UNITED INTERNATIONAL COLLEGE

REGRESSION ANALYSIS USING R

Beijing PM2.5 Data Analysis

Author:
Junjie LIU
Daichen YAO
Yun YANG

Supervisor:
Dr.YeHUA JUN

Group S
Division of Science and Technology

December 13, 2018

Declaration of Authorship

I, Junjie LIU
Daichen YAO
Yun YANG, declare that this thesis titled, “Beijing PM2.5 Data Analysis” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Junjie LIU Daichen YAO Yun YANG

Date: December 13, 2018

UNITED INTERNATIONAL COLLEGE

Abstract

Statistics
Division of Science and Technology

Bachelor

Beijing PM2.5 Data Analysis

by Junjie LIU
Daichen YAO
Yun YANG

I would like to dedicate this dissertation to our teachers ...

Acknowledgements

We would like to dedicate this dissertation to our teachers, Dr.Ye Hua Jun, Dr.He Ping, also, our Teacher Asistant Jenna, without their helps, we would not able finish this report. Also, Thanks to the Statistics Labtoray, which provides the computers…

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Data Set Selection

In the very beginning, we have been shown the data set website UCI Data Sets, we selected the **"Beijing PM2.5 Data Data Set"** which was donated by Prof.Song Xi Chen, from Guanghua School of Management, Center for Statistical Science, Peking University(Liang et al., 2015), for Beijing is known around the world by severe air pollution. The most widely used method to measure air pollution is PM2.5, the particulate matter smaller than or equal to 2.5 microns in diameter in the air, which is a kind of air pollutant that people and our daily life pay great attention to. Although this data set is time-dependent, as long as the data pre-processing is reasonable, we can still treat the data at each time point as independent and unrelated data for regression analysis.

1.2 Data Sets Introduction

In the original data sets, we can found that it contains almost every hour's PM2.5, Dew Point, Temperature, Pressure, Combined wind direction, Cumulative wind speed, Cumulative hours of snow and Cumulative hours of rain data, and in this project, our purposes are to determine the major factors responsible for this pollution, discuss whether these factors have been targeted by recent initiatives, and predict future PM2.5 levels.

1.3 Multiple Regression Estimation

In statistical modeling, regression model is one of the most important models. It is estimating the relationship between dependent variable and independent variables by the observation data. Also, the parameters are unbiased.(Montgomery, Peck, and Vining, 2012).

Regression model involve the following parameters and variables:

- **unknown parameters**, denoted as β, which may represented as a vector;
- **independent variables**, denoted as $\mathbf{X}$, which is represented as a matrix;
- **dependent variable**, denoted as $\mathbf{Y}$, which is represented as a vector.

The **Ordinary Multiple Regression Model** is

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

We think this kind of estimation may suitable for our data.