电子科技大学

University of Electronic Science and Technology of China

计量经济学小组课题

基于某信贷公司的个人贷款逾期行为与坏账金额影响因素分析

摘　要

本文针对金融领域信用贷款的审核、评估、预测这一复杂难题，运用计量经济学方法进行了深入研究。

首先，本文对现有信贷模型的理论基础和文献内容进行了研讨分析，以初步了解影响还款行为的潜在因素。随后，我们对某公司的两个公开数据集的基本数据特征、分布特征、相关系数进行了简要分析。

在此基础上，我们分别针对这两个数据集构建了还款能力模型和坏账金额预测模型。

在还款能力模型中，我们考虑了多个影响因素，如申请人的年龄、婚姻状况、子女个数、教育程度和所在城市 GDP 评级等。我们采用 Logit 模型分析申请人的还款能力，以帮助贷款机构在审核贷款申请时更准确地评估申请人的信用风险。

坏账金额预测模型则旨在预测贷款违约时的坏账金额。通过分析违约贷款的债务资产比、信用评级、年龄等因素，我们拟合出了一个线性回归模型。并对模型进行了多重共线性、模型设定、序列相关性、异方差的检验及修正。

本文的研究结论具有一定的理论和实践意义。首先，还款能力模型为贷款机构提供了一个有效的评估申请人信用风险的工具，有助于降低贷款违约风险。其次，坏账金额预测模型为贷款机构在面临坏账时提供了预警和应对策略，有助于减轻贷款损失。此外，这两个模型还可为其他金融机构和企业在信贷审核、评估和预测方面提供借鉴和参考。

总之，本文通过深入研究信用贷款的审核、评估、预测问题，为金融领域提供了一种有效的信用风险管理手段。

关键词：信用贷款, 还款预测, 坏账估计, Logit 模型, 线性回归

目　录

1 选题背景与调研

本节将首先阐述本题目的现实背景，从社会现象分析课题的现实意义。接着，通过文献调研，总结提出了初步的研究见解。此外，本节还将对一些基础理论进行详细说明，为后续的实证研究奠定理论基础。

1.1 选题背景与社会现状

近年来，我国经济持续高速发展，人民生活水平不断提高，消费观念发生了深刻变化。越来越多的人开始倾向于”提前消费”，即将未来的财富提前到现在消费，这已成为一种新的消费趋势。

这种消费观念的转变推动了金融信贷市场的快速发展。用未来的钱来满足现在的需求，金融信贷市场也得以日益强大。阿里的花呗、蚂蚁积分、人人贷等金融网贷平台凭借高效便捷的服务迅速吸引力大量的群众。中国人民银行发布数据显示，2021 年末，金融机构人民币各项贷款余额 192.69 万亿元，同比增长 11.6%；全年人民币贷款增加 19.95 万亿元，同比多增 3150 亿元。[1]

贷款数量的增加必然会带来金融风险的增加，由于信息不对称导致金融网络借贷信用风险也日益凸显。一些信用良好且还款能力强的客户往往也因此被拒之门外。同时每年都有众多金融信贷机构因为资金链断裂、违约、骗贷等原因而倒闭停业，其数量已高达上千家。因此，如何回答金融机构该不该借钱？如何有效管理贷款人情况，做出准确风险判断？当预期贷款人已经不能正常还款，如何准确预估坏账的金额大小？这一系列问题，成为了亟需解决的问题。

基于以上现象，我们小组决定对以下两个主要问题进行探究：

(a) 判断借款人是否有能力还款的可能影响因素有哪些？

(b) 如何估计某一贷款的坏账金额？

1.2 文献调研与理论基础

为了初步探究逾期行为与坏账大小的影响因素，小组进行了相关文献调查并进行了理论研究。

一些国外的研究认为影响贷款逾期和坏账的因素包括贷款政策、信誉衡量标准、绩效指标、宏观经济环境和银行特定因素等。Ikram 等人 (2016) 以及 Nikolopoulos 和 Tsalas (2017) 对不良贷款的影响因素进行了讨论，提到了经济增长、通货膨胀、房地产价格、利率和汇率等宏观经济因素，以及银行对借款人或抵押品的评估、法律和监管环境的影响 [2, 3]。Fennee Chong (2021) 对影响贷款拖欠的因素进行了调查，发现借款人与贷款人的距离、抵押品、教育水平和每月预算具有显著影响，但是收入和性别则没有显著影响 [4]。Matteo Foglia (2022) 调查了宏观经济因素对意大利银行系统不良贷款的影响，发现生产总值和公共债务对不良贷款有负面影响，而失业率和国内信贷对减值贷款产生积极影响 [5]。

国内关于信用贷款逾期行为影响因素的研究也十分广泛。在《互联网贷款个人信用风险影响因素研究综述》[6] 中，作者将贷款逾期行为的主要因素简单分为三类。

一是如**借款人个体显著特征**，如学历、年龄、婚姻，该类影响因素占主要成分。我们在文献《流量覆盖风险——网络小额信贷风险控制新思路》[7] 中也找到了对这一观点的实证。其实证显示女性相比男性更谨慎信用风险更低、年龄较大的借款人违约率更低、受教育程度与违约率呈负相关、已婚人士比未婚和单身人士信用更优，婚姻稳定者违约率较低，收入较高的借款人拥有更低的违约率。

二是**借款人关联特征**，如家庭、资产、居住城市，该类影响因素对一个人的贷款信用高低有着一定的关联。根据文献《城市信用环境对个人贷款保证保险违约风险影响的研究》[8] 的调研显示，良好的城市信用环境，会降低违约风险

三是**历史记录与贷款产品**特征，这类数据通常是由信贷机构基于对用户画像的评分和用户申请时填写，如历史违约次数、申请金额高低。文献《P2P 网贷借款人信用风险因素分析与对策》[9] 谈到逾期次数与违约概率直接正相关，成功次数多者申请次数越多，违约比例越小，同时，违约较多发生在申请次数较少的新借款人中。

在相关文献中，我们也看到了两个比较新颖的评判指标——`ltv` 和 `ita`。

其相关计算公式如下。

$$ltv = \frac{loanAmount}{Asset}$$

$$ita = \frac{income}{annAmount}$$

可以看到，ltv 反映了借款人借款金额和拥有资产的比率，ita 反映了借款人收入和每年还款金额的比例。这两个指标在信用贷款的评估中是一个被广泛使用的度量标准。例如，在文献《个人住房抵押贷款的市场风险研究》[10]显示 LTV 值越大，违约风险越大。在这种情况下，若借款人违约，银行收回的资产可能无法弥补贷款损失，从而导致银行的风险暴露越大。此外，较高的 ltv 值也可能意味着借款人负担过重，还款压力较大，进一步增加了违约的可能性。因此，银行和金融机构在审批贷款时，通常会密切关注 ltv 值，对过高 ltv 值的贷款申请进行更严格的审核，以确保风险可控。

ita 值则是衡量借款人收入偿还能力的重要指标。较高的 ita 值意味着借款人在还款过程中承担了较大的负担，可能会影响其生活质量并加大违约风险。银行在评估贷款申请时，会通过分析 ita 值来判断借款人的还款能力。若 ita 值过高，银行可能会认为借款人承受不了如此大的还款压力，从而拒绝贷款申请或采取更高的利率措施以降低风险。

2 数据收集与分析

本节重点阐述了数据集的来源，并对数据集进行了简洁的概述。同时，我们对数据集实施了初步的预处理工作。之后，将预处理后的数据导入到 Eviews 软件中，对其数据特征进行了分析。在这个过程中，我们初步探讨了数据的内在规律和联系，为后续的研究奠定了坚实的基础。

2.1 数据来源与初步处理

2.1.1 数据集概述

为了深入研究借款人还款能力的影响因素以及估算某一贷款的坏账金额，我们从 Kaggle 网站的大数据集中获取了两个较为有用的数据集信息。这两个数据集分别涵盖了各类借款人的信用记录和贷款情况，为我们提供了丰富的数据样本。通过对这些数据进行深入分析，我们期望能够发现借款人还款能力的关键影响因素，并为金融行业提供有效的风险防控措施。相关数据集概况如下：

I. 还款能力：贷款审批数据集

第一个数据集包含了大量借款人的信用记录，包括借款人的基本信息、财务状况、资产情况等多个方面的数据。这个数据集为我们研究借款人还款能力的影响因素提供了丰富的数据支持。我们可以通过构建 Logit 模型，探究哪些因素是影响申请人还款的重要因素。数据集见参考文献 [11]

II. 坏账金额：某信贷公司的两轮贷款业务中的坏账数据集

第二个数据集聚焦于贷款的违约情况，包含了贷款的坏账金额 (违约损失) 等关键指标。利用这个数据集，我们可以对不同情况的贷款申请进行坏账风险评估，并预测可能的坏账金额。这对于金融机构在审慎放贷、防范金融坏账风险方面具有重要的指导意义。数据集见参考文献 [12]。

2.1.2 数据预处理

在将数据导入 Eviews 之前，我们需要进行一系列简单的处理，以便为后续的分析做好准备。

首先，我们需要剔除数据集中的异常值。以第二个数据集为例，其中的性别信息被分为三个变量，分别为“MALE”表示男性、“FEMALE”表示女性，以及“OTHER”表示其他。第三个变量“OTHER”的出现，是由于数据集在性别这一项的录入存在不足，部分用户并未填写性别信息。因此，我们需要将这些异常值排除在外。

其次，我们需要对数据集中的某些变量设置虚拟变量。以第一个数据集中的堆对城市评级为例，其评级由字符串“GOOD”，“NORMAL”和“BAD”表示。然而，Eviews 无法直接处理和标记这类字符串变量。为此，我们需要事先将这些变量转换为 1 和 0 的数值形式，设定两个虚拟变量“POOR”和“RICH”，分别标识该地区评级是否为“GOOD”、该地区评级是否为“BAD”。以便在后续的分析过程中能够顺利进行。

综上所述，在数据集处理过程中，我们需要进行两项主要操作：一是排除异常值，二是设置虚拟变量。这样，在完成这些基础工作之后，我们就可以继续进行下一步的数据分析了。

2.2 特征解读与简要分析

2.2.1 数据集 I：还款能力数据集

我们将数据集 I 导入到 Eviews 软件中，进行初步的数据分布分析。

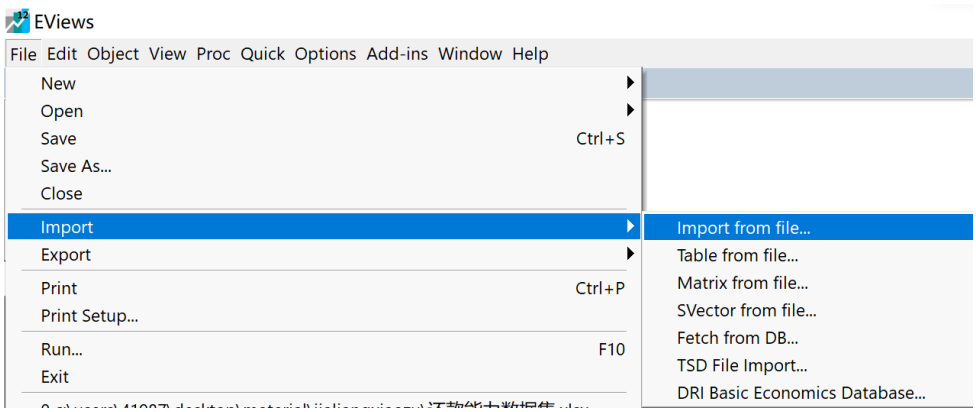

图 1: 导入数据到 Eviews 中

本数据集的全部变量的解释如表1所示：

表 1: 数据集 I 变量组成描述

变量名	解释	取值
target	借款人是否存在还款困难*	虚拟变量，是为 1，不是则为 0
annamount	借款人每年还款金额	整数
age	借款人的年龄	整数
income	借款人的收入	整数
gend	借款人的性别	虚拟变量，男性为 1，女性则为 0
family	家庭人口数量	整数
children	抚养子女的数量	整数
graduate	是否大学毕业	虚拟变量，是为 1，不是则为 0
marry	当前是否已婚	虚拟变量，是为 1，不是则为 0
car	借款人是否拥有车辆	虚拟变量，是为 1，不是则为 0
house	借款人是否拥有公寓或者住宅	虚拟变量，是为 1，不是则为 0
population	借款人居住城市的人口数量	整数
rich	借款人居住城市的 GDP 评级是否为优**	虚拟变量，是为 1，不是则为 0
poverty	借款人居住城市的 GDP 评级是否为差**	虚拟变量，是为 1，不是则为 0

* 相比“借款人是否最终还上款”，我们更关注借款人**在还款期间**是否存在还款困难的情况，**而不是只是最后的时间点**。所以变量的设定不是“借款人是否还清”，请注意。

** 城市的评级分为“优”、“中”、“差”，以信贷机构给出的评级为准

如图2导入结果所示，数据集共含有 27569 个有效数据。值得说明的是，根据教材，在大样本下，极大似然估计量（Logit 模型）仍有无偏性和最小方差性，因此在这里不必考虑大样本对显著性的影响。图3显示了整数字段（包括 age、annamount、children、family、income、population）的分布情况。可

以看到，年龄字段分布最为均衡，年度还款金额和收入水平较为集中。借款人申请的贷款年金和收入水平差距不大。

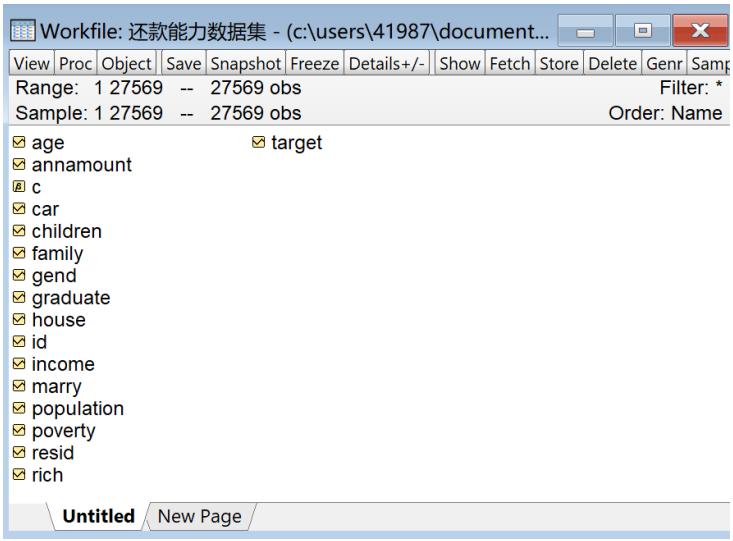

图 2: 数据集 I: 还款能力数据集变量导入结果

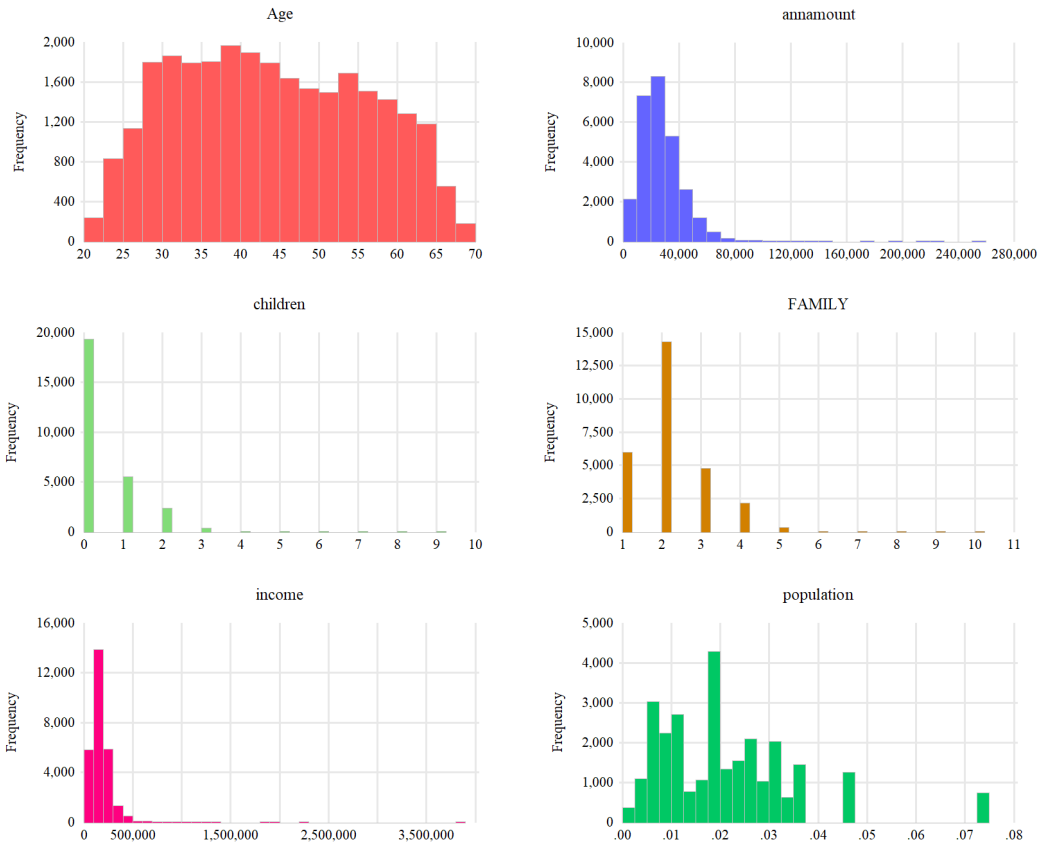

图 3: 数据集 I: 整数字段分布

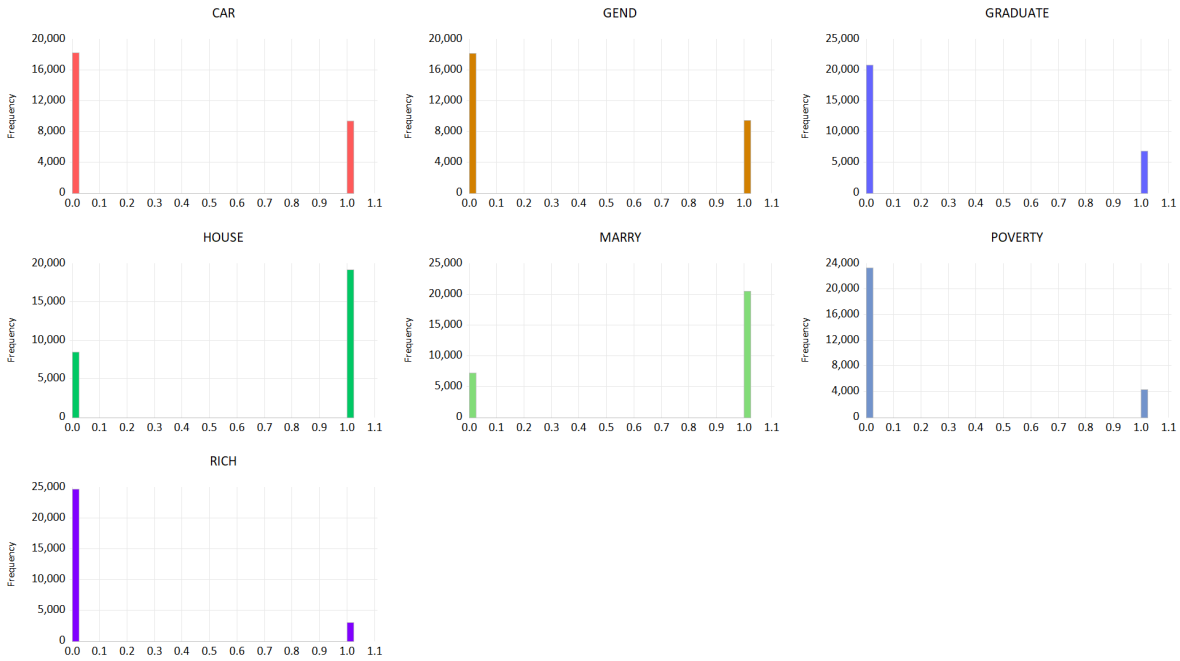

图 4: 虚拟变量字段分布

图4展示了虚拟变量字段（包括 car、gend、graduate、house、marry、proverty、rich）的分布情况。可以看到，性别字段分布最为均衡，rich 变量的分布较为集中。可能的原因是信贷机构评估的“好”的城市数量较少，我们在后续会结合 proverty 字段一同来对目标变量作回归分析。

综合来看，数据集数据分布均衡。下面对数据集特征作简要分析。

表2展示了数据集 I 中整数数据样本特征。例如，我们可以看到，数据集中平均年龄为 43 岁，借款样本中最高收入为 258025.5，最低收入为 2052.0，分布广泛。各个数据的均值、方差、标准差等均在正常范围以内，认为导入的数据正确无误。

表3展示了数据集 I 中整数数据之间的的相关系数，可以看到，除 family 和 children 外，其他变量的相关系数均低于 0.8。family 和 children 都表征了样本数据家庭人口情况，因此其相关系数较大，在后续变量选取中，将纳入最好的一个。由于版面原因，我们没有办法展示虚拟变量之间的相关系数和数据描述，但是经过我们的实操，发现各个虚拟变量与其他变量的相关系数均小于 0.8。

表 2: 数据集 II：整数数据样本特征

	AGE	ANNAMOUNT	CHILDREN	FAMILY	INCOME	POPULATION
Mean	43.92707	27177.27	0.416156	2.157858	168485.5	0.020786
Median	43.16438	24997.50	0.000000	2.000000	148500.0	0.018850
Maximum	68.99178	258025.5	9.000000	10.00000	3825000.	0.072508
Minimum	21.04110	2052.000	0.000000	1.000000	25650.00	0.000533
Std. Dev.	11.94650	14662.61	0.722796	0.908802	98940.63	0.013777
Skewness	0.117600	1.816131	1.901573	0.959529	5.043504	1.478870
Kurtosis	1.951234	14.72883	7.599771	4.649344	100.0198	6.226274
Jarque-Bera	1327.023	173177.9	40919.07	7355.318	10929496	22005.89
Probability	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Sum	1211025.	7.49E+08	11473.00	59490.00	4.64E+09	573.0511
Sum Sq. Dev.	3934473.	5.93E+12	14402.44	22769.00	2.70E+14	5.232924
Observations	27569	27569	27569	27569	27569	27569

表 3: 数据集 II：整数数据样本相关系数表

	AGE	**ANNAMOUNT**	**CHILDREN**	**FAMILY**	**INCOME**	**POPULATION**
AGE	1.000000	-0.008217	-0.332034	-0.282121	-0.065086	0.033135
ANNAMOUNT	-0.008217	1.000000	0.027175	0.078604	0.449707	0.115412
CHILDREN	-0.332034	0.027175	1.000000	**0.880504**	0.032303	-0.025432
FAMILY	-0.282121	0.078604	**0.880504**	1.000000	0.038767	-0.022765
INCOME	-0.065086	0.449707	0.032303	0.038767	1.000000	0.176098
POPULATION	0.033135	0.115412	-0.025432	-0.022765	0.176098	1.000000

* 受限与版面原因，无法展示各虚拟变量与其他变量的相关系数，但是他们都是小于 0.8 的

值得说明的是，poplutaion 与 rich、proverty 看似均反映了借款人所在城市的发展情况，但实际上，城市人口数量与城市富有的关联度并不高，其相关系数均小于 0.8。例如，在各个省份（或国外的州）中，城市人口较少的地方也可能出现富裕的情况。这可能是由于经常遭受自然灾害等原因，导致人们不愿意在那里居住。因此，在评估借款人的信用风险时，不应仅依赖城市人口数量来判断其所在城市的经济发展状况。

除此以外，根据文献调研结果，我们还引入了收入与债务比（ita）这一关键变量。该变量代表个人债务与收入之间的比值，旨在衡量个体所承受的还款压力。根据众多研究成果，收入债务比作为一个重要指标，能够直观地反映人们在承担债务时的财务状况和还款能力。通过对收入债务比的分析，我们可以更好地了解一个人的财务健康状况，以及他们在面临还款压力时可能会遇到的困难。这一指标有助于金融机构更加精确地评估借款人的信用风险，从而制定更为合理的风险控制策略，降低不良贷款率。

2.2.2 数据集 II：坏账金额数据集

与数据集 I 方法相同，我们将数据集 II 导入到 Eviews 软件中，进行初步的数据分布分析。

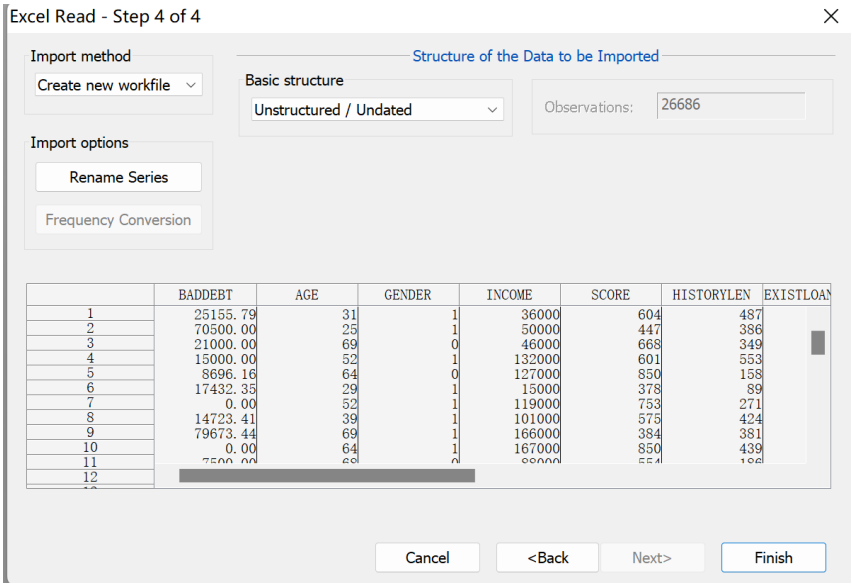

图 5: 导入数据步骤

本数据集的全部变量的解释如表4所示：

表 4: 数据集 II 变量组成描述

变量名	**解释**	**取值**
BadDebt	坏账金额	浮点数
Gend	借款人的性别	虚拟变量，男性为 1，女性则为 0
Age	借款人年龄	整数
Income	借款人的收入	整数
Score	信用评分*	整数
HistoryLen	距首次登记时长（单位：月）	整数
OldCustomer	是否是本机构老用户	虚拟变量，是为 1，否则为 0
ExistLoan	名下共有贷款数**	整数
Asset	个人拥有的资产	整数
Amount	申请金额	整数
T	偿还时间（单位：月）	整数
Job	是否拥有工作	虚拟变量，是为 1，否则为 0

* 根据申请人的信用记录量化得出的信用分数，**分数量化标准与贷款人情况不相干，因此不考虑其与如年龄等个人因素存在多重共线性，但可能和名下贷款数等财务情况相关，需检验**，范围在 300 至 850，由信贷机构给出。

** 名下共有贷款数包括在其他机构的还款中的贷款数量，不包括已经还清的贷款。

如图6导入结果所示，数据集共含有 26686 个有效数据。图7显示了各个字段的分布情况。从数据分布来看，年龄、信用评分和名下贷款字段的分布相对均衡，这说明该机构的信用评分体系具有一定的鉴别能力，能够通过不同的借款人信息，较好地区分出信用优良者和潜在风险承担者。另一方面，还款期数和借款金额的分布相对集中，且主要集中在较大金额和较长期限的贷款上。这表明在贷款金额较高和借款周期较长的情况下，违约风险可能更高。这是因为，对于借款人来说，大金额和长期限的贷款意味着更高的还款压力，可能导致还款困难，从而增加违约风险。

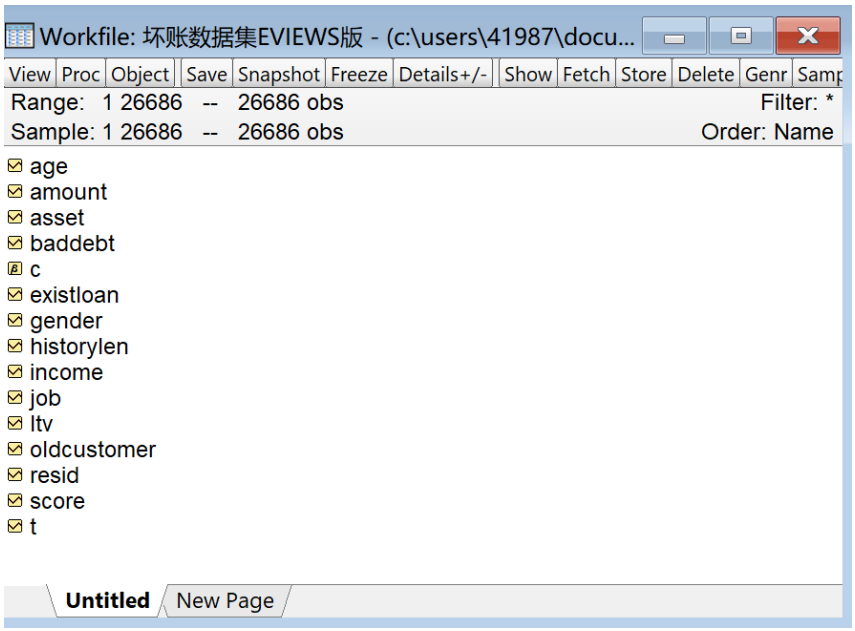

图 6: 导入数据结果

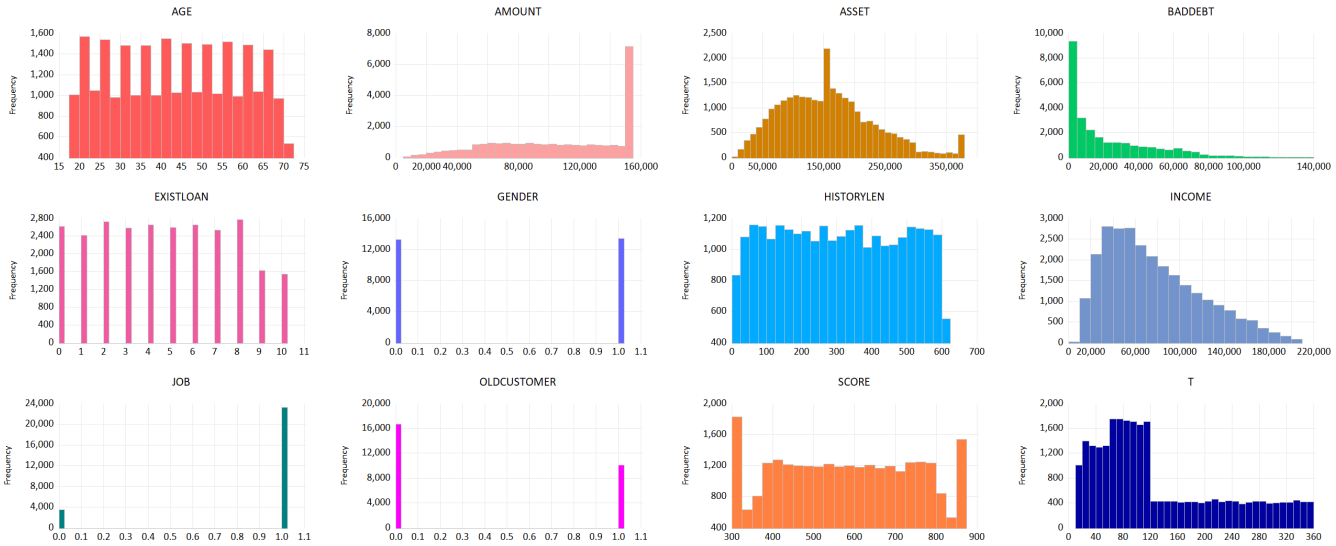

图 7: 数据集 II: 字段数据分布

在下页表5中展示了数据集 II 中整数数据样本特征。例如，我们可以看到，数据集中平均年龄为 43 岁，与数据集 I 基本一致；信用分均值为 581 左右，位于分布中心；资产情况分布广泛，标准差达到 76390.42，数值较大。各个数据的均值、方差、标准差等均在正常范围以内，认为导入的数据正确无误。

表 5: 数据集 II：整数数据样本特征

	AGE	AMOUNT	ASSET	EXISTLOAN	HISTORYLEN	INCOME	SCORE	T
Mean	43.88361	105421.1	157018.3	4.680469	307.4901	76062.32	581.5645	132.8584
Median	44.00000	110556.5	155200.4	5.000000	306.0000	67000.00	582.0000	100.0000
Maximum	70.00000	150000.0	375000.0	10.00000	611.0000	209000.0	850.0000	359.0000
Minimum	18.00000	5985.000	6878.159	0.000000	6.000000	9000.000	300.0000	12.00000
Std. Dev.	15.30358	40447.73	76390.42	3.004889	175.7420	42881.25	164.3482	95.96814
Skewness	0.004908	-0.422880	0.584653	0.064732	0.017156	0.723864	-0.025944	0.849593
Kurtosis	1.801662	1.926964	3.150689	1.873832	1.790392	2.788630	1.855095	2.521296
Jarque-Bera	1596.834	2075.633	1545.549	1428.831	1628.212	2380.159	1460.503	3465.169
Probability	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Sum	1171078.	2.81E+09	4.19E+09	124903.0	8205681.	2.03E+09	15519631	3545459.
Sum Sq. Dev.	6249616.	4.37E+13	1.56E+14	240948.4	8.24E+08	4.91E+13	7.21E+08	2.46E+08
Observations	26686	26686	26686	26686	26686	26686	26686	26686

表 6: 数据集 II：整数数据样本相关系数表

	AGE	**AMOUNT**	**ASSET**	**EXISTLOAN**	**HISTORYLEN**	**INCOME**	**SCORE**	**T**
AGE	1.000000	0.271171	0.238275	0.137138	0.003965	0.625455	0.137470	0.073296
AMOUNT	0.271171	1.000000	**0.801459**	0.086326	0.006340	0.390543	0.086059	0.049266
ASSET	0.238275	**0.801459**	1.000000	0.297150	0.007703	0.351110	0.296299	0.192803
EXISTLOAN	0.137138	0.086326	0.297150	1.000000	0.001809	0.223387	**0.994621**	0.645556
HISTORYLEN	0.003965	0.006340	0.007703	0.001809	1.000000	0.008078	0.002491	-0.007034
INCOME	0.625455	0.390543	0.351110	0.223387	0.008078	1.000000	0.223245	0.109349
SCORE	0.137470	0.086059	0.296299	**0.994621**	0.002491	0.223245	1.000000	0.649629
T	0.073296	0.049266	0.192803	0.645556	-0.007034	0.109349	0.649629	1.000000

* 受限与版面原因，无法展示各虚拟变量与其他变量的相关系数，但是他们都是小于 0.8 的

在上页表6展示了数据集 II 中整数数据之间的的相关系数，可以看到，存在两个变量的相关系数较高。

首先，在分析数据时，我们发现 ExisitLoan 和 Score 两个变量之间存在高度相关性。具体来说，这两个变量之间的相关系数高达 0.99。在统计学中，相关系数的取值范围在 -1 到 1 之间，值越接近 1，表示两个变量之间的正相关性越强。在这种情况下，我们可以认为评分者在对借款人进行评分时，很大程度上参考了借款人名下的贷款数量。因此，在后续变量选取时，只能选择其中之一。

其次，对于 Asset 和 Amount 两个变量，其相关系数达 0.801459，也表明其存在一定的相关性。这是因为借款人的资产状况也可能影响其借款需求。资产总额较大的借款人可能更有动力借款来进行投资、扩大生产等，从而使得资产总额与贷款金额呈正相关。而资产总额较小的借款人，借款需求可能相对较低，贷款金额也较小。

因此，在后续变量选取时，也只能选择其中之一，或者根据相关文献所述，通过 ltv 值 (债务资产比) 进行变量代换。该变量代表个人债务与资产之间的比值，与模型 I 的 ita 相同，均旨在衡量个体所承受的还款压力。较低的债务资产比意味着借款人承担的还款压力较小，坏账风险较低；相反，较高的债务资产比则表示借款人承受较大的还款压力，坏账风险较高。

由于版面原因，我们没有办法展示虚拟变量之间的相关系数和数据描述，但是经过我们的实操，发现各个虚拟变量与其他变量的相关系数均小于 0.8。

3 模型建立与诊断

本节主要讲述了还款能力和坏账估计两个模型的具体建立流程。对于每个模型，我们会从变量选取触发，进行初步建立，随后诊断模型中存在的各类可能问题（包括多重共线性、异方差、遗漏变量等），进行调整修正，得出最后的回归结果。

3.1 模型 I: 还款能力模型

3.1.1 变量预期

首先,我们对各个变量的预期符号提出假设预期。预期结果与解释如表7所示。注意，被解释变量 TARGET 取值为 1 代表存在还款困难的情况。

表 7: 模型 I 各变量符号预期

变量名	预期符号	解释
age	−	随着年龄的增长，个人的收入稳定性降低，坏账比例增加。
annamount	−	每年需要还款金额越多,那么还款压力越大，越容易逾期
income	+	收入越高，那么还款压力越小，越不容易逾期
ita	−	收入债务比，值越高，债务越少，收入越高，还款压力小，不容易逾期
gend	?	无法预期
family	+	家庭人口数量越多,照顾家庭开支可能更大，还款压力大，更可能发生逾期
children	+	子女数量越多，照顾子女的开支可能更大，还款压力大，更可能发生逾期
接下一页		

续表：

变量名	预期符号	解释
graduate	?	无法预期，一方面受过高等教育的对自己的财务可能有严格规划，其法律意识、信用意识等更加清晰；另一方面，受到高等教育可能在求学过程中承受较大的经济压力，为完成学业而借款，并且高学历也不一定能获得高收入（如文科类专业）。
marry	?	无法预期，一方面婚后需要照顾家庭，还款压力高，容易逾期；另一方面，结婚后的人群相比未结婚的人群对消费更具有谨慎性和理性，不会毫无计划的乱花钱。
car	−	拥有汽车证明其财务水平较好。
house	−	拥有公寓证明其财务水平较好。
population	−	人口越多，意味着整体经济发展水平较高，居民收入水平相对较高。较高的收入水平有利于居民按时还款，降低逾期风险。
rich	−	整体经济发展水平较高，居民收入水平相对较高。较高的收入水平有利于居民按时还款，降低逾期风险。
poverty	+	整体经济发展水平低，居民收入水平相对不高。不利于居民按时还款。

鉴于本研究的自变量为虚拟变量（最终是否成功还款），为了更为精确地预测和解析这一结果，我们将采用 Logit 回归模型进行进一步的实证分析。

3.1.2 初步回归

首先，为了充分利用数据尽量纳入全部变量作为解释变量。对于具有多重线性的变量，我们选择其中之一纳入（如 family 和 chilren，ita 和 annamount 和

income), 在 Eviews 中输入指令 `LS TARGET C AGE ANNAMOUNT CAR CHILDREN GEND GRADUATE HOUSE INCOME MARRY POPULATION RICH POVERTY` 进行第一次回归。

由于本模型为含有虚拟应变量的模型，因此还需将模型调整为 Logit 模型，调整步骤如图8所示：

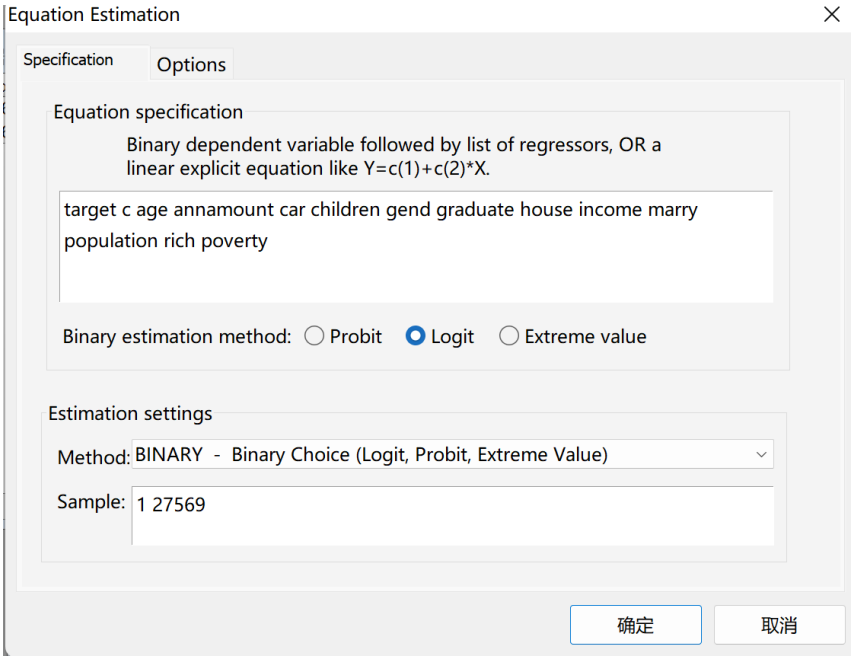

图 8: logit 模型步骤

回归的结果如下一页图9所示。

回归结果为：

$$\begin{aligned} L: Pr(\widehat{TARGET}_i = 1) = &-1.25 - 0.03 \times AGE + 4.70 \times 10^{-6} \times ANNAMOUNT \\ &- 0.29 \times CAR + 0.06 \times CHILDREN \\ &+ 0.42 \times GEND - 0.63 \times GRADUATE \\ &+ 0.04 \times HOUSE - 5.41 \times 10^{-7} \times INCOME \\ &- 0.19 \times MARRY + 1.23 \times POPULATION \\ &- 0.63 \times RICH + 0.39 \times POVERTY \end{aligned} \tag{1}$$

Equation: UNTITLED Workfile: 还款能力数据集::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: TARGET
Method: ML - Binary Logit (Newton-Raphson / Marquardt steps)
Date: 11/11/23 Time: 14:49
Sample: 1 27569
Included observations: 27569
Convergence achieved after 5 iterations
Coefficient covariance computed using observed Hessian

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-1.252612	0.118740	-10.54922	0.0000
AGE	-0.026563	0.002061	-12.88603	0.0000
ANNAMOUNT	4.70E-06	1.78E-06	2.640889	0.0083
CAR	-0.292832	0.052721	-5.554391	0.0000
CHILDREN	0.060870	0.031024	1.962026	0.0498
GEND	0.415727	0.049804	8.347196	0.0000
GRADUATE	-0.630190	0.062786	-10.03708	0.0000
HOUSE	0.041535	0.048777	0.851530	0.3945
INCOME	-5.42E-07	3.16E-07	-1.714471	0.0864
MARRY	-0.193973	0.052077	-3.724727	0.0002
POPULATION	1.233803	2.291633	0.538395	0.5903
RICH	-0.629109	0.117216	-5.367113	0.0000
POVERTY	0.391314	0.056710	6.900282	0.0000

McFadden R-squared	0.036085	Mean dependent var	0.079800
S.D. dependent var	0.270988	S.E. of regression	0.268119
Akaike info criterion	0.537420	Sum squared resid	1980.941
Schwarz criterion	0.541298	Log likelihood	-7395.065
Hannan-Quinn criter.	0.538669	Deviance	14790.13
Restr. deviance	15343.81	Restr. log likeli...	-7671.904
LR statistic	553.6764	Avg. log likelihood	-0.268238
Prob(LR statistic)	0.000000		

Obs with Dep=0	25369	Total obs	27569
Obs with Dep=1	2200		

图 9: 模型 I 初步回归结果

聚焦单独每个变量来看————

首先关注变量 house 和 population。从预期符号来看，这两个变量的预期符号与实际结果不一致。同时，从显著性水平来看，这两个变量的 z 统计量较低，是所有变量中唯一两个不具备显著性的变量。因此，我们有理由怀疑这两个变量与目标变量之间的关联性不强，可以考虑将它们视为不相干变量。在

下一部分将做具体的检验。

接下来，观察变量 annamount 和 income。虽然这两个变量在模型中具有显著性，但它们的系数较小，对目标变量的影响微乎其微。根据文献资料，更合理的方程设定应该是以收入债务比 ita 代替 annamount 和 income。

从模型整体来看来看————

模型整体拟合优度 McFadden R-squared（即 $\overline{R^2_{MCF}}$）为 0.036085，拟合程度一般，表明模型中选取的变量对模型有一定的拟合作用，这个模型可以解释目标变量变异的 3.6%；模型整体显著性水平 LR statistic 为 553.6764，数值远高于临界值，因此，认为模型整体具备显著性。说明模型整体拟合效果较好。

3.1.3 问题诊断

通过刚刚的分析，我们认为 house 和 population 这两个变量与目标变量的关联性较弱，可以考虑在构建模型时将其排除。而 annamount 和 income 的系数较小，用收入债务比 ita 代替它们能够更好地反映变量间的关联。下面具体来探究这两个优化的正确性与科学性。

A. 不相干变量：house 和 population

下面，依次检验 house 和 population 是否为冗余变量。在 Eviews 软件中，我们按照顺序选择 Coefficient Diagnostics -> Redundant Variables Test

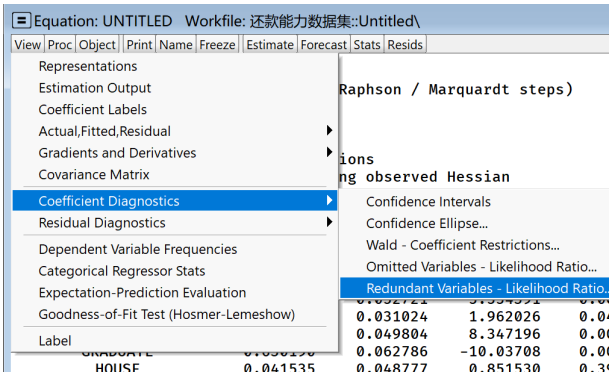

图 10: 冗余检查步骤

建立假设

H_0 :变量$house$和$population$是冗余变量

H_A :H_0不成立

在输入框中输入 house 和 population, 进行检验

检验结果如图11所示:

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Redundant Variables Test
Equation: UNTITLED
Redundant variables: HOUSE POPULATION
Specification: TARGET C AGE ANNAMOUNT CAR CHILDREN GEND GRADUATE HOUSE INCOME MARRY POPULATION RICH POVERTY
Null hypothesis: HOUSE POPULATION are jointly insignificant

	Value	df	Probability
Likelihood ratio	1.032284	2	0.5968

LR test summary:

	Value
Restricted LogL	-7395.581
Unrestricted LogL	-7395.065

Restricted Test Equation:
Dependent Variable: TARGET
Method: ML - Binary Logit (Newton-Raphson / Marquardt steps)
Date: 11/11/23 Time: 15:36
Sample: 1 27569
Included observations: 27569
Convergence achieved after 7 iterations
Coefficient covariance computed using observed Hessian

Variable	Coeffic...	Std. Error	z-Statistic	Prob.
C	-1.212457	0.109577	-11.06492	0.0000
AGE	-0.026326	0.002046	-12.86650	0.0000
ANNAMOUNT	4.69E-06	1.78E-06	2.632826	0.0085
CAR	-0.291011	0.052685	-5.523610	0.0000
CHILDREN	0.061379	0.030995	1.980291	0.0477
GEND	0.414651	0.049788	8.328357	0.0000
GRADUATE	-0.630715	0.062783	-10.04598	0.0000
INCOME	-5.38E-07	3.16E-07	-1.703184	0.0885
MARRY	-0.192632	0.052060	-3.700200	0.0002
RICH	-0.596787	0.099109	-6.021521	0.0000
POVERTY	0.384866	0.055627	6.918695	0.0000

McFadden R-squared	0.036017	Mean dependent var	0.079800
S.D. dependent var	0.270988	S.E. of regression	0.268113
Akaike info criterion	0.537312	Sum squared resid	1981.001
Schwarz criterion	0.540594	Log likelihood	-7395.581
Hannan-Quinn criter.	0.538369	Deviance	14791.16
Restr. deviance	15343.81	Restr. log likel...	-7671.904
LR statistic	552.6442	Avg. log likelihood	-0.268257
Prob(LR statistic)	0.000000		

Obs with Dep=0	25369	Total obs	27569
Obs with Dep=1	2200		

图 11: 模型 I 冗余变量检查结果

根据检验结果，我们可以发现，似然比低于临界值，p 值远高于显著性水平 5%，因此不能拒绝原假设 H_0，**所以可以认为 house 和 population 是一个不相干变量**。应该将其剔除。

从理论再次分析，首先，house 变量原本应作为借款人财务状况的衡量标准，然而在实际操作中，鉴于年轻人在购房方面的特殊处境，这一变量并未充分体现其财务实力。购房对于年轻人来说通常是必需的，而这方面的资金往往来源于父母、亲戚、朋友等资助。这样一来，拥房与否并不能直接反映一个人的财务状况。相反，汽车作为非必需消费品，一般情况下是在购房之后，年轻人在经济允许的情况下自费购买，受他人资助较少。因此，拥有汽车能够更好地体现出一个人的经济实力。

根据 [13] 的研究，这一观点得到了证实。尽管 house 这个变量看似合理，但实际上却没有解释效力。综上所述，我们在分析借款人财务状况时，应当关注其是否拥有汽车，而不仅仅是房子。

其次，population 这一因素用于反映借款人所处城市收入水平并不恰当。因为人口数量与城市财富水平之间并非完全正相关，人口众多并不代表人均收入较高。根据文献 [14] 的研究，对于大部分中等发达地区和一部分欠发达地区，人口数量与 GDP 确实存在正相关关系。然而，在大多数发达地区和部分欠发达地区，人口增长率与 GDP 增长率呈现负相关。因此，人口并非是衡量城市收入水平、物价等经济因素的理想指标。

B. 变量设定：ita 代替 annamount 和 income

我们首先检测 ita 是否是遗漏变量。

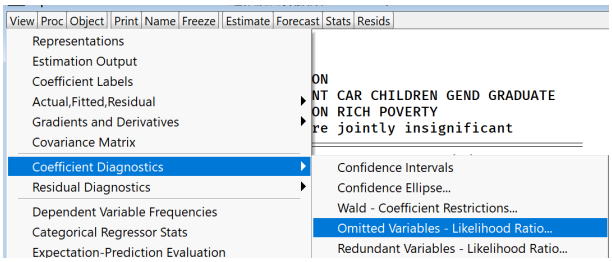

图 12: 遗漏变量检测步骤

在 Eviews 软件中，我们按照顺序选择 Coefficient Diagnostics -> Omitted

Variables Test 菜单。建立假设

$$H_0 \text{:变量}ita\text{不是遗漏变量}$$

$$H_A \text{:}H_0\text{不成立}$$

检验结果如图13所示：

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Omitted Variable Test
Equation: UNTITLED
Omitted Variables: ITA
Specification: TARGET C AGE ANNAMOUNT CAR CHILDREN GEND GRADUATE HOUSE INCOME MARRY POPULATION RICH POVERTY
Null hypothesis: ITA is not significant

	Value	df	Probability
Likelihood ratio	12.60713	1	0.0004

LR test summary:

	Value
Restricted LogL	-7395.065
Unrestricted LogL	-7388.762

Unrestricted Test Equation:
Dependent Variable: TARGET
Method: ML - Binary Logit (Newton-Raphson / Marquardt steps)
Date: 11/11/23 Time: 16:08
Sample: 1 27569
Included observations: 27569
Convergence achieved after 7 iterations
Coefficient covariance computed using observed Hessian

Variable	Coeffic...	Std. Error	z-Statistic	Prob.
C	-0.988890	0.142219	-6.953274	0.0000
AGE	-0.027035	0.002068	-13.07339	0.0000
ANNAMOUNT	-3.37E-06	2.96E-06	-1.140130	0.2542
CAR	-0.289741	0.052723	-5.495484	0.0000
CHILDREN	0.058777	0.031019	1.894847	0.0581
GEND	0.420132	0.049796	8.437067	0.0000
GRADUATE	-0.624594	0.062830	-9.941002	0.0000
HOUSE	0.048909	0.048839	1.001422	0.3166
INCOME	6.54E-07	4.43E-07	1.476010	0.1399
MARRY	-0.195436	0.052071	-3.753262	0.0002
POPULATION	1.160878	2.292969	0.506277	0.6127
RICH	-0.622293	0.117299	-5.305190	0.0000
POVERTY	0.389009	0.056729	6.857385	0.0000
ITA	-0.031894	0.009331	-3.418179	0.0006

McFadden R-squared	0.036906	Mean dependent var	0.079800
S.D. dependent var	0.270988	S.E. of regression	0.267992
Akaike info criterion	0.537035	Sum squared resid	1978.992
Schwarz criterion	0.541212	Log likelihood	-7388.762
Hannan-Quinn criter.	0.538381	Deviance	14777.52
Restr. deviance	15343.81	Restr. log likel...	-7671.904
LR statistic	566.2836	Avg. log likelihood	-0.268010
Prob(LR statistic)	0.000000		

图 13: 遗漏变量检测结果

根据检验结果，我们可以发现，p 值 < 显著性水平 5%，所以**可以认为**

ita 是一个遗漏变量。

ita 这个变量具有深刻的理论意义，它通过衡量人们的收入与还款金额之间的比例关系，揭示了一种衡量借款人还款压力的有效方法。这一点已经在第一部分的理论中阐述，不再此处赘述。这个变量在一定程度上反映了借款人的经济状况和对债务的承担能力，从而为贷款评估有力的支持。因此，将其纳入模型是必然的选择。

当然，由于 ita 是由 income、annamount 计算而来，因此 ita 和 income、annamount 存在一定的多重共线性。基于此，**我们选择用 ita 代替 annamount 和 income**。

3.1.4 模型修正

下面，根据我们的诊断结果，我们对模型 I 进行进一步的调整和修正，使得其拟合效果更加、经济理论意义更好。

在 Eviews 中输入指令 LS TARGET C AGE CAR CHILDREN GEND GRADUATE MARRY RICH POVERTY ITA 进行修正回归。整个修正中，我们删掉了 house 和 population 两个冗余变量，同时用 ita 代替 annamount 和 income。由于本模型为含有虚拟应变量的模型，因此还需将模型调整为 Logit 模型，回归结果如下一页图14所示。

模型回归结果为————

$$\begin{aligned} L: Pr(\widehat{TARGET}_i = 1) = &- 0.70 - 0.01 \times AGE - 0.14 \times CAR \\ &+ 0.03 \times CHILDREN + 0.22 \times GEND \\ &- 0.30 \times GRADUATE - 0.10 \times MARRY - 0.28 \times RICH \\ &+ 0.19 \times POVERTY - 0.01 \times ITA \end{aligned} \tag{2}$$

Equation: UNTITLED Workfile: 还款能力数据集::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: TARGET
Method: ML - Binary Probit (Newton-Raphson / Marquardt steps)
Date: 11/11/23 Time: 16:32
Sample: 1 27569
Included observations: 27569
Convergence achieved after 6 iterations
Coefficient covariance computed using observed Hessian

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-0.699797	0.055825	-12.53561	0.0000
AGE	-0.013186	0.001018	-12.95846	0.0000
CAR	-0.143283	0.026003	-5.510223	0.0000
CHILDREN	0.027331	0.015814	1.728237	0.0839
GEND	0.215774	0.024896	8.667142	0.0000
GRADUATE	-0.295346	0.029184	-10.12011	0.0000
MARRY	-0.098638	0.026054	-3.785877	0.0002
RICH	-0.275048	0.044400	-6.194805	0.0000
POVERTY	0.194673	0.028874	6.742056	0.0000
ITA	-0.009397	0.002497	-3.763865	0.0002

McFadden R-squared	0.036315	Mean dependent var	0.079800
S.D. dependent var	0.270988	S.E. of regression	0.268030
Akaike info criterion	0.537074	Sum squared resid	1979.839
Schwarz criterion	0.540057	Log likelihood	-7393.298
Hannan-Quinn criter.	0.538035	Deviance	14786.60
Restr. deviance	15343.81	Restr. log likeli...	-7671.904
LR statistic	557.2101	Avg. log likelihood	-0.268174
Prob(LR statistic)	0.000000		

Obs with Dep=0	25369	Total obs	27569
Obs with Dep=1	2200		

图 14: 修正后的回归结果

从模型设定的四个角度来看：

(a) 理论： 去除冗余变量 house 和 population、用 ita 代替 annamount 和 income 这两个操作均具有实际的理论意义。ita 可以直观的反应借款者的还款压力，从而更好的衡量借款者的还款能力。具备经济理论基础。

(b) 显著性： 在 5% 的显著性条件下，所有变量其余的变量均变得完全显著。(因为 Eviews 是双侧检验，对于 children 变量，$p/2 < 5\%$ 仍然成立)。单个变量的显著性上提升较大；同时，LR 值也由原来的 553 提升到现在的 557，整体显著性上也有一定的提升。

(c) 拟合优度: 整体拟合优度由原来的 0.036085 提升到 0.036315，有略微的提升，模型的拟合效果更好了。

(d) 偏误: 对比前后两个模型，各个系数的偏误不大。

综上，修复后的模型解决了我们刚刚检查出的两个问题，并在显著性、拟合优度上有一定程度提升，更具备理论意义。

模型的具体经济意义我们将在第 4 节详细论述。

3.2 模型 II: 坏账金额模型

3.2.1 变量预期

首先,我们对各个变量的预期符号提出假设预期。预期结果与解释如表8所示。

表 8: 模型 II 各变量符号预期

变量名	预期符号	解释
age	+	随着年龄的增长，个人的收入稳定性降低，坏账金额增加。
annamount	+	每年需要还款金额越多,那么还款压力越大，坏账金额增加。
asset	−	资产数量越多，可用于抵押偿还贷款的金额越多，可能的坏账金额越少
ltv	+	较低的债务资产比意味着借款人承担的还款压力较小，坏账风险较低；相反，较高的债务资产比则表示借款人承受较大的还款压力，坏账风险较高
income	−	收入越高，那么还款压力越小，不容易造成欠款，坏账金额低
gender	?	无法预期

接下一页

续表：

变量名	预期符号	解释
score	−	信用分越低，其偿还贷款的可能性越低，拥有坏账的可能性越高
t	?	无法预期，一方面，还款周期越长，借款人有更多的时间去周转，每月还款压力小；同时另一方面，时间越长，不确定性越多，风险可能越高。
existloan	+	现有贷款数越多，需要偿还的数量越多，还款压力越大，坏账比例越高
job	−	拥有稳定工作的人比没有稳定工作的人收入更稳定
historylen	+	信用记录长，信用越可靠，那么坏账的风险越小
oldcustomer	?	无法预期，一方面，老顾客具有贷款经验，且在过去的交易中建立了良好的信用记录；另一方面，信贷机构可能会对他们产生过度信任，导致在贷款审批过程中放松警惕。

这是一个典型的可以使用最小二乘法的普通模型，下面，我们对模型进行初步的建立。

3.2.2 初步回归

在初次建立模型时，因为遗漏变量往往比不相干变量更严重，因此我们尽可能考虑更多的解释变量，以防遗漏变量带来的一些风险。由于受到多重共线性的影响，我们去掉相关系数最为严重的一组中的 existloan，保留了 score。因为相比前者，score 更能全面反应借款者的用户个人情况。

在 Eviews 中，我们输入指令 LS baddebt c age asset amount score oldcustomer t gender income job historylen 开始第一次回归常识

回归的结果如图15所示。

Equation: UNTITLED Workfile: 坏账数据集EVIEWS版::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: BADDEBT
Method: Least Squares
Date: 11/11/23 Time: 20:42
Sample: 1 26686
Included observations: 26686

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	68230.27	630.2585	108.2576	0.0000
AGE	3.908504	7.572885	0.516118	0.6058
ASSET	-0.150456	0.002170	-69.33077	0.0000
AMOUNT	0.444790	0.004033	110.2795	0.0000
SCORE	-111.2397	1.069164	-104.0436	0.0000
OLDCUSTOMER	9558.468	351.5986	27.18574	0.0000
T	-9.027165	1.265254	-7.134668	0.0000
GENDER	-17.07695	180.7215	-0.094493	0.9247
INCOME	0.000106	0.002886	0.036783	0.9707
JOB	-8755.957	271.1848	-32.28778	0.0000
HISTORYLEN	0.159827	0.514197	0.310829	0.7559

R-squared	0.656806	Mean dependent var	21808.91
Adjusted R-squared	0.656678	S.D. dependent var	25188.62
S.E. of regression	14758.95	Akaike info crite...	22.03750
Sum squared resid	5.81E+12	Schwarz criterion	22.04088
Log likelihood	-294035.3	Hannan-Quinn criter.	22.03859
F-statistic	5105.077	Durbin-Watson stat	1.973625
Prob(F-statistic)	0.000000		

图 15: 模型 II 初步回归结果

回归结果为:

$$\begin{aligned} BADDEBT = {} & 68230.27 + 3.91 \times AGE - 0.15 \times ASSET + 0.44 \times AMOUNT \\ & - 111.24 \times SCORE + 9558.47 \times OLDCUSTOMER \\ & - 9.03 \times T - 17.08 \times GENDER + 0.000106 \times INCOME \\ & - 8755.96 \times JOB + 0.16 \times HISTORYLEN \end{aligned} \tag{3}$$

下面对回归结果做简要分析。

聚焦单独每个变量来看————

从符号来看，income 变量出现了和预期相背离的情况，这可能意味着我们在构建模型时存在一些问题。从显著性水平来看，gender、income、historylen 这三个变量，其 t 值均小于临界值，无法拒绝原假设，这说明我们在当前的模型中并未观察到显著性水平，这可能表明这些变量在预测结果中的作用并不明显，或者我们的模型尚未足够完善。

从模型整体来看来看————

模型整体拟合优度调整的 R 平方（即 $\overline{R^2}$）为 0.656806，拟合程度较好，表明模型中选取的变量对模型有不错的拟合作用，这个模型可以解释目标变量变异的 65.6%；模型整体显著性水平 F 值为 5105.077，数值远高于临界值，因此，认为模型整体具备显著性。说明模型整体拟合效果较好。

模型的整体显著性水平很高，但出现了 3 个不显著变量与 1 个预期符号相反的变量，因此，有必要对模型进行诊断。

3.2.3 问题诊断

A. 序列相关性

我们对数据采用 LM 检验法。首先建立假设：

$$H_0 \text{:模型存在序列相关性}$$

$$H_A \text{:}H_0\text{不成立}$$

在线性回归的结果中，依次点击 view->Residual Diagnostcis->Serial Correlation LM Test。进行 LM 检验。我们设定检验的阶数为 2。

最终，得到如下页图16的检验结果。

从结果可以看到，nR^2 的结果为 4.701187，p 值为 0.0953，其统计量小于卡方临界值且 p 值 >5%，因此不能原假设，因此，可以得出结论————**认为该模型并不存在序列相关性**。

Equation: UNTITLED Workfile: 坏账数据集EVIEWS版::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Breusch-Godfrey Serial Correlation LM Test:
Null hypothesis: No serial correlation at up to 2 lags

F-statistic	2.349863	Prob. F(2,26673)	0.0954
Obs*R-squared	4.701187	Prob. Chi-Square(2)	0.0953

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Date: 11/11/23 Time: 20:56
Sample: 1 26686
Included observations: 26686
Presample missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-11.34608	630.2490	-0.018003	0.9856
AGE	0.176673	7.572953	0.023330	0.9814
ASSET	-1.88E-05	0.002170	-0.008684	0.9931
AMOUNT	5.90E-05	0.004033	0.014619	0.9883
SCORE	0.009506	1.069122	0.008891	0.9929
OLDCUSTOMER	-1.795504	351.5818	-0.005107	0.9959
T	-0.009304	1.265232	-0.007354	0.9941
GENDER	0.832444	180.7180	0.004606	0.9963
INCOME	-8.05E-05	0.002886	-0.027895	0.9777
JOB	-0.414777	271.1722	-0.001530	0.9988
HISTORYLEN	0.009082	0.514191	0.017662	0.9859
RESID(-1)	0.013156	0.006124	2.148262	0.0317
RESID(-2)	0.001609	0.006124	0.262725	0.7928

R-squared	0.000176	Mean dependent var	2.23E-12
Adjusted R-squared	-0.000274	S.D. dependent var	14756.18
S.E. of regression	14758.20	Akaike info crite...	22.03747
Sum squared resid	5.81E+12	Schwarz criterion	22.04146
Log likelihood	-294033.0	Hannan-Quinn criter.	22.03876
F-statistic	0.391644	Durbin-Watson stat	1.999976
Prob(F-statistic)	0.967251		

图 16: 序列相关检验

B. 多重共线性

方程可能存在序列相关性，因为在第二节的相关系数分析中，我们发现 amount 和 asset 的相关系数较高。下面对其使用 VIF 检验。

在 Eviews 软件中，我们选择 View->Coefficient Diagnostics->Variance Inflation Factors 进行方差膨胀因子分析。

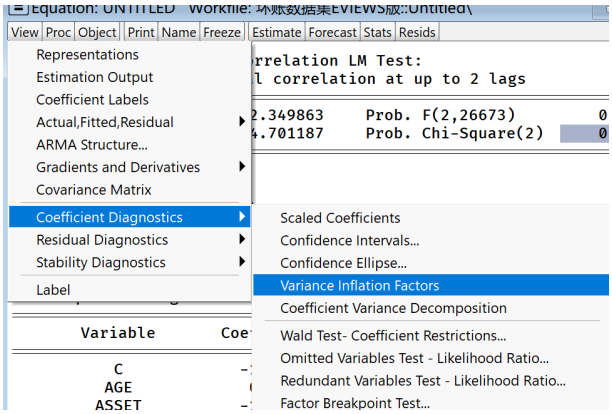

图 17: 方差膨胀因子查看步骤

分析结果如图18所示。

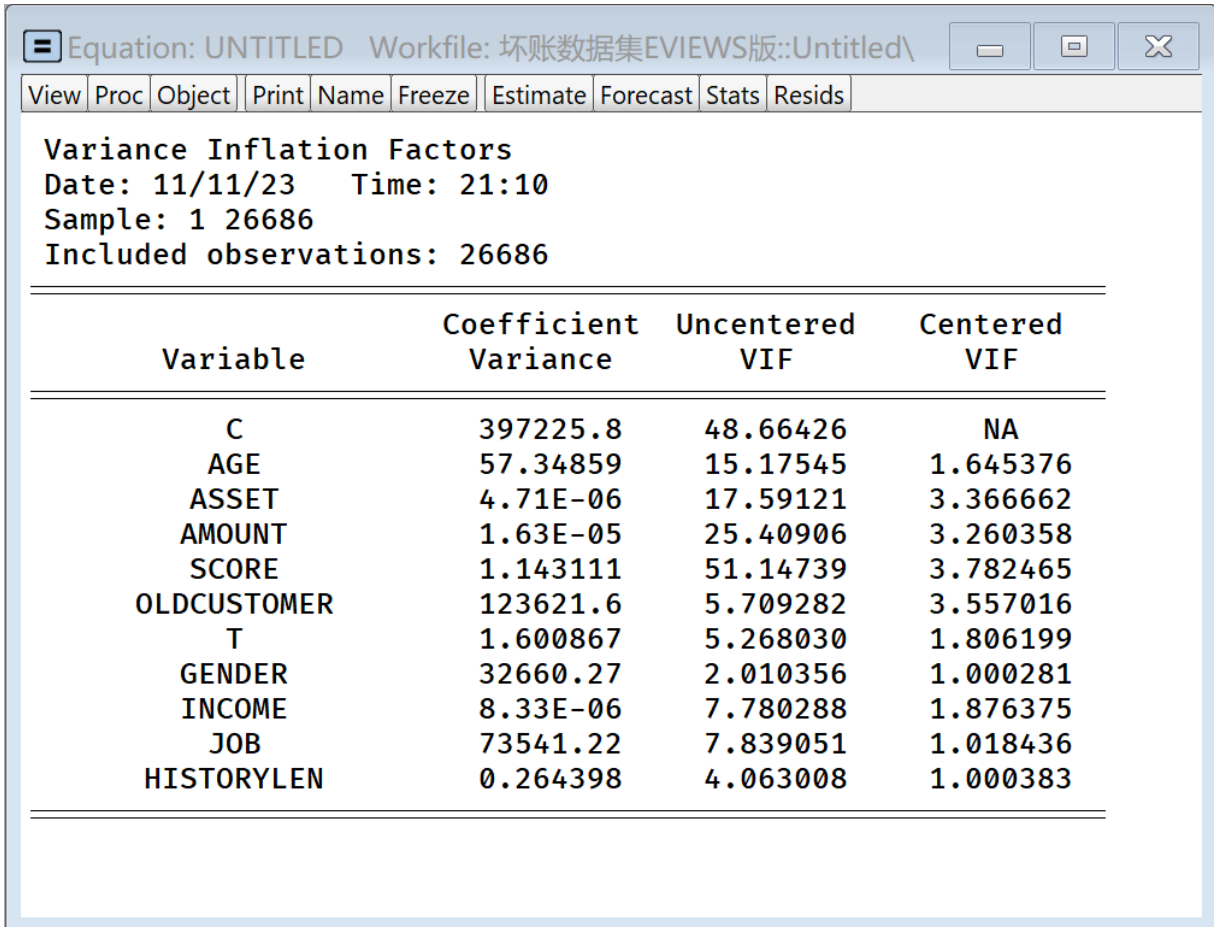

图 18: 方差膨胀因子检验结果

可以看到，所有系数的 VIF 值均小于 5。一般而言，只有 VIF（方差膨胀

因子）大于五才认为存在明显的多重线性。综上，可以认为**该模型不存在多重共线性**。

C. 不相干变量

下面，依次检验 gender、income、historylen 是否为冗余变量。在 Eviews 软件中，我们按照顺序选择 Coefficient Diagnostics -> Redundant Variables Test。在输入框中输入 gender、income、historylen，进行检验

检验结果如图19所示：

Equation: UNTITLED Workfile: 坏账数据集EVIEWS版::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Redundant Variables Test
Equation: UNTITLED
Redundant variables: GENDER INCOME HISTORYLEN
Specification: BADDEBT C AGE ASSET AMOUNT SCORE OLDCUSTOMER T GENDER INCOME JOB HISTORYLEN
Null hypothesis: GENDER INCOME HISTORYLEN are jointly insignificant

	Value	df	Probability
F-statistic	0.035826	(3, 26675)	0.9909
Likelihood ratio	0.107521	3	0.9909

F-test summary:

	Sum of Sq.	df	Mean Squares
Test SSR	23411371	3	7803790.
Restricted SSR	5.81E+12	26678	2.18E+08
Unrestricted SSR	5.81E+12	26675	2.18E+08

LR test summary:

	Value
Restricted LogL	-294035.4
Unrestricted LogL	-294035.3

Restricted Test Equation:
Dependent Variable: BADDEBT
Method: Least Squares
Date: 11/11/23 Time: 21:20
Sample: 1 26686
Included observations: 26686

Variable	Coeffic...	Std. Error	t-Statistic	Prob.
C	68264.66	587.3098	116.2328	0.0000
AGE	4.068581	6.191599	0.657113	0.5111
ASSET	-0.150454	0.002170	-69.33407	0.0000
AMOUNT	0.444818	0.003972	111.9805	0.0000
SCORE	-111.2340	1.058880	-105.0487	0.0000
OLDCUSTOMER	9558.514	351.2574	27.21228	0.0000
T	-9.033201	1.264316	-7.144733	0.0000
JOB	-8754.311	269.9349	-32.43119	0.0000

R-squared	0.656805	Mean dependent var	21808.91
Adjusted R-squared	0.656715	S.D. dependent var	25188.62
S.E. of regression	14758.15	Akaike info crit...	22.03728
Sum squared resid	5.81E+12	Schwarz criterion	22.03973
Log likelihood	-294035.4	Hannan-Quinn cri...	22.03807

图 19: 模型 II 冗余变量检查结果

根据检验结果，我们可以发现，F 统计量为 0.035826，低于临界值，p 值远高于显著性水平 5%，因此不能拒绝原假设 H_0，**所以可以认为 gender、income、historylen 三者均是不相干变量**。应该将其剔除。

从理论角度来看，首先，性别对贷款坏账的影响一直存在广泛的争议，没有明确的结论。性别差异在很大程度上可能是由其他变量所解释的，如收入、职业等。因此，从理论角度来看，性别作为一个独立变量对贷款坏账的预测能力有限，将其剔除可能更为合适。

其次，收入这个变量在理论上也是冗余的。这是因为收入的解释能力可能已经被其他更能反映借款人财务水平的变量所替代，如资产、信用评分等。这些变量能更好地预测借款人的还款能力，从而降低贷款坏账的风险。因此，在模型中包含收入这个变量可能导致信息冗余，影响模型的预测效果。

最后，信用记录（historylen）这个变量也是一个冗余变量。虽然理论上信用历史较长的借款人可能具有较高的还款能力，从而降低贷款坏账的风险，但这个变量仅仅简单地衡量了首次登记距今天的时间长短，并未反映出这期间的信用状况好坏。在其他变量如信用评级、收入等已经能够解释信用历史差异的情况下，这个变量可能显得冗余。

综上所述，从理论角度来看，gender、income、historylen 这三个变量都被确认为是冗余变量。需要进行剔除。

D. 异方差性

该方程可能存在异方差性，因为随着贷款金额的增加，还款金额也就随之越高，如果他的信用较低，那么坏账的总金额也越高，呈现一种规模效应。

我们首先通过图解法简单观察,在 Eviews 中输入 SCAT LOG(AMOUNT) RESID 绘制出贷款金额的对数与残差的散点分布图。绘图结果如下一页图20所示。

从图结果可以看到，随着贷款金额的增加，残差也逐步增大。整体呈现着一种上升趋势。综上，可以初步认为该模型存在异方差性

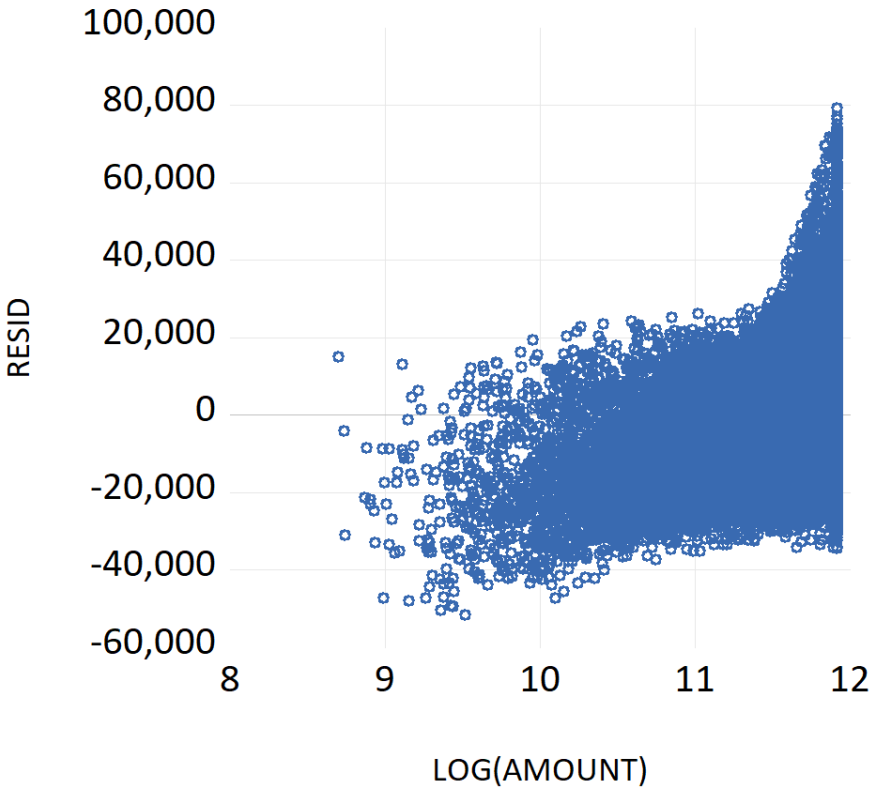

图 20: 贷款金额的对数与残差的散点分布图

下面，我们通过 white 检验法进行进一步检验。

在 Eviews 中，我们选择 View 菜单中的 Residual Diagnostics 的 Heteroskedasticity tests 选项，接着选择 White 检验。

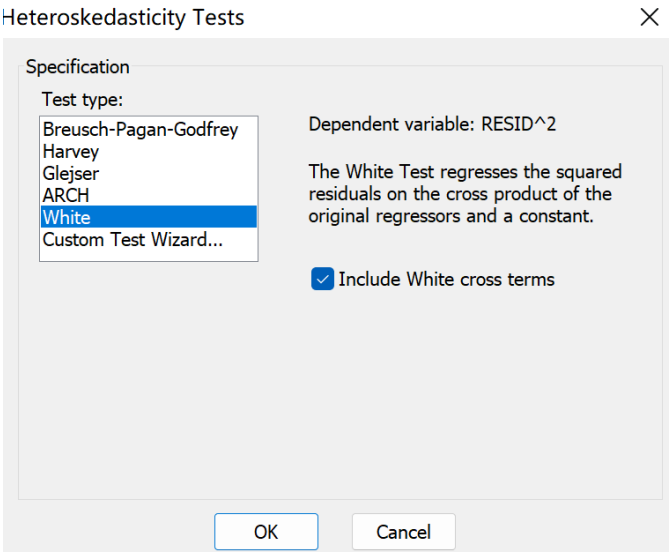

图 21: 怀特检验的步骤

怀特检验的结果如图22所示。

Equation: UNTITLED Workfile: 坏账数据集EVIEWS版::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Heteroskedasticity Test: White
Null hypothesis: Homoskedasticity

F-statistic	183.6358	Prob. F(62,26623)	0.0000
Obs*R-squared	7993.788	Prob. Chi-Square(62)	0.0000
Scaled explained SS	10485.29	Prob. Chi-Square(62)	0.0000

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 11/11/23 Time: 21:39
Sample: 1 26686
Included observations: 26686
Collinear test regressors dropped from specification

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.04E+09	71936353	14.45793	0.0000
AGE^2	1319.680	12205.44	0.108122	0.9139
AGE*ASSET	1.114439	3.745759	0.297520	0.7661
AGE*AMOUNT	-3.335467	6.855388	-0.486547	0.6266
AGE*SCORE	5330.652	1765.761	3.018898	0.0025
AGE*OLDCUSTOMER	-1557041.	584315.8	-2.664725	0.0077
AGE*T	-2121.063	2143.203	-0.989670	0.3223
AGE*GENDER	-163878.9	303847.2	-0.539346	0.5897
AGE*INCOME	0.009320	7.194938	0.001295	0.9990
AGE*JOB	716856.5	429615.8	1.668599	0.0952
AGE*HISTORYLEN	-274.7763	860.1222	-0.319462	0.7494
AGE	-2668497.	1284205.	-2.077937	0.0377
ASSET^2	0.016737	0.000944	17.72245	0.0000
ASSET*AMOUNT	-0.104234	0.003568	-29.21395	0.0000
ASSET*SCORE	19.81427	0.602381	32.89326	0.0000
ASSET*OLDCUSTOMER	-5175.357	202.8241	-25.51648	0.0000
ASSET*T	-2.054707	0.559613	-3.671654	0.0002
ASSET*GENDER	178.1943	87.01853	2.047774	0.0406
ASSET*INCOME	0.004618	0.001340	3.446322	0.0006
ASSET*JOB	1536.591	138.0644	11.12952	0.0000
ASSET*HISTORYLEN	0.430317	0.248348	1.732720	0.0832
ASSET	-5942.672	396.2539	-14.99713	0.0000
AMOUNT^2	0.157654	0.003776	41.74675	0.0000
AMOUNT*SCORE	-41.61816	1.019870	-40.80731	0.0000
AMOUNT*OLDCUSTOMER	9497.355	343.3804	27.65841	0.0000
AMOUNT*T	3.469512	1.106619	3.135238	0.0017

图 22: 怀特检验的结果

从结果可知，nR^2 的结果为 7993.788，p 值 >5%，因此拒绝原假设，因此，可以得出结论——**该模型存在异方差**。

3.2.4 模型修正

下面，根据我们的诊断结果，我们对模型 II 进行进一步的调整和修正。

首先，需要删除冗余变量 gender、income、historylen; 其次，对于异方差性，基于文献，我们首先尝试使用经济理论意义更强的变量 lvt 代替 asset、amount，来消除贷款金额的增加引发的规模效应导致的异方差。lvt 的经济理论意义已经在第一部分讲述。

在 Eviews 中输入指令 LS baddebt c age ltv score oldcustomer t income job 进行修正回归。回归结果为——

Equation: UNTITLED Workfile: 坏账数据集EVIEWS版::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: BADDEBT
Method: Least Squares
Date: 11/11/23 Time: 21:49
Sample: 1 26686
Included observations: 26686

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	53644.12	861.8871	62.24031	0.0000
AGE	27.86515	8.609320	3.236626	0.0012
LTV	42667.94	668.5530	63.82132	0.0000
SCORE	-110.2494	1.216573	-90.62290	0.0000
OLDCUSTOMER	9268.519	399.8927	23.17751	0.0000
T	-8.156161	1.439437	-5.666216	0.0000
INCOME	0.076735	0.003143	24.41142	0.0000
JOB	-8981.047	308.4924	-29.11270	0.0000

R-squared	0.555734	Mean dependent var	21808.91
Adjusted R-squared	0.555617	S.D. dependent var	25188.62
S.E. of regression	16791.24	Akaike info crite...	22.29540
Sum squared resid	7.52E+12	Schwarz criterion	22.29786
Log likelihood	-297479.6	Hannan-Quinn criter.	22.29619
F-statistic	4767.374	Durbin-Watson stat	1.982890
Prob(F-statistic)	0.000000		

图 23: 模型 II 修正结果 (第一次)

可以看到，所有变量均呈现较强的显著性。这意味着这些变量在所研究的模型中具有较高的影响力，对模型结果的解释具有重要意义。下面我们仍然使用 white 检验对模型的异方差性进行修正后的检验。检验结果如图24所示。

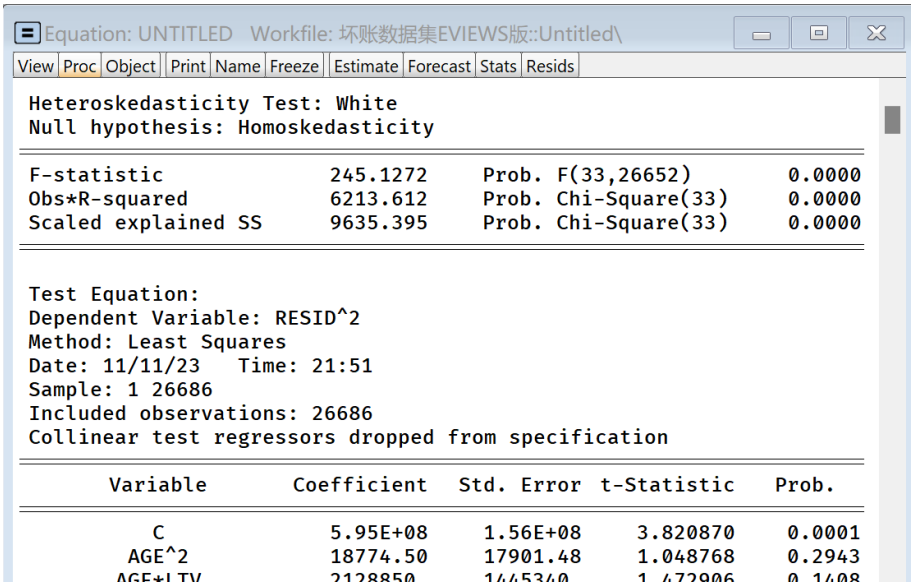

图 24: white 检验第一次修正后的模型 II 结果

很遗憾，仍然存在异方差性，究其原因是 `lvt` 在一定程度上仍然反应了借款金额大小。但是相比之前的模型，nR^2 值有所下降。

由于方程仍然存在异方差性，我们使用 white 调整的方法进行调整。通过依次点击 Estimate -> Options -> Huber-white 进行校正。

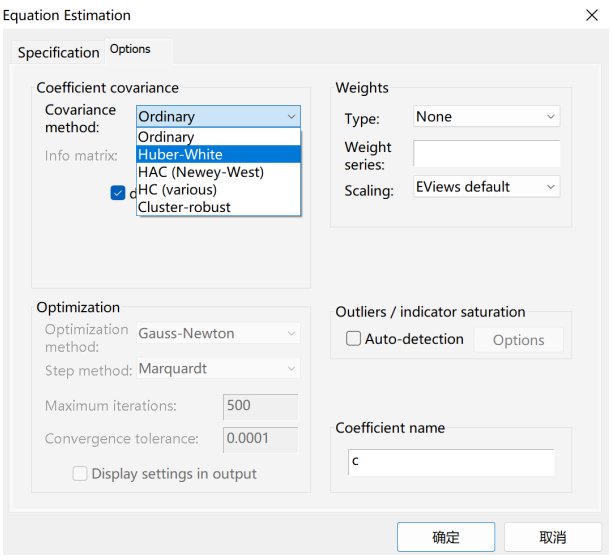

图 25: white 调整步骤

调整后的结果如图26所示

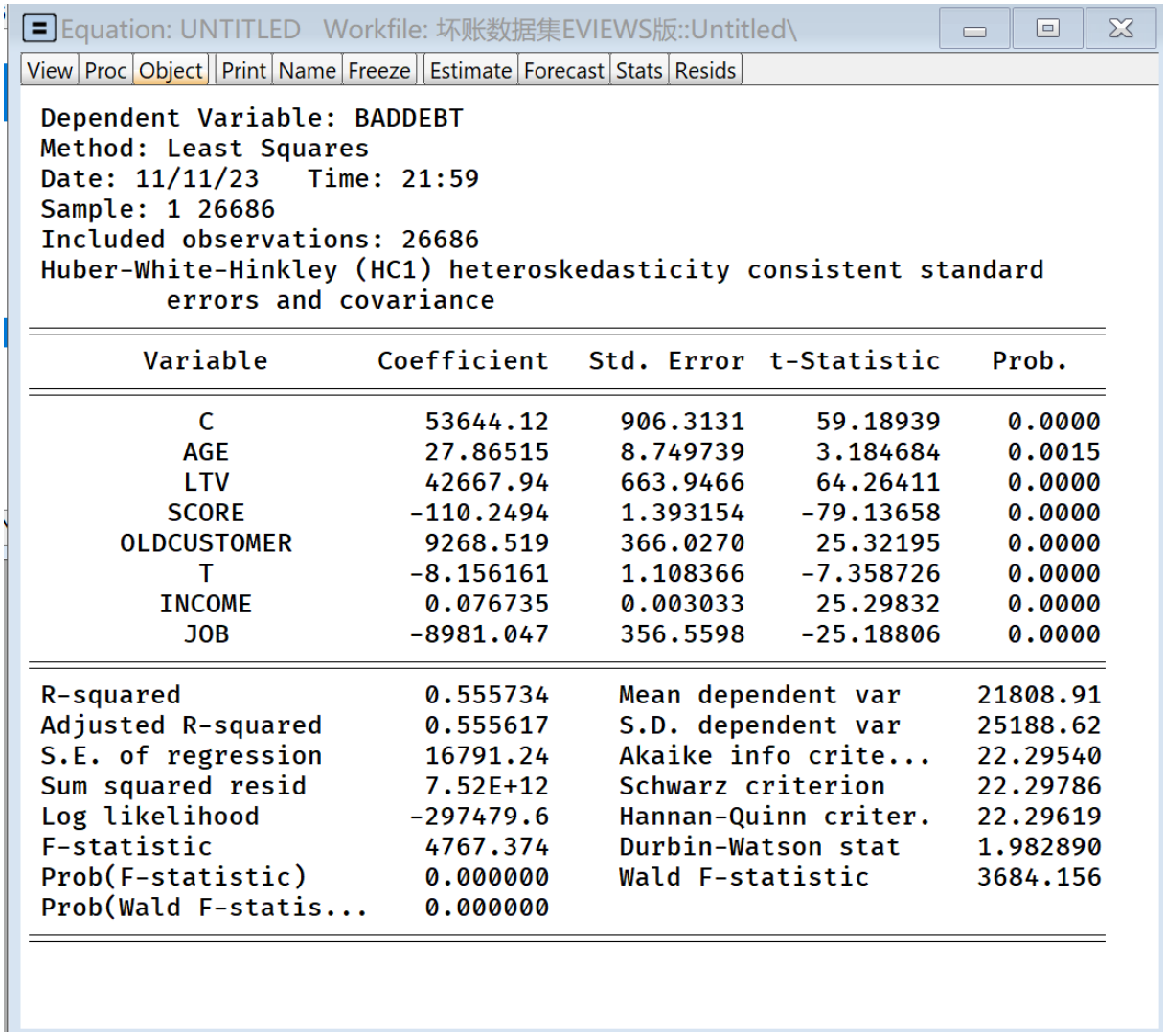

图 26: white 调整后的结果

可以看到，white 调整只修正了标准误，对各个参数的系数没有改变。最终的模型调整为

$$\begin{aligned} BADDEBT = {} & 53644.12 + 27.87 \times AGE + 42667.94 \times LTV \\ & - 110.25 \times SCORE + 9268.52 \times OLDCUSTOMER - 8.16 \times T \\ & + 0.0767 \times INCOME - 8981.05 \times JOB \end{aligned} \tag{4}$$

4 结果分析与解释

基于第三节对两个模型的回归结果，本节主要聚焦在模型结果的分析解释上，我们将深入分析并解释这两个模型的结果，以期更好地理解自变量与因变量之间的关系和经济意义。从而让其指导实际生活。

4.1 模型 I: 还款能力模型

根据我们在第三节所得到的回归结果，我们得到了这一样一个还款能力的 Logit 模型。**请注意，被解释变量取值为 1 时表明该借款者在借款期间存在还款困难**。

$$\begin{aligned} L: Pr(\widehat{TARGET}_i = 1) = &- 0.70 - 0.01 \times AGE - 0.14 \times CAR \\ &+ 0.03 \times CHILDREN + 0.22 \times GEND \\ &- 0.30 \times GRADUATE - 0.10 \times MARRY - 0.28 \times RICH \\ &+ 0.19 \times POVERTY - 0.01 \times ITA \end{aligned}$$

如表9，我们可以将评价用户还款能力的影响因素大致分为三类。

表 9: 三类评价用户还款能力的影响因素

个人信息	资产情况	外部环境
年龄	是否有车	城市 GDP 评级
性别	收入债务比	
教育水平		
婚姻情况		
子女个数		

可以看到，有关个人信息的变量占到了总变量个数的 62.5%，这表明我们在衡量一个人是否具备还款能力时，主要考究其个人与家庭的基本信息，这是因为这些信息能够很好的反映出借款人的信用状况、消费习惯和生活稳定性

而具有独立的个体差异。当然，我们也会同时参考其资产情况和外部环境因素，这两者也是很重要的因素。

值得我们注意的是，对于我们之前无法预测符号的变量 graduate，其实际符号为负。其**系数的经济意义是：在保持其他变量不变的前提下，拥有更高的学历比低学历人群拥有还款困难情况的概率低 7.5%**(-0.3×0.5 × 0.5 = 0.075)。也就是说，高学历的人往往比低学历的人信用更好。

这也符合我们的常识，在一般银行，学历更高者往往也会拥有更高的信用额度。根据 [15] 的研究显示，高学历人群的收入水平较高，这使得他们在还款方面有更强的能力和稳定性。收入越高，还款压力相对较小，违约的可能性就越低；同时，高学历人群往往具备较强的信用意识，明白信用对于个人和社会的重要性。因此，他们在借款时会更加谨慎，还款时也会更加遵守承诺，降低逾期和违约行为的发生。当然，高学历人群在财务规划方面通常具备较强的能力。他们可以更好地管理个人财务，合理安排贷款还款，从而降低还款逾期及违约的风险。因此，学历之于信用是个不错的隐形“度量衡”。

我们也注意到，收入债务比这一个新引入的变量表现出较好的解释效力。其经济意义是：**在其他变量保持不变的情况下，收入债务比每增加 1 个单位，这个借款者拥有还款困难情况的概率就会降低 2.5%**。这启示我们，了解自己的收入债务比非常重要。这有助于他们更加清晰地认识到自己的财务状况，从而做出更为明智的消费和投资决策。在一定程度上，降低债务水平、优化负债结构，有助于提高还款能力，降低信用风险。

4.2 模型 II: 坏账金额模型

基于在第三节的结果，我们得到了这样一个估计坏账金额的线性模型。

$$\begin{aligned} BADDEBT = {} & 53644.12 + 27.87 \times AGE + 42667.94 \times LTV \\ & - 110.25 \times SCORE + 9268.52 \times OLDCUSTOMER - 8.16 \times T \\ & + 0.0767 \times INCOME - 8981.05 \times JOB \end{aligned}$$

上述包含了多个影响因素，如借款人的年龄、贷款价值比率（债务资产

比)、信用评分、是否为老客户、贷款期限、收入和是否在职。下面就几个典型的系数分析。

首先，模型中的系数代表了各个变量对坏账金额的影响程度。从这个模型中，我们可以看出年龄、贷款价值比率、信用评分、是否为老客户、贷款期限、收入和职业这七个因素都对坏账金额有显著影响。

贷款价值比率（债务资产比）为 42667.94，这意味着**在其他条件不变的情况下，债务资产比每增加一个单位，坏账金额就有可能增加 42667.94**。因此，借款人在申请贷款时，应合理评估自己的债务承受能力，避免过度负债。借款人应根据自身的收入、支出、还款能力等因素，合理规划贷款金额，以降低违约风险。

信用评分为 -110.25，表明信用评分越高，坏账金额越低。这一现象符合生活实际，因为信用评分是对借款人信用状况的一种量化衡量，评分越高，说明借款人的信用状况越好，违约风险越小。基于评分卡的贷款审批模型目前也在各大银行、信用贷款 APP 问世，如蚂蚁信用分、微信支付分等也已经有了成熟的评分体系。

是否为老客户的系数为 9268.52，表示老客户的坏账金额要低于非老客户。这个系数的大小超出我们的逾期。但通过分析，我们认为，银行在与老客户长期合作的过程中，积累了丰富的信用数据，可以更准确地评估老客户的信用状况。这有助于银行更加了解老客户的风险特征，从而有针对性地制定风险控制措施，降低坏账风险；同时，银行为了维护老客户关系，通常会为他们提供一定的优惠政策，如更低的利率、更高的额度等。这些政策有助于降低老客户的贷款成本，减轻还款压力，从而降低违约风险。这也就造成了老客户的坏账金额较少的缘故，**因此银行也乐意挽留老客户，而对于新客户的审查更加苛刻**。

这个模型的主要用途是预测坏账金额。通过对借款人的年龄、贷款价值比率、信用评分、是否为老客户、贷款期限、收入和职业等因素进行评估，可以预测出借款人的坏账风险，从而为金融机构提供风险管理的依据。

5 结论、感悟与展望

5.1 结论与展望

信用贷款的审核、评估、预测一直以来都是一个复杂的难题。本文基于某公司的两个公开数据集，通过计量经济学方法回归拟合出了还款能力模型以及坏账金额预测模型。

在还款能力模型中，我们考虑了多个影响因素，如申请人的年龄、性别、婚姻状况、子女个数、教育程度、所在城市 GDP 评级等。通过对这些变量进行回归分析，我们得出了对应的系数，从而可以根据申请人的信息预测其还款能力。这一模型有助于贷款机构在审核贷款申请时，更加准确地评估申请人的信用风险。

坏账金额预测模型则针对已经发生违约的贷款，通过分析违约贷款的债务资产比、信用评级、年龄等因素，拟合出了一个预测模型。该模型可以预测出贷款发生违约时的坏账金额，为贷款机构在发生坏账时提供了一定的防范和应对措施。

通过建立这两个模型对于信贷机构开展信贷业务具有重要作用，其一方面可以消除高风险借款人减少个人消费信贷违约风险，其另一方面可以寻找“高质量”的信贷人，实现双赢局面。在未来，这种应用将进一步发挥其潜力，为信贷市场带来更多积极影响。相信随着大数据和人工智能技术的发展，模型的设定和调整会更加精确，信贷机构能更全面地了解借款人的信用状况，提高贷款审核的精准度。

5.2 感想与感悟

本次计量经济学小组作业由小组三名同学齐心完成，第一节由***负责编写，第二节由***负责编写，第四节由***负责编写，第三节由三人共同完成。在完成本次作业的过程中，收获颇多。

首先，我们对计量经济学的方法有了更深入的理解和掌握。通过本次小组作业，我们对回归分析、预测模型等方法的应用有了更为直观的认识，明白了

如何将理论知识运用到实际问题中。

其次，我们在数据处理和分析方面积累了宝贵的经验。在完成作业的过程中，我们学会了如何使用统计软件进行数据处理，如何从数据集中提取有用信息，以及如何基于模型存在的问题进行诊断、修正、弥补。这些经验将在未来的学术研究和实际工作中发挥重要作用。

此外，本次作业使我们认识到团队合作的重要性。在小组讨论和分工合作的过程中，我们学会了如何有效沟通、如何协调不同观点，以及如何高效完成任务。这种团队协作能力对于我们未来的学术和职业生涯都具有深远影响。

同时，本次作业也让我们意识到计量经济学在实际问题中的应用价值。通过构建还款能力模型和坏账金额预测模型，我们发现计量经济学方法可以帮助信贷机构更好地评估借款人的信用风险，从而降低违约风险。这使我们更加坚定了将理论知识与实际问题相结合的决心。

最后，本次作业使我们认识到自身在知识体系和技能方面的不足。在完成作业的过程中，我们发现自己在某些方面的理论知识尚不完善，对某些实际问题的理解仍有待提高。这激发了我们在今后的学习中更加努力，不断提升自己的综合素质。

总之，本次计量经济学小组作业为我们带来了丰富的收获。我们将以此为契机，继续深入学习和探索计量经济学领域，为未来的学术和职业生涯打下坚实基础。

参考文献

[1] 中国人民银行. 年度金融机构贷款投向统计报告. 人民日报, 2022.

[2] Amir Ikram, Qin Su, Faisal Ijaz, Muhammad Fiaz, et al. Determinants of non-performing loans: An empirical investigation of bank-specific microeconomic factors. *Journal of Applied Business Research (JABR)*, 32(6):1723–1736, 2016.

[3] Konstantinos I Nikolopoulos and Andreas I Tsalas. Non-performing loans: A review of the literature and the international experience. *Non-performing loans and resolving private sector insolvency: Experiences from the EU periphery and the case of Greece*, pages 47–68, 2017.

[4] Fennee Chong. Loan delinquency: Some determining factors. *Journal of Risk and Financial Management*, 14(7):320, 2021.

[5] Matteo Foglia. Non-performing loans and macroeconomics factors: the italian case. *Risks*, 10(1):21, 2022.

[6] 李焱文 , 刘小勇. 互联网贷款个人信用风险影响因素研究综述. 时代金融(上旬) , (12):89–91,94, 12 2020.

[7] 蔡闽. 流量覆盖风险——网络小额信贷风险控制新思路. 金融研究, (9):131–144, 1 2016.

[8] 栾伊娜. 城市信用环境对个人贷款保证保险违约风险影响的研究. 硕士论文, 上海财经大学, 2020.

[9] 雷舰. P2p 网贷借款人信用风险因素分析与对策. 金融理论与实践, (12):31–39, 1 2019.

[10] 杨青. 个人住房抵押贷款的市场风险研究. 硕士论文, 对外经济贸易大学, 2011.

[11] https://www.kaggle.com/competitions/home-credit-default-risk/data?select=HomeCredit_columns_description.csv. Loan Approval Data Set.

[12] https://www.kaggle.com/datasets/yashkmd/credit-profile-two-wheeler-loan-dataset. 全面了解贷款申请人概况 (共计两轮贷款).

[13] 方兆本, 李红星, 李健伦 , 雷娜. 汽车消费、住房抵押贷款违约特性及对策. 运筹与管理, 13(6):69–73, 1 2004.

[14] 吕安民. 中国省级人口增长率和 gdp 增长率及其相关关系研究. 郑州大学学报（理学版）, 38(1):110–114, 1 2006.

[15] 王先柱, 吴义东 , 吴景. 住房按揭贷款逾期风险识别与管理研究——基于借款人学历视角的实证检验. 浙江工商大学学报, (4):125–137, 7 2020.