Interpreting Heteroscedasticity

George W. Downs; David M. Rocke


Stable URL:
http://links.jstor.org/sici?sici=0092-5853%28197911%2923%3A4%3C816%3AIH%3E2.0.CO%3B2-U

*American Journal of Political Science* is currently published by Midwest Political Science Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR’s Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR’s Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/mpsa.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.
Interpreting Heteroscedasticity

George W. Downs, University of California, Davis
David M. Rocke, Governors State University

Heteroscedasticity is generally viewed as no more than a barrier to the accurate estimation of regression coefficients which requires the application of weighted least squares or variable transformations. Here it is argued that its analysis can also provide political scientists with significant substantive information that would ordinarily go undetected as well as evidence of model misspecification. Several empirical examples are presented which illustrate the kinds of conclusions which can be drawn.

Heteroscedasticity, or inequality of error variance, is a topic that is treated in some detail by most statistics and econometric texts (cf. Theil, 1971; Draper and Smith, 1966; Kmenta, 1971; Hanushek and Jackson, 1977; Pindyck and Rubinfeld, 1977). While there is generally some discussion of the factors that can bring it about, it is treated almost exclusively as a problem which must be overcome if optimal estimates of regression coefficients are to be made. Such an approach is consistent with the commonly held idea that the importance of variance lies entirely in its functioning as a barrier to the accurate estimation of means and coefficients. That is, one needs to know something about the variance only in order to assess the precision of estimates. In this paper a very different point of view is set forth. The argument is made that differences in variances can often have important theoretical and substantive implications over and above those for accurate estimation.

Traditional Treatments

The requirements for inference in ordinary least squares (OLS) are that the error terms of the various observations be independent and normally distributed with zero mean and equal variances $\sigma^2$. Violations of these assumptions—such as serial correlation, long-tailed error distributions, and inequality of variances—require changes in the estimation techniques. It is well known that when the equality-of-error-variance assumption is violated the use of OLS results in inefficient estimates and the usual tests of significance are invalid (Goldfeld and Quandt, 1972, p. 81). Because of this much effort has been spent in developing techniques to “cure” heteroscedasticity.
There are roughly two schools of thought concerning the sources of and remedies for heteroscedasticity: the econometric school and the data analytic school. Economists have developed an enormous array of weighted least squares (WLS) techniques that provide more accurate estimates than OLS in the presence of heteroscedasticity. In WLS, observations are given weight inversely proportional to their estimated variance. While the ultimate justification for employing this technique rests on technical grounds, it has strong intuitive appeal if—as certainly can be the case—the data points consist of the aggregation of differing numbers of units or are subject to differing measurement error.

There are other circumstances with which economists are familiar. One occurs when percentage errors are more constant than absolute errors. Budget changes can fall into this category. Because they are more often considered in percentage than absolute terms, deviations from a predictive model of actual expenditures will tend to be far greater for larger budgets. Another interesting case (Theil, 1971; Kmenta 1971) involves the operation of selective constraints. If family consumption is regressed on family income one finds far greater consumption variation in high income families than in low. This is not surprising, for we expect families with low incomes to spend almost all of their funds on the basic necessities of life. The requirement that these necessities be provided together with income-related limitations on credit lead us to anticipate that major departures from this pattern will be rare. As incomes rise, however, families enjoy an increasingly greater range of choice concerning consumption, savings, and investment. Hence we expect that variation in the amount of income devoted to consumption will increase with income.

Data analysts and statisticians less closely tied to formal econometric modeling take a different approach to overcoming the effect of heteroscedasticity. Rather than developing special methods for each situation (e.g., types of WLS for varieties of heteroscedasticity) they try to transform the independent variables in order to simplify their relationships and make them consistent with the usual assumptions. They have noticed that the three problems of nonlinearity of regressions, nonnormality of marginal distributions, and heteroscedasticity often occur together and that a transformation calculated to eliminate the skewness of the individual distributions will often simultaneously remedy the other two problems.

The point of view of the data analyst is generally that the characteristics of relationships between variables depends greatly on the scale on which they are measured. As Box and Cox put it "we are concerned not merely to find a transformation which will justify assumptions but rather to
find, where possible, a metric in terms of which the findings may be succinctly expressed” (1964 p. 211). Thus heteroscedasticity is seen to arise from measurement on the “wrong” scale.¹

While the approaches of econometricians and data analysts to the problem of heteroscedasticity are divergent, they share the common attitude that the goal of any analysis lies in detecting average values (including regression coefficients). Variance is important only insofar as it interferes with the estimates of these averages and in suggesting how much faith should be placed in the results. However, there are other grounds for believing that heteroscedasticity merits attention.

**Substantive Interpretations**

Recall the example cited by Theil and Kmenta about the relationship between income and consumption. Here they have quite sensibly used an a priori suspicion—that consumption patterns of low income people are more constrained than those with high incomes—to suggest the need for WLS.

¹ WLS can be thought of as involving transformations of a kind also. What is meant here is applying the logarithm, square root, or other appropriate transformations to each variable individually.
The fact that such theoretical and substantive knowledge is capable of anticipating the existence of heteroscedasticity raises the interesting possibility that the opposite can also be true; that the presence of heteroscedasticity and its subsequent exploration can yield various types of substantive knowledge.

To continue with the income/consumption example, if we did not already know that the consumption level of low income people is strongly bounded on the low side we would discover this and quite probably infer the existence of "necessary expenditures for survival." Similarly, the fact that the upper bound of the distribution was closer to the regression line in the case of low income people might lead us to think about differential access to credit and accumulated assets. More interestingly perhaps, the rate at which the lower bound of residuals (as opposed to the regression line) rose with income might suggest something useful (by implication) about the relationship between permanent and transitory income and that between income and wealth (Ando and Modigliani, 1963).

In this particular case, the availability of substantive knowledge and theory makes such an analysis largely unnecessary, although detailed empirical confirmation and aid in parameter estimation are always welcome. Frequently, however, such knowledge needs to be derived directly from the data or never discovered at all. Because so much data analysis in political science begins not with an established theory but with a large data set and a few suspicions, the possibility that heteroscedasticity can have substantive implications suggests that it should be explored. Several examples follow in which the significance of a finding of heteroscedasticity lies more in its interpretation and implications for model specification than in its implications for the efficiency of estimation.

Figure 2 is a scatterplot depicting the relationship between state deinstitutionalization rates in juvenile corrections and the amount of interest which the legislature has traditionally shown in this policy area. Deinstitutionalization rate represents the percentage of incarcerated juveniles being held in more "innovative" community-based facilities as opposed to institutions. The level of legislative interest was estimated by the state director of juvenile corrections on the basis of his or her experience in the past five years. Note the presence of heteroscedasticity: as legislative interest rises, the variance of deinstitutionalization increases. The correlation between the two variables is .44, and this value together with a WLS estimate of the regression line would normally be considered an adequate summary of the relationship—especially for exploratory purposes.
FIGURE 2
The Relationship between Deinstitutionalization and Legislative Interest in Juvenile Corrections.
However, the nature of the heteroscedasticity in Figure 2 suggests something more. While the mean or median level of deinstitutionalization rises as the amount of legislative interest increases, a substantial number of states at each activity level have achieved no significant amount of deinstitutionalization. Put somewhat differently, there is no noticeable relationship between the lower boundary of the distribution around the regression line and the independent variable. On the other hand, a glance at the upper boundary indicates a very strong relationship.

One substantively meaningful interpretation of this is that legislative interest appears to function as a necessary but not sufficient condition for the extensive adoption of this particular policy innovation. That is, the triangular appearance of the scatterplot suggests that increased legislative interest permits but does not require a greater degree of deinstitutionalization. Perhaps legislative interest functions as a kind of resource that permits juvenile corrections—a policy area which does not enjoy the continued salience of education or transportation—to be included on the legislative agenda. The low deinstitutionalization rates of high “interest” states could be attributed to a lack of motivation on the part of the bureaucracy to utilize the resource for initiating innovation that this legislative interest represents. The acknowledged marginality of juvenile corrections (except for periodic crises) and the increased realization that bureaucracies must frequently take the lead in initiating innovations make such an interpretation attractive. In any event, this type of scatterplot indicates that there is quite likely an unobserved variable with which legislative interest interacts and which needs to be included in any satisfactory model. The presence of heteroscedasticity has suggested what could be an important line of investigation.\(^2\)

Figure 3 depicts the relationship between deinstitutionalization rate and the percent of adults in the state with less than five years education. For the sake of convenience the latter will be referred to as “low education” in what follows. The correlation between the two variables is \(-.42\), which again does not tell the whole story. This figure resembles Figure 2 except that in this case low values appear to facilitate implementation of

\(^2\) Many will immediately suspect that this is an artifact of the boundedness of any percentage figure. However, considerable experimentation with transformations of the form

\[(p + a)^b - (1 - p + a)^b,\]

including logits, has failed to yield any which substantially changes the heteroscedastic appearance of the plot.
FIGURE 3

The Relationship between Deinstitutionalization Rate and Percentage of Population with Less than Five Years Education.
the innovation. As values increase the upper bound of the distribution decreases, but the range of the lower bound remains the same. In the comparative policy literature the correlation would, by itself, be taken as further corroboration of the long-recognized association between level of education and liberality or reform-mindedness. The shape of the scatterplot encourages further reflection, however, because such an explanation does not account for all features of the data, specifically the heteroscedasticity.

One alternative explanation—consistent with the character of the heteroscedasticity and precipitated by its discovery—is that the reluctance of communities to integrate a large minority population acts as a barrier which limits the maximum amount of deinstitutionalization that can take place but is unrelated to the minimum that can occur. Low education is a statistical surrogate for racial and income diversity, and this particular reform involves the treatment of offenders in the community rather than in institutions. In states where class and racial differences are deep, one might anticipate a more punitive (i.e., institutional) approach to deviancy, particularly when offenders often belong to a powerless minority group.\(^3\)

Yet there is no reason to believe that a reduction in the sort of barrier toward community acceptance that this variable represents will necessarily result in reform unless other barriers are also removed and/or motivation on the part of decision makers is present. The shape of the scatterplot implies that no statistical model of the determinants of deinstitutionalization can employ low education in a strictly additive fashion. There ought to exist some variable(s) representing barriers/motivations whose inclusion in multiplicative terms should greatly increase explanatory power.

In addition to the type of plots just discussed there are at least two other general shapes worth considering. The first is depicted in Figure 4. Here the most conspicuous feature is that the lower boundary of \(Y\) appears to be much more closely related to \(X\) than either the conditional mean of \(Y\) or the upper boundary of the distribution around what would be the regression line. Extending the metaphor of the previous examples, \(X\) appears to be functioning as a sufficient but not necessary condition. An increase in \(X\) is sufficient to ensure an increase in the minimal value of \(Y\) but where above that minimum level \(Y\) will fall depends on some other

---

\(^3\) The fact that the percentage of the adult population with less than five years of education is more strongly related to the dependent variable than median education (\(-.42\) versus \(.34\)) adds further credence to the argument that the uneven distribution of educational benefits may be more important than the absolute level.
Sufficient But Not Necessary Condition

unspecified factors that interact with $X$ (as well as some stochastic element, of course). It is the sort of plot, for example, that might be anticipated when examining the relationship between the size of a city or organization and the extent to which it employs electronic data processing. A small organization may or may not utilize it depending on the tastes and resources of management, but as the organization becomes larger it is going to be forced to employ some minimal amount. This sort of relationship has considerable relevance to many policy questions where the attempt is made simultaneously to permit discretion while ensuring that some baseline of service or compliance exists.

Figure 5 depicts still another variety of heteroscedastic relationship. Here $X$ is simultaneously related—in different directions—to the upper and lower boundaries but not at all to the mean of the distribution. In fact, the correlation is 0. Yet while there is no correlation between the two variables it would be precipitous to charge that $X$ has no impact. If manipulating $X$ contracts or expands the variance of $Y$, the fact is worth thinking about. Suppose one is interested in examining the relationship between relaxed zoning restrictions and housing prices. Evidence that relaxing restrictions increases the range of prices is important for policy making regardless of whether the mean or median price goes up or down. We
might expect the same sort of relationship to exist between rules for party participation and ideological representation. Fundamentally, the relationship in Figure 5 is closely related to both that discussed by Kmenta and the empirical examples presented here. What makes these interesting is not the fact that they contain statistically significant relationships (this may or may not be the case) but that the heteroscedasticity is a function of the independent variable.

The examination of substantively important heteroscedasticity makes it apparent that summary statistics such as regression coefficients or correlations can be misleading in a number of ways. When heteroscedasticity exists $R^2$ provides only an average estimate of the variance explained. Because percent low education provides good predictions for high values and poor predictions for low values, the $R^2$ obtained from a simple regression simultaneously over- and underestimates the accuracy of the prediction.\(^4\) Exclusive emphasis on the significance of regression coefficients can also be a problem. In the cases of legislative interest and low education, their coefficients are significant but underestimate their true substan-

\(^4\) Using weighted least squares one can obtain uncertainty estimates for each value of low education, but this information is not easily summarized and tends to be ignored.
tive value. Even worse, a variable like that in Figure 5 will be dismissed as being totally unimportant despite the fact that the change in variance provides useful information. Exploratory data analysis which is directed by \( t \) scores and \( F \) statistics is likely to ignore the latter entirely and give the former less attention than they deserve.

Perhaps the most important reason to identify and study heteroscedasticity is that it may provide the only available evidence of interacting variables. Existing computer programs can search data sets for evidence of simple interaction; however, a finding of heteroscedasticity can provide a rationale for suspecting interaction with variables not included in the data set. The discovery of such variables is necessary both to provide stability of predictions and accurate specification. If an interactive model is incorrectly specified as linear, the values of the regression coefficients will change depending on what ranges of the independent variables are examined in a given sample. While it is true that much interaction may disappear after transformation, it is necessary to find it before it can be eliminated. Furthermore, the detection of possible interaction may suggest that a variable which has previously been ignored needs to be included in the analysis; as when the resource interpretation of legislative interest suggests that some measure of bureaucratic motivation is needed. Substantive conclusions of this kind are independent of the technical problems of specification and provide a good reason to examine heteroscedasticity and interaction.

**Transformation and Multivariate Analysis**

The interpretation of heteroscedasticity in the two empirical examples is substantively fruitful, but can the pictures be trusted? Many data analysts would argue that the shape of the plots is merely the result of a failure to transform the variables to a suitable scale. Certainly it is true that any systematic heteroscedasticity can be eliminated by transformation, and for quantitative modeling this is a necessary step. It is essential to obtain the best estimate of a particular coefficient, and in any of the examples the application of WLS or the appropriate transformation will eventually be necessary. However, this is not to say that conclusions drawn from the heteroscedasticity of the original variables are any less valid. Researchers commonly use the log of family income rather than family income itself for model building because the former is roughly symmetrical. However, no one suggests that the skewness of income distribution is an unimportant triviality brought about by the unfortunate choice of dollars as a measure
of income. Taking logarithms does not redistribute income. Justification for the examination of variance, whether in the univariate or bivariate case, need not be based on optimal estimation. The potential discovery of missing variables, the existence of important interaction effects, necessary but not sufficient conditions, multiple barriers, and so forth, is fully as important as preparing variables for entry into a multiple regression equation.

Another issue involves the interpretation of heteroscedasticity in multivariate analysis. To this point we have been dealing with simple regression. Most exploratory data analysis in political science, however, takes place within the context of multiple regression and it is appropriate to consider how this might complicate matters. In one sense, the fact that there are a number of variables that are related to the dependent variable presents no problem at all. Heteroscedasticity cannot be produced by linear relations with other variables. A finding of heteroscedasticity in a plot of \( Y \) vs. \( X \) has the same interpretation whether or not one or both is related to other variables.

A real problem, however, might be a loss of sensitivity. Just as a scatterplot of two variables may fail to disclose a relationship which would be revealed by a multiple regression analysis, so heteroscedasticity may be hidden by the noise from other variables. Fortunately, the elimination of this noise presents no special difficulty. Given the independent variable \( Y \) and the independent variables \( X_1, X_2, \ldots, X_p \) one plots \( \hat{Y} \) vs. \( \hat{X}_1 \) where \( \hat{Y} \) and \( \hat{X}_1 \) are the residuals of the separate regressions of \( Y \) and \( X_1 \) on \( X_2, \ldots, X_p \). This procedure removes the linear effect of \( X_2, \ldots, X_p \), and the interpretation does not differ from the simple bivariate case.

**Conclusion**

In sum, there are good reasons to believe that heteroscedasticity is more than simply a barrier to the correct estimate of coefficients. Attention to differences in variance can provide both an important supplement to present exploratory analysis techniques and suggest ways to evaluate theory. Quantitative models require precisely estimated coefficients, but a single-minded emphasis on curing heteroscedasticity can lead one to ignore valuable information. The suspicion that a variable may function as a nec-

---

\(^5\) If the distribution of \( Y \) given \( X_1, \ldots, X_p \) is normal with mean \( B_0 + B_1X_1 + \cdots + B_p X_p \) and variance \( \sigma^2 \), then the distribution of \( Y \) given \( X_1 \) is normal with mean \( B_0 + B_1 X_1 \) and constant variance \( \sigma^2 \).
necessary but not sufficient condition is no less important than knowing that it is "related" to the dependent variable. Moreover, without knowledge of the former the latter information can be quite misleading. It is one thing to believe that an increase in $X$ will ensure—within a consistent margin or error—an increase in $Y$, and another to learn that $X$ is more strongly related to the maximum value of $Y$ than the minimum, and that $X$ is likely to have an interactive effect with some unspecified variable.

These sorts of findings can be valuable. To restrict interest in heteroscedasticity to the problem of optimal estimation can lead us to miss the implications that its existence may hold for substantive insight, model specification, and the discovery of stable estimates. Obviously there is no guarantee that all discoveries of heteroscedasticity will contribute to theory building; many will not. Yet the same thing is true of any statistical finding.

Manuscript submitted 30 October 1978
Final manuscript received 3 April 1979

REFERENCES


