Inference using Shape-Restricted Regression Splines

April 27, 2006

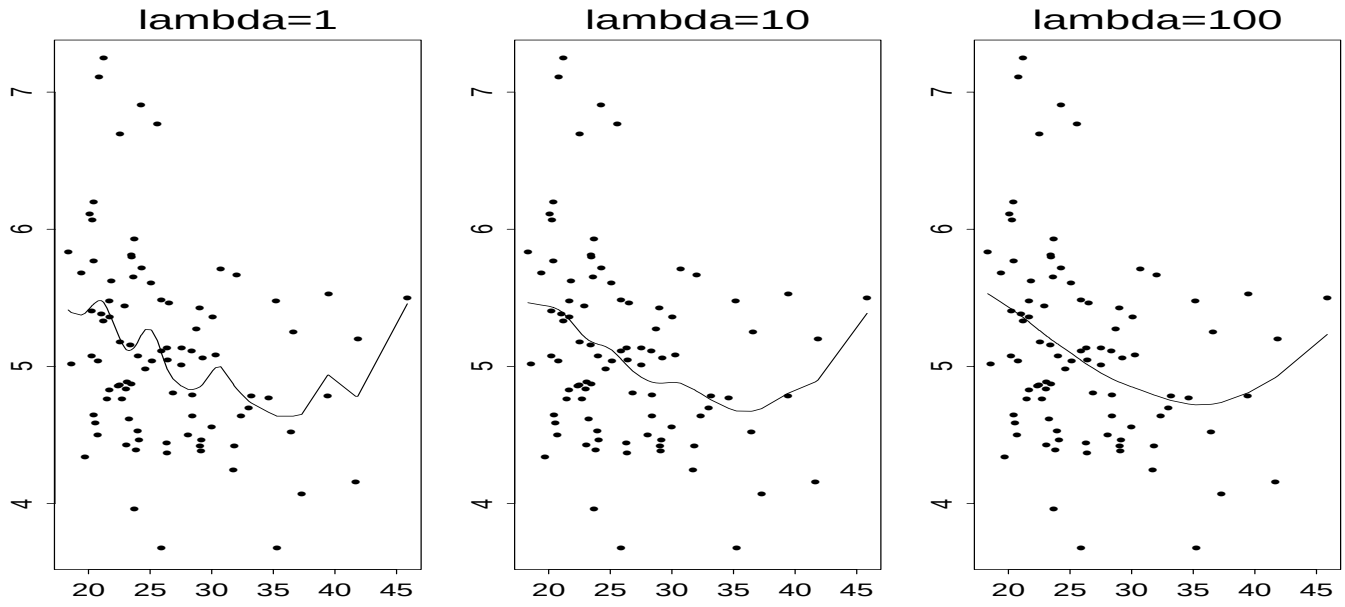
Nonparametric regression provides fits to data with minimal assumptions, and is used when a parametric version of the regression function is not known.

Popular methods require a choice of user-defined parameters in the fit:

- Kernel smoother: bandwidth
- Smoothing spline: smoothing parameter
- Regression splines: number and placement of knots

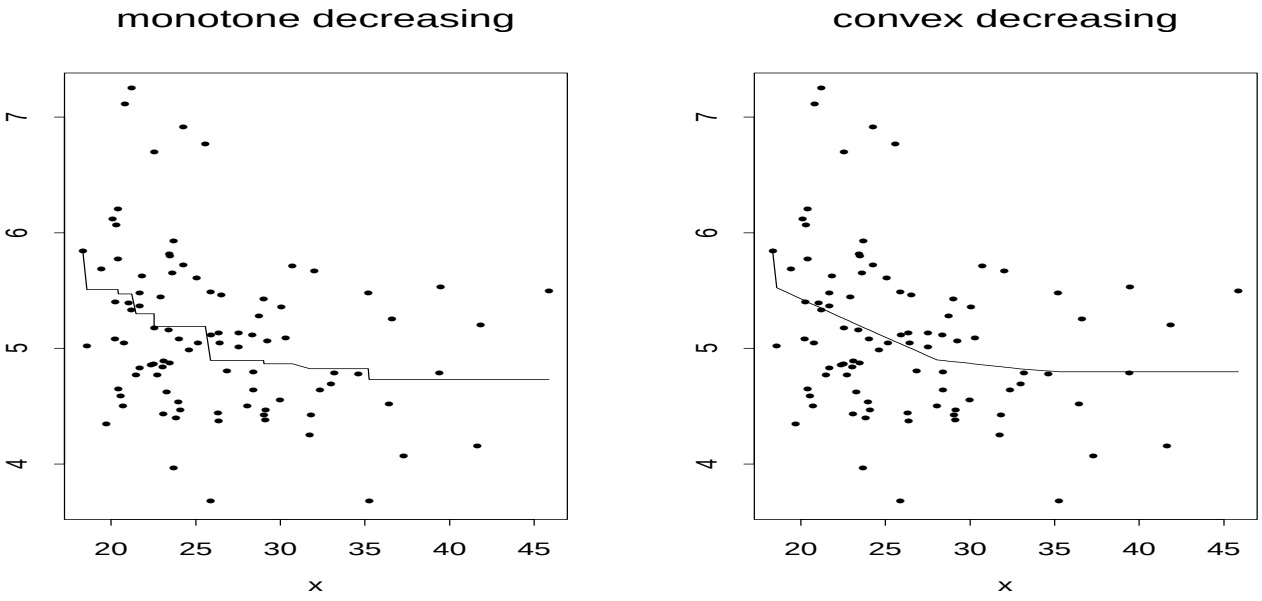
If the fits are sensitive to the user-defined parameters, it's hard to make practical inference about the underlying regression function.

Examples of natural cubic smoothing splines with different smoothing parameters.



Observational data regarding blood plasma micronutrients. The response is the logarithm of the level of beta carotene the subject's blood plasma (high levels are believed to be protective against cancer). The predictor is the body mass index.

Fits using only shape restrictions such as monotonicity or convexity do not require user-defined parameters, but are not smooth, and the fits are not parsimonious, in that the model degrees of freedom is in some sense large.



Some hypothesis tests:

H_0 : function is constant

H_1 : function is increasing

H_0 : function is linear

H_1 : function is convex

CAN'T do:

H_0 : function is linear

H_1 : function is increasing

Smoothing and Shape Assumptions

When shape assumptions can be combined with smoothing, we get the best of both worlds:

The fits are nice-looking.

Although we still use user-defined parameters, the fit is more robust to choices.

We can do some types of inference about the regression function.

The I-splines from Ramsay (1988) are constrained to be monotone.

These are integrals of positive M-splines, and so are increasing.

The M-splines are given recursively for subsequently higher order. Order-1 splines are the piecewise constant (step functions):

$$M^{(1)}(x) = \begin{cases} \frac{1}{t_{i+1}-t_i} & \text{for } \; t_i \leq x \leq t_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

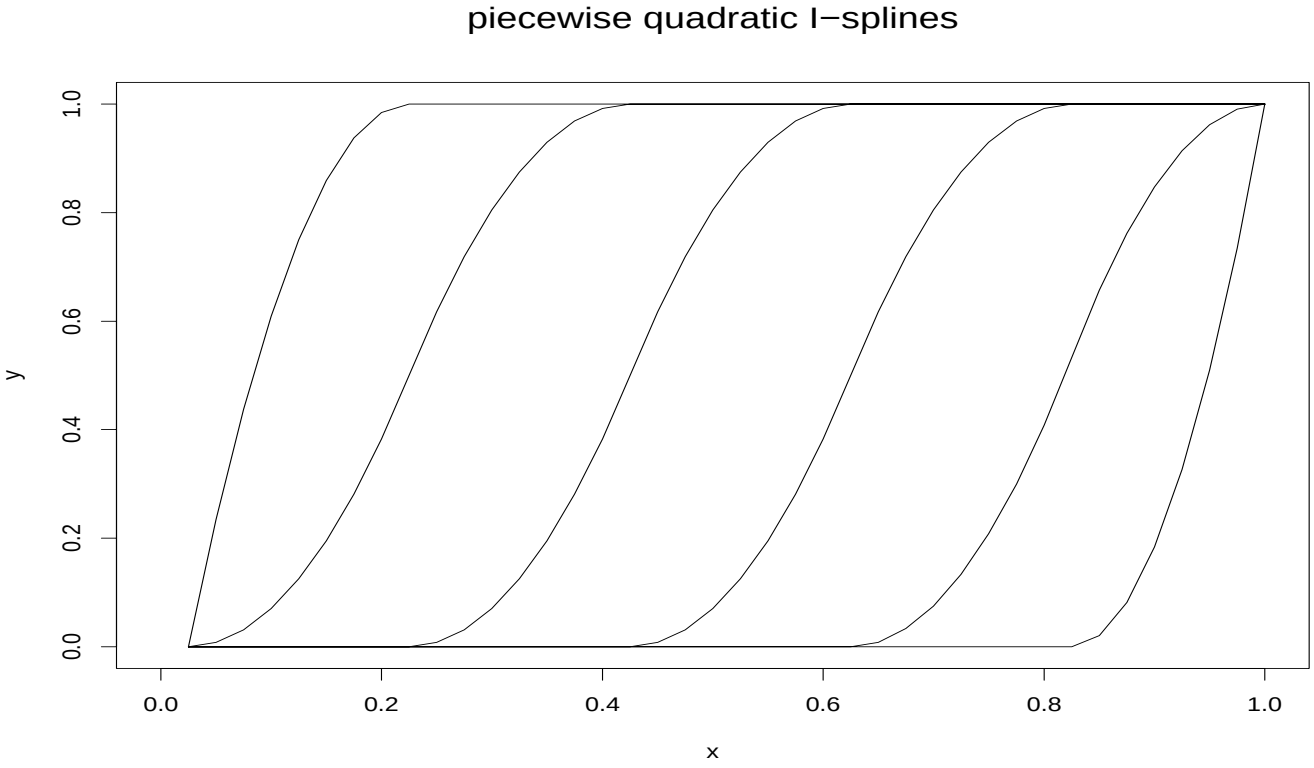
Order-k M-splines are computed from the lower orders, recursively:

$$M_i^{(k)}(x) = \begin{cases} \frac{k[(x-t_i)M^{(k-1)}(x)+(t_{i+k}-x)M^{(k-1)}(x)]}{(k-1)(t_{i+k}-t_i)} & \text{for } \; t_i \leq x \leq t_{i+k} \\ 0 & \text{otherwise} \end{cases}$$

The I-splines are

$$I_i^{(k)}(x) = \begin{cases} 0 & \text{for } x < t_i \\ \int_{t_i}^{x} M_i^{(k)}(u)du & \text{for } t_i \le x \le t_{i+k} \\ \int_{t_i}^{t_{i+k}} M_i^{(k)}(u)du + x - t_{i+k} & \text{for } x > t_{i+k} \end{cases}$$

The I-splines are integrals of positive functions so they are increasing. The first order I-splines are piecewise linear and continuous; the second order I-splines are piecewise quadratic with continuous first derivative, etc.

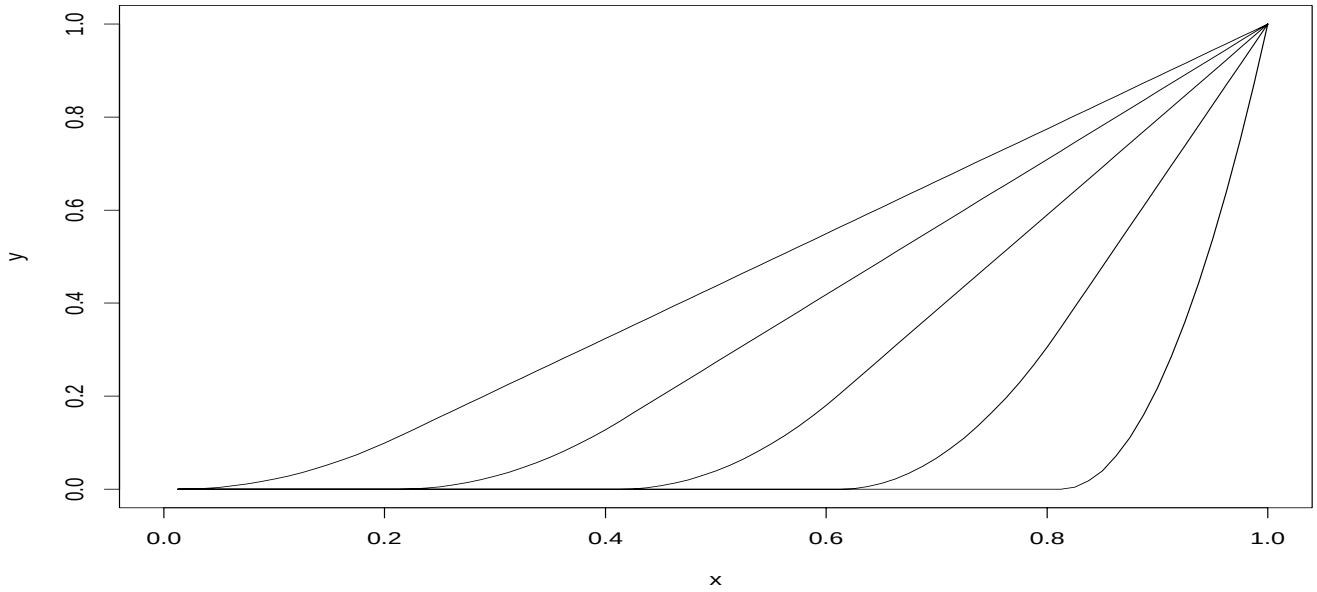


The I-splines can easily be adapted to monotone decreasing constraints.

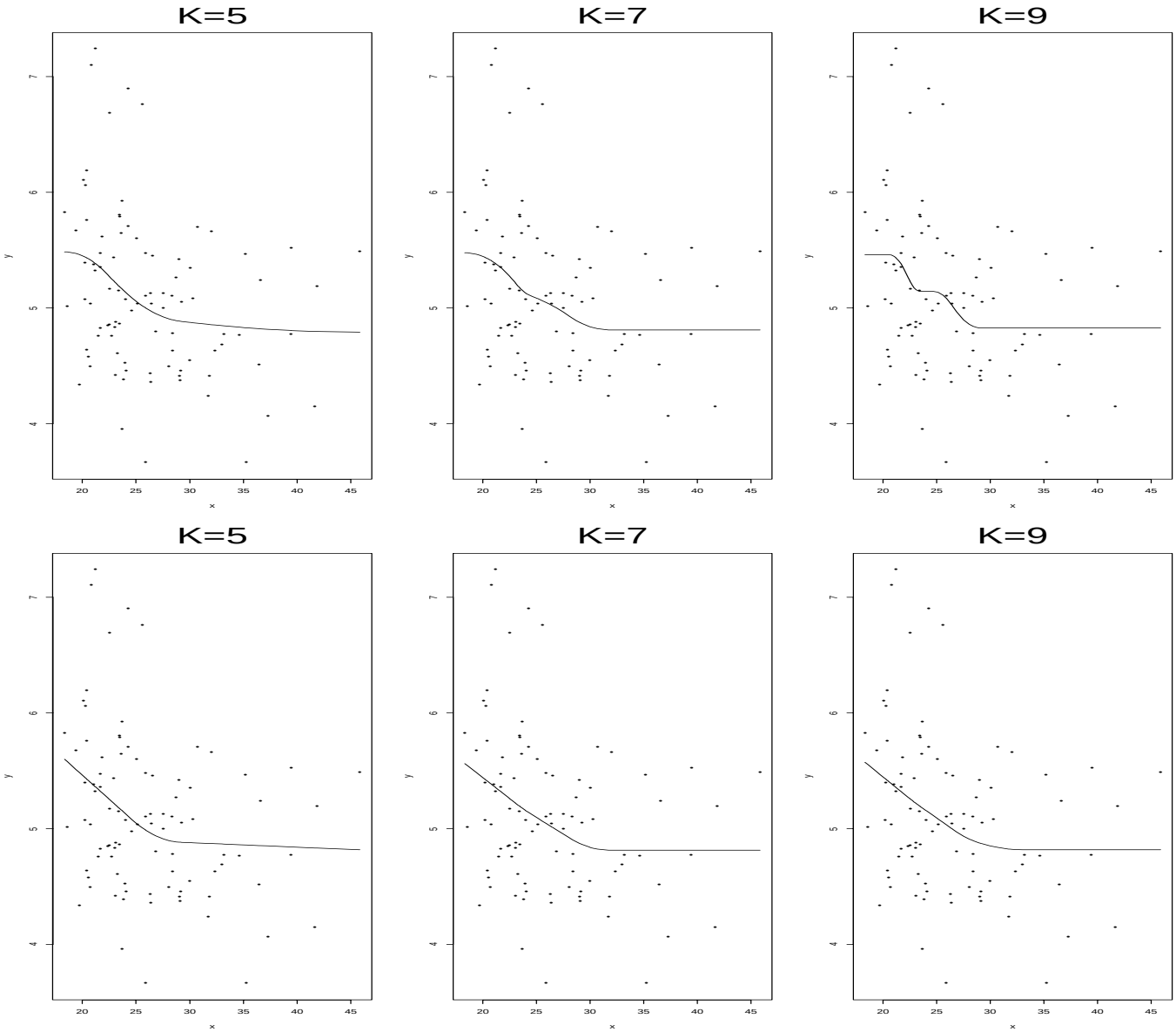
We can get convex C-splines by integrating the increasing I-splines. Concave C-splines can be obtained by integrating the decreasing I-splines.

It is easy to get increasing concave, decreasing convex, etc.

Convex C-splines are shown below:



Piecewise quadratic regression splines: monotone decreasing and convex decreasing fits to the beta carotene data:



Algorithm for regression spline fitting

Ramsay provides a iterative gradient-based algorithm for finding the solution. This converges to the solution in "infinitely many" steps, meaning that there is a stopping criterion defining "close enough."

The algorithm proposed here is a simple cone-projection. The algorithm obtains the exact solution in a few iterations, and each iteration has little computation.

The ideas behind the algorithm motivate inference methods.

We'll quickly review important ideas in shape-restricted regression.

Shape-Restricted Regression

The model is

$$y_i = f(x_i) + \sigma\epsilon_i, \quad i = 1, \ldots, n,$$

where ϵ_i are *iid* standard normal random variables and f is unknown except for its "shape," such as increasing, convex, etc.

Vector form: $\boldsymbol{y} = \boldsymbol{\theta} + \sigma\boldsymbol{\epsilon}$, where $\theta_i = f(x_i)$.

Then the shape restrictions are a set of linear inequality constraints such as

$$\theta_1 \leq \theta_2 \leq \cdots \leq \theta_n.$$

The convex constraints are:

$$\frac{\theta_2 - \theta_1}{x_2 - x_1} \leq \frac{\theta_3 - \theta_2}{x_3 - x_2} \leq \ldots \leq \frac{\theta_n - \theta_{n-1}}{x_n - x_{n-1}},$$

The shape restrictions can be written in the form $\boldsymbol{A\theta} \geq \mathbf{0}$.

The constraint matrix A is $m \times n$ where $m = n - 1$ for monotone and $m = n - 2$ for convex constraints.

The problem is to find $\boldsymbol{\theta}$ to minimize $\| \boldsymbol{y} - \boldsymbol{\theta} \|^2$ over constraints $\boldsymbol{A\theta} \geq \mathbf{0}$.

The m inequality constraints form a convex polyhedral cone C in $I\!R^n$.

The smallest linear space that contains the cone is $I\!R^n$.

There may be a linear space inside the constraint set; call this V and its dimension is r.

For monotone constraints, V is all multiples of the one-vector. For convex constraints, $V = \boldsymbol{\mathcal{L}}(\mathbf{1}, \boldsymbol{x})$.

Let $\Omega = C \cap V^{\perp}$.

The set Ω is a convex polyhedral cone contained in an $n - r = m$ dimensional subspace of $\mathbb{R}^n$. It has m "edges" that are uniquely defined (up to a scalar multiple).

The least-squares estimator of $\boldsymbol{\theta}$ is the projection of $\boldsymbol{y}$ onto C, or equivalently the sum of the projections of $\boldsymbol{y}$ onto V and Ω.

The "edges" of the cone, called $\boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^m$, are also the generators of the cone; these can be found using A and the basis vectors for V.

Any vector in Ω can be written as a sum of the edge vectors with non-negative coefficients; any vector in C can be written as

$$\boldsymbol{\theta} = \sum_{j=1}^{r} c_j \boldsymbol{v}^j + \sum_{j=1}^{m} b_j \boldsymbol{\delta}^j,$$

where $b_j \geq 0$.

Any subset of the edges indexed by $J \subseteq \{1, \ldots, m\}$ forms a face of the cone F_J. The projection lands on one of the faces.

Proposition: If J is known, then the least-squares estimator $\hat{\boldsymbol{\theta}}$ is the projection of $\boldsymbol{y}$ onto the linear space spanned by the set of edges indexed by J, plus the $\boldsymbol{v}^j$.

Therefore, the algorithm for projection involves finding J. Although the algorithm is iterative, it is finite because the number of faces is finite (2^m).

Sectors

Any subset of the edges indexed by $J \subseteq \{1, \ldots, m\}$ forms a *sector* Ω_J, another polyhedral cone in $V^\perp$.

All vectors in the sector project onto $\mathcal{F}_J$.

It can be shown that these 2^m sectors partition $V^\perp$.

Sectors C_J are similarly defined and partition $I\!R^n$.

Inference

If $\boldsymbol{\theta} \in V$, then the conditional distribution of SSE/σ^2 given $\boldsymbol{y} \in C_J$ is $\chi^2(n-d)$, where d is the number of edges or the size of J, plus the number of dimensions in V.

This can be used to construct test statistics for $H_0 : \boldsymbol{\theta} \in V$ versus $H_1 : \boldsymbol{\theta} \in C$. Examples are constant versus monotone regression function or linear versus convex.

The test statistic for the known σ^2 case

$$\chi^2_{01} = \frac{SSE_0 - SSE_1}{\sigma^2}$$

has a density equal to that of mixture of chi-squares densities, under H_0.

For $H_0 : \boldsymbol{\theta} \in V$ versus $H_1 : \boldsymbol{\theta} \in C$,

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

is distributed under H_0 as a mixture of beta densities

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

where $B(p, q)$ is a beta random variable with parameters p and q.

The mixing distribution parameters $P(D = d)$ are found numerically.

Back to regression splines

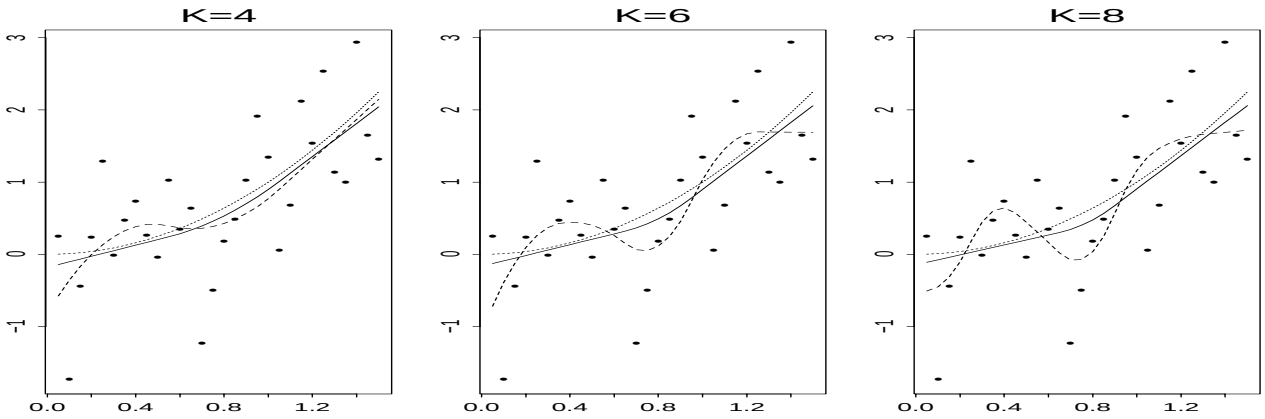
These ideas can be used in the regression spline arena:

The spline basis functions are the edges of the constraint cone.

There are a comparatively small number of edges, so that the fit is more parsimonious than the standard shape-restricted regression estimator, and it is smooth.

The cone projection algorithm typically converges in only a few iterations.

Comparing shape-restricted regression splines with the unrestricted version



The true regression function is $f(x) = x^2$ (dotted lines); $n = 30$; *iid* normal errors.

Piecewise quadratic regression splines are shown for various numbers of interior knots k, chosen at equal x-percentiles.

The solid lines show fits constrained to be non-decreasing and convex, while the dashed lines are the unconstrained fits.

Test of constant versus increasing f:

We can use the shape-restricted regression spline for the alternative fit, and we get again a test statistic with a mixture of betas distribution.

linear regression function				"ramp" regression function			
n	F-test	SRRS B-test	MREG B-test	n	F-test	SRRS B-test	MREG B-test
20	0.250	0.225	0.213	20	0.250	0.294	0.286
20	0.500	0.449	0.424	20	0.500	0.655	0.624
20	0.750	0.691	0.661	20	0.750	0.928	0.903
40	0.250	0.224	0.201	40	0.250	0.294	0.281
40	0.500	0.447	0.402	40	0.500	0.642	0.610
40	0.750	0.692	0.640	40	0.750	0.909	0.884
80	0.250	0.223	0.190	80	0.250	0.292	0.276
80	0.500	0.445	0.382	80	0.500	0.633	0.594
80	0.750	0.691	0.643	80	0.750	0.899	0.869

Power comparisons for the test of constant vs. monotone regression function. For the regression spline, the number of interior knots is the smallest integer larger than $\log n$, so $K = 5$, 6, and 7 corresponding to $n = 20$, 40, and 80. The results for the tests using the ordinary shape-restricted regression estimators are labeled as MREG.

Similar power results for piecewise quadratic convex regression splines and the test of linear versus convex regression function.

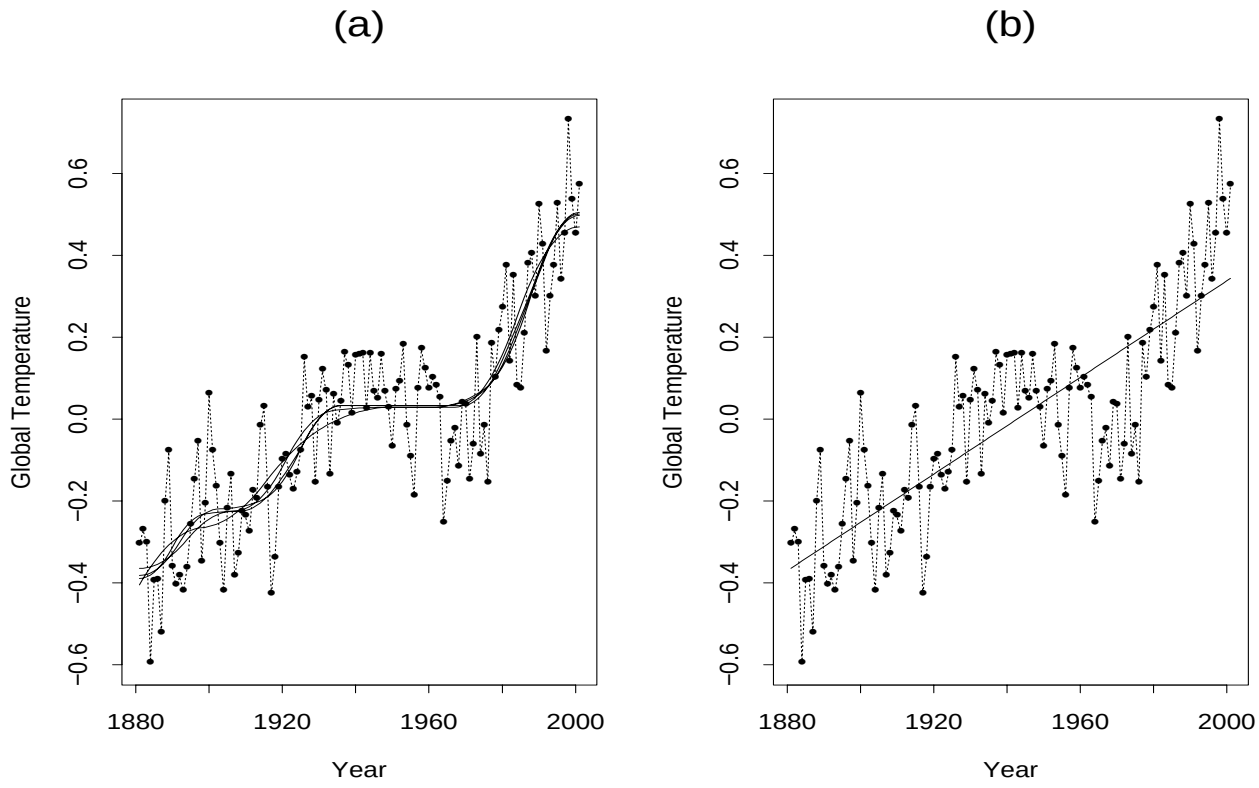
quadratic regression function				"ramp" regression function			
n	F-test	SRRS B-test	CREG B-test	n	F-test	SRRS B-test	CREG B-test
20	0.250	0.227	0.213	20	0.250	0.251	0.229
20	0.500	0.451	0.424	20	0.500	0.521	0.480
20	0.750	0.697	0.662	20	0.750	0.786	0.751
40	0.250	0.224	0.201	40	0.250	0.249	0.219
40	0.500	0.449	0.402	40	0.500	0.518	0.462
40	0.750	0.695	0.639	40	0.750	0.786	0.732
80	0.250	0.222	0.188	80	0.250	0.247	0.209
80	0.500	0.445	0.382	80	0.500	0.516	0.442
80	0.750	0.691	0.618	80	0.750	0.784	0.712

Concurrence of test results with different knot choices. The underlying regression function is linear, and the model variance is chosen so that the power for the test is approximately 0.50.

n	k_1	k_2	average power	% concurrence
40	3	5	0.25	92.3%
40	3	5	0.50	90.8%
80	3	6	0.25	88.9%
80	4	6	0.50	92.3%
80	4	7	0.25	92.2%
80	4	7	0.50	88.1%
80	5	7	0.25	94.7%
80	5	7	0.50	93.7%
120	5	7	0.25	93.7%
120	5	7	0.50	93.1%

Example: Global warming dataset

Average global temperatures (Hansen & Lebedeff) are plotted against year for 121 years:



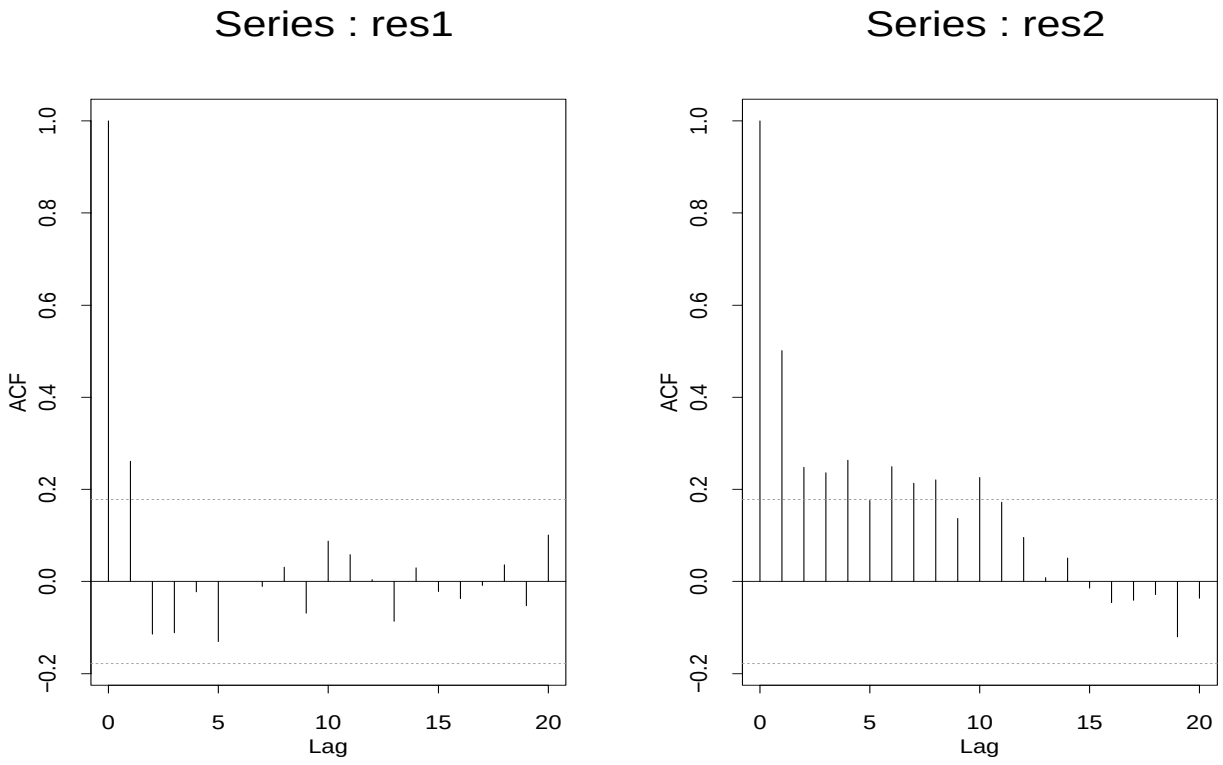
(a) Four piecewise quadratic monotone regression splines with 6, 8, 10, and 12 knots.

(b) The linear fit.

The auto-correlation plots of the residuals for both fits are shown below.

Left: residuals from the monotone regresion spline with 10 knots.

Right: residuals from the linear fit.



Fitting an AR(1) model with shape-restricted regression splines:

The covariance matrix for $\boldsymbol{\epsilon}$ is $\boldsymbol{A}$, where

$$\boldsymbol{A} = \frac{1}{1-\phi^2}\begin{pmatrix} 1 & \phi & \phi^2 & \phi^3 & \cdots & \phi^{n-1} \\ \phi & 1 & \phi & \phi^2 & \cdots & \phi^{n-2} \\ \phi^2 & \phi & 1 & \phi & \cdots & \phi^{n-3} \\ & & & \vdots & & \\ \phi^{n-1} & \phi^{n-2} & \cdots & \phi^2 & \phi & 1 \end{pmatrix}.$$

We can use the Cholesky decomposition $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}'$ to transform the model to an *iid* errors model:

$$\boldsymbol{L}^{-1}\boldsymbol{y} = \boldsymbol{L}^{-1}\boldsymbol{\theta} + \boldsymbol{L}^{-1}\sigma\boldsymbol{\epsilon}$$

becomes

$$\boldsymbol{y}^* = \boldsymbol{\theta}^* + \sigma\boldsymbol{\epsilon}^*$$

where now $\boldsymbol{\epsilon}^*$ has identity covariance matrix.

We find $\hat{\boldsymbol{\theta}}^*$ to minimize $\| \boldsymbol{y}^* - \boldsymbol{\theta}^* \|^2$ with the constraints $\boldsymbol{A}^*\boldsymbol{\theta}^* \geq \boldsymbol{0}$, where $\boldsymbol{A}^* = \boldsymbol{A}\boldsymbol{L}$.

Inference is done in the transformed model, and the solution $\hat{\boldsymbol{\theta}}$ to the original problem is obtained through the reverse transformation.

For known ϕ we get an exact test for constant versus increasing trend; if ϕ is unknown, we can estimate it using standard methods, and do a Cochran-Orcutt iteration.

Shape-restrictions are especially useful for time series trends. If the trend function is misspecified, the residuals are often more strongly correlated than the actual errors.

The power for tests of constant versus increasing trend is reduced for more strongly positively correlated errors.

Minimizing the assumptions about the trend function will guard against correlations induced by the incorrect trend function.

Goodness of fit

			Regression Spline			Smooth Spl	Shape-Restricted		
n	K	σ	IQ	CQ	ICQ	Cross-Valid	Isot	Conv	Inc-Cnvx
20	5	0.1	.0511	.0487	.0479	.0508	.0810	.0520	.0501
			.015	.0141	.0137	.0172	.0129	.0143	.0137
20	5	0.5	.221	.202	.188	.219	.272	.216	.195
			.0687	.0724	.0685	.0858	.0659	.0747	.0699
20	5	1.0	.396	.315	.343	.397	.471	.411	.357
			.137	.153	.140	.184	.142	.158	.143
40	6	0.1	.0391	.0363	.0358	.0372	.0671	.0395	.0379
			.0108	.00993	.00973	.0119	.00870	.0102	.00975
40	6	0.5	.169	.149	.140	.162	.218	.164	.148
			.0483	.0508	.0483	.0628	.0471	.0531	.0498
40	6	1.0	.302	.280	.255	.291	.376	.309	.271
			.0965	.106	.0986	.123	.101	.111	.102
80	7	0.1	.0298	.0272	.0268	.0273	.0542	.0300	.0287
			.00766	.00702	.00692	.00835	.00604	.00724	.00698
80	7	0.5	.129	.111	.105	.117	.173	.124	.112
			.0342	.0360	.0341	.0463	.0333	.0380	.0354
80	7	1.0	.231	.206	.189	.215	.298	.233	.205
			.0680	.0732	.0689	.0903	.0717	.0773	.0721

Table 1: Comparison of square root of average squared error loss. The underlying regression function is $f(x) = \exp(x-1)$, with design points equally spaced on $(0, 2)$ and i.i.d. normal errors. The

We see that the more restrictions on the shape, the better the fit, for both the standard shape-restricted estimator and for the regression splines. The convex and increasing-convex regression splines have typically lower SASEL than the smoothing spline, but the SASEL for the isotonic regression spline is a bit higher. The variation of the fit tends to be highest for the smoothing splines.

Estimating the model variance

The MLE SSE/n is as usual too small.

If d is the dimension of the face on which the projection lands (including the $\boldsymbol{v}$ vectors), then we can show that $SSE/(n-d)$ is also too small.

In fact, we can show that

$$n - 2E(D) \leq \frac{E(SSE)}{\sigma^2} \leq n - E(D),$$

This is important when D can range from zero to $n-1$, but when the number of edges is small compared to n, the estimator $SSE/(n-d)$ is close to correct and much more stable than for standard shape-restricted regression.

Simulations results for estimating the model variance.

MLE uses standard shape-restricted regression and SSE/n.

IQRS uses increasing quadratic regression spline; $SSE/(n-d)$.

M-W uses standard shape-restricted regression and $SSE/(n-1.5d)$.

		IQRS		MLE		M-W	
n	σ	mean	std dev	mean	std dev	mean	std dev
20	0.1	.0994	.018	.035	.015	.240	.096
40	0.1	.0996	.012	.054	.011	.161	.062
80	0.1	.0998	.0083	.069	.0078	.111	.013
20	0.5	.488	.089	.341	.078	.574	.167
40	0.5	.494	.060	.397	.056	.523	.074
80	0.5	.497	.041	.433	.039	.509	.046
20	1.0	.972	.173	.611	.249	1.20	.547
40	1.0	.986	.118	.736	.192	1.08	.284
80	1.0	.992	.082	.825	.144	1.03	.180

ANCOVA modeling with regression splines

We can add a categorical covariate to the regression model. The simplest example is adding a dummy variable so that the model is

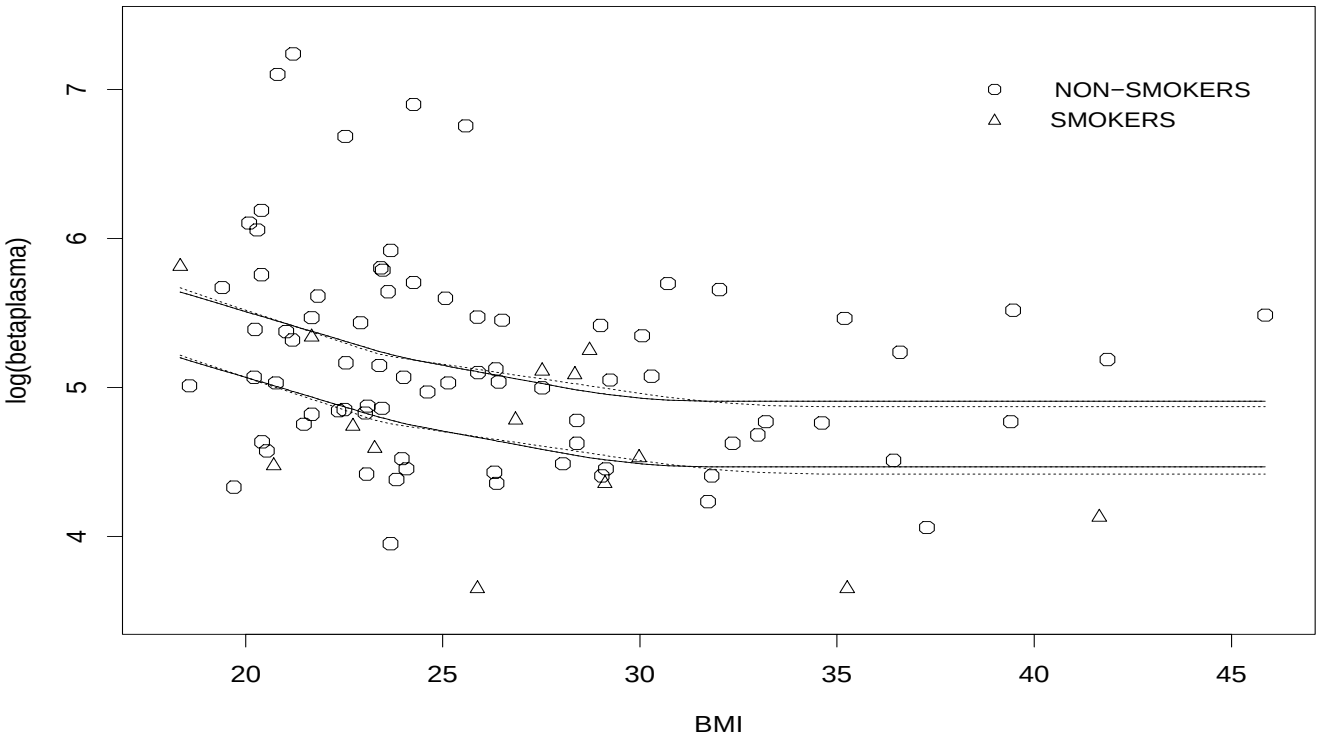
$$y_i = f(x_i) + \beta d_i + \sigma\epsilon_i, \quad i = 1, \ldots, n,$$

where d_i is either zero or cone according to the group of the ith observation. (For more groups, we add more dummy variables.)

The least-squares estimators for f and β can be found simultaneously.

Suppose we are also interested in whether smoking is associated with blood plasma levels of beta carotene. We can fit parallel decreasing convex regression splines.

The solid lines represent splines using five interior knots, and the dashed lines are splines using eight interior knots.



If the interest is in testing the null hypothesis that the categorical variable has no effect on the response, we can get an exact test statistic for balanced design models and known σ^2.

Consider the statistic

$$F = \frac{(SSE_0 - SSE)/(g-1)}{SSE/(n-g-d-r)}$$

where g is the number of groups, and SSE_0 is for the model without the categorical variable.

This would have an $F(g-1, n-g-d-r)$ density under H_0 if the shape restrictions were absent or not binding, and if the true regression function were contained in the linear space spanned by the edges and the vectors in V.

Assuming that the statistic has approximately the correct null density gives good results in simulations.

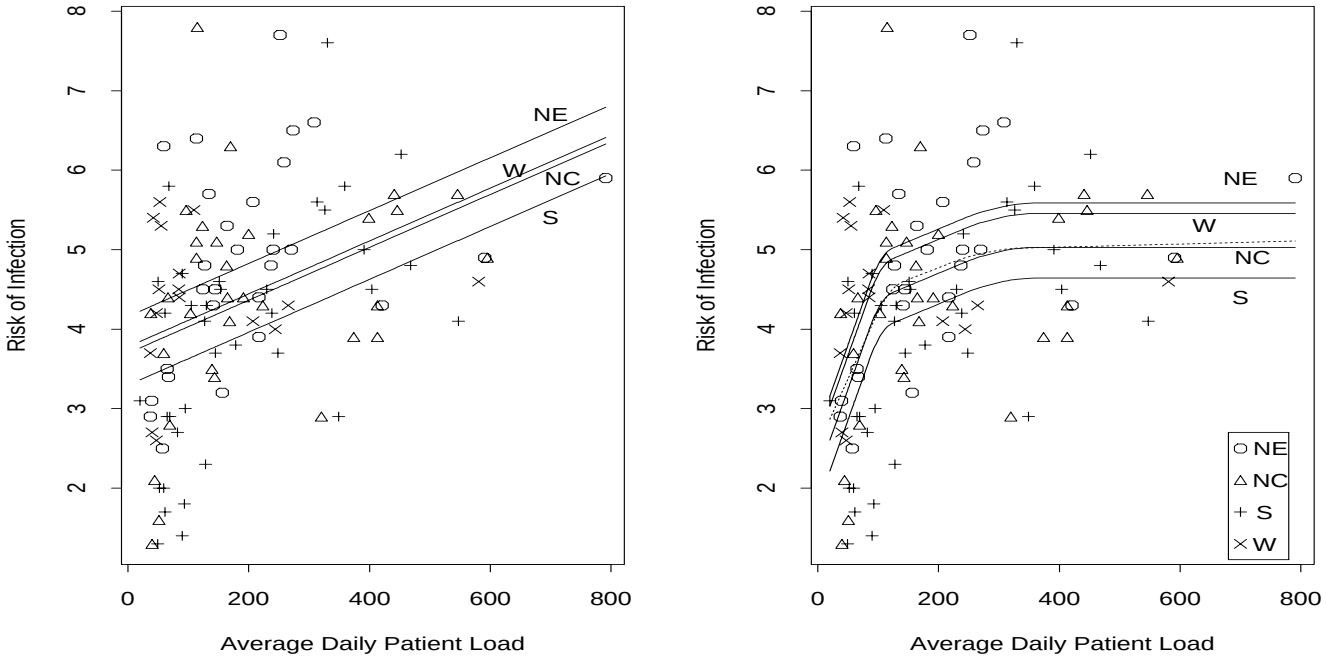
true f	n	g	σ	effect size	OLS $Pr(rejH_0)$	NP $Pr(rejH_0)$
linear	10	2	1	0	0.050	0.053
linear	20	2	1	0	0.050	0.051
linear	40	2	1	0	0.050	0.051
$\exp(x-1)$	10	2	1	0	0.046	0.053
$\exp(x-1)$	10	2	1	0	0.047	0.051
$\exp(x-1)$	10	2	1	0	0.047	0.051
$\exp(x-1)$	10	2	.4	0	0.032	0.051
$\exp(x-1)$	10	2	.4	0	0.033	0.051
$\exp(x-1)$	10	2	.4	0	0.034	0.050
linear	10	3	1	0,0	0.050	0.053
linear	20	3	1	0,0	0.050	0.051
linear	40	3	1	0,0	0.050	0.051
$\exp(x-1)$	10	3	1	0,0	0.046	0.053
$\exp(x-1)$	20	3	1	0,0	0.046	0.051
$\exp(x-1)$	40	3	1	0,0	0.046	0.050

Simulations results to determine test sizes for the ANCOVA models. Several sample sizes are chosen for either two or three levels of the categorical variable.

true f	n	g	effect size	OLS $Pr(rejH_0)$	NP $Pr(rejH_0)$
linear	10	2	1	0.500	0.505
linear	20	2	1	0.500	0.504
linear	40	2	1	0.500	0.503
$\exp(x-1)$	10	2	1	0.500	0.517
$\exp(x-1)$	20	2	1	0.500	0.509
$\exp(x-1)$	40	2	1	0.500	0.505
$\exp(x-1)$	10	2	0.5	0.500	0.552
$\exp(x-1)$	20	2	0.5	0.500	0.522
$\exp(x-1)$	40	2	0.5	0.500	0.512
$\exp(x-1)$	10	2	0.25	0.500	0.778
$\exp(x-1)$	20	2	0.25	0.500	0.591
$\exp(x-1)$	40	2	0.25	0.500	0.538
linear	10	3	0.25,0.25	0.500	0.498
linear	20	3	0.25,0.25	0.500	0.500
linear	40	3	0.25,0.25	0.500	0.501
$\exp(x-1)$	10	3	0.25,0.25	0.500	0.854
$\exp(x-1)$	20	3	0.25,0.25	0.500	0.618
$\exp(x-1)$	40	3	0.25,0.25	0.500	0.547

Simulations results for ANCOVA models. Power is compared with the F-test using a linear relationship between the response and the continuous predictor.

Hospital infection risk data, where infection risk is modeled as an increasing concave function of average daily patient load, for four regions of the U.S.



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. *Scandanivian Journal of Statistics*, **16**, 65-74.

[2] Haley R.W., Quade D., Freeman H.E., Bennett J.V. (1980) The Senic Project: Study on the efficacy of nosocomial infection control. *Am. J. Epidemiol.* **111(5)**, pp 472-85.

[3] Hastie T.J. and Tibshirani, R.J. (1990) *Generalized Additive Models*, Chapman & Hall/CRC, Boca Raton.

[4] Meyer M.C. (1999) An extension of the mixed primal-dual bases algorithm to the case of more constraints than dimemsions. *J. Statist. Plan. Infer.* **81**, pp 13-31.

[5] Meyer M.C. (2003) A test for linear versus convex regression function using shape-restricted regression, *Biometrika*, **90(1)** 223-232.

[6] Nierenberg DW, Stukel T.A., Baron J.A., Dain B.J., Greenberg E.R. (1989). Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology* **130** pp511-521.

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