A TEST FOR LINEAR VS CONVEX REGRESSION FUNCTION USING SHAPE-RESTRICTED REGRESSION

by

Mary C. Meyer

Technical Report No. 2001-20
August 2001

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305

A TEST FOR LINEAR VS CONVEX REGRESSION FUNCTION USING SHAPE-RESTRICTED REGRESSION

by

Mary C. Meyer
Department of Statistics
The University of Georgia

Technical Report No. 2001-20
August 2001

This research was supported in part by the National Science Foundation
grants DMS 9631278

Department of Statistics
Sequoia Hall
STANFORD UNIVERSITY
Stanford, California 94305

http://www-stat.stanford.edu

A Test for Linear vs. Convex Regression Function using Shape-Restricted Regression

Mary C. Meyer
The University of Georgia
Athens, GA 30602
mmeyer@stat.uga.edu

August 17, 2001

Summary

A test for the appropriateness of the simple linear regression model is presented. The null hypothesis is that the underlying regression function is indeed a line, and the alternative is that it is convex. An exact distribution for a likelihood ratio test statistic is derived. This is a mixture of beta densities, with the mixing distribution calculated from relative volumes of polyhedral convex cones determined by the convex shape restriction. A table is given for the critical values of the test statistic, along with a URL for the web page containing the fortran code and documentation. Simulations show that for smaller sample sizes or large model variance, the power of the test is favorable compared with the usual F-test against a quadratic model, even when the underlying function is actually a parabola.

Key Words: convex regression, effective dimension, likelihood ratio test

1 Statement of the Problem

Consider the problem of fitting a function to a scatterplot of data (x_i, y_i), for $i = 1, \ldots, n$, using the model

$$y_i = f(x_i) + \sigma\epsilon_i, \tag{1}$$

where the errors ϵ_i are assumed to be independent standard normal. The simplest relationship between the response variable y and the predictor variable x is a straight line. After obtaining a simple linear regression estimate of f, a residual analysis is standard practice to check the assumptions. A pattern in the residual plot might indicate some curvature in the relationship between the variables. Often a practitioner will check for curvature in the regression function by fitting a

parabola as well as a line, using an F-test to see if the parabola explains significantly more of the variation in y than the line. In this paper a more general alternative hypothesis is proposed:

$H_0 : f(x) = a + bx$; vs.

$H_1 : f(x) \in \mathcal{F}$,

where $\mathcal{F}$ is the class of convex functions. A likelihood ratio test statistic is shown to have a mixture of Betas density. The convex regression estimator is described in the next section, and the test statistic is developed in Section 3. Simulations results are in Section 4 with discussion in the Section 5.

2 Convex Regression

The ordinary least-squares regression estimator is the projection of the data vector y onto a smaller dimensional linear subspace of $\mathcal{R}^n$, whereas the convex regression estimator can be obtained through the projection of y onto an $n-2$ dimensional polyhedral convex cone in $\mathcal{R}^n$, where the dimension of a cone is understood to be the dimension of the smallest linear subspace containing the cone. The two estimators share some properties; for example, the residual vector is orthogonal to the projection, but there are important differences. The notion of "error degrees of freedom" for the cone projection is not so straight-forward as for the projection onto a linear space.

The constraint set C over which we maximize the likelihood or minimize the sum of squared residuals is constructed as follows. If we consider piecewise linear approximations to the regression function f, with knots at the x-values, the shape restrictions can be written as a set of linear inequality constraints. Suppose the x-values are sorted and distinct, and let $\theta_j = f(x_j)$, $j = 1, \ldots, n$. The convex requirement can be written as a set of linear inequality constraints. The piecewise linear fit is convex at x_2 if the slope is nondecreasing, i.e.,

$$\frac{\theta_2 - \theta_1}{x_2 - x_1} \leq \frac{\theta_3 - \theta_2}{x_3 - x_2};$$

or rather, $\theta_1(x_3 - x_2) - \theta_2(x_3 - x_1) + \theta_3(x_2 - x_1) \geq 0$. This inequality defines a half-space in $\mathcal{R}^n$. The other inequalities, constraining the fit to be convex at x_3 through x_{n-1}, also define half-spaces, so that the problem is to find θ to minimize the sum of squared errors $\sum_{i=1}^n (y_i - \theta_i)^2$ over the intersection of these $n-2$ half-spaces. This intersection is a closed convex polyhedral set, and can be expressed as $C = \{\theta : A\theta \geq 0\}$ for an $(n-2) \times n$ constraint matrix A. The least-squares estimator $\hat{\theta}$ is called the projection of the data vector y onto C, i.e., $\hat{\theta}$ is the point in C with smallest Euclidean distance from y.

For example, if $n = 6$ and the x-coordinates are equally spaced, the convex constraints are

summarized by the constraint matrix

$$A = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \end{pmatrix}. \tag{2}$$

The problem of finding the least-squares estimator $\hat{\theta}$ over C is a quadratic programming problem. There is no known closed-form solution, but $\hat{\theta}$ can be found using the mixed primal-dual bases algorithm of Fraser and Massam (1989) or the hinge algorithm of Meyer (1999a).

Let V be the space spanned by the vectors $v^1 = (1, \ldots, 1)'$ and $v^2 = (x_1, \ldots, x_n)'$, and let $V^\perp$ be the $n-2$ dimensional linear subspace of $\mathcal{R}^n$ that is orthogonal to V. Let Ω be the set $C \cap V^\perp$; this Ω is called the "constraint cone" as distinguished from the constraint set C. The following results from convex analysis describing the constraint cone Ω and its polar cone Ω_0 will be used in the development of the test statistic.

Define vectors γ^j, $j = 1, \ldots, n-2$, as the negative rows of A, i.e., $[\gamma^1 \ldots \gamma^{n-2}] = -A'$. The constraint set may be written as $C = \{\theta : \langle \gamma^i, \theta \rangle \leq 0, \text{ for } i = 1, \ldots, n-2\}$, where the notation $\langle a, b \rangle$ refers to the vector inner product of a and b. The "polar cone" (see Rockafellar (1970)) is defined as

$$\Omega^0 = \{\rho : \langle \theta, \rho \rangle \leq 0, \ \forall\, \theta \in \Omega\}$$

Clearly, $\gamma^1 \ldots \gamma^{n-2} \in \Omega^0$. These vectors are *edges* of the polar cone, i.e., each $\rho \in \Omega^0$ can be written as a non-negative linear combination of the γ^i, and furthermore, an edge cannot be written as the sum of two or more linearly independent vectors in the cone. For a more detailed discussion of the constraint and polar cones in the context of shape-restricted regression, see Meyer (1999b) or Fraser and Massam (1989).

Note that v^1 and v^2 are orthogonal to the edges γ^i, so that the polar cone Ω^0 is contained in $V^\perp$. Vectors $\delta^1, \ldots \delta^{n-2}$, also orthogonal to V, can be found so that any $\theta \in \Omega$ can be written as $\sum_{i=1}^{n-2} b_j \delta^j$ where the scalars b_j are nonnegative. These δ^j are the edges of the constraint cone Ω. For the constraint matrix A given by (2), the constraint cone edges are the negative rows of the matrix A^0:

$$A^0 = \begin{pmatrix} -10 & 8 & 5 & 2 & -1 & -4 \\ -20 & 2 & 24 & 11 & -2 & -15 \\ -15 & -2 & 11 & 24 & 2 & -20 \\ -4 & -1 & 2 & 5 & 8 & -10 \end{pmatrix}.$$

Note that the lengths of the row vectors are arbitrary. The edges can be "normalized" so that $A^0 A' = -I_{n-2}$; i.e., $\langle \delta^j, \gamma^i \rangle = 0$ if $i \neq j$ and $\langle \delta^i, \gamma^i \rangle = -1$. The matrix A^0 is a constraint matrix for Ω^0, i.e., $\rho \in \Omega^0$ if and only if $A^0 \rho \geq 0$.

The subspace $V^\perp$ can be partitioned into "sectors" that are determined by subsets of the integers $1, \ldots, n-2$. For any $J \subseteq \{1, 2, \ldots, n-2\}$, let Ω_J be the set of all $y \in V^\perp$ such that

$$y = \sum_{j \in J} b_j \gamma^j + \sum_{j \notin J} b_j \delta^j$$

where the $b_j > 0$ for $j \in J$, and $b_j \geq 0$ for $j \notin J$. Note that the interior of the polar cone is a sector defined by $J = \{1, 2, \ldots, n-2\}$ and the constraint cone is a sector with $J = \emptyset$. Each sector Ω_J is a convex polyhedral cone with edges γ^j, $j \in J$, and δ^j, $j \notin J$. Let $d = \#(J)$, or the number of elements in the set J, and let S_J be the space spanned by the γ^j, $j \in J$. The cone Ω_J has a constraint matrix A_J such that the rows of A_J span $V^{\perp}$. Further, A_J contains d rows which are in S_J and $n-2-d$ rows orthogonal to S_J. This can be seen by constructing a matrix A_J^0 with rows $-\gamma^j$, $j \in J$, and $-\delta^j$, $j \notin J$, and noting that $A_J^0 A_J' = -I_{n-2}$.

Let C_J be the set of all y in $\mathcal{R}^n$ such that

$$y = \sum_{j \in J} b_j \gamma^j + \sum_{j \notin J} b_j \delta^j + c_1 v_1 + c_2 v_2 \tag{3}$$

where the $b_j > 0$ for $j \in J$, and $b_j \geq 0$ for $j \notin J$, and c_1 and c_2 are any real numbers. Then the C_J partition $\mathcal{R}^n$ and representation (3) is unique. For proof of these and the following results, see Meyer (1999b).

Proposition 1 *Given $y \in \mathcal{R}^n$ with representation (3), the projection of y onto the constraint set C is*

$$\hat{\theta} = \sum_{j \notin J} b_j \delta^j + c_1 v_1 + c_2 v_2,$$

and the residual vector $\hat{\rho} = y - \hat{\theta}$, or $\sum_{j \in J} b_j \gamma^j$, is the projection of y onto Ω^0.

Proposition 2 *If $y \in C_J$, then $\hat{\theta}$ is the projection of y onto the* linear space *spanned by the vectors δ^j, $j \notin J$, v^1, and v^2. Similarly, $\hat{\rho}$ is the projection of y onto the linear space spanned by the vectors γ^j, $j \in J$.*

The mixed primal-dual bases algorithm and the hinge algorithm find $\hat{\theta}$ by determining the set J such that $y \in C_J$ and performing an ordinary least squares regression of y on the vectors v^1, v^2, and δ^j, for $j \notin J$. Alternatively, an ordinary least squares regression on the γ^j, $j \in J$, produces $\hat{\rho}$, from which $\hat{\theta} = y - \hat{\rho}$ is easily obtained.

An example of a scatterplot with data generated from a quadratic function with independent mean-zero normal errors is shown in Figure 1. The simple linear regression estimate is shown as a dotted line, the quadratic regression estimate is shown as a dashed line, and the convex regression estimate is the solid line. The fortran code for the convex regression and documentation are provided at the following web page. The code provides a p-value for the hypothesis test given in Section 1.

`www.stat.uga.edu/~mmeyer/convexreg.html`

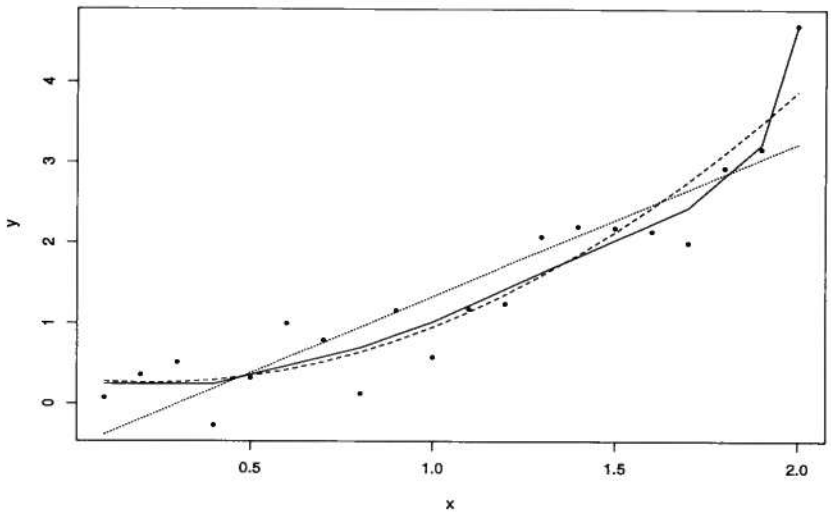

Figure 1: Example of fits to scatterplot. The dotted line is the simple linear regression estimate, the dashed line is quadratic regression estimate, and the solid line is the convex regression estimate.

3 The test statistic and its distribution

Let $\hat{y}$ be the simple linear regression estimator, i.e., the projection of y onto V. Let $SSE_0 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ and $SSE_1 = \sum_{i=1}^n (y_i - \hat{\theta}_i)^2$. If the model variance σ^2 is known, then a likelihood ratio test statistic for the hypotheses given in Section 1 is

$$\chi^2_{01} = \frac{1}{\sigma^2}(SSE_0 - SSE_1). \tag{4}$$

For the unknown variance case, the test statistic is

$$B_{01} = \frac{\chi^2_{01}}{\chi^2_{01} + SSE_1/\sigma^2} = \frac{SSE_0 - SSE_1}{SSE_0}. \tag{5}$$

The notation for the test statistics is borrowed from Robertson, Wright, and Dykstra (1988), Chapter 2. Large values indicate support for the alternative hypothesis.

To find the distributions of these test statistics under the null hypothesis that the true regression function is a straight line, the model dimension and error degrees of freedom for the convex regression model are important. For ordinary least squares regression, the dimension of the model is the dimension k of the linear space spanned by the columns of the design matrix, and the error degrees of freedom is $d = n - k$, so that SSE/σ^2 has a $\chi^2(d)$ distribution if the true θ is actually in the linear space. Since by Proposition 2, $\hat{\theta}$ is the projection of y onto the linear space spanned

by v^1, v^2, and δ^j, $j \notin J$, one might conjecture that the dimension of the model is $n - \#(J)$, where $\#(J)$ is the number of elements in the set J, and hence $d = \#(J)$ is the error degrees of freedom. However, this J is a random set, since another value of the error vector ϵ might put the data vector y in a different set C_J.

The derivation of the distribution of the test statistics (4) and (5) begins with:

$$P(\chi^2_{01} \leq a) = \sum_{subsets J} P(\chi^2_{01} \leq a, y \in C_J) = \sum_{subsets J} P(\chi^2_{01} \leq a | y \in C_J) P(y \in C_J).$$

It will be argued that, under H_0, the conditional distribution of SSE_1/σ^2 given $y \in C_J$ is $\chi^2(d)$, where $d = \#(J)$. From this it is argued that the null distribution of χ^2_{01} is a mixture of Chi-square densities, and since χ^2_{01} and SSE_1 are independent, the distribution of B_{01} has a mixture of Betas density. The mixing distribution in each case is determined by the relative volumes of the sets C_J, that is, by the probabilities that a standard multivariate normal random vector falls in C_J.

The following lemmas are general results about multivariate normal random vectors. The first is Lemma B of Robertson, Wright, and Dykstra (1988), pp71-72.

Lemma 1 *Let $Z = (Z_1, \ldots, Z_n)'$, where the Z_i are independent standard normal random variables, and let A be an $m \times n$ matrix. Then the conditional distribution of $\| Z \|^2$, given $AZ \geq 0$, is $\chi^2(n)$, provided the conditioning set is not empty.*

The lemma is proved by writing the components of Z in polar coordinates and observing that the value of $\| Z \|^2$ depends only on the radius R, and the condition $AZ \geq 0$ concerns the angles, which are independent of R.

Lemma 2 *Let $Z = (Z_1, \ldots, Z_n)'$, where the Z_i are independent standard normal random variables, and let $\hat{Z}$ be the projection of Z onto a linear space S of $d < n$ dimensions. Let A be an $m \times n$ matrix such that each row of A is orthogonal to S. Then the conditional distribution of $\| \hat{Z} \|^2$, given $AZ \geq 0$, is $\chi^2(d)$, provided the conditioning set is not empty.*

Proof: Let $v_1, v_2, \ldots, v_n$ be an orthonormal set of vectors in $\mathcal{R}^n$ such that $v_1, \ldots, v_d$ span S. Write $Z = \sum_{i=1}^n \langle v_i, z \rangle v_i =: \sum_{i=1}^n a_i v_i$. Then the a_i are independent standard normal random variables, for $i = 1, \ldots, n$, and $\hat{Z} = \sum_{i=1}^d a_i v_i$. Further, $\| \hat{Z} \|^2 = a_1^2 + \cdots a_d^2$, which has a $\chi^2(d)$ density. If V is a matrix such that the columns of V are the vectors $v_1, \ldots, v_n$, and $a = (a_1, \ldots, a_n)'$, then $Z = Va$. This can be written as $V_1 \alpha^1 + V_2 \alpha^2$ where V_1 is the $n \times d$ matrix such that $V = [V_1 | V_2]$ and $\alpha^1 = (a_1, \ldots, a_d)'$. The condition $AZ \geq 0$ can be written as $AV_1 \alpha^1 + AV_2 \alpha^2 \geq 0$. This does not affect the density of $a_1^2 + \cdots a_d^2$, since $AV_1 = 0$ by construction, and α_1 and α_2 are independent.

Lemma 3 *Let $Z = (Z_1, \ldots, Z_n)'$, where the Z_i are independent standard normal random variables, and let $\hat{Z}$ be the projection of Z onto a linear space S of $d < n$ dimensions. Let A be an $m \times n$ matrix such that each row of A is a vector in S. Then the conditional distribution of $\| \hat{Z} \|^2$, given $AZ \geq 0$, is $\chi^2(d)$, provided the conditioning set is not empty.*

Proof: Define the vectors v_i and a as in the proof of Lemma 2. Note that $AV_2 = 0$ in this case, so that the condition $AV \geq 0$ becomes $AV_1\alpha^1 \geq 0$. Then the distribution of $a_1^2 + \cdots + a_d^2$, given $AV_1\alpha^1 \geq 0$, is equal to its unconditional distribution, by Lemma 1.

The following result follows from the uniqueness of the representation (3). If $y \in C_J$, then $y + av^1 + bv^2 \in C_J$, for all $a, b \in \mathcal{R}$, and furthermore, the projection $\hat{\rho}$ of y onto Ω^0 is unchanged if any vector in V is added to y. Then when H_0 is true, without loss of generality take the underlying θ to be the origin, or $f(x) = 0$ and $y = \epsilon$ in model (1).

For any realization of ϵ in $\mathcal{R}^n$, let J be the index set so that $\epsilon \in C_J$, and let S_J be the space spanned by γ^j, for $j \in J$. Let $\hat{\epsilon}_J$ be the projection of ϵ on S_J. The next proposition is a consequence of Lemmas 2 and 3, since the constraint matrix A_J for sector Ω_J can be partitioned vertically into two submatrices A_1 and A_2. The d rows of the matrix A_1 are in S_J, and the $n - 2 - d$ rows of the matrix A_2 are orthogonal to S_J. Further, both $A_1\epsilon \geq 0$ and $A_2\epsilon \geq 0$ when $\epsilon \in C_J$.

Proposition 3 *The conditional distribution of $\| \hat{\epsilon}_J \|^2 / \sigma^2$ given J, is $\chi^2(d)$, the unconditional distribution of $\| \hat{\epsilon}_J \|^2 / \sigma^2$.*

Since SSE_1 is $\| \hat{\rho} \|^2$, and under H_0 $\| \hat{\rho} \|^2$ is equivalent to $\| \hat{\epsilon}_J \|^2 / \sigma^2$ given J, the corollary follows from the proposition.

Corollary 1 *If the null hypothesis $\theta \in V$ is true, the conditional distribution of SSE_1/σ^2, given $y \in C_J$, is $\chi^2(d)$, where $d = \#(J)$.*

Let D be the random variable $\#(J)$. From the corollary, we have that the conditional distribution of SSE_1/σ^2, given $D = d$, is $\chi^2(d)$. Note that since $\langle y - \hat{\theta}, \hat{y} \rangle = 0$, and $\langle y - \hat{\theta}, \hat{\theta} \rangle = 0$ by Proposition 2, the random vectors $y - \hat{\theta}$ and $\hat{\theta} - \hat{y}$ are independent. Further, SSE_0/σ^2 has a $\chi^2(n-2)$ density, so the conditional distribution of χ_{01}^2 given $D = d$ is $\chi^2(n - d - 2)$. Similarly, by the independence of χ_{01}^2 and SSE_1, the conditional distribution of B_{01} given $D = d$ is $Beta((n - d - 2)/2, d/2)$.

Theorem 1 *Under* H_0,

$$P(\chi^2_{01} \leq a) = \sum_{d=0}^{n-2} P\left[\chi^2(n-d-2) \leq a\right] P(D=d)$$

and

$$P(B_{01} \leq a) = \sum_{d=0}^{n-2} P\left[Beta\left(\frac{n-d-2}{2}, \frac{d}{2}\right) \leq a\right] P(D=d).$$

where $\chi^2(0) \equiv 0$, $Beta(0,\beta) \equiv 0$, and $Beta(\alpha,0) \equiv 1$. Note that when $D = n-2$, the convex regression estimate is identical to the simple linear regression estimtate, and when $D = 0$, $\hat{\theta} = y$, so that the definitions of the test statistics make sense when a degree of freedom has value zero.

The mixing distribution, or values of $P(D=d)$, for $d = 0, \ldots, n-2$, is obtained from the relative volumes of the sets C_J. The probability that y is in C_J, when H_0 is true, is equivalent to the probability that the value of a standard multivariate normal random vector taking values in $\mathcal{R}^n$ falls into C_J. This is equivalent to calculating the positive orthant probability for a mean-zero multivariate normal with covariance matrix calculated using the constraint matrix A_J. There are some papers in the literature on calculating and approximating orthant probabilities that could be used to calculate the mixing distributions. However, there are 2^{n-2} sectors, so when n is large, there are a prohibitive number of orthant volume calculations. Instead, the mixing probabilities can be found numerically by generating N standard multivariate normal random vectors, and determining the value of D for each vector. The probabilities $P(D=d)$ can be found to about three decimal place accuracy when N is one million. Tables of these mixing distribution values for $n = 6$ through $n = 80$ can be found at the web page listed in the last section. Table 1 provides critical values of the test statistic B_{01} for selected sample sizes and test sizes.

4 Simulations

The power of the convex alternative hypothesis test is compared with the quadratic alternative F-test for several underlying regression functions and sample sizes, and for two values of the model variance. One million datasets were generated for each regression function, sample size, and model variance combination; the test statistics for both alternatives were calculated for each dataset and compared with the critical values. The proportion of times the null hypothesis was rejected is reported in the tables as the power. The large number of simulations provides about three decimal place accuracy in the tables.

The x-values are chosen to be equally spaced in $(0,2)$, and the regression functions are chosen to correspond to the null hypothesis ($f_0(x) = x$), the quadratic alternative ($f_1(x) = x^2$), a convex

	α						
n	0.80	0.85	0.90	0.95	0.975	0.99	0.995
6	0.569	0.670	0.779	0.898	0.958	0.991	0.998
7	0.476	0.567	0.672	0.803	0.887	0.950	0.975
8	0.411	0.491	0.590	0.721	0.814	0.895	0.934
9	0.364	0.436	0.527	0.653	0.749	0.839	0.887
10	0.327	0.392	0.476	0.595	0.691	0.785	0.838
11	0.298	0.358	0.435	0.548	0.641	0.737	0.793
12	0.273	0.329	0.400	0.508	0.597	0.692	0.750
13	0.253	0.304	0.371	0.473	0.559	0.653	0.711
14	0.236	0.284	0.347	0.443	0.526	0.617	0.675
15	0.222	0.266	0.325	0.417	0.496	0.585	0.643
16	0.209	0.251	0.306	0.393	0.469	0.556	0.612
17	0.198	0.237	0.290	0.373	0.446	0.530	0.586
18	0.188	0.225	0.275	0.354	0.424	0.506	0.560
19	0.179	0.214	0.262	0.337	0.405	0.484	0.537
20	0.171	0.205	0.250	0.322	0.388	0.464	0.516
21	0.164	0.196	0.240	0.309	0.372	0.446	0.496
22	0.157	0.188	0.230	0.296	0.357	0.429	0.478
23	0.151	0.180	0.221	0.285	0.343	0.413	0.461
24	0.145	0.174	0.212	0.274	0.331	0.399	0.446
25	0.140	0.168	0.205	0.264	0.319	0.385	0.431
26	0.135	0.162	0.198	0.255	0.309	0.373	0.418
27	0.131	0.156	0.191	0.245	0.298	0.361	0.405
28	0.127	0.151	0.185	0.239	0.289	0.350	0.392
29	0.123	0.147	0.179	0.232	0.281	0.340	0.381
30	0.119	0.143	0.174	0.225	0.272	0.330	0.371
35	0.104	0.124	0.152	0.196	0.238	0.289	0.325
40	0.093	0.111	0.135	0.174	0.211	0.257	0.290
45	0.084	0.100	0.121	0.157	0.190	0.232	0.262
50	0.077	0.091	0.111	0.143	0.173	0.212	0.239
55	0.071	0.084	0.101	0.131	0.159	0.194	0.220
60	0.066	0.078	0.094	0.121	0.147	0.180	0.204
65	0.061	0.072	0.088	0.113	0.137	0.168	0.190
70	0.057	0.068	0.082	0.106	0.128	0.160	0.178
75	0.054	0.064	0.077	0.099	0.121	0.148	0.167
80	0.051	0.060	0.073	0.094	0.114	0.139	0.158

Table 1: Percentiles of beta mixture test statistic

function that is "close to" a quadratic over the range given ($f_2(x) = \exp(x)$), and a convex function that is *not* close to a quadratic ($f_3(x) = \exp(x-1)$). The results in Tables 2 and 3 show that for small samples or large variance, when the power is not expected to be large, the more general alternative results in more power, even when the underlying function is quadratic. When power for both tests is larger, as in $n = 80$ and $\sigma^2 = 1$, the quadratic alternative has higher power for the underlying $f_1(x)$ and $f_2(x)$, but not for the underlying $f_3(x)$. However, since this test is more likely to be performed for scatterplots that look "borderline" it can be argued that higher power for the smaller-power situation is more important.

	$f_0(x) = x$		$f_1(x) = x^2$		$f_2(x) = \exp(x)$		$f_3(x) = \exp(x-1)$	
n	convex	quadratic	convex	quadratic	convex	quadratic	convex	quadratic
10	0.050	0.050	0.189	0.125	0.339	0.246	0.117	0.076
20	0.050	0.050	0.298	0.239	0.539	0.480	0.155	0.109
40	0.050	0.050	0.473	0.450	0.787	0.781	0.215	0.172
80	0.050	0.050	0.724	0.750	0.931	0.938	0.320	0.298

Table 2: Power of test compared with that of F-test for quadratic model and $\sigma^2 = 1$, for equally spaced x values on $(0, 2)$.

	$f_0(x) = x$		$f_1(x) = x^2$		$f_2(x) = \exp(x)$		$f_3(x) = \exp(x-1)$	
n	convex	quadratic	convex	quadratic	convex	quadratic	convex	quadratic
10	0.050	0.050	0.073	0.054	0.091	0.062	0.063	0.052
20	0.050	0.050	0.085	0.061	0.112	0.077	0.085	0.053
40	0.050	0.050	0.102	0.074	0.141	0.106	0.075	0.057
80	0.050	0.050	0.128	0.100	0.194	0.163	0.085	0.064

Table 3: Power of test compared with that of F-test for quadratic model and $\sigma^2 = 16$, for equally spaced x values on $(0, 2)$.

5 Discussion

A similar hypothesis testing problem is solved in Robertson, Wright, and Dykstra (1988). Consider

$H_0 : f(x) = c$; vs.

$H_1 : f(x) \in \mathcal{F}$,

where $\mathcal{F}$ is the class of nondecreasing functions. There is a closed form solution for the monotone regression problem, but the nondecreasing constraints can be written in a $n - 1 \times n$ constraint

matrix and the distribution of the test statistic (see Robertson, Wright, and Dykstra (1988), p69) can be derived analogously to methods given in Section 3. For $n = 6$, the constraint matrix for nondecreasing constraints is

$$A = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}.$$

Note that the rows of the constraint matrix are orthogonal to the vector $v = (1, 1, \ldots, 1)'$, the edges of the constraint cone can be chosen to be also orthogonal to v, and that, under the null hypothesis, the underlying θ is a multiple of v. Then Proposition 3 holds for monotone regression, and the derivation of the mixture of betas test statistic is analogous. The web site given at the end of the second section also contains code and (numerically calculated) mixing distribution values for testing the above hypotheses concerning monotone regression.

Consider next the hypothesis test

$H_0 : f(x) = a + bx + cx^2$; vs.

$H_1 : f(x) \in \mathcal{F}$,

where $\mathcal{F}$ is the class of convex functions. This testing problem is of interest in the case when the practitioner knows that the regression function is convex, and wishes to fit a quadratic to the data, but is not sure if this is the correct convex function. However, in this case, the distribution of the likelihood ratio test statistic can *not* be calculated analogously to Section 3 methods, since the mixing distribution depends on the true value of θ, which can not be taken to be the origin in this case. Further, even if the probabilities $P(D = d)$ were known, Lemmas 2 and 3 and hence Proposition 3 do not hold for θ in the interior of the constraint cone.

The error degrees of freedom for shape-restricted regression was explored in Meyer and Woodroofe (2000). For θ in the interior of the constraint cone, it is shown that

$$n\sigma^2 - 2\sigma^2 E(D) \leq E\left[\| y - \hat{\theta} \|^2\right] \leq n\sigma^2 - \sigma^2 E(D),$$

where E denotes expectation, and this expectation depends on the values of the underlying θ and σ^2. For monotone regression, it is shown that

$$E\left[\| y - \hat{\theta} \|^2\right] = n\sigma^2 - c_1\sigma^2 E(D) + o(n^{1/3}),$$

where c_1 is about 1.5. This shows that the distribution of the error sum of squares given in this paper is in fact not correct when the true value of θ can not be taken to be at the origin.

Note that the condition of distinct x-values is restrictive. If there are more than one observation at a given value of x, a pracitioner would like to average those observations and perform a weighted

analysis. More generally, suppose that for the model (1), the error vector ϵ has a known covariance matrix Σ, so that the decomposition $\Sigma = BB'$ can be found with full rank matrix B. Then for the transformations $\tilde{y} = B^{-1}y$, $\tilde{\theta} = B^{-1}\theta$, and $\tilde{A} = AB$, the maximum likelihood estimate of θ is found by minimizing $\| \tilde{y} - \tilde{\theta} \|^2$ under the restriction $\tilde{A}\tilde{\theta} \geq 0$. If $\tilde{V}$ is the space spanned by the vectors $\tilde{v}_1 = B^{-1}v_1$ and $\tilde{v}_2 = B^{-1}v_2$, then under the null hypothesis, $\tilde{\theta} \in \tilde{V}$, and the problem reduces to a projection onto a cone. The methods of Section 3 are used to find the maximum likelihood estimator for $\tilde{\theta}$, then the inverse transformation $\theta = B\tilde{\theta}$. The mixing distribution values of $P(D = d)$ depend on the matrix Σ and have to be recalculated. These numerical calculations are time consuming but not difficult. Code is also provided at the web address given in Section 2 for the $\Sigma \neq I\sigma^2$ case, but it may take several hours to run for moderate-sized samples.

Acknowledgements

This research was supported by the National Science Foundation.

References

[1] Fraser, D.A.S. and Massam, H., (1989). A mixed primal-dual bases algorithm for regression under inequality constraints. Application to convex regression. *Scand. J. Statist*, **16** 65-74.

[2] Meyer, M. C., (1999a) An Algorithm for Projections onto Convex Cones with Applications of Nonparametric Regression and Quadratic Programming. Technical Report **99-13**, University of Georgia.

[3] Meyer, M. C., (1999b) An extension of the mixed primal-dual bases algorithm to the case of more constraints than dimemsions. *JSPI*, **81**; 13-31.

[4] Meyer, M. and Woodroofe, M. (2000) On the Degrees of Freedom in Shape-Restricted Regression. *Annals of Statistics* **28(4)** pp1083-1104

[5] Robertson, T., Wright, F. T., and Dykstra, R. L. (1988) *Order Restricted Statistical Inference.* John Wiley & Sons, New York.

[6] Rockafellar, R. T. (1970). *Convex Analysis.* Princeton University Press, New Jersey