is illustrated in the graph below, which shows the effect of the interaction variable. Without interaction being present, all groups have lines with the same slope α_1. The introduction of the interaction variable changes the slope in the k-th group from α_1 to $\alpha_1+\alpha_3 x_k$. The intercept for the group does not change, it is determined by the group effect and remains at δ_{0k}. The graph shows two lines, one without the interaction effect present and one with the interaction effect present. The effect of the interaction variable are shown in the vertical lines.

Centering moves the lines onto the Y-axis. In the centered space the effect of the interaction variable is also shown as vertical lines. But since the intercept is maintained, the effect of the interaction variable now amounts to a pivoting of the line around the mean point located on the Y-axis. Comparing the two sets of vertical lines, we see that the effect of the interaction variable takes different forms depending upon whether we look at the centered data or the original data.

The method of centering was first proposed by Boyd and Iversen (1979). For a critique of the this method see Tate (1985).


Y
$\alpha_1+\alpha_3\bar{x}_k$
α_1
δ_{0k}
X

Example anchored model

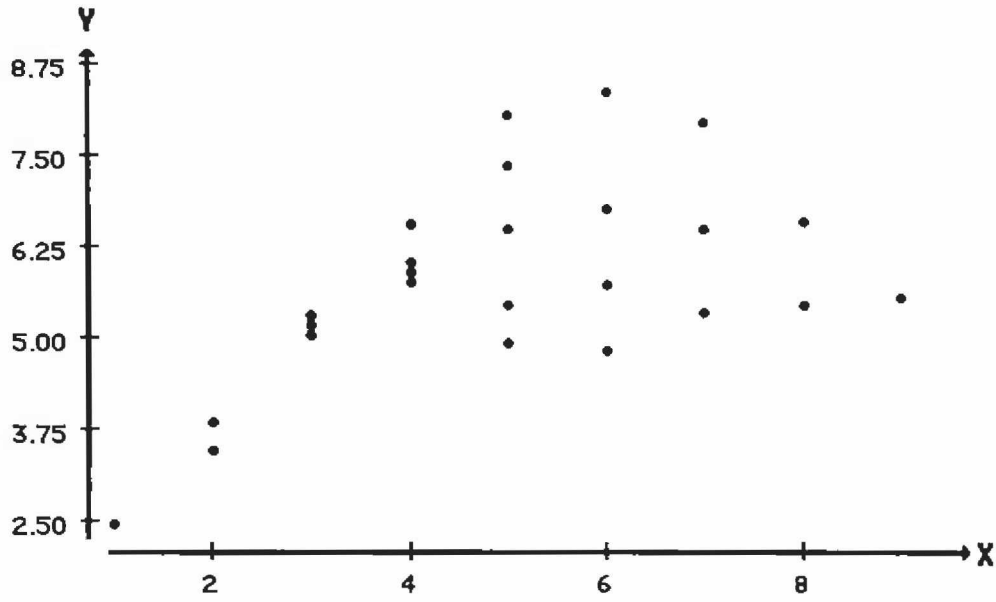
The table shows an example with data from five different groups. Each group contains five individuals, and observations with the same group mean are in the same group. The data are constructed according to the anchored model with the true parameter values $\alpha_0 = 0.00$, $\alpha_1 = 2.00$, $\alpha_2 = 0.50$ and $\alpha_3 = -0.25$.

y_{ik}	$x_{ik} - x_k$	$x_k - x$	$(x_{ik} - x_k)(x_k - x)$
2.42	1	3	3
3.44	2	3	6
5.28	3	3	9
6.52	4	3	12
8.02	5	3	15
3.80	2	4	8
4.99	3	4	12
5.88	4	4	16
7.32	5	4	20
8.31	6	4	24
5.13	3	5	15
5.90	4	5	20
6.45	5	5	25
6.72	6	5	30
7.92	7	5	35
5.72	4	6	24
5.41	5	6	30
5.68	6	6	36
6.44	7	6	42
6.54	8	6	48
4.89	5	7	35
4.78	6	7	42
5.29	7	7	49
5.39	8	7	56
5.51	9	7	63

If we were only given the first two columns in the data matrix; that is, if we were given the observations on the two variables X and Y without any regards for the groups, we might consider studying the overall relationship between those two variables by regressing Y on X. This analysis results in the regression line with the equation $y = 3.93 + 0.36x$. While the slope is significantly different from zero, the correlation coefficient for the relationship between X and Y is only $r = 0.54$.

In order to examine the relationship between the two variables more carefully we can make a scatterplot of Y on X. Such a scatterplot is shown below, and we find a somewhat unusual pattern of points. The relationship does not look linear, and Y has a larger variance for middle values of X than for extreme values of X. These patterns are both due to the fact that the data come from five different groups. When we identify what groups the observations belong to, it becomes clear that the relationship between X and Y is very different in

the different groups and this is the reason for the pattern found in the scatterplot.



Pursuing this point further, the next step consists of regressing Y on X within each of the five groups. These analyses give the following regression lines:

Group	d_0 d_1
1	$y = 0.85 + 1.43x$
2	$y = 1.52 + 1.14x$
3	$y = 3.22 + 0.64x$
4	$y = 4.36 + 0.27x$
5	$y = 3.88 + 0.18x$

These equations show that the intercepts range in values from 0.85 to 4.36 and the slopes from 1.43 to 0.18. Because the lines have such different intercepts and slopes, group membership should be taken into account in order to understand more fully the nature of the relationship between X and Y.

In the following we use the model which specifies that group intercepts and slopes are linear functions of the group mean. When we regress the observed intercepts and slopes on the group means, we get these estimates of the effect parameters:

$$d_{0k} = -1.68 + 0.89x_k \qquad d_{1k} = 2.41 - 0.34x_k$$
$$(1.08)\quad(0.21) \qquad (0.20)\quad(0.04)$$

The numbers in parentheses below the coefficients are the standard deviations of the coefficients, as estimated from the two separate equations. The coefficient $a_1 = 2.41$ shows that there is an individual level effect of X on Y, the coefficient $a_2 = 0.89$ shows that there is a group effect, and $a_3 = -0.34$ shows that there is an interaction effect present as well. All three effects are significantly different from zero.

Regressing Y on the individual, group and interaction variables gives us these single equation estimates of the effect parameters:

$$y_{ik} = \underset{(0.47)}{-1.65} + \underset{(0.10)}{2.34x_{ik}} + \underset{(0.11)}{0.88x_k} - \underset{(0.02)}{0.32x_{ik}x_k} \qquad R^2 = 0.97$$

From these estimates, $A_1 = 2.34$, $A_2 = 0.88$ and $A_3 = -0.32$ we see that all three effects are present. These effect parameters are also significantly different from zero, as seen from the standard deviations of the coefficients given in parentheses below.

We get the following sums of squares by regressing of the various variables:

RegrSS(ind,gr,int) = 44.38
RegrSS(ind,gr) = 26.72
RegrSS(ind) = 13.25

From these sums of squares we can compute the following measures of the effects of the three variables using sequential sums of squares:

Source	Sequential SS effect	Proportion
Individual	13.25	0.29
Group	26.72 - 13.25 = 13.47	0.30
Interaction	44.38 - 26.72 = 17.66	0.39
Residual	1.28	0.03
Total	45.65	1.01

According to these computations, the interaction variable is the most important while the individual and group variables have about the same effects.

Alternatively, since the sum of the X's equals 125, the effect of the individual level variable can be taken to be the product of the individual effect coefficient and the sum of the individual level variable. This product becomes (2.34)(125) = 293, and similarly for the other variables. This gives us the table of effects measuring sums of absolute distances shown below. From the table we see that the individual level effect is the largest, followed by the interaction and then the group effect.

Source	Abs.dist.effect	Proportion
Individual	293	0.47
Group	110	0.18
Interaction	217	0.35
Residual	5	0.01
Total	625	1.01

The effect of the residual variable is very small.The individual variable has the most effect, followed by the interaction variable and then the group variable.

Using the centering procedure we get the sums of squares and sums of absolute values as measures of the effects as seen in the table below. Looking at the effects as measured with centering we find the group variable to have the largest effect both when we use sums of squares and sums of absolute values, with the individual effect next and the interaction effect the smallest. Sums of squares and sums of absolute values do not give radically different re-

sults.

Source	Sum of sq.	Proportion	Sum abs. val.	Proportion
Individual	26.72	0.35	22	0.34
Group	36.72	0.48	26	0.40
Interaction	11.25	0.15	12	0.19
Residual	1.26	0.02	5	0.08
Total	75.94	1.01	65	1.01

The four ways of measuring individual, group and interaction effects give different results. If we look at the proportions measured the various ways we find:

Effect	Sequential SS	Abs. value	Centering SS	Cen.abs. value
Individual	0.29	0.47	0.35	0.34
Group	0.30	0.18	0.48	0.40
Interaction	0.39	0.35	0.15	0.19
Residual	0.03	0.01	0.02	0.08
Total	1.01	1.01	1.00	1.01

The four distributions are different, which is to to be expected since they measure different aspects of the data. The sequential sums of squares is the least satisfying way of measuring effects since it is so dependent upon the order in which the effects are measured, unless we have strong substantive reasons for a particular ordering of the variables. The methods using centering sums of squares and sums of absolute values are not much more than minor variations of each other. They are based on the same terms, and in one case they are squared while in the other case we take their absolute values. In order to use either of them we have to decide that centering makes substantive sense for our data. It may be that the sum of absolute values for the original data comes the closest to measuring what we mean by the various effects.

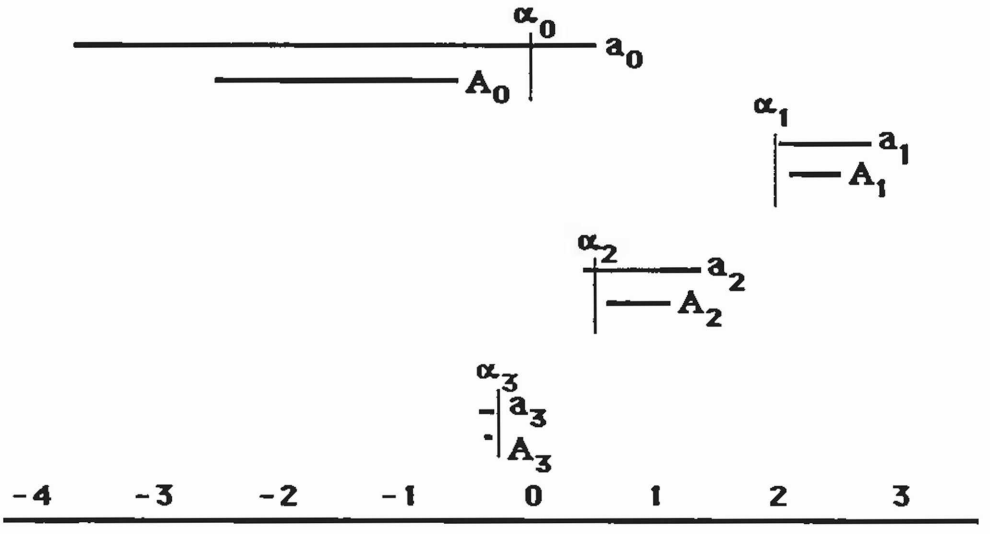
In this example the true values of the parameters are known, and that makes it possible to examine how well the estimation works. For the effect parameters we have the following results:

	Intercept	Individual	Group	Interaction
True values	$\alpha_0 = 0.00$	$\alpha_1 = 2.00$	$\alpha_2 = 0.50$	$\alpha_3 = -0.25$
Sep. est.	$a_0 = -1.68$	$a_1 = 2.41$	$a_2 = 0.89$	$a_3 = -0.34$
Single est.	$A_0 = -1.65$	$A_1 = 2.34$	$A_2 = 0.88$	$A_3 = -0.32$

Across all four parameters the estimates from the single equation are closer to the true values than the estimates from the separate equations.

When we take into account the standard deviations of the estimates and construct confidence interval, we find intervals as shown in the picture below. The intervals are illustrated by lines where the endpoints are two standard deviations away from the point estimate. For example, $a_1 = 2.41$ with a standard deviation $s(a_1) = 0.20$. Two standard deviations equals 0.40, and the line for a_1 is drawn from 2.01 to 2.81. The graph does not take into account that the a's are based on only 3 degrees of freedom while the A's are based on 21 degrees of freedom.

The single equation produces estimates with smaller standard deviations and thereby shorter confidence intervals, even without the adjustment for degrees of freedom. But even if the intervals are shorter, we see that in the case of the single equation none of the intervals contain the true values of the parameters.



It is also possible to compare the various estimates of the effects of the three variables when using sums of absolute values. The sum of the individual variable equals 125, the sum of the group variable also equals 125 and the sum of the interaction variable equals 675. Multiplying these sums by the absolute values of the parameters give us the effects of the different variables. These effects are seen in the table, with proportions in parentheses:

Variable	True effects	Single eqn. est.	Sep. eqn. est.
Individual	250 (0.51)	293 (0.47)	301 (0.46)
Group	62 (0.13)	110 (0.18)	111 (0.17)
Interaction	169 (0.35)	217 (0.35)	229 (0.35)
Residual	6 (0.01)	5 (0.01)	6 (0.01)
Total	487 (1.00)	625 (1.01)	647 (0.99)

Since the single equation estimates are closer to the true parameters than the separate equation estimates, it follows that the single equation effects are also closer to the true effects. Both methods overestimate the actual effects in this example, but at the same time both methods come close to finding the true proportional effects.

It is also possible to use the various estimated coefficients to study the relationship between X and Y within each group. When we substitute the estimated effect parameters into the model equations, we can use those equations to estimate the intercepts and slopes within the groups. We also have estimates of these intercepts and slopes from the regression analysis of Y on X within each group. We get the results for the five groups as shown in the table below.

Group	True rel.ship	Single eqn. est.	Sep. eqn. est	Within group
1	y = 1.50 + 1.25x	y = **0.99** + **1.38**x	y = **0.99** + 1.40x	y = 0.85 + 1.43x
2	y = 2.00 + 1.00x	y = 1.87 + **1.06**x	y = **1.88** + 1.07x	y = 1.52 + 1.14x
3	y = 2.50 + 0.75x	y = **2.75** + **0.74**x	y = 2.77 + 0.73x	y = 3.22 + 0.64x
4	y = 3.00 + 0.50x	y = **3.63** + **0.42**x	y = 3.66 + 0.40x	y = 4.36 + 0.27x
5	y = 3.50 + 0.25x	y = 4.51 + 0.10x	y = 4.55 + 0.06x	y = **3.88**+ **0.18**x

For each group there are three estimates of the line relating X and Y. Since we know the true values of the intercepts and slopes from the true values of the alphas, we can identify which estimation method gives the best estimates. The estimates that are closest to the true values are written out in bold numbers above. The comparison shows that the estimates found from using the A's from the single equation method are better in most of the cases. Only in the fifth group is it better to estimate the intercept and slope directly from the actual observations in that group. For the other groups we do better using the estimated alphas in the model equations for establishing the relationship between X and Y within the groups.

When we only have aggregate data available for the groups, meaning the group means of X and Y, we cannot estimate all the effect parameters. We can regress the group mean of Y on the group mean and square of the group mean of X. This gives the equation

$$y_k = -1.22 + 3.03x_k - 0.30x_k^2$$

where -1.22 is an estimate of α_0, 3.03 is an estimate of $\alpha_1 + \alpha_2$, and -0.30 is an estimate of α_3. These estimates compare well with the earlier estimates of the same parameters, but the major difference is that there is now no way of obtaining separate estimates of the alphas for the individual and group effects. We know that the sum of the two alphas is estimated to be equal to 3.03, but there are infinitely many ways two numbers can add up to a fixed sum. This is another example of how aggregate (ecological) data cannot be used alone to make conclusions about individual level relationships.

Balanced model

When the relationship between X and Y is linear within a group, this relationship can be expressed in the equation

$$y_{ik} = \delta_{0k} + \delta_{1k}x_{ik} + \varepsilon_{ik}$$

By adding and subtracting the term $\delta_{1k}x_k$ this equation can be rewritten in the form

$$y_{ik} = (\delta_{0k} + \delta_{1k}x_k) + \delta_{1k}(x_{ik} - x_k) + \varepsilon_{ik}$$

The sum in the first parentheses is constant within the k-th group, and to simplify the notation let $\delta_{0k} + \delta_{1k}x_k = \mu_{0k}$ and $\delta_{1k} = \mu_{1k}$. That way the relationship between the two variables can be expressed in the equation

$$y_{ik} = \mu_{0k} + \mu_{1k}(x_{ik} - x_k) + \varepsilon_{ik}$$

The only difference is that we now have subtracted the group mean of X from each of the X-values. This does not change the slope of the regression line, but the intercept becomes the Y-value when X is equal to the mean rather than when X is equal to zero.

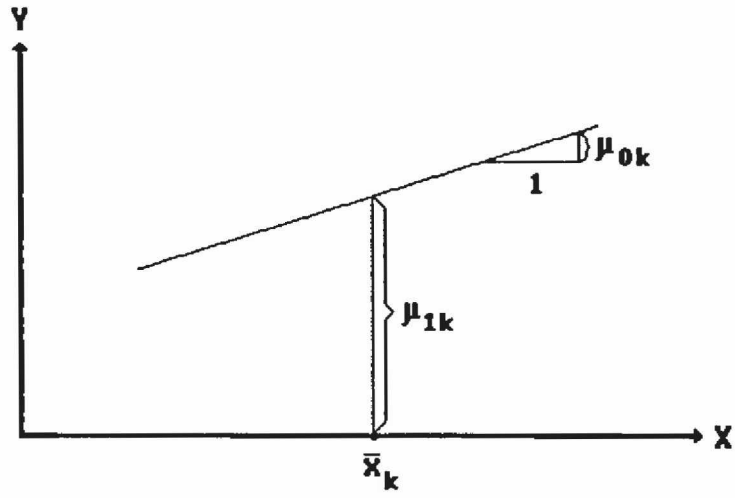
With many groups we have a set of intercepts $\mu_{01}, \mu_{02}, \ldots, \mu_{0k}, \ldots$ and a set of slopes $\mu_{11}, \mu_{12}, \ldots, \mu_{1k}, \ldots$, one intercept and one slope for each group. Again, if the intercepts and slopes vary across the groups, then there must be something about the groups that affect the way X and Y are related in the different groups. We want to determine why they differ across the groups, and that can be expressed in the <u>model</u> equations

$$\text{intercept } \mu_{0k} = \text{function of something}$$
$$\text{slope } \mu_{1k} = \text{function of something}$$

One possibility is that the intercepts and slopes are both linear functions of the group means. This is only one of many possibilities, just like the linear function used for the anchored model as an important but only one of many models. The linear model with the group mean can be expressed in the equations

$$\mu_{0k} = \beta_0 + \beta_2(x_k - x)$$
$$\mu_{1k} = \beta_1 + \beta_3(x_k - x)$$

The model is expressed using four betas as parameters. The model looks very much like the model we use for the anchored model, but for reasons discussed below we subtract the overall mean of X from the various group means. This is also a deterministic model, in the sense that there is no residual term in either of the equations. But the mus and betas are all unknown parameters, and a deterministic model may well be appropriate. Residuals enter the analysis when we replace the unknown mus on the left sides by the estimated intercepts and slopes from the groups.



In this model the center balance point of each line is specified by the model, and we express this by saying that we have a balanced model. The balanced model is illustrated in the figure above. The graph shows that it is the balance point and slope in each group that is determined by the model, and this is what gives the model its name.

The four parameters β_0, β_1, β_2 and β_3 in the model equations determine the presence of the various types of effects. In order to get a better understanding of this model, let us examine certain combinations of parameter values.

$\beta_0 \neq 0$, $\beta_1 = \beta_2 = \beta_3 = 0$. In that case $\mu_{0k} = \beta_0$ and all the slopes are equal to zero. Each group line has the same height above the X axis, and all the lines are horizontal. When we here substitute for the model back into the equation for the relationship between X and Y in each group, we get the equation

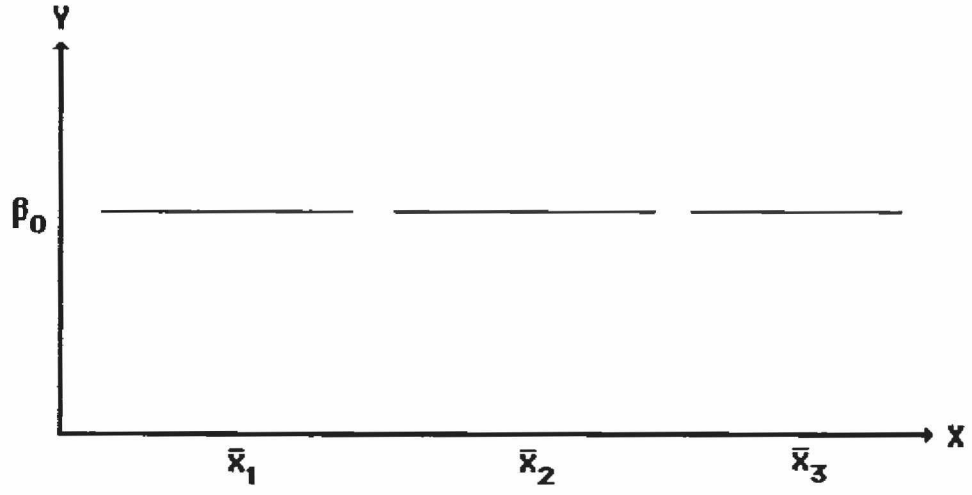
$$y_{ik} = \beta_0 + 0(x_{ik} - x_k) + \varepsilon_{ik} = \beta_0 + \varepsilon_{ik}$$

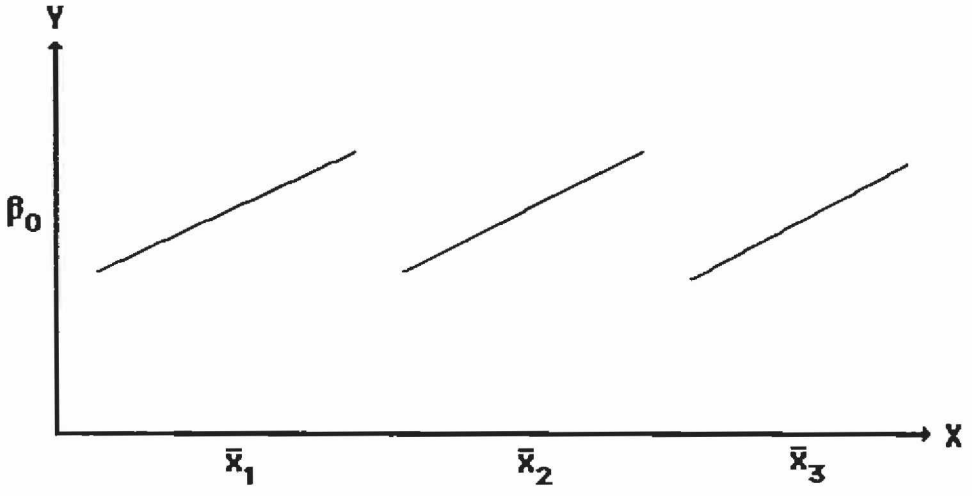
This says that aside from the epsilon residuals, all the observations in all the groups are the same and equal to β_0 in this case. The group lines are shown in the graph above. The graph shows the horizontal lines, all with intercept β_0. There is no relationship between X and Y in each of the groups, and the level of Y is the same for all the groups. In this case there are no effects of X on Y.

$\beta_0 \neq 0$, $\beta_1 \neq 0$, $\beta_2 = \beta_3 = 0$. In this case $\mu_{0k} = \beta_0$ and $\mu_{1k} = \beta_1$. When we substitute this model back into the relationship between X and Y in the k-th group, we get the relationship expressed in the equation

$$y_{ik} = \beta_0 + \beta_1(x_{ik} - x_k) + \varepsilon_{ik}$$

This equation says that all the lines have the same slope β_1, and when X is equal to the group mean then Y is equal to the same value β_0 aside from residuals. This model is illustrated in the second graph below.



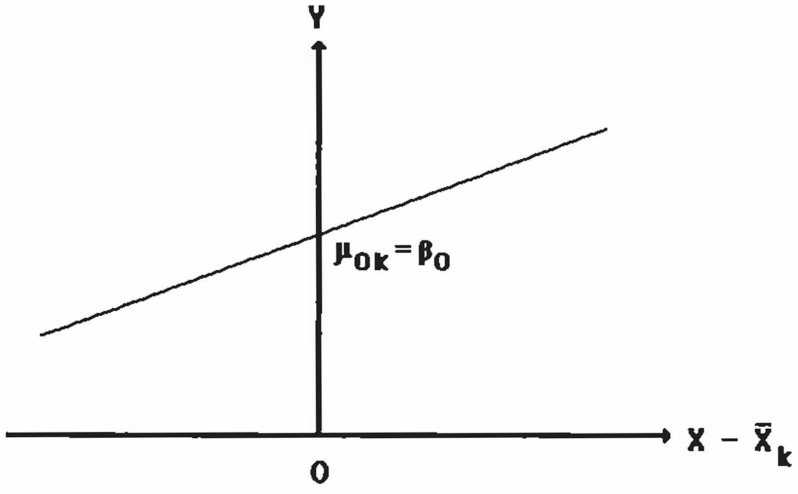


The lines all have the same slope β_1 and the level of Y is the same in all the groups. In particular, when X is equal to the group mean then the value of Y equals β_0. In this case there is a relationship between X and Y, since the slopes are different from zero, and X has an effect on Y. This is the effect of X on the level of the individual. Thus, the coefficient β_1 represents the individual level effect of X.

When we know that the model containing only the individual level effect is the correct one to use, we can regress Y on X as specified above. It can then be shown that the estimate of the parameter β_1 becomes the average of the slopes within the groups. Since the model specifies that the groups have the same slope, it is not surprising that the best estimate of this common slope is the mean of the group slopes.

It is also possible to illustrate this case in a slightly different way. The values of the independent variable are found by subtracting the group mean from every observation. Within each group the mean is therefore replaced by the value 0, all the observations smaller

than the group means will have negative differences and all the observations larger than the group means will have positive differences. This way the groups will be superimposed upon each other, and the three lines in the graph above will be superimposed as well since they have the same intercept and the same slope. The three lines are shown as one line in the graph below.



From this graph we see that what matters for an individual is where the value of X is located relative to the group mean. Two individuals in different groups but with the same distance to the group mean are thought of as having the same X-value. Thus, it is not the observed value of X that matters, but where the individual is located in the group relative to the group mean. The effect on Y of being a certain distance below the group mean on X is the same no matter what the actual value is of X, when we use the balanced model. This relative feature of the balanced model is what distinguishes this model from the anchored model where it is the actual value of X that influences Y.

This graph also shows that it does not matter for Y what group an individual is in, since the lines overlap. The Y value is determined only by the rescaled X value of the person, which means that in this case we have only an individual and not a group effect of X on Y.

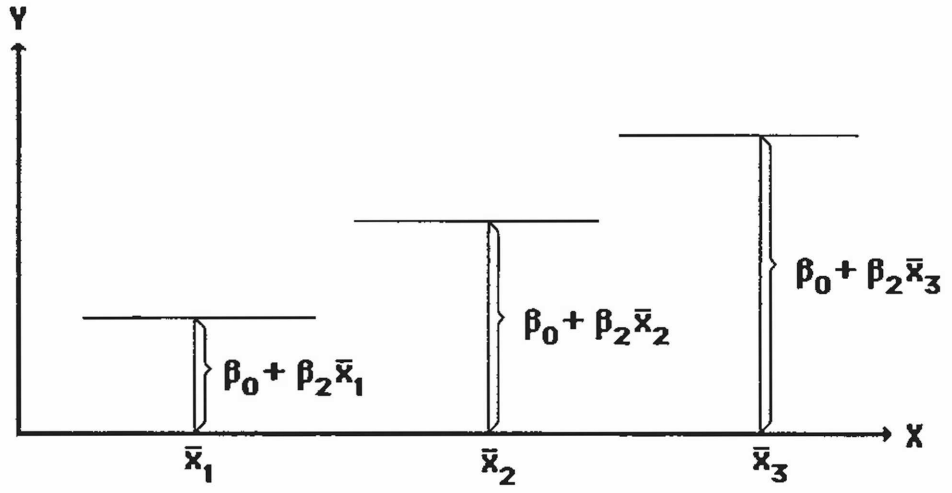
$\beta_0 \neq 0$, $\beta_1 = 0$, $\beta_2 \neq 0$, $\beta_3 = 0$. In this case $\mu_{0k} = \beta_0 + \beta_2(x_k - x)$ and $\mu_{1k} = 0$. When we substitute this model back into the relationship between X and Y in the k-th group we get the relationship expressed in the equation

$$\begin{aligned} y_{ik} &= \beta_0 + \beta_2(x_k - x) + 0x_{ik} + \varepsilon_{ik} \\ &= \beta_0 + \beta_2(x_k - x) + \varepsilon_{ik} \end{aligned}$$

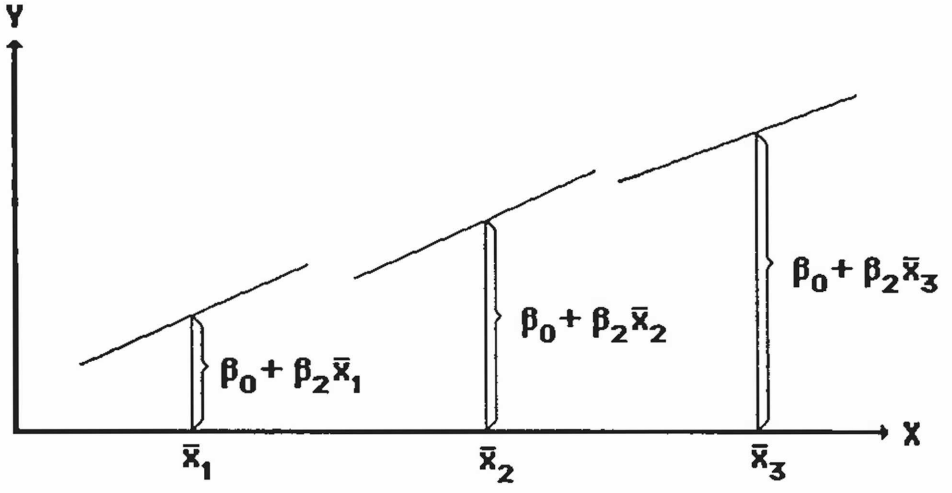
Thus, aside from the residual, all values of Y in a group are equal, and this common value is determined by the group mean. This case is illustrated in the graph below.

In the graph all the lines are horizontal, meaning that within each group X is not related to Y. Thus, in this case there is no individual level effect of X on Y. But it make a difference for Y what group an individual belongs to, and we express that by saying that

there is a group effect of X on Y. There is a group effect of X since it is the value of the the group mean of X which determines Y. This effect is present because the parameter β_2 is different from zero, and β_2 is therefore the group effect parameter.



When we regress Y on the group mean, as expressed in the model above, it can be shown that the estimate of the group coefficient β_2 is the slope of the line that best fits the group mean points with coordinates (x_k,y_k).



$\beta_0 \neq 0$, $\beta_1 \neq 0$, $\beta_2 \neq 0$, $\beta_3 = 0$. In this case we have both an individual and a group level effect of X on Y. If we look at the graph for the group effect, the addition of the individual level effect amounts to pivoting the lines around their balance points.

Within the k-th group we have the following model:

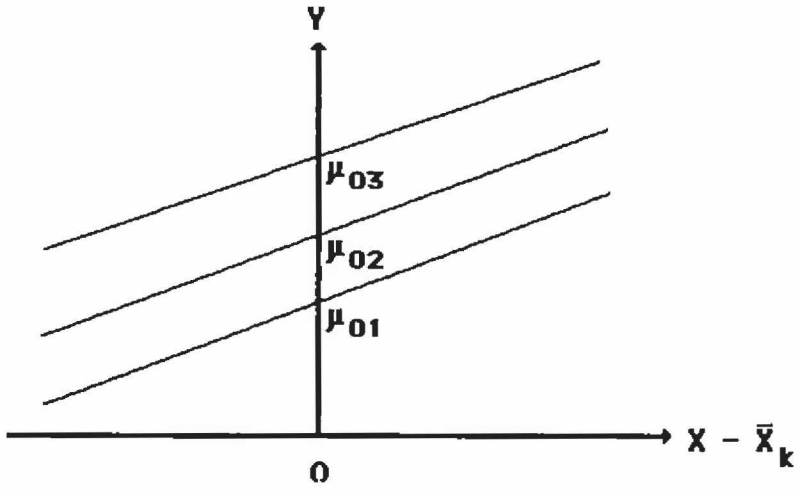
$$\mu_{0k} = \beta_0 + \beta_2(x_k - x)$$
$$\mu_{1k} = \beta_1$$

When we substitute for this model in the equation for the relationship between X and Y in the k-th group, we get

$$y_{ik} = \beta_0 + \beta_1(x_{ik} - x_k) + \beta_2(x_k - x) + \varepsilon_{ik}$$

The graph of the relationship between X and Y looks like the graph above. The group lines are parallel and have the common slope β_1. The fact that the lines are not horizontal shows the presence of the individual level effect of X. In addition, the mean points on the lines are at different levels, and that shows the presence of the group level effect of X.

This case can also be illustrated with a different graph. It is possible to plot Y against the difference between X and the group mean instead of plotting Y against X. In this case the graph above changes and becomes:



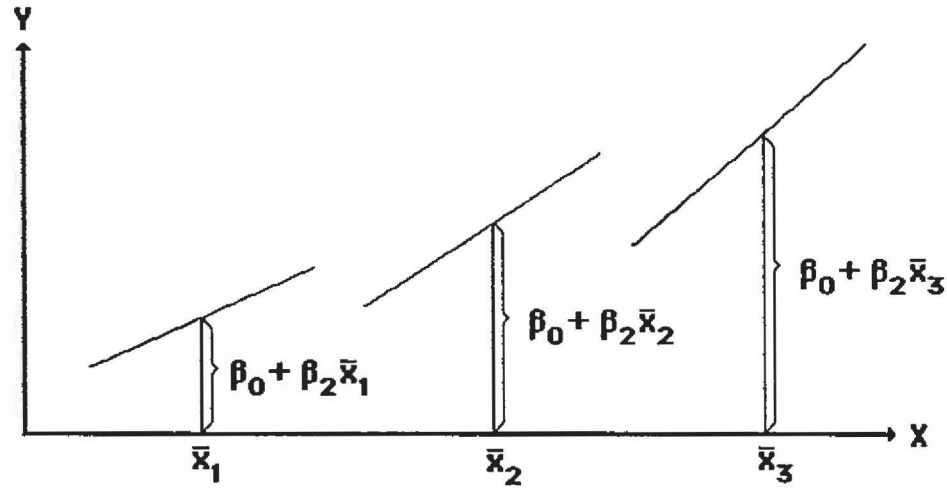
We see the presence of the individual level effect since the lines are non horizontal, and we see the presence of the group effect since the level of Y is different in the various groups.

All betas different from zero. In this case the slope μ_1 is also a function of the group mean and the lines for the different groups will no longer have the same slope. Such a situation is shown in the graph below.

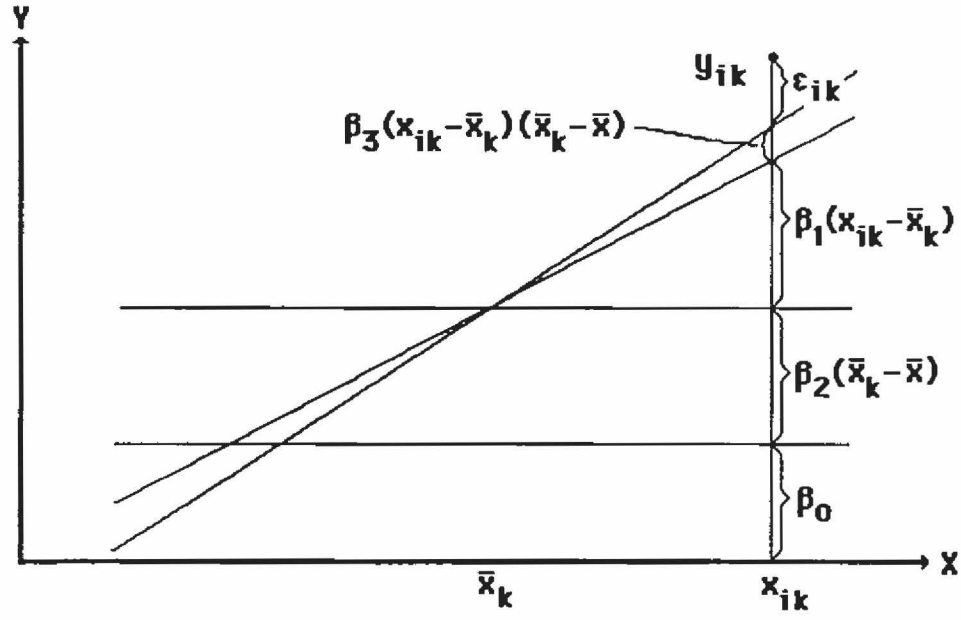
The different slopes show that in addition to the individual and group effects there is now also an individual-group interaction effect present. The unequal slopes are produced by the interaction effect which comes from having β_3 different from zero.

When the model equations are substituted into the equation for the relationship between X and Y, we get the resulting equation

$$y_{ik} = \beta_0 + \beta_1(x_{ik} - x_k) + \beta_2(x_k - x) + \beta_3(x_{ik} - x_k)(x_k - x) + \varepsilon_{ik}$$



The equation shows how the observed value of Y is related to the position of the individual relative to the mean of the group the individual belongs to, to the position of the group relative to the overall mean, and to the product of these two factors.



The different effects can be shown in the graph above. This graph shows how the observed value y_{ik} is decomposed into the various components due to the individual, group and interaction effects. We can think of this graph as a representation of the process that took place in order to determine the observed value of Y for an individual. The individual started with an original Y-value equal to β_0. First, this individual was exposed to the effect of the group effect variable by being in a group with mean x_k, and this added an amount

$\beta_2(x_k - x)$ to the original Y-value. Next, this individual was exposed to the individual effect variable by having an X-value of x_{ik}, and this added an amount $\beta_1(x_{ik} - x_k)$ to the value of Y. Finally, the individual was exposed to the interaction effect variable, which added the term with β_3. After this individual was exposed to the residual variable, which added another ε_{ik}, we finally observe the value y_{ik} for this individual. The only reason the group effect is introduced here before the individual effect is that the graph is slightly easier to draw this way.

Example balanced model

The table shows data for 25 individuals arranged in 5 different groups. The first column shows the observed values of the dependent variable Y. The second column shows the individual level variable computed as the difference between the original X-values and the group mean. The third column is the group variable, computed by subtracting the overall mean from the group mean. Finally, the interaction variable is the product of the individual and the group variable. These data are constructed from these parameter values: $\beta_0 = 2.00$, $\beta_1 = 1.00$, $\beta_2 = 0.50$ and $\beta_3 = 0.25$.

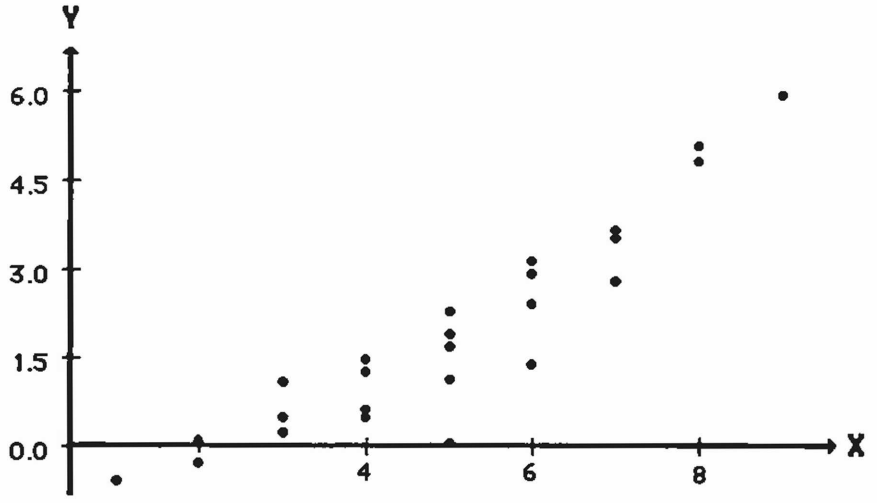
y_{ik}	$x_{ik}-x_k$	x_k-x	$(x_{ik}-x_k)(x_k-x)$
-0.50	-2	-2	4
0.67	-1	-2	2
2.19	0	-2	0
3.06	1	-2	-2
3.98	2	-2	-4
-0.21	-2	-1	2
0.81	-1	-1	1
1.86	0	-1	0
3.12	1	-1	-1
4.21	2	-1	-2
0.32	-2	0	0
0.60	-1	0	0
1.20	0	0	0
3.03	1	0	0
3.59	2	0	0
0.71	-2	1	-2
1.52	-1	1	-1
1.99	0	1	0
3.00	1	1	1
4.16	2	1	2
0.15	-2	2	-4
0.96	-1	2	-2
1.90	0	2	0
3.38	1	2	2
4.00	2	2	4

These data have been rescaled in such a way that they show the variables as they enter the different analyses. It is worth noting how the individual level variable is measured as deviations from the group mean, not as the originally observed values of X. What matters in this model is therefore not the absolute observed value, but where the value is relative to the group mean. As an example, an observed value of 9 in a group where the mean is 7 is the same as an observed value of 5 in a group where the mean is 3. In both cases the difference $x_{ik} - x_k$ is equal to 2. In the example above there are five observations with a value of 2 on the individual variable, but even though they are equal, originally they were different observed values of X. It is this conceptualization of what it means to belong to a certain group

that distinguishes the balanced from the anchored model. In the anchored model it is the actual observed value of X that enters the analysis while in the balanced model it is the value of X relative to the group mean which enters the analysis.

The same difference between the two models shows up in the group variable. In the anchored model the absolute value of the group mean enters the analysis, while in the balanced model it is the relative value of the group mean to the overall mean that matters. What matters for the group effect is where the group mean is located relative to the overall mean.

An examination of the last three columns in the table above shows that those three variables are uncorrelated. That means that we do not have the problem of collinearity when it comes to the estimation of the parameters in the single equation regression analysis. That also means there will be unique sums of squares for each of the three variables, and those sums can be used as measures of the effects of the variables. It can be shown that the individual and group variables are always uncorrelated, as are the group and interaction variables. The individual and interaction variables are uncorrelated as long as the variance of X is the same in every group.



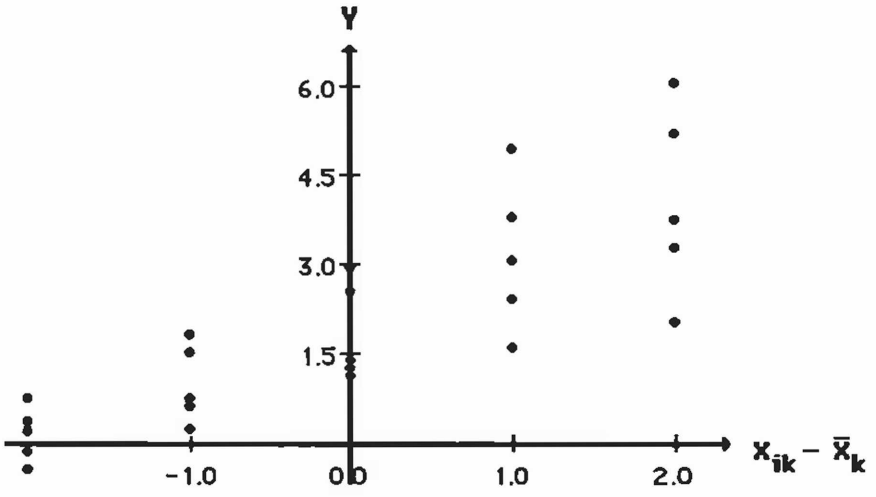
The original values of X are the same as those for the example of the anchored model. When we examine the relationship between Y and the original X-values, we get the scatterplot shown above. The points display a linear relationship, but we want to examine these data from the point of view of the balanced contextual model instead of just regressing Y on X.

The first step consists of examining the relationship between Y and X within each of the five groups. We get the following regression lines:

Group	m_0 m_1
1	$y = 0.88 + 0.64(x - x_1)$
2	$y = 1.46 + 0.86(x - x_2)$
3	$y = 1.77 + 0.92(x - x_3)$
4	$y = 2.78 + 1.09(x - x_4)$
5	$y = 3.08 + 1.51(x - x_5)$

The lines have different slopes as well as different Y-values when X is equal to the group mean.

The scatterplots for the five groups can be shown if we plot Y against $X - X_k$ instead of Y against X. This give the second scatterplot shown below. The five lower points belong to group 1, and so on up to the top five points which belong to group 5. This scatterplot reveals how the five groups differ in a way that was not apparent in the original scatterplot. Since all five groups have nonhorizontal lines, we know that there is an individual level effect of X present in these data. Also, since the level of Y is different in the five groups, we know that there is a group effect present, and since the lines have different slopes we know there is an interaction effect present as well.



In the following we use the model which specifies that group intercepts and slopes are linear functions of the group mean. When we regress the observed intercepts and slopes on the group means, we get these estimates of the effect parameters:

$$m_{0k} = \underset{(0.08)}{1.99} + \underset{(0.06)}{0.57}(x_k - 5) \qquad m_{1k} = \underset{(0.05)}{1.00} + \underset{(0.04)}{0.20}(x_k - 5)$$

The numbers in paretheses below the coefficients are their standard deviations, and each coefficient is based on 3 degrees of freedom since there are five units in each analysis. The coefficient $b_1 = 1.00$ shows the presence of the individual level effect, $b_2 = 0.57$ shows the presence of the group effect, and $b_3 = 0.20$ shows the presence of the interaction effect. From the magnitudes of the standard deviations we see that the coefficients are all significantly different from zero.

Regressing Y on the three effect variables in a multiple regression analysis gives us these single equation estimates of the effect parameters:

$$y_{ik} = \underset{(0.06)}{1.99} + \underset{(0.04)}{1.00}(x_{ik}-x_k) + \underset{(0.04)}{0.57}(x_k-5) + \underset{(0.03)}{0.20}(x_{ik}-x_k)(x_k-5) \qquad R^2 = 0.97$$

From this analysis we find the estimates $B_1 = 1.00$, $B_2 = 0.57$ and $B_3 = 0.20$. These estimates are based on 21 degrees of freedom, since there are 25 units in this analysis. From the standard deviations in parentheses below the estimates we see that all the estimates are significantly different from zero.

The two sets of estimates are identical in this case. The two sets of estimates are the same because the three explanatory variables in the single equation are uncorrelated. The only difference is that the estimates from the single equation have smaller standard deviations than the estimates from the separate equations.

Since those variables are uncorrelated we get unique sums of squares for the three effects. The expressions for the sum of square for the individual variable becomes $B_1^2\Sigma\Sigma(x_{ik}-x_k)^2$, and this is the sum of the squared distances the observations are moved by the individual level effect. For the group and interaction effects there are similar sums of squared distances. If we do not square these distances, but simply add them up the same way effects are measured with the anchored model, we get the effect of the individual level variable as $|B_1|\Sigma\Sigma|x_{ik} - x_k|$. The absolute values are needed since we are only interested in how far a point is moved, not whether it is moved up or down. For the group and interaction variables there are similar sums of absolute values. The estimated squared and absolute-value effects together with their proportions are shown in the following table:

Source	Est. squared effects		Est. abs. value effects	
Individual	50	(0.69)	30	(0.50)
Group	16	(0.22)	17	(0.29)
Interaction	4	(0.05)	7	(0.02)
Residual	2	(0.03)	5	(0.09)
Total	72	(0.99)	59	(1.00)

The proportions work out a little differently in the two cases, mainly with the individual effect being somewhat larger using squared effects.

In this example the true values of the parameters are known, and it is therefore possible to examine how well we are able to estimate the various effects and parameters. We find that all the estimates are within two standard deviations of the true parameter values in this case. The true effects are seen in the table below. Comparing the two tables of effects, we see that for both the squared and absolute value effects the estimates are quite close to the true values. The main difference seems to be that the true interaction effect is larger that the estimated interaction effects.

Source	True squared effects		True abs. value effects	
Individual	50.00	(0.71)	30	(0.51)
Group	12.50	(0.18)	15	(0.25)
Interaction	6.25	(0.09)	9	(0.15)
Residual	2.53	(0.04)	5	(0.08)
Total	70.78	(1.02)	59	(0.99)

It is also possible to use the various estimated coefficients to study the relationship between X and Y within each group. When we substitute the estimated effect parameters into the model equations, we can use those equations to estimate the intercepts and slopes within the groups. We also have estimates of these intercepts and slopes from the regression analysis of Y on X within each group. We get the following results for the five groups:

Group	True rel.ship	Model est.	Within group
1	$y = 1.00+0.50(x-x_1)$	$y = 0.85+\mathbf{0.60}(x-x_1)$	$y = \mathbf{0.88}+0.64(x-x_1)$
2	$y = 1.50+0.75(x-x_2)$	$y = 1.42+\mathbf{0.80}(x-x_2)$	$y = \mathbf{1.46}+0.86(x-x_2)$
3	$y = 2.00+1.00(x-x_3)$	$y = \mathbf{1.99}+\mathbf{1.00}(x-x_3)$	$y = 1.77+0.92(x-x_3)$
4	$y = 2.50+1.25(x-x_4)$	$y = \mathbf{2.56}+\mathbf{1.20}(x-x_4)$	$y = 2.78+1.09(x-x_4)$
5	$y = 3.00+1.50(x-x_5)$	$y = 3.13+1.40(x-x_5)$	$y = \mathbf{3.08}+\mathbf{1.51}(x-x_5)$

The estimates closest to the true values are written in bold numbers for each group. From these results we see that particularly for the middle groups it is better to estimate the relationships between Y and X using the model equations than regressing Y on X within each group. For the more extreme groups the term $x_k - x$ in the model equations is larger and when it is multiplied by the estimated parameters we get a magnification of the error in the estimate through the multiplication.. But when $x_k - x$ is small, it seems as if we do better with the model equations than with the within group regressions.

With aggregate data only we can estimate only two of the parameters. When we aggregate the individual level equation across individuals in a group we get the equation

$$y_k = \beta_0 + \beta_2(x_k - x) + \varepsilon_k$$

where y_k is the mean of Y in the k-th group. This equation shows the relationship between the group means. But since $y_k = m_{0k}$, this equation is the same as the first model equation used for the separate equation method for estimation of the parameter. Thus, when we are restricted to group data, we can only estimate the two parameters β_0 and β_2, and we get no information about the other two parameters.

Residuals balanced model

Because of the orthogonal nature of the balanced model, the residuals lend themselves to a more extended analysis. Residuals occur in the analysis of Y on X within the groups, in the separate equation estimation of the effect parameters, and in the single equation estimation of the same parameters. Here we examine how these residuals relate and give an additional interpretation of the different residuals.

Within the groups. When we regress Y on X in the k-th group, we get the residuals f_{ik} as shown in the equation

$$y_{ik} = m_{0k} + m_{1k}(x_{ik} - x_k) + f_{ik}$$

Within a group, the lack of fit measured by the residual can be seen as the lack of fit of individual level variables. Within the group the magnitudes of these residuals can be measured by the sum of their squares, Σe_{ik}^2. Adding these sums across all the groups we get the sum $\Sigma\Sigma e_{ik}^2$ as an overall measure of the lack of fit of individual level variables.

Separate equations. When we regress the intercepts and slopes on the group means in the model equations, we get the residuals u_k and v_k, as seen in the equations

$$m_{0k} = b_0 + b_2(x_k - x) + u_k$$
$$m_{1k} = b_1 + b_3(x_k - x) + v_k$$

The u's measure the extent to which b_2 can be used as a slope, and we therefore take the u's as measures of the lack of fit for the group effect since b_2 is the coefficient for the group effect of X. Similarly, b_3 is the interaction measure, and we take the v's as the lack of fit for the interaction effect. Let us measure the magnitudes of the u's by computing the sum $\Sigma n_k u_k^2$ and the magnitudes of the v's by the sum $\Sigma\Sigma(x_{ik} - x_k)^2 v_k^2$.

Single equation. When we regress Y on the three effect variables we get the residuals e_{ik} as seen in the equation

$$y_{ik} = B_0 + B_1(x_{ik} - x_k) + B_2(x_k - x) + B_3(x_{ik} - x_k)(x_k - x) + e_{ik}$$

The magnitudes of these residuals can be measured by the residual sum of squares $\Sigma\Sigma e_{ik}^2$. This sum measures the unexplained effects of variables on all three levels.

If we now substitute for the m's from the model equations into the equation for the relationship between Y and X in the k-th group, we get the equation

$$y_{ik} = b_0 + b_1(x_{ik} - x_k) + b_2(x_k - x) + b_3(x_{ik} - x_k)(x_k - x) + [f_{ik} + u_k + (x_{ik} - x_k)v_k]$$

When the b's are equal to the B's, as they often are in the balanced model, the residuals in this equation and the single equation above must be equal. Thus, we get the following relationship between the residuals,

$$e_{ik} = f_{ik} + u_k + (x_{ik} - x_k)v_k$$

When we square the two sides of this equation and add up across all the observations, we get the equality

$$\Sigma\Sigma e_{ik}^2 = \Sigma\Sigma f_{ik}^2 + \Sigma n_k u_k^2 + \Sigma\Sigma[(x_{ik} - x_k)v_k]^2$$

The square of the right hand side involves all three possible cross products in addition to all the squares, but it can be shown that the sums of the various cross products are all equal to zero. This means that we have partitioned the residual sum of squares from the single equations into components which can be thought of as the unexplained parts of the individual, group and interaction effects respectively.

The various sums of squares can be summarized in the following table:

Effect	Explained	Unexplained	Total
Individual	$b_1^2\Sigma\Sigma(x_{ik} - x_k)^2$	$\Sigma\Sigma f_{ik}^2$	Sum
Group	$b_2^2\Sigma n_k(x_k - x)^2$	$\Sigma n_k u_k^2$	Sum
Interaction	$b_3^2\Sigma\Sigma[(x_{ik}-x_k)(x_k-x)]^2$	$\Sigma\Sigma[(x_{ik}-x_k)v_k]^2$	Sum
Total	Regression sum of sq.	$\Sigma\Sigma e_{ik}^2$	Total sum of sq.

For the numerical example of the balanced model we get the following figures:

Effect	Explained	Unexplained	Total
Individual	50.34	1.13	51.47
Group	16.32	0.50	16.82
Interaction	3.91	0.37	4.28
Total	70.57	2.00	72.57

The table shows, if we look at the columns, that most of the unexplained variation is due to the individual level effect. But while the explained individual effect accounts for more than two thirds of the overall explained effect, it only accounts for about half of the overall unexplained effect. If we look at the rows, we see that more of the interaction effect is unexplained by X than is the case for the other two effects.

Recovering individual data

There are times when only the group level data are available. A common example is when we have census data published for geographical subdivisions on various levels, like city blocks or counties. These data typically consist of means or totals for interval level variables and frequencies with percentages for nominal level variables. Ecological data of this kind permit the study of aggregate relationship, but they do not contain enough information to permit the study of individual, group and interaction effect. This raises the question of whether it is possible to recover the data on the level of the individual, after the individual data have been aggregated and presented on the group level, and thereby make a complete contextual analysis possible. Earlier attempts at such recovery are found in Boyd and Iversen (1979), Goodman (1959), Iversen (1973, 1981), Lee et al. (1967) and Telser (1963), among others.

Such recovery of individual level data is possible if there is additional information available. This information can take two different forms. One possibility is that we know the individual level data to contain certain regularities, the other possibility is that we have a limited amount of individual level data available in addition to the group level data. Without either of these forms of additional information recovery of the individual level data is impossible.

Contingency tables. Data on two nominal level variables can be arranged in contingency tables, one table for each group, as shown below in the case of two rows and two columns:

n_{111}	n_{121}	$n_{1\cdot1}$	n_{112}	n_{122}	$n_{1\cdot2}$	. . .	n_{11k}	n_{12k}	$n_{1\cdot k}$	. . .
n_{211}	n_{221}	$n_{2\cdot1}$	n_{212}	n_{222}	$n_{2\cdot2}$		n_{21k}	n_{22k}	$n_{2\cdot k}$	
$n_{\cdot11}$	$n_{\cdot21}$	$n_{\cdot\cdot1}$	$n_{\cdot12}$	$n_{\cdot22}$	$n_{\cdot\cdot2}$		$n_{\cdot1k}$	$n_{\cdot2k}$	$n_{\cdot\cdot k}$	

When we only have group data and not the individual data, it means that the margins $n_{i\cdot k}$ and $n_{\cdot jk}$ are known, but the cell entries n_{ijk} are unknown. In each table we know the marginal distributions for the two variables, but we do not know their joint distribution.

The group may be a geographical subdivision with one table for each area. It is also possible to arrange data over time in a series of contingency tables. For example, in the first table the columns may represents election results for two parties at time 1, and the rows may represent the election results at time 2. In the second table the columns represent the election results at time 2 and the rows the results at time 3, etc. In that case the known margins are the reported election statistics while the unknown cell entries are the turnover votes for each pair of elections. This example assumes that the same people all voted in all elections. While this is unrealistic, changing to proportions partly gets around this restriction.

If the two column proportions $p_{11k} = n_{11k}/n_{\cdot1k}$ and $p_{12k} = n_{12k}/n_{\cdot2k}$ do not vary much from table to table, then it is possible to recover the missing cell entries from the observed frequencies on the margins of the tables. This is the case discussed above on the sec-

tion on contingency tables of an individual level effect only. To see how the cell entries can be recovered from the margins in a set of contingency tables with (almost) constant column proportions, we start by noting that in the k-th table the cell entries in the first row add up to the total in the first row. This can be expressed in the equation

$$n_{1\cdot k} = n_{11k} + n_{12k}$$

From the definition of the column proportions above we have that each cell frequency can be written as a product of the column proportion and the column total. When we substitute for the two cell frequencies we get the equation

$$n_{1\cdot k} = p_{11k}n_{\cdot 1k} + p_{12k}n_{\cdot 2k}$$

As a next step let us divide both sides of the equation by the table total $n_{\cdot\cdot k}$. That changes the marginal frequencies to marginal proportions, and we get

$$p_{1\cdot k} = p_{11k}p_{\cdot 1k} + p_{12k}p_{\cdot 2k}$$

Since the marginal proportion in the second column equals 1.00 minus the proportion in the first column, we can rewrite the equation and get

$$p_{1\cdot k} = p_{11k}p_{\cdot 1k} + p_{12k}(1 - p_{\cdot 1k})$$
$$p_{1\cdot k} = p_{12k} + (p_{11k} - p_{12k})p_{\cdot 1k}$$

So far this is simply an identity for the k-th table, where the column proportions p_{11k} and p_{12k} are unknown and the marginal proportions $p_{1\cdot k}$ and $p_{\cdot 1k}$ are known.

The next step consists of adding up all the tables. The resulting sumtable has the column proportions p_{11} and p_{12}, where

$$p_{11} = \Sigma n_{11k}/\Sigma n_{\cdot 1k} = \Sigma n_{\cdot 1k}p_{11k}/\Sigma n_{\cdot 1k}$$
$$p_{12} = \Sigma n_{12k}/\Sigma n_{\cdot 2k} = \Sigma n_{\cdot 2k}p_{12k}/\Sigma n_{\cdot 2k}$$

From the form of these expressions we see that p_{11} and p_{12} are the weighted means of the column proportions in the tables.

The column proportions in a particular table are different from the column proportions in the sum table. In the k-th table the first column proportion p_{11k} differs from the mean column proportion p_{11} by some amount u_k, and the second differs from the mean by some amount v_k. This can be expressed in the equations

$$p_{11k} = p_{11} + u_k \qquad\qquad p_{12k} = p_{12} + v_k$$

By substituting for these expressions in the basic identity above we get

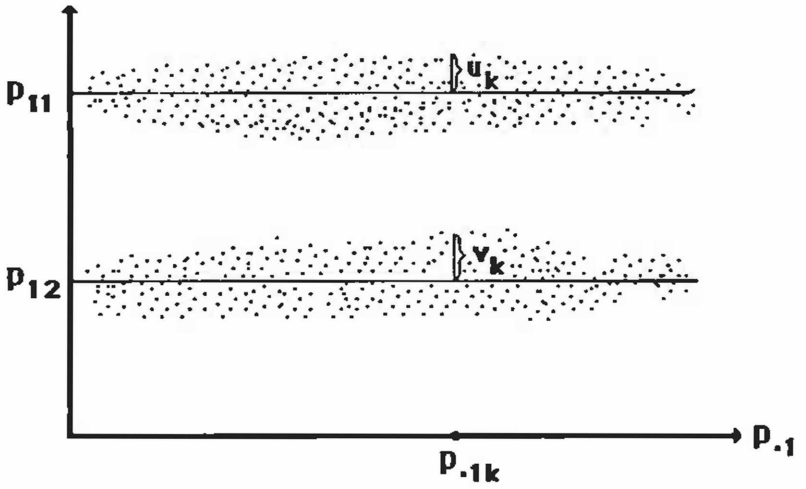
$$p_{1\cdot k} = p_{12} + (p_{11} - p_{12})p_{\cdot 1k} + [u_k p_{\cdot 1k} + v_k(1 - p_{\cdot 1k})]$$

or rewritten,

$$p_{1\cdot k} = p_{12} + (p_{11} - p_{12})p_{\cdot 1k} + e_k \qquad \text{where} \qquad e_k = u_k p_{\cdot 1k} + v_k(1 - p_{\cdot 1k})$$

It is now beginning to look like we can use the set of observed marginal proportions to estimate the two unknown column proportions p_{11} and p_{12} in the sumtable. This is because the equation above says that the known marginal proportion $p_{1\cdot k}$ equals a constant intercept $a = p_{12}$ plus a constant slope $b = p_{11} - p_{12}$ times the known marginal proportion $p_{\cdot 1k}$ plus a residual term e_k. Both intercept and slope are unknown, but the marginal proportions are known, and the situation we are in very much resembles the usual simple regression of a Y variable on an X variable. The equation above suggests that we should regress the marginal proportion in the first row ($p_{1\cdot k}$) on the marginal proportion in the first column ($p_{\cdot 1k}$).

In order to apply the usual least squares methods to estimate p_{12} and $p_{11} - p_{12}$, the residuals must satisfy certain conditions. The first condition is that the mean of the residuals must equal zero. One way this condition is satisfied is for the mean of both the u's and the v's to be zero and for the u's and v's to be uncorrelated with the $p_{\cdot 1k}$'s. Scatterplots of the column proportions versus the marginal proportions would look like the plot below when this condition is satisfied. This is the case we have identified earlier as the case when there is only an individual level effect of X on Y. The difficulty is, however, that the column proportions p_{11k} and p_{12k} are unknown, and we cannot make this plot and see if the conditions are satisfied. Instead, we have to have the knowledge from some other source that there is only an individual effect present in our data.



When we know there is an individual effect only, we can regress the marginal proportions according to the model expressed in the equation

$$p_{1\cdot k} = a + bp_{\cdot 1k} + e_k$$

This gives the expressions below for a and b, and they can be solved for p_{11} and p_{12}.

This is not estimation in the usual statistical sense where we have data from a sample as a subset of a larger population and use that data to estimate parameters in the population. We want to compute the statistic p_{11} according to the expression $\Sigma n_{11k}/\Sigma n_{\cdot 1k}$, but the frequencies in this numerator are unknown since they are the unobserved cell entries and the sta-

$$\hat{a} = \bar{p}_{1\cdot} - \hat{b}\bar{p}_{\cdot 1} \qquad \hat{b} = \frac{\Sigma(p_{\cdot 1k} - \bar{p}_{\cdot 1})(p_{1\cdot k} - \bar{p}_{1\cdot})}{\Sigma(p_{1\cdot k} - \bar{p}_{1\cdot})^2}$$

$$\hat{p}_{11} = \hat{a} + \hat{b} \qquad \hat{p}_{12} = \hat{a}$$

tistic cannot be computed. Instead, we have developed another statistic p_{11} which can be computed since it depends upon the observed margins. Thus, here it is more a question of not having the right data than having a subset of the data. To emphasize this point we call the p_{11} we can compute a proxy statistic for p_{11} rather than an estimator.

The additional issue of estimation comes up if the contingency tables we have are a sample of tables from a larger population of tables and/or if the observations in a table are a sample from a larger population of observations. For example, if we had a multistage sample design where we first drew a sample of counties from the population of all counties and within each chosen county drew a sample of individuals, we would have both a sample of tables and a sample of individuals in each table. In that case we could construct estimators and be concerned with standard errors and other issues that come up in statistical estimation.

More formally, the case of individual effect only with constant column proportions can be expressed in the model equations

$$p_{11k} = \alpha + u_k \qquad \text{and} \qquad p_{12k} = \gamma + v_k$$

By substituting these model equations into the basic table identity we get

$$p_{1\cdot k} = \gamma + (\alpha - \gamma)p_{\cdot 1k} + e_k$$

This is the same equation we used above to get the proxy statistics for p_{11} and p_{12}.

It could be that the data contain a group effect only of X instead of an individual level effect. In that case the column proportions p_{11k} and p_{12k} vary from table to table, and within a particular table they are equal. With a linear relationship this case can be expressed in the model equations

$$p_{11k} = \alpha + \beta p_{\cdot 1k} + u_k$$
$$p_{12k} = \alpha + \beta p_{\cdot 1k} + v_k$$

When these expressions are substituted into the basic table identity we get the equation

$$p_{1\cdot k} = \alpha + \beta p_{\cdot 1k} + e_k$$

This equation tells us that when there is only a group effect of X on Y, there will be a linear relationship between the marginal proportions across a set of contingency tables. The only problem is that with only an individual level effect, the relationship between the marginal proportions is also linear. Thus, when we only have a set of marginal proportions

and find that they are linearly related, there is no way of finding out whether we have the case of individual level or group level effect.

It turns out that the marginal proportions are linearly related as well when there are both an individual and a group level effect present. With both these effect present, the column proportions are related to the marginal proportions according to the model equations

$$p_{11k} = \alpha + \beta p_{\cdot 1k} + u_k$$
$$p_{12k} = \gamma + \beta p_{\cdot 1k} + v_k$$

Substituted into the basic table identity we find that the marginal proportions are related according to the equation

$$p_{1\cdot k} = \gamma + (\alpha + \beta - \gamma)p_{\cdot 1k} + e_k$$

This is again a linear equation. The major difference between this equation and the previous other two equations is that here there are three parameters while the others had only two. The data can be used to find the intercept and slope, but these two quantities do not contain enough information to give us estimates of all three parameters.

From the marginal proportions alone there is no way of distinguishing between the case of individual effect only, group effect only, and both individual and group effect. But we are able to see the presence of the individual-group interaction effect. This is because with the interaction effect the column proportions are related to the marginal proportions according to the model equations

$$p_{11k} = \alpha + \beta p_{\cdot 1k} + u_k$$
$$p_{12k} = \gamma + \delta p_{\cdot 1k} + v_k$$

The difference now is that there are different slopes β and δ for the two column proportions. When this model is substituted into the basic table identity we get the following equation for the relationship between the marginal proportions:

$$p_{1\cdot k} = \gamma + (\alpha - \gamma + \delta)p_{\cdot 1k} + (\beta - \delta)p^2_{\cdot 1k} + e_k$$

This is no longer a linear equation. With the presence of the interaction effect comes a term with the square of the marginal proportion. Thus, when we find a nonlinear relationship between the marginal proportions, we take that as a sign that there is an interaction effect present in the data. But least squares methods give us three estimates only, for the intercept and the coefficients for the linear and quadratic terms, and the model contains four parameters. This means we are not able to estimate the parameters from group data alone.

Up to this point we have seen that the available group data alone do not contain enough information to recover the individual data. If we know that the data contain only an individual level effect or only a group level effect, the individual data can be recovered. If there are two or more effects present, there are too many parameters in the model and the individual data cannot be recovered. In most cases we do not know what effects are present in the data; instead, this is something we want to use the data to find out.

But we are always free to assume any model we want, and if we assume an indi-

vidual effect only, individual level data can be recovered. These results are dependent upon two sources: the group level data and the assumptions in the model. This then is a situation where the model plays an unusually strong role, and it may defeat the purpose of the study to use such a powerful model.

A more realistic solution to obtaining the missing cell entries in a set of contingency tables lies with the presence of partial individual level data. Perhaps we have survey data for a few of the groups and can thereby fill in the tables for those groups. Such additional information can then be combined with the marginal distributions and used to estimate all the parameters in our model and thereby recover the missing individual data in the remaining tables. Since contextual analysis have such extensive data requirements, it is also possible to design the data collection in such a way that we gather individual and group data for only some of the groups and only group data from the remaining groups.

There are many approaches that can be followed when we have partial individual level data. We are working with the two separate model equations

$$p_{11k} = \alpha + \beta p_{\cdot 1k} + u_k$$
$$p_{12k} = \gamma + \delta p_{\cdot 1k} + v_k$$

and the single equation

$$p_{1\cdot k} = \gamma + (\alpha - \gamma + \delta)p_{\cdot 1k} + (\beta - \delta)p^2_{\cdot 1k} + e_k$$

One thing we can do is the following. The group data can be used in the single equation to estimate the three parameter combinations γ, $\alpha - \gamma + \delta$ and $\beta - \delta$. The second single equation suggests that one way to estimate δ would be to observe the column proportion p_{12k} for a few groups where the marginal proportion $p_{\cdot 1k}$ is large, and then use the second single equation with the existing estimate of γ to estimate the slope δ. This estimate of δ can then be used together with the estimate of $\beta - \delta$ to get a separate estimate of β. Finally, we get the separate estimate of α by using the estimate of $\alpha - \gamma + \delta$ and the separate estimates of γ and δ.

When we have group data for a set of tables and individual cell entries for some of the tables, it is possible to use both the single equation and the separate equations together and minimize an overall sum of squares for the estimation of the parameters. The combined sum of squares becomes

$$Q = \Sigma[p_{1\cdot k} - \gamma - (\alpha - \gamma + \delta)p_{\cdot 1k} - (\beta - \delta)p^2_{\cdot 1k}]^2 + \Sigma[p_{11k} - \alpha - \beta p_{\cdot 1k}]^2 + \Sigma[p_{12k} - \gamma - \delta p_{\cdot 1k}]^2$$

By taking the derivatives with respect to the four unknown parameters and setting the resulting expressions equal to zero, we get four normal equations that can be solved for estimated parameters.

Metric variables. The situation is much the same with data for interval level variables. Without individual data there is not enough information in the group data alone to do a contextual analysis. From the group data alone it is not possible to distinguish between the presence of the various effects, and when there are two or more effects there are more parameters in the model than we can estimate.

With additional, partial individual data it is usually possible to estimate the parameters

in the models we have discussed. These estimates can then be used to estimate the within group relationships between the variables. But because of the effects of the residual variable it is not possible to recover actual values of the variables.

References

Blalock, H. M. (1984), "Contextual-effects models: Theoretical and methodological issues," *Annual Review of Sociology,* 10: 353-372.

Blau, P. M. (1960), "Structural effects," *American Sociological Review,* 25 (2): 178-93

Boyd, L. H.Jr. and G. R. Iversen (1979), *Contextual Analysis: Concepts and Statistical Techniques*, Belmont, CA: Wadsworth Publishing Co.

Davis, J. A., J. L. Spaeth and C. Huson (1961), "A technique for analyzing the effects of group composition," *American Sociological Review,* 26(2): 215-25.

Goodman, L. A. (1959), "Some alternatives to ecological correlation," *American Journal of Sociology* 64(6): 610-25.

Hero, R. E. and R. Durand (1985), "Explaining citizen evaluations of urban services: A comparison of some alternative models," *Urban Affairs Quarterly,* 20(3): 344-354.

Iversen, G. R. (1973), "Recovering individual data in the presence of group and individual effects," *American Journal of Sociology,* 79(2): 213-23.

Iversen, G. R. (1981), "Group data and individual behavior," in J. M. Clubb, W. H. Flanigan and N. H. Zingale (eds.), *Analyzing Electoral History: A Guide to the Study of American Voting Behavior*, pp. 267-302,Sage Focus Editions 33, Beverly Hills: Sage Publications.

Kendall, P. L. and P. F. Lazarsfeld (1950), "Problems of survey analysis," in R. K. Merton and P. F. Lazarsfeld (eds.), *Continuities in Social Research: Studies in the Scope and Method of The American Soldier'* , pp. 133-96, Glencoe, IL: The Free Press.

Knoke, D. (1981), "Commitment and detachment in voluntary associations," *American Sociological Review*, 46(2): 141-58.

Lee, T. C., G. G. Judge and A. Zellner (1967), "Maximum likelihood and Bayesian estimation of transition probabilities," Report 6733, Center for Mathematical Studies in Business and Economics, University of Chicago.

Schuessler, K. (1969), "Covariance analysis in sociological research," in E. F. Borgatta (ed.), *Sociological Methodology,* pp. 219-44, San Francisco: Jossey-Bass.

Telser, L. G. (1963), "Least squares estimates of transition probabilities," in C. Christ et al. (eds.), *Measurement in Economics* pp. 270-92, Stanford: Stanford University Press.

Tate, R. L. (1985), "Limitations of centering for interactive models," *Sociological Methods and Research,* 13 (2): 221-234.

van den Eeden, P. and H. J. M. Hüttner (1982), "Multi-level research," *Current Sociology,* 30(3): 1-117.

ZUMA-Arbeitsberichte

80/15 Gerhard Arminger, Willibald Nagl, Karl F. Schuessler
Methoden der Analyse zeitbezogener Daten. Vortragsskripten der ZUMA-Arbeitstagung vom 25.09. - 05.10.79

81/07 Erika Brückner, Hans-Peter Kirschner, Rolf Porst, Peter Prüfer, Peter Schmidt
Methodenbericht zum "ALLBUS 1980"

81/19 Manfred Kuchler, Thomas P. Wilson, Don H. Zimmerman
Integration von qualitativen und quantitativen Forschungsansätzen

82/03 Gerhard Arminger, Horst Busse, Manfred Küchler
Verallgemeinerte Lineare Modelle in der empirischen Sozialforschung

82/08 Glenn R. Carroll
Dynamic analysis of discrete dependent variables: A didactic essay

82/09 Manfred Küchler
Zur Messung der Stabilität von Wählerpotentialen

82/10 Manfred Küchler
Zur Konstanz der Recallfrage

82/12 Rolf Porst
"ALLBUS 1982" - Systematische Variablenübersicht und erste Ansätze zu einer Kritik des Fragenprogramms

82/13 Peter Ph. Mohler
SAR - Simple AND Retrieval mit dem Siemens-EDT-Textmanipulationsprogramm

82/14 Cornelia Krauth
Vergleichsstudien zum "ALLBUS 1980"

82/21 Werner Hagstotz, Hans-Peter Kirschner, Rolf Porst, Peter Prüfer
Methodenbericht zum "ALLBUS 1982"

83/09 Bernd Wegener
Two approaches to the analysis of judgments of prestige: Interindividual differences and the general scale

83/11 Rolf Porst
Synopse der ALLBUS-Variablen. Die Systematik des ALLBUS-Fragenprogramms und ihre inhaltliche Ausgestaltung im ALLBUS 1980 und ALLBUS 1982

84/01 Manfred Küchler, Peter Ph. Mohler
Qualshop (ZUMA-Arbeitstagung zum "Datenmanagement bei qualitativen Erhebungsverfahren") - Sammlung von Arbeitspapieren und -berichten, Teil I + II

84/02 Bernd Wegener
Gibt es Sozialprestige? Konstruktion und Validität der Magnitude-Prestige-Skala

86/11 Günter Rothe
Bootstrap in generalisierten linearen Modellen

87/01 Klaus Zeifang
Die Test-Retest-Studie zum ALLBUS 1984 - Tabellenband

87/02 Klaus Zeifang
Die Test-Retest-Studie zum ALLBUS 1984 - Abschlußbericht

87/03 Michael Braun, Rolf Porst
ALLBUS-Bibliographie (6. Fassung, Stand: 30.06.87)

87/04 Barbara Erbslöh, Michael Wiedenbeck
Methodenbericht zum "ALLBUS 1986"

87/05 Norbert Schwarz, Julia Bienias
What mediates the Impact of Response Alternatives on Behavioral Reports?

87/06 Norbert Schwarz, Fritz Strack, Gesine Müller, Brigitte Chassein
The Range of Response Alternatives May Determine the Meaning of the Question: Further Evidence on Informative Functions of Response Alternatives

87/07 Fritz Strack, Leonard L. Martin, Norbert Schwarz
The Context Paradox in Attitude Surveys: Assimilation or Contrast?