全流程统计分析虚拟仿真实验教学项目

机动车车牌号优选策略与限行政策分析

实 验 指 导 书

目　录

实验概述

统计学是实践性很强的数据科学，从实际问题出发，以解决实际问题为归宿。随着传感等技术的兴起，社交网络等媒体的涌现，数据正以前所未有的形式、速度、广度不断累积和增长。在大数据时代，问题的分析规模越来越庞大。面向实际开展全流程统计建模分析，涉及海量数据采集、处理、存储和分析，成本很高。进一步针对社会经济问题研究来讲，特别是政策效果评估和机制设计，限于伦理或法律的约束，难以通过自然实验进行比较研究。为此，在教育教学过程中，学校很难提供一个面向现实的全流程统计分析环境，导致学生所学知识的碎片化，动手能力不强，难以用于现代统计实践。建立一套基于虚拟仿真的，易于实施的，从数据到结论的统计学实验教学方法体系势在必行。

该实验平台设计以“一切皆可量化”、“废旧数据重用”的大数据思维为基础，构建从数据到结论的统计模拟及建模全流程仿真实验项目，应用于统计学、经济学类多门课程，试图实现数据分析类教学流程再造。具体实验以兰州市城市问题为例，以车牌号优选策略为切入点，以机动车限行政策为实验对象，开发低成本、高时效的全流程仿真统计建模实验项目，用以解决政策效果评估和机制设计问题。

限行政策在交通上具有“治堵”效果，在国内外许多城市实施。对于“智慧”程度相对不高的城市，由于监控与传感设备的覆盖不够广，或图像数据处理成本过高，直接通过交通图像研究限行政策具有一定的局限性。鉴于限行政策在生态环境上具有“防污”效果，据此，可以利用互联网上发布的微观空气质量数据，间接测算机动车限行政策效应，开展政策模拟，进行机制设计。

进一步来讲，由于政策效应取决于政策本身以及随着时间推移人们对政策的适应。为了使得实验更加贴近日常生活，就地取材，更好地说明人的行为对政策的适应机制，该实验从两个视角出发开展设计：第一，通过交互操作，模拟实现多出行、少受限的个人购车车牌策略选择，进而从个体视角讨论政策适应，并过渡到公共政策效应讨论。第二，对限行的政策效应进行测算的虚拟仿真实验，并结合车牌选号策略，从公共政策视角开展政策优化设计实验。

实验目标

实验后应该达到的知识水平:

1.掌握频数统计方法的编程实现;

2.掌握用绘图工具及编程;

3.掌握简单的因果推断方法及政策评估手段方法实现。

4.掌握数据处理技巧和 Python(或 R)代码编写;

实验后应该达到的能力水平:

1.了解从实际生活中发现并提出统计研究问题的方式;

2.掌握从问题到结论的统计设计思维模式;

3.具备从数据到结论的统计分析全流程的设计能力;

4.初步掌握通过统计分析进行政策模拟及机制设计能力。

实验步骤

（1）学生交互性操作步骤，共 11 步

步骤序号	步骤目标要求	步骤合理用时	目标达成度赋分模型	步骤满分
1	车牌选号模拟	5（min）	**步骤目标:** 观察并分析总结两种限行政策下不同尾号车牌限行天数及规律。 **判定条件:** (1)操作用时 > 1 分钟; (2)表单改变 == TRUE。 **赋分模型:** 条件(1)、(2)同时为真，得 5 分。	5
2	车牌选号效益分析	15（min）	**步骤目标:** 分别对两种规则下，每个尾号组合限行天数进行排序。 **判定条件:** (1)操作用时 > 1 分钟; (2)操作次数 == 1 次; (3)操作次数 == 2 次; (4)规则 1 排序正确; (5)规则 2 排序正确。 **赋分模型:** 1.条件(1)、(2)、(4)同时为真，得 5 分; 2.条件(1)、(2)、(5)同时为真，得 5 分; 3.条件(1)、(3)、(4)同时为真，得 3 分; 4.条件(1)、(3)、(5)同时为真，得 3 分。	10
3	限行政策尾数组合建议	15（min）	**步骤目标:** 在星期限号规则下，针对每对尾号组合限行天数不均匀问题，启发学生分析，并重	10

			新形成组合对。 **判定条件：** (1)操作用时 ＞1 分钟； (2)操作次数 == 1 次； (3)操作次数 == 2 次； (4)组合结果正确。 **赋分模型：** 1.条件(1)、(2)、(4)同时为真，得 10 分； 2.条件(1)、(3)、(4)同时为真，得 5 分。	
4	空气质量模型观察	5 （min）	**步骤目标：** 观察并分析限行政策等因素对空气质量的影响。 **判定条件：** (1)操作用时 ＞1 分钟； (2)表单改变 == TRUE。 **赋分模型：** 条件(1)、(2)同时为真，得 5 分。	5
5	空气质量建模模型选择	15 （min）	**步骤目标：** 学生通过分析实际问题、了解模型原理，选择建模模型。 **判定条件：** (1)操作用时 ＞1 分钟； (2)操作次数 == 1 次； (3)操作次数 == 2 次； (4)建模模型选择正确。 **赋分模型：** 1.条件(1)、(2)、(4)同时为真，得 10 分； 2.条件(1)、(3)、(4)同时为真，得 5 分。	10
6	模型优化分析-稳健估计	15 （min）	**步骤目标：** 学生通过观察模型初步结果，通过表单输入形式选择对模型进行稳健性估计。 **判定条件：** (1)操作用时 ＞1 分钟； (2)分析题目选择正确。 **赋分模型：** 条件(1)、(2)同时为真，得 10 分。	10
7	模型优化分析-控制变量	15 （min）	**步骤目标：** 学生通过观察稳健估计结果，通过表单输入形式对模型进行控制变量选择。 **判定条件：** (1)操作用时 ＞1 分钟； (2)分析题目选择正确。 **赋分模型：** 条件(1)、(2)同时为真，得 10 分。	10
8	模型优化分析-交互效应	15 （min）	**步骤目标：** 学生通过控制变量分析结果，通过表单输入形式对模型进行自变量交互效应分析。 **判定条件：** (1)操作用时 ＞1 分钟； (2)分析题目选择正确。 **赋分模型：** 条件(1)、(2)同时为真，得 10 分。	10
9	模型优化分析-安慰剂检验	15 （min）	**步骤目标：** 学生在上一步交互效应分析的基础上，进行安慰剂检验。 **判定条件：** (1)操作用时 ＞1 分钟； (2)分析题目选择正确。 **赋分模型：** 条件(1)、(2)同时为真，得 10 分。	10
10	空气质量建	120 （min）	**步骤目标：** 依据建模结果，学生通过自己编写 Python 或 R 程序，对模型进行优化完善。	20

	模进阶分析		**判定条件：** (1)操作用时 ＞1 分钟； (2)程序结果文件成功生成； (3)程序结果文件内容正确。 **赋分模型：** 1.条件(1)、(3)同时为真，得 10 分； 2.条件 (1)、(2)同时为真，且条件 (3)为假，得 5 分。	
11	政策建议报告生成及实验结果查阅	5 （min）	系统根据学生前期所有实验步骤完成情况，生成该生政策建议报告，并展示学生实验结果得分。	0

（2）交互性步骤详细说明

该实验源于社会经济现实问题，由课程团队的科研成果融合、串联、简化设计形成。

图 1：实验背景介绍

图 2：实验原理介绍

第一步：车牌选号模拟

操作：以兰州为例（甘 A），系统提供车牌号码、用车计划时间、限行规则等网页表单供学生输入，仿真形成学生购车时自己心仪的车牌号码和用车计划。系统根据用户输入内容，动态生成限行频数图和拥堵情况散点图，供学生观察思

考。系统页面交互图如下:

图 3：第 1 步——机动车模拟选号

说明 1: 直接讨论政策效果、进行统计建模，不易引起学生们的兴趣。我们从日常生活中的事情导入。学生们大学毕业后，可能在较短的时间内买车，对购车选号是有潜在需求的。我们从购买机动车选车牌号码入手，让学生了解，选号除了图喜欢、图吉利，还能图“实惠”；也即通过车牌号的优化选择，可以让自己的机动车在既定的限行政策下，多出行、少堵车，以此引起学生的思考和好奇。

说明 2: 学生在输入车牌号码、用车计划等表单时，屏幕上提示兰州市的机动车限行规则，目的在于暗示学生要把限行规则“吃”透。

屏幕提示:

表 1　兰州市机动车限行规则

限行类型	尾号日期限行	尾号星期限行
描述	自 2010 年 9 月开始，常年实行每日两个尾号的限行措施(不含周末和法定节假日)，机动车按照车牌尾数对应的日期依次限行。限行组合分别为：1 和 6、2 和 7、3 和 8、4 和 9、5 和 0。	自 2020 年 4 月 13 日起，实行尾号星期限行措施(不含周末和法定节假日)，每天限两个号，以周为循环，星期数对应车牌号最后一位数字，星期一至星期五限行车牌尾号分别为：1 和 6，2 和 7，3 和 8，4 和 9，5 和 0。

对表 1 的限行规则解释如下：对于尾号日期限行而言，倘若您的车牌尾号为 1 或 6，则每月 1 日、6 日、11 日、16 日、21 日、26 日、31 日您不能驾车出行，其他尾号以此类推；对于尾号星期限行而言，倘若您的车牌尾号为 2 或 7，则每周星期二您不能驾车出行，其他尾号以此类推。

说明 3: 汽车拥堵模拟采用 Nagel-Schreckenberg 交通流模型，由服务端按客户端请求的参数模拟数据提供 API 服务，前端进行图像渲染

说明 4: 系统根据用户操作用时长、是否有输入等条件，自动调用赋分模型，产生本步骤学生得分并记录。

第二步：车牌选号效益分析

操作：页面针对两种限号规则，分别提供一组尾号限行天数排序表单。启发学生总结尾号限行规律，对各尾号组合的限行天数进行由高到低排序。系统页面交互图如下：

图 4：第 2 步——车牌选号效益分析

说明 1：此步骤通过让学生对尾号组合限行天数进行排序的方式，促使学生认识到，即使在同一限号规则下，不同尾号的车牌限行天数亦有不同。从而引导学生分析限号规律，总结如何“吃透”规则，选择一个可以“多出行、少限行”的车牌号，使得车牌选号效益最大。

说明 2：系统根据限号规则自动计算尾号限行天数排序结果，与学生作答内容进行对比，自动调用赋分模型产生本步骤学生得分并记录。

第三步：限行政策尾数组合建议

操作：页面针对两种限号规则，分别提供一组尾号重新组合表单，由学生根据屏幕提示的“兰州市车辆尾号抽样调查分布图表”，自行总结尾号分布规律，进行限行“尾号对”重新组合。系统页面交互图如下：

图 5：第 3 步——限行政策尾数组合建议

说明 1：启发学生从政策制定角度思考，充分认识人们存在的车辆选号偏好，以每日限号车辆数分布更加均匀、减少拥堵为出发点，分析总结车辆尾号数据分布规律，进而重新组合限行“尾号对”，形成政策建议。

屏幕提示：

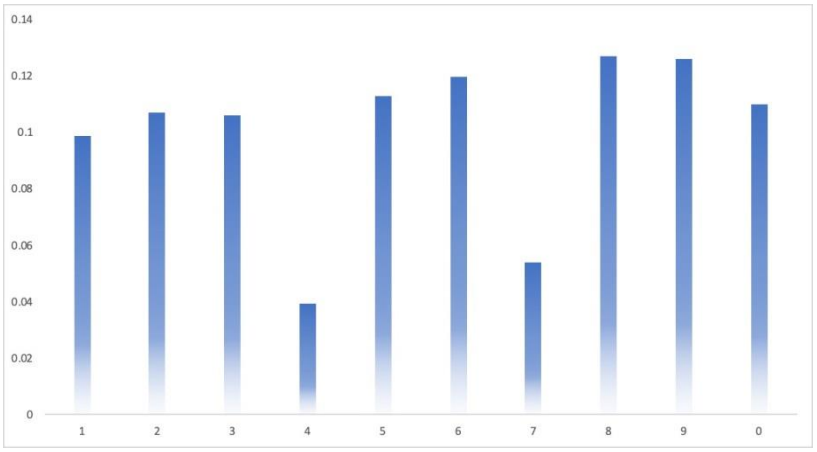


图 6：兰州市车辆尾号抽样调查分布图表

说明 2：系统根据限号规则自动计算尾号限行天数排序结果，依据首尾组合原则形成正确“尾号对”，并与学生作答内容进行对比，自动调用赋分模型，产生本步骤学生得分并记录。

第四步：空气质量模型观察

操作：页面提供可能影响空气质量（AQI）的各因素变量表单，学生通过不断调整各因素变量取值，观察系统动态生成的空气质量指标图表变化情况，分析总结空气质量影响变量及权重，做好下阶段空气质量建模准备。其中，可调节因素变量包括是否限行、生产型企业数量、小煤炉使用数量、有氟冰箱使用数量、温度和风力等 6 项内容，空气质量指标包括 AQI、PM2.5、PM10、SO2、NO2、CO、O3、Tmpt、Wndfrc 等 9 项内容。系统页面交互图如下：

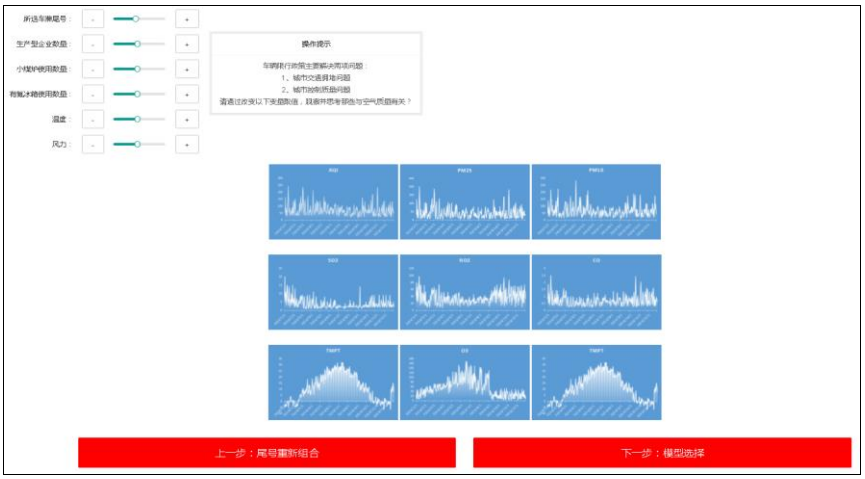


图 7：第 4 步——空气质量模型观察

说明 1：团队成员以兰州市近五年（2015 年-2019 年）气象数据为模型数据基础，建立断点回归模型。大气污染的形成是一个包含人类活动和自然力作用的复杂现象，我们构建机动车限行政策的空气质量效应模型如下：

$$\ln Y = \gamma_0 + \gamma_1 Res + \boldsymbol{\alpha}'\mathbf{H} + \boldsymbol{\beta}'\mathbf{W} + \varepsilon \quad (1)$$

其中，因变量Y代表空气质量指标，本实验为 NO_2；ε为随机扰动，表示未

能观测因素的综合影响；研究变量为Res，其取值为

$$Res = \begin{cases} 1, & \text{单双限行} \\ 0, & \text{尾号限行} \end{cases}$$

$\boldsymbol{W}$是自然因素对空气质量影响变量的可能集合：$\boldsymbol{W}$ = (HUMI, TEMP, WIND, SEA_EFT)。其中，HUMI 为湿度、TEMP 为气温、WIND 为风力、SEA_EFT 为季节效应虚拟变量。$\boldsymbol{H}$是人类活动对空气质量影响变量的可能集合：$\boldsymbol{H}$ = (HOUR_EFT, DAY_EFT, WEEK_EFT)。其中 HOUR_EFT 为限行时段效应虚拟变量，DAY_EFT 是日期尾数虚拟变量，WEEK_EFT 为工作日效应虚拟变量。

说明 2: 系统根据用户操作用时长、是否有输入等条件，自动调用赋分模型，产生本步骤学生得分并记录。

说明 3: 数据模拟采用深度学习模型在服务器端生成，并提供 API 接口。

第五步：空气质量建模模型选择

操作: 页面提供四个模型供学生进行建模模型选择，包括人工神经网络模型、支持向量机模型、断点回归模型和决策树模型，并且展示每个模型的名称及介绍内容。当学生作出正确选择时，系统进入后续步骤；当选择错误并点击“下一步”按钮时，提示“模型选择错误！原因：所选模型不适用于该问题。”，学生可进行重新选择，直到模型选择正确，系统允许继续后续步骤。

图 8：第 5 步——空气质量建模模型选择

说明 1: 学生可根据对问题域的了解和对每种不同统计模型的已有知识经验，分析不同模型的适用场景，从而作出正确判断。

说明 2: 系统会根据“正确答案”，对学生选择结果进行判断，自动调用赋分模型，产生并记录学生该步骤得分。其中，学生在作出正确答案前错误尝试的次数，在赋分模型中进行体现，以防止通过随意尝试排除错误选项的作弊行为。

第六步到第九步总体说明

第六步到第九步为以建模初步结果为基础进行的模型持续优化步骤。系统展示模型建模初步结果，以表单形式逐步提供模型结果优化方法，供学生分析并选择。模型结果分析内容包括：模型优化分析-稳健估计、模型优化分析-控制变量、模型优化分析-交互效应、模型优化分析-安慰剂检验等。系统根据用户输入内容，

通过调用后台接口方式进行逐步系统优化，并以交互性方式展现优化结果，且最终优化步骤结果将在该学生政策建议方案中体现。系统会根据“正确答案”，对学生选择选择的优化选项进行判断，自动调用赋分模型，产生并记录学生该步骤得分。第六步到第九步系统页面交互图如下:

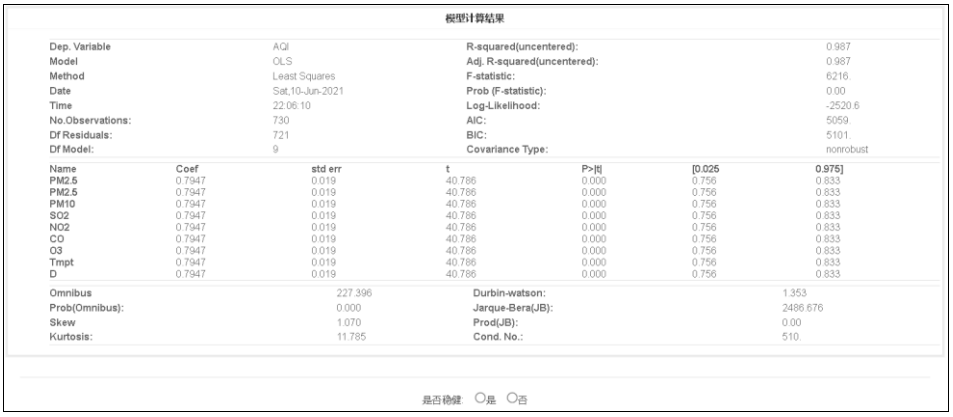


图 9：第 6-9 步——空气质量建模模型优化

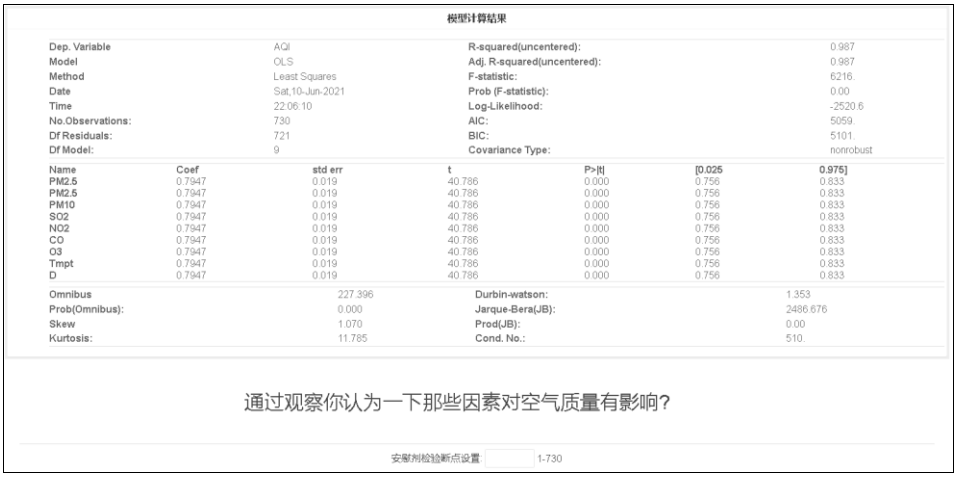


图 10：第 6-9 步——空气质量建模模型优化

第十步：空气质量建模进阶分析

操作：系统在页面提供在线编程环境，学生根据建模结果及模型存在问题，对模型进行编程优化。编程环境同时支持数据科学最主要的两大编程语言：Python 语言和 R 语言。学生在编写代码完毕并运行程序后，系统后台自动检查学生程序完成度和完成质量，并提示学生进行后续实验步骤。系统页面交互图如下:

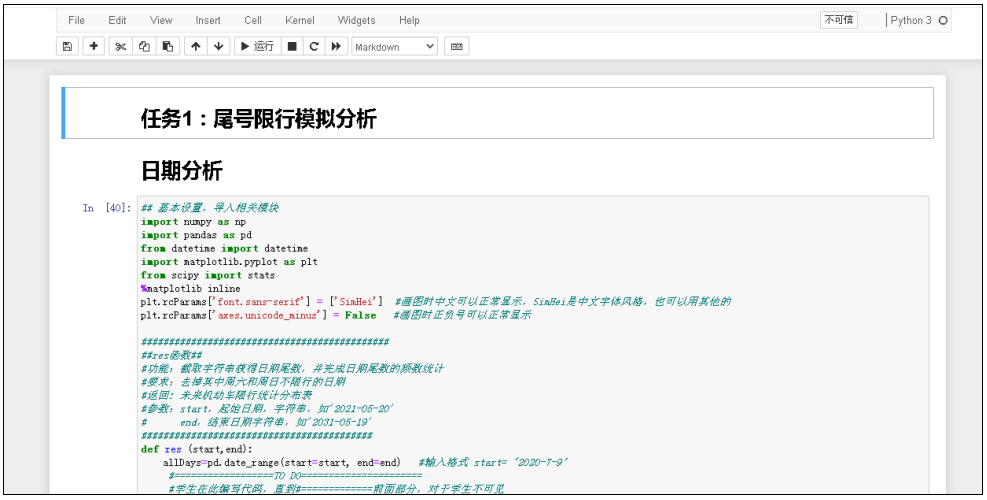


图 11：第 10 步——空气质量建模进阶分析

说明 1： 本步骤以激发学生使用程序语言解决具体问题兴趣，提升学生模型优化、程序编写、数据处理等统计学科基础实践能力为主要目标，着力培养学生的动手实践意识和模型应用能力。

说明 2： 为了使学生将主要时间精力运用在统计建模核心问题上，避免学生因录入过多与模型问题处理相关度不高的基础性程序语言代码而降低实验效率，系统会根据模型具体优化需求，自动创建基础代码文件，学生须根据代码上下文，仅在基础代码文件中指定空缺位置进行填充式代码录入。在学生完成代码填充并运行后，会出现以下三种可能情形：（一）程序存在语法错误，无法成功执行；（二）程序无语法错误，可成功执行，但模型优化结果不正确；（三）程序无语法错误，可成功执行，并且模型优化结果正确。系统分别针对以上三种情形的处理逻辑如下：（一）无结果性文件产生；（二）成功生成结果性文件，但文件内容有误；（三）结果性文件内容正确。系统会自动调用赋分模型，判别三种不同结果，并产生学生该步骤得分。

第十一步：政策建议报告生成及实验结果查阅

操作： 系统根据学生前期所有实验步骤完成情况，自动累加计算学生各步骤得分，以列表形式展示学生实验最终结果得分，并生成政策建议报告，且提供下载链接，供学生下载查看。

（3）进阶建模分析代码说明

(以 R 为例进行说明，详见 http://202.201.80.252:8888/notebooks/R_Code.ipynb

Python 版有类似的情况，详见 http://202.201.80.252:8888/notebooks/Exercise.ipynb)

日期尾号数据生成

操作: 通过 R 语言的字符串截取功能，提取日期尾数，要求形成可执行代码。学生可见提示及基础代码如下:

```
##########################################
##res函数##
#功能：截取字符串获得日期尾数，并完成日期尾数的频数统计
#要求：去掉其中周六和周日不限行的日期
#返回：未来机动车限行统计分布表
#参数：start，起始日期，字符串，如'2021-05-20'
#      end，结束日期字符串，如'2031-05-19'
##########################################
    res <- function(start, end)
    {
      allDays <- seq(start, end, by = 'day')
     #==================TO DO=====================
     #学生在此编写代码#
     #学生在此编写代码#
     #============================================
      df <- data.frame(dayend=lastNum, week=week)
      df <- df[week!='星期六' , ]
      df <- df[week!='星期日' , ]
      return(table(df$dayend))
    }
    ##============================================
```

说明 1: 以上实验材料中提供基础代码，减少学生不必要的输入，学生根据提示，完成“TO DO”部分的核心代码。参考答案代码如下:

```
    ##==================参考代码=================
    res <- function(start, end)
    {
      allDays <- seq(start, end, by = 'day')
      week <- weekdays(allDays)
      allDays <- as.character(allDays)
      n <- nchar(allDays)
      lastNum <- substr(allDays, n, n)
```

```
  df <- data.frame(dayend=lastNum, week=week)
  df <- df[week!='星期六' , ]
  df <- df[week!='星期日' , ]
  return(table(df$dayend))
}
##=========================================
```

说明 2: 学生通过编写代码，若能够正确运行，则可以在既定限行政策下，获得依据日期进行车牌号码优选的数据。

数据分析 1：日期尾号分析与车牌优选策略

操作: 通过 R 语言编程，调用第步中的 res 函数，形成用车计划期间各个日期被限行的天数统计结果。要求形成可执行代码。学生可见提示及基础代码如下:

```
##########################################
##模拟形成用车计划时间序列
##调用res函数
##形成被限行的天数统计结果
##########################################
start <- #学生在此编写代码#
end <- #学生在此编写代码#
#==================TO DO=====================
#学生在此编写代码#
#学生在此编写代码#
#=========================================
```

说明 1: 以上实验材料中提供基础代码，减少学生不必要的输入，学生根据提示，完成“TO DO”部分的核心代码。

参考答案代码如下:

```
##==================参考代码=================
start <- as.Date('2021-05-20')
end <- as.Date('2031-05-19')
result <- res(start, end)
result
barplot(result, col = c(4,2,rep(4,7)))
##=========================================
```

说明 2: 学生通过编写代码，若能够正确运行，则可以针对既定限行政策以及既定的用车计划时期，形成各个日期被限行的日期统计。学生可选的参数设置包括：通过不同的用车计划时期设置，得到不同限行结果；采用不同的绘图方式展现统计结果。

该步骤为了让学生了解，往往最简单的频数统计方法就足以实现分析，关键

是问题提出的能力。

数据分析 2：车牌尾号分析与车牌优选策略

操作：通过 R 语言编程，学生根据自己的感觉，按照既定分布生成车牌尾号分布数据，（模拟）形成用车计划期间不同车牌尾号被限行的天数统计结果。进而加载兰州机动车车牌号实际分布数据，重复以上分析。最后，通过卡方检验，判断车牌选号的数字偏好。要求形成可执行代码。学生可见提示及基础代码如下：

```
##########################################
##按照限行规则，统计兰州市机动车尾号比例，并检验数字偏好
##按照限行规则合并'0-5','1-6', '2-7','3-8'以及'4-9'
##统计机动车尾行比例
##绘制图形
##检验数字偏好
##########################################
    ##数字偏好假设检验，按照指定分布生成车牌号比例
    #==================TO DO====================
    #学生在此编写代码#
    #学生在此编写代码#
    pCar <- car/sum(car)
    barplot(pCar, col = 4)
    chisq.test(car)
    #===========================================

    ##数字偏好假设检验，按照兰州车牌实际情况
    carNum <- read.csv('./car_num.csv')#按照兰州市车牌分布实际加载车牌号比例
    #==================TO DO====================
    #学生在此编写代码#
    #学生在此编写代码#
    #===========================================

    ##单双号分析
    #==================TO DO====================
    s1 <- sum(carNum$car[carNum$num %% 2 == 0])
    s2 <- sum(carNum$car[carNum$num %% 2 == 1])
    #学生在此编写代码#
    #学生在此编写代码#
    #===========================================
```

说明 1：以上实验材料中提供基础代码，减少学生不必要的输入，学生根据提示，完成“TO DO”部分的核心代码。参考答案代码如下：

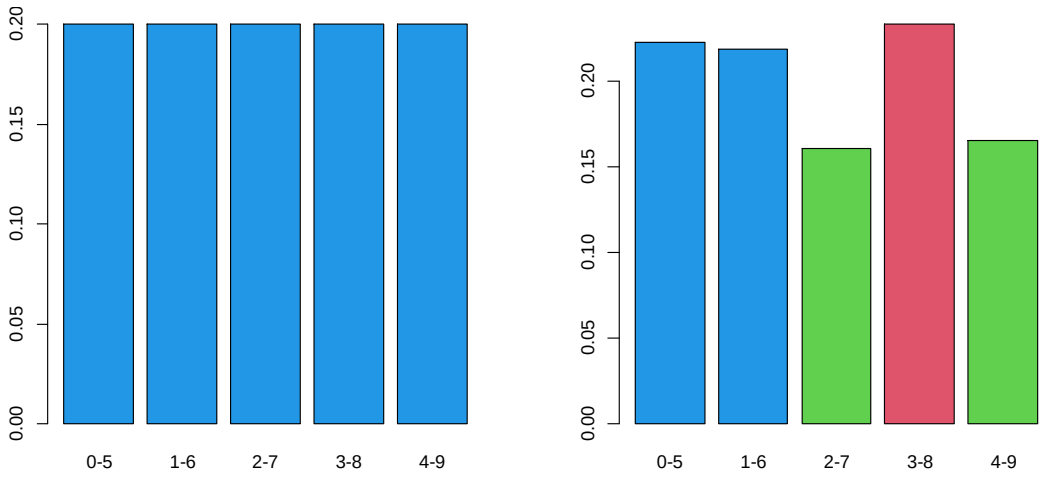
```
    ##数字偏好假设检验，按照指定分布生成车牌号比例
    carNum <- rep(0.1, 10)                    #按照均匀分布生成车牌号比例
```

```
carNum <- data.frame(car = carNum)
car <- carNum$car
car <- car[1:5] + car[6:10]
names(car) <- c('0-5','1-6', '2-7','3-8','4-9')
pCar <- car/sum(car)
barplot(pCar, col = 4)
chisq.test(car)

##数字偏好假设检验，按照兰州车牌实际情况
carNum <- read.csv('./car_num.csv')#按照兰州市车牌分布实际加载车牌号比例
car <- carNum$car
car <- car[1:5] + car[6:10]
names(car) <- c('0-5','1-6', '2-7','3-8','4-9')
pCar <- car/sum(car)
barplot(pCar, col = 4)
chisq.test(car)

##单双号限行分析
s1 <- sum(carNum$car[carNum$num %% 2 == 0])
s2 <- sum(carNum$car[carNum$num %% 2 == 1])
sc <- c(s1, s2)
sc <- as.table(sc)
barplot(sc)
```

结论：根据统计输出结果，如下图所示。



车牌尾号分布：无数字偏好模拟结果(左)与兰州市实际车牌结果 (右)

理论上来看，若不存在数字偏好，左图中尾号比例均应接近 20%，车牌尾号均匀分配。而事实上，右图的结果显示出了明显的数字偏好特征。按照“反其道而行”的选号原则，符合“实惠”标准的可选尾号是右图中显示为绿色的部分。

若按此选择尾号，在你的车限行时，道路上的机动车相对多，而在你的车不限行时，道路上的机动车相对少，何乐而不为。通过对日期尾数和车牌尾号的简单频数统计分析可以看出，从多出行、少堵车的“实惠”角度来看，机动车尾号为 2、7、4 或 9，在兰州市是比较不错的选择。

网络数据采集

操作：通过 R 语言编程，利用网络爬虫，采集、解析指定网页中的大气污染物小时浓度数据。需要说明的是，这里采集到数据是单一时间点的，本实验中实际应用的数据为课程团队自动化不间断实时采集形成的。学生可见提示及基础代码如下：

```
##########################################
##针对空气质量网页，通过网络爬虫，定向采集并解析得到空气质量数据
##过程包括：
##1．下载网页
##2．清洗数据：利用正则表达式过滤非内容信息
##########################################
    ##读取网页
    #==================TO DO====================
    url <- 'http://www.pm25.in/lanzhou'
    #url <- './web.html'  #若网络不畅，我们在服务器上提供离线版本
    #学生在此编写代码#
    #学生在此编写代码#
    #==========================================

    ##数据清洗
    #初步模式筛选，查找需要的数据
    #==================TO DO====================
    pt1 <- '<table.*>'
    pt2 <- '</table>'
    #学生在此编写代码#
    #学生在此编写代码#
    #==========================================

    ##进一步清洗处理，并输出清洗后的数据
    #==================TO DO====================
    pt <- '<[^>]*?>'       #匹配HTML标签的正则表达式
    #学生在此编写代码#
    #学生在此编写代码#
    #==========================================
```

说明 1: 以上实验材料中提供基础代码，减少学生不必要的输入，学生根据提示，完成“TO DO”部分的核心代码。参考答案代码如下：

```
##读取网页
url <- 'http://www.pm25.in/lanzhou'
#url <- './web.html'  #若网络不畅，我们在服务器上提供离线版本
web <- readLines(url, encoding = 'UTF-8')

##数据清洗
#初步模式筛选，查找需要的数据
pt1 <- '<table.*>'
pt2 <- '</table>'
web <- web[grep(pt1, web):grep(pt2, web)]

##进一步清洗处理，并输出清洗后的数据
pt <- '<[^>]*?>'       #匹配HTML标签的正则表达式
cont <- gsub(pt, '', web)
cont <- gsub('\\s+', '', cont)
cont <- cont[cont!='']
cont                   #数据展现
```

断点回归模型构建

建模操作：建立断点回归模型。大气污染的形成是一个包含人类活动和自然力作用的复杂现象，我们构建机动车限行政策的空气质量效应模型如下：

$$\ln Y = \gamma_0 + \gamma_1 Res + \boldsymbol{\alpha}'\mathbf{H} + \boldsymbol{\beta}'\mathbf{W} + \varepsilon \tag{1}$$

其中，因变量Y代表空气质量指标，本实验为 NO_2；ε为随机扰动，表示未能观测因素的综合影响；研究变量为Res，其取值为

$$Res = \begin{cases} 1, & \text{单双限行} \\ 0, & \text{尾号限行} \end{cases}$$

$\boldsymbol{W}$是自然因素对空气质量影响变量的可能集合：$\boldsymbol{W}$ = (HUMI, TEMP, WIND, SEA_EFT)。其中，HUMI 为湿度、TEMP 为气温、WIND 为风力、SEA_EFT 为季节效应虚拟变量。$\boldsymbol{H}$是人类活动对空气质量影响变量的可能集合：$\boldsymbol{H}$ = (HOUR_EFT, DAY_EFT, WEEK_EFT)。其中 HOUR_EFT 为限行时段效应虚拟变量，DAY_EFT 是日期尾数虚拟变量，WEEK_EFT 为工作日效应虚拟变量。

代码操作：通过 R 语言编程，根据模型式(1)，在$\boldsymbol{W}$和$\boldsymbol{H}$选择变量子集，构建不同的空气质量效应评估模型，并运行出估计结果。学生可见提示及基础代码如下：

```
##############################################
##针对限行政策的空气质量效应建模
##过程包括：
##1. 载入数据：本案例为NO2，学生可以采用data文件
##   夹下的O3、CO、PM10、PM2.5和SO2自行建模
##2. 数据处理：筛选表2中的样本
##3. 建模分析：最小二乘估计及逐步回归
```

```
##############################################
    ##载入数据
    df <- read.csv('./data/no2_Int.csv')  #数据可修改
    start <- ISOdatetime(2013, 8, 3, 0,0,0)
    end <- ISOdatetime(2014, 8, 31, 23,0,0)
    dt <- seq(start, end, by = 'hour')

    ##数据处理：
    dep <- df[c(9)]                          #研究变量，包括应变量
    stu <- df[c(22,23,16)]                   #研究变量，包括限行政策
    wea <- df[12:14]                         #气象控制变量，包括湿度、温度和风力
    month <- df[24:34]                       #月份虚拟变量，12月为基础
    time <- df[36:58]                        #时间虚拟变量，24点为基础
    ctrl <- df[c(1)]                         #控制变量
    num <- df[c(17:20)]                      #尾号
    n <- nrow(df)
    wea1 <- wea[-c(n,1),]
    names(wea1) <- c('sd_1', 'wd_1', 'wse_1')
    wea2 <- wea[-c(n,n-1),]
    names(wea2) <- c('sd_2', 'wd_2', 'wse_2')
    df1 <- cbind(dt, dep, stu, wea)
    df1 <- df1[-c(1,2),]
    df1 <- cbind(df1, wea1)
    #限行时间res及左右外推时间rnd
    res <- c(ISOdatetime(2013, 11, 17, 0,0,0),
             ISOdatetime(2014, 1, 10, 23,0,0),
             ISOdatetime(2014, 5, 25, 0,0,0),
             ISOdatetime(2014, 7, 20, 23,0,0)
             )
    rnd <- c(ISOdatetime(2013, 10, 17, 0,0,0),
             ISOdatetime(2014, 2, 10, 23,0,0),
             ISOdatetime(2014, 4, 25, 0,0,0),
             ISOdatetime(2014, 8, 20, 23,0,0))
    winter <- df1$dt>=rnd[1] & df1$dt <= rnd[2]
    summer <- df1$dt>=rnd[3] & df1$dt <= rnd[4]
    df1$season[winter] <- 0
    df1$season[summer] <- 1
    df2 <- subset(df1, winter | summer)

    ## 建模分析(采用最小二乘估计及逐步回归进行变量筛选)
    #==================TO DO=====================
    #以下公式可在实验中修改
    st <- "log(no2, base=exp(1)) ~ . + I(sd^2) + I(wd^2) + I(wse^2) - dt"
    #学生在此编写代码#
```

```
#学生在此编写代码#
#==========================================
```

说明 1: 以上实验材料中提供基础代码，减少学生不必要的输入，学生根据提示，完成“TO DO”部分的核心代码。运行成功后，得到如下结果:

```
Call:
lm(formula = log(no2, base = exp(1)) ~ resSD + resDaily + wd +
    wse + sd_1 + wd_1 + wse_1 + season + I(sd^2) + I(wd^2) +
    I(wse^2), data = df2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.91278 -0.28913  0.04599  0.30811  1.97262

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.216e+00  3.536e-02 119.233  < 2e-16 ***
resSD       -2.111e-01  1.289e-02 -16.374  < 2e-16 ***
resDaily     4.295e-02  1.337e-02   3.212  0.00133 **
wd           1.669e-02  2.544e-03   6.561 5.82e-11 ***
wse         -2.733e-01  2.437e-02 -11.217  < 2e-16 ***
sd_1         4.083e-03  9.504e-04   4.296 1.77e-05 ***
wd_1        -7.735e-03  2.308e-03  -3.351  0.00081 ***
wse_1       -2.046e-01  1.252e-02 -16.340  < 2e-16 ***
season       3.598e-01  2.376e-02  15.144  < 2e-16 ***
I(sd^2)     -8.315e-05  9.030e-06  -9.209  < 2e-16 ***
I(wd^2)     -9.648e-04  5.509e-05 -17.513  < 2e-16 ***
I(wse^2)     3.707e-02  6.349e-03   5.839 5.55e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4523 on 5628 degrees of freedom
Multiple R-squared:  0.2302,    Adjusted R-squared:  0.2287
F-statistic:   153 on 11 and 5628 DF,  p-value: < 2.2e-16
```

从结果上看，resSD 的系数为-0.21，且效应显著。结合模型式(1)，说明限行政策在短期内有助于改善空气质量，负号表示改善。根据研究设计和等比例假设，依据公式

$$\text{限行效应} = resSD \times \frac{\text{理论限行量}}{0.3} \times 100\%$$

推算得到尾号限行(理论限行量 20%)和单双号限行(理论限行量 50%)的空气质量改善效应分别为：14%和 35%。

说明2: 建模分析采用了最小二乘估计及逐步回归。实际上是简化了断点回归的局部估计方法，本实验中采用类似于Davis L W. The effect of driving restrictions on air quality in Mexico City[J]. Journal of Political Economy. 2008, 116(1): 38-81.的全局估计，可以用最小二乘实现。这样本科生只要掌握基本的最小二乘方法，就可以理解。参考答案代码如下:

```
##载入数据
df <- read.csv('./data/no2_Int.csv')
start <- ISOdatetime(2013, 8, 3, 0,0,0)
end <- ISOdatetime(2014, 8, 31, 23,0,0)
dt <- seq(start, end, by = 'hour')
##数据处理:
```

```
dep <- df[c(9)]                          #研究变量，包括应变量
stu <- df[c(22,23,16)]                    #研究变量，包括限行政策
wea <- df[12:14]                         #气象控制变量，包括湿度、温度和风力
month <- df[24:34]                       #月份虚拟变量，12月为基础
time <- df[36:58]                        #时间虚拟变量，24点为基础
ctrl <- df[c(1)]                         #控制变量
num <- df[c(17:20)]                       #尾号
n <- nrow(df)
wea1 <- wea[-c(n,1),]
names(wea1) <- c('sd_1', 'wd_1', 'wse_1')
wea2 <- wea[-c(n,n-1),]
names(wea2) <- c('sd_2', 'wd_2', 'wse_2')
df1 <- cbind(dt, dep, stu, wea)
df1 <- df1[-c(1,2),]
df1 <- cbind(df1, wea1)
#限行时间res及左右外推时间rnd
res <- c(ISOdatetime(2013, 11, 17, 0,0,0),
         ISOdatetime(2014, 1, 10, 23,0,0),
         ISOdatetime(2014, 5, 25, 0,0,0),
         ISOdatetime(2014, 7, 20, 23,0,0)
         )
rnd <- c(ISOdatetime(2013, 10, 17, 0,0,0),
         ISOdatetime(2014, 2, 10, 23,0,0),
         ISOdatetime(2014, 4, 25, 0,0,0),
         ISOdatetime(2014, 8, 20, 23,0,0))
winter <- df1$dt>=rnd[1] & df1$dt <= rnd[2]
summer <- df1$dt>=rnd[3] & df1$dt <= rnd[4]
df1$season[winter] <- 0
df1$season[summer] <- 1
df2 <- subset(df1, winter | summer)
## 建模分析(采用最小二乘估计及逐步回归进行变量筛选)
#以下公式可在实验中修改
st <- "log(no2, base=exp(1)) ~ . + I(sd^2) + I(wd^2) + I(wse^2) - dt"
fm <- as.formula(st)
fit <- lm(fm, data = df2)
summary(fit)
f <- step(fit)
summary(f)
```