Exploration of Real Estate in Taiwan using Multivariate Smooth Spline Regression

By: Cesar Rene Pabon Bernal

OUTLINE

- Introduction
  - Theoretical
  - Practical
- Data
  - Exploratory Analysis
  - Feature Selection
  - Comparisons: Multiple Linear, Polynomial Regression, Regression Spline
- Results
  - Model Building: GAM Smoothing Spline Regression
  - Model Performance
- Conclusions

INTRODUCTION: THEORETICAL

- Smooth spline non-parametric regression is a well documented technique for analysis of multivariate data
- They are an extension of *Single variance smoothing splines*
- Single predictor smoothing splines were proposed by Whittaker(1923), Schoenberg (1964) and Reinsch (1967)
  - $$y_i = f(x_i) + E_i, 1 = 1,\ldots,n$$ eq. 1
    - *f* is an unknown smooth function
    - E_i are random errors

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. 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.

INTRODUCTION: THEORETICAL

- *Natural Cubic Smoothing Splines*
  - $$\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$ eq. 2
  - Minimize Loss Function/RSS using *g(X)* & Roughness by tuning non-negative parameter λ
    - *g"(t)* is a measure variability & *λ* controls the bias-variance
      - *λ* →∞, *g(t)* is perfectly smooth
    - df_λ is a measure of the smoothing spline
      - High df_λ, more flexible model (lower bias/higher variance)
    - df_λ and *λ* are found computationally using (LOOCV)

3. *James, Gareth, et al. An Introduction to Statistical Learning: with Applications in R. Springer, 2017.*

INTRODUCTION: THEORETICAL

- *Multi-variance in Smoothing Splines*
- Generalized additive models (GAMs) were proposed by Trevor Hattie and Robert Tibshirani (1986)
  - $y_i = B_o + B_i x_i + B_2 x_2 + \ldots + B_p x_{i\ p} + E_i$ eq. 3
  - $y_i = B_o + f_i(x_{i\ 1}) + f_2(x_{i\ 2}) + \ldots + f_p(x_{i\ p}) + E_i$ eq. 4
    - The linear component of *eq. 3* $B_p x_{ip}$ is replaced with a smooth non-linear function of *eq. 4* $f_p x_{ip}$
    - Each f_p is calculated separately for each x_{ip} then added together, we call the process <u>additive</u>

4. Hastie, T.j., and R.j. Tibshirani. "Generalized Additive Models." Generalized Additive Models, 2017, pp. 136–173., doi:10.1201/9780203753781-6.

INTRODUCTION: PRACTICAL

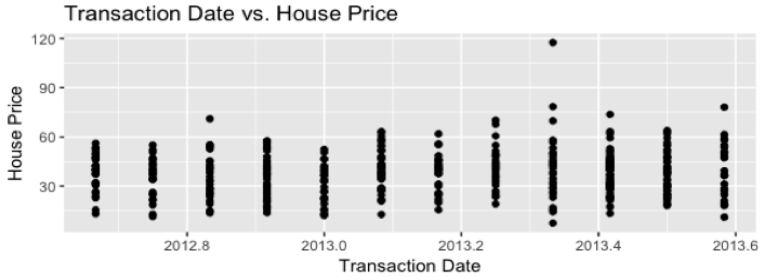
- Using *R Programming, we can* extend a univariate model using multiple predictors through additivity
- An an attractive technique:
  - Fitting smooth non-linear f_p to each x_{ip} is automatic
  - Due to additivity, f_p & x_{ip} can be studied separately
  - Smoothness of f_p can be summarized via degrees of freedom
- **We propose a flexible 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, house age, latitude in New Taipei City, Taiwan

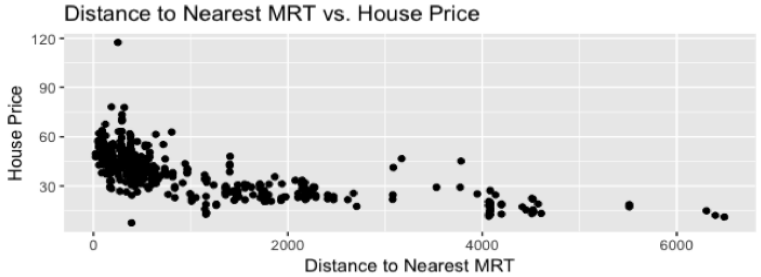
DATA

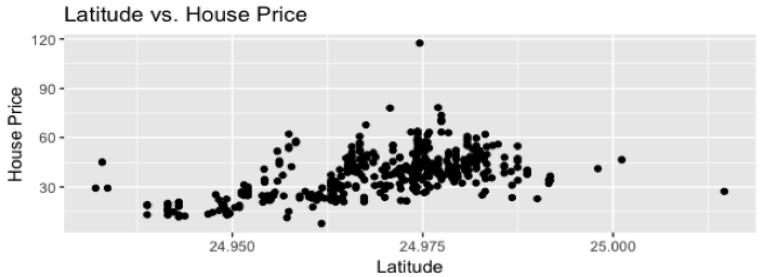
- Data originally collected in 2018 by Yeh and Hsu for their research in building real estate valuation models in Sindian District, New Taipei City, Taiwan.
- The raw data contained 415 observations and 8 variables
  - ID
  - Transition date
  - House age
  - Distance to the nearest metro statioN
  - Number of convenience stores within walking distance of house
  - Latitude
  - Longitude
  - House price of unit area

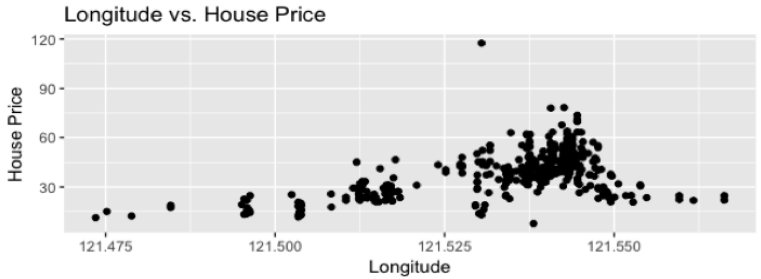
5. Yeh, I-Cheng, and Tzu-Kuang Hsu. "Building Real Estate Valuation Models with Comparative Approach through Case-Based Reasoning." *Applied Soft Computing*, vol. 65, 2018, pp. 260-271., doi:10.1016/j.asoc.2018.01.029.

DATA: EXPLORATORY ANALYSIS



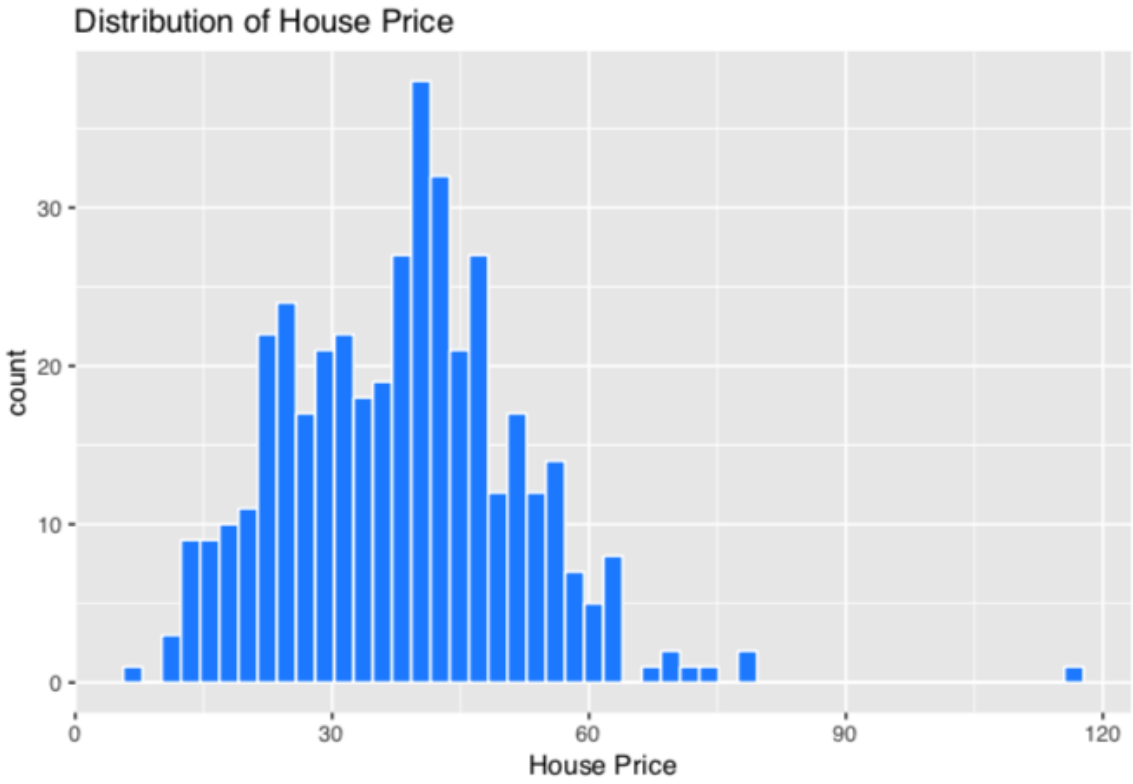






DATA: EXPLORATORY ANALYSIS

- Sales Price Distribution
  - Largely normally distributed with a few outliers.
  - No transformation necessary.



DATA: FEATURE SELECTION

- Forward, Backward, & Stepwise
- All methods selected had identical feature sets.
- Feature Correlations:
  - Mostly moderate correlations between predictors and house price.
  - Potential for multicollinearity.

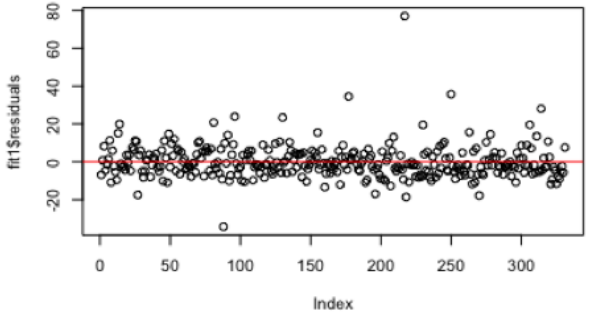
	row	col	corr
8	dist.nearest.mrt	longitude	-0.80631677
12	dist.nearest.mrt	house.price	-0.67361286
3	dist.nearest.mrt	num.convenience.stores	-0.60251914
5	dist.nearest.mrt	latitude	-0.59106657
11	house.age	house.price	-0.21056705
7	house.age	longitude	-0.04852005
1	house.age	dist.nearest.mrt	0.02562205
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

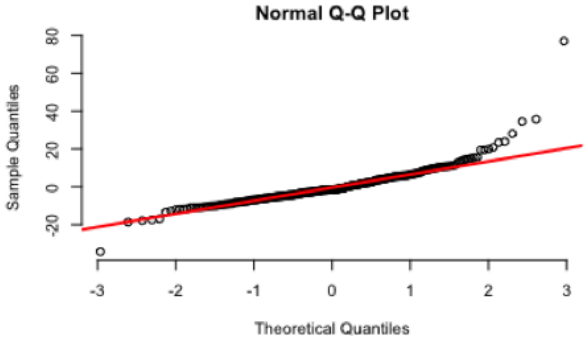
DATA: MULTILINEAR & POLYNOMIAL REGRESSION

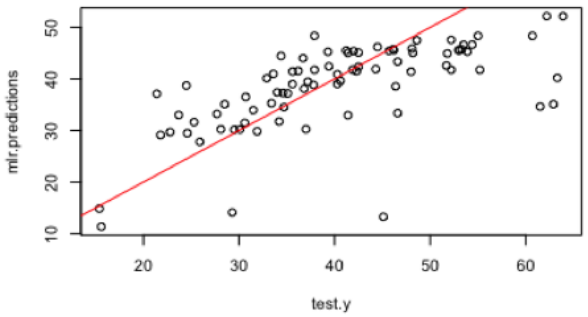
```
Family: gaussian
Link function: identity

Formula:
house.price ~ dist.nearest.mrt + num.convenience.stores + house.age +
    latitude

Parametric coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -5.942e+03  1.276e+03  -4.655 4.72e-06 ***
dist.nearest.mrt      -4.108e-03  5.385e-04  -7.629 2.63e-13 ***
num.convenience.stores  1.226e+00  2.107e-01   5.818 1.42e-08 ***
house.age             -2.786e-01  4.441e-02  -6.274 1.12e-09 ***
latitude               2.397e+02  5.112e+01   4.688 4.05e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


R-sq.(adj) =   0.59   Deviance explained = 59.5%
GCV = 82.455  Scale est. = 81.209    n = 331
```

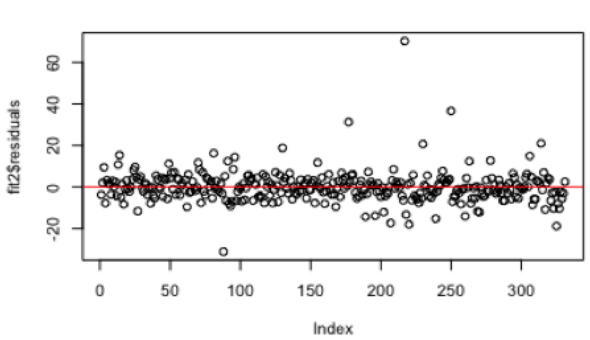
```
Family: gaussian
Link function: identity

Formula:
house.price ~ poly(dist.nearest.mrt, 3) + poly(num.convenience.stores,
    3) + poly(house.age, 3) + poly(latitude, 3)

Parametric coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      37.3900     0.4307  86.810  < 2e-16 ***
poly(dist.nearest.mrt, 3)1      -98.4244    14.7990  -6.651 1.27e-10 ***
poly(dist.nearest.mrt, 3)2       63.6456     9.1510   6.955 2.01e-11 ***
poly(dist.nearest.mrt, 3)3      -49.6626     9.4089  -5.278 2.42e-07 ***
poly(num.convenience.stores, 3)1 16.3939    11.4183   1.436  0.15205
poly(num.convenience.stores, 3)2 -3.9411     8.8205  -0.447  0.65532
poly(num.convenience.stores, 3)3  5.3925     9.0596   0.595  0.55212
poly(house.age, 3)1             -57.9937     7.9652  -7.281 2.62e-12 ***
poly(house.age, 3)2              32.9417     8.5833   3.838  0.00015 ***
poly(house.age, 3)3               9.6565     7.9150   1.220  0.22335
poly(latitude, 3)1               76.2536    12.8608   5.929 7.93e-09 ***
poly(latitude, 3)2              -10.8714    11.4684  -0.948  0.34388
poly(latitude, 3)3              -15.0388     8.0702  -1.864  0.06331 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


R-sq.(adj) =   0.69   Deviance explained = 70.1%
GCV = 63.914  Scale est. = 61.404    n = 331
```

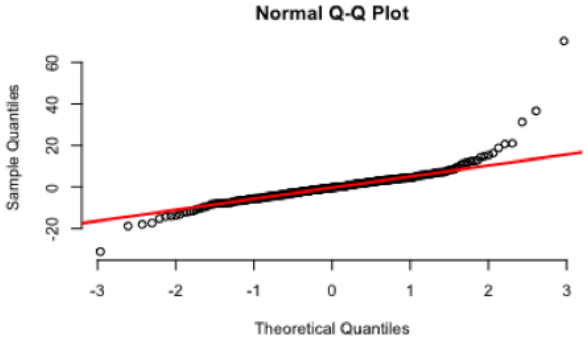
DATA: MULTILINEAR & POLYNOMIAL REGRESSION

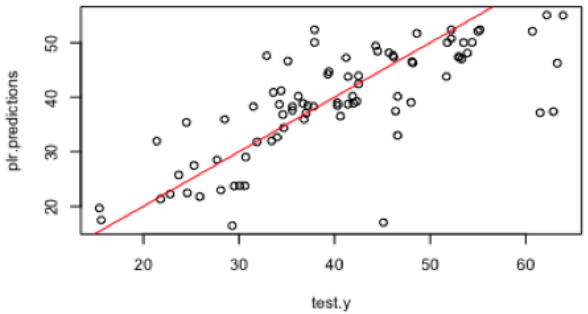












Top: Multiple Linear Regression Model; Bottom: Polynomial Regression Model

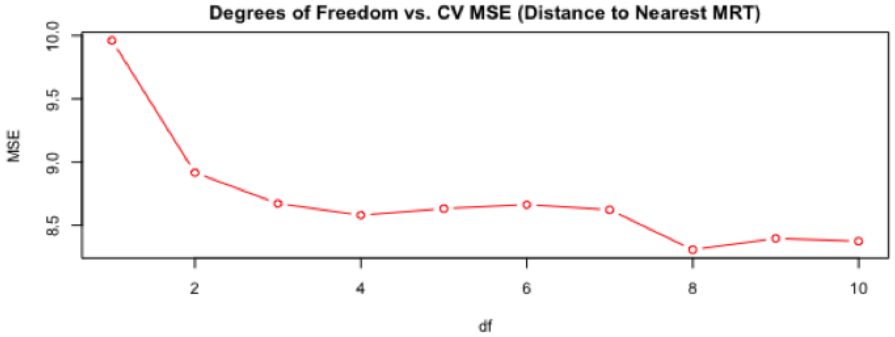
DATA: REGRESSION SPLINES

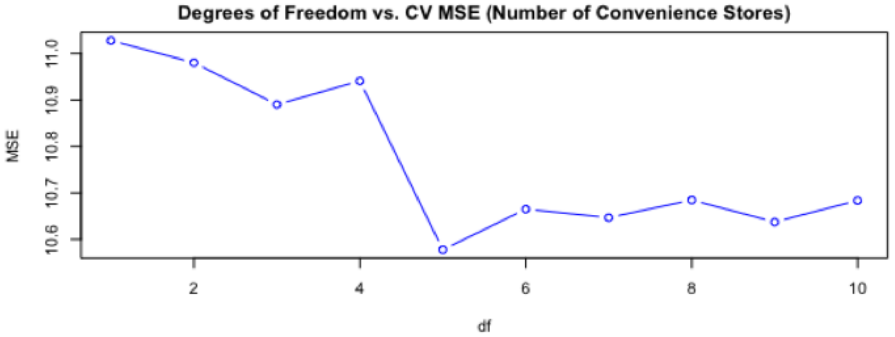
- Instead of fitting a single linear function or polynomial over the entire range of X, regression splines fit separate polynomials over different regions of X.
- Regions of the X domain are separated by *'knots'*.

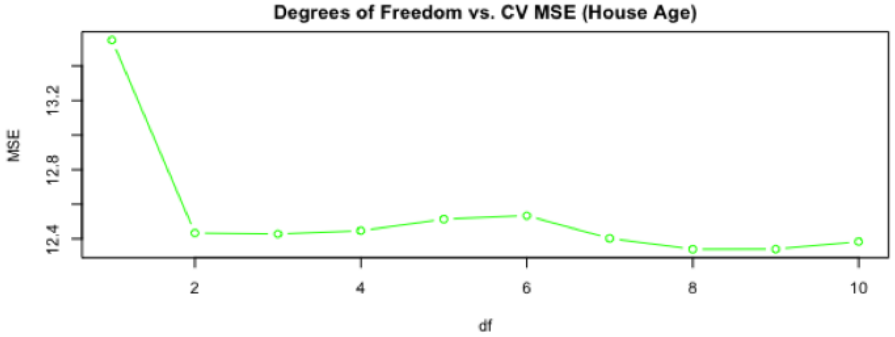
$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \epsilon_i & \text{if } x_i < c; \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \epsilon_i & \text{if } x_i \geq c. \end{cases}$$

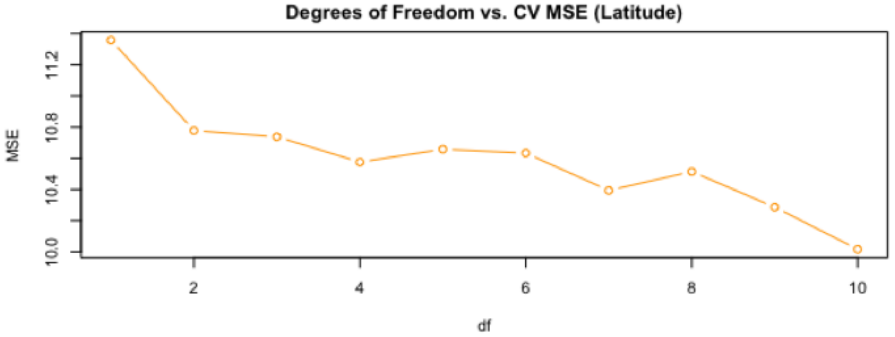
DATA: REGRESSION SPLINES

- Determine best degrees of freedom, function selects best knots.
- 10-Fold CV

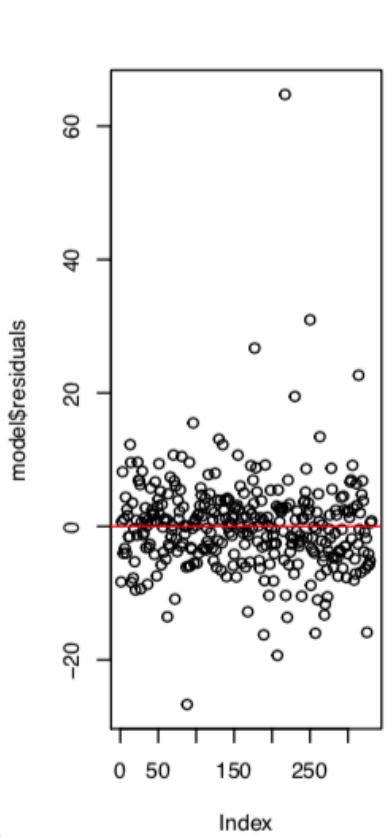


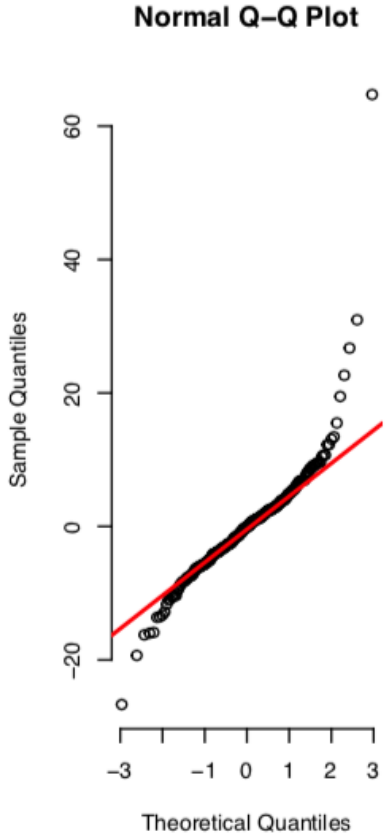


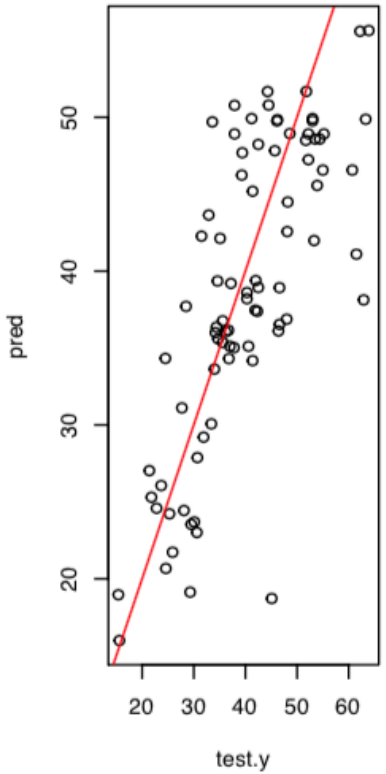




DATA: REGRESSION SPLINES







RESULTS: MODEL BUILDING (GAM)

- Using the best four predictors, *yi* is updated in equation 3:

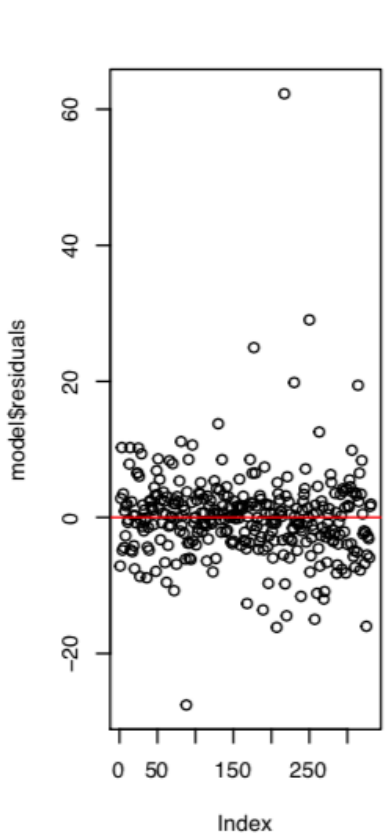
$$y_i = B_o + f_i(x_{i\ 1}) + f_2(x_{i\ 2}) + \ldots + f_p(x_{i\ p}) + E_i$$

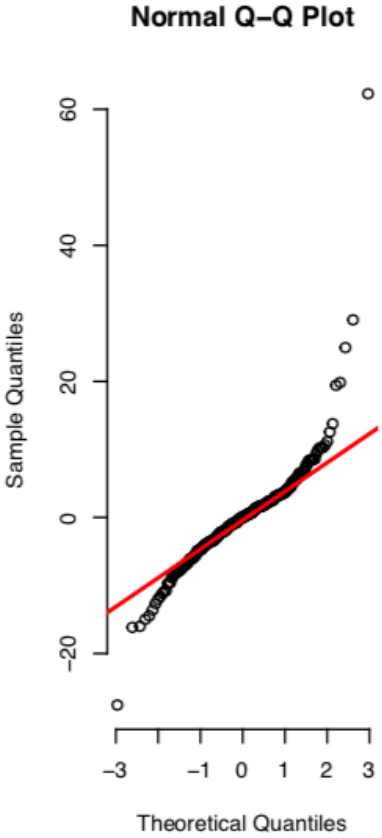
$$price = -7.822e03 - 4.11e03(nearestmetro) + 3.684e01(conviencestore) - 2.342e01 + 3.15(houseage) + 3.15e02(latitude)$$

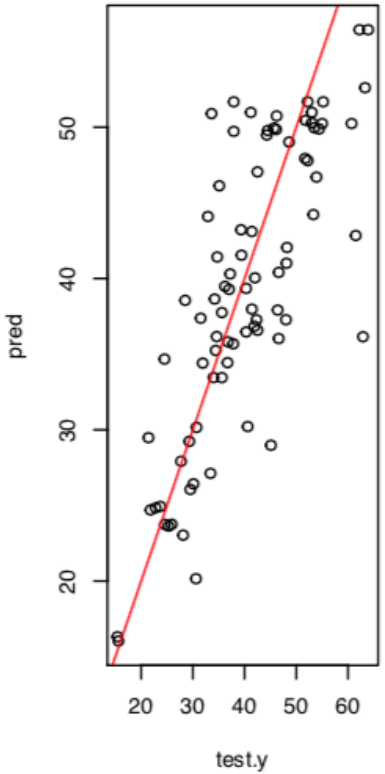
and gam() is used to find df_λ in order to tune the parameter λ

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

RESULTS: MODEL BUILDING (GAM)



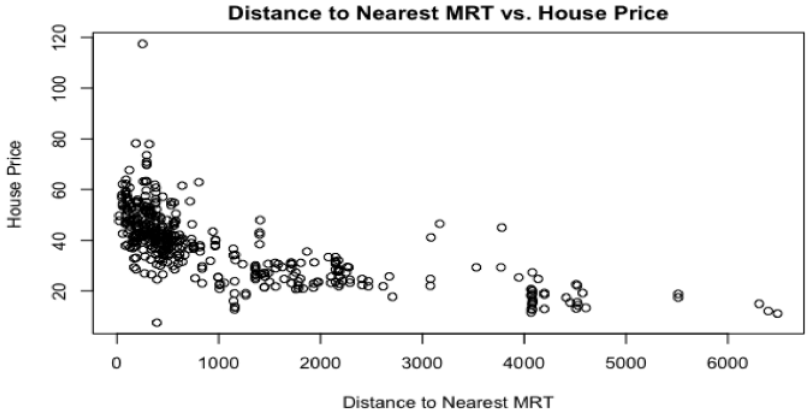


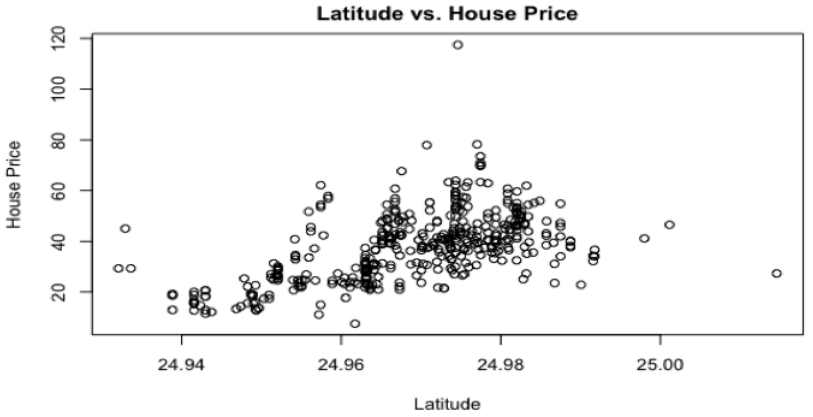


RESULTS: MODEL BUILDING (SMOOTH SPLINE)

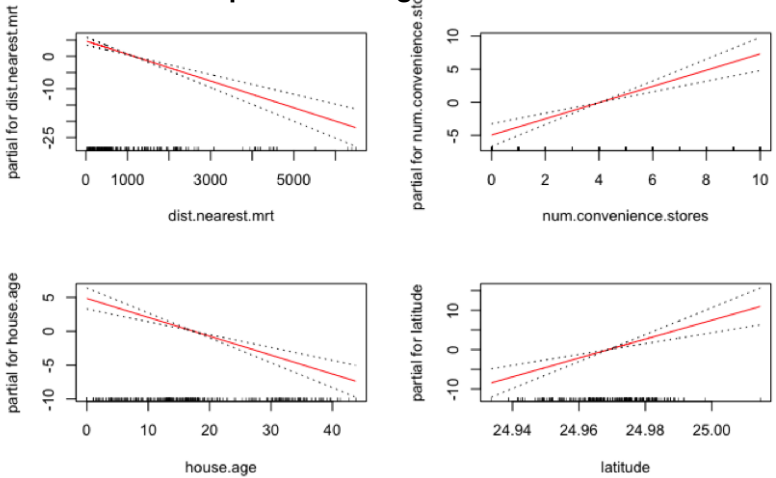
- smooth.spline( ) function uses LOOCV to automatically select the ‘best’ degrees of freedom to fit the data.
- Specifying degrees of freedom, function automatically selects value of smoothing parameter that leads to the df.
- We incorporate our findings from GAM and test our model performance

RESULTS: MODEL BUILDING (SMOOTH SPLINE)

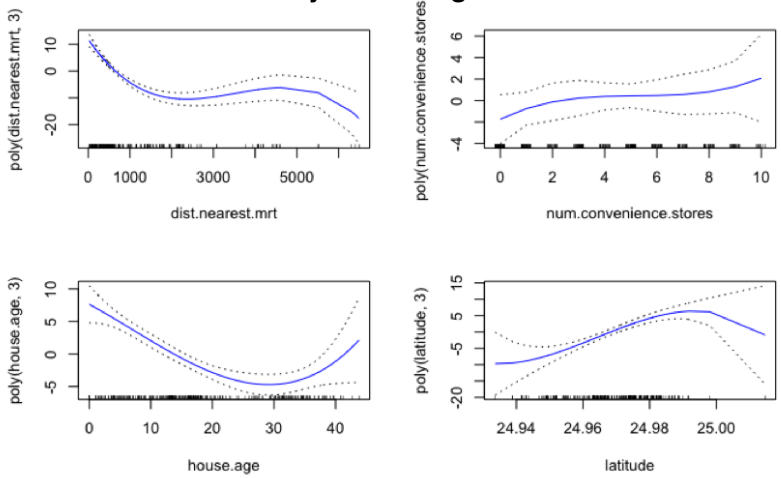




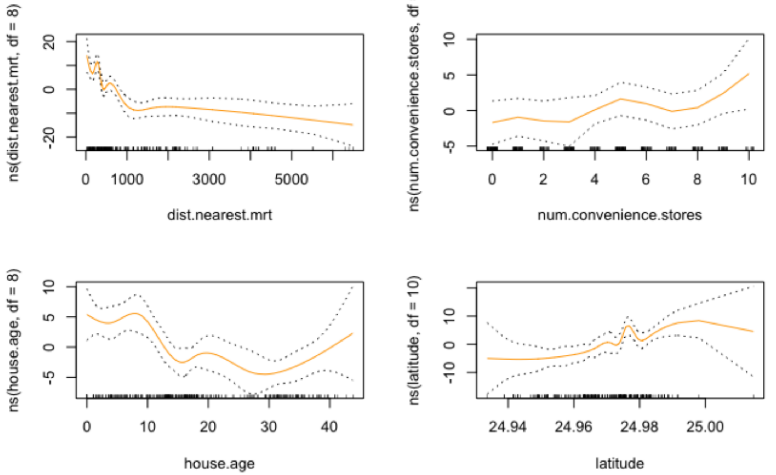
Multiple Linear Regression



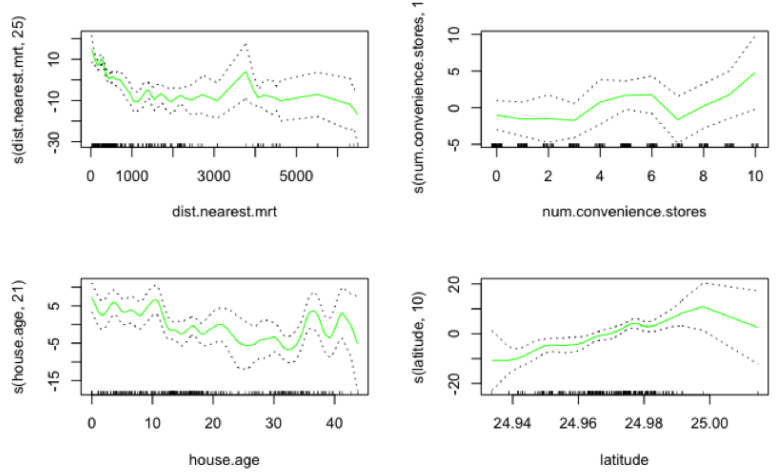
Polynomial Regression



Regression Spline



Smoothing Spline



RESULTS: MODEL PERFORMANCE

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

CONCLUSIONS

- Best predictors: distance to the nearest metro station, number of convenience stores within walking distance of house, house age, and latitude.
- Smoothing Spline exhibited the greatest performance out of all four models; lowest test MSE and RMSE.
- Further research including random forests and boosting.

REFERNCES

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. 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.
3. Hastie, T.j., and R.j. Tibshirani. “Generalized Additive Models.” *Generalized Additive Models*, 2017, pp. 136–173., doi:10.1201/9780203753781-6.
4. James, Gareth, et al. An Introduction to Statistical Learning: with Applications in R. Springer, 2017.
5. 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.
6. Image 1: “HALF DAY TAIPEI CITY TOUR.” *Shoretrips*, www.shoretrips.com/excursion/country-taiwan (china)-kee-079791/keelung-taipei-taiwan/half-day-taipei-city-tour.

SPECIAL THANKS

- Professor Iordan Slavov for his guidance
- Residents of New Taipei City, Taiwan. Without you, we'd have no data