Multivariate Adaptive Regression Splines (MARS)

BC2407 ANALYTICS II SEMINAR 6

NEUMANN CHEW C. H.

Recall the Linear Regression Model

$$y = \underbrace{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m}_{\hat{y}} + e$$

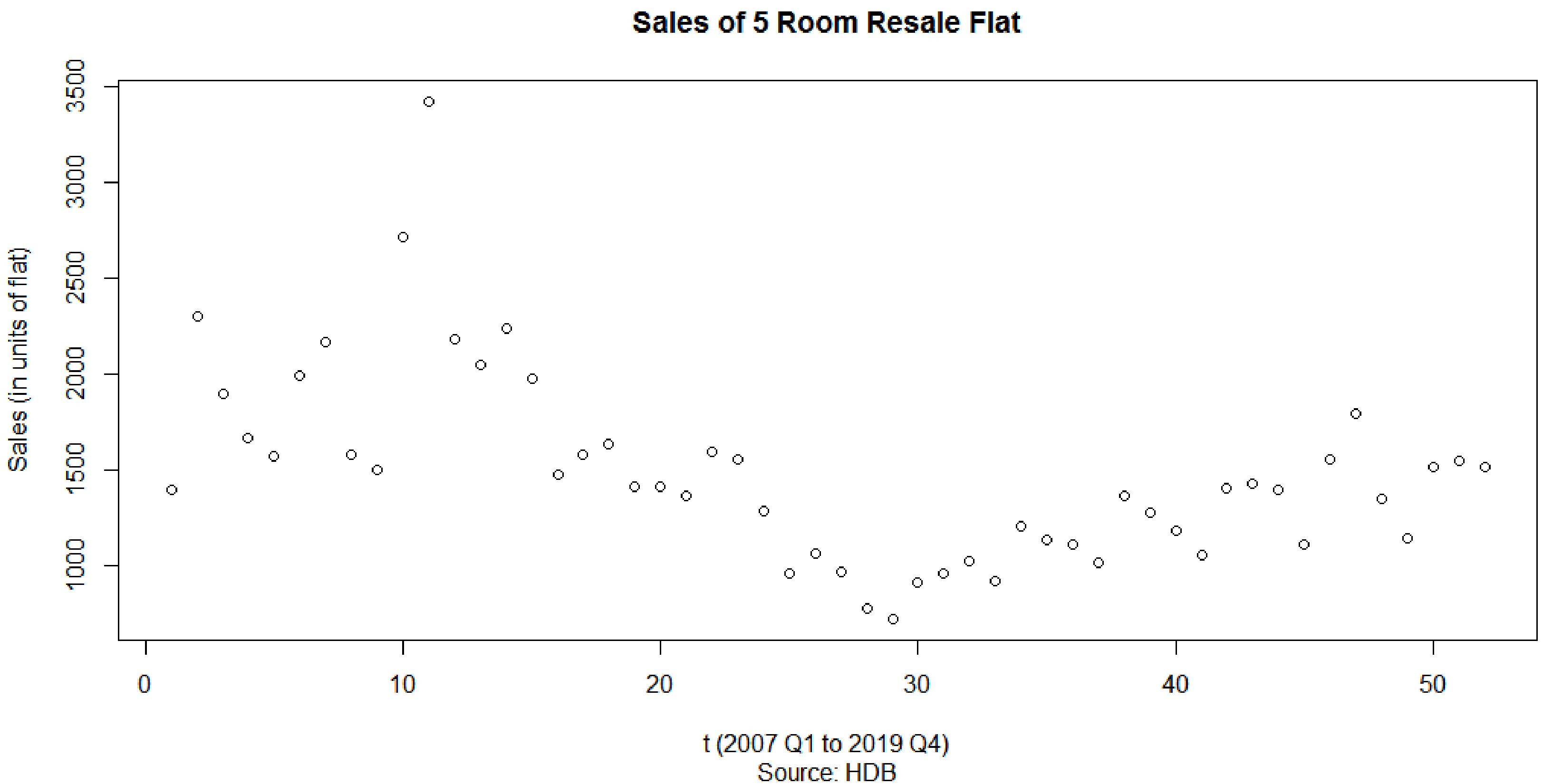
Straight Line Equation

Assumes LINEAR trend relating Y to all the Xs.
What if non-linear?
Fit Quadratic? Cubic?

e ~ N(0, σ)

Errors (aka Residuals) follow a Normal Distribution with mean 0 and constant standard deviation.

HDB 5-room Flat Resale data
Dataset: 5 room flat resale applications.csv



- Linear Trend? Quadratic? Cubic?

Fitting Linear, Quadratic and Cubic Trends in R

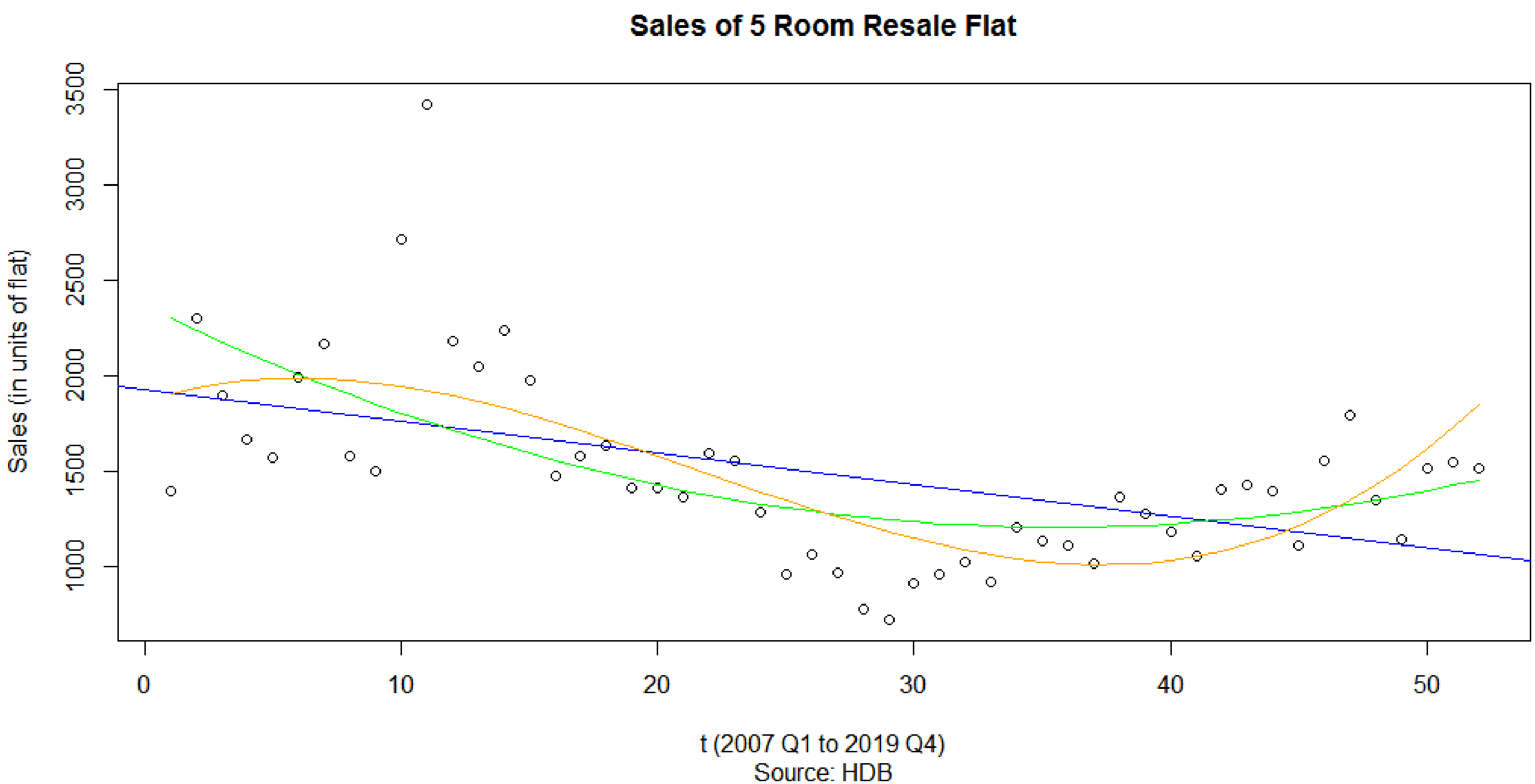
```
m.sales.lin1 <- lm(Sales.5rm ~ t, data = data.sales)

m.sales.lin2 <- lm(Sales.5rm ~ t + I(t^2), data = data.sales)

m.sales.lin3 <- lm(Sales.5rm ~ t + I(t^2) + I(t^3), data = data.sales)
```

Note: Necessary to use I()

Fitted Linear, Quadratic and Cubic Trends



- What is the problem with using Linear/Quadratic/Cubic reg?

- Ans: The trend applies globally throughout the data.

MARS Theory in ESL vs earth() Implementation

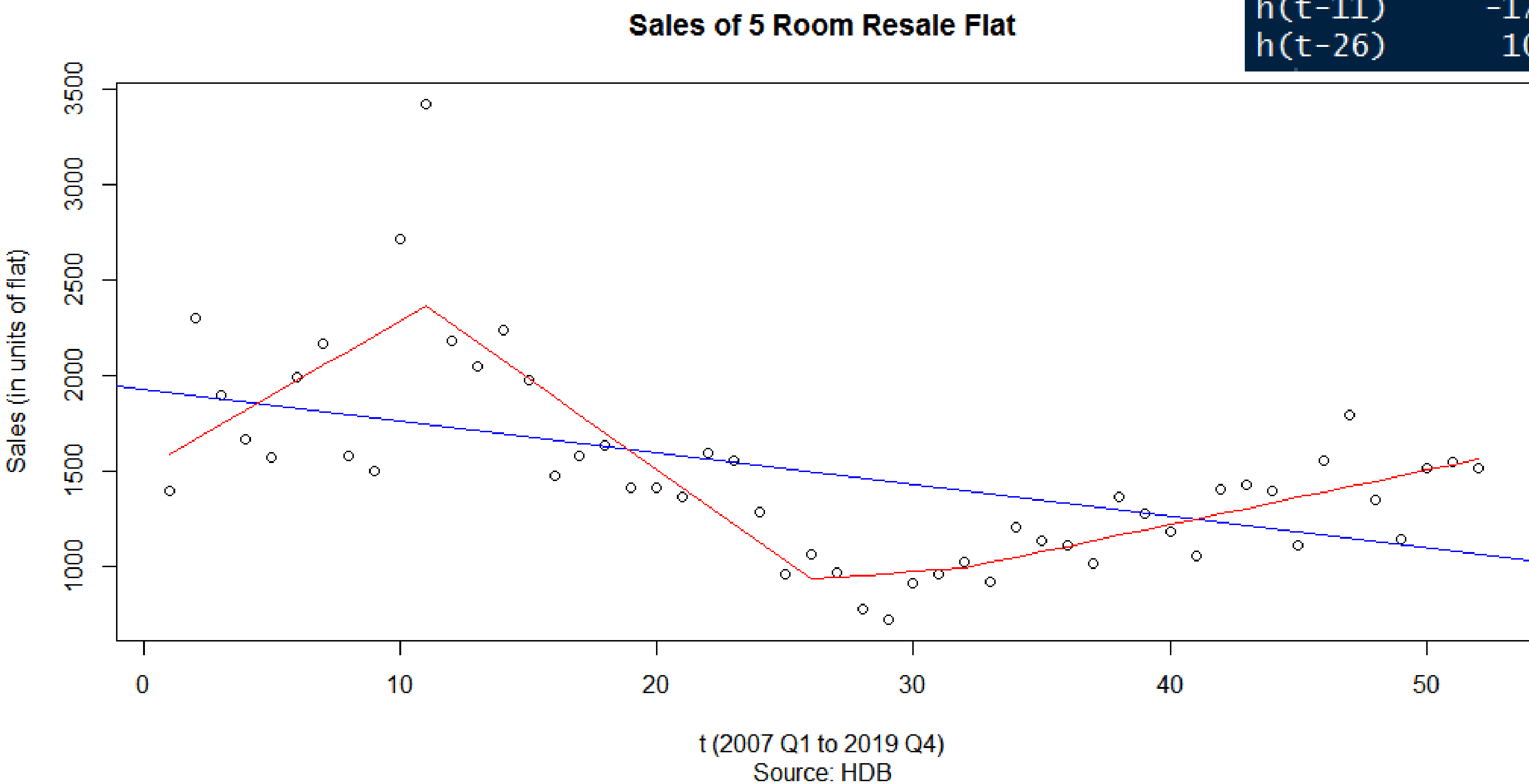
- MARS theory summarized in the textbook Elements of Statistical Learning 2nd Edition [ESL] Section 9.4 pp.321 – 329.

- Link to download free PDF textbook ESL given in Main Site Announcement.
  - https://web.stanford.edu/~hastie/ElemStatLearn/

- Different software (R, Python, SAS, etc) have different implementation of MARS and thus results might differ.

Linear vs MARS
Rscript: flatsales-mars1.R

```
> m.mars1$coefficients
                Sales.5rm
(Intercept) 3994.20425
h(t-32)       96.05913
h(32-t)      -77.57353
h(t-11)     -172.75760
h(t-26)      105.03626
```



- MARS fit local hinge functions adaptively i.e. not global trend.
  - What is a hinge function?
  - The knots (aka cuts) t = 11, 26, 32 are found automatically in MARS. How?

Hinge functions in MARS

```
> m.mars1$coefficients
              Sales.5rm
(Intercept) 3994.20425
h(t-32)       96.05913
h(32-t)      -77.57353
h(t-11)     -172.75760
h(t-26)      105.03626
```

- The Mars model is a weighted combination of hinge functions:
- $\hat{y} = 3994.2 - 172.8\mathrm{h}(\mathrm{t}-11) + 105\mathrm{h}(\mathrm{t}-26) + 96.1\mathrm{h}(\mathrm{t}-32) - 77.6\mathrm{h}(32-\mathrm{t})$
- Hinge function: h(s) ≡ max(s, 0).
- Q: What is the MARS model predicted value of y if
  - t = 10
  - t = 20
  - t = 40

Answers

- At t = 10: $\hat{y} = 3994.2 - 172.8(0) + 105(0) + 96.1(0) - 77.6(22) = 2287$

- At t = 20: $\hat{y} = 3994.2 - 172.8(9) + 105(0) + 96.1(0) - 77.6(12) = 1507.8$

- At t = 40: $\hat{y} = 3994.2 - 172.8(29) + 105(14) + 96.1(8) - 77.6(0) = 1221.8$

RMSE Results (on trainset) in Ascending Order

Model	RMSE
MARS degree 1	273
Linear Reg degree 3	350
Linear Reg degree 2	388
Linear Reg degree 1	430

- earth() function default: degree = 1
- degree = k will consider up to kth power in the growing phrase but insignificant terms will be pruned away. Thus final result will only show significant terms and may or may not show any kth power term. This is a valuable feature in MARS.
- No interaction effects are found in MARS degree 2 as this dataset contains one X only.

MARS degree 1 Results

```
# MARS Degree = 1
m.mars1 <- earth(Sales.5rm ~ t, degree = 1, data=data.sales)
```

```
> 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
```

- Growing Phrase (aka Forward Pass): Selected 6 terms (incl. intercept).
- Pruning Phrase (aka Backward Pass): Removed one term. Hence 5 terms left.
- Use Trace = 3 to view the Growing and Pruning sequence.

R documentation: ?earth at console

```
## S3 method for class 'formula'
earth(formula = stop("no 'formula' argument"), data = NULL,
    weights = NULL, wp = NULL, subset = NULL,
    na.action = na.fail,
    pmethod = c("backward", "none", "exhaustive", "forward", "seqrep", "cv"),
    keepxy = FALSE, trace = 0, glm = NULL, degree = 1, nprune = NULL,
    nfold=0, ncross=1, stratify=TRUE,
    varmod.method = "none", varmod.exponent = 1,
    varmod.conv = 1, varmod.clamp = .1, varmod.minspan = -3,
    Scale.y = NULL, ...)
```

`degree`	Maximum degree of interaction (Friedman's *mi*). Default is 1, meaning build an additive model (i.e., no interaction terms).
`pmethod`	Pruning method. One of: `backward none exhaustive forward seqrep cv`. Default is `"backward"`. Specify `pmethod="cv"` to use cross-validation to select the number of terms. This selects the number of terms that gives the maximum mean out-of-fold RSq on the fold models. Requires the `nfold` argument. Use `"none"` to retain all the terms created by the forward pass.
`trace`	Trace `earth`'s execution. Values: 0 (default) no tracing .3 variance model (the `varmod.method` arg) .5 cross validation (the `nfold` arg) 1 overview 2 forward pass 3 pruning 4 model mats summary, pruning details 5 full model mats, internal details of operation

Trace = 3 in earth() to view growing and pruning sequence

```
> # trace = 3 to view the MARS growing and pruning sequence
> earth(Sales.5rm ~ t, degree = 1, trace = 3, data=data.sales)
x[52,1] with colname t, and values 1, 2, 3, 4, 5, 6, 7, 8, 9, 10...
y[52,1] with colname Sales.5rm, and values 1402, 2305, 1901, 1667, 1574,...
Forward pass: minspan 3 endspan 7   x[52,1] 416 Bytes   bx[52,21] 8.53 kB

         GRSq    RSq      DeltaRSq Pred     PredName         Cut  Terms   Par Deg
1      0.0000 0.0000                      (Intercept)
2      0.3979 0.4886        0.4886    1            t          32  2   3         1
4      0.5455 0.6461        0.1575    1            t          11  4             1
6      0.5731 0.6965       0.05038    1            t          26  5             1
8      0.5340 0.6988      0.002353    1            t          38  6             1
10     0.4857 0.6992     0.0003907    1            t          41  7             1 reject (small DeltaRSq)

RSq changed by less than 0.001 at 9 terms, 6 terms used (DeltaRSq 0.00039)
After forward pass GRSq 0.486 RSq 0.699
Forward pass complete: 9 terms, 6 terms used

Subset size        GRSq     RSq  DeltaGRSq nPreds
          1      0.0000  0.0000     0.0000      0
          2      0.3536  0.4033     0.3536      1
          3      0.5343  0.6045     0.1807      1
          4      0.5572  0.6553     0.0229      1
chosen    5      0.5731  0.6965     0.0159      1
          6      0.5340  0.6988    -0.0390      1
```

```
Prune backward penalty 2 nprune null: selected 5 of 6 terms, and 1 of 1 preds
After pruning pass GRSq 0.573 RSq 0.696
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
```

FAQ: Why trace reveal only one pair added in forward pass?

```
        GRSq    RSq      DeltaRSq Pred      PredName      Cut  Terms    Par Deg
1     0.0000 0.0000                      (Intercept)
2     0.3979 0.4886        0.4886    1             t       32  2   3          1
4     0.5455 0.6461        0.1575    1             t       11  4              1
6     0.5731 0.6965       0.05038    1             t       26  5              1
8     0.5340 0.6988      0.002353    1             t       38  6              1
10    0.4857 0.6992     0.0003907    1             t       41  7              1 reject (small DeltaRSq)
```

Ans: This is another different implementation compared to theory. In earth-notes documentation p.6:

“ the forward pass discards one side of a term pair if it adds nothing to the model — but the forward pass counts terms as if they were actually created in pairs,”

i.e. after the first pair is added (checking that each term in the pair contributes significantly to reducing RSS), the future pairs are actually added solo as the other side was checked and found to contribute “nothing” to reducing RSS. For software, we need to set a threshold to define “nothing” e.g. less than 0.0000001 for RSS or < 0.001 for Rsq

Overview of MARS Model Development

Initialize MARS → Grow MARS → Prune MARS

- Similar strategy to CART.
- Grow to add the “best” hinge (in reflected pairs) at each step.
- Prune to remove weakest hinge terms when complexity penalty is included to reduce risk of overfitting, one term at a step.
- Note: Implementation differences with different software.

MARS Implementation in R earth() vs Python py-earth

Initialize MARS

- Choose degree. Default = 1.
- Choose pruning method (GRsq , GCV or k-fold CV).
- R Default : GRsq.
- Python Default: GCV.

Grow MARS

- Add the “best” hinge functions in reflected pairs, one pair per step.
- Best determined by largest increase in Rsq (R) or largest decrease in RSS (Python) to current model.
- Update current model coefs by lm fit, after each addition.
- Stop when Rsq increase < 0.001 or no new term increases Rsq.

Prune MARS

- Remove weakest term, one at each step.
- The weakest term defined as the term whose removal result in smallest increase in RSS.
- Compute complexity adjusted model error GRsq, GCV or mean OOF CV error.
- Use GRSq, GCV, or k-fold CV to estimate optimal number of terms in final model.

“The GRSq normalizes the GCV in the same way that the Rsq normalizes the RSS”.
– R earth notes documentation p.29

If Degree > 1

- In MARS growth phrase, candidate hinge functions may be multiplied to existing hinge functions in the current model.
- The maximum number of multiplications is the Degree.
- If Degree = 1, no hinge functions are multiplied. i.e. additive model.

GCV and GRsq
Source: earth notes documentation p.57.

```
GCV = RSS / (n * (1 - nparams / n)^2))
```

Effective Number of Parameters in MARS

```
GRSq = 1 - GCV / GCV.null,
```

where `GCV.null` is the GCV of an intercept-only model.

- GCV actually adjust the RSS to account for model complexity i.e. complexity adjusted RSS. No Cross Validation is used.
- Both GCV and GRsq are opinionated estimation of model predictive performance on future unseen data, by adding a penalty for model complexity (number of hinges in MARS).
- GRSq is the normalized version of GCV, scaled to between 0 to 1 under normal conditions.
- A means for more direct-than-k-fold CV estimation of testset error.

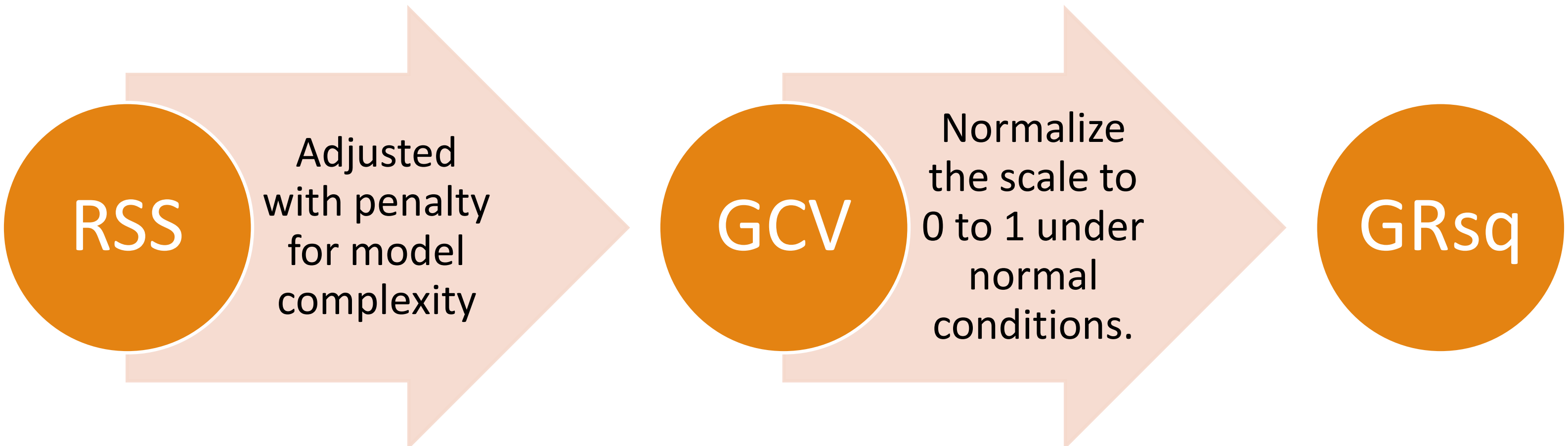
FAQ: How to compute nparams?

Effective number of parameters =

(num of terms) + (penalty) * ½ (num of terms − 1)

- num of terms – 1 is the number of hinge terms
- Take ½ to compensate for double counting of terms added in pairs during forward pass.

FAQ: What's the relationship between RSS, GCV, GRsq?



- RSS has no penalty for model complexity and thus will select the most complex model as RSS will be lowest.
- GCV and GRsq are ways to incorporate penalty for model complexity into RSS. Using either will select a model neither too big nor too small, and thus avoid overfitting (hopefully). This is a substitute for 10-fold CV.

Pruning alternative using 10-fold CV

- Default in earth():
  - nfold = 0
  - ncross = 1
- 10-fold CV: set nfold = 10
- Since CV is sensitive to data partition, try ncross > 3, to repeat 10-fold CV multiple times with another random partition of data.
- CVRsq is the mean Rsq from out-of-fold (OOF) data.
  - Note: CVRsq still use terms selected by GCV – earth notes p.38
- Use pmethod = "cv" to select the select optimal model using cross-validation.
  - Based on max mean OOF Rsq, not 1SE rule.

Class Activity 1

Single Variate MARS

Est. Duration: 30 mins

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

Class Activity 1

Single Variate MARS

Est. Duration: 30 mins

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.

Instructor answers in 5 room flat resale applications solution.xlsx will be posted by end of week.

```
> 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

- Q2: ncross = 5 is more stable than ncross = 1 as MARS Model and RMSE remains the same despite change in seed. [See RScript solution.] This shows CV is sensitive to data partition.
- Q3: See workings and solution in excel file solution.

	A	B	C	D	E
1	SUMMARY OUTPUT				
2					
3	*Regression Statistics*			MARS	
4	Multiple R	0.834563		Eff nparams	9
5	R Square	0.696496		GCV	109300.01
6	Adjusted R Square	0.670666		GCV.null	256007.34
7	Standard Error	287.5597		GRsq	0.5730591
8	Observations	52			
9					
10	ANOVA				
11		*df*	*SS*	*MS*	*F*
12	Regression	4	8918834	2229708.51	26.964488
13	Residual	47	3886456.2	82690.5572	
14	Total	51	12805290		
15					
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*
17	Intercept	3994.204	538.99859	7.41041691	1.956E-09
18	h(t-11)	-172.758	30.585599	-5.6483314	9.131E-07
19	h(t-26)	105.0363	37.603984	2.7932216	0.0075262
20	h(32-t)	-77.5735	21.552891	-3.59921689	0.0007656
21	h(t-32)	96.05913	38.009446	2.52724351	0.0149191

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	[illegible]	[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 2

Multi-Variate MARS

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 instead of default.
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?

Categorical Y

- Set as factor.
- use glm() function.
- Check earth-notes documentation.

Python Implementation of MARS

- py-earth
  - https://github.com/scikit-learn-contrib/py-earth
  - Learnt from R earth package.
  - Incl. missing data support (R earth has no support for NAs).
  - Some Differences in py-earth compared to R earth:
    - Penalty = 3 (In R earth, penalty = 2 for additive model, 3 otherwise.)
    - Use MSE (i.e. same as using RSS) in forward pass (R earth use Rsq)
    - Use GCV in pruning pass (R earth use GRsq)
    - Did not discard any term during forward pass (R earth may discard terms)
    - Others…

Python code for HDB 5-rm Sales data

```
from pyearth import Earth
import pandas as pd

df = pd.read_csv (r'D:/Dropbox/Datasets/HDB Resale Prices/subset2017-19/5 room flat resale applications.csv')
df['t'] = pd.Series(range(1,53))
```

```
mars = Earth(penalty=2)
mars.fit(X=df['t'], y=df['Sales 5rm'])
print(mars.summary())
## Result is not the same as R. Different implementation?
## Increase max_terms does not craete more terms as threshold reached.
```

```
Earth Model
--------------------------------------
Basis Function  Pruned  Coefficient
--------------------------------------
(Intercept)     No      392.019
h(x0-29)        Yes     None
h(29-x0)        No      113.126
h(x0-12)        Yes     None
h(12-x0)        No      -173.099
h(x0-16)        No      33.9763
h(16-x0)        Yes     None
x0              Yes     None
--------------------------------------
MSE: 75943.3561, GCV: 101407.8197, RSQ: 0.6916, GRSQ: 0.6039
```

Recall from BC2406: There is often more than one correct model. It's fine to use this MARS but be aware of the differences between Python vs R implementations.

Trace py-earth MARS reveals different criteria compared to R earth

```
print(mars.trace())
```

```
Forward Pass
--------------------------------------------------------------------
iter  parent  var  knot  mse            terms  gcv          rsq    grsq
--------------------------------------------------------------------
0     -       -    -     246255.581361  1      256007.340   0.000  0.000
1     0       0    28    126746.046454  3      155147.718   0.485  0.394
2     0       0    11    79422.110668   5      116147.857   0.677  0.546
3     0       0    15    71450.868851   7      127023.767   0.710  0.504
4     0       0    -1    71450.868851   8      141127.209   0.710  0.449
--------------------------------------------------------------------
Stopping Condition 2: Improvement below threshold

Pruning Pass
--------------------------------------------------
iter  bf  terms  mse        gcv         rsq    grsq
--------------------------------------------------
0     -   8      71450.87   141127.209  0.710  0.449
1     7   7      71450.87   127023.767  0.710  0.504
2     6   6      71450.87   114933.462  0.710  0.551
3     1   5      71450.87   104490.616  0.710  0.592
4     3   4      75943.36   101407.820  0.692  0.604
5     5   3      115080.20  140867.751  0.533  0.450
6     4   2      141750.54  159639.094  0.424  0.376
7     2   1      246255.58  256007.340  0.000  0.000
--------------------------------------------------
Selected iteration: 4
```

Note the inconsistency in term names and knots between the summary() results and trace().

Summary

- MARS:
  - Non-parametric
  - Can handle continuous Y and categorical Y too.
  - Construct hinge functions that adapts to the data via knots.
  - Automated selection of:
    - Significant X variables.
    - Knots.
    - Significant Interaction terms.
  - Rpackage: earth
    - Has some differences in implementation compared to MARS theory.
  - SAS Stat Procedure: adaptivereg
  - Python: py-earth