NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Multivariate Adaptive Regression Splines (MARS)

IN-CLASS SESSION 6

Intended Learning Outcomes

- Identify aspects of business problems that cause standard analytics models to become useless or less effective.
- Apply advanced techniques to overcome or mitigate the weaknesses of standard analytics models.
- Evaluate performance of the advanced predictive techniques.
- Explain the workings and results of the advanced predictive techniques in the context of the business problem to client/employer.
- Propose business solutions/recommendations based on the advanced predictive techniques.

Quiz

Ungraded. Check your understanding of this Session Content.
Use your real name (not nickname) in the quiz.

Activity 1

Single Variate MARS

Pre-class Exercise

1. Run flatsales-mars.R
2. What is the MARS model coefficients and RMSE if 10-fold CV is used to prune instead of GRsq? Which ncross level is more stable?
   - Seed = 2 vs 2020
   - pmethod="cv“
   - nfold = 10
   - ncross = 1 vs 5

[Continue with Q3 and 4 in next slide]

Activity 1

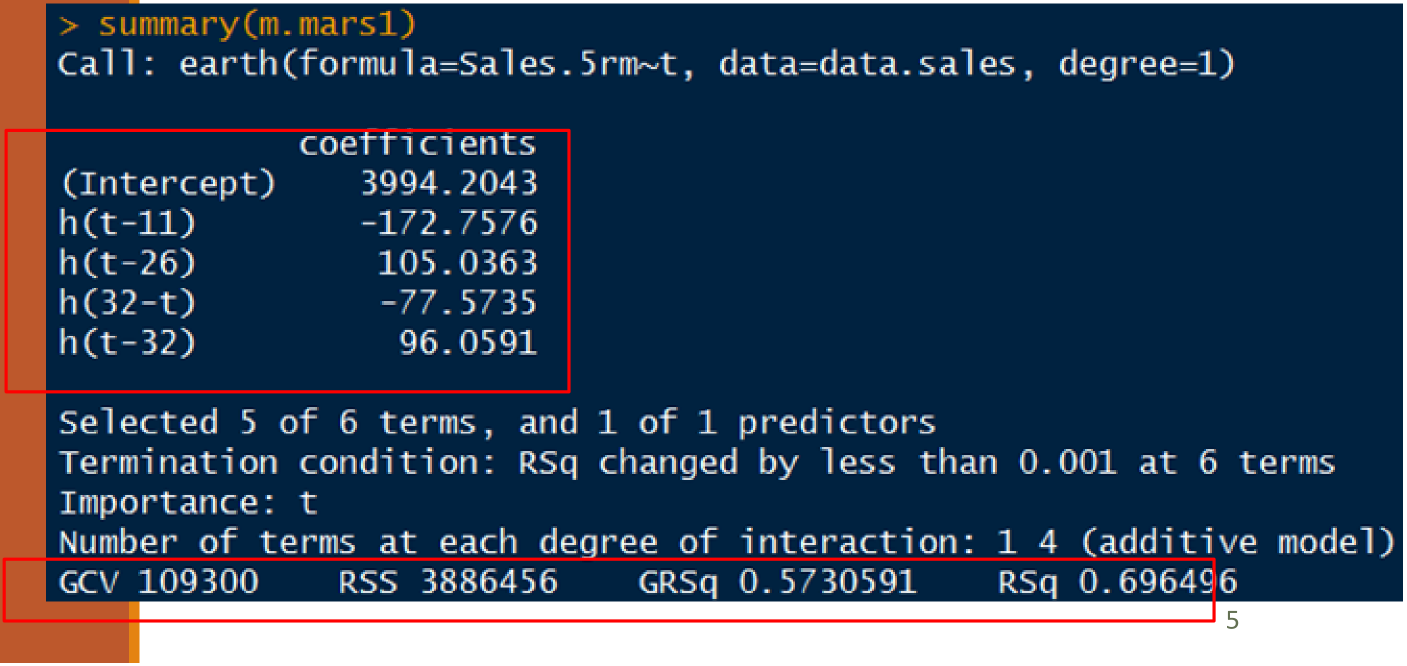
Single Variate MARS

Pre-class Exercise

3. Create a copy of the sales dataset as an Excel workbook. Using Excel, show that the linear regression model with the selected 4 hinge functions has the same model coefficients as R output.

4. Advanced option: Compute GCV, GCV.null and GRsq in excel.

See Instructor answers in flatsales-mars2.R & 5 room flat resale applications solution.xlsx.



```
> summary(m.mars1)
Call: earth(formula=Sales.5rm~t, data=data.sales, degree=1)

             coefficients
(Intercept)     3994.2043
h(t-11)         -172.7576
h(t-26)          105.0363
h(32-t)          -77.5735
h(t-32)           96.0591

Selected 5 of 6 terms, and 1 of 1 predictors
Termination condition: RSq changed by less than 0.001 at 6 terms
Importance: t
Number of terms at each degree of interaction: 1 4 (additive model)
GCV 109300    RSS 3886456    GRSq 0.5730591    RSq 0.696496
```

Answers to Class Activity 1

Activity 2

MARS Application

Pre-class Exercise

In Chang et. al. (2012) Analysis of freeway accident frequency using multivariate adaptive regression splines p. 827 states *"…(5) if the degree of horizontal curve is greater than 8.20, the accident frequency will increase by 0.992 for additional increase in degree of horizontal curve (indicated by BF7)."*

Does this mean that if horizontal curve degree = 9.2, then accident frequency will increase by 0.992?

PDF article provided in the NTULearn content folder.

Answer for Activity 2

MARS with Multiple Xs

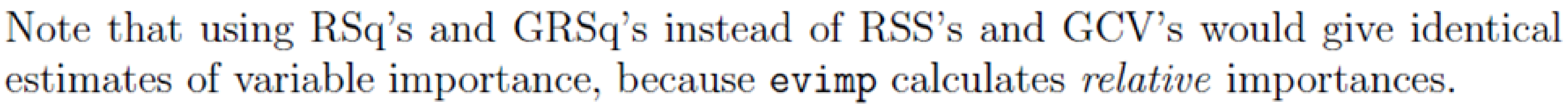
Multiple Xs

- Knots (i.e. hinges) only created for continuous X
- Categorical X should be “factor” type and treated the same as in linear regression
    - Dummy variables auto-created.

Which Xs are impt? Variable Importance via evimp()

Source: earth notes documentation p.50

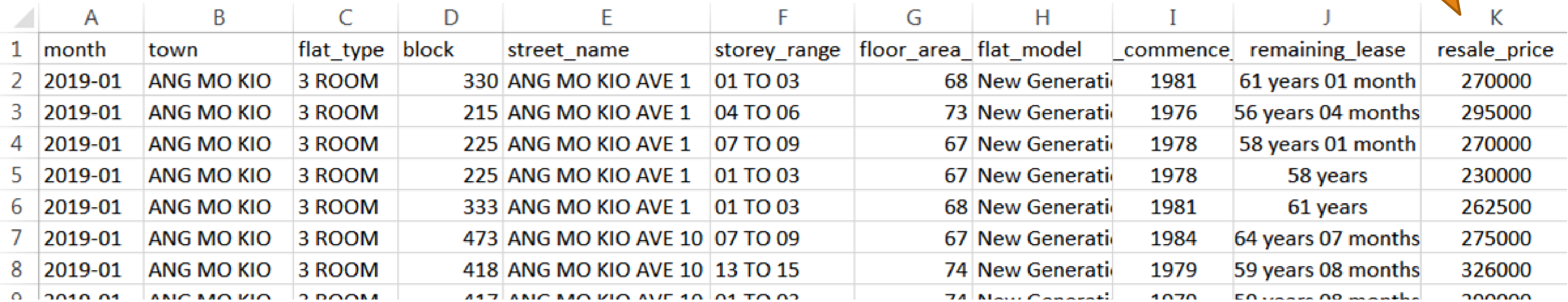
- 3 Criteria:
  - nsubsets criterion:
    - counts the number of model subsets that include the variable.
    - Variables that are included in more subsets are considered more important.
  - RSS:
    - Variables which cause larger net decreases in the RSS are considered more important.
  - GCV:
    - Variables which cause larger net decreases in the GCV are considered more important.



Note that using RSq's and GRSq's instead of RSS's and GCV's would give identical estimates of variable importance, because `evimp` calculates *relative* importances.

Resale Flat Prices
Dataset: resale-flat-prices-2019.csv

Y variable



	A	B	C	D	E	F	G	H	I	J	K
1	month	town	flat_type	block	street_name	storey_range	floor_area_	flat_model	_commence_	remaining_lease	resale_price
2	2019-01	ANG MO KIO	3 ROOM	330	ANG MO KIO AVE 1	01 TO 03	68	New Generati	1981	61 years 01 month	270000
3	2019-01	ANG MO KIO	3 ROOM	215	ANG MO KIO AVE 1	04 TO 06	73	New Generati	1976	56 years 04 months	295000
4	2019-01	ANG MO KIO	3 ROOM	225	ANG MO KIO AVE 1	07 TO 09	67	New Generati	1978	58 years 01 month	270000
5	2019-01	ANG MO KIO	3 ROOM	225	ANG MO KIO AVE 1	01 TO 03	67	New Generati	1978	58 years	230000
6	2019-01	ANG MO KIO	3 ROOM	333	ANG MO KIO AVE 1	01 TO 03	68	New Generati	1981	61 years	262500
7	2019-01	ANG MO KIO	3 ROOM	473	ANG MO KIO AVE 10	07 TO 09	67	New Generati	1984	64 years 07 months	275000
8	2019-01	ANG MO KIO	3 ROOM	418	ANG MO KIO AVE 10	13 TO 15	74	New Generati	1979	59 years 08 months	326000
9	2019-01	ANG MO KIO	3 ROOM	417	ANG MO KIO AVE 10	01 TO 03	74	New Generati	1979	59 years 08 months	[illegible]

- 4 Main Xs to apply in MARS:
  - Floor Area [continuous]
  - Remaining lease in Years (Max 99 for new flat) [continuous]
  - Town [categorical]
  - Storey Range [categorical]

Class Activity 3

Multi-Variate MARS

Q1 & Q2 done in Session 2.

Est. Duration: 30 mins

1. Create a new continuous X variable remaining lease in years.
2. Change the Baseline Reference level for Town to Yishun.
3. Use only the 4 input X variables used in S2. (floor_area_sqm, remaining_lease_years, town, and storey_range).
4. Develop 2 MARS models and compare their RMSE and model coefficients.
   - degree = 1
   - degree = 2
5. Using the 2 MARS models, predict the resale price of a flat in Clementi, 100 square metres, 19-21 storey & 80 yrs lease remaining. Verify your calculations using hinge functions in Excel.
6. Which X variables are relatively more impt in MARS degree 2 model? Hint: evimp()

Answers to Class Activity 3

- flatprice-mars solution.R
- flatprice-mars predictions.xlsx

Reminder

Please complete the Pre-Class Learning Activities before next class.

Reflection on your Learning

Go

NTULearn Class Site > Journal

Post

Read the instructions and post entry on this week's learning.

- Reply on the 3 questions as stated in the Journal Instructions.