数据分析与处理结课论文

论文题目：基于 Bootstrap 的银行借款、贷款与盈利关系研究

姓　　名：侯圳

学　　号：21341040103

专　　业：应用统计学

2023 年 11 月 15 日

基于 Bootstrap 的银行借款、贷款与盈利关系研究

摘要

本文旨在利用 Bootstrap 方法，研究银行借款、贷款与盈利之间的关系。首先，我们整理了数据，并对对借款、贷款和盈利三个变量进行了详细的描述性统计分析，揭示了它们的基本特征和相互关系。然后，利用 Bootstrap 方法，我们对样本数据进行了重抽样，以估计模型参数，并利用得到的参数对借款、贷款和盈利之间有关数据的关系进行了建模。

关键词：Bootstrap；借款；贷款；盈利；关系研究；回归模型

目录

表格与插图清单

基于 Bootstrap 的银行借款、贷款与盈利关系研究

一、问题描述

银行作为金融体系的核心组成部分，其经营状况对整个经济运行有着重要影响。借款、贷款和盈利是银行经营活动的三个主要方面，它们之间的关系复杂且相互影响。因此，研究这三者之间的关系对理解银行经营行为、优化资源配置、提高金融稳定性具有重要意义。

近年来，许多学者采用不同的统计方法对银行的借款、贷款和盈利之间的关系进行了研究。然而，由于银行数据的复杂性和不确定性，这些方法往往无法准确估计模型参数，从而影响了模型的预测能力和稳健性。因此，寻找一种能够准确估计模型参数的方法，一直是学术界和实务界关注的焦点。

Bootstrap 方法是一种基于重复抽样的统计方法，通过对样本数据进行重抽样，构造出一个新的样本数据集，从而实现对模型参数的准确估计。Bootstrap 方法具有操作简单、适用范围广、准确度高等优点，已被广泛应用于各种统计问题的研究.本文将采用 Bootstrap 方法，对银行的借款、贷款和盈利之间的关系进行深入研究。首先，我们对这三个变量进行描述性统计分析，以揭示它们的基本特征和相互关系。然后，利用 Bootstrap 方法对样本数据进行重抽样，以估计模型参数。最后,我们利用得到的参数对借款、贷款和盈利之间的关系进行建模，并通过预测银行未来的盈利状况，检验模型的预测能力和稳健性。

二、指标选择

（1）包括与借款、贷款和盈利等相关基础数据的描述；

（2）基础数据的深入统计分析，包括均值、方差等；

（3）利用统计学、数学模型等方面知识来获取其的分布规律；

（3）利用 Bootstrap 方法对借款、贷款和盈利之间的关系的参数。

三、数据描述

（一）数据的整理

基于附件“bank.marketing.training、bank_reg_training”的数据，本文获取了相关的量，包括银行客户的年龄、工作等基本信息，也包含了与借款、贷款和盈利等相关基础的数据。但是由于文本格式，我们不能直接使用其数据，通过 except 分列功能将其自动分列为正常数据集。

（二）数据预处理

数据预处理是指在进行数据分析或建模之前，对原始数据进行清洗、转换和整理的过程。数据预处理的目标是提高数据质量、减少错误和噪音对分析结果的影响，以及使数据适应所需的方法和模型。

本题选择用 spsspro 软件进行数据的预处理，基于 3sigma 原则，由于数据样本量大于 100，属于统计学意义上的大样本，故将各项数据异常值直接替换为均值即可。

$$u_i = \bar{x_i} \tag{1}$$

其中 i 为数据类别数量，$\bar{x_i}$为各项属性数据的平均值，u_i为各项异常值。

（三）数据的分析

现在本文选取其中的某些数据进行统计分析。

1. 被查者基本信息统计

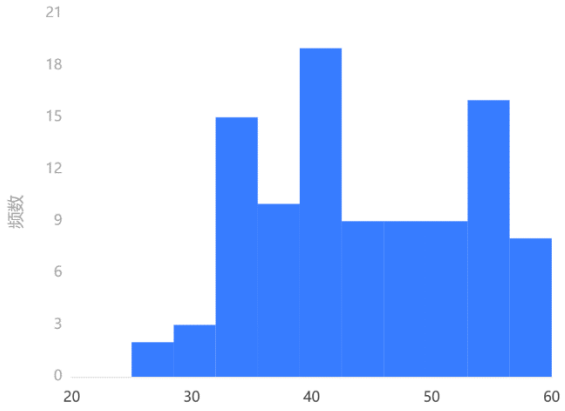

图 14　年龄频数图

表 5　年龄分析表

类型	定量	最大值	60
样本量	100	最小值	25
缺失值	0	中位数	43
去重量	33	变异系数	0.199
平均值	44.29	方差	77.582
标准差	8.808	S-W 正态检验	不满足 (P=0.006***)

上述数据分析显示被调查者年龄大都集中在中年附件的阶段，此阶段工作比较稳定，且生活压力大多数情况下比其它年龄段小，因此会更有意愿和机会去进行金融方面的交易，比如说贷款、借贷、投资、理财等活动。

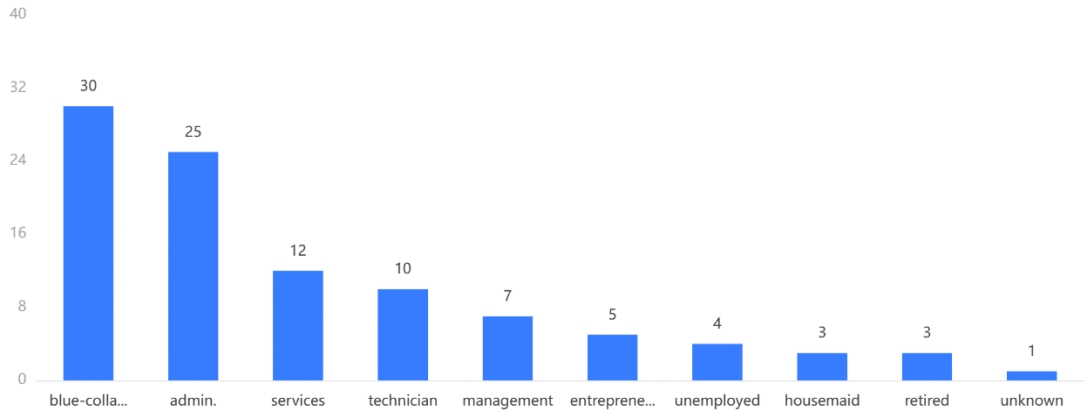

图 15　工作数据概述

在统计到的数据中，蓝领、管理的类型远远大于其它工作类型的数量，可能是因为此类工作平均薪酬更高，且接触到的资源及想法更超前，因此进行金融交易的占比较高。而退休、保姆、无业者等群体，可能对金融活动的意愿不是很强烈。

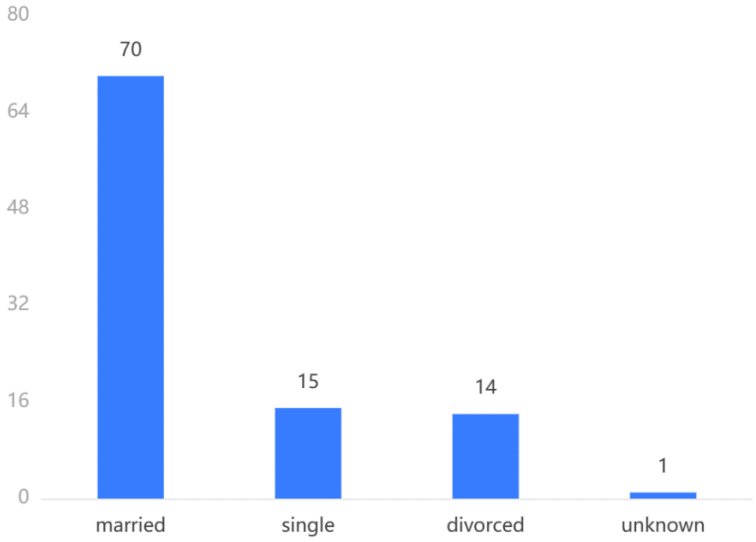

图 16　结婚状态数据图

此项数据表明了参与金融活动的人大多数是成婚状态，这与年龄、经济状况有着密切的联系。

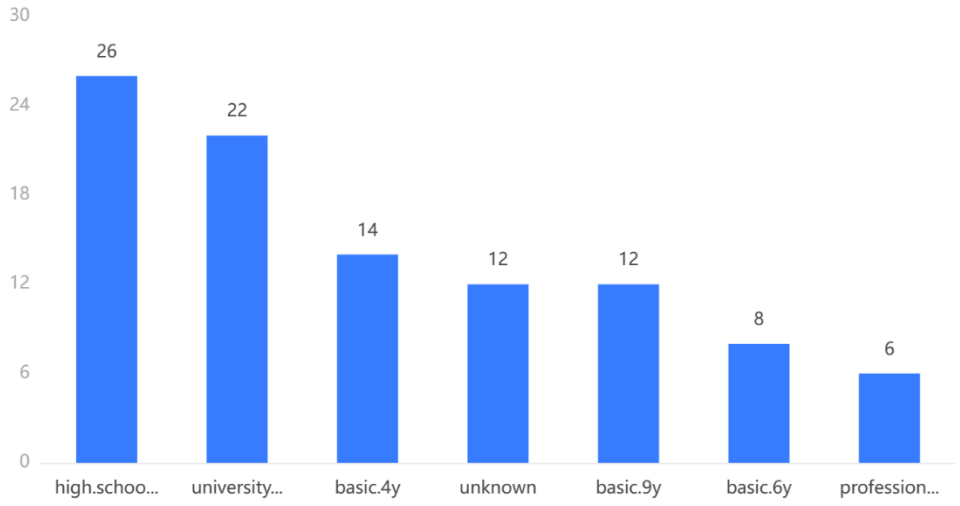

图 17　受教育程度分布图

结果显示受中等及高等教育的人数占了较大的比例，这与实际情况是一致的。

2. 相关性检验

相关性检验（correlation test）是一种统计方法，用来衡量两个变量之间的相关程度。它可以帮助确定两个变量是否具有统计上显著的相关性，以及相关性的强弱和方向。

常见的相关性检验方法有皮尔逊相关系数（Pearson correlation coefficient）、斯皮尔曼秩相关系数（Spearman rank correlation coefficient）和肯德尔秩相关系数（Kendall rank correlation coefficient）等。

皮尔逊相关系数适用于两个连续变量之间的线性相关性检验，其取值范围在-1 到 1 之间。当相关系数接近 1 时，表示两个变量呈正相关；当相关系数接近-1 时，表示两个变量呈负相关；当相关系数接近 0 时，表示两个变量之间没有线性相关性。

斯皮尔曼秩相关系数适用于两个变量之间的非线性相关性检验，其基于变量的秩次而非原始数值进行计算，取值范围同样在-1 到 1 之间。

肯德尔秩相关系数也适用于非线性相关性检验，它主要用于评估两个变量的等级之间的关系，取值范围同样在-1 到 1 之间。通过进行相关性检验，我们可以了解到两个变量之间是否存在显著的相关性，从而帮助我们理解它们之间的关系及可能的影响。

在进行相关性检验前，本文需要对其数据进行选择，基于 excel 数据筛选功能，在 loan 中选取 yes 的数据，在 previous_outcome 中选取 success 的数据量。

数据部分展示如下：

表 6　筛选后的数据部分展示表

50	self-employed	married	basic.9y	no
28	technician	single	high.school	no
48	blue-collar	married	basic.4y	no
40	admin.	married	basic.9y	no
27	technician	single	basic.9y	no
28	admin.	single	basic.9y	no
34	management	married	university.degree	no
43	blue-collar	married	basic.4y	no
39	technician	married	unknown	no

接下来进行如下操作：

（1） 数据线性检验

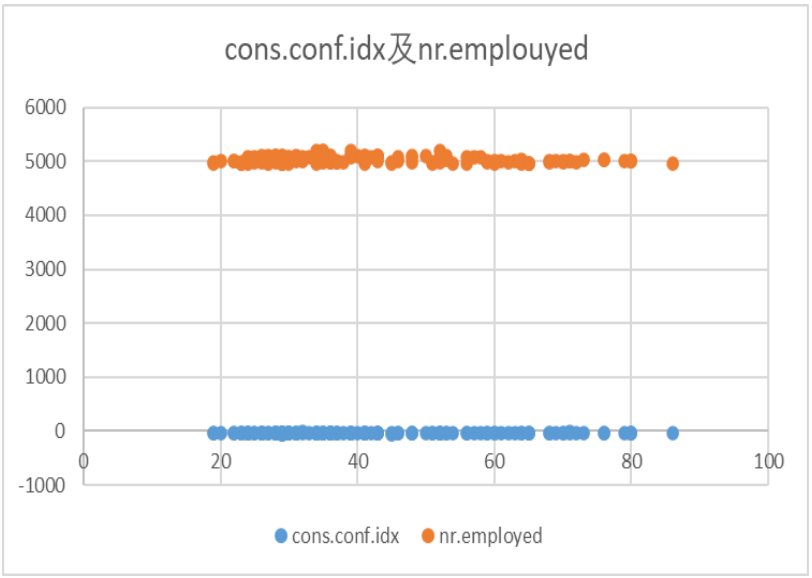

图 18 cons.conf.idx 和 nr.emplouyed 线性图

（2） 数据正态检验

通过数据直方图可直接观察出其的正态性