Predicting Expenditure Per Person for Cities

Group Gamma

Daniel Eck, Lian Hortensius, Yuting Sun

Bo Yang, Jingnan Zhang, Qian Zhao

Background

Client

A government organization of State Planning Commission (SPC).

Objective

Three housing projects have been proposed by three companies for new development in three different cities. The client wanted to get some advice on approving these housing projects, specifically, they wanted to know **how the expenditure/person will change in 2020 and 2040 for these three cities.**

What data look like?

Data for estimation – information for 914 cities in 2010

City ID	Expenditure	Wealth	Population	Percent Intergovernmental	Density	Income	Growth Rate
680	316	41692	3157	6.3	49	13146	20.6
710	345	58689	28203	7.1	578	21032	13.4
1800	298	55629	4572	24	94	15783	-4
1870	480	58867	64628	9	1306	16887	3.9
3550	301	52979	27649	10.5	483	20227	8

What data look like?

Data for prediction provided by the developing team of housing project

City Name	City ID	Year	Expenditure	Population	Wealth	Percent Intergovernmental	Density	Income	Growth Rate
Warwick	8730	2020		20442	85000	24.7	214	19500	35
Warwick	8730	2040		31011	89000	26	325	20000	40
Monroe	5420	2020		10496	58000	8.8	695	17100	35
Monroe	5420	2040		13913	60000	10.1	959	18000	35
Tuxedo	8400	2020		10685	116000	6.1	249	28300	300
Tuxedo	8400	2040		29246	115000	7	656	25000	100

Convert Real Problem into Statistical Problem

Models of Interest

Linear Regression Model

Local Polynomial Model

Random Forest

Principal Components Analysis

Method for Comparison

Cross-Validation

Measure

Prediction sum of squares error

Some Concerns About All Methodologies

- Correlation among cities

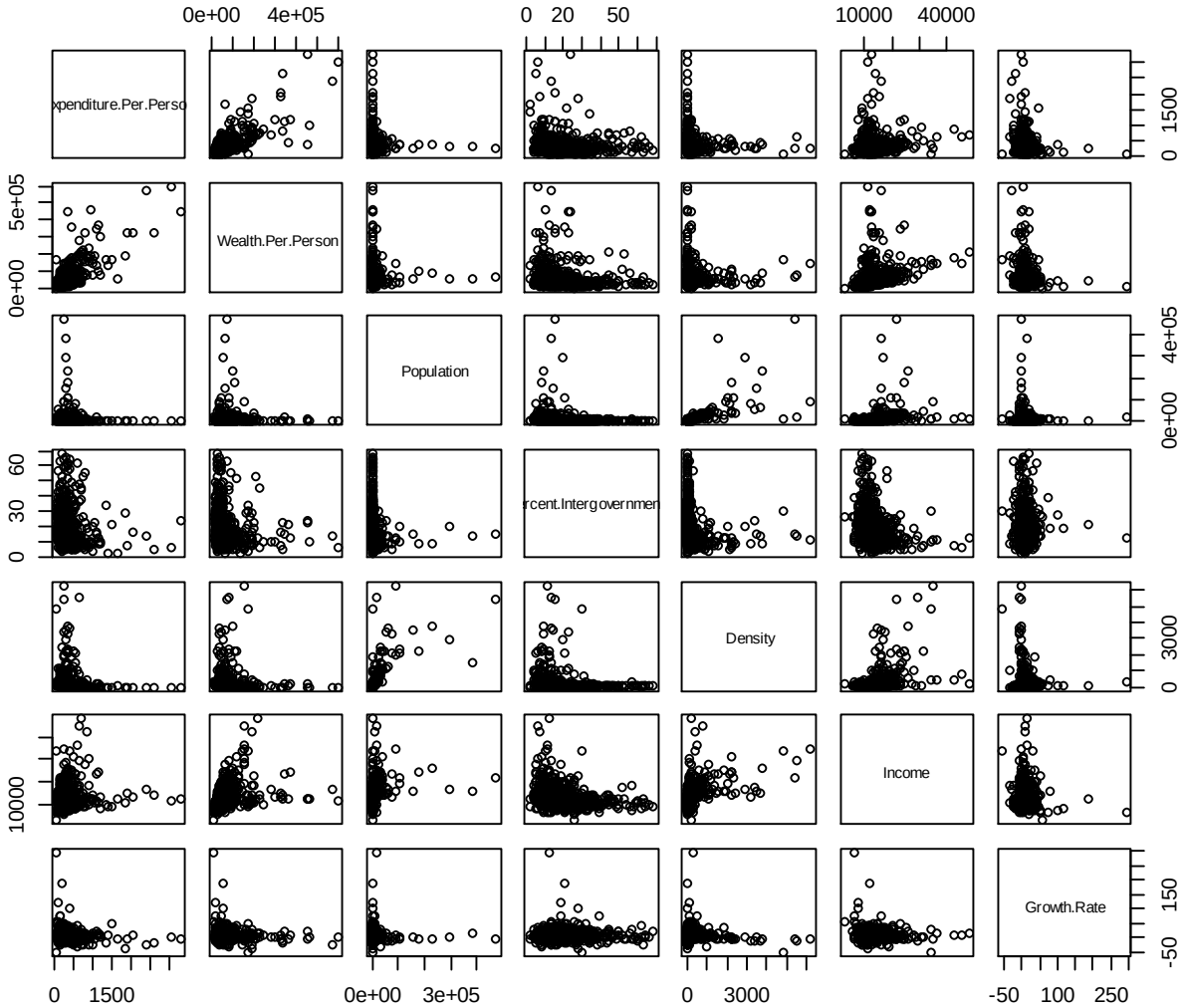
  Spatial relationship

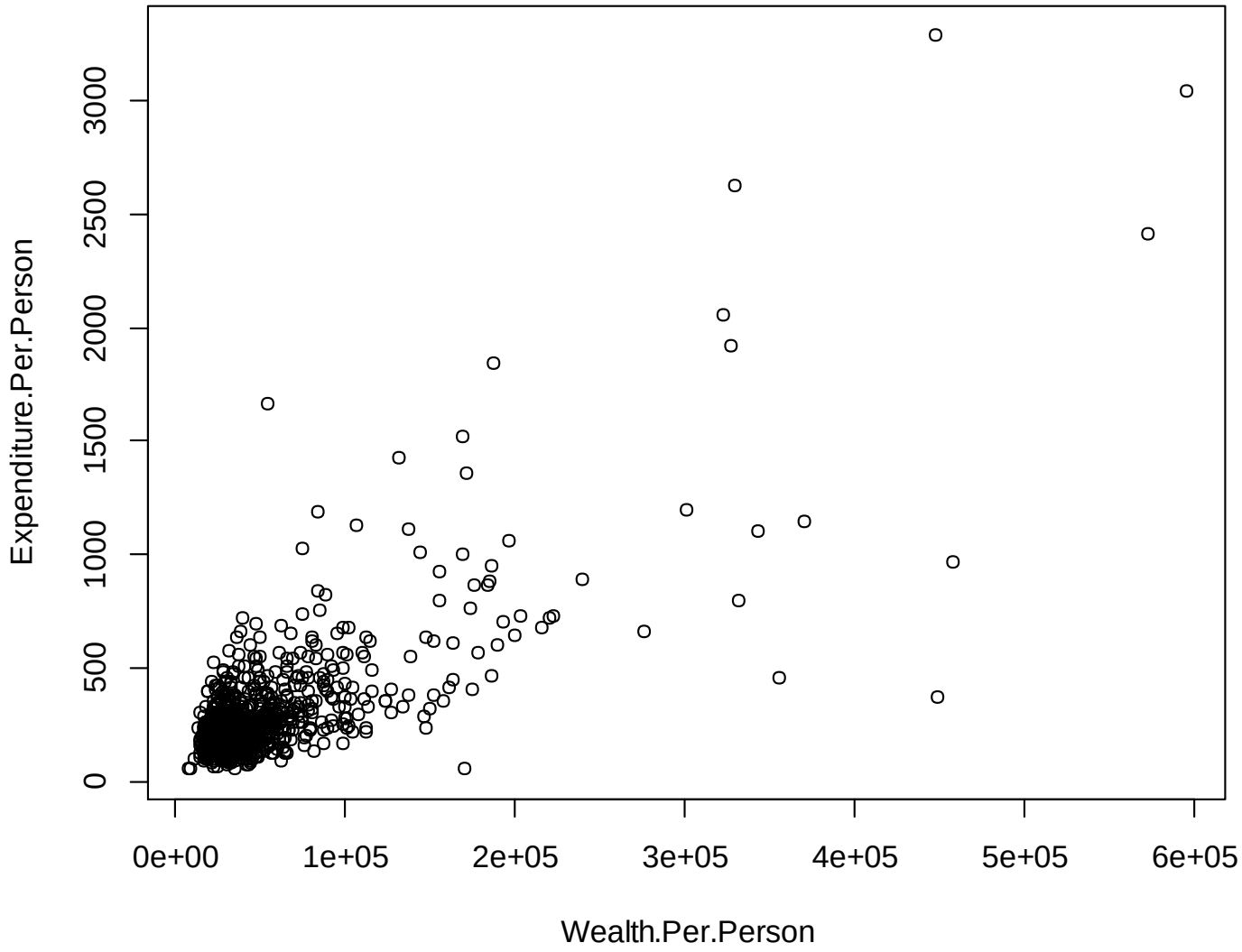
  Time relationship

- Assuming the relationship between predictor and responses are stable from 2010 to 2040.

- Information provided for three cities in 2020 and 2040 is accurate.

Linear Regression Models




Expenditure.Per.Person
0
500
1000
1500
2000
2500
3000
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
Wealth.Per.Person

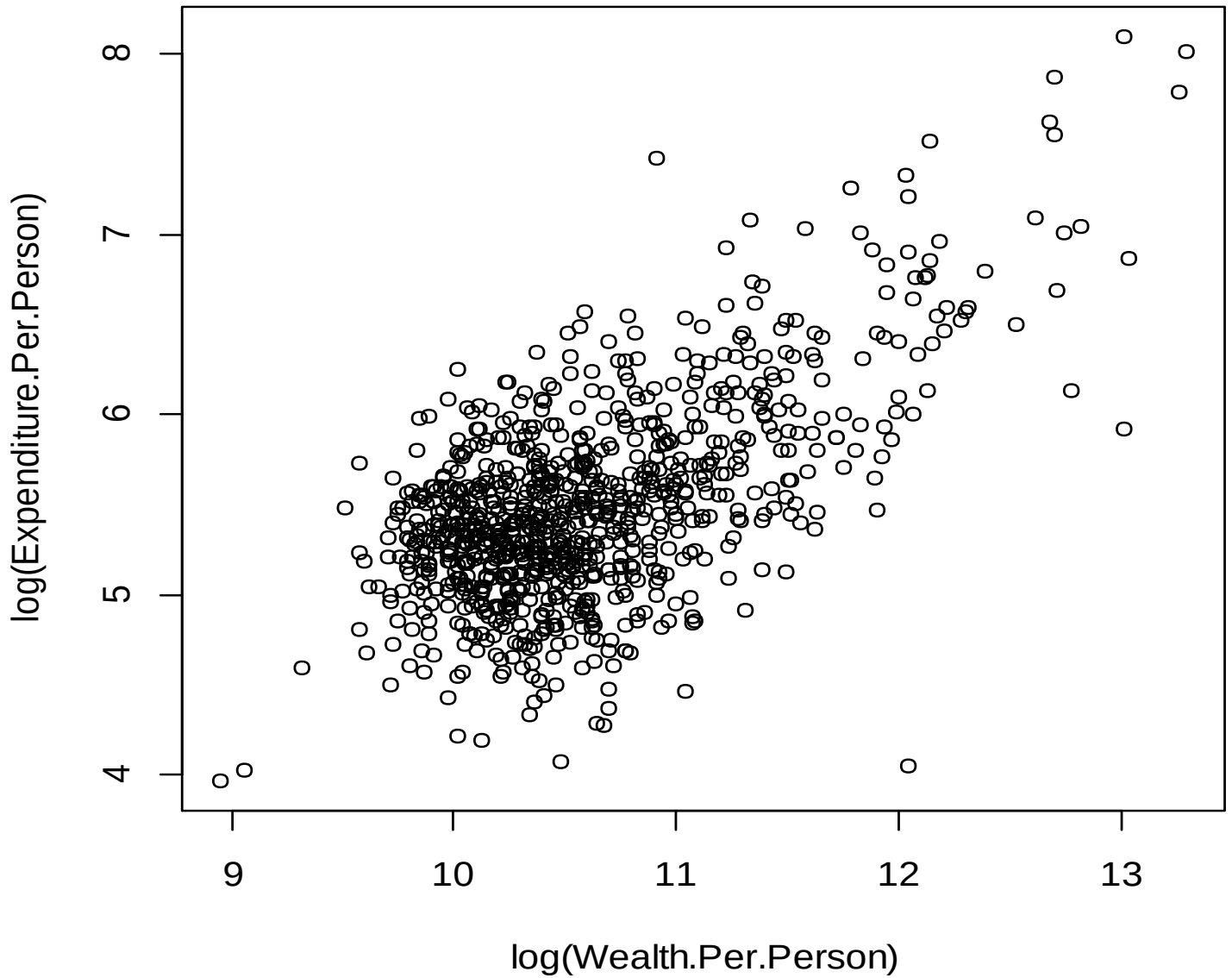
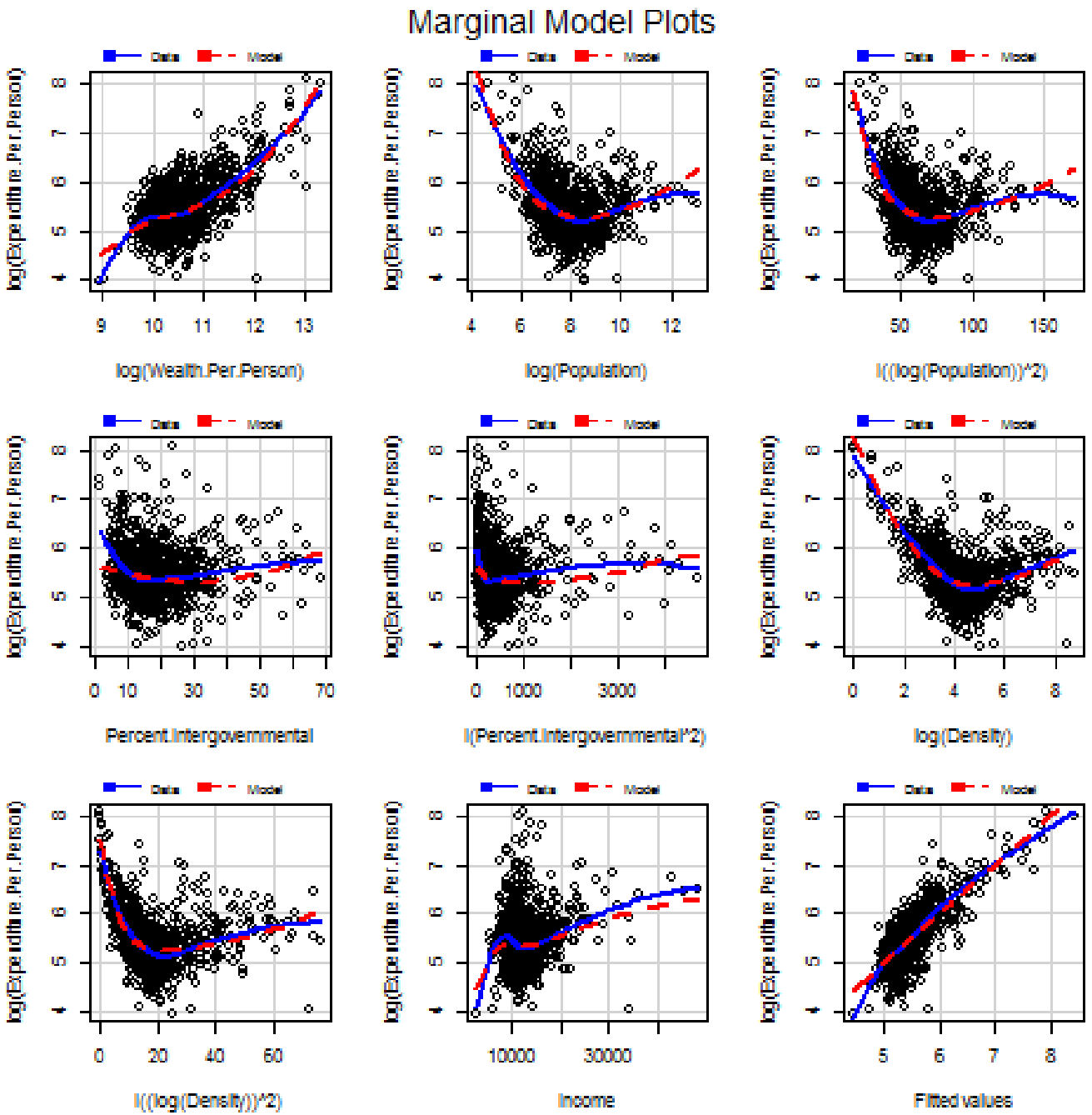
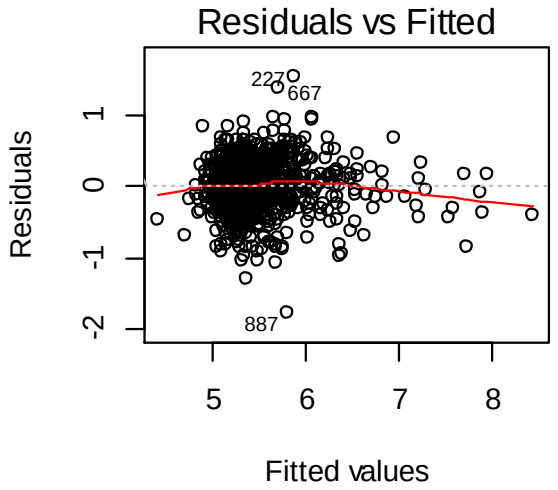

log(Expenditure.Per.Person)
4
5
6
7
8
9
10
11
12
13
log(Wealth.Per.Person)

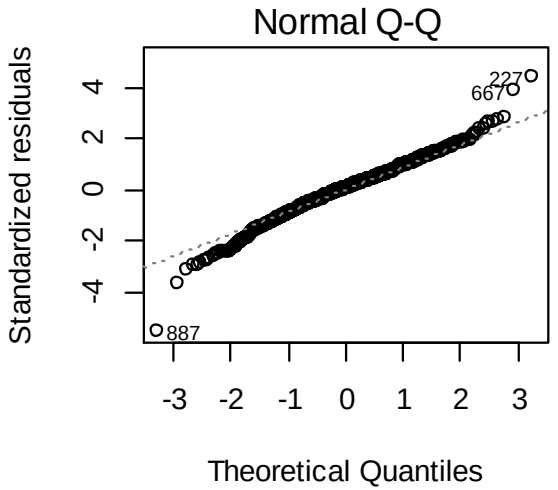
Table 1: Transformations

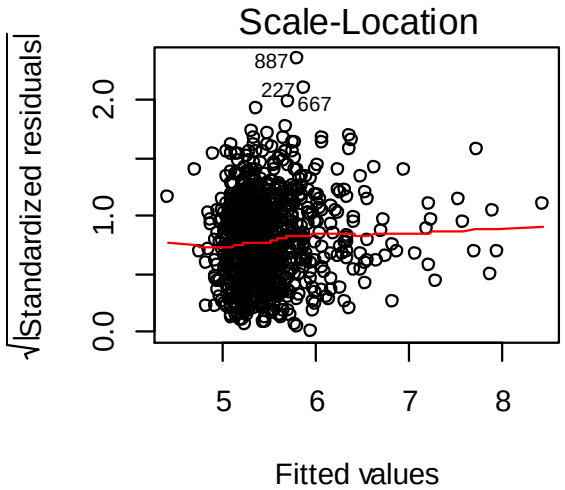
Variable	Transformation
Expenditure.Per.Person	log
Wealth.Per.Person	log
Population	log , square of log
Percent.Intergovernmental	none , square
Density	log , square of log
Income	none
Growth.Rate	not used

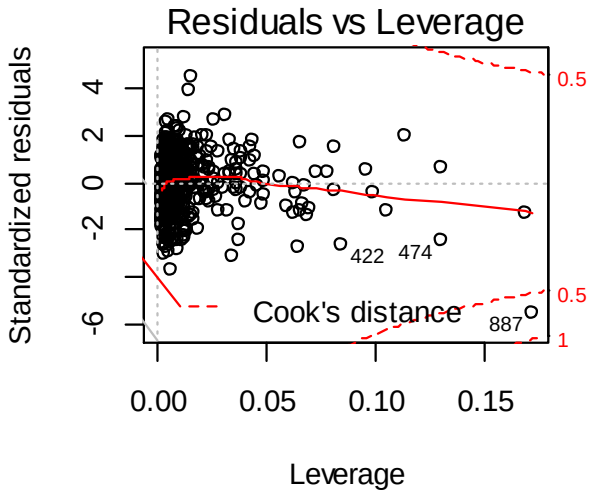

Marginal Model Plots
Data
Model
log(Expenditure.Per.Person)
log(Wealth.Per.Person)
log(Population)
I((log(Population))^2)
Percent.Intergovernmental
I(Percent.Intergovernmental^2)
log(Density)
I((log(Density))^2)
Income
Fitted values

Check Assumption

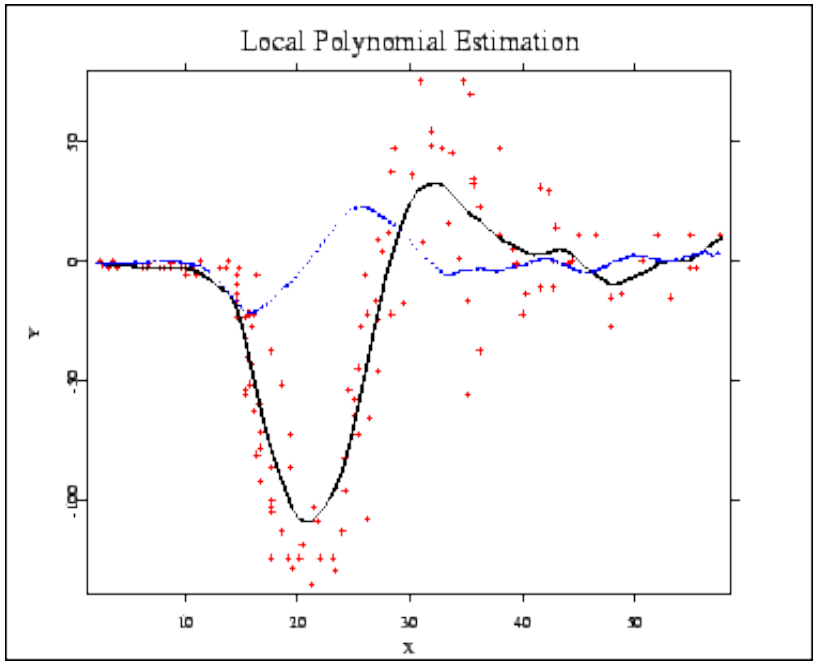








Local Polynomial Model



Local polynomial model is one of the non-parametric regression model.

General steps to build a local polynomial model

- Take localized subsets of data.
- Fit polynomial regression models for each subset.
- Use weight function to make point estimation by giving the most weight to the data points nearest the point of estimation and the least weight to the data points that are furthest away.

Strength

No restraint on the data. No assumptions for model.

Local Polynomial Model Ctd.

Packages for fitting LPM in R: car, locfit.

Predictors: Wealth.Per.Person, Percent.Intergovernmental, Density, Income.

Fitting: Used four predictors at a time.

The Local Fit Model

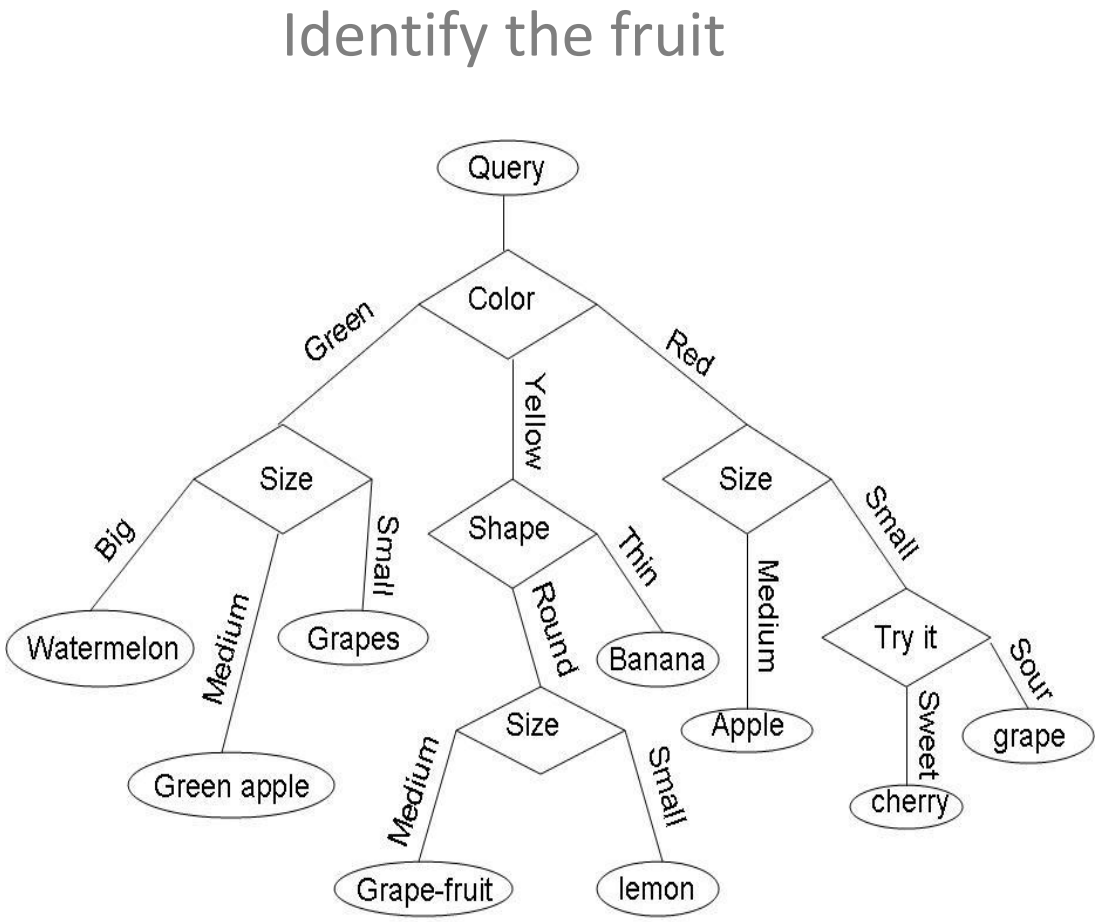
log(Expenditure.Per.Person)~ lp (Wealth.Per.Person, Density, Income,
Percent.Intergovernmental , degree=2)

Random Forests Model

Random Forest is a multitude of random decision trees.

Example of a decision tree

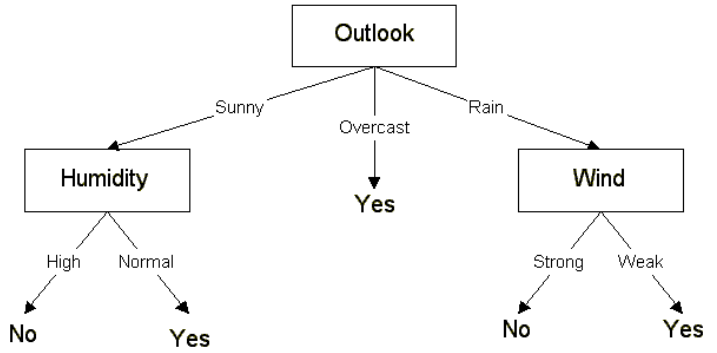
A set of 7 categories {watermelon, apple, grape, lemon, grapefruit, banana, cherry}

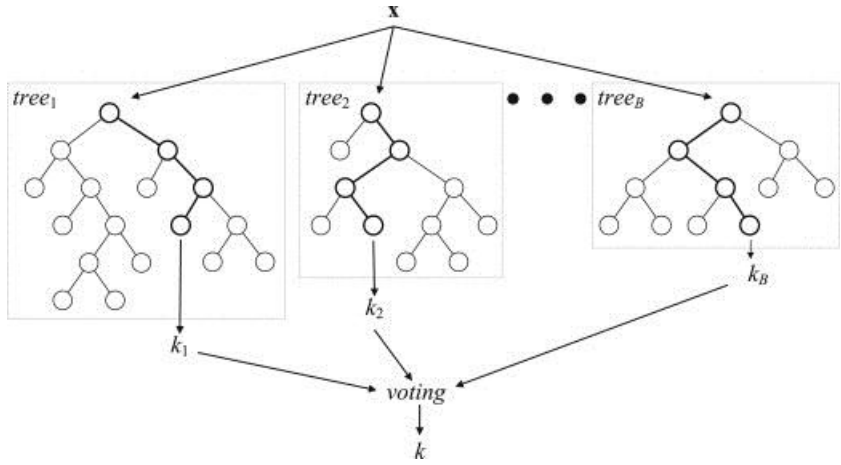


Random Forests Model Ctd.

General steps

- Sample the cases in the training set with replacement at random.
- Build a decision tree depending the data set get by step 1.
- Repeat step 1 and step 2 to get a forest.
- Get the final result according to the output of each tree (the mean for example).





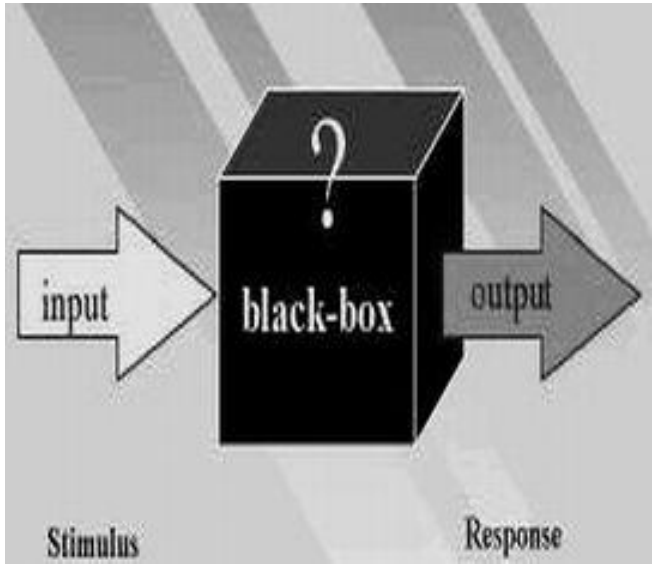
Random Forests Model Ctd.

Strength

- Does not have assumptions, be flexible for any type of data from unknown distribution.

Weakness

- Difficult to interpret/understand the model.
- Difficult to control.



Principal Component Analysis (PCA)

Goal: Dimension reduction

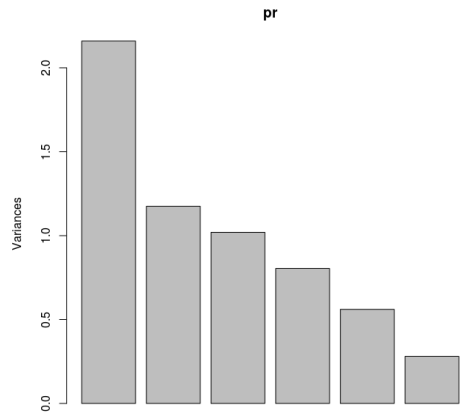
Assume y is dependent on p predictors $x_1, x_2, \ldots, x_p$.

- Find principal components (uncorrelated lineaer combinations) $PC_1, PC_2, \ldots, PC_p$ of original predictors;
- List the PCs from highest variance to least;
- Find a subset $PC_1, PC_2, \ldots, PC_q$ $(q < p)$ of the PCs, such that the subset makes a high proportion of the total variance;
- Fit linear model

$$y \sim PC_1 + PC_2 + \ldots + PC_q$$

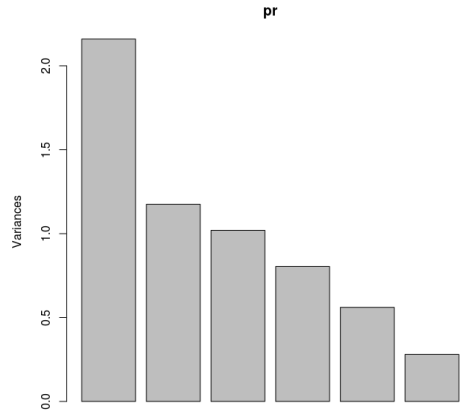
PCA

The first 4 PCs have 86% of the variance, and the first 5 PCs have 95%:



PCA

The first 4 PCs have 86% of the variance, and the first 5 PCs have 95%:



Model:

$$\log y \sim PC_1 + PC_2 + \ldots + PC_5$$

Cross Validation

Goal: **Compare models within the given data**

Assume we have n ($n \gg 1$) observations, and would like to validate a model.

- Partition the data into two subests;
- Choose one subset (called training set), fit the model based on it;
- Validating the model on the other subset (called testing set).

In our computation, we performed log transformation on y, compared the corresponding prediction variance, and found the most suitable model.

Cross validation results

We used 11 folds to compute the four models above.

Cross validation results

We used 11 folds to compute the four models above.

1. Partition the data into 11 groups randomly;

Cross validation results

We used 11 folds to compute the four models above.

1. Partition the data into 11 groups randomly;
2.
   - Choose testing set = Group 1 and training set = the others;

Cross validation results

We used 11 folds to compute the four models above.

1. Partition the data into 11 groups randomly;
2.
   - Choose testing set = Group 1 and training set = the others;
   - Fit the model on the training set;

Cross validation results

We used 11 folds to compute the four models above.

1. Partition the data into 11 groups randomly;
2.
   - Choose testing set = Group 1 and training set = the others;
   - Fit the model on the training set;
   - Run on the testing set;

Cross validation results

We used 11 folds to compute the four models above.

1. Partition the data into 11 groups randomly;
2.
   - Choose testing set = Group 1 and training set = the others;
   - Fit the model on the training set;
   - Run on the testing set;
   - Calculate the sum of squared errors (SSE);

Cross validation results

We used 11 folds to compute the four models above.

1. Partition the data into 11 groups randomly;
2.
   - Choose testing set = Group 1 and training set = the others;
   - Fit the model on the training set;
   - Run on the testing set;
   - Calculate the sum of squared errors (SSE);
3. Repeat Step 2 with testing set = Group 2, 3, . . . , 11.

Cross validation results

We used 11 folds to compute the four models above.

1. Partition the data into 11 groups randomly;
2.
   - Choose testing set = Group 1 and training set = the others;
   - Fit the model on the training set;
   - Run on the testing set;
   - Calculate the sum of squared errors (SSE);
3. Repeat Step 2 with testing set = Group 2, 3, . . . , 11.

Cross validation score = Sum of 11 SSEs.

Cross validation results

Compare all models:

The lower score, the better model!

Cross validation results

Compare all models:

The lower score, the better model!

Model	Cross validation score
Linear	1.361803
Local fit	1.415833
PCA	4.818902
Random forest	0.2460014

Cross validation results

Compare all models:

The lower score, the better model!

Model	Cross validation score
Linear	1.361803
Local fit	1.415833
PCA	4.818902
Random forest	0.2460014

Random forest won!

Final result

Prediction given by random forest:

City	Year	Median	Low	High
Warwick	2020	254.0	168.0	475.7
Warwick	2040	260.0	168.0	489.2
Monroe	2020	241.0	159.0	368.8
Monroe	2040	238.0	162.0	355.6
Tuxedo	2020	464.7	275.0	795.0
Tuxedo	2040	444.2	290.1	637.8

Final result

Prediction given by random forest:

City	Year	Median	Low	High
Warwick	2020	254.0	168.0	475.7
Warwick	2040	260.0	168.0	489.2
Monroe	2020	241.0	159.0	368.8
Monroe	2040	238.0	162.0	355.6
Tuxedo	2020	464.7	275.0	795.0
Tuxedo	2040	444.2	290.1	637.8

In above:

- We exponentiated back to the normal scale;
- Medians were used in the prediction, not averages;
- 80% confidence intervals were computed.

Summary

Goal: Predicting expenditure per person of 3 cities in future.

Summary

Goal: Predicting expenditure per person of 3 cities in future.

Methods considered:

- Linear regression models;
- Local polynomial;
- Random forest;
- Principal component analysis.

Summary

Goal: Predicting expenditure per person of 3 cities in future.

Methods considered:

- Linear regression models;
- Local polynomial;
- Random forest;
- Principal component analysis.

Final model:

Random forest

Summary

Goal: Predicting expenditure per person of 3 cities in future.

Methods considered:

- Linear regression models;
- Local polynomial;
- Random forest;
- Principal component analysis.

Final model:

Random forest

Prediction:

Year	Warwick	Monroe	Tuxedo
2020	254.0	241.0	464.7
2040	260.0	238.0	444.2

Thank you!

(From left: Dan, Bo, Jingnan, Yuting, Qian, Lian)