

The University of Texas MD Anderson Cancer Center

From the SelectedWorks of Jeffrey S. Morris

September 2001

Parametric and Nonparametric Methods for Understanding the Relationship Between Carcinogen-Induced DNA Adduct Levels in Distal and Proximal Regions of the Colon.

Contact
Author

Start Your Own
SelectedWorks

Notify Me
of New Work



Available at: http://works.bepress.com/jeffrey_s_morris/31

Parametric and Nonparametric Methods for Understanding the Relationship Between Carcinogen-Induced DNA Adduct Levels in Distal and Proximal Regions of the Colon

Jeffrey S. MORRIS, Naisyin WANG, Joanne R. LUPTON, Robert S. CHAPKIN, Nancy D. TURNER, Mee Young HONG, and Raymond J. CARROLL

An important problem in studying the etiology of colon cancer is understanding the relationship between DNA adduct levels (broadly, DNA damage) in cells within colonic crypts in distal and proximal parts of the colon, following treatment with a carcinogen and different types of diet. In particular, it is important to understand whether rats who have elevated adduct levels in particular positions in distal region crypts also have elevated levels in the same positions of the crypts in proximal regions, and whether this relationship depends on diet. We cast this problem as estimating the correlation function of two responses as a function of a covariate for studies where both responses are measured on the same experimental units but not the same subsampling units. Parametric and nonparametric methods are developed and applied to a dataset from an ongoing study, leading to potentially important and surprising biological results. Theoretical calculations suggest that the nonparametric method, based on nonparametric regression, should in fact have statistical properties nearly the same as if the functions nonparametrically estimated were known. The methodology used in this article can be applied to other settings when the goal of the study is to model the correlation of two continuous repeated measurement responses as a function of a covariate, whereas the two responses of interest can be measured on the same experimental units but not on the same subsampling units. In our example, the two responses were measured in two different regions of the colon.

KEY WORDS: Adaptive inference; Asymptotics; Bootstrap; Colon cancer; Correlation; Functional data analysis; Local polynomial regression; Mixed models; Nutrition; Semiparametric estimation.

1. INTRODUCTION

In studying the etiology of colon cancer, it is important to understand the mechanisms of DNA damage experienced within distal (back) and proximal (front) colon cells in rats after exposure to a carcinogen [azoxymethane (AOM)] and having been fed certain types of diet. This damage can lead to the formation of tumors in affected crypts if not either repaired or removed. The magnitude of DNA damage suffered by a cell can be characterized by the response DNA adduct level. (Rogers and Pegg 1977; Swenberg et al. 1979). Carcinogen-induced colon cancer in rodent models has been used extensively to delineate molecular mechanisms of colon tumorigenesis.

There are strong epidemiological, clinical, and experimental indications of the existence of links between diets and colon cancer incidence. In particular, a fish oil-supplemented diet, compared to a corn oil-supplemented diet, was shown to have a protective effect against colon tumorigenesis (Haenszel and Kurihara 1968; Blot et al. 1975; Bang, Dyerberg, and Hijorne 1976; Boyle, Zaridze, and Smans 1985; Chang, Chapkin, and Lupton 1997; Hong et al. 2000). In addition, evidence from

epidemiological and rodent-based studies suggests that the anatomical distribution of tumors within the colon is affected by diet (Jacobs and Lupton 1986; Weisburger 1991; Chang et al. 1994; Zhang and Lupton 1994; Holt et al. 1996). Much is still unknown concerning the biological mechanisms contributing to these effects.

Given the general agreement that the etiologies of proximal and distal colon cancers differ, there is interest in understanding the anatomically site-specific molecular mechanisms regulating colon carcinogenesis. One area of interest is to investigate the relative expression of the DNA damage within the distal and proximal sites within the colon and compare this relative expression for rats fed either fish oil or corn oil diets.

This raises various interesting questions, one of which involves comparing the anatomical distribution of damage between the distal and proximal colon regions for the two diets. We pose this problem as estimation of the correlation between the distal and proximal DNA adduct level responses for rats fed fish oil and corn oil diets. A positive, zero, or negative correlation corresponds to relative damage distributions that are uniform throughout, unrelated, or localized within one of the two regions of the colon. Diet is expected to influence this correlation, because it affects the chemical environment contributing to the activation patterns of the carcinogen in the colon.

Estimation of these correlations is considered an exploratory analysis. Potentially, it could yield insights into the effects of dietary lipids on biological processes regulating the formation of carcinogen-induced DNA adducts in distal and proximal

Jeffrey S. Morris is Assistant Professor, Department of Biostatistics, University of Texas, M.D. Anderson Cancer Center, Houston, TX, 77030-4095. Naisyin Wang is Associate Professor, Department of Statistics, Texas A&M University, College Station, TX 77845-3143. Joanne R. Lupton is: Professor of Animal Science, of Food Science and Technology, of Nutritional Sciences and of Veterinary Anatomy and Public Health and Holder of the William W. Allen Endowed Chair in Nutrition. Robert S. Chapkin is Professor of Nutritional Sciences and Veterinary Anatomy and Public Health, Nancy D. Turner is Research Assistant Professor and Mee Young Hong is Postdoctoral Researcher, Department of Animal Science, Texas A&M University, College Station, TX 77843-2471. Raymond J. Carroll is Distinguished Professor, Departments of Statistics, Professor, Department of Epidemiology and Biostatistics, Nutrition and Toxicology, Texas A&M University, College Station, TX 77845-3143. This work was supported in part by grants NIH CA57030, CA74552, CA59034, CA61750, and NIEHS P30-ES09106.

regions of the colon, and it could further suggest hypotheses to investigate more rigorously in future studies.

Colonic cells replicate and grow completely within discrete units called crypts. Within each crypt, there are stationary, permanent cells called stem cells that generate all the cells within that crypt. Daughter cells are formed at the crypt depths where the stem cells are located, typically the bottom; as more cells are created, they move up the crypt until, finally, they are exfoliated into the lumen. Once they reach the uppermost third of the crypt, cells normally will not divide (Lipkin, Bell, and Sherlock 1963; Lipkin 1974). Cells near the bottom of the crypt are younger, and cells near the top of the crypt are older. Thus, a cell's relative depth within the crypt, hereafter referred to as relative cell position, is closely related to its age. Although this is a rough assessment, without further knowledge it is still the most reasonable action by which to measure the relative cell position from the bottom of the crypt.

This special cell-life sequence within colonic crypts suggests two important facts. First, cells at different depths are in different biological stages, and, second, cells at the same relative cell positions of different crypts share common characteristics. As a result, it is important to study biological mechanisms in the colon as a function of cell depth, because averaging over all crypt cells obscures any potential depth-specific effects. Thus, in this article we estimate the aforementioned correlation as a function of relative cell position.

The methodology developed in this article was motivated by and is applied to data from a colon carcinogenesis study performed on rats (Hong et al. 2000). In this study, 30 rats were randomized to a diet supplemented with either fish oil or corn oil. Three rats from each diet group were euthanized to serve as controls, and the others were injected with AOM, a carcinogen known to induce colon cancer, and then euthanized at 3, 6, 9, or 12 hours after exposure to the carcinogen. For each rat, roughly 20 crypts were then selected from the proximal and distal regions of the colon, and the DNA adduct level measurement was taken for cells lining the left side of a cross section of each crypt (roughly 40 cells per crypt). There was also a corresponding level of the covariate, relative cell position, for each cell. The values of the covariate were equally spaced within each crypt, ranging from zero at the very bottom of the crypt to one at the very top.

A difficulty that prevents us from estimating the correlation of interest by using conventional methods is that the two responses are not measured simultaneously on the same basic units, cells within crypts. They are, however, both measured on the same experimental units, rats, so we can obtain an estimate of the correlation at the experimental unit level.

Several papers in the statistical literature examined the estimation of a correlation as a function of a covariate. Bjerve and Doksum (1993) and Doksum et al. (1994) introduced a correlation curve as a function of a covariate X . Their goal differed from ours, however, in that they tried to model a correlation between a single response Y and the covariate X as a function of X , whereas we are modeling correlation between two responses as a function of a common covariate. Leurgans, Moyeed, and Silverman (1993) effectively considered a correlation between two responses as a function of a covariate, but their data are not of a mixed-model nature. Also, unlike in

our case, they observed both responses on the same basic unit of measurement. Given the structure of our data, there are no methods in the existing statistical literature that address our specific problem, so a primary goal of this article is to develop approaches that allow us to estimate this quantity. Both parametric and nonparametric methods are used.

A general model describing the data is proposed in Section 2, and the correlation function of interest is defined in detail. The parametric methods for estimating the correlation function, described in Section 3, basically rely on hierarchical mixed models that are carefully tailored to fit the design of the AOM experiment. Likelihood ratio and parametric bootstrap approaches to inference are used.

As described in Section 5, analysis of the AOM data yields surprising results. This naturally leads us to question whether the results so obtained are a result of the parametric modeling or are something intrinsic to the data. To check model robustness, we develop a new nonparametric correlation function estimate that is an extension of standard functional data analysis (Ramsay and Silverman 1997) appropriate for the design. In effect, what we did was first fit nonparametric regressions at the crypt level and then fit a mixed model locally to estimate the correlation at each cell location. Our analysis of the AOM experiment yields results that are similar to those of the parametric method. Surprisingly, parametric bootstrap simulations indicate that the nonparametric method is very nearly efficient, i.e., there is essentially no effect due to nonparametrically estimating the adduct level–cell position function at the crypt level. In Appendix A and Appendix B, we develop a theoretical justification for this empirical result, which suggests that at least in terms of rates of convergence, given the setup of the AOM experiment one would expect that the nonparametric method should be nearly efficient.

In Section 2, a general model is proposed and the correlation function is described. Sections 3 and 4 describe the parametric and nonparametric methods, respectively. Section 5 contains the details of the data analysis of the AOM experiment. Simulations are carried out to examine the performance of the methods. Section 6 contains a discussion of the implications of these results and the general applicability of the methodology studied. The theory of nonparametric correlation function estimation is outlined in Appendix A. Appendix B contains a higher order expansion of local linear kernel nonparametric regression and is of some independent interest.

2. THE GENERAL MODEL AND CORRELATION FUNCTION

Each of the two responses is assumed to follow a nested mixed model with a fixed mean structure and random, mean zero processes at the experimental unit, subsampling, and residual error levels. In our example, the two responses are the DNA adduct levels in the distal and proximal regions of the colon; the experimental units are rats, and the subsampling units are crypts. The mean structure and the processes are functions of the covariate X , the relative cell position. The fixed mean structure may differ for different diets as well as for times since carcinogen exposure. The response (DNA adduct level) is denoted by Y . Consider time group

$j = 1, \dots, 5$, rat $r = 1, \dots, R$, and cell number i . The models for the DNA adduct levels at relative cell position X for distal crypt c_d and proximal crypt c_p of rat r , time group j are

$$Y_{jrc_p}(X) = m_j(X, p) + m_{jr}(X, p) + m_{jrc_p}(X, p) + \epsilon_{jrc_p}(X, p), \quad (1)$$

$$Y_{jrc_d}(X) = m_j(X, d) + m_{jr}(X, d) + m_{jrc_d}(X, d) + \epsilon_{jrc_d}(X, d). \quad (2)$$

In (1) and (2), $m_j(\cdot, p)$ and $m_j(\cdot, d)$ denote fixed mean functions for time group j in the proximal and distal regions, respectively. The functions $m_{jr}(\cdot, p)$ and $m_{jr}(\cdot, d)$ are the realizations of rat level random effects, the functions $m_{jrc_p}(\cdot, p)$ and $m_{jrc_d}(\cdot, d)$ are the realizations of crypt level random effects, and the ϵ 's are residual errors. Conceptually, these models give us curves for the two responses as functions of relative cell position for each crypt. The different subscripts for the distal and proximal crypts emphasize that although both measurements are taken from the same rat, they are from different crypts. We assume that except for a possible nonzero correlation $\rho(X_1, X_2)$, depending on cell position, between $m_{jr}(X_1, p)$ and $m_{jr}(X_2, d)$, the random functions $m_{jr}(\cdot, p)$, $m_{jr}(\cdot, d)$, $m_{jrc_p}(\cdot, p)$, and $m_{jrc_d}(\cdot, d)$ and the ϵ 's are mutually independent. For simplicity, we assume that the ϵ 's are also independent within crypts, which effectively holds in our example, although this assumption is not necessary and the methods can be modified to weaken the assumption. All random processes have mean zero. Note that the independence assumptions imply that measurements from the same rat but different crypts are dependent on each other only through the rat level random effect. Although obviously the fixed effects $m_j(X, p)$ and $m_j(X, d)$ are of interest, we primarily are interested here in estimating the correlation between the distal adduct level and the proximal adduct level as a function of relative cell position.

If adduct levels could be measured on the same cells from the same crypts, then we could just calculate the estimated correlation function by using a simple generalization of the usual variance and covariance formulas (Ramsay and Silverman 1997). Obviously, in our case adduct levels cannot be measured for the same cells or crypts. Making use of the fact that we have both proximal and distal measurements for each rat, however, we can estimate their correlation at the rat level.

The function of interest, namely, the correlation at the rat level having accounted for overall trends, as a function of relative cell position X is

$$\rho(X) = \text{corr}\{m_{jr}(X, p), m_{jr}(X, d)\}. \quad (3)$$

Because we do not observe measurements at the same relative cell positions for each crypt, we cannot estimate this function directly without specifying and fitting the model (2). Our first approach is to estimate the correlation assuming parametric structure on the various cell positions functions of this model.

3. PARAMETRIC APPROACH

This section describes our parametric approach, which is based on a Gaussian hierarchical mixed model.

3.1 Model and Correlation Function

In our Gaussian mixed model, parametric structure is assumed for the fixed mean function as well as for the rat- and crypt-level random effects. For the proximal response, we let $\beta_{p,j}$, $\beta_{p,jr}$, and $\beta_{p,jrc}$ denote the fixed effect and rat- and crypt-level random effects, respectively. For the vector of responses \underline{Y}_{jrc_p} within a particular crypt c_p , at time j and for rat r , we can in general write the parametric model version of (1) as

$$\underline{Y}_{jrc_p} = Z_{1,jrc_p}\beta_{p,j} + Z_{2,jrc_p}\beta_{p,jr} + Z_{3,jrc_p}\beta_{p,jrc} + \underline{\epsilon}_{jrc_p}, \quad (4)$$

where Z_{1,jrc_p} , Z_{2,jrc_p} , and Z_{3,jrc_p} are the corresponding design matrices. The random variables $\beta_{p,jr}$ and $\beta_{p,jrc}$ are assumed to be independent mean zero Gaussian random variables with covariance matrices $\Sigma_{1,p}$ and $\Sigma_{2,p}$, respectively. Equivalent model and structures are assumed for the distal region. Whereas, as described above, the crypt-level random effects $\beta_{p,jrc}$ and $\beta_{d,jrc}$ are assumed to be mutually independent, the rat-level random effects $\beta_{p,jr}$ and $\beta_{d,jr}$ have covariance matrix $\Sigma_{p,d}$. In our example, we assume that $\underline{\epsilon}_{jrc_p}$ and $\underline{\epsilon}_{jrc_d}$ are mean zero Gaussian random variables with diagonal covariance matrices. As mentioned, more general assumptions allowing for autocorrelation in the errors within the crypts are easily accommodated.

If at a given value of X the component of the design matrix Z_{2,jrc_p} is denoted by the vector $Z_p(X)$, and $Z_d(X)$ is defined similarly, then the correlation function $\rho(X)$ defined in (3) becomes

$$\rho(X) = \{Z_p^T(X)\Sigma_{p,d}Z_d(X)\} \{Z_p^T(X)\Sigma_{1,p}Z_p(X)\}^{-1/2} \times \{Z_d^T(X)\Sigma_{1,d}Z_d(X)\}^{-1/2}. \quad (5)$$

3.2 Estimation Issues

To estimate this correlation function, we need to estimate the rat-level covariance matrices $\Sigma_{1,p}$, $\Sigma_{1,d}$, and $\Sigma_{p,d}$. For this article, restricted maximum likelihood (REML) estimates of these matrices were obtained by using PROC MIXED in SAS. Pointwise confidence bounds for the estimate of (5) were found by using a parametric bootstrap procedure (Efron and Tibshirani 1993). To make this procedure computationally feasible, we used MIVQUE (Swallow and Searle 1978) rather than REML estimates of the rat-level covariance matrices for the bootstrap. This gave error bounds that were conservative for REML estimates, but spot checks indicated that the two estimates were nearly identical.

4. NONPARAMETRIC APPROACH

As described in Section 1, and as detailed in Section 5, the parametric analysis of the AOM data yielded surprising results. Our aim here is to provide a nonparametric estimate of the correlation function, used in effect as a model robustness check on the parametric approach but of considerable interest itself.

There is a variety of ways to fit the correlation function nonparametrically. We adopt a functional data analysis approach (Ramsay and Silverman 1997), modified for our

particular problem. In scalar form, for a cell i with the relative cell position $X_{jrc_p,i}$, our models are

$$Y_{jrc_p,i} = m_j(X_{jrc_p,i}, p) + m_{jr}(X_{jrc_p,i}, p) + m_{jrc_p}(X_{jrc_p,i}, p) + \epsilon_{jrc_p,i}, \quad (6)$$

$$Y_{jrc_d,i} = m_j(X_{jrc_d,i}, d) + m_{jr}(X_{jrc_d,i}, d) + m_{jrc_d}(X_{jrc_d,i}, d) + \epsilon_{jrc_d,i}, \quad (7)$$

where $\epsilon_{jrc_p,i}$ and $\epsilon_{jrc_d,i}$ are independent and identically distributed mean-zero random variables with variances $\sigma_{\epsilon,p}^2$ and $\sigma_{\epsilon,d}^2$, respectively.

Our approach is to estimate the crypt-level functions on a fixed grid of cell positions and then separately fit a mixed model at each point on the grid. Specifically, within each crypt we use nonparametric regression techniques to estimate the crypt-level function, based on a sample of approximate size 40 cells, and we then evaluate the fit on the grid. The results in (A.16) and (A.18) indicate that, at each grid point, the obtained estimates follow a linear mixed model with solely rat-level random effects and crypt-level random errors, that is, (6)–(7) without the ϵ 's; see (8)–(9). We thus obtain REML estimates of $\rho(X)$ by pointwise fitting a mixed model with the suggested structure to the nonparametrically estimated crypt-level functions. Both our numerical analysis (Section 5) and our theoretical analysis (Appendix A) suggest that the variability of the nonparametric regression at the crypt level is sufficiently small compared to the rat-to-rat variability so that the estimation error can be ignored for practical purposes.

We provide here a few important details of the methodology. In the AOM data the method of nonparametric regression is not crucial, but the numerical results displayed in this article used a Gasser–Müller estimator using the Epanechinov kernel to estimate the crypt-level curves. Boundary kernels were used (Gasser and Müller 1979) to minimize the well-known boundary bias effect (see Hart 1997 for discussion). Our proposed crypt-by-crypt method also was performed by using locally weighted regression (lowess) rather than Gasser–Müller smoothers, with very little change in the results.

Along with a choice of kernel estimator must come a choice of bandwidth. To simplify our discussion, we assume that there are R rats, C crypts per rat, and n cells per crypt, assumptions that do not have a substantive bearing on the principles discussed. The most convenient approach to bandwidth selection would be to optimize the bandwidth at the crypt level, i.e., estimate the bandwidth to minimize mean squared error within each crypt. Recall, however, that our aim is not to estimate the crypt level regression function but to estimate the correlation function $\rho(X)$ in (3). We claim that optimizing the bandwidth at the crypt level is the wrong strategy in general. The results of Appendix A show that for there to be no effect of the nonparametric regression on the estimate of the correlation function $\rho(X)$, a situation we call efficiency, asymptotically as $R, n \rightarrow \infty$ we require that $Rh^4 \rightarrow 0$. In general, at the crypt level the optimal bandwidth is well known to be of order $h \sim n^{-1/5}$. Thus, for efficiency to be obtained, we require $Rn^{-4/5} \rightarrow 0$. In our example, $R = 15$, $n \approx 40$,

and $Rn^{-4/5} = .78$, which even charitably is not particularly near zero. Thus, from rates of convergence calculations, optimizing at the crypt level generally does not lead to an efficient estimate of the correlation function in the sense described.

The basic theoretical problem with choosing the bandwidth to optimize the function at the crypt level is that the resulting function is too smooth, because additional smoothing is done when effectively averaging across crypts as the mixed model is applied. In fact, one should undersmooth at the crypt level, allowing the averaging over crypts to provide the correct level of further smoothing for optimal estimation of the correlation function. For example, if we were to pool the data across all crypts within a rat, ignoring the crypt-to-crypt variability, then the optimal bandwidth for estimating the rat level function is of the order $(nC)^{-1/5}$, or approximately the optimal bandwidth at the crypt level multiplied by $C^{-1/5}$, that is, $h_{\text{opt}} C^{-1/5}$. In our case, $nC \approx 800$, $R(nC)^{-4/5} \approx .07$, and $R(h_{\text{opt}} C^{-1/5})^4 \approx .004$. This bandwidth selection procedure is an improvement over optimizing at the crypt level.

To implement this idea, we estimated the optimal bandwidth separately for each crypt, selected the median of this set h , and multiplied by $(C_{jrp})^{-1/5}$ for all proximal crypts for rat r of time group j and by $(C_{jrd})^{-1/5}$ for the distal crypts. This procedure adjusts for the differing number of crypts per rat and approximates the amount of undersmoothing necessary to optimally estimate the correlation function. Obviously, if the results of inferences were sensitive to the choice of bandwidth, one would need to derive a more sophisticated bandwidth selection procedure. Our sensitivity analysis in Section 5.2, however, indicates that such a practice is not necessary for our example. Appendix A provides the theoretical details necessary to implement such an approach and makes a brief suggestion about numerical implementation for cases where it may be considered necessary.

After obtaining nonparametric estimates of the crypt-level functions in the proximal and distal crypts, $\tilde{Y}_{jrc_p}(X_i)$ and $\tilde{Y}_{jrc_d}(X_i)$, we fit the following mixed model at each fixed grid point X_i :

$$\tilde{Y}_{jrc_p}(X_i) = m_j(X_i, p) + m_{jr}(X_i, p) + \tilde{m}_{jrc_p}(X_i, p), \quad (8)$$

$$\tilde{Y}_{jrc_d}(X_i) = m_j(X_i, d) + m_{jr}(X_i, d) + \tilde{m}_{jrc_d}(X_i, d), \quad (9)$$

where $m_{jr}(X_i, p)$ and $m_{jr}(X_i, d)$ are mean-zero multivariate normal random variables with variances $\sigma_r^2(X_i, p)$ and $\sigma_r^2(X_i, d)$, respectively, and covariance $\sigma_r(X_i, p, d)$, and $\tilde{m}_{jrc_p}(X_i, p)$ and $\tilde{m}_{jrc_d}(X_i, d)$ are independent and identically distributed mean-zero random variables with variances $\sigma_m^2(X_i, p)$ and $\sigma_m^2(X_i, d)$, respectively. For each value on the grid, each model is simply a linear mixed model with rat-level random effects and independent and identically distributed crypt-level random errors. The estimated correlation function, $\hat{\rho}(X_i)$, was then obtained by using the REML estimates of the rat-level covariance parameters.

An alternate approach involves pooling the data and fitting a single nonparametric regression at the rat level. It can be shown that when the ratio of crypts to rats is small, such a method attenuates the correlation estimate in an unpredictable way, depending on the cell positions. Regressing at the crypt

level and then fitting a mixed model properly accounts for the between-crypt covariance structure and thus avoids this attenuation.

We developed a nonparametric bootstrap procedure to obtain confidence bounds on the estimate of the correlation function and to test various hypotheses of interest. Our bootstrap datasets were constructed by using models (8) and (9) and samples taken with replacement from a set of unbiased nonparametric predictors of the rat-level random effects and crypt-level residuals, which we now describe. Let R_j be the number of rats per time period j : $R_j \equiv 3$ in our example. Let $C_{jr}(p)$ be the number of proximal crypts for rat r of time period j . Then consider the following expressions for proximal data:

$$\hat{m}_{jr}(X_i, p) = \left(\frac{R_j}{R_j - 1} \right)^{1/2} \times \left(\frac{\sigma_r^2(X_i, p)}{\sigma_r^2(X_i, p) + \sigma_m^2(X_i, p)} C_{jr}(p) \right)^{1/2} \times \left\{ \bar{Y}_{jr\bullet}(p) - \bar{Y}_{j\bullet\bullet}(p) \right\}, \quad (10)$$

$$\hat{m}_{jrc_p}(X_i, p) = \left(\frac{C_{jr}(p)}{C_{jr}(p) - 1} \right)^{1/2} \left\{ \tilde{Y}_{jrc_p}(p) - \bar{Y}_{jr\bullet}(p) \right\}. \quad (11)$$

Similar expressions are used for the distal region. The definitions (10) and (11) ensure that $\text{var}\{\hat{m}_{jr}(X_i, p)\} = \sigma_r^2(X_i, p)$ and that $\text{var}\{\hat{m}_{jrc_p}(X_i, p)\} = \sigma_m^2(X_i, p)$. The unknown covariance parameters were substituted by the REML estimates from the full dataset.

To preserve the correlation of the curves across relative cell positions, this bootstrapping was done for all positions simultaneously. Furthermore, bootstrap observations of $\hat{m}_{jr}(X_i, p)$ and $\hat{m}_{jr}(X_i, d)$ were constructed by sampling matching pairs together.

For each bootstrap dataset, we obtained the REML estimates of the rat-level covariance parameters $\sigma_r^2(X_i, p)$, $\sigma_r^2(X_i, d)$, and $\sigma_r(X_i, p, d)$. The estimated correlation was then calculated from these REML estimates. Straightforward calculations show that $\text{cov}\{\hat{m}_{jr}(X_i, p), \hat{m}_{jr}(X_i, d)\}$ equals

$$\left\{ \frac{\sigma_r^2(X_i, p)}{\sigma_r^2(X_i, p) + \sigma_m^2(X_i, p)} C_{jr}(p) \right\}^{1/2} \times \left\{ \frac{\sigma_r^2(X_i, d)}{\sigma_r^2(X_i, d) + \sigma_m^2(X_i, d)} C_{jr}(d) \right\}^{1/2} \sigma_r(X_i, p, d). \quad (12)$$

Therefore, a correction factor based on (12) was multiplied to each bootstrap correlation estimate to warrant consistency.

5. ANALYSIS OF THE RAT AOM AND DIET DATA

The methods described in the previous sections were applied separately for the fish oil and corn oil diet groups, with the intention of comparing them. Subsections 5.1 and 5.2 present the results from each of these methods, and the two methods are compared in Section 5.3.

By using both methods, we obtained results that were somewhat surprising. Some of these results were only marginally

statistically significant, but considering the low power resulting from our small sample sizes, that is, only three rats from each of five time levels, this was not surprising. Section 5.4 describes a simulation done to estimate the power of some of these tests and provide sample size recommendations for future studies.

5.1 Parametric Approach

Initial analyses suggested that the covariate effect was quadratic in shape, so the initial full model we considered had the fixed mean, random rat, and random crypt portions of the model all involving full quadratics in relative cell position. The joint covariance matrix of the random rat coefficients was unstructured in form, as were the covariance matrices for the random crypt coefficients $\Sigma_{2,p}$ and $\Sigma_{2,d}$. The parabolic coefficients in the fixed mean were allowed to differ for the five time levels. To minimize collinearity in the coefficients, orthogonal polynomials were used in the design matrices. After examining various reduced models and performing likelihood ratio tests, we arrived at final models for Y and W for both diet groups. In all cases, the best fitting model involved full quadratics for both the fixed mean and the random crypt portion, with covariance matrices left unstructured. There were differences, however, in the random rat level portions. For both diets, the random rat portion of the distal response was fully quadratic, but that of the proximal response was only linear.

Results from PROC MIXED in SAS and crypt-level plots (not shown) suggested a slight first-order autocorrelation ($< .10$) in the within-crypt residuals. This autocorrelation had no practical impact on the example, a fact we confirmed by rerunning our analyses on even less autocorrelated data, namely, first the even-numbered cells and then the odd-numbered cells. Both reduced datasets had negligible autocorrelation, and both gave results that were virtually the same as those reported here.

Figure 1 shows the parametric estimates of the correlation functions between DNA adduct levels in the distal and proximal regions of the colon for both fish oil and corn oil diets, as well as pointwise 90% confidence bounds from the parametric bootstrap. All the parametric bootstrap results reported in this section were obtained by using 2,000 bootstrapped datasets.

For rats fed fish oil diets, the correlation estimate was positive near the bottom of the crypt and close to zero in the middle and upper portions of the crypts. There was weak evidence of the existence of nonzero correlation ($p = .056$), based on a likelihood ratio test. The parametric bootstrap results, showed there was also evidence of a position effect, with the correlation at the very bottom of the crypts significantly greater than the correlation at the very top of the crypts ($p = .028$). A positive correlation here would suggest that in the bottom of the crypts, fish oil-fed rats with higher-than-average DNA adduct levels in the distal region of their colons also tended to experience higher-than-average DNA adduct levels in the proximal region of their colons. Such effects diminished as cells further divided.

The correlation estimate for the rats fed corn oil was negative throughout the crypts. From a likelihood ratio test, there was evidence of nonzero correlation ($p = .018$), and pointwise bootstrap confidence intervals were marginally statistically significant for the middle 60% of the crypts. There was

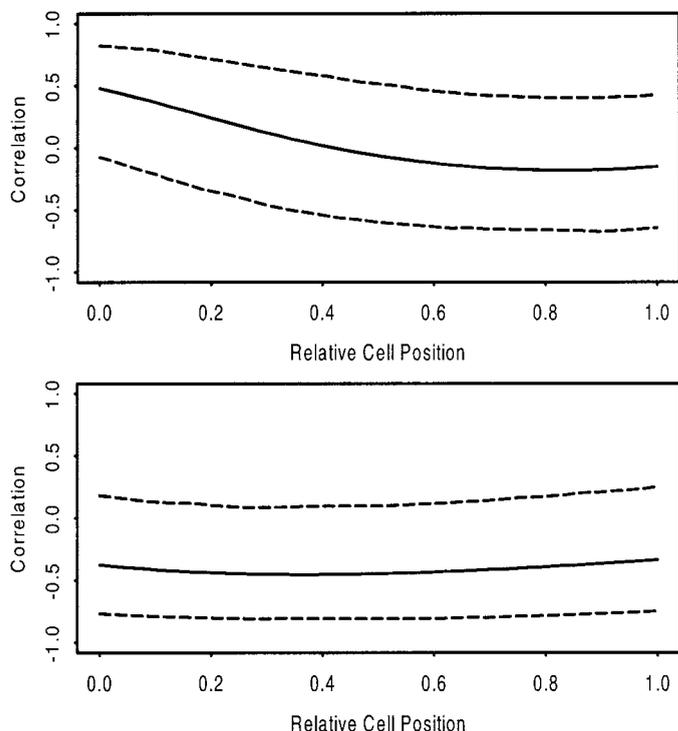


Figure 1. Parametrically Estimated Correlation Function Between DNA Adduct Level in Distal and Proximal Sections of Colon, Function of Relative Cell Position, for Fish Oil and Corn Oil Diets With 90% Parametric Bootstrap Confidence Bounds. $B = 2,000$.

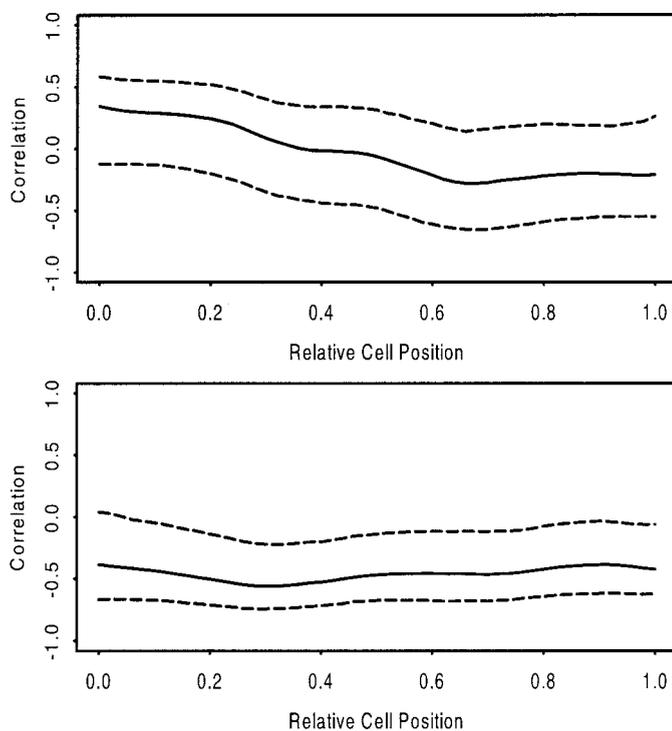


Figure 2. Nonparametrically Estimated Correlation Function Between DNA Adduct Level in Distal and Proximal Sections of Colon, Function of Relative Cell Position, for Fish Oil and Corn Oil Diets With 90% Nonparametric Bootstrap Confidence Bounds. $B = 2,000$.

no evidence from the parametric bootstrap of any position effect. A negative correlation like that observed here would mean that uniformly down the crypts, corn oil-fed rats with higher-than-average DNA adduct levels in the distal region of their colons tended to experience lower-than-average DNA adduct levels in the proximal region of the colon, and vice versa. On the basis of the parametric bootstrap results, the difference between the fish oil and corn oil correlations was significant ($p < .05$) for cells belonging to the bottom 10% of the crypts.

5.2 Nonparametric Approach

Here, our nonparametric approach is applied to the AOM data. Its development was motivated by a desire to check whether we obtain the same results in the AOM example when the parametric assumptions on the functions are removed.

We estimated the optimal bandwidth at the crypt level for all crypts: the median optimal bandwidth over all crypts was $h = .23$. This bandwidth was then adjusted for the number of crypts in each rat during estimation, see Section 4. To check how sensitive the nonparametric estimator is to the choice of bandwidth, the estimation procedure was repeated by using various fixed and variable bandwidths ranging from .05 to .40, but it reached nearly identical results.

Figure 2 shows the nonparametric estimates of the correlation functions between DNA adduct levels in the distal and proximal regions of the colon for both fish oil and corn oil diets, as well as 90% confidence intervals from the nonparametric bootstrap.

For rats fed the fish oil diet, the nonparametric correlation function estimate behaved similarly to that from the parametric method. There was a positive estimated correlation at the bottom of the crypts that decreased to near zero for the middle and upper portions of the crypts. The statistical significance of the results was similarly marginal.

For rats fed the corn oil diet, the nonparametric method also yielded correlation estimates that were negative at all depths within the crypts, with no position effect. The negative correlation encountered was significant ($p < .05$) for all cell positions in the middle 60% of the crypts. The fish oil and corn oil correlation functions were marginally significantly different ($p < .10$) over the bottom 35% of the crypts.

As a verification of the proposed nonparametric bootstrap estimated confidence bounds, 90% confidence bounds on the function of interest were constructed on the basis of Fisher's Z transformation (Fisher 1954). The estimated correlation function, $\hat{\rho}(X)$, was transformed by using Fisher's $Z = (1/2) \log\{1 + \hat{\rho}(X)\} / \{1 - \hat{\rho}(X)\}$ at each value over the same fixed grids of X . The approximate standard error of these Z statistics, independent of $\rho(X)$, is known to be $(R - 3)^{-1/2}$, leading to a simple t -type confidence interval. These intervals matched very closely with those obtained by using the proposed nonparametric bootstrap procedure.

5.3 Comparison of Parametric and Nonparametric Methods

It is of interest to compare the performance of these two methods in a situation where a particular parametric model holds. A simulation study was performed by using the 2,000

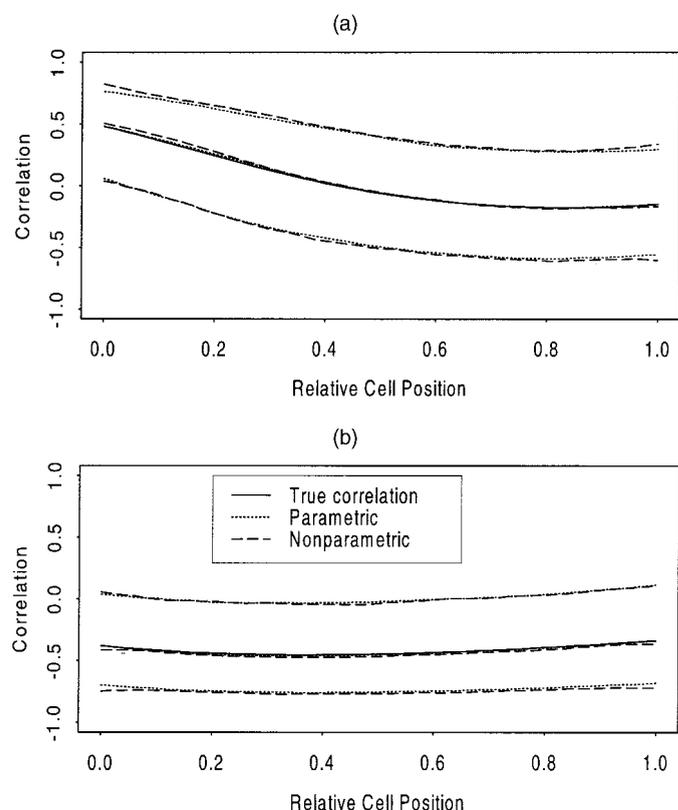


Figure 3. Parametric Bootstrap Results ($B = 2,000$) for Parametric and Nonparametric Methods. The mean correlation curves and 90% confidence bounds for both methods are given for (a) fish oil and (b) corn oil diets. The true correlation curves for each diet were also plotted.

parametric bootstrap samples generated in Section 3.2. A correlation function estimate using each method was obtained for each of the bootstrap samples.

In Figure 3, (a) and (b) show the mean and 90% confidence bounds for the parametric and nonparametric estimates for fish oil and corn oil, respectively, as well as the true correlation function with which the data were generated.

Figure 3 indicates that, even for efficiency consideration, the nonparametric method is competitive with the parametric method using MIVQUE estimates for this parametric model. The means of the estimated functions were close to the true function for both methods. The precision of the two methods was also comparable: note the almost identical 90% confidence bounds.

This result, although surprising at first glance, is to be expected from the rate of convergence theory developed in Appendix A. In effect, with roughly 15 rats, 20 crypts, and 40 cells per rat, and with our bandwidth selection method, we expect the nonparametric method to give results essentially the same as if we knew the true crypt-level functions exactly. Thus, in this case we effectively pay a small price in efficiency for the flexibility of nonparametric modeling. For both parametric and nonparametric approaches, the estimated correlations are root- R consistent with the rat-level variation as the dominating factor.

5.4 Power Simulation Results

Many of our analyses found only marginal levels of statistical significance. A simulation was therefore done to check

Table 1. Simulated Power for Detecting Positive Fish Oil Correlation, Negative Corn Oil Correlation, and Positive Difference Between Fish and Corn Oil Correlation at Bottom of Crypts

H_A :	$\rho_{FISH} > 0$	$\rho_{CORN} < 0$	$\rho_{FISH} - \rho_{CORN} > 0$
$n = 3$.3114	.3812	.5852
4	.4894	.5872	.8086
5	.6265	.7264	.9115
6	.7425	.8282	.9626
7	.8089	.8875	.9845
10	.9345	.9708	.9994

NOTE: The correlations and differences were set to be the estimates from the current dataset. Numbers of rats per time group, n , are chosen to be 3–7 and 10.

the power for testing for positive fish oil correlations, negative corn oil correlations, and differences between fish oil and corn oil correlations at the very bottom of the crypts. The setup was aimed at detecting correlations and differences like those encountered in the data, assuming that the nonparametric BLUP estimates from the data reflect the true random effects.

The simulation involved calculating test statistics and p -values for the three hypotheses for each of $M = 10,000$ simulated data sets. At each step of the simulation, data were generated from the multivariate normal distribution that best fit the empirical distribution of the BLUP's for the rat-level functions at the very bottom of the crypts. For computational feasibility, t -tests based on the Fisher's Z transformation, as discussed in Section 5.2, were used to replace the bootstrap-based tests.

The simulation was performed for datasets containing $n = 3, 4, 5, 6, 7$, and 10 rats per time since carcinogen exposure group. The estimated power for each hypothesis test is reported in Table 1.

We can see that with three rats per time group, our power was rather low for detecting positive fish oil and negative corn oil correlations (.31 and .38, respectively). The power for detecting a larger fish correlation, comparing to corn, was slightly higher (.59). We would need at least four rats per time group to have at least 80% power for detecting a difference between fish oil and corn oil like that encountered in the data. To have 80% power to detect the specified alternatives for fish and corn, respectively, we would need at least seven rats per group.

These results demonstrate that, although the dataset is itself large, the effective sample size for estimating the correlation function is quite small, and thus we have low power for detecting whether the correlations encountered are significantly different from zero. The results also provide guidelines for choosing sample sizes in designing future studies.

6. DISCUSSION

Our analysis of the rat AOM and diet data suggest that for rats fed fish oil-supplemented diets, there may be a positive relationship between the DNA adduct level measurements in the distal and proximal regions of the colon near the bottom of the crypts. For rats fed corn oil-supplemented diets, however, the observed relationship between DNA adduct level measurements in the distal and proximal regions of the colon was negative at all depths. Such a negative relationship implies that rats who experienced higher-than-average DNA adduct

levels in their distal crypts tended to experience lower-than-average DNA adduct levels in their proximal crypts, and vice versa. This result was surprising because negative correlations are rarely encountered in biology.

A variety of statistical tools, including a parametric mixed-model-based method and nonparametric FDA methods, were used to reach these conclusions. The various nonparametric procedures used involved kernel smoothing done either at the crypt level or marginally at the rat level, by using either Gasser–Müller or lowess smoothers, over a range of fixed and variable bandwidths. Confidence bounds were calculated and hypothesis tests were performed by using various bootstrapping procedures as well as asymptotic results involving Fisher’s Z transformation. These methods all generally agreed and suggested the same results.

As mentioned, the estimation of these correlations was meant to be an exploratory analysis, potentially yielding some hypotheses concerning the effects of dietary lipids on some of the biological mechanisms regulating colon carcinogenesis. In discussing the possible implications of the results observed here, however, we must be cautious considering the marginal significance level of some of the observed effects. At the same time, we believe that this caution should not prevent us from searching for possible explanations for these results, given the somewhat large magnitude of the observed effects and the low power for detecting nonzero correlations due to the small effective sample sizes (see Section 5.4). This is especially true considering that we have a rather unique case where, at the bottom of the crypts, we observe an apparent diet effect in which the directions of the correlations actually differ. If this result is replicable, it could yield significant insights into the biological mechanisms regulating the observed fish oil–corn oil effects in colon carcinogenesis.

The carcinogen AOM, after metabolism in the liver, interacts with the chemical environment of the colon to induce the formation of the DNA adducts. A possible implication for our results is that the differing chemistries resulting from the fish oil or corn oil enhanced diets might have caused different activating patterns of the carcinogen. A negative relationship between the distal and proximal DNA adduct level measurements like that observed would suggest that the activation of the carcinogen may occur more locally for rats fed corn oil–enhanced diets, which could in principle intensify its effect on the colonic cells in that region. A chemical environment conducive to diffusion of the carcinogen activation throughout the colon would be one explanation for a positive correlation, such as that observed at the bottom of the crypts for rats fed fish oil. Carefully designed future studies involving a larger number of rats (see Section 5.4) are needed before these hypotheses can be verified. A better understanding of how the molecular environment imposed by diet affects the colon carcinogenesis process could lead to more targeted and accurate prevention efforts.

Estimation of this correlation function was nontrivial given the longitudinal hierarchical structure of the data, in which we never observed proximal and distal measurements on the same cells, and we did not observe measurements at the same relative cell positions for each crypt. Our work considered

the correlation of two responses as a function of a longitudinal covariate. Along with the nonparametric methods themselves and a fairly simple bandwidth selection procedure, in Appendix A we develop an asymptotic theory concerning rates of convergence. The theory is important because it indicates the feasibility of applying simple linear mixed modeling locally in our nonparametric approach. Furthermore, the theory gives a theoretical explanation to a phenomenon we observed but did not expect, namely, that there is little effect on the efficiency of the correlation function estimate due to performing nonparametric regressions at the crypt level. This result is important because it exemplifies a case in which there is little price paid in efficiency for the flexibility of a nonparametric estimator. A few calculations also suggested that the nonparametric estimate should be fairly insensitive to the choice of the bandwidth, something we also observed in the example.

The methodology used in this article can be applied to other settings when the goal of the study is to model the correlation of two continuous repeated measurement responses as a function of a covariate, whereas the two responses of interest can be measured on the same experimental units but not on the same subsampling units. In our example, the two responses were measured in two different regions of the colon. Other examples include cases in which the responses can be measured only through a destructive procedure.

APPENDIX A: THEORY FOR KERNEL METHODS

This appendix derives conditions under which the correlation function estimated by using kernel regression methods is asymptotically the same as if the regression functions were known. We derive the results when there are $r = 1, \dots, R$ rats, $c_p = 1, \dots, C$ and $c_d = 1, \dots, C$ crypts in the proximal and distal regions, respectively, and $i = 1, \dots, n$ cells per crypt. The adduct levels in the proximal region are $Y_{rc_p,i}$ the cell positions are $X_{rc_p,i}$, and similarly for the distal region.

In what follows, for convenience we drop the dependence on j , the time level. This makes little difference to the results while greatly simplifying already complex calculations. Also, to make the calculations as straightforward as possible, we assume that cell positions are independent and identically distributed with density function $f(\cdot)$ and that within a crypt the regression by function of DNA adduct levels is fit via local linear regression by using a symmetric kernel density function $K(\cdot)$ with variance 1. For a bandwidth h , define $K_h(x) = h^{-1}K(x/h)$. This is slightly different from the method actually used in the application, but the numerical results are not much affected by the change in procedure, and the theoretical development is far cleaner. In our asymptotics, we require that $R \rightarrow \infty$ and that $n \rightarrow \infty$: C can remain fixed or can also converge to ∞ . Because the nonparametric function is fit at the crypt level, we immediately require that $nh \rightarrow \infty$.

Because of notational complexity inherent in (6)–(7), we write $G_{rc_p}(x, p) = m(x, p) + m_r(x, p) + m_{rc_p}(x, p)$, where $m(x, p)$ is a fixed function, $m_r(x, p)$ are mean zero random functions at the rat level, and $m_{rc_p}(x, p)$ are mean zero random functions at the crypt level. The j th derivative of any function F is denoted by $F^{(j)}$.

Let $\widehat{G}_{rc_p}(x, p, h)$ be the local linear kernel estimator computer for rat r and crypt c_p , and make the definitions

$$v_r(x, p) = m_r^{(2)}(x, p) - E\{m_r^{(2)}(x, p)\}, \tag{A.1}$$

$$g_r(x, p) = E\{m_{rc_p}^{(2)}(x, p)|\text{Rat} = r\} - E\{m_{rc_p}^{(2)}(x, p)\}, \tag{A.2}$$

$$b_r(x, p) = v_r(x, p) + g_r(x, p), \tag{A.3}$$

$$d_{rc_p}(x, p) = m_{rc_p}^{(2)}(x, p) - E\{m_{rc_p}^{(2)}(x, p)|\text{Rat} = r\}, \tag{A.4}$$

$$\mathcal{S}_r(x, p, h) = \{nCf(x)\}^{-1} \sum_{c_p=1}^C \sum_{i=1}^n K_h(X_{rc_{pi}} - x) \epsilon_{rc_{pi}}, \tag{A.5}$$

$$\mathcal{L}_r(x, p) = \overline{G}_r(x, p) - \overline{G}_*(x, p), \tag{A.6}$$

$$\mathcal{V}_{rc_p}(x, p) = G_{rc_p}(x, p) - \overline{G}_*(x, p), \tag{A.7}$$

$$\begin{aligned} \widehat{\mathcal{L}}_r(x, p, h) &= C^{-1} \sum_{c_p=1}^C \widehat{G}_{rc_p}(x, p, h) \\ &\quad - (RC)^{-1} \sum_{c_p=1}^C \sum_{r=1}^R \widehat{G}_{rc_p}(x, p, h), \end{aligned} \tag{A.8}$$

$$\widehat{\mathcal{V}}_{rc_p}(x, p, h) = \widehat{G}_{rc_p}(x, p, h) - C^{-1} \sum_{c_p=1}^C \widehat{G}_{rc_p}(x, p, h). \tag{A.9}$$

By using the results of Appendix B, there are functions $\phi_{rc_{pj}}(x, p)$, $j = 1, \dots, 4$, which are random because of (r, c_p) but are otherwise fixed, and asymptotically bounded random variables $B_{rc_{pj}}(x, p)$, $j = 1, 2, 3$, which are independent of the ϵ 's such that

$$\begin{aligned} &\widehat{G}_{rc_p}(x, p, h) - G_{rc_p}(x, p) - (h^2)2G_{rc_p}^{(2)}(x, p) \\ &= \{nf(x)\}^{-1} \sum_{i=1}^n K_h(X_{rc_{pi}} - x) \epsilon_{rc_{pi}} \\ &\quad + \left[\{nf(x)\}^{-1} \sum_{i=1}^n K_h(X_{rc_{pi}} - x) \epsilon_{rc_{pi}} \right. \\ &\quad \times \{h^2 \phi_{rc_{p1}}(x, p) + (nh)^{-1}2B_{rc_{p1}}(x, p)\} \\ &\quad + \{nf(x)\}^{-1} \sum_{i=1}^n K_h(X_{rc_{pi}} - x) \{ (X_{rc_{pi}} - x) h \} \epsilon_{rc_{pi}} \\ &\quad \times \{h \phi_{rc_{p2}}(x, p) + h^3 \phi_{rc_{p3}}(x, p)\} \\ &\quad + \{nf(x)\}^{-1} \sum_{i=1}^n K_h(X_{rc_{pi}} - x) \{ (X_{rc_{pi}} - x) h \} \epsilon_{rc_{pi}} \\ &\quad \times \{h(nh)^{-1}2B_{rc_{p2}}(x, p)\} \left. \right] \\ &\quad + [h^4 \phi_{rc_{p4}}(x, p) + \{h^2(nh)^{-1}2\} B_{rc_{p3}}(x, p)] \\ &\quad + O_p\{h^5 + h^3(nh)^{-1}2\}. \end{aligned} \tag{A.10}$$

Now define $\mathcal{F}_r(x, p, h)$ and $\mathcal{G}_r(x, p, h)$ to be the averages across crypts of the terms inside the first and the second set of square brackets in (A.10), respectively. Then, we have that

$$\begin{aligned} \widehat{\mathcal{L}}_r(x, p, h) &= \mathcal{L}_r(x, p) + (h^2)2\mathcal{L}_r^{(2)}(x, p) + \mathcal{S}_r(x, p, h) - \overline{\mathcal{S}}_*(x, p, h) \\ &\quad + \mathcal{F}_r(x, p, h) - \overline{\mathcal{F}}_*(x, p, h) + \mathcal{G}_r(x, p, h) - \overline{\mathcal{G}}_*(x, p, h) \\ &\quad + O_p\{h^5 + h^3(nh)^{-1}2\}. \end{aligned} \tag{A.11}$$

Define $\Omega_{pd}(x) = \text{cov}\{m_r(x, p), m_r(x, d)\}$, $\Omega_{pp}(x) = \text{var}\{m_r(x, p)\}$ and $\Omega_{dd}(x) = \text{var}\{m_r(x, d)\}$. Our goal is to find conditions under

which there is no cost in estimating these functions because of the nonparametric regression, that is, if we could actually observe $G_{rc_p}(x, p)$ and $G_{rc_d}(x, d)$. If we could observe these terms, then unbiased estimates would be

$$\widehat{\Omega}_{pd}(x, \text{ideal}) = (R-1)^{-1} \sum_{r=1}^R \mathcal{L}_r(x, p) \mathcal{L}_r(x, d), \tag{A.12}$$

$$\begin{aligned} \widehat{\Omega}_{pp}(x, \text{ideal}) &= (R-1)^{-1} \sum_{r=1}^R \mathcal{L}_r^2(x, p) - \{RC(C-1)\}^{-1} \\ &\quad \times \sum_{r=1}^R \sum_{c_p=1}^C \mathcal{V}_{rc_p}^2(x, p), \end{aligned} \tag{A.13}$$

and similarly for $\widehat{\Omega}_{dd}(x, \text{ideal})$. The obvious estimates are

$$\widehat{\Omega}_{pd}(x, h) = (R-1)^{-1} \sum_{r=1}^R \widehat{\mathcal{L}}_r(x, p, h) \widehat{\mathcal{L}}_r(x, d, h), \tag{A.14}$$

$$\begin{aligned} \widehat{\Omega}_{pp}(x, h) &= (R-1)^{-1} \sum_{r=1}^R \widehat{\mathcal{L}}_r^2(x, p, h) - \{RC(C-1)\}^{-1} \\ &\quad \times \sum_{r=1}^R \sum_{c_p=1}^C \widehat{\mathcal{V}}_{rc_p}^2(x, p, h), \end{aligned} \tag{A.15}$$

and similarly for $\widehat{\Omega}_{dd}(x, h)$.

Define $z(R, C, n, h) = (nC)h^{-1} + Rh^4$. If we replace $\widehat{G}_{rc_p}(x, p, h)$ and $\widehat{G}_{rc_d}(x, d, h)$ by their asymptotic expansions (A.10), then long, detailed, and tedious calculations show that if $nh \rightarrow \infty$, then

$$\begin{aligned} &R\{\widehat{\Omega}_{pd}(x, h) - \widehat{\Omega}_{pd}(x, \text{ideal})\}^2 \\ &= O_p\{z(R, C, n, h)\}, \end{aligned} \tag{A.16}$$

$$\begin{aligned} &R\{\widehat{\Omega}_{pd}(x, h) - \Omega_{pd}(x)\}^2 \\ &= O_p\{1 + h^2 + z(R, C, n, h)\}, \end{aligned} \tag{A.17}$$

$$\begin{aligned} &R\{\widehat{\Omega}_{pp}(x, h) - \widehat{\Omega}_{pp}(x, \text{ideal})\}^2 \\ &= O_p\{z(R, C, n, h) + R(nCh)^{-2}\}, \end{aligned} \tag{A.18}$$

$$\begin{aligned} &R\{\widehat{\Omega}_{pp}(x, h) - \Omega_{pp}(x)\}^2 \\ &= O_p\{1 + h^2 + z(R, C, n, h) + R(nCh)^{-2}\}. \end{aligned} \tag{A.19}$$

Consider, for example, (A.16). One checks this by making use of (A.11), writing each term in the difference $R^{1/2}\{\widehat{\Omega}_{pd}(x, h) - \widehat{\Omega}_{pd}(x, \text{ideal})\}$, and checking the rate of convergence to zero. For example, one of the terms is

$$\begin{aligned} &R^{-1/2} \sum_{r=1}^R \mathcal{L}_r(x, p) \{ \mathcal{S}_r(x, p, h) - \overline{\mathcal{S}}_*(x, p, h) \} \\ &= R^{-1/2} \sum_{r=1}^R \mathcal{L}_r(x, p) \mathcal{S}_r(x, p, h). \end{aligned} \tag{A.20}$$

This is easily seen to have mean zero and variance $O\{(nC)^{-1}\}$, and hence it contributes $O_p\{(nC)h^{-1}\}$ to (A.16).

Here are some of the consequences of these calculations:

1. By using (A.16) and (A.18), we see that there is no effect to the nonparametric regressions for estimating the correlation function if $nh \rightarrow \infty$, $Rh^4 \rightarrow 0$, and $R(nhC)^2 \rightarrow 0$. In our example, $R = 15$, $C \approx 20$, $n \approx 40$, and $h \approx .20$, so that $nh \approx 8$, $Rh^4 \approx .02$, and $R(nhC)^2 \approx .00$. These calculations suggest that there is essentially no cost in our example due to estimating the functions at the

crypt level nonparametrically, when interest focuses on the correlation function at the rat level. Thus, it is not at all surprising that in our example we observed that the kernel method is fairly insensitive to the bandwidth and that the kernel method gives results in the bootstrap very close to the parametric method.

2. It is possible, at least in principle to estimate the bandwidth that minimizes the mean squared error for the correlation function, because by (A.17), (A.19), and the delta method, this bandwidth minimizes a function of the form $a_1 + a_2 h^2 + a_3 R(nCh)^{-2} + a_4 (nCh)^{-1} + a_5 R h^4$. The variance part of the mean squared error can be estimated by resampling rats and then crypts within rats, and the squared bias can be estimated by using the method of Ruppert (1997). This is certainly interesting and deserves further study, but for the carcinogenesis example, it seems more than sufficient to work under conditions for which the nonparametric regression has no real impact on the estimated correlation function.

APPENDIX B: BASIC EXPANSION FOR LOCAL LINEAR REGRESSION

This section derives the asymptotic expansion (A.10) for local linear regression, an expansion that is of order high enough to allow the difficult calculations of Appendix A to be performed. For local linear regression in a sample of size n , we write $Y_i = m(X_i) + \epsilon_i$. Define $\beta_j = h^j m^{(j)}(x) j!$. We are interested in estimating the intercept $\beta_0 = m(x)$ at a given value x . If we define

$$G_{nh}(x) = n^{-1} \sum_{i=1}^n K_h(X_i - x) \{1, (X_i - x)\} h^T \{1, (X_i - x)\} h, \quad (B.1)$$

it is easily shown that analytically

$$\hat{m}(x) - m(x) = (1, 0) G_{nh}^{-1}(x) n^{-1} \sum_{i=1}^n K_h(X_i - x) \{1, (X_i - x)\} h^T \times \{\epsilon_i + m(X_i) - m(x) - m^{(1)}(x)(X_i - x)\}. \quad (B.2)$$

It follows from standard calculations that

$$\sum_{i=1}^n K_h(X_i - x) \{m(X_i) - m(x) - m^{(1)}(x)(X_i - x)\} = A_1 + A_2 + A_3 + O_p(h^4), \quad (B.3)$$

$$A_1 = (h^2) 2 m^{(2)}(x) f(x), \quad (B.4)$$

$$A_2 = (h^2) 2 m^{(2)}(x) \left[\sum_{i=1}^n K_h(X_i - x) \{(X_i - x)\} h^2 - f(x) \right], \quad (B.5)$$

$$A_3 = (h^3) 6 m^{(3)}(x) \sum_{i=1}^n K_h(X_i - x) \{(X_i - x)\} h^3. \quad (B.6)$$

It is easily seen that $E(A_3) = O(h^4)$ and that $\text{var}(A_3) = O\{h^6 (nh)^{-1}\}$, so that $A_3 = O_p\{h^4 + h^3 (nh)^{-1/2}\}$. Similarly, $A_2 = O_p\{h^4 + h^2 (nh)^{-1/2}\}$. By the structure of A_1 , A_2 , and A_3 , we thus see that there is a constant c_1 and a bounded random variable B_1 , independent of the ϵ 's, such that

$$\begin{aligned} n^{-1} \sum_{i=1}^n K_h(X_i - x) \{m(X_i) - m(x) - m^{(1)}(x)(X_i - x)\} \\ = (h^2) 2 m^{(2)}(x) f(x) + h^4 c_1 + h^2 (nh)^{-1/2} B_1 \\ + O_p\{h^5 + h^3 (nh)^{-1/2}\}. \end{aligned} \quad (B.7)$$

Throughout the appendix, we let B_j represent $O_p(1)$ random variables that are independent of the ϵ 's. By a similar calculation, it follows

that for a constant c_2 and a bounded random variable B_2 ,

$$\begin{aligned} n^{-1} \sum_{i=1}^n K_h(X_i - x) \{(X_i - x)\} h \\ \times \{m(X_i) - m(x) - m^{(1)}(x)(X_i - x)\} \\ = h^3 c_2 + h^2 (nh)^{-1/2} B_2 + O_p\{h^4 + h^3 (nh)^{-1/2}\}. \end{aligned} \quad (B.8)$$

Define $\kappa_j = n^{-1} \sum_{i=1}^n K_h(X_i - x) \{(X_i - x)\} h^j$. Then $(1, 0) G_{nh}^{-1}(x) = (\kappa_2, -\kappa_1) (\kappa_2 \kappa_0 - \kappa_1^2)^{-1}$. For constants d_1, \dots, d_6 , it can be shown that $\kappa_0 = f(x) + h^2 d_1 + h^4 d_2 + (nh)^{-1/2} B_3 + O_p(h^5)$, $\kappa_1 = h d_3 + h^3 d_4 + (nh)^{-1/2} B_4 + O_p(h^5)$, and $\kappa_2 = 1 + h^2 d_5 + h^4 d_6 (nh)^{-1/2} B_5 + O_p(h^5)$. It then follows that

$$\begin{aligned} (1, 0) G_{nh}^{-1}(x) n^{-1} \sum_{i=1}^n K_h(X_i - x) \{1, (X_i - x)\} h^T \\ \times \{m(X_i) - m(x) - m^{(1)}(x)(X_i - x)\} \\ = (h^2) 2 m^{(2)}(x) + h^4 c_3 + h^2 (nh)^{-1/2} B_6 \\ + O_p\{h^5 + h^3 (nh)^{-1/2}\}. \end{aligned} \quad (B.9)$$

Similarly,

$$\begin{aligned} (1, 0) G_{nh}^{-1}(x) n^{-1} \sum_{i=1}^n K_h(X_i - x) \{1, (X_i - x)\} h^T \epsilon_i \\ = \left\{ \kappa_0^{-1} + \frac{\kappa_1^2}{\kappa_0 (\kappa_2 \kappa_0 - \kappa_1^2)} \right\} n^{-1} \sum_{i=1}^n K_h(X_i - x) \epsilon_i \\ - \frac{\kappa_1}{\kappa_2 \kappa_0 - \kappa_1^2} n^{-1} \sum_{i=1}^n K_h(X_i - x) \{(X_i - x)\} h \epsilon_i. \end{aligned} \quad (B.10)$$

Because $\kappa_0^{-1} - 1) f(x) = h^2 d_7 + (nh)^{-1/2} B_7 + O_p(h^4)$, by the given expressions of κ_0 , κ_1 , and κ_2 , it follows that

$$\begin{aligned} (1, 0) G_{nh}^{-1}(x) n^{-1} \sum_{i=1}^n K_h(X_i - x) \{1, (X_i - x)\} h^T \epsilon_i \\ = \{nf(x)\}^{-1} \sum_{i=1}^n K_h(X_i - x) \epsilon_i \{1 + h^2 c_4 + (nh)^{-1/2} B_8\} \\ + h \{nf(x)\}^{-1} \sum_{i=1}^n K_h(X_i - x) \epsilon_i \{(X_i - x)\} h \\ \times \{hc_5 + h^2 c_6 + (nh)^{-1/2} B_9\} + O_p(h^4). \end{aligned} \quad (B.11)$$

Combining (B.2)–(B.11) yields (A.10).

[Received November 1999. Revised April 2001.]

REFERENCES

- Bang, H. O., Dyerberg, J., and Hjorne, N. (1976), "The Consumption of Food Consumed by Greenland Eskimos," *Acta Medicine Scandanavia*, 200, 69–73.
- Bjerve, S., and Doksum, K. (1993), "Correlation Curves: Measures of Association as Functions of Covariate Values," *Annals of Statistics*, 21, 890–902.
- Blot, W. J., Lanier, A., Fraumeni, J. F., and Bender, T. R. (1975), "Cancer Mortality Among Alaskan Natives, 1960–69," *Journal of the National Cancer Institute*, 55, 547–554.
- Boyle, P., Zaridze, D. G., and Smans, M. (1985), "Descriptive Epidemiology of Colorectal Cancer," *International Journal of Cancer*, 36, 9–18.
- Chang, W. C. L., Chapkin, R. S., and Lupton, J. R. (1997), "Predictive Value of Proliferation, Differentiation and Apoptosis as Intermediate Markers for Colon Tumorigenesis," *Carcinogenesis*, 18, 721–730.
- Chang, W. C. L., Lupton, J. R., Frolich, W., Schoeffler, G. L., Peterson, M. L., and Chen, X. Q. (1994), "A Very Low Intake of Fat Is Required to Decrease Colonic Cell Proliferation and Fecal Bile Acid Concentrations in Rats," *Journal of Nutrition*, 124, 181–187.

- Doksum, K., Blyth, S., Bradlow, E., Meng, X. L., and Zhao, H. (1994), "Correlation Curves as Local Measures of Variance Explained by Regression," *Journal of the American Statistical Association*, 89, 571–582.
- Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.
- Fisher, R. A. (1954), *Statistical Methods for Research Workers*, (12th ed.), New York: Hafner.
- Gasser, T., and Müller, H. G. (1979), "Kernel Estimation of Regression Functions," in *Smoothing Techniques for Curve Estimation*, eds. Th. Gasser and M. Rosenblatt, Berlin: Springer-Verlag, pp. 23–68.
- Haenszel, W., and Kurihara, M. (1968), "Studies of Japanese Migrants. I. Mortality from Cancer and Other Diseases Among Japanese in the United States," *Journal of the National Cancer Institute*, 40, 43–68.
- Hart, J. D. (1997), *Nonparametric Smoothing and Lack-of-Fit Tests*, New York: Springer-Verlag.
- Holt, P. R., Mokuolu, A. O., Distler, P., Liu, T., and Reddy, B. S. (1996), "Regional Distribution of Carcinogen-Induced Colonic Neoplasia in the Rat," *Nutrition and Cancer*, 25, 129–135.
- Hong, M. Y., Lupton, J. R., Morris, J. S., Wang, N., Carroll, R. J., Davidson, L. A., Elder, R. H., and Chapkin, R. S. (2000), "Dietary Fish Oil Reduces O⁶-methylguanine DNA Adduct Levels in the at Colon in Part by Increasing Apoptosis During Tumor Initiation," *Cancer Epidemiology, Biomarkers, and Prevention*, 9, 819–826.
- Jacobs, L. R., and Lupton, J. R. (1986), "Relationship Between Colonic Luminal pH, Cell Proliferation, and Colon Carcinogenesis in 1,2-Dimethylhydrazine Treated Rats Fed High Fiber Diets," *Cancer Research*, 46, 1727–1734.
- Leurgans, S. E., Moyeed, R. A., and Silverman, B. W. (1993), "Canonical Correlation Analysis when the Data are Curves," *Journal of the Royal Statistical Society, Ser. B*, 55, 725–740.
- Lipkin, M. (1974), "Proliferative Changes in the Colon," *Digestive Diseases*, 19, 1029–1032.
- Lipkin, M., Bell, B., and Sherlock, P. (1963), "Cell Proliferation Kinetics in the Gastrointestinal Tract of Man. I. Cell Renewal in Colon and Rectum," *Journal of Clinical Investigation*, 42, 767–776.
- Ramsay, J. O., and Silverman, B. W. (1997), *Functional Data Analysis*, New York: Springer-Verlag.
- Rogers, K. J., and Pegg, A. E. (1977), "Formation of O⁶-Methylguanine by Alkylation of Rat Liver, Colon and Kidney DNA Following Administration of 1,2-Dimethylhydrazine," *Cancer Research*, 37, 4082–4087.
- Ruppert, D. (1997), "Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation," *Journal of the American Statistical Association*, 92, 1049–1062.
- Swallow, W. H., and Searle, S. R. (1978), "Minimum Variance Quadratic Unbiased Estimation (MIVQUE) of Variance Components," *Technometrics*, 20, 265–272.
- Swenberg, J. A., Cooper, H. K., Bucheler, J., and Kleihues, P. (1979), "1,2-Dimethylhydrazine-Induced Methylation of DNA Bases in Various Rat Organ and the Effect of Pretreatment With Disulfiram," *Cancer Research*, 39, 465–467.
- Weisburger, J. (1991), "Causes, Relevant Mechanisms, and Prevention of Large Bowel Cancer," *Seminars in Oncology*, 18, 316–336.
- Zhang, J., and Lupton, J. R. (1994), "Dietary Fibers Stimulate Colonic Cell Proliferation by Different Mechanisms at Different Sites," *Nutrition and Cancer*, 22, 267–276.