

4 Generalized Additive Model

died in childbirth, and the following year, Gauss married her best friend, Minna Waldeck, although it appears that this marriage was not a happy one.

Gauss's astronomical work resulted in inventions such as the heliotrope, an instrument for accurate direction-finding by means of reflected sunlight. Gauss also experimented with magnetometers; photometers; and, some 5 years before Samuel Morse, the telegraph. He died peacefully at Göttingen in the early morning of February 23, 1855.

—Graham Upton

See also Normal Curve

Further Reading

Dunnington, G. W. (2004). *Carl Friedrich Gauss: Titan of science*. Washington, DC: Mathematical Association of America.

Carl Friedrich Gauss article: http://en.wikipedia.org/wiki/Carl_Friedrich_Gauss

GENERALIZED ADDITIVE MODEL

Estimating the linear model $Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + e_i$ is at the core of many of the statistics conducted today. If you allow the individual X variables to be products of themselves and other variables, the linear model is appropriate for factorial ANOVAs and polynomial regressions, as well as estimating the mean, t tests, and so on. The flexibility of the linear model has led authors of some textbooks and software to call this the *general linear model*. I try to avoid this phrase because it can be confused with the *generalized linear model*, or GLM. The GLM is an important extension that allows researchers to analyze efficiently models where the responses are proportions and counts, as well as other situations. More on this later.

The main focus of this entry is extending the linear model into an additive model. In the linear model, each X variable is multiplied by a scalar, the β value. This is what makes it a linear model, but this restricts

the relationship between X and Y (conditioned on all the other X s). With additive models, the β values are replaced by usually fairly simple (in terms of degrees of freedom) functions of the X variables. The model can be rewritten as: $Y_i = \alpha + f_1(X_{1i}) + \dots + f_k(X_{ki}) + e_i$. The functions are usually assumed to be splines with a small number of knots. More complex functions can be used, but this may cause the model to overfit the observed data and thus not generalize well to new data sets. The typical graphical output shows the functions and the numeric output shows the fit of the linear and nonlinear components. The choice of functions, which often comes down to the type and complexity of the splines, is critical.

To illustrate this procedure, data from 534 respondents on hourly wages and several covariates (experience in years, gender, and education in years) are considered. One outlier with an hourly wage of \$44 ($z = 6.9$) is removed, but the data remain skewed ($1.28, se = 0.11$). Logging these data removes the skew ($0.05, se = 0.11$), so a fairly common approach is to use the logged values as the response variable and assume that the residuals are normally distributed. Suppose the researchers' main interests are with the experience variable, and whether it steadily increases or whether it increases rapidly until some point and then increases but less rapidly. For argument's sake, let us assume that the increases are both linear with the logged wages. The researchers accept that wages increase with education and believe that the relationship is nonlinear, and so they allow this relationship to be modeled with a smoothing spline. Because the variable *female* is binary, only a single parameter is needed to measure the difference in earnings between males and females. Although categorical variables can be included within generalized additive models (GAMs), the purpose of GAMs is to examine the relationships between quantitative variables and the response variable. The first model is

$$\ln \text{wages}_i = \beta_0 + \beta_1 \text{female}_i + \beta_2 \text{Exper}_i + f_1(\text{Educ}_i) + e_i.$$

This is like a normal multiple linear regression for the variables *female* and *Exper*; the model fits both as

conditionally linear with the log of wages, but the relationship for education is allowed to be curved. This was fit with the GAM package for S-Plus with the default spline (smoothing spline with $df = 4$). The residual deviance is 103.74. There is a positive linear relationship between experience and the log of wages and a positive curved relationship between education and the log of wages. The effect for *female* is negative, meaning that after controlling for experience and education, females earn less than their male counterparts. The top three plots in Figure 1 show this model. With GAMs, people usually rely on plots to interpret the models and compare the deviance values, or they use methods of cross-validation to decide how complex the model (including the complexities of the individual f s) should be. Here, the deviance values will be compared.

The bottom three plots of Figure 1 show the model:

$$\ln wages_i = \beta_0 + \beta_1 female_i + f_1(Exper_i) + f_2(Educ_i) + e_i.$$

f_1 has been set to a piecewise linear model, so two lines connected at a knot determined by the algorithm. The residual deviance drops to 97.55, which is statistically significant ($\chi^2(1) = 6.20, p = .01$). The package used (GAM) allows different types of curves (including loess) to be included in the model, although the efficiency of the algorithm works best if the same type is used. What is clear from this model is that a single linear term of experience is not sufficient to account for these data. If we allow the relationship between experience and the log of wages to be a $df = 4$ spline,

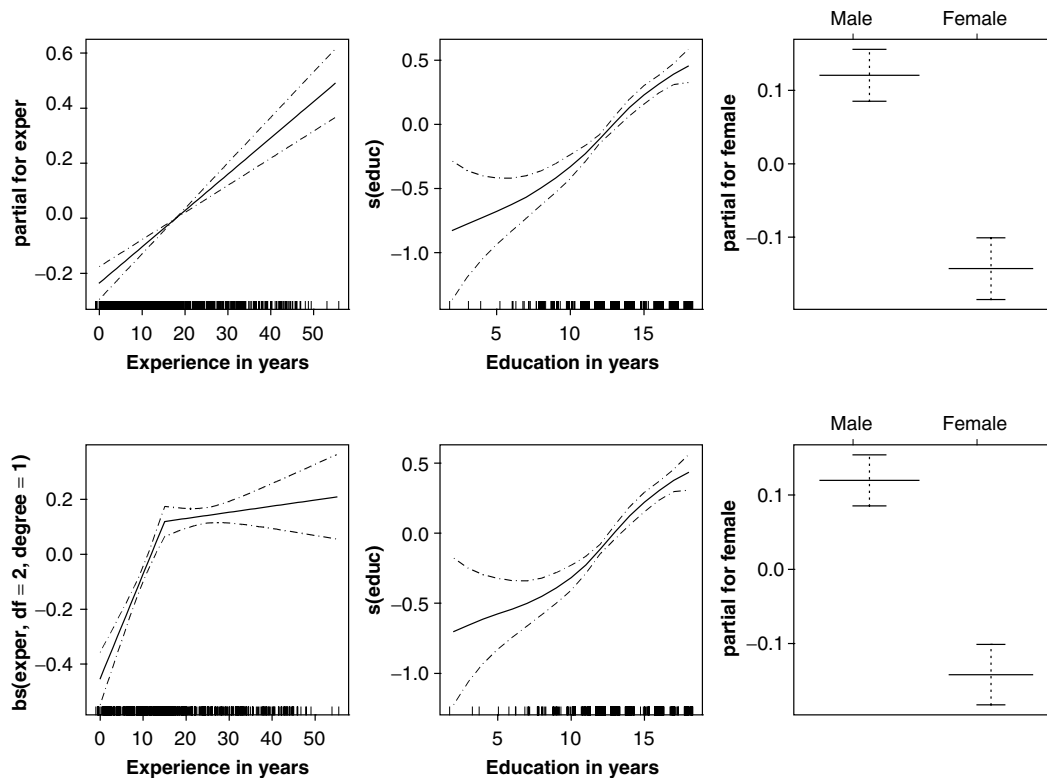


Figure 1 Plots for the Predictor Variables for Two Models Predicting the Log of Wages Using Experience, Education, and Gender

Notes: The top row shows the GAM where experience is a linear predictor. The bottom row has experience as a piecewise linear relationship with single knot estimated at approximately 15 years' experience. The dashed lines are for standard errors and rugplots are used to show the univariate distributions of experience and education.

6 Generalized Additive Model

the residual deviance drops only to 96.85. Although this is not an improvement in the fit of the model, the researcher might still opt for this unless he or she has a theoretical explanation for the sudden change in slope for the previous model.

Just as the generalized linear model allows researchers to model proportions and counts efficiently using the basic concepts of the linear model, the generalized additive model also allows this. The user chooses an error distribution from the exponential family and an associated link function, denoted $g(\cdot)$. Popular distributions are the normal distribution (associated link is the identity function), the binomial distribution (associated link is the logit function), and the Poisson distribution (associated link is the natural log, or \ln , function). In fact, the above example could have been modeled with the log link and assuming Poisson error, and this approach shows that a smooth spline does fit experience better than the piecewise linear model. Generalizing the additive model can be done in the same way as generalizing the linear model. If μ_i is the value of the response variable, then $E(g(\mu_i)) = \eta_i$, where η_i is an additive model of the form $\sum f_k(X_{k_i})$ where one of the X variables is a constant so that an intercept is included. Including the error term, this is $g(\mu_i) = \eta_i + e_i$, where the e_i are assumed to follow some distribution. The wages example could be fit with the following GAM:

$$\begin{aligned} \ln(\mu_i) &= \eta_i + e_i \\ \eta_i &= \alpha + f_1(\text{Exper}_i) + f_2(\text{Educ}_i) + \beta_1 \text{female}_i \\ e_i &\sim \text{Poisson}(\mu_i). \end{aligned}$$

To illustrate a logistic additive model, data inspired by truth and lie detection using criteria-based content analysis (CBCA) will be used. This is a method used in several countries to try to determine whether a child is telling the truth or a lie when questioned, usually in connection with cases of child sexual abuse. There are 19 criteria, and each statement can be given a 0, 1, or 2. These are summed so that each person can get a score from 0 to 38, with high scores indicating more truthfulness. One problem with this procedure is that people with more linguistic skills tend to have higher scores than people with fewer linguistic skills.

Because of this, there is assumed to be a complex relationship between age, CBCA score, and truth.

Suppose there are 1,000 statements from people who are 3 to 22 years old. All of the statements have CBCA scores, and it is known whether or not they are truthful. For these data, age and truth were created independently, so age, on its own, does not predict truth ($t(998) = 1.04, p = .30$). Three GAMs were estimated. The first has just CBCA to predict truth. This uses the logit link function and assumes binomial variation. The default smoothing function for the gam package is used, and the result (in the upper left hand corner of Figure 2) shows that the probability of truth increases with CBCA scores. The deviation from linear is statistically insignificant ($\chi^2(3) = 44.18, p < .01$). The residual deviance of this model is 1251.65.

The next model has $\eta_i = \alpha_i + f_1(\text{CBCA}_i) + f_2(\text{age}_i)$, where both f_1 and f_2 are $df = 4$ smoothing splines, and the resulting curves are shown in the second row of Figure 2. CBCA is again positively related to truth. However, age is negatively related (because it is conditional on CBCA). Both curves show marked nonlinearity. The residual deviance is 1209.60, which is a large improvement in fit on the previous model ($\chi^2(3.82) = 42.05, p < .001$). The final row in Figure 2 shows the GAM, which includes an interaction term. The graph of the interaction effect (new residual deviance 1201.51, change $\chi^2(3.96) = 28.09, p = .09$) shows that the predictive value of the CBCA scores increases with age. To examine this interaction further, values of the age variable were placed into four approximately equal sized bins, and separate GAMs were run on each. The resulting ogives for these are shown in Figure 3. Simple monotonic curves appear to represent the relationship between CBCA and truthfulness for the older people, but not for the younger groups. It appears either that the relationship between CBCA and truthfulness is different for the age groups, or that the CBCA is only diagnostic of truthfulness above about 16 or 17 points (which the older people do not score below for either true or false statements). Given that these are data created for illustration, it is not appropriate to speculate further about either explanation.

GAMs are useful generalizations of the basic regression models. Like GLMs, they allow different link functions and distributions that are appropriate

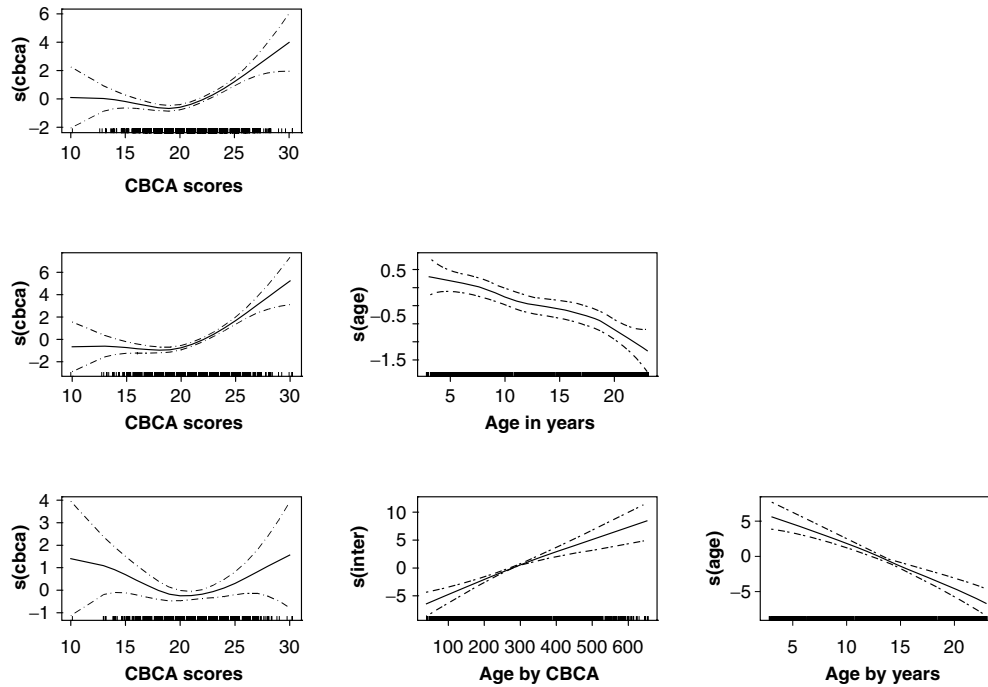


Figure 2 Plots for the Predictor Variables for Three Models Predicting the Probability of a Statement Being Truthful Based on CBCA Score and Age

Notes: The first (upper left hand corner) uses just CBCA score. The second model (row two) also uses age. The third model (row three) also includes the interaction. The dashed lines are for standard errors and rugplots are used to show the univariate distributions.

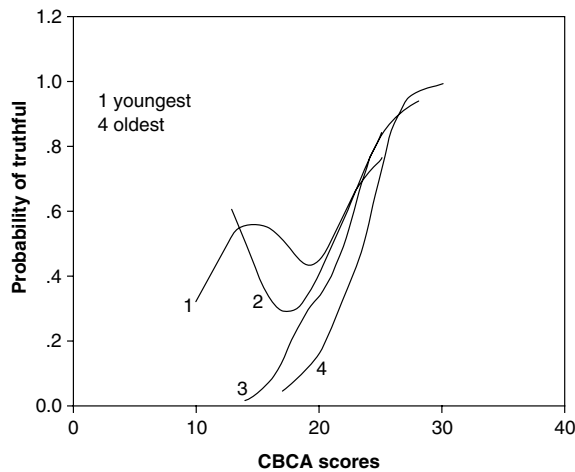


Figure 3 Individual GAMs for Four Different Age Groups

Notes: 1 = 3.0 to 7.9 years, 2 = 7.9 to 13.0 years, 3 = 13.0 years to 18.1 years, and 4 = 18.1 years and higher. There are relatively smooth monotonic curves for the two older age groups. However, the curves for the younger age groups appear more complex, particularly for low CBCA values.

for a large amount of data collected in science. Furthermore, the additive components allow an extremely flexible approach to data modeling. There are several extensions to GAMs not discussed here, such as model selection and regularization techniques, multilevel GAMs, and different types of estimation. Current software allows many different types of curves to be fit within GAMs. Two were illustrative, a theory-driven example where a linear model was compared with a piecewise linear model, and a data-driven example that included an interaction. As algorithms and software advance, these models should become more flexible and more widely used.

—Daniel B. Wright

See also Ogive; Smoothing

Further Reading

Berndt, E. R. (1991). *The practice of econometrics*. New York: Addison-Wesley.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. London: Chapman & Hall.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman and Hall.

Vrij, A. (2005). Criteria-based content analysis—A qualitative review of the first 37 studies. *Psychology, Public Policy, & Law*, 11, 3–41.

GENERALIZED METHOD OF MOMENTS

When information on a set of parameters is given in the form of moments (expectations), equations containing this information are called the *moment conditions*. For example, if $y_i = x_i'\theta + u_i$ is considered and the statistician knows a priori that x_i and u_i are uncorrelated, then the corresponding moment conditions are $E x_i(y_i - x_i'\theta) = 0$. Alternatively, if it is believed that z_i and u_i are uncorrelated for some random variables z_i , then the moment conditions would be $E z_i(y_i - x_i'\theta) = 0$. In the above examples, the functions $x_i(y_i - x_i'\theta)$ and $z_i(y_i - x_i'\theta)$, whose expectations are set to zero, are called the *moment functions*. In general, for some functions $g(X_i, \theta)$ of random variables X_i and unknown parameter vector q , the moment conditions are expressed as $Eg(X_i, \theta) = 0$.

Identification and Overidentification

For a given set of moment functions $g(X_i, \theta)$, the true parameter sets the expected moment functions to zero by definition. When $Eg(X_i, \theta) = 0$ at only the true parameter, we say that the true parameter vector is *identified* by the moment conditions. A necessary condition for the identification of the true parameter is that the number of moment conditions should be at least as large as the number of parameters. When the number of moment conditions is exactly equal to the number of parameters (and when the true parameter is identified), we say that the true parameter is *exactly identified*. On the other hand, if there are more moment conditions than necessary, we say that the true parameter is *overidentified*.

Generalized Method of Moments

When there is a set of moment conditions that exactly identifies a parameter vector, *method of moments estimation* is widely used. As the true parameter sets the population moments to zero, the method of moments estimator sets the sample moments to zero. More precisely, when the true parameter is exactly identified by $Eg(X_i, \theta) = 0$, the method of moments estimator $\hat{\theta}$ satisfies $T^{-1}\sum_{i=1}^T g(X_i, \hat{\theta}) = 0$.

If the true parameter is overidentified, that is, if there are more moment conditions than are necessary to identify q , then it is usually impossible to set the sample moment vector to zero (because there are more equations than parameters). The *generalized method of moments* (GMM) was introduced by Lars Peter Hansen in 1982 in order to handle this case. Let $\bar{g}(\theta) = T^{-1}\sum_{i=1}^T g(X_i, \theta)$ for notational simplicity. Instead of setting the sample moment functions simultaneously to zero (which is usually impossible), Hansen proposed to minimize the quadratic distance of the sample moment vector from zero, that is, to minimize $\bar{g}(\theta)'\bar{g}(\theta)$ with respect to q over the parameter space. The minimizer is called the *generalized method of moments (GMM) estimator*.

The GMM estimator is consistent and asymptotically normal. In addition, the GMM procedure contains method of moments estimation as a special case. The method of moments estimator sets $\bar{g}(\hat{\theta}) = 0$, in which case the criterion function $\bar{g}(\theta)'\bar{g}(\theta)$ attains the minimal value zero at $\theta = \hat{\theta}$.

Weighted GMM and Optimal GMM

A symmetric and positive definite constant matrix W can be used in the criterion function to form a weighted criterion function $\bar{g}(\theta)'W\bar{g}(\theta)$, whose minimizer is called the *weighted GMM estimator* using the matrix W as weights. Because any symmetric and positive definite matrix can be decomposed into $A'A$ for some nonsingular matrix A (e.g., by a Cholesky decomposition), we observe that any weighted criterion function can be regarded as the (unweighted) quadratic distance of the transformed sample moment vector $A\bar{g}(\theta) = T^{-1}\sum_{i=1}^T Ag(X_i, \theta)$ from