Data Analytics

Yong Zheng

Illinois Institute of Technology
Chicago, IL, 60616, USA

School of Applied Technology
ILLINOIS INSTITUTE OF TECHNOLOGY

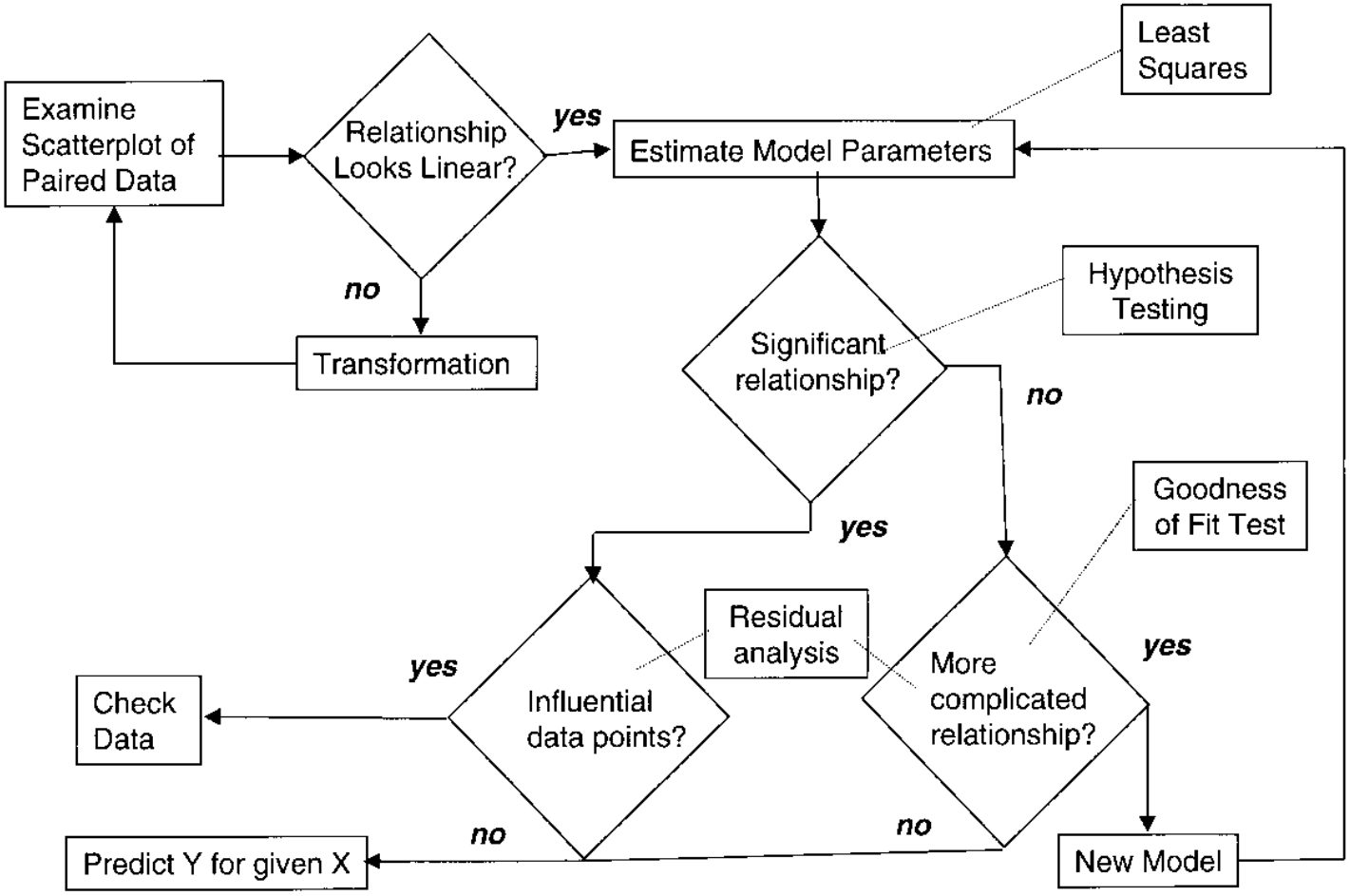
Schedule

- Linear Regression: Workflow
- Issue of Missing Value
- In-Class Practice

Schedule

- Linear Regression: Workflow
- Issue of Missing Value
- In-Class Practice

Multiple Linear Regression



11-16

Schedule

- Linear Regression: Workflow
- Issue of Missing Value
- In-Class Practice

Data Cleaning: Missing Values

- Data is not always available (missing attribute values in records)
  - equipment malfunction
  - deleted due to inconsistency or misunderstanding
  - not considered important at time of data gathering
- Solving Missing Data
  - Ignore the record with missing values;
  - Fill in the missing values manually;
  - Fill in the missing values automatically;
    - Use a global constant to fill in missing values (NULL, unknown, etc.);
    - Use the attribute mean value to filling missing values of that attribute;
    - Use the attribute mean for all samples belonging to the same class to fill in the missing values;
    - Infer the most probable value to fill in the missing value
      - may need to use methods such as Bayesian classification or decision trees to automatically infer missing attribute values

Data Cleaning: Missing Values

- Fill in Missing Data if it is numerical variable, Exp: age
  - Use a global constant to fill in missing values
  - Use the attribute mean value to filling missing values of that attribute;
  - Use the attribute mean for all samples belonging to the same class to fill in the missing values;
  - Build a predictive model (e.g., regression model) to predict missing values
- Fill in Missing Data if it is nominal variable, Exp: gender
  - Use a global constant to fill in missing values, e.g., NULL
  - Use the most frequent value to filling missing values of that attribute;
  - Use the most frequent value belonging to the same class to fill in the missing values;
  - Build a predictive model (e.g., classification model) to predict missing values

Data Preprocessing by Using R

- Replace missing values by R

```
##    Country Age Salary Purchased
## 1   France  44  72000        No
## 2    Spain  27  48000       Yes
## 3  Germany  30  54000        No
## 4    Spain  38  61000        No
## 5  Germany  40     NA       Yes
## 6   France  35  58000       Yes
## 7    Spain  NA  52000        No
## 8   France  48  79000       Yes
## 9  Germany  50  83000        No
## 10  France  37  67000       Yes
```

```
dataset$Age <- ifelse(is.na(dataset$Age),
                      ave(dataset$Age, FUN = function(x)
                        mean(x, na.rm = TRUE)),
                      dataset$Age)

dataset$Salary <- ifelse(is.na(dataset$Salary),
                      ave(dataset$Salary, FUN = function(x)
                        mean(x, na.rm = TRUE)),
                      dataset$Salary)
```

Data Preprocessing by Using R

- Replace missing values by R

ifelse

Conditional Element Selection

`ifelse` returns a value with the same shape as `test` which is filled with elements selected from either `yes` or `no` `TRUE` or `FALSE`.

Keywords programming, logic

Usage

```
ifelse(test, yes, no)
```

Arguments

test	an object which can be coerced to logical mode.
yes	return values for true elements of `test`.
no	return values for false elements of `test`.

https://www.rdocumentation.org/packages/base/versions/3.5.2/topics/ifelse

Data Preprocessing by Using R

- Replace missing values by R

ave

Group Averages Over Level Combinations Of Factors

Subsets of `x[]` are averaged, where each subset consist of those observations with the same factor levels.

Keywords univar

Usage

```
ave(x, …, FUN = mean)
```

Arguments

x	A numeric.
...	Grouping variables, typically factors, all of the same `length` as `x`.
FUN	Function to apply for each factor level combination.

https://www.rdocumentation.org/packages/stats/versions/3.5.2/topics/ave

Data PreProcessing

- Convert Nominal Variable to Dummy variables in R

F1	F2	F3	F4	Class
C3	0	0	2	—
C2	1	0	5	+
C1	0	1	8	—
C2	1	1	16	—
C1	1	0	23	+
C3	0	1	11	+

```
install.packages("dummies")
library(dummies)
data=read.table("book1.csv", head=T, sep=',')
df=dummy.data.frame(data,names=c("F1"))
```

```
> df=dummy.data.frame(data,names=c("F1"))
>
> df
  F1C1 F1C2 F1C3 F2 F3 F4 Class
1    0    0    1  0  0  2     —
2    0    1    0  1  0  5     +
3    1    0    0  0  1  8     —
4    0    1    0  1  1 16     —
5    1    0    0  1  0 23     +
6    0    0    1  0  1 11     +
```

Note that it will create N dummy variables if there are N values in the nominal variable

Schedule

- Linear Regression: Workflow
- Issue of Missing Value
- In-Class Practice

Where to Find Data

- UCI Data, https://archive.ics.uci.edu/ml/datasets.html
- Kaggle, http://www.kaggle.com

In-Class Practice

- UCI Data: Auto-Mpg Data ➔ Assignment
  https://archive.ics.uci.edu/ml/datasets/auto+mpg
- Kaggle Data: Car Fuel Consumption ➔ In-Class
  https://www.kaggle.com/anderas/car-consume

In-Class Practice

- Kaggle Data: Car Fuel Consumption https://www.kaggle.com/anderas/car-consume
  - Figure out a predictive task. Which variable can be used as y and which ones are x
  - Preprocessing on Excel/csv documents. Should we ignore the variable or fill in missing values. Which way we should use?
  - Build multiple linear regression model and improve them step by step