THE EFFECTS OF VARIANCE FUNCTION ESTIMATION ON PREDICTION AND CALIBRATION : AN EXAMPLE

Raymond J. Carroll

University of North Carolina at Chapel Hill

Research supported by the Air Force Office of Scientific Research Contract AFOSR-F-49620-85-C-0144.

KEY WORDS AND PHRASES : Weighted least squares, Heteroscedasticity, Regression.

ABSTRACT

We consider fitting a straight line to data when the variances are not constant. In most fields, it is fairly common folklore that how one estimates the variances does not matter too much when estimating the regression function. While this may be true, most problems do not stop with estimating the slope and intercept. Indeed, the ultimate goal of a study may be a prediction or a calibration. We show by an example that how one handles the variance function can has large effects. The point is almost trivial, but so often ignored that it is worth documenting. Additionally, this points out that one ought to spend time trying to understand the structure of the variability, a theoretical field that is not particularly well developed.

1 : INTRODUCTION

Consider a heteroscedastic regression model, in which we observe N pairs (y,x) following the model

$$E(y|x) = f(x,\beta) \; ; \tag{1.1}$$

$$\text{Standard Deviation}(y|x) = \sigma g(x,\beta,\theta) \; . \tag{1.2}$$

While our remarks hold generally, in what follows it suffices to consider the special case of linear regression for the mean and the power of the mean model for the standard deviation, i.e.,

$$f(x,\beta) = \beta_0 + \beta_1 x \quad : \quad g(x,\beta,\theta) = f(x,\beta)^{\theta} \; . \tag{1.3}$$

When $\theta = 0$, we have the homoscedastic regression model, and unweighted least squares will ordinarily be used to estimate β. For other values of θ, generalized least squares can be used to estimate β, see Carroll & Ruppert (1987) for a discussion and a review of the literature. Generalized least squares is weighted least squares with estimated weights. The version of generalized least squares used here for each θ is fully iterated reweighted least squares, sometimes called quasi-likelihood, see McCullagh & Nelder (1983). In practice, θ is unkown and must be estimated. The theory of such estimation is given by Davidian & Carroll (1986).

The common folklore theorem of generalized least squares states that as long as one's estimate $\hat{\theta}$ of θ is root-N consistent, the resulting generalized least squares estimate has the same asymptotic distribution as if θ were known.

See Judge, et al (1985) and Carroll & Ruppert (1982, 1987) for references and proofs. Indeed, any generalized least squares estimate has the same limit distribution as weighted least squares based on the correct weights, i.e., the inverse of the square of (1.2).

The folklore theorem has an analogue in practice. In the linear regression model with a reasonably sized data set, since unweighted least squares is consistent its fitted values rarely differ much from the fitted values from a generalized least squares fit. Consequently, the usual practice is to treat the estimation of the variance function $g(x,\beta,\theta)$ fairly cavalierly, if at all. To quote Schwartz (1979), "there is one point of agreement among statistics texts and that is the minimal effect of weighting factors on fitted regression curves. Unless the variance nonuniformity is quite severe, the curve fitted to calibration data is likely to be nearly the same, whether or not the variance nonuniforming is included in the weighting factors". The narrow focus on estimating the mean is misplaced, as Schwartz later notes, see also Garden, et al (1980). Sometimes the variance function is itself of importance. Box & Meyer (1985) state that "one distinctive feature of Japanese quality control improvement techniques is the use of statistical experimental design to study the effect of a number of factors on variance as well as the mean". Other times the variance function essentially determines the quantity of interest. This occurs, for example, in the estimation of the sensitivity of a chemical or biochemical assay, see Carroll, Davidian & Smith (1986). However, there are even more basic problems where the variance function is of considerable importance, namely prediction and calibration.

It is perhaps trite to state that how well one estimates the variance function has a large effect on how well one can do prediction and calibration. It is, however, a point that is rarely taken into account in practice, as any

review of the (rudimentary) techniques in the assay literature will quickly show. There are two ways to see this point. The first is through an asymptotic theory outlined in section 3, where we show that the difference in the length of a prediction interval between θ known and unknown is asymptotically distributed with variance a monotone function of how well one estimates θ. The second and probably more useful way to see the effect of variance function estimation is through an example. The large costs involved in not weighting at all will be evident in this example, and will serve as an object lesson.

2. : CALIBRATION AND PREDICTION

Calibration experiments start with a training or calibration sample $(y_1,x_1),\ldots,(y_N,x_N)$ and then fit models to the mean and variance structures. The real interest lies in an independent pair $(y_\square,x_\square)$. Sometimes $x_\square$ is known and we wish to obtain confidence intervals for $y_\square$; this is prediction. Other times, $y_\square$ is easily measured but $x_\square$ is unknown and inference is to be made about it, see Rosenblatt & Spiegelman (1982).

For example, in an assay x might represent the concentration of a substance and y might represent a counted value or intensity level which varies with concentration. One will have a new value $y_\square$ of the count or intensity and wish to draw inference about the true concentration $x_\square$. The calibration sample is drawn so that we have a good understanding of how the reponse varies as a function of concentration. The regression equation relating the response to

concentration is then inverted to predict the concentration from the observed response.

For the remainder of this section we will assume that the responses are normally distributed, although this can be relaxed. Given a value $x_\square$, the standard point estimate of the response $y_\square$ is $f(x_\square,\beta)$. Let $\hat{\beta}_G$ be a generalized least squares estimate, and define

$$S_G = S_G(\theta) = N^{-1} \sum_{i=1}^{N} f_\beta(x_i,\hat{\beta}_G)\, f_\beta(x_i,\hat{\beta}_G)^T \,/\, f(x_i,\hat{\beta}_G)^{2\theta} ,$$

$$\hat{\sigma}^2(\theta) = N^{-1} \sum_{i=1}^{N} \{y_i - f(x_i,\hat{\beta}_G)\}^2 \,/\, f(x_i,\hat{\beta}_G)^{2\theta} ,$$

where f_β is the derivative of f with respect to β. For large calibration data sets, the variance in the error made by prediction is

(2.1) $$\text{Variance}\{y_\square - f(x_\square,\hat{\beta}_G)\} \cong \sigma^2\, q_N^2(x_\square,\beta,\theta), \text{ where}$$

$$q_N^2(x_\square,\beta,\theta) = g^2(x_\square,\beta,\theta) + N^{-1} f_\beta(x_\square,\beta)^T\, S_G^{-1}\, f_\beta(x_\square,\beta).$$

Note that if the size N of the calibration data set is large, then the error in prediction is determined predominately by the variance function

$$\sigma^2\, g^2(f(x_\square,\beta),x_\square,\theta),$$

and not by the calibration data set itself. An approximate $(1-\alpha)100\%$ confidence interval for the response $y_\square$ is given by

$$(2.2) \qquad I(x_{\square}) = \{\text{all values y in the interval } f(x_{\square},\hat{\beta}_G) \pm t^{N-p}_{1-\alpha/2}\,\hat{\sigma}_G\, q_N(x_{\square},\hat{\beta}_G,\theta)\ \}\ ,$$

where $t^{N-p}_{1-\alpha/2}$ is the $(1-\alpha/2)$ percentage point of a t-distribution with N-p degrees of freedom. For large sample sizes, this interval becomes

$$(2.3) \qquad I(x_{\square}) \cong \{\text{all y in the interval } f(x_{\square},\hat{\beta}_G) \pm t^{N-p}_{1-\alpha/2}\,\hat{\sigma}_G\, g(f(x_{\square},\hat{\beta}_G),x_{new},\theta)\}\ .$$

The prediction interval (2.2) is only an approximate $(1-\alpha)100\%$ confidence interval because the function q_N is not known but rather must be estimated.

The effect of ignoring the heterogeneity can be seen through examination of (2.3). If $\hat{\sigma}_L^{\,2}$ is the unweighted mean squared error, then for large samples we have the approximation

$$\hat{\sigma}^2_L \cong \sigma^2\, g^2_{mean} = \sigma^2\, N^{-1}\, \Sigma^N_{i=1}\, g^2(x_i,\beta,\theta)\ .$$

Thus the unweighted prediction interval for large sample sizes is approximately

$$(2.4) \qquad I_L(x_{\square}) \cong \{\text{ all y in the interval } f(x_{\square},\hat{\beta}_L) \pm t^{N-p}_{1-\alpha/2}\,\sigma\, g_{mean}\ \}\ .$$

Comparing (2.3) and (2.4) we see that where the variability is small, the unweighted prediction interval will be too long and hence pessimistic, and conversely where the variance is large.

Now suppose that we are given the value of the response $y_{\square}$ and wish to estimate and make inference about the unknown $x_{\square}$. The estimate of $x_{\square}$ is that

value which satisfies $f(x_0, \hat{\beta}_G) = y_0$. The most common interval estimate of x_0 is the set of all values x for which y_0 falls in the prediction interval I(x), i.e.

$$\text{Calibration interval for } x_0 = \{ \text{ all x such that } y_0 \in I(x) \text{ where I(x) is given by (2.3) } \}.$$

The effect of not weighting is too long and pessimistic confidence intervals for x_0 where the variance is small and the opposite where the variance is large. As far as we know, little work has been done to determine whether one can shorten the calibration confidence interval by making more direct use of the variance function.

3. : ASYMPTOTICS

Assume throughout that the data are symmetrically distributed about their mean. Let $\hat{\beta}_G$ be any generalized least squares estimate of β based on an estimate of θ, call it $\hat{\theta}$ say. Davidian & Carroll (1986) introduce a class of estimators which depend on the data only through $\hat{\beta}_G$, the design $\{x_i\}$, and either sample variances from replicates at each design point or on transformations of the squared residuals

$$\{y_i - f(x_i, \hat{\beta}_G)\}^2 \quad .$$

This class of estimators includes most methods in the literature, see Judge, et

al (1985). Davidian & Carroll (1986) show that all members of their class of estimators have the asymptotic expansion

$$N^{1/2}(\hat{\theta} - \theta) = W_N + a^T N^{1/2}(\hat{\beta}_G - \beta) + o_p(1) \quad . \tag{3.1}$$

In (3.1), a is a fixed vector and W_N is asymptotically normally distributed. Because the observations have symmetric distribution, W_N is asmptotically uncorrelated with $\hat{\beta}_G$.

Let $\hat{\beta}_G(\theta)$ and $\hat{\beta}_G(\hat{\theta})$ be generalized least squares estimates of β with θ known and unknown respectively, and let $\hat{\sigma}(\theta)$ and $\hat{\sigma}(\hat{\theta})$ be the correspnding estimates of σ. The length of the prediction intervals with θ known and unknown are proportional to $L(\theta)$ and $L(\hat{\theta})$ respectively, where

$$L(\theta) = \hat{\sigma}(\theta)\ q_N(x_\square, \hat{\beta}_G(\theta), \theta) \quad .$$

The random variable

$$\Delta L = N^{1/2}\{\ L(\hat{\theta}) - L(\theta)\ \} / \sigma$$

describes how well one approximates the length one would use if θ were known. Intuitively, we would like ΔL to have smallest possible variability.

THEOREM : Suppose that W_N in (3.1) is asymptotically normally distributed with mean zero and covariance $\mathbf{C} = \mathbf{C}(\hat{\theta})$ depending on the method of estimating θ. Then, under regularity conditions, ΔL is asymptotically normally distributed with variance an increasing function of $\mathbf{C}(\hat{\theta})$.

NOTE : The Theorem remains valid if ΔL is the normalized difference in length between the interval with θ unknown and the interval with completely specified variance function.

PROOF (Sketch) : It is easily seen that in the definition of L we may replace q_N by g. Further, $\Delta L = A_1 + A_2 + A_3$, where

$$A_1 = N^{1/2} g(x_\square, \hat{\beta}(\hat{\theta}), \hat{\theta}) \{\hat{\sigma}(\hat{\theta}) - \hat{\sigma}(\theta)\}/\sigma$$
$$A_2 = N^{1/2} \{\hat{\sigma}(\theta)/\sigma\} \{g(x_\square, \hat{\beta}(\hat{\theta}), \hat{\theta}) - g(x_\square, \hat{\beta}(\hat{\theta}), \theta)\}$$
$$A_3 = N^{1/2} \{\hat{\sigma}(\theta)/\sigma\} \{g(x_\square, \hat{\beta}(\hat{\theta}), \theta) - g(x_\square, \hat{\beta}(\theta), \theta)\} .$$

Now, $A_3 \xrightarrow{p} 0$ since, from Carroll & Ruppert (1987), we have that

$$N^{1/2} \{ \hat{\beta}(\hat{\theta}) - \hat{\beta}(\theta) \} / \sigma \xrightarrow{p} 0,$$

By a Taylor series,

$$A_2 = N^{1/2} g_x(x_\square, \beta, \theta) (\hat{\theta} - \theta) + o_p(1) .$$

Lemma A.3 of Carroll, Davidian & Smith shows that for some constant $b(x_\square)$,

$$A_1 = b(x_\square) N^{1/2} (\hat{\theta} - \theta) + o_p(1) .$$

This shows that for some constant $c(x_\square)$,

$$\Delta L = c(x_\square) N^{1/2} (\hat{\theta} - \theta) + o_p(1)$$

The proof is completed by applying (3.1) .

4. : AN EXAMPLE

In Chapter 2, section 8, Carroll & Ruppert (1987) present the results of an assay for the concentration of an enzyme (esterase). There were 113 observations, of which 5 were deleted. The observed concentration of esterase was recorded and then a binding experiment was undertaken, so that the response is the count of the number of bindings. These data were given to us by another statistician and we are unable to give further detail into the background of the experiment. We do not know whether the recorded concentration of esterase has been accurately measured, although we will assume it has been and that there is little if any measurement error in this predictor. The lack of replicates in the reponse is rather unusual in our experience. Since the response is a count, one might expect Poisson variation, i.e., the power of the mean model holds with $\theta = 0.50$. In our experience with assays, such a model almost always underestimates θ, with values between 0.60 and 0.90 being much more common: see Finney (1976) and Raab (1981a).

The eventual goal of the study is to take observed counts and infer the concentration of esterase, especially for smaller values of the latter. As is typical in these experiments, a calibration or training data set is taken for which the predictor variable esterase is known as is the counted response. Carroll & Ruppert (1982) plot the data, which appears reasonably although not perfectly linear. Actually, the logarithm of the response plotted against the

logarithm of the predictor may appear more linear to some, and less heteroscedastic. As in evident from that plot, the data exhibit rather severe heterogeneity of variance. The Spearman correlation between absolute studentized residuals and predicted values from an unweighted least squares fit is $\rho = 0.39$ with formal computed significance level ≤ 0.0001. Analysis as in Carroll & Ruppert (1982) indicate that the constant coefficient of variation model $\theta = 1.0$ is reasonable, although a value $\theta = 0.9$ might be even better. For $\theta = 1.0$, the Spearman correlation between absolute studentized residuals and predicted values is $\rho = -0.10$, with significance level 0.29. In Figure 1, we plot kernel regression estimates of the Anscombe studentized residuals, i.e., the absolute studentized residuals to the power 2/3, see McCullagh & Nelder (1983). Note that the plots indicate that $\theta = 1.0$ does a far better job of accounting for the heteroscedasticity.

In these data, the effect of not weighting should be to have prediction and calibration confidence intervals which are much too large for small amounts of esterase and conversely for large amounts. In Figure 2 we plot the 95% prediction intervals for the count response for unweighted versus weighted regression: the effect is clear. A similar plot for the calibration intervals shows the same effect: the unweighted analysis is much too conservative for small amounts of esterase, and much too liberal for larger amounts. As Oppenheimer, et al (1983) state, "Rather dramatic differences have been observed depending on whether a valid weighted or invalid unweighted analysis is used".

This example shows that the actual prediction intervals are sensitive to misspecification of the variance function. It should be clear by inference and the previous section that one should make efforts to estimate the structural variance parameter θ as well as possible.

REFERENCES

Box, G. E. P. & Myer, R. D. (1985). Dispersion effects from fractional designs. Technometrics, 28, 19-28.

Carroll, R. J., Davidian, M. & Smith, W. (1986). Variance function estimation and the minimum detectable concentration in assays. Preprint.

Carroll, R. J., and Ruppert, D. (1982). Robust estimation in teroscedastic linear models, Annals of Statistics 10, 429-441.

Carroll, R. J., and Ruppert, D. (1987). Transformations and Weighting in Regression. Chapman & Hall, London.

Davidian, M. and Carroll, R.J. (1986) Variance function estimation in regression. Preprint.

Finney, D. J. (1976). Radioligand assay. Biometrics 32, 721-740.

Garden, J. S., Mitchell, D. G. & Mills, W. N. (1980). Nonconstant variance regression techniques for calibration curve based analysis. Analytical Chemistry 52, 2310-2315.

Judge, G. G., Grifffiths, W. E., Hill, R. C., Lutkepohl, H. & Lee, T. C. (1985). The Theory and Practice of Econometrics, Second Edition. John Wiley and Sons, New York.

McCullagh, P. & Nelder, J. A. (1983). Generalized Linear Models. Chapman & Hall, New York.

Oppenheimer, L., Capizzi, T.P., Weppelman, R.M. and Mehto, H. (1983) Determining the lowest limit of reliable assay measurement. Analytical Chemistry 55, 638-643.

Raab, G. M. (1981). Estimation of a variance function, with application to radioimmunoassay. Applied Statistics 30, 32-40.

Rosenblatt, J. R. & Spiegelman, C. H. (1981). Discussion of the paper by Hunter & Lamboy. Technometrics 23, 329-333.

Schwartz, L. M. (1979). Calibration curves with nonuniform variance. Analytical Chemistry 51, 723-729.

FIGURE 1

The esterase assay data. This is a plot of the kernel regression fits to the Anscombe absolute residuals against the logarithms of the predicted values. The unweighted least squares fit is the solid line, while the generalized least squares fit for the constant coefficient of variation model is the dashed line. Endpoint effects have been ajusted for by selective deletion.

Figure 2

The esterase assay data. These are the 95% prediction intervals for a new response. The dashed line is unweighted least squares, while the solid line is the constant coefficient of variation fit. The lower part of the least squares interval has been truncated at zero where necessary.