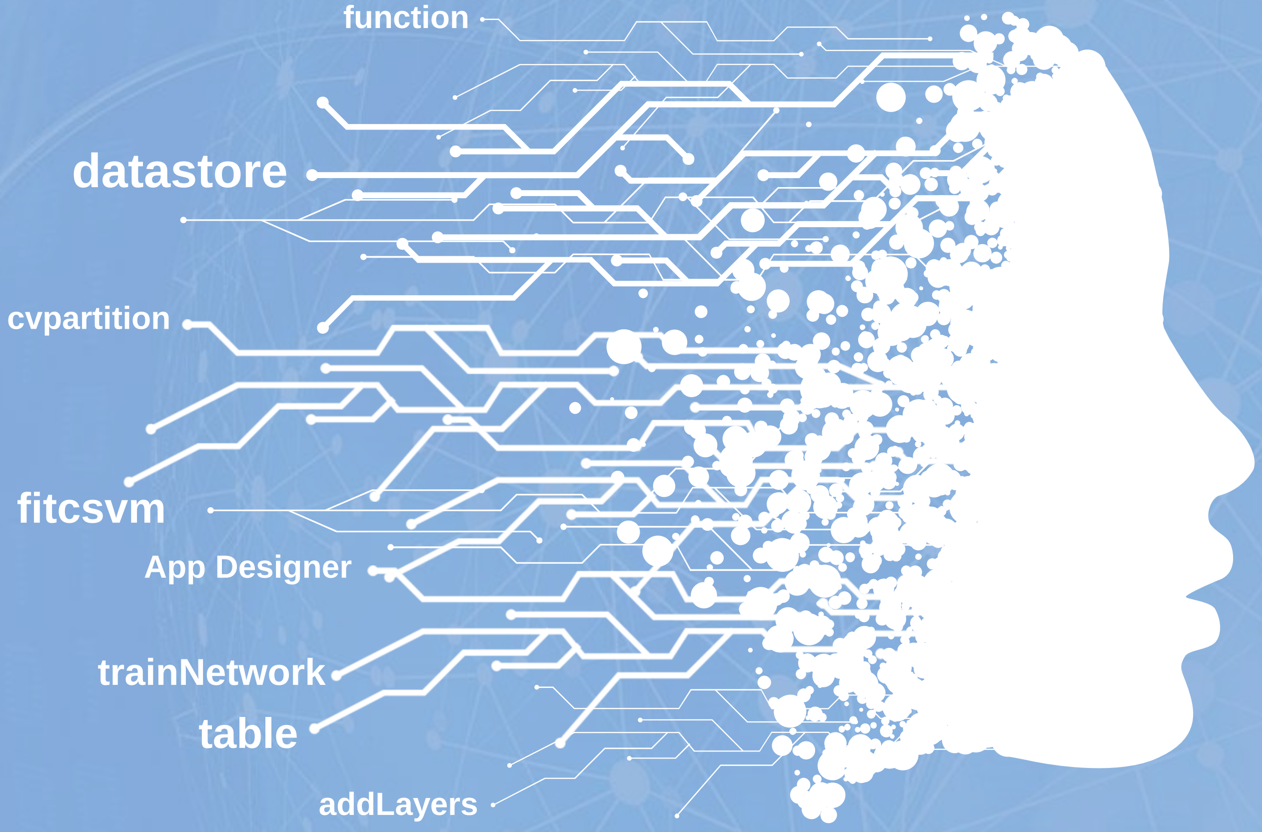
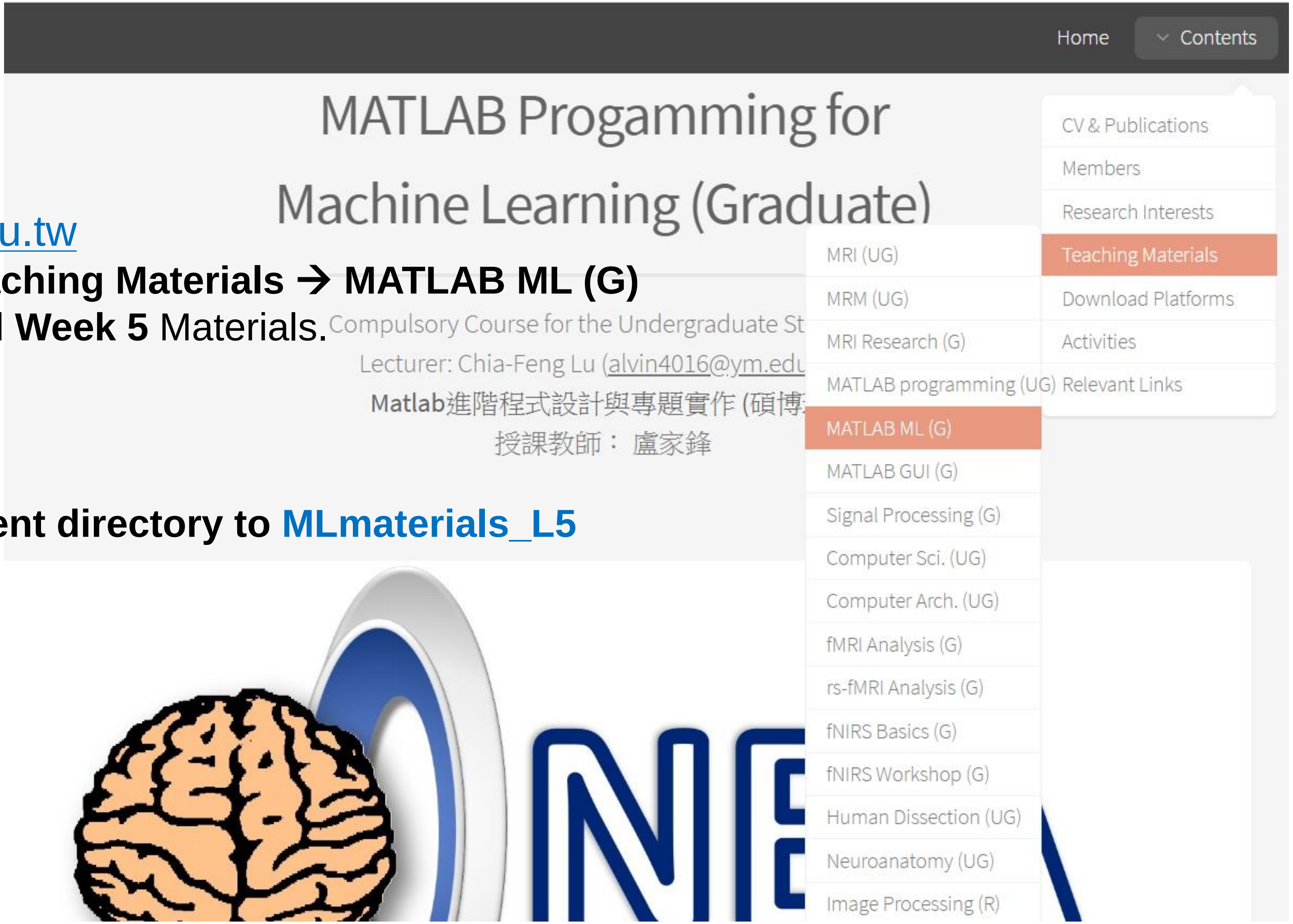


Regression Analysis

MATLAB進階程式語言與實作

盧家鋒 Chia-Feng Lu, Ph.D.
Department of Biomedical Imaging and Radiological Sciences, NYCU
alvin4016@nycu.edu.tw

Teaching Materials



cflu.lab.nycu.edu.tw

Contents → Teaching Materials → MATLAB ML (G)

Please download **Week 5** Materials.

Please set current directory to MLmaterials_L5

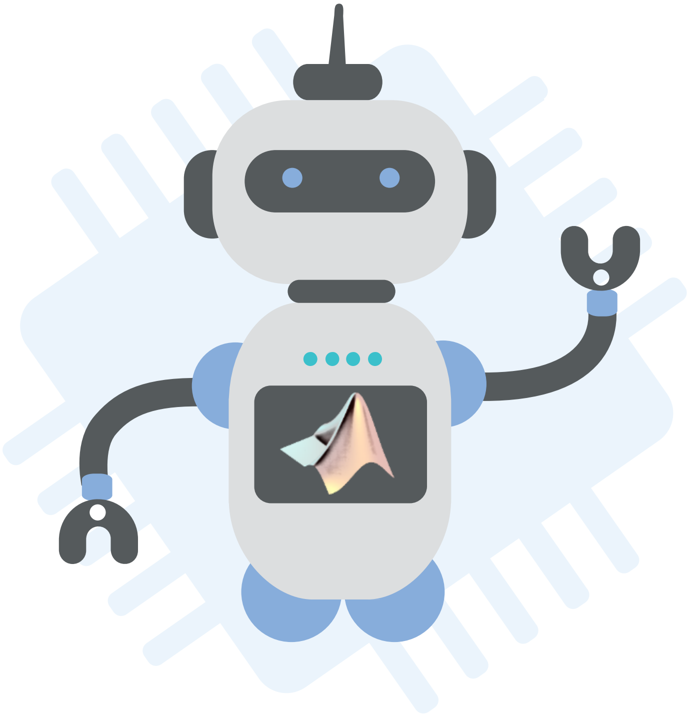
Contents in this Week

01 Parametric Regression

Linear and nonlinear regression

02 Nonparametric Regression

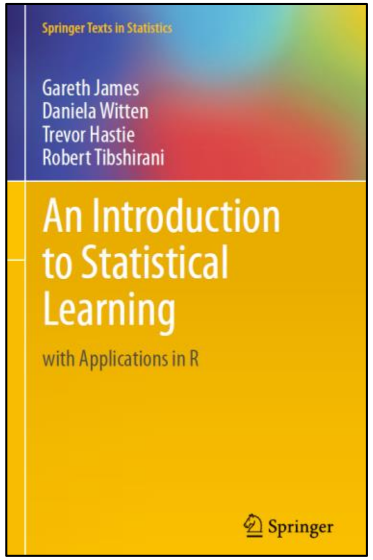
Regression tree



References

[Textbook 3]

- **An Introduction to Statistical Learning, 2nd edition, 2013**
  Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
- **Online resources:** https://github.com/rghan/ISLR
- **Online resources:** https://github.com/JWarmenhoven/ISLR-python
- **Linear and nonlinear regression (Ch.3)**
- **KNN regression (Ch.3.5), regression tree (Ch.8.1)**



Parametric Regression

Linear and nonlinear regression

Linear Regression

- Linear regression is a useful tool for predicting **a quantitative response**.

$$\boldsymbol{Y \approx \beta_0 + \beta_1 X}$$

$$\boldsymbol{\hat{y} \approx \hat{\beta}_0 + \hat{\beta}_1 x}$$

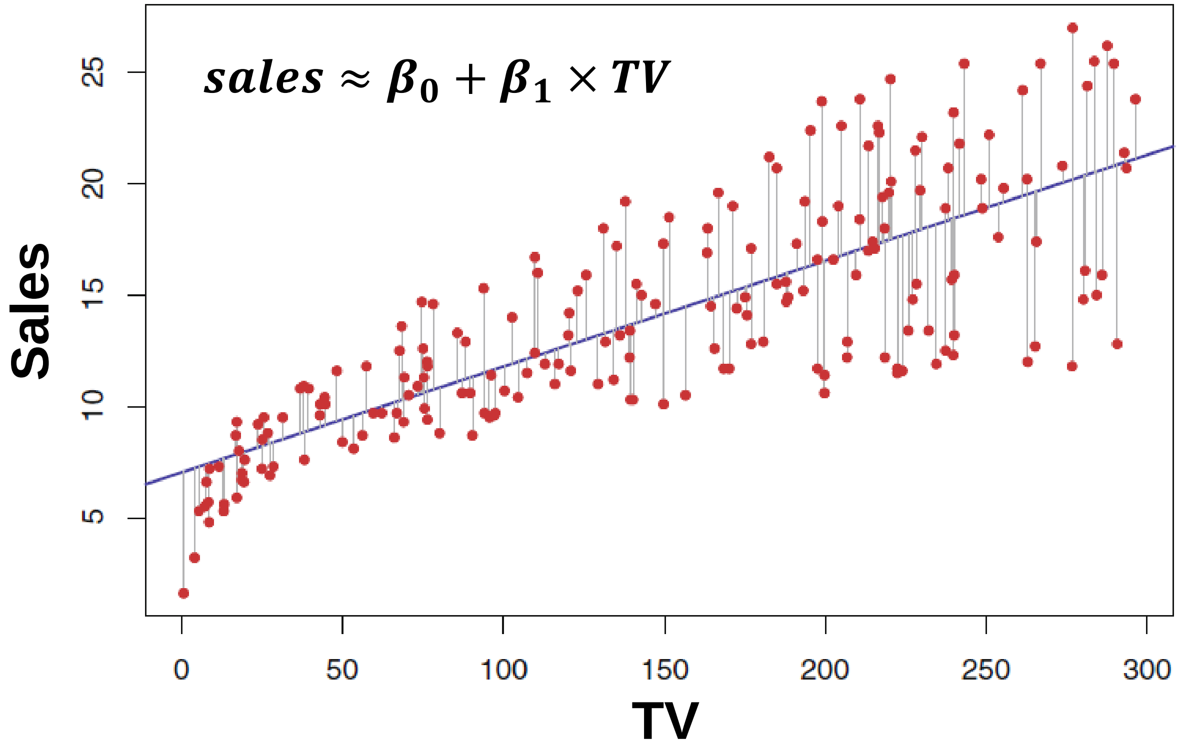
Estimated model coefficients

Predicted response

$$\boldsymbol{e_i = y_i - \hat{y}_i}$$

$$\boldsymbol{RSS = e_1^2 + e_2^2 + \cdots + e_n^2}$$

Residual sum of squares (RSS)



Advertising Dataset

- **Sales** (in thousands of units) for a particular product as a function of **advertising budgets** (in thousands of dollars) for **TV**, **radio**, and **newspaper** media.
- Including 200 observations (markets).

	A	B	C	D	E
1		TV	Radio	Newspape	Sales
2	1	230.1	37.8	69.2	22.1
3	2	44.5	39.3	45.1	10.4
4	3	17.2	45.9	69.3	9.3
5	4	151.5	41.3	58.5	18.5
6	5	180.8	10.8	58.4	12.9
7	6	8.7	48.9	75	7.2
8	7	57.5	32.8	23.5	11.8
9	8	120.2	19.6	11.6	13.2
0	9	8.6	2.1	1	4.8
1	10	199.8	2.6	21.2	10.6
2	11	66.1	5.8	24.2	8.6
3	12	214.7	24	4	17.4
4	13	23.8	35.1	65.9	9.2
5	14	97.5	7.6	7.2	9.7
6	15	204.1	32.9	46	19
7	16	195.4	47.7	52.9	22.4

MLmaterials_L5\Advertising.csv

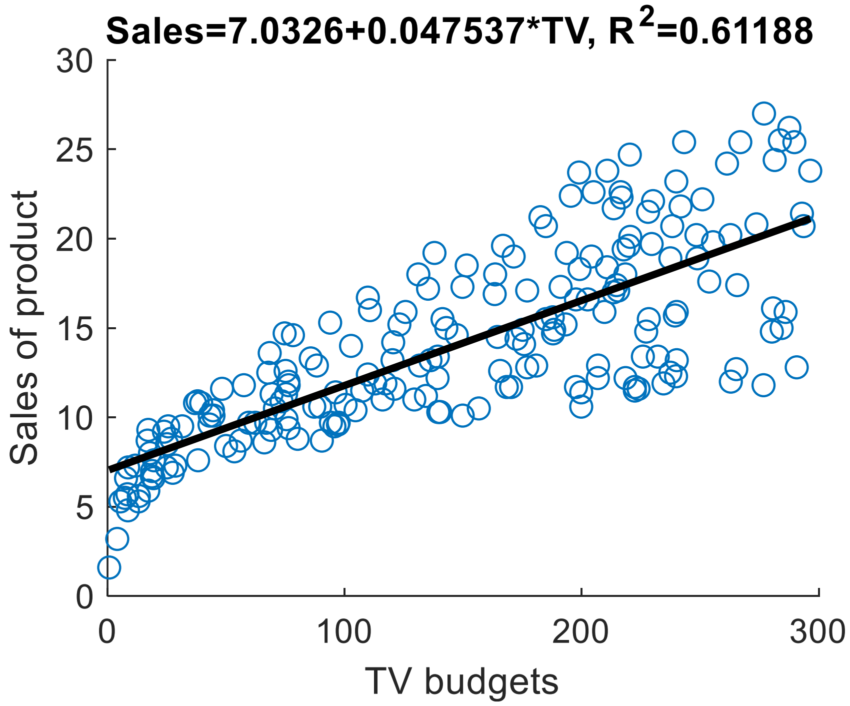
Questions to be Addressed

- *Is there a **relationship** between advertising budget and sales?*
- ***How strong** is the relationship between advertising budget and sales?*
- *Which media **contribute** to sales?*
- *How **accurately** can we estimate the effect of each medium on sales?*
- *How accurately can we **predict future sales**?*
- *Is the relationship **linear**?*
- *Is there **synergy among the advertising media**?*

Exercise – Univariate Linear Regression

- Regress "Sales" on "TV" budget

```
mdl_tv = fitlm(TV,Sales)

figure, scatter(TV,Sales), hold on
plot([min(TV), max(TV)],…
      [predict(mdl_tv,min(TV)), predict(mdl_tv,max(TV))],…
      'k-','linewidth',2)
xlabel('TV budgets'), ylabel('Sales of product')
title(['Sales=' num2str(mdl_tv.Coefficients.Estimate(1)) '+' ...
         num2str(mdl_tv.Coefficients.Estimate(2)) '*TV',...
         ', R^2=' num2str(mdl_tv.Rsquared.Ordinary)])
```



Lines 15 to 28 in MLmaterials_L5\Ex_LinearRegression.m

Exercise – Univariate Linear Regression

- Information of the linear regression model

mdl_tv =

Linear regression model:

y ~ 1 + x1

Estimated Coefficients:

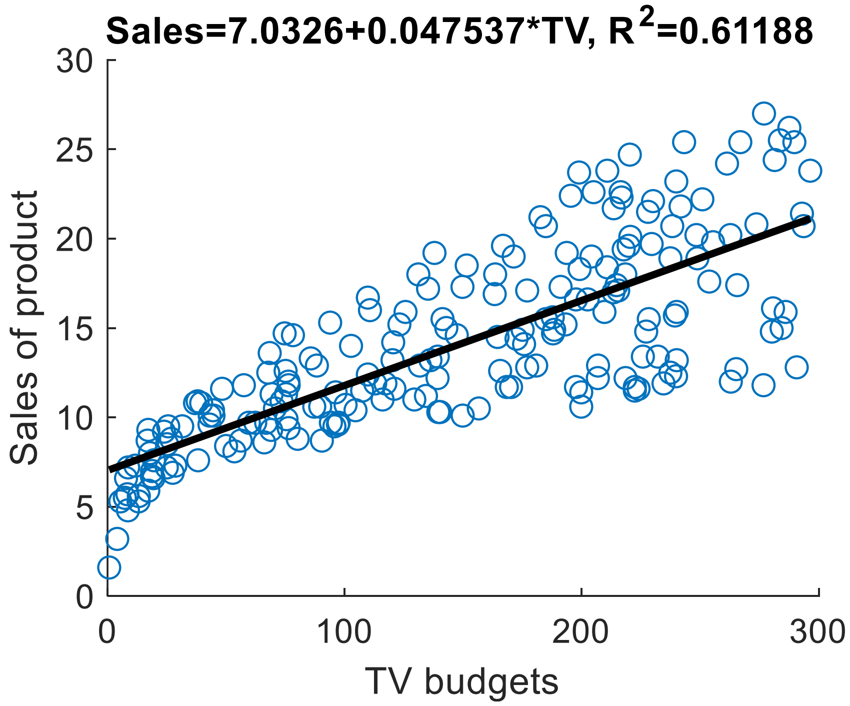
	Estimate	SE	tStat	pValue
(Intercept)	7.0326	0.45784	15.36	1.4063e-35
x1	0.047537	0.0026906	17.668	1.4674e-42

Number of observations: 200, Error degrees of freedom: 198

Root Mean Squared Error: 3.26

R-squared: 0.612, Adjusted R-Squared: 0.61

F-statistic vs. constant model: 312, p-value = 1.47e-42



Exercise – Multiple Linear Regression

$$\boldsymbol{Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}$$

1. *Is at least one of the predictors X1,X2, . . . , Xp useful in predicting the response?*
2. *Do all the predictors help to explain Y , or is only a subset of the predictors useful?*
3. *How well does the model fit the data?*
4. *Given a set of predictor values, what response value should we predict, and how accurate is our prediction?*

- Regress "Sales" on "TV", “Radio", and “Newspaper" budgets

```
mdl_multiple = fitlm([TV,Radio,Newspaper],Sales)
```

Lines 31 to 34 in MLmaterials_L5\Ex_LinearRegression.m

Exercise – Multiple Linear Regression

- Information of the multiple linear regression model

mdl_multiple =

Linear regression model:

y ~ 1 + x1 + x2 + x3

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	2.9389	0.31191	9.4223	1.2673e-17
x1	0.045765	0.0013949	32.809	1.51e-81
x2	0.18853	0.0086112	21.893	1.5053e-54
x3	-0.0010375	0.005871	-0.17671	0.85992

Number of observations: 200, Error degrees of freedom: 196

Root Mean Squared Error: 1.69

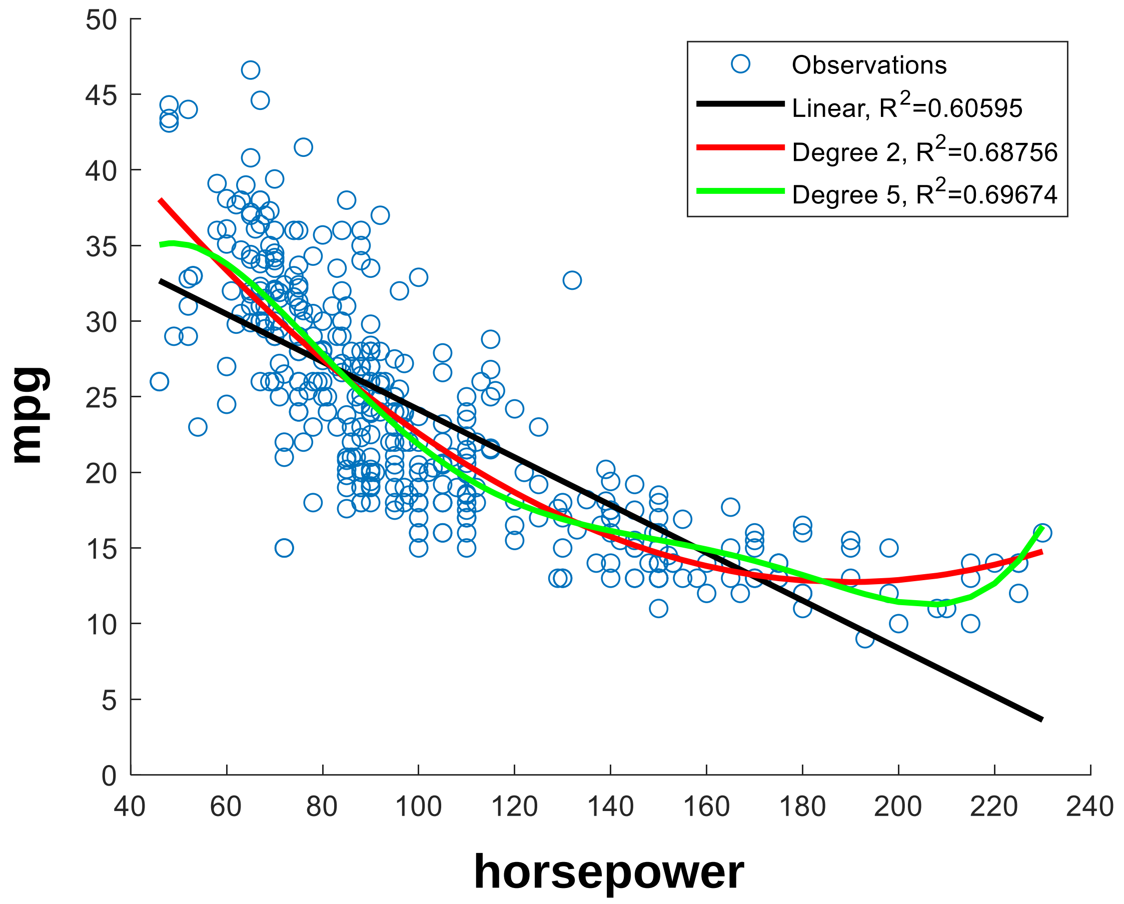
R-squared: 0.897, Adjusted R-Squared: 0.896

F-statistic vs. constant model: 570, p-value = 1.58e-96

Non-Linear Polynomial Regression

$$Y \approx \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p$$

```
p_d2=polyfit(horsepower,mpg,2);

% residual sum of square
RSS=sum((polyval(p_d2,horsepower)…
        -mpg).^2);
% total sum of square
TSS=sum((mpg-mean(mpg)).^2);
R_square_d2=1-RSS/TSS;
```



MLmaterials_L5\Auto.csv
MLmaterials_L5\Ex_PolyRegression.m

2 Nonparametric Regression

Regression tree

Parametric vs. Nonparametric

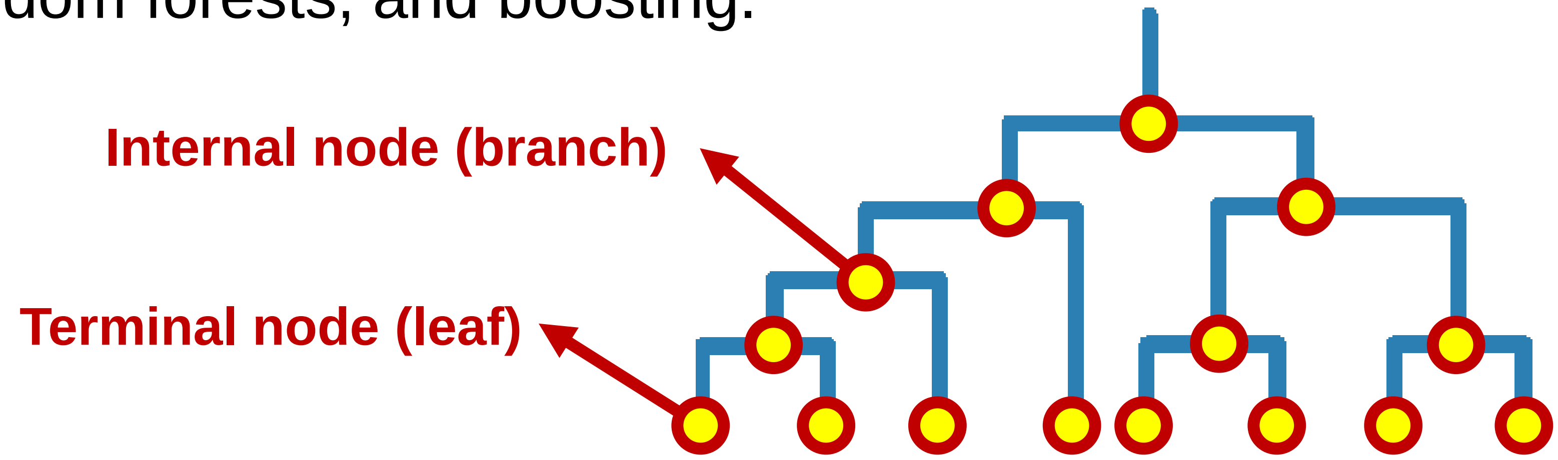
- **Parametric methods**
  - Such as the **linear regression** and **polynomial regression**.
  - **Pro:** They are often easy to fit, because one need estimate only a small number of coefficients.
  - **Con:** They make strong assumptions about the form of f(X).

- **Nonparametric methods**
  - Such as the **KNN regression**, **regression tree**, and **SVM regression**
  - They do not explicitly assume a parametric form for f(X), and thereby provide an alternative and more flexible approach for performing regression.

Decision Trees

- Decision trees can be applied to both **regression** and **classification** problems.
- **Tree-based methods** are simple and useful for interpretation.

- Combining a larger number of trees can often result in improvements in prediction accuracy.
  - Bagging, random forests, and boosting.



Hitters Dataset

- Major League Baseball Data for 322 players from 1986 and 1987 seasons.

- **Salary**: 1987 annual salary on opening day in thousands of dollars
- **Years**: Number of years in the major leagues
- **Hits**: Number of hits in 1986 (安打)
- **RBI**: Number of runs batted in in 1986 (打點)
- **PutOuts**: Number of put outs in 1986 (出局)
- **Walks**: Number of walks in 1986 (保送)
- **Runs**: Number of runs in 1986 (得分)

	A	B	C	D	
1	Player	AtBat	Hits	HmRun	Ru
2	Andy Allanson	293	66	1	
3	Alan Ashby	315	81	7	
4	Alvin Davis	479	130	18	
5	Andre Dawson	496	141	20	
6	Andres Galarraga	321	87	10	
7	Alfredo Griffin	594	169	4	
8	Al Newman	185	37	1	
9	Argenis Salazar	298	73	0	
10	Andres Thomas	323	81	6	
11	Andre Thornton	401	92	17	
12	Alan Trammell	574	159	21	
13	Alex Trevino	202	53	4	
14	Andy VanSlyke	418	113	13	

MLmaterials_L5\Hitters.csv

Hitters Dataset

- Preprocessing steps after reading Hitters.csv
  - Converting data format to the **Table** array.
  - Removing the players with missing Salary data.
  - Log-transform of Salary to have a typical bell-shape distribution.

17	18	19	20	2
Outs	Assists	Errors	Salary	NewLe
~~446~~	~~33~~	~~20~~	~~NaN~~	~~'A'~~
632	43	10	475	'N'
880	82	14	480	'A'
200	11	3	500	'N'
805	40	4	91.5000	'N'
282	421	25	750	'A'
76	127	7	70	'A'

Lines 5 to 16 in MLmaterials_L5\Ex_RegressionTree.m

Algorithm of Regression Tree

1. Use **recursive binary splitting** to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations (**'MinLeafSize'**).

2. Apply **cost complexity pruning** to the large tree in order to obtain a sequence of best subtrees, as a function of α (**'PruneAlpha'**).

3. Return the subtree from Step 2 that corresponds to the chosen value of α. (**prune**)

Algorithm of Regression Tree

- **Recursive binary splitting**

The *j*th predictor

*Cutpoint **s** to split the predictor space*

$$R_1(j,s)=\{X|X_j<s\}\ and\ R_2(j,s)=\{X|X_j\geq s\}$$

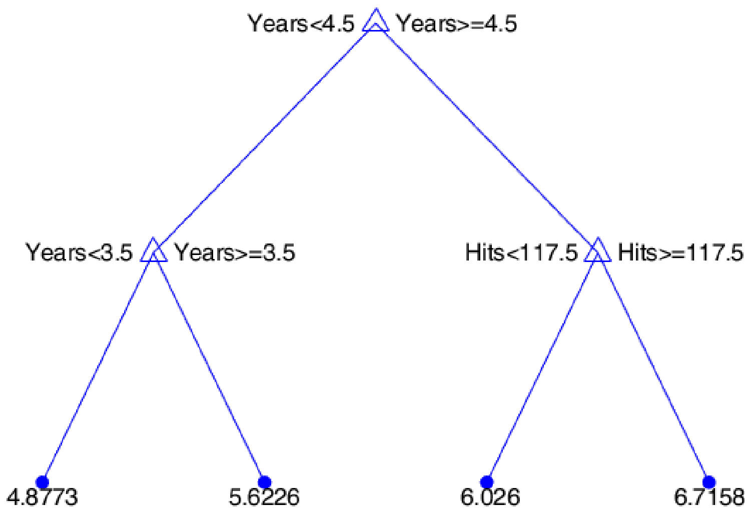
Identify ***j*** and ***s*** to minimize: $$\sum_{x_i\in R_1(j,s)}(y_i-\hat{y}_{R_1})^2+\sum_{x_i\in R_2(j,s)}(y_i-\hat{y}_{R_2})^2$$

- **Cost complexity pruning**

Number of terminal nodes

Given an α, prune the tree to minimize: $$\sum_{m=1}^{|T|}\sum_{x_i\in R_m}(y_i-\hat{y}_{R_m})^2+\alpha|T|$$

$$a\ subtree\ T\subset T_0$$



Exercise – Data Separation

- **Separate data into training (70%) and test (30%) datasets**

```
rng(0,'twister')  % For reproducibility

C = cvpartition(size(data,1),'holdout',0.30); % hold out 30% for test
dataTrain = data(C.training,:);
dataTest = data(C.test,:);
```

185 players (70%) for training

78 players (30%) for test

Lines 18 to 23 in MLmaterials_L5\Ex_RegressionTree.m

Exercise – Regression Tree

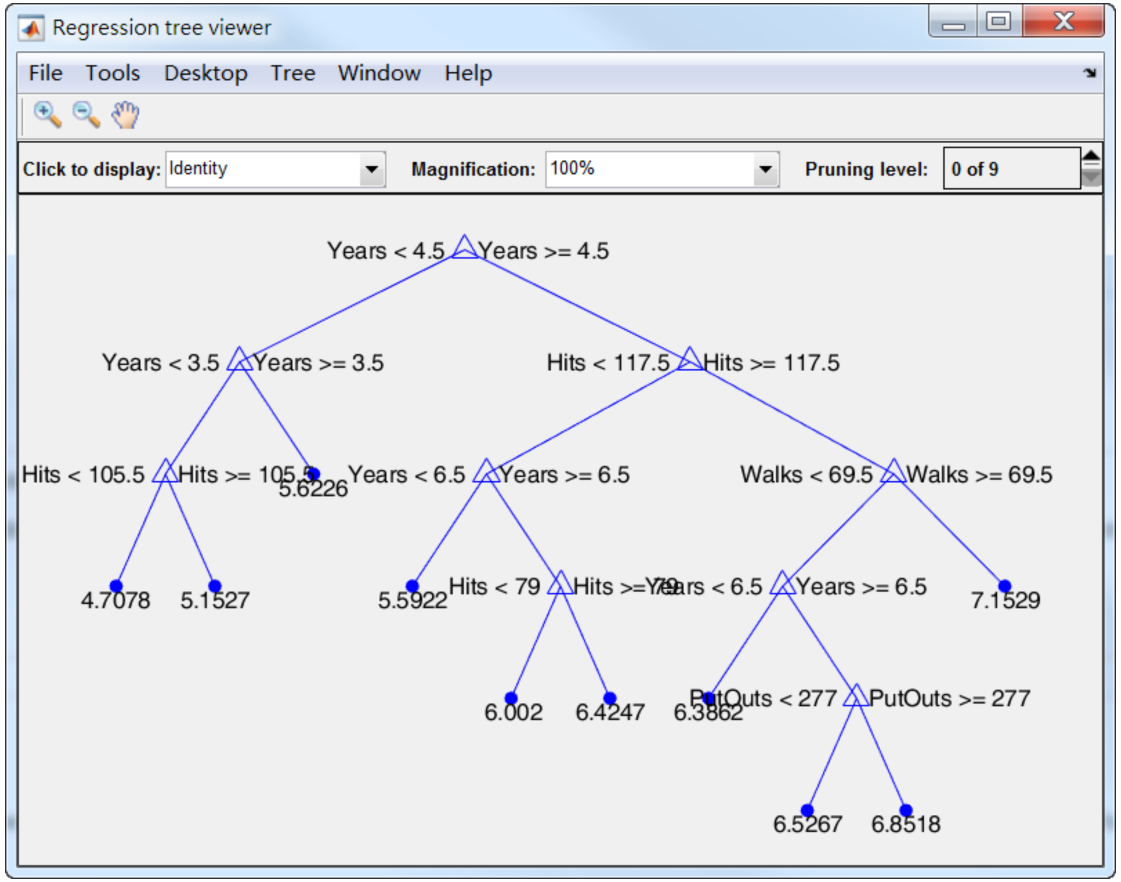
- **[Method 1]** Construct a regression tree using six variables/features

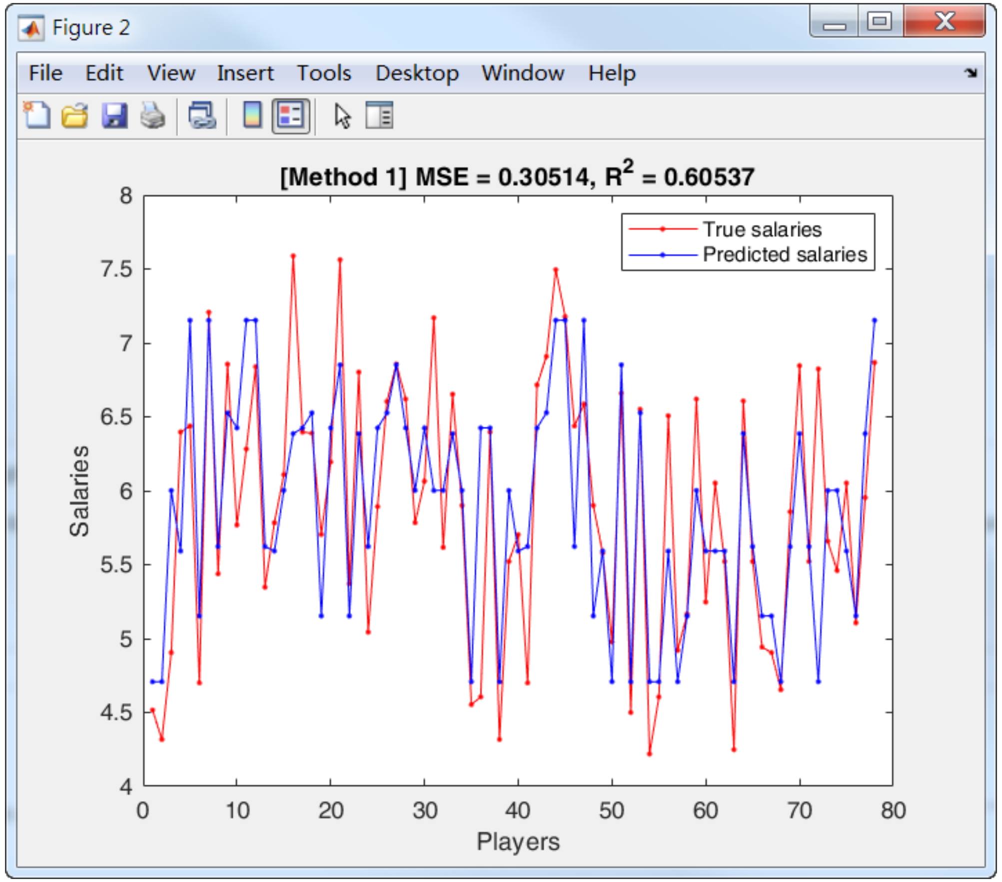
```
predictors={'Years','Hits','RBI','PutOuts','Walks','Runs'};
tree_6v = fitrtree(dataTrain,'Salary','PredictorNames',predictors,...
    'OptimizeHyperparameters','all',...
    'HyperparameterOptimizationOptions',...
    struct('AcquisitionFunctionName','expected-improvement-plus','kfold',5));

view(tree_6v,'Mode','graph')
```

Lines 25 to 46 in MLmaterials_L5\Ex_RegressionTree.m

Exercise – Regression Tree

- **[Method 1]** Construct a regression tree using six variables/features

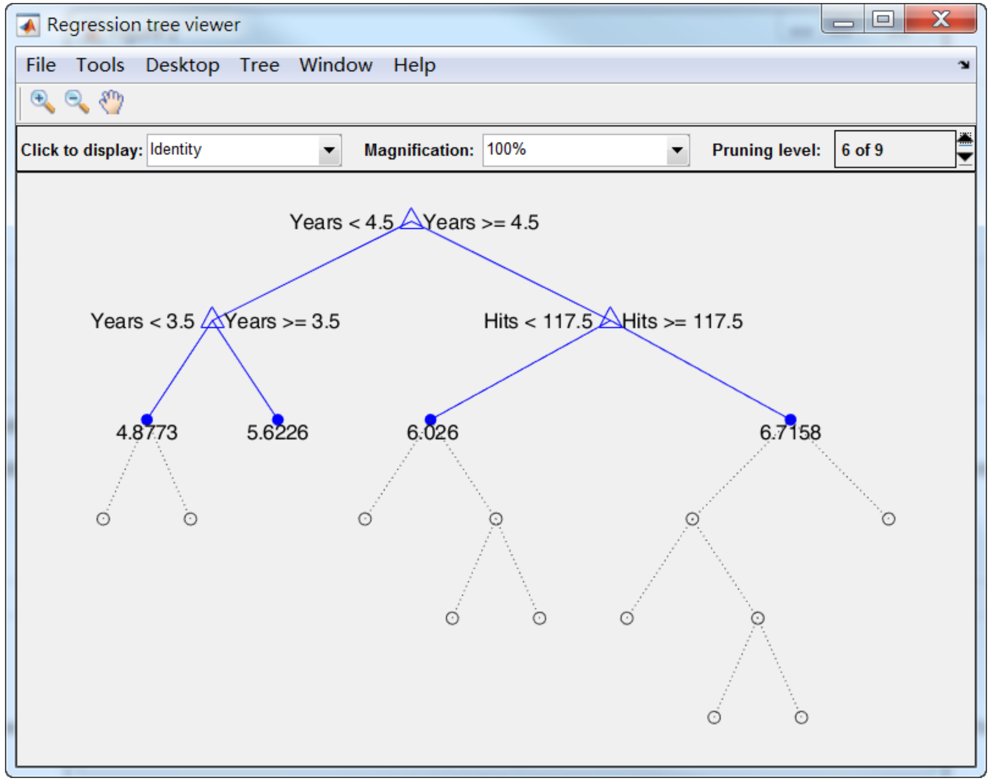




Lines 25 to 46 in MLmaterials_L5\Ex_RegressionTree.m

Exercise – Tree Pruning

- **[Method 2]** Prune the constructed regression tree to reduce the complexity

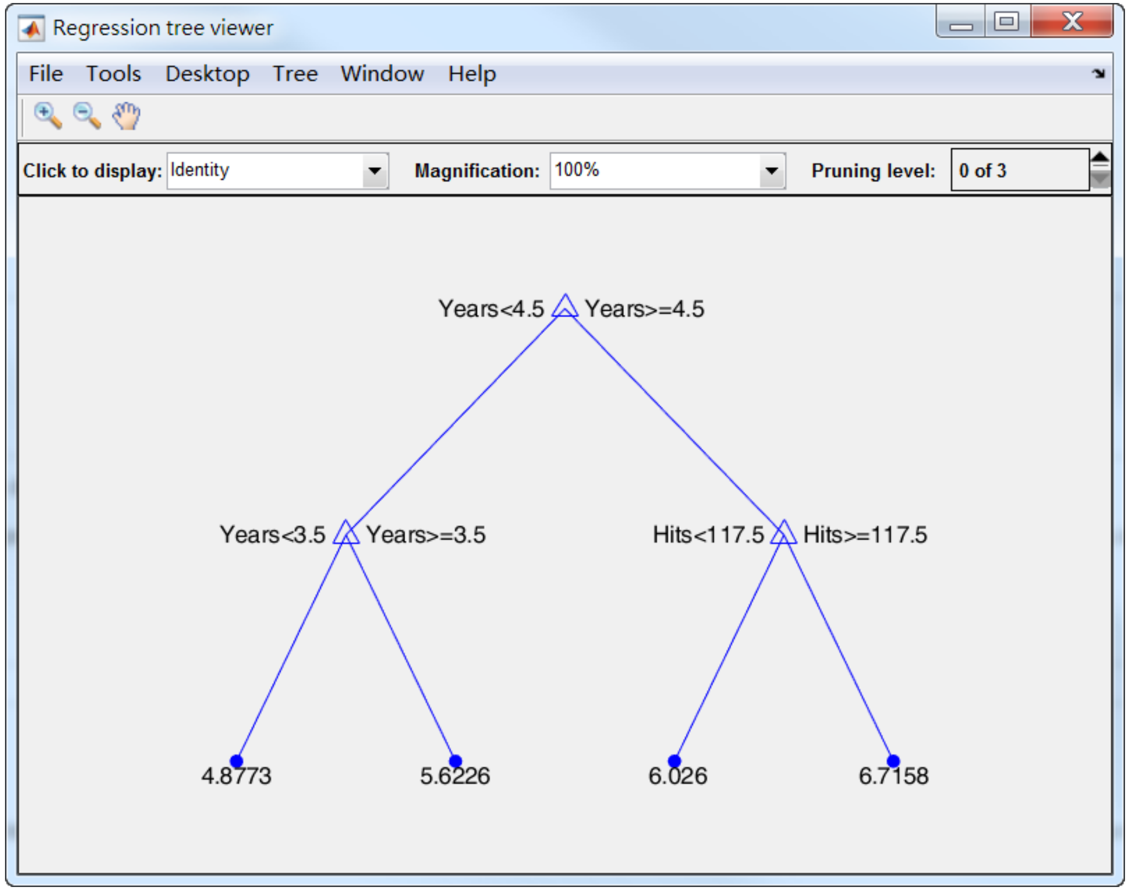


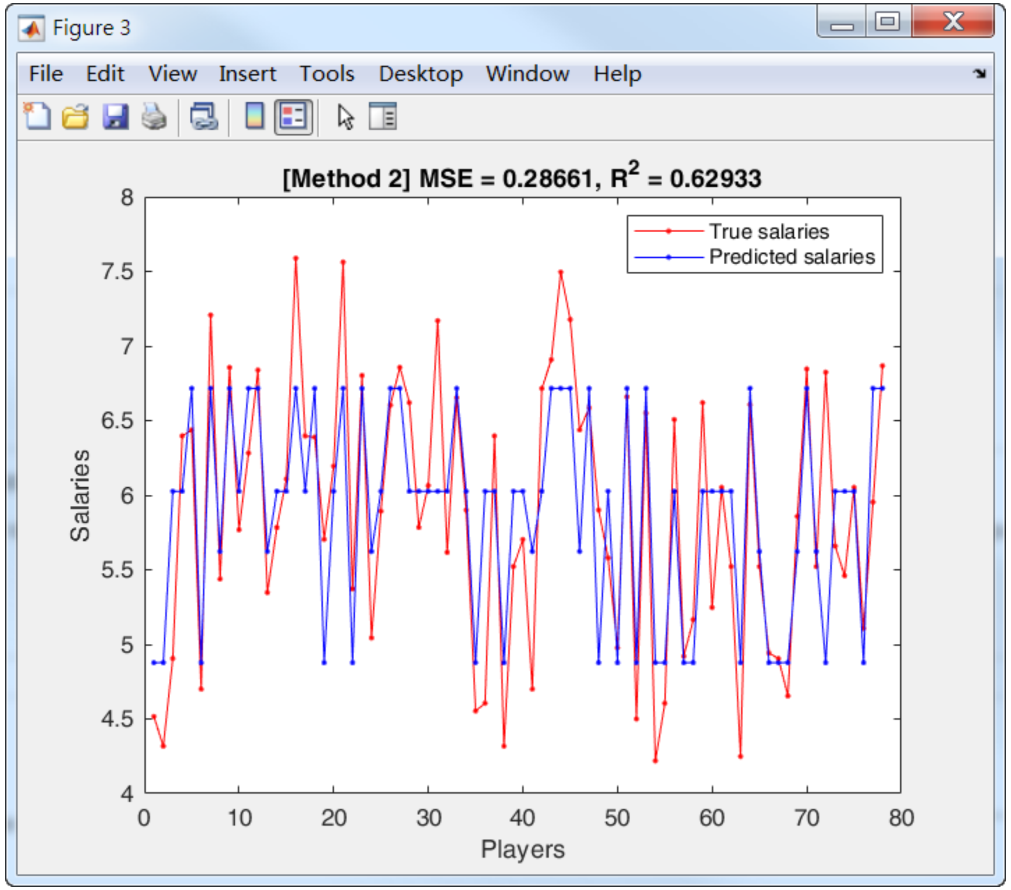

```
prunelevel=6;
tree_prune = prune(tree_6v,'level',prunelevel);

% prunealpha=0.03;
% tree_prune = prune(tree_6v,'alpha',prunealpha);

view(tree_prune,'Mode','graph')
```

Lines 47 to 68 in MLmaterials_L5\Ex_RegressionTree.m

Exercise – Tree Pruning

- **[Method 2]** Prune the constructed regression tree to reduce the complexity



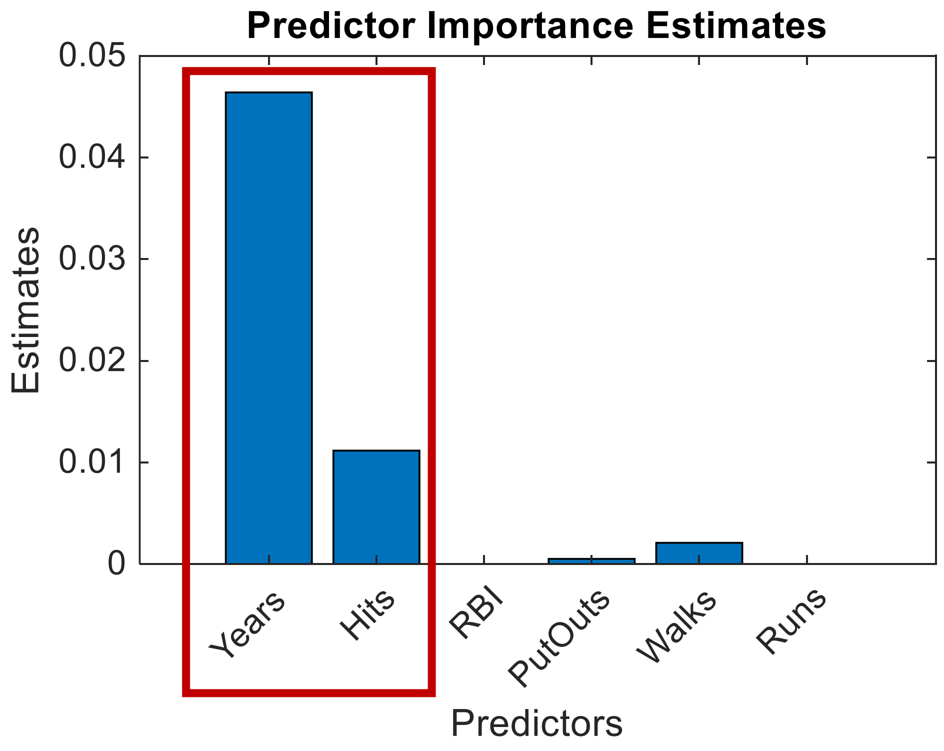


Lines 47 to 68 in MLmaterials_L5\Ex_RegressionTree.m

Exercise – Identify Key Variables

- Variable/feature selection based on the importance scores

```
imp = predictorImportance(tree_6v);

figure; bar(imp);
title('Predictor Importance Estimates');
ylabel('Estimates'); xlabel('Predictors');
h = gca;
h.XTickLabel = tree_6v.PredictorNames;
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
```



Lines 69 to 81 in MLmaterials_L5\Ex_RegressionTree.m

Exercise – Regression Tree

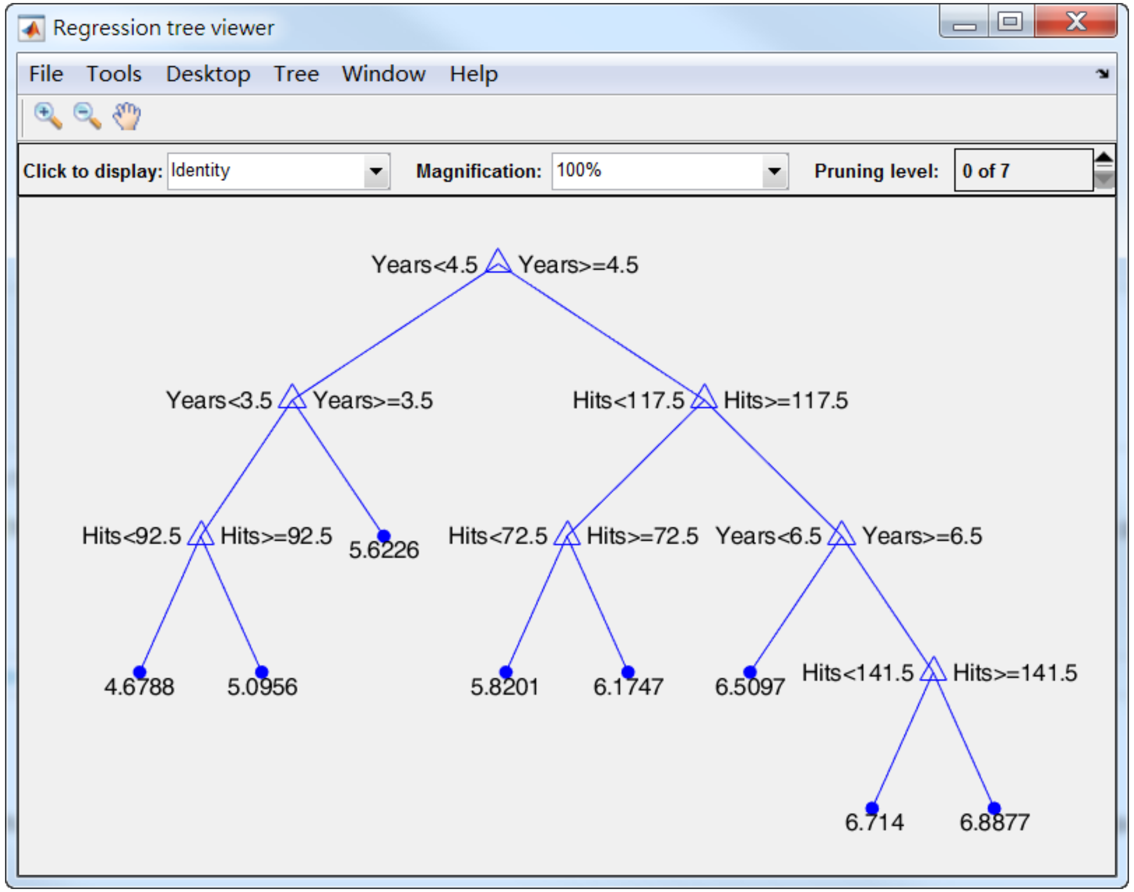
- **[Method 3]** Construct a regression tree using key variables/features

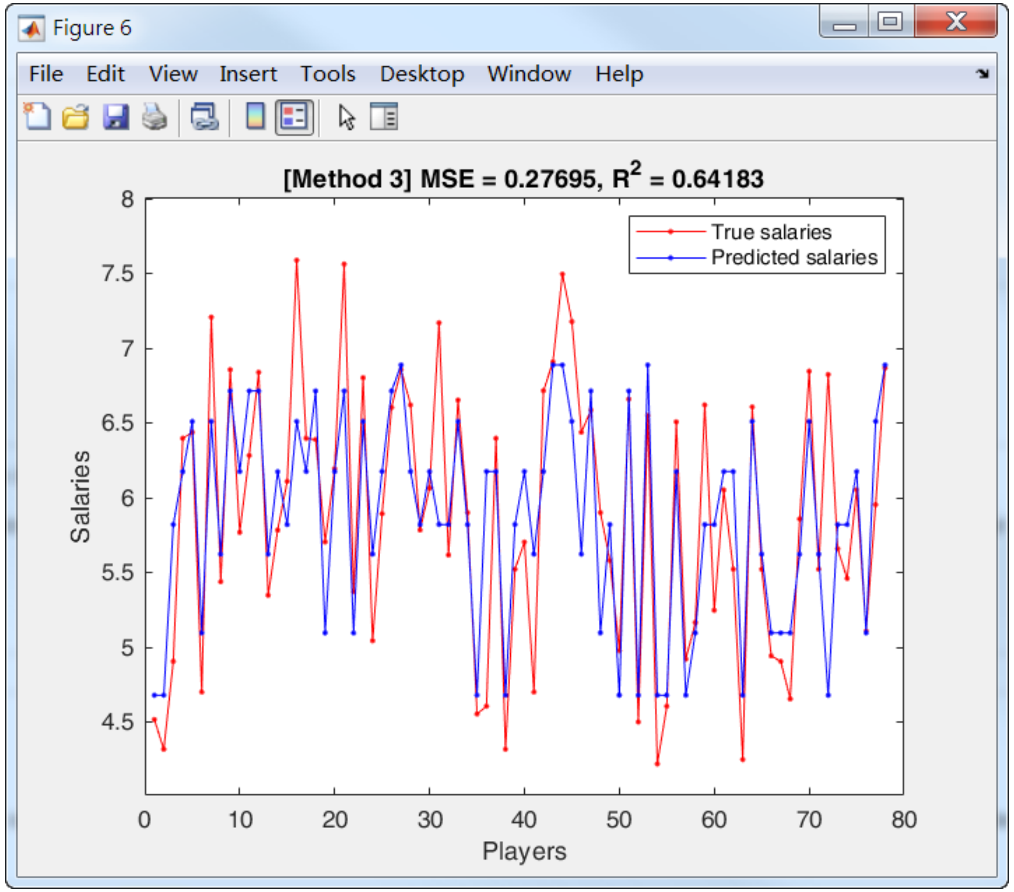
```
predictors={'Years','Hits'};  % only use the two key variables/features
tree_2v = fitrtree(dataTrain,'Salary','PredictorNames',predictors,...
    'OptimizeHyperparameters','all',...
    'HyperparameterOptimizationOptions',...
    struct('AcquisitionFunctionName','expected-improvement-plus','kfold',5));

view(tree_2v,'Mode','graph')
```

Lines 82 to 102 in MLmaterials_L5\Ex_RegressionTree.m

Exercise – Regression Tree

- **[Method 3]** Construct a regression tree using key variables/features

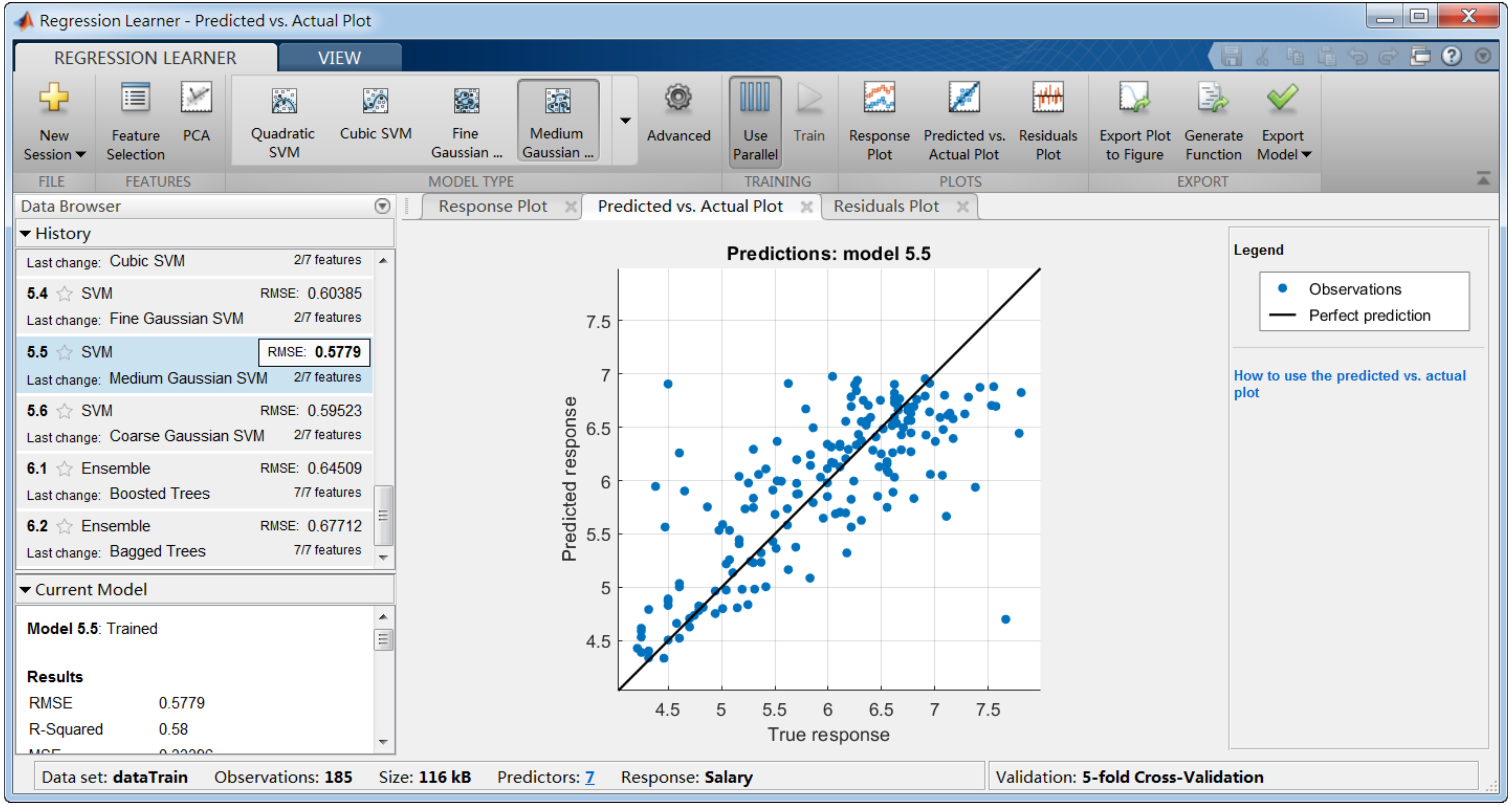




Lines 82 to 102 in MLmaterials_L5\Ex_RegressionTree.m

MATLAB Regression Learner

>> **regressionLearner**



THE END

Contact:
盧家鋒 alvin4016@nycu.edu.tw