April 29, 2019

Exploration of Real Estate in Taiwan using Multivariate Smooth Spline Regression

Cesar Rene Pabon Bernal,*[a] Peter Salamon *[a]

Smooth spline non-parametric regression is a well documented technique for analysis of multivariate data.[1] By fitting a smooth curve to a function of a data set, the intention is to minimize the variability of its “loss” and “roughness” through control of the tuning parameter λ. We propose a flexible generalized additive model (GAM) for determining the quantitive response of house prices per unit area on the basis of four predictors (distance to the nearest metro station, number of convenience stores within walking distance of house, house age, latitude) in New Taipei City, Taiwan.[2] Results reveal that a smooth spline regression model offers better results than polynomial regression, normal regression spline, and multiple linear regression. Proposed alternatives and further studies are explored in the conclusions section.

A. Introduction

In this report, we analyze multivariate smooth spline regression as a generalized additive model (GAM). We first review uni-variance in smoothing splines, their extension into multi variance and application in additivity for quantitative responses.

i. Single Variance in Smooth Splines

Single predictor smoothing splines were proposed by Whittaker(1923), Schoenberg (1964) and Reinsch (1967).[3] Given $y_i = f(x_i) + E_i, 1 = 1,...,n$ where f is an unknown smooth function and E_i are random errors, a natural cubic smoothing spline of $g(x;\lambda)$ is the function that minimizes:

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt \qquad \textbf{eq. 1}$$

Our objective is to minimize the first term, known as RSS or the “loss function” of *eq.1:* $\sum (y_i - g(x_i))^2$, using $g(x)$ and the second term, known as the “roughness penalty term” of *eq 1:* $\lambda \int g''(t)^2 dt$, by tuning the non-negative parameter λ. We state that the second term of the second derivative of the function, $g''(t)$, is a measure variability or “wiggle” and λ controls the bias-variance of the smoothing spline. Hence when:

a) $\lambda = 0$,function g can interpolate the data
b) $\lambda \to \infty$, g is perfectly smooth & be least squares line[4]

Since the tuning parameter λ controls the roughness of the smoothing spline, it also has great effective on the degrees of freedom and thus, df_λ is a measure of the smoothing spline— the greater df_λ is, the more flexible (lower bias/higher variance) the smoothing spline becomes.

In this study, both df_λ and λ are found computationally using *leave-one-out* cross validation (LOOCV).

i. Multi Variance in Smooth Splines as a GAM

Generalized additive models were proposed by Trevor Hattie and Robert Tibshirani (1986).[5] GAMs provide a natural framework to extend multiple variance in smoothing spline non-parametric regression for quantitive responses by replacing terms in y_i of a multiple linear regression model:

$$y_i = B_o + B_i x_i + B_2 x_2 + \ldots + B_p x_{i\ p} + E_i \qquad \textbf{eq. 2}$$

to a more flexible and exotic model of y_i:

$$y_i = B_o + f_i(x_{i\ 1}) + f_2(x_{i\ 2}) + \ldots + f_p(x_{i\ p}) + E_i \qquad \textbf{eq. 3}$$

The linear component of eq 2 $B_p x_{ip}$ is replaced with a smooth non-linear function of eq 3 $f_p x_{ip}$. Since each f_p is calculated separately for each x_{ip} then added together, we call the process additive. The new y_i in eq 3 is then replaced with the y_i in eq 1. The ability to extend a univariate model using multiple predictors through additivity are attractive and offers several advantages:

a) Fitting smooth non-linear f_p to each x_{ip} is automatic
b) Due to additivity, f_p & x_{ip} can be studied separately
c) Non-linear fits can offer more accurate responses
d) Smoothness of f_p can be summarized via degrees of freedom

B. Experimental

In this report, R software was utilized for a multivariate analysis of a market historical set in 2018 of real estate valuation from Sindian District, New Taipei City, Taiwan. The raw data contained 415 observations and 8 variables:

1.DOI: 10.1097
2.DOI: 10.1016
3.DOI: 10.1214
4.DOI: James, Gareth, et al. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2017.
5.DOI: 10.1201

1. ID
2. Transition date
3. House age
4. Distance to the nearest metro station
5. Number of convenience stores within walking distance of house
6. Latitude
7. Longitude
8. House price of unit area

i. Feature Selection

Forward, Backward and Stepwise Feature Selection was utilized for feature selection.

ii. Application of a GAM using Smoothing Spline Regression

Once the best predictors were chosen, a smoothing spline was individually fit to each variable and LOOCV was utilized to determine the ideal smoothing parameter λ. Specifying details in the R program, df_λ can then be extracted and applied to a GAM smoothing spline multivariate function.

iii. Model Performance

Analysis of Variance (ANOVA) was performed on fours models: GAM smoothing spline, regular multi linear regression, polynomial regression, and regression spline. We hypothesize that due to its flexibility it performs best versus other regression methods.

C. Results

According to literate, smooth spline non-parametric regression is a well documented technique for analysis of multivariate data. In this report, we implement this method using R software in real estate valuation data from Sindian District, New Taipei City, Taiwan.

i. Feature Selection

Forward, Backward and Stepwise Feature Selection was applied for feature selection. Fig. 1 shows the correlation results for the best four predictors, they are:

1. Distance to the nearest metro station
2. Number of convenience stores within walking distance of house
3. House age
4. Latitude

	row	col	corr
8	dist.nearest.mrt	longitude	-0.80631677
12	dist.nearest.mrt	house.price	-0.67361286
3	dist.nearest.mrt	num.convenience.stores	-0.60251914
5	dist.nearest.mrt	latitude	-0.59106657
11	house.age	house.price	-0.21056705
7	house.age	longitude	-0.04852005
1	house.age	dist.nearest.mrt	0.02562205
2	house.age	num.convenience.stores	0.04959251
4	house.age	latitude	0.05441990
10	latitude	longitude	0.41292394
6	num.convenience.stores	latitude	0.44414331
9	num.convenience.stores	longitude	0.44909901
15	longitude	house.price	0.52328651
14	latitude	house.price	0.54630665
13	num.convenience.stores	house.price	0.57100491

Figure 1: Correlation Results for Best Predictors

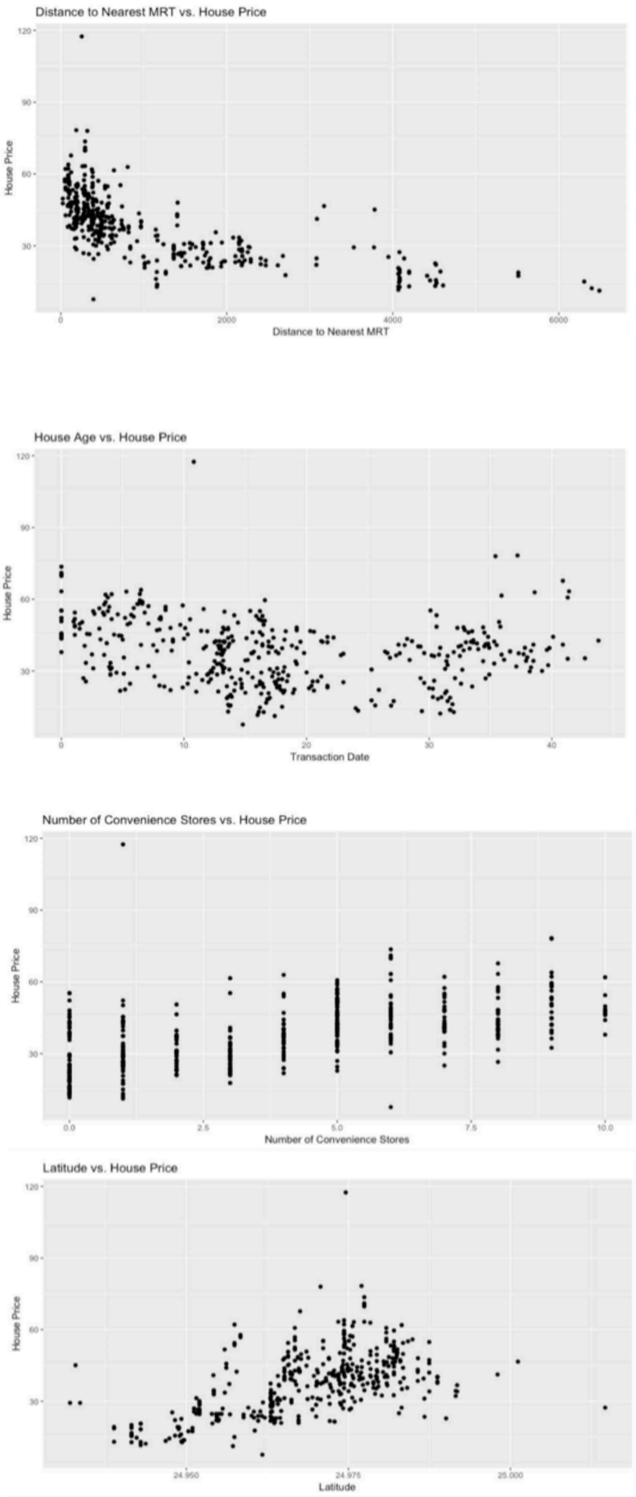


Figure 2(a-d): Graphs for Best Predictors

Please do not adjust margins

ii. Application of a GAM using Smoothing Spline Regression

In the introductory section of this report, we stated the equation for smooth spline regression as:

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt \qquad \textbf{eq. 1}$$

and its extension as a GAM for analysis of multi-variance for y_i as:

$$y_i = B_o + f_i(x_{i\ 1}) + f_2(x_{i\ 2}) + \ldots + f_p(x_{i\ p}) + E_i \qquad \textbf{eq. 3}$$

Utilizing R software and the function gam(), the output for a new y_i of eq 3 using the four best predictors is:

$$houseprice = -7.822e03 - 4.11e03(nearestmetro) + 3.684e01(conviencestore) - 2.342e01 + 3.15(houseage) + 3.15e02(latitude)$$

In order to capture the best smoothness of the regression curve, we use this new y_i, and tune the parameter λ by exploring ideal df_λ for smoothing the spline allowing minimization of the first term of *eq.1:* $\sum (y_i - g(x_i))^2$, and the roughness of the second term of *eq 1:* $\lambda \int g''(t)^2 dt$.

We find the tuning smoothing parameter λ by individually fitting each variable and implementing LOOCV. Specifying details in the R software, df_λ can then be extracted, and applied to the GAM smoothing spline function. Ultimately, this calculates the best smoothing parameters in multivariate regression. Table 1 lists the results:

Variable	Degrees of Freedom
Distance to the Nearest Metro Station	24.74904
Number of Convenience Stores (Within Walking Distance)	11.00001
House Age	21.69036
Latitude	10.62293

Table 1: Ideal df_λ for GAM smoothing spline

Based on the additive nature of GAMs, we are able to individually analyze, the fit spline for each independent variable versus the dependent variable. Because GAMs take similar statistical properties as linear models, we can extend its definition to include other linear regression methods and explore their outcome. Figue 3 are model fit plots for multi linear regression, polynomial regression, smooth spline regression and regular spline regression.

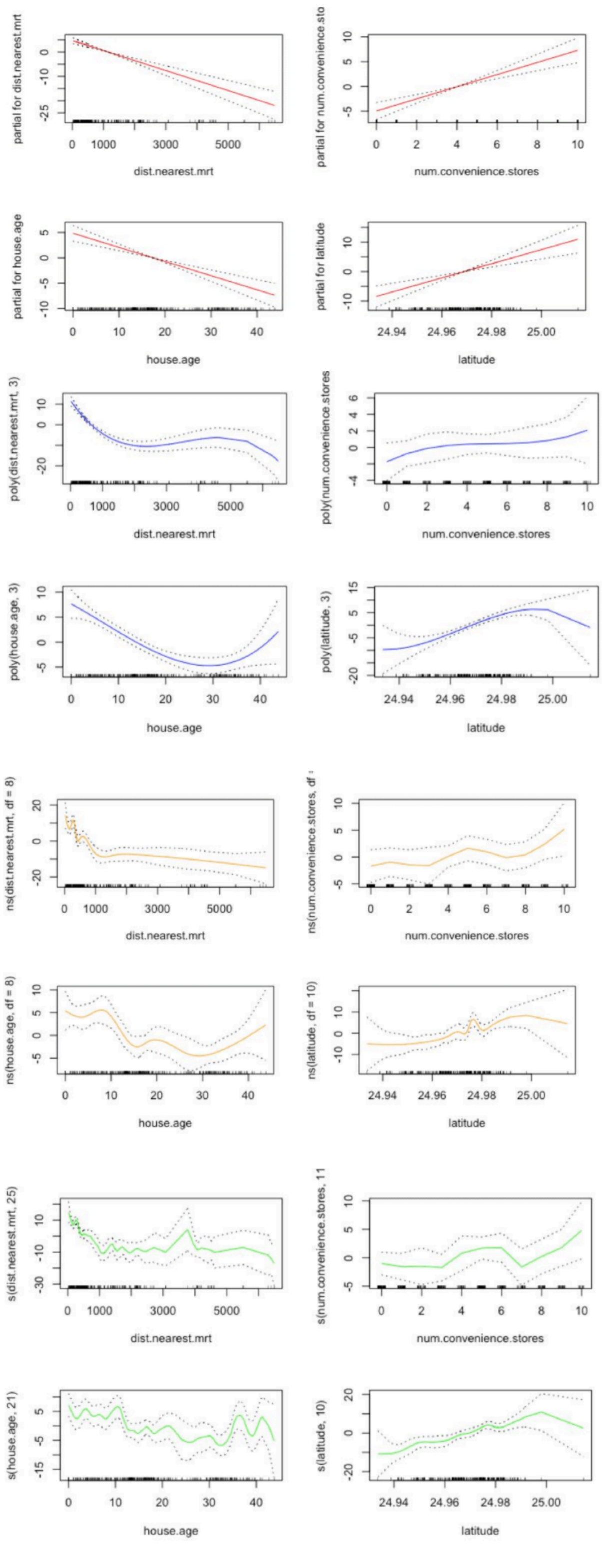


Figure 3: df_λ model fit plot of GAMs of multi linear regression, polynomial regression, smooth spline regression and regular spline regression.

Visually, we can state that implementation of additivity in polynomial regression, gave us the most smoothest curve versus multi linear regression, smooth spline regression and regular spline regression. However, according to literature, we know that when gradual addition of parameters is given to y_i of eq 3 (GAM), smoothing spline provides better flexibility and ultimately, depicts a more accurate response.

iii. Model Performance

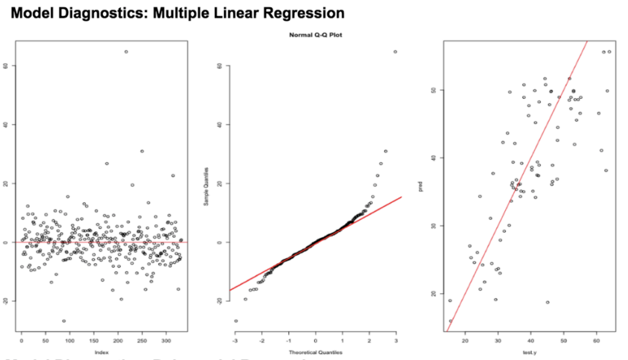
Table 2 provides the results of Analysis of Variance (ANOVA) of fours models: multi linear regression, polynomial regression, smooth spline regression and regular spline regression.

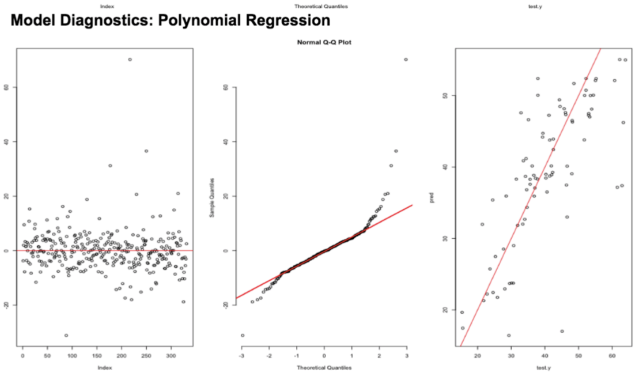
Model Performances

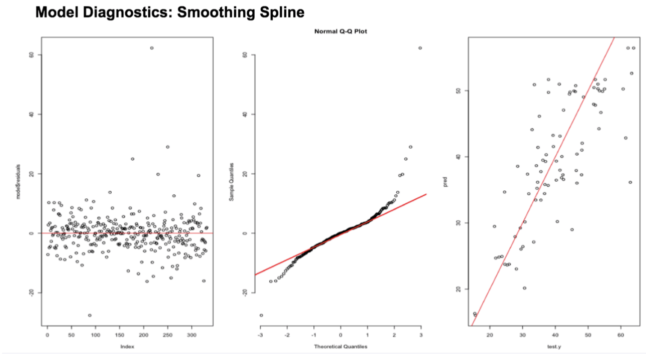
Model	Test MSE	Test RMSE
Average Train House Price	135.8761	11.65659
Multiple Linear Regression	76.78634	8.762781
Polynomial Regression	59.64175	7.722807
Smoothing Spline	51.8825	7.202951
Regression Spline	60.88662	7.802988

Table 2: ANOVA results for model performance

Figure 4 provides model diagnostics of fours models: multi linear regression, polynomial regression, smooth spline regression and regular spline regression.







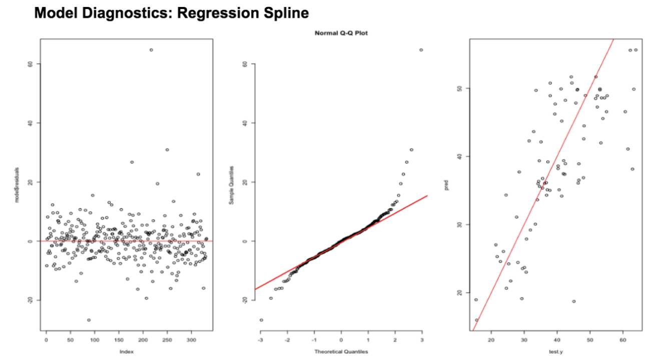


Figure 4: model diagnostics of fours models: multi linear regression, polynomial regression, smooth spline regression and regular spline regression..

The model performance of four models confirms that smooth spline regression provides the greatest flexibility. It has the lowest Test MSE and Test RMSE. This confirms the attractive nature of smooth spline regression for multivariate analysis using GAMs in comparison to other linear regression methods

Conclusions

In this report, we have analyzed multivariate smooth spline regression as a generalized additive model (GAM) in real estate valuation data from Sindian District, New Taipei City, Taiwan. Appropriate feature selection was made and the four best predictors were chosen; distance to the nearest metro station, number of convenience stores within walking distance of house, house age, and latitude. Tuning of smoothing parameter λ was utilized to lower roughness of the smoothing spline. When compared with other regression methods, our hypothesis was confirmed as smooth spline regression demonstrated the lowest test MSE and RMSE scores. Further studies include more flexible approaches such as random forests and boosting.

References

1. Howe, Chanelle J., et al. “Splines for Trend Analysis and Continuous Confounder Control.” *Epidemiology*, vol. 22, no. 6, 2011, pp. 874–875., doi:10.1097/ede.0b013e31823029dd.
2. Yeh, I-Cheng, and Tzu-Kuang Hsu. “Building Real Estate Valuation Models with Comparative Approach through Case-Based Reasoning.” *Applied Soft Computing*, vol. 65, 2018, pp. 260–271., doi:10.1016/j.asoc.2018.01.029.

Please do not adjust margins

3. Rice, John, and Murray Rosenblatt. “Smoothing Splines: Regression, Derivatives and Deconvolution.” *The Annals of Statistics*, vol. 11, no. 1, 1983, pp. 141–156., doi: 10.1214/aos/1176346065.
4. James, Gareth, et al. An Introduction to Statistical Learning: with Applications in R. Springer, 2017.
5. Hastie, T.j., and R.j. Tibshirani. “Generalized Additive Models.” *Generalized Additive Models*, 2017, pp. 136–173., doi:10.1201/9780203753781-6.