Exploring Curvilinearity through Fractional Polynomials in Management Research

Ralitza Nikolaeva
Business Research Unit, ISCTE-Lisbon University Institute, Av. das Forças Armadas, 1649-026 Lisboa, Portugal
ralitza.nikolaeva@iscte.pt
Ph: +351 21 7903437

Amit Bhatnagar
Lubar School of Business, University of Wisconsin Milwaukee, Milwaukee, WI 53201
amit@uwm.edu
Ph: 414-229-2520

Sanjoy Ghose
Lubar School of Business, University of Wisconsin Milwaukee, Milwaukee, WI 53201
sanjoy@uwm.edu
Ph: 414-229-4224

Accepted at *Organizational Research Methods*

Exploring Curvilinearity through Fractional Polynomials in Management Research

Abstract

Imprecise theories do not give enough guidelines for empirical analyses. A paradigmatic shift from linear to curvilinear relationships is necessary to advance management theories. Within the framework of the abductive generation of theories, the authors present a data exploratory technique for the identification of functional relationships between variables. Originating in medical-research, the method uses fractional polynomials to test for alternative curvilinear relationships. It is a compromise between non-parametric curve fitting and conventional polynomials. The multivariable fractional polynomial (MFP) technique is a good tool for exploratory research when theoretical knowledge is non-specific and thus, very useful in phenomena discovery. The authors conduct simulations to demonstrate MFP's performance in various scenarios. The technique's major benefit is the uncovering of non-traditional shapes that cannot be modeled by logarithmic or quadratic functions. While MFP is not suitable for small samples, there does not seem to be a downside of overfitting the data as the fitted curves are very close to the true ones. The authors call for a routine application of the procedure in exploratory studies involving medium and large sample sizes.

Keywords: fractional polynomials, curvilinear relationships, non-monotonic curves, abductive method

Regardless of several recent calls (Edwards & Berry, 2010; Pierce & Aguinis, 2013) for re-examining and extending existing management theories to make them more precise and testable, the field still falls short of putting such recommendations into practice. One reason for the slow diffusion of these ideas might be their lack of critical mass. There are various studies that equate numbers with legitimacy (Carroll & Hannan, 1989). Thus, it can be argued that the application of novel methodologies to help in theory fine-tuning has not reached an adequate legitimacy level. The goal of the current study is to add to the critical mass necessary for raising visibility and, consequently, legitimacy of the call for progress in management research through methodological improvements (Aguinis & Edwards, 2014). Relatedly, Haig (2005) places the importance for scientific progress on phenomena discovery, which often starts with data exploration and notes that more Nobel prizes have been awarded for the *discovery* of a phenomenon rather than its theoretical explanation.

A significant barrier for advancement in management research is the reliance of scholars on linear relationships. Exceptions to the linearity default are some popular functions such as logarithmic or quadratic used to describe well documented phenomena like diminishing returns. However, in the absence of concrete subject-knowledge, linearity is the golden standard. Is this a problem? Apart from the obvious case when an effect may be over(under)stated due to non-linearity, a mismodeled variable may be even rejected as insignificant in a linear model (Johnson, 2014). Further, a different problem may arise when an influential mismodeled factor introduces correlated variables spuriously. Non-accounting for curvilinearity can have even more subtle effects like masking an interaction effect for a synergistic relationship when, in fact, it might be offsetting (Ganzach, 1997). Hence, Pierce and Aguinis (2013) suggest that "a paradigmatic shift from linear to curvilinear models is needed to improve management theory and practice" (p. 317). Such a shift also involves the recognition that non-linearity can take different shapes that are not necessarily

monotonic. S-shaped phenomena like the diffusion of innovation or advertising effects are examples that would be mismodeled by logarithmic or quadratic functions.

The current study's objective is to illustrate how methodological advancements can be used as an exploratory tool in empirical analysis to extend theories to include curvilinearity. As such, it fits in the context of Haig's (2005, 2008) abductive theory of scientific method, which he calls ATOM. At the core of the method is the clear distinction between phenomena, data, and theories. As opposed to the predominant hypothetico-deductive approach in the social sciences, the ATOM starts with the detection of phenomena. The detection of phenomena is facilitated by the availability of data sets that are analyzed to extract empirical generalizations. The discovery phase is followed by abductively inferring the causal mechanisms of the phenomenon. The elaboration of constructing explanatory theories of these causal mechanisms goes through building analogies to ideas that are well understood and accepted. Finally, theories are assessed for their explanatory goodness by selecting the best of several competing explanations. Haig (2005, p.374) states: "The methodological importance of data lies in the fact that they serve as evidence for the phenomena under investigation. In detecting phenomena, one extracts a signal (the phenomenon) from a sea of noise (the data). … It is for this reason that, when extracting phenomena from the data, one often engages in data exploration and reduction by using graphical and statistical methods." To this statement, we add that in the age of Big Data, the importance of data exploration increases exponentially. Accordingly, our objective is to add to the exploratory toolbox of management researchers a curvilinearity detection method that has been successfully applied in medical research.

There is an old tradition in management where "theoretical propositions are usually silent about functional form, implying that the proposed relationship is simply some monotonic function" (Edwards & Berry, 2010, p. 675). Venturing beyond the identification of linear relationships in empirical studies can be used as a way to build richer theories. As management employs a lot of psychology research, Rozin's (2009) call for prioritizing research which identifies functional

relationship between variables is also relevant. Researchers have already written about the absurdity of linear assumptions as in the "too much of a good thing" effect (Pierce & Aguinis, 2013). Sporadically, authors have observed that empirical processes tend to be complex and curvilinear (Agustin & Singh, 2005). These observations lead to the conclusion that phenomena manifesting non-linear relationships, while prevalent, are under theorized and tested (Johnson, 2014). Describing a functional form of the relationship offers a more precise way of hypothesis development as it is easier to disprove a hypothesis. Therefore, we propose that empirical tests of non-linearity can be used as exploratory research in help of sharpening theories, which is in line with the abductive method philosophy (Haig, 2005, 2008). Alba (2012) emphasizes that abduction should not be viewed as less serious than the hypothetico-deductive approach and that actually the lack of abduction in research can lead to incrementalism.

The current study contributes to the toolbox of management researchers by familiarizing them with a method that can be easily applied to explore non-linear relationships beyond the usual logarithmic or quadratic test. In essence, the proposed method is a discovery tool, which should be useful to management scientists as they try to emulate the mature sciences that appreciate alternative avenues to discovery (Alba, 2012). Some management scholars, in fact, claim that too much theory can stifle valuable discoveries (Alba, 2012; Hambrick, 2007). While there are various methods for modeling and fitting non-linear relationships, there are not too many guidelines among which to choose. In the absence of concrete theoretical recommendations, exploration should lead us to estimate different relationships and present the best fitting one. Since management theories rarely offer precision, they are hard to disprove. In order to get more traction of theory, management scholars have two options: either follow the example of economists and build precise mathematical models of functional relationships that can be empirically tested or follow the example of medical research, which relies a lot on exploratory studies that can be later used to enrich theories. Medical

research is an example of a field that has registered tremendous progress due to evidence based studies (Armstrong, 2011). Not surprisingly, it is more equipped with tools for the purpose.

The multivariable fractional polynomial (MFP) algorithm proposed by Royston and Altman (1994) is designed to help researchers in exploratory situations when subject-knowledge is insufficient or vague. The method, which has been successfully applied in medical statistics, chooses the best fitting shape among various parameterizations to differentiate between competing functional forms. It is designed to be applied in a multivariable model building setting, which is the case with many management empirical estimations. The algorithm is built on the principle that a more complex model should be retained only when there is enough evidence that it is better than a simpler one. The best applications are where subject matter knowledge is deficient. Since management theories are often quite generic, we would like to invite researchers to use more exploratory data examination in search of richer hypotheses as prescribed by Haig's (2005) abductive approach.

The best applications of the fractional polynomial method are in cases where the functional form of a relationship of interest might be non-traditional (i.e. not linear, logarithmic, or quadratic) or non-monotonic. Of course, in the management field, this is almost a catch-22 situation, because we rarely encounter relationships of other shapes. Is it because they do not exist or because we do not have adequate tools to model them? We hope that by adding the MFP method to the toolbox, researchers would feel better equipped. The defining features of MFP modeling are: a) a systematic search for better fitting fractional polynomial functions starting from a default linear relationship; b) a function selection process which is transferable as opposed to non-parametric curve fitting; c) centering and scaling variables to reduce the effect of extreme values (Royston & Sauerbrei, 2008). We perform a series of simulations on data to illustrate various scenarios of the application of the method including small and noisy samples. The results of the estimations give a clear idea in what cases the use of the methodology is not advisable.

We suggest that exploratory function fitting can be helpful in sharpening existing theories. However, as opposed to empirical curve fitting, we would have better chances of external validation if the curves are represented by interpretable functions that we can test in different samples. This is one of the advantages of the proposed MFP method. As Wilson (1998, p.7), quoted in Alba (2012), says: "There is no fixed way to make and establish a scientific discovery. Throw everything you can at the subject, so long as the procedures can be duplicated by others."

Model-building with Fractional Polynomials

The go-to remedy for modeling non-linearity by management researchers is expressing the covariate by either a quadratic term or a logarithmic one. Theories are rarely specific enough to point to one or the other. In such cases, it might be justifiable to turn to the data and let them speak. If non-linearity is detected (e.g. through graphical examination), the next logical question is how to model it. It is especially relevant when we have reasons to suspect non-monotonic behavior.

Royston and Altman (1994) propose a multivariable model-building algorithm based on fractional polynomials as a flexible remedy to the curvilinearity problem. The method, further developed in Royston and Sauerbrei (2008), offers a compromise between precision and generalizability. If the objective is data fitting and precision of estimation at each point (what is referred to as the "local-influence" property), then non-parametric models are preferred (these include regression splines, smoothing splines, etc.). On the other hand, if the purpose of the model is to represent a more generalizable relationship and to be transferable to other datasets, then models that are parametric in nature (retain the "global-influence" property) are preferred (these include polynomials, exponential and logistic functions, etc.). What Royston and Sauerbrei (2008) suggest is a method of modeling continuous or rank-ordered covariates through fractional polynomial (FP) functions, which is parametric in nature, but has more flexibility. They also discuss extensively its

advantages vs. other methods. For example, smoothing splines and kernel-based estimates are not transferable and are usually difficult to interpret. This is so because the smoothed curve may be too "choppy", which is impossible to transfer to a different setting. On the other hand, fractional polynomials offer a middle ground between non-parametrical curve-fitting and high dimensional polynomials in the sense that FPs are quite flexible with low-dimensional curves (see Fig. 1). Further, the advantage of polynomial models, in general, is the transparency in terms of the contribution of each factor to the model. The reason why Royston and Sauerbrei (2008) prefer to limit the modeling of FPs to second degree is precisely the tractability and the ease of replicating the model estimation in different samples (external validity). An FP of a higher degree comes closer to empirical curve fitting making it harder to interpret and to transport to future studies. Restrictions to a second degree FP may sacrifice some model fit and predictive ability, but they gain in generalizability compared to empirical curve fitting. Thus, FPs are especially useful in cases when researchers are interested in the discovery or the explanation of a phenomenon rather than the predictive capability of a model. It is exactly the compromise between flexibility and transportability that make the proposed FP method a good tool in Haig's (2005) abductive theory framework: "The appropriate task here is not to determine which model best fits a single set of data but to ascertain whether the model holds across different datasets" (Haig, 2013, p.147).

The basic description of a fractional polynomial model is the following (Royston & Sauerbrei, 2008). For a given variable X, the non-linear relationship can be represented by a polynomial function of degree m and powers p_m, such that:

$$\mathrm{FP}m(X) = \beta_1 X^{p}{}_1 + \ldots + \beta_m X^{p}{}_m \qquad (1)$$

Based on simulation studies the authors have documented that even very complex relationships can be adequately represented with a polynomial of second degree (m=2) with a set of powers p from

{−2, −1, −0.5, 0, 0.5, 1, 2, 3} where 0 represents the natural logarithm (i.e. $X^0 = \ln(X)$). The authors call the polynomials of first and second degree FP1 and FP2 models. Some examples of the shapes that can be generated within this set are shown in Figure 1. This set of fractional polynomials often gives better fit than conventional polynomials even when they are of a higher degree. For example, a log transformation of X, results in an FP1 model: $\text{FP1}(X) = \beta_0 + \beta_1 \ln(X)$; an FP2 model is the result of an FP2 transformation of X: $\text{FP2}(X) = \beta_0 + \beta_1 X^{-1} + \beta_2 X^2$. Thus, for an FP1 model we estimate 8 functions (the powers p in the set above) and for FP2 all the combinations (including repeated powers expressed as $\beta_1 X^p + \beta_2 X^p \ln(X)$) result in 36 functions.

[Insert Figure 1 about here.]

While FP1 functions are monotonic, FP2 do not have to be. In fact, one of the most useful features of the suggested method is the ability of FP2 models to estimate a variety of (non-monotonic) curves, even some with sharp angles. It should be noted, though, that the FP method is not suitable for every type of non-linear relationship. Sigmoid curves (logistic functions), time-series, complex temporal-spatial relationships among others are examples where FP estimations would not fare well. Royston and Sauerbrei (2008) recommend that the use of higher degree polynomials (FP3 and greater) is not advisable as the models become quite unstable and difficult to interpret. In general, the advice for any type of non-linearity modeling is to plot the resulting function. This is the best way to evaluate whether it makes sense from the subject-knowledge perspective and whether it contains some data artefacts like strong influential outliers.

The best fitting FP model is selected on the basis of the powers among the set of 8 {−2, −1, −0.5, 0, 0.5, 1, 2, 3} that give the highest likelihood. If we have a single covariate, this means that an FP1 results from the test of 8 models, whereas an FP2 is selected from 36 models. Since the selection is made through maximum likelihood estimation in order to accommodate generalized

linear models and Cox regression, the corresponding statistic used for model comparison is the deviance (-2*log-likelihood). Royston and Altman (1994) show that the difference between the deviance of the estimated model and the null model is chi-square distributed on 2*m* d.f. (e.g. an FP2 models implies 4 d.f.).

The estimation algorithm, which is available in Stata under the prefix command 'mfp' (followed by the standard regression commands) fulfills two functions – it combines variable selection with function selection. The advantage of the method is that the variable selection is accompanied by a search for non-linear functions, which reduces greatly the possibility of dismissing an influential variable from the model due to strong curvilinearity. While variable selection is one of the features of MFP, we currently do not discuss it and we assume minimal theoretical guidance for variable inclusion in the model. Several authors (e.g., Leigh, 1988 and Whittingham et al., 2006) have mentioned that data driven approaches like stepwise regression are less preferred compared to theory driven approaches. Whittingham et al. (2006) use real data to affirm that use of stepwise regression is bad practice in view of its many biases. In similar lines, Antonakis and Dietz (2011) use Monte Carlo simulation to highlight how stepwise regression can contribute to wrong validity estimates. Therefore, we call for future research to investigate the issue in the context of MFP and at this point we recommend that scholars use theoretical guidelines for variable inclusion.

The MFP algorithm proceeds in the following steps[1].

1. The p-values for variable and function selection should be determined. The default values in the Stata program are 1 for variable selection meaning that all variables are forced in the model and 0.05 for function selection meaning that a non-linear function is selected over the default linear one if the likelihood ratio statistic is significant at 5%. Researchers can manipulate these values in Stata through the options menu.

[1] The procedure is available in SAS and R as well. For more details readers can turn to Royston and Sauerbrei (2008).

2. The maximum polynomial degree is to be selected. As Royston and Sauerbrei (2008) have determined that a wide array of complex relationships can be expressed by second degree polynomials, this is the default option in Stata. It is expressed by the maximum degrees of freedom permitted (for an FP1 model – 2 d.f. and for an FP2 model – 4 d.f.; as the models are tested against a null model, the linear function has 1 d.f.).
3. After the pre-selection of criteria, the algorithm fits the full linear model. (The linear model is the default one, but it can be substituted by other functions, e.g. logarithmic, if necessary.)
4. The function selection procedure starts by testing the best fitting FP2 model for a given independent variable (starting with the most significant one based on the linear fit) against the null model at the pre-specified significance level (step 1). If the χ^2 statistic with 4 d.f. is not significant, the variable is dropped. Otherwise, the FP2 model is tested against a linear function. If the χ^2 statistic with 3 d.f. is not significant, the relationship is linear. If the test statistic is significant, at the last step the FP2 is compared to the best fitting FP1. If the χ^2 statistic with 2 d.f. is not significant, the best fitting model is FP1. Only if all of the above tests are significant, the retained model is FP2. The algorithm is built on the idea that a more complex model should be chosen only when there is sufficient evidence that it provides a better fit. Royston and Sauerbrei (2008) call this approach a closed test procedure.
5. The best fitting function is retained and step 4 is repeated for the next variable (the one with the second highest significance level in the full linear model). The iterative fitting procedure is used for all the variables of interest. This is the end of the first cycle.
6. The next cycle starts with the best fitting model from the previous cycle instead of the full linear model in the first cycle. Steps 4 and 5 are repeated. The cycles continue until the new model does not offer improvement over the model in the previous cycle (the deviance is not significant).

7. Occasionally, the algorithm may fail to converge. The default number of cycles in the Stata program is 5 (this can be changed in the options menu). The outcome of the algorithm is that it chooses the best fitting fractional polynomial function while simultaneously controlling for simplicity and relevancy.

Another feature of the algorithm is that it scales and centers the variables to ensure that a) the influences of extreme cases is diminished and b) covariates are all in positive values so that all powers of the polynomial set can be fitted. Scaling involves dividing the value of the variable by an integer power of 10. If necessary, centering is performed on the mean of the scaled values of *X*. While the estimation of the powers in the FP model is not affected by scaling, depending on the range of *X*s, scaling them can give more accurate results.

Though modeling with fractional polynomials has many advantages, it is not a panacea for all kinds of problems. The most serious disadvantages of the method are the lack of enough power to detect non-linearities in small samples and the influence of extreme values in the distribution of a variable (also more of a problem in small samples). In management research, we are often interested in existence and the direction of an effect rather than the precise quantification. Therefore, we advise not doing any variable selection with small samples (having too many variables to choose from is also not a particular problem in management research). Further, power loss may be limited if only a first degree polynomial is estimated. This is only relevant for the cases where subject knowledge is strong to assume a monotonic relationship. The loose recommendation of Sauerbrei, Royston, and Binder (2007) for a minimum sample size when applying FP modeling should be at least 10 events per number of parameters (in the context of regular regression, events imply the sample size; in logistic regression – the minimum of the number of occurrences or non-occurrences; in Cox regressions – the number of uncensored observations). Regarding the influence of extreme values of covariates, Royston and Sauerbrei (2008) suggest preliminary transformations of variables. While the default scaling and centering included in the Stata program may do the job for many cases (these

are not available in the SAS and R versions), sometimes the researcher needs to do a different type of transformation[2].

Simulations

As mentioned, one of the weaknesses of the fractional polynomial method is the not so good performance in small samples. Hence, we demonstrate the performance of the method in several simulation studies including small samples. In this way, researchers can get a better idea to what extent they can rely on fractional polynomials for exploratory studies. We strongly discourage the application of the method in small samples.

First, we start by the simple case of one predictor. Following Royston and Sauerbrei (2008), we look at two non-linear relationships exemplifying the FP1 case (underlying logarithmic function) and the FP2 case (underlying quadratic function). In addition, we add a case where the underlying function is a third degree polynomial (a theoretically plausible S-shape), but the fitting is restricted to a polynomial of second degree. For each case we generate[3] three scenarios of not very strong explained variation of $R^2 = 0.1$, $R^2 = 0.2$, and $R^2 = 0.3$. Within each of these scenarios we apply the 'mfp' estimation procedure in Stata to 1000 samples with 50, 100, 200, 500, and 2000 observations. The results appear in Table 1A (the correctly estimated FP transformations corresponding to the underlying function are in bold). We start with the worst example (row 1) – a very small sample of 50 and weak correlation of the independent and dependent variables. In the FP1 case (logarithmic function), we can see that the fractional polynomial estimation resulted in a FP1 function in only 99 of the 1000 samples and out of these 99 only 21 were the logarithmic function (other transformations within the rest of the 78 FP1 cases would be, for example $X^{.5}$). In 878 of the estimations, the MFP

[2] For a detailed discussion, see chapter 5 in Royston and Sauerbrei (2008).
[3] Data generation details are given in the Appendix.

algorithm did not find convincing evidence of non-linearity. In the FP2 case, there were only 298 samples where non-linearity was not detected. However, among the 220 samples where estimation suggested a FP2 transformation, only in 18 of them it was a quadratic one. As can be seen in Table 1A, these numbers improve with the increase of the sample size and the correlation between the dependent and independent variables. With a sample size of 200 and an R^2 of 20%, the recovery of the FP1 and FP2 models is quite reliable. The recovery of the exact function used in the data generation process seems to be better, however, in the FP1 scenario (the logarithmic function) where it is correctly estimated in 695 of the 874 cases indicating an FP1 transformation. With $R^2 = 0.3$ and n=500, the algorithm estimates a logarithmic transformation in 925 out of 1000 samples and with n=2000, the correct recovery is in 937 cases. In the case of the underlying quadratic function, while the algorithm performs pretty well in terms of identifying a non-linear relationship by fitting a second degree polynomial (with n>=200, almost 100% across all three correlations), the precise quadratic function of powers 1 and 2 is not estimated that often. Nevertheless, the shapes of the other polynomials from the MFP estimation are very similar to the original function. In the case of the weekly S-shaped function in the last three columns, we observe that the detection of more complex curvilinearity improves better with sample size (e.g. n>=200 exhibits strong curvilinearity detection) than with higher correlation. Comparing the S-shaped to the quadratic case, we see that curvilinearity detection in small sample sizes and growing correlation coefficient does not improve as well as in the quadratic case (e.g. when n=50, in the quadratic case the MFP algorithm fails to reject a linear relationship in 702 samples with R^2=.10, in 367 with R^2=.20, and in 161 with R^2=.30; for the S-shape case these numbers are 635 with R^2=.10, 430 with R^2=.20, and 306 with R^2=.30).

[Insert Table 1A about here.]

Table 1B illustrates comparisons of MFP estimations and OLS estimations (the default option for management scholars) with randomly selected samples from scenarios simulated in Table 1A. The table is to be read in the following way. Within each underlying function, a different sample was selected to represent the various levels of *n* and R^2. Then, for each of these samples, we estimated an OLS and a MFP regression. Table 1B lists the values of the estimated coefficients and, in the case when the MFP estimation selected a function different from the true underlying one, the variable (X) transformations suggested by the algorithm. For example, when the underlying function is quadratic and n=50 and R^2=.20, the MFP estimated a second degree polynomial with powers (-1, -1)[4] with corresponding coefficients of 17.1 and -63.4. What can be seen in Table 1B is that the problems of fitting an OLS estimation are more severe when the non-linearity is stronger as in the case of the underlying quadratic function. Most of the OLS estimations in the quadratic case result in non-significant coefficients, which can wrongly lead the researcher to the conclusion of the absence of an effect of the independent variable.

[Insert Table 1B about here.]

In sum, there are two observations to be made based on these simulations. While the use of the fractional polynomial estimation in small samples does not provide stable results, there is no compelling argument for shunning the method. Why? Mainly, due to the design of the algorithm, which prefers simplicity and recommends more complex models only with strong evidence to do so. Thus, if a researcher uses the method with a small sample (i.e. less than 100 observations), s/he runs the risk of not detecting a non-linear relationship if the non-linearity is not strong. Currently, most researchers use the default linear function anyway and there are valid arguments for using linear functions as approximations for weak non-linearities. Consequently, "the cost" of the "erroneous"

[4] When the MFP estimation returns a second degree polynomial with the same powers, the second power is multiplied by the logarithm of the variable as well.

result from the MFP estimation would be sticking to the default option. Further, in small samples, computational time for MFP is not an issue, which augments the argument for its use. On the other hand, using the default linear estimation may prevent the discovery of interesting and potentially important non-linear functional relationships.

Next, we proceed with scenarios of many predictors and different levels of noise and sample sizes. We simulated data sets with 10 control variables and one independent variable. The results show that due to the conservative nature of the MFP algorithm, the estimations "err" on not rejecting the linear hypothesis when there is too much noise in the samples. This is the case with the first row in Table 2. As in the univariate case, we estimate the model over 1000 samples in each scenario. Thus, in the first row, we have the estimates of a model (generated either from an underlying logarithmic or quadratic relationship between the independent and the dependent variables) with a hypothesis variable X, 10 control variables, a normal error term multiplied by 10 to make the samples really noisy, and a sample size of 120. In the case when the true function is logarithmic, the model indicated not enough evidence to reject a linear fit in 914 out of the 1000 samples; in 74 samples the MFP algorithm suggested as fractional polynomial of degree 1 (FP1) as the best fit, but only 29 of these indicated the logarithmic function (power 0 in the Stata set of FP powers). In 12 samples, the algorithm estimated a second degree polynomial (FP2) as the best fit suggesting possible over-fitting. We simulated a weak quadratic relationship as shown in Figure 2. Under this scenario, the algorithm also sticks overwhelmingly to linear relationship with 961 out of 1000 samples estimating a linear function. The second row where the sample size is increased to 200 observations, we can see a slight improvement in the logarithmic case where in 62 cases the algorithm recommends the correct logarithmic relationship between X and Y. In the quadratic case, non-linearity is detected in only 57 (45 FP1 and 12 FP2) of the 1000 samples and none of these estimated the correct powers of 1 and 2 in the FP2 model. Increasing the sample size exhibits a significant improvement in the logarithmic case. However, due to the weak curvature in the

quadratic case, we see considerable improvement only when the noise is reduced substantially. As can be seen in Figure 2, the curvature in the logarithmic case is much stronger compared to the quadratic case. Thus, when we have a case of a markedly curvilinear functional relationship, the MFP algorithm performs quite well even in very noisy samples as long as the sample size is big enough to accommodate the number of covariates.

[Insert Table 2 about here.]

[Insert Figure 2 about here.]

Overall, we observe the following development in the case of small samples (n=120, 200) with many covariates. When the true function is logarithmic, there is not much improvement than estimating the correct function in about 50% of the cases as we decrease the noise and increase the explained variation (going down the rows in Table 2). Only when the relationship is very strong as in the lowest set of rows in Table 2 (R^2>.80), the MFP performs well in small samples. We do not see a sizeable danger of over-fitting even with more noise. When the true function is quadratic (albeit with a weak curvature) and we keep the default option in Stata of rejecting linearity only when the FP fit is better at p=5% significance, then the MFP algorithm performs quite badly when the samples are noisy. Even in the last row of Table 2 where the fit is pretty high (not much noise in the data), while non-linearity is detected in 100% of the samples, the correct function of FP2 with powers 1 and 2 is estimated in only 27% of the samples of size 200 observations.

The problem of over-fitting relates to modeling too closely the particular data set rather than the underlying true relationship. Simulations allow us to investigate the severity of the situation, i.e. to what extent a model that fits particularly well a certain data set would fit another sample where the underlying relationship is the same. To investigate the over-fitting detected when the underlying function is logarithmic, we did some investigation of the case of small samples with high fit (the last

rows scenario in Table 2). Out of 38 samples where an FP2 model provided the best fit according to the MFP algorithm with n=200, 3 resulted in FP2 model with powers -2 and 0 (i.e. FP2(X) = $\beta_1 X^{-2}$ + $\beta_2 \ln(X)$). We applied the same FP2 transformation (powers -2 and 0) to 5 samples outside of the set of the 3 where the MFP algorithm suggested it as the best fitting model (i.e. 5 samples chosen at random from the remaining 997). The idea was to examine the "transportability" of the model if we apply it to a different sample. We estimated the fit over the samples of 200 observations and compared it to a model with a covariate transformation ln(X). In all the five pairs, the differences between the R^2, the Root MSE, and the mean absolute error were minute – beyond the second decimal (e.g. FP2 vs. logarithmic fit: adjusted R^2 0.9752 vs. 0.9754, RMSE .33476 vs. .33393, MAE .5447019 vs. .5421821). We get similar results when we replicate the model in more noisy samples – where we have noise of 8 times the error term. The fit between the logarithmic and the FP2 model are virtually indistinguishable. This is visually illustrated in Figure 3. There we compare the predictions from the fitting of a logarithmic function (the true function in the data generating process) and the predictions resulting from the fitting of the FP2 model – a second degree polynomial with powers -2 and 0. We fit the FP2 models to samples that did not generate that model in the MFP estimation. In the upper panel we have the case of very little noise. In the lower panel, we have the same situation, but with much more noise (8*error). In both cases the predictions based on the logarithmic function and the FP2 function are very similar. These results appear to indicate that the over-fitting is not worrisome when the MFP estimation is restricted to second degree polynomials.

[Insert Figure 3 about here.]

Based on two sets of simulation studies involving small samples – one with a single covariate and another one with 11 covariates – we confirm the conclusions of prior studies (see Royston &

Sauerbrei, 2008) that the use of fractional polynomial estimations in small, noisy samples should be exercised with great caution if at all. A multivariable model with a small sample size might not have sufficient power to detect non-linearities in the context of fractional polynomial estimation. When subject-knowledge indicates that we are clearly dealing with a monotonic relationship, then one recommendation is to restrict the algorithm to the estimation on only FP1 functions. In this way, we can increase the power by not testing for a FP2 fit. As an example, when the estimation for the logarithmic function was repeated for line 5 in Table 2 (i.e. high noise level – 8*error term, sample size 120), the MFP algorithm suggested and FP1 model for 234 samples (almost twice the 127 in Table 2 where the algorithm is not restricted to the estimation of only FP1 functions) of which 104 – logarithmic relationship. On the other hand, if there are reasons to suspect that the relationship of interest might be non-monotonic, we would need a larger sample to increase the power of a MFP estimation.

While the overall conclusion based on the above simulations is that the fractional polynomial estimation cannot offer stability in the estimated models with small and noisy samples, it is helpful to compare it against the commonly used alternatives. In general, such characteristics of a sample are bound to make any estimation model unstable. The common alternative of just using a linear (or if theoretical arguments call for non-linearity – logarithmic or quadratic) functions are already embedded in the 'mfp' algorithm in Stata. Therefore, the downside of using the MFP procedure for exploratory purposes is that it might fail to detect non-linearities and thus give a false sense of security in modeling a linear relationship. In larger samples, though, such a probability is very small; hence, the procedure can be used to uncover relationships researchers were not familiar with, which can be a starting point in building/expanding theory.

Further, we illustrate the flexibility of the MFP method to model irregular shapes. As recommended by Royston and Altman (1994), it is better to restrict the algorithm to fitting second degree polynomials due to tractability issues. Therefore, we visually demonstrate how a second

degree polynomial fits data generated from a third degree polynomial – a theoretically plausible S-shaped function. In addition to the results presented in Table 1, we compare the MFP fits with fitting a quadratic function – the go-to approach to model non-linearities in management. The results are shown in Table 3 and Figures 4A-B. Data were generated from the same underlying S-shaped function with variables drawn from Normal distributions (the function appears in the top graphs of the panels as well as in the table). The models in Figure 4A were estimated in different sample sizes and Figure 4B represents data from the same function, but with more noise – 5 times the error term in Figure 4A. It is obvious in all cases that the FP2 models represent a better fit compared to a quadratic model.

[Insert Figure 4A about here.]

[Insert Figure 4B about here.]

In addition to the visual comparisons, we assess the estimation results of different samples of the same scenarios as in Fig. 4 in Table 3. It can be observed that the more expressed the curvilinearity is, the worse the performance of the OLS estimation. In the low error scenario (Table 3A), when the sample size is small, while the MFP still offers the best fit, the difference between the quadratic and the OLS fit is not big. The differences are demonstrated graphically in the last column of the table. With the increase of the sample size, both the OLS and the quadratic estimations offer markedly worse fits compared to the MFP estimation. For example, with n=500, the R^2 drops by 10% (MFP - .75 vs. quadratic - .66 vs. OLS - .64) and the root mean square error increases (MFP - 1.938 vs. quadratic - 2.284 vs. OLS - 2.320).

[Insert Table 3A about here.]

We go through the same exercise in Table 3B, but we increase the error to five times the error in Table 3A. Compared to the previous case, we can see that the differences in terms of the fit between the MFP, quadratic, and OLS estimations are not so dramatic even with larger sample size due to the increased level of noise. This example, though, illustrates the usefulness of graphical representations of the estimated fit. Looking at the last section of Table 3B (n=2000), while it might be argued that the gain in 3% in R^2 (MFP - .31 vs. quadratic - .28) is not outstanding, we can see that the quadratic estimation grossly overestimates the effect of X in the lower values of the variable compared to the MFP estimation.

[Insert Table 3B about here.]

Similarly, we increase even more the error in Table 3C – to 10 times the error in Table 1A. Due to the high level of noise, the estimates become quite unstable. Therefore, in this case we also include the values of the mean absolute error of the estimates to be able to more adequately compare the fits between the three estimation methods. In the first two cases with low sample sizes (n=100 and n=200), the MFP estimation does not have enough evidence to reject a linear fit (we present in the table the best-fitting fractional polynomial for comparison purposes). With that much noise and a small sample, none of the methods provides a good fit. Even when the sample size is increased to n=2000, the gain from the MFP estimation is not that sizeable in terms of fit statistics. Not surprisingly, there is no magic treatment for noisy samples. What these results demonstrate, though, is that when the MFP is restricted to two degrees, there is no danger of funky shapes over-fitting noisy samples. Thus, while we may need to look in a messier stack of hay for the needle, we should not worry that we would be finding strings of needles.

[Insert Table 3C about here.]

Finally, we turn to the problem described by Ganzach (1997) – how ignoring non-linearity can cause an interaction effect to be interpreted as synergistic when the true effect is offsetting. We estimated Ganzach's simulation example[5] both with OLS and MFP. As reported in the original simulation, all the OLS estimations resulted in a positive coefficient for the interaction term XZ, whereas the true relationship is offsetting. In contrast, all of the 500 MFP estimations resulted in the identification of a second degree polynomial transformation of the variables X and Z (481 of those correctly reported the quadratic relationship with powers 1 and 2) and a negative coefficient for the interaction XZ indicating an offsetting interaction effect. We proceed further to investigate what happens with different sizes and noisier samples. The results are presented in Table 4. The lines in the MFP estimation represent, in addition to the sign of the interaction term (offsetting vs. synergistic), the number of samples estimating a particular functional relationship of X and Z. For example, when n=500 combined with 5*error, the MFP algorithm estimates an offsetting relationship in 119 samples (0 when we use OLS). The algorithm included X and Z as a linear terms in 310 and 323 samples respectively, as FP1 transformations in 84 and 82 samples respectively, as FP2 transformations in 106 and 95 samples respectively. Out of the FP2 transformations, 31 resulted in the correct polynomial with powers 1 and 2 for both X and Z. As in the previous cases, it can be seen that the MFP algorithm performs better in larger samples with less noise.

[Insert Table 4 about here.]

Applying the MFP Methodology in Empirical Studies

[5] "As an example, consider the following simulation in which n = 1,000; X, Z, and E (an error term) are normally distributed with a mean of 0 and standard deviation of 1, the correlation between X and Z is .7, and the true value of Y is given by $Y = X + Z + X^2 + Z^2 - XZ + E$ (that is, the true interaction is negative). In each of 500 such simulations, the regression $Y = 130 + 13X + 132Z + 133XZ$ yielded a significantly positive coefficient for XZ." (Ganzach, 1997, p. 237)

Multivariable model-building with fractional polynomials is a data exploration technique at its core. We propose that exploratory techniques can be utilized for refining existing theories or coming up with new ones as suggested by Haig (2005, 2008). Haig (2005) develops his abductive methodology in the context of psychology, but it is equally relevant for management research. While management scholars may be slightly more open to empirical generalizations as a source of theoretical insights (the marketing science sub-discipline of management is probably the most open to taking seriously empirical generalizations – e.g. Alba (2012)), there is no doubt that the hypothetico-deductive method is the predominant approach. Haig (2005, 2013), on the other hand, proposes an alternative bottom-up course of proper data-to-theory scientific investigations. It consists of two major stages – the detection of phenomena and the ensuing abductive generation of theory to explain those phenomena (Haig, 2013). A prerequisite for the working application of the bottom-up approach is the proper distinction between phenomena and data. As opposed to data, which are contextual and depend on many factors, phenomena are stable and stubborn. Further, unlike data, phenomena are not easily observable and they need to be abstracted and extracted from data analyses (Haig, 2013). It is exactly at this stage that we need to engage in data exploration. Data exploration allows us to look for phenomena in a sea of noise (Haig, 2013). We propose that the MFP procedure is added to the arsenal of statistical techniques in data analysis that can help us discover phenomena leading to new explanatory theories or augmentation of existing theories. It would fall in Haig's (2005, 2013) data exploratory stage, which is to be followed by close replication and constructive replication. In Haig's (2013, p.141) words, the exploratory methods "are concerned with the effective organization of data, the construction of graphical displays, and the examination of distributional assumptions and functional dependencies". Consequently, we suggest the following procedure for the use of MFP in detection of functional dependencies:

- Start by building models based on existing theories and estimate them testing for non-linearities, i.e. instead of running an ordinary regression, run an MFP regression. For

example, many innovation adoption imitation theories just claim the existence of an imitation effect, but do not specify any functional relationship between prior innovation adopters and subsequent adoption (Lieberman & Asaba, 2006). Similarly, the literature on first mover advantages (FMA) has been concerned with their existence (as such, the effect has been mostly modeled in linear terms), but not so much with the functional relationship of FMA on firm performance (Lieberman & Montgomery, 2013; Suarez & Lanzolla, 2007). Thus, exploratory studies may start by estimating the existing imitation and FMA models testing for non-linear effects of the independent variables. Alternatively, a purely exploratory study does not need to rely on any existing theories to uncover interesting relationships. Such a study, though, would require more close and constructive replications to validate the findings.

- If non-linearities are documented, conduct bootstrap simulations to assess the stability of the detected non-linearity (Royston & Sauerbrei, 2009). This step corresponds to Haig's (2005, 2013) close replication.
- Close replications are to be followed by more tests with different data sets to see if similar non-linearities appear in other contexts. Haig (2005, 2013) calls such constructive replications validity checks.
- If the non-linear pattern is persistent, see if it can be supported by the logic of existing theories. As it happens, the management field is not short on theories (Hambrick, 2007), which usually means that a phenomenon can be described by various theoretical perspectives. Consequently, the detected non-linearity can serve to propose fine-tuning of existing theories.
- Test sharpened theory with other samples, if possible, following the hypothetico-deductive approach.
- If the empirical findings are not supported by existing theories, do more tests and simulations to ensure that the non-linearities uncovered in empirical tests are reliable. If the results are consistent, revise/build new theory in light of new findings.

Discussion

The major goal of the current study is to add to management research's toolbox an exploratory method that can be used in an abductive manner to discover new phenomena or sharpen existing theories. Namely, it contributes to the literature on the importance of accounting for non-linearities by reviewing for management scholars a method for testing for curvilinear relationships that is easy to apply and has transportable qualities. It is our hope that scholars will start utilizing it routinely in exploratory studies to gain insights about the shape of relationships that can be used to augment existing theories or come up with new ones.

For the purpose, we started with a brief description the fractional polynomial method for modelling continuous variables. This method has been successfully used in the medical sciences (Royston, 2000) and very limitedly in economics (Henley & Peirson, 1997; Jones & Weinberg, 2011). Its main features, as opposed to other curve-fitting techniques like spline modelling, are generalizability, transportability, and practicality (Sauerbrei et al., 2007). As part of the presentation of the method to the management scientists, we conduct simulations to demonstrate its performance in multiple scenarios. Not surprisingly, the method works better for larger and less noisy samples. Nevertheless, when the true underlying function is notably curvilinear, the method is quite powerful even with noisy and not so big samples.

What is the overall assessment of the MFP model building method? We think it is important that management scholars are aware of it for the following reasons. Management is a discipline where theories rarely provide specific guidelines for the functional form of a relationship (Edwards & Berry, 2010). While the linear assumption of an effect may work in many cases, there are instances when it may be misleading or where we may need more precision in the estimation. The MFP procedure can be a very useful tool when subject-knowledge is vague and data exploration is

desirable. Moreover, while not popular among management researchers, evidence-based research can be instrumental for the advancement of a discipline (Armstrong, 2011). This idea is the impulse behind Haig's (2005, 2013) abductive theory of scientific method. It emphasizes the discovery of phenomena through data exploration later leading to abductive generation of theories.

The best applications of the MFP procedure to model continuous variables in regression models (linear, logistic, Cox) are when: a) theoretical knowledge is not concrete and there are reasons to suspect non-linearity; b) sample size is not too small; c) full information of the variables is to be used, meaning that the variable is not to be dichotomized or grouped in any other number of levels; d) improved model fit is desirable, meaning that it is not the result of inflated model complexity and over-fitting; e) the goal is the model to be transportable to other data sets. The application of the procedure is very easy as it is available as a prefix command in Stata ('mfp'). Thus, even researchers who are not well versed in the methodology can use it as a diagnostic tool to evaluate the functional form of relationships between variables.

Haig's (2005) abductive theory of method (ATOM) is not only applicable to, but also desirable for management scholarship. The problems of psychological science as described by Haig (2013) exist in the management disciplines as well. The stickiness of the hypothetico-deductive approach can lead to incrementalism (Alba, 2012). Therefore, data exploration should be accepted as a valid scientific approach to phenomena detection. We suggest that the explanation of complex phenomena should be preceded by empirical investigation of the shape of relationships. Very often management scholars become over reliant on a particular theory/paradigm at the expense of other plausible explanations of observed market behavior. This can be especially problematic when (as is often the case with empirical studies) we have imperfect constructs and measure effects indirectly. While there is acceptance that subject-knowledge should direct model-building, management as a discipline usually provides rather general subject guidelines, which makes it hard to disprove

theories. Therefore, turning to data for help in model building can be a way to generate better explanatory theories.

Answering the calls of Edwards and Berry (2010) and Pierce and Aguinis (2013) for testing for non-linearity, we propose data exploration through fractional polynomial modeling as a way of gaining insights into a phenomenon and extending existing theories to conjecture distinguishable functional forms. In doing so, our hope is to alert researchers of the applicability of the method and the consequences of mismodeling non-linearity of explanatory variables. The use of this tool allows us to expose the dangers of reading too much into results that happen to support a particular (vague) theory. We contend that testing for different functional shapes should become a standard practice in empirical studies, because the significance of a linear effect or the lack of it, may tell only part of the story. Keeping in mind the existence of multiple stories provides for a healthy scholarly attitude in our humble opinion.

References

Aguinis, H., & Edwards, J. R. (2014). Methodological wishes for the next decade and how to make wishes come true. *Journal of Management Studies*, *51*(1), 143–174.

Agustin, C., & Singh, J. (2005). Curvilinear effects of consumer loyalty determinants in relational exchanges. *Journal of Marketing Research*, *42*(1), 96–108.

Alba, J. W. (2012). In defense of bumbling. *Journal of Consumer Research*, *38*(April), 981–987.

Antonakis, J., & Dietz, J. (2011). More on testing for validity instead of looking for it. *Personality and Individual Differences*, *50*(3), 418-421.

Armstrong, J. S. (2011). Evidence-based advertising: an application to persuasion. *International Journal of Advertising*, *30*(5), 743-767.

Carroll, G.R., & Hannan, M.T. (1989). Density dependence in the evolution of populations of newspaper organizations. *American Sociological Review*, *54*, 524-541.

Edwards, J. R., & Berry, J. W. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods*, *13*(4), 668–689.

Ganzach, Y. (1997). Misleading interaction and curvilinear terms. *Psychological Methods*, *2*(3), 235–247.

Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods*, *10*(4), 371–388.

Haig, B. D. (2008). Precis of "An abductive theory of scientific method." *Journal of Clinical Psychology*, *64*(9), 1019–1022.

Haig, B. D. (2013). Detecting psychological phenomena: Taking bottom-up research seriously. *American Journal of Psychology*, *126*(2), 135–153.

Hambrick, D. C. (2007). The field of management's devotion to theory: Too much of a good thing? *Academy of Management Journal*, *50*(6), 1346–1352.

Henley, A., & Peirson, J. (1997). Non-linearities in electricity demand and temperature: Parametric versus non-parametric methods. *Oxford Bulletin of Economics and Statistics*, *59*(1), 149–162.

Johnson, J. S. (2014). Nonlinear analyses in sales research: Theoretical bases and analytical considerations for polynomial models. *Journal of Personal Selling & Sales Management*, *34*(4), 302-317.

Jones, B. F., & Weinberg, B. A. (2011). Age dynamics in scientific creativity. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(47), 18910–4.

Leigh, J. P. (1988). Assessing the importance of an independent variable in multiple regression: is stepwise unwise?. *Journal of Clinical Epidemiology*,*41*(7), 669-677.

Lieberman, M. B., & Asaba, S. (2006). Why do firms imitate each other? *Academy of Management Review*, *31*(2), 366–385.

Lieberman, M. B., & Montgomery, D. B. (2013). Conundra and progress: Research on entry order and performance. *Long Range Planning*, *46*(4-5), 312–324.

Pierce, J. R., & Aguinis, H. (2013). The too-much-of-a-good-thing effect in management. *Journal of Management*, *39*(2), 313–338.

Royston, P. (2000). A strategy for modelling the effect of a continuous covariate in medicine and epidemiology. *Statistics in Medicine*, *19*(14), 1831-1847.

Royston, P., & Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics*, *43*(3), 429-467.

Royston, P., & Sauerbrei, W. (2008). *Multivariable Model-Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modeling Continuous Variables.* New York: John Wiley and Sons.

Royston, P., & Sauerbrei, W. (2009). Bootstrap assessment of the stability of multivariable models. *The Stata Journal*, *9*(4), 547–570.

Rozin, P. (2009). What kind of empirical research should we publish, fund, and reward? *Perspectives on Psychological Science*, *4*, 435–439.

Sauerbrei, W., Royston, P., & Binder, H. (2007). Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Statistics in Medicine*, *26*(30), 5512-5528.

Suarez, F. F., & Lanzolla, G. (2007). The role of environmental dynamics in building a first mover advantage theory. *Academy of Management Review*, *32*(2), 377–392.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour?.*Journal of Animal Ecology*, *75*(5), 1182-1189.

Wilson, Edward O. (1998). Scientists, scholars, knaves and fools. *American Scientist*, *86* (January–February), 6–7.

Table 1A

Simulation Results – single covariate, 1000 samples. For each function, the table includes the number of MFP estimations where the best fitting function is linear (Lin), first degree fractional polynomial (FP1), and second degree fractional polynomial (FP2).

R^2	*n*	*Underlying Function Logarithmic*			*Underlying Function Quadratic*			*Underlying Function S-shaped*		
		Lin	**FP1 (Ln)**	**FP2**	**Lin**	**FP1**	**FP2(X,X^2)**	**Lin**	**FP1**	**FP2**
0.1	50	878	**99(21)**	23	702	78	**220(18)**	635	278	87
	100	730	**249(108)**	21	322	107	**571(55)**	371	375	254
	200	388	**575(372)**	37	38	38	**924(139)**	73	225	702
	500	28	**931(770)**	41	0	0	**1000(221)**	0	13	987
	2000	0	**934(931)**	66	0	0	**1000(417)**	0	0	1000
0.2	50	722	**251(107)**	27	367	106	**527(45)**	430	369	201
	100	404	**573(373)**	23	66	51	**883(111)**	135	316	549
	200	78	**874(695)**	48	0	0	**1000(195)**	3	77	920
	500	0	**951(894)**	49	0	0	**1000(288)**	0	1	999
	2000	0	**934(934)**	66	0	0	**1000(573)**	0	0	1000
0.3	50	538	**429(247)**	33	161	74	**765(76)**	306	383	311
	100	159	**811(606)**	30	7	8	**985(151)**	47	226	727
	200	11	**939(839)**	50	0	0	**1000(250)**	0	28	972
	500	0	**933(925)**	67	0	0	**1000(360)**	0	0	1000
	2000	0	**937(937)**	63	0	0	**1000(683)**	0	0	1000

Table 1B

MFP estimation results illustrating each condition for the underlying logarithmic and quadratic functions in Table 1A compared to OLS estimations

R^2	*n*	*Underlying Function Logarithmic*				*Underlying Function Quadratic*			
		OLS		**MFP***		**OLS**		**MFP***	
		β	R^2	β	R^2	β	R^2	β	R^2
0.1	50	ns	.00	ns	.00	ns	.00	$-2.37+.30$	.19
	100	.13	.07	.13X	.07	.21	.03	$8.59X^{-1}+.01X^3$	.17
	200	.16	.06	.34	.10	ns	.00	$15.81X^{-2}+2.66X^{.5}$	.10
	500	.16	.07	.32	.09	ns	.00	$-1.67+.21$	.06
	2000	.16	.07	.32	.09	-.05	.00	$6.42X^{-.5}+0.1X^3$	.12
0.2	50	.22	.11	$-.32X^{-.5}$	.29	.36	.06	$17.1X^{-1}-63.4X^{-1}\ln(X)$	.29
	100	.12	.03	$-.03X^{-1}$	.16	ns	.00	$-2.31+.29$	.11
	200	.17	.12	.47	.21	ns	.00	$-2.51+.32$	.17
	500	.15	.09	.43	.15	-.15	.02	$-4.68X^{.5}+.02X^3$	.26
	2000	.22	.13	.49	.19	-.07	.00	$-3.34\ln(X)+.02X^3$	.23
0.3	50	.20	.14	$-.5X^{-.5}$	.46	ns	.00	$20.16X^{-.5}+.02X^3$	.37
	100	.37	.32	.90	.51	ns	.00	$-.57X^2+.09X^3$	.28
	200	.28	.25	.67	.32	ns	.00	$-3.44+.44$	.15
	500	.30	.23	.70	.33	ns	.00	$-1.95X+.038X^3$	.36
	2000	.28	.20	.65	.28	-.08	.00	$-4.36\ln(X)+.02X^3$	.34

** Where the MFP estimated transformation of X is not shown in the table, it correctly corresponds to the underlying true function – ln(X) in the scenario of a logarithmic underlying function and a second degree polynomial of X with powers 1 and 2 in the scenario of a quadratic underlying function..*

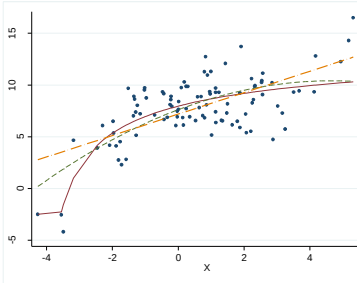
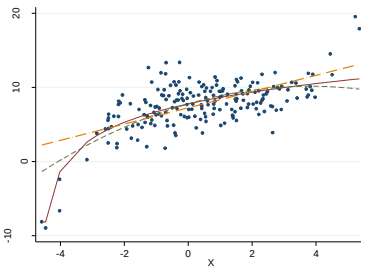
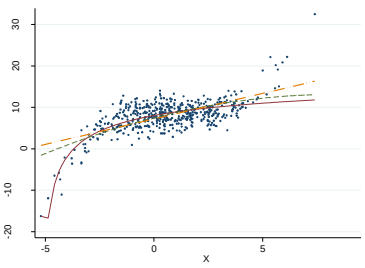
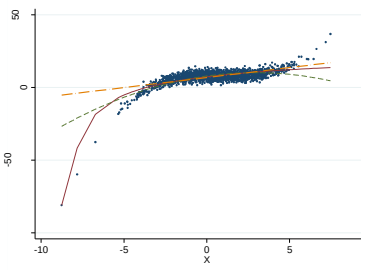
Table 2

Simulation Results – 11 covariates, 1000 samples. For each function, the table includes the number of MFP estimations where the best fitting function is linear (Lin), first degree fractional polynomial (FP1), and second degree fractional polynomial (FP2).

error	*n*	*Underlying Function Logarithmic*			*Underlying Function Quadratic*		
		Lin	**FP1 (Ln)**	**FP2**	**Lin**	**FP1**	**FP2(X,X^2)**
10*err	120	914	**74(29)**	12	961	28	**11(1)**
	200	828	**148(62)**	24	943	45	**12(0)**
	500	610	**349(285)**	41	852	120	**28(0)**
	2000	20	**920(885)**	60	440	510	**50(0)**
8*err	120	853	**127(67)**	20	953	30	**17(0)**
	200	749	**213(140)**	38	907	73	**20(1)**
	500	419	**545(500)**	36	776	200	**24(0)**
	2000	0	**960(936)**	40	250	620	**130(0)**
5*err	120	644	**330(199)**	26	896	91	**13(0)**
	200	440	**521(435)**	39	794	182	**24(0)**
	500	44	**896(853)**	60	457	483	**60(0)**
	2000	0	**931(931)**	69	4	375	**621(76)**
err	120	0	**940(937)**	60	0	227	**773(155)**
	200	0	**962(962)**	38	0	58	**942(267)**
	500	0	**956(956)**	44	0	0	**1000(478)**
	2000	0	**962(962)**	38	0	0	**1000(777)**

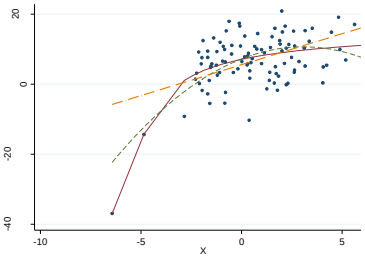
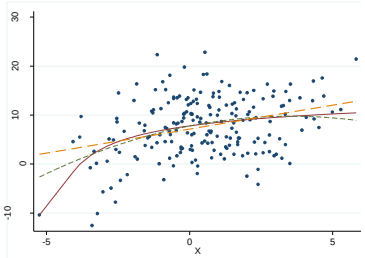
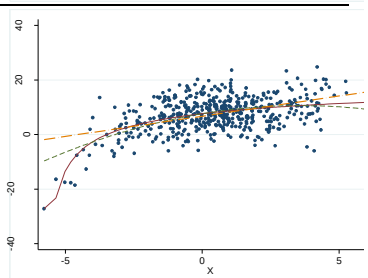
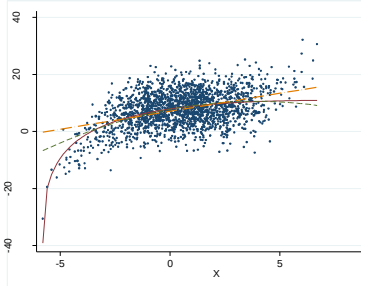
Table 3A

MFP estimation results compared to OLS and quadratic estimations for the cases in Fig.4, different samples – low error, high fit

n	*Underlying Function S-shaped – low error* $Y= 8+0.3X-0.3X^2+0.1X^3 + 0.8Z_1 +0.3Z_2 -0.2Z_3 -0.1Z_4+0.4Z_5+ 0.1Z_6 + 0.3Z_7-0.1Z_8-0.2Z_9+e$							
	OLS		**Quadratic**		**MFP**			
	RMSE	R^2	*RMSE*	R^2	*Powers of MFP fit*	*RMSE*	R^2	
100	1.673	.71	1.606	.74	-.5 -.5	1.367	.81	
200	1.980	.65	1.749	.73	-2 0	1.316	.84	
500	2.320	.64	2.284	.66	-.5 -.5	1.938	.75	
2000	3.183	.52	2.706	.65	-2 -.5	1.981	.81	

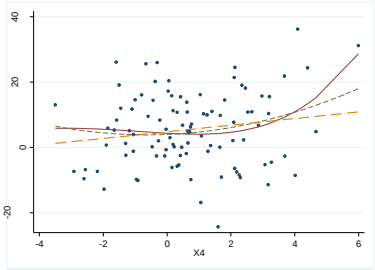
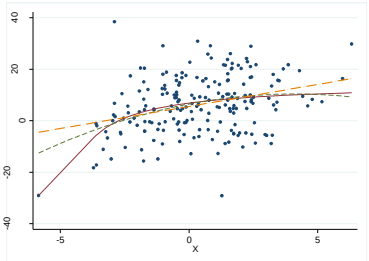
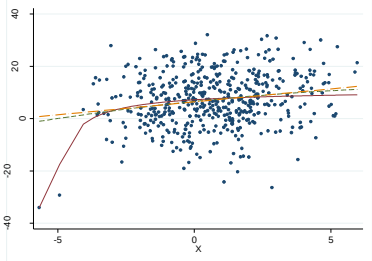
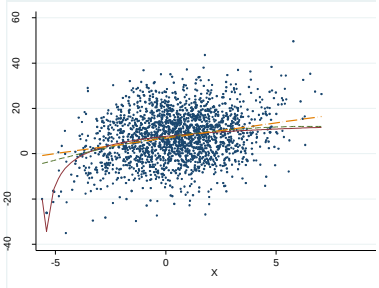
RMSE – root mean square error

Table 3B
MFP estimation results compared to OLS and quadratic estimations for the cases in Fig.4, different samples – higher error

n	*Underlying Function S-shaped – higher error* $Y= 8+0.3X-0.3X^2+0.1X^3 + 0.8Z_1 +0.3Z_2 -0.2Z_3 -0.1Z_4+0.4Z_5+ 0.1Z_6 + 0.3Z_7-0.1Z_8-0.2Z_9+e*5$							
	OLS		**Quadratic**		**MFP**			
	RMSE	R^2	*RMSE*	R^2	*Powers of MFP fit*	*RMSE*	R^2	
100	6.432	.35	5.971	.44	-1 -1	5.710	.48	
200	5.441	.21	5.355	.23	.5	5.341	.24	
500	5.457	.30	5.285	.35	-1 -.5	5.048	.41	
2000	5.385	.26	5.298	.28	.5 .5	5.175	.31	

RMSE – root mean square error

Table 3C
MFP estimation results compared to OLS and quadratic estimations for the cases in Fig.4, different samples – very high error

n	*Underlying Function S-shaped – highest error* $Y= 8+0.3X-0.3X^2+0.1X^3 + 0.8Z_1 +0.3Z_2 -0.2Z_3 -0.1Z_4+0.4Z_5+ 0.1Z_6 + 0.3Z_7-0.1Z_8-0.2Z_9+e*10$							
	OLS		**Quadratic**			**MFP**		
	RMSE	R^2	*RMSE*	R^2	*Powers of MFP fit*	*RMSE*	R^2	
100	10.517 *MAE* 8.281	.13*	10.39 *MAE* 8.041	.15*	3 3	10.294 *MAE* 7.998	.16**	
200	10.705 *MAE* 8.218	.14	10.669 *MAE* 8.150	.15	-2 -1	10.535 *MAE* 8.037	.17**	
500	10.05 *MAE* 7.893	.10	10.055 *MAE* 7.893	.10*	0	9.912 *MAE* 7.844	.12	
2000	10.391 *MAE*l 8.242	.09	10.372 *MAE* 8.227	.09	-.5 -.5	10.297 *MAE* 8.165	.11	

*One or both of X and X^2 n.s..
** Not enough evidence to reject a linear model.
RMSE – root mean square error; MAE – mean absolute error

Table 4

Ganzach's (1997) example: interactions – true relationship offsetting, 500 replications.

error	*n*	**OLS**		**MFP**		
		Offsetting (Synergistic)	Offsetting (Synergistic)	Lin	FP1	FP2(X,X^2; Z, Z^2)
err	100	0 (500)	**422**(78)	50, 48	70, 66	**380, 386 (213, 217)**
	200	0 (500)	**500** (0)	0, 0	1, 1	**499, 499 (353, 365)**
	500	0 (500)	**500** (0)	0, 0	0, 0	**500, 500 (441, 458)**
	1000	0 (500)	**500** (0)	0, 0	0, 0	**500, 500 (481, 481)**
5*err	100	**23** (477)	**49** (451)	469, 480	14, 10	**17, 10 (7, 2)**
	200	**4** (496)	**44** (456)	442, 456	27, 18	**31, 26 (4, 6)**
	500	0 (500)	**119** (381)	310, 323	84, 82	**106, 95 (31, 31)**
	1000	0 (500)	**270** (230)	137, 141	110, 98	**253, 261 (103, 106)**
7*err	100	**40** (460)	**59** (441)	481, 476	5, 10	**14, 14 (0, 2)**
	200	**7** (493)	**36** (464)	467, 474	16, 10	**17, 16 (1, 6)**
	500	**1** (499)	**67** (433)	430, 412	23, 38	**47, 50 (14, 14)**
	1000	0 (500)	**134** (366)	300, 321	79, 50	**121, 129(41, 37)**

Figure 1
Representations of different shapes of FP1 (row 1) and FP2 (row 2) with the numbers indicating the power of the polynomials

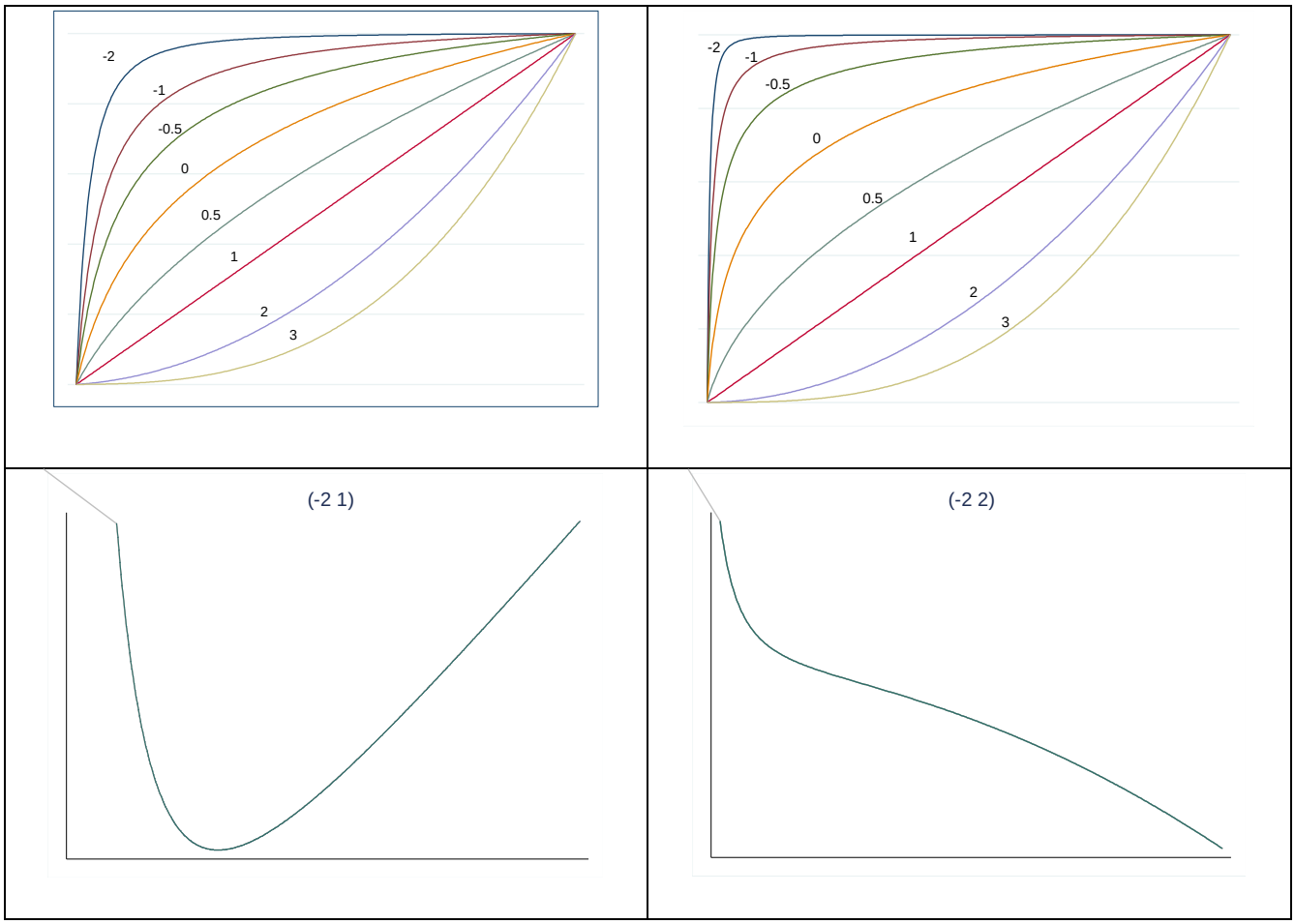


Figure 2
True underlying logarithmic and quadratic functions in Table 2.

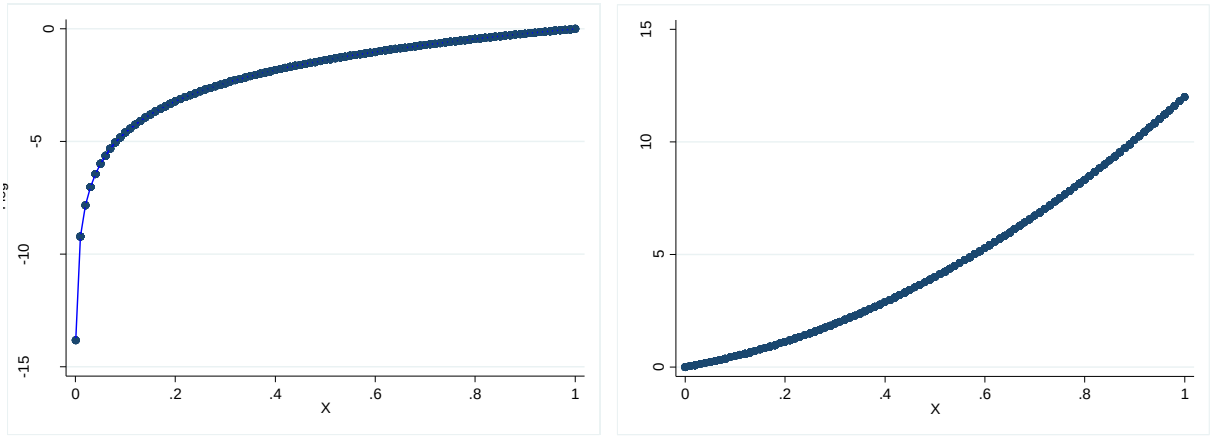
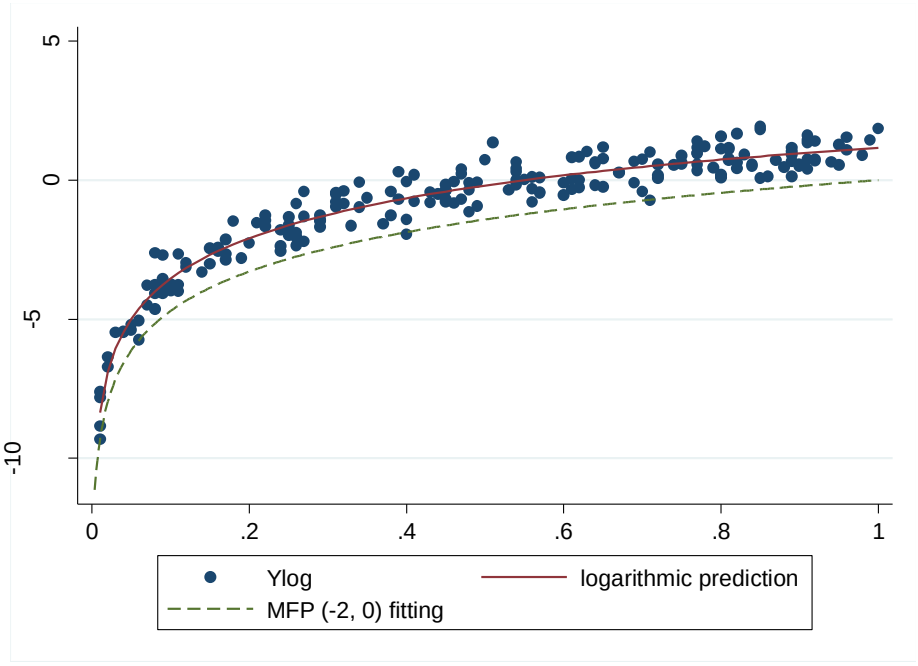
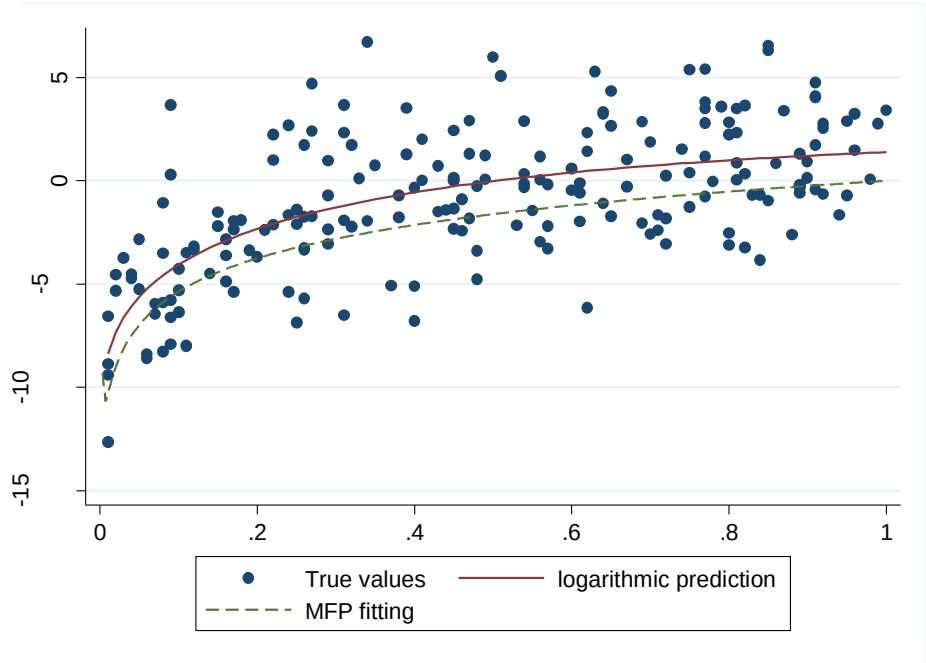


Figure 3
Predictions from FP2 and FP1 (logarithmic) fittings – n=200.





Upper panel – high fit, small error; lower panel – more noise – 8*error.

Figure 4A

Comparisons of FP2 and quadratic fittings to an underlying S-shape relationship – high fit

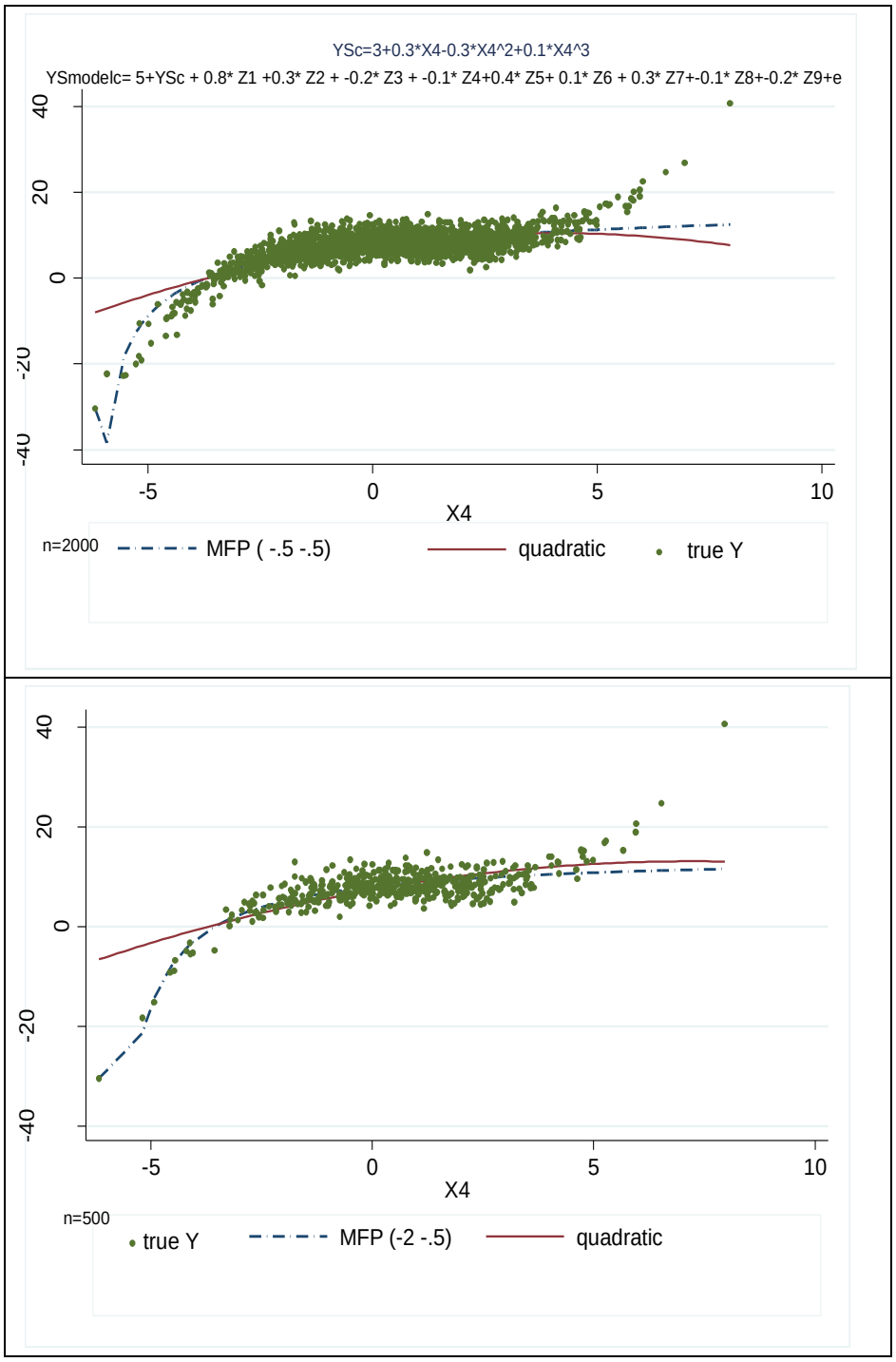


Figure 4A – (continued)

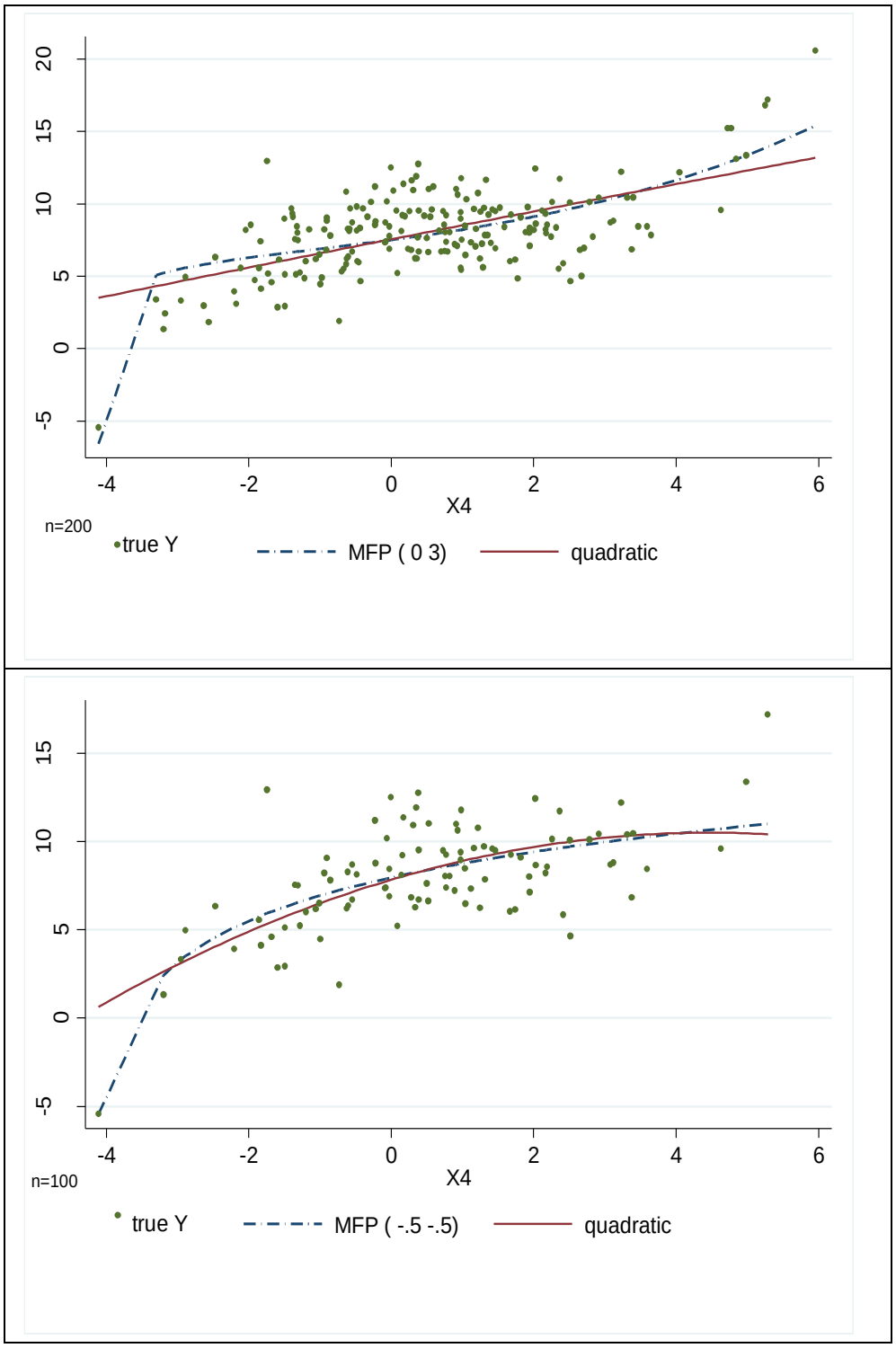


Figure 4B

Comparisons of FP2 and quadratic fittings to an underlying S-shape relationship – lower fit

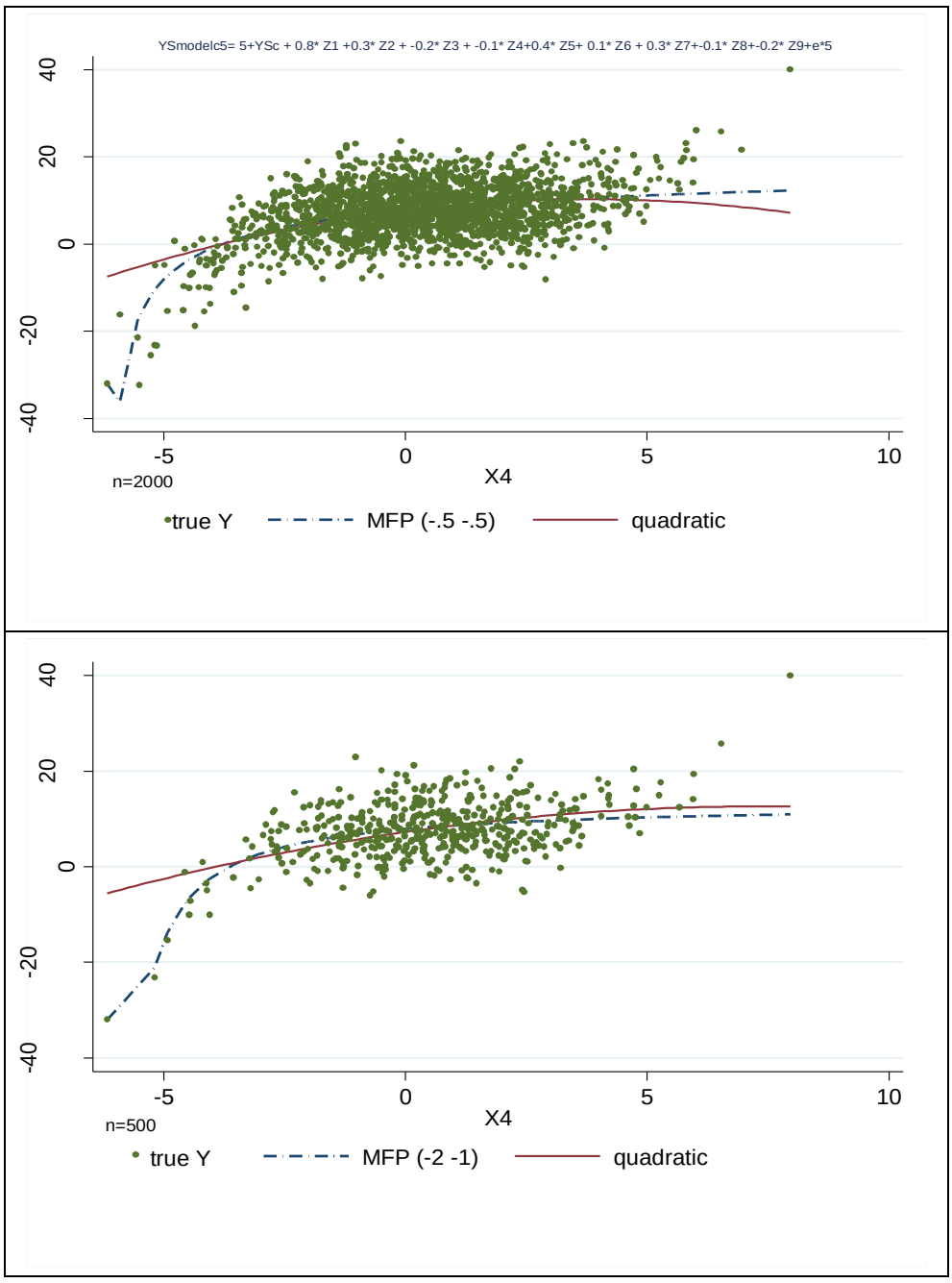
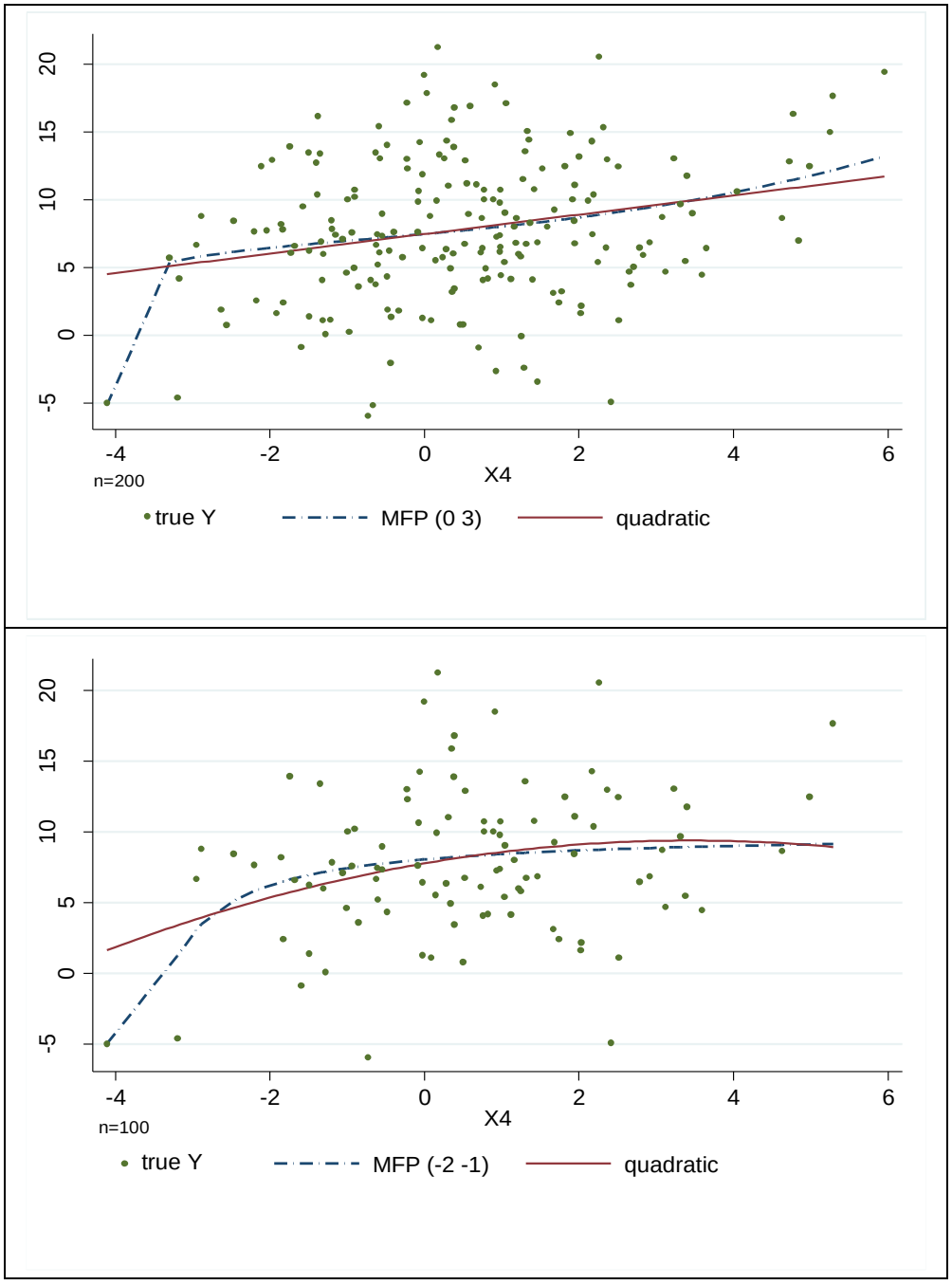


Figure 4B – (continued)



Appendix

Simulations Data Generation

Table 1: For the logarithmic function, ln(X) is drawn from $N(0,1)$, and $Y = \beta\log(X) + e$, where β is .33 for R^2=.1, .5 for R^2=.2, and .65 for R^2=.3. For the quadratic and S-shape cases, X is drawn from $N(4,1)$ and $Y = \beta(X-4)^2 + e$ and $Y = \beta(X-4)^3 + e$ respectively.

Table 2: Four of the control variables varied between 0 and 1. They were random draws from a uniform distribution with boundaries at 0 and 1. Two of the control variables were simulated to be indicator variables. They were simulated by again taking random draws from a uniform distribution with the supports (0,1). The random draws were then rounded to integer values. Two control variables were random draws from a uniform distribution supported on the interval (-2,2). The last two control variables could take integer values between 1 and 10. They were simulated by taking random draws from a uniform distribution with the support being (1,10) and then the values were truncated to integer values. The independent variable X was also a random draw from U(0,1). The dependent variable Y was determined from the equation, $= \gamma Z + f(X) + \epsilon$, where *Z* is the vector of 10 control variables, γ is (0.8,0.3, -0.2, -0.1,0.4, 0.1, 0.3,-0.1,-0.2,0.01,), $\epsilon \sim N(0,1)$, and f(*X*) for log function is $2\ln(X)$ and for quadratic function is $4X+8X^2$.

Figure 4: X was drawn from N(0.5,2), Z1-9 are drawn from normal distributions with the following means and SDs: means(0,0,2,0,1,2,1,1,1) SDs(1,1,3, 1.5, 2,2,4, 1, 1.5). The S-shape function is $Y=8+0.3X-0.3X^2+0.1X^3+ \gamma Z + \epsilon$ where γ is (0.8, 0.3, -0.2, -0.1,0.4, 0.1, 0.3,-0.1,-0.2), $\epsilon \sim N(0,1)$.