北京大学

硕士研究生学位论文

题目：空气污染的统计和计量经济学实证研究

——数据挖掘北京 PM2.5 浓度的主要影响因素

姓　　名：冯婧

学　　号：1601214577

院　　系：国家发展研究院

专　　业：金融学

研究方向：环境经济学

导师姓名：胡大源

二零一八年四月

版权声明

【内容摘要】近些年来，政府部门及普通民众对于空气污染问题给予了越来越多的关注。PM2.5（可吸入颗粒物）是造成雾霾天气、降低能见度、影响交通安全的重要原因。随着有关部门公开环境数据，使得应用计量经济学方法分析环境问题成为可能。本文通过对 2016 年北京市农展馆监测点小时数据的分析，研究影响 PM2.5 形成及扩散的重要因素，希望能够找到刻画 PM2.5 影响因素的最简模型。首先从最小二乘入手初步观察回归结果，之后通过逐步回归法（stepwise）、Lasso 等新方法来帮助进行变量选择，并在此基础上研究 PM2.5 的形成和扩散机制，在自然科学理论的指导下进一步进行变量选择和模型修正。最终发现风速变量对 PM2.5 的影响存在滞后，诸多风向中北+北东北风的影响最为明显，能够显著降低 PM2.5 浓度值，而二氧化氮、二氧化硫、湿度等变量则会起到正向的作用。

【关键词】PM2.5　气体污染物　气象条件　变量选择　逐步回归　Lasso

Empirical econometric study of air pollution——Analysis about main influencing factors of PM2.5 in Beijing

Jing Feng

Directed by Dayuan Hu

Abstract

In recent years, government departments and people have paid more and more attention to the problem of air pollution. PM2.5 (inhalable particulate) can cause smog, reduce visibility and affect traffic safety. Its influencing factors and forming mechanism have been discussed in various subjects. After the relevant data has been published, it is possible to apply econometric methods to analyze environmental problems. This article through to the 2016 hours of Beijing nongzhanguan monitoring data analysis, studies the important factors influencing the formation and spread of PM2.5, hope to be able to select the variables, the optimal model to find the right. First of all, from the perspective of the least-squares regression, based on the regression results and basic principle of variable selection, through the method of stepwise regression, Lasso and new methods for variable selection, comparing with the results before choice, comparing the applicability and accuracy of different methods. Then, based on the selected basic variables, the model is modified to find the optimal model that affects PM2.5.In this process, the guidance of practical theory is very important, which enables us to clear the thread in complex variables and selection results, so as to help us better discover the laws of the real world.

Key Words: PM2.5 gaseous pollutants variable selection model setting stepwise Lasso

目录

空气污染的计量经济学实证研究

——北京市 PM2.5 浓度主要影响因素分析

一. 引言

近些年来，随着经济发展和人民生活水平的提高，人们对环境问题给予了越来越多的关注。然而事实上空气污染并不是近几年才出现的问题，空气污染早已存在，二三十年前的严重程度更甚于今日，只是那时的关注点仍更多地放在经济增长和提高收入上，随着生活条件的改善，环境及空气问题开始进入人们的视野，引起了广泛的讨论。研究表明：灰霾很可能取代吸烟，成为肺癌头号致病“杀手”，这也让大家对灰霾的危害格外关注。此外，气溶胶细粒子可对人体呼吸系统、心血管系统、免疫系统、生殖系统、神经系统和遗传系统产生有害影响，但机制非常复杂，目前很多影响和机制尚不清楚。[1]

细颗粒物（粒径小于 2.5 微米，PM2.5）是霾天气形成的主要因素之一。PM2.5 指空气动力等效直径等于或者小于 2.5 微米的大气颗粒。这样的颗粒物是造成雾霾天气、降低能见度、影响交通安全的主要因素。它还会吸附空气中的有毒有害物质，随呼吸进入肺部，长期处于 PM2.5 超标的环境会严重危害人体健康。最新医学研究表明，粒径在 2.5 微米及以下的细颗粒物不容易被鼻腔内的绒毛阻挡，细粒子绝大部分能通过人体支气管，直达人体肺部，甚至可以进入人体血液循环。[2]但是我国空气质量信息公开的历史还很短，监测指标近些年有所变化和调整，PM2.5 浓度值这个指标也是从 2013 年才开始公开发布。国外一些发达国家早在上世纪中期就开始研究空气颗粒物的形成并评估其负面影响。美国环保局于 1971 年发布总悬浮颗粒物（TSP）的控制标准，在长时间深入研究和慎重评估的基础上，1984 年美国环保局建议采用可吸入颗粒物(PM10)代替 TSP，于 1987 年发布 PM10 标准，并开始对其进行网络化监测。1997 年制定了 PM2.5 控制标准，随后开展网络化监测，公开监测数据，以便学术研究和公众监督。

我国早在上世纪 80 年代就开始对北京等主要城市空气中的颗粒物进行监测。1993 年北京申办第 27 届奥林匹克运动会以两票之差输给了悉尼，这一失败在很大程度上激励了政府有关部门下决心改善北京的空气质量，也推动了空气质量监测信息的公开。自 1994 年开始，北京开始每年发布环境状况公报，公布空气质量监测指标的年度平均值。从 1994 年到 1996 年，北京市环境状况公报中的空气质量监测指标包括总悬浮颗粒物（TSP）、二氧化硫（SO_2）和降尘。1997 年增加了氮氧化物（NOx）和一氧化碳（CO）。2000 年增加了可吸入颗粒物

[1] 吴兑：《探秘 PM2.5》，气象出版社，2013 年 3 月，第 53 页
[2] 吴兑：《探秘 PM2.5》，气象出版社，2013 年 3 月，第 53 页

（PM10），取消了降尘，用二氧化氮（NO_2）取代氮氧化物。TSP 的浓度数据一直延续到 2003 年。PM10 这个指标从开始发布一直持续至今，2013 年开始发布 PM2.5 浓度指标。

对于大气污染问题，政府部门也一直试图从源头出发来解决，希望通过有效遏制污染物的前体物来治理污染。在北京市各项空气污染治理措施中，耗资巨大的“煤改气”作用最为显著，空气质量监测指标二氧化硫持续改善。1994 年为 83 每立方米微克，1998 年曾上升到 120 微克，此后持续下降，平均每年降幅为 11.5%，2015 年北京年平均二氧化硫浓度仅每立方米 15 微克，低于我国环境空气污染物浓度限值 20 微克的一级标准。二氧化硫在空气中会通过一系列反应形成硫酸盐，而硫酸盐正是颗粒物的构成组分之一。因此，有效分析 PM2.5 的来源和形成机制，对于我们更高效和准确地治理空气污染问题具有重要的意义。

近年来，国内各地环境监测部门也逐渐将源解析纳入污染防治的核心业务。各界对于 PM2.5 的形成原因都有广泛的研究和讨论，污染气体排放与颗粒物的形成密切相关，而气象因素又会对颗粒物的迁移和扩散产生重要影响，因此寻找最佳模型来刻画 PM2.5 的形成过程就显得尤为重要。颗粒物的构成较为复杂，其中含有硫酸盐、硝酸盐、含碳组分等，而这些构成成分由于我们日常监测到的污染气体密切相关，如燃煤会产生二氧化硫，而汽车尾气中则含有二氧化氮和一氧化碳等气体。随着煤改气在华北地区的推进，燃煤减少从而导致二氧化硫排放下降，继而减少硫酸盐的形成，由此我们也对二氧化硫对颗粒物形成的重要性展开思考，各类环保举措是否能对颗粒物的治理卓有成效。

此外，气象条件对于天气状况也有较为重要的作用，风是其中最重要的影响因素，大风往往可以将颗粒物吹散，从而改善空气质量，北方来的风比南风更加洁净，能起到更好的净化作用，风速和风向都会对颗粒物的扩散产生很明显的影响，但是风速与污染物浓度之间还涉及一定的滞后关系，我们日常的观察中往往也可以发现，起风后仍需要一段时间，雾霾才可以渐渐消散。自从政府公开环境检测数据以来，我们能够更加容易地获取大量的观测数据，尽管各学科也都采用不同的方法来研究 PM2.5 的成因，但是迄今为止，对于空气污染问题，仍未能看到一个理想的模型，因此数据的公开为我们使用计量经济学方法对 PM2.5 的影响因素进行实证检验提供了可能性。

我们正处在 IT 技术高速发展的网络时代，每时每刻都会产生大量数据，大数据、云计算、机器学习等新技术也为处理海量数据创造了便利条件，各种数据挖掘方法在解决实际问题的应用中取得了预期的成效，获得了广泛的关注。与一些数据挖掘新方法相比，计量经济学已经过上百年的发展，各种计量经济模型的构建方面也逐渐走向精细化。

本文试图将计量经济学方法与传统的和新兴的数据挖掘方法相结合，应用于分析近年来

广泛收集的大量环境质量监测数据和气象观测数据，通过对比分析探索描述北京 PM2.5 浓度变化的既有效又精炼的影响因素模型。与此同时，我们还广泛查阅和整理了环境与气象学科领域的相关文献，跟踪了解有关 PM2.5 浓度变化的自然科学研究成果。这些研究成果在一定程度上为我们的计量经济模型的设定提供了指导。然而，由于许多研究成果是在特定环境下的实验室研究结果，有待于在现实环境中得到进一步的验证。近年来大量实时空气污染监测结果的发布，也为在错综复杂的现实环境下分析和验证 PM2.5 浓度变化的主要影响因素提供了良机。

二. PM2.5 影响因素研究现状

现在关于PM2.5的影响因素仍在讨论中，不同领域对于PM2.5的成分构成的研究方法存在一定差异。环境化学中比较普遍的有PM2.5的元素组成分析。如朱光磊等[3]，2000-2001年收集北京市区PM2.5颗粒，在元素分析的基础上进行有机物成分（PAHs）的分析，在获得源谱的基础上，利用CMB方法得出燃煤、机动车排放、二次硫酸盐等是PM2.5的主要来源。与此研究方法类似的还有陈添等[4]对北京市PM10的源解析研究。刘保献等[5]于2012年8月至2013年7月间在北京市城区石景山、车公庄、东四和通州四个点位开展了为期一年的PM2.5化学组分研究，得出结论主要组分为OM、EC、SO_4^{2-}、NO_3^-、NH_4^+、CI^-、地壳元素、微量元素，SO_4^{2-}成为本次污染过程中最主要的组分。

环境因素分析方面，首先在市区、郊区于不同季节和时间对PM2.5进行采样，然后利用AES等分析化学方法分析颗粒物元素组成，以及离子、有机物的含量。最后，或利用富集因子分析、或利用CMB等方法进行源解析。此类文章已大量发表，然而各个研究者使用的收集、分析仪器各异，采样的地点和时间段也各不相同，导致结论不一，权威性没有保证。值得注意的是，研究者们大多关注一次排放对PM2.5的贡献，比如土地扬尘、汽车尾气排放等。对于二次排放，研究者们通常只使用PM2.5颗粒中硫酸根的含量来代表。而PM2.5在大气中的二次转化生成受到很多条件的影响，比如光照、O_3浓度、SO_2浓度、NO_2浓度、湿度等，这些数据都是国家气象台长期观测的数据，比较容易获得。

除了研究化学组分构成，也有学者从气象因素去考察PM2.5的转移、扩散等物理过程，研究天气情况与PM2.5浓度的关系。在前人关于气象条件对PM2.5影响的研究中，我们发现

[3] 朱光磊，张远航，曾立民等.北京市大气细颗粒物 PM2.5 的来源研究[J]. 环境科学研究，2005，18(5):1-5.
[4] 陈添，华蕾，金蕾等.北京市大气 PM10 源解析研究[J]. 中国环境监测，2006，22(6)：59-63
[5] 刘保献，杨懂艳，张大伟等.北京城区大气 PM2.5 主要化学组分构成研究[J]. 环境科学，2015，36(7): 2346-2352

了一些重要的气象变量，比如湿度，湿度和PM2.5浓度有显著的正相关关系，但当湿度过高时可能会由于降水反而使PM2.5受到冲刷导致浓度下降。此外，风速则与PM2.5的浓度有比较明显的负相关关系，其中西北风对PM2.5浓度的降低有最显著的作用。另外有学者认为在不同季度，气象条件对于PM2.5的影响程度有所不同。比如冬季，PM2.5的浓度与气温、湿度、日照长度正相关、与风速负相关；春季PM2.5与气压显著负相关，夏季则与水汽浓度关系较为密切。朱珠[6]等利用Mann-Kendall趋势检验，表明PM2.5秋冬浓度较高，春夏浓度较低；其余降水量、风速呈现出负相关且与气压呈显著正相关，但与气温的显著性不强。该文作者只对PM2.5浓度和其他天气因素做了单因子的分析，并没有获得各个因素的偏效应，只相当于计量经济学当中的简单一元回归，由于解释变量缺失，这种分析所得出的结果是有偏的，因此模型还需要改进。此外还有学者提出不同城市应该区别分析，因为不同城市的气象条件对PM2.5的影响程度不同。同时不同城市的对比看出，PM2.5浓度越高的地区，气象因素能够解释的部分就越少。此外，还有一部分文章是以定性分析为主，通过观察特定天气条件下PM2.5的变化来归纳各个天气对PM2.5的清除或积累效果。此类研究只针对为数不多的案例进行研究，因此主观性过强，缺乏对各种天气共性的分析，也缺乏不同天气对PM2.5贡献率的定量分析。[7]

从前人的研究中，我们大概对于影响PM2.5形成的因素有了初步的判断，也对影响其变化的气象条件有了一定的了解，但是之前的研究中作者多是寻找变量之间的相关关系，没有将计量经济学的方法严谨地应用于气象数据中进行分析，也并无对PM2.5影响因素的模型设定进行讨论的前例，虽然这些研究已经描绘出了一个基本的轮廓，但仍还没有特别合意的模型出现，因此本文会在前人的基础上进行更深入探讨，寻找更为合适且简约的模型。接下来的部分会介绍本文所使用的数据及模型设定和构建的过程，首先会使用经典计量经济学方法初步探讨变量的选择。之后，随着现在处理大量数据能力的提高，其他的一些方法也可以为我们的变量选择和模型设定提供借鉴和参考。

统计学教授Efron在2016年出版的《计算机时代的统计推断》（“Computer Age Statistical Inference”）一书中认为：21世纪见证了统计学方法的惊人扩展，无论是规模还是影响上，随着统计学方法应用于大规模的现代科学和商业数据当中，‘大数据’，‘数据挖掘’，‘机器学习’这些概念已经耳熟能详。Efron从频率派经典统计推断方法入手，梳理了半个多世纪

[6] 朱珠等，《PM2.5/PM10 浓度变化规律及其气象条件分析——以深圳市龙岗区为例》，2014 中国环境科学学会学术年会（第六章），2014 年 8 月

[7] 李思麒，曹瀚尹，胡大源：《北京市 PM2.5 与相关污染性气体影响因素探究》》

以来统计推断的发展思路，特别是随着计算机技术的迅速发展而兴起的若干数据挖掘新方法，并且对未来数据科学的发展方向做出了判断。

Efron在介绍数据挖掘时对照比较了逐步回归法和Lasso这两种回归模型变量选择的方法。我们首先对近年来收集到的大量北京空气污染和气象因素的每小时观测数据进行描述性统计分析和简单关联分析；然后，结合已有的自然科学研究结果，估计、对比和探讨上述方法在解释变量选择和回归模型设定方面的作用和局限性。在此基础上，根据计量经济学关于时间序列数据建立回归模型的假设，对模型做进一步的检验和修正，以期得到精炼且更为有效的的PM2.5影响因素解释模型。

三. 数据及影响因素初探

本文主要使用的是 2016 年 1-12 月北京市农展馆监测点的大气浓度小时数据及相应时间段的气象小时数据，采用了互联网数据抓取方法获得，每天共 24 个观测值，总观测值数为 8712 个，主要变量及相应解释见表 1。PM2.5 浓度为被解释变量，PM10、二氧化硫、二氧化氮、臭氧、一氧化碳等污染气体及风速、相对湿度、温度等气象条件为解释变量，同时控制风向的影响作用，根据风向 16 方位图，共有 17 种风（包括静风），由于涉及虚拟变量过多，因此将相邻方向合并，最终合并成八种风向，每一种类型作为一个虚拟变量。

表 1 主要变量与解释

变量	解释
PM2.5	PM2.5 细颗粒物质量浓度，单位 μg/ m^3
PM10	PM10 颗粒物质量浓度，单位 μg/ m^3
SO2	SO_2 质量浓度，单位 μg/ m^3
NO2	氮氧化物质量浓度，单位 μg/ m^3
O3	臭氧质量浓度，单位 μg/ m^3
CO	一氧化碳质量浓度，单位 mg/ m^3
windspeed	风速 m/s
humi	湿度（百分比）
temp	温度（摄氏度）
pres	气压（帕）
风向（8 种）	虚拟变量
N_NNE（北+北东北）	0/1
NE_ENE（东北+东东北）	0/1
E_ESE（东+东东南）	0/1
SE_SSE（东南+南东南）	0/1

S_SSW（南+南西南）	0/1
SW_WSW（西南+西西南）	0/1
W_WNW（西+西西北）	0/1
NW_NNW（西北+北西北）	0/1

下面将对本文关注的被解释变量 PM2.5 及不同类型的解释变量进行详细介绍，讨论不同解释变量与 PM2.5 可能存在的解释关系。

1. 被解释变量

PM2.5 是本文所关注的被解释变量，其浓度值在 2016 年的变化如图 1 所示，横轴为时间，纵轴为 PM2.5 浓度值，单位是 μg/m^3。从上图中可以看出，PM2.5 在 2016 年最高浓度达到接近 600μg/m^3，主要分布在年初的 1-2 月及年末的 12 月，这几个月份 PM2.5 浓度值波动较大，有较高的极端值，而 12 月不仅波动幅度大，整体浓度值也相对较高。6-8 月 PM2. 浓度值整体较低，始终保持在 300μg/m^3 之下，且波动并不明显。

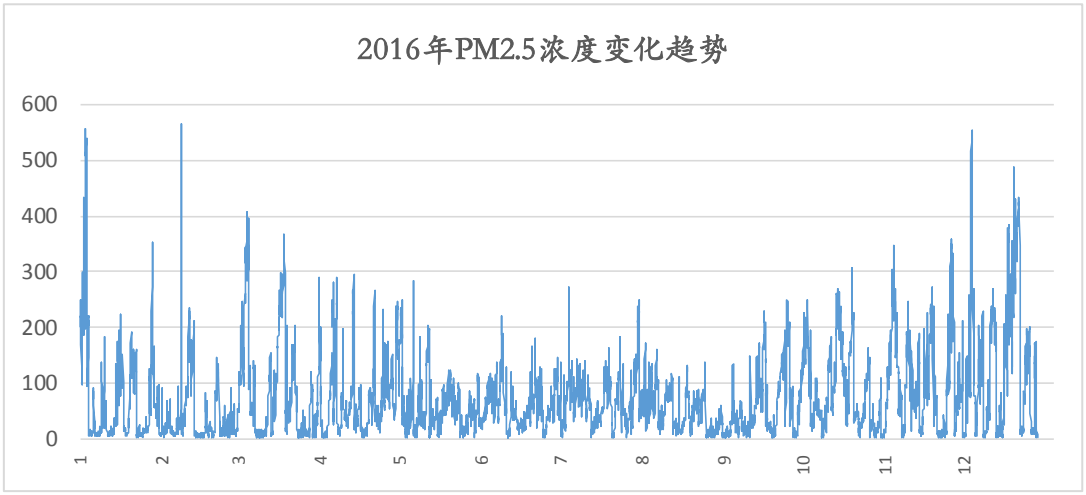


图 1 ： 2016 年 PM2.5 浓度变化趋势

2. 解释变量

（1） 气体污染物

PM10 是指空气动力学当量直径≤10 微米的颗粒物，许多影响 PM2.5 的变量也会对 PM10 造成影响，在前人研究中普遍将 PM10 作为重要的解释变量，然而二者存在包含关系，相关系数非常高，为 0.9044，尽管 PM10 与 PM2.5 有相同的组成部分，但是 PM10 并不是 PM2.5 的影响因素，因此将 PM10 作为解释变量并不合适，在很大程度上会分散其他相关解释变量对 PM2.5 的解释力度，对参数造成影响，因此不应将该变量包含在最终模型之中。

此外，硫酸盐与硝酸盐是 PM2.5 的重要组成部分，而硫酸盐与硝酸盐的形成主要来源是气体转化，因此对应的 SO_2 与 NO_2 浓度也是重要的影响变量，关于 PM2.5 的形成机制，后文会作更为详细的介绍。从图 2 及图 3 中可以看出，SO_2 和 NO_2 与 PM2.5 整体呈正相关关系，NO_2 与 PM2.5 的关系相对更加明显，在 NO_2 浓度值小于 135μg/m^3 时，波动较小，正向关系较为稳定，当浓度值大于 135μg/m^3 时，开始产生较大幅度的波动，这可能是由于高浓度观测值数量较少的缘故。SO_2 与 PM2.5 的关系在下图看来不如 NO_2 明显，在浓度超过 45μg/m^3 时，关系的波动愈发明显，SO_2 浓度在 45~100μg/m^3 之间时，PM2.5 的浓度值基本上围绕在 200μg/m^3 左右，并未随着 SO_2 浓度的增加而增加，当 SO_2 浓度值超过 100μg/m^3 时，PM2.5 的浓度值陡然增加。

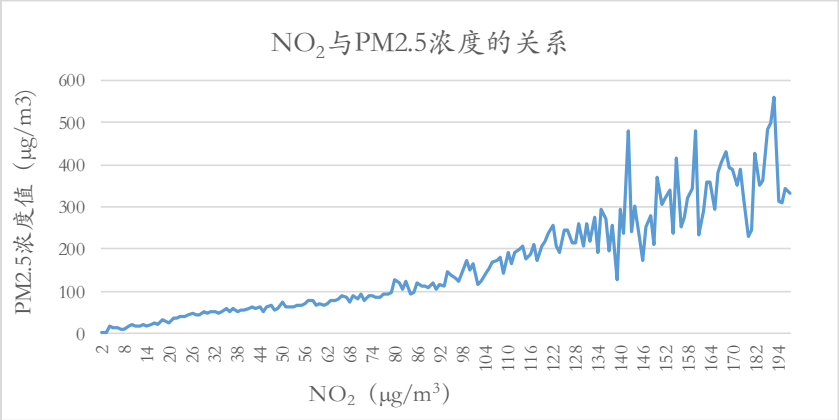


图 2：二氧化氮与 PM2.5 浓度的关系

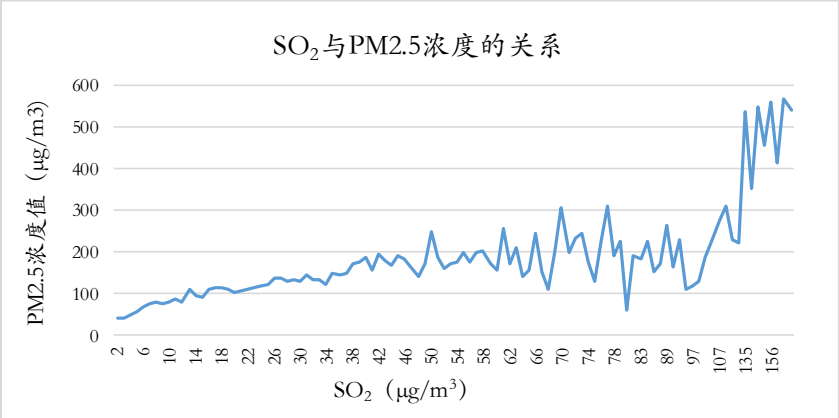


图 3：二氧化硫与 PM2.5 浓度的关系

关于 O_3 和 CO，臭氧作为一种污染气体，会刺激呼吸道，引起气道反应和气道炎症增加、哮喘加重等。与 PM2.5 有所不同，臭氧作为一种氧化物，会对人体产生更为直接的刺激作用和急性危害。臭氧的形成主要是由于以下要素：由发电厂、燃煤锅炉等排放出的“氮氧化物”和“挥发性有机物”；高温和光照。O_3 与 PM2.5 并列作为两种污染物，呈现出一定的

相关关系可能是由于在二者的形成过程中，“氮氧化物”和温度都发挥了一定的作用，但 O_3 本身并不是 PM2.5 的影响因素。CO 是一种有毒性气体，其来源主要是工矿企业、交通运输、家庭炉灶等，而城市大气中 86%的一氧化碳是由汽车排出，汽车车速越大，一氧化碳的排放量越小。虽然碳元素是 PM2.5 的重要组成部分，但是根据相关同位素组成研究表明，PM2.5 中的碳元素主要是来自于机动车尾气烟尘或燃油锅炉烟尘[8]，且其二次转化是由碳氢化合物氧化，与一氧化碳并无关系。但从下图中可以看出，CO 与 PM2.5 有比较明显的正相关关系，尤其是在 CO 浓度小于 6.4mg/m^3 时，而当浓度超过 6.4mg/m^3 时，PM2.5 浓度值则稳定在 400μg/m^3 左右，不再具有明显的正相关关系。而 O_3 与 PM2.5 在下图中并不能看到明显的线性关系。

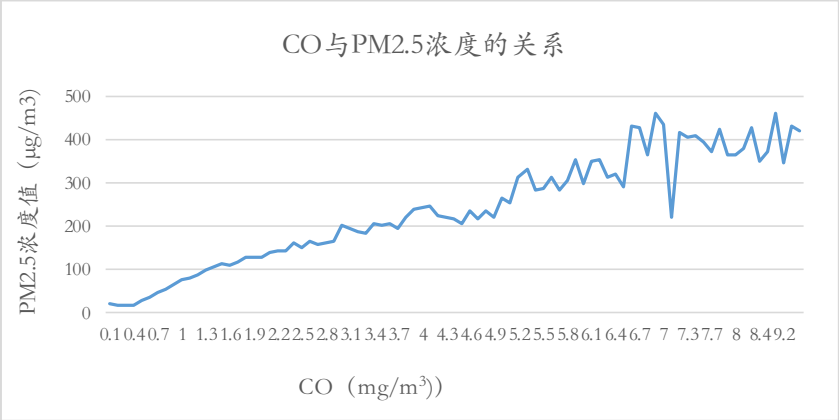


图 4：一氧化碳与 PM2.5 浓度的关系

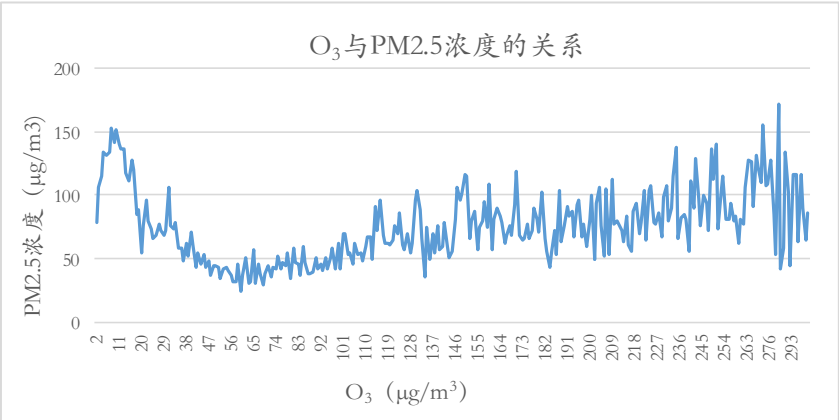


图 5：臭氧与 PM2.5 浓度的关系

（2） 气象变量

至于气象因素，风是影响 PM2.5 最重要的变量，无论是从风速还是风向来看，风在污

[8] 刘刚等，《杭州市大气 PM2.5 中有机碳元素碳的同位素组成》，《科学通报》，2007 年 8 月

染物扩散和消除的过程中都有不容小觑的作用。灰霾天气出现时，一般都伴随着静风、小风、强日照和合适的相对湿度（60%~85%）。总的来说，扩散条件不好、空气流动性不强、风速不大，使城市中各种污染物无法得到及时扩散，并在近地面积聚，若再加上日照强烈、湿度合适，污染物之间就容易发生各种光化学反应，最终导致灰霾。自然界中，霾和雾是可以互相转化的，当相对湿度增加达到或接近饱和时，霾粒子吸附析出的液态水成为雾滴；而相对湿度降低时，雾滴脱水后霾粒子又再[9]悬浮在大气中。

从图 5 中，可以看到风速与 PM2.5 浓度的关系，二者总体上看呈负向的线性关系，在风速小于 13m/s 的时候，PM2.5 的浓度严格随风速的增大而降低，但从 13m/s 到 14m/s 时，出现了反向的提升。这与现实中风对于 PM2.5 的作用原理是一致的，在小风速的条件下，风速对 PM2.5 的扩散有影响，同时也能对空气起到搅拌作用，有利于使气溶胶中 PM2.5 进行聚沉；而当风速更大时，空气将呈现湍流的特性，从而地面上的 PM2.5 将被吹起，并且速度较大的风很有可能会从其他地区带来沙土从而导致 PM2.5 浓度值的升高。

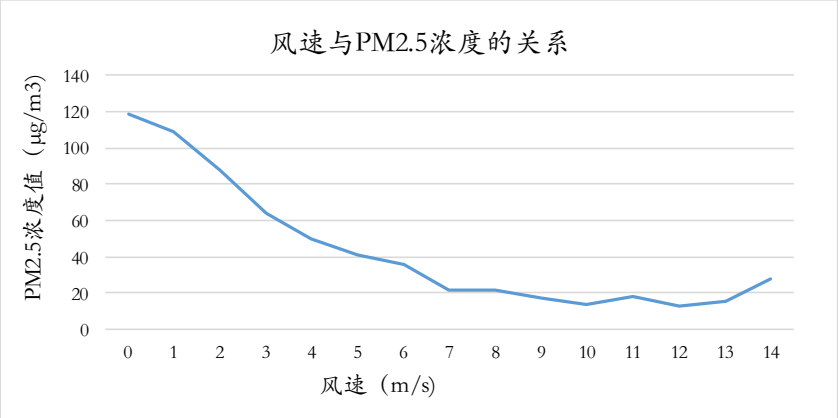


图 6：风速与 PM2.5 浓度的关系

图 6 为湿度与 PM2.5 浓度的关系，当相对湿度在 0~60%之间时，随着相对湿度的提高，PM2.5 浓度值有比较明显的增加，这是由于一方面水分子的存在能够催化空气中氮氧化物反应的发生，另一方面 PM2.5 具有吸水性，有可能会导致测量浓度大于其真实浓度。当湿度大于 60%之后，二者的关系不再稳定，有比较明显的波动，相对湿度为降水满足了水汽输送的条件，因此有一定概率会增加降水的可能性，从而对 PM2.5 产生了冲刷作用。

[9] 吴兑：《探秘 PM2.5》，气象出版社，2013 年 3 月，第 40 页

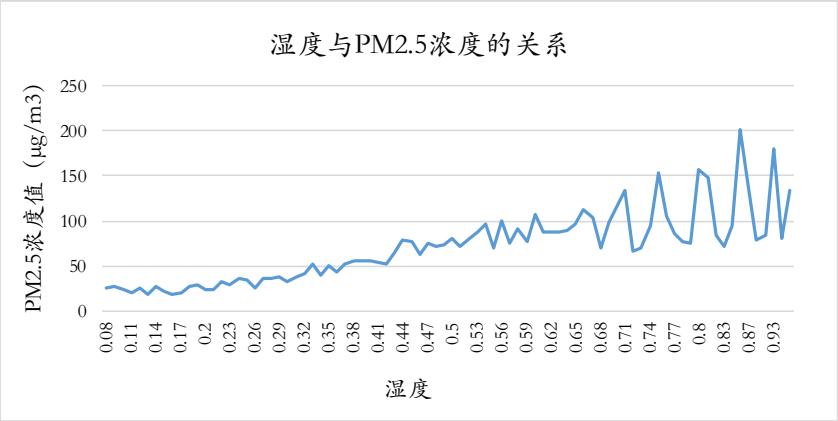


图 7：相对湿度与 PM2.5 浓度的关系

除风速外，风向也是 PM2.5 的重要影响变量，对于北京来说，不同方向的风对 PM2.5 的扩散作用不同，北京北部由于没有污染源，污染程度较低，而京南为工厂聚居地污染源较多，因此从北方来的风会更好地起到净化作用。此处通过雷达图来更为清晰地观察 2016 年风向的分布，右图为风向 16 方位图，风向的划分共有 17 种（包括静风），由于涉及虚拟变量过多，因此将相邻的两种风向合并为一种，这样以静风为对照组，就划分了 8 种风向，每种风向都作为一个虚拟变量。总体看来，北风频率相对较高，由于北风污染程度较低，往往可以起到净化空气的作用。

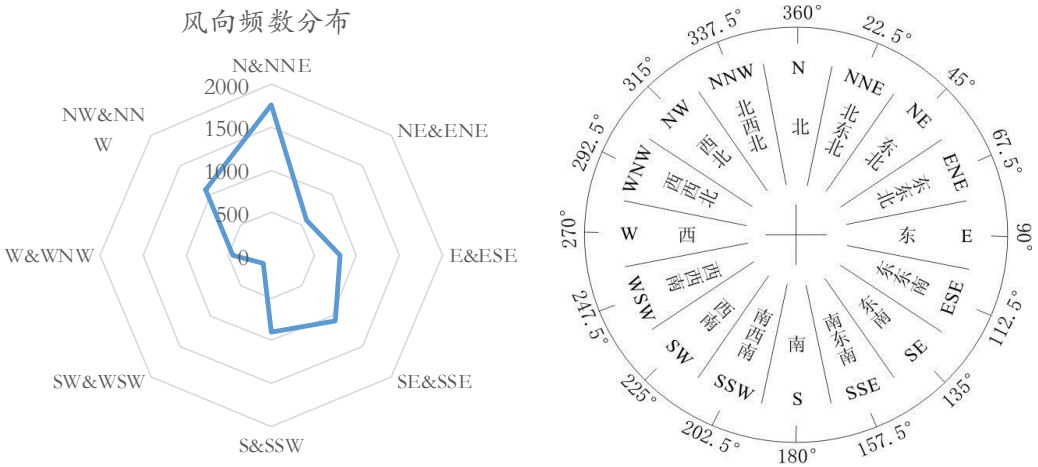


图 8：风向频数分布

（3） 控制变量

供暖：在图 1 中可以看出 PM2.5 在不同季节间的浓度值有非常明显的差异，总体趋势为上半年逐渐降低，下半年再逐渐升高的抛物线状。因此应该对 PM2.5 时间上的差异加以控制，由于北京市秋冬季节燃煤供暖，造成 PM2.5 前体物大量排放，从而导致 PM2.5 浓度较

高，所以将供暖季 11-3 月作为控制变量。

温度：下图则是温度与 PM2.5 浓度的关系，总体上看二者并无明显的线性关系，在 0 摄氏度以下，温度与 PM2.5 整体是正向关系，在 0 摄氏度以上，PM2.5 浓度开始随着温度的升高缓慢下降，通过以往的研究可以发现：温度较高时有利于大气垂直对流，加快颗粒物扩散，颗粒物的日均浓度低；而温度较低时，近地面容易形成逆温层，不利于颗粒物扩散，颗粒物的日均浓度高。[10]但持续性的高温会导致光化学反应等持续性光污染发生，生成 O3 等强氧化性物质，容易出现高浓度 PM2.5[11]。因此导致温度对 PM2.5 浓度的影响方向并不明确。但温度在颗粒物形成和扩散的过程中仍是不可忽视的重要变量，因此将其作为控制变量。

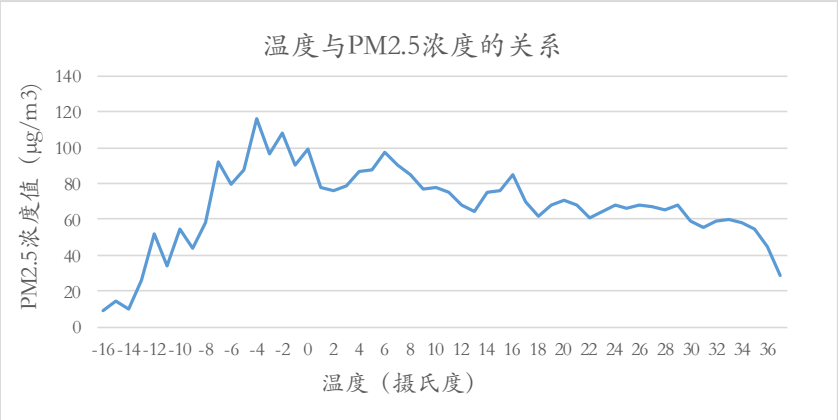


图 9：温度与 PM2.5 浓度的关系

3. 主要解释变量与 PM2.5 关系初步总结

表 2：解释变量与 PM2.5 关系总结

变量名称	关系总结
NO_2	（正向）NO_2 与 PM2.5 浓度整体呈正相关关系，虽然当 NO_2 浓度值大于 135μg/m^3 时，开始产生较大幅度的波动，但是正向关系总体比较明显。
SO_2	（正向）SO_2 与 PM2.5 的关系不如 NO_2 那样明显，随着 SO_2 浓度值的增加，PM2.5 的浓度先增加，然后保持稳定，在超过 100μg/m^3 时有了陡然增加。
CO	（正向）CO 与 PM2.5 有着非常明显的正相关关系，这可能是由二者的排放来源高度一致所造成的。
O_3	（方向不明）未呈现出明显的线性关系。
风速	（负向）随着风速增大，PM2.5 的浓度值不断下降，但当超过 13m/s 时，风速的增大反而使 PM2.5 浓度值增加。

[10] 田谧：《京津冀地区霾污染过程大气 PM2.5 及前体物变化特征研究》，北京化工大学，2013 年
[11] 宋宇，唐孝炎，张远航等：《夏季持续高温天气对北京市大气细粒子（PM2.5）的影响》，《环境科学》，2002（4），33-36 页

风向	（方向不明，北风应为负向）北京市全年中北风相对较多，而污染程度较低的北风可以对空气起到良好的净化作用。
湿度	（正向）相对湿度在 0~60%之间时，随着相对湿度的提高，PM2.5 浓度值有明显的增加，当湿度大于 60%之后，二者的关系不再稳定，有较大波动。
温度	（方向不明）未呈现出明显的线性关系，不同方向的影响可能会产生抵消。

4. 数据描述性统计

表 3 是文中主要被解释变量及解释变量的描述性统计结果，PM2.5、PM10、NO_2、SO_2、O_3 的单位均为 μg/m^3，CO 样本均值较小是由于其单位为 mg/m^3，与前几类气体污染物选取的衡量单位有所差异。PM10 的缺失值较多，因此样本观测值较其他变量更小。PM2.5 的均值为 74.89μg/m^3，最大值为 566μg/m^3。风速从 0m/s 到 14m/s 不等，平均风速为 2.98m/s，风向均为虚拟变量。

表 3：颗粒物、污染性气体浓度及气象变量分布统计结果

变量	观测值	均值	标准差	最小值	最大值
PM2.5	8,708	74.89	76.40	3	566
PM10	5,759	96.81	89.53	5	995
CO	8,437	1.20	1.15	0.1	9.4
NO2	8,488	52.10	32.63	2	206
SO2	8,550	11.51	14.47	2	187
O3	8,543	60.65	58.00	2	390
windspeed	8,712	2.98	2.22	0	14
humi	8,712	0.56	0.25	0.08	1
temp	8712	13.34	12.10	-16	37
heat	8712	0.41	0.49	0	1
风向（8 种）	8712	-	-	0	1

四. 模型构建

1. 初步回归结果

首先，我们将全部解释变量置于模型中，来考察进行多元线性回归时，各解释变量对于 PM2.5 的影响情况。表 5 中第（1）栏包含了 PM10、CO、NO_2、SO_2、O_3 等气体污染物及风速、风向、湿度、温度等全部气象变量，同时也控制了供暖季的影响。从（1）中结果可以看出，除风向外的解释变量均对 PM2.5 浓度有正向的影响，其中 CO 和相对湿度系数绝对值较大是由于其单位分别为 mg/m^3 和百分比转化，与其他变量的单位有所差别。气体污

染物的影响方向与我们之前所观察的一致，但是风向的符号也为正，该结果与现实情况严重违背。此外，共有两种风向显著，但是西+西西北风的符号却为正。当将全部变量置于模型中时，我们发现一些基本变量的符号与事实不相符合。

表 4：回归结果 1

VARIABLES	(1) PM2_5	(2) PM2_5
PM10	0.463***	
	(0.00552)	
CO	18.53***	37.25***
	(0.506)	(0.592)
NO2	0.260***	0.789***
	(0.0183)	(0.0226)
SO2	0.372***	0.661***
	(0.0297)	(0.0389)
O3	0.154***	0.276***
	(0.00962)	(0.0120)
windspeed	0.555**	4.047***
	(0.221)	(0.295)
humi	33.86***	56.66***
	(2.004)	(2.467)
temp	0.127**	0.754***
	(0.0545)	(0.0726)
N_NNE	-0.0832	-11.13***
	(1.066)	(1.370)
NE_ENE	0.689	-3.430*
	(1.422)	(1.855)
E_ESE	-1.288	-2.824*
	(1.256)	(1.634)
SE_SSE	-2.723**	-3.877**
	(1.215)	(1.556)
S_SSW	-0.453	-4.888***
	(1.292)	(1.724)
SW_WSW	-0.546	-3.930
	(2.669)	(3.567)
W_WNW	4.047**	-6.131***
	(1.748)	(2.334)
NW_NNW	1.777	-6.776***
	(1.483)	(1.925)
heat	3.400***	9.335***
	(1.182)	(1.655)
Constant	-49.26***	-87.94***

	(2.178)	(2.879)
Observations	5,661	8,329
R-squared	0.897	0.768

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

虽然前人研究中普遍将PM10作为主要解释变量，但是由于PM10与PM2.5的包含关系，这种做法应该谨慎。PM10 的存在会分散和干扰其他解释变量的解释力度。第（2）栏展示了去掉PM10后的回归结果，所有变量系数的绝对值都有所提高，更多风向结果显著，前一个回归中系数为正的西+西西北风符号变为负，西北+北西北风符号也从正转负，可以看出去掉PM10后对模型有一定的修正作用，但是风速变量的符号为正，这一点仍有待进一步研究和解决。

2. 经典计量经济学中关于模型设定的讨论

当直接面临大量的环境数据和相关变量时，该选择哪些进入模型着实令人困惑。“现代计量经济学的关注点更多是在给定模型框架下进行估计和推断，却较少地涉及到变量选择和模型构建的问题”[12]，主要是由于这一问题本身就较为复杂，正如 Hendry 和 Richard（1983，p.112）所说：“数据生成过程是复杂的，数据是稀缺的且具有不确定的关联，实验是难以控制的，可用的理论是高度抽象的且很少是没有争议的”，此外，很难找到评价模型是否正确设定的最佳方式，因此很多经济学家认为这是一个颇具创造力和想象力的过程，“为计量经济学家建造‘系统神学’可能会扼杀创造力”（Pagan，1987，p.20）。

尽管设定问题非常复杂，但在实际研究中，在模型设定这一环节仍有一些方法作为指导。最主要的两种方法是平均经济回归（AER）和检验，检验，检验（TTT）法。术语 AER 是从 Gilbert（1986）的研究中得到的，该方法是指从一个被认为正确的设定开始，使用主要数据来确定未知参数大小的顺序，如果这些复杂的方法不能解决问题，那么就会进行设定检验，使用正确的符号、高的 R^2 值、“已知”异于零的系数的显著 t 值等标准来进行判断，这种方法也被称为“向上检验”，从一个简单模型出发至一种特定的更一般的模型。对于 AER 模型，Johnston（1984，pp.498-510）强调研究者需要同模拟这一领域的专家交流、熟悉相关的研究所、切实去看数据、认识到数据的局限、避免数据的挖掘、使用经济学理论，最重要的是利用有经验的批评家的判断。

[12] Peter Kennedy, *“A Guide to Econometrics”*, 2008, p.71

第二种方法是 TTT 法，这种方法让初始设定的模型比最终选择的模型更加一般化，并且进行不同约束的检验，来简化这个一般化的设定，这种检验也被称为“向下检验”，从一个一般的模型到一个更特殊的模型。然后，模型遵从一组诊断检验，寻找该模型被错误设定的迹象，模型不断地被再设定，直到一组诊断检验允许研究者得出在特定标准下令人满意的结论。术语 TTT 的来源是参考 Hendry（1980，p.403）的研究中常用的引语：“计量经济学的三条黄金法则是检验、检验和检验。”TTT 方法的一个重要的方面是数据应该被允许帮助确定模型设定，尤其是如滞后长度等模型特性，关于这些经济学理论几乎没有提供指导。然而，数据挖掘早先的评论认为让数据为自己说话是很危险的事情。Belsley（1986a）赞成在设定分析中对先验信息的使用，Belsley 对此认为要在让数据帮助设定和不让数据支配设定之间寻找一个平衡，但是这种抽象的概念又再一次使得模型设定“艺术化”，难以找到一种明确的标准来衡量。

计量经济学建议我们在熟悉相关领域知识的基础上，在模型设定中使用先验信息，避免数据的挖掘和让数据自己说话。但是大数据时代，在一些新兴领域中，往往很难得到对某一问题的先验信息，这时让数据说话就成了解决问题的重要突破口，人工智能和数据挖掘等新方法为科学研究提供了思路和方向。那么数据挖掘的新方法能否给我们对于环境数据的研究提供一些变量选择方面的借鉴？受到 Efron 在关于变量选择问题相关讨论的启发，接下来我们尝试使用逐步回归和 Lasso 进行变量选择，来考察这两种方法能否为我们在变量选择的过程中提供一定的解决思路。

3. 数据挖掘与变量选择

（1） 逐步回归（stepwise）

逐步回归法出现的时间较早，分为向前和向后两种单方向筛选法，也可以将两种方法结合筛选。许多统计计算软件中均有现成的程序可以调用，在医学、工程学等学科的应用研究中广为使用。向前逐步回归的基本思想是：将变量逐个引入模型，每引入一个解释变量后都要进行 F 检验，并对已经选入的解释变量逐个进行 t 检验，当原来引入的解释变量由于后面解释变量的引入变得不再显著时，则将其删除。以确保每次引入新的变量之前回归方程中只包含显著性变量。这是一个反复的过程，直到既没有显著的解释变量选入回归方程，也没有不显著的解释变量从回归方程中剔除为止。以保证最后所得到的解释变量集是最优的。其基本步骤为先选定一个标准，然后按自变量对 y 的贡献大小由大到小依次挑选进入方程，每选入一个变量进入方程，则重新计算方程外各自变量对 y 的贡献，直到方程外变量均达不到入

选标准，没有自变量可被引入方程为止。但问题是一开始变量太少可能会缺失主要变量，从而得到有偏的结果。向后逐步回归与之相反，先将变量全部引入，再逐步缩减变量，这种方法则有可能会带来多重共线性的问题。

逐步回归往往用于缺乏成熟理论指导情况下的回归方程解释变量选取。然而，无论是向前还是向后逐步回归，在理论上均有其无法克服的弱点。因此，长期以来，逐步回归方法在计量经济学应用中备受争议。本世纪以来，随着计算机技术和网络经济的迅速发展，面对大量数据和许多缺乏理论指导的新问题，机器学习和数据挖掘备受关注。作为机器学习和数据挖掘的主要方法之一的逐步回归在金融、营销和许多企业管理均有许多新的应用实例。

近年来大数据、云计算和人工智能的兴起使我们有必要进一步审视逐步回归方法在解决实际问题时的可用之处及其局限性。

表 5：逐步回归结果

Significance	**0.1**			**0.01**		
Step	Variable entered	Parameter estimate	Standard error	Variable entered	Parameter estimate	Standard error
1	CO	37.32***	0.590	CO	37.30***	0.590
2	NO2	0.785***	0.0225	NO2	0.788***	0.0224
3	O3	0.272***	0.0115	O3	0.275***	0.0114
4	Humi	56.78***	2.432	Humi	56.82***	2.432
5	SO2	0.661***	0.0386	SO2	0.666***	0.0385
6	Windspeed	3.620***	0.264	Windspeed	3.428***	0.240
7	Temp	0.744***	0.0724	Temp	0.757***	0.0720
8	N_NNE	-8.332***	1.106	N_NNE	-7.759***	1.056
9	heat	9.135***	1.650	heat	9.194***	1.650
10	NW_NNW	-2.646*	1.520			

表 5 为逐步回归的最终结果，左侧序号为依次进入模型的顺序。将显著性水平定为 0. 1 时，有 10 个变量被选入了模型，当显著性水平定位 0.01 时，有 9 个变量进入了模型，二者的差别在于西北+北西北风向这个变量，而其他变量的进入次序并没有变化。进入的变量依次为一氧化碳、二氧化氮、臭氧、相对湿度、二氧化硫、风速、温度、北风+北东北风、供暖。逐步回归的结果与 OLS 相似，CO 由于和 PM2.5 的高度相关性首先被选入了模型。另外，各解释变量的系数及数值也与 OLS 的结果非常一致，但是逐步回归为我们选择哪种风向进入模型提供了一定帮助。

尽管逐步回归提供了一种看似便捷的变量选择方法，但是，它的选择结果仍是有比较明

显的问题，并不能避免 OLS 回归中带来的困惑，从上面的回归结果中可以看出，风向的符号仍为正向，因此变量选择和模型设定还是存在问题。因为逐步回归是一个逐步引入变量的过程，首先会引入 t 值最大的变量，即 CO，随后再逐步引入变量，以保证最终模型中的每个变量都是显著的，但是却不能说明这些变量是服从同一个 F 分布的，因此也很难基于此结果再进行下一步的推断。从前文的统计性描述中可以看出，CO 与 PM2.5 的浓度值有高度的相关性，PM2.5 对 CO 进行单变量回归时，其 t 值也达到了 133.78，因此 CO 出现了与 PM10 相类似的问题。尽管逐步回归在面临大量变量时可以帮助我们快速筛选损失最小的模型，但是对于其结果并不能够完全依赖，还应加以现实情况进行综合判断。

（2） **Lasso**

逐步回归帮助我们在诸多风向的虚拟变量中做出了选择，但是模型的设定仍存在一定问题，导致风向的符号与现实情况不符。下面我们用一种更为新颖的变量选择的方法来进行尝试。

Lasso（the least absolute shrinkage and selection operator）是一种压缩估计。它通过构造一个罚函数得到一个较为精炼的模型，使得它压缩一些系数，同时设定一些系数为零。因此保留了子集收缩的优点，是一种处理具有复共线性数据的有偏估计，它较好地解决了普通最小二乘中存在的一些问题。普通最小二乘（OLS）通过最小化残差平方和来进行估计，但 OLS 存在一些不令人满意的问题，一是预测精度，OLS 往往偏差较低但是方差大；二是可解释性的问题，在大量的预测值中，我们通常想确定一个展现出最强影响的更小的子集。两个公认优秀的改善 OLS 估计的方法是子集选择（subset selection）和岭回归（ridge regression）。子集选择提供了可解释的模型但是可变性非常强，因为它是一个离散的过程——回归变量要么保留要么从模型中去掉。小的数据变化就会使得模型的选择改变，这会降低预测准确度。岭回归是连续缩小参数的过程，因此更稳定，然而它不会使得任何参数为 0，没办法得出简单的可解释的模型。Lasso 就此提出同时缩小和设置参数为 0，则同时保持了子集选择和岭回归的特征。Lasso 计算损失函数的具体约束式为：

$$\text{Minimize}\ \ \frac{1}{n}\sum_{i=1}^{n}(y_i-\beta_0-x_i^{'}\beta)^2\ \ \text{subject to}\ \|\beta\|_1\le \text{t} \qquad (1)$$

Lasso 与逐步回归的差别主要表现在两个方面，一是当面临数目庞大的变量时，Lasso 比逐步回归的处理能力更强，另一个方面是逐步回归没有号称最优化的过程。而 Lasso 则是

有更加标准的最优化过程。$\|\beta\|_1 \leq t$ 通过将系数拉向 0 来进行限定，这种方式可以减小方差，避免过渡拟合。

Lasso 和岭回归的差别在于 Lasso 设定了所有系数绝对值之和小于某个常数，而岭回归则是设定了所有系数平方之和小于某个常数，在图 5 中可以看出二者的差别，Lasso 的限定范围是一个四边形，而岭回归是圆形，这样就导致了 Lasso 方法下会出现某些变量系数为 0 的情况，从而相当于做出了变量选择，而岭回归则不会出现系数等于 0 的情形。

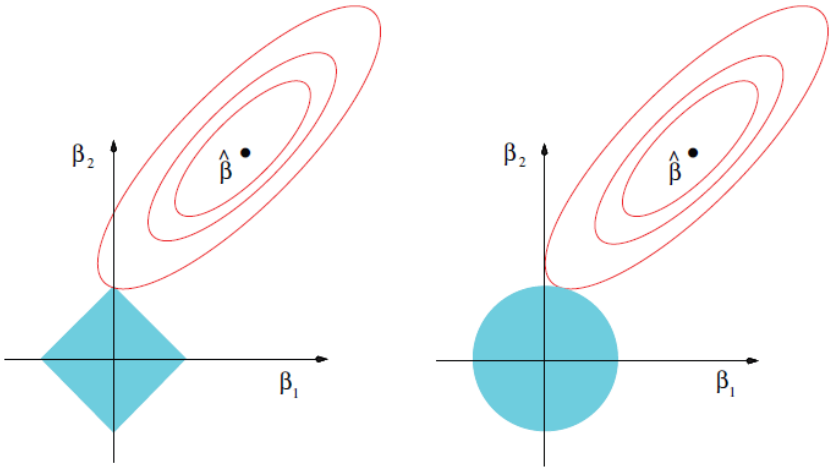


图 10：Lasso 与岭回归的比较

为了增强预测精度，模型会将一部分变量的系数设置为 0，而其他系数不为 0 的变量则会在不同的节点上被依次选入。从上表中可以看出，在第一个节点上，一氧化碳最先进入，随后是二氧化氮和二氧化硫，在之前的逐步回归中，一氧化碳和二氧化氮也是最先进入的变量。之后是相对湿度、温度和臭氧在接下来的三个节点依次进入，且系数为正，这里的结果与之前的 OLS 和逐步回归结果均一致。在此之后，北+北东北方向的风、风速最后进入模型，风速的符号问题仍未被解决。

表 6：Lasso 结果（1）

变量	系数	节点
CO	0.37506	1
NO2	0.6697	2
SO2	0.56477	3
humi	0.40584	4
temp	0.37382	5
O3	0.23948	6
N_NNE	-5.4913	7
windspeed	1.0072	8

下表是逐步回归和 Lasso 两种方法变量选择情况的对比，二者选出的变量总体上看较为一致，从第 3 个变量开始有顺序的差异，在 Lasso 中 O3 进入顺序较晚，SO2 和温度变量相对更加靠前，风向变量在两种方法中的选择结果一致，均选出了北+北东北风，供暖变量在 Lasso 中未被选入。

表 7：逐步回归和 Lasso 变量选择结果对比

Step	1	2	3	4	5	6	7	8	9
Stepwise	CO	NO2	O3	humi	SO2	Windspeed	Temp	N_NNE	heat
Lasso	CO	NO2	SO2	humi	Temp	O3	N_NNE	Windspeed	

上面涉及到的两种方法中，逐步回归已具有较久的使用历史，而 Lasso 是逐渐为人所熟知的一种较为新颖的方法，二者均涉及到了变量选择的问题，从上面的结果中看，二者的选择大体相似，也为我们变量的筛选提供了一定思路，接下来的模型中，可以保留北+北东北风这一种风向，将其他风向统一作为对照组，可以减少模型中虚拟变量的个数，达到简化模型的目的，但是我们最初面临的风向符号相反的问题仍未解决。通过数据挖掘的方法让数据说话，的确提供了一些思路，但是模型的设定仍让人存有困惑。因此，下面我们回到对 PM2.5 形成机制的讨论上来，进一步分析各解释变量对于 PM2.5 的真实影响。

4. PM2.5 形成机制分析

研究表明，细粒子 PM2.5 成因复杂，约 50%是来自燃煤、机动车、扬尘、生物质燃烧等直接排放的一次细颗粒物；约 50%是空气中的二氧化硫、氮氧化物、挥发性有机物、氨等气态污染物，经过复杂的化学反应和光化学反应形成的二次细颗粒物。细颗粒物的来源十分广泛，既有火电、钢铁、水泥、燃煤锅炉等工业源的排放，又有机动车、船舶、飞机、工程机械、农机等移动源的排放，还有餐饮油烟、装修装潢等量大面广的面源排放。也有一小部分是植物排放的挥发性有机物通过光化学反应转化而来的。灰霾的本质是细粒子气溶胶污染，主要是 PM1，考虑到标准的引用和现阶段的科技发展水平，当前可以界定为 PM2.5.在人类活动强度不太大的时候，霾主要是自然现象，霾的前身主要是尘卷风、扬沙、沙尘暴吹起的沙尘，当风速减小之后，有下降末速度的巨大颗粒物很快沉降，留下较细的尘粒子在空中就会出现浮尘，以上天气现象都可以追溯到明显的沙尘源，再演变下去，经过一段时间，

或者高浓度尘粒子原理沙尘源区之后，所谓没有明显能识别的沙尘源时，当层结稳定使尘粒浓度增加到一定程度而影响能见度时，就出现了霾。

大气颗粒物由各种人为源和自然源所排放的大量化学组成复杂的物质构成，主要包括有水溶性离子组分、含碳组分和无机元素的化合物，[13]同时还含有一定的地壳物质和痕量元素。

PM2.5 的来源可分为自然源及人为源，而从形成上看，PM2.5 中既包括直接排放的一次颗粒物，也包括气态污染物反应形成的二次颗粒物。其中一次颗粒物分为自然源与人为源，自然源包括沙尘天气、地面扬尘、海盐、这些污染源受天气和风速的影响较为明显。自然源排放出的颗粒物本身毒性较小，对人体危害不大，而人为源包括工厂等固定源和汽车尾气等移动源，含有有毒有害物质较多，同时排放出的污染物有可能会吸附在颗粒物上，进而增强其毒性。二次源则来自空气中污染性气体化合、氧化过程生成的颗粒性物质，其中包含氮氧化物气体的氧化、二氧化氯的氧化成盐等过程。水分子在空气中的存在能够催化氧化反应的发生，同时 PM2.5 的吸水性也可能造成其测量浓度大于真实浓度。

表 8：PM2.5 化学组分的主要来源(贺克斌等,2011)

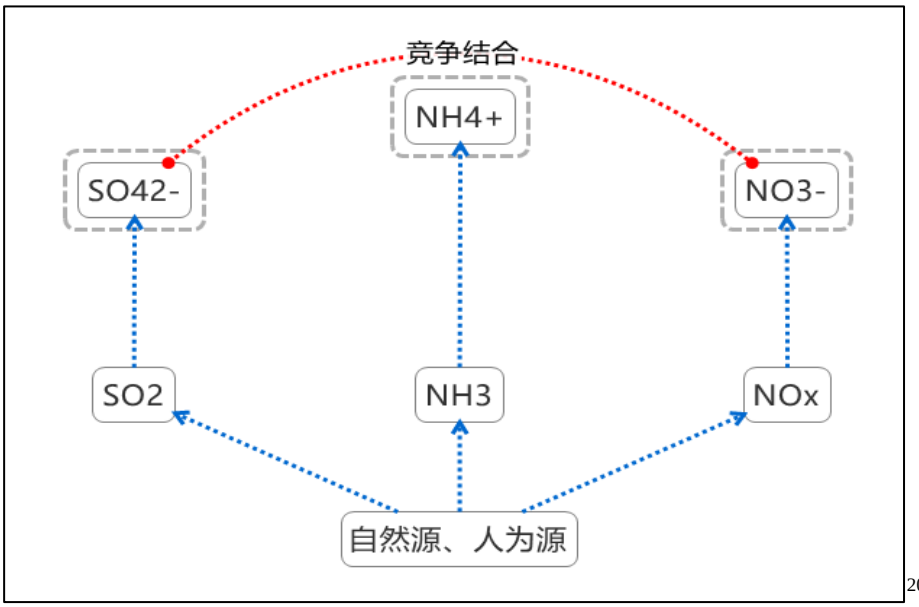
成分	一次源		二次源	
	自然源	人为源	自然源	人为源
SO_4^{2-}（硫酸盐）	海浪沫	化石燃料燃烧	S, SO_2, H_2S（海洋，湿地，火山）的氧化	化石燃料排放的 SO_2 氧化
NO_3^-（硝酸盐）	-	机动车排放，大型燃烧源	土壤，森林火灾产生的 NO_x 氧化	化石燃料、机动车排放的 NO_x 氧化
NH_4^+（铵盐）	-	机动车排放	野兽，未开垦土地	饲养，施肥土地释放
OC（有机碳）	野火	露天燃烧，烹调，机动车排放	植物释放的碳氢化合物氧化（如萜、蜡）	机动车，露天燃烧，碳氢化合物氧化
EC（元素碳）	野火	机动车排放，木材燃烧	-	-

水溶性离子是 PM2.5 的重要组成部分，主要包括硫酸盐、硝酸盐、铵盐、氯化钠、有机酸等离子组分，一般占 PM2.5 质量的 30-50%。[14]大气颗粒物中的水溶性离子直接影响大气降水的酸度，同时也是许多大城市能见度降低的主要原因，是导致大气复合污染的重要物种之一。[15]在水溶性离子组分中，SO_4^{2-}、NO_3^-、NH_4^+（合称 SNA）是三种最主要的组分，主要由气粒转化形成，其浓度与相应的气态前体物、反应转化率有关，并受温度、湿度等气

[13] 贺克斌等：《大气颗粒物与区域复合污染》，科学出版社，2014
[14] He K, Yang F, Ma Y, Zhang Q, Yao X H, Chan C K, Cadle S H, Chan T, Mulawa P A. The characteristics of PM2.5 in Beijing, China[J]. Atmospheric Environment 2001, 35, 4959–4970.
[15] 贺克斌等：《大气颗粒物与区域复合污染》，科学出版社，2014

象因素影响，[16]构成一个复杂的无机气溶胶体系。

主要水溶性组分 SO_4^{2-}、NO_3^-、NH4+之间互相影响，与 H2O 构成复杂的二次无机气溶胶体系。夏季大气中氧化剂浓度高（如臭氧）、气温较高，光学氧化性增强，二次无机气溶胶转化率通常较高。[17]也有研究发现，SO_4^{2-}、NO_3^-在夏季浓度比较低，而冬季则较高，这可能与冬季燃烧源的排放加强、污染物集聚不易扩散有关。[18]在 PM2.5 的水溶性组分中，我们关注三种最重要的离子组分 SO_4^{2-}、NO_3^-、NH4+。在相关性分析中，发现 SO_4^{2-}、NO_3^-与 NH4+高度相关，推测 PM2.5 中的 SNA 可能有相同的形成过程，即由气态前体物经由气相或液相反应形成。[19] SO_4^{2-}（硫酸盐）主要通过 SO_2 气体氧化形成，通常以铵盐、硫酸形式存在；NO_3^-（硝酸盐）主要是由 NO_x 在大气中发生均相反应形成 HNO_3，之后再与 NH_3 气体或已有颗粒物反应生成，但其中的 NH4NO3 在高温低湿条件下也容易分解为气态；NH4+（铵盐离子）由氨(NH3)在酸性颗粒表面反应或凝结形成，多与硫酸根、硝酸根结合。



[20]

图 11：二次污染形成机制

含碳组分是大气颗粒物的重要化学组分，也是大气复合污染的关键化学物种之一，一般占 TSP 浓度的 15-25%，占 PM2.5 浓度的 20-60%(贺克斌等,2011)[21]，主要包括 OC（有机碳）与 EC（元素碳）。与水溶性离子相比较，含碳组分组成复杂，含有各种结构复杂的化合物；

[16] Sailesh N. Behera, et al. Insights into Chemical Coupling among Acidic Gases, Ammonia and Secondary Inorganic Aerosols. Aerosol and Air Quality Research, 2013.13:1282-1296.
[17] Robarge W P, lker J T, McCulloch R B. Atmospheric concentrations of ammonia and ammonium at an agricultural site in the southeast United States[J]. Atmospheric Environment, 2002,36: 1661-1674.
[18] 曹军骥等：《PM2.5 与环境》，科学出版社, 2014.
[19] 贺克斌等：《大气颗粒物与区域复合污染》，科学出版社，2014
[20] 杨笑寒：《中国 PM2.5 组分关系与时空差异研究》
[21] 贺克斌等：《大气颗粒物与区域复合污染》，科学出版社，2014

来源及形成途径复杂，包含多种自然源和人为源，同时一次排放的气态前体物还会在大气中进行化学转化；环境影响复杂，对气候、能见度、人体健康都有影响。

从源头上来看，灰霾的成因，主要与化石燃料的燃烧相关。人类活动排放颗粒态污染物，比如水泥厂、发电厂、冶炼厂、工业窑炉都会直接排放颗粒物，汽车尾气会直接排放黑碳粒子，人类活动也会排放二氧化硫、氮氧化物、挥发性有机物（或者说碳氢化合物）等气态污染物。二氧化硫被氧化后会形成硫酸盐，氮氧化物和挥发性有机物在太阳紫外光的照射下发生光化学反应，这些反应导致臭氧浓度升高，最终生成如过氧乙酰硝酸酯（PAN）等新的气态污染物，进而转化为硝酸晶粒和有机硝酸盐等二次气溶胶，这些物质都是气溶胶细颗粒物，可造成能见度的恶化，也就是所谓的灰霾天气。[22]

PM2.5 的消除过程主要是通过扩散与沉降，两种条件均与风速显著相关。不同的风速条件对 PM2.5 的影响必然不同。在无风条件下 PM2.5 的扩散可按照菲克定律确定，扩散速率远小于风速。而在小风速的条件下，风速不仅对 PM2.5 的扩散有影响，同时小风速对空气有搅拌作用，从而有利于气溶胶中 PM2.5 的聚沉。而当风速更大的条件下空气将呈现湍流的特性，从而地面上的 PM2.5 将被吹起，并且风速较大的风也有一定的可能性会带来其他地区的沙土导致 PM2.5、PM10 的增高，而风向也会起到比较明显的作用，从空气较为清洁地区吹来的风可以对当地的雾霾天气起到净化作用。

5. 模型的重新设定

在模型构建部分最先使用的 OLS 是经典线性回归模型（CLR 模型）下被认为最理想的估计量。CLR 模型由五个关于数据生成方式的假设组成，一旦改变这些假设，可以得出不同的估计情况，其中很多情况下，OLS 估计量不再被认为是最佳估计量。这五个假定对 CLR 模型至关重要，如果某一假定发生改变，模型也要进行相应的调整来得到更加准确的结果。

环境数据由于其自身特点，对这五个假定中的部分假定存在违背现象，也因此导致了模型的不准确，部分重要变量发生符号的变化。主要违背的是 CLR 模型的第三个和第五个假设，这两个假设分别要求：所有的干扰项具有同方差且彼此不相关；观测结果的个数多于自变量的个数，而且在自变量之间不存在确切的线性关系，即不存在多重共线性问题。当违背这两个假设时，尽管 OLS 估计量仍然是无偏的，但是它的有效性会受到很大的影响，方差会变得非常大。我们首先来讨论多重共线性的问题，下面考察一下各主要变量之间的相关关

[22] 吴兑：《探秘 PM2.5》，气象出版社，2013 年 3 月，第 11 页

系，具体数值见表 9。PM2.5 与 PM10 的相关系数为 0.9044，与 CO 的相关系数为 0.8245，与 NO_2 的相关系数为 0.6784，与 SO_2 的相关系数为 0.5225，PM2.5 与 PM10 及 CO 的相关程度非常高。风速对于气体污染物普遍具有负向作用，除了臭氧之外，与其他几种气体之间的相关系数均为负，说明风对于污染物有明显的驱散作用，但是相关系数值并不算高。风速和湿度之间是负向关系，风速越大，湿度越低。除了臭氧之外，温度变量与其他解释变量相关性非常低。

表 9：主要变量相关系数表

	PM2.5	**PM10**	**CO**	**NO2**	**SO2**	**O3**	**风速**	**湿度**	**温度**
PM2.5	1.0000								
PM10	0.9044	1.0000							
CO	0.8245	0.7331	1.0000						
NO2	0.6784	0.6238	0.7097	1.0000					
SO2	0.5225	0.4702	0.5414	0.5402	1.0000				
O3	-0.1423	-0.2066	-0.3076	-0.5430	-0.1872	1.0000			
风速	-0.3542	-0.2801	-0.3804	-0.4907	-0.2227	0.2495	1.0000		
湿度	0.4032	0.3714	0.4148	0.3645	-0.0347	-0.3220	-0.5434	1.0000	
温度	-0.1144	-0.1221	-0.3047	-0.3449	-0.3562	0.6169	-0.0301	0.0419	1.0000

前文讨论过 CO 与 PM2.5 的关系，可以看出 CO 与 PM2.5 呈现高度的相关性，从上表的相关系数中也可以看出这一特点，但在实际中 CO 对 PM2.5 的形成并无直接影响，二者的高度相关性很有可能是由于排放源一致所导致的。CO 与其他几种气体污染物的相关程度较高，很容易造成多重共线性的问题，因此，出于简化模型的考虑，首先将 CO 从模型中手动删去，那么删掉 CO 变量之后，会对其他变量的选择造成怎样的影响呢？下面用 Lasso 再次尝试进行变量选择，结果如下：

表 10：Lasso 结果（2）

变量	系数	节点
NO2	1.3646	1
SO2	1.2393	2
humi	0.9816	3
O3	0.40001	4
N_NNE	-7.7112	5
heat	10.9863	6
windspeed	1.8286	7

可以看出手动去掉 CO 之后，Lasso 变量选择的结果也发生了变化，温度变量未被选入，取而代之的是供暖变量，这两个变量均是我们之前考虑置入模型的控制变量。下面根据以上各种方法进行变量选择的结果再次进行 OLS 估计，来观察相应系数发生的变化，具体结果见表 11：

表 11：回归结果 2

VARIABLES	(1) PM2_5	(2) PM2_5	(3) PM2_5
CO	37.25***		
	(0.592)		
NO2	0.789***	1.465***	1.158***
	(0.0226)	(0.0238)	(0.0230)
SO2	0.661***	1.322***	1.657***
	(0.0389)	(0.0452)	(0.0465)
O3	0.276***	0.432***	
	(0.0120)	(0.0134)	
windspeed	4.047***	4.717***	5.058***
	(0.295)	(0.291)	(0.307)
humi	56.66***	127.0***	114.7***
	(2.467)	(2.627)	(2.749)
temp	0.754***	0.930***	2.355***
	(0.0726)	(0.0872)	(0.0794)
heat	9.335***	34.56***	43.85***
	(1.655)	(1.948)	(2.038)
N_NNE	-11.13***	-10.63***	-15.55***
	(1.370)	(1.279)	(1.343)
其他风向省略			
Constant	-87.94***	-152.2***	-129.8***
	(2.879)	(3.227)	(3.334)
Observations	8,329	8,447	8,465
R-squared	0.768	0.653	0.611

第一栏为最初变量选择前的回归结果，包含一氧化碳和全部的风向，第二栏去掉了一氧化碳，并且只保留了北+北东北这一种风向，可以看出去掉一氧化碳后，二氧化氮、二氧化硫、湿度等变量的系数都有所增加。之后，我们再继续根据前面讨论的 PM2.5 形成机制来分析是否还能对模型进行简化和调整。O3 是一种与 PM2.5 并列的污染物，其形成同样受到光照和温度的影响，但是它本身并非 PM2.5 的组成部分或是前体物，因此在第三栏中去掉了 O3 变量，考察一下模型的变化，二氧化氮和湿度的系数值降低，二氧化硫、风速、温度、

供暖、风向这几个变量系数的绝对值有所增加，正如之前所说温度是影响 O3 形成的重要变量，因此 O3 一定程度上分散了温度对 PM2.5 的影响，因此当去掉 O3 之后，温度的系数值有所提高。

然而，我们最初的困惑仍为解决，最重要的风速变量符号却一直为正。通过对线性模型变量的增减调整模型，改善多重共线性，但问题仍未解决，因此接下来考虑改变线性模型的设定，加入风速变量的二次型，回归结果如下：

表 12：回归结果 3

VARIABLES	Coef.	Std.Err.
CO	38.06***	(0.575)
NO2	0.788***	(0.0225)
SO2	0.677***	(0.0385)
O3	0.280***	(0.0115)
wind2	-0.0643	(0.0594)
windspeed	4.036***	(0.647)
humi	52.14***	(2.290)
temp	0.459***	(0.0496)
N_NNE	-8.082***	(1.092)
Constant	-81.02***	(2.533)
Observations	8,329	-
R-squared	0.767	-

然而我们发现二次项的符号为负，一次项符号为正，这说明抛物线开口向下，与实际情况不符，改变线性模型的设定仍未能解决问题。那么又重新回到 PM2.5 的生成和扩散机制，模型中除风之外的其他变量，比如二氧化氮、二氧化硫、湿度、温度等都是 PM2.5 形成的前体物或催化条件。而风与这些变量不同，与其他变量相比，风速的影响存在一定的滞后性。

根据生活中的观察，静稳天气条件下往往需要若干小时才会形成空气中 PM2.5 浓度的显著升高。对空气污染监测数据和气象数据的关联分析也表明，PM2.5 浓度与风速存在若干小时的时间差，因此有必要在模型中加入风速的滞后项。

表 13：回归结果 4

	(1)	(2)
VARIABLES	PM2_5	PM2_5
NO2	1.158***	1.172***
	(0.0230)	(0.0235)
SO2	1.657***	1.664***
	(0.0465)	(0.0469)
windspeed	5.058***	3.638***
	(0.307)	(0.444)

L.windspeed		1.769***
		(0.536)
L2.windspeed		1.024*
		(0.536)
L3.windspeed		-0.932**
		(0.439)
humi	114.7***	116.2***
	(2.749)	(2.820)
temp	2.355***	2.377***
	(0.0794)	(0.0806)
heat	43.85***	43.88***
	(2.038)	(2.041)
N_NNE	-15.55***	-15.27***
	(1.343)	(1.345)
Constant	-129.8***	-133.2***
	(3.334)	(3.652)
Observations	8,465	8,462
R-squared	0.611	0.611

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

从上表结果中发现当加入滞后三期的风速变量时，风速符号为正的问题被解决了，而且除风速外，其他变量的系数均保持稳定，几乎没有发生变化。事实上，从起风到颗粒物浓度降低再到雾霾显著消散，往往需要一段时间，而 PM2.5 和瞬时风速呈现正向关系正是由于这种时滞，当洁净的空气从北方吹来时，首先会经过污染较重的北京市区，而这一部分空气将会首先到达观测点，所以会呈现出瞬时风速与瞬时 PM2.5 浓度正相关的关系，但是这一关系并不具有现实的解释意义。从最终模型看来，NO_2 每增加 1μg/m^3，PM2.5 的浓度上升了 1.20μg/m^3；SO2 每增加 1μg/m^3，PM2.5 的浓度上升了 1.80μg/m^3；而相对湿度、温度及风向的回归结果也都比较稳定。

本文在之前提到所使用的环境数据存在违背 CLR 模型五项基本假设的情况，前面的部分讨论了多重共线性，通过去掉无关解释变量来降低多重共线性的影响，除此之外，违反球形扰动项也是面临的另一项问题。我们所使用的数据为 2016 年农展馆监测点 PM2.5、气态污染物、气象变量的小时数据，是标准的时间序列数据，在时间序列中，由于观测变量具有一定的连续性，因此自相关现象在时间序列数据中比较常见。如果存在“自相关”，即存在 $i \neq j$，使得 $E(\varepsilon_i \varepsilon_j \mid X) \neq 0$，即扰动项的协方差阵 $Var(\varepsilon \mid X)$ 的非主对角线元素不全为 0

时，因此我们猜测可能存在自相关（或序列相关）的问题。存在自相关的情况时，虽然 OLS 估计量仍然无偏且一致，但是已不再是最有效的估计，因此下文主要讨论和解决自相关的问题。

首先构造模型如下：

$$\text{PM2.5}_t=\beta_0+\beta_1 SO_{2t}+\beta_2 NO_{2t}+\beta_3 windspeed_t+\beta_4 windspeed_{t-1}+\beta_5 windspeed_{t-2}+\beta_6 windspeed_{t-3}$$

$$+\beta_7 humi_t+\beta_8 \text{temp}+\beta_9 heat_t+\beta_{10} N\_NNE_t+u_t \qquad (2)$$

首先将残差与滞后残差制成散点图，来观察其是否存在明显的线性关系。下方的散点图中，纵坐标为残差 u_t ，横坐标为滞后一期的残差 u_{t-1} 。从图中可以看出二者存在高度的线性相关性，因此怀疑模型存在自相关问题。

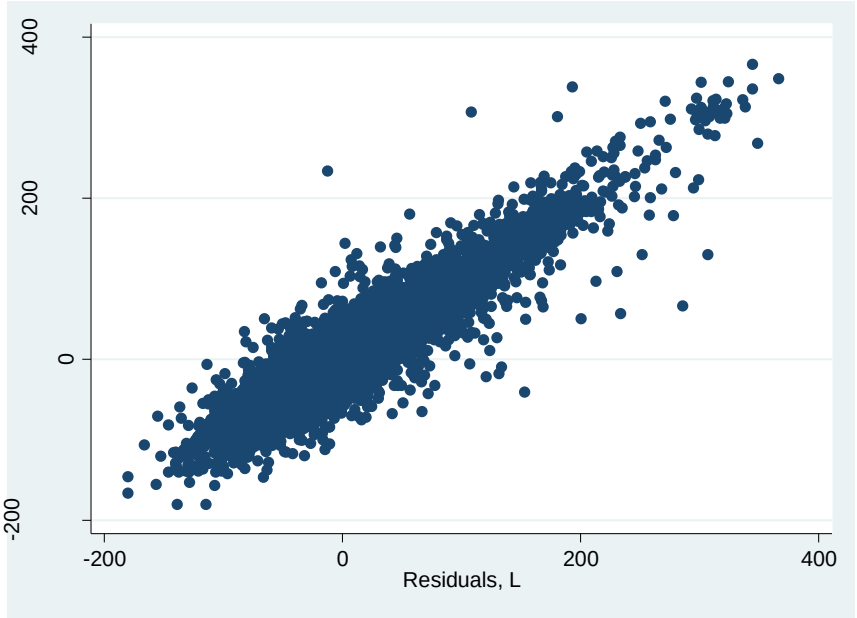


图 12：自相关检验

然后对模型进行 DW 检验，假定：

$$u_t=\phi u_{t-1}+a_t \qquad (3)$$

$H_0:\phi=0$ 不存在自相关关系

$H_1:\phi>0$ 存在正向的自相关关系

$$DW=d=\frac{\sum_{t=2}^{n}(u_t-u_{t-1})^2}{\sum_{t=1}^{n}u_t^{\,2}}=\frac{\sum_{t=2}^{n}u_t^{\,2}-2\sum_{t=2}^{n}u_t u_{t-1}+\sum_{t=2}^{n}u_{t-1}^{\,2}}{\sum_{t=1}^{n}u_t^{\,2}}\approx 2-2\frac{\sum_{t=2}^{n}u_t u_{t-1}}{\sum_{t=1}^{n}u_t^{\,2}}=2(1-\phi_1)$$

=0.150516

$d \approx 0, \phi_1 \approx 1$，我们拒绝原假设 $H_0 : \phi = 0$，存在一阶正自相关关系。

当存在自相关问题时，OLS 估计量仍然是无偏、连续、渐近正态的，但却不再有效，所以我们应该重新对模型进行有效估计。对于自相关问题，有四种处理方法，下文使用可行广义最小二乘法（FGLS）来处理。

通过变换，每个观测值的回归方程如下：

$$\sqrt{1-\phi^2}\, y_1 = \sqrt{1-\phi^2}\,\beta_0 + \sqrt{1-\phi^2}\,\beta_1 x_{1,1} + ... + \sqrt{1-\phi^2}\,\beta_{10} x_{1,10} + u_1$$

$$y_2 - \phi y_1 = (1-\phi)\beta_0 + \beta_1 (x_{2,1} - \phi x_{1,1}) + ... + \beta_{10} (x_{2,10} - \phi x_{1,10}) + u_2$$

.....　　（4）

$$y_n - \phi y_{n-1} = (1-\phi)\beta_0 + \beta_1 (x_{n,1} - \phi x_{n-1,1}) + ... + \beta_{10} (x_{n,10} - \phi x_{n-1,10}) + u_n$$

用 OLS 估计上式中变换后的模型，为 Prais-Winsten 估计法。实际操作中，常使用迭代法进行估计：首先用 OLS 估计原模型，然后通过辅助回归得到对 ϕ 的第一轮估计，再用其进行 FGLS 估计，然后使用新的残差进行第二轮估计，以此类推，直到最终收敛。迭代结果如下表所示：

表 14：PW 迭代法结果

	(1)	(2)
VARIABLES	PM2_5	PM2_5
NO2	1.172***	0.502***
	(0.0235)	(0.0165)
SO2	1.664***	1.486***
	(0.0469)	(0.0432)
windspeed	3.638***	0.199
	(0.444)	(0.141)
L.windspeed	1.769***	-0.101
	(0.536)	(0.142)
L2.windspeed	1.024*	-0.200
	(0.536)	(0.141)
L3.windspeed	-0.932**	-0.158
	(0.439)	(0.136)
humi	116.2***	35.46***
	(2.820)	(3.555)
temp	2.377***	1.259***
	(0.0806)	(0.163)

heat	43.88***	12.90**
	(2.041)	(6.528)
N_NNE	-15.27***	-0.525
	(1.345)	(0.413)
Constant	-133.2***	-9.762
	(3.652)	(6.746)
Observations	8,462	8,462
R-squared	0.611	0.278

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

使用 PW 迭代估计法使得 DW 统计量的值从 0.15 提高到 1.52，而且我们发现通过 FGLS 方法解决自相关问题之后，大部分变量系数的绝对值都减小了，也就是说在控制了自相关问题之后，各解释变量的解释力有所减弱，在各变量中，受到影响最小的是二氧化硫，系数并没有太明显的变化，与之相比，每增加 1μg/m^3 二氧化氮，对 PM2.5 浓度值的影响从 1.172μg/m^3 降到了 0.502μg/m^3。湿度、温度、供暖变量的影响程度都有所下降。风速的变化最为明显，瞬时风速的系数值大幅减小，说明风从监测点北面市区带来的污染空气对监测点处 PM2.5 浓度值的影响有所降低，而滞后一期至滞后三期的风速虽然并不显著，但符号均为负，说明滞后一期的风已开始逐渐对 PM2.5 起到驱散作用。

五. 总结

本文将计量经济方法应用于环境经济学数据当中，在自然科学领域相关研究和已有成果的基础上，通过变量选择和模型的调整来寻找最精简且最合适的模型，来验证自然科学领域中的理论关系。

在逐步优化模型的过程中，我们先从普通最小二乘法入手，将全部变量置于模型中，发现这种办法会造成混乱，一些无关变量会对参数估计造成干扰，比如风速符号为正，这也是我们逐步探讨优化模型的动因。因此，希望能通过合适的办法来帮助进行变量选择，受到 Efron 的启发，在变量选择的部分使用了逐步回归和 Lasso 来解决这个问题。两种方法均将变量缩减到了十个以内，主要是对风向变量的筛选，虽然对其他解释变量的筛选作用不明显，但是将风向变量从八种缩减到一种，也很大程度上帮助我们精简了模型。然而最初所面临的风速符号问题仍未解决，且其余解释变量也都包含在模型内。在此基础上，展开对 PM2.5 形成机制的分析，从自然科学的角度深入了解 PM2.5 形成及扩散的原因，从而对变量选择给予更多启发。最终将无关解释变量 CO 和 O_3 从模型中删去，解释变量系数有所变化，但

风速系数仍为正。此时开始考虑改变线性模型的设定，加入风速的二次项，但结果仍不理想，与现实情况不符。最终发现，当加入风速的滞后项之后，滞后三期的风速符号显著为负，说明风对于 PM2.5 的驱散需要一定时间，在模型中表现为三个小时的时滞，而瞬时风速符号为正主要是由于北风会先将观测点北侧市区的污染物带到观测点，造成其浓度的上升，随后才会带来洁净的空气从而起到净化作用。完成变量选择之后，对模型可能涉及的自相关问题作出了修正，发现风速影响的滞后期数减少，滞后一期的风速就能够造成 PM2.5 浓度值的下降，此外，其他解释变量系数的绝对值都有所减小，二氧化氮的变化较二氧化硫的变化更为明显。但即便是修正了自相关问题后，风速滞后影响的情况仍未发生改变。根据自然科学的实验结果，风速是影响 PM2.5 积聚和扩散过程最重要的因素，然而通过数据挖掘方法由机器来选择的结果却并未全部得到理论的支持。

文章内很重要的一部分在讨论变量选择的问题，也使用了逐步回归这种颇有历史及 Lasso 这样相对新颖的方法进行尝试，它们在精简模型方面发挥了一定的作用，但是结果也并不尽如人意。在网络时代，我们常常会面临庞大数目的变量，如何在众多变量中发掘出最重要的因素，构建最合适的模型？已经有前人给出了一些准则，“在没有理解现实生活系统的非统计学性质时，不要尝试建立模型并对其进行统计学分析。在忽略了主体事物的情况下做出的统计学分析只能是无知的统计分析。”（Belsley 和 Welch，1988，p.447）Burderkin 和 Burkett（1998），Breuer 和 Wohar（1996）等都说明了了解问题的背景或者数据的生成机制会对计量经济学分析起到良好的辅助作用，提高实证工作的质量。另外，使用最合适的简单模型，许多复杂的和最新的技术有时并不是最合适的，Wilkinson（1999，p.598）强调：“不要选择解析方法来打动你的读者或者歪曲批评。如果一个更简单的模型的假设和解释力对于你的数据和研究对象是合理的，那么就可以使用它。”

每种依据数据来建立模型的方法都难免有一些不可避免的局限性，专业领域的知识则是强大的变量选择工具。我们应当基于对事物基本原理的把握，了解每种方法的适用条件，使用这些工具，与专业领域的知识相结合，更准确地把握现实世界的规律。

六. 参考文献

1. 吴兑：《探秘 PM2.5》，气象出版社，2013 年
2. 胡大源：《大气污染治理研究课题报告》
3. 贺克斌等：《大气颗粒物与区域复合污染》，科学出版社，2014 年
4. 曹军骥等：《PM2.5 与环境》，科学出版社，2014 年
5. 蒋魏楣，孙鉴泞，曹文俊，蒋瑞宾：《空气污染气象学教程》，气象出版社，2004

6. 黄丽坤，王广智：《城市大气颗粒物组分及污染》，化学工业出版社，2015 年
7. 白志鹏，王宝庆，王秀艳，姬亚芹等：《空气颗粒物污染与防治》，化学工业出版社，2011 年
8. 吴昱，《大数据精准挖掘》，化学工业出版社，2015 年
9. 朱光磊，张远航，曾立民等：《北京市大气细颗粒物 PM2.5 的来源研究》，《环境科学研究》，2005 年
10. 陈添，华蕾，金蕾等：《北京市大气 PM10 源解析研究》，《中国环境监测》，2006 年
11. 刘保献，杨懂艳，张大伟等：《北京城区大气 PM2.5 主要化学组分构成研究》，《环境科学》，2015 年
12. 朱珠等：《PM2.5/PM10 浓度变化规律及其气象条件分析——以深圳市龙岗区为例》，2014 中国环境科学学会学术年会（第六章），2014 年 8 月
13. Linoff, Gordon S., and Berry, Michael J.A.,《数据挖掘技术》（第 3 版），清华大学出版社，2013 年
14. Belsley, D. *A. Model Selection in Regression Analysis, Regression Diagnostics and Prior Knowledge*. International Journal of Forecasting, 1986
15. Bibby, J. and H. Toutenburg *Prediction and Improved Estimation in Linear Models,* 1977
16. Binkley, J. K. and P. C. Abbott *The Fixed X Assumption in Econometrics: Can the Textbooks be Trusted?* American Statistician, 1987
17. Conlisk, J. *When Collinearity Is Desirable.* Western Economic Journal, 1971
18. Dreze, J. *Nonspecialist Teaching of Econometrics: A Personal Comment and Personalistic Lament.* Econometric Reviews, 1983
19. Efron, Trevor Hastie. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science,* 2016
20. Feldstein, M. *Multicollinearity and the Mean Square Error of Alternative Estimators.* Econometrica,1973
21. Feldstein, M. S. *Inflation, Tax Rules and Investment: Some Econometric Evidence*, 1982
22. Gilbert, C. L. *Professor Hendry's Econometric Methodology.* Oxford Bulletin of Economics and Statistics, 1986
23. Greenberg, E. and R. P. Parks *A Predictive Approach to Model Selection and Multicollinearity.* Journal of Applied Econometrics, 1997

24. He K, Yang F, Ma Y, Zhang Q, Yao X H, Chan C K, Cadle S H, Chan T, Mulawa P A. *The characteristics of PM2.5 in Beijing, China*. Atmospheric Environment 2001
25. Hendry, D. F. *Econometrics - Alchemy or Science*, 1980
26. Hendry, D. F. and J. F. Richard. *The Econometric Analysis of Economic Time Series. International Statistical Review,* 1983
27. Johnston, J. *Econometric Methods*, 1984
28. Leamer, E. E. *Are the Roads Red? Comments on "Size Matters."* Journal of Social Economics, 2004
29. Maddala, G. S. and 1-M. Kim *Unit Roots, Cointegration, and Structural Change*, 1998
30. Magnus, J. R. *The Missing Tablet: Comment on Peter Kennedy's Ten Commandments*. Journal of Economic Surveys, 2002
31. Pagan, A. R. *Three Econometric Methodologies: A Critical Appraisal*. Journal of Economic Surveys, 1987
32. Peter Kennedy. *A Guide to Econometrics*,2008
33. Quah, D. *Business Cycle Empirics: Calibration and Estimation*, 1995
34. Robarge W P, lker J T, McCulloch R B. *Atmospheric concentrations of ammonia and ammonium at an agricultural site in the southeast United States.* Atmospheric Environment, 2002
35. Sailesh N. Behera, et al. Insights into Chemical Coupling among Acidic Gases, Ammonia and Secondary Inorganic Aerosols. Aerosol and Air Quality Research, 2013
36. Theil, H. *Principles of Econometrics*, 1971
37. Worswick, G. D. N. *Is Progress in Science Possible*, 1972

致谢

完成本文的写作，首先最要感谢的是我的导师胡大源老师。从论文的选题到写作思路，再到每一步的修改，胡老师都给予了耐心而充分的指导。每当遇到瓶颈时，胡老师都会给予我关于接下来研究方向的启发，在写作的过程中，对于从未接触过的新领域和新方法都做了充分的尝试。除了论文写作，胡老师在学习、未来的方向、为人处世的道理等方方面面给予的指导都让我们受益匪浅。很感谢胡老师不辞辛苦地利用周末的休息时间来与我们分析和讨论研究的成果，每当我们有任何疑问时，都会第一时间给予回应，并尽力帮助我们解决问题。胡老师不断钻研的治学精神和严谨的学术态度也一直是我们学习的标准。很幸运能成为胡老师的学生，今后的学习和工作中，会一直谨记胡老师的谆谆教诲，诚恳做人，认真做事。

还要感谢同组的同学们，他们的研究为我后续的深入讨论和写作奠定了非常扎实的基础，其中很多内容也给我带来了灵感和启发。此外，非常感谢组内的师弟师妹在新方法尝试及编程应用方面给予的帮助，他们牺牲了自己的闲暇时间来帮助我运行程序，使得论文写作能够非常顺利地进行。

最后要感谢院里的培养和各位老师的教导，大师身旁宜聆教，短暂几年的研究生时光，很荣幸能在朗润园度过，上过的每一次课和听过的每一次讲座都令人难忘。各位老师的智慧让人钦佩，也让我们时刻鼓舞自己，以他们为榜样，成为善于思考的智者。

再次感谢大家！

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明： 所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名： 日期： 年 月 日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名： 导师签名：

日期： 年 月 日