Data Science

Homework 07

1. 한동 honor code 에 맞게 과제를 진행하여 주세요.
2. 과제의 경우 팀당 1 개의 결과물을 제출하면 됩니다.
3. 과제 제출 기한은 ~5/9 23:59 입니다. (1 분당 0.1 감점)
4. 제출은 LMS>과제 및 평가>Homework07 로 하시면 됩니다. (팀 내 1 명이 제출)
5. LMS 제출이 안되는 경우는 TA 이메일로 제출하시기 바랍니다. (22200479@handong.ac.kr)

1. Please proceed with the assignment following the Handong honor code.
2. For assignments, one submission per team is sufficient.
3. The deadline for assignment submission is until 5/9, 23:59. (0.1 points deducted per minute late)
4. Submissions should be made to LMS>Assignments>Homework07. (One member of the team should submit)
5. If you cannot submit via LMS, please submit to the TA email. (22200479@handong.ac.kr)

모든 학생들은 아래의 링크에 접속하여, 본 과제물에 대한 Peer Evaluation을 진행하여 주시기 바랍니다.
제출시간 마감 이전에 응답하지 않으면 불이익이 있을 수 있습니다.
All students are requested to access the link below and conduct a Peer Evaluation for this assignment.

01분반(KOR) - https://forms.gle/mxHcGcGRrGNrrD5KA
2rd Class(ENG) - https://forms.gle/ZUgXrpCDCs4LNV7x8

Practice 7: Single Variable model for regression

Loading data into R

```
PRSA_data <- read.csv("PRSA_data.csv")
```

```
## Rows: 43,824
## Columns: 13
## $ No    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18...
## $ year  <int> 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2...
## $ month <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ day   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ hour  <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,...
## $ pm2.5 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ DEWP  <int> -21, -21, -21, -21, -20, -19, -19, -19, -19, -20, -19, -18, -...
## $ TEMP  <dbl> -11, -12, -11, -14, -12, -10, -9, -9, -9, -8, -7, -5, -5, -3,...
## $ PRES  <dbl> 1021, 1020, 1019, 1019, 1018, 1017, 1017, 1017, 1017, 1017, 1...
## $ cbwd  <chr> "NW", "NW", "NW", "NW", "NW", "NW", "NW", "NW", "NW", "NW", "...
## $ Iws   <dbl> 1.79, 4.92, 6.71, 9.84, 12.97, 16.10, 19.23, 21.02, 24.15, 27...
## $ Is    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Ir    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
```

PRSA dataset: 2010년 1월1일부터 2014년 12월 31일까지의 중국 베이징의 미세먼지 농도 및 날씨 정보

PRSA dataset: Fine dust concentration level and weather record of Beijing China from 2010 - Jan-01 ~ 2014-Dec-31

variable	description
No	Row Index
year	관측연도
month	관측월
day	관측일
hour	관측시간 0h~23h
pm2.5	미세먼지 농도 Fine dust concentration (ug/m^3)
DEWP	Dew Point (이슬점)
TEMP	Temperature (기온)
PRES	Air pressure (기압) hPa
cbwd	Wind Direction (풍향)
Iws	Cumulated wind speed (m/s) (누적 풍속)
Is	Snowfall per hour (시간당 누적 강설량)
Ir	precipitation per hour(시간당 누적 강수량)

Question 1

미세먼지 농도(**pm2.5**)를 예측하는 Single variable Regression 모델을 만들어보려고 한다.
가장 먼저 전체 데이터를 train과 test 용도로 분할한다.

- 2010년부터 2013년 데이터를 train 데이터로 하고, 2014년 데이터를 test 데이터로 분할하여라.
- 그리고 목적 변수인 pm2.5 값에 NA인 것이 있다면 삭제하고 필요한 전처리 과정이 있다면 수행하여라.
- train 데이터와 test 데이터의 sample 수는 어떻게 나누어졌으며 비율은 어떠한가?
- train 데이터의 pm2.5 값과 test 데이터의 pm2.5 값의 분포(평균, 분산)를 비교하여 보고 비슷한지 확인하여라.

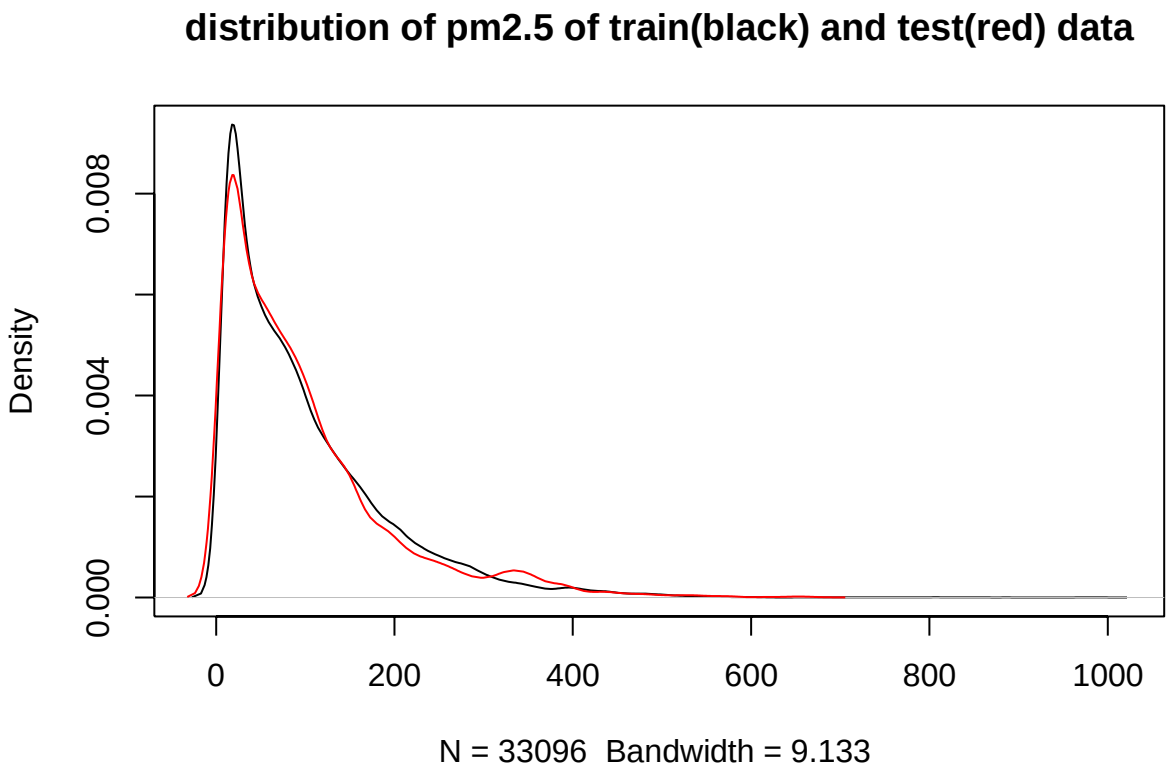
We build a prediction model to predict the fine dust concentration(**pm2.5**) using single variable.
First, we partition the dataset into two, one for training model and one for testing.

- Use the dataset from year of 2010 to 2013 for train, the rest of year 2014 for testing.
- If we have missing values in our target variable **pm2.5**, remove those NAs and perform any necessary data-preparation.
- What are the ratio of the number of samples for training and testing?
- compare the distribution of variable **pm2.5** in training and testing dataset in terms of mean and variance. do you find their distribution are almost identical?

distribution of pm2.5 in both datasets

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.         var
##     0.00000    29.00000    73.00000    98.84315   138.00000   994.00000  8401.34863
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.         var
##     2.00000    28.00000    72.00000    97.73456   133.00000   671.00000  8748.14674
```

*참고사진(필수X)


distribution of pm2.5 of train(black) and test(red) data
Density
0.000
0.004
0.008
0
200
400
600
800
1000
N = 33096 Bandwidth = 9.133

Question 2-1

month를 사용하여 pm2.5를 예측하는 단일 변수 모델(single variable model)을 만들어라.
예측한 pm2.5 값과 실제 값과의 차이를 error(residual)로 데이터에 추가하여라
Build a single variable model to predict **pm2.5** using the variable **month.**
add predicted pm2.5 and their error (residual = actual value - predicted value) to the dataset.

Sample Result

##	month	pm2.5	pred	error
## 1	1	129	114.1159	14.884053
## 2	1	148	114.1159	33.884053
## 3	1	159	114.1159	44.884053
## 4	1	181	114.1159	66.884053
## 5	1	138	114.1159	23.884053
## 6	1	109	114.1159	-5.115947
## 7	1	105	114.1159	-9.115947
## 8	1	124	114.1159	9.884053
## 9	1	120	114.1159	5.884053
## 10	1	132	114.1159	17.884053

Question 2-2

Question 2-1에서 구한 모델의 MSE와 RMSE 구하라
이 모델을 test data에도 적용하여 MSE와 RMSE를 구하라
Find the MSE and RMSE for the prediction model we found in Question 2-1.
Find the MSE and RMSE for the test dataset as well as the training dataset.

sample result

[1] "train data: (MSE 8254.242) (RMSE 90.853)"

[1] "test data : (MSE 8397.312) (RMSE 91.637)"

Question 2-3

Question 2-1에서 구한 모델을 적용하여 train data와 test data의 R^2 값을 구하고,
그것을 바탕으로 만들어진 단일변수 모델이 pm2.5의 변동을 얼마나 잘 설명하고 있는지 이야기해보자.
Find the R^2 of model from from Question 2-1 for both train and test dataset.
Explain how well your model predicts the variance Of pm2.5 using R^2.

***R*² for train dataset and test dataset**

[1] "R2 for train data: 0.017"

[1] "R2 for test data: 0.040"

Question 3

hour 변수를 사용해서 Question 2번의 과정을 반복하라.

hour를 어떤 구간으로 나누어서 모델을 만드는 것이 효과적인가?

Repeat the question 2-1 ~ 2-3 using a variable of **hour.**

What would be the best way to categorize **hour** to train prediction model?

Question 4

동일한 과정을 DEWP 변수를 사용해서 수행하라.

Repeat the question 2-1 ~ 2-3 using variable of **DEWP**

Question 5

위에서 시도한 다양한 단일 변수 모델 중 어떤 모델이 가장 예측 성능이 뛰어난가?

예측 성능이 높은 이유가 무엇인지 생각해보자.

Among those attempts above, which model was the best to predict **pm2.5**?

State your idea why the model outperforms others.