Pak. J. Statist.
2021 Vol. 37(4), 315-367

WHEN WOULD HETEROSCEDASTICITY IN REGRESSION OCCUR?

James R. Knaub, Jr.
1890 Winterport Cluster, Reston, VA 20191-3624, USA
Email: jamesrknaub@gmail.com

ABSTRACT

In Ken Brewer's book, Brewer(2002), he made a strong argument for the ubiquitous presence of a specific range of the coefficient of heteroscedasticity when modeling for 'survey populations.' Basically, if each y-value were a cluster of smaller ones, the variance within the cluster would generally increase with larger predicted-y-values, in a systematic fashion. Thus the variance of Y_i, or of ε_i, tends to be larger for larger predicted-y-values. (Here we often use e_i as a substitute for ε_i.) So why don't we always see or discuss this heteroscedasticity? Here it is argued that there are several reasons: (1) Often it is not obvious, but it is there. (2) Often there is heteroscedasticity in the estimated residuals, but it is not considered important to the application. (3) It may be countered by model and/or data issues, which effectively reduce the coefficient of heteroscedasticity. Or (4) you may be looking at an example using artificial data designed to be homoscedastic. The point is that the dominance of OLS regression may be largely due to habit. Some informal papers have been placed on ResearchGate, under a "project" covering various issues related to this, as well as the estimation of the coefficient of heteroscedasticity in regression, which is another topic that was of interest to Ken Brewer, and sometimes may be used to definitively determine when there is heteroscedasticity and approximately to what degree. These issues are discussed here.

KEYWORDS

Cluster, coefficient of heteroscedasticity, measure of size, predicted-y, essential heteroscedasticity, achieved predicted-y, ideal predicted-y.

1. INTRODUCTION

Brewer(2002), mid-page 111, discussed in Knaub(2017b), shows predicted-y, and consequently y, as a conglomerate of small elements of predicted-y, plus their "errors," so that each y, and predicted-y-value, y_i^*,[1] is a cluster, and the resulting variance of the total

[1] y* is chosen here, as G.S. Maddala appeared to relatedly use it for the weighted least square predicted-y. See Maddala(2001). However, in practice here, it is used for different levels of heteroscedasticity, including γ=0, which would be for homoscedasticity, usually written as $\hat{y}$. Also, when estimating the coefficient of heteroscedasticity, γ, a preliminary predicted-y must be used, and the homoscedastic one is convenient. This is a size measure, and it is often referred to as z.

for the within cluster error of Y_i is approximately $e_0^2 y_i^{*2\gamma}$,[2] corresponding to equation 9.16, page 243 in Cochran(1977), S_w^2=AM^g, discussed below, where g is twice γ (gamma), and γ is the coefficient of heteroscedasticity. M in Cochran(1977) is a cluster size. Here, M, y_i^*, or z_i, may be used as a size measure. Ken Brewer showed why gamma ranges from 0.5 to 1.0, as historically observed (Cochran(1953), page 212). For the ratio estimator, which Ken and the author discussed, in practice you see this holds up very well. (At that point in the 1990s, the author tested this for electric power official statistics and found this range to generally be correct. Often, however, lower quality data for smaller establishments tended to artificially bring γ closer to 0.5, when sampling on a frequent basis, as it inflated the estimated residuals for those smaller respondents. See Knaub(2017a), slide 19.)

But for multiple regression, it appears that complexity makes it more difficult for predicted-y, as a size measure, to behave as predicted-y should with regard to heteroscedasticity of the Y_i, or the estimated residuals, e_i, as a substitute for the Y_i. Even the need for an intercept term can be problematic. Any model and/or data issue might impact heteroscedasticity, sometimes to enhance the effect, and sometimes to dampen it. Perhaps over-complexity might hide heteroscedasticity (Longley example in Section 8), and under-complexity might dampen it (arm strength example in Section 8). That would fall under point number 3 in the abstract. Model-unbiasedness might enter into consideration. However, there are other possibilities, as noted in points numbered 1, 2, and 4. Points 1 and 2 indicate that we often may just not notice it if it isn't glaringly obvious. Some examples are explored in Knaub(2019), and here we explore further.

Here is an excerpt from a response on March 13, 2021 to a question on ResearchGate: https://www.researchgate.net/post/What_is_the_best_way_of_selecting_predictors_for_a_regression_model_and_is_it_good_to_use_many_predictors_in_a_regression_model: "So, more predictors is not necessarily better or worse. The best set of predictors, working together, is best. (By the way, I think that when you have the best set of predictors, that you are also most likely to observe heteroscedasticity....)"

Knaub(2017b) and Knaub(2019) discuss the nature and magnitude of heteroscedasticity for regressions of form $Y_i = y_i^* + e_{0_i} z_i^{\gamma'}$, where γ' is for $V(e_i)$, where γ is for $V(Y_i)$, and written informally and approximately here as $y_i = y_i^* + e_{0_i} z_i^{\gamma}$, as a follow up to Brewer(2002), mid-page 111. Heteroscedasticity in regression is to be expected. Here we look at some examples to see where we find it, and speculate as to why it occurs, or does not occur.

[2] The general regression forms $Y_i = y_i^* + e_{0_i} y_i^{*\gamma}$ and $Y_i = y_i^* + e_{0_i} z_i^{\gamma}$ will follow, the simplest case being $Y_i = bx_i + e_{0_i} x_i^{\gamma}$, model-based ratio estimation. But this form will be approximate in an interesting way. Gamma, γ, is the coefficient of heteroscedasticity. Weisberg(1980), pages 100 and 101, notes the difference between $V(\boldsymbol{\varepsilon})$ and $V(\mathbf{e})$. γ actually is for heteroscedasticity of $V(\mathbf{Y})$. When we do not have model misspecification, $V(\mathbf{Y}) = V(\boldsymbol{\varepsilon})$, as illustrated in an example in Thompson(2012), on page 106. When part of the variance of $V(\mathbf{Y})$ is for the model, the coefficient of heteroscedasticity estimated for $V(\mathbf{e})$, say, γ', is slightly different from γ. Then $Y_i = y_i^* + e_{0_i} z_i^{\gamma'}$. However, the approximate form is usually employed here: $y_i = y_i^* + e_{0_i} z_i^{\gamma}$, dropping the "prime" mark, but it usually should be there. Note that the difference between γ and γ' may generally be too small to distinguish.

2. STATEMENT OF THE ISSUE: RETAIL EXAMPLE BY KEN BREWER

In the middle of page 111, in Brewer(2002), he provides an example, illustrating why we can expect that $0.5 \leq \gamma \leq 1.0$. (There appears to have been an editorial confusion at the end of the page with a comment about an exercise, but this is after the example and not needed.) Brewer's succinct example illustrates that not only should we see heteroscedasticity, but that an empirical form found for what Cochran(1977) calls "agricultural surveys," seen on page 243, is appropriate. The within cluster or within agricultural/crop plot variance shown there as equation 9.16, the "empirical formula" given as S_w^2=AM^g, is approximately equivalent to $e_0^2 y^{*2\gamma}$, or $e_0^2 z^{2\gamma}$ as noted in the introduction. Brewer(2002), page 111 states that $\sigma_i^2 \propto x_i^{2\gamma}$ where $0.5 \leq \gamma \leq 1$ (to be explained below), and points out that possible values of gamma were given at the end of section 5.2 (actually 5.3), where it said that "…for most business populations…," $\gamma = 0.75$ was often useful. For official energy statistics, the author found that would often be close, except that for a great many populations dealt with simultaneously, and frequently published, data quality problems from the smallest responders could artificially reduce the coefficient of heteroscedasticity (gamma, or γ) to 0.5 or even smaller, because estimated residuals were artificially inflated near the origin by the data quality issues. At the end of section 5.3, page 87, Brewer also said that for tree populations, $\gamma = 1$ appeared to be useful. He also encouraged very large sample sizes when trying to find an estimate of γ. However, one can imagine that to do better than the default most people use, $\gamma = 0$, when we should have $0.5 \leq \gamma \leq 1$, should not be difficult.

Knaub(2017b) goes through some detail to discuss why $0.5 \leq \gamma \leq 1$, and notes that in general, instead of x as the size measure, z, we can use predicted-y. In the case of a ratio estimator, predicted-y is bx. Because b is a constant, we can use x. But in general, we should use predicted-y, the closest to the best predicted-y that we can obtain. This expands the notion that heteroscedasticity should be "the rule" beyond just "sample survey populations," though that is not always apparent. Brewer(2002), page 111 shows how he explains his idea, which definitely is demonstrable for sample survey populations, and he writes further about this on pages 87, 126, 137, 142, 200, 201, and 203. This paper addresses when this might be expanded to other applications, but first, let us consider Brewer's illustration of his idea on page 111 as to why γ falls in the range that it does, which does not include 0, *i.e.* homoscedasticity is not an option:

Brewer said to consider retail stores of various sizes, where a larger unit (store) could be considered to be an aggregation of smaller elements (stores). If all the smaller elements acted independently within a larger unit, then Brewer argues that the variances would be additive, so $\sigma_i^2 \propto X_i$, where X_i represents the size of the cluster of smaller, independent stores. Brewer noted that $0.5 \leq \gamma \leq 1$ for $\sigma_i^2 \propto X_i^{2\gamma}$ is "commonly used" (page 111, referring to page 87), so this implies that the lower bound is $\gamma = 0.5$, since Brewer noted that no smaller value for γ could be reached unless the elements within the cluster differ from each other more than from elements of other clusters. Brewer also argues, all in the middle of page 111, that an upper bound would be $\gamma = 1$. Included in his argument is the thought that a case where σ_i increases faster than X_i does not seem plausible. (Consider $\sigma_{\epsilon_0}^2 x_i^{2\gamma}$. From Knaub(2017b), on page 3, we have approximately $\sigma_{e_0}^2 x_i^{2\gamma} = Var(y_i|x_i) =$

$Var[\sum_{j=1}^{x_i} y_{i,j}]$, and an argument for the two bounds is explained from there, though the argument for the upper bound is somewhat different.) Thus we think of each y-value as coming from a predicted-y sized cluster with accompanying estimated residual, where $\sigma_i^2 = \sigma_{e_0}^2 x_i^{2\gamma}$. (Consider the "gamma population model" in Chambers and Clark(2012), page 49.[3]) This paper considers extension to more complex models, where size measure z_i is no longer just bx_i or just x_i, but we strive for the best predicted-y or a good approximation to it, for our size measure. Note that when considering cluster sampling on page 243 of Cochran(1977), $z_i = M$ is the size measure in $S_w^2 = AM^g$. On page 111 in Brewer(2002), $Z_i = X_i$. Here $Z_i = y_i^*$ where y_i^* is predicted-y. It should be the best weighted least squares predicted-y, but there are some problems with that. In the process of estimating γ, we first use the homoscedastic predicted-y. (See Knaub(2019).) Further, below we will argue that one reason for not finding $\gamma \geq 0.5$ may be a problem with the selected model, and thus we distinguish the achieved predicted-y from the ideal predicted-y.

3. IMPLICATIONS

Ironically, using the homoscedastic predicted-y to find an estimate of γ (gamma) is usually going to make little difference, as the coefficient of heteroscedasticity does not usually change the predicted-y values very much, especially with larger sample sizes where there is better symmetry for y-values about the regression. Changes for some individual predictions could be substantial, but will have less impact on γ-estimation. However, it is proposed here that changes in the number and makeup of independent variables used, as well as other complexities in the model selection, may very well influence the value found for γ. That will be discussed. For now, consider the simplest case of a proportional relationship between y and one independent variable, x, with no intercept, a ratio-type model, where we will write, $y_i = bx_i + e_{0_i} x_i^{\gamma}$. Here predicted-y is bx_i, and since only relative size matters, we use x_i as the size measure. Cochran(1953), page 205, discusses measures of size in surveys. For a household survey, the cluster size, M_i, may be the number of people in a household. But for "Farms, banks, and restaurants ...," in other words, establishments, and perhaps other cases, Cochran notes that the best measure of size may be the same data "item" on a previous census survey. Thus x_i may be the same data item as measured by y_i, where the former is from a "previous census" and the latter is from a current sample. This works very well for $y_i = bx_i + e_{0_i} x_i^{\gamma}$, as demonstrated in cases found in Knaub(2017a). In such ratio model cases, γ is clearly in the range $0.5 \leq \gamma \leq 1$, except in the most extreme cases of data quality issues for the smaller respondents, which increase estimated residuals near the origin. But when data quality is good, having a size measure, x_i, to function in the role of the predicted-y-values, as bx_i, is nearly a perfect/ideal

[3] Chambers and Clark(2012), pages 49-52, discusses the "*gamma population model*," where we assume proportionality "…between Y and the size variable, Z…," and independent y-values. There, equations 5.2a and 5.2b, respectively are $E(y_i|z_i) = \beta z_i$ and $Var(y_i|z_i) = \sigma^2 z_i$. When $\gamma = 0.5$ we have the "*ratio population model*." β would then be such that we have the model-based classical ratio estimator, CRE, as reflected on page 126 in Brewer(2002). In Särndal, Swensson, and Wretman(1992), on pages 255-258, we see the general gamma population model presented as "alternative ratio models," of which the (classical) ratio model is one case (with $\gamma = 0$).

predicted-y. Then γ is often in the range of 0.7 to 0.9. Using $\gamma = 0.5$ appeared robust for many cases found in Knaub(2017a). Results were generally not the best, but good, and in cases where data quality became problematic, useful when a great many small populations with small samples were considered on a frequent basis, using establishment surveys for purposes of producing official energy statistics. (It is very useful, as noted in Cochran(1953), at the bottom of page 205, to be able to use a different size measure, x_i, for each y_i, as opposed to a "general size measure," which makes prediction so much better than unequal probability sampling for a multipurpose sample survey.) This is the type of situation where we have highly visible evidence that indeed it is true that $0.5 \leq \gamma \leq 1$. So why would anyone assume homoscedasticity, i.e., $\gamma = 0$, for some other situation, when clearly we should have heteroscedasticity any time that predicted-y values vary? That is, we should have heteroscedasticity unless our model is the common mean model, as in Särndal, Swensson, and Wretman(1992), page 258-260, and also shown in Chambers and Clark(2012), page 20, "A Model for a Homogeneous Population," where the y-values are independent, and equations 3.1a and 3.1b show that $E(y_i) = \mu$, and $Var(y_i) = \sigma^2$. There $\gamma = 0$ because the predicted-y are all identical, but elsewhere, it makes no sense.[4] Yet homoscedasticity is often insisted upon, and may exist. How can that be? We will look at some examples later.

4. SUMMARY OF REASONING FOR ESSENTIAL HETEROSCEDASTICITY

In Brewer(2002), on page 111, Ken Brewer showed why there should not be homoscedasticity in "sample survey populations," but instead we should have $0.5 \leq \gamma \leq 1$, and $\sigma_i^2 \propto x_i^{2\gamma}$. Cochran(1953), page 205 says that "…the best measure of size…" is often the same "item" from a previous census survey, x_i. The efficacy of this is certainly supported by Knaub(2017a), where this is and has been applied to massive numbers of small populations for the production of official energy statistics. Here we wish to show that the same logic extends to more complex regression, in general, but that certain other issues may arise. Examples will be considered to demonstrate these issues. The logic which Ken Brewer produced, which works well for $y_i = bx_i + e_{0_i}x_i^{\gamma}$, also extends to the more general case[5] $y_i = y_i^* + e_{0_i}y_i^{*\gamma}$, whose better working form might be $y_i = y_i^* + e_{0_i}z_i^{\gamma}$. Heteroscedasticity here is such that every predicted-y value, being a cluster of infinitesimal elements, is 'essential,' owing only to the difference in sizes of predicted-y-values. See Knaub(2017b) regarding "essential heteroscedasticity," and Knaub(2018) with regard to other influences on heteroscedasticity, data quality issues, and model issues, some enhancing and some degrading for γ. Here we will discuss how model issues can actually

[4] Särndal, Swensson, and Wretman(1992), Section 7.3.3. "Optimal Sampling Design for the π Weighted Ratio Estimator," on page 254 tells us that simple random sampling with a ratio estimator is most efficient if $\gamma = 0$, not the classical ratio estimator, where $\gamma = 0.5$ here. Thus, if x is a good predictor, we should not use simple random sampling with a classical ratio estimator, yet that has historically been the norm.

[5] Note that $e_{0_i}x_i^{\gamma}$, $e_{0_i}y_i^{*\gamma}$, and $e_{0_i}z_i^{\gamma}$, would each be divided by $\sqrt{1-h_i}$, with the appropriate hat-value, h_i, in each case, depending upon the term representing the model variance part of $V(Y_i)$. See hat-values, Section 5.7.

impact the quality of predicted-y regarding its ability to cause the associated estimated residuals to mimic the behavior that they ideally should. Thus they actually impact essential heteroscedasticity as well.

Accordingly – remembering that $\gamma\ should\ really\ be\ \gamma'$:

For all population y-values for a given item, consider

$$y_i = y_i^* + e_{0_i} y_i^{*\gamma}$$

y_i^* *would be the "ideal" predicted value. It is a measure of size.*

In practice, however, we may have inferior measures of size, but in any case, the measure of size is often designated as z. Thus,

$$y_i = y_i^* + e_{0_i} z_i^{\gamma}$$

For the simplest case,

$$y_i = bx_i + e_{0_i} x_i^{\gamma}$$

In that case, b is absorbed in the e_0 *in the estimated residuals. We can say* $z = x$ *instead of* $z = bx$ *here, because size is only relative.*

Any predicted y, *say* y^*, *size measure can be considered to be a cluster of smaller parts, and thus within cluster variance accounts for the variance behavior of* Y_i, *and approximately that of the estimated residuals. The size measure corresponds to M when looking at cluster variance behavior.*[6] *Therefore, any time there is more than one predicted-y value, thus more than one measured size, which would be the norm, there is heteroscedasticity.*

Note: A single predicted-y is compared here to a cluster in cluster sampling, or an entire agricultural crop plot. That is an interesting transfer of application, but one that has long been known. See Cochran(1953), pages 199 and 212.

5. OTHER CONSIDERATIONS

There are various interactions between variables in multiple regression, and model and data complexities, which might interfere with the straightforward interpretation of predicted-y as a cluster of small elements. Here we will mention some thoughts on this. The examples will demonstrate some of them. Some or all of these issues may possibly change the variance related nature of predicted-y, as in the case of omitted variables in the "arm strength" example in Section 8.

[6] In Cochran(1977), on page 234, under "Single-Stage Cluster Sampling" for clusters of equal size, he lets "$M_u =$ relative size of unit." Then on page 238 he says that one may divide a "survey unit" "…into M smaller units…." When he gets to unequal size cluster, he uses M_i, starting on page 249. Here we use x_i, or y_i^*, or z_i, in place of M_i.

5.1 e Not Independent of x

Because we should expect essential heteroscedasticity, the estimated residuals are a function of the predicted-y values, so for any x, e is not independent of x, contrary to the usual assumption. This has implications. For example, the way that omitted variables may bias coefficients. On pages 111 and 112 of Fox(2008), we see bias in a regression coefficient possible when a correlated variable is omitted. However, the argument is written assuming homoscedasticity, but the residuals are not independent of the 'independent' variables. Results there are then altered.

5.2 Omitted Variables

In Fox(2008), at the top of page 274, Fox notes that an omitted variable which is categorical, such as urban versus rural, could mean a different interaction with another variable which has a different slope for urban as opposed to rural, and without that categorical variable, there would be the appearance of heteroscedasticity. That would be an example of nonessential heteroscedasticity. However, another concern regarding omitted variables that would mean a dampening rather than enhancement of heteroscedasticity is a concern for one of the examples to be shown below (Penn State(2021a), alcohol-arm strength).

In that example, given later, the dependent variable is arm strength, and the only independent variable, aside from an intercept term, is a certain measure of alcohol consumption among alcoholics. But the independent variable hardly seems adequate. As Brewer(2002) points out on pages 109 and 110, and is a considerable topic in Hastie, Tibshirani, and Friedman(2009), adding variables can increase variance. Brewer(2002) noted that if added, such a variable should have high "explanatory power." Here it would seem that too much is unexplained just by the independent variable used, and ideally, more information is required. Age of the person whose alcoholism is being measured, as well as how long it took to reach that level, *i.e.* years as an alcoholic, might vary more greatly for the middle levels of the size measurement. That could explain why, if anything, sigma of estimated residuals seems larger in the middle of the graph, horizontally speaking, than at the ends, where sigma is considered in the vertical direction. Thus, the predicted-y here does not behave as the gold standard. Because the predicted-y, as a size measure, could be improved, the natural (essential) heteroscedasticity that should occur has been masked. Thus the absence of heteroscedasticity of the estimated residuals is problematic. Generally, it is proposed here that if the predicted values are good enough, one should have heteroscedasticity. (Note: This example, borrowed with permission, was only intended as a simple exercise for students to first learn regression analysis. We take it a step further here.)

5.3 Model-Unbiasedness

At the bottom of page 158 in Cochran(1977), he states that in Brewer(1963) and Royall(1970), the concept of 'model-unbiasedness' is considered. This means that the expected sum of the estimated residuals should be zero. (In Särndal, Swensson, and Wretman(1992), on pages 231 and 232, they show under what conditions we will have the sum of the estimated residuals be exactly zero, not just in expectation. Also see Brewer(2002), near the top of page 111.) One might think that an intercept term is always needed to have model-unbiasedness, but that thought may implicitly assume OLS (homoscedastic) regression. This is an example of how the assumption of homoscedasticity permeates statistical (economic, and perhaps other) literature, but here we note that that is not always a good idea, and actually should generally be incorrect. When we allow heteroscedasticity, there is much more flexibility.

Please see my last response (March 7, 2021) to
https://www.researchgate.net/post/What_to_do_In_Linear_regression_model_intercept_is_not_coming_significant_residual_error_mean_is_non_zero_for_no_intercept_model

5.4 Beyond Collinearity

Moderator, confounding, and suppressor variables, any correlation between "independent" variables in multiple regression, complicates how they work together to arrive at the predicted-y. Suppressor variables seem particularly interesting, as no one would use a variable uncorrelated with the dependent variable as a sole predictor, but when it enhances another predictor's relationship to y, that sounds useful. In Ludlow and Klein(2014), they consider the usefulness of introducing a suppressor variable based on theoretical considerations, as opposed to just discovering that one of your variables performs that way.

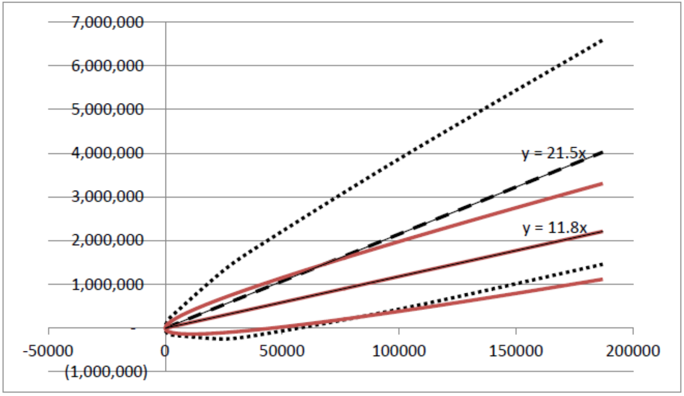
5.5 Stratification (Chambers and Clark, Section 5.5)

This is when one needs to be careful that results from regression do not appear heteroscedastic, or more heteroscedastic than they really are, because you have data that should have been considered in separate groups, strata, or subpopulations, depending on whether you are looking at small area estimation, overall results, or subpopulation results. Therefore, instead of one regression for the entire data set, we could have separate regressions by region, as in Chambers and Clark(2012), pages 49-58. In Knaub(2012), pages 12 and 13, we see that if we are unsure that separate regressions should be used, we can put prediction interval bands about the regression lines. Figure 3 from page 13 is shown here. Note that the estimated slopes, 21.5 and 11.8, are quite different, but there is not enough data to be sure of that, as the 90-percent prediction intervals strongly overlap, with one set of bounds almost completely within the other.

Excerpted from Knaub(2012):

Figure 3.

Enhancement of a figure like Figure 1. Here we have 90% Prediction Bounds.



(Notice also that near the origin, the prediction intervals dip below zero for y, which indicates the need for asymmetric prediction intervals.)

Also, see the graph in the example, “Kenya,” in Section 8.

5.6 Autocorrelation (Aside)

This paper is not about time series, but rather finite populations, and basically does not consider autocorrelation, though spatial autocorrelation could be possible. (See the end of this subsection for a comment regarding the Spanish shops example of Section 8.) However, one of the examples in Section 8, the Longley employment data example, is one where OLS is assumed, but the y-data do come from different years. One might assume year does not matter, that the mechanism which produces the y-data is identical to one producing a finite population, but the full model does specifically have a small contribution by year, and Faraday(2002) below, does consider autocorrelation, even for a reduced model that does not include the one independent variable specifically about the year. We do take an ancillary glance at this, knowing that the Longley data are at least technically a short time series.

Faraway(2002), pages 59-62, used the Longley data, but the only predictors used were GNP, population, and the intercept. These variables are said to be highly correlated, adjusting the large intercept up and down, in a reduced view of the up and down adjustments we see in the full linear regression model. There is high variability in coefficients, and when also considering autocorrelation, that makes a substantial difference as well. Further rho is highly variable, so there may be substantial autocorrelation in that reduced model, or none at all. The residual standard error is very uncertain and grows with autocorrelation. He shows a 95-percent confidence interval for “residual standard error” from 0.24772 to 1.91748, with a point estimate of 0.68921. Previously, without autocorrelation, the point estimate was 0.546. (Apparently Faraway worked in different units, perhaps by a factor of 1000. A check of the data showed this to make sense.)

When you look at the estimated residuals in the full model, OLS example below, their absolute values range from 16 to 440. The sigma (residual standard error) is about 169. Thus we are comparing 169 in the full model to 546 in the reduced model, as we are likely fitting the full model too tightly to a sample of only 16 data points.

The y-values range, from highest to lowest, by a factor of only about 1.17. One would expect a similar range for predicted-y, whatever model was used. So instead of a constant residual standard error, one would expect it to range (on a graphical residual analysis scatterplot, from left to right) by a factor as little as 1.17^0.5 = 1.08, to as much as 1.17, if $e = e_0 y^{*\gamma}$. Among 16 data points, with so much relative variability, a tendency for a 10- to 15-percent change from left to right may not be apparent when we have estimated residuals ranging from 16 to 440.

Whether or not there is autocorrelation here, which is not specifically considered in this paper, adherence to the range for the coefficient of heteroscedasticity given by Brewer for sample survey populations may be partially hidden due to the relatively small impact for the short y-range, and may be partially dampened by the substantial difference between achieved predicted-y and ideal predicted-y, possibly at least partially due to too few degrees of freedom.

As for spatial autocorrelation, the example of Spanish shops, Guadarrama, Molina, and Tille(2020), used a random intercept in an attempt to accommodate regional differences. However, since what was being measured was estimated by use of a ratio, perhaps a random slope would be better. But if we could obtain a different slope in each region, we could stratify that way. At any rate, that example showed substantial heteroscedasticity, adjusting for the intercept terms so that all estimated residuals could be considered together. The article is quite interesting.

5.7 Hat-values Used to Adjust Graphical Residual Analyses

For estimation of the coefficient of heteroscedasticity, γ, one suggestion which Ken Brewer made was basically something this author later found suggested independently in Carroll and Ruppert(1988). Further, Ken suggested obtaining a standard error for γ. Ken Brewer noted to the author that by taking the log of both sides of $y_i - y_i^* = e_{0_i} y_i^{*\gamma'}$, one could estimate and find a standard error for γ'. If we plot $log|y_i - y_i^*|$ on the y-axis and $log y_i^*$ on the x-axis, then a simple linear regression will have slope γ', and one can even estimate the standard error of γ', as suggested by Dr. Brewer. (Actually, this author had previously only written $y_i = y_i^* + e_{0_i} y_i^{*\gamma}$, and Ken may have thought studentized residuals were meant, or that it mattered little because he suggested very large sample sizes were needed.) Note that there are examples, such as the one on pages 49 and 50, in Carroll and Ruppert(1988), where they considered leverages, not considered in Knaub(2019), to account for the sample size and its distribution, with regard to a model, when estimating γ. Similarly Carroll and Ruppert(1988), on page 49, for instance, have a scatterplot with “Log Absolute Residual” on the y-axis, and “Logarithm of Predicted Value” on the x-axis. They justify this on pages 12 and 89 by noting that when variability is a ‘function of the mean,’ it is usual to give a ‘standard deviation’ as $\sigma\mu_i(\beta)^\theta$, where θ is our γ, and μ_i is size measure, z_i (or y_i^*).[7] Taking logs, they get $log\sigma_i = \theta log\mu_i + log\sigma$, where again the slope is $\theta = \gamma$. However, the label “Log Absolute Residual” used in Carroll and Ruppert(1988) apparently means log absolute studentized residual. They use studentized residuals, as is also used in graphical residual analyses in Fox(2008), on page 273, to account for the sample size and predictor(s) distribution(s) impact on estimated residuals, and to make their distributions more like a t-distribution. The hat-value is used to account for leverage. However, in the large sample sizes Ken was considering, as large as the forestry example in Section 8 below, this would not have been necessary. In some cases, however, the hat-value will have a greater presence. However, it seems unlikely that use of the hat-value will make a practical difference in estimating γ, and could be messy for some applications. Though situations will differ, see Section 8.6 for an example with a sample size of $n = 9$.

[7] To complete the comparison: The within cluster variance from Cochran(1977), page 243, and Cochran(1953), page 199 is $S_w^2 = AM^g$. (In scatterplot labels, g will mean γ, not this g.) In regressions here σ_i^2 is $\left(e_{0_i} z_i^\gamma\right)^2$ or $\left(e_{0_i} y_i^{*\gamma}\right)^2$. From Carroll and Ruppert(1988), page 12, equation 2.5, the model declared “most common” when “variance depends on the mean,” gives us the following variance: $\left(\sigma\mu_i(\beta)^\theta\right)^2$. Therefore, $AM^g \equiv \left(e_{0_i} z_i^\gamma\right)^2 \equiv \left(\sigma\mu_i(\beta)^\theta\right)^2$, so $g \equiv 2\gamma \equiv 2\theta$. γ is the coefficient of heteroscedasticity, so θ is also, and $\frac{g}{2}$ is equivalent. The size measures are M, z_i, y_i^*, and μ_i.

For small sample sizes it could be problematic to ignore the hat-value, but it was not really a problem there.

For simple linear regression, the hat-value, h_{ii} or h_i, is, from Fox(2008), page 245,

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

Leverage is usually discussed, as in Penn State(2021c), as a diagnostic tool. It helps determine where an extreme y-value would be most influential, were you to have one, based solely on the model predictor (x) values, and the intercept value. One might compare the hat-value, h_i, pages 244 and 245 in Fox(2008), for simple linear regression, to the square root of the estimated variance of the prediction error such as that found in Penn State(2021b), also for simple linear regression, with $\gamma = 0$, which they call the "standard error of the prediction error," and they write it as follows:

$$\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$$

Here they have MSE = $\hat{\sigma}$ as an estimate of σ, using the estimated (raw) residuals. The estimated variance of the prediction error becomes the "expected prediction error," EPE, if you add the square of the bias due to any model misspecification. Penn State(2021c) notes the difference between $MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)$, the estimated variance of the prediction error, and $MSE\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)$, the variance of the means. The difference involves only MSE, as an estimate of σ^2. Therefore $(MSE)\ h_i$ accounts, approximately,[8] for the variance due to the model, for the given sample. Recall that this is for one predictor, an intercept term, and homoscedasticity. (The middle of the one page of notes found at Lind(2004), for simple linear regression, may add clarity.)

Both the standardized and studentized residuals (see Penn State(2021d,e)) result from dividing the raw residuals by a measure of standard error which is the square root of an estimate of the variance of Y, from which was *subtracted* an estimate of the variance due to the model. This seems odd, but is the result of the variance of the estimated residuals being an underestimate of σ^2. Here we see it is underestimated by the amount assigned above to the model variance: $\sigma^2 h_i$, estimated as $(MSE)\ h_i$, also an underestimate with a finite sample. Also note that the studentized residuals delete the i^{th} case to better achieve a t-distribution.

Here is some reasoning: A derivation comparing $\boldsymbol{e}$ and $\hat{\boldsymbol{e}}$ is given in Weisberg(1980), pages 100 to 103. There he used **V**, apparently before **H** came into use, as well as the name, "hat matrix." As noted in Weisberg(1980) and elsewhere, this means that our estimated residuals have different variances, even when $V(\mathbf{Y})$ is homoscedastic, and also are

[8] By including the estimated variance due to the estimated regression coefficients in the estimated variance of the prediction error, we multiply the expression for the estimated variance of the means by a factor of $(1 + h_i)$. However, using MSE could be corrected by dividing by $(1 - h_i)$. Note that multiplying by $(1 + h_i)$ is slightly less of a factor than dividing by $(1 - h_i)$.

correlated. In place of his e_i and $\hat{e}_i$,[9] we will use ε_i and e_i, respectively. On page 100 Weisberg has $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, and $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$, and from page 100 and 101, and using **H** for **V**, we can see that $V(\mathbf{e}) = V(\mathbf{Y}) - V\left(\mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \sigma^2(\mathbf{I} - \mathbf{H})$. In Thompson(2012), on page 106, we see a special case, for a ratio estimator, where he has $\mathrm{E}(Y_i) = \beta x_i$, that $\mathrm{var}(Y_i) = v_i\sigma_R^2$. (We can use $v_i = x_i^{2\gamma}$.) In general then, for $\gamma = 0$, $V(\mathbf{Y}) = \sigma^2\mathbf{I}_n = \mathrm{var}(\boldsymbol{\varepsilon})$, and the variance due to the estimated model coefficients, $V\left(\mathbf{X}\widehat{\boldsymbol{\beta}}\right) = V\left(\widehat{\mathbf{Y}}\right)$, is *subtracted* from $V(\mathbf{Y})$ to obtain $V(\mathbf{e})$.

Using $Y_i = \beta x_i + \varepsilon_i$ as Thompson(2012) does on page 105 for a ratio estimator, we have $var(Y_i) = \sigma_i^2$. Then for $Y_i = \beta^* x_i + e_i$, $var(Y_i) = var(Y_i^*) + var(e_i)$, so $\sigma_i^2 = var(Y_i^*) + var(e_i)$, which means that $var(e_i) = \sigma_i^2 - var(Y_i^*) = \sigma_i^2(1 - h_i)$. Thus $\sigma_i^{*2} = \sigma_i^2(1 - h_i)$, and $\sigma_i = \sigma_i^*/\sqrt{1 - h_i}$.

$V(\mathbf{Y}) = \mathrm{var}(\boldsymbol{\varepsilon})$ is for $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ where there is a fixed $\mathbf{X\beta}$, which is "correct." When we have $\mathbf{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{e}$, we have an estimate for the model parameters, so now there is a variance contribution from the model.

To examine the use of the hat-value, h_i, here, let us look for a moment at $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, and $\mathbf{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{e}$. Let us say $\mathbf{Y_I} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ and $\mathbf{Y_{II}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{e}$. So when is $V(\mathbf{Y_I}) = V(\mathbf{Y_{II}})$? That would mean $V(\boldsymbol{\varepsilon}) = V\left(\mathbf{X}\widehat{\boldsymbol{\beta}}\right) + V(\mathbf{e})$, so $V(\mathbf{e}) = V(\boldsymbol{\varepsilon}) - V\left(\mathbf{X}\widehat{\boldsymbol{\beta}}\right)$ means that $V(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$. Therefore, for any adjustment such as $\sigma_i = \sigma_i^*/\sqrt{1 - h_i}$ to hold, the model must be perfectly specified.

Consider the model-based approach to survey sampling (see Royall(1992) for an overview), where we often may use $Y_i = bx_i + e_{0_i} x_i^{\gamma'}$ for model-based ratio estimation. See, for example, Thompson(2012), pages 105-109, where predictions are based on $\mathrm{var}(Y_i) = x_i^{2\gamma}\sigma_R^2$, and an unbiased[10] estimate is given as $\hat{\sigma}_R^2 = \frac{1}{n-1}\sum_{i=1}^{n}\frac{\left(Y_i - \widehat{\beta}x_i\right)^2}{x_i^{2\gamma}}$. (Technically, it becomes nearly unbiased with large enough n.) More complex models are used for surveys, but the idea is that the σ^2 for the Y_i, or better, the σ_i^2 for the Y_i, are for Y_i, not e_i. So, for example, for ratio estimators, when we estimate σ_R^2 by using $\hat{\sigma}_R^2$, say for $\gamma = 0.5$, we underestimate, as seen in pages 131-133 in Valliant, Dorfman and Royall(2002). However, note that besides the leverages shown in factors in alternatives provided there, $\frac{1}{n-1}\sum_{i=1}^{n}\frac{\left(Y_i - \widehat{\beta}x_i\right)^2}{x_i}$ is not the same as $\frac{1}{n\bar{x}_s}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}x_i\right)^2$, where $\bar{x}_s$ is the mean of the x_i's associated with the sample, so there is that alteration (see footnote 14). Using $\frac{1}{n-1}\sum_{i=1}^{n}\frac{\left(Y_i - \widehat{\beta}x_i\right)^2}{x_i}$ with the hat-value adjustment would be more exact.

[9] Also, instead of u_i and $\hat{u}_i$, as in Maddala(2001), pages 64 and 65.

[10] That this is an unbiased estimate is confirmable from pages 70 and 76 in Maddala(2001). The plot in Penn State(2021f) helps as a reminder of when we are looking at the conditional distribution of $Y_i|X_i$, and when we are considering the unconditional distribution. Here we have $Y_i - \hat{\beta}x_i$, where we consider the conditional distribution, with often relatively little variance from the estimated model coefficient.

It should be noted that Carroll and Ruppert(1988), page 31, states that it is possible, without the adjustment due to leverage, for these graphs to appear heteroscedastic, even when there is homoscedasticity.[11] At any rate, a comparison is made in an example in Section 8.6, on a segment of residential electric sales, where leverage and heteroscedasticity are considered, with a very small sample where it should matter if leverage is going to matter, but it made little difference to the estimate of γ. Also, what appeared to be a large, possible "outlier" was completely removed, as an experiment, in the example of Section 8.12 on baseball payrolls, where there was again a small sample size such that it could matter, and again, conclusions about γ, for practical purposes, were barely altered. Potential outliers could matter, however, especially ones with larger leverage. When estimating γ, perhaps they should be removed to better capture a likely more accurate estimate of γ, though actually deleting the data should take a thorough review. One should be careful not to easily completely throw away inconvenient data that may be from the tails of a distribution, or may tell us something about the distribution that we did not know.

On page 31 of Carroll and Ruppert(1988), we see their decision to use studentized residuals. (They noted that Cook and Weisberg[12] suggested studentized residuals in graphical residual analysis scatterplots as an improvement.) In Penn State(2021d), we see that the square of the standardized residuals, using $\hat{\sigma}^2$ for MSE, are $\frac{e_i^2}{\hat{\sigma}^2(1-h_i)}$, so we divide the e_i by $\hat{\sigma}\sqrt{(1-h_i)}$ to standardize in such a way that the “design,” as noted by Carroll and Ruppert(1988), page 31, is given consideration. Using $h_i = \frac{1}{n} + \frac{(X_i-\bar{X})^2}{\sum_{j=1}^{n}(X_j-\bar{X})^2}$ for simple linear regression, and thus $\gamma = 0$, and there is an intercept term, does this. The $\frac{1}{n}$ part is in deference to the intercept term, and $\frac{(X_i-\bar{X})^2}{\sum_{j=1}^{n}(X_j-\bar{X})^2}$ is due to the single predictor term.

However, in general, *i.e.*, when γ can vary (*i.e.*, when we can have heteroscedasticity, whereas $\gamma = 0$, homoscedasticity, is just a special case), and when we can have any number of predictors, we can think of the hat-value as the factor of the variance due to the design that is multiplied by σ_i^2 to thus represent the variance due to the model coefficients.

Consider the estimated variance of the prediction error for $y_i = bx_i + e_{0_i}x_i^{\gamma}$, for an individual prediction, from slide 16 in Knaub(2017a). The estimated variance due to the estimated model coefficients is $x_i^2V^*(b)$. From Maddala(1977), pages 259 and 260, $V^*(b) = \sigma_{e_0}^{*2} / \sum_{i=1}^{n} x_i^{2-2\gamma}$. We subtract that from $x_i^{2\gamma}\sigma_{e_0}^{*2}$ to generalize $\sqrt{\hat{\sigma}^2(1-h_i)}$ for γ,

[11] However, the direction of the heteroscedasticity may not be as usual. Consider Weisberg(1980), pages 100 – 104. For simple linear regression, the variances of the estimated residuals would become smaller as one travels further from the mean of x, in either direction. If the intercept term is removed, however, those variances are smaller with larger x, or larger predicted-y, so in the case of ‘true’ heteroscedasticity of Y_i, it will be underestimated. This happened, as shown in the electric sales example of Section 8.6. However, the impact of h_i^* was small, even though the sample size was only n=9, where one could expect a larger impact.

[12] Cook, R.D. and Weisberg, S.(1982), Residuals and Influence in Regression, Chapman and Hall, New York and London.

but only for $y_i = bx_i + e_i$, to become $\sqrt{x_i^{2\gamma}\sigma_{e_0}^{*2} - x_i^2\sigma_{e_0}^{*2}/\sum_{i=1}^n x_i^{2-2\gamma}} = x_i^\gamma\sigma_{e_0}^*\left(1 - x_i^{2-2\gamma}/\sum_{i=1}^n x_i^{2-2\gamma}\right)^{0.5}$. Thus we divide "$e_i$" by $x_i^\gamma\sigma_{e_0}^*\left(1 - x_i^{2-2\gamma}/\sum_{i=1}^n x_i^{2-2\gamma}\right)^{0.5}$, and to move from those standardized residuals to studentized residuals, y_i^*, and $\sigma_{e_0}^*$ would need to be estimated deleting the current case each time. See Penn State(2021e).

For $\gamma = 0$, $x_i^\gamma\sigma_{e_0}^*\left(1 - x_i^{2-2\gamma}/\sum_{i=1}^n x_i^{2-2\gamma}\right)^{0.5}$ becomes $\hat{\sigma}(1 - x_i^2/\sum_{i=1}^n x_i^2)^{0.5}$, so for $y_i = bx_i + e_{0_i}x_i^{\gamma'}$, when $\gamma = 0$, $y_i = bx_i + e_i$, which should not happen with a good choice of x, and good data quality,

$h_i = x_i^2/\sum_{i=1}^n x_i^2$, and[13] for any γ, $h_i^* = x_i^{2-2\gamma}/\sum_{i=1}^n x_i^{2-2\gamma}$.

Below, example 8.6, Electric Sales, has a very small sample size. The data are very well-behaved with no potential 'outliers' (see Penn State(2021d)), so we are only concerned with leverage. We will compare the two results for this example, i.e., two methods for estimating γ, illustrated more thoroughly in Knaub(2019), with them run again here, first dividing the e_i by $(1 - x_i^2/\sum_{i=1}^n x_i^2)^{0.5}$ for this new experiment. Actually it is $\sqrt{\hat{\sigma}^2(1 - h_i)}$ which should be the divisor, rather than just $\sqrt{1 - h_i}$, but because $\hat{\sigma}^2$ is constant there, that will not matter here. However, if one were to want to explore the impact of the hat-value when applying weighted least squares (WLS) regression, which we will, then the use of σ_i^{*2} in the divisor is necessary, or rather it will turn out that only a factor of it will be needed. Differences between results from different methods of estimating γ, the standard error for γ suggested by Ken Brewer for the second method, consideration of potential outliers as discussed below for the baseball example, 8.12, which is the reason Carroll and Ruppert(1988) used studentized residuals, all are considerations. However, we see that the major issue, as noted in Brewer(2002), on page 137, is that $\gamma = 0.75$, or 1.0, or as in Knaub(2017a), slides 13, 14, 17, 18, and 19, we see that $\gamma = 0.5$, may be best for specific applications. Here we are noting that the choice of $\gamma = 0$, which is commonly used without any consideration, is often not a good practice. Ignoring this will usually have much less impact on prediction, but a large impact on variance, and may be important to your application. It may even be important to some of the predicted-y-values. That is especially possible with small sample sizes, where study of the application may help you determine a better default value for γ than 0. Otherwise one is automatically using perhaps the worst value for γ possible.

In the example on pages 49 and 50, Carroll and Ruppert(1988) estimate θ, here γ, the coefficient of heteroscedasticity, as 0.79, but consider that an underestimate because of using unweighted least squares predicted-y values (something they note on page 51 that they consider in other methods in their following chapter). An adjustment for weighted fits to this method was studied in Knaub(2019), and found not to matter very much. However, this would vary somewhat by application.

[13] Although h_i is not shown as $\hat{h}_i$, h_i^* can be used to denote heteroscedasticity of the ε_i, or $h_i(\gamma)$ could be used in all cases.

Below, example 8.6, Electric Sales, has a very small sample size. The impact of the hat-value, h_i, is considered. There we use a zero intercept, which is appropriate in that case, where the lone predictor is the same data item from a previous census survey. The expression for the hat-value used first is $h_i = \frac{x_i^2}{\sum_{j=1}^{n} x_j^2}$. That is for $\gamma = 0$. However, in general, *i.e.*, when γ can vary, the variance due to the estimated regression coefficients, within the estimated variance of the prediction error, for the case of $y_i = bx_i + e_{0_i} x_i^{\gamma}$, from slide 16 in Knaub(2017a), for an individual prediction, is $x_i^2 V^*(b)$. From Maddala(1977), pages 259 and 260, $V^*(b) = \sigma_{e_0}^{*2} / \sum_{i=1}^{n} x_i^{2-2\gamma}$. Thus, h_i here, accounting for heteroscedasticity, for one predictor and no intercept term, would be $h_i(\gamma) = \frac{x_i^2}{\sum_{j=1}^{n} x_j^{2-2\gamma}}$. For the example in section 8.6, estimates without using a hat-value were 0.7 and 0.9, for the two methods used. Follow up estimates are shown here, dividing e_i by $\sqrt{1 - h_i(\gamma)}$, ignoring σ as the comparison of interest there is the hat-value, and using $\gamma = 0$ for fitted values. A follow up is done where the fitted values use $\gamma = 0.8$, so then σ_i has to be used. Dividing by $x_i^{\gamma} \sigma_{e_0}^{*} \left(1 - x_i^{2-2\gamma} / \sum_{i=1}^{n} x_i^{2-2\gamma}\right)^{0.5}$ to accommodate leverage and heteroscedasticity,[14] and dropping $\sigma_{e_0}^{*}$ as division or multiplication by a constant will not change anything, means we only need to divide for this purpose by $x_i^{\gamma} \left(1 - x_i^{2-2\gamma} / \sum_{i=1}^{n} x_i^{2-2\gamma}\right)^{0.5}$. However, we are already dividing by $y_i^{*\gamma}$ to account for heteroscedasticity, and $y_i^{*\gamma}$ is just a constant multiple of x_i^{γ}, so we only need to divide by $\left(1 - x_i^{2-2\gamma} / \sum_{i=1}^{n} x_i^{2-2\gamma}\right)^{0.5}$ to account for the leverage.

Note that using studentized residuals would also make them approximately follow the t-distribution. However here we are concerned with leverage and with outliers, both of which were considered in one example or another and found not to interfere substantially with our study of when heteroscedasticity is found in regression with the magnitude range noted in Brewer(2002), except that a suspected outlier, though it should remain in your data unless there is more reason to remove it than that it may be in the tail of a distribution, may be removed temporarily for the purpose of estimating inherent heteroscedasticity. Leverage does not seem to be a large complication factor for this study. Therefore, the simple spreadsheet tool accompanying Knaub(2019) can be used to quickly check for the

[14] Here, for $y_i = bx_i + e_{0_i} x_i^{\gamma\prime}$, where $e_{i=} e_{0_i} x_i^{\gamma\prime}$, but in terms of the notation in Weisberg noted earlier, it is $\hat{e}_i$, we see that $V(\hat{e}_i)$ is $x_i^{2\gamma} \sigma_{e_0}^{*2} \left(1 - x_i^{2-2\gamma} / \sum_{i=1}^{n} x_i^{2-2\gamma}\right)$. Note that this is consistent with the case for $\gamma = 0.5$ in Valliant, Dorfman, and Royall(2000), page 131, where they have $r_i = Y_i - \hat{\beta} x_i$, and the expected value of r_i^2 under this model is $\sigma^2 x_i (1 - x_i / \sum_s x_k)$. There they are examining variance for the classical model-based ratio estimator when predicting totals, and how to adjust for leverage. One result using leverage, υ_D there, is compared with the not-leverage-adjusted υ_R on pages 131-133. However, Figure 1, page 879 in Knaub(1992) does not show this to be particularly helpful. On pages 132 and 133 of Valliant, Dorfman, and Royall(2000), we see that υ_R regarding predicted totals there is generally an underestimate. But υ_R, and alternatives for it involving a hat-value adjustment, all for $\gamma = 0.5$, each make a further approximation by applying an average impact of the weight to each case, rather than individual ones, apparently prompting Thompson(2012), page 108, to comment on when there is overestimation or underestimation.

substantial heteroscedasticity one might expect, by inputting only the y-values and corresponding homoscedastic predicted-y-values. An effort was made to make the tool as easy as possible to use.

Potential outliers were considered in Knaub(2019), and in another case here. Leverage is considered in an example here. Two different methods for estimating the coefficient of heteroscedasticity are exercised. But the main emphasis for this paper is to consider when we may or may not see heteroscedasticity of the magnitude discussed mid-page 111 in Brewer(2002). It would seem that such heteroscedasticity should be a frequent occurrence. Is it generally a good sign? It is proposed here that it is generally an indicator of a good model, but model adequacy depends on various other factors as well.

6. THE COEFFICIENT OF HETEROSCEDASTICITY, GAMMA, AND ITS ESTIMATION – A BASIC REVIEW

The format used in Cochran(1977), on page 243, and also in Cochran(1953), on page 199, for within cluster variance, included an exponent, g: S_w^2=AM^g. By using twice the coefficient of heteroscedasticity, two gamma (2γ), for that exponent, using X in place of cluster size M, and using σ^2 for S_w^2, the within cluster variance, we see that the model that Ken Brewer noted for variance on pages 87 and 111 in Brewer(2002), $\sigma_i^2 \propto X_i^{2\gamma}$, is of the same structure, but designed for more than one cluster or data point. When we use variance for estimated residuals as an approximation in place of $V(Y_i)$, the structure is the same: $e_{0_i}^2 z_i^{2\gamma}$. Thus, the estimated residuals are a product of a random factor, we will call e_{0_i}, and a nonrandom factor, z_i^γ.

At the top of page 2 in Knaub(2017b) we see an argument for the simplest case, $y_i = bx_i + e_{0_i}x_i^\gamma$, which can be applied to the general case, $y_i = y_i^* + e_{0_i}z_i^\gamma$. (The γ here is an approximation as noted earlier.) There we see $y_i = bx_i + e_{0_i}w_i^{-0.5}$, $e_{0_i}^2 = w_i(y_i - bx_i)^2$, and $w_i = x_i^{-2\gamma}$. Note that w_i is the regression weight, which your software may allow you to enter. In multiple linear regression, you may minimize the sum of the $e_{0_i}^2$ to find the parameters (regression coefficients, including the intercept, if appropriate), using the same normal equation approach, employing a little calculus and solving simultaneously for all parameters needed, just as illustrated in Lohr(2010), on pages 430-432 (where it says “simple random samples,” but that is not necessary to proceed). We just have to use $w_i = x_i^{-2\gamma}$ or more generally $w_i = z_i^{-2\gamma}$ and minimize $\sum e_{0_i}^2$ instead of $\sum e_i^2 = \sum e_{0_i}^2 z_i^{2\gamma}$. But to obtain w_i we need to estimate gamma, γ.

In Knaub(1992, 1993), the author estimated γ using Fortran. A simpler version is done in Excel in the spreadsheet that accompanies Knaub(2019). There the idea is to find a value for γ which will cause a simple regression line through the points $|e_i|/z_i^\gamma$ to have a slope satisfactorily close to zero, where “γ” is a working, and changing value until it becomes an approximation for the actual γ. In the examples in Knaub(2019), y^* is used in the graphs instead of z. Another way to estimate γ used there, and suggested to this author by Ken Brewer, and used elsewhere, is noted in Knaub(2019), which involves taking logs of both sides of $y_i - y_i^* \approx e_{0_i}z_i^\gamma$, where you then have γ as a regression coefficient, and can estimate it and find an estimate of the standard error for γ. Comparison of the two estimates of γ may be helpful in considering accuracy, and in more than one example, the standard error for γ, found by the second method, was estimated, and provided interesting information, as will be shown. We will call these two methods the graphical analysis method and the PDQ ("pretty darn quick") method.

At the bottom of page 5, Section 3 of Knaub(2019), it is also noted that some estimate the impact of heteroscedasticity by simply running a regression line through the points $(\hat{y\imath},|ei|)$, where $\hat{y}_i$ is the homoscedastic predicted-y, and $|e_i|$ is a crude estimate of σ_i, such that a regression through them would smooth this out and provide predictions of σ_i for other cases, in an *ad hoc* manner. From comments on ResearchGate, this appears to be in common use in statistical software. (Also, see Penn State(2021g).) But the better method is to provide regression weights, w_i as a 'formula'/mathematical expression, based on the coefficient of heteroscedasticity, γ, as above. SAS PROC REG, for example, allows entry of such a regression weight, w_i, such as $w_i = 1/x_i$ for the classical ratio estimator. Note that both Penn State(2021g), and Särndal, Swensson, and Wretman(1992), on pages 231 and 232 regarding *exact* model-unbiasedness, say that when there are multiple predictors, one may make calculations based on one of them, or any combination of them. But here we are insisting that in looking for the best size measure, if there are multiple predictors, that best size measure should be the ideal predicted-y, and thus a combination of predictors. Therefore, $w_i = z_i^{-2\gamma}$, where we want the z_i to be the weighted least square predicted-y, y^*, which is ideal, but in first estimating γ, z_i may need to first be the homoscedastic (OLS) predicted-y, which appears to work well in practice, but could be improved iteratively.

Section 4, "Summary of Reasoning for Essential Heteroscedasticity," and this one, Section 6, "The Coefficient of Heteroscedasticity, Gamma, and its Estimation," refer, respectively, to the nature and magnitude of heteroscedasticity of estimated residuals in regressions of form $Y_i = y_i^* + e_{0_i} z_i^{\gamma\prime}$, for finite populations. This is covered, again respectively, in Brewer(2002), page 111 plus Knaub(2017b), and Knaub(2019), including the Excel spreadsheet tool. An ongoing project on ResearchGate, https://www.researchgate.net/project/OLS-Regression-Should-Not-Be-a-Default-for-WLS-Regression, contains a number of project updates, in reverse chronological order, relevant here.

Please note that Knaub(1993) compared a graphical method to the iterated reweighted least squares (IRLS) method for estimating γ. When artificial data were used, for a given value of γ, both methods estimated γ extremely closely, but for real data, the IRLS method sometimes would not converge, and the graphical method might indicate a solution without showing it to be exactly correct. Real data of a finite sample size will have some randomness and perhaps model misspecifications and/or data quality issues which will impact different γ estimation methods differently. That occurred here and in Knaub(2019) as well.

7. WHY HYPOTHESIS TESTS FOR HETEROSCEDASTICITY IN REGRESSION ARE NOT PRACTICAL

Reference is again to this project:
https://www.researchgate.net/project/OLS-Regression-Should-Not-Be-a-Default-for-WLS-Regression.

An update there, from August 9, 2019 is titled "No need for an hypothesis test." In there an excerpt from a response the author made to a question on ResearchGate was noted:

"Once you test for heteroscedasticity, then what? Heteroscedasticity is a matter of degree. Do you have too much of it to ignore? If so, what do you do? Heteroscedasticity is natural and occurs because the different predictions are not the same size. Some model specification problems and/or data issues can cause it to be made artificially larger or smaller, but it is to be expected, not a problem to be fixed, but something to be handled routinely. OLS is a special case of WLS, where the coefficient of heteroscedasticity is zero. If you estimate the coefficient of heteroscedasticity, or use a reasonable default value …, you can use that in the regression weight that you enter into your software. For SAS PROC REG, for example, you enter the regression weight ('formula'), ‘w,’ to change from OLS to WLS. (If you enter a constant, that means the weights are the same, and you still have OLS.)

"The thing is, if you test for heteroscedasticity, it does not tell you how much there is, or what to do about it."

8. REAL DATA EXAMPLES TO CONSIDER VARIOUS REASONS FOR THE PRESENCE OR ABSENCE OF HETEROSCEDASTICITY

Here we look at a variety of **real data** examples of regression of the form $Y_i = y_i^* + e_{0_i} z_i^{\gamma}{}'$, where ideally $y_i^* = Z_i$, and y_i^* strives to be the ideal weighted least square predicted-y, where one should expect to find $0.5 \leq \gamma \leq 1$.

This is for finite populations, though autocorrelation for a time series is mentioned, as well as spatial autocorrelation, but we are really only concerned here with weighted least squares (WLS) regression. Ordinary least squares (OLS) regression is a special case which is generally taken as the default, but here we assert that heteroscedasticity of estimate residuals is to be expected, and try to explain what might be wrong if we do not have $0.5 \leq \gamma \leq 1$. The following information on real data examples varies in the level of detail available and/or presented, but is meant to stir discussion on various applications.

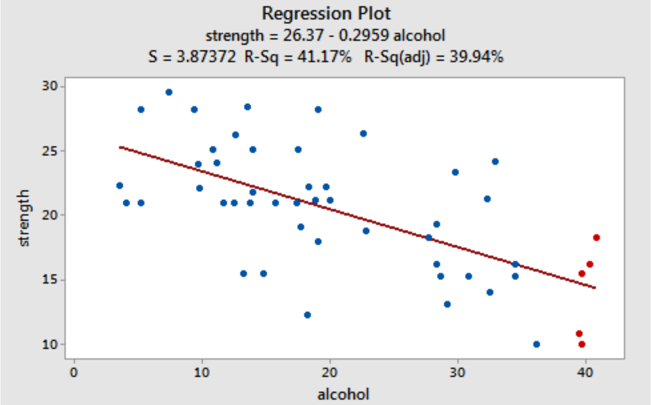
8.1 Arm Strength and Alcohol Consumption

In Penn State(2021a) we see arm strength as our dependent variable, and alcohol use as a predictor. Because this is a negative relationship, an intercept term is involved and is quite important. The "total lifetime consumption of alcohol" is X. But two people with the same total lifetime consumption who are of very different ages would have an impact on variation. If a younger person had the same lifetime consumption as an older person, then the younger person would be drinking much more heavily. That may overcome or match the youth factor regarding strength. So another 'independent' variable might be age of the responder, and/or another might be number of years as an alcoholic. It might be difficult to determine that last variable but then it also may have been difficult to determine total lifetime consumption. Of those three independent variables, age would appear to be the easiest to accurately obtain. Also, one might expect that there would be some collinearity. However, with only total alcohol consumption, one may not know that the reason we do not see increasing sigma for estimated residuals becoming larger as we get to larger predicted-y, here going right to left on the scatterplot graph below, and even perhaps larger sigma towards the mid-alcohol range, is that the age and rate of alcohol consumption related variables are missing. Since the one variable and the intercept term in predicted-y here must absorb this information, this increases mid-range sigma where there is more impact from age and rate of drinking. (The hat-value adjustment would push estimated residuals that way a little also.) At the largest values for predicted-y, with near zero "total lifetime consumption of alcohol," there is less variability of age and years as an alcoholic, so less variability from those factors, precisely where sigma should be greatest. That is, with the lowest lifetime alcohol consumption, you tend to have younger people with fewer years as an alcoholic. Thus we have a smaller mix of categories than we do at larger total lifetime consumption of alcohol values.

In summary, when we have an oversimplified model, a result may sometimes be to produce an inferior 'size' measure (predicted-y, the model, often “z” is used for the size measure) which may not support the natural/essential heteroscedasticity. Perhaps the negative slope, and intercept are a clue as to one situation where we might look for this.

Here is the graph (3rd party) from Penn State(2021a), reproduced with permission:

The red data points were put there to show that they are at the other end of the scatterplot when the predicted values are on the x-axis, and the estimated residuals are on the y-axis.
Estimated residuals are measured vertically, above and below the regression "line."
The original data were taken from the following:
The Effects of Alcoholism on Skeletal and Cardiac Muscle, by Alvaro Urbano-Marquez,M.D., Ramon Estruch, M.D., Francisco Navarro-Lopez, M.D., Jose Maria Grau, M.D., Lluis Mont, M.D., and Emanuel Rubin, M.D.
February 16, 1989 N Engl J Med 1989; 320:409-415
DOI: 10.1056/NEJM198902163200701



n=50
Penn State(2021a)

8.2 Home Natural Gas Energy Use: Multiple Linear Regression Prediction

In Roberts, *et al.* (2012), on page 29, in a section on modeling, Table 8 identifies 11 variables, 8 "numerical" and 3 binary, "significant" in multiple linear regression modeling for obtaining predictions for "Site NG" (natural gas) energy use. Many other variables had apparently also been considered. (There appear to have been over 100 in total.) They noted that collinearity is a consideration. (See the bottom of page 21.) We know that that can even change the sign of a coefficient when ‘independent’ variables are used together. Apparently good predictors were chosen. But do we know that the ideal predicted-y would have more predictors, or fewer predictors, or a different set of predictors? It seems that several experts approved 11 out of more than 100 variables, which sounds promising. The way variables work together makes t-values less important, but identifying three binary variables could help avoid nonessential heteroscedasticity (Knaub(2018)), and hopefully reducing the model to such a degree would leave only what is needed for a good size measure.

Consider the scatterplot below, reprinted with permission from the US National Renewable Energy Laboratory. The dashed line indicates a linear regression. However, considering that predicted-y and y should approach zero together, if one were to drop the intercept term, and consider the data points shown in the graph, vertically above and below the line marked "Line of Perfect Agreement," heteroscedasticity is apparent.

Note here a criterion proposed by this paper: If the achieved predicted-y is close enough to the ideal predicted-y, we should expect heteroscedasticity. It appears that this could be the case with this example of multiple linear regression. Suggestions: Use a regression weight here, and $Y_i = y_i^* + e_{0_i} z_i^{\gamma}{}'$, with $y_i^* = Z_i$, and in the scatterplot below, switch axes so that predicted-y is on the x-axis. The estimated residual σ_i^* values are larger than one would generally hope to obtain, but perhaps this is unavoidable given the complex nature of the contributions to natural gas use. At least the “confidence interval for μ_Y” could be meaningful, if not the prediction interval for a new y-value. (See the homoscedastic version of this at Penn State(2021b).)

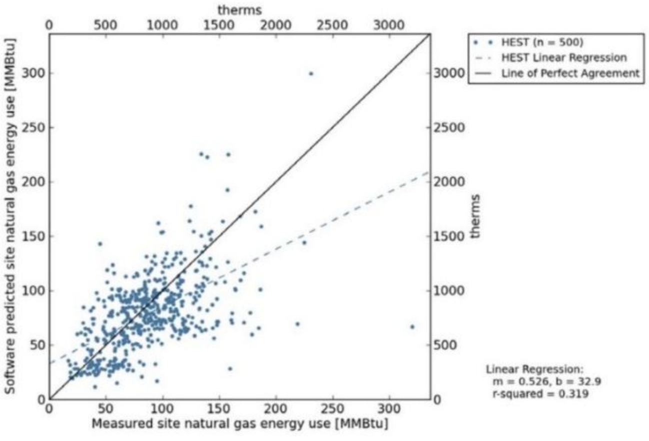


Figure 7. HEST-predicted site NG energy use versus weather-normalized measured site NG energy use

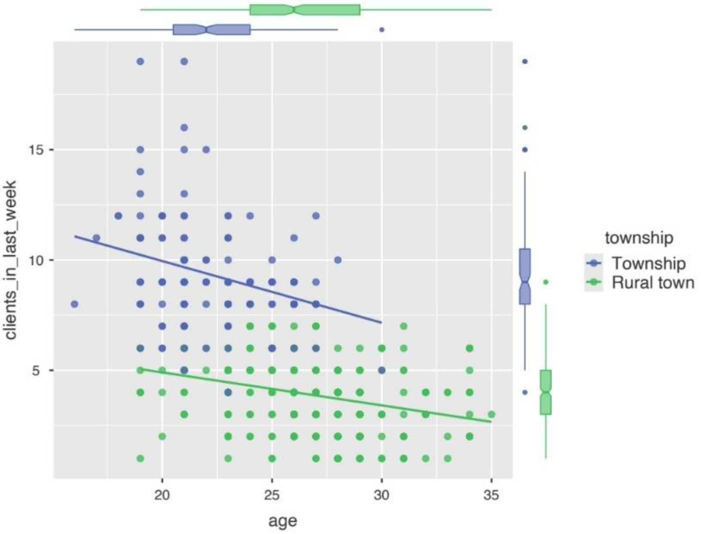
9

Figure 7, on page 9, **n=500**. (3rd party graph)
Reprinted with permission from the National Renewable Energy Laboratory from https://www.nrel.gov/docs/fy12osti/54074.pdf, accessed April 24, 2021.

8.3 Kenyan Sex Worker Example

In this example (Elmore-Meegan, Conroy, and Agala(2004)), the number of clients which sex workers in Kenya had in the previous week is predicted by the worker's age. [*This third-party scatterplot was separately supplied for use in this paper by one of the authors, Dr. Conroy.*] It shows two simple linear regression lines, one for townships, and another for rural towns. The line for workers in rural towns has a smaller slope. Notice that if the range for age were restricted to 23 years and up, that that line would have been even flatter. Thus age is not strongly related to the response variable for that stratum or subpopulation. The range of predicted-y would be small, thus the size measure would vary little, and there would be near homoscedasticity. That is the case for rural towns. For a slope of zero perhaps we could have the common means model described earlier. The paper explains the different circumstances in the two strata. What we see here is that there is heteroscedasticity in the township stratum, but as age is less of a factor in the rural towns, especially once you move above age 23, there homoscedasticity and a common means model (Särndal, Swensson, and Wretman(1992), page 258-260, and Chambers and Clark(2012), page 20) appear more reasonable.

From the article (Elmore-Meegan, Conroy, and Agala(2004)), age does seem more important in the townships. Given the premise here that an achieved predicted-y that is nearer to the ideal will be more likely to show heteroscedasticity of the estimated residuals, one wonders if there are no important omitted variables, and if there are none, perhaps that is why we see heteroscedasticity, indicating a good model, in this stratum. Is the model very good there, in spite of a σ_i that is relatively very large? Perhaps so, or perhaps there are other categorical variables which are missing, and thus causing nonessential heteroscedasticity. (See Knaub(2018).) At any rate, the importance of age in this stratum is shown in the graph, which was discussed in the article. The fact that the slope is downward and variance is reduced is reasonable. (Note that variance is still increased for larger predicted-y.)



n = 475 (139 in Nairobi townships; 336 in rural towns)
[Third party graph: Dr. Ronán Conroy]

8.4 Spanish Shops: Example in Guadarrama, Molina, and Tillé(2020)

A nested error model, page 59, with random domain effects is used in an example here. In Section 9, the authors look at data from Spanish tobacco shops, for one particular "product," by province, where the size measure, "z_{ij}," for shops by province, is the past three months revenue for that product, for that shop, for the population they considered. Further, for a sample of the largest shops, they were able to obtain "v_{ij}," the more current month's revenue values, for which they wish to estimate a total. This is noted at the bottom of page 68. On page 69 they note that heteroscedasticity is apparent, and they use a transformation to ameliorate the impact. However, using the coefficient of heteroscedasticity, γ, to account for this would be a more direct approach to the actual feature. (Heteroscedasticity is a not a problem to be 'fixed,' but a feature which should be included in the error structure of the model.)

The relationship here, data from a previous census used as a size measure variable for a variable representing current data on the same item, is an excellent choice. (See the bottom of page 205 in Cochran(1953).) This is generally the case throughout many applications at the US Energy Information Administration (EIA). (See Knaub(2017a).) However, the relationship of interest is a ratio, but a random domain effect means using a change in intercept to account for a change in ratio between provinces. It is unclear whether that is helpful. Borrowing strength by collapsing some groups of provinces, but not all, would mean less of a compromise on ratio estimates, but samples smaller than that from the entire population. If the difference in underlying province ratios in a collapsed group is substantial, that would increase nonessential heteroscedasticity to some extent, but may not be problematic. (See slide 35 in Knaub(2017a).)

On page 69, v_{ij} and z_{ij} are said to have right-skewed distributions, which is also true of EIA establishment survey data, and establishment survey data in general. (There are often many small operations, and a few very large companies.) Heteroscedasticity is often more apparent because there is an obvious size difference. The few large companies, with possibly large estimated residuals, will have smaller regression weights. They are very important to sample when 'predicting' a total, since their part of that total is large.

At the bottom of page 69 in Guadarrama, Molina, and Tillé(2020), the estimated residuals are given as the difference between a sampled y-value, and the model prediction, which includes the random domain effect as an intercept, or adjustment to one already present. (There seems no reason for an intercept term here. See Brewer(2002), pages 109-110).

Overall, the example in Guadarrama, Molina, and Tillé, (2020) is very interesting and clearly presented so that results can be understood.

The first paragraph in the conclusions on page 72 note that business applications (say, establishment surveys) can trade higher cost for lower accuracy with cutoff samples. A cutoff sample may entail bias from excluding the smallest members of the population. However, variance is decreased tremendously. Accuracy is therefore often improved with a cutoff sample. (So it is often less expensive and more accurate.) The authors also note some data will not be available. That could be more of a problem for design-based (probability-of-selection-based) sampling. The key is to have good predictor data for the entire population. Weighted least squares regression, using a reasonable coefficient of heteroscedasticity, is very good for such data. See Knaub(2017a) and references there. (If there are a few substantial cases with no predictor data, collect them as “add-ons”).

The last paragraph of the article notes how small area estimation can be used, piecing parts together to cover domains. This is illustrated in Knaub(2017a) in slide number 31.

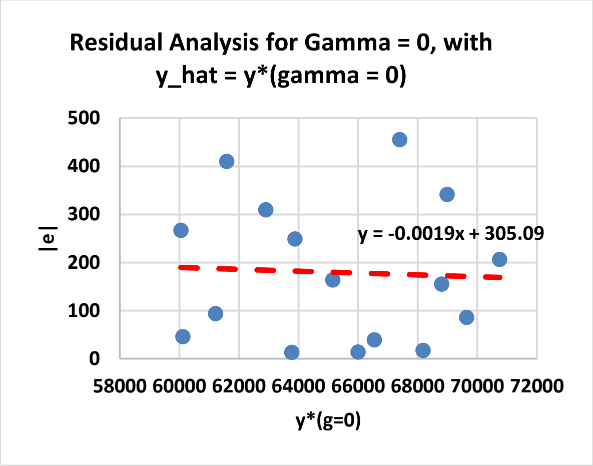
Overall sample size is **n = 1842** in 48 provinces, with N = 12,791.

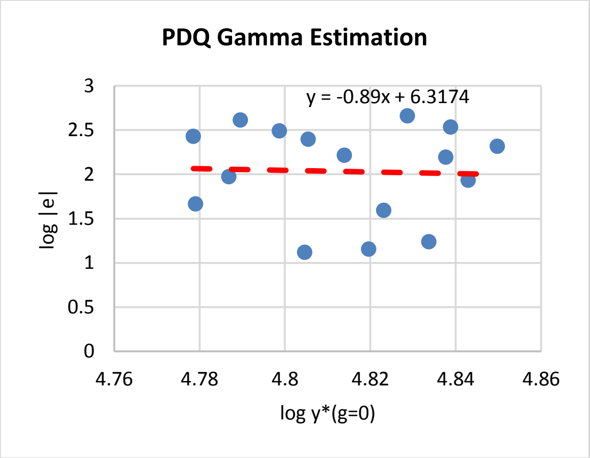
8.5 Longley Data – Multiple Linear Regression

The National Institute of Standards and Technology (NIST), US Commerce Department, has collected some real data for use in testing statistical software for accuracy in producing regression results. One well-known data set was considered in NIST Information Technology Laboratory(2021).[15] We will call this NIST ITL(2021). We will refer to that as the Longley data, which is employment and related data, for which they used multiple linear regression with six independent variables and an intercept term of very large relative magnitude, and only 16 data points. NIST ITL(2021) shows a graphical analysis, but it only involves the first predictor, which as part of the full model has an estimate of 15.06, and “standard deviation of estimate” of 84.91. With n = 16, one might visually assume homoscedasticity, but there is a great deal of variance, little data, and that one predictor may not be a good measure of size. However, full model predicted-y values may be unsatisfactory measures of size as well. Statsmodels(2021) tells us that the

[15] The reference from which these data came is given as Longley, J. W. (1967), “An Appraisal of Least Squares Programs for the Electronic Computer from the Viewpoint of the User,” Journal of the American Statistical Association, 62, pp. 819-841.

variables are "highly collinear." The data set is known for the difficulties it presents to computer algorithms. See Weisberg(1980), page 178, on "ill-conditioning" as a consequence of extreme collinearity.



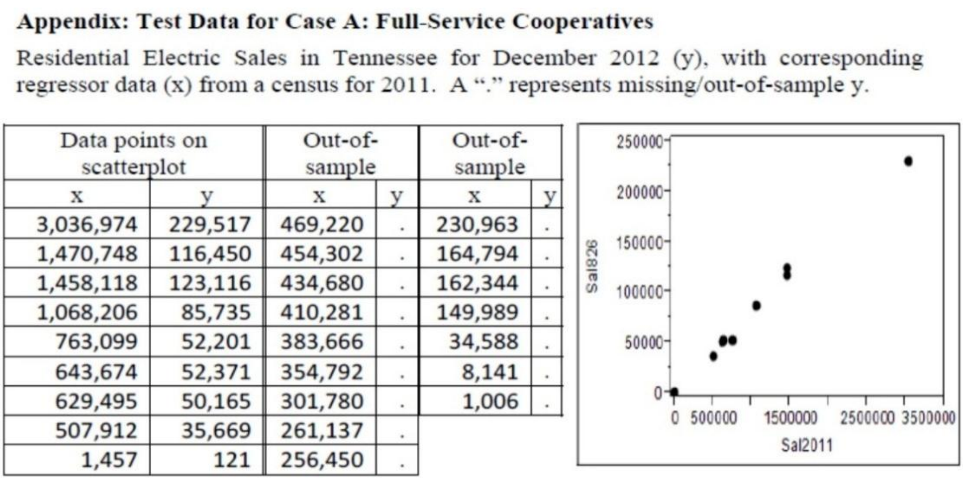


These graphs are for the two methods of estimating γ in Knaub(2019).
n=16.

The first graph, with a negative slope, tells us that the estimated γ is less than 0. The second graph says it is -0.89, and the standard error for γ is the standard error of the slope in that graph, which is estimated from simple linear regression as 5.83. So, we cannot really say anything about it. Brewer(2002), page 137, called for very large sample sizes. Here, with apparently too many variables for a small sample size, we cannot obtain useful information on γ, and because the range of predicted-y is small, that means there is little change in measure of size.

8.6 Residential Electric Sales by Full-Service Cooperatives in Tennessee

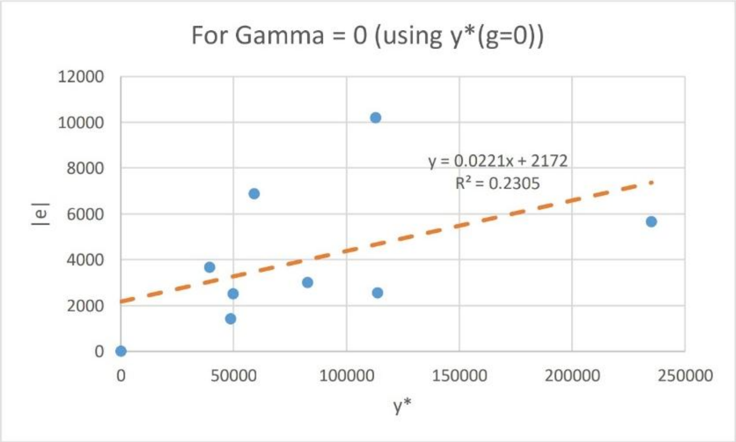
This figure is found on page 9 of Knaub(2019), where also is found an easy reference/reminder as to how the two methods of estimating the coefficient of heteroscedasticity used here can be explained. **n=9**.



Appendix: Test Data for Case A: Full-Service Cooperatives

Residential Electric Sales in Tennessee for December 2012 (y), with corresponding regressor data (x) from a census for 2011. A "." represents missing/out-of-sample y.

Data points on scatterplot		Out-of-sample		Out-of-sample	
x	y	x	y	x	y
3,036,974	229,517	469,220	.	230,963	.
1,470,748	116,450	454,302	.	164,794	.
1,458,118	123,116	434,680	.	162,344	.
1,068,206	85,735	410,281	.	149,989	.
763,099	52,201	383,666	.	34,588	.
643,674	52,371	354,792	.	8,141	.
629,495	50,165	301,780	.	1,006	.
507,912	35,669	261,137	.		
1,457	121	256,450	.		

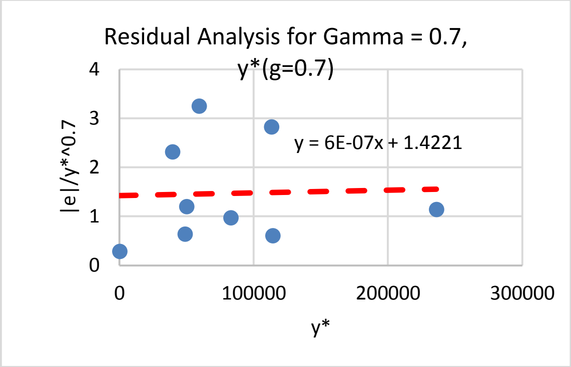


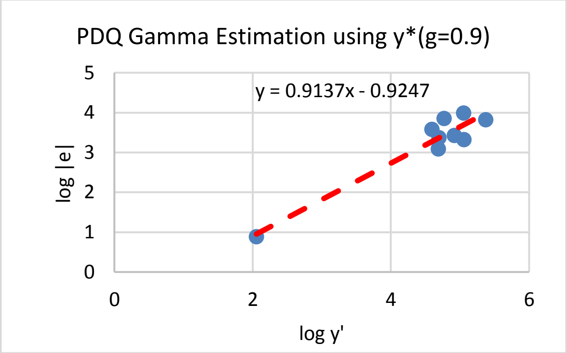
From page 10 in Knaub(2019) - preliminary **Residual Analysis**

In these figures, "g" means γ

Here we have a very small sample size, n=9, like many actually used for the many small populations of electric power and other data at the US Energy Information Administration (EIA). (Small area estimation is often used for "borrowing strength" at the EIA, but one should be very careful.)

For this example, were it used, γ = 0.5 would have been assumed during a data production cycle - monthly in this case - to accommodate possible data quality issues. Even though the sample size is extremely small for estimating γ, results were in the range Ken Brewer confirmed, just as census data checks, perhaps well over N=100, this author used for experiments in the 1990s also confirmed. This graph shows, by the first of two methods used, that gamma appears to be greater than zero.





The first graph shows an estimate of gamma to be a little more than 0.7, found on page 13 of Knaub(2019).

The second graph is from page 17 in that same reference. Gamma is estimated to be 0.9 with a standard error of 0.1. (Note that a small change in one data point might change this substantially).

Making an adjustment for the hat-value, we start with the following for $\gamma = 0$:

$h_i = x_i^2 / \sum_{i=1}^{n} x_i^2$	$(1-h_i)^{0.5}$
5.657E-01	0.659
1.327E-01	0.931
1.304E-01	0.933
6.999E-02	0.964
3.572E-02	0.982
2.541E-02	0.987
2.430E-02	0.988
1.582E-02	0.992
1.302E-07	1.000

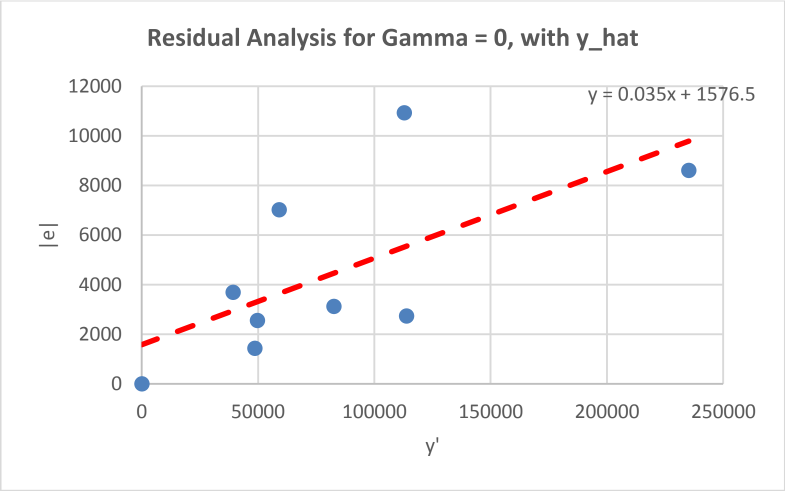
Note that σ is for $Y_i|X_i$, and σ^* is for the estimated residuals. σ^*underestimates σ by the variance due to the model coefficients:

$$\sigma \ = \frac{\sigma^*}{\sqrt{1-h_i}}$$

So if $\gamma = 0$ were the case here, from the lists on the left, one 'true' ε would be much larger, but most would barely change when translating from e to ε, if model specification is good.

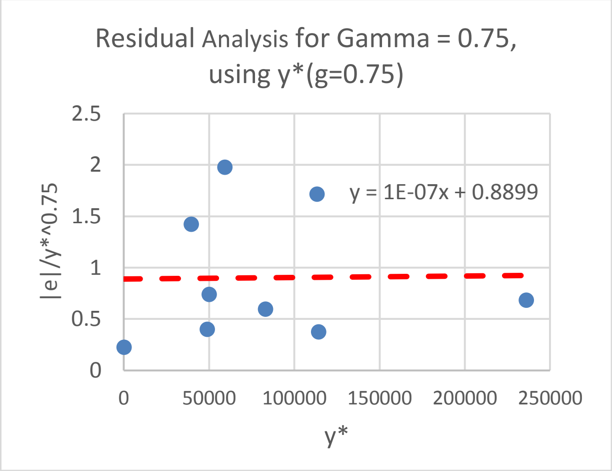
For $\gamma = 0$, $Y_i = bx_i + e_i$, we have $h_i = x_i^2 / \sum_{i=1}^{n} x_i^2$, used above. To adjust for leverage, were γ to be zero, we obtain the values above. The data point with the largest value for x_i, more than twice the second largest value for x_i, responsible for the values at the top of the two lists above, had the greatest change to the adjusted e_i value in this real data example. Here, however, heteroscedasticity for $var(Y_i)$ should be considered. The key is not σ^2, but rather σ_i^2, where h_i (perhaps we should say h_i^*) considers heteroscedasticity. Thus the key for $Y_i = bx_i + e_{0_i} x_i^{\gamma}{}'$, when directly considering γ, is not $h_i = x_i^2 / \sum_{i=1}^{n} x_i^2$, but rather $h_i^* = \ x_i^{2-2\gamma} / \sum_{i=1}^{n} x_i^{2-2\gamma}$, which is the same thing when $\gamma = 0$. The idea of adjustment by the hat-value for 'ordinary' least squares (OLS) regression is that σ is a constant, and the estimate for it is reduced by the influence of the model part as determined by the sample taken. But with heteroscedasticity for $\mathrm{var}(\boldsymbol{\varepsilon})$, the actual σ_i becomes larger with larger z_i. When studying a scatterplot of $|e_i|$, "predicted" by x_i, or y_i^*, or z_i, if there is a simple linear regression line through those points, the slope indicates heteroscedasticity, and thus the size measure can predict the $|e_i|$, a substitute for σ_i. In fact these predictions are often used. (See the second of four suggestions in Penn State(2021g), for estimating σ_i, when weights are not 'known' (or estimated using γ).) However, it is better, considering 'essential heteroscedasticity,' to estimate γ, or at least consider a reasonable default value, preferably where $0.5 \leq \gamma \leq 1$, though nonessential heteroscedasticity may have a roll. This fits with the nature of essential heteroscedasticity, following the logic of Brewer(2002), mid-page 111.

For the first method of estimating γ in Knaub(2019), adjusting for leverage, as suggested in Carroll and Ruppert(1988), but potential outliers will be considered in the baseball example, 8.12, we see below that γ is slightly larger than 0.75 now, where it had been approximately 0.7 for this method.



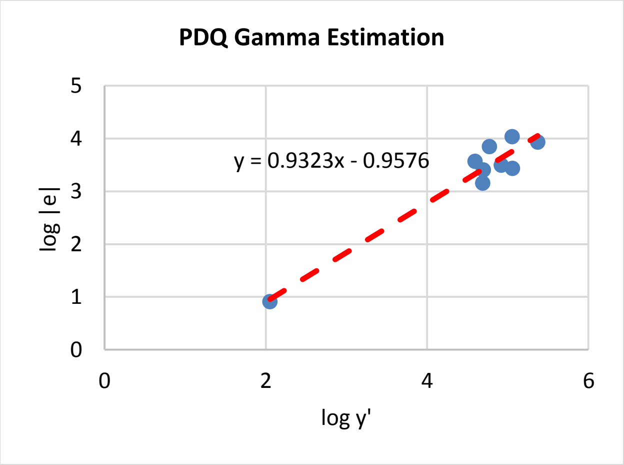
We follow up using the fitted values with $\gamma = 0.75$, and dividing $\left|e_{0_i}\right| = |e_i|/y_i^{*0.75}(\gamma = 0.75)$ further by $\left(1 - x_i^{2-2\gamma}/\sum_{i=1}^{n} x_i^{2-2\gamma}\right)^{0.5}$, with $\gamma = 0.75$.

When accounting for ‘true’ heteroscedasticity, *i.e.*, for Y_i, the impact of the new “hat-values,” h_i, is shown here.



$\left(1 - \frac{x_i^{2-2\gamma}}{\sum_{i=1}^{n} x_i^{2-2\gamma}}\right)^{0.5}$
8.904752E-01
9.251544E-01
9.254894E-01
9.365908E-01
9.466840E-01
9.511453E-01
9.517001E-01
9.567260E-01
9.977298E-01

Using the second method in Knaub(2019) for estimating γ, including the estimate of standard error, all suggested by Ken Brewer, and adjusting for leverage as part of what is shown in Carroll and Ruppert(1988) - see pages 31, and, for example, pages 49 and 50 - we estimate $\gamma = 0.93$ with the standard error shown below the scatterplot, in an excerpt from a spreadsheet, as approximately 0.087 ≅ 0.1. So, as before the adjustment for leverage, we will still use $\gamma = 0.9$. (Note that the estimate changed from 0.91 to 0.93. This is because of the hat-value adjustment used, as shown in Carroll and Ruppert.)



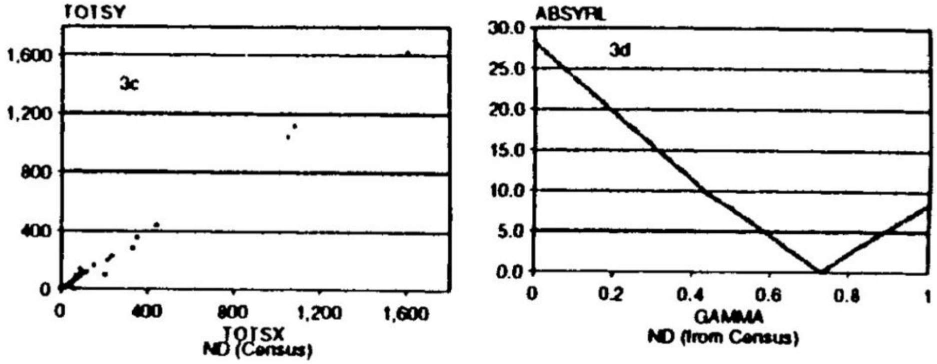
log|y-y*|	**log y***			
new y	**new x**	**new e**	**e^2**	**(x-xmean)^2**
3.934915	5.371419	-0.11526	0.013285	0.629387558
3.437788	5.056516	-0.3188	0.101635	0.228901667
4.038774	5.052771	0.285676	0.081611	0.225331636
3.494429	4.917633	-0.13268	0.017604	0.115296536
3.846437	4.771559	0.355513	0.126389	0.037434243
3.407584	4.697644	-0.01443	0.000208	0.014295644
3.156254	4.68797	-0.25674	0.065916	0.012075968
3.567518	4.594766	0.241417	0.058282	0.000278453
0.912045	2.052437	-0.04384	0.001922	6.378867636
mean x =	4.578079	RSS=	0.466852	
		sigma^2=	0.058356	
			Sxx=	7.64186934
	var(b)=	sigma^2/Sxx=	0.007636412	
			se(b)=	0.087386567

The estimated standard error for γ is 0.087, but the confidence interval may be very skewed. See the corresponding (but not hat-value adjusted) estimates and scatterplot for the body fat example (for quadratic linear regression) in Section 8.8.

Estimates of γ with and without the hat-value adjustment were close above, for such a small sample size. Residual variance was very small. Compatible accuracy in estimating γ is found in the forestry example, Section 8.11, where the sample size is over 1000. The difference in results between the residual analysis and PDQ methods might at least partially be due to a change in apparent value of estimated γ across ranges of the x-axis. This may be more easily pictured with the forestry data, as there are more data.

8.7 North Dakota Total Electric Sales

On page 880 in Knaub(1992), we have the following, among other graphs:

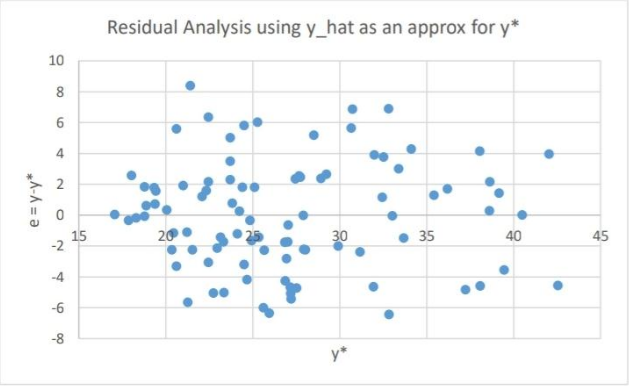


n = 42

Here a census (N=42) of North Dakota retail sellers of total (all economic end-use sectors) electric sales was predicted by a previous annual census. The goal was to find estimates of γ. These are Figures 3c and 3d. The point on the right-hand graph where the line touches y=0 is the point where a value for γ would satisfy the first of the two estimation methods explained and used above. When a cutoff sample of n=8 was used, the line curved and dipped well over the x-axis, y = 0 line, hovering over about the same solution, but did not come near it. In practice it was noted that $\gamma = 0.5$ often did well, especially to offset possible data quality problems from the smaller respondents. With such small sample sizes, γ not only matters for variance, it also matters for prediction. This is part of research done for publication of official energy statistics on a frequent basis for a great many small populations. For more information, see Knaub(1992, 1993, 2017a, and 2017b).

8.8 Percent Body Fat Predicted By BMI

The first graph in the second link in Frost(2019) is a regression for percent body fat predicted by a quadratic function of body mass index (BMI). There Frost provided parameters for that regression. In Knaub(2019), at the top of page 19, we have the following graphical residual analysis for this quadratic regression:

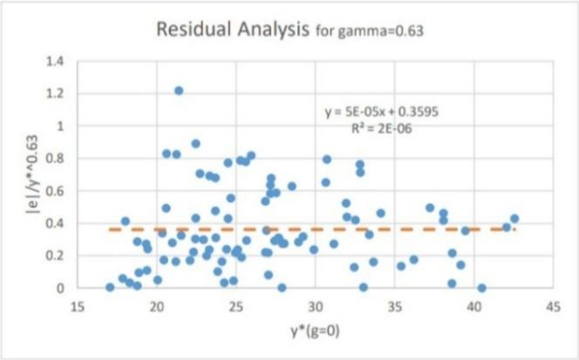


Top of page 19, Knaub(2019)
n = 92

The above graph is the usual graphical residual analysis with predicted-y on the x-axis, and estimated residual on the y-axis. We use $\hat{y}_i$ for the homoscedastic predicted-y, and y_i^* for the heteroscedastic version to be used later, though there is usually very little difference, especially for purposes of this routine for estimating γ. But here we often use y_i^* for any γ, including $\gamma = 0$, as the value of γ was changed in succeeding tables used for Knaub(2019) until it was determined to be adequate. Meanwhile, the actual predicted-y values on the x-axis are often left as $\hat{y}_i$ as it was found to make very little difference, and using $\hat{y}_i$ as an approximation to y_i^* makes it easy for anyone to use only their y-values and $\hat{y}_i$-values in the spreadsheet which goes with Knaub(2019), to look at γ for their data. That was done here, and in other cases. The title on the graph above from page 19 of Knaub(2019) is misleading in that y_i^* there is exactly $\hat{y}_i$. On later tables, it is usually a very close approximation on the x-axis, for the purpose of estimating γ.

Note that the graphical residual analysis above does not have a wider range in the y-direction for estimated residuals as we move to larger $\hat{y}_i$ values. The often expected "fan-shape" is not present. However, the variance, $\hat{\sigma}_i^2$, of the estimated residuals does increase as the density of estimated residuals in the y-direction decreases with larger predicted-y values, $\hat{y}_i$. This assessment was verified by the two methods of estimating the coefficient of heteroscedasticity noted earlier. Sample size is small, but both methods did agree with this visual analysis.

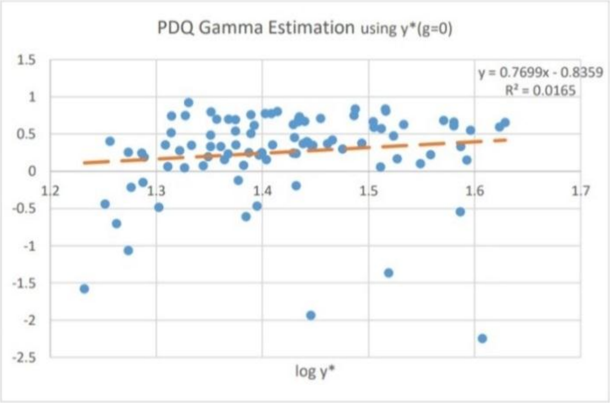
Following are two scatterplots showing those results, as before. Again, the first one was one of several used to find when a regression line through approximate $\left|e_{0_i}\right|$ values is closer to horizontal than the alternative that is different by one unit in the lowest significant digit considered. This graph appeared to come closest. The second scatterplot is one where logs were taken so that γ would be in the usual place for b, the slope, when the log of the homoscedastic estimated residuals is on the y-axis, and the log of $\hat{y}_i$ is on the x-axis.



Above, let us say $\gamma \cong 0.63$.

Top of page 21, Knaub(2019)

To the nearest 0.01, the best estimate here is $\gamma = 0.63$. Note that g in y*(g=0) on the graph is gamma (γ), and indicates that the homoscedastic predicted-y, $\hat{y}$, is used in this approximation to find an estimate of γ.



Top of page 22, Knaub(2019).
Note the slope is 0.7699, or about 0.77

With estimates of 0.63 and 0.77 for the coefficient of heteroscedasticity, one might use $\gamma = 0.7$.

8.9 Motor Fuel Consumption Example Multiple Linear Regressions

On page 23 in Knaub(2019), we see "The examples to follow … are for a motor fuel consumption data set in Weisberg(1980), pages 32 – 47."[16] First we look at the case of two

[16] On page 34 in Weisberg(1980), it notes that Christopher Bingham drew these data from the American Almanac for 1974, with the exception of total fuel consumption per State. Those data were from the 1974 World Almanac and Book of Facts.

predictors, and then we will add two more predictors, as was done in Weisberg(1980). However, any improvement seen is not necessarily due to the number of predictors, but rather which predictors are used. At the bottom of page 34 in Weisberg(1980), we see that motor fuel consumption per person is the y-value, and of the four regressors shown, the first multiple regression example there only used motor fuel tax and the proportion of the population holding a license to drive in each US State. Left out were highway and personal income data. It would seem that personal income would play a large role in the personal consumption of motor fuel. (Please remember that below.) Adding superfluous variables does not help, and can add variance, but omitting an important predictor would not appear to lead to the "ideal" predicted-y (model) either. (In the appendix to Shmueli(2010) there is an example shown where a biased model, having dropped a variable from the ideal model, gave better predictions. However, it would seem that we are considering the better explanatory case here instead.) Considering Brewer(2002), mid-page 111, it would seem logical then that of the two fuel consumption models in Weisberg(1980) and considered in Knaub(2019), that Weisberg's two-predictor model may not show heteroscedasticity, but the four-predictor model might. Other factors might enter into consideration, but that is what seemed apparent in Knaub(2019), which we reference here. (For more information, see Knaub(2019) regarding the examples which are also among those used here.) In fact, scatterplots with y (fuel consumption) on the y-axis, and predicted-y on the x-axis, show a closer fit[17] for the four-predictor model, versus the two-predictor model. So, the four-predictor model had a tighter fit, and exhibited heteroscedasticity. But … actually looking at a scatterplot of income as a predictor for motor fuel consumption was disappointing. There is a large intercept from which the income term is subtracted. In both the two- and four-predictor cases, the intercept was a large part of the predictions, but much more so in the four-predictor case. These models hardly appear "ideal."

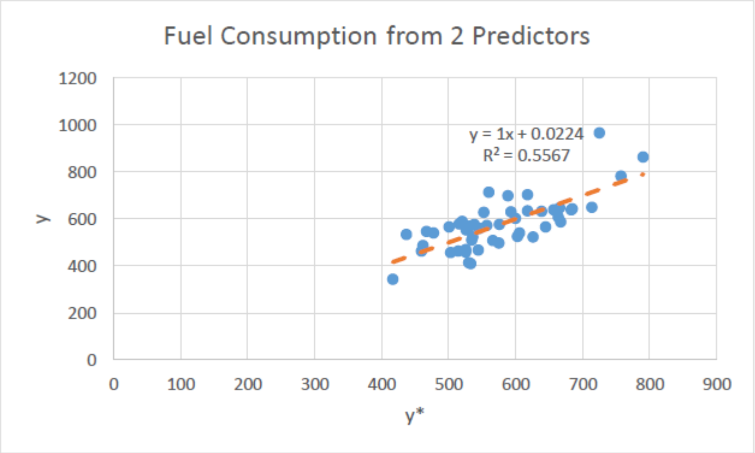
For the case of these motor fuel consumption regressions, it is apparent that the model with the best fit also had heteroscedasticity. However, with the sample size n=48, a small number, one case for each Continental US State, and various factors which might impact each State differently, results may not be very meaningful. Here, as with the Longley data, the variables appear to be a somewhat eclectic collection, for which a subject matter theory describing why they should be used together appears to be missing.

On page 47 in Weisberg(1980), Table 2.4 shows the t-values for all four predictors and the intercept term. The apparent relative importance of predictors will change, depending upon relationships between them, in different combinations, but in that table, we see that, as with the Longley data, there are multiple variables with negative signs which tend to make prediction calculations bounce around the final value as you add terms. This all makes the four-predictor scatterplots less impressive than as was discussed in Knaub(2019). Perhaps the heteroscedasticity seen there, and not seen with the two-predictor case, may not be very meaningful, or at least it is harder to explain.

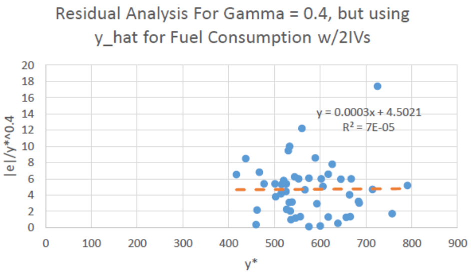
[17] Fitting too many variables to a small sample could be a problem, which might also be the case in the Longley data example. However, there it seemed it might be a reason for not having heteroscedasticity.

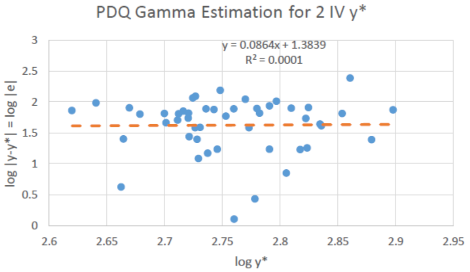
Following are some of the scatterplots found in Knaub(2019) for first the two-predictor, plus intercept term, and then the four-predictor, plus intercept term, Weisberg fuel consumption models:

For the two predictor model with an intercept term:



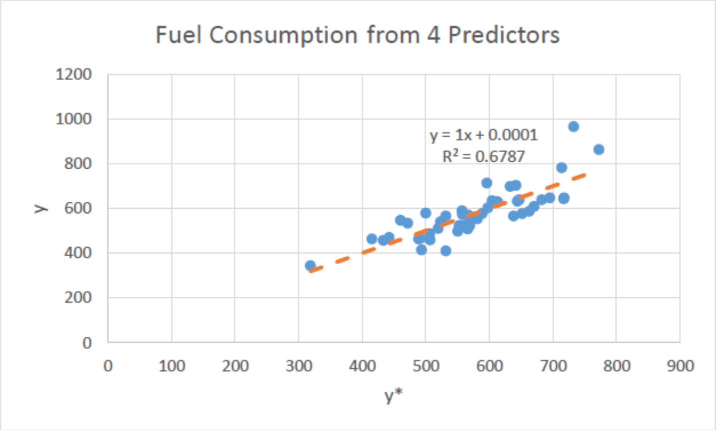
y^* here is just $y^*(\gamma = 0) = \hat{y}$

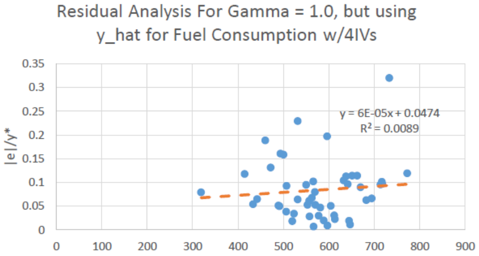


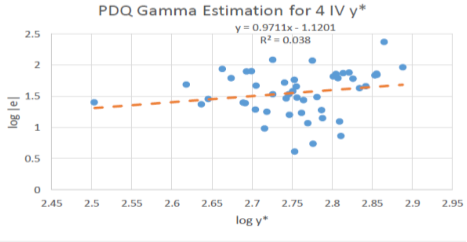


For two-predictors, the residual analysis method estimated $\gamma \cong 0.4$, and the PDQ method estimated $\gamma \cong 0.1$. About all we can say is that both estimates are less than 0.5, so heteroscedasticity has been dampened, or the sample size is not adequate for this application. Note: **n = 48**.

Next: For the four predictor model with an intercept term:







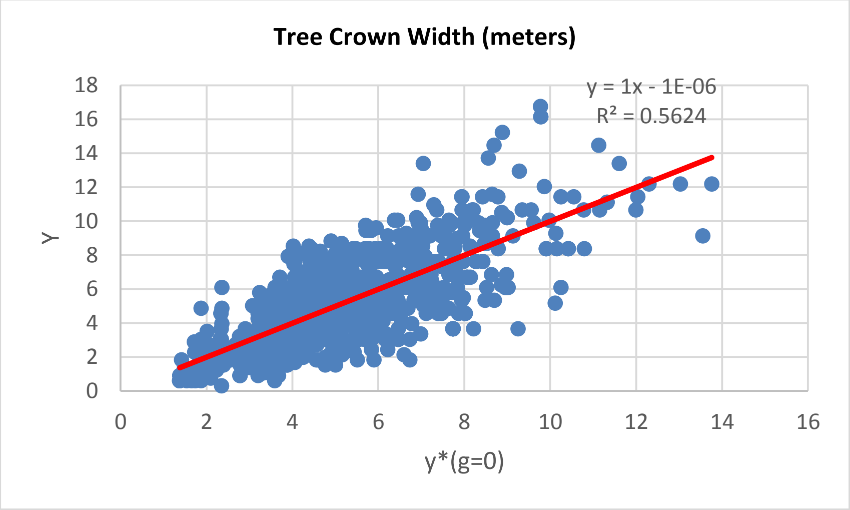
Above we see a PDQ estimate of $\gamma \cong 0.97$.

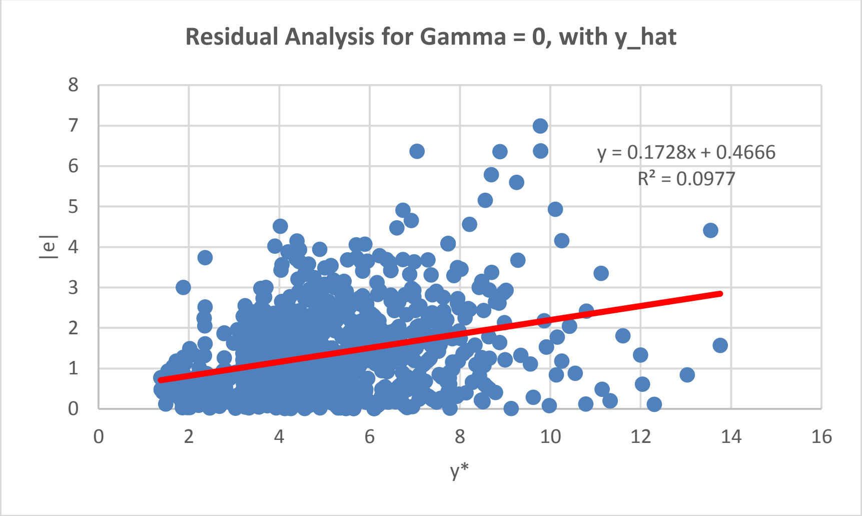
For the graphical residual analysis method, we estimate $\gamma > 1$, and the PDQ estimate here is 0.97. **n=48**.

Note: In Knaub(2019), when a potential outlier was removed as an experiment, the PDQ estimate of γ changed from 0.97 to 0.64.

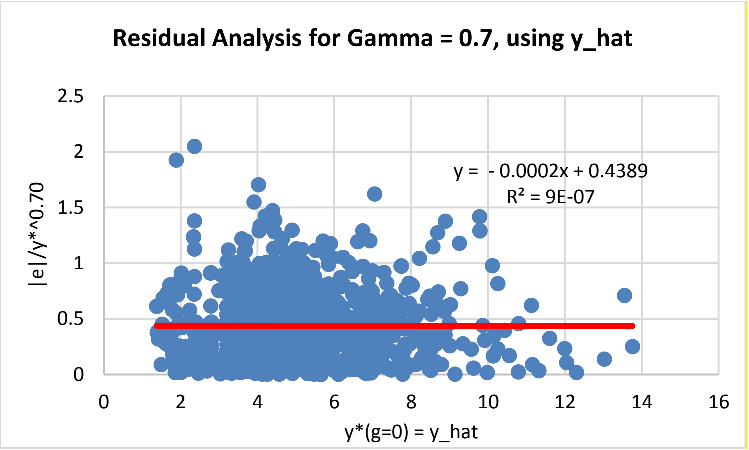
8.10 Forestry

Here we have tree crown width (for 1165 loblolly pines and 3 shortleaf pines) predicted by tree diameter, tree height, and their interaction term. The only input were y-values and the homoscedastic predicted-y values. For more details, see Knaub(2019), pages 30 to 33. (Data were provided by Dr. James A. Westfall, US Forest Service. His response to the author’s request for data is greatly appreciated.)

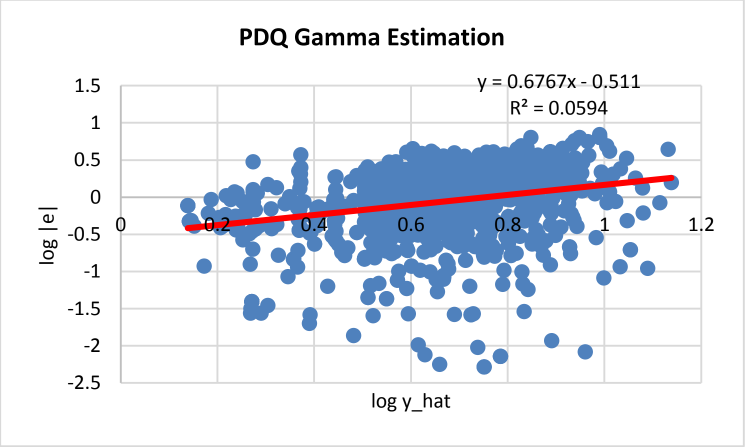




The upward slope indicates heteroscedasticity. Accounting for what coefficient of heteroscedasticity will give us a slope of zero?



$\gamma = 0.7$ is just a bit more than enough to give us a horizontal line for the first method of estimating γ (really, γ').



The two estimates of γ here are 0.70 and 0.68.

Compare Section 8.6 on Tennessee electric cooperatives data to this, Section 8.10, on forestry data. Both models clearly show heteroscedasticity. We know that the former is an example of a case where the size measure is highly desirable, according to the bottom of page 205 in Cochran(1953). The latter depends upon the relationship between different parts of a given tree, with regard to growth. The latter has much larger σ_i^*, but about the same heteroscedasticity. It appears to be a fairly simple model. Adding a variable that is not very helpful should just increase variance (Brewer(2002), pages 109-110, and Hastie, Tibshirani, and Friedman(2009), page 223), though your sample may be good. It does not appear that there are any categories to distinguish, as the data are for almost all one specific type of pine tree (1165 loblolly pines and 3 shortleaf pines). Perhaps this is as good as the model can be, and is just more useful for finding confidence of means than individual predictions. (However, geographic information might be considered.)

The second gamma estimation (PDQ) method, from Knaub(2019), proposed by Ken Brewer, using n = 1168 trees, yielded γ = 0.6767, s.e.(γ) = 0.0789, or better, γ = 0.68, s.e.(γ) = 0.08. (The first method yielded greater than 0.7.) For the Tennessee electric cooperatives data, using n = 9 establishments, and adjusting by the hat-value, yielded γ = 0.932, s.e.(γ) = 0.087, or better, γ = 0.93, s.e.(γ) = 0.09. (The first method yielded approximately 0.75 using the hat-value adjustment.) It is proposed that one would expect heteroscedasticity, generally such that $0.5 \leq \gamma \leq 1.0$, when we have a good size measure/predicted-y-values/model. For the Tennessee electric cooperatives we have a very good size measure, the current sampled data item in a previous census (Cochran(1953), page 205), where $Z = x$, or $y^* = bx$, but the sample size is very small. One case of poor data quality, which may be expected to more likely come from among the smaller such establishments, could lead to a change, such that in Knaub(2017a), the general default when performing many such regressions on a frequent basis for purposes of publishing official energy statistics was/is $\gamma = 0.5$. However, for the forestry data, one can expect more accurate results with the large sample size, unless data should have been stratified. It is proposed here that it appears that relevant data were used, such that one might expect a good model/size measure, for a narrow category of tree, though one might wonder about any other factors.

It seems promising that heteroscedasticity was so undeniable. Perhaps others with data sets they could check for heteroscedasticity will do so, using the Excel spreadsheet tool made available with Knaub(2019), or something more sophisticated.

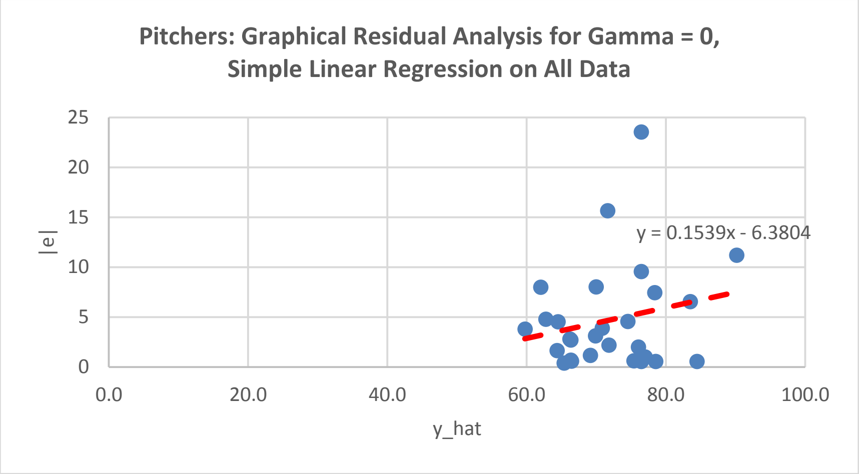
8.11 Baseball

Samaniego and Watnik(1997) is an interesting paper on Major League Baseball player payroll data as predictors for number of baseball games won. The pitcher and batter/hitter payroll data are available for all 28 teams for the year 1995.[18] The paper is with regard to separating the pitcher payroll data from the total player payroll data. Here, however, we are looking at the presence or absence of heteroscedasticity in the range Brewer justified. In general, ridiculously large coefficients of heteroscedasticity were indicated whether it was when only pitcher data were used, and an intercept term, or only the less useful ‘hitters’ payroll data, with an intercept term, or a multiple regression, with both payrolls and an intercept term. An exception was for the second (PDQ) method, for the multiple regression case, where γ was estimated to be -0.16 with a standard error of 2.69. For hitters with an intercept, the second method showed an estimate of $\gamma = 8.97$, with a standard error of 6.10! Obviously this is problematic. The sample size is small, and the year may have been unusual enough to account for this. One team, Cleveland, had an unusually large 100 games won. Toronto had a low number of wins, 56. For heteroscedasticity, the larger predicted-y values have larger estimated residuals, but neither of those teams were unusual for predicted wins. Still, first Cleveland data, and then Cleveland and Toronto data were removed from use by the pitcher payroll with an intercept term model, but this did not help.

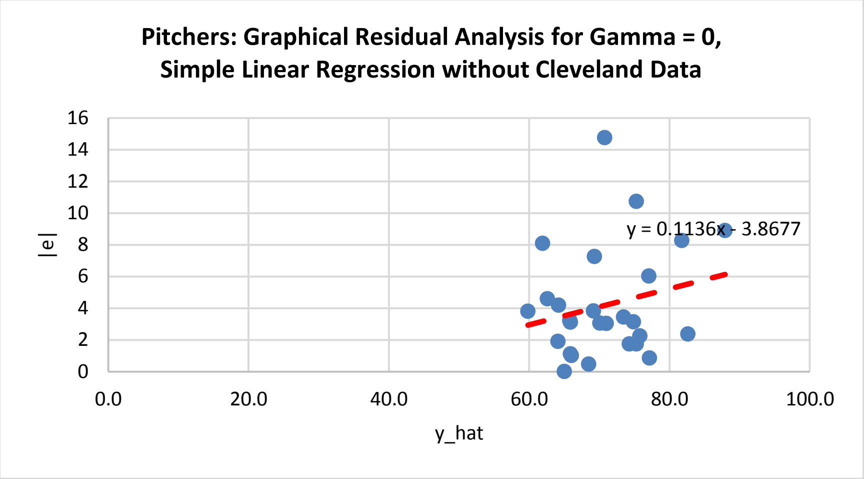
[18] The data source for Samaniego and Watnik(1997) was given as “…the November 17, 1995, issue of *USA Today*.”

Results for both methods for γ were still very high, beyond reasonable. However, for multiple regression, removing Cleveland, the first method estimated $\gamma = 0.75$, and the second method yielded $\gamma = 0.21$. Thus there may be a problem with the sample size and perhaps volatile data, or perhaps important information is missing. By several measures, accounting and not accounting for heteroscedasticity, the multiple regression method was only slightly better at predicting wins than the pitcher model, and not by nearly enough to make the additional variable worthwhile, according to the AIC and BIC. For all models, there is a large intercept, which reduced the range of predicted values greatly, which may help account for heteroscedasticity not being noticeable on all scatterplots. (Perhaps the unusual restriction that a change from a win to a loss for one team means a change for another team as well indicates an even larger than usual sample size is needed.)

Here we show the basic, starting graphical residual analysis, indicating that we do have heteroscedasticity. The scatterplot for a simple linear regression using the pitcher payroll data is given, followed by one where Cleveland is left out, and then Cleveland and Toronto. This is not an endorsement of dropping data without better data quality review procedures, it is just to see what the data might be saying about heteroscedasticity.

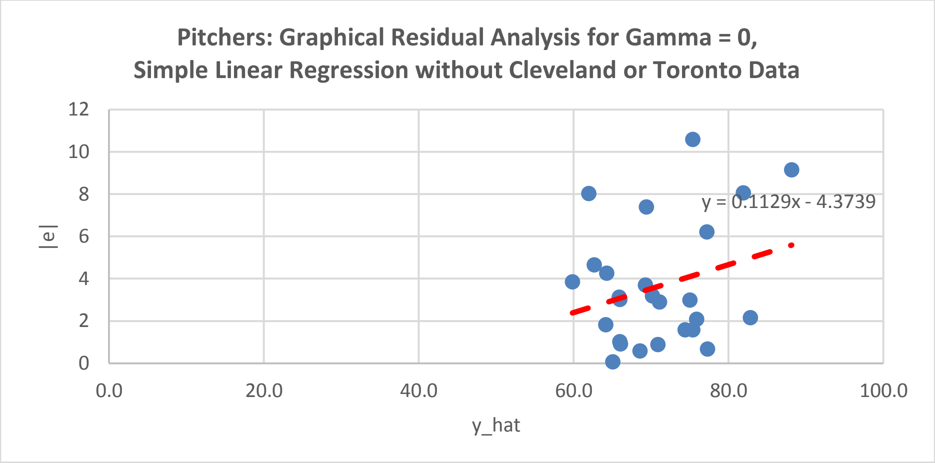


One might think that this is all driven by the data from Cleveland, which is that largest point for |e|. However, please note that the predicted value is in about the middle of the range. The next scatterplot shows that removing that point still leaves a steep slope indicating increase in the absolute value of the estimated residuals with increasing values of predictions, *i.e.*, heteroscedasticity.



Now the highest |e| belongs to Toronto, and it has a smaller predicted-y value.

This seems like a familiar, undesirable treatment of potential outliers! Throw one out, and another looks just like it. This could indicate none should be dropped. But more data would possibly be very helpful. There are two more major league teams now, but it would take a great deal more than that! This is a situation where the sample size is likely too small, and one would need to use a default, possibly $\gamma = 1$ for this application. A better predictor variable might help. We do not know.



9. SEARCHING FOR THE IDEAL PREDICTED-Y

The ideal predicted-y, unlike that in the arm strength example, will be associated with estimated residuals which vary more greatly with larger predicted-y. For the arm strength example, perhaps mid-range values of predicted-y had larger sigma of the estimated residuals because variables which would tell us about the age of the person, and his/her years to reach a given alcohol consumption level, and/or perhaps other such relevant variables, were omitted. (The principle of the hat-value adjustment tells us that a small increase in sigma of estimated residuals at mid-range of the predicted-y-values is to be expected with homoscedasticity of $V(Y_i)$ with an intercept term.) This tendency for homoscedasticity is more subtle than a missing categorical variable which could cause nonessential heteroscedasticity, but perhaps far more prevalent, though often unnoticed. Statistical learning, for example, will have complex situations where the achieved predicted-y may not often very closely approach the ideal predicted-y, and thus estimated residuals might be impacted by various factors along the range of predicted-y, say in perhaps a more complex version of the arm strength example.

Once, when asked to comment upon some ratio models another researcher was considering, the estimated variance of the prediction error apparently became larger with larger predicted-y, even though the coefficient of heteroscedasticity had been set at zero in the case of that one model. (Using γ = 0 for those models is something not generally to be encouraged unless there is a reason such as very low data quality for the smallest respondents, which artificially increases sigma for e_i for smaller predicted-y.) But why would this happen? Heteroscedasticity for e_i and heteroscedasticity for ε_i are two different things, as shown by Weisberg(1980), pages 100-106,[19] and it is $V(Y_i)$ which actually matters when specifying the model. But this would not explain increasing estimated variance of the prediction error. For that, we have the variance from the model coefficients, and that is likely ignored when estimating sigma.

Thus, when looking for the ideal predicted-y, y_i^*, we need to remember that when we estimate $V(\mathbf{Y})$, it is impacted by both $V(\mathbf{e})$, and $V(\mathbf{X\beta}^*)$. So we look for an impact of $\mathbf{X\beta}^*$ on $V(\mathbf{e})$ but it also relates to $V(\mathbf{X\beta}^*)$.

For the Longley employment example, we are told that ill-conditioning is a problem, which is apparently the reason Longley choose these data when testing computer algorithms for precision. (See SAS(1999), and IBM Support(2021).) Such a great deal of collinearity might make variable selection more difficult. (See Statsmodels(2021), and Weisberg(1980), pages 178-179.) Further, the intercept is the largest part of the regression for that example, by far, which does not appear to be ideal. The predicted-y-values range by only about 17%, providing little change in sigma, even for an ideal situation.

It is proposed that for predicted-y to be a good size measure, there should be an increase in $V(Y_i)$ as predicted-y becomes larger. Further, as we can see from the graph in Penn State(2021f), the coefficient of determination is based on comparing the part of the variance of y_i within a population which can be explained by a model (the sum of squares of differences between predicted-y and the unconditional mean of y), with the conditional variance of y_i, given the model ($V(Y_i)$). Usually we just care that predicted-y is close to y_i. A high coefficient of determination is not really a requirement, and it can be highly

[19] This prompted a note in the ResearchGate abstract for Knaub, J. (2007). Heteroscedasticity and homoscedasticity. In N. Salkind (Ed.), Encyclopedia of measurement and statistics. (pp. 431-432). Thousand Oaks, CA: SAGE Publications, Inc. doi: http://dx.doi.org/10.4135/9781412952644.n201 at https://www.researchgate.net/publication/262972023_HETEROSCEDASTICITY_AND_HOMOSCEDASTICITY: “Erratum: ‘variances of the predictions’ should be ‘variances of the Y_i.’"

misleading by itself, but it is good to know when the conditional $V(Y_i)$ are relatively small. Similarly, for the conditional $V(Y_i) = \sigma_i^2 = V(Y_i^*) + V(e_i)$, it would be nice if the $V(e_i)$ were not too large, but sometimes they just are large, as in the forestry example. The model, however, can still be good. It is proposed that the model in the forestry example behaves well because it does show heteroscedasticity. If another model were found with lower $\sigma_i^2 = V(Y_i^*) + V(e_i)$, that would be better, but the Y_i^* here are behaving as a good size measure. The Tennessee electric cooperatives data had a well behaved size measure and small σ_i^2. However, though the model for the Longley data had tight fits to the sample data, and questionable heteroscedasticity, model usefulness in general there may be questionable. So like a high R^2, heteroscedasticity, with $0.5 \leq \gamma \leq 1$, is just one indication that something has gone well. However, without a very large sample size, or enough other experience with the same model and application, γ may be difficult to determine.

10. MORE ON EXAMPLES

Real data examples above were considered to see if there was appreciable heteroscedasticity, with the baseline consideration that it was expected as long as predicted values differed in size, and to see what features of the models and data may have contributed to the appearance or disappearance of heteroscedasticity. In a thesis, Gelfand(2015), pages 6 through 11, it was noted that of 42 data sets representing a wide range of applications, 25 showed a general increase in absolute residuals from the smallest to the largest predicted-y values. (Note that an optimal random forest was first applied.)

There are many real data examples in which heteroscedasticity is either obvious, or present but not obvious. Some models are better than others. The arm strength example indicates that a missing variable (not categorical) can make the model miss heteroscedasticity. In the baseball example, an additional variable which does not clearly help (see Brewer(2002), pages 109 and 110) may not be justified, and there it may reduce heteroscedasticity, though the sample size is clearly not adequate to know this.

Considering which model is best in statistical learning: In Dalpiaz(2020), Section 6.1, Assessing Model Accuracy, Section 6.3, Test-Train Split, and Section 6.5, Choosing a Model, he uses RMSE, which he can be seen to base on sigma for the estimated residuals, except that he does not use the degrees of freedom of the model, but uses the sample size instead.

Let us consider this for a moment: As complexity of the model is increased, we can force a model to come closer to the training set (sample) of data points. That will reduce the sigma of the estimated residuals for the manifestation of the changed model, and thus reduce estimated RMSE for the training set. However, the sample should be 'representative' of the population or subpopulation to which it is to be applied. A more complex model determined from the training set may be more or less well suited to the test data set. If increased complexity of the model using the training data set can cause lower RMSE for the test data set, then it seems justified. Eventually, with more complexity, though the RMSE for the training data set continues to be reduced, the RMSE for the test data set will start to increase, because there is too much complexity to handle the eligible data in general. This emphasizes the importance of these two data sets. Neither can be atypical of the population or subpopulation to be modeled, or there is a problem of one kind or another. (See Dalpiaz(2020).)

There are three components of the expected [actually *squared*] prediction error, Err, or EPE, to consider: the sigma-squared of the ε_i (for the "irreducible"[20] part), the to-be-considered bias-squared of the model due to misspecification, and the variance of the estimated model due to its estimated coefficients. (See Hastie, Tibshirani, and Friedman(2009), page 223.) This is the variance of the prediction error, plus the unknown bias due to misspecification of the model. It is easy to consider one part and start to forget about the influence of another, and they are all estimated, the results of groping in the dark. There is a "bias-variance tradeoff" tendency as well. Increased complexity, such as an extraneous variable, tends to increase variance, but reduced complexity, such as in the example in the appendix to Shmueli(2010), where one predictor of two was lost, tends to increase bias. In that example, however, under some conditions, the biased model might have lower expected prediction error (EPE). In Knaub(2017a), on slides 39 and 40, a case was noted where increased complexity was helpful because it sometimes added a good deal of "explanatory power" as Brewer(2002), pages 109 and 110 noted would be a reason to do this.

Using RMSE is something like sigma for the estimated residuals, but without considering degrees of freedom. Kutner, Nachtsheim, and Neter(2004), page 424, use a subscript w to indicate when they are looking at the mean of the sum of the weighted squared estimated residuals, MSE_w. There $MSE_w = \frac{\sum w_i e_i^2}{n-p}$, where in the case of $e_i = e_{0_i} z_i^{\gamma'}$, $e_{0_i}^2 = w_i e_i^2$. So when we compare model results, perhaps we should compare the sum of weighted sigma-squared estimates. This is basically what the chi-square statistic essentially does, with a weight consistent with γ = 0.5.

In the baseball example, results for the simple linear regression using an intercept term and the payroll for only pitchers was compared to that using only hitters and an intercept term, and to the multiple regression using hitter payroll, and pitcher payroll separately as two predictors, plus an intercept term. Using both unweighted sum of squared errors and the chi-square, there appeared to be a big difference in each case between hitters only and pitchers only, and very little apparent "improvement" between the simple linear regression using only the pitcher payroll predictor and intercept term, and the multiple regression. (This seems consistent with the "separation principle" discussed in Samaniego and Watnik(1997).) Whether or not the multiple regression is better might be determined by test data, if there were any. The data set is so small that splitting it to form a training set and a test set would be problematic. As it is, the results of attempting to estimate gamma for each model using all data are dubious. Sample sizes are small, and one can surmise that a change to one data point might possibly have a substantial impact. As it is, it appears that gamma may be very large for the simple linear regression cases, but there is less evidence for the multiple regression case. Could collinearity have an impact? There is not enough information to tell, which may often be the situation. Looking at the scatterplots, nothing is obvious. Thus one might not even consider heteroscedasticity, though assuming

[20] See Singh(2018).

substantial heteroscedasticity may be better than assuming homoscedasticity, as is generally automatically done.

There are many interesting relationships between predictors. A suppressor variable,[21] for example, which may have little or no correlation with the response variable, will "suppress"[22] variance in one or more predictors, indicating that the relationships here are complicated. Shmueli(2010), page 6, notes in the first column that when we are only looking for the best predictions, we may want a biased model. (Biased in the sense of being misspecified, not necessarily model-biased where the expectation of the sum of estimated residuals is not zero. But here a misspecified model is meant.) We might surmise that the ideal predicted-y is more like the one without bias (the "correct" model, Shmueli(2000), page 6), though that may then mean dealing with collinearity and issues between predictors, such as suppression. The idea of considering the "ideal predicted-y" or best model then might be quite problematic, whether or not it might actually relate to when we would have heteroscedasticity. One might want to consider principle components.

Note that the sample size needed to accurately estimate the coefficient of heteroscedasticity may vary greatly by application. In some cases, sample size needs might not be so large. Cochran(1953), page 205 notes how helpful it can be if there is one predictor, which is the same data "item" from a previous census. See examples 8.6 and 8.7 above. There, sample sizes are n=9 and n=N=42, respectively. Knaub(2017a) and some references there, are with regard to many such small samples from numerous small populations of establishments, for official energy statistics, where γ can be expected to fall such that $0.5 \leq \gamma \leq 1.0$, but $\gamma = 0.5$ is generally the default used to hedge against the possibility of data quality issues for smaller respondents to the frequently occurring sample. Brewer(2002), notably on page 137, considers that very large sample sizes are appropriate for estimating γ, but that certain default values for γ might be used, depending upon the application. The supplementary spreadsheet tool for Knaub(2019) has a sheet commenting on this.

11. CIRCUMSTANCES UNDER WHICH HOMOSCEDASTICITY IN REGRESSION MIGHT APPEAR TO BE APPROPRIATE

OLS regression may be the first stop when most new models requiring a continuous response are developed. Certainly this is encouraged in the literature. When one first learns regression, you are told that one of the main requirements is homoscedasticity. Other developments may depend upon it. But there are situations which cause heteroscedasticity (Knaub(2018)) which can be 'corrected,' such as omitted categorical variables, which might be identified, or one may be encouraged to apply a transformation to "fix" the natural, 'essential' heteroscedasticity (Knaub(2017b)) that is present. This might or might not work very well, and may or may not cause problems in interpreting the results. In this paper, essential heteroscedasticity is considered to be an important feature to be treated as part of the error structure. First, we will review why this should be the case.

[21] Suppressor variable complications can be quickly noted in Ludlow and Klein(2014), in the first part of the Introduction, pages 1 and 2, and the Conclusions on pages 21 and 22.

[22] Lancaster(1999), page 5.

In summary: In William Cochran's 1953 textbook, Sampling Techniques, Cochran(1953), on page 199 he notes an already by then well-established empirical expression for the within agricultural plot or cluster variance: S_w^2=AM^g. On page 212, Cochran notes that when z_i is a relative size measure, and there is a straight line relationship 'through' the origin, between y_i and z_i, that only a few studies had reported the relationship, between z_i and the variance for y_i, see page 211, but those that did report this had found this variance to increase at a rate "…between az_i and az_i^2." There Cochran noted that this was the variance for e_i, but wrote this such that no other random variable was involved, and thus it is ε_i in regression, and the variance is for Y_i. There z_i is $M_i/\sum M_j$, where M_i is the size of unit i, so z_i is a relative size measure. Thus, Brewer(2002), mid-page 111, explains why a long established empirical relationship was found. This indicates that for $Y_i = bx_i + e_i$, and from there $Y_i = y_i^* + e_i$, can be described such that $\sigma_i^2 = V(Y_i) = az_i^{2\gamma}$, where "…between az_i and az_i^2" means that we have $0.5 \le \gamma \le 1$. Thus, for the "gamma population model" on page 49 in Chambers and Clark(2012), where they have $E(y_i|z_i) = \beta z_i$ and $Var(y_i|z_i) = \sigma^2 z_i^{2\gamma}$, we have $0.5 \le \gamma \le 1$. Brewer(2002), page 111, shows why this was found to be the case.

So this is for the model of form $Y_i = bx_i + e_i$. Why should this not carry over to every model of form $Y_i = y_i^* + e_i$?

There are obvious cases where the researcher might just note heteroscedasticity. Consider the examples of Sections 8.4 and 8.10. In 8.4, the authors of the article for that example of Spanish shop data simply stated that there was heteroscedasticity, and how they addressed it. In 8.10, the forestry data was supplied for purposes of the experiment there. (That worked well, and it was for multiple regression.)

Below are some ideas as to why heteroscedasticity may not be seen, or we may just think it is not seen, or perhaps do not consider it.

11.1 Sometimes there is heteroscedasticity, but it may not be apparent if you are not looking for it. Consider the Tennessee electric power cooperatives example in Section 8.6. We first look at the scatterplot of y and x. The estimated residuals, which will generally give us an idea of the actual residuals, are small enough that heteroscedasticity may not be obvious from the first scatterplot. This was also true for Section 8.7, North Dakota total electric sales. Further, with small sample sizes, heteroscedasticity might not be obvious even from a graphical residual analysis, or even completely determinable, as in the case of the baseball example in Section 8.11, where there are indications of possible heteroscedasticity, or the Longley employment and related data, Section 8.5, where ill-conditioning and perhaps fitting too many variables to a small sample may only make it appear to be homoscedastic. Also for the baseball example, and particularly for the Longley data, there is a fairly short range of the predicted-y (or size measure). This makes heteroscedastic differences smaller, such that they may be overwhelmed by general randomness, and thus less noticeable. (This seems plausible for the baseball example, but we will put the Longley example under 11.3, below.) The range of predicted-y values for the percent body fat example of Section 8.8 is relatively short, and even with a sample size of 92, the graphical residual analysis did not show the typical 'fan-shaped' pattern. The density of estimated residuals in the y-direction, as we move to larger predicted values is decreased, and therefore variance increases, but without causing a 'fan-shape' in that case.

11.2 Heteroscedasticity might be apparent, but not considered important to the current application. This could be the case with the home natural gas use and Kenyan sex worker examples of Sections 8.2 and 8.3, respectively. It might be considered indirectly.

11.3 Heteroscedasticity may be countered by model and/or data issues.

This appears to be the case in the example of the effect of alcoholism on arm strength given in Section 8.1, where missing information may have dampened heteroscedasticity, as discussed there. In the case of official energy statistics discussed in Knaub(2017a), one might often estimate γ to be between 0.7 and 0.9, but routinely use 0.5 to guard against data quality issues with the smallest respondents in the sample. This artificially increases variance near the origin, lowering the effective value of γ.

However, in Section 8.9, the motor fuel consumption case, it was not clear why the addition of two more predictors made heteroscedasticity of the order expected as noted in Brewer(2002). Perhaps that selection of predictors worked well together and the collinearity was not so extreme as in the Longley case of Section 8.5, but it is not clear. The sample size may have been too small to make results reproducible. Also, in the Longley example, autocorrelation may have had an impact.

11.4 Data May be Artificial and Designed to be Homoscedastic

Examples above were picked in a search for real data examples, while avoiding copyright issues. Many times one might encounter an example that looks promising, only to discover that it was manufactured as an example to illustrate how to do regression analysis. Real data examples are often messy. One may wish to illustrate a given point only to have it obscured by other issues in a real data example. Thus, real data examples, especially outside of your work experience, may be difficult for an individual to obtain. Examples here were located and requested for this purpose.

11.5 Suggestion

Perhaps readers may wish to look for heteroscedasticity in regression using their own data. The idea is to encourage everyone to think of heteroscedasticity in regression as the norm.

12. CONCLUSIONS

Think of each y-value as having at its core, a predicted value made up of infinitely many smaller, equal size elements. (If one element were bigger, it could be broken down. Therefore, there are infinitely many infinitesimal, equal size elements. Each also has its own associated “random error” to consider.) So each y-value is a cluster of elements which has variance, as it is a realization of a random variable. The within cluster variance is then the square of sigma for the y-value, $V(Y_i) = S_w^2 = AM^g = \sigma_i^2$, which can be said to be “irreducible.” (For the simplest case, $\sigma_i^2 = \sigma_{\epsilon_0}^2 x_i^{2\gamma}$= $Var(y_i|x_i)$.) Thus every different predicted-y value would be associated with a y-value and therefore an estimated random 'error,' e, which would generally, but not always, be of larger magnitude for larger predicted-y. Ken Brewer explained this for survey populations using retail stores for illustration. So why wouldn't it always be true? Perhaps it is true when model selection is at its best. So instead of thinking of heteroscedasticity as a problem that needs to be 'fixed,' we should really be thinking that there may be a problem if you have homoscedasticity.

Yes, textbooks often say that linear (and other) regressions call for homoscedasticity, or seem to encourage it, but that is just a reflection of the mathematics traditionally used, not reality.

Inherently, for complicated cases in machine learning for example, though a balance between bias and variance is sought, it may often not be possible to have the achieved predicted-y accurately mimic the behavior of the ideal predicted-y with regard to associated variance. As in the simpler example of arm strength, variability will be impacted by contradictions between the ideal and achieved predicted-y. This will complicate the issue, but under it all is the concept of essential heteroscedasticity, which will show itself whenever it is not suppressed. As we see on page 111 in Brewer(2002), "… homoscedasticity is the exception…," at least with regard to "…sample survey populations," and here we contend that that underlies all applications. However, more complex regressions tend to make it harder to obtain "achieved" predicted-y values which accurately mimic the behavior of the "ideal" predicted-y values, in terms of associated variance. Larger predicted-y values should generally correspond to larger sigma for the estimated residuals. This is due primarily to essential heteroscedasticity. If this is not actually the case, then it is proposed that the achieved predicted-y may not be functioning closely to the ideal predicted-y, with regard to variance.

Many times, OLS (homoscedastic) regression is assumed, but when checked, is easily shown to not be the case. On other occasions, it may be less conspicuous, but nevertheless true. Yet other applications which might stay in one's memory from learning regression may have been contrived cases, where artificial data were used which were designed to be homoscedastic. To automatically assume homoscedasticity is to use gamma = 0 as a default, which may be convenient, but far from the truth.[23]

So, When Would Heteroscedasticity in Regression Occur? The answer proposed is that it should occur when you have the ideal predicted-y, or one that is close enough. It appears that the more complex the model that is used or that needs to be used, the less likely this is to be achieved.

ACKNOWLEDGMENTS

Thank you to the late Dr. Kenneth R.W. Brewer, friend and mentor, for many helpful (mostly email and previously FAX) conversations from which I benefitted greatly. Any misconceptions I might have developed would be due to wanting on my part. Ken went to great lengths to support and fortify my efforts. He was also a good friend to me and to others. He will be missed. Please see the appendix for a tribute to Ken.

My thanks also to Prof. Muhammad Hanif who was a co-author and good friend to Ken Brewer, and is an internationally recognized statistician who encouraged me with regard to this paper.

[23] Consider this project once again: https://www.researchgate.net/project/OLS-Regression-Should-Not-Be-a-Default-for-WLS-Regression.
There, an update, "When Would Heteroscedasticity Occur/Not Occur?" was posted on March 22, 2020. It was an earlier effort to consider ideas for this paper.

Also -

Thank you to the following people for their assistance in obtaining permission to use graphics or obtain data, and/or provide information so that real data could be used for examples here:

Ronán M Conroy, Royal College of Surgeons in Ireland
David Dalpiaz, University of Illinois, Urbana-Champaign
Rob Finger, US National Renewable Energy Laboratory
Jim Frost, Statistics by Jim
John Haubrick, Penn State Department of Statistics
Frank Samaniego, University of California, Davis
Mitch Watnik, Cal State, East Bay
Jim Westfall, US Forest Service

Thanks to Phil Kott, now of RTI International, for the Cochran(1953), page 212 reference to the historical, empirical range of γ, explained by Ken Brewer.

And thank you to Karen, my wife, for her love and support, and to our son, Seth, for his.

REFERENCES

1. Brewer, K.R.W. (1963). Ratio Estimation in Finite Populations: Some Results Deducible from the Assumption of an Underlying Stochastic Process. *Australian Journal of Statistics*, 5, 93-105.
2. Brewer, K.R.W. (2002). *Combined survey sampling inference: Weighing Basu's elephants*, Arnold London and Oxford University Press.
3. Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*, Chapman and Hall.
4. Chambers, R. and Clark, R. (2012). *An Introduction to Model-Based Survey Sampling with Applications*, Oxford Statistical Science Series.
5. Cochran, W.G. (1953). *Sampling Techniques*, 1st ed., John Wiley & Sons.
6. Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed., John Wiley & Sons.
7. Dalpiaz, D. (2020). R for Statistical Learning, Book in progress, *This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License*. Retrieved from https://daviddalpiaz.github.io/r4sl/, and https://daviddalpiaz.github.io/r4sl/r4sl.pdf, May 4, 2021.
8. Elmore-Meegan, M., Conroy, R.M. and Agala, C.B. (2004). Sex Workers in Kenya, Number of Clients and Associated Risks: An Exploratory Survey, *Reproductive Health Matters*, 12(23), 50-57. https://www.tandfonline.com/doi/abs/10.1016/S0968-8080(04)23125-1 (Graph provided to this author separately by Dr. Ronán M. Conroy.)
9. Faraway, J.J. (2002). *Practical Regression and ANOVA using R*, https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf, pages 59-62.
10. Fox. J. (2008). *Applied Regression Analysis and Generalized Linear Models*, 2nd ed, 2008, Sage Publications, Inc.

11. Frost, J. (2019). Statistics by Jim, “Making Predictions with Regression Analysis,” https://statisticsbyjim.com/regression/predictions-regression/, found under “Regression Tutorial with Analysis Examples,” https://statisticsbyjim.com/regression/regression-tutorial-analysis-examples/, downloaded April 25, 2021.
12. Gelfand, S. (2015). *Understanding the impact of heteroscedasticity on the predictive ability of modern regression methods*, Master’s thesis, Simon Fraser University, Canada. https://www.stat.sfu.ca/content/dam/sfu/stat/alumnitheses/2015/SharlaGelfandProject.pdf
13. Guadarrama, M., Molina, I. and Tillé, Y. (2020). Small area estimation methods under cut-off sampling, *Survey Methodology*, Statistics Canada, Catalogue No. 12-001-X, 46(1), 51-75. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2020001/article/00004-eng.htm
14. Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, 2nd Ed (corrected at 7th printing 2013). Springer.
15. IBM Support (2021). *Diagnosing ill conditioning*, IBM, retrieved April 25, 2021, https://www.ibm.com/support/pages/diagnosing-ill-conditioning
16. Knaub J.R., Jr. (1992). More Model Sampling and Analyses Applied to Electric Power Data, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 876-881. http://www.asasrms.org/Proceedings/papers/1992_148.pdf
17. Knaub J.R., Jr. (1993). Alternative to the Iterated Reweighted Least Squares Method – Apparent Heteroscedasticity and Linear Regression Model Sampling. *International Conference on Establishment Surveys*, https://www.researchgate.net/publication/263809034_Alternative_to_the_Iterated_Reweighted_Least_Squares_Method_-_Apparent_Heteroscedasticity_and_Linear_Regression_Model_Sampling, or in Section VII in https://ww2.amstat.org/meetings/ices/1993/
18. Knaub, J.R., Jr. (2012). Use of Ratios for Estimation of Official Statistics at a Statistical Agency, *InterStat,* May 2012, http://interstat.statjournals.net/. https://www.researchgate.net/publication/261508465_Use_of_Ratios_for_Estimation_of_Official_Statistics_at_a_Statistical_Agency
19. Knaub, J.R., Jr. (2017a). Quasi-Cutoff Sampling and the Classical Ratio Estimator - Application to Establishment Surveys for Official Statistics at the US Energy Information Administration - Historical Development. Invited Presentation, *Conference: EIA Math/ Stats Lunch*, September 2017, DOI: 10.13140/RG.2.2.33300.60803/1, *ResearchGate*, https://www.researchgate.net/publication/319914742_Quasi-Cutoff_Sampling_and_the_Classical_Ratio_Estimator_-_Application_to_Establishment_Surveys_for_Official_Statistics_at_the_US_Energy_Information_Administration_-_Historical_Development
20. Knaub, J.R., Jr. (2017b). *Essential Heteroscedasticity*, November 2017, DOI: 10.13140/RG.2.2.20928.64005, *ResearchGate*, https://www.researchgate.net/publication/320853387_Essential_Heteroscedasticity
21. Knaub, J.R., Jr. (2018). *Nonessential Heteroscedasticity*, April 2018, DOI: 10.13140/RG.2.2.27972.73603, *ResearchGate*, https://www.researchgate.net/publication/324706010_Nonessential_Heteroscedasticity

22. Knaub, J.R., Jr. (2019). Estimating the Coefficient of Heteroscedasticity, June 2019, https://www.researchgate.net/publication/333642828_Estimating_the_Coefficient_of_Heteroscedasticity with a tool for implementing this found at https://www.researchgate.net/publication/333659087_Tool_for_estimating_coefficient_of_heteroscedasticityxlsx
23. Kutner, M., Nachtsheim, C. and Neter, J. (2004). *Applied Linear Regression Models*, 4th ed., McGraw-Hill/Irwin.
24. Lancaster, B. (1999). Defining and Interpreting Suppressor Effects: Advantages and Limitations, Presented at the *Annual Meeting of the Southwest Educational Research Association*, San Antonio, January, 1999. Posted by the Educational Resources Information Center (ERIC) to https://eric.ed.gov/?id=ED426097
25. Lind, J.T. (2004). *The Variance of the Prediction Error*, University of Oslo, ECON4150, Spring 2004, Retrieved from https://www.uio.no/studier/emner/sv/oekonomi/ECON4150/v04/seminar/, Var_f.pdf, May 21, 2021.
26. Lohr, S.L. (2010). *Sampling: Design and Analysis*, 2nd ed., Brooks/Cole.
27. Ludlow, L. and Klein, K. (2014). Suppressor Variables: The Difference between 'Is' versus 'Acting As'. *Journal of Statistics Education*, 22(2), DOI: 10.1080/10691898.2014.11889703
28. Maddala, G.S. (1977). *Econometrics*, McGraw-Hill.
29. Maddala, G.S. (2001). *Introduction to Econometrics*, 3rd Ed., Wiley.
30. NIST Information Technology Laboratory (2021). *NIST Standard Reference Database 140*, National Institute of Standards and Technology, US Commerce Department, Retrieved from http://www.itl.nist.gov/div898/strd/lls/data/Longley.shtml, April 25, 2021. [Last update to Data Content: 2003 DOI: http://dx.doi.org/10.18434/T43G6C]
31. Penn State (2021a). *STAT 462: Applied Regression Analysis, 4.2 - Residuals vs. Fits Plot*. Retrieved from https://online.stat.psu.edu/stat462/node/117/, April 16, 2021.
32. Penn State (2021b). *STAT 462: Applied Regression Analysis, 4.11 - Prediction Interval for a New Response*. Retrieved from https://online.stat.psu.edu/stat462/node/127/, May 9, 2021.
33. Penn State (2021c). *STAT 462: Applied Regression Analysis, 9.2 - Using Leverages to Help Identify Extreme X Values*. Retrieved from https://online.stat.psu.edu/stat462/node/171/, May 9, 2021.
34. Penn State (2021d). *STAT 462: Applied Regression Analysis, 9.3 - Identifying Outliers (Unusual Y Values)*. Retrieved from https://online.stat.psu.edu/stat462/node/172/, May 18, 2021.
35. Penn State (2021e). *STAT 462: Applied Regression Analysis, 9.4 - Studentized Residuals*. Retrieved from https://online.stat.psu.edu/stat462/node/247/, May 18, 2021.
36. Penn State (2021f). *STAT 500: Applied Statistics, 9.3 – Coefficient of Determination*. Retrieved from https://online.stat.psu.edu/stat500/lesson/9/9.3, May 23, 2021.
37. Penn State (2021g). *STAT 501: Regression Methods, 13.1 - Weighted Least Squares.* Retrieved from https://online.stat.psu.edu/stat501/lesson/13/13.1, May 13, 2021.
38. Roberts, D., Merket, N., Polly, B., Heaney, M., Casey, S. and Robertson, J. (2012). *Assessment of the U.S. Department of Energy's Home Energy Scoring Tool*, National Renewable Energy Laboratory (NREL). https://www.nrel.gov/docs/fy12osti/54074.pdf.
39. Royall, R.M. (1970). On Finite Population Sampling Theory under Certain Linear Regression Models. *Biometrika*, 57, 377-387.

40. Royall, R.M. (1992). The model based (prediction) approach to finite population sampling theory. *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, 17, 225-240. The paper is available under Project Euclid, open access: https://projecteuclid.org/euclid.lnms/1215458849.
41. Samaniego, F.J. and Watnik, M.R. (1997). The Separation Principle in Linear Regression. *Journal of Statistics Education*, 5(3), DOI: 10.1080/10691898.1997.11910599.
42. Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlang.
43. SAS (1999). *Longley Data*, OnlineDoc, Version 8, SAS Institute Inc., Cary, NC, USA, https://v8doc.sas.com/sashtml/stat/chap48/sect3.htm
44. Shmueli, G. (2010). *To Explain or to Predict*? Electronic reprint version on ResearchGate, found at https://www.researchgate.net/publication/48178170_To_Explain_or_to_Predict. Original article published by the *Institute of Mathematical Statistics in Statistical Science*, 2010, 25(3), 289-310.
45. Singh, S. (2018). *Understanding the Bias-Variance Tradeoff*, Towards Data Science. Retrieved from https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229, May 23, 2021.
46. Statsmodels (2021). *Longley Dataset*, statsmodels.org. Retrieved from https://www.statsmodels.org/stable/datasets/generated/longley.html, April 25, 2021.
47. Valliant, R, Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*, Wiley Series in Probability and Statistic.
48. Weisberg, S. (1980). *Applied Linear Regression*, John Wiley & Sons, Inc.

APPENDIX A

A.1 Remembering Ken Brewer:

During the first International Conference on Establishment Surveys (ICES). June 1993 in Buffalo, New York, USA, there was a special session in the program which I found to be very interesting. It was the talk given by Ken Brewer which was later published in a book, Business Survey Methods, Wiley, pages 589-606, 1995, "Combining Design-Based and Model-Based Inference." I did not introduce myself to Ken that day, but thought about what he had to say, and was intrigued. Though he advocated combining these two methods, I was trying to learn more about the modeling aspect, which had great potential for the many small populations monitored by energy establishment surveys for official statistics in which I was involved and would become involved at the US Energy Information Administration (EIA). So I decided to contact Ken to see if he would discuss this with me.

As it turned out, no one could have been more helpful, in spite of his very busy schedule. Ken described the application of the coefficient of heteroscedasticity to me in great detail. This started before email was used as exclusively as it is now, and before the use of attachments, so Ken sent FAXed pages of equations and handwritten descriptions. We eventually used email, exclusively. A faux pau on my part occurred when I tried to discuss the way the estimated residuals are factored, but instead of using the word "factor," I used the word "component," which threw off the discussion. Ken thought I was referring to a different format for the nonrandom part of the residuals, and for the regression weights. When Ken eventually realized my mistake, breaking my habit of using "component" when I mean "factor" became a high priority for me. (Ken also was a stickler about using the term "estimated residuals," rather than "residuals" when that is what you mean, which again I can see as an important distinction.)

Ken and I carried on email discussions for many years, seldom seeing each other in person. (The first time we saw each other, he said he had me confused with someone else, but I told him that we had not met in Buffalo, 1993. I only heard him speak.) But finally, in the summer of 2002, Ken was in the US again, perhaps for both the Joint Statistical Meetings that year in New York, and other meetings afterward, I do not exactly recall his schedule, but he was in Washington DC in August, and visited my family and me at our home one evening. His book had come out, but he was puzzled about something. That anecdote follows.

A.2 Anecdote

At the Joint Statistical Meetings, Ken told me that his (2002) book was to be found at the conference book display and sales area, and suggested I go see it. When I got to that store, the sales person showed me his book. It was quite thick as it was printed on only one side of each page, for some reason. She suggested that I purchase it immediately, but I just wanted to place an order for delivery, and not have to squeeze it into my luggage. However, she was very insistent, and wore down my resistance until I finally agreed.

When Ken later visited us at home, he said the following, or something very similar: "We don't know what happened to the two preprint copies of my book." It was a mystery! I said that I knew where one of them could be found! I showed the book to him and he autographed it for me. Apparently, the insistent salesperson was not supposed to have sold that "preprint' at all!

Attached is a picture of Ken with that autographed preprint. I bought another, regular copy, for regular use.

Also note his Waksberg Award paper: Brewer (2014). “Three controversies in the history of survey sampling,” Survey Methodology, Dec 2013, https://www150.statcan.gc.ca/n1/pub/12-001-x/2013002/article/11883-eng.htm. (Note that Ken comments there on modeling and small sample sizes.)

A.3 Conclusion

So, that is my memory of Ken Brewer: A fervent statistician, helpful and eager to discuss the area in which he chose to work. He, as a British-Australian, also sent a large book of marsupials, and a boomerang to our son. He was an A.A. Milne fan. He was a good friend to many.

Ken Brewer
August 2002
Arlington, Virginia, USA

Also please note that Ken Brewer wrote about his own mentor, Ken Foreman, in his very entertaining style: Brewer, K.R.W. (2005). Anomalies, probings, insights: Ken Foreman's role in the sampling inference controversy of the late 20th century. Australian and New Zealand Journal of Statistics, 47, 4, 385-399.

Over the years I have talked to people who met or worked with Ken Brewer, and everyone respected him. More importantly, Ken was a decent person who cared about others. Condolences to his wife, Maggie. Ken will be missed.