# 中国人主观幸福感综合评价

15051036 刘星超

15231030 李子睿

# 1 问题的重述与分析

## 1.1 问题重述

本题要求基于全国社会幸福感调查的数据，分析影响个人主观幸福感的因素，找出各个因素对居民主观幸福感的影响及其影响能力大小；并进一步分析对于不同人群，不同因素的影响能力的变化；最后，就如何提高个人主观幸福感，向政府、公司等工作单位给出建议。

实际上，问题要求给出的是一个预测模型，也就是求出一组关于影响因素的参数，使得在该组参数下模型可以最准确的预测居民主观幸福感。所谓的主要影响因素，我们认为，就是该组数据的有无会对模型的预测能力有较大影响的因素。

## 1.2 难点与关键

问题的关键在于对所给数据的处理、分析，从庞杂的大数据中获取有效的主要信息。（1）排查删除不合理数据，并判断数据可靠性；（2）对各已知因素的相关性进行分析，确定主要因素；（3）将各主要因素的数据通过均值，归一化进行整合，以 1—5 分为标准表示为水平值，对缺失的重要数据进行补充；（4）建立幸福感与各主要因素间的预测模型；（5）检验模型正确性；（6）针对不同性别、年龄及教育程度，分析不同主导因素；（6）给出提高个人主观幸福感的建议。

# 2　模型假设

## 2.1 如果在一条调查结果中有一个调查项目明显不合理，或者出现前后明显矛盾，我们认为整条调查结果无效，会将其删除。

显然，如果一位被调查者的某一条调查结果是不合理的，我们有理由相信该调查者没有客观真实的填写问卷，从而他其余的调查结果也是不可信的。为了不被噪声影响模型的准确性，我们将其删除。

## 2.2 大部分调查者客观真实的填写了问卷。

本调查数据的来源可靠，被调查人群基数大，可以认为大部分数据是可靠的，不可靠数据很少。经过筛查，不合理数据确实是极少部分。

# 3　数据处理

## 3.1 数据来源

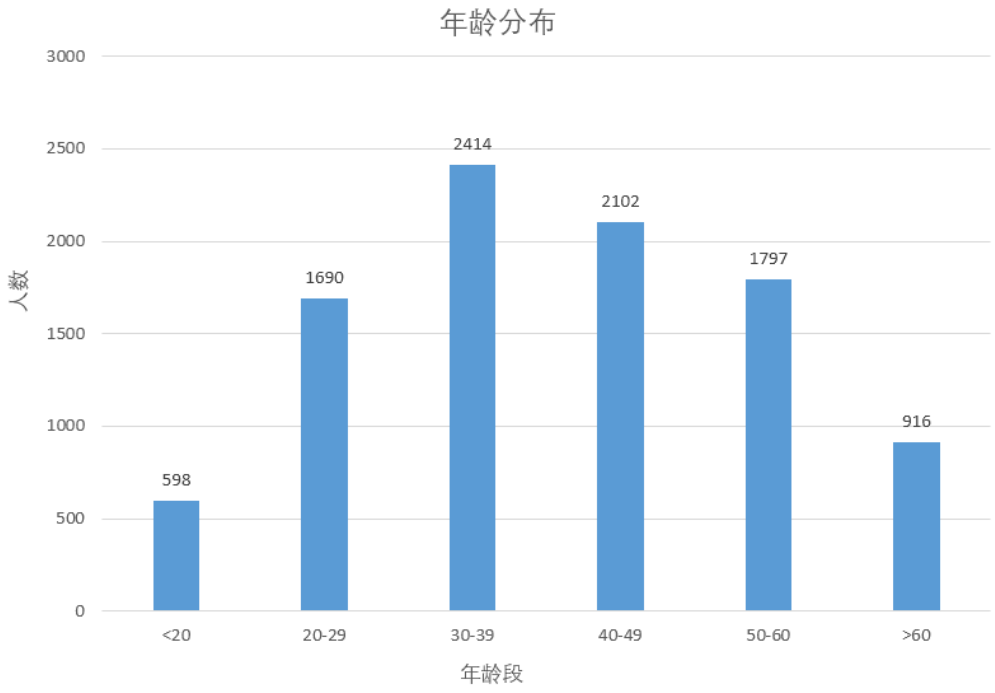
数据由中国综合社会调查项目收集、测查，随题给出。

## 3.2 数据的初步描述

该数据共有 9517 个调查结果。

### 3.2.1　年龄范围

经统计处理，截至 2005 年，调查者年龄分布如下：

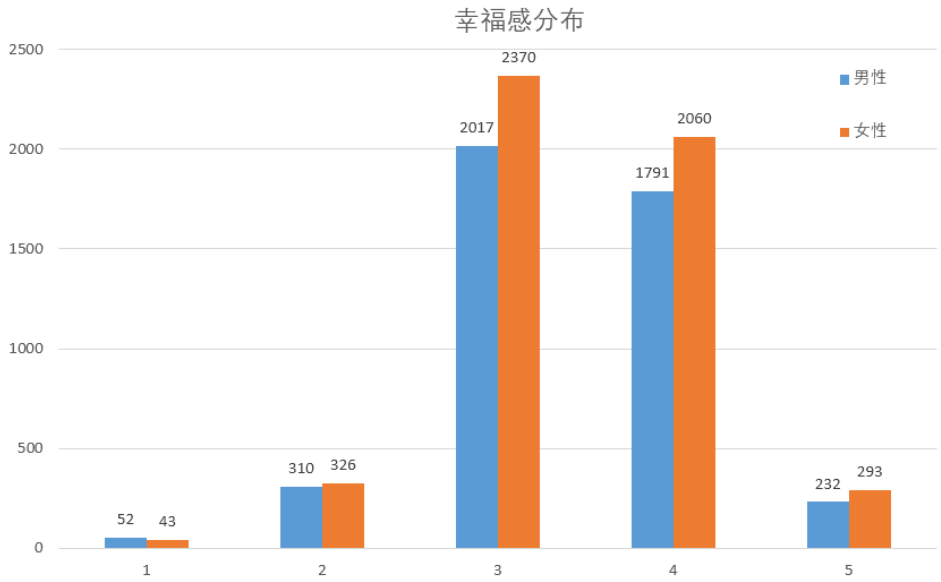



查阅中国人口年龄分布相关资料可知，所给数据与实际人口的年龄分布比例大致相同。同时，由中国人口年龄分布相关资料可知，20 至 70 岁人群约占总人口 71%，可以代表大多数人。因此，所给数据的年龄范围合理，不做删改。

### 3.2.2 性别比例

经过统计处理，该数据中男性 4412 人，女性 5105 人，男女比例约为 0.86，尽管女性相对较多，但总体比例与 1 的偏差不大。

另外，我们初步的考察了性别与幸福感的关系，结果如下。




可以看出，尽管女性人数较多，但男女幸福感的分布比例基本相同。也就是说，性别对幸福感的影响不大。本着最大化利用数据的原则，我们也不对数据根据性别进行删选。

### 3.2.3 幸福感的总体特征

9517 组数据中，全部的调查结果均有效回答了“主观幸福感”(qe49)这一问题。其平均值为 3.43，方差为 0.744。在当前中国的大环境下，居民的幸福感应该总体呈现略高于“一般”的现象。数据确实符合我们的主观感受，侧面印证了数据来源的正确性。

## 3.3 对数据的理解

查阅资料得知，幸福感受到多个方面的影响：经济因素、社会因素、人口因素、文化因素、心理因素、政治因素。给出的数据中，主要包括了经济因素、社会因素和人口因素。因此，我们也将把数据分为如上的三个方面：经济因素（如就业状况、收入水平等）、社会因素（如教育程度、婚姻质量等）、人口因素（如性别、年龄等）。显然，各个方面内部的因素互相影响，各个方面之间互相影响，同时它们也影响着居民的主观幸福感。

另一方面，在问卷调查的问题中，也有一些问题是有所关联或者相互制约的，这些问题间的逻辑关系可以判断一份数据的可靠性。

第三，有些调查结果可以整合（比如家人的教育程度等可以用平均值和标准差来描述），降低数据的维度，为后续的分析提供方便。

最后，数据中面临大面积的数据缺失（包括回答“不知道”、“拒绝回答”等），需要进行数据补全。

在数据的预处理部分，我们将进行如下工作：删除一部分不合常识或逻辑自相矛盾的数据；检查各个方面内部因素之间的关系，并据此再删除一部分数据；；整合补全数据；考察各个方面和幸福感之间的关系，初步考察哪个方面对幸福感的影响最大。

## 3.4 数据的预处理

### 3.4.1 明显不合理数据的删除

（1） 家庭社会经济地位与家庭收入情况
① 家庭的社会经济地位：下层；
和社会上其他家庭比较起来，您的家庭收入：高很多；
② 家庭的社会经济地位：上层；
和社会上其他家庭比较起来，您的家庭收入：低很多；

通过 EXCEL 筛选功能，共删除 1 个调查结果。

（2） 健康状况与个人健康状况满意度
① 健康状况 - 您自己：很好；
满意程度 - 个人健康状况：非常不满意；
② 健康状况 - 您自己：很不好；
满意程度 - 个人健康状况：非常满意；

通过 EXCEL 筛选功能，共删除 11 个调查结果

（3） 对目前生活满意度与主观幸福感
① 满意程度 - 总体而言，您对目前的生活状况是否满意：非常不满意
总体而言，您对自己所过的生活的感觉：非常幸福
② 满意程度 - 总体而言，您对目前的生活状况是否满意：非常满意
总体而言，您对自己所过的生活的感觉：非常不幸福

通过 EXCEL 筛选功能，共删除 9 个调查结果

（4） 个人全年总收入>家庭全年收入

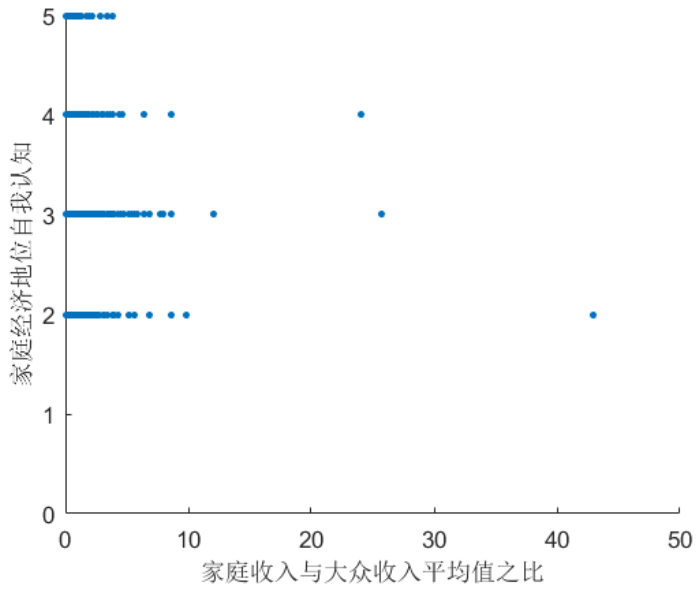
利用 EXCEL 处理，插入判断列，有判断语句：=IF(CWi>CYi,"0","1")，其中 i 遍历所有数据的行数，将个人全年总收入>家庭全年总收入的调查结果标记为“0”，其余为“1”，再通过筛选功能，查出不合理数据，共删除 2 个调查结果。

该部分共计删除了 23 组数据，余下 9494 组数据。

### 3.4.2 经济方面内部因素之间的考察

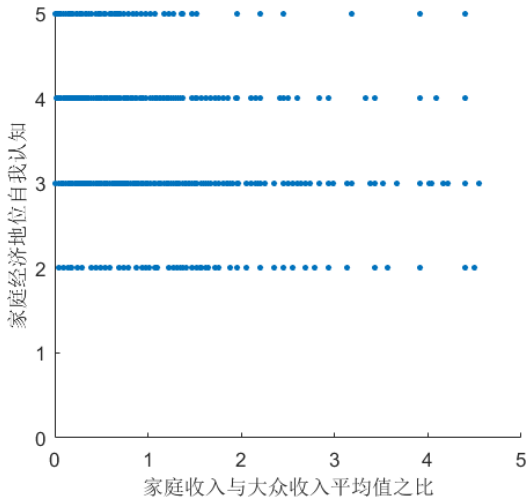
（1） 家庭收入与自我经济地位认知的关系

我们在这部分考察了 qd36a 和 qhj2 调查结果之间的关系。首先，我们筛选出了 qd36a 和 qhj2 均为有效答案的调查结果 2684 个。为了更好地体现数据的现实意义，我们把家庭收入通过家庭收入与大众收入平均值(即 2684 个收入数据的平均值)之比描述。散点图如下。




可以看出，大部分家庭的收入与平均值的比在 10 以下，同时大部分人对自己收入的认知在“3-差不多”这一档，这是符合我们常识的。在结果中出现了一个奇异点，其收入是平均值的 24 倍，但自我认知属于“4-低”。

为了去除富人对平均值的强烈影响，我们删除了家庭年收入高于 10 万的数据，余下 2636 组数据。

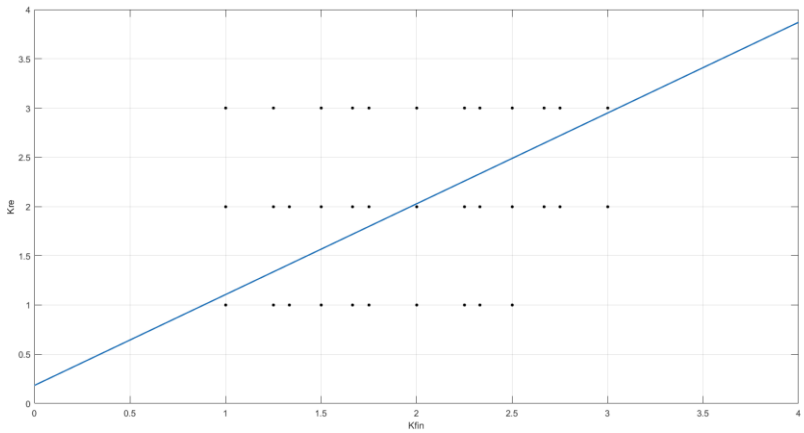



可以看出，大部分人对自己的收入认知偏差不大。基于常识，我们又去除了不合理数据：家庭收入与大众收入平均值之比小于 0.5 但家庭经济地位自我认知为高(2)的；家庭收入与大众收入平均值之比大于 2 但家庭经济地位自我认知为低很多(5)的。

（2） 对个人经济状况变化的认知

在这一部分我们考察问题 qe11_1 到 qe11_5 之间的关系。为了体现数据的实际意义，我们对数据做如下处理：若 qe11_1 到 qe11_4 的调查结果均为“4-不好说”或 qe11_5 的调查结果为“4-不好说”，则舍弃该组数据；用 qe11_1 到 qe11_4 的平均值代表这四个数据，若这 4 个数据中有 x 个“4-不好说”，我们将用其余 4-x 个数据的平均值 a 代表这 4 个数据。

我们观察数据，发现 qe11_5 的值 b 与 a 有一定的线性关系，于是用 matlab 对 a 与 b 进行了线性回归拟合。得到的结果为 b=0.9217a+0.1835,R^2 = 0.6056。由此可以看出，尽管 a 与 b 的线性关系不是非常强，但仍然具有一定的线性性。据此，我们去掉离拟合直线较远的点。

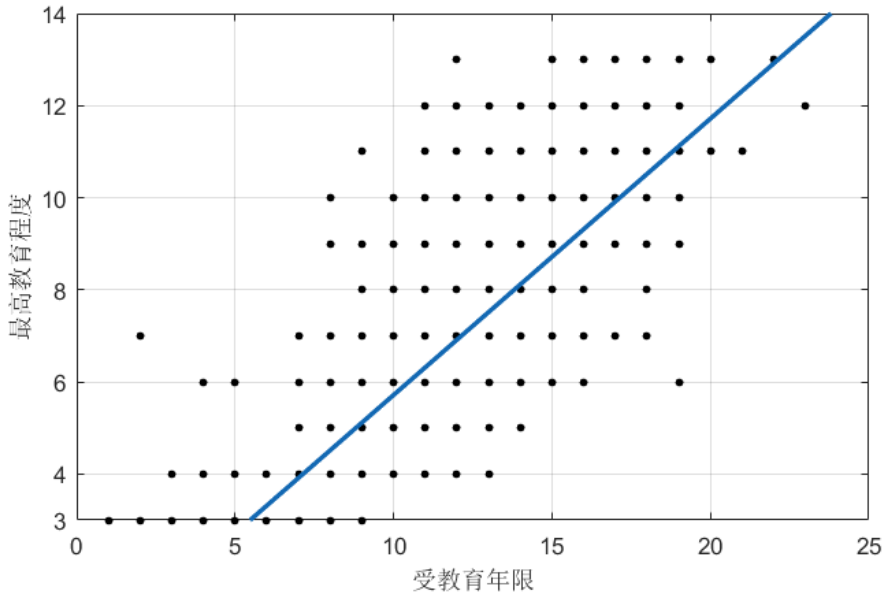


### 3.4.3 社会方面内部因素之间的考察

（1） 最高教育程度与受教育年限间的关系

这部分我们考察最高教育程度与受教育年限(自开始上小学算起)间的关系，也就是 qa05a 与 qa05d 之间的关系。显然按照常识，最高教育程度越高，受教育年限也应该越长。

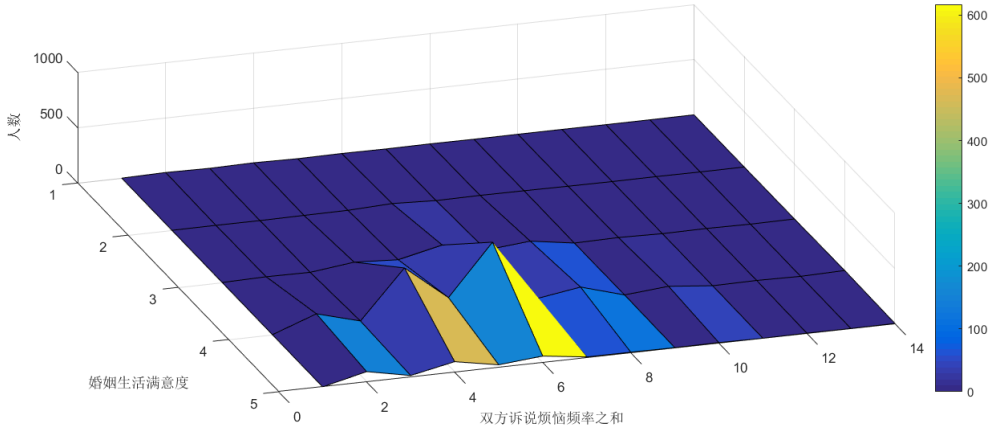
我们先对数据进行了如下处理：去除 qa05a 调查结果为“14-其他”或 qa05d 调查结果为 97、98、99 等无效答案的调查数据。随后我们对受教育年限 a 与最高教育程度 b 进行了线性拟合。拟合结果为 b=0.6002a-0.286，R^2 = 0.7537。可以看出，b 和 a 有较强的线性性，据此，我们去除离直线偏差值较大的数据。




（2） 婚姻相关问题间的关系

这部分我们考察问题 qhh01a、qhh01b 和 qhh05 之间的关系。根据常识，如果婚姻生活中二人交流较多的话，婚姻生活质量也应该更好。

我们先对数据进行了如下处理：去除没有回答或者有无效回答的数据；用 qhh01a 和 qhh01b 的和代表这两个数据。有效数据共有 2412 组。下图为数据处理结果。




可以看出，大部分数据集中在双方诉说烦恼频率之和在 1 到 12，婚姻生活满意度在 2 到 5 的区间内，据此我们删除了离群点 38 个。

### 3.4.4 人口方面内部因素之间的考察

一般而言，人口因素如年龄、性别、民族等并不受人自己的控制，一般也没有明显的关系，是相互独立了，故这一部分不做考察。

### 3.4.5 不同方面对幸福感的影响

在这一部分，我们使用支持向量机模型。支持向量机是一种机器学习模型，被广泛的应用于分类问题。

在这一部分，我们用每组数据中的 7 个调查结果(qe48_1 到 qe48_7)构成一个 7 维向量，并给每组数据一个由主观幸福感决定的标签——考虑到是粗略的估计，我们仅将数据分为两类：主观幸福感为 1-3(幸福感低于平均值的)的标签为“0”，主观幸福感为 4-5(幸福感高于平均值的)的标签为“1”。7 维向量的每一维的范围都是相同的(均为 1-5)，所以不对数据进行预处理。凡是有无效回答的数据均被舍弃。共得到 7844 组有效数据。

接下来，我们用全部数据作为训练集对支持向量机进行训练，随后用训练得到的模型再重新对全部数据进行预测。核函数采用 Sigmoid 核函数，其余设置均采用 libsvm 的默认设置。得到的模型的预测成功率是 72.8%(5710/7844)。

在没有更深入的处理的情况下，这个预测成功率相对而言已经比较高了。我们认为可以把预测正确的和预测错误的分为两组，在后续的建模中分别处理。另外，各个方面对幸福感的大致影响也可以通过支持向量机模型的参数看出。

### 3.4.6 数据的整合和补全(未来的工作)

囿于时间限制，我们并没能完成这部分工作，但仍在此处对未来工作的思路进行叙述。

这一部分我们将进行数据的整合和补全工作。

（1） 数据的整合

数据的整合主要进行了如下方面的整合：

i. 家人的婚姻状况、家人的教育程度、家人的学业完成状况等家人相关的调查结果都用调查结果的平均值和标准差描述(在计算过程中舍弃所有无效回答，包括不知道和拒绝回答等)。如果被调查者没有家人，那么考虑到此时被调查者的家人有且只有他自己，我们将用他自己的这些调查数据当作其家人调查结果的平均值。

ii. 娱乐休闲活动 9 项调查结果用它们中有效答案的平均值描述。

iii. 家人支持项中 16 项调查结果用它们有效答案的平均值和标准差描述。

通过数据整合，我们将降低数据的维度，为后续处理提供便利。

（2） 数据的补全

数据的补全主要将进行一些关键数据的补全，包括：健康状况、个人收入、家庭收入等。这些数据将基于已有的数据，通过极大似然估计或支持向量机等方法生成出。在补全这些数据并去除不合理数据后，我们将进行后续模型的处理。