Geographically Weighted Regression as a Statistical Model

Chris Brunsdon Stewart Fotheringham
Martin Charlton

October 6, 2000

Spatial Analysis Research Group
Department of Geography
University of Newcastle-upon-Tyne
Newcastle-Upon-Tyne UK
NE1 7RU

Abstract

Recent work on Geographically Weighted Regression (GWR) (Brunsdon, Fotheringham, and Charlton 1996) has provided a means of investigating spatial non-stationarity in linear regression models. However, the emphasis of much of this work has been exploratory. Despite this, GWR borrows from a well founded statistical methodology (Tibshirani and Hastie 1987; Staniswalis 1987a; Hastie and Tibshirani 1993) and may be used in a more formal modelling context. In particular, one may compare GWR models against other models using modern statistical inferential theories. Here, we demonstatrate how Akaike's Information Criterion (AIC) (Akaike 1973) may be used to decide whether GWR or ordinary regression provide the best model for a given geographical data set. We also demonstrate how the AIC may be used to choose the degree of smoothing used in GWR, and how basic GWR models may be compared to 'mixed' models in which some regression coefficients are fixed and others are non-stationary.

1 Introduction

In this short article we intend to do two things - firstly to show how GWR can be considered as a "proper" statistical model and secondly to consider how

different GWR models may be compared. One matter that we will single out for special consideration is the "effective number of parameters" or "effective degrees of freedom" in a GWR model. Here we consider these quantities in terms of the expected value of the residual sum of squares. This concept is important, as it plays a key rôle in defining the Akaike Information Criterion (AIC) (Akaike 1973), which we will use as a model comparison tool. Initially, we will only consider the basic GWR model (Brunsdon, Fotheringham, and Charlton 1996) rather than the spatially autocorrelated model variant described in Brunsdon, Fotheringham, and Charlton (1998), although we intend to consider autocorrelated models eventually.

2 GWR as a Statistical Model

Suppose we have a set of observations $\{x_{ij}\}$ for $i = 1...n$ cases and $j = 1...k$ explanatory variables, and a set of dependent variables $\{y_i\}$ for each case. This is a standard data set for a global regression model. Now suppose that in addition to this we have a set of location coordinates $\{(u_i, v_i)\}$ for each case. The underlying model for GWR is

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{k} x_{ij}\beta_j(u_i, v_i) + \epsilon_i \tag{1}$$

where $\{\beta_0(u, v)...\beta_k(u, v)\}$ are $k+1$ continuous functions of the location (u, v) in the geographical study area. The ϵ_i's are random error terms. In the basic GWR model we assume that these are independently normally distributed with mean zero and common variance σ^2. The aim of GWR is to obtain non-parametric estimates of these functions. A related technique, the *expansion method* (Casetti 1972) attempts to obtain parametric estimates. Both are special cases of the very general *varying coefficient model* of Hastie and Tibshirani (1993). More specifically, GWR attempts to obtain estimates of the functions using kernel-based methods.

The log-likelihood for any particular set of estimates of the functions may be written as

$$L(\beta_0(u, v)...\beta_k(u, v)|\mathbf{D}) = -\frac{\sigma^{-2}}{2}\sum_{i=1}^{n}\left(y_i - \beta_0(u_i, v_i) - \sum_{j=1}^{k} x_{ij}\beta_j(u_i, v_i)\right)^2 \tag{2}$$

where $\mathbf{D}$ is the union of the sets $\{x_{ij}\}$, $\{y_i\}$ and $\{(u_i, v_i)\}$. As with many situations involving non-parametric regression, choosing function estimates to maximise this expression is not very helpful. With the distribution assumptions for the error terms above, this maximum likelihood approach is

equivalent to choosing the functions using least squares. However, since the functions are arbitary we can simply choose them to obtain a residual sum of squares of zero, with an associated and rather unconvincing estimate of σ^2 of zero. For this reason a straightforward ML approach to calibrating equation 1 is not used.

3 Local Likelihood

A more useful way forward is to consider *local likelihood*. Rather than attempting to minimise equation 2 globally, we consider the problem of estimating $\{\beta_0(u,v)...\beta_k(u,v)\}$ on a pointwise basis. That is, given a specific point in geographical space (u_0, v_0) (which may or may not correspond to one of the observed $\{(u_i, v_i)\}$'s) we attempt to estimate $\{\beta_0(u_0,v_0)...\beta_k(u_0,v_0)\}$. We do this by assuming that if these functions are reasonably smooth, we can assume that a simple regression model

$$y_i = \gamma_0 + \sum_{j=1}^{k} x_{ij}\gamma_j + \epsilon_i \tag{3}$$

holds close to the point (u_0, v_0), where each γ_j is a constant valued approximation of the corresponding $\beta_j(u,v)$ in model 1. We can calibrate a model of this sort by considering observations close to (u_0, v_0). An obvious way to do this is to use weighted least squares, that is to choose $\{\gamma_0...\gamma_k\}$ to minimise

$$\sum_{i=1}^{n} w(d_{0i}) \left(y_i - \gamma_0 - \sum_{j=1}^{k} x_{ij}\gamma_j \right)^2 \tag{4}$$

where d_{0i} is the distance between the points (u_0, v_0) and (u_i, v_i). This gives us the standard GWR approach. We simply set $\hat{\beta}_j(u_0, v_0)$ as $\hat{\gamma}_j$ to obtain the familiar GWR estimates. At this stage it is worth noting that 4 may be multiplied by $-\sigma^{-2}$ and be considered as a *local log-likelihood* expression:

$$\mathrm{WL}(\gamma_0...\gamma_k|\mathbf{D}) = \sum_{i=1}^{n} w(d_{0i}) L(\gamma_0...\gamma_k|\mathbf{D}) \tag{5}$$

The properties of such estimators have been studied fairly comprehensively over the past decade or so. Typically[1] this is in the context where the weighting function is applied to the $\{x_{ij}\}$'s, but this does not have to be the case - see for example Hastie and Tibshirani (1993). In particular, Staniswalis

[1] In the early work of Joan Staniswalis, for example

(1987a) notes that if $w()$ is scaled to sum to unity (which it may be without loss of generality) then WL($\gamma_0 ... \gamma_k | \mathbf{D}$) is an empirical estimate of the expected log-likelihood (not the local log-likelihood) at the point of estimation. Further work by Staniswalis (1987b) shows that under certain conditions - which will apply for any bounded $\beta_j(u, v)$ functions with bounded first second and third derivatives - the γ_j's do provide pointwise consistent estimators for the $\beta_j(u_0, v_0)$'s. Furthermore, the distribution of the estimates for the γ_j's is asymptotically normal and asymptotically unbiased. Thus, the referee's statement that the "framework used ... is simply not applicable" is incorrect — it overlooks some recent very useful practical and theoretical work by a large number of statisticians, only a few of which I have cited here. For a more in-depth view of this body of work, see for example Bowman and Azzalini (1997).

Thus, GWR does provide a reasonable calibration technique for model 1. On a historical note, it must be admitted that GWR was first devised as an exploratory technique, and not as an explicit attempt to fit model 1, but with hindsight it now seems that the approach does have a more formal interpretation. What is interesting to note, however, is that although model 1 has non-stationary regression coefficients, σ^2, the variance of the error term, is a global constant. However, the local likelihood interpretation can also be used to generalise GWR to non-normal error models (for example Poisson or Binomial models), or to heteroskedastic or non-independent normal error models.

4 Inference

4.1 Inference and Local Likelihood

Here inference will be regarded more in terms of confidence intervals for estimated values, rather than significance testing. This simply reflects trends in the statistical community over the past few years – see Nester (1996) for a discussion of this. To establish (pointwise) confidence intervals for the regression coefficients we need to know the form for the asymptotic variance-covariance matrix. In a GWR context, this is given by inverting the local information matrix: the expression may be found by re-casting a result from Staniswalis (1987b):

$$I(\gamma_0 ... \gamma_k) = \text{outer}(E(\{\frac{\partial L(\gamma_0 ... \gamma_k)}{\partial \gamma_i} | u_0, v_0)\})) \tag{6}$$

Where outer() denotes a multiplicative outer product. Note that although the estimates of the $\beta_j(u_0, v_0)$ are the local ones for (u_0, v_0), the likelihood

function is the global one. Since we do not know the true values of these partial derivatives, we could use the fact that the local likelihood is an estimator of the expected global likelihood and 'plug in' the local likelihood estimates of the functions in the likelihood expression. In fact, although this is a general result which could be applied to a variety of models, there is a more direct approach for model 1. To see this, we note that for any pointwise model calibration, we may write in matrix form

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y} \tag{7}$$

where the matrix $\mathbf{X}$ and the vectors $\mathbf{y}$ and $\hat{\boldsymbol{\beta}}$ correspond to the x's y's and γ's used previously in this article, and $\mathbf{W}$ is the diagonal matrix of the local weights around (u_0, v_0). Thus, the vector of local coefficient estimates $\hat{\boldsymbol{\beta}}$ is a linear function of the observed dependent variable vector, $\mathbf{y}$. Thus, equation 7 could be simplified to $\hat{\boldsymbol{\beta}} = \mathbf{C}\mathbf{y}$, where $\mathbf{C} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$. But, in model 1 we have assumed that the y_i's are independently distributed with the same variance σ^2. Thus, $\mathrm{var}(\mathbf{y}) = \sigma^2\mathbf{I}$. Thus, the pointwise variance of the vector $\hat{\boldsymbol{\beta}}$ is just $\mathbf{C}\mathbf{C}'\sigma^2$. Thus, we can obtain pointwise confidence intervals for the surface estimates once we have a way of estimating σ^2. As it turns out, this has a lot to do with the "degrees of freedom" issue, and so discussion of this will be deferred until section 5.which we discuss in the next section.

5 Degrees of Freedom

In the non-parametric framework set out here, the concept of "number of parameters" or "degrees of freedom" seems meaningless. However, as is the case with many nonparametric regression problems, the related idea of "effective degrees of freedom" can be considered. In global linear regression models, the idea of degrees of freedom relates to the expected value of the *residual sum-of-squares* (RSS) of the model. In particular, for a global model with k linear parameters, $\mathrm{E(RSS)} = (n-k)\sigma^2$. Here we regard k as the degrees of freedom of the model, and so write

$$\mathrm{E(RSS)} = (n - \mathrm{D.F.})\sigma^2 \tag{8}$$

Note that this also gives the usual estimate for σ^2,

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n - \mathrm{D.F.}} \tag{9}$$

Now we consider the distribution for the RSS in the GWR situation. First we note that the *fitted* values for the y_i's, denoted by $\{\hat{y}_i\}$ can be expressed as

a matrix transform of the "raw" y_i's (Brunsdon, Fotheringham, and Charlton 1999). In matrix form we write this as

$$\hat{\mathbf{y}} = \mathbf{S}\mathbf{y} \tag{10}$$

for some n by n matrix $\mathbf{S}$. Thus, fitted residuals are just $(\mathbf{I} - \mathbf{S})\mathbf{y}$, and

$$RSS = \mathbf{y}'(\mathbf{I} - \mathbf{S})'(\mathbf{I} - \mathbf{s})\mathbf{y} \tag{11}$$

Following Cleveland (1979) and Tibshirani and Hastie (1987) we then note that

$$\mathrm{E(RSS)} = (n - [2\mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{S}'\mathbf{S})])\sigma^2 + \mathrm{E}(\mathbf{y})'(\mathbf{I} - \mathbf{S})'(\mathbf{I} - \mathbf{S})\mathrm{E}\mathbf{y} \tag{12}$$

The first term of this expression relates to the variance of the fitted values, and the second to the bias. However, if we assume that the bandwidth for the GWR is chosen so that bias is negligible - which is reasonable for a large sample, as the asymptotic results suggest - then we have the approximation

$$\mathrm{E(RSS)} = (n - [2\mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{S}'\mathbf{S})])\sigma^2 \tag{13}$$

which is analogous to equation 8. In this sense, the degrees of freedom is in fact $2\mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{S}'\mathbf{S})$. This result is often not an integer, but does in fact vary from k (as the bandwidth tends to infinity) to n (as the bandwidth tends to zero). In many cases, $\mathrm{tr}(\mathbf{S})$ is very close to $\mathrm{tr}(\mathbf{S}'\mathbf{S})$ so an approximate value for the degrees of freedom is $\mathrm{tr}(\mathbf{S})$.

Note that this then provides a method of estimating σ^2, as required in the previous section, by substituting the effective degrees of freedom into equation 9. This in turn may be used in conjunction with the methods in section 4.1, to obtain standard errors for the pointwise γ_i estimates. Note that the standard error estimates obtained in this way differ from those obtained in the original, exploratory formulation of GWR (Brunsdon, Fotheringham, and Charlton 1996), since they exploit the fact that σ^2 is spatially stationary.

6 Model Comparison

The previous sections have outlined a framework in which GWR models may be regarded as "real" statistical models, where certain inferential techniques such as estimation of confidence intervals may be applied. Another important inferential process is the *selection* of statistical models. An approach we suggest here is based on the Akaike Information Criterion (AIC) (Akaike 1973). In the following sections we will provide a short overview of the AIC, and argue that it may be used as a tool for model comparison in GWR.

7 An Overview of the AIC

The Akaike Information Criterion is best understood by first considering the *Kullback-Liebler information distance* (KLID) between two statistical distributions. If f and g are two continuous (possibly multivariate) probability distribution functions, then this quantity is defined by

$$\mathrm{I}(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx \tag{14}$$

This can be thought of as the *information loss* when approximating the distribution f with the distribution g.

Note that $\mathrm{I}(f,g)$ may be thought of as a *distance* between f and g — the lower the value it takes, the closer f is to g. It can also be shown (although it takes some effort) that

$$\begin{array}{rcl} \mathrm{I}(f,g) & \geq & 0 \\ \mathrm{I}(f,g) = 0 & \text{iff} & f \equiv g \end{array} \tag{15}$$

The relationships in (15) imply some form of distance measure, although distance here is not symmetric since $\mathrm{I}(f,g) \neq \mathrm{I}(g,f)$.

The measure is particularly useful when comparing two distributions, say g_1 and g_2 as approximators of a distribution f. Given the interpretation of $\mathbf{I}$ as a distance between functions, we need to find which of g_1 and g_2 is 'closest' to f. We are therefore interested in whether the statement

$$\int f(x) \log\left(\frac{f(x)}{g_1(x)}\right) dx < \int f(x) \log\left(\frac{f(x)}{g_2(x)}\right) dx \tag{16}$$

is true. If it is, then g_1 is the best approximator, otherwise g_2 is the better choice. Note that 16 can be written as

$$\int f(x) \left(\log(f(x)) - \log(g_1(x))\right) dx < \int f(x) \left(\log(f(x)) - \log(g_2(x))\right) dx \tag{17}$$

or, more simply

$$\int f(x) \log(g_1(x))\, dx > \int f(x) \log(g_2(x))\, dx\ . \tag{18}$$

If the true distribution of the variable x is $f(x)$, then 18 can be written in terms of expected values:

$$\mathrm{E}[\log(g_1(x))] > \mathrm{E}[\log(g_2(x))] \tag{19}$$

Equation 19 is useful if g_1 and g_2 are competing models of an unknown true distribution f. Since f is not known, theoretical expectations cannot be computed. However, if we have a set of observed x-values $\{x_1, x_2, \ldots, x_n\}$ the expectations can be estimated by sample means:

$$\mathrm{E}[\log(g_j(x))] \approx n^{-1} \sum_{i=1,n} \log(g_j(x_i)) \text{ for } j = 1, 2 \tag{20}$$

Thus, we have a method for choosing between two models g_1 and g_2 given a data set $\{x_1, x_2, \ldots, x_n\}$. Note that the method does not require the unknown true distribution f to be specified. Information relating to f is extracted from the data sample. The assymetry of the KLID can now be justified — the respective rôles of model and reality are not interchangeable in this theory.

There is still one issue which must be addressed. Here we assume that g_1 and g_2 are fully specified, but in practice each of these models would have a number of unspecified parameters which must also be estimated from $\{x_1, x_2, \ldots, x_n\}$. Denote these vectors of parameters by $\tilde{\mathbf{a}}_1$ and $\tilde{\mathbf{a}}_2$ respectively. For example, g_1 and g_2 could both be regression models with different predictor variables. Then $\tilde{\mathbf{a}}_1$ and $\tilde{\mathbf{a}}_2$ would be the regression coefficients for each model. In this case, one might expect the estimates in (20) to take the form

$$\mathrm{E}[\log(g_j(x))] \approx n^{-1} \sum_{i=1,n} \log(g_j(x_i|\tilde{\mathbf{a}}_j)) \tag{21}$$

where $\tilde{\mathbf{a}}_j$ is the maximum likelihood estimate of $\tilde{\mathbf{a}}_j$. Up to a constant of proportionality, the left hand side of expression 21 is just the log-likelihood of $\tilde{\mathbf{a}}_j$. However, the likelihood maximising procedure introduces bias into the estimate 21 — this should be apparent from the fact that this estimate is in fact dependent on the likelihood. In fact Akaike (1973) demonstrates that this bias is roughly equal to k_j, the dimension of $\tilde{\mathbf{a}}_j$. Correcting for this bias leads to the definition of the *Akaike Information Criterion* (AIC)

$$\mathrm{AIC} = -2\log(\mathcal{L}(\tilde{\mathbf{a}}_j|x_1 \ldots x_n)) + 2k_j \tag{22}$$

where $\mathcal{L}(\tilde{\mathbf{a}}_j|x_1 \ldots x_n)$ is the likelihood of $\tilde{\mathbf{a}}_j$ given the data sample $\{x_1, x_2, \ldots, x_n\}$. This gives a number which, up to multiplication by one constant and multiplication by another, gives an estimate of $\mathrm{I}(f, g_j)$. Thus, comparing the AIC for each model g_j gives a method for deciding which model is best: the smallest AIC corresponds to the smallest estimated value of $\mathrm{I}(f, g_j)$ and hence the 'closest' model to f, the true situation.

8 Use in Practice

Note that although the discussion above refers to the comparison of *two* models, the argument holds for any number. This gives a general methodology for comparing several competing models for the same data set $\{x_1, x_2, \ldots, x_n\}$:

1. Identify a set of l models which it is thought may apply to the data.
2. Calibrate these models using maximum likelihood.
3. Compute the AIC for each model.
4. Select the model with the smallest AIC.

This method has a number of advantages over the more conventional use of hypothesis tests for model selection. Firstly, models being compared do not have to be nested. For example, it is possible to compare two linear regression models involving entirely different predictor variables, or two non-linear regression models with entirely different functional forms. Secondly, there are no problems with multiple hypothesis tests, as it is quite reasonable to compare several AICs for a given data set. However, it would be wrong to give the impression that the model selection procedure is error-proof. Since the AICs are in fact sample estimates of the true information distances, they are subject to sampling variation. Thus, a degree of uncertainty surrounds any estimated AIC, and if two competing models have very close AIC values it is unclear which one is truly the closest to f.

9 A Refinement to the AIC

One refinement to the AIC is the *corrected* AIC, or AIC_c. Recall that k_j was used as an approximation of the bias in the estimate of $\text{E}[\log(g_j(x))]$ for the AIC. An improved, but more complex estimate of this bias is the expression

$$k_j \left(\frac{n}{n - k_j - 1} \right) . \tag{23}$$

Note that when there are much fewer parameters than observed data points, this is approximately equal to k_j, the original expression. However, when this is not the case, it is recommended that this expression is substituted for k_j in the expression for the AIC, giving the AIC_c. The latter may be easily computed from the former, using the relationship

$$\text{AIC}_c = \text{AIC} + \frac{2k_j(k_j + 1)}{n - k - j - 1} \tag{24}$$

10 Least Squares Problems AIC and GWR

In the case of least squares problems, such as ordinary linear regression, the likelihood model assumed is one of normally distributed error terms, so that a random variable x_i is assumed to be distributed as

$$x_i = \mu_i + \epsilon_i \text{ where } \epsilon_i \sim \mathrm{N}(0, \sigma^2) \tag{25}$$

where μ_i is the expected value of x_i, and is assumed to depend on a number of model parameters. Here, the AIC may be written as

$$\mathrm{AIC} = n \log(\hat{\sigma}^2) + 2k_j \tag{26}$$

where

$$\hat{\sigma}^2 = \frac{\sum \hat{\epsilon}_i^2}{n} \tag{27}$$

and the $\hat{\epsilon}_i$'s are the residuals after fitting the model. Care should be taken here regarding k_j - this is the total number of parameters in the model *including* σ. Thus, in a regression model with v variables, iSf there is an intercept term then $k_j = v + 2$. These ideas can also be applied to *local likelihood* situations, such as those set out in section 3. In situations where the fitted y-values can be found by pre-multiplying the observed y values by a matrix $\mathbf{S}$ the AIC can be reasonably estimated by the expression

$$AIC_c = 2n \log(\hat{\sigma}) + \frac{n + \mathrm{Tr}(\mathbf{S})}{n + 2 - \mathrm{Tr}(\mathbf{S})} \tag{28}$$

(Hurvich and Simonoff 1998). Since GWR falls into this general category of model, we may compute the AIC of a GWR model in this way. In particular, we may compare GWR models with different bandwidths - or different predictor variables. Additionally, we may compare GWR models with ordinary global regression models in order to determine which gives the better model fit.

11 Mixed GWR models and AIC

One interesting class of GWR models are *mixed* GWR models (Brunsdon, Fotheringham, and Charlton 1999), where some parameters are stationary and others vary geographically. If $\mathbf{S}^*$ is the S-matrix associated with the geographically varying parameters in the model, assuming the stationary parameters are known, then it can be shown that the relationship between $\hat{\mathbf{y}}$ and $\mathbf{y}$ for a mixed GWR model is given by

$$\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$$

where

$$S = X(X'(I - S^*)X)^{-1}X'(I - S^*) + I - W$$

Therefore mixed GWR models may also be expressed using an **S**-matrix equation, and so their AIC may also be defined using equation 28. This allows mixed GWR models to be compared with full GWR models. This provides a framework for further model comparison. For example, one could conmpare a model in which a specific parameter is allowed to vary spatially against one where it is fixed.

12 Conclusion

This article has attempted to demonstrate the statistical properties of the basic and mixed GWR models, and also to suggest ways for analysing more complex GWR-type models. Theoretical work by a number of statistical workers implies that GWR will provide consistent estimates of models in the form of equation 1. Of course, these will provide less efficient estimates than "global" regression models in the case when there is no spatial nonstationarity, but it should be noted that when stationarity is present no global model, with or without spatially autocorrelated errors, could ever provide a consistent estimate of the true model. In addition to this, the AIC provides a method for comparing competing GWR models. This method works on a realistic framework which, rather than testing for the absolute truth of a particular hypothesis, compares several models as approximations of reality. Its is hoped that these methods may be generalised further, to provide a flexible set of model assessment tools.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Csaki (Eds.), *2nd Symposium on Information Theory*, pp. 267–281. Budapest: Akademiai Kiado.

Bowman, A. and A. Azzalini (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford: Oxford University Press.

Brunsdon, C., A. S. Fotheringham, and M. Charlton (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis 28*, 281–289.

Brunsdon, C., A. S. Fotheringham, and M. Charlton (1998). Spatial nonstationarity and autoregressive models. *Environment and Planning A 30*, 957–973.

Brunsdon, C., A. S. Fotheringham, and M. Charlton (1999). Some notes on parametric significance tests for geographically weighted regression. *Journal of Regional Science 39*, 497–524.

Casetti, E. (1972). Generating models by the expansion method: Applications to geographic research. *Geographical Analysis 4*, 81–91.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association 74*, 829–836.

Hastie, T. J. and R. J. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society (B) 55*, 757–796.

Hurvich, C. M. and J. S. Simonoff (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B 60*, 271–293.

Nester, M. (1996). An applied statistician's creed. *Applied Statistics 45*, 401–410.

Staniswalis, J. G. (1987a). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association 84*, 276–283.

Staniswalis, J. G. (1987b). A weighted likelihood formulation for kernel estimators of a regression function with biomedical applications. Technical Report 5, Medical College of Virginia Department of Biostatistics, Virginia Commonwealth University.

Tibshirani, R. J. and T. J. Hastie (1987). Local likelihood estimation. *Journal of the American Statistical Association 82*, 559–567.