2019 年江苏省研究生数学建模科研创新实践大赛

题 目　　乘坐高铁还是传统火车的行为分析

摘　　要

本文关于乘坐高铁还是火车的出行选择问题属于二分类问题，我们引入微观经济学中的效用理论对两种出行选择的效用进行了数学建模。首先筛选出对模型效用起决定作用的影响因素，并对因素进行了量化分析；其次根据已有的数据信息，使用该模型对购买何种票做出预测，通过与实际结果进行比对验证了模型的合理性，证明了此模型能够解决此类出行方式选择的决策问题；然后使用此模型和决策树模型分别对另一组数据做出购票预测；最后对该数学模型的优缺点做出讨论，并研究乘客的购票行为和铁路客运的规律，提出相关建议供铁路部门参考。

对问题 1，首先将影响出行方式选择的因素分成三类：主观因素、客观因素、随机因素，并筛选出 6 种主要因素；然后根据微观经济学中的效用函数，利用 6 种影响因素建立了以出行总广义费用最小为最优目标的旅客效用最大化模型，并对该模型进行改进，得到了旅客购票意愿模型。

对问题 2，首先对目标数据进行预处理，从中剔除了不合理的数据；其次对各个影响因素与购票选择的相关度进行数据分析，并结合题目要求：只考虑附件一中的因素，将问题一模型中多余的因素剔除，从而得到行驶时间、舒适度、票价三个主要影响因素，进而得到了新的模型；然后利用层次分析法对模型参数求解，并根据权重建立因素之间的排序准则；还依次对旅行时间价值 T、由舒适度产生的出行广义费用 G、出行费用 Q 三个因素进行了详细的量化分析；使用该模型基于附件一已知的信息对出行结果进行了预测，实际结果中购票人数比例为：高铁人数/火车人数=111/49 =2.2653，而预测正确的人数比例为高铁人数/火车人数=99/38=2.605，高铁人数预测准确率达 89.19%，火车人数预测准确率达 77.55%，所有人数预测准确率达 137/160=85.63%，证明该模型可以在很大程度上基于已知信息对购票结果做出正确预测，说明了该模型具有合理性和科学性。

对问题 3，使用了两种方法对附件二之中每个学生下一年的购票行为作出了预测。第一种方法是使用第二问的模型来预测，预测结果为：选择高铁出行的学生为 59 位，选择火车出行的人数为 26 位，两者比例为 2.2692，与附件一中高铁人数/火车人数=111/49=2.2653 非常接近。其中购买高铁票人数所占比例为 69.4%，

购买火车票人数所占比例为30.6%。第二种方法是通过决策树模型对购票行为作出预测，预测结果为：选择高铁出行的学生为62位，选择火车出行的人数为23位，两者比例为62/23=2.6957，其中高铁人数占比为72.9%；火车人数占比为27.1%。通过分析附件一中更大量的已知数据可发现：购买高铁票人数占比69.4%，购买火车票人数占比30.6%，由此可知，两种方法中，本文所建立的模型比决策树模型更贴合实际，预测的可靠度更高。

对问题4，基于旅客效用最大化的目标，通过对北京到上海方向的160份调查问卷进行数据分析，研究了铁路运行市场的客运规律，向铁路部门提出了三点建议：1)清晰客户定位、2)灵活售票时间、3)丰富换乘组合。

关键词：出行总广义费用 旅客效用最大化模型 层次分析法 因素量化分析 决策树模型

目 录

一、问题重述

随着我国交通工具的不断更新换代，我国进入了高速铁路时代，目前高速铁路的运营里程已超过 2.5 万公里，占世界的三分之二。高速铁路大大减少了人们的出行时间，提升了出行品质，同时以安全，换乘方便，乘坐舒适等特点受到广大群众的欢迎。

对于选择高铁还是火车出行，每个人的想法都有所不同。每个人在做出行选择时会考虑到经济状况、出行目的、里程长度、时间成本、购票方便程度、个人爱好、追求舒适意愿等多种因素，对于不同的人来说是选择高铁还是火车出行都与个人所关注的因素有关，以下是需要考虑的问题：

1. 筛选出影响购买高铁票还是火车票的主要因素，并说明理由；然后建立乘客购票行为（结果）与这些因素之间关系的数学模型。

2. 附件一给出了 2019 年寒假期间某高校本科生从南京回程购票信息调查表，只考虑该附件给出的因素，并根据表中数据估计问题 1 模型中的参数，建立可供计算的具体乘客购票行为数学模型；着重对影响乘客购票行为的因素及其因素之间的关系进行量化分析并给予解释；建立某种准则对影响顾客购票行为的因素按从高到低排序。

3. 附件二给出了该高校本科生另一组从南京回程的信息调查表，假设学生们的经济状况、购票观念变化不大，预测下一年寒假每个学生的购票行为。要求列表给出每个人的购票结果，并给出购买高铁与火车票的具体人数及百分比。

4. 铁路部门对车次、车辆计划、票务管理等工作通常需要提前规划。通过研究一定区域、特定阶段、并具有代表性人群的购票行为，并分析高铁客运量与传统火车客运量的规律；最后写成建议书供铁路管理部门参考与决策。

二、模型假设

1．本文每段里程区间都开通了高铁路线和火车路线，可供选择；

2．从里程的起点到终点不考虑换乘另一种交通方式；

3．假定旅客均是理性消费者，在一定社会经济约束条件下，旅客总会对各因素做出理性综合判断，选择效用最大的出行方式；

4．铁路服务部门能提供正常的运输服务，且不考虑因自然因素和社会因素产生的突发事件而影响交通的情况；

5．从里程起点到终点的出行费用仅考虑购买高铁票或火车票的价格，不考虑其他交通费用和享受铁路部门提供的额外服务费用；

6．问题三中假设学生们的经济状况、购票观念无太大变化。

三、符号说明

符号	意义	符号	意义
U	铁路旅客选择某种出行方式的总广义费用（元）	X	影响旅客购票选择的量化因素
N	影响旅客购票选择的因素种类	β^n	某种量化因素 X 所占的权重
T	旅行时间价值（元）	G	由舒适度产生的出行广义费用（元）
Q	出行费用（元）	w	单位旅行时间价值（元/h）
g	旅客旅行疲劳恢复时间（h）	t	列车行驶时长
U_{ik}	第 K 个旅客选择高铁出行的效用	U_{jk}	第 K 个旅客选择火车出行的效用
P_{ik}	第 K 个旅客选择高铁出行的意愿	P_{jk}	第 K 个旅客选择火车出行的意愿
Gini	基尼指数	D	数据集

四、问题分析

针对乘坐高铁还是火车的选择问题，共有四个待解决的问题。我们分别对各个问题进行了分析：

在问题 1 中，首先要寻找哪些因素可能对选择结果造成影响，这些因素对做出决策的影响程度有大有小，确定影响大的主要因素用来建立数学模型。然后利用已确定的这些影响因素，结合效用函数建立一个可在高铁和火车之间做出选择的意愿模型。

在问题 2 中，由于附件一的数据信息比较多，首先应当从中剔除掉不合理的数据，以免后续计算会产生较大误差；其次应该对各个影响因素与购票选择的相关度进行数据分析，从而得到行驶时间、舒适度、票价三个主要影响因素；得到了新模型之后便可对模型参数求解，并根据权重建立因素之间的排序准则，还应该对这三个主要因素进行量化分析；并使用该模型基于附件一已知的信息做出预测，与实际结果对比，验证模型的合理性。

在问题 3 中，可以直接利用问题 2 中的模型对购票行为做出预测，并于另一种方法决策树模型预测结果进行对比，以确定本文的模型是否更优

在问题 4 中，通过查阅文献资料，对乘客的购票行为进行更深度的分析，了解高铁和火车的客运量规律，为铁路运输部门提供我们的建议。

五、 问题一

5.1 主要因素筛选

题目将铁路旅客出行方式分为两种：一种为高铁，另一种为传统火车（含直达、普客、普快、特快、动车，后文统称火车）。

查阅相关文献资料[1]可知，影响铁路旅客出行方式选择的主要因素，可归纳为以下三类，其包含的具体因素见表 5.1。

1，主观因素：指铁路旅客的主体特性。

2，客观因素： 指与旅客本身不直接相关，且旅客无法控制和决定的外部因素。

3，随机因素：指旅客因知识和认识上的差异而造成的对出行方案的理解偏差及交通流的异常变化等。

表 5.1 影响铁路旅客出行方式选择的主要因素

主观因素	客观因素	随机因素
性别、年龄、职业、收入水平、受教育程度以及出行目的、出行距离、费用来源、消费观念等	车票价格、行驶时间、列车舒适程度、购票方便程度、安全性、发车准时性、到达准时性等	突发事件对交通出行的影响、不可直接观测因素等

其中客观因素在旅客选择过程中起着决定性影响作用，并且本文的研究对象为学生，他们为同一个群体，主观因素对购票决策的影响不大，故本文考虑的主要因素是行驶时间、车票价格、舒适程度、准时性、购票方便程度、安全性。

5.2 模型建立

在微观经济学的理论体系中，效用理论是在假定消费者为了实现自身效用最大化的前提下，研究其消费心理和行为的理论。效用是指某消费主体在消费某种商品或劳务的过程中，从中获得的满足程度，它是一种心理感觉，是消费者对某种物品的主观心理评价[2]。

铁路客运部门提供的运输产品是一种无形的服务，它可以满足旅客从出发地到达目的地的空间位移需要。因此，从经济学角度来说，铁路部门的服务具有“效用”，其由诸多具有效用的客观因素组成，所以，消费者效用理论同时也适用于旅客购买铁路客运服务产品这一过程。

综上所述，旅客出行效用最大化就是旅客在出行过程中的个人满足程度达到最大化。由消费者效用理论可知，铁路旅客出行效用最大化就是旅客选择出行总广义费用最小的出行方案来满足自身空间位移需要的过程。

5.2.1 旅客效用最大化模型

铁路出行方式选择集合 V={i,j}，i 代表高铁，j 代表火车。某铁路方式对乘客的效用值可以表示为：

$$\mathrm{U}=\sum_{n=1}^{N}\beta^{n}X+\varepsilon(X)=\beta^{1}T+\beta^{2}Q+\beta^{3}\mathrm{G}+\beta^{4}S+\beta^{5}H+\beta^{6}\mathrm{R}+\varepsilon(X)$$

$\beta^{n}X$为效用固定项，$\varepsilon(X)$为效用概率项，反映的是旅客对各种主客观特性的

理解偏差及乘车偏好、不可直接观测因素的影响，在建模量化的计算中可忽略。T、Q、G、S、H、R 分别代表行驶时间、车票价格、舒适程度、准时性、购票方便程度、安全性这些主要因素产生的出行广义费用。

因此，基于铁路旅客效用最大化的出行方式效用函数可以表示为：

$$\mathrm{U}=\sum_{n=1}^{N}\beta^{n}X=\beta^{1}T+\beta^{2}Q+\beta^{3}\mathrm{G}+\beta^{4}S+\beta^{5}H+\beta^{6}R$$

对于选购高铁票，其效用函数为：

$$U_i=\beta_i^1\mathrm{T}+\beta_i^2\mathrm{Q}+\beta_i^3\mathrm{G}+\beta_i^4S+\beta_i^5H+\beta_i^6R$$

对于选购火车票，其效用函数为：

$$U_j=\beta_j^1\mathrm{T}+\beta_j^2\mathrm{Q}+\beta_j^3\mathrm{G}+\beta_j^4S+\beta_j^5H+\beta_j^6R$$

由于符合旅客效用最大化的出行方式是以出行总广义费用最小为最优目标，故旅客在而二者中做出选择之后，最终效用模型为：

$$\mathrm{U}=\min(U_i,U_j)$$

5.2.2 旅客购票意愿模型

以上模型可让旅客直观地看到其选择某种出行方式的效用，而旅客在做出决定之前，必须对多种出行方式的效用进行比较，从中选择效用最大的出行方式购票。而上述模型还不能直接达到这种目的，因此需要对模型进行改进，以使得模型函数输出可以让乘客直接在高铁和火车之间做出购票选择，改进后选择高铁出行的概率函数如下：

$$P_{ik}=\frac{U_{ik}}{U_{ik}+U_{jk}}$$

为了方便地让数据结果在二维平面上显示，更直观地对结果进行判断，对此模型进行归一化处理得：

$$P_{ik}=\frac{U_{ik}/\sum_{k=1}^{K}U_{ik}}{U_{ik}/\sum_{k=1}^{K}U_{ik}+U_{jk}/\sum_{k=1}^{K}U_{jk}}$$

同理：

$$P_{jk}=\frac{U_{jk}/\sum_{k=1}^{K}U_{jk}}{U_{ik}/\sum_{k=1}^{K}U_{ik}+U_{jk}/\sum_{k=1}^{K}U_{jk}}$$

称其为旅客购票意愿模型，P_{ik}表示第 k 个旅客对第 i 种出行方式（高铁）的选择概率。

六、 问题二

6.1 模型建立与求解

首先对附件一中的所有数据进行进行预处理，以列车行驶速度和可支配收入为指标，剔除不合理的数据元组，在进行后续步骤。

原问题附件一中给出的因素包括：起点与终点、里程长度、行驶时间、购票方便程度、可支配收入、购票价格、舒适程度。由于该旅客群体为学生，他们职业身份一致、年龄相仿、消费习惯大体相似、出行目的相同(均为放假回家)，因此就影响购票的主观因素而言，该学生群体的差异不大。对附件一中提及的因素进行数据分析，处理结果如表 6.1 所示

表 6.1 附件一各属性数据预处理分析

	平均可支配收入	返程平均里程	返程平均时长	注重舒适度	注重时间成本	平均票价
高铁	1669	1049	5.487	84.68%	84.68%	388
火车	1661	974	10.847	30.61%	40.82%	200

表 6.1

由该表可知，里程长度、可支配收入，对学生购买高铁票还是火车票的选择影响不大，而另外三个因素行驶时间、舒适程度、购票价格是影响学生购票选择的主要因素。

只考虑这三个主要因素，结合问题一中的模型，学生旅客效用函数为：

$$\text{选购高铁票：} U_i = \beta_i^1 \mathrm{T} + \beta_i^2 \mathrm{Q} + \beta_i^3 \mathrm{G}$$

$$\text{选购火车票：} U_j = \beta_j^1 \mathrm{T} + \beta_j^2 \mathrm{Q} + \beta_j^3 \mathrm{G}$$

学生旅客购票意愿函数为：

$$P_{ik} = \frac{U_{ik} / \sum_{k=1}^{K} U_{ik}}{U_{ik} / \sum_{k=1}^{K} U_{ik} + U_{jk} / \sum_{k=1}^{K} U_{jk}}$$

$$P_{jk} = \frac{U_{jk} / \sum_{k=1}^{K} U_{jk}}{U_{ik} / \sum_{k=1}^{K} U_{ik} + U_{jk} / \sum_{k=1}^{K} U_{jk}}$$

6.2 层次分析模型

6.2.1 建立层次结构模型

层次分析法（Analytic Hierarchy Process，简称 AHP）是对一些较为复杂、较为模糊的问题作出决策的简易方法，它特别适用于那些难于完全定量分析的问题。它是美国运筹学家 T. L. Saaty 教授于上世纪 70 年代初期提出的一种简便、灵活而又实用的多准则决策方法。本文中我们利用层次分析法对选购高铁和选购火车票的效用函数各属性系数进行分析求解。目标层：分析问题的预定目标或理想结果，本文中为“铁路出行方式选择”。准则层：该层包含为了实现目标所涉及的中间环节，本文主要分析“旅行时间价值”,“旅客出行费用”,“旅客乘车舒适度”。

方案层：该层包括为了实现目标可供选择的各种实施措施和决策方案等，本文为“高铁”和“火车”。具体模型结构如图 6.1 所示。

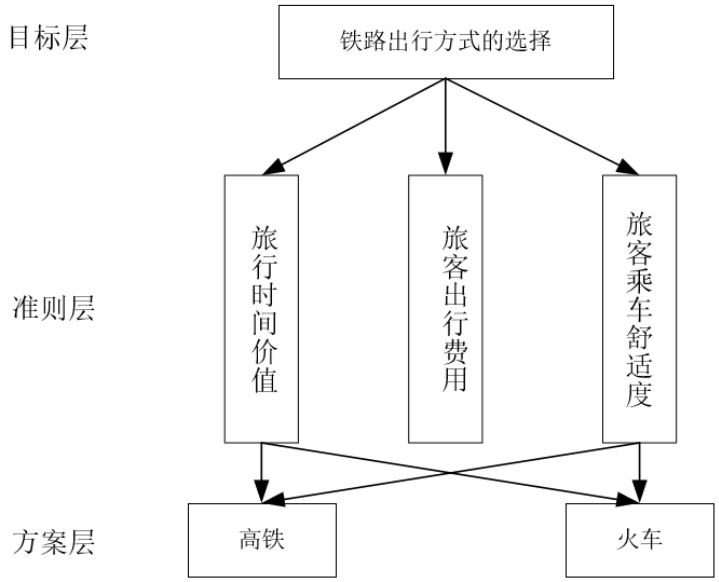

图 6.1 层次分析模型系统框图

6.2.2 构造判断矩阵

建立了铁路出行分层结构后，下面计算每一层内各要素对上一层有关因素的重要程度，即权重。各因素指标之间逐对地进行两两比较判断，同一层次的各要素两两对比的定量评价值可构成一个判断矩阵 A。这里 $A=(a_{ij})_{m\times n}, a_{ij}>0, a_{ji}=1/a_{ij}$。

A 体现了同一层次中第 i 个元素用第 j 个元素相对上一层某一个因素的重要性，并用数量化的相对权重 a_{ij} 来描述，a_{ij} 的取值采用 Satty 的 1~9 的标度方法，见表 6.1。

表 6.1　1~9 标度含义

标度	含义
1	表示两个因素相比，具有相同重要性
3	表示两个因素相比，前者比后者稍重要
5	表示两个因素相比，前者比后者明显重要
7	表示两个因素相比，前者比后者强烈重要
9	表示两个因素相比，前者比后者极端重要
2、4、6、8	表示上述两个相邻等级之间
倒数	为上面标度的倒数

6.2.3 建立判断矩阵

（1）建立判断矩阵：判断矩阵是各元素针对上一层某个元素建立起同一层任意两个元素之间评测的数据矩阵。参考文献[3]对本文研究的三个因素构造判断矩阵如下：

$$A=\begin{bmatrix}1 & 1/4 & 1/2\\ 4 & 1 & 3\\ 2 & 1/3 & 1\end{bmatrix}$$

（2）构造方案层对第二层的每一层准则的成对比较矩阵 B_n，在这里 B_n 中的元素是影响同学铁路出行方案的因素（旅行时间价值，旅客出行费用和旅客乘车舒适度）。

$$B_1=\begin{bmatrix}1 & 3\\ 1/3 & 1\end{bmatrix} \qquad B2=\begin{bmatrix}1 & 4\\ 1/4 & 1\end{bmatrix} \qquad B3=\begin{bmatrix}1 & 1/3\\ 3 & 1\end{bmatrix}$$

判断矩阵的最大特征值 λ_{max}；衡量判断矩阵偏离一致性的指标为 $CI=\frac{\lambda_{max}-n}{n-1}$（n 为判断矩阵的阶数）。一致性比率 $CR=\frac{CI}{RI}<0.1$ 时满足一致性要求。

表 6.2：随机一致性指标 RI 的数值

n	1	2	3	4	5	6	7	8	9	10	11
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51

运用 matlab 程序软件求准则层判断矩阵 A 的特征向量与特征值，求得最大特征值：λ_{max}=4.12。一致性比率为：$CR=0.0158<0.1$ 通过一致性检验。

表 6.3：高铁出行方式层次分析权重系数

参数（高铁）	β_i^1	β_i^2	β_i^3
权重	0.25	0.8	0.75

参数（火车）	β_j^1	β_j^2	β_j^3
权重	0.75	0.2	0.25

三个主要因素对购票选择影响程度的排序结果：
对于乘坐高铁出行的同学：舒适度>购票价格>行车时间
对于乘坐火车出行的同学：行车时间>购票价格>舒适度

运用求好参数的模型，将附件一中的相应的主要因素数据代入模型函数的自变量，数据在 matlab 中如图 6.2 所示，直线上方数据是经归一化处理后每个学生购买火车票的效用，直线下方数据是经归一化处理后每个学生购买高铁票的效用，红线代表分界线，数据点在红线以下表示购买高铁票出行的总广义费用小，宜选购高铁票，数据点在红线以上表示购买火车票出行的总广义费用小，宜选购火车票。

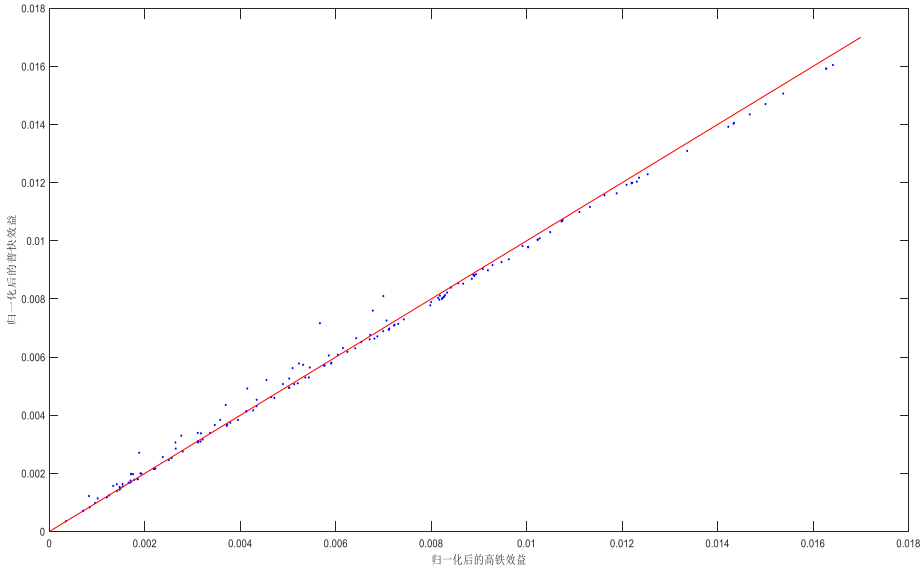

图 6.2 高铁效益/火车效益散点图

将图 6.2 中选购高铁票的乘客与选购火车票的乘客分离，如图 6.3 和图 6.4 所示。

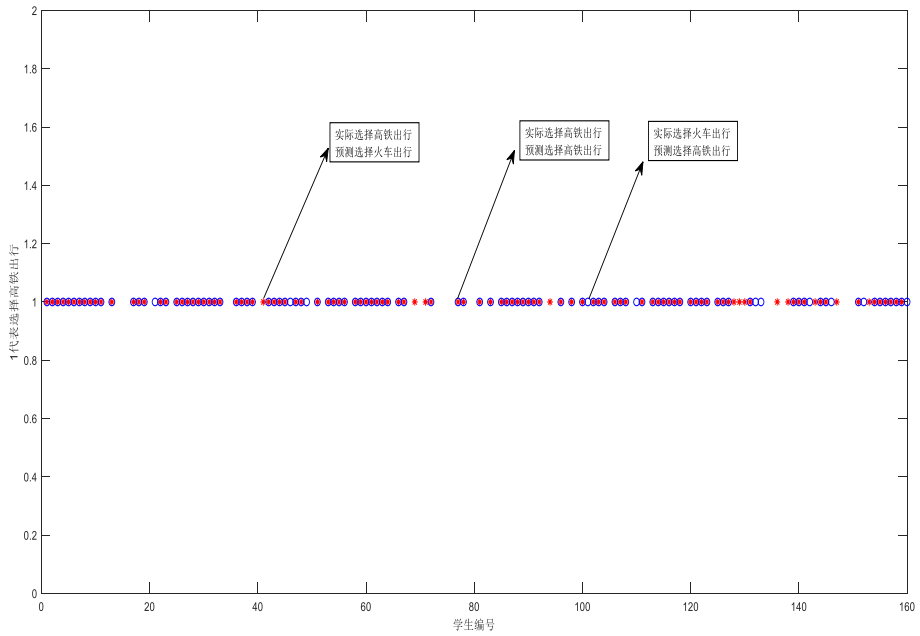

图 6.3 模型预测结果与附件一实际购票结果对比图（高铁）

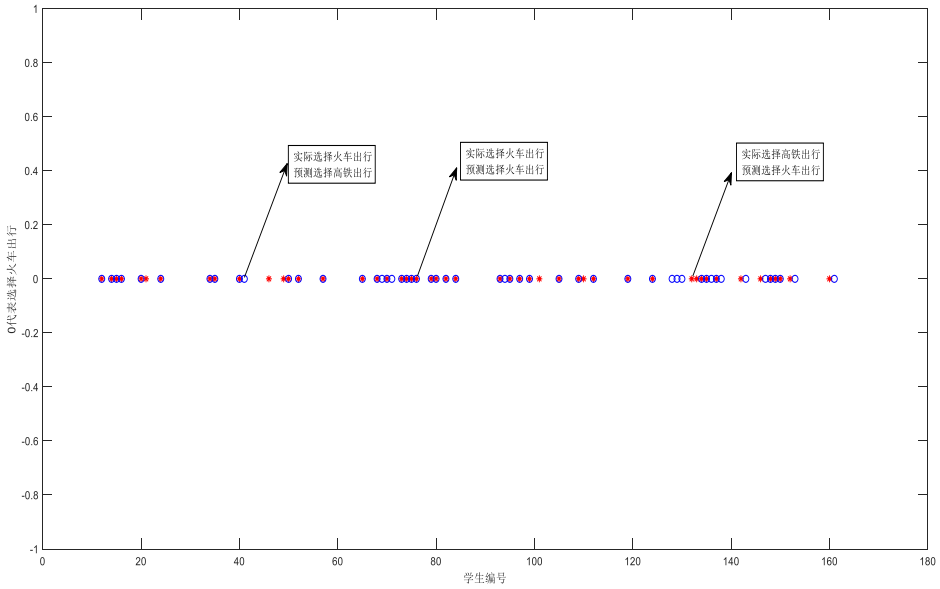

图 6.4 模型预测结果与附件一实际购票结果对比图（火车）

将上图的预测结果与附件一中每个乘客实际购票结果做对比分析可得：

实际结果：高铁人数/火车人数=111/49 = 2.2653

购买高铁票人数占比：111/160=69.4%

购买火车票人数占比：49/160=30.6%

预测正确的人数结果：高铁人数/火车人数=99/38=2.605

模型预测准确率：高铁人数预测准确率：99/111=89.19%

火车人数预测准确率：38/49=77.55%

所有人数预测准确率：137/160=85.63%

由此预测结果可知，该模型很大程度上能够正确预测购票结果，从而验证了此模型建立的合理性。

6.3 因素量化分析

1，旅客旅行时间价值 T 的量化分析

关于在交通运输项目经济评价中应用比较普遍的旅客时间价值的研究比较多[4]，本文只对典型的机会成本法进行分析说明。

假设旅客把旅行过程中节省下来的时间都用在工作和休闲娱乐两个部分，那么理论上旅客的单位旅行时间价值可表示为如下公式：

$$w= P_wB_w+ (1- P_w) B_l$$

公式中：P_w 为旅客将节约时间用于工作的概率; B_w 为旅客将节约时间用于工作的收益; $1-P_w$ 为旅客把旅行过程中的节省时间用于休闲的概率；B_l 为旅客将节约时间用于休闲的收益(闲暇时间价值)。

旅行时间价值与旅客个人工资率相关，旅行过程中节省下来的时间用在工作上产生的收益是由工作创造的实际收益减去因为工作而无法进行休闲娱乐活动所产生的收益。因此，旅客单位时间的工作收益可表示为：

$$B_w= B_{wage}- B_l$$

$$B_{wage}=Y_{wage}/(50x40)$$

其中，B_{wage} 为小时工资率，Y_{wage} 为年平均工资。

因此，可以将旅客的单位旅行时间价值表示为：

$$w= P_wB_{wage}+ B_l-2P_wB_l$$

一般情况下，用于工作的时间占旅行节约总时间的比例在 0.5 左右，为了方便计算不妨取 P_w=0.5。那么旅客的单位旅行时间价值即可简化为：

$$w=0.5xY_{wage}/(50x40)$$

故旅客的旅行时间价值:

$$T=wt$$

2，旅客出行费用 Q 的量化分析

出行费用是影响旅客出行的重要因素之一，包括：票价以及由旅行时间延长旅客的其他必要花费等。为方便研究，本文仅考虑列车票价对旅客出行选择的影响，其他费用忽略不计。即出行费用 Q=购票价格。

3，旅客乘车舒适度 G 的量化分析

随着人们生活水平的提高，旅行舒适度越来越受到重视，出行方式舒适度水平的差异对旅客出行选择的影响力日益增大。铁路旅客列车舒适度的评价标准，可归纳为以下五点：旅客列车人均占有坐卧面积、旅行时间、旅行环境、列车平稳性、客运部门服务质量。

关于如何对旅客舒适度进行量化[5]：通过旅客恢复旅行疲劳所需的时间来描述乘坐某一车次的旅客的舒适程度。基本思想为：用疲劳恢复时间量化旅客的疲劳度，然后利用旅行时间价值把疲劳恢复时间转化为疲劳恢复的时间价值，即转化为广义费用。旅行时间(t)和环境直接决定疲劳恢复时间的长短，可以利用列车类别描述车内环境，则旅客旅行疲劳恢复时间可表示为：

$$g=T_{max}/[1+\alpha_T exp(-\beta_T t)]$$

其中：

T_{max}——最大恢复疲劳时间，通常取 14～15；

α_T——旅行时间 t=0 时，乘坐某类列车的疲劳恢复时间（最小疲劳恢复时间）；

β_T——单位旅行时间的疲劳恢复时间强度系数(h^{-1})，β_T 与疲劳恢复时间成正比，β_T>0；

若 T_{max} 等于 15，那么 α_T、β_T 的取值如表 6.2 所示：

列车类别	α_T	β_T
普速列车	59	0.28
高速列车	39	0.40

表 6.2

图 6.5 是当 T_{max} 等于 15 时，两类列车对应的疲劳恢复时间曲线图。本文将火车视为普速列车，高铁视为高速列车来分析计算。

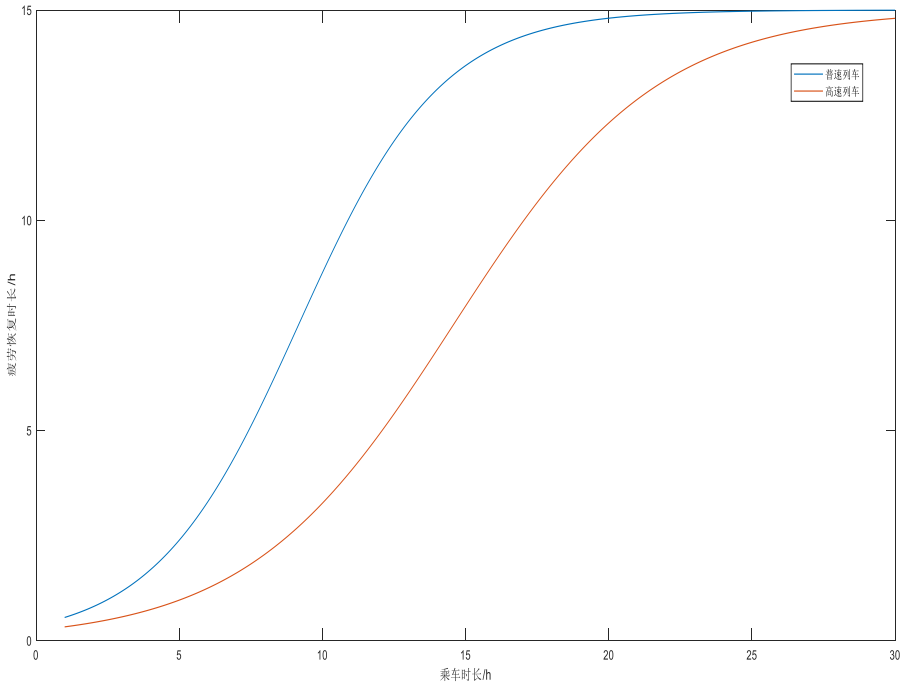

图 6.5 高铁和火车的疲劳恢复曲线

综上所述，疲劳程度越高旅客所需的恢复时间就越长，出行广义费用就越大，这将导致旅客出行的效用减小。因此，本文利用疲劳恢复时间的价值来表示铁路旅客选择某出行服务方案的舒适度价值，由舒适度产生的出行广义费用可以用下式计算：

$$G=wg$$

七、 问题三

7.1 用问题二模型预测下一年购票行为

假设学生们的经济状况、购票观念变化不大，使用问题二中的学生旅客购票意愿模型，基于附件二中的学生信息，可对下一年寒假每个学生的购票行为作出预测。

每个人的预测结果见下表：

学号	购票选择	学号	购票选择
9171XXXX0422	高铁	9171XXXX0215	高铁
9171XXXX0117	高铁	9181XXXX1000	高铁
9171XXXX0214	火车	9171XXXX0114	高铁
9171XXXX0428	火车	9171XXXX0218	高铁
9171XXXX0526	高铁	9171XXXX0426	高铁
9171XXXX0410	火车	9171XXXX0301	高铁
9171XXXX0350	火车	9171XXXX0129	高铁

9171XXXX0323	高铁	9171XXXX0110	高铁
9171XXXX0305	火车	9171XXXX0211	高铁
9171XXXX0439	高铁	9171XXXX0403	高铁
9171XXXX0147	高铁	9171XXXX0309	高铁
9171XXXX0304	高铁	9171XXXX0121	火车
9171XXXX0142	高铁	9171XXXX0308	高铁
9161XXXX0316	高铁	9171XXXX0430	火车
9171XXXX0448	高铁	9171XXXX0223	高铁
9171XXXX0243	高铁	9171XXXX0502	高铁
9171XXXX0428	高铁	9171XXXX0108	火车
9171XXXX0444	火车	9181XXXX0620	火车
9171XXXX0147	高铁	9181XXXX0809	高铁
9171XXXX0143	火车	9181XXXX0104	高铁
9171XXXX0116	高铁	9181XXXX1023	高铁
9171XXXX0429	火车	9181XXXX0513	高铁
9171XXXX0348	火车	9181XXXX0206	高铁
9171XXXX0222	高铁	9181XXXX0114	火车
9171XXXX0326	高铁	9181XXXX0415	火车
9171XXXX0310	火车	9181XXXX1002	高铁
9171XXXX0330	高铁	9181XXXX0220	火车
9171XXXX0103	高铁	9181XXXX0127	高铁
9171XXXX0111	高铁	9181XXXX0707	火车
9161XXXX0211	火车	9181XXXX0116	高铁
9171XXXX0103	高铁	9181XXXX0216	高铁
9171XXXX0111	高铁	9181XXXX0326	高铁
9171XXXX0121	火车	9181XXXX0127	高铁
9171XXXX0125	火车	9181XXXX0439	高铁
9171XXXX0316	高铁	9181XXXX0632	火车
9171XXXX0123	高铁	9181XXXX0113	火车
9171XXXX0416	火车	9181XXXX0118	高铁
9171XXXX0333	高铁	9181XXXX0503	高铁
9171XXXX0105	高铁	9171XXXX0230	火车
9171XXXX0316	高铁	9171XXXX0210	高铁
9171XXXX0141	高铁	9171XXXX0104	火车
9171XXXX0203	高铁	9171XXXX0213	高铁
9171XXXX0407	高铁		

以上数据中，选择高铁出行的学生为 59 位，选择火车出行的人数为 26 位，两者比例为：高铁/火车=59/26=2.2692；其中购买高铁票人数所占比例为 69.4%，购买火车票人数所占比例为 30.6%。

在问题二中，高铁人数/火车人数=111/49=2.2653；购买高铁票人数占比：

111/160=69.4%；购买火车票人数占比：49/160=30.6%。

此预测数据与问题二中相应的数据比例几乎完全一致，说明模型合理，预测结果可靠度高。

7.2 用决策树模型预测下一年购票行为

若将预测乘客是选择高铁还是火车的问题看作是一个分类问题，决策树通过使用基尼指数来选择最优的属性特征。假设有 K 个类，样本点属于第 K 类的概率为 p_k，则概率分布的基尼指数定义为

$$Gini(p) = \sum_{k=1}^{k} p_k(1 - p_k) = 1 - \sum_{k=1}^{k} p_k^2$$

本文关于乘坐高铁还是火车的问题是属于二分类问题，假设样本点属于第一类（乘坐高铁）的概率为 P，则概率分布的基尼指数为 $Gini(p)=2p(1-p)$，对于给定的样本集合 D，其基尼指数为

$$Gini(D) = 1 - \sum_{k=1}^{k} \left(\frac{|C_k|}{|D|} \right)^2$$

如果样本集合 D 根据特征 A 是否取某一可能值 a 被分割成 D_1 和 D_2 两部分，即

$$D_1 = \{(x, y) \in D | A(x) = a\}, \quad D_2 = D - D_1$$

则在特征属性 A 的条件下，集合 D 的基尼指数定义为

$$Gini(D, A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

直观来说，基尼指数用于表示不确定性，基尼指数值越大，样本集合的不确定也就越大，数据集的纯度也就越低。于是我们在候选属性特征集合 A 中选择使得基尼指数最小的属性为最优属性特征。

CART 生成算法：

输入：训练数据集附件 1 的数据，停止计算的条件；

输出：*CART* 决策树。

（1）设结点的训练数据集为 D，计算现有特征对该数据集的基尼指数，对每个特征属性 A 可能取的每个值 a，根据分类特征“高铁”或“火车”将数据集 D 分割成 D_1 和 D_2 两部分，并计算出 $A = a$ 时的基尼指数；

（2）选择有关特征属性 A 以及所有对应的切分点 a 中的最优特征属性和切分点，从而生成两个子结点，并实现对训练数据集的分类；

（3）对两个子结点递归到（1）和（2），直至满足停止条件；

（4）生成 *CART* 决策树。

通过在 R 软件上运行有关 *CART* 决策树算法的代码，可以得到下表中学生出行选择的预测结果：

学号	购票选择	学号	购票选择
9171XXXX0422	高铁	9171XXXX0215	高铁
9171XXXX0117	高铁	9181XXXX1000	高铁
9171XXXX0214	火车	9171XXXX0114	高铁
9171XXXX0428	火车	9171XXXX0218	高铁
9171XXXX0526	高铁	9171XXXX0426	高铁
9171XXXX0410	火车	9171XXXX0301	高铁
9171XXXX0350	火车	9171XXXX0129	高铁
9171XXXX0323	高铁	9171XXXX0110	高铁
9171XXXX0305	火车	9171XXXX0211	高铁
9171XXXX0439	高铁	9171XXXX0403	高铁
9171XXXX0147	高铁	9171XXXX0309	高铁
9171XXXX0304	高铁	9171XXXX0121	火车
9171XXXX0142	高铁	9171XXXX0308	高铁
9161XXXX0316	高铁	9171XXXX0430	高铁
9171XXXX0448	高铁	9171XXXX0223	高铁
9171XXXX0243	高铁	9171XXXX0502	高铁
9171XXXX0428	高铁	9171XXXX0108	火车
9171XXXX0444	火车	9181XXXX0620	火车
9171XXXX0147	高铁	9181XXXX0809	高铁
9171XXXX0143	高铁	9181XXXX0104	高铁
9171XXXX0116	高铁	9181XXXX1023	高铁
9171XXXX0429	火车	9181XXXX0513	高铁
9171XXXX0348	火车	9181XXXX0206	高铁
9171XXXX0222	高铁	9181XXXX0114	火车
9171XXXX0326	高铁	9181XXXX0415	火车
9171XXXX0310	火车	9181XXXX1002	高铁
9171XXXX0330	高铁	9181XXXX0220	高铁
9171XXXX0103	高铁	9181XXXX0127	高铁
9171XXXX0111	火车	9181XXXX0707	火车
9161XXXX0211	火车	9181XXXX0116	高铁
9171XXXX0103	高铁	9181XXXX0216	高铁
9171XXXX0111	火车	9181XXXX0326	高铁
9171XXXX0121	火车	9181XXXX0127	高铁
9171XXXX0125	高铁	9181XXXX0439	高铁
9171XXXX0316	高铁	9181XXXX0632	火车
9171XXXX0123	高铁	9181XXXX0113	火车
9171XXXX0416	火车	9181XXXX0118	高铁
9171XXXX0333	高铁	9181XXXX0503	高铁
9171XXXX0105	高铁	9171XXXX0230	火车

9171XXXX0316	高铁	9171XXXX0210	高铁
9171XXXX0141	高铁	9171XXXX0104	高铁
9171XXXX0203	高铁	9171XXXX0213	高铁
9171XXXX0407	高铁		

由上表的汇总结果可知，选择高铁出行的学生为 62 位，选择火车出行的人数为 23 位，两者比例为 62/23=2.6957，其中高铁人数占比为 72.9%；火车人数占比为 27.1%。

八、 问题四

快速客运网的迅猛发展，使大量的铁路客票数据随之产生。这些数据具有规模巨大、信息丰富、维度较多、结构复杂的特点，合理利用技术手段，整合数据库中存储的铁路客票信息资源，对席位的管理发售、客流数量统计分析和辅助决策提供理论依据，可以为铁路客运管理和营销人员提供决策服务，对高效组织运输调度具有重要意义；同时也为提高铁路运营水平、增强铁路在运输市场竞争能力提供有力支持。本文参考[6]对北京到上海方向的160份调查问卷进行数据分析。

旅客属性分析：男生占比53.2%，女生占比46.8%。年龄分布主要集中在23及以下以及24-30岁区间，所占比重分别为29.6%和21.5%。职业分布主要以学生和企事业单位人员为主，所占比重分别为24.7%和34.4%。通过利用效用最大化理论对已知数据进行多属性决策，旅客出行车次选择行为中，效用可以理解为车次的属性、服务水平等因素满足旅客出行需求的程度。根据效用最大化假设，旅客会确定所有可能的备选车次，并为每一车次赋予相应的效用值，最后选择一种他认为最能满足自己需求的车次，即效用最大的车次。小于30岁的群体占主导地位，与此同时职业分布也主要集中在学生和企事业单位人员，通过数据分析可知，购票时间段主要分布在距离出发日期1-2天和距离出发日期3-10天两个时间段内。这与铁律票价及退改签规定也有一定程度的联系，距离出发日期1-2天时间段内，车票保持原价，退票费为票价的10%，不可改签。距离出发日期3-10天时间段内，车票打九折，退票费为票价的20%,不可改签。旅客的职业特性会对购票时间的选择产生影响。由于旅客在确定具体出行需求的时间上的不差异，所以导致了旅客购票时间上的差异。对于学生来说，放假时间很早就确定了，学生可以提前很长时间购票；而对于工作人员来说，由于工作原因要临时出差，可能提前一周左右确定出差的时间，所以购票时间距离发车日期会比较近。根据数据统计分析现对铁路部门给出以下建议：

1：清晰客户定位

在市场发展的今天，市场的定位与营销日益精细化，仅满足铁路旅客的出行需求已经不能满足现代发展的要求。目前，单一的铁路旅客运输营销模式，很难满足广大铁路旅客的需求。所以，这需要铁路票务管理部口对铁路旅客以明确的定位，进行有针对性的服务可以很好地解决这一问题。

2：灵活售票时间

售票时间有两个含义，一个是指一天中销售火车票的时间，另一个指的提前售票车票的天数。售票天数的改变也可以影响顾客对售票方式的选择，二者相互作用。通过对提前售票时间的调整，从而诱导更多的旅客选择通过自动售票和代售方式售票，减少车站售票的压力。

3：丰富换乘组合

我国地域辽阔，经济发展参差不齐，客选择车次时，换乘系统不仅提供到达一线城市的直达铁路客票，还提供在一部分二、三线城市换乘的列车选择组合，用“1+1<2”组合给予一定的优惠措施。使到达一线城市的客流得到有效转移，一定程度上避免购票难或者突发事件发生的情况。

九、 模型评价

模型优点：

1)利用微观经济学中的效用理论建立模型，模型中的每个因素均可进行详细的量化分析，

2)模型对购票行为的结果预测贴合实际，准确度高，符合乘客购票的意愿选择。

3)创新性地将旅客效用最大化模型进行优化，进而得到旅客购票意愿模型，通过这个概率模型的结果，可直观而迅速的做出决策。

模型缺点：

1)此模型的每个主要因素均可量化，进而可求解；若针对的问题改变，因素难以量化，则求解困难。

2)由于本文的研究对象属于一个群体(学生)，主观因素对模型求解影响不大，若将对象范围扩大，则影响因素变多，模型的适用程度就有待进一步提高。

十、 参考文献

[1] 于冉冉.基于铁路旅客需求异质性的换乘服务方案优化研究与系统[D],北京：北京交通大学,2011.

[2] 赵延风.基于消费者效州理论的图书馆服务模式分析[J].科技情报开发与经济，2008,18(22):25-26.

[3] 朱桃杏.基于 AHP 方法的高速铁路旅客出行特征和影响因素分析[J].石家庄铁道大学学报（社会科学版），2017,11（3）:1-6.

[4]刘东坡.旅客旅行时间价值分析方法研究[J].华东经济管理,2003,17(4):155-156.

[5] 史峰,邓连波,黎新华,等.客运专线相关旅客列车开行方案研究[J].铁道学报,2004,26(2):16-20.

[6] 李云峰. 高速铁路平行列车条件下旅客购票行为研究[D]. 2017.

附件

R 语言实现 *CART* 算法的代码：

```
> Train<-read.csv(file.choose(), header=T, sep=“ , ”)
>Test<-read.csv(file.choose(), header=T, sep= “ , ”)
>install.packages(“rpart”)
>library(c(rpart, rpart.plot))
>formula_gt<-gt.hc~jl+sj+hc+sr+zf.bx+pj+ssd+sjcb
>rp_gt_cla<-rpart(formula_gt, Train, method=“class”)
>print(rp_gt_cla)

>rpart.plot(rp_gt_cla)
>pre_gt_cla<-predict(rp_gt_cla, Test, type=“class” )
>pre_gt_cla

>summary(pre_gt_cla)
```