1

WHAT IS HETEROSKEDASTICITY AND WHY SHOULD WE CARE?

For concreteness, consider the following linear regression model for a quantitative outcome (y_i) determined by an intercept (β_1), a set of predictors ($x_2, x_3, \ldots, x_K$) and their coefficients ($\beta_2, \beta_3, \ldots \beta_J$), and a random error ($\varepsilon_i$):

$$y_i = \beta_1 + \sum_{k=2}^{K} \beta_k \, x_{ik} + \varepsilon_i, \; for \, i = 1, 2, 3, \ldots, N \tag{1.1}$$

or in matrix notation,

$$\underline{y} = \mathrm{X}\underline{\beta} + \underline{\varepsilon}$$

where $\underline{y}$ is an $N \times 1$ column vector of the values of the outcome, X is an $N \times K$ matrix whose columns are the values of the predictors,[1] $\underline{\beta}$ is a $1 \times K$ column vector of the coefficients, and $\underline{\varepsilon}$ is an $N \times 1$ column vector of the values of the error term.

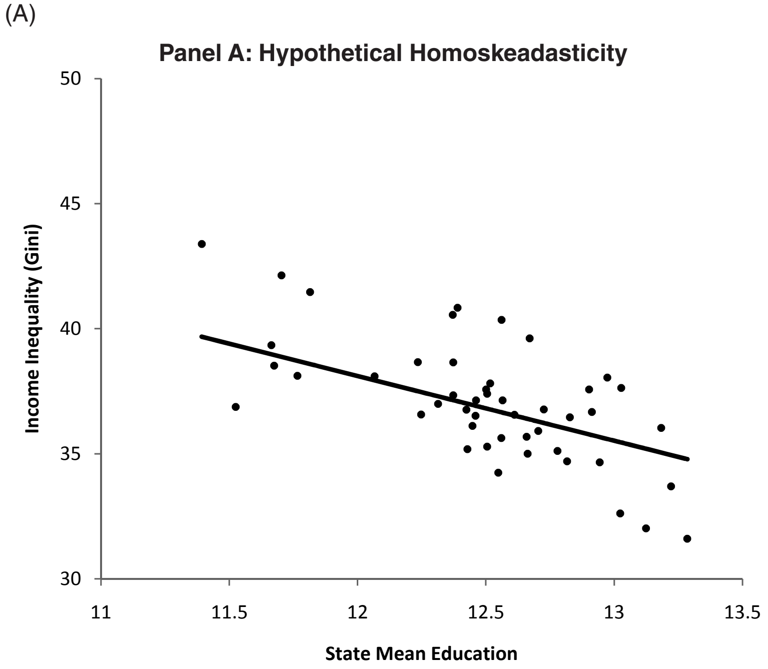
One of the usual ordinary least squares (OLS) assumptions is that the variance of ε_i is constant (see QASS # 50 by Berry & Feldman, 1985, for a good discussion of all of the OLS assumptions): $\mathrm{Var}(\underline{\varepsilon}) = \sigma^2 I$. Heteroskedasticity means that this assumption is incorrect. Instead, the error variance differs in some systematic fashion across the cases in the analysis. As I discuss in more detail later in this chapter, there are three common and easily identifiable situations in which social science researchers should strongly suspect the potential for heteroskedasticity: (1) when analyzing an aggregate dependent variable, (2) when making comparisons among social groups, and (3) when the distribution of the dependent variable is far from a symmetric and bell-shaped distribution. One easily recognized instance of the last situation is the analysis of a dummy dependent variable, which is described in many textbooks. Although it is certainly true that an OLS model for a dummy outcome will have heteroskedastic errors, logistic or probit models are more appropriate for such analyses (see, e.g., Greene, 2008, pp. 772–773; Hanushek & Jackson, 1977, pp. 181–189). Similarly, generalized linear models (GLMs; briefly discussed in the concluding chapter) are suitable for the analysis of

[1] By convention, the values in the first column of X are all equal to 1 and the corresponding coefficient in $\underline{\beta}$ is the intercept term.

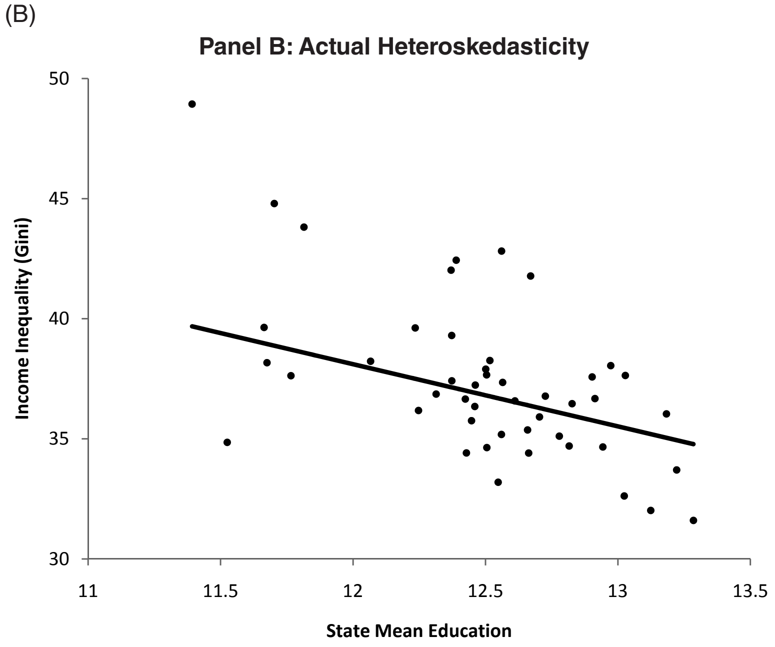
outcomes with distributions such as the Poisson for count models or the exponential for proportional hazards models. Other forms of heteroskedasticity are possible (e.g., a function of the square of the dependent or an independent variable), but these are somewhat less readily apparent in an analysis.

As an example, suppose you are interested in understanding why income inequality varies across the states of the United States. One explanation might be the differences across states in their stock of human capital; specifically that inequality declines as the level of mean education rises. Figure 1.1 shows the scatterplot of state-level income inequality by state mean education as well as the regression line, with Panel A illustrating homoskedasticity and Panel B demonstrating the actual heteroskedasticity.[2] Note that in Panel A,

Figure 1.1 Plots showing (A) Hypothetical Homoskedasticity and (B) Actual Heteroskedasticity for State-Level Analysis of Income Inequality Regressed on State Mean Education



[2] The data in Panel B were constructed using the data for application Example 1 described at the end of this chapter. The data in Panel A are an idealized modification of these data to portray the lack of heteroskedasticity. An outlier is excluded from the plots and calculations in this chapter because it exaggerates the differences between the various estimates of the coefficient variance.



the data points are similarly scattered around the regression line for all levels of education as is expected when the data are homoskedastic. In Panel B, the data points are scattered farther from the regression line for lower levels of mean education but closer to the regression line at higher levels of mean education. Such systematic patterning of the *y* values around the regression line is potentially indicative of heteroskedasticity. But as I discuss in the next chapter, the reason that the error variance differs across cases is not necessarily (or even likely) due to states' levels of mean education.

Consequences of Heteroskedasticity

If there is heteroskedasticity, the good news is that using OLS to estimate Equation 1.1 provides unbiased estimates of the coefficients. But it also creates two different problems: (1) the OLS estimates of the coefficients (β) are inefficient and (2) ignoring heteroskedasticity leads to biased estimates of the OLS standard errors in practice and hence biased statistical tests of the coefficients. Saying that OLS is inefficient means that there is an alternative unbiased estimator (the generalized least squares [GLS] estimator described in Chapter 5), whose coefficient estimates vary less from sample to sample than do the OLS coefficients. When the errors are heteroskedastic, the usual formula for calculating the variance of the OLS coefficients does not apply. Instead, it is given by (Greene, 2008, p. 150)

$$\mathrm{Var}(\underline{b}_{\mathrm{OLS}}) = \sigma^2 (X'X)^{-1} X'\Omega X (X'X)^{-1} \tag{1.2}$$

where $\sigma^2\Omega$ is an $n \times n$ diagonal matrix with the varying values of the variance of ε_i on the diagonal.

The coefficients' standard errors are the square roots of the diagonal elements of $\mathrm{Var}(\underline{b}_{\mathrm{OLS}})$. Similarly, a different estimator of σ^2 is also required (Theil, 1971, p. 256):

$$\hat{\sigma}^2 = \frac{(\underline{y} - \hat{\underline{y}})'(\underline{y} - \hat{\underline{y}})}{\mathrm{Trace}\,[\Omega] - \mathrm{Trace}[(X'X)^{-1} X'\Omega X]} \tag{1.3}$$

where the elements of $\hat{\underline{y}}$ are the predicted values of the outcome from the OLS regression.

A reasonable question to ask is whether the inefficiency of OLS as estimated using Equations 1.2 and 1.3 makes any difference in practical terms given that many analyses use large samples. The loss of efficiency of OLS is given by[3]

$$\sigma^2 [(X'\Omega^{-1}X)^{-1} - (X'X)^{-1} X'\Omega X (X'X)^{-1}] \tag{1.4}$$

To illustrate concretely the degree of relative inefficiency of OLS estimates, consider the case of a single predictor x_i, with both x_i and the dependent variable y_i measured as standardized variables with a mean of 0 and a variance of 1.0. Let ω_i be the error variance of the ith case (the ith diagonal element of Ω). The variances of the OLS and GLS coefficients and their ratio are given by (see the appendix for the derivation):

$$\mathrm{Var}(\underline{b}_{\mathrm{OLS}}) = \sigma^2 \frac{\sum_i \omega_i x_i^2}{(n-1)^2}$$

$$\mathrm{Var}(\underline{b}_{\mathrm{GLS}}) = \sigma^2 \frac{1}{\sum_i \frac{x_i^2}{\omega_i}}$$

[3] This formula results from subtracting the variance of the OLS coefficients given by Equation 1.2 from the variance of the efficient GLS estimator (described in Chapter 5) given by Equation 5.1.

$$\frac{\text{Var}(\underline{b}_{\text{OLS}})}{\text{Var}(\underline{b}_{\text{GLS}})} = \frac{\sum_i \omega_i x_i^2 \sum_i \frac{x_i^2}{\omega_i}}{(n-1)^2} \tag{1.5}$$

Thus, the degree of inefficiency of OLS depends on the squared values of the predictor x_i, the degree of heteroskedasticity as indexed by ω_i, the degree of association between x_i^2 and ω_i, and the sample size. For the example shown in Panel B of Figure 1.1, the ratio of the variance of the OLS coefficient to the variance of the GLS coefficient is 1.79. That is, the OLS coefficient variance is 79% larger than the efficient alternative estimator. This means that the coefficient standard error (which is the square root of the variance) is 34% larger in OLS than in GLS, and correspondingly that the confidence interval for the OLS coefficient is 34% larger. Such a loss of efficiency is more than large enough to affect conclusions about the significance or lack of significance of findings.

A second problem is what happens if a researcher were to ignore heteroskedasticity and use the OLS results as estimated by any standard statistical software. In this case, the variances of the OLS coefficients are improperly estimated. The problem is that the formulas used to calculate the variance of the OLS coefficients are not what are shown in Equations 1.2 and 1.3 (which takes the heteroskedasticity into account). Instead, what is used is the usual OLS formula for calculating the variances (standard errors) of the OLS coefficients, assuming a constant error variance (Greene, 2008, p. 48):

$$\text{Var}(\underline{b}_{\text{OLS}}) = \hat{\sigma}^2 (X'X)^{-1} \text{ and } \hat{\sigma}^2 = \frac{(\underline{y}-\hat{\underline{y}})'(\underline{y}-\hat{\underline{y}})}{N-K} \tag{1.6}$$

The difference between using Equation 1.6 and Equation 1.2 is not in general known in the sense that the coefficient variances as calculated by Equation 1.6 could be either larger or smaller than the correct values estimated using Equation 1.2. Consequently, the reported hypothesis tests are biased in unknown directions. In the single standardized predictor case, the ratio of the incorrect estimate of the coefficient variance to the correct estimate is given by (see the appendix for derivation of this formula)

$$\frac{\text{Incorrect Var}(\underline{b}_{\text{OLS}})}{\text{Correct Var}(\underline{b}_{\text{OLS}})} = \frac{\sum_i \omega_i}{\sum_i \omega_i x_i^2} - \frac{1}{n-1} \tag{1.7}$$

Whether this ratio is greater than or less than 1 depends on the degree of heteroskedasticity and the degree of association between the error variance and the square of the predictor. When there is relatively little heteroskedasticity, the value of this expression will be close to 1 (see the appendix). For the data portrayed in Figure 1.1, the incorrect estimate of the coefficient variance is 57% of the size of the correct estimate. Such a large underestimate creates the potential for a substantial error in drawing conclusions about the effect of mean education on income inequality.

Common Forms of Heteroskedasticity

A final reason to care about heteroskedasticity is that the potential for it to exist is more common than is usually recognized by researchers, although there has been a growing use of estimation methods in sociology which model heteroskedasticity in some fashion. For example, I reviewed all the 227 articles, comments, and replies in the *American Sociological Review* during a recent 5-year period.[4] I found that on average somewhat less than one article per issue (0.73) reported at least one analysis using some method of correction for heteroskedasticity. But this same review also indicated that much less attention is paid to this issue than the type of data and hypotheses analyzed would warrant. As I describe next, there are several readily identifiable scenarios in which a researcher should suspect heteroskedasticity. Table 1.1 tabulates the type of heteroskedasticity correction (if any) by two potential likely types of heteroskedasticity. In my assessment, nearly one third of all the articles (32.2%) included an analysis that fit into these situations. Of these, 38% ignored the potential for heteroskedasticity, 32% included some method of correction for heteroskedasticity, and the remaining 30% were indeterminate but most likely ignored the possibility of heteroskedasticity. Virtually all of the analyses in the latter category were multilevel models in which insufficient detail on modeling was provided to determine whether or not the model included a heteroskedastic specification. To me, this review suggests that about twice as many articles should be testing and possibly correcting for heteroskedasticity than currently do.

What are the situations in which the potential for heteroskedasticity should be readily apparent? Consider first the analysis of a set of units that are aggregations of (typically) a set of n_A individual cases within each unit A, such as a country or some smaller geographic unit. The dependent variable (y_A) in such analyses is often constructed as an aggregation of individual cases

[4] This review included all 227 articles, research notes, comments, and replies in the issues from October 2005 to August 2010.

Table 1.1 Type of Heteroskedasticity Correction in *ASR* articles ($N = 227$) by Type of Potential Heteroskedasticity and Estimation Technique

Type of Potential	*Estimation*	*Type of Correction*			
Heteroskedasticity	*Technique*	*None*	*HCCM*[a]	*EGLS*[b]	*Unknown*
Aggregate ($n = 25$)	OLS	6	7		
	Non-OLS	6	3	1	2
Social categories ($n = 48$)	OLS	11	10		
	Non-OLS	5	2		20
None apparent	Various	154			

Note: ASR, *American Sociological Review*; HCCM, heteroskedasticity-consistent covariance matrix; EGLS, estimated generalized least squares; OLS, ordinary least squares.

a. See Chapter 4.

b. See Chapter 5.

within the unit rather than a sui generis global property of the unit. It might take the form of an aggregate summary statistic:

$$y_A = f(y_{Ai}, n_A) \tag{1.8}$$

where $f(\ )$ is an aggregate function such as the mean or median, y_{Ai} is a characteristic of individual i in aggregate unit A, and n_A is the number of individuals in aggregate unit A.

Or it could be the form of an aggregate sum:

$$y_A = \sum_i y_{Ai} \tag{1.9}$$

An example of the former would be a city's median family income, while the number of refugees entering a country in a year would be an example of the latter.

In the case of an aggregate statistic, the variance of the outcome measure will be inversely proportional to n_A. If y_A is defined as the mean of individual cases, $y_A = \sum_i y_{Ai} / n_A$, then the variance of y_A is $\sigma^2_{yA} = \sigma^2_y / n_A$, where σ^2_y is the variance of y in the population of individuals. Similarly, if y_A is the median of individual cases, then it has variance

$$\sigma^2_{yA} \approx \frac{1.57\sigma^2_y}{n_A}$$

If y_A is defined as the proportion of individual cases with a particular characteristic c,

$$y_A = \frac{n \text{ for which } y_{Ai} = c}{n_A} = P_A$$

then the variance of y_A is

$$\sigma^2_{y_A} = \frac{P_A(1 - P_A)}{n_A}$$

In the case of an aggregate sum, the variance of y_A is directly proportional to the size of the aggregate unit, $\sigma^2_{yA} = n_A \sigma^2_y$. Although the fact that the variance of y_A varies with n_A across cases does not guarantee that the variance of ε_i does so in the same manner, it is likely that ε_i will be heteroskedastic. Of the 73 articles I identified as having the potential for heteroskedasticity, 34% were such aggregate analyses.[5]

The second situation in which the potential for heteroskedasticity is easy to identify is when the analysis involves a comparison among social groups (categories). Suppose that one of the predictors X_j is a dummy indicator for a case's membership in a social group (e.g., race/ethnicity or gender) or more broadly a contrast between two social categories (e.g., married vs. not married or employed vs. not employed). Such a situation suggests the potential for heteroskedasticity. Conceptually, if there is a process creating systematic differences in y_i between the two social categories, then the same process may also create random, nonsystematic differences in ε_i corresponding to the two social categories. A careful reading of the existing empirical literature in an area may sometimes indicate this. For example, it has long been argued in the labor market literature that there is less explained variation in the determination of rewards for minority groups than for nonminority groups in the United States (e.g., Featherman & Hauser, 1976, p. 636). This is tantamount to saying that the error variance is different for the groups:

$$\text{Var}(\varepsilon_i) = \sigma_1^2 \text{ for Group 1 and } \text{Var}(\varepsilon_i) = \sigma_2^2 \text{ for Group 2} \qquad (1.10)$$

Similarly, the literature might suggest that the range of outcomes is more restricted for one group than another. A common argument is that the sex segregation of occupations restricts the range of occupational

[5] Five of these were also characterized by the social categories form but were counted only as an aggregate analysis form to avoid double counting.

choices and hence the variation in the earnings of women (e.g., Reskin & Padavic, 1994, p. 46). Again, this suggests the potential for unequal error variance in earnings for men and women. The logic of this can be readily extended to more than two social categories, such as social classes or regions of a country. In the articles I reviewed, 66% of those with the potential for heteroskedasticity were such social group or category comparisons.[6]

The third common situation is when the distribution of the dependent variable is far from bell shaped. A highly skewed distribution or the presence of many extreme values (whether high or low) potentially could lead to different error variances across the range of values. Workers' earnings is an example of the former in which the magnitude of reporting errors is likely to be much larger for high earners than for low earners. Similarly, it is reasonable to argue that there could be greater random variation in a respondent's answers concerning assets and liabilities at the opposite ends of the net worth distribution (extreme wealth or high net debt) than in the middle parts of the distribution. Reporting errors (or rounding tendencies) in the tens of thousands of dollars seem highly plausible for someone with millions of dollars of assets or with hundreds of thousands of dollars of net debt but highly unlikely for someone with modest assets. Or consider the analysis of a dichotomous dependent variable that is guaranteed to be heteroskedastic as a function of the predictors. Avoiding heteroskedasticity is one of the reasons that logit or probit analyses are commonly preferred to OLS regression analysis of the dummy dependent variable (Greene, 2008, pp. 772–773; Hanushek & Jackson, 1977, pp. 181–189). I did not try to count how many would fall into this category because too many articles did not provide sufficient information to determine if the distribution of the outcome was unusual. And I did not count logit or probit analyses as potentially heteroskedastic (but see Allison, 1999; Williams, 2009; for a discussion of heteroskedasticity within logit and probit analyses).

There are certainly many other forms of heteroskedasticity that are often less readily apparent on a priori grounds, such as variance proportional to the square of the mean or multiplicative heteroskedasticity (Greene, 2008, p. 170). So how do we know if heteroskedasticity is an actual problem in a given analysis, and not just a potential problem? The next chapter presents diagnostic tools for determining the presence of heteroskedasticity in an analysis and applications of these tools to three examples. Each example illustrates one or more of the three common "obvious" situations. In the

[6] Five of these were also characterized by the aggregate analysis form and were counted only as an aggregate analysis form to avoid double counting.

next section, I describe the topic studied, the data source, and the measurement of the dependent and independent variables for each example. I have also posted on the web (www.sagepub.com/kaufman) a read-me file describing the contents of the posting and how to use them, a Stata do-file with two special-purpose programs (commands) that I've written, and a data file and its associated annotated Stata do-file for all of the analyses presented in this monograph.

Application Examples

Example 1: A State-Level Analysis of Income Inequality. The research question for this example is to explain why income inequality is higher in the Southern region of the United States than it is elsewhere. It requires an aggregate analysis because conceptually income inequality is an aggregate property of groupings of families or persons. This example uses states as the units of analysis, and I measure income inequality by the Gini coefficient calculated from the distribution of family income within a state. It is an attempt to determine if an initial significant difference between Southern states and other states in family income inequality (Gini) can be explained by state differences in the following:

- Human capital as measured by the average years of education completed by school-age adults in the state; education is expected to decrease income inequality
- Economic structure as measured by the prevalence of small businesses (proportion of the labor force in establishments employing 19 or fewer workers); this is predicted to increase income inequality
- Family structure as measured by the proportion of households headed by single mothers; this is hypothesized to increase income inequality

These variables were constructed by aggregating data from the March 1990 current population survey to the state level for the 48 coterminous states.[7] This data set was extracted from a larger pooled cross section created by a former graduate student of mine, Dr. Hyun-Song Lee, for his dissertation research, which he has graciously made available for instructional uses.

[7] Thus, the data exclude Alaska, Hawaii, and the District of Columbia.

Example 2: An Individual-Level Analysis of Voluntary Association Memberships. The second example uses persons as the units of analysis and focuses on understanding what factors can explain why people have greater or lesser connectivity (social bonds) to the larger society. The outcome measure is the number of voluntary associations to which a person belongs. There are three sets of predictors, all of which would be hypothesized to increase social bonds:

- Two indicators of socioeconomic status, years of education completed and self-reported social class rank on an integer scale from 1 = *lowest* to 10 = *highest*
- Two indicators of exposure to membership opportunities, population size of the place the person lives (in millions of persons) and whether or not the respondent has attended college coded as a dummy variable with 1 = *some college* and 0 = *no college*
- Two sociodemographic controls, age in years and sex coded as a dummy indicator with 1 = *male* and 0 = *female*

These measures were created from the 1987 NORC (National Opinion Research Center) General Social Survey, and the sample consists of 1,374 respondents with nonmissing values for all the variables.[8]

Example 3: A Household-Level Analysis of Wealth. The data for this final example are drawn from a more extensive published analysis of race and ethnic differences in wealth (Campbell & Kaufman, 2006). For simplicity, I use a subset of the predictors and restrict the sample to the white non-Hispanic households (*N* = 14,237). The dependent variable is household net worth, defined as the value of all assets minus the value of all debts held by household members. The predictors fall into four sets:

- Geographic location, measured by a dummy variable for metropolitan residence
- Household structure, measured by dummy indicators for type of household head (dual heads, male headed, female headed) and by the number of children (logged because its effect should be smaller at higher family sizes)

[8] This data set was originally constructed for another purpose with several additional variables, which might have affected the count of cases with nonmissing values.

- Life cycle stage, measured by age in years and its square
- Socioeconomic statuses, measured by dummy variables for educational credentials (less than high school, high school completion, some college, bachelor's degree, and postgraduate degree), dummy indicators of labor force status (in the labor force, retired, and not otherwise in the labor force), and monthly household income in $1000s

My coauthor and I constructed this data set from the 1992 panel of the Survey of Income and Program Participation using information from the Wave 4 core panel, the Wave 2 topical module on migration history, and the Wave 4 topical module on assets and liabilities. For details on measurement and sample selection see Campbell and Kaufman (2006, pp. 138–141).