

**FISHER LECTURE: The 2002 R. A. Fisher Lecture:
Dedicated to the Memory of Shanti S. Gupta**

Variations Are Not Always Nuisance Parameters

Raymond J. Carroll

Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.
email: carroll@stat.tamu.edu

SUMMARY. In classical problems, e.g., comparing two populations, fitting a regression surface, etc., variability is a nuisance parameter. The term “nuisance parameter” is meant here in both the technical and the practical sense. However, there are many instances where understanding the *structure of variability* is just as central as understanding the mean structure. The purpose of this article is to review a few of these problems. I focus in particular on two issues: (a) the determination of the validity of an assay; and (b) the issue of the power for detecting health effects from nutrient intakes when the latter are measured by food frequency questionnaires. I will also briefly mention the problems of variance structure in generalized linear mixed models, robust parameter design in quality technology, and the signal in microarrays. In these and other problems, treating variance structure as a nuisance instead of a central part of the modeling effort not only leads to inefficient estimation of means, but also to misleading conclusions.

KEY WORDS: Calibration; Heteroscedasticity; Immunoassays; Marginal models; Measurement error; Microarray; Mixed models; Quality technology; Robust parameter design; Variance functions.

1. Introduction

As statisticians, we owe a great debt to R. A. Fisher, whose formulations of statistical modeling and inference continue to form the basis of much of statistical theory and practice. An especially important aspect of his work was the partitioning of variance in designed experiments, highlighting the issue that ignoring important factors can lead to loss of statistical power. This article is based on the notion that understanding the structure of variation remains of major importance.

1.1 Variance Structure

In much of statistics, variability is treated as a nuisance, either formally as an uninteresting parameter or more informally as just that: a nuisance. The classic example of course is comparing two population means, wherein one thinks about variances, if at all, only as one or two lines in computer output depending on whether one wishes to assume equal variance or unequal variance in testing hypotheses on the mean.

My thesis here is that there are many problems, increasingly many of them, where variance structure is not a nuisance, but is fundamental to the proper solution of scientific problems.

What do I mean by the term *variance structure*? My definition is fairly broad, but encompasses two main ideas:

- Systematic dependence of variability on known factors: this includes the classic regression phenomenon of heteroscedasticity.

- Random effects: their inclusion, exclusion, or dependence on covariates.

My points are: (a) variance structure can be important in itself; and (b) variance structure can have a major impact on downstream analyses.

In this article, I will focus on two main examples to illustrate the importance of the two types of variance structure. The first example, involving *systematic dependence of variability on known factors*, comes from the world of drug discovery, and more particularly the validation of an immunoassay. This is seemingly a straightforward nonlinear regression application with heteroscedasticity, but the context is vital. Indeed, it can almost be said that one could pick 20 people off the street to hand-draw a line through the data, and that their hand-drawn lines would be nearly identical. In other words, this is a problem for which all regression fits to the line will be essentially identical. However, assay validation is much more than simply fitting a line: it is showing the ability to measure actual concentrations of substances. I will show that naively ignoring the variance structure leads to incorrect conclusions about the validity of the assay.

The second type of variance structure, *random effects: their inclusion, exclusion, or dependence on covariates*, is a more recent phenomenon, arising from the increasing popularity of linear, nonlinear, and generalized linear mixed models. Here the context I will discuss is from a recent controversy in nutritional epidemiology, where a host of nonstatistically

significant studies has led to questions about the validity of measures of dietary intake. I will show in this context that variance structure arises via the inclusion or exclusion of a random effect, and that the conclusions about the instrument differ depending on whether one does or does not include this random effect.

Finally, I will mention, either in passing or in some detail, a series of other problems for which consideration of variance structure is crucial. These include DNA microarray experiments, robust parameter design in quality technology, generalized linear mixed models, management of fisheries, and the consideration of health effects arising from environmental exposures.

1.2 Parameter Efficiency: What This Article Is Not About

It is worth mentioning what this article is *not* about, namely, the attempt to increase efficiency of parameter estimation and inference for means in regression, repeated measures, and longitudinal studies via consideration of variance structure. This choice is not because the problem is unimportant, but because it is not my main point.

Increasingly, the default analysis in such studies is to ignore the variance structure, e.g., pretend the data are independent and homoscedastic (what is now known as working independence) and then “fix up” the estimated sampling variances of the parameters at the end via sandwich-type methods. Of course, ignoring variance structure in this context causes a loss of efficiency, but no bias. It seems to be worth pointing out, however, that the loss of efficiency in parameter estimation should not be taken so lightly. Many of the biological experiments with which I work are small but expensive, and ignoring the variance structure leads to higher variability of an important nature, with sign changes in parameter estimates common and changes in statistical significance even more common. The idea that one should ignore variance structure in these contexts seems to have taken hold in many quarters, without consideration of the real and important effects that the resulting increases in standard errors of parameter estimates have.

1.3 Outline

The outline of the article is as follows. Section 2 describes the problem of assay validation, with special reference to the work of David Finney. Section 3 discusses variance structure and its implications in dietary instruments, with reference to work of Karl Pearson and William Cochran. Section 4 briefly mentions two other problems: robust parameter design in quality technology (Section 4.1) and microarray technology in genetics and bioinformatics (Section 4.2). Section 5 has concluding remarks.

2. Assay Validation

2.1 Introduction, Working Ranges, and %-Recovery Plots

From Finney (1978): “*Here the weighted analysis has also disclosed evidence of invalidity . . . This needs to be known and ought not to be concealed by imperfect analysis.*”

In immunoassays, the idea is to estimate the concentration of a substance such as a drug from observations on plasma (blood) samples, such observations being either counts as in

a radioimmunoassay, optical densities as in an ELISA assay, etc.

In drug discovery, the hope is to discover a method (an assay) that can *reliably* detect small concentrations of the substance. The key term here is “reliably.” Over the years, the drug discovery process has been formalized into a series of steps so that manufacturers can announce with some degree of precision that their assay can detect concentrations in a particular, useful range. We emphasize that the general aim is to detect small concentrations of the substance.

Briefly, and not completely accurately, the typical idea is to fit a regression model of known concentrations (X) to some sort of response (Y): these are often called the standards. Then, in practice, all one will have is what are called unknowns, i.e., the response Y will be observed, but not the main target, the concentration X . One will estimate the concentration X by an inverse regression, i.e., finding the concentration X such that the fitted value of the regression function equals the observed Y . The question is importance is whether such a procedure can reliably detect and estimate small concentrations of the substance.

In this section, I will illustrate a two-step process for assay validation, recognizing that this is simplistic. More details and background are given in Smith and Sittampalam (1998) and Findlay et al. (2000). I will use specific numbers below, although these can be and often are changed. The two steps in my simple paradigm are as follows:

- Determination of the Working Range. The working range of the data is, in my setup, the range of concentrations for which the coefficient of variation of the fitted concentrations is <0.2 . Typically, validation of the assay will be attempted only on this working range, assuming the range is satisfactory to the scientists running the study.
- Determination of Bias, with the %-recovery Plot. The %-recovery plot forces a 90% confidence interval for the mean percentage bias in an estimated concentration to be $<30\%$.

The first step, determination of the working range, is crucial in narrowing down where one tries to validate the assay (the second step).

I will illustrate the following feature, namely, that not accounting for variance structure can result in an unusually pessimistic, high-concentration working range, and that even if one tries to validate the assay at lower and more useful concentrations, one will not be able to do so. A weighted analysis that accounts for variance structure will often lead to a determination that the assay is valid at much lower concentrations than can be confirmed in an unweighted analysis.

The ideas of this section will be discussed by reference to an experiment described in O’Connell, Belanger, and Haaland (1993). Figure 1 gives a plot of the data, with the response being an optical density and the predictor being concentration. At each concentration, there were three replicates. The figure clearly shows that as the concentration increases, the variability of the response increases.

2.2 The Mean Model and Variance Structure

In many immunoassays, it has been found (Finney, 1976, 1978) that the four-parameter logistic model fits the mean

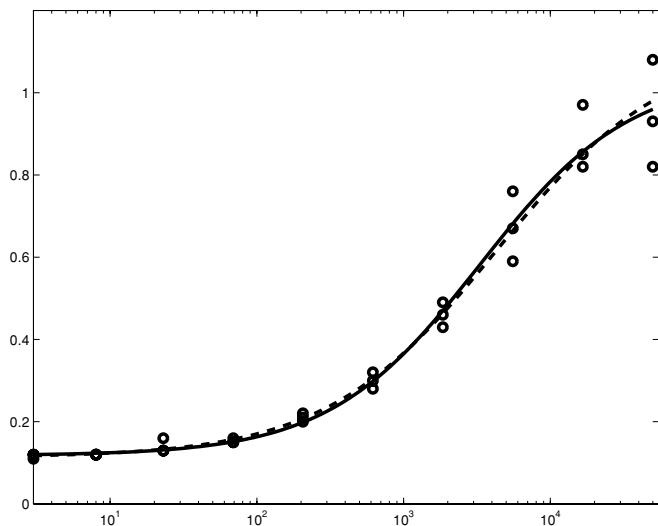


Figure 1. Immunoassay data from O’Connell, Belanger, and Haaland (1993). The x-axis is concentration in the log scale, while the y-axis is intensity. The data are the circles, the unweighted fit to the four-parameter logistic mean model (1) is the solid line, and the weighted fit is the dashed line.

response of the data reasonably well:

$$E(Y | X) = \beta_2 + \frac{\beta_1 - \beta_2}{1 + (X/\beta_3)^{\beta_4}}. \quad (1)$$

This model is sigmoidal in shape, and plateaus as X becomes large.

The parameters of this model are easily interpreted. When the concentration of the substance is $X = 0$, then the mean is β_1 . Referring to Figure 1, one can see visually that $\beta_1 \approx 0.10$. The term β_2 is the mean response at full saturation ($X \approx \infty$). Figure 1 has clearly not reached full saturation, although there is sufficient curvature to estimate β_2 . The term β_3 is the concentration at which the response is midway between that of no concentration and that of full saturation. Finally, β_4 is effectively the slope of the linear part of the curve.

The nonconstant variability seen in Figure 1 is characteristic of immunoassays, and is often modeled by a power of the mean function:

$$\text{var}(Y | X) = \sigma^2 \{E(Y | X)\}^\theta;$$

see Rodbard (1978) and Munson and Rodbard (1978), Carroll and Ruppert (1988) and Davidian, Carroll, and Smith (1988). Generally, the power parameter θ is between Poisson-like variability ($\theta = 1$) and Gamma-like variability ($\theta = 2$).

2.3 Weighted and Unweighted Fits

Figure 1 also displays the unweighted least squares fit and a weighted fit, both using the four-parameter logistic mean function and the latter accounting for variance structure via a power of the mean model. Note how the two lines are almost identical: this is typical of immunoassays. Thus, as the level

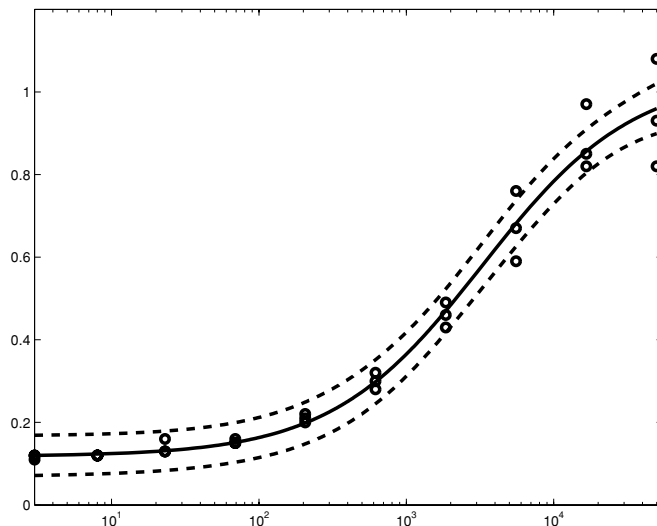


Figure 2. Immunoassay data. The x-axis is concentration in the log scale, while the y-axis is intensity. The data are the circles, the unweighted fit to the four-parameter logistic is the solid line, and the dashed lines are 95% prediction limits.

of simply fitting the mean structure, *it does not matter what method is used to account for the variance structure.*

2.4 Prediction Intervals

An important component of assay validation is prediction intervals for a response. This is reasonably clear. The idea of an assay is to take an observed response and from it make inference about the concentration that led to the response. Since the basic starting point is a single response, the prediction interval of interest is a prediction interval for a single response.

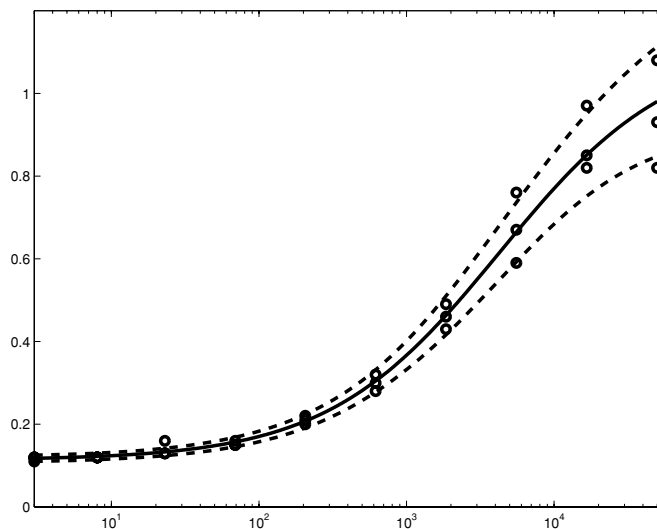


Figure 3. Immunoassay data. The x-axis is concentration in the log scale, while the y-axis is intensity. The data are the circles, the weighted fit to the four-parameter logistic is the solid line, and the dashed lines are 95% prediction limits.

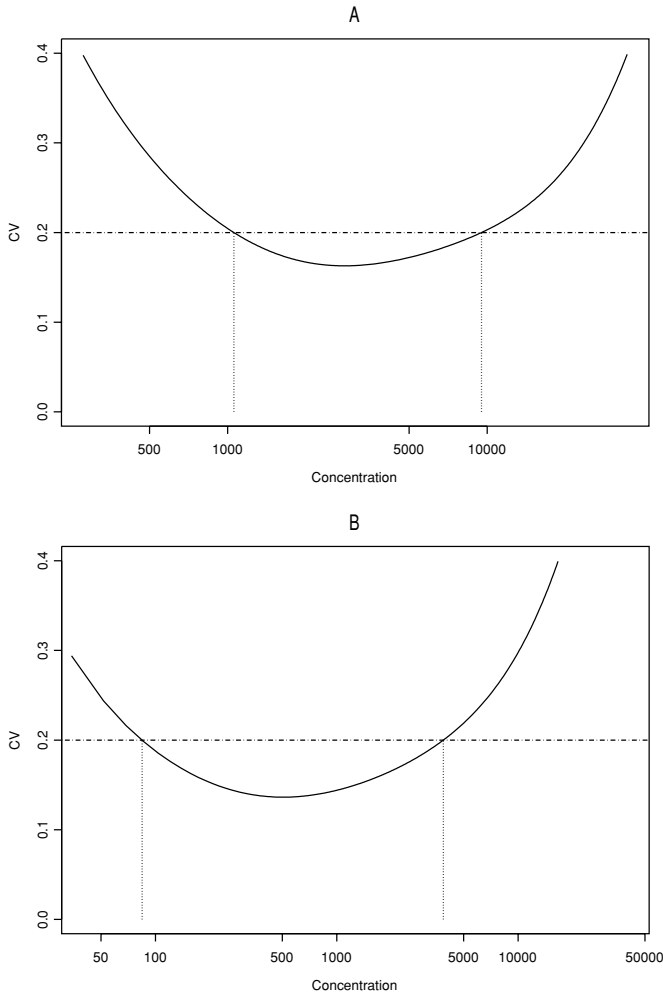


Figure 4. Immunoassay assay. Coefficient of variation (y -axis) plotted against concentration for the unweighted (A) and weighted (B) fits. The scales are not the same: see Figure 5 for a more direct comparison.

It is important here to note that the goal is to form a prediction interval for the actual response, and not for the regression function. Because the goal is an interval for the actual response, it is not the fit itself that matters, but the variance assumption underlying it. This is illustrated in two figures. Figure 2 gives the prediction interval for a response using the unweighted fit, hence *assuming* a constant variance. Here it is obvious that the prediction interval from the constant variance assumption is far too conservative for small concentrations, and too liberal for large concentrations. In contrast, Figure 3 gives the prediction interval for the weighted fit, hence assuming nonconstant variance. Here the intervals more closely conform to the data, because the variance assumption is more appropriate.

All this matters when one recalls that what will be done is to take an observed response Y and use inverse regression to estimate the unobserved concentration X . One can see immediately that confidence intervals for the unobserved concentration will be very different depending on whether one assumes constant or nonconstant variability.

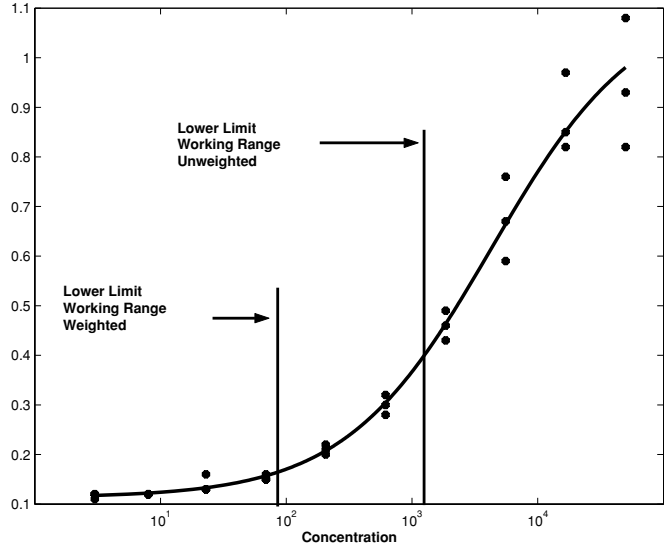


Figure 5. Immunoassay assay. Data fit by weighted method and lower limits of the working range for the unweighted and weighted fits.

2.5 *Variance Structure Matters for Constructing Working Ranges*

As described previously, the concentrations are estimated by inverse regression. Given an observed response Y , the concentration X is estimated so that the value of the regression function (1) equals Y .

As described previously, the *working range* is the range of those concentrations for which the coefficient of variation for the estimated concentrations is below a fixed bound, say 0.2. Figure 4 plots the coefficient of variation versus the concentration based on the unweighted and weighted fits. While the shapes of these plots are fairly similar, the careful reader will note that the ranges in which the two plots have coefficient of variation <0.2 are vastly different; this is illustrated in Figure 5. In fact, for the unweighted fit the working range is from 1057 to 9505, while the weighted fit suggests a working range from 84 to 3866.

The upshot of this is important. If one uses an unweighted fit, and hence ignores the variance structure, then one will only try to validate the assay on a range of concentrations greater than 1000. As will be shown later later, if one were to try to validate the assay for smaller concentrations using an unweighted fit, one will not be able to do so.

In contrast, the weighted fit suggests a working range that includes much smaller concentrations, even down to 84 units.

2.6 *Variance Structure Matters for %-Recovery Plots*

Having decided upon a working range, in my simple paradigm, it would then be typical to perform a validation experiment. While the practice varies, one common approach is to look at %-recovery plots, where as described above what are measured are the actual concentration X , the predicted concentration \hat{X} , and the %-recovery $R = \hat{X}/X$, using a series of replicates. One method of validating an assay is to insist that a 90% confidence interval for the mean formed from the replicated R 's at each concentration be entirely contained in

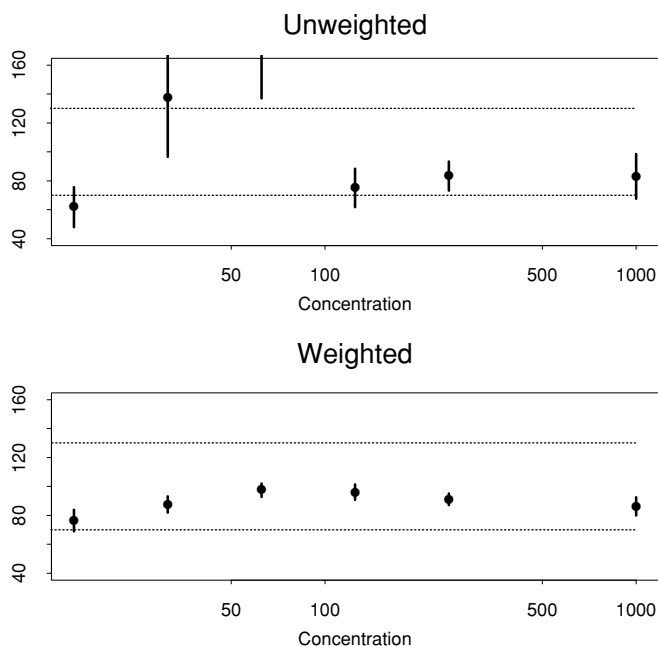


Figure 6. IL-10 experiment, %-recovery plots with 90% confidence limits. % recovery (y-axis) plotted against concentration for the unweighted and weighted fits. The assay is invalid once the confidence interval bars fall outside the dashed lines.

a prescribed range, say 70%–130%. The range at which the assay is validated is computed back from high concentrations and stops at the first time the range constraint is violated.

Figure 6 shows what happens in an Interleukin-10 (IL-10, a biomarker for inflammation) experiment, for concentrations in the working range formed by a weighted fit. Because at a concentration of 1000, the unweighted fit’s interval is not in the prescribed range, users of an unweighted fit would conclude that the assay is invalid for concentrations at 1000 and below. On the other hand, the weighted fit suggests that the assay is valid for all concentrations except the smallest, and that just barely.

What is happening here? The effect is subtle, but can be seen in Figures 2–3. The unweighted fit treats low concentrations as equally variable with high concentrations, and hence does not tightly fit the data at the low concentrations. The weighted fit insists on fitting the low concentrations very well. When attention turns to %-recovery plots, what happens is that the predictions at low concentrations are not particularly good for the unweighted fit, and thus the range constraints are violated.

2.7 Conclusions

Assay validation cannot be done accurately without consideration of variance structure. Unweighted fits ignore the variance structure, and often result in working ranges and %-recovery plots that suggest the assay is valid only for high concentrations. Because in many (but not all) assays one wants to measure low concentrations (see the beginning of this section), this means that in practice an unweighted anal-

ysis will suggest that the assay is invalid. This will lead either to misguided abandonment of the project, or expensive, possibly unnecessary modifications of the assay itself.

3. Dietary Instruments

3.1 The Problem

Much of the recent literature on the relationship between diet and cancer has been based on analytic epidemiologic studies using food frequency questionnaires (FFQs). A number of large prospective studies of this kind have failed to find a consistent relationship between dietary components (such as fat, fiber, fruits, and vegetables) and cancers of the breast, colon, or rectum (Hunter, Spiegelman, and Adami, 1996; Fuchs, Giovannucci, and Colditz, 1999; Michels, Giovannucci, and Joshipura, 2000). This may be explained by a true lack of diet-cancer associations or, alternatively, by serious methodological limitations of the studies themselves, especially due to FFQ measurement error.

Over the years, investigators have recognized that the reported values from FFQs are subject to substantial error, both systematic and random, that can profoundly affect the design, analysis, and interpretation of nutritional epidemiologic studies (Beaton, Milner, and Corey, 1979; Freudenheim and Marshall, 1988; Freedman, Schatzkin, and Wax, 1990). Dietary measurement error often attenuates (biases toward one) the estimates of disease relative risks (RRs) and reduces statistical power to detect their significance. An important relation between diet and disease, therefore, may be obscured.

These considerations have prompted many investigations into the properties of dietary intake instruments, and especially the FFQ. A review of such work, with many source references, is described by Kipnis et al. (2001) and Kipnis, Subar et al. (2003, to appear).

With this as the starting point, the issue can be phrased as follows. There is a primary instrument, an FFQ, that is a measure of diet and denoted by Q . This instrument is a proxy for true, long-term dietary intake, denoted by T : truth can never be measured for all individuals in a large sample. A study has now been done, possibly using logistic regression, of disease outcome on the questionnaire. The study was null, i.e., diet was not statistically significantly related to disease. The question: *how much power did the study have to detect an effect of practical importance, e.g., a relative risk of 1.8?*

This *post hoc* calculation obviously really matters in practice, because if there is indeed 80% power to detect a relative risk of 1.8, yet simultaneously there are a series of null-study results, then the case for a lack of a relationship between disease and a specific nutrient intake is strengthened mightily.

3.2 Reference Instruments and Attenuation

The question then is how to perform the *post hoc* power calculation. One way to phrase this question is to ask: what sample size would have been needed to achieve a given power (say 80%) for a given relative risk (say 1.8)? Given that the study has been done, and the variability of the FFQ Q has been determined, it turns out that this sample size is essentially determined in a *post hoc* calculation by the *attenuation* λ :

$$\begin{aligned} \lambda &= \text{attenuation} \\ &= \text{slope of regression of true intake } T \text{ on FFQ intake } Q. \end{aligned} \quad (2)$$

The sample size required for fixed power is

$$\text{sample size required} = \frac{\text{constant determined by the FFQ}}{\lambda^2}. \quad (3)$$

Although (3) is the key issue in this article, it is also true that the relationship between observed relative risk using the FFQ and the true relative risk is determined by λ :

$$\text{observed relative risk} \approx (\text{true relative risk})^\lambda. \quad (4)$$

I now return to (3). Suppose one has done an experiment and estimated that the attenuation $\lambda_{obs} = 0.50$, and that with this attenuation your sample size is large enough to achieve 80% power to detect a relative risk of 1.8 between, say, two fixed values of intake. Suppose, however, that in fact the real attenuation is $\lambda_{true} = 0.30$. Since sample size is determined by the *square* of the attenuation, this means that the sample size actually needed to achieve the given power needs to be increased by the factor

$$\text{sample size inflation} = \frac{\lambda_{obs}^2}{\lambda_{true}^2} = 2.56. \quad (5)$$

Thus, quite a lot is riding on the estimation of the attenuation. How is this done?

I should point out that (3)–(4) are primary factors in the statistical power available from an instrument. My discussion here is not one about how to test for disease-diet relationships when diet is measured with error, even though I have written many papers on that topic and, in some cases, valid tests are achieved by ignoring measurement error. Instead, the issue here is how much power we actually have to detect a disease-diet relationship.

3.3 Variance Structure: Pearson and Cochran

In most of the literature, the attenuation is estimated via a so-called calibration/validation study. In this setup, another, presumably better, dietary intake instrument is used, e.g., multiple 24-hour recalls, 7-day diaries, or 4-day weighed food records. These are typically referred to as reference instruments; see Kipnis et al. (2001) and Kipnis, Subar et al. (2003).

Call the reference instruments F . Let i denote the individual, and j denote the replicate of the instrument. Then the common assumption is that F_{ij} is the truth T_i , except for random measurement error, i.e.,

$$F_{ij} = T_i + \epsilon_{ij}^F. \quad (6)$$

Under the assumption that the presumptive measurement errors ϵ_{ij}^F have mean 0 and are unrelated to true diet T_i , it is easily seen that the attenuation λ defined in (2) is the slope of the regression of F on Q . Almost all calibration/validation studies estimate the attenuation in this fashion, and the conclusion typically is that there is plenty of statistical power for detecting interesting nutrient effects such as relative risks of 1.8.

It is at this point that Pearson (1902) and Cochran (1968) weigh in. Pearson was interested in issues of self-report, just as I am, although of course in a different context. He gave different individuals a series of lines drawn on paper, had them bisect the lines by eye, and then measured the errors (positive or negative) that were made in the bisection. As pointed out by Cochran (1968), what Pearson found was that the simple measurement error model (6) did not hold.

Note that Pearson knew the truth T_i and he was able to measure the errors ϵ_{ij}^F exactly. He noted that even if he averaged these errors for an individual, the mean errors were clearly different for different individuals. The simplest way to explain this phenomenon is to change the variance structure by adding a *random effect* that is the systematic error made by the individual. Following this reasoning, Cochran (1968) suggested that a more reasonable error model is

$$F_{ij} = \beta_0^F + \beta_1^F T_i + r_i^F + \epsilon_{ij}^F. \quad (7)$$

Equation (7) has three types of errors:

- The terms ϵ_{ij}^F represent the fact that an individual bisecting the same line multiple times will make different errors. In our context, these random errors can be thought of as combining a person's random fluctuations in their diet coupled with random mistakes in recording the diet.
- The term r_i^F represents what I call *person-specific biases*. In other words, two people who eat exactly the same diet (same T_i) will not report the same diet on average over many replications of the experiment. Informally, people with the same diet may be more or less prone to report their intake of ice cream.
- The term $\beta_0^F + \beta_1^F T_i$ represents a systematic bias at the population level. Thus, informally, the population of people who eat a great deal of ice cream may for reasons of societal pressure tend to underreport their ice cream consumption.

Clearly, if such considerations apply to reference instruments F_{ij} , they should also apply to the FFQ, since it too is a self-report instrument, so that a reasonable model is

$$Q_{ij} = \beta_0^Q + \beta_1^Q T_i + r_i^Q + \epsilon_{ij}^Q. \quad (8)$$

Kipnis et al. (2001) and Kipnis, Subar et al. (2003) recognized the importance of models (7)–(8), and they also recognized that it is likely that the person-specific biases r_i^Q and r_i^F would be correlated. In order to identify these models, and hence estimate the attenuation λ , they recognized the need for a true reference instrument, namely, a biomarker not subject to reporting biases. In other words, they described the need for physical measurements M_{ij} that satisfy the classic assumption:

$$M_{ij} = T_i + \epsilon_{ij}^M. \quad (9)$$

Two such biomarkers are known to exist. Under fairly reasonable assumptions, Protein intake can be measured by urinary nitrogen, and energy (caloric) intake can be measured by doubly-labeled water. Details are given in Kipnis, Subar et al. (2003), including the identifiability of the model parameters.

With biomarkers such as this, Kipnis et al. (2001) suggested that one fit models (7)–(9) and compare the

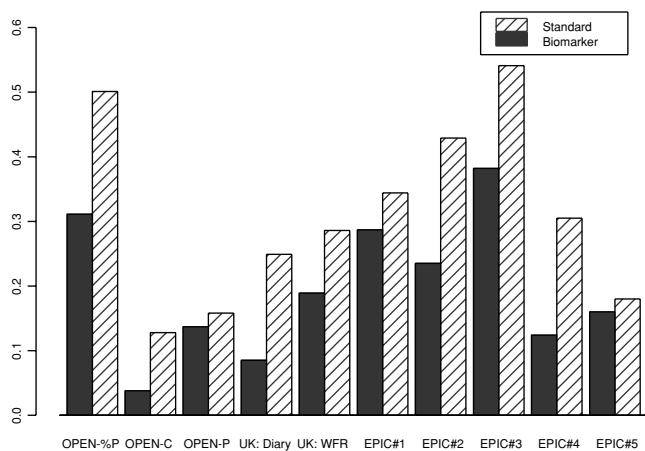


Figure 7. The estimated attenuations from a series of biomarker studies with different reference instruments. Solid bars: attenuations with naive variance structure that ignores person-specific bias random effects. Crosshatched bars: attenuations when person-specific bias random effects are included. The studies are *OPEN-%P*: OPEN study, % calories from protein. *OPEN-C*: OPEN study, energy (calorie) intake. *OPEN-P*: OPEN study, protein intake. *UK Diary*: U.K. study using a diary. *UK WFR*: U.K. study using a weighed food record. *EPIC#1-EPIC#5*: five European studies using 24-hour recalls.

attenuation that researchers thought they had, namely, pretending that (6) held, with the attenuations that they actually have.

3.4 Results

I now discuss the results of these considerations over a series of studies (Kipnis, Midthune, et al., 2003). There are five studies from European cohorts that measured protein intake, labeled EPIC #1–#5, that used a 24-hour recall as a reference instrument. Two studies from the U.K. also measured protein intake, with one using a diary as the reference instrument and the other using a weighed food record as the reference instrument. Finally, the OPEN study measured both protein and energy, with a 24-hour recall as the reference instrument. Along with protein and energy, they also considered protein density, the % of calories coming from protein.

The estimated attenuations are given in Figure 7. Note that in every case, the attenuations estimated from the nominal reference instrument are larger than the attenuations estimated using the biomarkers. In other words, the common calibration/validation studies make measurement error appear to be less of an issue than it really is.

How much less of an issue is given in Figure 8. Recall that what matters in a *post hoc* calculation of the sample size required to achieve a fixed power is the square of the attenuation; see (3) and (5). Figure 8 shows the following. If the reference instrument suggests that the current sample size is large enough to achieve 80% power for detecting a meaningful relative risk, by what factor does the sample size have to be increased to actually achieve that power?

For example, consider energy intake in the OPEN study. Figure 8 shows that using 24-hour recalls as the reference in-

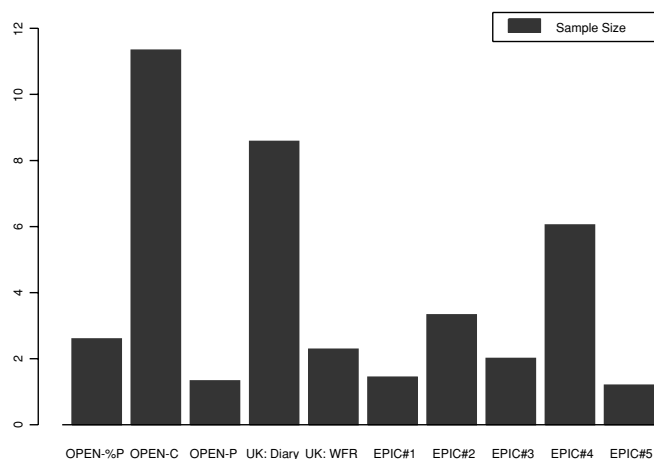


Figure 8. The estimated variance inflation factors from a series of biomarker studies with different reference instruments. The *variance inflation factor* is defined as the proportional increase in sample size required for a fixed power in a post hoc calculation when the reference instrument used is not a biomarker, but is instead an instrument subject to person-specific biases. The studies are *OPEN-%P*: OPEN study, % calories from protein. *OPEN-C*: OPEN study, energy (calorie) intake. *OPEN-P*: OPEN study, protein intake. *UK Diary*: U.K. study using a diary. *UK WFR*: U.K. study using a weighed food record. *EPIC#1-EPIC#5*: five European studies using 24-hour recalls.

strument underestimates the required sample size by a factor of 11(!). The OPEN study for protein density shows that the factor is 2.56, still an impressive figure.

3.5 Conclusions

This example has illustrated that consideration of variance structure through the inclusion or exclusion of random effects can have important implications for analysis. By including person-specific biases, and allowing them to be correlated, I have shown that measurement error is often much more severe in nutritional epidemiology than standard reference instrument calibration/validation studies would suggest.

4. Other Topics

4.1 Robust Parameter Design

According to Wu and Hamada (2000, p. 436), robust parameter design “*aims at reducing the performance variation of a system . . . by choosing the setting of control factors to make it less sensitive to noise variation.*” Somewhat simplistically, the basic idea is to distinguish between the factors that are under one’s immediate control (control factors), and the factors that are less easily controlled without great effort (noise factors), and to find settings of the control factors so that the output is not much affected by changes in the noise factors.

For example, one might desire a specific mean response, but in a quality technology application one would want the variability of the response to be as small as possible. Thus, one would want to consider settings of the control factors that satisfy the constraint on the mean, but which have small variability under ideal conditions, and have variability that is not

much affected by changes in the noise factors. In other words, the operation is to set the target, then minimize variance.

Modeling variability is clearly an intrinsic part of the method, and not simply a nuisance. This can be done in at least three ways, across experiments where the noise factors are allowed to vary systematically: these issues are discussed in detail by Wu and Hamada (2000).

- The so-called Taguchi method maximizes the signal-to-noise ratio across the control factor levels. Variability is often modeled here indirectly via replication.
- One can model location and dispersion across the control factors separately as two separate functions, finding the settings of the control factors that reach a target while minimizing variance.
- Assuming constant variance given control and noise factors, one can model location across the control and noise factors, and then minimize the variance transmitted through the noise factors.

Clearly, in robust parameter design, variance structure matters.

4.2 Microarray Technology

DNA microarray technologies, such as cDNA array and oligonucleotide array, provide a means of measuring the expression of thousands of genes simultaneously. These technologies have attracted much excitement in the biological and statistical literature, and promise to revolutionize biological research and further our understanding of biological processes. A recent introduction to the area is given by Nguyen et al. (2002), along with numerous references.

There is a large and growing statistical literature on this topic, one that can easily fill many books. The measurements of gene expression involve multiple steps (sample preparation, imaging, etc.) that affect the quality of the results. The entire process could clearly benefit from robust parameter design, but surprisingly there has been little work along these lines, with the notable exceptions of Kerr, Martin, and Churchill (2000), and Kerr and Churchill (2001a, 2001b). More specifically, the question becomes: how much of the observed variability in gene expression measurements is noise versus signal, and how does that impact on statistical analysis?

Such questions may not matter much in situations such as the comparisons of two groups of patients in which one is affected with cancer, since these experiments involve such massive changes in gene expression that even a great deal of variability in the measurements is overwhelmed by signal. On the other hand, questions of the structure of variability begin to matter greatly when the groups are not so dissimilar.

I have been involved in an experiment in which 30 rats were first fed a standard diet, and then the diets were enhanced with either corn oil, fish oil, or olive oil. Ten rats were allocated to each dietary group. Gene expression was measured as in Ramakrishnan, Dorris, and Lublinsky (2002). The effects here would be expected to be subtle, and in some sense they are. In our analysis, each gene (one-at-a-time) was analyzed to test for statistically significant differences in mean gene expression across the diet groups. Using the usual false-discovery-rate criterion (Benjamini and Hochberg, 1995) with

a 5% false discovery rate, only 93 genes out of 8038 showed differential expression between the diet groups.

A second experiment was also undertaken, again with 30 rats and 10 rats per diet group. In this case, the rats were exposed to a potent carcinogen, for which it is known that many phenotypes show a clear protective effect of fish oil, as compared to corn oil. However, under this regime, and because of the phenotypic data very surprisingly, *no* genes showed differential expression across diets.

One can argue that the lack of statistically significant gene expression is the outcome of the one-at-a-time analysis, and that the phenotypic-level protective effects of a fish oil diet are conferred by groups of genes acting in tandem. On the other hand, one can also argue that microarray technology is so fraught with random (and systematic) errors that it can only find the strongest of signals.

To understand this to even a reasonable extent, we have been in the lucky situation of having true replicates, i.e., 1/2 the rats had a second microarray experiment performed. This gave us an (unbalanced) repeated measures design, from which we could extract the between-animal variability, the within-animal variability, and the intraclass correlation. We found that overall, the intraclass correlation estimates had a median value of 1/3, indicating that on average, across animals, 2/3 of the observed variability in the data is due to noise. The observed histograms of estimated intraclass correlations were remarkably similar to those arising from an experiment in which genes acted at random with common intraclass correlation equal to 1/3; see Figure 9.

If one believes then that 2/3 of the variability at the gene level across animals is noise, then decreasing the noise by 50% using robust parameter design could have a substantial impact. For example, if there are two groups, each with 10 subjects, and the intraclass correlation is 1/3, and if the

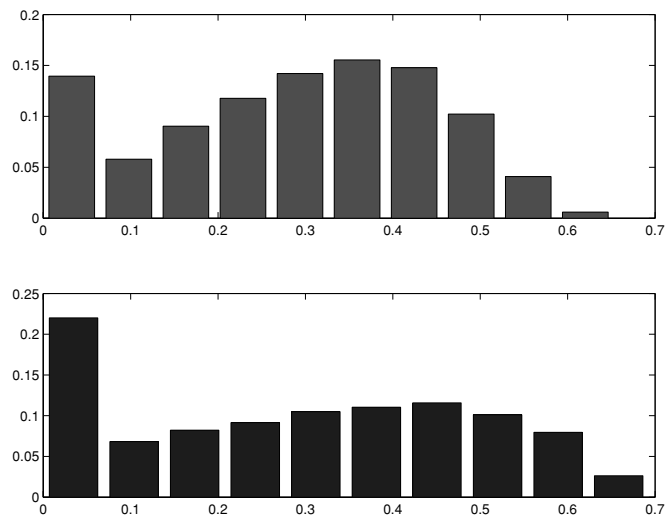


Figure 9. Histograms of intraclass correlation estimates in the nutrition microarray experiment for animals not given a carcinogen. There were 30 animals and 15 of them had replicated microarrays. *Top:* simulated data with common intraclass correlation $\rho = 0.35$. *Bottom:* actual data.

difference of the means between the two groups is 1 unit of the within and between variability, then there is 61% power for detecting a difference. Decreasing the measurement errors in the microarray by 50%, thus making the intraclass correlation equal to $1/2$, would increase the power to 78%.

5. Discussion

The list of problems for which variance structure is important is much longer than that given here. Topics with which I have been concerned that are within the theme of this article include the management of the Atlantic Menhaden Fishery (see Ruppert et al., 1984, 1985; Reish et al., 1985; Ruppert and Carroll, 1985) and understanding the effect of radiation on the development of thyroid cancer (Schafer, Stefanski, and Carroll, 1999; Schafer et al., 2001; Mallick, Hoffman, and Carroll, 2002), both problems for which naive consideration of variance structure can lead conclusions that are scientifically invalid. Heagerty (1999) and Heagerty and Kurland (2001) also discuss subtleties of inference in generalized linear mixed models when random effects have variance that depends on cluster-level covariates, wherein naive consideration of variance structure can lead to incorrect inferences. I have illustrated in detail the importance of consideration of variance structure in assay validation and in understanding nutrient intake instrument quality. I have also briefly described the importance of variance structure in robust parameter design and microarrays.

In this article, I have described the importance of *variance structure*, by which I mean two main ideas:

- Systematic dependence of variability on known factors: this includes the classic regression phenomenon of heteroscedasticity.
- Random effects: their inclusion, exclusion, or dependence on covariates.

My points are: (a) variance structure can be important in itself; and (b) variance structure can have a major impact on downstream analyses.

ACKNOWLEDGEMENTS

This article is dedicated to the memory of my major professor, Shanti S. Gupta, who headed the Department of Statistics at Purdue University for 20 years. His kind manner and constant encouragement made graduate school a great pleasure for me, and all his students.

Much of my research (3 books and 38+ articles) has been written with David Ruppert, and I gratefully acknowledge his being willing to work with me over the past 25 years. I also thank my many coauthors and friends for their inspiration, a partial list of which includes Mitchell Gail, Marie Davidian, Naisyin Wang, Xihong Lin, Peter Hall, Alan Welsh, Cliff Spiegelman, Matt Wand, Wendell Smith, Victor Kipnis, Lawrence Freedman, and Douglas Midthune. Devan Devanarayan was especially helpful in working with me on Section 2.

The site <http://stat.tamu.edu/~carroll/recenttalks.html> contains the talk upon which this article is based. It includes a map of my home state of Texas and photographs of many of the people mentioned in the paper. This research was supported by a grant from the National Cancer Institute

(CA57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106).

RÉSUMÉ

Dans des problèmes classiques comme la comparaison de deux populations, l'estimation d'une surface de régression ou d'autres divertissements de même nature, la variabilité constitue un paramètre de nuisance. La structure de la variabilité, dans bien des cas, s'avère pourtant aussi importante que la structure de la moyenne. Me proposant, dans cet article, de rendre compte d'une petite partie d'entre eux, je me concentrerai plus particulièrement sur deux sujets : a) la validation d'un dosage par immunoessai, b) la probabilité de détecter les effets, sur la santé, d'apports de nutriments mesurés par des questionnaires sur l'alimentation. Par ailleurs, j'évoquerai brièvement les problèmes de structure de la variance dans les modèles mixtes (modèles linéaires généralisés), les plans expérimentaux robustes (élaborés en contrôle de la qualité) et l'identification du signal (microréseaux). Dans ces problèmes comme dans d'autres, le fait de considérer la structure de la variance comme une astreinte au lieu de lui accorder une place centrale dans l'élaboration de la modélisation n'a pas pour seule conséquence d'entraîner une estimation peu efficace des moyennes; procéder de la sorte, en effet, peut aussi conduire à des conclusions erronées.

REFERENCES

- Beaton, G. H., Milner, J., and Corey, P. (1979). Sources of variance in 24-hour dietary recall data: Implications for nutrition study design and interpretation. *American Journal of Clinical Nutrition* **32**, 2546–2559.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.
- Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. New York: Chapman and Hall CRC Press.
- Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics* **10**, 637–666.
- Davidian, M., Carroll, R. J., and Smith, W. (1988). Variance functions and the minimum detectable concentration in assays. *Biometrika* **75**, 549–556.
- Findlay, J. W. A., Smith, W. C., Lee, J. W., Nordblom, G. D., Das, I., DeSilva, B. S., Khan, M. N., and Bowsher, R. R. (2000). Validation of immunoassays for bioanalysis: A pharmaceutical industry perspective. *Journal of Pharmaceutical and Biomedical Analysis* **21**, 1249–1273.
- Finney, D. J. (1976). Radioligand assay. *Biometrics* **32**, 721–740.
- Finney, D. J. (1978). *Statistical Method in Biological Assay*, 3rd edition. High Wycombe U.K.: Charles Griffin.
- Freedman, L. S., Schatzkin, A., and Wax, Y. (1990). The impact of dietary measurement error on planning a sample size required in a cohort study. *American Journal of Epidemiology* **132**, 1185–1195.
- Freudenheim, J. L. and Marshall, J. R. (1988). The problem of profound mismeasurement and the power of epidemiologic studies of diet and cancer. *Nutrition and Cancer* **11**, 243–250.

- Fuchs, C. S., Giovannucci, E. L., and Colditz, G. A. (1999). Dietary fiber and the risk of colorectal cancer and adenoma in women. *New England Journal of Medicine* **340**, 169–176.
- Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics* **55**, 688–698.
- Heagerty, P. J. and Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalized linear mixed models. *Biometrika* **88**, 973–985.
- Hunter, D. J., Spiegelman, D., and Adami, H.-O. (1996). Cohort studies of fat intake and the risk of breast cancer—A pooled analysis. *New England Journal of Medicine* **334**, 356–361.
- Kerr, M. K. and Churchill, G. A. (2001a). Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–202.
- Kerr, M. K. and Churchill, G. A. (2001b). Statistical design and analysis of gene expression microarrays. *Genetical Research* **77**, 123–128.
- Kerr, M. K., Martin, M., and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**, 819–837.
- Kipnis, V., Midthune, D., Freedman, L. S., Bingham, S., Schatzkin, A., Subar, A., and Carroll, R. J. (2001). Empirical evidence of correlated biases in dietary assessment instruments and its implications. *American Journal of Epidemiology* **153**, 394–403.
- Kipnis, V., Midthune, D., Freedman, L. S., Bingham, S., Day, N. E., Riboli, E., and Carroll, R. J. (2003). Bias in dietary-report instruments and its implications for nutritional epidemiology. *Public Health Nutrition* **5**, 915–923.
- Kipnis, V., Subar, A. F., Midthune, D., Freedman, L. S., Ballard-Barbash, R., Troiano, R., Bingham, S., Schoeller, D. A., Schatzkin, A., and Carroll, R. J. (2003). The structure of dietary measurement error: Results of the OPEN biomarker study. *American Journal of Epidemiology*, to appear.
- Mallick, B., Hoffman, F. O., and Carroll, R. J. (2002). Semi-parametric regression modeling with mixtures of Berkson and classical error, with application to fallout from the Nevada Test Site. *Biometrics* **58**, 13–20.
- Michels, K. B., Giovannucci, E., and Joshipura, K. J. (2000). Prospective study of fruit and vegetable consumption and incidence of colon and rectal cancers. *Journal of the National Cancer Institute* **92**, 1740–1752.
- Munson, P. J. and Rodbard, D. (1978). Computerized analysis of quality control for radioimmunoassays. *Proceedings of Computer Science and Statistics: 10th Annual Symposium on the Interface*, 288–291.
- Nguyen, D., Arpat, A. B., Wang, N., and Carroll, R. J. (2002). DNA microarray experiments: Biological and technological issues. *Biometrics* **58**, 701–717.
- O’Connell, M. S., Belanger, B. A., and Haaland, P. D. (1993). Calibration and assay development using the four-parameter logistic model. *Chemometrics and Intelligent Laboratory Systems* **20**, 97–114.
- Pearson, K. (1902). On the mathematical theory of errors of judgment. *Philosophical Transactions of the Royal Society of London A* **198**, 235–299.
- Ramakrishnan, R., Dorris, D., and Lublinsky, A. (2002). An assessment of Motorola CodeLink microarray performance for gene expression profiling applications. *Nucleic Acids Research* **30**, e30.
- Reish, R. L., Deriso, R. B., Ruppert, D., and Carroll, R. J. (1985). An investigation of the population dynamics of Atlantic menhaden (*Brevoortia tyrannus*). *Canadian Journal of Fisheries and Aquatic Sciences* **42**, 147–157.
- Rodbard, D. (1978). Statistical estimation of the minimum detectable concentration (“sensitivity”) for radioligand assays. *Analytical Biochemistry* **90**, 1–12.
- Ruppert, D. and Carroll, R. J. (1985). Data transformations in regression analysis with applications to stock recruitment relationships. In *Resource Management*, M. Mangel (ed). Lecture Notes in Biomathematics 61. New York: Springer-Verlag.
- Ruppert, D., Reish, R. L., Deriso, R. B., and Carroll, R. J. (1984). Monte-Carlo optimization by stochastic approximation, with application to harvesting of Atlantic menhaden. *Biometrics* **40**, 353–546.
- Ruppert, D., Reish, R. L., Deriso, R. B., and Carroll, R. J. (1985). A stochastic model for managing the Atlantic menhaden fishery and assessing managerial risks. *Canadian Journal of Fisheries and Aquatic Sciences* **42**, 1371–1379.
- Schafer, D. W., Stefanski, L. A., and Carroll, R. J. (1999). Consideration of measurement errors in the international radiation study of cervical cancer. In *Uncertainties in Radiation Dosimetry and Their Impact on Dose Response Analysis*, E. Ron and F. O. Hoffman (eds). Bethesda, Maryland: National Cancer Institute Press.
- Schafer, D. W., Lubin, J. H., Ron, E., Stovall, M., and Carroll, R. J. (2001). Thyroid cancer following scalp irradiation: A reanalysis accounting for uncertainty in dosimetry. *Biometrics* **57**, 689–697.
- Smith, W. C. and Sittampalam, G. S. (1998). Conceptual and statistical issues in the validation of analytic dilution assays for pharmaceutical applications. *Journal of Biopharmaceutical Statistics* **8**, 509–532.
- Wu, C. F. J. and Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. New York: Wiley.

Received January 2003. Revised January 2003.

Accepted January 2003.