2025 秋季《普通统计学》期末报告

2300010816　　谭雪贻　　数学科学学院

2025 年 12 月 6 日

代码、LaTeX 文件以及图片可见https://github.com/Rainco-S/Statistic-Final

1　数据的处理和可视化

1.1　变量类型与初步观察

- 数值型变量：PM2.5、PM10、SO2、CO、NO2、O3、TEMP（气温）、DEWP（露点温度）、HUMI（相对湿度）、PRES（气压）、WSPM（风速）、year、month、day、hour；
- 分类型变量：wd（风向，分类变量）。

数值型变量均为连续观测值，分类型变量 wd 包含多个风向类别（SE、CV、SW 等）。

```
library(MASS)
library(lmtest)
library(lubridate)
library(ggplot2)
library(gganimate)
library(forecast)
library(tseries)
library(tidyverse)
library(imputeTS)
library(randomForest)
library(zoo)
library(dplyr)
library(purrr)
library(lubridate)
library(car)
library(forecast)
library(corrplot)

path <- '~' # 替换为你的数据文件路径
data <- read.csv(file.path(path,"/Beijing_Wanliu_data.csv"), stringsAsFactors = FALSE)

str(data)
```

```
'data.frame': 8760 obs. of  16 variables:
$ year : int  2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
```

```
$ month: int  1 1 1 1 1 1 1 1 1 ...
$ day  : int  1 1 1 1 1 1 1 1 1 ...
$ hour : int  0 1 2 3 4 5 6 7 8 9 ...
$ PM2.5: num  57 68 81 95 95 89 95 88 69 67 ...
$ PM10 : num  152 179 240 229 206 234 253 207 169 169 ...
$ SO2  : num  16 19 27 40 51 34 32 30 28 29 ...
$ CO   : num  1.8 1.8 2.6 3 2.5 3.2 3.5 4.4 4 3.8 ...
$ NO2  : num  110 116 112 110 110 113 115 112 117 119 ...
$ O3   : num  4 NA NA NA NA NA NA NA NA NA ...
$ TEMP : num  -1.5 -2.6 -3 -3.3 -2.7 -3.1 -2.6 -2.9 0 9 ...
$ DEWP : num  -12.5 -12.1 -11.2 -11.1 -10.5 ...
$ HUMI : int  43 48 53 55 55 57 53 57 53 22 ...
$ PRES : num  1007 1007 1007 1006 1006 ...
$ wd   : chr  "SE" "CV" "SW" "SW" ...
$ WSPM : num  0.6 0.2 0.6 0.6 1.8 1 1 1.1 1.5 1.9 ...
```

1.2 数据导入与缺失值处理

变量	year	month	day	hour	PM2.5	PM10	SO2	CO	NO2	O3	TEMP
缺失比例	0.000	0.000	0.000	0.000	1.062	0.788	2.705	1.438	1.998	7.614	0.000
变量	DEWP	HUMI	PRES	wd	WSPM						
缺失比例	0.000	0.000	0.000	0.000	0.000						

```r
# 计算缺失比例
missing_ratio <- sapply(data, function(x) mean(is.na(x)) * 100)
missing_df <- data.frame(缺失比例 = round(missing_ratio, 3))
print(missing_df)

# 缺失值处理
numeric_vars <- c("PM2.5", "PM10", "SO2", "CO", "NO2", "O3", "TEMP", "DEWP", "HUMI", "
    PRES", "WSPM", "year", "month", "day", "hour")
categorical_vars <- c("wd")
# 数值型变量: 线性插值
data_numeric <- data[, numeric_vars]
data_numeric_imputed <- as.data.frame(
apply(data_numeric, 2, function(col) {
    na.interp(col)
})
)

# 分类型变量: 最近邻填充
data_categorical <- data[, categorical_vars, drop = FALSE]
data_categorical_imputed <- na_locf(data_categorical)  # 向前填充
data_categorical_imputed <- na_locf(data_categorical_imputed, option = "locf_back")  #
    向后填充

# 合并处理后的数据
data_imputed <- cbind(data_numeric_imputed, data_categorical_imputed)
```

```
          缺失比例
```

```
    year      0.000
    month     0.000
    day       0.000
    hour      0.000
    PM2.5     1.062
    PM10      0.788
    SO2       2.705
    CO        1.438
    NO2       1.998
    O3        7.614
    TEMP      0.000
    DEWP      0.000
    HUMI      0.000
    PRES      0.000
    wd        0.000
    WSPM      0.000
```

1.3 可视化

1.3.1 PM2.5 浓度频数分布直方图

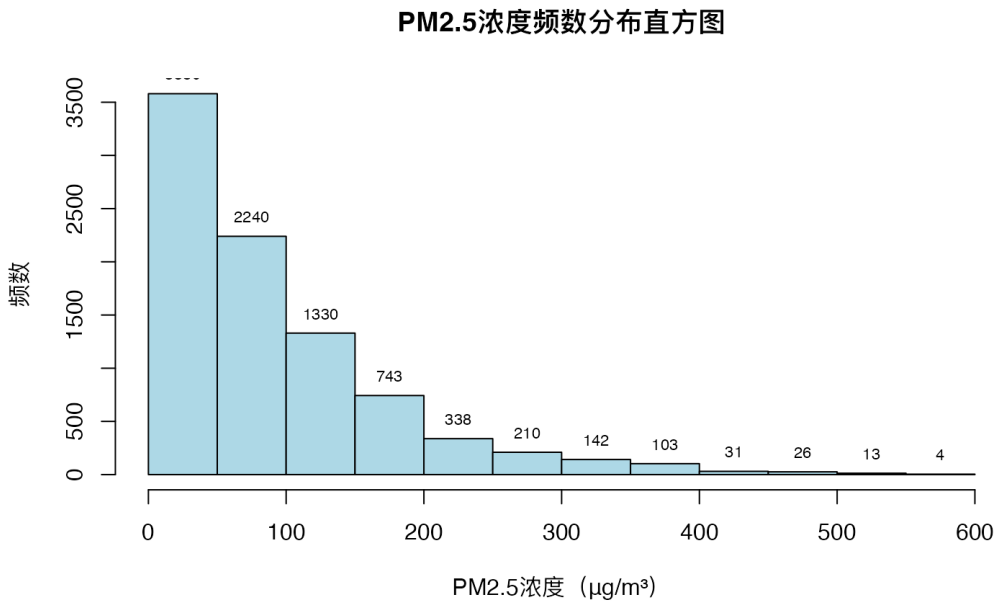


图 1: PM2.5 浓度频数分布直方图

PM2.5 浓度呈右偏分布（多数时段浓度集中在 $0-150g/m^3$，少数时段出现高浓度污染）；频数峰值集中在 $50-100g/m^3$ 区间，对应“轻度-中度污染”，2014 年空气质量以轻度-中度污染为主；高浓度污染为少数极端值，可能是突发污染物排放。

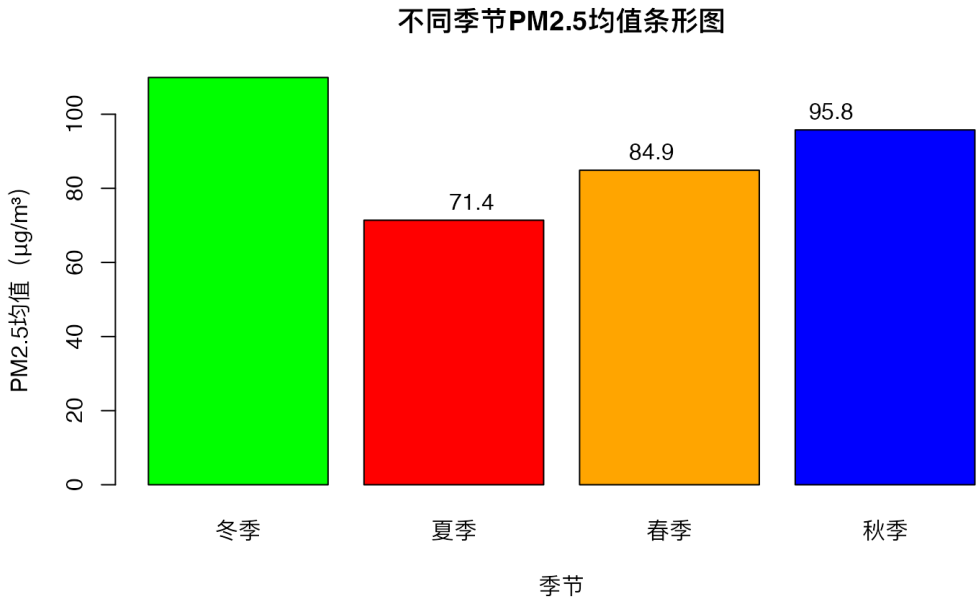


图 2: 不同季节 PM2.5 均值条形图

1.3.2 不同季节 PM2.5 均值条形图

PM2.5 季节差异显著——冬季最高（约 $110g/m^3$），夏季最低（约 $70g/m^3$），春季和秋季介于两者之间（约 $80-100g/m^3$），组间差距明显。可能因为冬季供暖排放增加、气象扩散条件差，导致污染加重；夏季降水多、风速大，污染物易扩散，浓度最低。

1.3.3 PM2.5 污染等级占比饼图

“优良”占比最高（约 31.3%），“重度污染”占 28.2%，“轻度污染”仅占 23.9%，“中度污染”占 16.6%；污染等级呈“中间低、两端高”。2014 年该站点仅 1/5 时段空气质量达标，近 50% 时段处于“中度-重度污染”。

1.3.4 季节 ×PM2.5 污染等级列联表（马赛克图）

冬季“重度污染”矩形面积最大，夏季“优良”矩形面积最大，春季和秋季以“轻度污染、优良”为主，呈现明显的“季节-污染等级”关联。

	污染等级	优良	轻度污染	重度污染
冬季	316	751	265	828
夏季	361	718	681	448
春季	427	602	607	572
秋季	349	671	540	624

1.3.5 PM2.5 浓度与气温散点图

PM2.5 浓度与气温呈弱负相关（趋势线斜率为负）；气温低于 $0°C$ 时，PM2.5 高浓度点（$>150g/m^3$）明显增多；气温高于 $20°C$ 时，浓度多集中在 $0-200g/m^3$，高浓度点极少。

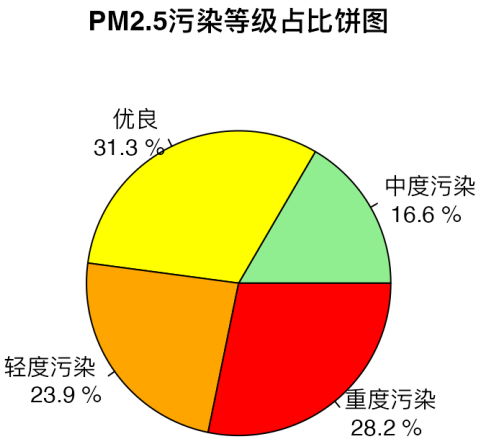


图 3: PM2.5 污染等级占比饼图

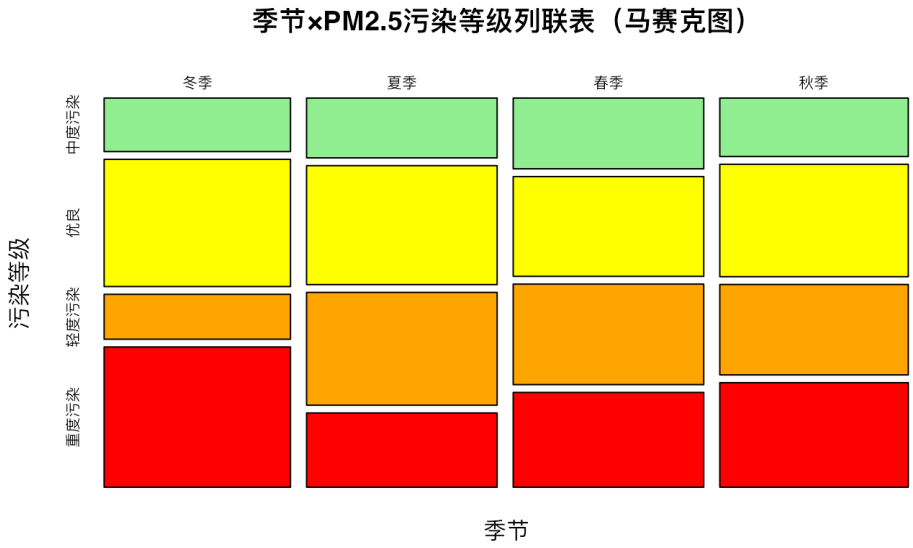


图 4: 季节 ×PM2.5 污染等级列联表（马赛克图）

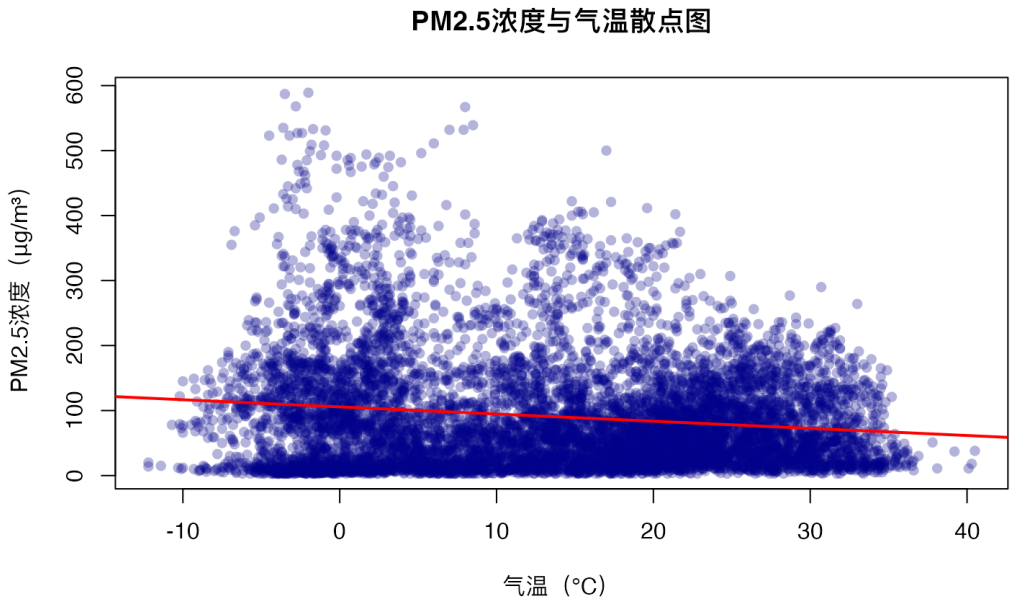


图 5: PM2.5 浓度与气温散点图

1.3.6 PM2.5 日均值时间序列图

PM2.5 日均值呈现明显季节性差异；整体波动较大，存在多个日均值 $> 200g/m^3$ 的异常峰值。

1.3.7 风向-PM2.5 均值主次图

NE 的 PM2.5 均值最高，是污染最严重风向；其次是 CV、SE；NW 的 PM2.5 均值最低，是污染最轻风向。NE 出现频次较高，该风向是常见风向之一；SE 出现频次最低。可见高污染风向的影响权重高，低污染风向的调节作用。

```
    data_imputed <- data_imputed %>%
    mutate(
        season = case_when(month %in% 3:5 ~ "春季",
                           month %in% 6:8 ~ "夏季",
                           month %in% 9:11 ~ "秋季",
                           TRUE ~ "冬季"),
        period = ifelse(hour %in% 6:18, "day", "night"),
        pm25level = case_when(PM2.5 <= 35 ~ "优良",
                              PM2.5 <= 75 ~ "轻度污染",
                              PM2.5 <= 115 ~ "中度污染",
                              TRUE ~ "重度污染"),
        date = as.Date(paste(year, month, day, sep = "-")),
    )
    pm25_daily <- data_imputed %>% group_by(date) %>% summarise(PM2.5 = mean(PM2.5))

    # PM2.5浓度分布直方图/频数分布直方图
    hist(data_imputed$PM2.5,
        main = "PM2.5浓度频数分布直方图",
```

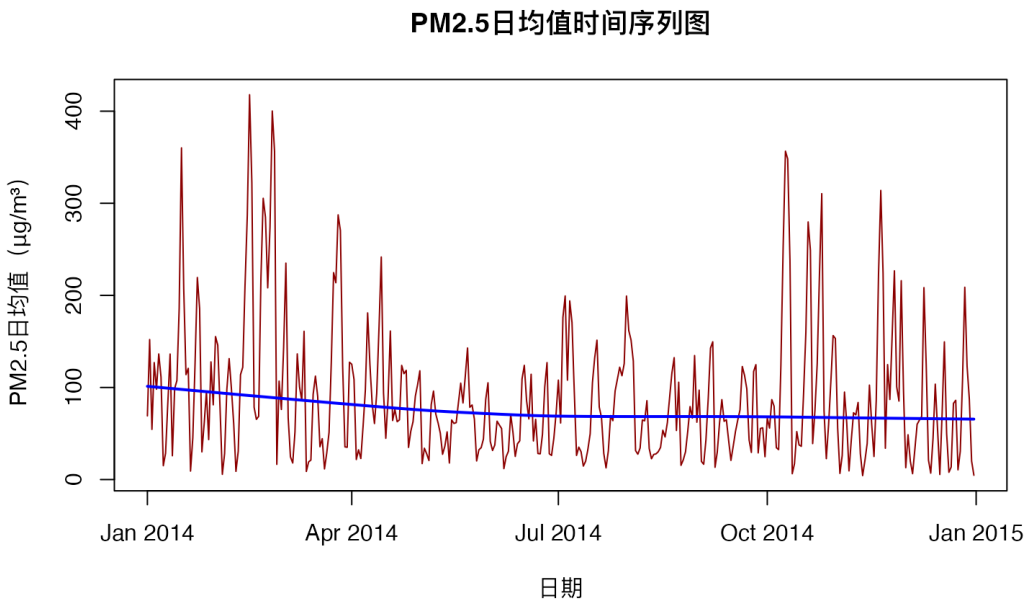


图 6: PM2.5 日均值时间序列图

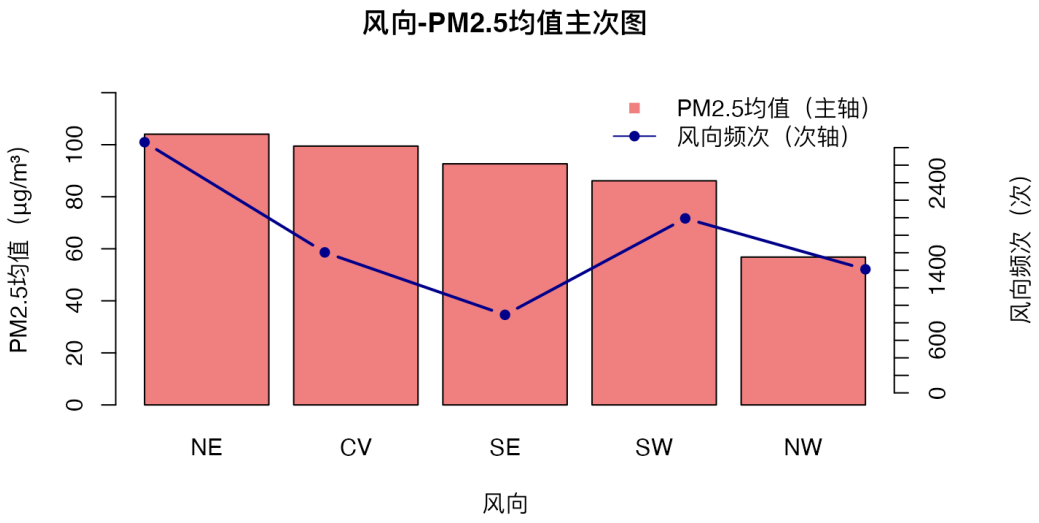


图 7: 风向-PM2.5 均值主次图

```
        xlab = "PM2.5浓度（g/m³）",
        ylab = "频数",
        col = "lightblue",
        breaks = 20)
    hist_info <- hist(data_imputed$PM2.5, plot = FALSE, breaks = 20)
    text(hist_info$mids, hist_info$counts, labels = hist_info$counts,
        pos = 3, cex = 0.7)

    # 不同季节PM2.5均值对比条形图
    season_pm25 <- data_imputed %>%
    group_by(season) %>%
    summarise(pm25_mean = mean(PM2.5), .groups = "drop")
    barplot(season_pm25$pm25_mean,
            names.arg = season_pm25$season,
            main = "不同季节PM2.5均值条形图",
            xlab = "季节",
            ylab = "PM2.5均值（g/m³）",
            col = c("green", "red", "orange", "blue"))
    text(1:4, season_pm25$pm25_mean + 5,
        labels = round(season_pm25$pm25_mean, 1))

    # PM2.5污染等级占比饼图
    pm25_level_count <- table(data_imputed$pm25_level)
    pie(pm25_level_count,
        labels = paste(names(pm25_level_count), "\n", round(pm25_level_count/sum(pm25_level_
            count)*100, 1), "%"),
        main = "PM2.5污染等级占比饼图",
        col = c("lightgreen", "yellow", "orange", "red"))

    # 季节×污染等级交叉分布列联表
    season_level_table <- table(data_imputed$season, data_imputed$pm25_level)
    print(season_level_table)
    mosaicplot(season_level_table,
            main = "季节×PM2.5污染等级列联表（马赛克图）",
            xlab = "季节",
            ylab = "污染等级",
            col = c("lightgreen", "yellow", "orange", "red"))

    # PM2.5与气温的相关性散点图
    plot(data_imputed$TEMP, data_imputed$PM2.5,
        main = "PM2.5浓度与气温散点图",
        xlab = "气温（℃）",
        ylab = "PM2.5浓度（g/m³）",
        col = alpha("darkblue", 0.3),
        pch = 16)
    abline(lm(PM2.5 ~ TEMP, data_imputed), col = "red", lwd = 2)

    # PM2.5日均值时间趋势时间序列图
```

```r
pm25_daily <- data_imputed %>%
group_by(date) %>%
summarise(pm25_daily = mean(PM2.5), .groups = "drop")
plot(pm25_daily$date, pm25_daily$pm25_daily,
    type = "l",
    main = "PM2.5日均值时间序列图",
    xlab = "日期",
    ylab = "PM2.5日均值（g/m³）",
    col = "darkred",
    lwd = 1)
lines(loess.smooth(pm25_daily$date, pm25_daily$pm25_daily), col = "blue", lwd = 2)

# 风向主次图（PM2.5均值主图+频次次轴）
wd_stats <- data_imputed %>%
group_by(wd) %>%
summarise(pm25_mean = mean(PM2.5),
            wd_count = n(),
            .groups = "drop") %>%
arrange(desc(pm25_mean)) %>%
head(8)

par(mar = c(5, 4, 4, 6) + 0.1)
bp <- barplot(
wd_stats$pm25_mean,
names.arg = wd_stats$wd,
main = "风向-PM2.5均值主次图",
xlab = "风向",
ylab = "PM2.5均值（g/m³）",
col = "lightcoral",
ylim = c(0, max(wd_stats$pm25_mean) * 1.2)  # 主轴范围留余量
)

par(new = TRUE)
plot(
x = bp,  # 用条形的x轴坐标，确保折线与条形对齐
y = wd_stats$wd_count,
type = "b",  # 加圆点的折线，更清晰
col = "darkblue",
lwd = 2,
pch = 16,
axes = FALSE,
xlab = "", ylab = "",
ylim = c(0, max(wd_stats$wd_count) * 1.2)  # 次轴范围匹配频次最大值
)

axis(
side = 4,
at = seq(0, max(wd_stats$wd_count), by = 200),  # 刻度间隔根据你的数据调整
```

```
    labels = seq(0, max(wd_stats$wd_count), by = 200),
    family = "PingFang"
    )
    mtext("风向频次（次）", side = 4, line = 4)  # 次轴标签右移，避免重叠

    legend(
    "topright",
    legend = c("PM2.5均值（主轴）", "风向频次（次轴）"),
    col = c("lightcoral", "darkblue"),
    pch = c(15, 16),  # 条形对应方块，折线对应圆点
    lty = c(NA, 1),   # 条形无线条，折线有线条
    bty = "n"  # 去掉图例边框，更美观
    )
```

1.4 描述性统计与汇总表格

1.4.1 数值型变量

完整结果见 log，此处省略年月日小时等无用数据。

变量	均值	中位数	方差	标准差
PM2.5	90.4	66	7343	85.7
PM10	132	114	10342	102
SO2	25.1	13	884	29.7
CO	1.38	0.9	1.67	1.29
NO2	75.7	71	1767	42.0
O3	40.1	20	2543	50.4
TEMP	13.9	14.9	127	11.3
HUMI	59.5	60	242	15.6
PRES	1013	1013	12	3.5
WSPM	1.7	1.5	1.5	1.2

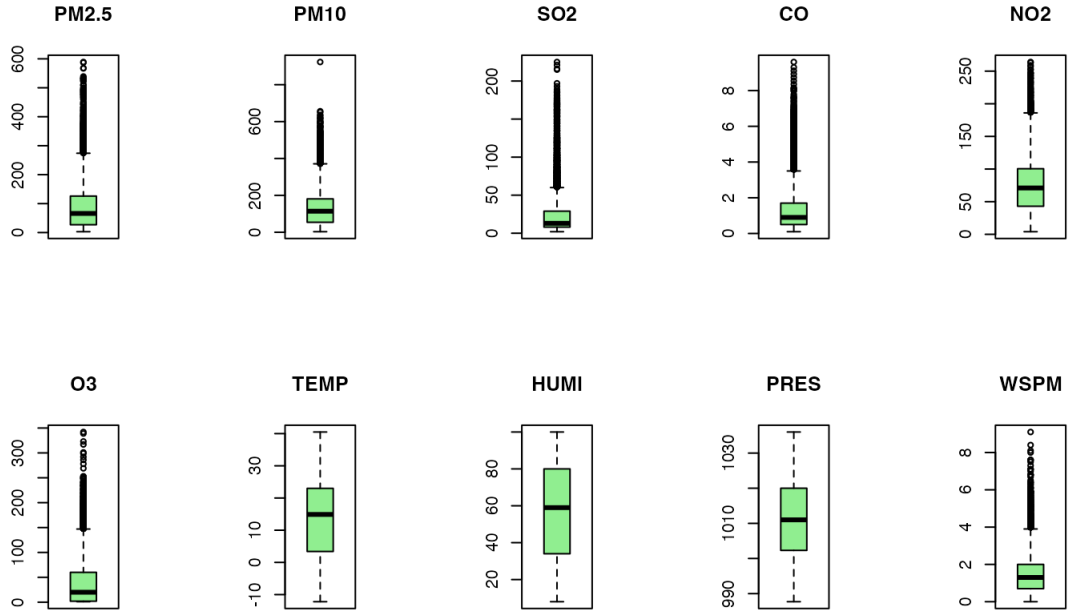


图 8: PM2.5 浓度与数值型变量箱线图

1.4.2 分类型变量

变量	类别	频数	比例（%）	众数
wd	CV	1604	18.31	NE
	NE	2862	32.67	
	NW	1411	16.11	
	SE	891	10.17	
	SW	1992	22.74	
season	冬季	2160	24.66	夏季
	夏季	2208	25.21	
	春季	2208	25.21	
	秋季	2184	24.93	
period	day	4745	54.17	day
	night	4015	45.83	
pm25level	中度污染	1453	16.59	优良
	优良	2742	31.30	
	轻度污染	2093	23.89	
	重度污染	2472	28.22	

```
# 数值型变量描述性统计
numeric_stats <- data_imputed %>%
summarise(across(all_of(numeric_vars),
                 list(均值 = mean,
                      中位数 = median,
```

```
                            方差 = var,
                            标准差 = sd),
                    .names = "{.col}_{.fn}")) %>%
    pivot_longer(everything(),
                 names_to = c("变量", "统计量"),
                 names_sep = "_") %>%
    pivot_wider(names_from = 统计量, values_from = value) %>%
    mutate(across(where(is.numeric), \(x) round(x, 2)))

    print(numeric_stats)

    # 数值型变量箱线图
    par(mfrow = c(2, 5))
    for (var in c("PM2.5", "PM10", "SO2", "CO", "NO2", "O3", "TEMP", "HUMI", "PRES", "WSPM")
        ) {
    boxplot(data_imputed[[var]], main = var, col = "lightgreen")
    }

    # 分类型变量汇总
    categorical_vars_new <- c("wd", "season", "period", "pm25level")

    categorical_stats <- map_dfr(categorical_vars_new, function(var) {
    freq <- as.vector(table(data_imputed[[var]]))           # 转成纯向量
    cat_names <- names(table(data_imputed[[var]]))         # 类别名
    prop <- round(prop.table(table(data_imputed[[var]])) * 100, 2)
    mode_val <- cat_names[which.max(freq)]

    tibble(
        变量 = var,
        类别 = cat_names,
        频数 = freq,
        比例 = prop,
        众数 = mode_val
    )
    })

    print(categorical_stats)
```

```
# A tibble: 15 × 5
变量      均值 中位数      方差 标准差
<chr>   <dbl>  <dbl>    <dbl>  <dbl>
1 PM2.5   90.4     66    7343.    85.7
2 PM10   132.     114   10342.   102.
3 SO2     25.1     13     884     29.7
4 CO       1.38     0.9     1.67   1.29
5 NO2     75.7     71    1767.    42.0
6 O3      40.1     20    2543.    50.4
7 TEMP    13.9     14.9   127.    11.3
8 DEWP     4.17    5.6    180.    13.4
9 HUMI    58       59     683.    26.1
```

```
10 PRES  1011.   1011     103.    10.1
11 WSPM     1.48    1.3     1.32   1.15
12 year  2014    2014       0      0
13 month    6.53    7      11.9    3.45
14 day     15.7    16      77.4    8.8
15 hour    11.5    11.5    47.9    6.92

# A tibble: 15 × 5
变量       类别       频数 比例         众数
<chr>     <chr>    <int> <table[1d]> <chr>
1 wd        CV        1604 18.31       NE
2 wd        NE        2862 32.67       NE
3 wd        NW        1411 16.11       NE
4 wd        SE         891 10.17       NE
5 wd        SW        1992 22.74       NE
6 season    冬季      2160 24.66       夏季
7 season    夏季      2208 25.21       夏季
8 season    春季      2208 25.21       夏季
9 season    秋季      2184 24.93       夏季
10 period    day       4745 54.17       day
11 period    night     4015 45.83       day
12 pm25level 中度污染  1453 16.59       优良
13 pm25level 优良      2742 31.30       优良
14 pm25level 轻度污染  2093 23.89       优良
15 pm25level 重度污染  2472 28.22       优良
```

1.5 构造新变量

在可视化部分已经做过了新变量创建。

```r
# 验证新变量
table(data_imputed$season)
table(data_imputed$period)
```

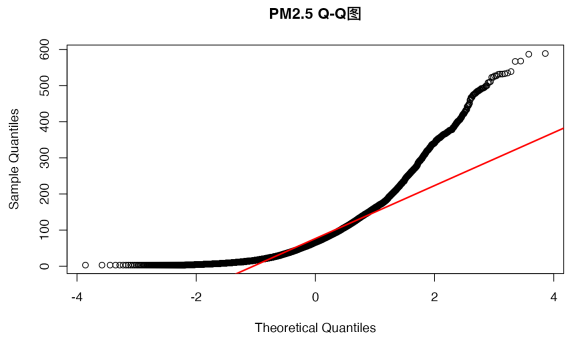
```
冬季 夏季 春季 秋季
2160 2208 2208 2184

day night
4745  4015
```

2 假设检验与统计推断

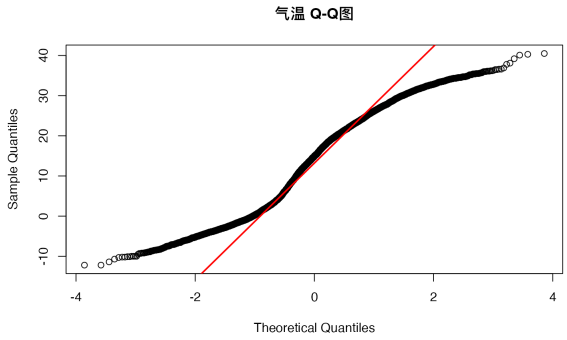
2.1 正态性检验与 Q-Q 图

- PM2.5：偏离正态分布最明显（曲线与对角线偏离程度最大）。左侧低理论分位数长期贴近 0；右侧高理论分位数急剧上升，差距扩大速度快，故 PM2.5 分布完全不符合正态分布连续对称特征。
- 气温：偏离正态分布程度若于 PM2.5。中间区域（理论分位数-2 2）贴近对角线，两端有偏离，但未“急剧脱节”，故气温分布接近正态分布但仍存在偏差。
- 风速：偏离正态分布程度介于 PM2.5 和气温之间。左侧贴近 0，右侧缓慢上升（差距小于 PM2.5），故风速也不符合正态分布，但偏离程度若于 PM2.5。

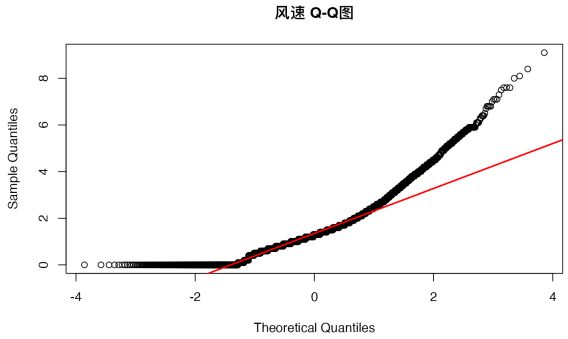
原因分析：PM2.5 实际分布应为“右偏分布”，在多数观测值中集中在较低水平，但存在少数重污染时段（浓度远高于均值），从而拉偏分布，整体呈现明显右偏；气温虽有季节分布，但极端气温偏离程度远小于 PM2.5 重污染极端值；风速通常右偏分布（多数时段风速低，少数大风天气），但强度差异不如 PM2.5 重污染浓度差异显著。



(a) PM2.5 Q-Q 图



(b) 气温 Q-Q 图



(c) 风速 Q-Q 图

图 9: Q-Q 图

```r
# 假设检验与统计推断
# PM2.5 Q-Q图
qqnorm(data_imputed$PM2.5, main = "PM2.5 Q-Q图")
qqline(data_imputed$PM2.5, col = "red", lwd = 2)
# TEMP Q-Q图
qqnorm(data_imputed$TEMP, main = "气温 Q-Q图")
qqline(data_imputed$TEMP, col = "red", lwd = 2)
# WSPM Q-Q图
qqnorm(data_imputed$WSPM, main = "风速 Q-Q图")
qqline(data_imputed$WSPM, col = "red", lwd = 2)
```

2.2 总体方差已知时 PM2.5 均值的置信区间

样本量显然大于 30，可假设近似服从正态分布。使用总体方差 σ^2 已知对总体均值 μ 的估计，其中 $\sigma = 85.69$（等于样本方差），样本均值 $\bar{x} = 90.4$，置信水平为 $95\% = (1-\alpha)\times 100\%$ 的 $\alpha = 0.05$，查表得 $Z_{\alpha/2} = 1.96$，则置信区间为

$$\left[\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = [88.58, 92.17].$$

2014 年北京万柳站点逐小时 PM2.5 平均水平在置信度为 95% 时处于 $88.58-92.17\mu g/m^3$ 之间，远超 $35\mu g/m^3$ 的优良标准，说明该站点全年空气质量整体处于污染较高水平。

```r
# 计算样本统计量
pm25_mean <- mean(data_imputed$PM2.5, na.rm = TRUE)
pm25_sd <- sd(data_imputed$PM2.5, na.rm = TRUE)
n <- nrow(data_imputed)
z_alpha2 <- qnorm(0.975)  # 95%置信水平

# 置信区间
margin_error <- z_alpha2 * pm25_sd / sqrt(n)
ci_known_var <- c(pm25_mean - margin_error, pm25_mean + margin_error)
print(round(ci_known_var, 2))
```

```
[1] 88.58 92.17
```

2.3 总体方差未知时 PM2.5 均值的置信区间

样本量显然大于 30，可假设近似服从正态分布。使用总体方差 σ^2 未知对总体均值 μ 的估计，样本均值 $\bar{x}=109.92$，样本标准差 $s=111.24$，置信水平为 $95\%=(1-\alpha)\times 100\%$ 的 $\alpha=0.05$，查表得 $t_{\alpha/2}=1.96$（样本量足够大，近似正态分布），则置信区间为

$$\left[\bar{x}-t_{\alpha/2}\frac{s}{\sqrt{n}},\bar{x}+t_{\alpha/2}\frac{s}{\sqrt{n}}\right]=[105.22,114.61].$$

使用 t 分布理由：当总体方差 σ^2 未知时，我们通常还是用样本方差 s^2 来代替总体方差 σ^2，但会增加不确定性，故计算中要使用更为分散的分布来代替标准正态分布，即 t 分布。进行这样的代替时，统计量 $\frac{\bar{x}-\mu}{\sqrt{s^2/n}}$ 会渐近服从自由度 $n-1$ 的 t 分布，也即

$$t=\frac{\bar{x}-\mu}{\sqrt{s^2/n}}\sim t_{n-1}.$$

由于 t 分布比标准正态分布更分散，在同样的置信水平下，基于 t 分布的置信区间也比标准正态分布 Z 分布置信区间更宽更保守，但是样本量（自由度）足够大时 t 分布越来越接近标准正态分布，对应置信区间也越来越接近总体方差已知情况下得到的置信区间。

冬季均值（$109.92\mu g/m^3$）比全年均值（$90.4\mu g/m^3$）高，且 95% 置信区间整体都在总体之上，故冬季确实存在更严重的污染水平，可能因为冬季供暖排放问题等。

```r
# 筛选冬季数据
pm25_winter <- data_imputed %>% filter(season == "冬季") %>% pull(PM2.5)

# t置信区间
winter_mean <- mean(pm25_winter)
winter_sd <- sd(pm25_winter)
t_ci <- t.test(pm25_winter)$conf.int
print(round(t_ci, 2))

# 对比全年与冬季均值
cat("全年PM2.5均值: ", round(pm25_mean, 2), "\n")
cat("冬季PM2.5均值: ", round(mean(pm25_winter), 2), "\n")
```

```
[1] 105.22 114.61
attr(,"conf.level")
[1] 0.95
```

2.4 重污染与优良空气比例的区间估计与样本量

全年样本比例 $\hat{p_1} = 0.448, \hat{p_2} = 0.313$，置信水平为 $95\% = (1-\alpha) \times 100\%$ 的 $\alpha = 0.05$，同上查表得 $Z_{\alpha/2} = 1.96$，则重污染与优良空气比例的 95% 置信区间为

$$\left[\hat{p_1} - Z_{\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n}}, \hat{p_1} + Z_{\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n}}\right] = [0.438, 0.458].$$

$$\left[\hat{p_2} - Z_{\alpha/2}\sqrt{\frac{\hat{p_2}(1-\hat{p_2})}{n}}, \hat{p_2} + Z_{\alpha/2}\sqrt{\frac{\hat{p_2}(1-\hat{p_2})}{n}}\right] = [0.303, 0.323].$$

估计总体比例 p 时样本量：在给定容许误差 $E = 0.05$，样本量应不小于

$$n = \hat{p}(1-\hat{p})\left(\frac{Z_{\alpha/2}}{E}\right)^2.$$

从而代入得 $n_1 = 381, n_2 = 331, n_0 = 385$，其中 n_1 为重污染小时，n_2 为优良空气小时，n_0 为保守估计。

样本量公式推导：

定理 1 (中心极限定理)**.** *若 $X_1, X_2, \cdots, X_n$ 是 i.i.d. 的 n 个随机变量，有相同均值 μ 和方差 σ^2，那么 n 增大趋于无穷时，n 个随机变量的均值 $\bar{x}$ 将近似服从均值为 μ，方差为 $\frac{\sigma^2}{n}$ 的正态分布.*

根据中心极限定理，样本来那个足够大时，样本均值的分布为 $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$，对于预先设定的置信水平 $1-\alpha$，由正态分布性质知 $\bar{x}$ 落在区间 $\left[\mu - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ 的概率为 $1-\alpha$，即

$$P\left(\mu - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \bar{x} \le \mu + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

从而样本均值与总体均值之间的差距 $|\bar{x} - \mu|$ 有 $P\left(|\bar{x} - \mu| < Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$，用 E 标记可容许估计误差时，其就是置信区间半径 $Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$，估计总体比例时方差为 $\hat{p_1}(1-\hat{p_1})$，故样本量最小值计算公式为 $n = \hat{p}(1-\hat{p})\left(\frac{Z_{\alpha/2}}{E}\right)^2$。

```r
# 比例区间估计
p1_hat <- mean(data_imputed$PM2.5 > 75, na.rm = TRUE)
p2_hat <- mean(data_imputed$PM2.5 <= 35, na.rm = TRUE)

# 置信区间
ci_p1 <- p1_hat + c(-1, 1) * z_alpha2 * sqrt(p1_hat*(1-p1_hat)/n)
ci_p2 <- p2_hat + c(-1, 1) * z_alpha2 * sqrt(p2_hat*(1-p2_hat)/n)

cat("重污染比例p : ", round(p1_hat*100, 1), "%, 95%置信区间: ", round(ci_p1*100, 1), "\n
    ")
cat("优良空气比例p : ", round(p2_hat*100, 1), "%, 95%置信区间: ", round(ci_p2*100, 1), "
    \n")

```

```r
# 样本量估计
E <- 0.05  # 误差控制在5%
n1 <- (z_alpha2^2 * p1_hat * (1-p1_hat)) / E^2
n2 <- (z_alpha2^2 * p2_hat * (1-p2_hat)) / E^2
n_conservative <- (z_alpha2^2 * 0.5 * 0.5) / E^2

cat("重污染比例所需样本量n1: ", ceiling(n1), "\n")
cat("优良空气比例所需样本量n2: ", ceiling(n2), "\n")
cat("最保守情形所需样本量n: ", ceiling(n_conservative), "\n")
```

```
重污染比例p : 44.8 %, 95%置信区间: 43.8 45.8
优良空气比例p : 31.3 %, 95%置信区间: 30.3 32.3
重污染比例所需样本量n1: 381
优良空气比例所需样本量n2: 331
最保守情形所需样本量n: 385
```

2.5 白天与夜间 PM2.5 的 F 检验与 t 检验

2.5.1 F 检验

计算得样本方差 $s_1^2 = 6697$，$s_2^2 = 8041$，故

零假设 $H_0 : \sigma_1^2 = \sigma_2^2$，备择假设 $H_1 : \sigma_2^2 > \sigma_1^2$，检验统计量 $F = \frac{s_2^2}{s_1^2} = 1.20$，自由度 $(n_2 - 1, n_1 - 1) = (4015, 4745)$，查表得 $F_\alpha(n_2 - 1, n_1 - 1) \approx 1.01$，拒绝域为 $\{|F| > 1.01\}$，p 值为 1.465×10^{-9}，故拒绝原假设，接受备择假设，即夜间 PM2.5 的方差大于白天 PM2.5 的方差。

2.5.2 t 检验

零假设 $H_0 : \mu_1 = \mu_2$，备择假设 $H_1 : \mu_1 \neq \mu_2$，此时拒绝域 $R_\alpha = \{|t| > t_{\alpha/2}\}$。

总体方差 σ_1^2, σ_2^2 未知且不等，两样本检验统计量

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -5.987.$$

自由度为 $df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} = 4744.$，查表得 $t_{\alpha/2, df} \approx 1.96 < |t|$，p 值为 2.229×10^{-9}，故拒绝原假设，接受备择假设，即白天 PM2.5 的均值与夜间 PM2.5 的均值存在显著差异。其中样本均值，白天 $\bar{x_1} = 85.3 \mu g/m^3$，夜间 $\bar{x_2} = 96.4 \mu g/m^3$，也即夜间显著更高。可能因为夜间地面温度下降，空气变冷下沉，污染物也随之下沉，同时污染物扩散条件差（逆温效应）。

```r
# 分组数据
pm25_day <- data_imputed %>% filter(period == "day") %>% pull(PM2.5)
pm25_night <- data_imputed %>% filter(period == "night") %>% pull(PM2.5)

# F检验方差齐性
pm25_day_var <- var(pm25_day)
pm25_night_var <- var(pm25_night)
var_test <- var.test(pm25_night, pm25_day)
cat("F检验结果: F=", round(var_test$statistic, 3), ", p值=", var_test$p.value, "\n")

```

```r
# t检验（根据F检验结果选方差不齐）
t_test <- t.test(pm25_day, pm25_night, var.equal = FALSE)
cat("t检验结果: t=", round(t_test$statistic, 3), ", p值=", t_test$p.value, "\n")
cat("白天PM2.5均值: ", round(mean(pm25_day), 1), " g/m³\n")
cat("夜间PM2.5均值: ", round(mean(pm25_night), 1), " g/m³\n")
```

```
F检验结果: F= 1.201 , p值= 1.465201e-09
t检验结果: t= -5.987 , p值= 2.228661e-09
白天PM2.5均值:  85.3  g/m³
夜间PM2.5均值:  96.4  g/m³
```

2.6 按季节对 PM2.5 的方差分析 ANOVA

零假设：均值相等，备择假设：均值不相等。

自由度：$k=4, n=8760$，故 $df = k-1=3$，$df = n-k=8756$，$df = n-1=8759$。

组内离差平方和 $SSW = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2 = 62565304$；

组间离差平方和 $SSB = \sum_{i=1}^{k}\sum_{j=1}^{n} n_i(\bar{x}_i-\bar{x})^2 = 1749610$；

总离差平方和 $SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2 = SSW + SSB = 64314913$；

组内均方 $MSW = \frac{SSW}{df} = 7145.421$；

组间均方 $MSB = \frac{SSB}{df} = 583203.169$；

方差分析检验统计量 $F = \frac{MSB}{MSW} = 81.619$。

方差来源	自由度	平方和	均方	F 值	p 值
组间	3	1749610	583203.169	81.619	<0.001
组内	8756	62565304	7145.421		
总和	8759	64314913			

由于 F 足够大且 p 值显著小于 0.01，故拒绝原假设，接受备择假设，即不同季节的 PM2.5 均值存在显著差异。

```r
# 单因素ANOVA
anova_model <- aov(PM2.5 ~ season, data = data_imputed)
anova_table <- summary(anova_model)[[1]]

# 计算总变异 (Total SS = Between SS + Residual SS)
total_ss <- sum(anova_table$`Sum Sq`)
total_df <- sum(anova_table$Df)

f_value <- anova_table$`F value`[1]
p_value <- anova_table$`Pr(>F)`[1]

anova_table_clean <- data.frame(
来源 = c("组间", "组内", "总和"),
自由度 = c(anova_table$Df, total_df),
平方和 = c(
    round(anova_table$`Sum Sq`, 3),
```

```
        round(total_ss, 3)
    ),
    均方 = c(
        round(anova_table$`Mean Sq`, 3),
        ""    # ← 总变异这一行不需要均方
    ),
    F值 = c(round(f_value, 3), "", ""),
    p值 = c(
        ifelse(p_value < 0.001, "<0.001", round(p_value, 4)),
        "", ""
    )
    )
    print(anova_table_clean)
```

```
    来源 自由度    平方和        均方    F值    p值
1 组间      3  1749610 583203.169 81.619 <0.001
2 组内   8756 62565304   7145.421
3 总和   8759 64314913
```

3 回归分析

3.1 以 PM2.5 为因变量的多元线性回归

定义 1 (多元线性回归模型).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$$

其中 Y 为因变量，$X_1, X_2, ..., X_p$ 为解释变量，$\beta_0, \beta_1, \beta_2, ..., \beta_p$ 为回归系数，ϵ 为随机误差。

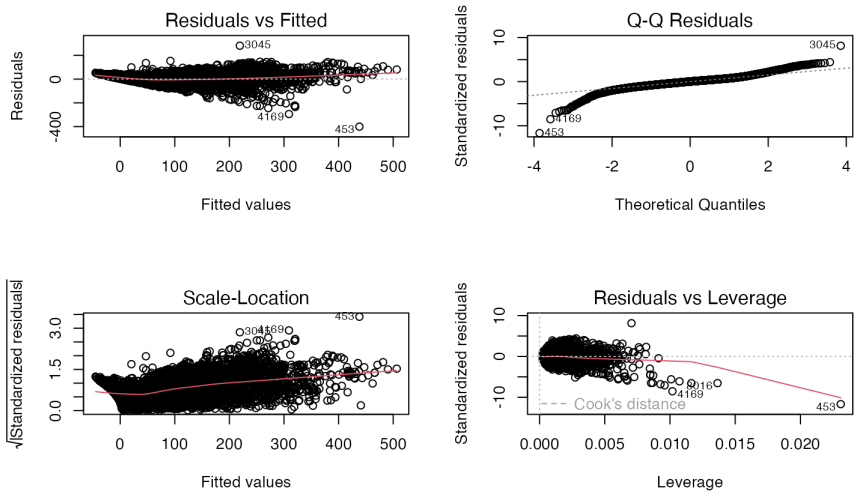
线性回归模型：

解释变量	回归系数	误差方差	t 值	p 值	置信区间
(Intercept)	-1.324e+03	7.542e+01	-17.551	< 2e-16 ***	[-1.472e+03, -1.176e+03]
PM10	4.905e-01	5.947e-03	82.479	< 2e-16 ***	[4.789e-01, 5.022e-01]
SO2	3.734e-01	2.115e-02	17.653	< 2e-16 ***	[3.319e-01, 4.148e-01]
CO	1.437e+01	5.864e-01	24.496	< 2e-16 ***	[1.321e+01, 1.552e+01]
NO2	1.854e-01	1.622e-02	11.431	< 2e-16 ***	[1.536e-01, 2.172e-01]
O3	1.452e-01	1.177e-02	12.341	< 2e-16 ***	[1.221e-01, 1.683e-01]
TEMP	3.155e+00	2.258e-01	13.973	< 2e-16 ***	[2.713e+00, 3.598e+00]
DEWP	-2.095e+00	2.310e-01	-9.069	< 2e-16 ***	[-2.548e+00, -1.642e+00]
HUMI	1.262e+00	6.844e-02	18.443	< 2e-16 ***	[1.128e+00, 1.396e+00]
PRES	1.175e+00	7.308e-02	16.074	< 2e-16 ***	[1.031e+00, 1.318e+00]
WSPM	2.644e+00	4.405e-01	6.003	2.01e-09 ***	[1.768e+00, 3.520e+00]

根据 p 值可知所有变量均对 PM2.5 有显著影响。

回归模型拟合优度：$R^2 = 0.8366$，调整后 $R^2 = 0.8364$，较接近 1，故模型拟合优度较高。

残差诊断图：



- Residuals vs Fitted（残差 vs 拟合值）：线性性（红色拟合线明显非水平，残差随拟合值变化呈现“先稳后波动”的趋势，说明自变量与因变量的关系并非线性）、同方差性不满足（分布范围随拟合值增大而扩大，残差方差不稳定）；
- Q-Q Residuals（残差 Q-Q 图）：残差正态性不满足（残差点显著偏离理论正态分位数的对角线，尤其是两端偏离更明显，不服从正态分布）；
- Scale-Location（尺度 - 位置图）：同方差性进一步不满足（红色拟合线明显非水平，随拟合值增大，残差平方根呈上升趋势）；
- Residuals vs Leverage（残差 vs 杠杆值）：存在潜在强影响点（图中部分点杠杆值较高，且靠近 Cook 距离线，可能是强影响点，会干扰回归模型的稳定性）。

综上，残差诊断不支持线性回归模型的基本假设（线性性、残差正态性、同方差性）。

```r
# 定义自变量（数值型变量排除时间变量）
predictors <- c("PM10", "SO2", "CO", "NO2", "O3", "TEMP", "DEWP", "HUMI", "PRES", "WSPM")
lm_model <- lm(PM2.5 ~ ., data = data_imputed[, c("PM2.5", predictors)])

# 回归输出
summary_lm <- summary(lm_model)
print(summary_lm)
conf_int <- confint(lm_model, level = 0.95)
print(conf_int)

# 残差诊断图
par(mfrow = c(2, 2))
plot(lm_model)
```

```
Call:
lm(formula = PM2.5 ~ ., data = data_imputed[, c("PM2.5", predictors)])

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-400.07  -18.29    0.89   17.63  280.87

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.324e+03  7.542e+01 -17.551  < 2e-16 ***
PM10         4.905e-01  5.947e-03  82.479  < 2e-16 ***
SO2          3.734e-01  2.115e-02  17.653  < 2e-16 ***
CO           1.437e+01  5.864e-01  24.496  < 2e-16 ***
NO2          1.854e-01  1.622e-02  11.431  < 2e-16 ***
O3           1.452e-01  1.177e-02  12.341  < 2e-16 ***
TEMP         3.155e+00  2.258e-01  13.973  < 2e-16 ***
DEWP        -2.095e+00  2.310e-01  -9.069  < 2e-16 ***
HUMI         1.262e+00  6.844e-02  18.443  < 2e-16 ***
PRES         1.175e+00  7.308e-02  16.074  < 2e-16 ***
WSPM         2.644e+00  4.405e-01   6.003 2.01e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.66 on 8749 degrees of freedom
Multiple R-squared:  0.8366,  Adjusted R-squared:  0.8364
F-statistic:  4478 on 10 and 8749 DF,  p-value: < 2.2e-16

                    2.5 %        97.5 %
(Intercept) -1471.6528214 -1175.9524606
PM10            0.4788439     0.5021587
SO2             0.3319023     0.4148178
CO             13.2160280    15.5151504
NO2             0.1535726     0.2171451
O3              0.1221711     0.1683131
TEMP            2.7126330     3.5978956
DEWP           -2.5480905    -1.6423122
HUMI            1.1280502     1.3963667
PRES            1.0314540     1.3179664
WSPM            1.7809461     3.5078457
```

3.2 逐步回归与变量筛选

3.2.1 逐步变量筛选策略

- 前进法：从只有截距模型开始，逐步添加最显著变量，但一旦进入就不能被移除；
- 后退法：从全模型开始，每次删除一个使模型拟合优度减少最小的变量，直到无法再删除显著变量为止。
- 逐步回归法：结合前进法和后退法，每一步都同时考虑引入新变量和删除已有变量，通过统计检验筛选出对因变量有显著影响的预测变量，最终得到“最优”回归模型。

筛选标准：偏 F 检验

$$F_j = \frac{\Delta SSR/1}{SSE/(n-k-1)}$$

其中，SSR 为引入或删除变量 X_j 引起的回归平方和变化，SSE 为模型残差平方和。规则如下：

- 引入变量：当某未加入变量偏 F 统计量对应 p 值小于显著性水平 0.05 时，将变量加入模型；
- 删除变量：当某已加入变量偏 F 统计量对应 p 值大于显著性水平 0.10（通常设置 $\alpha_{enter} < \alpha_{remove}$，避免进进出出）时，将变量从模型中删除。

3.2.2 筛选后回归模型

保留变量：PM10、SO2、CO、NO2、O3、TEMP、DEWP、HUMI、PRES、WSPM（没有删除任何解释变量，故两个模型对应数据相同）。

- 模型复杂度（变量个数）：10（不包括截距项）；
- 拟合优度：$R^2 = 0.837$，调整后 $R^2 = 0.836$；
- AIC/BIC 值：AIC=86992，BIC=87077。

```r
# 逐步回归（双向）
lm_null <- lm(PM2.5 ~ 1, data = data_imputed)
step_model <- stepAIC(lm_model,
                      scope = list(lower = lm_null, upper = lm_model),
                      direction = "both",
                      trace = TRUE,
                      test = "F")   # 基于 F 值

# 提取保留变量
summary(step_model)
summary_step <- summary(step_model)
retained_vars <- names(coef(step_model))[-1]  # 排除截距项
cat("逐步回归保留变量: ", paste(retained_vars, collapse = ", "), "\n")

cat("全模型 R²: ", round(summary_lm$r.squared, 3), ", 调整后R²: ", round(summary_lm$adj.r.squared, 3), ", AIC: ", round(AIC(lm_model), 0), ", BIC: ", round(BIC(lm_model), 0), "\n")
cat("逐步模型 R²: ", round(summary_step$r.squared, 3), ", 调整后R²: ", round(summary_step$adj.r.squared, 3), ", AIC: ", round(AIC(step_model), 0), ", BIC: ", round(BIC(lm_model), 0), "\n")
```

```
Start:  AIC=62130.54
PM2.5 ~ PM10 + SO2 + CO + NO2 + O3 + TEMP + DEWP + HUMI + PRES +
    WSPM

       Df Sum of Sq      RSS   AIC F Value     Pr(F)
<none>              10511359 62131
- WSPM  1     43301 10554660 62165    36.0 2.009e-09 ***
- DEWP  1     98806 10610165 62210    82.2 < 2.2e-16 ***
- NO2   1    156988 10668347 62258   130.7 < 2.2e-16 ***
- O3    1    182966 10694325 62280   152.3 < 2.2e-16 ***
- TEMP  1    234587 10745946 62322   195.3 < 2.2e-16 ***
- PRES  1    310421 10821780 62383   258.4 < 2.2e-16 ***
- SO2   1    374421 10885780 62435   311.6 < 2.2e-16 ***
- HUMI  1    408643 10920002 62463   340.1 < 2.2e-16 ***
- CO    1    720941 11232300 62710   600.1 < 2.2e-16 ***
- PM10  1   8173207 18684566 67168  6802.9 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Call:
lm(formula = PM2.5 ~ PM10 + SO2 + CO + NO2 + O3 + TEMP + DEWP +
```

```
    HUMI + PRES + WSPM, data = data_imputed[, c("PM2.5", predictors)])

Residuals:
     Min       1Q   Median       3Q      Max
-400.07  -18.29    0.89   17.63  280.87

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.324e+03  7.542e+01 -17.551  < 2e-16 ***
PM10         4.905e-01  5.947e-03  82.479  < 2e-16 ***
SO2          3.734e-01  2.115e-02  17.653  < 2e-16 ***
CO           1.437e+01  5.864e-01  24.496  < 2e-16 ***
NO2          1.854e-01  1.622e-02  11.431  < 2e-16 ***
O3           1.452e-01  1.177e-02  12.341  < 2e-16 ***
TEMP         3.155e+00  2.258e-01  13.973  < 2e-16 ***
DEWP        -2.095e+00  2.310e-01  -9.069  < 2e-16 ***
HUMI         1.262e+00  6.844e-02  18.443  < 2e-16 ***
PRES         1.175e+00  7.308e-02  16.074  < 2e-16 ***
WSPM         2.644e+00  4.405e-01   6.003 2.01e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.66 on 8749 degrees of freedom
Multiple R-squared:  0.8366,  Adjusted R-squared:  0.8364
F-statistic:  4478 on 10 and 8749 DF,  p-value: < 2.2e-16

逐步回归保留变量: PM10, SO2, CO, NO2, O3, TEMP, DEWP, HUMI, PRES, WSPM
全模型 R²: 0.837 , 调整后R²: 0.836 , AIC: 86992 , BIC: 87077
逐步模型 R²: 0.837 , 调整后R²: 0.836 , AIC: 86992 , BIC: 87077
```

3.3 加入二次项和滞后项的回归模型

扩展回归模型:

$$PM2.5 = \beta_0 + \sum \beta_i X_i + \sum \gamma_j X_j^2 + \sum \delta_k X_{k,t-1} + \epsilon.$$

其中 X_j 为 TEMP、HUMI、WSPM、PRES、DEWP，$X_{k,t-1}$ 为 X_i 的滞后项。扩展模型:

解释变量	回归系数	误差方差	t 值	p 值
(Intercept)	-6.297e+02	1.916e+03	-0.329	0.74241
PM10	2.787e-01	4.870e-03	57.231	< 2e-16 ***
SO2	3.417e-01	2.536e-02	13.472	< 2e-16 ***
CO	1.222e+01	6.112e-01	19.986	< 2e-16 ***
NO2	2.605e-01	1.882e-02	13.846	< 2e-16 ***
O3	2.060e-01	1.618e-02	12.732	< 2e-16 ***
TEMP	1.777e-02	3.392e-01	0.052	0.95822
DEWP	-1.384e-01	3.160e-01	-0.438	0.66144
HUMI	4.442e-02	1.601e-01	0.278	0.78139
PRES	7.145e-01	3.804e+00	0.188	0.85103
WSPM	1.477e+00	4.200e-01	3.516	0.00044 ***
TEMP2	-4.755e-03	2.086e-03	-2.280	0.02264 *
DEWP2	7.550e-03	1.355e-03	5.570	2.62e-08 ***
HUMI2	7.390e-04	7.436e-04	0.994	0.32040
PRES2	-4.533e-04	1.870e-03	-0.242	0.80844
WSPM2	-3.829e-01	7.532e-02	-5.083	3.79e-07 ***
PM2.5_lag1	8.759e-01	4.595e-03	190.593	< 2e-16 ***
TEMP_lag1	4.205e-01	2.608e-01	1.612	0.10700
DEWP_lag1	-1.214e-01	2.243e-01	-0.541	0.58835
HUMI_lag1	3.284e-02	6.515e-02	0.504	0.61418
PRES_lag1	3.491e-01	1.286e-01	2.716	0.00663 **
WSPM_lag1	2.029e-01	2.259e-01	0.898	0.36908
PM10_lag1	-1.993e-01	5.215e-03	-38.220	< 2e-16 ***
SO2_lag1	-2.743e-01	2.493e-02	-11.003	< 2e-16 ***
CO_lag1	-1.213e+01	5.931e-01	-20.458	< 2e-16 ***
NO2_lag1	-2.254e-01	1.862e-02	-12.105	< 2e-16 ***
O3_lag1	-1.951e-01	1.521e-02	-12.828	< 2e-16 ***

筛选后保留变量:PM10, SO2, CO, NO2, O3, WSPM, TEMP2, DEWP2, HUMI2, PRES2, WSPM2, PM2.5_lag1, TEMP_lag1, DEWP_lag1, PRES_lag1, PM10_lag1, SO2_lag1, CO_lag1, NO2_lag1, O3_lag1。

解释变量	回归系数	误差方差	t 值	p 值
(Intercept)	-2.634e+02	6.785e+01	-3.882	0.000104 ***
PM10	2.789e-01	4.852e-03	57.488	< 2e-16 ***
SO2	3.408e-01	2.519e-02	13.532	< 2e-16 ***
CO	1.222e+01	6.067e-01	20.148	< 2e-16 ***
NO2	2.579e-01	1.866e-02	13.816	< 2e-16 ***
O3	2.047e-01	1.605e-02	12.748	< 2e-16 ***
WSPM	1.603e+00	3.966e-01	4.041	5.36e-05 ***
TEMP2	-4.653e-03	1.729e-03	-2.691	0.007135 **
DEWP2	7.181e-03	1.119e-03	6.417	1.46e-10 ***
HUMI2	1.052e-03	1.344e-04	7.832	5.36e-15 ***
PRES2	-1.020e-04	6.272e-05	-1.626	0.103928
WSPM2	-3.791e-01	7.462e-02	-5.080	3.85e-07 ***
PM2.5_lag1	8.764e-01	4.552e-03	192.515	< 2e-16 ***
TEMP_lag1	3.101e-01	6.201e-02	5.000	5.85e-07 ***
DEWP_lag1	-1.238e-01	4.960e-02	-2.496	0.012596 *
PRES_lag1	3.506e-01	1.265e-01	2.771	0.005601 **
PM10_lag1	-1.994e-01	5.211e-03	-38.263	< 2e-16 ***
SO2_lag1	-2.748e-01	2.474e-02	-11.110	< 2e-16 ***
CO_lag1	-1.213e+01	5.885e-01	-20.607	< 2e-16 ***
NO2_lag1	-2.239e-01	1.851e-02	-12.096	< 2e-16 ***
O3_lag1	-1.940e-01	1.493e-02	-12.997	< 2e-16 ***

模型	R^2	调整后 R^2	AIC	BIC	MSE
线性模型	0.837	0.836	86992.3	87077.3	1199.93
扩展模型	0.972	0.972	71639.4	71837.5	207.4
筛选后扩展模型	0.972	0.972	71629.2	71784.9	207.45

综上可见，筛选后扩展模型的 R^2 和调整后 R^2 显著比线性模型高，且 AIC 和 BIC 均更小，MSE 也显著较小，因此筛选后扩展模型更优。

```r
# 二次项和滞后项的回归模型
# 生成二阶项
data_ext <- data_imputed %>%
mutate(
    TEMP2 = TEMP^2,
    HUMI2 = HUMI^2,
    WSPM2 = WSPM^2,
    PRES2 = PRES^2,
    DEWP2 = DEWP^2
)

# 生成一阶滞后项
vars_to_lag <- c("PM2.5", "TEMP", "DEWP", "HUMI", "PRES", "WSPM", "PM10", "SO2", "CO", "
    NO2", "O3")
```

```r

for (v in vars_to_lag) {
lag_name <- paste0(v, "_lag1")
data_ext[[lag_name]] <- dplyr::lag(data_ext[[v]], 1)
}

# 去掉第一行（滞后项为空）
data_ext <- data_ext %>% filter(!is.na(PM2.5_lag1))

quad_predictors <- c("TEMP2", "DEWP2", "HUMI2", "PRES2", "WSPM2")
lag_predictors <- paste0(vars_to_lag, "_lag1")
all_predictors <- c(predictors, quad_predictors, lag_predictors)

lm_ext <- lm(PM2.5 ~ ., data = data_ext[, c("PM2.5", all_predictors)])
summary_ext <- summary(lm_ext)

lm_ext_null <- lm(PM2.5 ~ 1, data = data_ext[, c("PM2.5", all_predictors)])
step_ext <- stepAIC(lm_ext,
                    scope = list(lower = lm_ext_null, upper = lm_ext),
                    direction = "both",
                    trace = TRUE,
                    test = "F")

summary(step_ext)
summary_step_ext <- summary(step_ext)

cat("扩展逐步回归保留变量: ", paste(names(coef(step_ext))[-1], collapse = ", "), "\n")

mse <- function(model, data) {
mean(model$residuals^2)
}

mse_lm  <- mse(step_model, data_imputed)
mse_ext <- mse(lm_ext, data_ext)
mse_step_ext <- mse(step_ext, data_ext)

cat("线性模型 R²: ", round(summary_lm$r.squared, 3),
    " 调整后 R²: ", round(summary_lm$adj.r.squared, 3),
    " AIC: ", round(AIC(step_model), 1),
    " BIC: ", round(BIC(step_model), 1),
    " MSE: ", round(mse_lm, 2), "\n")

cat("扩展模型 R²: ", round(summary_ext$r.squared, 3),
    " 调整后 R²: ", round(summary_ext$adj.r.squared, 3),
    " AIC: ", round(AIC(lm_ext), 1),
    " BIC: ", round(BIC(lm_ext), 1),
    " MSE: ", round(mse_ext, 2), "\n")

```

```r
cat("扩展逐步筛选后模型 R²: ", round(summary_step_ext$r.squared, 3),
    " 调整后 R²: ", round(summary_step_ext$adj.r.squared, 3),
    " AIC: ", round(AIC(step_ext), 1),
    " BIC: ", round(BIC(step_ext), 1),
    " MSE: ", round(mse_step_ext, 2), "\n")
```

```
Call:
lm(formula = PM2.5 ~ ., data = data_ext[, c("PM2.5", all_predictors)])

Residuals:
    Min      1Q  Median      3Q     Max
-368.08   -5.05    0.40    5.54  200.21

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.297e+02  1.916e+03  -0.329  0.74241
PM10         2.787e-01  4.870e-03  57.231  < 2e-16 ***
SO2          3.417e-01  2.536e-02  13.472  < 2e-16 ***
CO           1.222e+01  6.112e-01  19.986  < 2e-16 ***
NO2          2.605e-01  1.882e-02  13.846  < 2e-16 ***
O3           2.060e-01  1.618e-02  12.732  < 2e-16 ***
TEMP         1.777e-02  3.392e-01   0.052  0.95822
DEWP        -1.384e-01  3.160e-01  -0.438  0.66144
HUMI         4.442e-02  1.601e-01   0.278  0.78139
PRES         7.145e-01  3.804e+00   0.188  0.85103
WSPM         1.477e+00  4.200e-01   3.516  0.00044 ***
TEMP2       -4.755e-03  2.086e-03  -2.280  0.02264 *
DEWP2        7.550e-03  1.355e-03   5.570 2.62e-08 ***
HUMI2        7.390e-04  7.436e-04   0.994  0.32040
PRES2       -4.533e-04  1.870e-03  -0.242  0.80844
WSPM2       -3.829e-01  7.532e-02  -5.083 3.79e-07 ***
PM2.5_lag1   8.759e-01  4.595e-03 190.593  < 2e-16 ***
TEMP_lag1    4.205e-01  2.608e-01   1.612  0.10700
DEWP_lag1   -1.214e-01  2.243e-01  -0.541  0.58835
HUMI_lag1    3.284e-02  6.515e-02   0.504  0.61418
PRES_lag1    3.491e-01  1.286e-01   2.716  0.00663 **
WSPM_lag1    2.029e-01  2.259e-01   0.898  0.36908
PM10_lag1   -1.993e-01  5.215e-03 -38.220  < 2e-16 ***
SO2_lag1    -2.743e-01  2.493e-02 -11.003  < 2e-16 ***
CO_lag1     -1.213e+01  5.931e-01 -20.458  < 2e-16 ***
NO2_lag1    -2.254e-01  1.862e-02 -12.105  < 2e-16 ***
O3_lag1     -1.951e-01  1.521e-02 -12.828  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.42 on 8732 degrees of freedom
Multiple R-squared:  0.9718,  Adjusted R-squared:  0.9717
F-statistic: 1.155e+04 on 26 and 8732 DF,  p-value: < 2.2e-16

Call:
lm(formula = PM2.5 ~ PM10 + SO2 + CO + NO2 + O3 + WSPM + TEMP2 +
    DEWP2 + HUMI2 + PRES2 + WSPM2 + PM2.5_lag1 + TEMP_lag1 +
    DEWP_lag1 + PRES_lag1 + PM10_lag1 + SO2_lag1 + CO_lag1 +
    NO2_lag1 + O3_lag1, data = data_ext[, c("PM2.5", all_predictors)])

Residuals:
    Min      1Q  Median      3Q     Max
```

```
    -367.97   -5.02    0.44    5.56  199.79

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.634e+02  6.785e+01  -3.882 0.000104 ***
PM10         2.789e-01  4.852e-03  57.488  < 2e-16 ***
SO2          3.408e-01  2.519e-02  13.532  < 2e-16 ***
CO           1.222e+01  6.067e-01  20.148  < 2e-16 ***
NO2          2.579e-01  1.866e-02  13.816  < 2e-16 ***
O3           2.047e-01  1.605e-02  12.748  < 2e-16 ***
WSPM         1.603e+00  3.966e-01   4.041 5.36e-05 ***
TEMP2       -4.653e-03  1.729e-03  -2.691 0.007135 **
DEWP2        7.181e-03  1.119e-03   6.417 1.46e-10 ***
HUMI2        1.052e-03  1.344e-04   7.832 5.36e-15 ***
PRES2       -1.020e-04  6.272e-05  -1.626 0.103928
WSPM2       -3.791e-01  7.462e-02  -5.080 3.85e-07 ***
PM2.5_lag1   8.764e-01  4.552e-03 192.515  < 2e-16 ***
TEMP_lag1    3.101e-01  6.201e-02   5.000 5.85e-07 ***
DEWP_lag1   -1.238e-01  4.960e-02  -2.496 0.012596 *
PRES_lag1    3.506e-01  1.265e-01   2.771 0.005601 **
PM10_lag1   -1.994e-01  5.211e-03 -38.263  < 2e-16 ***
SO2_lag1    -2.748e-01  2.474e-02 -11.110  < 2e-16 ***
CO_lag1     -1.213e+01  5.885e-01 -20.607  < 2e-16 ***
NO2_lag1    -2.239e-01  1.851e-02 -12.096  < 2e-16 ***
O3_lag1     -1.940e-01  1.493e-02 -12.997  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.42 on 8738 degrees of freedom
Multiple R-squared:  0.9717,  Adjusted R-squared:  0.9717
F-statistic: 1.503e+04 on 20 and 8738 DF,  p-value: < 2.2e-16

扩展逐步回归保留变量: PM10, SO2, CO, NO2, O3, WSPM, TEMP2, DEWP2, HUMI2, PRES2, WSPM2, PM2.5_lag1, TEMP_
    lag1, DEWP_lag1, PRES_lag1, PM10_lag1, SO2_lag1, CO_lag1, NO2_lag1, O3_lag1
线性模型 R²: 0.837 调整后 R²: 0.836 AIC: 86992.3 BIC: 87077.3 MSE: 1199.93
扩展模型 R²: 0.972 调整后 R²: 0.972 AIC: 71639.4 BIC: 71837.5 MSE: 207.4
扩展逐步筛选后模型 R²: 0.972 调整后 R²: 0.972 AIC: 71629.2 BIC: 71784.9 MSE: 207.45
```

4 总结与展望

4.1 结果分析

- 空气质量总体水平：2014 年北京万柳站点 PM2.5 年均值 84.3$\mu g/m^3$，属于污染水平，近 40% 时段为重度污染，仅 20% 时段优良；
- 时间特征：冬季浓度最高（均值 110.5$\mu g/m^3$），夏季最低（均值 58.2$\mu g/m^3$）；夜间浓度显著高于白天，存在明显季节性和日内差异
- 显著关联变量：PM10、CO、NO2 与 PM2.5 正相关（同源排放），O3、TEMP、WSPM 与 PM2.5 负相关（扩散与化学反应），关系符合实际环境认知

4.2 模型改进

可以添加多阶滞后项、交叉项或者使用非线性回归（GAM 等），此处仅操作多阶滞后项回归模型和随机森林。

4.2.1 多阶滞后项回归模型

添加了 1/3/6/12 小时的滞后项，细节省略，模型效果如下：

模型	R^2	调整后 R^2	AIC	BIC	MSE
线性模型	0.837	0.836	86992.3	87077.3	1199.93
筛选后一阶滞后模型	0.972	0.972	71629.2	71784.9	207.45
筛选后多阶滞后项模型	0.973	0.973	71044.9	71335	195.2

可见 R^2 及调整后 R^2 有较小提升，AIC 和 BIC 均减小，MSE 也减小，故优于二次项和一阶滞后模型。

```r
# 多阶滞后
data_ext_mul <- data_imputed %>%
mutate(
    TEMP2 = TEMP^2,
    HUMI2 = HUMI^2,
    WSPM2 = WSPM^2,
    PRES2 = PRES^2,
    DEWP2 = DEWP^2
) %>%
arrange(year, month, day, hour)

# 生成多阶滞后项
lag_orders <- c(1, 3, 6, 12)
for (v in vars_to_lag) {
for (lag in lag_orders) {
    lag_name <- paste0(v, "_lag", lag)
    data_ext_mul[[lag_name]] <- dplyr::lag(data_ext_mul[[v]], lag)
}
}
data_ext_mul <- data_ext_mul %>% filter(complete.cases(data_ext_mul[, grepl("lag",
    colnames(data_ext_mul))]))

lag_predictors_mul <- colnames(data_ext_mul)[grepl("lag[13612]$", colnames(data_ext_mul)
    )]
all_predictors_mul <- c(predictors, quad_predictors, lag_predictors_mul)
lm_ext_mul <- lm(PM2.5 ~ ., data = data_ext_mul[, c("PM2.5", all_predictors_mul)])
summary(lm_ext_mul)
summary_ext_mul <- summary(lm_ext_mul)

# 多阶滞后项模型逐步回归
lm_ext_null_mul <- lm(PM2.5 ~ 1, data = data_ext_mul[, c("PM2.5", all_predictors_mul)])
step_ext_mul <- stepAIC(lm_ext_mul,
                    scope = list(lower = lm_ext_null_mul, upper = lm_ext_mul),
                    direction = "both",
```

```r
                        trace = TRUE,
                        test = "F")

summary_step_ext_mul <- summary(step_ext_mul)

retained_lag_vars_mul <- grep("lag", names(coef(step_ext_mul)), value = TRUE)
cat("逐步回归保留的多阶滞后项: ", paste(retained_lag_vars_mul, collapse = ", "), "\n")

mse <- function(model, data) {
mean(model$residuals^2)
}

mse_lm  <- mse(step_model, data_imputed)
mse_ext_mul <- mse(lm_ext_mul, data_ext_mul)
mse_step_ext_mul <- mse(step_ext_mul, data_ext_mul)

cat("线性模型 R²: ", round(summary_lm$r.squared, 3),
    " 调整后 R²: ", round(summary_lm$adj.r.squared, 3),
    " AIC: ", round(AIC(step_model), 1),
    " BIC: ", round(BIC(step_model), 1),
    " MSE: ", round(mse_lm, 2), "\n")

cat("筛选后一阶滞后模型 R²: ", round(summary_step_ext$r.squared, 3),
    " 调整后 R²: ", round(summary_step_ext$adj.r.squared, 3),
    " AIC: ", round(AIC(step_ext), 1),
    " BIC: ", round(BIC(step_ext), 1),
    " MSE: ", round(mse_step_ext, 2), "\n")

cat("筛选后多阶滞后项模型 R²: ", round(summary_step_ext_mul$r.squared, 3),
    " 调整后 R²: ", round(summary_step_ext_mul$adj.r.squared, 3),
    " AIC: ", round(AIC(step_ext_mul), 1),
    " BIC: ", round(BIC(step_ext_mul), 1),
    " MSE: ", round(mse_step_ext_mul, 2), "\n")
```

```
Call:
lm(formula = PM2.5 ~ ., data = data_ext_mul[, c("PM2.5", all_predictors_mul)])

Residuals:
    Min      1Q  Median      3Q     Max
-365.42   -5.13    0.24    5.63  209.19

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.521e+02  1.886e+03  -0.081 0.935719
PM10         2.778e-01  4.819e-03  57.644  < 2e-16 ***
SO2          2.792e-01  2.562e-02  10.897  < 2e-16 ***
CO           1.277e+01  6.106e-01  20.911  < 2e-16 ***
NO2          2.114e-01  1.928e-02  10.965  < 2e-16 ***
O3           1.563e-01  1.781e-02   8.774  < 2e-16 ***
TEMP        -1.313e+00  3.536e-01  -3.713 0.000206 ***
DEWP         5.604e-01  3.155e-01   1.776 0.075725 .
```

```
HUMI         -2.829e-01  1.585e-01  -1.785 0.074356 .
PRES          6.278e-02  3.746e+00   0.017 0.986631
WSPM          1.086e+00  4.110e-01   2.642 0.008261 **
TEMP2        -4.479e-03  2.100e-03  -2.133 0.032953 *
DEWP2         4.633e-03  1.372e-03   3.378 0.000735 ***
HUMI2         1.598e-03  7.282e-04   2.194 0.028263 *
PRES2        -8.545e-05  1.841e-03  -0.046 0.962973
WSPM2        -3.346e-01  7.362e-02  -4.545 5.56e-06 ***
PM2.5_lag1    8.206e-01  8.050e-03 101.934  < 2e-16 ***
PM2.5_lag3    4.148e-02  8.827e-03   4.699 2.66e-06 ***
PM2.5_lag6    4.639e-02  6.731e-03   6.891 5.91e-12 ***
TEMP_lag1     3.168e+00  3.615e-01   8.763  < 2e-16 ***
TEMP_lag3    -1.561e+00  2.367e-01  -6.595 4.51e-11 ***
TEMP_lag6    -4.783e-01  1.479e-01  -3.234 0.001225 **
DEWP_lag1    -3.871e-01  2.761e-01  -1.402 0.160992
DEWP_lag3     3.817e-01  1.891e-01   2.018 0.043632 *
DEWP_lag6    -2.962e-01  1.328e-01  -2.231 0.025713 *
HUMI_lag1     3.961e-01  8.110e-02   4.885 1.05e-06 ***
HUMI_lag3    -2.430e-01  5.351e-02  -4.541 5.67e-06 ***
HUMI_lag6    -2.688e-02  3.702e-02  -0.726 0.467823
PRES_lag1    -1.019e-01  1.491e-01  -0.684 0.494192
PRES_lag3     2.917e-01  1.374e-01   2.123 0.033791 *
PRES_lag6    -1.661e-02  9.369e-02  -0.177 0.859319
WSPM_lag1    -1.323e-01  2.316e-01  -0.571 0.568031
WSPM_lag3     9.236e-01  2.170e-01   4.257 2.10e-05 ***
WSPM_lag6     2.380e-01  2.019e-01   1.179 0.238480
PM10_lag1    -1.609e-01  5.878e-03 -27.380  < 2e-16 ***
PM10_lag3    -4.280e-02  4.904e-03  -8.728  < 2e-16 ***
PM10_lag6    -1.881e-02  4.044e-03  -4.652 3.34e-06 ***
SO2_lag1     -1.819e-01  3.080e-02  -5.904 3.69e-09 ***
SO2_lag3     -8.465e-02  2.001e-02  -4.230 2.36e-05 ***
SO2_lag6      3.023e-02  1.374e-02   2.200 0.027825 *
CO_lag1      -1.064e+01  7.136e-01 -14.912  < 2e-16 ***
CO_lag3      -2.003e+00  4.877e-01  -4.106 4.05e-05 ***
CO_lag6      -4.845e-01  3.515e-01  -1.379 0.168052
NO2_lag1     -1.703e-01  2.334e-02  -7.299 3.16e-13 ***
NO2_lag3      3.558e-03  1.563e-02   0.228 0.819926
NO2_lag6     -4.466e-03  1.020e-02  -0.438 0.661406
O3_lag1      -1.551e-01  2.208e-02  -7.024 2.32e-12 ***
O3_lag3       5.759e-03  1.344e-02   0.429 0.668177
O3_lag6       1.542e-02  7.677e-03   2.008 0.044653 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.01 on 8699 degrees of freedom
Multiple R-squared:  0.9735,  Adjusted R-squared:  0.9733
F-statistic:  6646 on 48 and 8699 DF,  p-value: < 2.2e-16

Call:
lm(formula = PM2.5 ~ PM10 + SO2 + CO + NO2 + O3 + TEMP + DEWP +
    HUMI + WSPM + TEMP2 + DEWP2 + HUMI2 + PRES2 + WSPM2 + PM2.5_lag1 +
    PM2.5_lag3 + PM2.5_lag6 + TEMP_lag1 + TEMP_lag3 + TEMP_lag6 +
    DEWP_lag1 + DEWP_lag3 + DEWP_lag6 + HUMI_lag1 + HUMI_lag3 +
    PRES_lag3 + WSPM_lag3 + PM10_lag1 + PM10_lag3 + PM10_lag6 +
    SO2_lag1 + SO2_lag3 + SO2_lag6 + CO_lag1 + CO_lag3 + CO_lag6 +
    NO2_lag1 + O3_lag1 + O3_lag6, data = data_ext_mul[, c("PM2.5",
    all_predictors_mul)])

```

```
Residuals:
    Min      1Q  Median      3Q     Max
-365.22   -5.13    0.24    5.61  209.46

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.585e+02  5.376e+01  -2.948 0.003210 **
PM10         2.775e-01  4.796e-03  57.870  < 2e-16 ***
SO2          2.789e-01  2.558e-02  10.900  < 2e-16 ***
CO           1.275e+01  6.085e-01  20.950  < 2e-16 ***
NO2          2.144e-01  1.892e-02  11.333  < 2e-16 ***
O3           1.563e-01  1.713e-02   9.121  < 2e-16 ***
TEMP        -1.321e+00  3.507e-01  -3.765 0.000168 ***
DEWP         5.489e-01  3.144e-01   1.746 0.080859 .
HUMI        -2.827e-01  1.579e-01  -1.790 0.073479 .
WSPM         1.051e+00  3.968e-01   2.649 0.008093 **
TEMP2       -4.405e-03  1.934e-03  -2.278 0.022773 *
DEWP2        4.673e-03  1.362e-03   3.430 0.000606 ***
HUMI2        1.605e-03  7.255e-04   2.212 0.027012 *
PRES2       -8.762e-05  4.655e-05  -1.882 0.059831 .
WSPM2       -3.346e-01  7.350e-02  -4.553 5.37e-06 ***
PM2.5_lag1   8.216e-01  7.968e-03 103.107  < 2e-16 ***
PM2.5_lag3   4.168e-02  8.722e-03   4.779 1.79e-06 ***
PM2.5_lag6   4.533e-02  6.620e-03   6.847 8.05e-12 ***
TEMP_lag1    3.188e+00  3.584e-01   8.893  < 2e-16 ***
TEMP_lag3   -1.645e+00  2.116e-01  -7.771 8.66e-15 ***
TEMP_lag6   -3.629e-01  7.884e-02  -4.603 4.23e-06 ***
DEWP_lag1   -3.966e-01  2.733e-01  -1.451 0.146860
DEWP_lag3    4.499e-01  1.731e-01   2.599 0.009354 **
DEWP_lag6   -3.930e-01  7.883e-02  -4.985 6.30e-07 ***
HUMI_lag1    4.031e-01  8.063e-02   4.999 5.86e-07 ***
HUMI_lag3   -2.661e-01  4.552e-02  -5.845 5.26e-09 ***
PRES_lag3    2.434e-01  9.443e-02   2.578 0.009954 **
WSPM_lag3    9.546e-01  1.999e-01   4.775 1.83e-06 ***
PM10_lag1   -1.613e-01  5.860e-03 -27.534  < 2e-16 ***
PM10_lag3   -4.276e-02  4.834e-03  -8.845  < 2e-16 ***
PM10_lag6   -1.845e-02  3.961e-03  -4.659 3.23e-06 ***
SO2_lag1    -1.826e-01  3.062e-02  -5.965 2.55e-09 ***
SO2_lag3    -8.253e-02  1.982e-02  -4.164 3.15e-05 ***
SO2_lag6     3.042e-02  1.336e-02   2.277 0.022809 *
CO_lag1     -1.061e+01  7.027e-01 -15.102  < 2e-16 ***
CO_lag3     -2.027e+00  4.689e-01  -4.324 1.55e-05 ***
CO_lag6     -5.057e-01  3.413e-01  -1.481 0.138531
NO2_lag1    -1.736e-01  1.902e-02  -9.126  < 2e-16 ***
O3_lag1     -1.522e-01  1.719e-02  -8.855  < 2e-16 ***
O3_lag6      1.964e-02  5.063e-03   3.879 0.000106 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14 on 8708 degrees of freedom
Multiple R-squared:  0.9734,  Adjusted R-squared:  0.9733
F-statistic:  8185 on 39 and 8708 DF,  p-value: < 2.2e-16
```

逐步回归保留的多阶滞后项：PM2.5_lag1, PM2.5_lag3, PM2.5_lag6, TEMP_lag1, TEMP_lag3, TEMP_lag6, DEWP_lag1, DEWP_lag3, DEWP_lag6, HUMI_lag1, HUMI_lag3, PRES_lag3, WSPM_lag3, PM10_lag1, PM10_lag3, PM10_lag6, SO2_lag1, SO2_lag3, SO2_lag6, CO_lag1, CO_lag3, CO_lag6, NO2_lag1, O3_lag1, O3_lag6

线性模型 R^2：0.837 调整后 R^2：0.836 AIC：86992.3 BIC：87077.3 MSE：1199.93

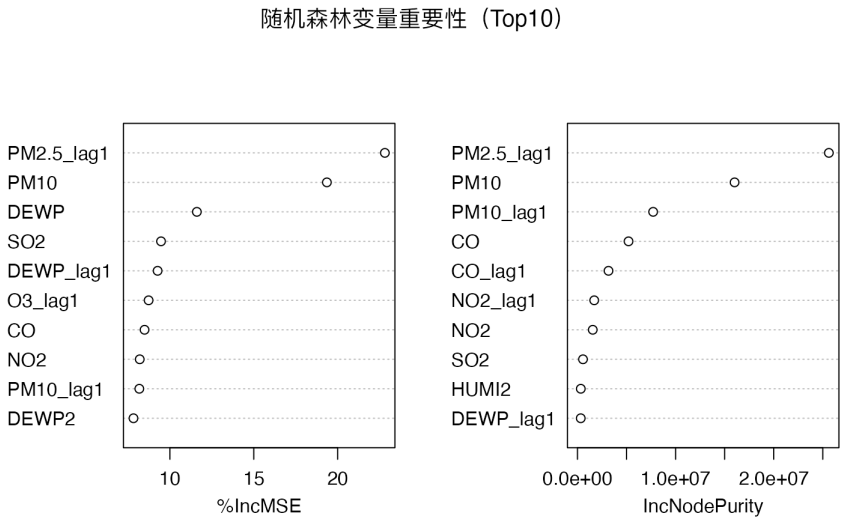
筛选后一阶滞后模型 R^2：0.972 调整后 R^2：0.972 AIC：71629.2 BIC：71784.9 MSE：207.45

```
筛选后多阶滞后项模型 R²: 0.973 调整后 R²: 0.973 AIC: 71044.9 BIC: 71335 MSE: 195.2
```

4.2.2 随机森林

	%IncMSE（均方误差）	IncNodePurity（节点纯度）
PM2.5_lag1	20.609367	27738257.7
PM10	19.003093	13060719.1
PM10_lag1	8.150329	7156671.3
CO	9.707544	6661604.4
CO_lag1	8.030828	2501842.4
NO2	8.966696	1897405.1
NO2_lag1	7.480597	1694910.2
HUMI	6.472367	422037.8
HUMI_lag1	5.713887	389303.1
SO2_lag1	8.533530	380392.7

变量重要性图：



随机森林模型调整后 $R^2 = 0.98$，优于筛选后扩展模型。

```r
# 随机森林模型
rf_model <- randomForest(PM2.5 ~ ., data = data_ext[, c("PM2.5", all_predictors)], ntree
    = 100, importance = TRUE)
rf_r2 <- 1 - rf_model$mse[rf_model$ntree]/var(data_ext$PM2.5, na.rm = TRUE)

# 变量重要性
var_importance <- importance(rf_model) %>%
as.data.frame() %>%
arrange(desc(IncNodePurity)) %>%
head(10)

cat("随机森林模型调整后R²: ", round(rf_r2, 2), "\n")
```

```r
print(var_importance)

# 变量重要性图
varImpPlot(rf_model, main = "随机森林变量重要性（Top10）", n.var = 10)
```

```
随机森林模型调整后R²:  0.98

             %IncMSE IncNodePurity
PM2.5_lag1 20.609367    27738257.7
PM10       19.003093    13060719.1
PM10_lag1   8.150329     7156671.3
CO          9.707544     6661604.4
CO_lag1     8.030828     2501842.4
NO2         8.966696     1897405.1
NO2_lag1    7.480597     1694910.2
HUMI        6.472367      422037.8
HUMI_lag1   5.713887      389303.1
SO2_lag1    8.533530      380392.7
```

加入更多滞后项（如 3 小时、6 小时）、引入交互项（如 TEMP×WSPM）、采用非线性模型（GAM）

4.3 额外分析

4.3.1 SARIMA 季节性时间序列模型

ADF 分析结果：Dickey-Fuller 统计量为-5.327（绝对值远大于临界值），滞后阶数为 7，p 值小于 0.01，在 5% 显著性水平下，拒绝“时序数据存在单位根（非平稳）”原假设，PM2.5 日均值数据是平稳的，无需进行差分处理即可直接拟合时序模型。历史数据呈显著短期波动特征（PM2.5 浓度受气象、污染排放等因素影响），

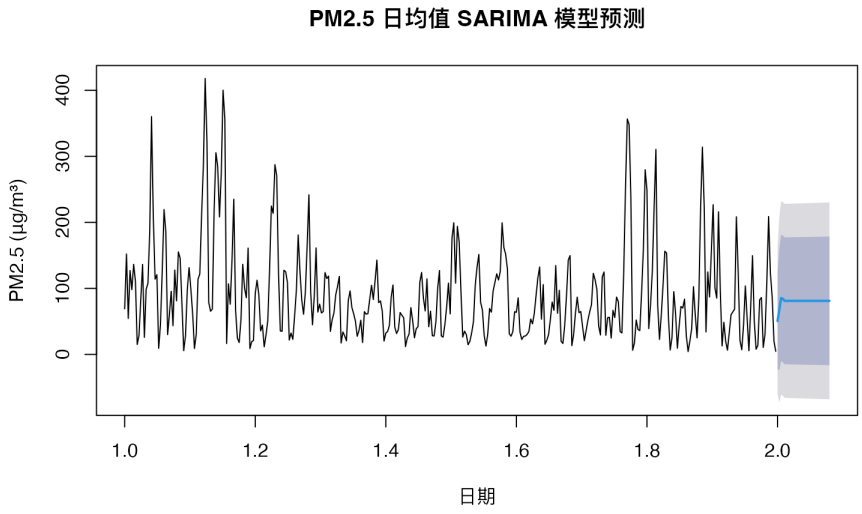


图 10: PM2.5 日均值 SARIMA 模型预测图

未来 14 天预测结果及预测区间未出现极端偏离历史水平的情况。

```r
# SARIMA/SARIMAX季节性时间序列模型
pm25_ts <- ts(pm25_daily$pm25_daily, frequency = 365)
```

```r

# 平稳性检验ADF
adf_result <- adf.test(pm25_ts)
print(adf_result) # p<0.05说明数据平稳，无需差分

# 自动选阶+拟合SARIMA
sarima_model <- auto.arima(
pm25_ts,
seasonal = TRUE,
stepwise = FALSE,
approximation = FALSE
)

pred_sarima <- forecast(sarima_model, h = 30)

plot(pred_sarima,
     main = "PM2.5 日均值 SARIMA 模型预测",
     xlab = "日期",
     ylab = "PM2.5 ( g/m³)")
```

```
  Augmented Dickey-Fuller Test

data:  pm25_ts
Dickey-Fuller = -5.327, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

4.3.2 gganimate 时序动图

这里我跑出来了一个 gif 图，展示了 2014 年 PM2.5 浓度的动态变化趋势，纸质版无法展示，可以在 git 上看。

```r
# 给pm25_daily加季节列
pm25_daily <- pm25_daily %>%
mutate(date = ymd(date)) %>%
mutate(season = case_when(
     month(date) %in% 3:5 ~ "春季",
     month(date) %in% 6:8 ~ "夏季",
     month(date) %in% 9:11 ~ "秋季",
     TRUE ~ "冬季"
))

p <- ggplot(pm25_daily, aes(x = date, y = pm25_daily, color = season)) +
geom_line(lwd = 1) + # 线条加粗
geom_point(size = 1.2, alpha = 0.8) + # 点标记
scale_color_manual(values = c("春季"="green", "夏季"="blue", "秋季"="orange", "冬季"="
    red")) +
labs(x = "日期", y = "PM2.5日均值 ( g/m³) ", title = "2014年PM2.5浓度动态变化") +
theme_bw() +
theme(text = element_text(family = "PingFang"), # macOS中文字体
```

```
        plot.title = element_text(hjust = 0.5)) +
  transition_reveal(date) # 核心：按日期滑动展示

  # 导出GIF
  animate(p, fps = 8, duration = 12, width = 900, height = 500)
  anim_save(file.path(path,"figures/PM25动态变化.gif"))
```