1

Introduction: Global versus Local Statistics

Statistical models are always simplifications, and even the most complicated model will be a pale imitation of reality. Given this fact, it might seem a futile effort to estimate statistical models, but George Box succinctly described the nature of statistical research: ‘All models are wrong, some are useful.’ Despite the fact that our models are always wrong, statistics provides us with considerable insight into the political, economic and sociological world that we inhabit. Statistical models become simplifications of reality because we must make assumptions about various aspects of reality. The practice of statistics is not shy about the need to make assumptions. A large part of statistical modeling is performing diagnostic tests to ensure that the assumptions of the model are satisfied. And in the social sciences, a great deal of time is devoted to checking assumptions about the nature of the error term: are the errors heteroskedastic or serially correlated? Social scientists are far more lax, however, when it comes to testing assumptions about the functional form of the model. In the social sciences, the linear functional (and usually additive) form reigns supreme, and researchers often do little to verify the linearity assumption. Much effort is devoted to specification and avoiding misspecification, but little is done to explore other functional forms when the incorrect functional form is in essence a specification error.

Semiparametric Regression for the Social Sciences Luke Keele

The reliance on linear functional forms is more widespread than many probably realize. For many analysts, the question of linearity is focused solely on the nature of the outcome variable in a statistical model. If the outcome variable is continuous, we can usually estimate a linear regression model using least squares. For discrete outcomes, analysts typically estimate generalized regression models such a logistic or Poisson regression. Researchers often believe that because they are estimating a logistic or Poisson regression model, they have abandoned the linear functional form. Often this is not the case, as the functional form for these models remains linear in an important way. The generalized linear model (GLM) notation developed by McCullagh and Nelder (1989) helps clarify the linearity assumption in models that many researchers think of as nonlinear.

In the GLM framework, the analyst makes three choices to specify the statistical model. First, the analyst chooses the stochastic component of the model by selecting a sampling distribution for the dependent variable. For example, we might choose the following sampling distribution for a continuous outcome:

$$Y_i \sim N(\mu_i, \sigma^2). \tag{1.1}$$

Here, the outcome Y_i follows a Normal sampling distribution with expected value μ_i and a constant variance of σ^2. This forms the stochastic component of the statistical model. Next, the analyst must define the systematic part of the model by choosing a set of predictor variables and a functional form. If we have data on k predictors, $\mathbf{X}$ is an $n \times k$ matrix containing k predictors for n observations, and η is an $n \times 1$ vector of linear predictions:

$$\eta = \mathbf{X}'\boldsymbol{\beta} \tag{1.2}$$

where $\boldsymbol{\beta}$ is a vector of parameters whose values are unknown and must be estimated. Both $\mathbf{X}'\boldsymbol{\beta}$ and η are interchangeably referred to as the linear predictor. The linear predictor forms the systematic component of the model. The systematic component need not have a linear functional form, but linearity is typically assumed. Finally, the analyst must choose a *link* function. The link can be as simple as the identity function:

$$\mu_i = \eta. \tag{1.3}$$

The link function defines the connection between the stochastic and systematic parts of the model. To make the notation more general, we can write the link function in the following form:

$$\mu_i = g(\eta_i), \tag{1.4}$$

where $g(\cdot)$ is a link function that must be monotonic and differentiable. When the stochastic component follows a Normal distribution and the identity function links the stochastic and systematic components, this notation describes a linear regression model. The GLM framework, however, generalizes beyond linear regression models. The stochastic component may come from any of the exponential family of distributions, and any link function that is monotonic and differentiable is acceptable.

As an example of this generality, consider the GLM notation for a model with a binary dependent variable. Let Y_i be a binary variable without outcomes 0/1, where 1 represents a success for each m_i trials and φ_i is the probability of a success for each trial. If this is true, we might assume that Y_i follows a Bernoulli distribution, which would imply the following stochastic component for the model:

$$Y_i \mid \varphi_i \sim B(m_i, \varphi_i). \tag{1.5}$$

The systematic component remains $\mathbf{X}'\boldsymbol{\beta}$, the linear predictor. We must now select a link function that will ensure that the predictions from the systematic component lie between 0 and 1. The logistic link function is a common choice

$$\varphi_i = \frac{1}{1 + e^{-\eta_i}}. \tag{1.6}$$

With the logit link function, no matter what value η_i takes the predicted value for Y_i, φ_i, will always be between zero and one.

What has been the purpose of recounting the GLM notation? The point of this exercise is to demonstrate that while the link function is a nonlinear transformation of the linear predictor, the systematic component of the model remains *linear*. Across the two examples, both the link function and stochastic component of the model differed, but in both cases we used the linear predictor $\mathbf{X}'\boldsymbol{\beta}$. The second model, a logistic regression, is thought of as a nonlinear model, but it has a linear functional form. Thus many of the models that analysts assume are nonlinear models retain a linearity assumption. There is nothing about the logistic regression model (or any other GLM) that precludes the possibility that the model is nonlinear in the variables. That is, instead of the effect of X on Y, which is summarized by the estimated β coefficient, being constant, that effect varies across the values of X. For example, one is as likely to use a quadratic term in a Poisson regression model as a linear regression model.

Why, then, is the assumption of linearity so widespread? And why is model checking for deviations from nonlinearity so rare? This question cannot be answered definitively. Quite possibly, the answer lies in the nature of social science theory. As Beck and Jackman (1998) note: 'few social scientific theories offer any guidance as to functional form whatsoever.' When researchers stipulate the predicted relationship between X and Y, they do not go beyond 'the relationship between X and Y is expected to be positive (or negative).' Given that most theory is silent as to the exact functional form, linearity has become a default, while other options are rarely explored. We might ask what are the consequences of ignoring nonlinearity?

1.1 The Consequences of Ignoring Nonlinearity

What are the consequences for using the wrong functional form? Simulation provides some useful insights into how misspecification of functional form can affect our inferences. First, we need some basic notation. We start with the linear and additive functional form for Y

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \tag{1.7}$$

The above equation is linear in the parameters in that we have specified that the mean of Y is a linear function of the variables X_1 and X_2. It is possible that the effect of X_2 on Y is nonlinear. If so, the effect of X_2 on Y will vary across the values of X_2. Such a model is said to be nonlinear in the variables but linear in the parameters. Estimation of models that are nonlinear in the variables presents few problems, which is less true of models that are nonlinear in the parameters. Consider a model of this type

$$Y = \beta_0 + \beta_1 X_1^{\beta_2}.$$

Here, the model is nonlinear in the parameters and estimation is less easy as nonlinear least squares is required. For a model that is nonlinear in the variables, we can still use least squares to estimate the model. What are the consequences of ignoring nonlinearity in the variables if it is present? For example, assume that the true data generating process is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2. \tag{1.8}$$

Suppose an analyst omits the quadratic term when estimating the model. The effect of omitting the nonlinear term from the right hand side of a regression model is easily captured with simulated data. To capture the above data generating process, we simulate X_1 as 500 draws from a uniform distribution on the interval 1 to 50. Y is a function of this X variable with an error term drawn from a Normal distribution with zero mean and constant variance and the β parameters are set to ones. What are the consequences of fitting a linear model without the quadratic term? Figure 1.1 contains a plot of the estimated regression line when only a linear term has been included on the right hand side of the estimated model and the true quadratic functional form.

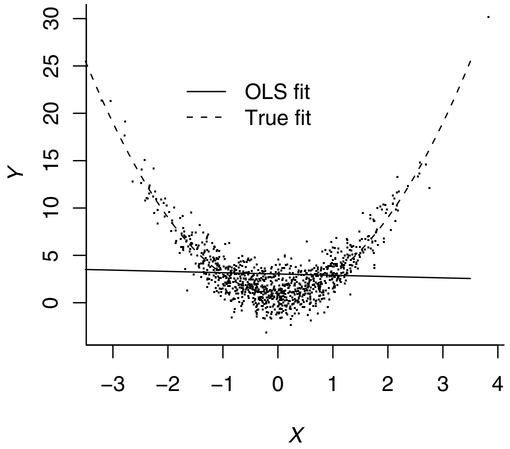

Figure 1.1 A linear fit and the true quadratic relationship with simulated data.

Here, the model with the incorrect functional form would lead one to conclude that X is unrelated to Y as the regression line is virtually flat, when in fact the curved line reflects the strong but nonlinear relationship between the two variables. This simulated example illustrates the misspecification that results from assuming linearity when the relationship between X and Y is actually nonlinear. In this particular example, the consequences are particularly severe, given that an analyst might conclude that there is no relationship between X and Y, when the two are strongly related. Moreover, as the GLM framework implies, models such as logistic or Poisson regression are equally prone to this misspecification from an incorrect functional form. If X^2 is in the true data generating process, no link function will correct the specification error. In these models, the systematic component of the model often remains linear and the failure to include a nonlinear term in the model will have equally deleterious effects on the model estimates. In fact, given the serious distortions such a misspecification can cause, it would be prudent to test that the effect of any continuous covariate on the right hand side of a model does not have a nonlinear effect. If we admit that defaulting to linearity might be a problematic method of data analysis, what alternatives are available? While there are a number of possibilities, analysts in the social sciences usually rely on power transformations to address nonlinearity.

1.2 Power Transformations

Power transformations are a simple and flexible means of estimating a nonlinear functional form. While a variety of power transformations are possible, most researchers restrict themselves to only one or two transformations. A common notation exists for all power transformations that is useful to outline since we will often use power transformations in the data analytic examples that follow. For a strictly positive variable, X, we can define the following set of power transformations:

$$X^{\lambda}. \tag{1.9}$$

Using different values for λ produces a wide variety of transformations. If λ is 2 or 3, the transformation is quadratic or cubic respectively, and if λ is $1/2$ or $1/3$, the transformation is either the squared or cubic root. By convention, a value of 0 for λ denotes a log transformation, and a value of 1 for λ corresponds to no transformation (Weisberg 2005).

Power transformations are often a reasonable method for modeling nonlinear functional forms. For example, it is well understood that older people vote more often than younger people (Nagler 1991). So as age increases, we expect that the probability of a person voting increases. But we do not expect the effect of age on voter turnout to be linear since once people reach a more advanced age, it often prevents them from voting. To capture such nonlinearity, most analysts rely on a quadratic power transformation. In this model, the analyst would square the age variable and then include both the untransformed and squared variable on the right hand side of the model. If the quadratic term is statistically significant at conventional levels ($p < 0.05$), the analyst concludes that the relationship between the two variables is nonlinear.

The process described above is often a reasonable way to proceed. Some care must be taken with the interpretation of the quadratic term as the standard error for the marginal effect must be calculated by the analyst, but the transformation method is easily to use. Moreover, the model is now a better representation of the theory, and power transformations avoid any misspecifiation due to an incorrect functional form. If the effect of age on voter turnout is truly quadratic, and we only include a linear term for age; we have misspecified the model and biased not only the estimate of age on voter turnout but all the other estimates in the model as well.[1]

Power transformations, however, have several serious limitations. First and foremost, power transformations are global and not local fits. With a global fit, one assumes that the statistical relationship between X and Y does not vary over the range of X. When an analyst estimates a linear relationship or uses a

[1] This is true since I assume the analyst has estimated a model such as a logit. In linear regression, misspecification only biases the coefficient estimates if the omitted variable is correlated with other variables, but for most models with nonlinear link functions, all the parameters will be biased if the model is misspecificed.

power transformation, the assumption is that the relationship between X and Y is exactly the same for all possible values of X. The analyst must be willing to assume a global relationship when he or she estimates linear functional forms or uses power transformations. Quite often the relationship between X and Y is local: the statistical relationship between two variables is often specific to local regions of X. The assumption of a global relationship often undermines power transformation.

For example, quadratic power transformations assume that the relationship between X and Y for all values of X is strictly quadratic regardless of whether this is true or not. When it is not, the power transformation can overcorrect the nonlinearity between X and Y. While there are a variety of power transformations, each assumes a global form of nonlinearity, which may or may not be correct. A variety of more complex nonlinear forms cannot be easily modeled with power transformations. Therefore, while power transformations can model some forms of nonlinearity, the global nature of power transformation fits often undermines them as a data analytic technique. In several of the data examples that follow, we see how power transformations often cannot adequately capture the nonlinear relationship in the data.

More seriously, the choice about which power transformation to use is often arbitrary. At best our theory might indicate that an effect is nonlinear, such as it does with age and voter turnout, but to have a theory that actually indicates that the effect is quadratic is unlikely. For example, how might one choose between a logarithmic or quadratic power transformation? One can estimate a model where the value of λ is estimated as the value that minimizes the sum of squared model errors. Unfortunately, this method often produces transformations that are uninterpretable (Berk 2006).

Moreover, the choice of which power transformation to use is not without consequences. For example, one might suspect a nonlinear relationship between X and Y and both a logarithmic and quadratic transformation seem like reasonable corrections. Does it matter which one the analyst uses? Another example with simulated data brings the problem into sharper focus. The simulated Y variable is a logarithmic function of a single X variable that is a drawn from a uniform distribution on the interval one to 50. The data generating process for Y is

$$Y_i = 0.50 \log(X_i) + \varepsilon \tag{1.10}$$

where ε is distributed IID Normal and the sample size is 500. We expect the estimate for β in the model to be 0.50. To better understand the consequences of an incorrect functional form, we estimate three different regression models with differing functional forms. The first model has the correct functional form with a logged X, the second model uses a squared X, and the last model has a linear functional form. The estimate from the model with logged X is 0.497 – very close to the true parameter value. The estimates from the models with the incorrect functional forms are 0.030 for the linear model and 0.049 for the quadratic, well under the true parameter value. The consequences of both ignoring the nonlinearity and using the wrong transformation are abundantly clear. For both the quadratic and linear fits, the parameter estimate is highly attenuated, so how might we choose the correct functional form? The model fit is not indicative of the wrong functional form as the R^2 values across the three models are: 0.168, 0.168, and 0.171. Perhaps a measure of nonnested model fit will provide some evidence of which model is to be preferred? The Akaike Information Criterion (AIC) is a measure of nonnested model fit estimated by many statistical software packages (Akaike 1973). Lower AIC values indicate better model fit to the data. The AIC values for the logarithmic, linear, and quadratic fits are respectively: 1409.35, 1409.47, and 1409.04, so the AIC points to the incorrect quadratic transformation. If the nonlinearity were quadratic instead of logarithmic, we could expect a similar quandary. This simulation demonstrates that the choice of power transformation is both important and difficult. The wrong power transformation can seriously affect the model estimates, but the analyst has few tools to discriminate between which transformation to use.

Moreover, since power transformations can only be used with strictly positive X variables, the analyst is often forced to arbitrarily add positive constants to the variable. Finally, power transformations are only effective when the ratio of the largest values of X to the smallest values is sufficiently large (Weisberg 2005).

In sum, power transformations are useful, but they have serious drawbacks and as such cannot be an analyst's only tool for modeling nonlinearity in regression models. What other alternatives are available? The most extreme option is to use one of several computationally intensive methods such as neural networks, support vector machines, projection pursuit, or tree based methods (Hastie, Tibshirani, and Friedman 2003). The results from such techniques, however, can be hard to interpret, and it can be difficult to know when the data are overfit. A less extreme alternative is to use nonparametric and semiparametric regression techniques.

1.3 Nonparametric and Semiparametric Techniques

Instead of assuming that we know the functional form for a regression model, a better alternative is to estimate the appropriate functional form from the data. In the absence of strong theory for the functional form, this is often the best way to proceed. To estimate the functional form from data, we must replace global estimates with *local* estimates. With global estimates, the analyst assumes a functional form for the model; with local estimates, the functional form is estimated from the data. The local estimators used in this text are referred to as nonparametric regression models or smoothers. Nonparametric regression allows one to estimate nonlinear fits between continuous variables with few assumptions about the functional form of the nonlinearity. Both *lowess* and splines are common nonparametric regressions models that rely on local estimates to estimate functional forms from data.

What is local estimation? With local estimation, the statistical dependency between two variables is described not with a single parameter such as a mean or a β coefficient, but with a series of local estimates. For local estimators, a estimate such as a mean or regression is estimated between Y and X for some restricted range of X and Y. This local estimate of the dependency between X and Y is repeated across the range of X and Y. This series of local estimates is then aggregated with a line drawing to summarize the relationship between the two variables. This resulting nonparametric estimate, which may be linear or quadratic, does not impose a particular functional form on the relationship between X and Y. Due to the local nature of the estimation process, nonparametric regression provides very flexible fits between X and Y. Local models such as nonparametric regression estimate the functional form between two variables while global models impose a functional form on the data.

Nonparametric regression models are useful for both the diagnosis of nonlinearity and modeling nonlinear relationships. A plot of the estimated residuals against the fitted values is a standard diagnostic for most statistical models. Trends in the variability of the residuals against the fitted values is, for example, a sign of heteroskedasticity. When examined in conjunction with nonparametric regression such plots are far more informative for the diagnosis of nonlinear functional forms. Any trend in the nonparametric regression estimate is a sign of various forms of misspecification including unmodeled nonlinearity. Figure 1.2 contains a plot of the residuals against the fitted values for the first example with simulated data. In that model, the true functional form was quadratic but we fit a linear model. The specification error is readily apparent in this plot. The plot also contains the estimate from a nonparametric regression model. The nonparametric regression estimate closely matches the quadratic functional form without the analyst having to assume he or she knows the true functional form. While such plots are useful diagnostics, a plot of this type can only tell us there is a problem. What is needed is a multivariate model that allows one to combine global and local estimates. The solution is the semiparametric regression model.

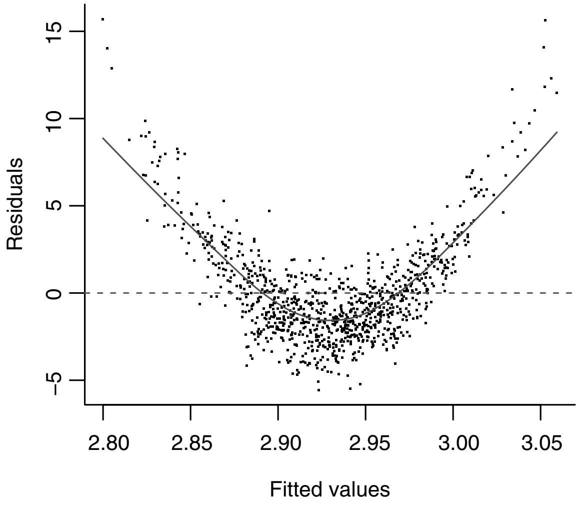

Figure 1.2 Residuals plotted against fitted values with a nonparametric regression estimate.

Semiparametric regression models are often referred to as either additive or generalized additive models (GAMs). Additive models and GAMs incorporate local estimation models like *lowess* and splines into standard linear models and GLMs. Given that the additive model is a special case of a GAM, throughout the text the term GAM will be used to refer to semiparametric regression models for both continuous and discrete outcome variables. These models allow the analyst to model some predictor variables with nonparametric regression, while other predictor variables are estimated in a standard fashion. A single continuous X variable suspected of having a nonlinear functional form may be modeled nonparametrically, while the rest of the model specification is estimated parametrically. Given that GAMs rely on nonparametric regression, the assumption of a global fit between X and Y is replaced with local fitting. While GAMs relax the assumption of a global fit, they do not dispense with the assumption of additive effects. The additivity assumption of GAMs makes the models easier to interpret than neural networks, support vector machines, projection pursuit, and tree based methods, but more flexible than a fully parametric model. The GAM framework is easily adapted to standard social science data analysis; GAMs work with a variety of dependent variables: interval, count, binary, ordinal, and time-to-event.

Most importantly, GAMs provide a framework for the diagnosis of nonlinearity. The simple linear and power transformation fits are nested within GAMs.

Therefore, the local estimates from a GAM can be tested against a linear, quadratic or any other transformation using either an F-test or a likelihood ratio test. If the semiparametric fit is superior to either a linear or quadratic fit, it should be used. If there is little difference between the local fit and the global fit, one can proceed with the global estimate. It is this ability to test for nonlinearity and if necessary model it that gives semiparametric regression models their power. In short, every analysis with continuous (or semicontinous) X variables requires the use of semiparametric regression model for diagnostic purposes.

At this point, one might ask if there are any objections to GAMs, and why are they not used more widely in the social sciences? A common objection is that GAMs are computationally intensive. In statistics, GAMs have often been classified as a computationally intensive technique, and semiparametric regression models are more computationally intensive than standard models since each nonparametric regression is the result of multiple local fits. The speed of modern computers has drastically reduced the time required for estimation. While a GAM may take longer than a standard parametric model, the difference in time is usually no more than a few seconds. Even with very large data sets, modern computers can now estimate most any GAM in less than 30 seconds.[2]

Some researchers object that the results from GAMs are more difficult to interpret than the results from standards models. This objection arises from the fact that GAMs do not produce a single parameter that summarizes the relationship between X and Y, but instead produces a plot of the estimated relationship between the two variables. Given that the estimation process is local as opposed to global, it is impossible to use a single parameter to describe the statistical relationship. While the absence of parameters does not allow for precise interpretation, a plot often does a far better job of summarizing a statistical relationship. With a plot, the scale of both variables should be obvious, and the eye can immediately discern the strength of the relationship along with the level of statistical precision so long as confidence bands are included.

It is often charged that GAMs overfit the data and produce highly nonlinear estimates that analysts are prone to overinterpret. It is true that an analyst may choose to undersmooth the nonparametric estimate, which can produce highly idiosyncratic and nonlinear fits between variables. Such overfitting, however, is the result of poor statistical technique; any method can be abused, and GAMs are no exceptions. Moreover, newer nonparametric regression models rely on penalized estimation which makes it more difficult to overfit the data.

[2] It should be noted that with very large data sets, those larger than $10,000$ cases, a GAM can take much longer to estimate than a standard model

Finally, a caveat. GAMs are not a panacea. They do not make up for poor theory or garbage can specifications. They cannot provide estimates of causal effects for observational data. The GAM is another model that should be in the toolkit of anyone that analyzes social science data. Like any statistical model, they must be used with care. Used correctly, GAM are a powerful tool for modeling nonlinear relationships. Used incorrectly, they can provide nonsensical results just as any model can.

1.4 Outline of the Text

Generalized additive models are inextricably linked to nonparametric regression, and thus one cannot understand semiparametric regression or competently estimate these models without a solid grasp of nonparametric regression. Consequently, Chapters 2–4 are devoted to various types of nonparametric smoothing techniques. These smoothers have a number of uses in their own right including exploring trends in scatterplots. As such, they are often used as a preliminary diagnostic to semiparametic regression. These chapters can be skipped if the reader is already familiar with nonparametric regression models.

The next two chapters explore how these nonparametric techniques can be married to the standard models that are used in the majority of quantitative social science research. While technical material is presented in these chapters, examples with real data from the social sciences are a point of emphasis. The next chapter explores the use of nonparametric smoothers in conjunction with mixed models, Bayesian estimation, and the estimation of propensity scores. The final chapter outlines the bootstrap – a nonparametric resampling method for estimating statistical precision – with special applications to semiparametric regression models. While bootstrapping is useful in its own right, it is critical for use with some types of semiparametric regression models. The appendix provides coverage of the software used for the analyses in the book.

The pedagogical philosophy behind this book is that new techniques are best learned through example. As such, each technique is explored with real data. In fact, the best argument for GAMs are examples of GAMs used with social science data. The majority of the examples are from published sources to demonstrate how semiparametric regression can alter the inferences made in research. All of the data sets and the computer code used in the analyses are available on the web at http://www.wiley.com/go/keele_semiparametric. There are also exercises at the end of each chapter. Many of these exercises require data analysis with applications from published research, so researchers can learn to use these models in their own work.