Smoothing Spline as a Guide to Elaborate Explanatory Modeling

Chon Van Le
School of Business
International University - VNU HCMC
Quarter 6, Linh Trung, Thu Duc Dist.
Ho Chi Minh City, Vietnam
Email: lvchon@hcmiu.edu.vn

July 15, 2017

Abstract

Although there are substantial theoretical and empirical differences between explanatory modeling and predictive modeling, they should be considered as two dimensions. And predictive modeling can work as a "fact check" to propose improvements to existing explanatory modeling. In this paper, I use smoothing spline, a nonparametric calibration technique which is originally designed to intensify the predictive power, as a guide to revise explanatory modeling. It works for the housing value model of Harrison and Rubinfeld (1978) because the modified model is more meaningful and fits better to actual data.

Keywords: Predictive econometrics, calibration, smoothing spline

1 Introduction

Econometric modeling has two main purposes: causal explanation and empirical prediction. Causal analysis determines whether an independent variable really affects the dependent variable and estimates the magnitude of the effect. And the second goal is to make predictions about the dependent variable, given the observed values of the independent variables.

However, the two objectives have received unequal treatment among academic researchers, especially those who work with cross section and panel data. The bulk of most econometric textbooks is given to issues relevant to causal explanation. Shmueli (2010) attributed this to the assumption that models with high explanatory power are expected to have high predictive power. He argued that there is clear distinction, both theoretically and empirically, between explanatory modeling and predictive modeling.

Although many academic statisticians may consider prediction as unacademic (for example, Kendall and Stuart (1977) and Parzen (2001)), it has played an increasingly important role in banking and financial services and other industries. Recent emergence of Big Data has fueled explosive growth of predictive modeling in the last decade. Large volumes, high velocity inflow, and a wide variety of data, including semi-structured and unstructured data such as text and images, have provided organizations with various opportunities for keeping track of and making prompt adjustments to their performance (Laney, 2001). Competent predictive modeling to analyze big data has become a key to gaining insights, forecasting the future, and automating non-routine decision making so that they can make big improvements such as better risk management, smoothened operations, personalized services, etc (ICAEW, 2014).

On the other hand, heavy emphasis placed on explanatory modeling has led to a widespread acceptance of regression results in many academic papers with highly signicant/insignificant coefficients but low R^2. A large proportion of variation in the dependent variable left unexplained may imply that several important independent variables are missing or the functional form that represents the relationship between the regressand and regressor(s) is not appropriate. In addition, since the linear regression model which is used most popularly is an approximation to some unknown, underlying function, such an approximation is likely to be useful only over a small range of variation of the independent variables around their means. It can be asserted that whether or not a factor has an impact on the dependent variable, but its magnitude, if any, cannot because it may vary across different observations.

Therefore, calibration techniques under predictive modeling should be used to improve model specification. Although they cannot explain what is wrong in the orginal model, they can work as a guide to elaborate explanatory modeling. In this paper, I focus particularly on smoothing spline and apply it to Harrison and Rubinfeld's (1978) study on housing prices for the Boston area. It is found that their model, though already good, can be fine-tuned to be more meaningful and to fit better to actual data.

The paper is structured as follows. Section 2 reviews predictive modeling in comparison with explanatory modeling. Section 3 presents calibration techniques, including smoothing splines. Section 4 outlines the data, the revised model and its regression results. Conclusions follow in Section 5.

2 Predictive Modeling[1]

In economics, statistical methods are used mainly to test economic theory. The theory specifies causal and effect relationships between variables. Explanatory modeling applies statistical models to data for testing these relationships, that is, whether a particular independent variable does influence the dependent variable. In this type of modeling, the theory takes a governing role and the application of statistical methods is done strictly through the lens of the theory (Shmueli, 2010).

In contrast, predictive modeling refers to the application of statistical models or data mining algorithms to data for predicting the dependent variable. According to Shmueli

[1] This part is to a large extent based on Shmueli (2010).

(2010), prediction involves point or interval prediction, predictive distribution, or ranking of new or future observations. There is a disagreement among statisticians on the value of predictive modeling. Many see it as unacademic. Berk (2008) indicates that researchers in social sciences only did "causal econometric style". As a consequence, many statistics textbooks wrote very little on predictive modeling. Others such as Geisser (1975), Aitchison and Dunsmore (1975), Friedman (1997) consider prediction as the foremost statistical application. This is true for most organizations and individuals outside of academia that are overwhelmed with exponential growth of data as a result of digital technology, mobile technology, social media, public sector open data, computer chips and sensors implanted in physical assets, etc. Increasingly large and rich datasets often consist of complex patterns and relationships that are beyond the reach of existing theories. Predictive modeling can help reveal new hypotheses and causality or propose improvements to existing explanatory models.

Shmueli (2010) clarifies that explanatory power can testify the strength of a causal hypothesis but cannot assess the distance between theory and empirics. Predictive modeling can work as a "fact check" to evaluate the relevance of theories in the light of actual data. Hence competing theories can be compared by examining their respective predictive power. Ehrenberg and Bound (1993) state that predictive modeling may create benchmarks of predictive accuracy under which scientific development can lead to substantial theoretical and practical gains.

The key element leading to the difference between explanatory and predicting modeling is errors of measurement. Observed data normally do not measure accurately their underlying constructs or variables. There has been considerable discussion in empirical studies on how to obtain reasonable measures of interest rates, profits, services from capital stocks, etc. For this reason, "the operationalization of theories and constructs into econometric models and measurable data creates a disparity between the ability to explain phenomena at the conceptual level and the ability to generate predictions at the measurable level" (Shmueli, 2010).

The disparity justifies important differences in several aspects. Firstly, in explanatory modeling, a statistical model "usually begins with a statement of a theoretical proposition" (Greene, 2011). It is built based on an economic model that consists of mathematical equations that describe deterministic relationships between independent variables and the dependent variable. The statistical model represents a causal relationship, and independent variables are assumed to cause the dependent variable. In predictive modeling, the statistical model is often built from the data. It shows the association between independent variables and the dependent variable. Interpreting the relationship between independent variables and the dependent variable is not required.

Secondly, Shmueli (2010) claims that explanatory modeling is backward-looking. A statistical model is used to test already existent hypotheses. On the contrary, predictive modeling is forward-looking. The statistical model is built to predict new observations.

Thirdly, explanatory modeling focuses on minimizing bias to secure reliable estimates of the "true" model coefficients. In contrast, since the goal of predictive modeling is to obtain optimal predictions of the dependent variable based on a regression model of whatever variables are available, it seeks to minimize the sum of bias and estimation variance (Shmueli, 2010). Therefore, in predictive modeling, bias is not a big problem and can be tolerated as long as empirical precision is improved. However, bias is a crucial issue in explanatory modeling. Several methods have been proposed to deal with the omission of variables that affect the dependent variable and are correlated with independent variables that are present in the statistical model.

Fourthly, explanatory modeling aims to estimate the theory-based statistical model with adequate statistical power for hypothesis testing. Consequently, multicollinearity is often a major concern in causal explanation. When two or more independent variables are highly correlated, the estimator is still unbiased but less precise. In predictive modeling, all these independent variables should be included if each variable contributes significantly to the predictive power of the model.

In addition, sufficient data are required for statistical inference. A variety of data imputation methods such as zero-order method, data augmentation and multiple imputation techniques, inverse probability weighting, etc. have been used to fill gaps in data sets. These methods seem to be constructed for parameter estimation and hypothesis testing. However, according to Shmueli (2010), "beyond a certain amount of data, extra precision is negligible for purposes of inference". Predictive modeling needs a larger sample to reduce bias and variance and to create holdout datasets for prediction testing.

Although there are essential differences between explanatory modeling and predictive modeling, they should be considered as two dimensions. And a statistical model should be evaluated based on its explanatory power and predictive power, whose weights depend on the question of interest. The predictive power can be intensified by calibration techniques which are discussed in the next section.

3 Calibration and Splines

Calibration dates back to the early 1980s in dynamic computable general equilibrium models (Kydland and Prescott, 1982). It is a procedure to select numerical values for the parameters of a model because sometimes no data are available to estimate its parameters. Canova (1994) points out that econometric estimation and calibration are two approaches to exposing general equilibrium models to data. They both begin with formulating a general equilibrium model and selecting funtional forms for production, utility, and exogenous factors. But they are different in choosing the parameters. The estimation approach believes that the model provides an accurate description of the data, or the model is true, is a data-generating process, and tests what attributes of the model are false, diverge significantly from the data. The calibration approach assumes the opposite view as Box (1976) states that "all models are wrong". As a result, the theoretical model should be modified or calibrated to gain a better approximation of the observed data. Otherwise, the model can produce systematically biased predictions, either too high or too low on average, so cannot be used for economic decisions.

According to Stine (2011), "a model is calibrated if its predictions are correct on average," or

$$\mathrm{E}(y|\hat{y}) = \hat{y}.$$

The adjustment procedure starts with the non-calibrated predicted value from a regression:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k.$$

The predictive ability of the model can be improved if $\hat{y}$ can be transformed to a better predictor, for example, $\hat{\hat{y}} = h(\hat{y})$ where h is a smooth function.

A popuplar smooth function is a spline function that was first applied by Poirier and Garber (1974). In order to be continuous and continuously differentiable at the 'joins' or 'knots', the spline must take a form of quadratic, cubic or a higher-degree polynomial. Suppose that there are two known knots $\hat{y}_a$ and $\hat{y}_b$ and that we use a cubic spline:

$$\hat{\hat{y}} = \alpha_{i0} + \alpha_{i1}\hat{y} + \alpha_{i2}\hat{y}^2 + \alpha_{i3}\hat{y}^3 + \varepsilon, \tag{1}$$

where the subsets are defined by

$$i = \begin{cases} 1 & \text{if } \hat{y} \leq \hat{y}_a, \\ 2 & \text{if } \hat{y}_a < \hat{y} \leq \hat{y}_b, \\ 3 & \text{if } \hat{y}_b < \hat{y}. \end{cases}$$

Continuity of $\hat{\hat{y}}$ and the first derivatives at the knots requires:

$$\begin{aligned}
&\alpha_{10} + \alpha_{11}\hat{y}_a + \alpha_{12}\hat{y}_a^2 + \alpha_{13}\hat{y}_a^3 = \alpha_{20} + \alpha_{21}\hat{y}_a + \alpha_{22}\hat{y}_a^2 + \alpha_{23}\hat{y}_a^3, \\
&\alpha_{20} + \alpha_{21}\hat{y}_b + \alpha_{22}\hat{y}_b^2 + \alpha_{23}\hat{y}_b^3 = \alpha_{30} + \alpha_{31}\hat{y}_b + \alpha_{32}\hat{y}_b^2 + \alpha_{33}\hat{y}_b^3, \qquad (2) \\
&\alpha_{11} + 2\alpha_{12}\hat{y}_a + 3\alpha_{13}\hat{y}_a^2 = \alpha_{21} + 2\alpha_{22}\hat{y}_a + 3\alpha_{23}\hat{y}_a^2, \\
&\alpha_{21} + 2\alpha_{22}\hat{y}_b + 3\alpha_{23}\hat{y}_b^2 = \alpha_{31} + 2\alpha_{32}\hat{y}_b + 3\alpha_{33}\hat{y}_b^2.
\end{aligned}$$

Additional restrictions should be imposed on the spline (1) before estimation if we want the second derivatives to be continuous:

$$\begin{aligned}
2\alpha_{12} + 6\alpha_{13}\hat{y}_a &= 2\alpha_{22} + 6\alpha_{23}\hat{y}_a, \\
2\alpha_{22} + 6\alpha_{23}\hat{y}_b &= 2\alpha_{32} + 6\alpha_{33}\hat{y}_b.
\end{aligned}$$

The cubic spline though allows discontinuities in the third derivatives at the knots.

So far we have assumed that we have specified the locations of the knots in advance. But in most cases, without further information on abrupt changes over time or size thresholds, it is impossible to determine the knots beforehand. Then we can resort to nonparametric smoothers. The oldest and simplest one is the smoothing spline that connects the medians of equal-width intervals. The number of intervals can be chosen by

$$\text{Number of intervals} = \max\{\min(b_1, b_2), b_3\},$$

where $b_1 = \text{round}\{10 \times \ln 10(N)\}$, $b_2 = \text{round}(\sqrt{N})$, $b_3 = \min(2, N)$, and N is the number of observations (StataCorp, 2011). In each interval, the median of y and the median of $\hat{y}$ are calculated. A spline is fit to these medians. If the spline appears to deviate much from the 45^0 line, then the original model can be modified to better capture the “true”, complex relationships between the dependent variable and independent variables. Consequently, the model would be more effective as an approximating function, that is, providing more rigorous hypothesis testing of a regressor's impact and an expectedly more accurate estimate of the magnitude of that impact. I suggest using smoothing splines as a guide to elaborate explanatory modeling. An example is presented in the next section.

4 Example of Harrison and Rubinfeld (1978)

In a study of the willingness to pay for air quality improvements in the Boston region, Harrison and Rubinfeld (1978) used data for 506 census tracts in the Boston Standard Metropolitan Statistical Area (SMSA) in 1970 to estimate a housing value equation

$$\begin{aligned}\text{Ln(Medianvalue)} &= \alpha_0 + \alpha_1\text{Room}^2 + \alpha_2\text{Age} + \alpha_3\text{Ln(Distance)} + \alpha_4\text{Ln(Highway)} + \alpha_5\text{Tax} \\ &+ \alpha_6\text{Pupil/Teacher} + \alpha_7(\text{Black - 0.63})^2 + \alpha_8\text{Ln(Lowstatus)} \qquad (3) \\ &+ \alpha_9\text{Crime} + \alpha_{10}\text{Zoning} + \alpha_{11}\text{Industry} + \alpha_{12}\text{Charles} + \alpha_{13}\text{Nox}^2 + \epsilon,\end{aligned}$$

Table 1: Variable Definitions

Variable	**Description**
Medianvalue	Median value of owner-occupied homes.
Room	Average number of rooms in owner-occupied homes.
Age	Proportion of owner-occupied homes built before 1940.
Black	Black proportion of population in the community.
Lowstatus	Proportion of population that is lower status = $\frac{1}{2}$(proportion of adults without some high school education and proportion of male workers classified as laborers).
Crime	Crime rate by town.
Zoning	Proportion of a town's residential land zoned for lots greater than 25,000 square feet.
Industry	Proportion of nonretail business acres per town.
Tax	Full value property tax rate ($/$10,000).
Pupil/Teacher	Pupil-teacher ratio by town school district.
Charles	Charles River dummy equals 1 if tract bounds the Charles River and 0 otherwise.
Distance	Weighted distance to 5 employment centers in the Boston area.
Highway	Highway access index.
Nox	Annual average nitrogen oxide concentration in pphm.

Source: Harrison and Rubinfeld (1978).

Table 2: Housing Value Models

	Harrison and Rubinfeld's (1978) Model[a] (3)		Revised Model (5)	
Constant	4.558***	(0.1544)	4.595***	(0.2921)
Room^2	0.0063***	(0.0013)	0.0231***	(0.0025)
Age	0.00009	(0.0005)	0.0052*	(0.0030)
Ln(Distance)	-0.1913***	(0.0334)	-0.2382***	(0.0301)
Ln(Highway)	0.0957***	(0.0191)		
Highway			0.0139*	(0.0076)
Tax	-0.0004***	(0.0001)		
Ln(Tax)			-0.1963***	(0.0427)
Tax600			-0.3011***	(0.0791)
Pupil/Teacher	-0.0311***	(0.0050)	-0.0316***	(0.0045)
(Black - 0.63)^2	0.3637***	(0.1031)	0.1240	(0.0934)
Ln(Lowstatus)	-0.3712***	(0.0250)		
Lowstatus			0.0547***	(0.0095)
Crime	-0.0119***	(0.0012)	-0.0225***	(0.0037)
Crime^2			0.00016***	(0.00005)
Zoning	0.00008	(0.0005)	0.0004	(0.0005)
Industry	0.0002	(0.0024)	-0.0074***	(0.0027)
Charles	0.0914***	(0.0332)	0.0474[b]	(0.0292)
Nox^2	-0.6380***	(0.1131)	-0.7763***	(0.1016)
Room×Age			-0.0009**	(0.00047)
Room×Lowstatus			-0.0118***	(0.0015)
Highway×Lowstatus			-0.0012***	(0.00016)
Highway×Industry			0.0019***	(0.0005)
Number of observations	506		506	
R^2	0.806		0.855	

Notes: ***, **, * significant at the 1%, 5%, 10% levels, respectively.

Standard errors in parentheses.

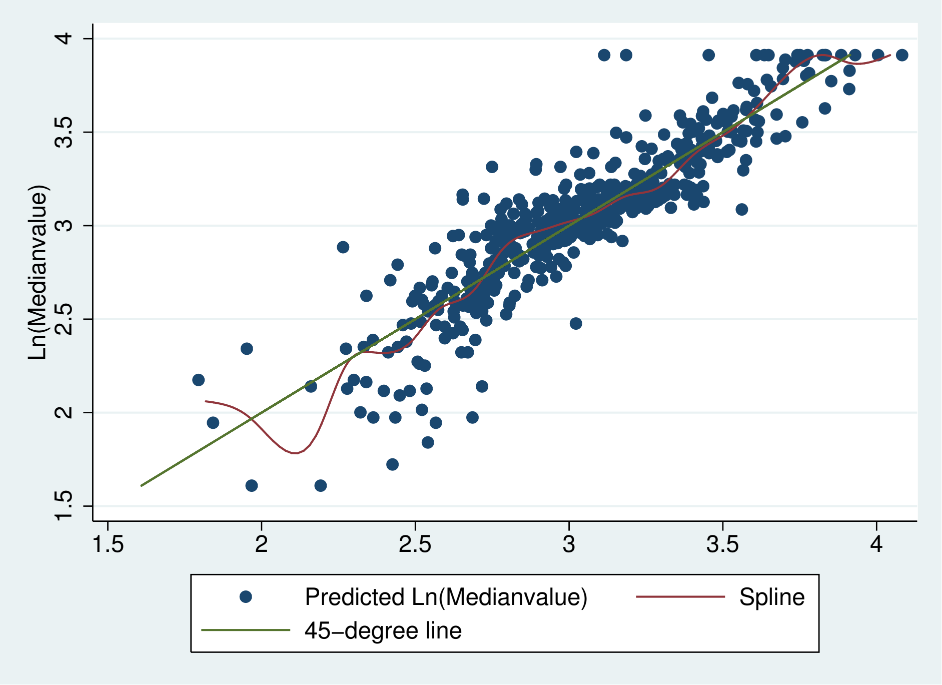
[a] Results are a little different from those in Harrison and Rubinfeld (1978).

[b] Charles River dummy has a p-value of 10.5.

Source: Author's calculation.

where variables are defined in Table 1. The results which are reported in the second and third columns of Table 2 seemingly provide strong evidence on the impacts of the independent variables, except Age, Zoning, and Industry. And R^2 is relatively high. However, Figure 1 indicates that the spline diverges considerably from the 45^0 line, especially at the small predicted values of the dependent variable.

Figure 1: Smoothing Spline based on Harrison and Rubinfeld's (1978) Model



The spline can be approximated by a 3^{rd} degree polynomial of the form:

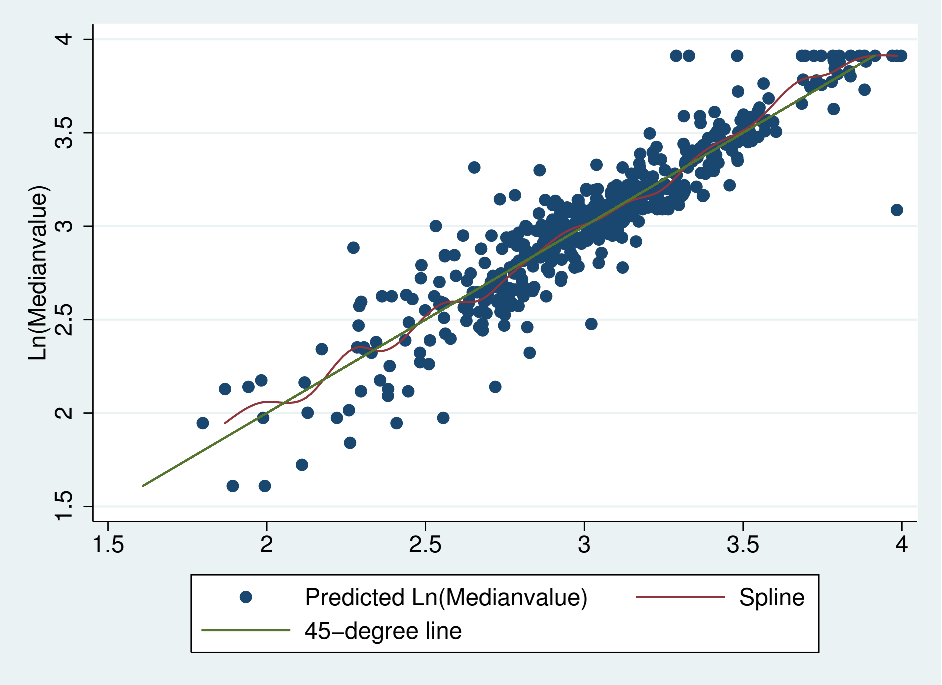
$$y = X\boldsymbol{\alpha} + \gamma_2(\hat{y} - \bar{y})^2 + \gamma_3(\hat{y} - \bar{y})^3 + \xi, \tag{4}$$

where X and $\hat{y}$ are the set of independent variables and the predict values of the original model.

The second and third columns of Table 3 show that the second term in equation (4) that approximates the spline of the Harrison and Rubinfeld's model is significant at 5% level. Moreover, the last two terms in equation (4) are jointly significant at 10% level (their F-statistics is 2.40), implying that the polynomial (4) fits the data better than the Harrison and Rubinfeld's model. It suggests that the model can be revised.

A closer look at the data set proposes several modifications as follows. Firstly, since the index of accessibility to radial highways takes nine discrete values, namely, 1, 2, 3, 4, 5, 6, 7, 8, and 24, the Highway index should be used instead of its log which may be less meaningful. Secondly, the Tax variable is replaced by its log. While most of the observations in the

Figure 2: Smoothing Spline based on Revised Model



sample have tax rates ranging from \$187 to \$469 per \$10,000, 137 tracts have unusually high tax rates of \$666 and \$711. A dummy variable, Tax600 which equals 1 if the tax rate is over \$600 and 0 otherwise, is included. Thirdly, as the lower status is measured as a percentage of the population in the community, the Lowstatus variable should be in original form, not in logarithmic form. Fourthly, squared crime rate, Crime^2, is added to allow for a changing impact of the crime rate on the housing value. Fifthly, house prices depend on various attributes which are considered not only separately but also together. Therefore, four interaction terms are included. The modified housing value equation is

$$\begin{aligned}
\text{Ln(Medianvalue)} &= \beta_0 + \beta_1\text{Room}^2 + \beta_2\text{Age} + \beta_3\text{Ln(Distance)} + \beta_4\text{Highway} + \beta_5\text{Ln(Tax)} \\
&+ \beta_6\text{Tax600} + \beta_7\text{Pupil/Teacher} + \beta_8(\text{Black - 0.63})^2 + \beta_9\text{Lowstatus} \\
&+ \beta_{10}\text{Crime} + \beta_{11}\text{Crime}^2 + \beta_{12}\text{Zoning} + \beta_{13}\text{Industry} + \beta_{14}\text{Charles} \qquad (5) \\
&+ \beta_{15}\text{Nox}^2 + \beta_{16}\text{Room}\times\text{Age} + \beta_{17}\text{Room}\times\text{Lowstatus} \\
&+ \beta_{18}\text{Highway}\times\text{Lowstatus} + \beta_{19}\text{Highway}\times\text{Industry} + \varepsilon.
\end{aligned}$$

Table 3: Polynomials Approximating Splines of the Two Models

	Harrison and Rubinfeld's (1978)		Revised Model	
Constant	4.635***	(0.1581)	4.531***	(0.2951)
Room^2	0.0072***	(0.0014)	0.0271***	(0.0033)
Age	0.0002	(0.0005)	0.0059**	(0.0030)
Ln(Distance)	-0.2048***	(0.0342)	-0.2519***	(0.0310)
Ln(Highway)	0.0997***	(0.0196)		
Highway			0.0151**	(0.0076)
Tax	-0.00046***	(0.0001)		
Ln(Tax)			-0.2048***	(0.0429)
Tax600			-0.2944***	(0.0820)
Pupil/Teacher	-0.0329***	(0.0051)	-0.0322***	(0.0045)
(Black - 0.63)^2	0.3314***	(0.1055)	0.1108	(0.0942)
Ln(Lowstatus)	-0.3880***	(0.0274)		
Lowstatus			0.0651***	(0.0111)
Crime	-0.0103***	(0.0023)	-0.0230***	(0.0040)
Crime^2			0.00016***	(0.00005)
Zoning	0.0003	(0.0005)	0.00055	(0.00046)
Industry	0.0006	(0.0024)	-0.0077***	(0.0028)
Charles	0.0985***	(0.0334)	0.0539*	(0.0295)
Nox^2	-0.6550***	(0.1139)	-0.7926	(0.1025)
Room×Age			-0.0010**	(0.00047)
Room×Lowstatus			-0.0134***	(0.0018)
Highway×Lowstatus			-0.0013***	(0.0002)
Highway×Industry			0.0019***	(0.0005)
$(\widehat{\text{Ln(MV)}} - \overline{\text{Ln(MV)}})^2$	-0.1371**	(0.0672)	-0.0913	(0.0675)
$(\widehat{\text{Ln(MV)}} - \overline{\text{Ln(MV)}})^3$	-0.0396	(0.0858)	-0.1096	(0.0668)
Number of observations	506		506	
R^2	0.808		0.856	

Notes: ***, **, * significant at the 1%, 5%, 10% levels, respectively.

Standard errors in parentheses.

Source: Author's calculation.

Figure 2 shows that the spline based on the revised model does not deviate much from the 45^0 line, even at the extreme predicted values. This is confirmed by the fact that the last two terms in the spline-approximating polynomial in the fourth and fifth columns of Table 3 are individually and jointly insignificant (their F-statistics is 1.68).

The regression results which are presented in the fourth and fifth columns of Table 2 differ in several aspects from those of Harrison and Rubinfeld. Age and Industry are now significant. Externalities associated with industrial activities such as noise, pollution, heavy traffic, and awful view negatively affect housing values as expected. Black proportion of population no longer has a positive impact on housing values, which makes more sense as black neighbors are often regarded as undesirable. In addition, the explanatory power of the revised model, though already good, still improves since R^2 increases by 5%. The example of Harrison and Rubinfeld demonstrates that smoothing spline helps elaborate modeling.

5 Conclusions

Greater weight has so far been put on causal explanation than on empirical prediction due to the wrong assumption that models having high explanatory power are supposed to have high predictive power. There are substantial theoretical and empirical differences between explanatory modeling and predictive modeling, but they should be considered as two dimensions. And predictive modeling can work as a "fact check" to propose improvements to existing explanatory modeling.

In this paper, I use smoothing spline, a nonparametric calibration technique which is originally designed to intensify the predictive power, as a guide to revise explanatory modeling. It works for the housing value model of Harrison and Rubinfeld (1978) as the modified model is more meaningful and fits better to actual data.

The world is producing enormous and complex amounts of data that often contain sophisticated patterns and relationships beyond the reach of current theories. Calibration techniques such as smoothing splines can help incorporate new data, new variables into explanatory models so that new hypotheses and causality can be revealed and tested. Furthermore, organizations and firms can capture and exploit new implications of big data for their own sake.

References

Aitchison, J., and Dunsmore, I. R., 1975, *Statistical Prediction Analysis*, New York: Cambridge University Press.

Berk, R. A., 2008, *Statistical Learning from a Regression Perspective*, New York: Springer-Verlag.

Box, G. E. P., 1976, "Science and Statistics," *Journal of the American Statistical Association*, 71, 791-799.

Canova, F., 1994, "Statistical Inference in Calibrated Models," *Journal of Applied Econometrics*, 9 (S), S123-S144.

Ehrenberg, A., and Bound, J., 1993, "Predictability and Prediction," *Journal of the Joyal Statisticial Society Series A*, 156 (2), 167-206.

Friedman, J. H., 1997, "On Bias, Variance, 0/1-loss, and the Curse-of-Dimensionality," *Data Mining and Knowledge Discovery*, 1, 55-77.

Geisser, S.,1975, "The Predictive Sample Reuse Method with Applications," *Journal of the American Statistical Association* 70, 320-328.

Greene, W. H., 2012, *Econometric Analysis*, 7th ed., Pearson.

Harrison, D., and Rubinfeld, D. L., 1978, "Hedonic Housing Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management* 5 (1), 81-102.

ICAEW, 2014, "Big Data and Analytics—What's New?" https://www.icaew.com/-/media/corporate/archive/files/technical/information-technology/technology/what-is-new-about-big-data-v2.ashx.

Kendall, M., and Stuart, A., 1977, *The Advanced Theory of Statistics*, 4th ed., New York: Macmillan.

Kydland, F. E., and Prescott, E. C., 1982, "Time to Build and Aggregate Fluctuations," *Econometrica* 50 (6), 1345-1370.

Laney, D., 2001, "3D Data Management: Controlling Data Volume, Velocity and

Variety," Gartner. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Parzen, E., 2001, "Comment on Statistical Modeling: The Two Cultures," *Statistical Science* 16 (3), 224-226.

Poirier, D. J., and Garber, S. G., 1974, "The Determinants of Aerospace Profit Rates 1951-1971," *Southern Economic Journal,* 41 (2), 228-238.

Shmueli, G., 2010, "To Explain or to Predict?" *Statistical Science* 25 (3), 289-310.

StataCorp, 2011, *Stata Release 12: Statistical Software,* College Station, TX: StataCorp LP.

Stine, R. A., 2011, *Lecture Notes on Advanced Quantitative Modeling*, The Wharton School at the University of Pennsylvania.