Testing Competing Theories with Shape Constrained Inference[*]

Jonathan Wand[†]

April 12, 2012

[*] Prepared for the Midwst Political Science Association Meetings, Chicago, 2012.

[†] Department of Political Science, Encina Hall West, Stanford University.
Stanford, CA 94305. http://wand.stanford.edu, wand@stanford.edu.

1 Introduction

This paper considers the comparison of theoretical models based on their ability to specify an (a priori unknown) functional relationship between variables. A common prediction of social science models is that there exists a monotonic relationship between two variables. A wide range of additional functional relationships arise in comparative statics and informal theories, including unimodal or U-shaped curves, the symmetry of curves, and convexity or concavity.

Adjudicating among functional predictions requires solving two distinct challenges. First, it requires a method of estimation to find the best fitting curve that satisfies the predictions of each theory. Second, it requires a method of inference for comparing the relative fits of multiple models.

There is considerable research in each of these areas separately. Research into the estimation of functions subject to shape restrictions has a long history, beginning with tests for ordered means. Comparisons of models have been limited to tests of single shape restricted function versus a simple alternative, such as linear versus unconstrained means; linear versus monotonic, or monotonic versus a constant.

Model comparisons have a long history as well. They generally find the best model among a set (model selection) or testing a benchmark model against a set of alternatives (model validation). In the context of competing social science theories, it is desirable to avoid privileging one theory over another by treating one as a benchmark. Picking among models, even if we can identify the best model, the question is which (if any) of the alternative models fit significantly worse.

Making comparisons between models where one has shape constraints is difficult. The dimensionality of the parameter space of a shape constrained model is probabilistic rather than fixed (the number of free parameters can vary across samples). as such, the distribution of test statistics are mixtures that are often difficult to characterize. The framework I present here provides a natural means by which to compare both nested and non-nested shape constraints with arbitrary differences in the distribution of their dimensionality.

This paper develops a method for combining shape constrained estimation with multiple comparisons. Estimation is based constrained ordered mean, and shape constrained splines. Adjudication of models is based on the Model Comparison Set of Hansen, Lunde, and Nason (2011). Comparing the models hinges on obtaining an estimate of the effective degrees of freedom (Shibata, 1997). Constrained estimation is achieved by a constrained optimization on ordered means (Barlow, Bartholomew, Bremer, and Brunk, 1972) or sequence of spline coefficients (Wand, 2010).

2 Functional relationships

Consider a bivariate functional relationship, where we seek to characterize the average mapping between x and y,

$$E[y_i \mid x_i] = f(x_i)$$

Insights from theories are commonly about the sign of a first derivative over the the range of the x, or subintervals, and sometimes higher order derivatives as well. In the case of discrete data, we can similarly describe constraints on the relative order of the sequence of means in terms of finite differences. Extensions beyond the bivariate case are easily incorporated into the framework that follows.

A few examples serve to illustrate the idea of thinking about theories in terms of shapes and issues of testing alternative shapes.

2.1 Example: Campaign finance

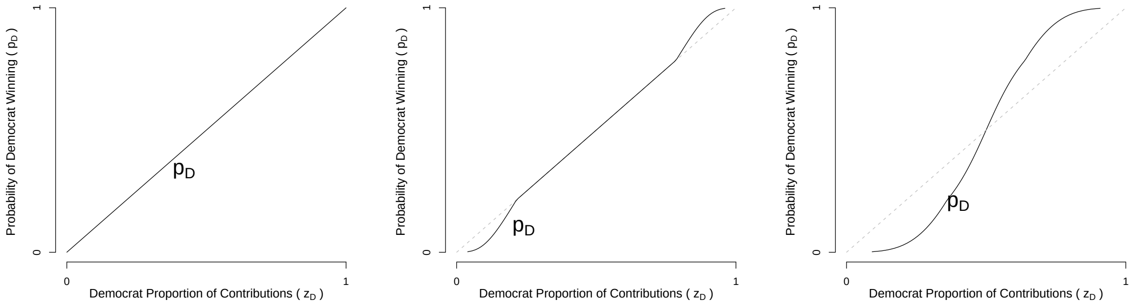
What can learn about the motives of PACs based on observing the relationship between the proportion of contributions raised by a Democratic candidate in a district, z_i, and the probability that the Democrat will win the election, p_i? The *shape* of the relationship between z_i and p_i offers the basis for distinguishing between a world exclusively of investors and another where there is a mixture of investor and partisan supporters.

Snyder (1990) investigated a model of investors and candidates, where candidates maximize their fundraising by selling all their available future services. The model implies that the equilibrium relationship between the proportion of money investors give to the Democrat in a district and the probability that she will win is. This function is illustrated in Figure 1(a). Snyder (1990) also considered another model of investors and candidates wherein candidates who have a high probability of winning reduce the amount of services they are willing to sell. As such, some candidates do not maximize their contributions from investors. The equilibrium relationship for this is illustrated in Figure 1(b).

An alternative to a model with only investor contributors is a model wherein candidates are funded by both investors and partisans. The equilibrium relationship between winning and proportions is illustrated in Figure 1(c). This comparative static can be derived generalizing a model with candidates supplementing investor money from parties (Baron, 1989), or from a joint game with both investors and partisan contributors.

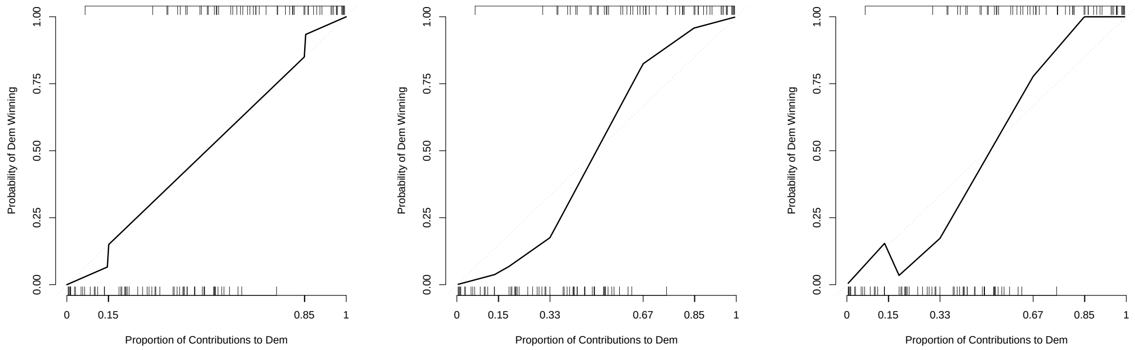
The differences between these three figures lie primarily in their qualitative shapes. The model in Figure (a) alone has an implication for parameter values and functional form: $p_i = \alpha + \beta z_i$, where $\alpha = 0$ and $\beta = 1$. In contrast, the precise form and location of the other figures will depend on details of the functional form of how money maps into votes, and the unknown ratio of investors to partisans.

Figure 2 shows the fitted values of three models; the most restrictive model, 1(a), actually



(a) Model 1(a) (b) Model 1(b) (c) Model 2

Figure 1: Equilibrium relationships between proportion of contributions to Democratic candidate and probability of Democrat winning in a district. Model 1(a) and (b) illustrate behavior among investors. Model 2 illustrates behavior among a group composed of both investors and partisans.



(a) Model 1(b): Symmetric, dip (b) Model 2: Symmetric, monotonic (c) Model 3: Unrestricted

Figure 2: Fitted probabilities from Bernoulli model of Democratic victory in a district by proportion of money raised by the Democratic candidate. Linear B-splines with knots at $k = (.148, .15, .85, .852)$ in 1(a) and 2(a); additional knots at (1/3,2/3) are included in 2(b) and 3(a). Rug at bottom and top of plots show location of observed Republican and Democratic victories, respectively.

requires no estimation, and simply equates the probability of winning to the proportion of money raised. Details of the method of fitting these curves is described in the applications section.

In each plot, the x-axis is the proportion of contributions in a district from economic PACs raised by the general election Democratic candidate. The y-axis is the associated fitted probability of the Democratic winning in the district. The rug of vertical dashes at the top at bottom and top of each plot shows the location of observed Republican and Democratic victories in districts, respectively. The dashed grey diagonal line marks the values at which proportions are equal to probabilities.

These curves were previously estimated and compared in an earlier paper (Wand, 2010). However, the means by which the models were compared left open questions, including how to address the issue of making multiple comparisons between pairs of models.

2.2 Example: child mortality and democracy

Testing theories based on the shape of a functional relationship extends to informal and qualitative hypotheses, as well as to functions of data that are not continuous. A common applied problem is testing theories that map ordinal scales to outcomes of interest. The types of theories that are tested are whether there is an significant positive or negative relationship between two variables.

For example, scholars have investigated whether there a connection between a country's democracy score and child mortality rate? Many have argued in the affirmative including Przeworski, Alvarez, Cheibub, and Limongi (2000) and De Mesquita, Smith, Siverson, and Morrow (2005). In contrast, Ross (2006) argues that with enough controls and imputed data, there is no relationship between democracy and child mortality rates.

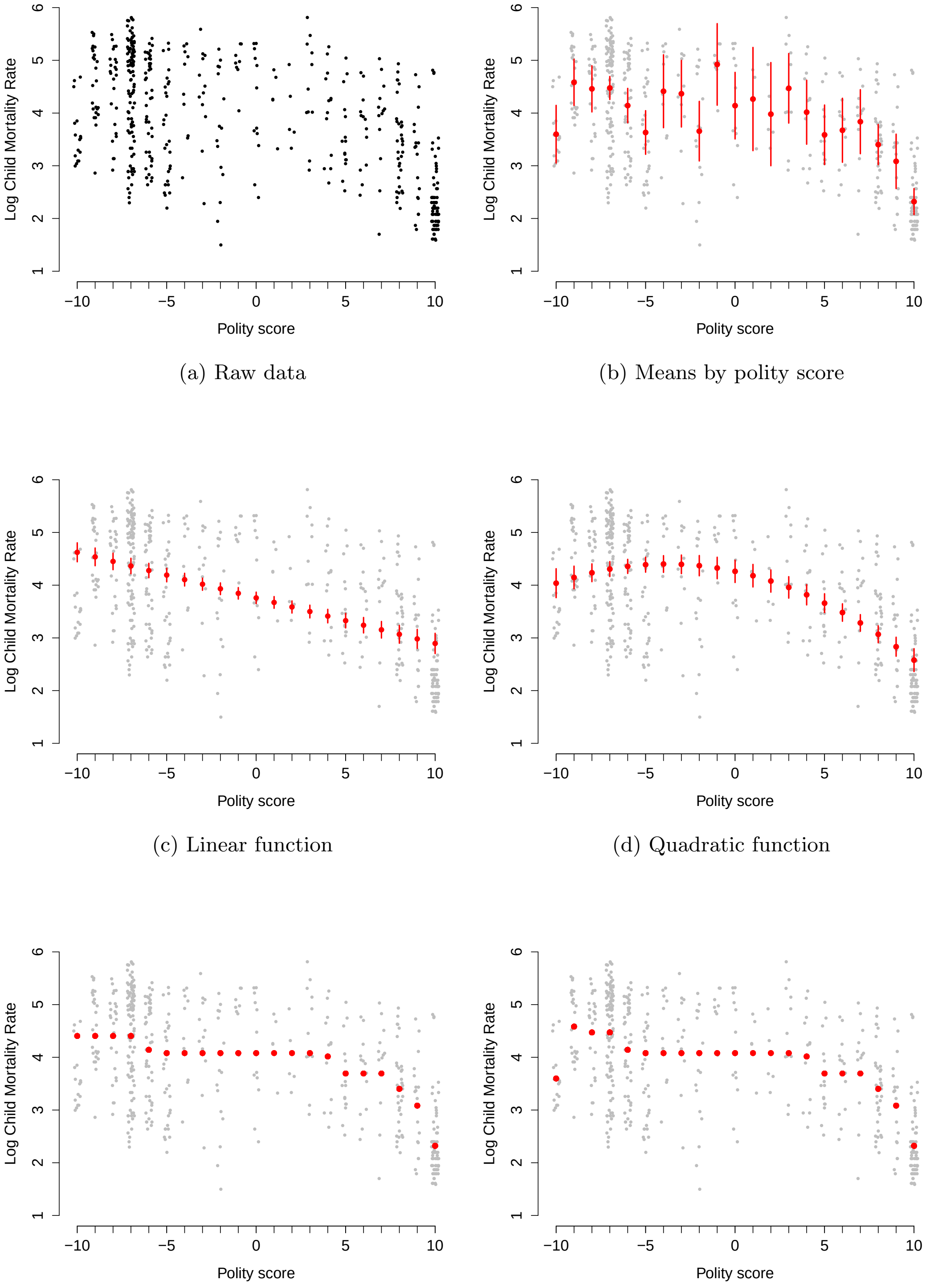
The raw data is plotted in Figure 3(a). The data is jittered along the horizontal axis in order to minimize obscuring data with similar values, but there is only 21 different polity score values. These are thought to be discrete and ordinal, with the least democratic countries at -10 and the most democratic countries at 10.[1]

of

With a fixed and finite set of categories, it is possible to summarize the data without imposing any particular structure, such as showing the means and confidence intervals of the means for each category. This is illustrated in (b). The appeal of this approach is that one can observe the variation over the polity scale in the average child mortality. The main reason that people do not use this, I would argue, is that there is not enough data reduction. Instead of a single summary statistic, there are 42 (21 means and CI). The undulating means moreover make describing the basic question of whether there is an increasing function difficult for a researcher to describe. A researcher is faced with the question of which, if any, of the observed differences in means are significant.

The simplest test would be to ask whether all the means are the same; this is a very useful null hypothesis to evaluate, but if rejected, it says nothing about which means are different or the shape of the relationship. One could make all possible pairwise comparisons, but this would require some

[1] There are also measurement problems with the ordinality of polity; see Treier and Jackman (2008).



(a) Raw data

(b) Means by polity score

(c) Linear function

(d) Quadratic function

(e) Best monotonic function

(f) Best unimodal function

Figure 3: Mapping polity to child mortality

adjustment for the multiple comparisons, and still there is the difficulty of correctly formulating tests using paired comparisons for evaluating the global question of a shape.

Ross (2006) considers whether the ordinal categories of Polity has a linear or quadratic relationship to child mortality rates (Figs (c) and (d). This is not special to Ross, and it turns out that treating ordinal data as if it contained metric information is not the biggest problem. The structure imposed by the parametric forms obscures a substantively important patterns in the data.

An alternative to the polynomial models is to ask whether the relationship is monotonic. Figure (e) shows the best fitting monotonic mapping. Monotonicity is what researchers actually say they want to test in most cases against the possibility that there is downturn. Figure (f) is the best fitting unimodal mapping. The question addressed in this paper is how to adjudicate which of the summaries in Figures (b)-(f) best describe the observed data.

3 Comparing models and the Model Confidence Set

In seeking to adjudicate among the shapes of curves, there are two challenges,

1. Often there are more than one possible model.

   As such, we are faced with making multiple comparisons among the competing models. Moreover, in many situations there is not a benchmark or leading candidate model that should be privileged as a null model, in a classical hypothesis testing sense.

2. Comparisons of shape constrained models are non-standard.

   In pairwise comparisons between models, the distribution of test statistics are complicated because models subject to shape constraints do not have fixed number of free parameters.

Here I sketch the key features of the method of selecting curves.

3.1 Model fit

The metric by which I assess the fit of the models is based on the Kullback-Leibler information criterion (KLIC). Let,

$$Q(\beta_j) = -2L(\beta_j)$$

the log-likelihood evaluated at parameters β_j. The population parameters β_{0j} minimizes the value of this function for model j. In the absence of shape constraints, for a correctly specified model, the difference between the fitted $\hat{\beta}_j$ and true Q value is distributed as,

$$Q(\hat{\beta}_j) - Q(\beta_{0j}) \sim \chi^2_k$$

with the expected degrees of freedom being k, the dimension of the parameter space β. This is the basis for the likelihood ratio test, AIC, and similar tests. In shape constrained models is probabilistic, $Q(\hat{\beta}_j) - Q(\beta_{0j})$ is a distributed as a mixture of χ^2 distributions; see the Appendix for details. I employ the bootstrap approach of Shibata (1997) to estimating the effective degrees of freedom of each model. For each bootstrap sample, $b = 1, ..., B$,

1. *treat $\hat{\beta}_j$ as the population parameter*
2. *sample $Z^*_b = (Y^*_b, X^*_b)$*
3. *calculate $Q(Z^*_b, \hat{\beta}_j) - Q(Z^*_b, \hat{\beta}^*_{b,j})$*

Then,

$$\hat{k}^*_j = B^{-1} \sum Q(Z^*_b, \hat{\beta}_j) - Q(Z^*_b, \hat{\beta}^*_{b,j})$$

The key here is to assume that we obtain a measure of the bias of KLIC under the assumption that the model is correctly specified by treating the estimated parameter values as the true values.

For shape constrained models, this provides the expected value of the mixture of χ^2 distributions under the null that the estimated model j is correct. In prior work, simulation approaches are also used to characterize the distribution of weights in the mixture of χ^2 distributions, yet the goal is to obtain the expected degrees of freedom; the method here avoids the intermediate steps and obtain this quantity directly.

Combining the effective degrees of freedom with the fitted Q, we have AIC$^* = Q(\hat{\beta}_j) + \hat{k}^*_j$, a variation on the classical AIC calculation. Like the classical AIC, this value could be used to rank the models. However, neither AIC nor AIC* characterize the confidence we have in the selected model being "best" among the alternatives.

3.2 Model comparison

Hansen et al. (2011) propose a method for identifying the set of best model(s), $\mathcal{M}^*$, from a set of potential models $\mathcal{M}_1$. The appealing properties of the MCS is that under weak regularity conditions, as the sample size increases,

- The estimated set of best models, $\hat{\mathcal{M}}^*$, contains the true best models with probability $1 - \alpha$, where α is the size of the test. If more than one best model, there is an α chance that at least one of the best will be rejected. I.e.,

$$\lim \inf_{n \to \infty} Pr(\mathcal{M}^* \subset \hat{\mathcal{M}}_{1-\alpha}) \geq 1 - \alpha$$

- As the sample size grows in the limit, all models not in $\mathcal{M}^*$ are eliminated. I.e., in the limit, the probability of the non-best model(s) being in the estimated set of best models, $\hat{\mathcal{M}}^*$ is zero.

Hansen observed that under the null hypothesis $H_{0,\mathcal{M}}$ that all models in $\mathcal{M}$ are equally good, then $E[Q(\beta_{0i}) - Q(\beta_{0j})] = 0$ for all $i, j \in \mathcal{M}$, and used this to motivate a test statistic

$$T_{\mathcal{M}} = \max_{i,j \in \mathcal{M}} \mid [(Q(\hat{\beta}_i) + k^*_i) - (Q(\hat{\beta}_j) + k^*_j)]$$

If $T_{\mathcal{M}}$ is large, then at least one of the models violates the null hypothesis. If the null hypothesis is rejected, then the model with the worst value $Q(\hat{\beta}_i) + k^*_i$ is eliminated from the MCS, and the process is repeated until the p-value associated with $\mathcal{M}$ is greater than α. The remaining models that were eliminated are the MCS, $\hat{\mathcal{M}}$

Characterizing the distribution $T_{\mathcal{M}}$ is achieved by simulation. The joint distribution for m models of

$$\{Q(\hat{\beta}_i) + k^*_i - Q(\beta_{0i}), ..., Q(\hat{\beta}_m) + k^*_j - Q(\beta_{0m})\}$$

is estimated by a bootstrap, taking the differences,

$$\{Q_b(\hat{\beta}^*_{b,i}) + k^*_i - Q_b(\hat{\beta}_i), ..., Q_b(\hat{\beta}^*_{b,m}) + k^*_j - Q_b(\hat{\beta}_m)\}$$

where $Q_b(\beta)$ is -2Log-likelihood for bootstrap sample b. Thus the bootstrap is generating the difference between Q fitted to the bootstrap data set and the fit conditional on the parameters estimated from the original sample $\hat{\beta}$. Thus the bootstrap is generated under the null assumption that all the models in the current set $\mathcal{M}_j$ are true.

4 Monte Carlo Study

I examine the finite sample performance of MCS and alternatives in a set of Monte Carlo studies. There are two main questions addressed here. First, how well do different methods perform in selecting true model? Here, I compare MCS using AIC^*, AIC^* without MCS, and AIC. Second, does a polynomial model correctly characterize the qualitative features of the curve?

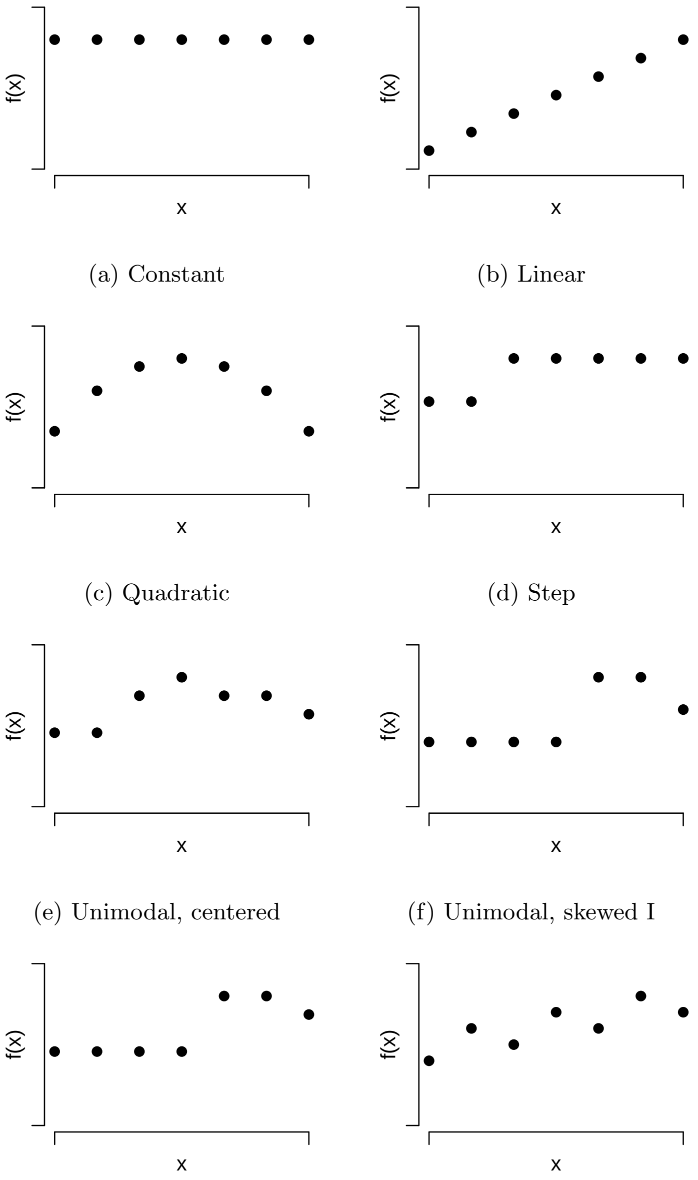
4.1 Design

The DGP are shown in Figure 4, and details are given in the Appendix. In all these cases, x takes on integer values between 1 and 7. Although the model $y = f(x)$ will be observed only at discrete values of x, the mean values in models (a)-(c) are consistent with polynomial models. Models (d)-(g) do not conform to (simple) polynomials but are either monotonic or unimodal. Model (g) has no simple structure to the sequence of means.

Matching the range of data generating processes, I estimate separate models which each adhere to one of the following hypotheses about the mean, $\mu_j = E(y \mid x = j)$:

$$\begin{array}{lll} H_0 & : \mu_1 = ... = \mu_j = \mu_{j+1} = ... = \mu_k & \text{(constant)} \\ H_1 & \mu = \beta_0 + x\beta_1 & \text{(linear)} \\ H_2 & \mu = \beta_0 + x\beta_1 + x^2\beta_2 & \text{(quadratic)} \\ H_{\nearrow} & : \mu_1 \leq ... \leq \mu_j \leq \mu_{j+1} \leq ... \leq \mu_k & \text{(mono inc.)} \\ H_{\cap} & : \mu_1 \leq ... \leq \mu_j \geq \mu_{j+1} \geq ... \geq \mu_k & \text{(umbrella)} \\ H_A & : \mu_j \gtreqless \mu_{j'} & \text{(unconstrained)} \end{array}$$

The sample size in each MC is $N = 500$. The number of simulations for each DGP is $B = 250$.



(a) Constant

(b) Linear

(c) Quadratic

(d) Step

(e) Unimodal, centered

(f) Unimodal, skewed I

(g) Unimodal, skewed II

(h) Oscillating

Figure 4: Shapes of $f(x) = E(y \mid x)$ in Monte Carlo studies. $y = f(x) + \epsilon$, where $\epsilon \sim N(0, 1)$

	Frequency selecting true model			Rejecting	
True shape	$\hat{\mathcal{M}}_{1-\alpha}$	AIC^*	AIC	linearity	monotonicity
constant	97.2	82.8	81.2	4.0	4.8
Linear	96.4	77.6	78.0	4.0	0.0
Step	93.2	72.8	0.0	100.0	99.2
Quadratic	96.0	93.2	93.2	100.0	100.0
Unimodal 0	95.2	83.6	0.0	100.0	100.0
Unimodal 1	95.2	81.2	0.0	19.6	0.0
Unimodal 2	95.2	81.2	0.0	12.8	0.0
Oscillating	100.0	100.0	100.0	99.2	0.8

Table 1: Monte carlo results for selecting correct model, and for hypothesis tests on polynomial models.
Notes: Test size is $\alpha = 0.05$ for MCS selection, and the tests of linearity and monotonicity.
Sample size of each MC is $N = 500$, number of simulations per model is $B = 250$.

4.2 Model selection

Results from the MC analysis are presented in table 1. The first three columns show the percentage of simulations wherein the correct model was identified. The last two columns show the percentage of simulations wherein the hypothesis of linearity is rejected, and the hypothesis of monotonicity is rejected.

MCS performs well, correctly identifying the functional relationship respecting the true shape with probability greater than $1-\alpha$. In contrast, using AIC^* performs considerably worse in almost every case. Simply correcting for the expected bias of KLIC for each model is not enough to provide the correct ordering of models. And yet it is important to note the correction alone is responsible for a vast improvement over a standard AIC calculation. For simulations which are not based on a polynomial form, AIC never selected the correct model.

The proper performance of the MCS is not simply a matter of whether the true model is included in $\hat{\mathcal{M}}_{1-\alpha}$, but also whether it excludes the wrong model. In short, MCS always excluded the wrong model in all but one set of the simulations, and in that one simulation it included the wrong model less than 1 percent of the time.

The "wrong" models is not simply all models other than the DGP, since some shapes encompass others. A monotonic curve can be fit by a unimodal curve which places the modal at an extreme of the range of x. A linear curve can be fit by a monotonic curve, although the later uses more parameters. Similarly an inverted-U shape curve can be fit by a unimodal curve. I enumerate here the models wrong models for each DGP in Table 2.

For no parameter values can these wrong models represent the DGP used in the particular simulation. Again, in almost no cases does the estimated best MCS include the wrong model.

True model	Wrong model(s)
Constant	none (all models have constant as special case)
Linear	constant
Step function	constant, linear, quadratic
Quadratic	constant, linear, monotonic
Unimodal	constant, linear, monotonic, quadratic
Oscillating	all models listed above

Table 2: Enumeration of wrong models for each true model.

4.3 Polynomials and inferring shapes

The logic of using polynomial in regressions is often used to argue for or against the presence of a non-monotonicity in the data. It is of some interest whether tests based on simple polynomials reliably identify the qualitative features of monotonicity or unimodality. The MC studies offer examples of just how poorly polynomials can perform in the task of providing inference about basic qualitative features of a functional relationship.

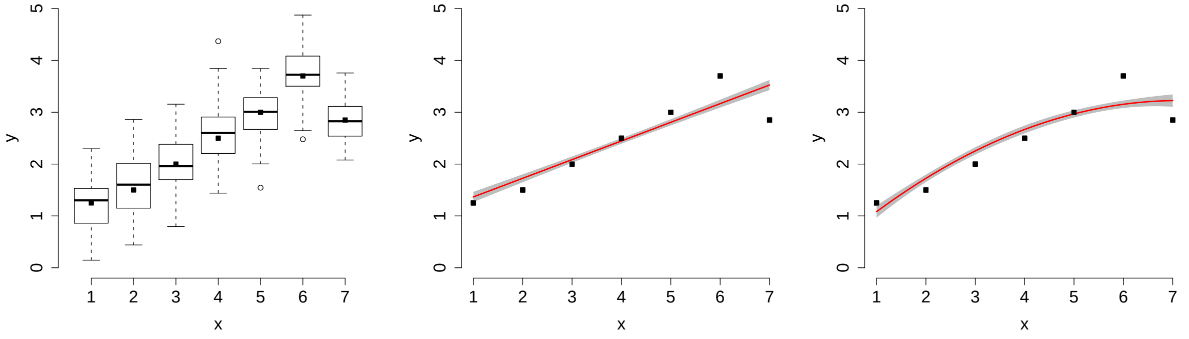
The performance is as we would expect in cases where the data is generated by a polynomial. In the case of a linear relationship, the polynomial performs as we would expect. Given a test level $\alpha = 0.05$, the true model was rejected against a quadratic alternative in the MC at about the correct rate (0.04). Even in these cases, however, monotonicity was never rejected. Conversely, in the case of a quadratic relationship, linearity and monotonicity are always rejected.

The polynomial based tests do not perform well more generally. In the case of step function, which is weakly monotonic, tests on a polynomial almost always falsely rejected monotonicity. In the lopsided unimodal curves, the polynomial models falsely fail to reject non-monotonicity. Figures illustrate what is going on in these case.

Data from a unimodal function MC is plotted in Figure 5(a). The true means are indicated by the solid squares, and are increasing and then decreasing. The box-plots show the median and the interquartile range of a simulated set of draws from these means. In (c), we observe that the quadratic function is incapable of both characterizing the increasing means and the downturn, and the hypothesis of monotonicity is not rejected. This difficulty in detecting a downturn is not simply a problem of small finite samples, but rather the relative inflexibility of the polynomial and the requirement that the same polynomial describe all the data. More flexible alternatives (e.g., splines, local regressions) exist, but they have their own issues in data with a small number of ordered categories in a regressor that is commonly used by researchers.

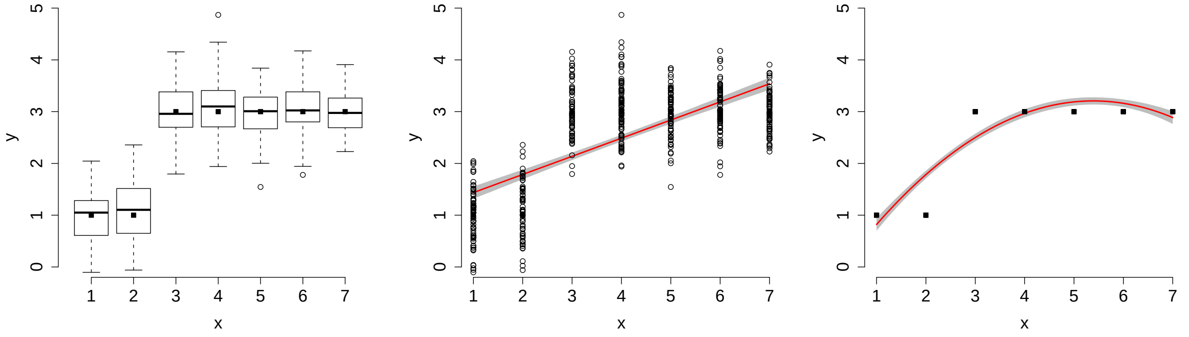
Failure to detect a downturn might seem a modest problem. However, including a quadratic term in a model can also lead us to the incorrect inference regarding the monotonicity of a relationship. Failing to find a pattern there is something we are accustomed to; finding a pattern that does not exist is far more troubling.

Consider a weakly monotonic process, indeed a single mean shift wherein, A summary of data



(a) Raw data and descriptives (b) Linear model (c) Quadratic model

Figure 5: Non-monotonic $f(x)$: skewed unimodal



(a) Raw data and descriptives (b) Linear model (c) Quadratic model

Figure 6: Weakly monotonic $f(x)$: step function

generated from this process is plotted in Figure 6(a). The appeal of a linear model is the simplicity of estimation, and ability to summarize and interpret the estimated quantity. In this case the linear model also happens to come to the correct view that the data is increasing (and monotonic), although it obviously misses the fact the the data is actually compose of only two distinct mean values.

Yet a researcher would not infer that the process is monotonic if they ran a higher order polynomial. The fitted values of the quadratic is able to come much closer to the means as shown in Figure 6(c), and yet would lead us to incorrect inference regarding the monotonicity of the mapping between x and y. Even with modest samples sizes, we would almost always reject the monotonicity of the fitted quadratic model.

Fitting polynomials via least squares (or other loss functions) is not concerned with the shape. The goals is to reduce the the difference between fitted and observed values, and this may imply shapes that are a property of the process generating the data.

There is rarely a good reason to use polynomials, particularly when a covariate takes on ordinal values. The appeal of polynomials mainly lies in the easy of summarizing the data in terms of direction of change and the average change by value of x. While these are useful pieces of information to report, the additional piece of information on the shape of the functional relationship hinges on either analyzing the shape of the sequence of means or evaluating the shape based on a flexible estimator, such as the B-spline (Wand 2010).

5 Applications

5.1 Theories of campaign contributions

Using data from 1982–1986 open House seats, Snyder (1990) argued that the pattern of contributions from economic Political Action Committees (PACs) were consistent with these groups acting as investors, like Figure 1(b).[2] In this section, I replicate Snyder's findings using the B-spline framework, and compare the shape implications of an investor-only hypothesis and an alternative hypothesis that contributions from economic PACs are motivated by a mixture of partisan and investor interests.

Here I consider five nested models are considered with varying constraints on the regressions parameters. The dip model, 1(b) impose $z = p$ only between (0.15, 0.85). The curve is constrained to be monotonic, symmetric, and have $p = 0$ at $z = 0$ (and by symmetry, $p = 1$ at $z = 1$). Additional knot locations (0.148, 0.852) were selected by maximizing the likelihood of the dip model; so in all knots are located at (0.148, 0.15, 0.85, 0.852). The models are fitted by maximizing a Bernoulli likelihood with the fitted probabilities produced by the linear B-spline regressions constrained to lie in [0,1]. The knot locations in this application are chosen to maximize comparability with the prior study by Snyder (1990). The knots includes (0.15, 0.85) which he argues are the location of the beginning of the dips, and (1/3, 2/3) which he uses to uses as an omnibus specification test against the linear model.

Model 2(a) removes the restriction that $z = p$ between (0.15, 0.85), but still requires the estimated probabilities to be monotonic. The second monotonic model, 2(b), allows the derivatives to change at (1/3, 2/3) but is otherwise the same as 2(a). Both of these models conform to the type of restrictions implied by Figure 1(c). Two unrestricted versions of the preceding models are also estimated, 3(a) and 3(b), differing respectively only by the exclusion and inclusion of knots at (1/3, 2/3). Note that all of the models are nested within 3(b), while 2(b) alone is not nested within 3(a).

While the estimated model 1(b) fits a noticeable dip, removing the restriction that $z_i = p_i$ between (.15, .85) with model 2(a) leads to a curve that moves in the direction toward the S shaped curve seen in Figure 1(b). Allowing changes in derivatives at (1/3, 2/3) with model 2(b) moves

[2] Economic PACs is used as a short hand for PACs that declare to the FEC sponsorship by a corporation, union, cooperative, or Trade/Health/Membership group.

			Probability of $\bar{\chi}^2 = -2(L_{\text{row}} - L_{\text{column}})$		
	Max. Number	Log-lik.	j vs 1(a)	j vs 1(b)	j vs 2(a)
Model (j):	of Parm (m_j)	L_j	Linear Null	Dips Null	Mono. Null
1(a): Linear Equality	0	−47.03			
1(b): w/ symmetric dips	1	−46.59	0.18		
2(a): Symmetric, monotonic	2	−44.44	0.06	0.03	
2(b): w/ knots at $(\frac{1}{3}, \frac{2}{3})$	3	−43.81	0.07	0.04	0.48
3(a): Unrestricted	6	−42.92	0.22	0.89	0.34
3(b): w/ knots at $(\frac{1}{3}, \frac{2}{3})$	8	−42.46	0.99	0.89	0.34

Table 3: Model fit statistics and comparisons.

further still toward an S shape. Removing all shape restrictions, Model 3(b) has an upward twist in districts where Democrats got a small proportion of contributions, but the shape of the curve is not fundamentally changed from 2(b) in districts where they received a majority of contributions.

The differences between the fits of the models are summarized in Table 3. The first column shows the maximum number of unconstrained parameters that are estimated, and the second column shows the log-likelihood of the model. The question is how to compare the fitted models.

Pairwise approach

One approach to testing would be to perform pairwise comparisons, ignoring the issue of multiple comparisons. The remaining cells in Table 3 show the probabilities from the hypothesis tests comparing the model in a row to the null model specified in the column heading.

Against the null of the linear model, one cannot reject the dip model or the unrestricted models. This is consistent with the conclusions of Snyder (1990). However, a researcher could reasonably treat the difference between the linear model and the monotonic models as significant, p=.06 for 2(a) and p=.07 for 2(b). How does one reconcile these results? As previously noted by Snyder (1990) the dip model only has a small fraction of the data identifying any difference between it and the linear model. The effective sample size of the comparison is only 36 because between (.15, .85) the observations contribute nothing to the difference in the likelihoods. The unconstrained models have 6 and 8 parameters, so the probability of seeing χ^2 values at least as big as the observed 8.2 and 9.1, respectively, is relatively large. Compared to the unrestricted models, the number of free parameters in symmetric monotonic models is cut by more than half while $\bar{\chi}^2$ values are reduced by less than half to 5.2 and 6.4. As such, we see the flexibility of the unrestricted models is unproductive, and the additional parameters impede our ability to discern, from a hypothesis testing perspective, a significant pattern in the data that deviates from the linear model.

I also consider as alternative hypotheses two other models in the last columns of the table. The null hypothesis of the dip model 1(b) is rejected against either of the monotonic models, $p = 0.03$ and $p = 0.04$, while again the over-parametrized and unrestricted model leads to no rejection of

Iteration j	H_{0,M_k} p-value	MCS p-value	$\hat{\mathcal{M}}_j$	Model to eliminate
1	0.034	0.034	Dip, Linear, Symmetric-mono, Unrestricted	Dip (1b)
2	0.032	0.034	Linear(1a), Symmetric-mono, Unrestricted	Linear (1a)
3	0.292	0.292	Symmetric-mono (2b), Unrestricted (3b)	

Table 4: Sequence of MCS tests

the null. The null hypothesis of the monotonic model 2(a) which omits knots at (1/3, 2/3) cannot be rejected against either the more flexible monotonic model 2(b) or the unrestricted models.

What one may note form this approach, beyond the length it requires to discuss the alternatives, is two troubling features. To make any comparison requires choosing a particular model as the null model, despite the fact is no natural order in which to perform the comparisons. Second, it is not clear how to adjust the size of the tests for the multiple comparisons being made, since the order in which the comparisons are made, and indeed the number of comparisons were idiosyncratic to the way in which I sought to compare results with prior research.

MCS approach

The MCS approach avoids these issues, and yet leads us to the same conclusion: the symmetric monotonic curve is among the best fitting models, and this is not worse fitting than an unrestricted curve. Table 4 lists the iterative sequence of hypothesis tests on a proposal for a set of best models, and the associated p-values, and the model to be eliminated. I restrict the analysis to the linear and dip models, along with the monotonic and unrestricted curves which add knots at (1/3,2/3).

The H_{0,M_k} p-value column reports the p-value for whether the set $\hat{\mathcal{M}}_j$ includes only the best models. the MCS p-value is the cumulative maximum of the test of the tests on each subset. $\hat{\mathcal{M}}_j$ reports the set of models that is evaluated by H_{0,M_k} p-value. When H_{0,M_k} p-value is less than $alpha = 0.05$, the worst model in the set eliminated; the final column reports the names of the model that will be eliminated prior to the tests in the next row.

In the first two iterations, the dip and then the linear model are eliminated. However, the symmetric-monotonic model is the best theoretically motivated curve, and is not worse than unrestricted model.

5.2 Democracy and Child welfare

Consider what we would learn from examining the relationship between polity scores and child mortality using polynomial regressions. The fits for each variant of the model are shown in Table 5, and the sequence of MCS tests are shown in Table 6

We find a significant decline in mortality as we move from the least democratic countries to the most democratic countries, as reflected in the linear model. The quadratic model fits better than

	Q	AIC	AIC*	k	k*
Constant	1575.5	1579.5	1578.5	2	1.5
Linear	1314.5	1320.5	1320.5	3	3.0
Quadratic	1253.2	1261.2	1261.9	4	4.4
Unrestricted	1184.3	1228.3	1230.6	22	23.1

Table 5: Fit statistics for polynomial models of child mortality and polity

Iteration				Model to
j	H_{0,M_k} p-value	MCS p-value	$\hat{\mathcal{M}}_j$	eliminate
1	< 0.001	< 0.001	Constant, Linear, Quadratic, Unrestricted	Constant
2	< 0.001	< 0.001	Linear, Quadratic, Unrestricted	Linear
3	< 0.001	< 0.001	Quadratic, Unrestricted	Quadratic
4	1	1	Unrestricted	

Table 6: MCS tests restricted to polynomial models of child mortality and polity

				Model to
j	H_{0,M_k} p	MCS p	$\hat{\mathcal{M}}_j$	eliminate
1	< 0.000	< 0.000	Constant, Mono, Linear, Quadratic, Unimodal, Unrestricted	Mono
2	< 0.000	< 0.000	Constant, Linear, Quadratic, Unimodal, Unrestricted	Constant
3	< 0.000	< 0.000	Linear, Quadratic, Unimodal, Unrestricted	Linear
4	< 0.000	< 0.000	Quadratic, Unimodal, Unrestricted	Quadratic
5	0.084	0.084	Unimodal, Unrestricted	

Table 7: Sequence of MCS tests for full set of models of child mortality and polity

the linear model with a reversal of the general increasing mortality as democracy diminishes, with the least democratic states appearing better off.

Despite the significance of the coefficients in these models, we can reject all of the polynomial models against an unrestricted set of means for each of the polity score categories. By an LRT test, AIC, AIC*, or the MCS test, the unrestricted model fits significantly better than the parametric models.

However, we come to a substantively important different conclusion when we add into the comparisons a monotonic and a unimodal curve as alternatives. The results of this expanded MCS test is shown in Table 7. The monotonic curve is eliminated first—the function is not simply increasing, and the less restrictive model is less efficient than the linear model at giving us a misspecified model of the data. The unimodal model remains standing in the end. While the MCS p-value is on the edge of significance, relative to the quadratic model it fits the data much better.

6 Concluding Remarks

In concluding, I address key questions about the applicability and generalizability of the methods in this paper.

How often do theories speak to shapes?

In the absence of a theory, imposing derivative/shape restrictions on an estimated relationship between variables might be puzzling. Commenting on early work employing shape restrictions in smoothers, Hastie and Tibshirani (1988, 451) bluntly ask "why monotonicity?" in the context of exploratory smoothing analysis. However, many theories that occupy the interest of political scientists posit a relationship beyond a single mean difference and (implicitly or explicitly) involve a hypothesis about a shape.

In terms of testing formal models, it is with regards to comparative statics that a flexible but shape constrained machinery is most clearly needed. The derivation of monotone comparative statics (Milgrom and Shannon, 1994) provides a fruitful approach to developing testable implication about relationships that avoid relying on ancillary functional form assumptions. However, the current number of political science applications are rare (Ashworth and Mesquita, 2006).

In more general terms, the key question is not whether our current theories provide adequate guidance as to the shape of relationships, but rather whether raising the possibility of testing shapes bring to light new insights within existing theories or stimulates the formulation of new theoretical analysis.

When are fully automated non-parametric methods useful?

There exist sophisticated methods for fitting curves in a data driven manner. However a curve is estimated, a researcher will still face the challenge of testing shapes. When an automated procedure finds a pattern that does not conform to any known theory, the question at the core of this paper will remain: how different is the discovered shape from other shapes that are motivated by theory? In so far as the researcher has a hypothesis, it is these insights that need to be evaluated as part of the data analysis. The framework here is designed to elicit the researcher's insights and translate them into meaningful restrictions about which we would like to evaluate. Except in the context of purely exploratory summaries of the data, the empirical engagement of competing explanations for social, economic, and political processes will remain a "mixture of rigorous theory, experienced judgment, and inspired guesswork" (Achen, 1982, 79).

A lesson from the campaign finance application is the importance of comparing theoretically motivated alternative hypotheses. But conditional on selecting a best fitting model, it is also important to know whether any constraints imposed by the model are doing violence to the data. It is as a omnibus specification test that unrestricted semi- or non-parametric model are useful; see Yatchew (2003, section 6.4) for a recent review of specification testing.

Where should we look for complexity in our models?

The complexity in most empirical studies lies in the dizzying array of covariates that are included. The approach adopted here takes a different approach, which is to attempt to understand theories in so far as they imply non-trivial relationships between a small number of variables. In the case of campaign finance, this suggests researchers look at a subset of cases (open seat races) which are most directly applicable to the theory. In the case of models of roll rates, the key questions at stake hinge on only a handful of variables: future advances in refining our understanding are likely to follow from studying substantively interesting subsets of final passage votes rather than adding covariates.

It is possible to combine the method of testing shapes conditional on the partial effects of other covariates. The easiest approach to combining multiple unidimensional shape constraints or other unconstrained functions being the framework of generalized additive models (Hastie and Tibshirani, 1986). However, the approach of this essay compounds the virtue of studying parsimonious models Achen (2002). The importance and potential rewards to studying a tractable number of variables increases when seeking to understand complex, often non-linear relationship between variables.

7 Appendix

7.1 Linear approximations in regressions

The coefficients in the ordinary least squares equation

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}, \dots, \beta_k X_{ki} + \epsilon_i \tag{1}$$

can be interpreted as (approximately) the partial derivatives of a (possibly) non-linear function evaluated at the average value of explanatory variables $(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_w)$,

$$Y_i = f(X_{1i}, X_{2i}, \dots, X_{wi}) \quad i = 1, \dots, n \tag{2}$$

for $k \leq w$.

Achieving linearity hypothesis assumes that second and higher order derivatives are negligible in magnitude or that the observed changes in X around the means are small.

The exclusion of $w-k$ variables relies on either the zero deviation condition $(X_{ji} - \bar{X}_j = 0 \; \forall i)$ or an orthogonality assumption (the cross partial derivatives with the k included variables disappear). The intuition for this interpretation is presented in Cramer (1969,79–83).

Bivariate regression

We are are interested in summarizing how much a change in x alters the value of $y = f(x)$. We can approximate the value of the function near a particular value $\bar{x}$ based on the size of a small change $x - \bar{x}$ using,

$$f(x) \approx f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + f''(\bar{x})(x - \bar{x})^2/2 \tag{3}$$

which is a univariate second-order Taylor expansion of g around $\bar{x}$. If g is linear with slope β then the derivatives beyond the first are equal to zero, and the change in the function takes on the simple and exact form,

$$f(x) - f(\bar{x}) = f'(\bar{x})(x - \bar{x}) = \beta(x - \bar{x}) \tag{4}$$

and we can rearrange

$$y = f(\bar{x}) + \beta(x - \bar{x}) = \beta_0 + \beta(x - \bar{x}) \tag{5}$$

where $\beta_0 = f(\bar{x})$. This is illustrative of an interpretation of the slope coefficient in a bivariate regression.

Multivariate case

A useful point around which to expand the function is the point where each explanatory variable is set to its mean $(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_w)$, thereby defining the outcome value,

$$Y_0 = f(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_w)$$

The choice of these points around which to expand the function will become clear in the context of regression.

$$Y_i = Y_0 + \frac{\partial f(\bar{X})}{\partial X} dx + \frac{1}{2} dx' H f(\bar{X}) dx \tag{6}$$

where $dx = (X_1 - \bar{X}_1, \ldots, X_w - \bar{X}_w)$. Dropping the higher order derivatives, and expanding the notation we have

$$Y_i = Y_0 + \frac{\partial f(\bar{X})}{\partial X_1}(X_{1i} - \bar{X}_1) + \cdots + \frac{\partial f(\bar{X})}{\partial X_w}(X_{wi} - \bar{X}_w) + \epsilon_i \tag{7}$$

Replacing the notation of the partial derivatives,

$$\beta_j = \frac{\partial f(\bar{X})}{\partial X_j} \tag{8}$$

and

$$\beta_0 = Y_0 - \sum_{j=1}^{k} \beta_j \bar{X}_j \tag{9}$$

7.2 Testing location of maxima in polynomials

For the purpose of illuminating the limits of traditional methods, it is useful to consider the modal political science model of x and y, and the use of polynomial function in least squares regressions. The quadratic model has the standard form,

$$y_i = \beta_0 + x_i \beta_1 + x_i^2 \beta_2 + \epsilon_i \qquad x_i \in \{1, ..., k\} \tag{10}$$

Although simple in form, the quadratic model has posed challenges to researchers a number of challenges. Rarely is the point estimate of the location of the extremum determined, even less often is there an attempt to characterize the uncertainty about the location of the extremum. Scholars are content to evaluate the individual significance of β_1 and β_2, which are necessary but not sufficient for answering questions about the shape of the curve.

The quadratic model has extremum where slope of curve is zero as a point on real number line (rather than an index value). The extremum is found at the solution to the FOC of the quadratic function,

$$\arg\max_x f(x) = \{x : dy/dx = 0\} = -\beta_1/(2\beta_2)$$

If $\hat{\jmath}_{max} < 1$ or $\hat{\jmath}_{max} > K$ then there is no interior extremum, and instead the extrema on teh support of the data is at each end of the ordinal scale, and curve if monotonic. This point is often missed. The suggestion by Stimson, Carmines, and Zeller (1978) for an alternative more intuitive representation of the quadratic model has not been widely adopted; likely because of the additional complexity for the researcher.

Note that the location of the maximum is a function of two estimated quantities. Confidence interval around point estimate can be calculated by the Fieller or Delta methods. Can also motivate tests based on significance of derivatives of the quadratic function, finding interval where first derivative is different from zero. The derivative of quadratic is a linear function of the parameters, and hence the the point wise standard error is actually easy to calculate; for an application see Lind and Mehlum (2010). The Union-Intersection test of derivatives at endpoints of x is,

$$\min |\frac{\hat{\beta}_1 + 2\hat{\beta}_2 x}{V(\hat{\beta}_1 + 2\hat{\beta}_2 x)}| > t_\alpha$$

7.3 Defining functions with B-splines

The B-spline decomposition of a variable provides a flexible method for describing an arbitrary function, $y_i = f(x_i)$. In general, a B-spline has the virtues that (a) it is able to describe a broad class of relevant shapes; (b) it is able to define the shape constraints by a relatively small number of parameter restrictions; and (c) in some cases enables competing shapes to be defined by nested restrictions on parameters. These features make testing competing theories particularly tractable. The use of splines for approximating and estimating particular shapes has a long history in statistics, engineering, and economics (Laurent, 1969; Wahba, 1973; Wright and Wegman, 1980; Beliakov, 2000). I summarize here only the properties are that are key to their use in estimating and testing shapes. Comprehensive treatments of B-splines are provided by De Boor (1978) and Dierckx (1993).

A B-spline decomposition transforms a variable x into a set of truncated basis functions, with the m_{th} basis written as,

$$h_{m,k+1}(x) = (\lambda_{m+k+1} - \lambda_i) \sum_{j=0}^{k+1} \frac{(\lambda_{m+j} - x)_+^k}{\prod_{l=0,l\neq j}^{k+1} (\lambda_{m+j} - \lambda_{m+l})}$$

The k and λ parameters determine the smoothness and flexibility of the curve. The individual $\{\lambda_1, ..., \lambda_M\}$ are a sequence of "knot" locations which specify locations along x where the derivatives of $f(x)$ can change. Weighting the mth truncated power function $h_m(x)$ by a coefficient θ_m and

	Restriction	Description
1)	$\theta_m - \theta_{m-1} > 0$	$f(x)$ increasing between $(\lambda_{m-1}, \lambda_m)$
2)	$\theta_m - \theta_{m-1} = 0$	$f(x)$ flat between $(\lambda_{m-1}, \lambda_m)$
3)	$\theta_m - \theta_{m-1} < 0$	$f(x)$ decreasing between $(\lambda_{m-1}, \lambda_m)$
4)	$\frac{\theta_{m+1} - \theta_m}{\lambda_{m+1} - \lambda_m} = \frac{\theta_m - \theta_{m-1}}{\lambda_m - \lambda_{m-1}}$	$f(x)$ linear between $(\lambda_{m-1}, \lambda_{m+1})$
5)	$\frac{\theta_{m+1} - \theta_m}{\lambda_{m+1} - \lambda_m} > \frac{\theta_m - \theta_{m-1}}{\lambda_m - \lambda_{m-1}}$	$f(x)$ convex between $(\lambda_{m-1}, \lambda_{m+1})$
6)	$\frac{\theta_{m+1} - \theta_m}{\lambda_{m+1} - \lambda_m} < \frac{\theta_m - \theta_{m-1}}{\lambda_m - \lambda_{m-1}}$	$f(x)$ concave between $(\lambda_{m-1}, \lambda_{m+1})$

Table 8: Properties of $f(x)$ via restrictions on linear B-spline coefficients

taking the sum defines a curve in terms of x,

$$f(x_i) = \sum_{m=0}^{M+1} \theta_m h_m(x_i)$$

Of particular interest is the fact that the shapes of the curve $f(x)$ can be described in terms of linear functions of spline coefficients θ, and each coefficient θ_m has only effect on the sub-interval of x. This enables shapes to be flexibly fit over the ranges of the variable. Table 8 summarizes the mapping between the restrictions on linear B-spline coefficients and the shape of a curve $f(x)$. The sign of the slope of each segment is determined by the difference in adjacent parameters, Restrictions 1-3. A function $f(x)$ is linear between λ_{m-1} and λ_{m+1} if Restriction 4 holds. The function is globally linear (i.e., equivalent to $y_i \propto x_i\beta$) for all x_i if Restriction 4 holds for every knot λ_m; as such, linear functions are nested within the more flexible B-spline model. A unimodal curve follows from requiring the curve to be monotonically increasing over a set of knots, and then monotonically decreasing thereafter. Convexity or concavity can be described by differences of differences of parameters and knot locations as described in Restrictions 5 and 6.

Given a particular theory and set of data, the estimation challenge is to find the best fitting model subject to the constraint of respecting the shape constraints. In the empirical applications of this paper, curves $f(x)$ are fitted based on maximizing a likelihood criterion, and an adaptive logarithmic barrier approach is used to enforce the boundary constraints on parameter inequalities to achieve particular shapes (Lange, 2010). In the context of linear regressions, solving the first order conditions of the model subject to inequality restrictions on parameters can be formulated as a quadratic programming problem.

The values of k and λ determine the smoothness and flexibility of the curve. Given M interior

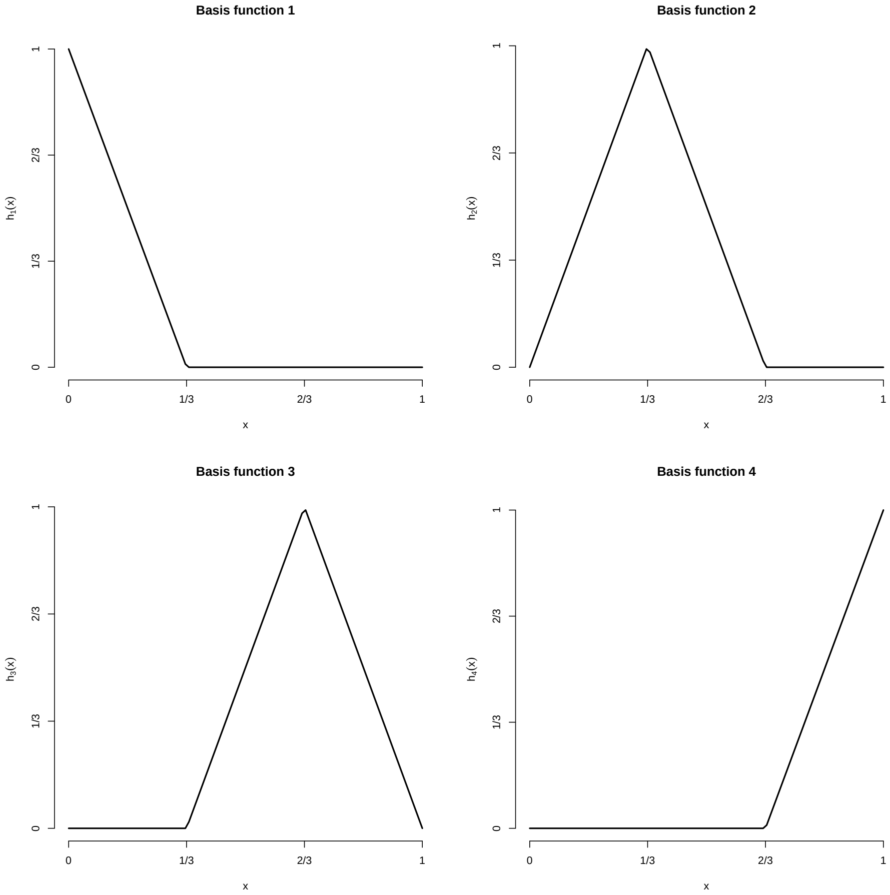


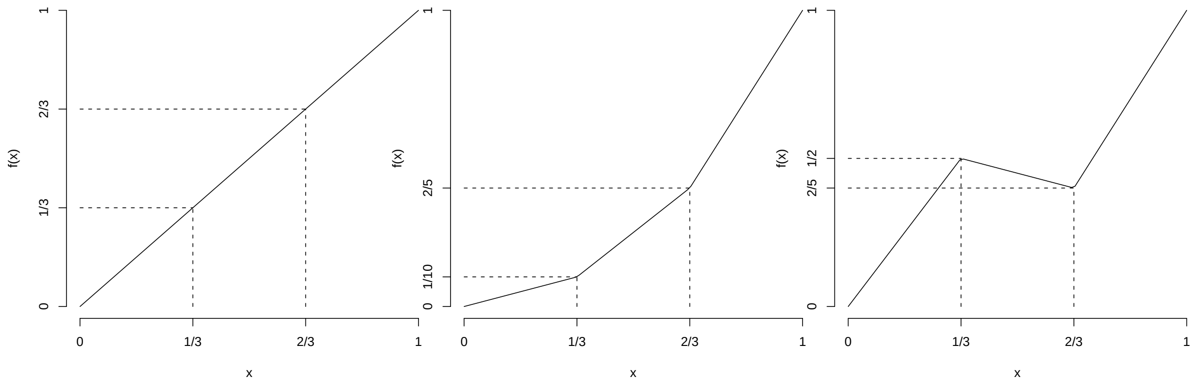
Figure 7: Linear B-spline decomposition of $x \in (0, 1)$ with $\lambda = (1/3, 2/3)$

knots, we have $M + 2$ basis functions (hence there are four distinct separate lines plotted). The function $f(x)$ has at most the number of sign changes in derivative as the number of coefficient sign changes (Schumaker, 1981, Theorem 4.76). Smoothness of the spline at knot locations is determined by the order of the basis function. The first k-1 derivatives are defined at a knot.

The shapes of the curve are determined by the relative values of adjacent coefficients in the vector θ, and many properties of curves can be specified in terms of restrictions on linear combinations of θ. Table 8 summarizes the mapping between the restrictions on linear B-spline coefficients and the shape of a curve $f(x)$.

Figure 7 provides an example of a B-spline decomposition of a continuous variable x on the interval (0,1), using a spline of order 1 (linear) and two interior “knots” located at $\lambda = (1/3, 2/3)$. In most circumstances there are implicit knots at the limits of a variable, either specified by prior knowledge or determined by the empirical bounds of the variable.

Figure 8(a) is a linear function built from the B-spline decomposition. A monotonic (weakly) increasing curve follows from having Restriction 1 or 2 hold for all m, as illustrated in Figure 8(b). Non-monotonicity in Figure 8(c) occurs because $\theta_3 - \theta_2 = -0.1 < 0$. Note the contrast between



(a) $\theta = [0, 1/3, 2/3, 1]$ (b) $\theta = [0, .1, .4, 1]$ (c) $\theta = [0, .5, .4, 1]$

Figure 8: $f(x)$ for values of θ, given $\lambda = (1/3, 2/3)$,

figures (b) and (c) illustrates the local effects of the parameters: changing $\theta_2 = 0.5$ to $\theta_2 = 0.1$ between affects only the first two line segments but not the third.

The figures in 8 highlight the issue of smoothness (or lack thereof) at knot locations of linear B-splines, depending on the rate of change of the slope across adjacent segments. Neither (b) or (c) are smooth functions, since the first derivatives of $f(x)$ are discontinuous at the knots.

Greater smoothness at knot points, and greater flexibility elsewhere in the function are achievable using higher order polynomials in specifying basis functions. In the applications that follow, however, there is no significant gain in using higher order B-splines in terms of fit and hence does not warrant the additional analytical and computational complexity needed to impose shape constraints. The best fitting curves in the applications look more like (b) than (c), and the former is not much outdone by using higher order B-splines. In most practical setting, adding flexibility and achieving better fits through additional knots may require fewer parameters and more easily derived constraints than moving to higher order B-splines. De Boor (1978) provides a succinct presentation of the formulation and properties of higher order B-splines; the core logic is nonetheless the same.

The dependence of splines on the choice of knot locations λ is a limitation for exploratory data analysis, but a virtue for testing specific shapes motivated by shapes. When theory provides guidance as to the location of knots one can use this prior knowledge. It is also possible to treat knots as parameters to be estimated Friedman (1991). In all cases, it is important feature of the data to know if the fitness of the model is sensitive to reasonable variations of knot locations and this should be checked as part of careful data analysis.

7.4 Inference with inequality constraints on parameters

The differences between testing multiple equality and inequality restrictions can most simply be described analytically and graphically in the context hypothesis tests on two independent and normally distributed variables. The logic of the framework generalizes to greater number of restrictions, and other distributional assumptions.

Let $\hat{\Delta} = (\hat{\Delta}_1, \hat{\Delta}_2)$ be a pair of parameters distributed as a bivariate normal ($\hat{\Delta} \sim N(\Delta, I)$), where Δ is unknown and I is an identity matrix. The reader might find it useful to think of these parameters as means estimated from a sample of points, or as regression coefficients. To make the direct connection to shape constrained inference consider that $\Delta_j = \mu_j - \mu_{j-1}$, a differences in spline coefficients. The "hat" on the parameters indicates simply that these are realized draws from the underlying population. It will be of interest to know whether $\hat{\Delta}$ could plausibly be observed under a particular hypothesis about the value of Δ. Consider testing $H_0 : \Delta = (0, 0)$ against $H_A : \Delta \neq (0, 0)$. Different draws of pairs of parameters are illustrated in Figure 9(a). In this simple example, the probability that each $\hat{\Delta}$ is consistent with H_0 is a function of its Euclidean distance from $(0, 0)$. For each point the distance of interest the length of the arrow. Since the sum of k squared of standard normals is distributed χ^2_k, the solution to $P(\hat{\Delta}_1^2 + \hat{\Delta}_2^2 < c) = 0.05$ is $c = 6.08$. The region of parameters wherein we would reject the H_0 are all points outside of the circle centered on (0,0) with radius $\sqrt{c}$. This is illustrated in Figure 9(b); for a point in the white circle, a research would not reject H_0 at the 0.05 level.

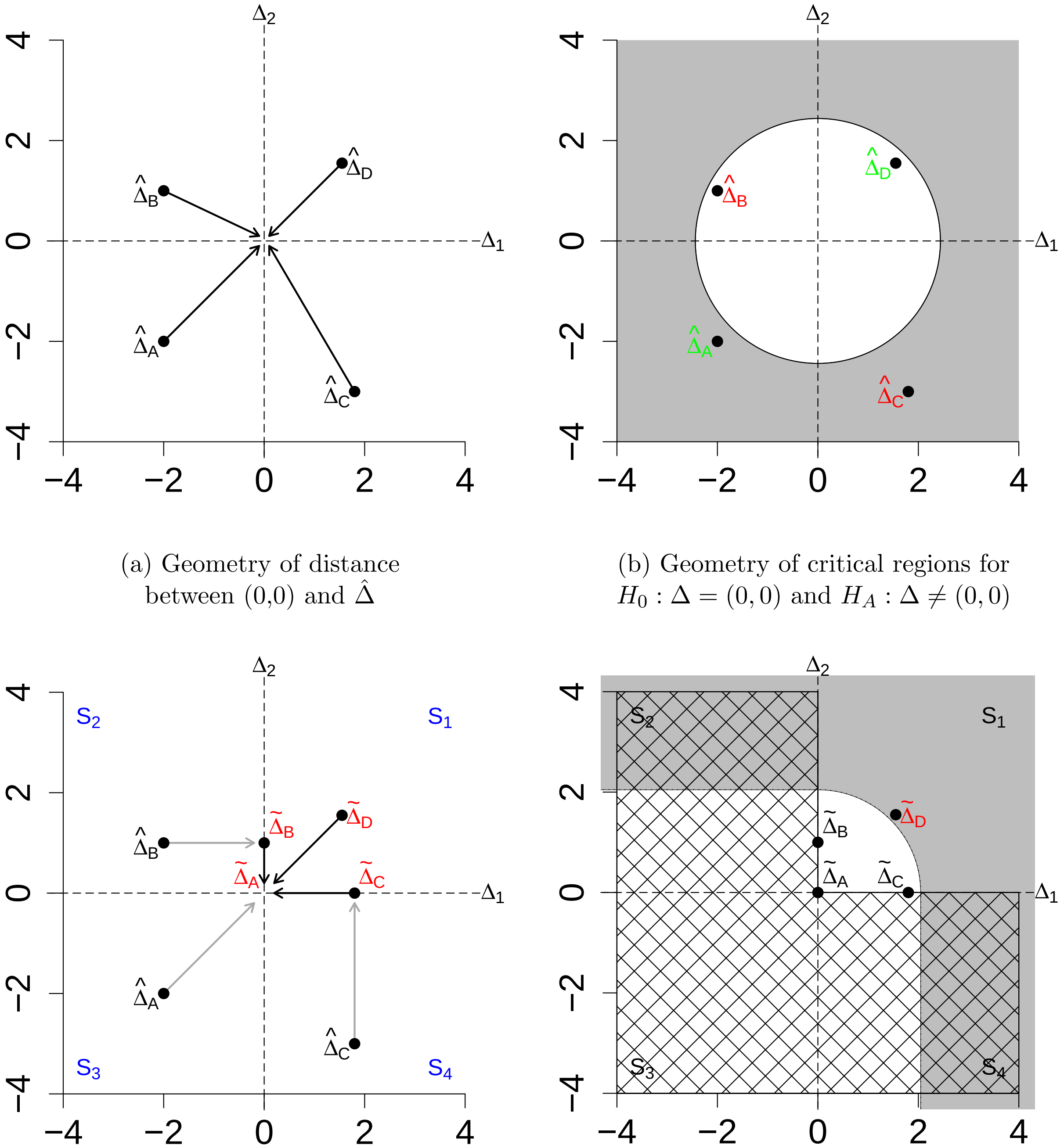
In the context of inequalities of parameters of B-splines, it will be of interest to test hypotheses of the form $H_0 : \Delta = (0, 0)$ against an inequality-constrained alternative $H'_A : \Delta > (0, 0)$. Again, consider $\Delta_1 = \mu_2 - \mu_1$ and $\Delta_2 = \mu_3 - \mu_2$, such that H'_A would imply that the three line segments associated with μ_1, μ_2, and μ_3 are monotonically increasing.

The parameter space under the H'_A is restricted to the first quadrant, and to undertake this test requires the additional step of finding the best fitting means $\tilde{\Delta}$ that satisfy the constraints of H'_A. If $\hat{\Delta} > 0$ then $\tilde{\Delta} = \hat{\Delta}$, but otherwise some constrained optimization is needed under this constraint, the closest points $\tilde{\Delta}$ to $\hat{\Delta}$,

$$\tilde{\Delta} = \begin{cases} (\hat{\Delta}_1, 0) & \text{if } \hat{\Delta}_1 > 0 \text{ and } \hat{\Delta}_2 < 0 \\ (0, \hat{\Delta}_2) & \text{if } \hat{\Delta}_1 < 0 \text{ and } \hat{\Delta}_2 > 0 \\ (0, 0) & \text{if } \hat{\Delta}_1 < 0 \text{ and } \hat{\Delta}_2 < 0 \end{cases}$$

For each of the $\hat{\Delta}$ plotted in Figure 9(a), the inequality constrained values are plotted as $\tilde{\Delta}$ in Figure 9(c).

The distribution of a standard test statistic for a hypothesis of a set of linear equality restriction on a parameters (e.g., $H_0 : R\Delta = 0$ versus $H_A : R\Delta \neq 0$ is a function of the difference of the dimensionality of models being compared. This difference in the number of free parameters is simply $r = \text{rank}(R)$. In the case of the $H_0 : \Delta = (0, 0)$ versus $H_A : \Delta \neq (0, 0)$, the difference in dimensionality is $r = 2$, and the distribution of the distance $\hat{\Delta}_1^2 + \hat{\Delta}_2^2$ is χ^2_2. How does the distribution of the test statistic of inequality constrained parameters differ from unconstrained tests of equality? When comparing models fit subject to inequality constraints the differences the difference in dimensionality (i.e., the number of free parameters) has a stochastic distribution rather



(a) Geometry of distance between (0,0) and $\hat{\Delta}$

(b) Geometry of critical regions for $H_0 : \Delta = (0,0)$ and $H_A : \Delta \neq (0,0)$

(c) Geometry of distance between $\hat{\Delta}$, $\tilde{\Delta}$, and $(0,0)$

(d) Geometry of critical regions for $H_0 : \Delta = (0,0)$ and $H'_A : \Delta > (0,0)$

Figure 9: Geometry of test statistics and critical values

than being a fixed number.

Cases where the unrestricted means lie in quadrant S_1, the estimator has two free parameters. Cases with $\hat{\Delta}$ in quadrant S_2 and S_4 have one free parameter ($\hat{\Delta}_1$ and $\hat{\Delta}_2$, respectively). Cases with unrestricted means that lie in quadrant S_3 have no free parameters. Thus, the distribution of $\tilde{\Delta}_1^2 + \tilde{\Delta}_2^2$ is not χ_2^2 in all four quadrants, but rather varies by quadrant. The conditional probability distribution in each quadrant is,

$$\begin{aligned}
P(\tilde{\Delta}_1^2 + \tilde{\Delta}_2^2 < c' \mid \hat{\Delta} \in S_1) &= P(\chi_2^2 < c') \\
P(\tilde{\Delta}_1^2 + 0 < c' \mid \hat{\Delta} \in S_2) &= P(0 + \tilde{\Delta}_2^2 < c' \mid \hat{\Delta} \in S_4) = P(\chi_1^2 < c') \\
P(0 + 0 < c' \mid \hat{\Delta} \in S_3) &= 1
\end{aligned}$$

So for a given test size α, again the critical value c' is determined by the solution to the equation

$$\begin{aligned}
P(\tilde{\Delta}_1^2 + \tilde{\Delta}_2^2 < c') &= P(\hat{\Delta} \in S_1)P(\chi_2^2 < c') + P(\hat{\Delta} \in \{S_2, S_4\})P(\chi_1^2 < c') + P(\hat{\Delta} \in S_3) \\
&= 1/4P(\chi_2^2 < c') + 1/2P(\chi_1^2 < c') + 1/4 \\
&= 1 - \alpha
\end{aligned}$$

where the second equality holds since under H_0, the probability of being in each of the four quadrants is $P(\hat{\Delta} \in S_i) = 1/4$.

The geometry of the critical region given $\sqrt{c'} = 2.05$ is plotted in the Figure 9(d). The crossed areas indicate the kportion of the parameter space that is excluded by H_A. For any point not in grey, we would again fail to reject H_0 against H_A.

To illustrate the differences in inference and the trade-offs that follow from using either H_A or H_A', Figure 10 overlays the two critical regions from Figures 9(b) and 9(d). The value of c' is less than the c in the unrestricted alternative, as such we gain power against H_A' when $\hat{\Delta} > 0$. However, we lose all power against cases where $\hat{\Delta} < 0$: any case in S_3 provides no information against the null when the alternative is H_A. Parts of S_2 and S_4 are also changed.

In general, the distribution of test statistics for comparing models subject to inequality constraints on parameters is a mixture of F or χ^2 distributions (Wolak, 1987, 1989; Silvapulle and Sen, 2005). The general form of the p-value of equality versus inequality tests is,

$$Pr(\bar{\chi}^2 > c) = \sum_{q=1}^{Q} Pr(\chi_q^2 > c) w_q$$

where Q is maximum number of unconstrained parameters, χ_q^2 is a chi-square random variable with q degrees of freedom, and $w_q = Pr(\text{q free parameters})$. Tests of inequality restrictions versus unrestricted alternatives follow a similar logic but in reverse: the number of restrictions in the null model is stochastic. Testing nested shapes implies that the number of parameters of both the null and alternative models are stochastic.

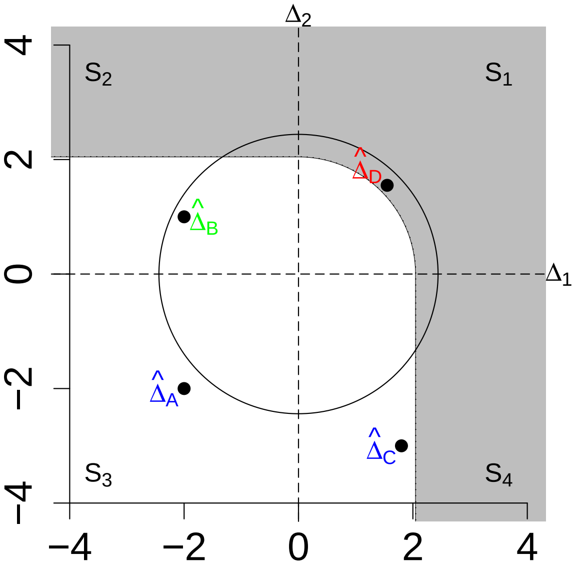


Figure 10: Comparison of unrestricted and restricted critical regions

7.5 Monte Carlo Details

The data generating process is of the form $y_i = I(x = j) \times \mu_j + \epsilon_{ij}$ where $\epsilon \sim N(0, 1)$. The value of the mean values for $j \in \{1, ..., 7\}$ are shown in Table 9

For each model the sample size

	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7
Constant	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Linear	1.00	2.00	3.00	4.00	5.00	6.00	7.00
Step	4.00	4.00	6.00	6.00	6.00	6.00	6.00
Quadratic	1.75	3.00	3.75	4.00	3.75	3.00	1.75
Unimodal 0	4.00	4.00	6.00	7.00	6.00	6.00	5.00
Unimodal 1	4.00	4.00	4.00	4.00	8.00	8.00	6.00
Unimodal 2	4.00	4.00	4.00	4.00	7.00	7.00	6.00
Oscillating	4.00	6.00	5.00	7.00	6.00	8.00	7.00

Table 9: Population mean values in MC

References

Achen, Christopher H. 1982. *Interpreting and Using Regression*. London: Sage.

Achen, Christopher H. 2002. Toward a New Political Methodology: Microfoundations and ART. *Annual Review of Political Science*, 5(1):423–50.

Ashworth, Scott and Ethan Bueno de Mesquita. 2006. Monotone Comparative Statics for Models of Politics. *American Journal of Political Science*, 50(1):214–231.
URL `http://www.jstor.org/stable/3694267`

Barlow, R. E., D. J. Bartholomew, J. M Bremer, and H.D. Brunk. 1972. *Statistical Inference Under Order Restrictions*. NY: Wiley.

Baron, David P. 1989. Service-Induced Campaign Contributions and the Electoral Equilibrium. *The Quarterly Journal of Economics*, 104(1):45–72.
URL `http://www.jstor.org/stable/2937834`

Beliakov, G. 2000. Shape preserving approximation using least squares splines. *Analysis in Theory and Applications*, 16(4):80–98.

De Boor, Carl. 1978. *A Practical Guide to Splines*. New York: Springer.

De Mesquita, B.B., A. Smith, R.M. Siverson, and J.D. Morrow. 2005. *The logic of political survival*. The MIT press.

Dierckx, Paul. 1993. Curve and Surface Fitting with Splines.

Friedman, J.H. 1991. Multivariate adaptive regression splines. *The annals of statistics*, 19(1):1–67.

Hansen, P.R., A. Lunde, and J.M. Nason. 2011. The model confidence set. *Econometrica*, 79(2):453–497.

Hastie, Trevor and Robert Tibshirani. 1986. Generalized Additive Models. *Statistical Science*, 1(3):297–310.

Hastie, Trevor and Robert Tibshirani. 1988. Comment on “Monotone Regression Splines in Action” by J. Ramsay. *Statistical Science*, 3(4):450–456.
URL `http://www.jstor.org/stable/2245398`

Lange, Kenneth. 2010. *Numerical analysis for statisticians*. Springer-Verlag.

Laurent, PJ. 1969. Construction of spline functions in a convex set. In I. J. Schoenberg, editor, *Approximation with Special Emphasis on Spline Functions*, NY: Academic Press. pages 415–46.

Lind, J.T. and H. Mehlum. 2010. With or Without U? The Appropriate Test for a U-Shaped Relationship*. *Oxford Bulletin of Economics and Statistics*, 72(1):109–118.

Milgrom, Paul. and Chris Shannon. 1994. Monotone Comparative Statics. *Econometrica*, 62(1):157–180.

Przeworski, A., M. Alvarez, J. Cheibub, and F. Limongi. 2000. *Democracy and Development: Political Institutions and Well-Being in the World, 1950-1990*. Cambridge University Press.

Ross, M. 2006. Is democracy good for the poor? *American Journal of Political Science*, 50(4):860–874.

Schumaker, Larry L. 1981. *Spline Functions: Basic Theory*. NY: Wiley.

Shibata, R. 1997. Bootstrap estimate of Kullback-Leibler information for model selection. *Statistica Sinica*, 7:375–394.

Silvapulle, Mervyn J. and Pranap P. Sen. 2005. *Constrained Statistical Inference*. New Jersey: Wiley.

Snyder, James M., Jr. 1990. Campaign Contributions as Investments: The U.S. House of Representatives, 1980-1986. *The Journal of Political Economy*, 98(6):1195–1227.
URL `http://www.jstor.org/stable/2937755`

Stimson, J.A., E.G. Carmines, and R.A. Zeller. 1978. Interpreting polynomial regression. *Sociological Methods & Research*, 6(4):515.

Treier, Shawn and Simon Jackman. 2008. Democracy as a Latent Variable. *American Journal of Political Science*, 52(1):201–217.
URL `http://www.jstor.org/stable/25193806`

Wahba, G. 1973. On the minimization of a quadratic functional subject to a continuous family of linear inequality constraints. *SIAM Journal on Control*, 11:64.

Wand, Jonathan. 2010. More than a Science of Averages: Testing Theories Based on the Shapes of Relationships.

Wolak, F.A. 1989. Testing inequality constraints in linear econometric models. *Journal of econometrics*, 41(2):205–235.

Wolak, Frank A. 1987. An Exact Test for Multiple Inequality and Equality Constraints in the Linear Regression Model. *Journal of the American Statistical Association*, 82(399):782–793.
URL `http://www.jstor.org/stable/2288787`

Wright, Ian W. and Edward J. Wegman. 1980. Isotonic, convex and related splines. *The Annals of Statistics*:1023–1035.

Yatchew, Adonis. 2003. *Semiparametric Regression for the Applied Econometrician*. Cambridge: Cambridge.