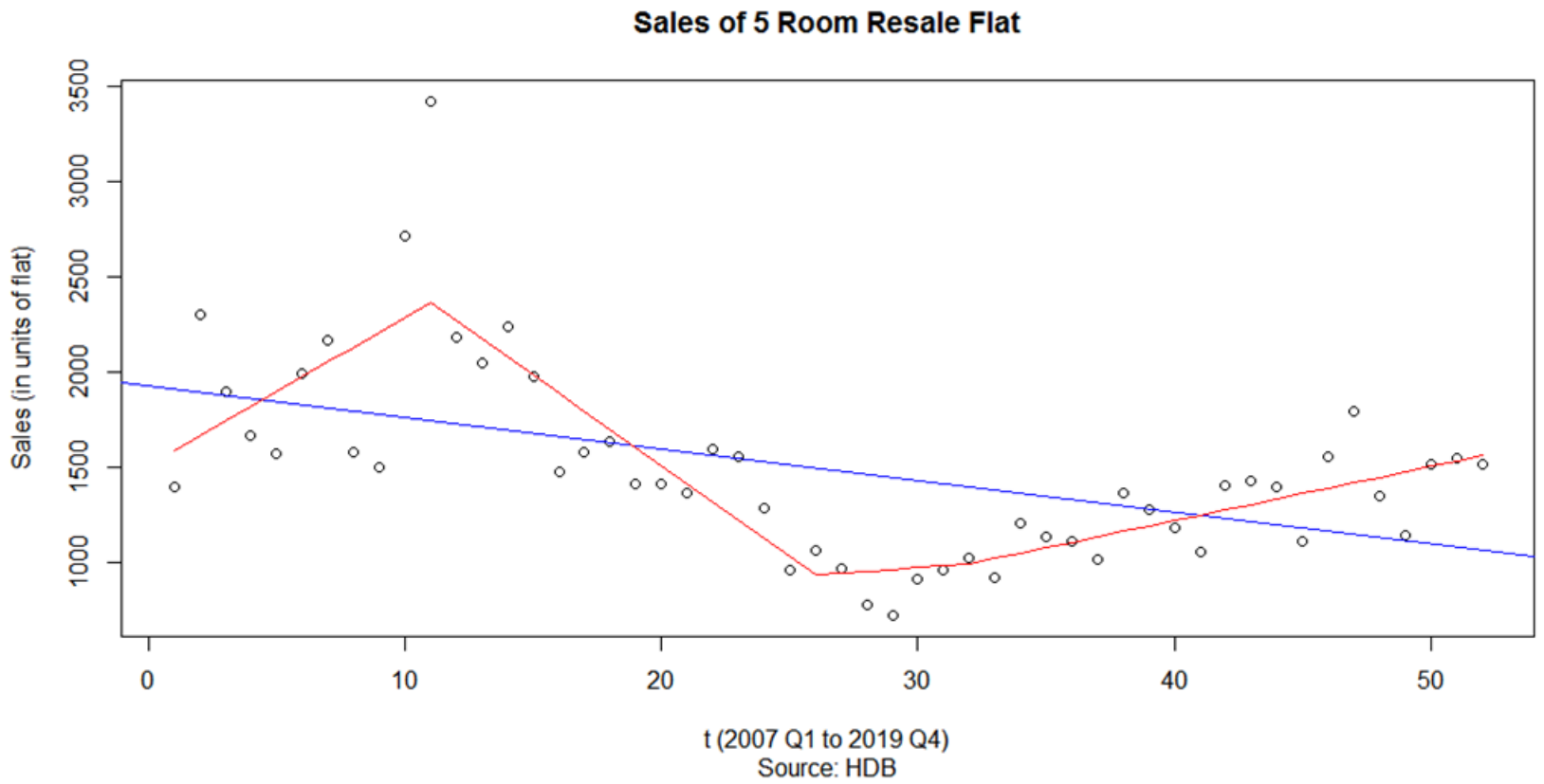
What is MARS

MARS (Part 2)

Lecture Video Slides

Linear vs MARS in Rscript: flatsales-mars.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 test and find best-fitting hinge functions

- Automatically from data
- The knots (aka cuts) t = 11, 26, 32 are found automatically in MARS. [How?]
- What is a hinge function?

MARS Model

```
> 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 linear combination of hinge functions:

- $\hat{y} = 3994.2 - 172.8 \times \mathrm{h}(\mathrm{t} - 11) + 105 \times \mathrm{h}(\mathrm{t} - 26) + 96.1 \times \mathrm{h}(\mathrm{t} - 32) - 77.6 \times \mathrm{h}(32 - \mathrm{t})$
- Hinge function: **h(s) ≡ max(s, 0), s = x – constant or constant – x.**
- Q: What is the MARS model predicted value of y if
  - t = 10
  - t = 20
  - t = 40

Answers

Note: 5 terms in MARS model (incl. intercept).

$$\hat{y} = 3994.2 - 172.8 \times \mathrm{h}(\mathrm{t} - 11) + 105 \times \mathrm{h}(\mathrm{t} - 26) + 96.1 \times \mathrm{h}(\mathrm{t} - 32) - 77.6 \times \mathrm{h}(32 - \mathrm{t})$$

Hinge Pair

- At t = 10: $\hat{y} = 3994.2 - 172.8(0) + 105(0) + 96.1(0) - 77.6(22) = 2287$
- At t = 20: $\hat{y} = 3994.2 - 172.8(9) + 105(0) + 96.1(0) - 77.6(12) = 1508$
- At t = 40: $\hat{y} = 3994.2 - 172.8(29) + 105(14) + 96.1(8) - 77.6(0) = 1222$

MARS Theory in ESL textbook vs R earth() Implementation

- MARS theory and math summarized in the textbook Elements of Statistical Learning 2nd Edition [ESL] Section 9.4 pp.321 – 329.
- Link to download free PDF textbook ESL: https://web.stanford.edu/~hastie/ElemStatLearn/
- Different software (R, Python, SAS, etc) have slightly different implementation of MARS and thus results might differ.
- https://cran.r-project.org/web/packages/earth/index.html

Overview of MARS Optimization Process

- Similar strategy to CART.
- Grow by adding the “best” hinge (in pairs) at each step. i.e. the hinge pairs that results in the greatest increase in RSq.
- Prune to remove weakest hinge terms when complexity penalty is included to reduce risk of overfitting, one term at a time.
- Note: Implementation differences with different software.

RMSE Results (on entire dataset) in Ascending Order

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

- R 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 not show any kth power term. This is a valuable feature in MARS.
- No interaction effects are found in MARS degree 2 as dataset contains only one X.

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 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 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 a 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 increasing RSq), the future pairs are actually added solo as the other side was checked and found to contribute “insignificantly” to increasing RSq. In Rpackage earth, the default threshold is less than 0.001 for RSq increase.

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 multiplied terms is the Degree.
- If Degree = 1, no hinge functions are multiplied. i.e. additive model.

The 4 statistics in MARS summary output

```
Number of terms at each degree of interaction: 1 4 (additive model)
GCV 109300    RSS 3886456    GRSq 0.5730591    RSq 0.696496
```

Residual Sum of Squares (aka SSR, SSE)

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

RSS ≥ 0

R Squared.

$$1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

min(RSS) = 0 max(RSS) = TSS

$0 \leq \text{RSq} \leq 1$

Both RSS and RSq do not have model complexity penalty.

Model Performance Metrics

Model	Without Model Complexity Penalty	With Model Complexity Penalty
CART	R(T)	$R(T) + \alpha\|T\|$

Model Performance Metrics

Model	Without Model Complexity Penalty	With Model Complexity Penalty
Linear Regression	RSq	Adjusted RSq $1 - \frac{n-1}{n-1-k}(1-RSq)$ k: num of X variables If k increases a lot but RSq does not change much, then the overall effect is a reduction in Adj RSq. [penalty]

Model Performance Metrics

Model	Without Model Complexity Penalty	With Model Complexity Penalty
MARS (in Python)	RSq	GCV $\frac{RSS}{n\left(1-\frac{nparams}{n}\right)^2}$ "Adjusted RSS" nparams: num of effective parameters If nparams increases a lot but RSS does not change much, then the overall effect is an increase in GCV. [penalty]

Calculating number of effective parameters (nparams)

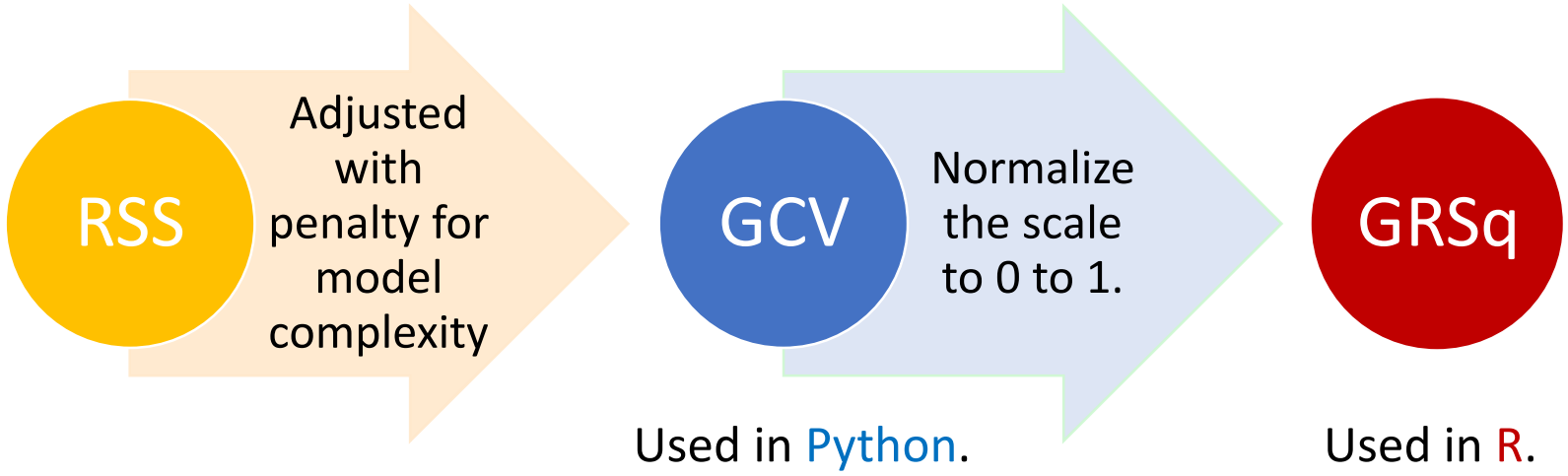
$$nparams = num.of.terms + penalty \times \frac{1}{2}(num.of.terms - 1)$$

- In R: If degree = 1, penalty = 2; Else penalty = 3.
- In Python: penalty = 3

Model Performance Metrics

Model	Without Model Complexity Penalty	With Model Complexity Penalty
MARS (in R)	RSq	GRSq $1 - \frac{GCV}{GCV.null}$ GCV.null: GCV of intercept-only MARS model. If GCV increases a lot then the overall effect is a reduction in GRSq. [penalty] $0 \leq \text{GRSq} \leq 1$

Relationship between RSS, GCV, GRSq



- RSS has no penalty for model complexity and thus will select the most complex model if minimizing RSS.
- 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.

Python Implementation of MARS

- py-earth
  - https://github.com/scikit-learn-contrib/py-earth
  - Learnt from R earth package and made some modifications.
  - Incl. missing data support (R earth has no support for NAs).
  - Some Notable Differences in py-earth compared to R earth:
    - Penalty = 3 (In R earth, penalty = 2 for additive model, 3 otherwise.)
    - Use MSE (same as using RSS or RSq) 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)

Pruning alternative using 10-fold CV

- Default settings:
  - nfold = 0
  - ncross = 1
- For 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.
- Use pmethod = "cv" to select optimal model using cross-validation.
  - Based on max mean OOF RSq, not 1SE rule.

Summary of MARS Model Building Process

Initialize MARS

- Choose degree. Default = 1.
- Choose pruning method.
- Default: Backward
- Option: k-fold CV.
- R Default : GRSq.
- Python Default: GCV.

Grow MARS

- Add the "best" hinge functions in pairs, one pair per step.
- Best determined by largest increase in RSq to current model.
- Update model coefs by linear reg fit if continuous Y (glm fit of categorical Y)
- Stop when Rsq increase < 0.001 or RSq ≥ 0.999.

Prune MARS

- Remove weakest term, one at each step, until only the intercept term remains.
- The weakest term defined as the term whose removal result in model with lowest RSS.
- For each subset, compute complexity adjusted model error GRSq, GCV or mean OOF CV error.
- Use GRSq, GCV, or mean OOF CV error to get optimal number of terms in final model.

Learning Activities

- Excel based exercises
  - Verify the R outputs.
  - Test assumptions (explicit or implicit).
  - Reinforce your understanding.
- Try 10-fold CV instead of the default optimal MARS model selection metrics (GRSq or GCV).
- Multiple X variables.