**中 国 地 质 大 学**

**实习报告**



课程名称 高级空间分析建模课程报告二

教师姓名 晁 怡

本科生姓名 童川博

本科生学号 20221000679

本科生专业 地理与信息科学

所在院系 未来技术学院

日 期 2024 年 6 月 4 日

目录

[1 实习准备 1](#_Toc168430799)

[1.1 实习任务 1](#_Toc168430800)

[1.2 数据来源 1](#_Toc168430801)

[1.3 数据说明 1](#_Toc168430802)

[2 数据预处理 1](#_Toc168430803)

[2.1 数据准备 1](#_Toc168430804)

[2.2 缺失值处理 2](#_Toc168430805)

[2.2.1 缺失值检查 2](#_Toc168430806)

[3 变量选择 4](#_Toc168430807)

[4 回归结果分析 7](#_Toc168430808)

# 实习准备

## 实习任务

1. 根据提供的数据分析全国地级市PM2.5浓度影响因素。
2. 进一步理解多元线性回归分析的原理，掌握利用软件工具（工具不限）进行多元线性回归分析的方法。
3. 能够对回归分析的结果结合问题背景进行解释。

## 数据来源

本次实习使用的所有数据均来自教师提供，未标明出处。

* 2018年286地级市PM2.5浓度数据
* 2019年中国城市统计年鉴(精修版)

## 数据说明

本次实习的目的是以城市为单位对全国城市的PM2.5影响因素进行探索性分析。统计结果为2018年全年数据，包括2018年全国城市行政区划，户籍人口与常驻人口，三次产业相关信息，工业相关排放和PM2.5浓度等指标。

# 数据预处理

## 数据准备

本次实习的所有数据处理均使用MATLAB实现，因此先将各数据表导入MATLAB并连接为一份表数据。之后再与PM2.5数据集进行连接，由于PM2.5是待回归的因变量，因此直接将缺失数据的城市舍去。同前次实习相同，对地名进行统一后连接，得到286个城市的数据。

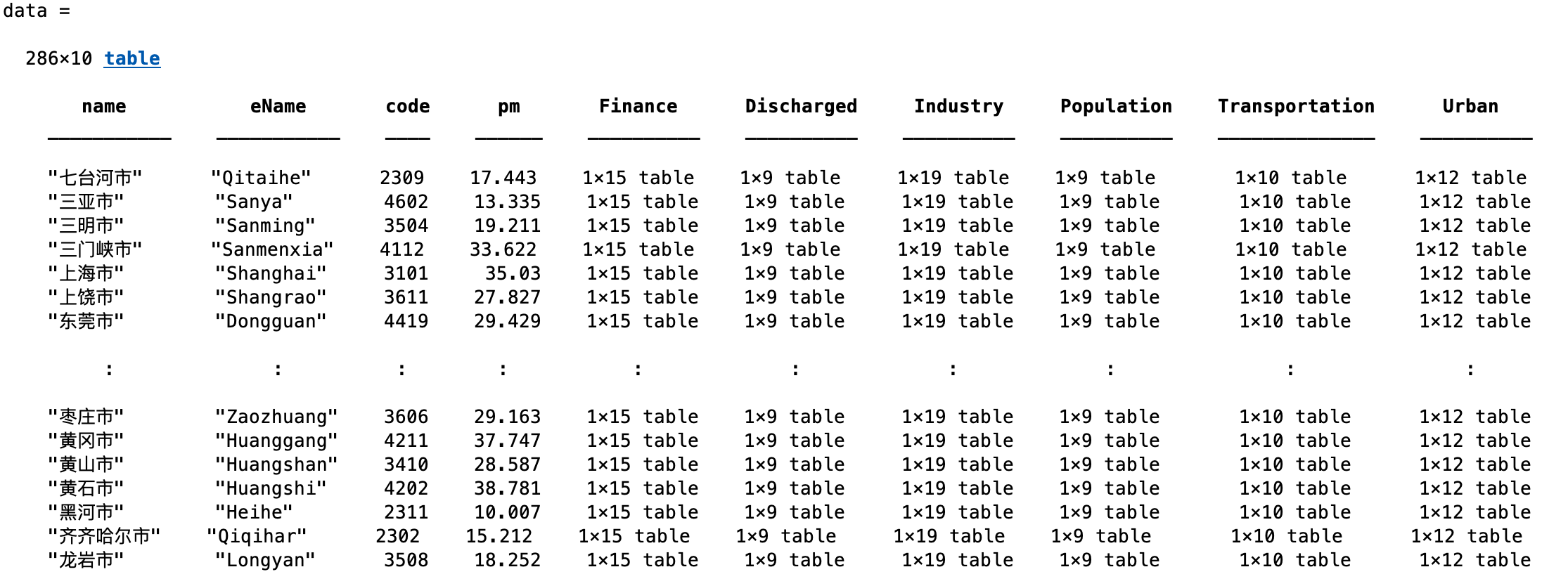


图 1初始数据集制作

其中初步将指标分为Finance, Discharged, Industry, Population, Transportation和Urban六类指标，具体指标如下：

表 1各类指标具体属性表

|  |  |
| --- | --- |
| 类别 | 指标 |
| Finance | GRP， GRPper ， GPgrowth， primaryGDPRate， secondaryGDPRate, teriaryGDPRate， bankDeposits， savingDeposits， residentialBuilding, volumeImport， volumeExport， currentAssets， insurance, telecommunication， internet |
| Industry | primaryPopulation, MiningPopulation, manufacturingPopulation, distributionPopulation, constructionPopulation, tradesPopulation, transportPopulation, servicesPopulation, informationPopulation, finacialPopulation, estatePopulation, bussinessPopulation, scientificPopulation, conservancyPopulation, otherPopulation, educationPopulation, headlthPopulation, culturePopulation, publicPopulation |
| Population | household, average, growthRate, , RD, employed, unemployed, primaryIndustry, secondaryIndustry, tertiaryIndustry |
| Transportation | busMaintain, busPassenger, busOperation, highwayPassenger, WaterwayPassenger, aviationPassenger, heighwayFreight, waterwayFreight, aviationFreight, postalFreight |
| Urban | birth, death, landArea, builtupArea, livingArea, constructionRate, greenRate, electricityTotal, electricityIndustrial, electrictyHousehold, gasSupply, liquefieldSupply |
| Discharged | waterDischarged, SO2Discharged, NO2Discharged, sootEmission, solide, concentration, primaryIndustry, secondaryIndustry, tertiaryIndustry |

## 缺失值处理

### 缺失值检查

首先检查各自变量的缺失情况，结果如图 2所示。



图 2各城市各指标缺失情况

可以发现大部分数据完整性较好，Finance, Population和Urban中的超过5%的缺失数据的指标直接舍去。

其中Discharged排放类指标大量缺失，而排放类指标本应当是计算PM2.5的重要数据。因此此时我将题目确定为“非排放因素对城市PM2.5浓度的影响”，并将所有排放类指标舍去。

而剩下的miningPopulation(从事采矿业人数), waterwayPassenger(水路客运量)，aviationPassenger(空运客运量)及对应的货运量也有大量缺失。但是这5个量的缺失却有很好的可解释性：有相当多的城市极少从事采矿业的人员，并有相当多的城市没有水港和航空港。因此此处将这5个指标填充0保留，其余数据填充均值。

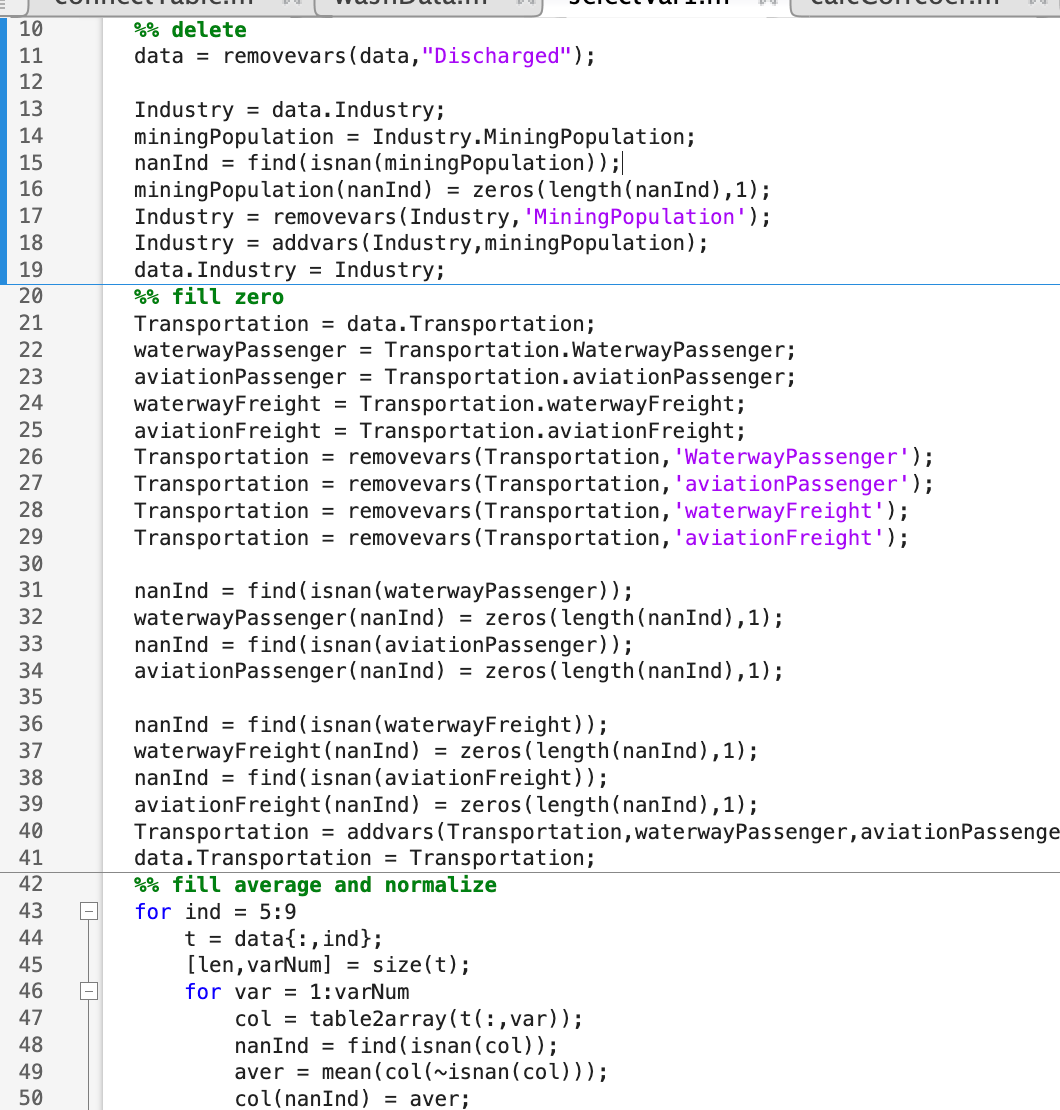


图 3数据预处理核心代码

# 变量选择

首先查看各类数据中不同指标的相关性，从而删去部分相关性很强的指标的若干个保留其中一个。对每一类数据进行皮尔逊相关系数分析，画出热力图如所示。并根据主成分分析结果确定每一类保留的变量数量，最终得到20个变量如表 2所示。

表 2根据相关性分析结果初步筛选变量

|  |  |
| --- | --- |
| 类别 | 指标 |
| Finance | GEPgrowth, primaryGDPRate, secondaryGDPRate, terinaryGDPRate. |
| Industry | primaryPopulation, manufacturingPopulation, servicesPopulation, |
| Population | household, growthRate, employed, secondaryIndustry |
| Transportation | busPassenger, highwayPassenger, aviationPassenger |
| Urban | birth, landArea, builtupArea, greenRate, electricityTotal , gasSupply, |

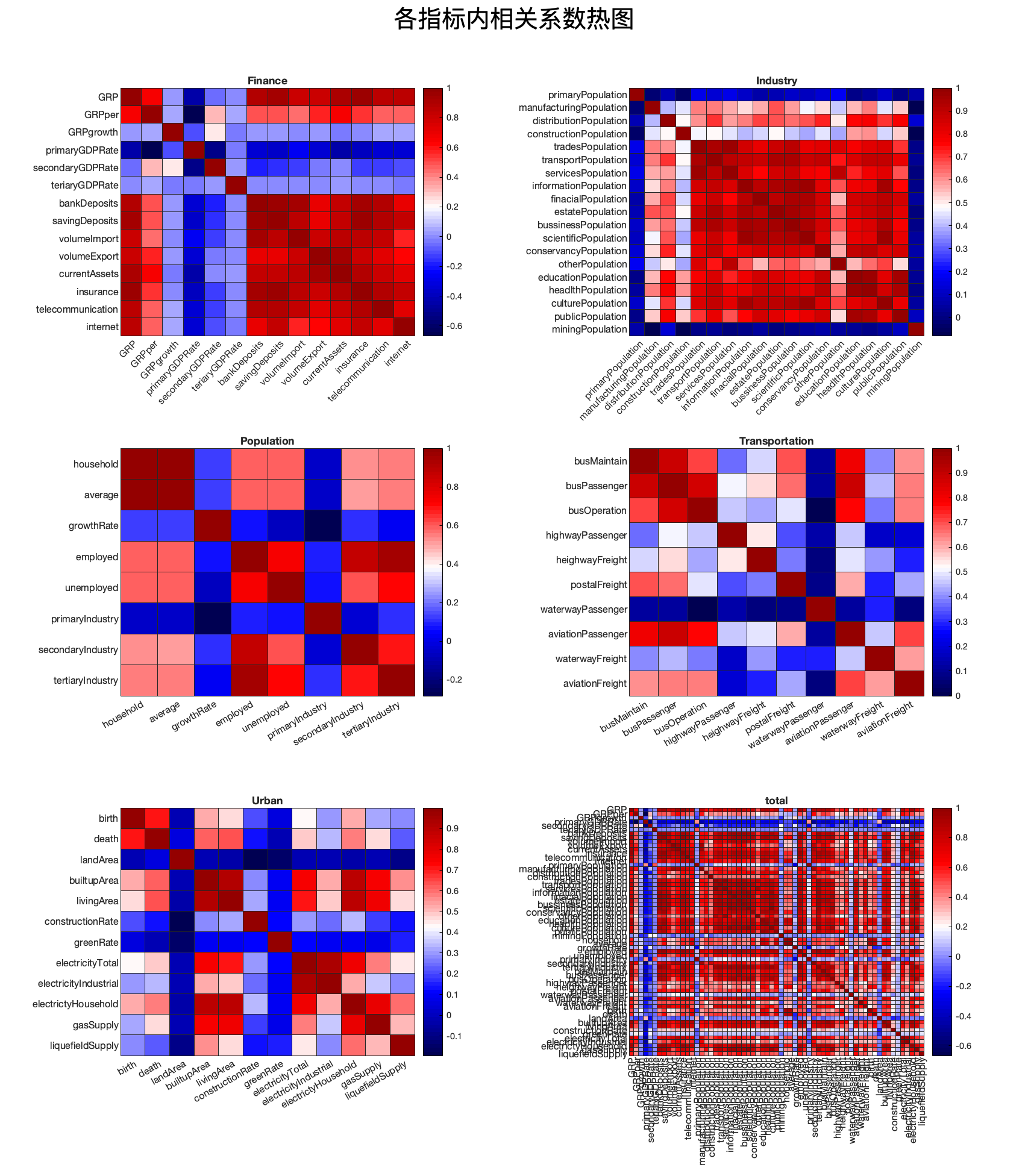


图 4各指标相关系数热力图

在筛选后合并各类别变量并统一考察，可以发现变量相关性得到明显改善，如图 5所示。

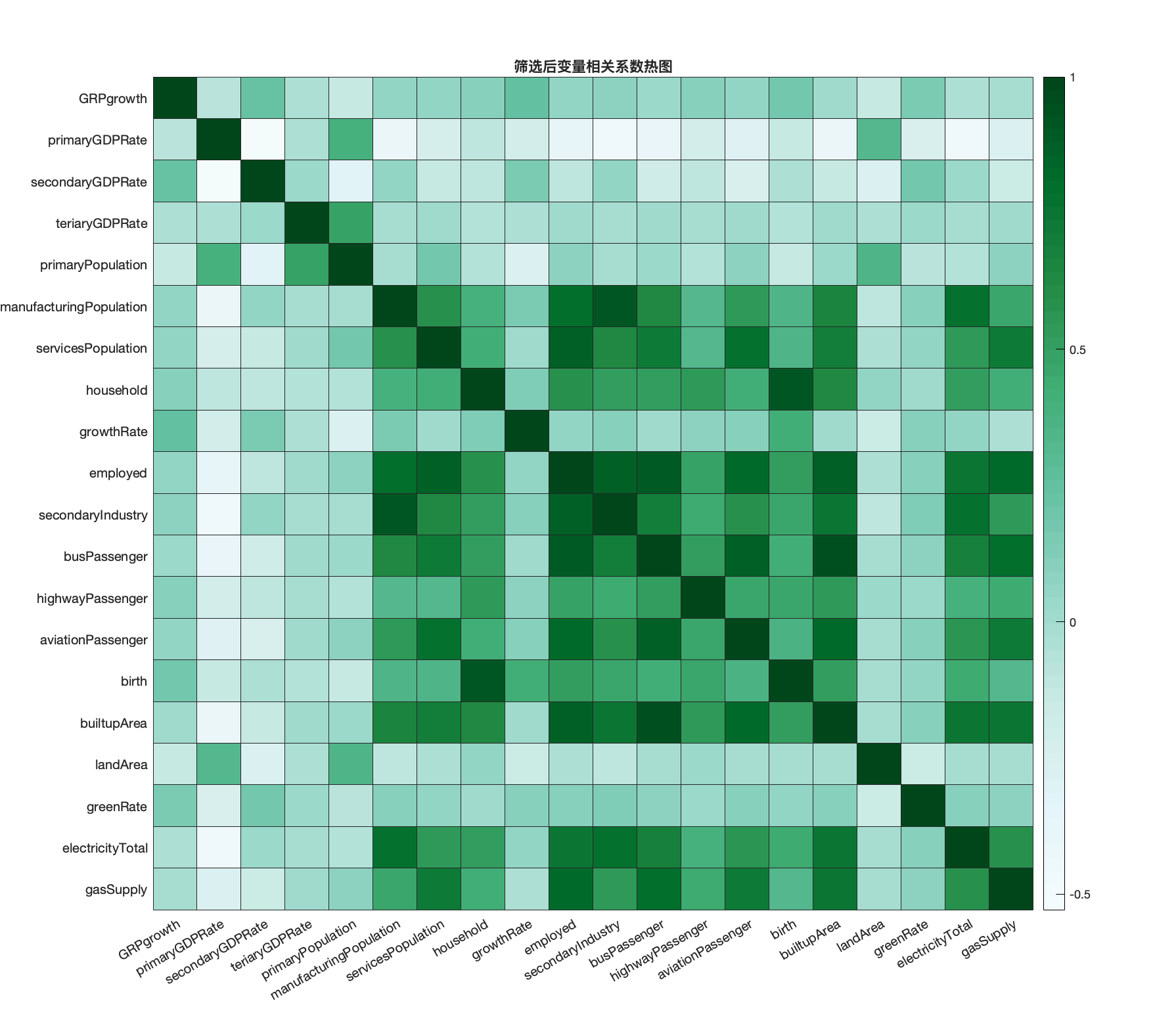


图 5筛选后相关系数热力图

之后建立最小二乘回归模型，对自变量逐个如图 6所示进行偏F检验。

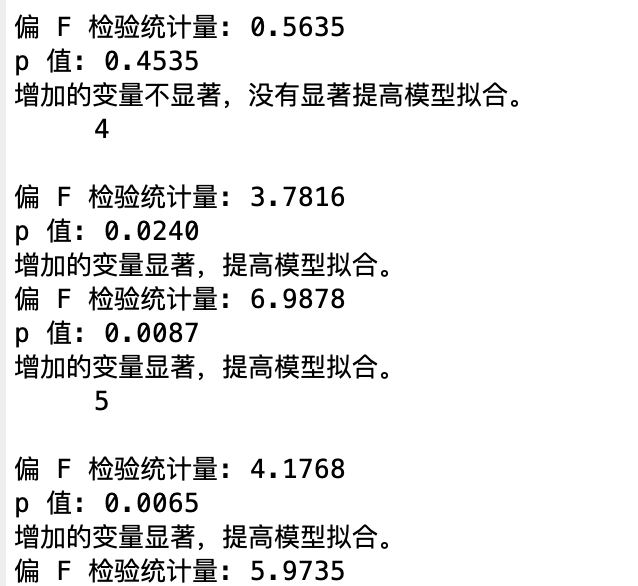
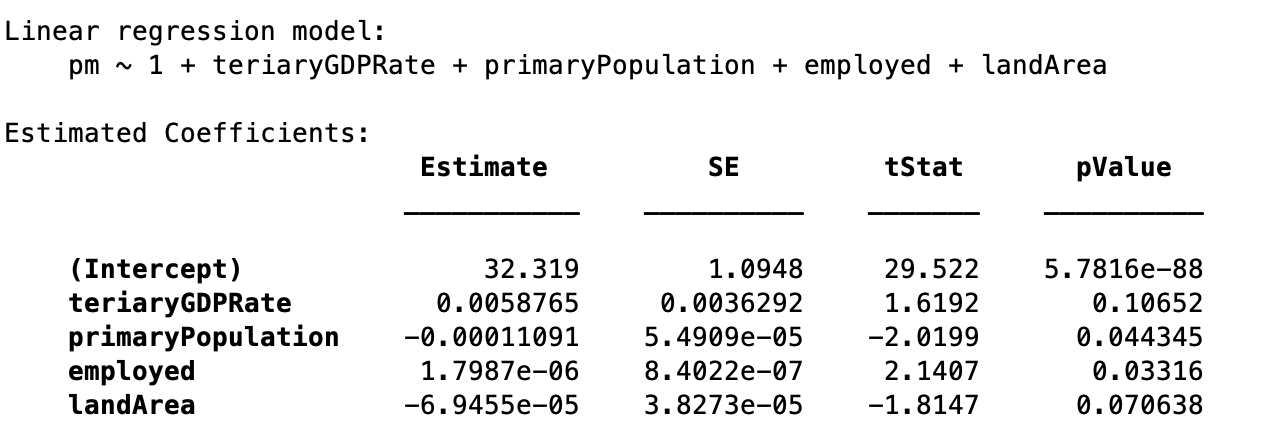


图 6偏F检验示意图

最终并综合变量的显著性指数，最终留下的变量为：



之后进行回归分析。

# 回归结果分析

拟合结果如图 7所示。可以发现拟合效果非常差，但是原数据却有明显的线性特征。说明“不考虑排放因素的模型”是无法有效反应PM2.5浓度的。

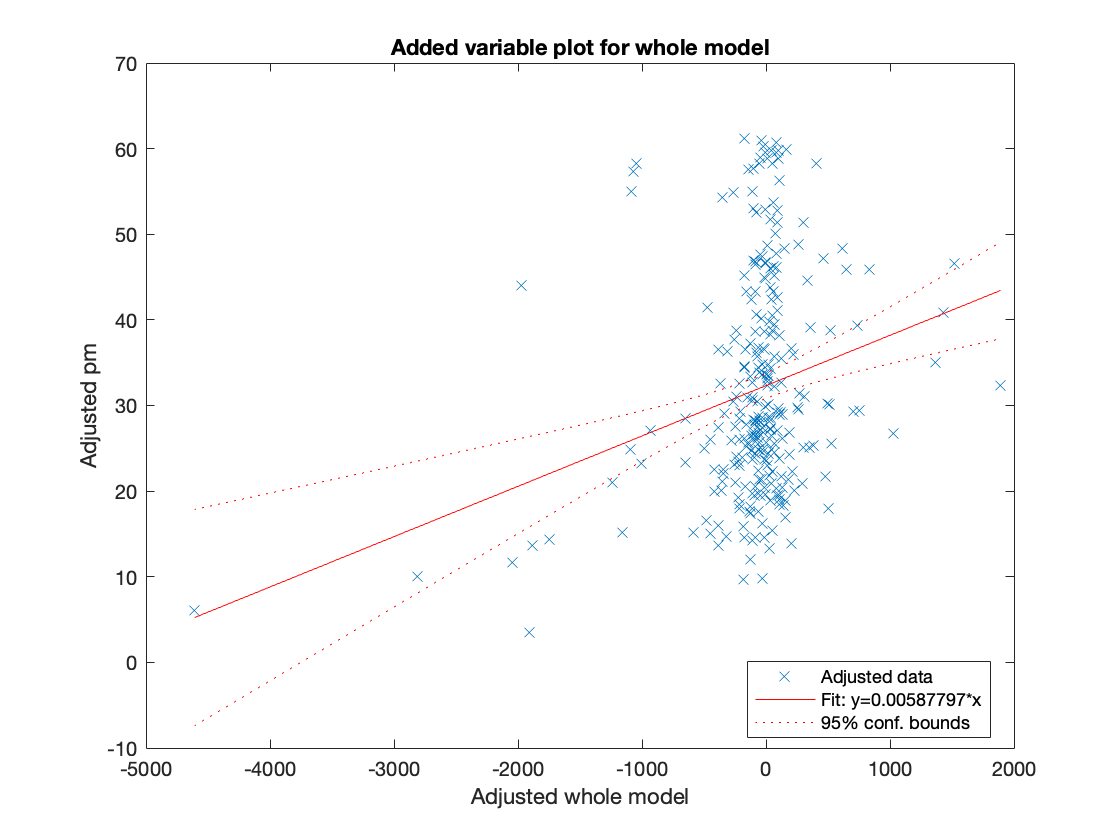


图 7最小二乘拟合结果

检验参数为：

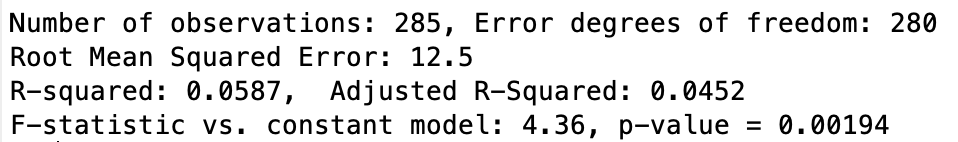


图 8模型有效性检验