分类号_____　　密级　公开
U D C _____　　编号　10741

硕 士 学 位 论 文

论文题目　基于 LASSO 的 PSM 变量选择与应用

研　究　生　姓　名：吴桐

指导教师姓名、职称：牛成英

学 科、专 业 名 称：应用统计硕士

研　　究　　方　　向：大数据分析

提　　交　　日　　期：2024 年 6 月 3 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：吴桐 签字日期：2024年6月3日

导师签名：毕成英 签字日期：2024年6月3日

导师(校外)签名：_____ 签字日期：_____

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：吴桐 签字日期：2024年6月3日

导师签名：毕成英 签字日期：2024年6月3日

导师(校外)签名：_____ 签字日期：_____

PSM Variable Selection And Application Based on LASSO

Candidate :Wu Tong

Supervisor:Niu Chengying

摘要

随机化实验是反事实框架下因果效应分析的黄金标准，但基于观察数据的实证研究中，由于各种原因，使得研究的样本单元无法满足随机化分配要求。倾向得分匹配（Propensity Score Matching，PSM）是一种将研究数据处理成“随机对照实验数据”的常用方法，目的在于减少观察数据偏差和混杂因素的干扰，目前在诸多领域有着广泛应用。

但基于高维数据的倾向得分匹配模型设定直接影响处理组与控制组样本匹配结果的平衡性，特别是为了达到控制潜在混淆变量、提高匹配质量和增强模型稳健性等目的而加入过多变量，造成变量之间存在相关性，给匹配带来维度灾难、支持度差异、多重检验等问题，最终导致匹配结果平衡性较差以及因果效应估计不可靠，因此使用 PSM 时需要合理选择模型变量来提高匹配结果的可靠性。

在变量选择方法中，最小绝对收缩和选择算子（Least Absolute Shrinkage and Selection Operator，LASSO）的独特优势是具有自动选择特征的能力，将 LASSO 变量选择的优势应用到 PSM 中，提出得到基于 LASSO 的 PSM 模型，即 LASSO-PSM 模型，解决了传统倾向得分匹配中模型设定主观性和维度灾难问题。结果表明，LASSO-PSM 模型的可行性及其匹配结果的平衡性优于 PSM 模型。

将 LASSO-PSM 模型应用到工作类型偏好与工作选择因素的实证研究中，利用 LASSO-PSM 模型对劳动者择业偏好的相应变量进行筛选，再利用筛选后的变量计算不同工作类型劳动者生活状况（经济收入、心理压力、健康状况和幸福感）的因果效应。研究结果发现：LASSO-PSM 模型选择变量更符合实际意义，不同类型工作劳动者的经济收入存在显著差异。

关键词：LASSO 倾向得分匹配（PSM） LASSO-PSM 变量选择 工作类型

Abstract

Randomized experiments are considered the gold standard for causal effect analysis within the counterfactual framework. However, in empirical studies based on observational data, various reasons often render the sample units unable to meet the requirements of random allocation. Propensity Score Matching (PSM) is a commonly used method to process study data into "randomized controlled trial data," aiming to reduce bias and interference from confounding factors in observational data. Currently, PSM is widely applied in numerous fields to mitigate the impact of observational data biases and confounding factors.

In high-dimensional data, the specification of the propensity score matching model directly affects the balance of matching results between the treatment and control groups. Particularly, adding too many variables to control potential confounders, improve matching quality, and enhance model robustness can lead to inter-variable correlations. This, in turn, results in issues such as the curse of dimensionality, support differences, and multiple testing, ultimately leading to poor balance in matching results and unreliable estimation of causal effects. Therefore, when using PSM, it is essential to judiciously select model variables to enhance the reliability of matching results.

In the variable selection process, the unique advantage of the Least

Absolute Shrinkage and Selection Operator (LASSO) is its ability to automatically select features. By incorporating the advantages of LASSO variable selection into Propensity Score Matching (PSM), a LASSO-based PSM model, namely LASSO-PSM model, is proposed to address the subjectivity and dimensionality issues in traditional propensity score matching. The results indicate that the feasibility of the LASSO-PSM model and the balance of matching results are superior to those of the PSM model.

In an empirical study on the preference for job types and factors influencing job selection, the LASSO-PSM model was applied to select relevant variables related to workers' job preferences. Subsequently, the selected variables were utilized to calculate the causal effects of different job types on workers' living conditions (economic income, psychological pressure, health status, and sense of happiness). The research findings indicate that the variables selected by the LASSO-PSM model are more aligned with practical significance, and there are significant differences in economic income among workers in different types of jobs.

Key words: LASSO；Propensity Score Matching(PSM)；LASSO-PSM；Variable Selection；Job Type

目 录

1 绪论

1.1 研究背景与研究意义

1.1.1 研究背景

随机化实验是反事实框架下因果效应分析的黄金标准，但基于观察数据的实证研究中，由于各种原因，使得研究的样本单元无法满足随机化分配要求。倾向得分匹配（Propensity Score Matching，PSM）是一种将研究数据处理成“随机对照实验数据”的常用方法，目的在于减少观察数据偏差和混杂因素的干扰，目前在诸多领域有着广泛应用。

然而，在运用 PSM 进行科学研究时，特别是基于高维数据的倾向得分匹配时，有时为了控制更多的潜在混淆变量、提高匹配质量和增强模型稳健性等目的，在倾向得分匹配模型中引入了过多变量，这导致变量之间呈现相关性，对 PSM 模型的参数估计产生影响，降低了匹配的可行性，造成匹配结果的平衡性较差，最终导致因果效应的估计结果不可靠。

因此，在运用 PSM 进行科学研究时，可通过对变量进行筛选来改进上述出现的问题，以提高 PSM 模型匹配结果的平衡性和模型的稳定性，最终使因果效应估计更可靠。在筛选变量时需要权衡变量的相关性，选取具有显著差异组合的协变量，可以使用合适的统计方法进行变量选择。例如，通过方差分析、主成分分析和逐步回归等方法进行变量筛选，当出现变量呈高度相关、数据集较小、变量呈非线性关系等情况时，这些方法筛选出的变量组合不够准确，从而影响 PSM 模型匹配结果的平衡性。

Robert（1996）提出的最小绝对收缩和选择算子（Least Absolute Shrinkage and

Selection Operator，LASSO）方法可以对变量进行选择，达到数据降维的目的，它保持了子集收缩和岭回归的良好特征，从而能在实现变量选择的同时对参数进行估计。LASSO 方法的引入在统计学和机器学习领域引起了广泛的关注和应用，它在高维数据分析、特征选择、信号处理等领域具有重要作用，并为处理大规模数据、稀疏数据和变量筛选等问题提供有效的解决途径。LASSO 的关键特点是具有自动特征选择能力，通过 L_1 正则化使变量选择和模型建立一体化，使部分不重要变量的系数参数收缩为 0，从而实现对不重要变量进行剔除。同时 LASSO 具有较强的灵活性，可以捕捉到变量与目标变量之间的非线性关系，只保留对目标变量预测具有显著影响的变量，有效地降低模型复杂度，提高模型的稳定性。

相比于方差分析、主成分分析和逐步回归等方法，LASSO 在变量筛选的问题上具有一定优势，因此将 LASSO 在变量选择上的优势应用到 PSM 中，来改善因加入过多变量给匹配带来地支持度差异、维度灾难、多重检验和数据过拟合等问题，从而提高 PSM 模型匹配结果的平衡性和因果效应估计的可靠性。

1.1.2 研究意义

运用 PSM 进行科学研究时，研究人员只能依据主观判断选取研究所需的变量组合，有时选取的变量组合存在高度的相关性或选取的变量过多，从而造成多重比较等问题，导致 PSM 模型匹配结果的平衡性和模型的稳定性较差。

于是有必要对所选取的协变量组合进行“缩简”，使用基于 LASSO 的 PSM 模型筛选出个数较少、且具有显著差异的变量组合，从而提高 PSM 模型匹配结果的平衡性和模型的稳定性，使因果效应的估计结果更加可靠。

基于 LASSO 的 PSM 模型筛选出个数较少、具有显著差异的变量组合，为

PSM 模型因加入过多变量而对匹配带来支持度差异、维度灾难、多重检验、数据过拟合等问题提供解决途径，增强 PSM 模型匹配结果的平衡性和模型的稳定性，使因果效应的估计结果更加可靠，同时也为其他学科的研究提供可靠支持，具有重要的社会意义和现实意义。

1.2 文献综述

Rubin（1983）首次提出 PSM，该方法主要解决因果推断中存在的选择性偏倚问题，PSM 将具有相似处理概率的个体进行匹配，从而实现处理组和控制组之间合理的比较，为因果推断提供强有力的工具，对倾向得分方法的发展和应用产生了深远的影响。研究者在使用 PSM 方法时，可能会遇到不同的问题，例如在变量选择上的困难。国外对变量选择方面的研究，经历了以下研究历程。

研究者使用 PSM 解决实际问题时，Rosenbaum 等（1985）首次提出全面变量法，该方法将所有可能与处理变量和结果变量相关的变量都纳入到模型中，以控制混淆变量的影响，并描述如何使用多变量匹配方法来构建控制组。Rosenbaum 将协变量全部纳入匹配过程中，全面考虑潜在的混淆变量，以提高匹配组与控制组之间的平衡性。Shen 等（2024）使用 PSM 中全变量法对中国不同地区农户的食物浪费情况进行调查，结果显示劳动力迁移显著增加农村家庭的食物浪费。Dehejia 等（2002）在研究就业培训项目的因果效应时，采用全面变量法将所有可能的混淆变量纳入到模型中，以控制混淆因素。Austin（2008）对医学文献中在 1996 年至 2003 年期间使用倾向得分匹配的研究进行批判性评估，其中包括对全面变量法的讨论和评价，强调 PSM 中全面变量法在提高匹配质量方面的重要性，但在使用 PSM 全面变量法进行匹配的过程中，发现存在匹配结果

不可靠的问题。Imai 等（2007）讨论全面变量法，当协变量维度较高时可能出现维度灾难的问题，导致匹配结果不可靠。Stuart（2010）发现全面变量法在处理高维数据时容易导致维度灾难，需要进行特征变量选择或降维等方法来解决这个问题。Zubizarreta（2015）讨论在 PSM 中使用稳定权重来平衡协变量，以便在不完整结果数据的情况下进行因果效应估计，其中也提到全面变量法可能会造成匹配过程过度纳入变量，导致因果效应估计结果失真。可见，全面变量法极有可能因为 PSM 模型中加入过多变量导致存在维度灾难和过度调整等问题，于是研究者便想通过一些方法对变量进行筛选。

根据前置知识法进行变量筛选，前置知识法是根据领域专家的先验知识，选择与处理变量和结果变量相关的变量。前置知识法中包括预处理变量、共同因果图（DAG）、可信度指数等方法，

预处理变量方法是在进行 PSM 匹配之前，根据领域专家的知识，选择与因果效应和处理选择相关的变量，剔除不相关的变量、进行变量变换或缩放等。Austin（2011）论证预处理变量方法可减少观察性研究中的混杂效应。Stuart（2010）发现预处理变量方法可减少模型复杂性，提高估计精度，但可能忽略一些重要的混淆变量。

共同因果图方法是使用因果图来表示变量之间的因果关系，并根据领域专家的知识来构建共同因果图。在图中，选择与处理选择相关的变量以及可能的混淆变量，以确保通过 PSM 进行匹配时可以控制潜在的混淆因素。Pearl（1988）详细介绍 DAG 的原理和应用，并将其作为因果推理和概率推理的有力工具进行阐述，对 DAG 的发展和应用产生深远的影响，并为后续因果推断和因果分析的研

究奠定基础。Spirtes（2000）总结因果图模型的理论、方法和应用，介绍有向无环的共同因果图作为表达因果关系的工具，可根据图示的因果关系对变量进行筛选进行深入讨论。Pearl 等（2018）介绍了因果推断的最新进展和应用，从哲学、统计学和人工智能等多个视角探讨了因果关系的本质，并解释 DAG 模型在因果推断中的作用。但是构建 DAG 需要依据领域专家的主观判断和先验知识，不同的专家可能对变量之间的因果关系有不同的见解，这可能导致在构建 DAG 时存在主观倾向性，从而影响到变量的选择和匹配结果的可靠性。Abadie 等（2010）在实证中发现构建 DAG 时存在多重比较问题，即选择一个符合预期结果的模型或图结构，而忽略其他可能的结构，导致结果出现选择性偏差。Pearl（2009）发现在变量较多或数据量较大的情况下，构建 DAG 可能变得非常复杂，并且需要耗费大量的时间和计算资源。

除上述两种方法，还有学者使用可信度指数等方法对 PSM 模型中的变量进行筛选，而这些方法都依赖于领域专家的知识和经验，会忽略一些重要的混淆变量，从而导致因果效应估计结果出现选择性偏差。

根据变量筛选法进行变量选择，从所有可能的变量中选择与因果效应和处理选择相关的最优子集，这种方法可减少模型复杂性，提高估计精度，但需要依赖于统计学假设和算法参数的选择。变量筛选法包括前向选择、后向消元、逐步回归、卡方检验等。

前向选择法是从空模型开始，逐步添加最相关的变量，直到达到预设的停止准则（如 AIC、BIC 等）或没有可加入的变量为止。Peres 等（2001）对比研究前向选择和其他变量选择方法，并指出前向选择的简单直观和计算效率高的优点，

但在实际使用中，发现前向选择法可能会陷入局部最优解。Austin 等（2015）讨论前向选择方法的局限性，如忽略变量间的交互作用和 PSM 模型中的噪声变量的敏感性。Brookhart 等（2006）研究发现前向选择法对停止准则的依赖以及对噪声和冗余变量较为敏感。

后向消元法是从包含所有变量的完整模型开始，逐步去除最不相关的变量，直到达到预设的停止准则或没有可剔除的变量为止。Zhang 等（2018）使用后向消元法来筛选出最佳的特征子集，并提到后向消元法的优点之一是能够识别不必要的变量，减少过拟合的风险。尽管后向消元法是一种常用的变量选择方法，但它也存在一些缺点。Guyon 等（2003）提到后向消元法可能导致模型丢失重要特征变量的问题。Chandrashekar 等（2014）发现后向消元法对噪声和冗余特征较为敏感，此外，后向消元法可能忽略特征变量间的关系，可能导致无法发现某些特征变量与响应变量之间的复杂关系，从而降低模型的预测性能。

Imbens 等（2009）提出一种基于逐步回归的变量选择方法，用于估计最优倾向得分，并应用于多个研究领域。逐步回归结合前向选择和后向消元的方法，根据预设的停止准则，同时考虑添加和删除变量，直到达到预设的停止准则为止。Zou 等（2005）经过实证发现逐步回归计算效率高。Hastie 等（2009）发现逐步回归可以通过逐步剔除不重要的特征变量来简化模型，减少过拟合的风险，提高模型的泛化能力。James 等（2013）发现逐步回归在每个步骤都会记录添加或剔除的变量，使得结果更具可解释性。逐步回归也存在缺点，Draper 等（1998）发现当特征数量较多时，逐步回归可能会在训练数据上过度拟合，并导致在新数据上的性能下降。逐步回归是基于单变量分析的方法，Fan 等（2010）实证中发现其忽略变量之间的相互作用和相关性，可能导致无法识别某些变量与目标变量之

间的复杂关系。逐步回归是一种贪婪算法，Guyon 等（2003）发现其每一步只关注当前最佳的变量选择，可能会剔除其他对目标变量有潜在影响的重要变量。

卡方检验用来评估变量与处理选择之间的关联程度，选择与处理选择显著相关的变量，具有简单易用、非参数性、检验结果直观等优点，但也存在一些缺点。Rosenbaum 等（1985）提出在 PSM 中使用单变量卡方检验进行变量筛选可能导致偏差和误导性问题。Luellen 等（2005）介绍使用单变量卡方检验进行变量筛选可能忽略某些重要性变量。Stuart（2010）指出在 PSM 中使用卡方检验进行变量筛选可能忽略变量之间的相互作用和复杂关系。

也有学者将其他方法应用到变量选择中。Xiao 等（2023）将 PSM 与 DID 结合对变量筛选并进行实证分析，研究发现中国碳排放交易体系的试点企业通过雇用更多劳动力和减少资本投资来提高劳动收入份额。Do 等（2023）使用表单规则进行特征变量选择。Liu（2024）使用自适应 LASSO 的惩罚剖面最大似然法，进行随机效应的半参数空间自回归面板数据模型中的变量选择。Cheng（2022）等提出一种基于 MI 结合 VIF 的变量选择方法，并验证 MI-VIF 变量选择方法在模型构建中是可行的。

上述不少学者将变量选择的各种方法运用到 PSM 中，除此之外，也有不少学者将 PSM 运用到社会学的各项研究中。

徐小兵等（2023）运用 PSM 研究农村中老年人慢性病共病对失能的影响，发现中老年人慢性病共病会增加失能的风险。赵梦阳（2023）分析农民合作社对农村集体经济的影响机理，利用民政部农村社会治理数据库村级调查数据，使用倾向得分匹配和广义倾向得分匹配（GPSM）等方法分析合作社对农村集体经济

的影响，研究结果显示：村内有合作社会对农村集体经济产生显著的负向作用，不过随着农户入社比率的增加，合作社对集体经济呈现出先抑制后促进的影响，机制检验发现，发展合作社会因增加村行政办公经费支出、挤占村干部精力、减少村集体经济组织财务状况公开次数、损害集体经济组织的公开透明性而抑制集体经济发展。

也有学者将 PSM 与其他方法相结合，例如 PSM-DID，邓小慧（2023）利用 PSM-DID 探究婚姻与幸福的关系，发现婚姻状态转变（未婚→已婚）能明显提升个体幸福感，婚姻同时带来的社会支持和经济增益等可能是婚姻促进幸福感的主因。黄寰（2023）使用 PSM-DID 方法并通过稳健性测试，多维度分析低碳试点城市建设对城市碳排放的影响，并采用中介模型考察低碳试点政策的影响机制。王玉荣（2022）运用 PSM-DID 方法，实证检验工业互联网对企业数字创新的影响。康娜（2023）使用 PSM-DID 方法探讨二孩生育惩罚机制的存在性和持续性是否有加深生育率低迷的可能。蒋青檀（2023）基于倾向得分构造处理组和对照组协变量的经验加权分布，采用能源距离度量加权经验分布与总体协变量经验分布的差异，通过最小化分布差异最大化协变量平衡，进而估计倾向得分和平均因果效应。

通过文献梳理可以发现，国外文献主要侧重于对 PSM 中变量筛选进行研究，国内文献主要侧重于运用 PSM 解决社会学的相关问题。并且上述的各种变量筛选的方法都不能合理地选出一组不相关的变量组合，本文在此基础上，提出将 LASSO 运用到 PSM 中进行变量选择，克服上述可能出现的问题。

1.3 研究思路框架与研究方法

1.3.1 研究思路框架

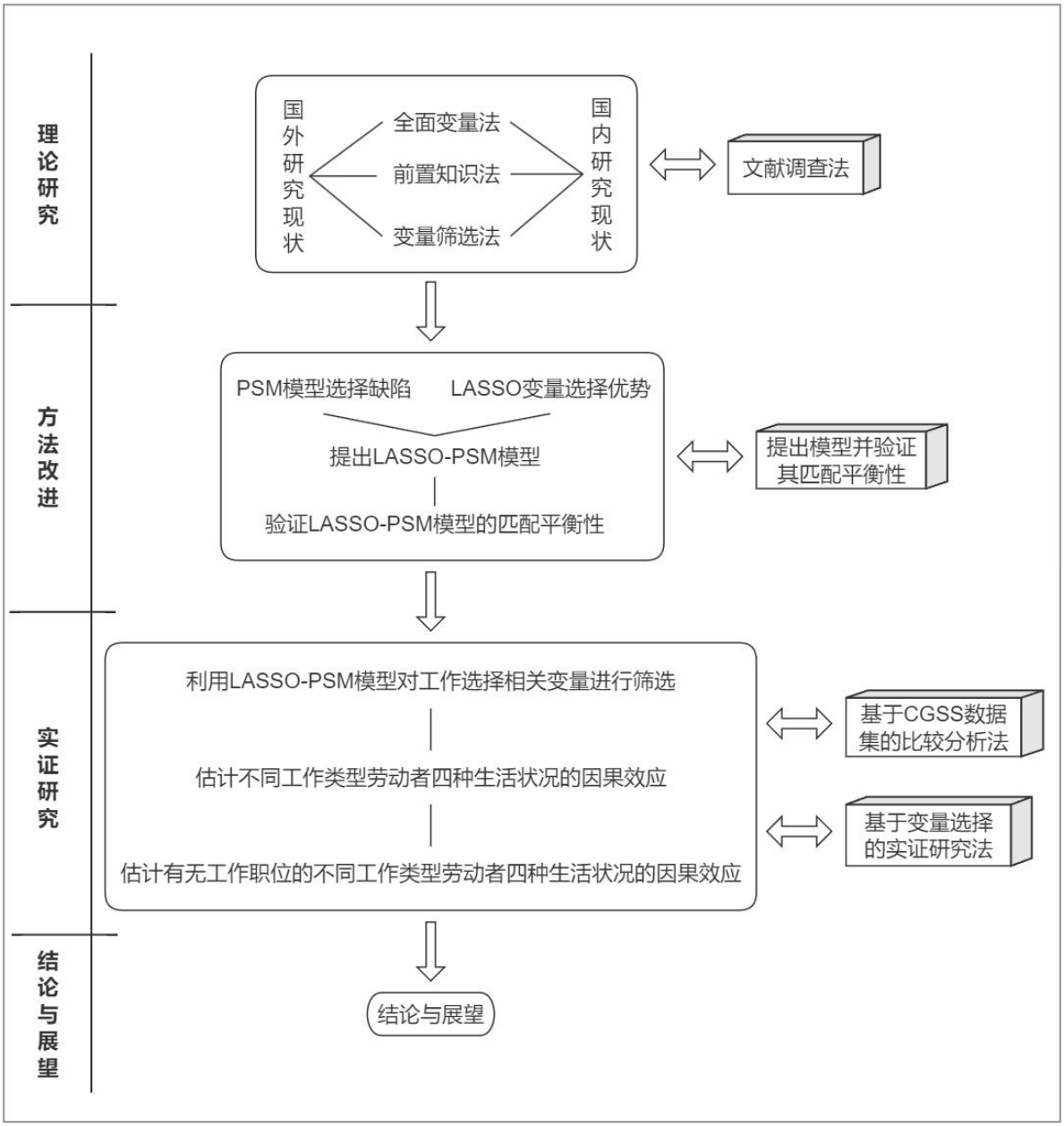

图 1.1 研究思路与技术路线图

1.3.2 研究方法

文献分析法：已有相关研究文献是本文研究的基础，通过全面细致的文献研

究，可以了解当前 LASSO 在变量选择上的优势，以及 PSM 因加入过多变量而带来支持度差异、维度灾难、多重检验、数据过拟合等问题。文献梳理同时也为实证分析所需变量的选取及处理提供相应的依据。

比较分析法：对比 LASSO-PSM 模型和 PSM 模型匹配后协变量的标准化平均值差异、对数标准差比以及倾向指数的标准化平均值差异，以验证 LASSO-PSM 模型的可行性以及模型匹配结果的稳定性。同时在实证研究分析中计算不同工作类型劳动者的生活状况（经济收入、心理压力、健康状况和幸福感）的平均因果效应进行比较分析。

定量分析法：基于 CGSS 实证数据，对变量进行处理时，将计数和评级的数据转化为数字形式数据。同时在实证研究中采用定量分析法分析工作类型偏好与工作选择因素的关系。

实证研究法：基于 CGSS（China General Social Survey，中国综合社会调查）实证数据和定量分析法，对大样本数据进行统计分析工作类型偏好与工作选择因素的关系。

机器学习法：使用 LASSO-PSM 模型进行变量筛选时，通过算法和模型来使计算机自动从数据中学习和改进，来筛选出最优的变量组合。

1.4 研究的贡献与创新

本文研究针对 PSM 中因加入过多变量而对匹配带来支持度差异、维度灾难、多重检验、数据过拟合等问题，提出将机器学习中的 LASSO 与 PSM 结合对变量进行筛选的方法，以解决 PSM 中模型变量选择和模型设定对匹配结果的影响。

实证研究中，使用 LASSO-PSM 模型对劳动者择业偏好影响因素进行实证分

析，对劳动者择业偏好影响因素进行定量分析，充实了国内外对劳动者择业偏好的分析研究。

2 相关理论基础

2.1 PSM 理论基础

PSM 旨在解决观察研究中的选择性偏倚问题。在观察研究中，个体的处理分配受某些外部因素影响，导致个体无法随机地被分配到处理组（Treatment，或称为干预组或实验组）或控制组（Control，或称为对照组）。这种非随机的处理分配可能导致处理组和控制组在协变量特征上存在差异，从而使直接比较处理组和控制组的结果产生系统性偏差。因此，PSM 主要解决以下两个问题：

处理组和控制组之间的观测特征不平衡：由于非随机的处理分配，处理组和对照组之间可能存在协变量的差异，这些差异可能会影响到研究结果。例如，在评估一种医疗政策干预措施的效果时，如果患者在接受干预前就已经存在偏高的医疗开销，则直接比较干预组和控制组的医疗开销可能出现误导性的结论。PSM 通过估计倾向得分进行样本的匹配，可以减少处理组和控制组之间协变量特征的差异，使处理组和控制组之间的比较更具可靠性。

选择性偏倚：选择性偏倚是指个体选择接受处理（处理是指个体分配到处理组或控制组）的概率与协变量之间存在相关性，这会导致在处理组和控制组之间存在系统性的差异。例如，在评估一种教育政策效果时，如果有教育意愿的家庭更倾向于接受该政策，则直接比较接受该政策的学生和未接受该政策的学生的学业成绩可能出现误导性的结论。PSM 通过估计倾向得分进行样本匹配，可以减少选择性偏倚的影响，从而提高因果效应估计的可靠性。

因此，PSM 主要解决观察研究中由于非随机处理分配导致的选择性偏倚问题，通过减少处理组和控制组之间的协变量差异，以提供更可靠的因果推断。

2.1.1 反事实结果

首先，需要明确研究目的和研究问题，依据个体的某些特征确定处理组和对照组，即将样本分为处理组和控制组。处理组是接受某种处理、干预或治疗的个体，控制组是未接受该处理、干预或治疗的个体。个体i接受某种处理D_i，令个体i的$D_i=1$，表示个体分配到处理组，其对应的结果Y_i为$Y_i(1)$；个体i未接受某种处理，令个体i的$D_i=0$，表示个体i分配到控制组，其对应的结果Y_i为$Y_i(0)$，例如将服用药物的个体，分到处理组（$D_i=1$），其身体健康结果为$Y_i(1)$；同理，未服用药物的个体，分到控制组（$D_i=0$），其身体健康结果为$Y_i(0)$。称这两种结果为反事实结果（或者称为潜在结果或结果变量），具体表示如下式（2.1）：

$$Y_i(\text{潜在结果})=\begin{cases}Y_i(0),\text{如果}D_i=0\\ Y_i(1),\text{如果}D_i=1\end{cases} \tag{2.1}$$

之所以称Y_i为反事实结果，是因为这两个结果是个体i本身一直具备的，只不过未必都显现出来，如果没有显现出来，就无法观测到。就如每个个体无论是否真的服用药物，都有服用药物和未服用药物两种情况下潜在的身体健康结果。对于未服用药物的个体，存在他如果服用药物的一个反事实结果，只不过由于未服用药物而没有显示出来；同理，对于已服用药物的个体，也存在他如果未服用药物的一个反事实结果。

2.1.2 倾向得分模型

此时，对于被分配到处理组的个体i，如何在控制组匹配与个体i相似的那个“他”？对所有可观测特征变量进行直接精确匹配是理想的匹配方法，因为它保证了处理组和控制组的协变量X（影响潜在结果Y的变量X被称为协变量）完

全相同。如果协变量 X 只包含少数几个非连续变量，就可以进行直接匹配。但是，当协变量 X 维数增加时，要在多维变量上进行直接匹配变得更加困难。即使协变量 X 只包含了 2 个变量，如果每个变量有 10 个值，需要匹配的特征组合就变成 100 个。当样本量有限、特征组合众多时，对于有些特征组合，可能无法同时找到完全一样的处理组和控制组样本。如果协变量 X 还包含连续变量，那么直接匹配方法不可行。

为了解决这个被称为“维数的诅咒”的问题，Rosenbaum and Rubin（1983）提出了 PSM 方法。PSM 的原理如下式（2.2）：

$$P（X）= F（D = 1 | X）\qquad（2.2）$$

其中，X 为协变量向量 $x = (x_1, x_2, \cdots, x_m)$，或者称为特征向量。

根据已知协变量向量的取值（x_i）计算的第 i 个个体进入处理组的条件概率 $P（X）$（即根据个体的协变量向量判定个体 i 进入处理组的概率），便通过函数关系 F 将多维协变量 X 变换为一维倾向得分（propensity score）$P（X）$，函数关系 F（也被称为倾向得分模型）常用的模型有 Logistic 回归模型（或者称为 Logit 回归模型）、Probit 模型等。主要介绍本次研究中使用的 Logistic 回归模型。

Logistic 回归模型是一种广义线性模型，常用于建立二分类问题的预测模型，它可以用于构建倾向得分模型，估计个体被分配到处理组的概率。以下详细介绍如何使用 Logistic 回归模型构建倾向得分模型：

首先，定义因变量：将因变量定义为二分类变量，通常用 0 表示控制组，1 表示处理组，这样可以将问题转化为一个二分类的预测问题。

然后，选择协变量：可以根据领域知识、统计意义和相关性等因素选择合适的协变量作为模型的输入特征，协变量应该与处理组分配有关，并且处理组和控制组之间可能存在协变量不平衡。

其次，建立模型：使用 Logistic 回归模型来建立倾向得分模型。Logistic 回归模型假设因变量服从二项分布，模型如下式（2.3）：

$$P(D_i=1|X=x_i)=\frac{e^{\beta_0+\beta x_i}}{1+e^{\beta_0+\beta x_i}} \tag{2.3}$$

其中，D_i 表示个体 i 被分配到处理组（$D_i=1$）还是控制组（$D_i=0$），x_i 是个体 i 的协变量向量，β 是模型系数。

最后，模型拟合与参数估计：使用最大似然估计来拟合 Logistic 回归模型，并估计出模型参数的值，在拟合过程中，根据样本数据的表现，调整模型参数以最大化似然函数。

对模型进行拟合和参数估计后，需要对模型进行评估，即评估所建立的 Logistic 回归模型的性能和拟合度。常见的评估指标包括 AIC（赤池信息准则）、BIC（贝叶斯信息准则）和模型的拟合度，也可以使用交叉验证等方法评估模型的稳定性和泛化性能。

此外，通过估计得到的模型系数 β 的正负，看出协变量对 P（X）存在正向影响还是负向影响，计算协变量的平均边际效应，可以直观地看出协变量对 P（X）影响程度的大小。

2.1.3 估计倾向得分

通过 2.1.2 建立含有未知参数的倾向得分模型，以及依据样本数据使用极大似然法对参数进行估计，最终得到不含未知参数的倾向得分模型。

再将每个个体i的协变量x_i带入到倾向得分模型中，得到每个个体i的倾向得分$P（x_i）$，其取值范围为(0,1)。倾向得分$P（x_i）$代表个体i被分配到处理组的概率的大小。

依据计算得出的倾向得分对样本进行分组，检验每一组内的处理组和控制组样本的协变量特征是否达到平衡，如果未达到平衡，说明构建的倾向得分模型存在一定的问题，此时需要加入高阶变量或交叉项以修正倾向得分模型。

2.1.4 样本匹配方法

计算出每个个体i的倾向得分，便可以进行样本匹配。具体实施方法有很多，以下介绍在此次研究中所使用的匹配方法。

近邻匹配法是对处理组中的样本，选择控制组中倾向得分最接近的n个样本作为其匹配样本：如果只取最近的1个样本，即$n=1$，称作最近邻匹配法；如果取最近的5个样本，即$n=5$。

使用这个方法要决定控制组样本是否可以重复使用。可重复使用是指控制组里的样本可以多次使用以作为处理组样本的匹配。如果不允许重复使用，控制组里的样本只能被用于匹配一次。如果允许重复使用，匹配的平均质量将增加，偏差会减少，代价是估计的方差会变大。这是在匹配时通常遇到的问题，必须在偏差与方差之间权衡。在实际运用中，可重复使用是比较常用的方法。

例如，假设有两个处理组样本个体，倾向得分分别为(0.6,0.7)。最近的3个控制组样本个体得分分别为(0.62,0.56,0.3)。如果$n=1$，并且控制组样本可以重复使用，那么处理组的两个样本都用控制组最接近的倾向得分为0.62的样本进行匹配。即0.6→0.62（→表示0.6的处理组样本与0.62的控制组样本配对），

0.7→0.62。如果样本不能重复使用，并让得分为 0.6 的处理组个体先匹配，匹配结果为 0.6→0.62，0.7→0.56。可以看到，第二个样本的匹配误差较大。不可重复使用的匹配结果还取决于匹配顺序。如果把匹配顺序倒过来（让得分为 0.7 的处理组个体先匹配），结果变为 0.7→0.62，0.6→0.56。

近邻匹配法的缺点是，即使可以重复使用，也存在着处理组样本的倾向得分和最近控制组样本的倾向得分可能相差较大的情况。

2.1.5 匹配效果诊断

匹配方法的目的是构造更加相似的样本，使处理组和控制组更具有可比性。事实上，在匹配之前就需要检验协变量的平衡性，如果协变量比较平衡，两组个体本来就具有比较好的可比性，也就没有必要进行匹配，可以直接利用回归等方法进行因果效应的估计。如果发现两组个体协变量存在有较大差异，直接回归往往会造成很大的估计偏差，进行匹配才是必要的，匹配方法相当于从观测数据中将隐藏的随机化实验样本寻找出来（King and Nielsen，2016），因而，对于匹配完成后形成的匹配样本，需要检验是否近似于随机化实验。常用的检验指标包括标准化平均值差异（Standardized Difference Averages）和对数标准差比（Log Ratio of Standard Deviations）等指标。

标准化平均值差异定义如下式（2.4）：

$$\Delta_{\alpha}=\frac{\overline{X}_{t}-\overline{X}_{c}}{\sqrt{(s_{t}^{2}+s_{c}^{2})/2}} \tag{2.4}$$

其中，$\overline{X}_{d}$、$d=t$，c 分别表示处理组和控制组某协变量的平均值，s_{d}^{2}表示处理组或控制组某协变量的样本方差，如下式（2.5）：

$$s_d^2 = \frac{\sum_i^{N_d} (X_i - \bar{X}_d)^2}{N_d - 1} \tag{2.5}$$

可以看到，如果两组个体协变量完全平衡，标准化平均值差异将接近于 0，因而 Δ_α 的值越接近于 0，说明样本越有可能平衡。

注意式（2.4）不同于 t 统计量，标准化平均值差异 Δ 与样本容量无关，考察的是给定样本两组协变量平均值是否相同，是样本内特征，和样本容量没有关系。Imbens and Rubin（2015）强调不要使用 t 检验来检验平衡性，因为 t 统计量与样本容量有关，当样本容量很大时，即使协变量平衡也可能发现显著差异的结果。另外，在匹配过程中，需要计算匹配前后的标准化平均值差异，以判断匹配效果。

标准化平均值差异主要考察了一阶矩，另一个指标对数标准差比考察的是二阶矩的差异，如下式（2.6）：

$$\hat{\Gamma}_\alpha = \ln(s_t) - \ln(s_c) \tag{2.6}$$

根据上式（2.6）可以看出，如果两组协变量分布平衡，那么两组协变量标准差将相同，从而两组协变量标准差的对数比将接近于 0。一般情况下，检验上述两个指标就可以大概地了解两组变量的协变量平衡性。在非正态分布中，前两阶矩不一定决定整个分布，因而，前两阶矩平衡并不一定代表整个分布平衡。为此，使用倾向得分的标准化平均值差异检验倾向得分的平衡性，如下式（2.7），Imbens and Rubin（2015）证明，如果两组倾向得分的期望值相同，那么两组个体的协变量分布将相同。因而，在匹配前后，可以比较两组个体的倾向得分平均值，计算倾向得分的标准化平均值差异，就可以检验两组协变量分布的平衡性。

$$\hat{\Delta}_\alpha^l = \frac{\bar{l}_t - \bar{l}_c}{\sqrt{(s_{l,t}^2 + s_{l,c}^2)/2}} \tag{2.7}$$

其中，$\bar{l}_t$ 是 $s_{l,t}^2$ 分别表示处理组倾向得分的平均值和倾向得分估计值的样本方差，$\bar{l}_c$ 是 $s_{l,c}^2$ 分别表示控制组倾向得分的平均值和倾向得分估计值的样本方差。

除上述三个检验指标外，还有一些直观的诊断方法，主要包括倾向得分分布图、分位数分布图（QQ 图）和标准化平均值差异变化图等。

倾向得分分布图是估计出倾向得分后，直接画出处理组和控制组的倾向得分分布图（直方图或概率密度图），观察两组个体倾向得分分布的差异，如果分布相似或图形重合部分较多，说明协变量平衡；如果分布差异较大或图形重合部分较少，说明协变量分布差异较大，匹配效果不好。可以同时画出匹配前后的倾向得分分布图，进行比较并判断匹配的效果。

QQ 图是将处理组和控制组的协变量或倾向得分按照分位由低到高分别画在横轴和纵轴上，以检验两组个体协变量或倾向得分分布是否相似，若完全相似，QQ 图形将与 45 度线重合，偏离越大，说明两组协变量差异越大。此外，皮尔逊相关系数（Pearson Correlation Coefficient）可以衡量 QQ 图中两个变量之间的线性关系的强度和方向，通过计算两个变量之间的协方差除以各自标准差的乘积得到皮尔逊相关系数，其取值范围在-1 到 1 之间，其中-1 表示完全的负相关，1 表示完全的正相关，0 表示没有线性相关性。

标准化平均值差异变化图是将每个协变量匹配前后的标准化的倾向得分均值差异用图形的方式呈现出来，从而可以直观地观察匹配效果。

完成配对后，应对配对结果进行检验，以验证处理组和对照组之间的协变量特征平衡是否得到改善。可以使用标准化差异（Standardized Difference）等指标和各种结果图检验协变量特征平衡性，以评估匹配结果。

2.1.6 因果效果估计

前面几阶段是设计阶段，在以上阶段中，没有考虑结果变量。通过定义相似性，运用合适的匹配方法得到匹配样本，并检验匹配样本的匹配质量。如果匹配样本使协变量平衡，从而匹配样本近似于随机化实验数据，则可以进入分析阶段。设计阶段相当于将隐藏于观测数据中的随机化实验样本寻找出来，后面的分析可以借鉴随机化实验数据的分析方法。

这里主要介绍处理组平均因果效应的匹配估计量，匹配估计量可以写成下式（2.8）：

$$\tau_{ATT}=\frac{1}{N_t}\sum_{i;D_i=1}[Y_i-\sum_{j\in M_{j(i)}}w(i,j)\ Y_j] \tag{2.8}$$

其中 N_t 表示处理组样本量，$w(i,j)$ 为权重，$0<w(i,j)\leq 1$，$M_{j(i)}$ 是上文定义的与处理组个体 i 相匹配的控制组个体的集合。

不同匹配方法主要差别在于权重差异，对于一对一最近邻匹配 $k(k=1)$，$w(i,j)=1$，不在匹配集 $M_{j(i)}$ 中控制组个体权重均为 0。对于近邻匹配 $k(k>1)$，$w(i,j)=1/k$，同样地，不在匹配集 $M_{j(i)}$ 中的控制组个体权重均为 0。

精确匹配可以保证匹配样本与处理组个体协变量完全相同，无论非精确匹配采用哪种匹配方法，所获得的匹配个体与处理组个体往往存在着一定的差异，尽管经过匹配后，这种差异比较小，但可能造成一定的估计偏差。

2.2 LASSO 的理论基础

在传统的最小二乘法中，通过最小化预测值与观测值之间的残差平方和来拟合线性模型。但是，随着变量的数量增加，模型可能会出现过度拟合的情况，导致模型的不稳定和低泛化能力等问题。基于传统的最小二乘法模型拟合过程中可

能出现的问题，Robert T 于（1996）提出了一种 LASSO（Least Absolute Shrinkage and Selection Operator）方法，该方法是一种用于线性回归和特征变量选择的统计学方法，旨在通过特征变量筛选解决回归中出现的过拟合等问题。

LASSO 在最小二乘法中引入了 L_1 正则化项，即将模型参数的绝对值之和添加到目标函数中，这表明除最小化残差平方和外，还应最小化模型参数的绝对值之和。其模型如下式（2.9）：

$$\hat{\beta} = \arg\min_{\beta} \left\{ \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \right\} \tag{2.9}$$

其中，y 是实际观测到目标变量的值；$\sum_{j=1}^{p} x_j \beta_j$ 是模型对目标变量的预测值；λ 是惩罚参数，用于控制惩罚力度；$\sum_{j=1}^{p} \left| \beta_j \right|$ 是 L_1 正则化项，它是所有系数绝对值之和。

LASSO 方法的关键特点是具有自动特征变量选择的能力，通过 L_1 正则化，LASSO 使部分不重要变量的系数参数迭代收缩为 0，从而实现对不重要的变量进行剔除，只保留对目标变量预测具有显著影响的变量。这样可以有效地降低模型复杂度，并提高模型的解释能力和泛化能力。

通过引入 L_1 正则化项，LASSO 能够更准确地控制不同变量的收缩程度，对于与目标变量高度相关的变量，减少对它们的压缩；对于与目标变量相关性较弱的变量，增加对它们的压缩。这样就使得 LASSO 能够适应数据集中的相关性和多重共线性，从而更准确地估计每个变量的系数。

LASSO 方法在许多实际问题中能显著提高预测准确性和模型解释性，还能够准确选择出与目标变量相关的重要特征变量，这使得 LASSO 在高维数据分析、特征选择、基因表达数据、图像处理、信号处理等领域发挥了重要作用。通过控

制正则化参数，LASSO 方法可以在压缩模型参数的同时保持一定的预测准确性，为实际问题的建模和分析提供了可靠选择。

3 基于 LASSO 的 PSM 变量选择

3.1 引言

在科学研究中，处理观测数据的方法十分关键。其中，基于反事实框架下因果效应的研究方法近年来备受关注，如倾向得分匹配（PSM）方法。PSM 方法旨在通过匹配处理组与控制组，来减少因混杂变量而导致的干扰，从而更加精准地估计因果效应。在借助倾向得分匹配等方法进行因果效应估计时，研究人员选择研究所需的协变量会受到多方面因素的影响。因此，需要依据领域知识、统计意义和相关性等因素来进行选择。然而，有时所选取的协变量存在相关性，或者选取的协变量过多，都可能给匹配过程造成支持度差异、维度灾难、多重检验以及数据过拟合等问题，甚至引发多重比较的困境，这些问题都可能导致因果效应的估计不够准确。

于是有必要对所选取的协变量进行“缩简”，即选取具有显著差异组合的变量。目前对 PSM 中变量进行筛选的方法有逐步回归、主成分分析、Bonferroni 校正等。这些方法都存在一定的局限性，逐步回归可能出现过度拟合的问题、主成分分析可能出现信息损失问题以及对结果的解释性较差的问题、Bonferroni 校正可能会忽略协变量之间的相关性问题，因此需要运用一种较为合理的方法对 PSM 中的协变量进行筛选。而 LASSO 方法的关键特点是具有自动特征变量选择的能力，通过 L_1 正则化，LASSO 使不重要变量的系数参数迭代收缩为 0，从而实现对不重要的变量进行剔除，较易筛选出个数较少、且具有显著差异组合的变量，只保留对目标变量预测具有显著影响的变量，极大限度地保留原始信息。这样可以有效地降低模型复杂度，并提高模型的解释能力和泛化能力。LASSO 方

法正好可以弥补上述变量筛选方法的缺陷，这便为 LASSO 应用到 PSM 中提供了可能。

3.2 LASSO-PSM 变量选择

3.2.1 LASSO-PSM 模型

通过对 PSM 方法中变量选取的缺陷和局限性以及 LASSO 方法在变量筛选方面具有的优越性进行讨论，可以在 PSM 中引入 LASSO 方法对变量进行筛选，有望提高 PSM 的匹配质量和因果效果估计的准确性。

传统的 PSM 方法在变量选取过程中存在一些缺陷和局限性。PSM 方法通常依赖于领域专家的主观判断或者基于经验法则选择变量，该方法容易导致遗漏重要的控制变量，从而影响到匹配质量和因果效果的准确性。此外，传统的 PSM 方法往往只考虑单个变量的影响，而忽略了变量之间的相互作用效应，这可能导致对变量的筛选不够精确。相比之下，LASSO 方法通过引入正则化项，可以自动地进行特征变量筛选和参数估计，使其具有很好的特征选择性能。LASSO 方法可以将不重要变量的系数参数迭代收缩为 0，从而实现对变量的自动筛选。LASSO 方法还能够处理变量之间的相互作用效应，以提供更加精确的变量选取。

因此，将 LASSO 方法引入 PSM 中对变量进行筛选，有望克服 PSM 方法的局限性，提高匹配质量和因果效果估计的准确性。通过 LASSO 在 PSM 中的应用，可以更准确地选择相关变量，并消除不必要的变量，从而提高研究的可靠性和准确性。

基于 PSM 模型的信息，对 LASSO 模型修改如下式（3.1）：

$$\hat{\beta}=\arg\min_{\beta}\left\{-\frac{1}{n}\sum_{i=1}^{n}(D_i\log(\frac{e^{\beta_0+\beta x_i}}{1+e^{\beta_0+\beta x_i}})+(1-D_i)\log(1-\frac{e^{\beta_0+\beta x_i}}{1+e^{\beta_0+\beta x_i}}))+\lambda\sum_{j=1}^{m}\left|\beta_j\right|\right\} \quad (3.1)$$

其中，$-\frac{1}{n}\sum_{i=1}^{n}(D_i\log(\frac{e^{\beta_0+\beta x_i}}{1+e^{\beta_0+\beta x_i}})+(1-D_i)\log(1-\frac{e^{\beta_0+\beta x_i}}{1+e^{\beta_0+\beta x_i}}))$ 称为交叉熵损失项，n 是样本量；D_i 表示第 i 个样本属于处理组还是控制组，当第 i 个样本属于处理组时，D_i 的取值为 1，当第 i 个样本属于控制组时，D_i 的取值为 0；p_i 表示第 i 个样本的倾向得分，根据样本数据，由式（2.3）计算得到；λ 是惩罚参数，用于控制惩罚力度；m 是协变量 x 的个数；β_j 是倾向得分模型中协变量系数，也即式（2.3）中的 β 。

当 D_i 的值为 1 时，$D_i\log(p_i)$ 表示第 i 个样本属于处理组且倾向得分模型预测为处理组情况下的损失贡献；当 D_i 的值为 0 时，$(1-D_i)\log(1-p_i)$ 表示第 i 个样本属于控制组且倾向得分模型预测为处理组情况下的损失贡献。交叉熵损失项可以衡量倾向得分模型的预测与实际结果之间的差异，该项的最小化可以使模型的预测概率更接近实际结果。

$\sum_{j=1}^{m}\left|\beta_j\right|$ 称为 L_1 正则化项，L_1 正则化项可使模型选择较少的变量，并倾向于将一些回归系数压缩为零，从而实现特征选择和模型简化。整个损失函数的目标是最小化交叉熵损失项以及限制回归系数的绝对值之和，从而在实现较好拟合的同时，保持模型的简洁性，这有助于提高模型的泛化能力，使其能够更好地适应新的样本，从而提高整体性能。

3.2.2 LASSO—PSM 模型优化算法

根据 3.2.1 中，对传统 LASSO 算法的更改，以及参考 LASSO 算法的具体步骤，提出的 LASSO-PSM 模型的伪代码如下表 3.1。

表 3.1 LASSO-PSM 模型算法

步骤	操作
Step1	初始化：处理数据，创建惩罚参数 λ、系数向量 $coef$ 和活跃集 Q。
Step2	迭代更新：将每个 λ 值代入到对系数向量 $coef$ 迭代中，满足一定条件则停止迭代，再将当前 λ 值以及迭代更新的 $coef$ 其添加到活跃集 Q 中。
Step3	路径选择：通过交叉验证（CV）选择最佳的 λ 值
Step4	输出结果：输出最优 λ 以及对应的模型系数

在初始化步骤中：加载数据并进行标准化：载入需要进行回归分析的数据集，对数据进行预处理，剔除缺失值和处理异常值，对变量进行标准化，以确保每个变量具有零均值和单位方差。这是为了避免不同变量尺度的影响，确保正则化项对各个变量的影响相对均等。创建 λ 值的网格：定义一组 λ 值，λ 值用于控制正则化项的强度。初始化系数向量和活跃集：创建一个长度等于变量个数的零向量作为初始的系数向量 $coef$。创建一个空的活跃集 Q，该集合存储当前被选中的变量的索引。最开始时，活跃集中没有被选中的变量。

在迭代更新步骤中：对模型系数进行循环迭代更新：在每次迭代中，选择一个变量进行更新，直到满足停止准则。采用的停止准则是残差小于某个阈值或迭代次数达到最大迭代次数。对于每个变量，根据当前的系数向量，计算该变量的得分，即将残差与该变量进行相关性计算。得分越大表示该变量对解释残差的贡献更大。根据 LASSO 惩罚项，考虑到该变量得分和 λ 值，使用坐标下降优化算法进行该变量系数的迭代。更新残差向量，将预测值和实际值之间的差异（残差）存储在一个新的残差向量中。当残差向量的数字之和小于某个阈值时停止，迭代结束。最后将当前 λ 数值和迭代更新的 $coef$ 其添加到活跃集 Q 中。

在路径选择步骤中：通过交叉验证（Cross-Validation,CV）选择最佳的λ值，首先，对于每一个λ值，使用其对应的倾向得分拟合模型进行预测。然后，计算每一个λ值对应的平均交叉熵损失函数。最后，根据最小平均交叉熵损失函数选择最优的λ值。

在输出结果步骤中：依据上一步骤中选择最小平均交叉熵损失函数，输出最优λ值及其对应的变量系数。

3.3 LASSO-PSM 匹配结果的平衡性检验

对 LASSO-PSM 模型及其算法伪代码进行详细介绍后，需要将 LASSO-PSM 模型与 PSM 模型匹配结果的平衡性进行对比，PSM 模型匹配结果的平衡性主要体现在其协变量标准化平均值差异、协变量对数标准差比以及倾向得分标准化平均值差异这三方面，于是将对比 LASSO-PSM 模型与 PSM 模型在协变量标准化平均值差异、协变量对数标准差比以及倾向得分标准化平均值差异这三方面平衡性的差异，以验证 LASSO-PSM 模型匹配结果的平衡性优于 PSM 模型匹配结果的平衡性，LASSO-PSM 模型相较于 PSM 模型确实能提高匹配质量和因果效果估计的准确性。

3.3.1 数据来源及介绍

为了得到 LASSO-PSM 模型匹配结果的平衡性与 PSM 模型匹配结果的平衡性差异，还需要一定的数据才能进行模型匹配结果的平衡性比较。在选取数据时，需要考虑到 PSM 和 LASSO 所适用的数据类型不同。

PSM 适用于观察研究下的大样本数据，且研究者无法对样本进行随机分组。此时通过大量样本计算得出倾向得分以确保相近倾向得分的样本匹配成功，以实

现处理组和控制组之间的比较。如果数据集过小，可能很难找到合适的匹配对。PSM 模型可以应用于各种类型的数据，包括但不限于横断面数据、纵向数据、面板数据、调查数据等，它在医学研究、社会科学研究和经济学研究等领域都有广泛的应用。需要注意 PSM 模型使用的是二分类数据，即处理组和对照组只有两种情况，因此需要选择符合这种特点的数据集，并且需要保证数据集中有足够的样本数量，以确保匹配结果的可靠性。此外，在使用 PSM 方法时，数据应当包含用于计算倾向得分的相关变量，这些变量应当能够捕捉到影响处理组分配的潜在混杂因素。因此，数据的质量和可用性对于 PSM 方法的有效应用非常重要。

LASSO 方法适用于高维数据、需要进行特征选择、存在多重共线性或追求模型可解释性等情况，LASSO 通过加入 L_1 正则化项，可以有效地降低共线性对模型估计的影响，并提供更稳定和可解释的结果，即从大量特征变量中选取最相关或最重要的变量。通过对系数进行稀疏化，LASSO 可以将不相关或冗余的变量的系数收缩至 0，从而实现对特征变量的筛选，使得仅保留一部分重要的特征变量，并将其他变量的系数缩减为 0。这种特性使得模型不仅更加简洁，而且具有可解释性，并有助于识别最重要的预测因素，该优势使 LASSO 在统计学、机器学习和数据分析等领域都有广泛应用。由于 LASSO-PSM 模型在变量选择过程中采用了 LASSO 回归方法，因此需要保证数据集中具有多个变量，且变量之间存在一定的线性关系，以便 LASSO 能够有效地筛选出影响处理结果的关键变量。

进行实验的数据集需要满足 PSM、LASSO 和 LASSO-PSM 的适用性，由于生成的模拟数据可能缺乏一定的真实性，于是在这里选取 Kaggle 官网（https://www.kaggle.com/）上公开数据集——心力衰竭预测数据集（Heart Failure Prediction Dataset，简称为 HFPD 数据集）进行模型匹配结果的平衡性对比。

HFPD 数据集是由匈牙利心脏病学研究所的 Andras 医学博士、瑞士苏黎世大学医院的 William 医学博士、瑞士巴塞尔大学医院的 Matthias 医学博士和弗吉尼亚州医疗中心、长滩和克利夫兰诊所基金会的 Robert 医学博士合作创建，通过组合已经独立可用，但以前未组合的不同数据集创建的。在这个数据集中，结合了 5 个独立可用的心脏数据集的 11 个共同特征，其中来自克利夫兰的 303 份观测样本、匈牙利的 294 份观测样本、瑞士的 123 份观测样本、弗吉尼亚州长滩的 200 份观测样本和 Stalog 数据集的 270 份观测样本，总计 1190 份观测样本，剔除重复的 272 个观测样本，该数据集共有 918 个观测样本，使其成为迄今为止可用于研究心脏病的最大数据集。

HFPD 数据集的具体变量相关信息如下表 3.2：

表 3.2 模数据集变量的相关信息

变量名称	变量	均值	最小值	最大值	标准差
年龄（x_1）	患者年龄（岁）	53.51	28	77	9.43
性别（x_2）	男：x_2=1；女：x_2=0	0.79	0	1	0.41
胸痛类型（x_3）	典型心绞痛：x_3=1 非典型心绞痛：x_3=2 非心绞痛：x_3=3 无症状：x_3=4	3.25	1	4	0.93
静息血压（x_4）	静息血压（mm/Hg）	132.40	68.23	200	18.51
胆固醇（x_5）	血清胆固醇（mm/dl）	198.80	89.72	603	109.38

续表 3.2

变量名称	变量	均值	最小值	最大值	标准差
BS（x_6）	空腹血糖＞120mg/dl：x6=1 空腹血糖≤120mg/dl：x6=0	0.23	0	1	0.42
静息心电图（x_7）	正常：x_7=1 非正常：x_7=0	0.81	0	1	0.40
MaxHR（x_8）	最大心率	136.81	60	202	25.46
运动性心绞痛（x_9）	具有运动诱发性心绞痛：x_9=0 没有运动诱发性心绞痛：x_9=0	0.60	0	1	0.49
Oldpeak（x_{10}）	以压抑为单位测量的数值	0.89	-2.6	6.2	1.07
ST_Slope（x_{11}）	斜率＞0：x_{11}=1 斜率＝0：x_{11}=2 斜率＜0：x_{11}=3	1.64	1	3	0.61
HeartDisease（D）	心脏病：D=1 正常：D=0	0.55	0	1	0.50

3.3.2 匹配结果的平衡性检验

为了精确地检验 LASSO-PSM 模型匹配结果的平衡性，需要与 PSM 模型匹配结果的平衡性进行比较。匹配结果的平衡性主要通过协变量标准化平均值差异、协变量对数标准差比和倾向得分标准化平均值差异来体现。先计算协变量间的皮尔逊相关系数矩阵，发现部分协变量间呈现中度相关性，再利用 LASSO-PSM 模型，也即式（3.1）对 HFPD 数据集的协变量进行筛选，剔除了年龄（x_1）性别（x_2）运动性心绞痛（x_9）这三个变量。

再利用式（2.4）、（2.5）、（2.6）和（2.7）计算出 LASSO-PSM 模型和 PSM 模型的协变量标准化平均值差异、协变量对数标准差比和倾向得分标准化平均值差异，在比较两模型匹配结果的平衡性时，需要对样本进行配对，样本配对时采用最近邻匹配，即采取一对一样本可重复配对方式。LASSO-PSM 模型和 PSM 模型匹配结果平衡性的数值如下表 3.3。

表 3.3 模型对比结果

模型	协变量标准化平均值差异	协变量对数标准差比	倾向得分标准化平均值差异
PSM	0.1548	0.1236	0.0721
LASSO-PSM	0.1462	0.0851	0.0137
提升百分比	5.58%	31.14%	80.95%

由表 3.3 可看出，当数据需要进行特征变量筛选时，LASSO-PSM 模型相较于 PSM 模型匹配结果的平衡性得到了较大幅度的提升。在协变量标准化平均值差异上，LASSO-PSM 模型相较于 PSM 模型的匹配结果平衡性提高了 5.58%；在协变量对数标准差比差异上，LASSO-PSM 模型相较于 PSM 模型的匹配结果平衡性提高了 31.14%；在倾向得分标准化平均值差异上，LASSO-PSM 模型相较于 PSM 模型的匹配结果平衡性提高了 80.95%。

为了更加直观的看出 LASSO-PSM 模型与 PSM 模型的匹配结果的平衡性差异，画出 LASSO-PSM 模型和 PSM 模型匹配结果的倾向得分分布直方图，见图 3.1 和图 3.2。

图 3.1 PSM 模型倾向得分分布直方图

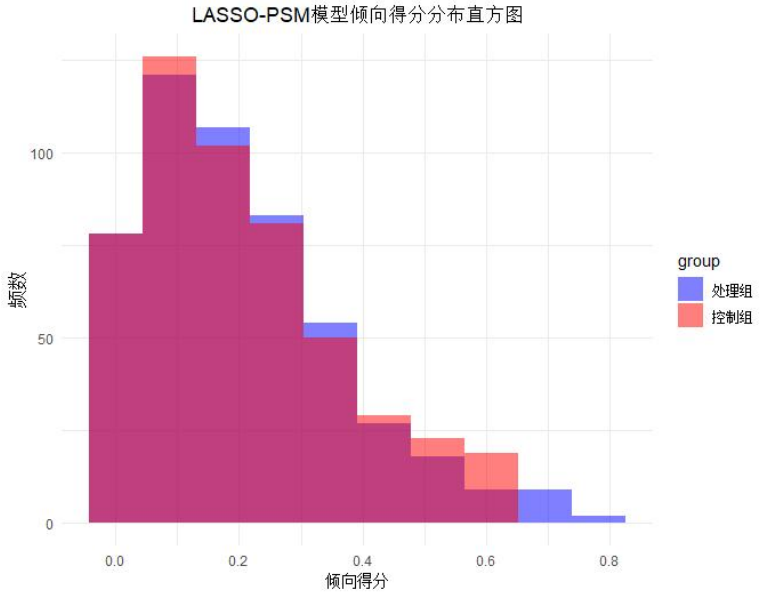

图 3.2 LASSO-PSM 模型倾向得分分布直方图

如图 3.1 和图 3.2 所示，LASSO-PSM 模型和 PSM 模型的倾向得分分布直方图大概都呈右偏分布，且 PSM 模型倾向得分分布直方图中处理组和控制组重叠的面积大致小于 LASSO-PSM 模型倾向得分分布直方图中处理组和控制组重叠的面积。这说明 LASSO-PSM 模型配对样本量更多，匹配结果的平衡性更好，为了精确比较两种模型的匹配结果的平衡性差异，画出其倾向得分概率密度分布图，见图 3.3 和 3.4。

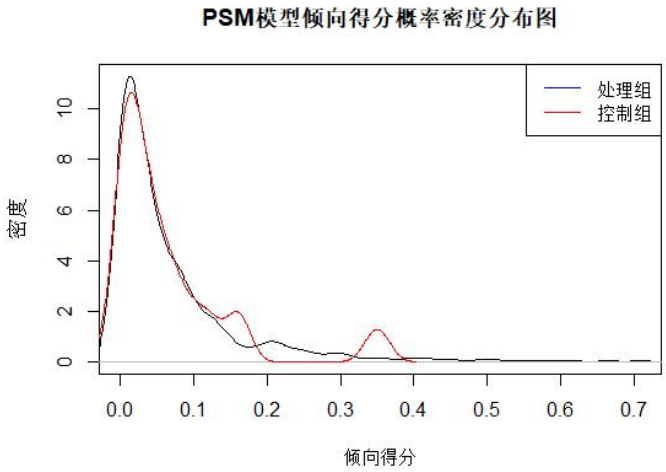

图 3.3 PSM 模型倾向得分概率密度分布图

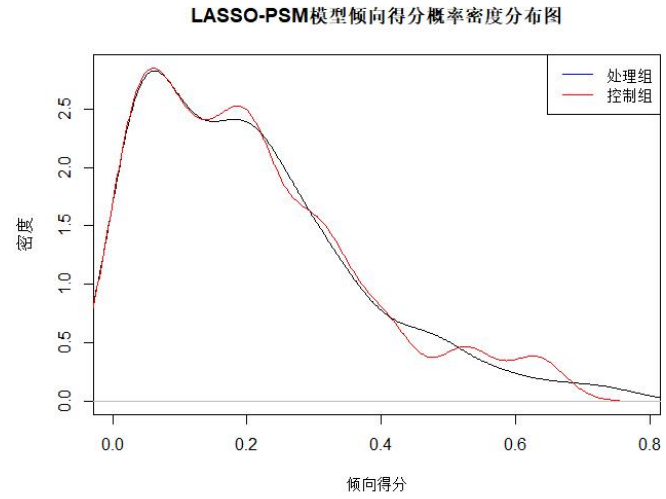

图 3.4 LASSO-PSM 模型倾向得分概率密度分布图

如图 3.3 和图 3.4 所示，PSM 模型和 LASSO-PSM 模型的倾向得分概率密度分布图大概都呈右偏分布。经计算，PSM 模型倾向得分概率密度分布图中，处理组和控制组重叠面积为 0.9215；LASSO-PSM 模型倾向得分概率密度分布图中处理组和控制组重叠面积为 0.9467，LASSO-PSM 模型相较于 PSM 模型模型的重叠的面积提升了 2.73%。这说明 LASSO-PSM 模型匹配效果更好，匹配结果的平衡性更好。也可以使用 QQ 图比较两种模型的匹配效果效果以检验匹配结果的平衡性差异，画出其模型倾向得分 QQ 图，见图 3.5 和图 3.6。

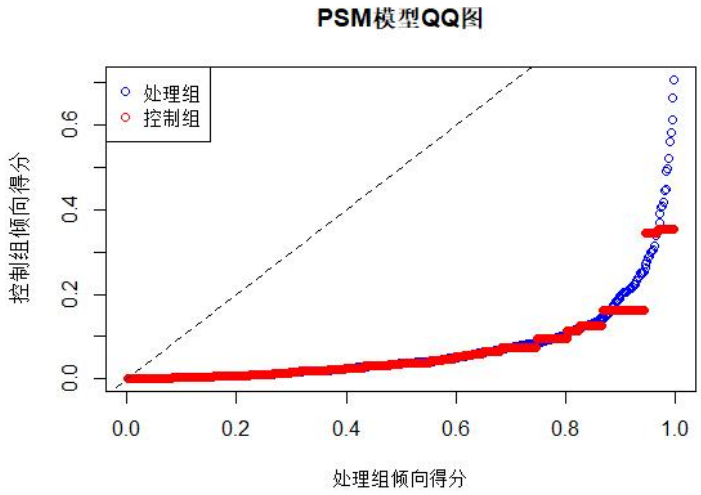

图 3.5 PSM 模型倾向得分 QQ 图

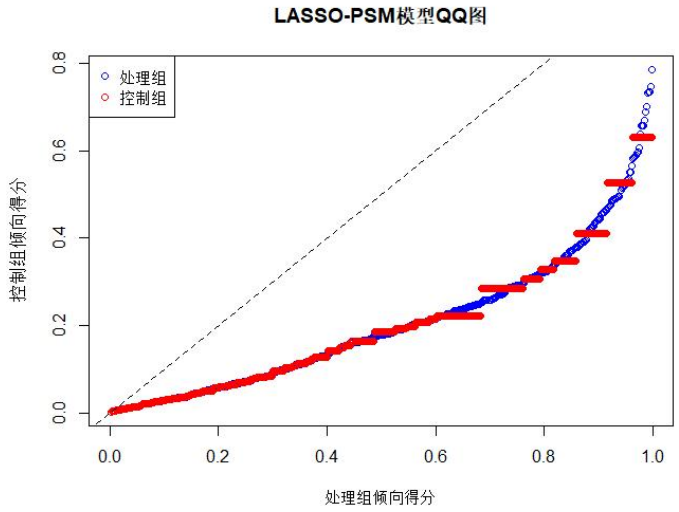

图 3.6 LASSO-PSM 模型倾向得分 QQ 图

如图 3.5 和图 3.6 所示，PSM 模型倾向得分 QQ 图与 45 度线的偏离程度较大，皮尔逊相关系数为 0.9543；LASSO-PSM 模型倾向得分 QQ 图与 45 度线的偏离程度较小，皮尔逊相关系数为 0.9953。说明 PSM 模型中处理组和控制组的协变量差异较大，LASSO-PSM 模型中处理组和控制组的协变量差异较小，LASSO-PSM 模型比 PSM 模型的匹配结果更平衡，模型的稳定性更好。也可以使用协变量标准化平均值差异变化图直观地观察匹配结果的平衡性，见图 3.7。

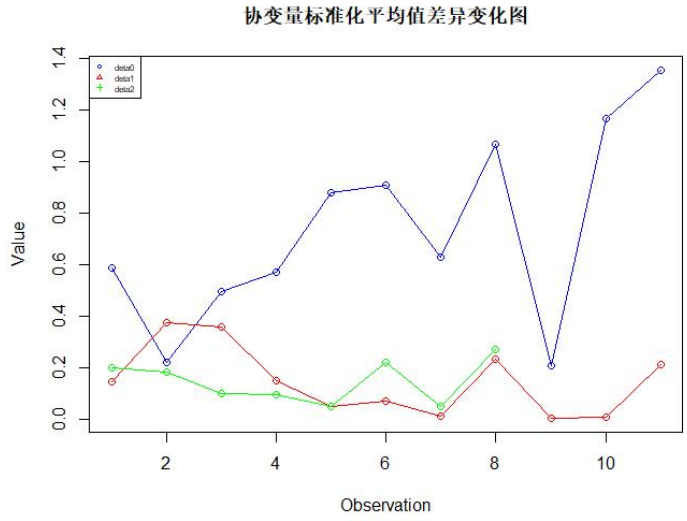

图 3.7 协变量标准化平均值差异变化图

如图 3.7 所示，蓝色实线表示配对前处理组和控制组的协变量标准化平均值差异变化，红色实线表示使用 PSM 模型进行配对后处理组和控制组的协变量标准化平均值差异变化，绿色实线表示使用 LASSO-PSM 模型进行配对后处理组和控制组的协变量标准化平均值差异变化。从图中可以看出，配对前处理组和控制组的协变量标准化平均值差异大于 PSM 模型和 LASSO-PSM 模型的协变量标准化平均值差异，LASSO-PSM 模型的协变量标准化平均值差异略小于 PSM 模型的协变量标准化平均值差异。协变量标准化平均值差异变化图也说明了 LASSO-PSM 模型比 PSM 模型的匹配结果平衡性更好。

3.4 小结

通过对 PSM 方法中变量选取的缺陷和局限性进行讨论，可以发现 LASSO 方法在变量筛选方面具有优越性，为提高 PSM 匹配质量提供可能。因此，根据 PSM 相关原理对 LASSO 的损失函数进行修改，再利用 HFPD 数据集，分析对比 LASSO-PSM 模型和 PSM 模型匹配结果的平衡性差异，发现面对需要进行变量

筛选的数据时，PSM 模型因为变量存在共线性而导致其模型匹配结果的平衡性较差。此时，LASSO-PSM 模型在变量中筛选出个数较少、且具有显著差异组合的变量组合，从而提高 PSM 模型匹配结果的平衡性，增强模型的稳定性。

4 基于 LASSO-PSM 劳动者择业偏好影响因素分析

4.1 引言

随着近年来高校毕业人数居高不下，以及受近三年新型冠状病毒疫情的影响，社会就业压力越来越大，以党政机关公务员、事业单位等为主的“铁饭碗”工作受到越来越多求职者的青睐，成为了求职者心目中的“香饽饽”。究其原因，“铁饭碗”类型工作被公认为收入稳定，且不易失业。但“铁饭碗”类型工作和其他类型工作对劳动者健康状况、心理状况、幸福感和经济收入的影响是否存在显著差异？“铁饭碗”类型工作是否一定是求职者眼中的“香饽饽”？采用 CGSS 调查数据，基于 LASSO-PSM 模型进行反事实检验，通过对比分析“铁饭碗”类型工作和其他类型工作对劳动者的健康状况、心理状况、幸福感和经济收入影响的差异，以期剖析两类型工作的优势与不足，为指导当前大学生就业，转变求职者就业思想提供实证参考。

基于 CGSS 调查数据，其中调查的工作类型有“党政机关”、“事业单位”、“军队”、“企业”、“无单位/自雇（包括个体户）”、“社会团体”和“其他”这七类。为了便于利用 LASSO-PSM 模型进行反事实检验，研究中将七类工作类型分为两大类：工作类型为“党政机关”、“事业单位”和“军队”的定义为“铁饭碗”类型工作；工作类型为“企业”、“无单位/自雇（包括个体户）”、“社会团体”和“其他”的定义为其他类型工作。

由于“铁饭碗”与其他类型工作性质不同，决定两种类型工作的时间与压力存在差异，因此大多数研究也从工作时间和压力角度探究其对劳动者健康状况、心理状况、幸福感和经济收入的影响。

从工作时间角度出发，任国强（2023）利用Regression方法和U检验方法，验证了工作时间对劳动者的健康状况和心理状况有重要的影响，呈倒“U”型关系，即在过长或过短的工作时长下，劳动者健康状况和心理状况较差。马红梅（2022）从工作时长角度出发，采用Grossman健康生产函数框架、Probit模型和中介效应模型，引入工作时间弹性这一调节变量，详细地研究工作时间长度与劳动者健康状况之间的关系，经研究发现，工作时间长度对劳动者健康状况和心理状况的影响呈倒“U”型关系，同时也发现该影响存在职业类型异质性。程蒙（2021）以社会比较理论、需求层次理论和资源保存理论为基础，利用有序Logit模型探究了职业幸福感的形成以及工作时间如何影响劳动者的职业幸福感，研究发现工作时间对劳动者职业幸福感的影响呈倒“U”型关系。司健敏（2018）通过问卷调查分析得出工作时间的长短与员工的工作绩效也呈现倒“U”型关系。总结发现，工作时间对劳动者健康状况、心理状况、幸福感和经济收入的影响都呈现倒“U”型关系。

从压力角度出发，两种类型的工作性质使劳动者面临的压力存在差异，住房条件、婚姻状况和经济状况都对劳动者的工作压力和生活压力存在影响。李金鑫（2017）通过问卷调查，选用回归分析方法进行研究，发现压力对健康状况和情绪存在惩罚效应。人们的幸福感与压力也息息相关，廖君瑶（2021）通过问卷调查研究，采用潜调节结构方程（LMS）研究工作压力对员工幸福感的影响，发现工作压力对员工幸福感存在惩罚效应。随着调查工作研究的深入，魏祥迁（2012）又把公司压力细分成家庭压力和工作压力，并对公司中员工开展调查研究分析各种特性人员的压力状况和健康状况，发现家庭压力和工作压力对身心健康状况的影响均存在惩罚效应，且已婚人员的压力高于未婚人员的压力，年青人员和基层

管理人员整体健康状况最差。陈春明（2006）研究发现锻炼会对生理和心理上的健康状况产生正向效应。程蒙（2021）研究发现社交可缓解心理压力，调节心理健康，心理健康又对身体健康、工作和幸福感有着至关重要的影响。除此之外，婚姻作为人类社会最重要的社会关系之一，对主观幸福感有着重要的影响，张云桥（2017）利用2013年中国综合调查（CGSS）数据，采用多分类logit模型研究个体的主观幸福感与婚姻的关系，探究了我国处于不同婚姻状况的个体主观幸福感差异，结果发现婚姻对主观幸福感有显著影响，处于在婚状态个体比处于非在婚状态个体的主观幸福感高。Drentea（2012）研究经济状况对老年人心理状况的影响，发现经济状况与老年人心理状况呈正相关。Luchman（2013）研究了工作环境与员工健康和工作绩效的关系，经大量研究综合分析，发现工作压力对健康和工作绩效存在惩罚效应，依此对工作设计和组织管理提出相关建议。Cooper（2012）研究了影响压力的因素、压力对身体和心理健康的影响以及如何管理和减轻压力。

尽管学者们从不同视角研究了工作时间和工作压力对劳动者健康状况、心理状况、幸福感和工作绩效的影响，但并未深入研究工作类型对劳动者健康状况、心理状况、幸福感和经济收入的影响以及其因果效应差异，并且不同工作类型对劳动者健康状况、心理状况、幸福感和经济收入影响的相关研究尚显不足。

4.2 理论假说

从Becker（2009）所提出的技能—工资不平等理论角度来看，一些“铁饭碗”类型工作通常出现在传统行业或公共部门，这些行业受到政府管制、市场竞争程度低等因素的影响，导致整体薪酬水平较低。与之相对应的一些其他类型的工作

出现在高科技、金融等领域，由于技能要求高、市场需求大，所以薪酬相对较高。从亚当·斯密在《国富论》中提出的劳动力供求关系来看，在某些行业或职业中，由于供给过剩或需求不足的情况，工资水平可能偏低。一些公共部门岗位会因为劳动力市场供给过剩而导致工资水平较低。最后从工作福利和稳定性角度来看，一些“铁饭碗”类型工作会提供稳定的福利待遇，如医疗保险、退休金等，以较高的福利弥补薪酬的不足。而其他类型的工作则更加注重薪酬本身，对福利和稳定性的要求相对较低。

总的来说，“铁饭碗”类型工作相对其他类型工作收入可能较低是由于行业性质、劳动力供求关系以及福利和稳定性权衡等因素的综合影响。这些因素使得“铁饭碗”类型工作薪酬相对较低，但在其他方面可能具有稳定性和福利待遇等优势，薪资水平又对劳动者的健康状况、心理状况和幸福感也存着相应的影响。因此不同类型工作对劳动者的健康状况、心理状况和幸福感的影响是否存在显著差异有待检验，因此提出以下假说一。

假说一：不同类型工作对劳动者经济收入方面的影响存在显著差异，“铁饭碗”类型工作比其他类型工作的平均经济收入水平低。

随着职业晋升，首先，劳动者的社会认可和自我价值感提升，因为工作职位晋升意味着在组织中拥有更高的地位和更大的权力，增强个体的社会认可和自我价值感，这种认可和价值感提升有助于改善个体的心理状况，促使其生活更加满意和幸福。其次，经济收入增加，由于工作职位的晋升伴随着薪资和福利的提升，继而增加个体的经济收入，经济收入的增加有助于改善个体的生活条件，提供更多的经济资源和机会，有助于满足物质需求和实现个人目标，进而促进幸福感的提升。除此之外，工作资源和控制权增加，因为随着职位晋升，个体会获得更多

的工作资源和控制权，更大的决策权、更高的自主性和灵活性，这种资源和控制权的增加可以提供更好的工作环境和条件，减少工作压力和冲突，有利于个体心理健康状况的改善。最后，获得职业发展和成长机会，由于职位晋升通常意味着更高级别的工作职责和更广阔的发展空间，个体在新的职位中面临更多的挑战和机遇，可以提升自己的专业技能、扩大人际关系网络，并获得更多的职业发展和成长机会，这种职业发展和成长的经历可以提升个体的自我实现感和满足感，从而促进幸福感和心理状况的改善。由以上论述提出假说二。

假说二：随着工作职位晋升，对劳动者健康状况、心理状况、幸福感和经济收入会产生正向效应。

晋升后提升了自我社会地位和职业认知，在社会学理论的视角下，一些“铁饭碗”类型工作可能受到职业认知和社会地位的影响，从而影响了其薪酬水平。这可能导致这些工作的工资水平相对较低，即使在晋升后也难以获得较高的薪酬。晋升后工作的技能要求和劳动强度也存在差异，从经济学的角度来看，一些“铁饭碗”类型工作可能对技能要求较低，劳动强度较大，导致相对较低的薪酬水平。相比之下，一些其他类型的工作可能需要更高级的技能或者承担更多的责任，因此在晋升后能够获得较高的薪酬。由以上论述提出假说三。

假说三：随着工作职位晋升，不同类型工作对劳动者的健康状况、心理状况和幸福感的影响不存在显著差异，但在经济收入方面存在显著差异，“铁饭碗”类型工作比其他类型工作的平均经济收入水平低。

4.3 数据与变量

4.3.1 数据来源

CGSS（China General Social Survey，数据来源网址：http://cgss.ruc.edu.cn/）数据库是我国一项重要社会调查项目，它由中国社会科学院社会学研究所主导，旨在了解和研究中国社会的现状、变迁和趋势。CGSS 数据库的数据来自于中国社会科学院社会学研究所组织的全国范围内的大规模社会调查。调查每年都会进行，涵盖多个领域和主题，包括人口与家庭、教育、就业、收入、健康、政治参与、价值观念等。CGSS 数据库包含大量的原始数据、问卷调查数据、访谈数据等。这些数据记录被调查者的背景信息（如性别、年龄、教育程度等）、生活状况、意见和态度等多个方面的内容。同时，CGSS 数据库还提供一系列衍生变量和整理过的数据集，方便研究者进行分析和研究。CGSS 数据库的样本覆盖全国各个地区和不同群体，具有较高的代表性和可比性。通过长期的追踪调查，CGSS 数据库也提供一些时间序列数据，可以用于观察和分析社会变迁和趋势。CGSS 数据库是中国社会科学领域的重要资源，它提供丰富的数据支持，帮助研究者深入了解中国社会的各个方面和层面。因此，实证部分的数据将采用 CGSS 数据库。

研究数据来源于 CGSS 数据库里 2010 年、2015 年、2017 年这三年的数据集，其他年份数据集或多或少缺少部分研究所需变量。

4.3.2 LASSO-PSM 模型的变量选取

首先，对本文实证研究的结果变量进行选取。选取劳动者的自我健康评分、自我心理压力评分、自我幸福感评分和年收入来度量劳动者的健康状况、心理状况、幸福感和经济收入，并作为研究的结果变量 Y。这四个方面是反应不同工作

类型对劳动者生活影响的最直接体现。

数据采用的样本为多期 CGSS 调查数据，在进行样本配对计算因果效应时，针对不同调查年份收入的可比性问题，主要采取加入调查年份时间虚拟变量的方式以实现对经济收入的调整。此处，设定 2010 年为基期，将其余调查年份的个体经济收入根据其所处年份全国人均收入与 2010 年对应全国人均收入的比值进行缩放，从而消除汇总体样本后收入分布的年度差异。

然后，对本文实证研究的处理变量进行选取。“铁饭碗”和其他类型工作主要区别在于：工作和收入是否稳定。根据数据中劳动者工作单位类型，工作单位类型为“党政机关”、“事业单位”和“军队”这三种类型的定义为“铁饭碗”类型工作，即处理变量 $D=1$；工作单位类型为“企业”、“社会团体”、“无单位/自雇（包括个体户）”和在“其他”这三种类型的定义为其他类型工作，即处理变量 $D=0$。

最后，对本文 LASSO-PSM 模型的协变量进行选取。当劳动者在劳动市场选择工作时，并不单单由劳动者选择所决定，它是由劳动力市场和劳动者双方的双向选择所决定。在这个双向选择中，劳动者的受教育程度起着决定性作用，这决定着劳动者能够有机会选择何种工作。去除受教育程度的影响，劳动者自身因素和原生家庭环境因素也对劳动者工作选择存在影响。依据劳动者自身因素和原生家庭环境因素选择 LASSO-PSM 模型的协变量。由相关研究文献可知，自身因素中，年龄、人均住房面积、体重指数 BMI、锻炼频率、社交频率、工作时长、家庭经济状况、婚姻状况以及个体受教育程度均对劳动者工作的选择存在影响。原生家庭环境因素中，一方面，经济条件决定后代的所见所闻；另一方面，父母所处的时代不同，接受的思想不同，对后代耳濡目染的思想也有所不同。因此这

两方面也对劳动者的工作选择有着一定的影响。故选取年龄、人均住房面积、体重指数 BMI、锻炼频率、社交频率、工作时长、家庭经济状况、婚姻状况、个体受教育程度、父母出生年代、14 岁时家庭社会地位这 12 个变量为 LASSO-PSM 模型的协变量 X。

4.3.3 变量处理及 LASSO-PSM 模型的变量选择

选取中国综合社会调查 2010 年、2015 年、2017 年这三年的数据集作为分析样本。利用 LASSO-PSM 模型对变量进行筛选，需要对变量进行相应的处理，其中，结果变量（Y）和协变量（X）及其相应的处理方式如下表 4.1：

表 4.1 变量处理

变量		处理方式
结果变量 Y	健康状况（y_1）	“很不健康”、“比较不健康”、“一般”、“比较健康”和“很健康”分别设定为 1－5 之间的整数
	心理状况（y_2）	将心情“总是”、“经常”、“有时”、“很少”和“从不”抑郁或沮丧程度分别设定为 1－5 之间的整数
	幸福感（y_3）	“非常不幸福”、“比较不幸福”、“介于幸福与不幸福之间”、“比较幸福”和“非常幸福”分别设定为 1－5 之间的整数
	经济收入（y_4）	对劳动者的经济收入取自然对数
协变量 X	年龄（x_1）	调查年份减去出生年份
	人均住房面积（x_2）	住房面积除以居住人数
	体重指数 BMI（x_3）	体重（kg）除以身高（m）的二次方得到体重指数 BMI，将 BMI 数值在正常范围 18.5～23.9 之间的设定为 1，其他数值设定为 0
	锻炼频率（x_4）	“从不”、“一年数次或更少”、“一月数次”、“一周数次”和“每天”分别设定为 1－5 之间的整数
	社交频率（x_5）	“从不”、“很少”、“有时”、“经常”和“总是”分别设定为 1－5 之间的整数
	工作时长（x_6）	个体平均每周工作小时数
	家庭经济状况（x_7）	“远低于平均水平”、“低于平均水平”、“平均水平”、“高于平均水平”和“远高于平均水平”分别设定为 1－5 之间的整数
	婚姻状况（x_8）	非在婚状态（同居、未婚、离婚和丧偶）和在婚状态（初婚有配偶、再婚有配偶和分居未离婚）分别设定为 0 和 1

续表 4.1

变量	处理方式
受教育程度（x_9）	“小学及以下”、“初中”、“高中及中专”、“大专及以上”分别设定为 1—4 之间的整数
原生家庭经济地位（x_{10}）	最底层到最顶层分别设定为 1—10 之间的整数，分为十个等级
父亲出生年代（x_{11}）	出生在 1949 年之前、1950-1959 年、1960-1969 年、1970-1979 年和 1980-1989 年分别设定为 1—5
母亲出生年代（x_{12}）	

在此需要说明的是，由于年龄超过六十岁的样本大多已经退出了劳动力市场，未成年的样本还未进入劳动力市场，为了保证检验结果的可靠性，因此，需要剔除年龄超过 60 岁和未成年的样本。总样本容量为 35333 个，剔除年龄不符合要求和变量存在缺失值的样本后，得到有效样本单元 10490 个，其中处理组样本单元 2058 个，控制组样本单元 8432 个。

使用所提出的 LASSO-PSM 模型进行协变量选择，最终的 LASSO-PSM 模型中剔除的变量有人均住房面积（x_1）、体重指数 BMI（x_3）、社交频率（x_5）、家庭经济状况（x_7）、原生家庭经济地位（x_{10}）、父亲出生年代（x_{11}）。再对比 LASSO-PSM 模型和 PSM 模型在协变量标准化平均值差异、协变量对数标准差比、倾向得分标准化平均值差异这三个指标上的差异，其结果如下表 4.2。

表 4.2 模型匹配平衡性对比结果

模型	协变量标准化平均值差异	协变量对数标准差比	倾向指数标准化平均值差异
PSM	0.0810	0.0840	5.3759×10^{-4}
LASSO-PSM	0.0649	0.0383	4.5956×10^{-5}
提升百分比	19.88%	54.40%	91.45%

由表 4.2 可看出，LASSO-PSM 模型较于 PSM 模型的匹配结果的平衡性得到了较大幅度的提升。在协变量标准化平均值差异上，LASSO-PSM 模型相较于 PSM 模型的匹配结果平衡性提高了 19.88%；在协变量对数标准差比差异上，

LASSO-PSM 模型相较于 PSM 模型的匹配结果平衡性提高了 54.40%；在倾向指数的标准化平均值差异上，LASSO-PSM 模型相较于 PSM 模型的匹配结果平衡性提高了 91.45%。可见在此次实证过程中，使用 LASSO-PSM 模型相较于 PSM 模型计算得到的因果效应更加可靠。

4.4 工作类型偏好与工作选择因素的分析

4.4.1 LASSO-PSM 模型估计结果

为了直观看出协变量 X 对劳动者选择职业的影响程度，计算出 LASSO-PSM 模型的估计结果和协变量的平均边际效应（dy/dx），结果如下表 4.3。

表 4.3 LASSO-PSM 模型的估计结果和协变量的平均边际效应

D	Coef.	dy/dx
年龄（x_1）	0.0160***	0.0037***
锻炼频率（x_4）	0.0439***	0.0102***
工作时长（x_6）	-0.0089***	-0.0021***
婚姻状况（x_8）	0.1512***	-0.0350***
受教育程度（x_9）	0.1657***	0.0384***
母亲出生年代（x_{12}）	-0.1357***	-0.0314***
constant	-2.5025***	

注：*、**、***分别表示在 0.1、0.05、0.01 显著性水平下显著。下表同。

根据表 4.3 中 LASSO-PSM 模型的估计结果，constant 为常数项，Coef.为系数，协变量系数全部在 0.01 显著性水平下显著。从个人特征来看，劳动者受教育程度的平均边际效应（dy/dx）数值为 0.0384，数值为正，表明协变量 x_9 对倾向得分存在正向影响，意味着劳动者学历越高，选择“铁饭碗”这类型工作的可能性越大，其平均边际效应绝对值最大，表明其对劳动者选择“铁饭碗”这类型

工作的影响程度最明显。从家庭背景特征来看，劳动者母亲的出生年代平均边际效应（dy/dx）数值为-0.0314，数值为负，表明协变量x_{12}对倾向得分存在负向影响，意味着，劳动者母亲的出生年代越晚，劳动者选择“铁饭碗”这类型工作的可能性越低。所有变量中，劳动者年龄的平均边际效应（dy/dx）数值为0.0037，数值为正，表明协变量x_1对倾向得分存在正向影响，意味着，劳动者的出年龄越大，劳动者选择“铁饭碗”这类型工作的可能性越大，但其平均边际效应绝对值最小，表明其对劳动者选择“铁饭碗”这类型工作的影响程度最弱。

4.4.2 LASSO-PSM模型的因果效应估计

根据LASSO-PSM模型计算工作类型对劳动者健康状况、心理状况、幸福感和经济收入影响的平均因果效应，如下表4.4。

表4.4 工作类型对劳动者生活状况影响的平均因果效应（ATT）

结果变量	结果信息	数值
健康状况（y_1）	处理组均值	4.0394
	控制组均值	3.9996
	ATT（1vs0）	0.0398
	t统计量	1.1200
	P值	0.2226
心理状况（y_2）	处理组均值	4.0612
	控制组均值	4.0986
	ATT（1vs0）	-0.0374
	t统计量	0.0900
	P值	0.2715
幸福感（y_3）	处理组均值	4.0077
	控制组均值	3.9531
	ATT（1vs0）	0.0546**
	t统计量	-0.2500
	P值	0.0467
经济收入（y_4）	处理组均值	10.4401
	控制组均值	10.6538
	ATT（1vs0）	-0.2137***
	t统计量	-5.6300
	P值	0.0017

在表 4.4 中，从工作类型对劳动者生活影响的平均因果效应的结果中，可看出，因果效应中只有幸福感和经济收入在存在显著差异，“铁饭碗”类型工作经济收入显著小于其他类型工作经济收入，但“铁饭碗”类型工作的幸福感显著高于其他类型工作的幸福感，健康状况和心理状况的平均因果效应不存在显著差异。由此假说一得证，且“铁饭碗”类型工作的幸福感显著高于其他类型工作的幸福感。总体上“铁饭碗”类型工作与其他类型工作对劳动者在健康状况和心理状况两个方面影响不存在显著差异。由此可见，“铁饭碗”类型工作和其他类型工作的性质存在一定差异。工作职位晋升，对劳动者健康状况、心理状况、幸福感和经济收入的影响程度也会有所不同。这时候工作职位晋升又如何影响两类型工作劳动者的健康状况、心理状况、幸福感和经济收入呢？将样本分组，分为工作无职务一组（j=1，工作级别为无级别），工作有职务一组（j=2，工作级别为股级及以上），样本的特征分布如下表 4.5，其平均因果效应（ATT）结果如下表 4.6。

表 4.5 样本特征分布

	工作无职务（j=1）	工作有职务（j=2）	累计频率
其他类型工作（$D=0$）	62.08%	18.30%	80.38%
“铁饭碗”类型工作（$D=1$）	12.74%	6.88%	19.62%
累计频率	74.82%	25.18%	100%

根据表 4.5 可看出，“铁饭碗”类型工作的人数约占工作人数总体的五分之一，工作有职务的人数约占工作人数总体的四分之一。根据表 4.6 工作类型对有无工作职务劳动者生活影响的平均因果效应的结果可看出，工作有职务时，劳动者的健康状况、心理状况、幸福感和经济收入的潜在结果均值均高于工作无职务时的潜在结果均值，即工作职位晋升后，对劳动者健康状况、心理状况、幸福感和经济收入产生正向效应，由此假说二得证。

表 4.6 工作类型对有无工作职务劳动者生活状况影响的平均因果效应（ATT）

结果变量	结果信息	工作无职务（j=1）	工作有职务（j=2）
健康状况（y_1）	处理组均值	4.0142	4.0859
	控制组均值	3.9566	4.0235
	ATT（1vs0）	0.0576	0.0624
	t 统计量	1.43	1.1100
	P 值	0.1530	0.2674
心理状况（y_2）	处理组均值	4.0382	4.1039
	控制组均值	3.9925	4.1662
	ATT（1vs0）	0.0457	-0.0623
	t 统计量	1.10	-1.1000
	P 值	0.2715	0.2717
幸福感（y_3）	处理组均值	3.9535	4.1080
	控制组均值	3.8705	4.0646
	ATT（1vs0）	0.0830	0.0434
	t 统计量	0.75	0.98
	P 值	0.4534	0.3274
经济收入（y_4）	处理组均值	10.2528	10.7868
	控制组均值	10.3202	11.3295
	ATT（1vs0）	-0.0674**	-0.5427***
	t 统计量	-2.42	-6.600
	P 值	0.0157	7.968×10^{-5}

在工作无职务时，工作类型对劳动者的健康状况、心理状况和幸福感的影响不存在显著差异，对劳动者的经济收入的影响存在显著差异。当工作有职务时，工作类型对劳动者的经济收入的影响也存在显著差异。工作晋升前，“铁饭碗”类型工作的收入水平相对于其他类型工作的收入水平低 0.65%（ATT 除以反事实结果均值），工作晋升后，“铁饭碗”类型工作的收入水平相对于其他类型工作的收入水平低 4.79%。由此假说三得证。还可以从表 4.6 中得出，工作晋升后，加大了“铁饭碗”类型工作与其他类型工作经济收入的差距。

4.5 小结

运用 LASSO-PSM 模型，对所选取的变量进行筛选，验证了在此次实证中使用 LASSO-PSM 模型比 PSM 模型得到的因果效应估计结果更可靠。利用

LASSO-PSM 模型计算平均因果效应结果得出，“铁饭碗”类型工作对劳动者经济收入的影响存在显著的惩罚效应，且“铁饭碗”类型工作劳动者的经济收入显著小于其他类型工作劳动者的经济收入。又将 CGSS 样本按工作有无工作职务进行划分，采用 LASSO-PSM 模型分别测算出“有工作职务”和“无工作职务”情况下两类型工作劳动者健康状况、心理状况、幸福感和经济收入的平均因果效应。基于实证结果，与其他类型工作相比，发现“铁饭碗”类型工作的优势和劣势。“铁饭碗”类型工作的优势：工作及其收入相对稳定。“铁饭碗”类型工作的劣势：“铁饭碗”类型工作对劳动者的经济收入存在显著的惩罚效应，惩罚效应在工作职位晋升后更为明显。工作职位晋升将对劳动者的健康状况、心理状况、幸福感和经济收入均产生正向效应。

可见“铁饭碗”类型工作并非一定是“香饽饽”，求职者应该正确地认识“铁饭碗”和其它类型工作的优和劣，重新建立对“铁饭碗”类型工作的看法，设立正确的就业观和恰当的就业预期。

5 结论与展望

5.1 结论

通过分析 PSM 目前所存在的缺陷以及 LASSO 的优势，将 LASSO 应用到 PSM 中，验证了 LASSO-PSM 模型提高了 PSM 模型匹配结果平衡性，拓展了 PSM 的应用场景。并将 LASSO-PSM 模型应用到劳动者择业偏好影响因素分析中，计算不同工作类型劳动者的经济收入、心理压力、健康状况和幸福感的平均因果效应，发现不同类型工作的劳动者的经济收入存在显著差异。与其他类型工作相比，“铁饭碗”类型工作对劳动者的经济收入存在显著的惩罚效应，惩罚效应在工作职位晋升后更为明显。通过对样本进行有无职务分组，利用 LASSO-PSM 模型计算不同工作类型劳动者经济收入、心理压力、健康状况和幸福感的平均因果效应，结果显示工作职位晋升将对劳动者的健康状况、心理状况、幸福感和经济收入产生正向效应。

5.2 展望

LASSO-PSM 模型提高了 PSM 模型匹配结果平衡性，但 LASSO-PSM 也存在着一定的局限性，LASSO-PSM 倾向于选择对目标变量具有较大影响的特征变量，可能忽略一些具有较小影响但仍然对目标变量有贡献的特征变量。当多个特征变量高度相关时，LASSO-PSM 倾向于选择其中一个特征变量，并将其他高度相关的特征变量系数压缩为 0，这可能导致特征变量选择结果的不稳定性。对于处理具有高度相关的特征变量时，LASSO-PSM 可能会产生不准确的估计，因此 LASSO-PSM 模型匹配结果的平衡性还有待提高。Zou（2006）提出了 LASSO 改进版——自适应 LASSO（Adaptive LASSO），它在 LASSO 的基础上引入了自适

应权重，可以调整变量的收缩力度，从而减小估计的偏差，提供更准确的参数估计。同时，它在变量选择上更加准确，能够增强模型的稳定性。未来可以尝试将改进后的 LASSO，即 Adaptive LASSO 应用到 PSM 模型当中，验证是否能提高 LASSO-PSM 模型匹配结果的平衡性。

参考文献

[1] Amato P R, Previti D.People's reasons for divorcing:Gender,social class,the life course,and adjustment[J].Journal of Family Issues,2003,24(5),602-626.
[2] Austin P C.A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003[J].Statistics in medicine,2008,27(12):2037-2049.
[3] Austin P C.An introduction to propensity score methods for reducing the effects of confounding in observational studies[J].Multivariate behavioral research, 2011,46(3):399-424.
[4] Abadie A,Diamond A,Hainmueller J.Synthetic control methods for comparative case studies:Estimating the effect of California's tobacco control program[J].Journal of the American statistical Association,2010,105(490):493-505.
[5] Austin P C,Stuart E A.Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies[J].Statistics in medicine,2015,34(28):3661-3679.
[6] Becker G S. Human capital:A theoretical and empirical analysis,with special reference to education[M].New York:University of Chicago press,2009.
[7] Brookhart M A,Schneeweiss S,Rothman K J,et al.Variable selection for propensity score models[J].American journal of epidemiology,2006,163(12):1149-1156.
[8] Cooper C,Quick J C.The Handbook of Stress and Health: A Guide to Research and Practice[M].Chichester,West Sussex:Wiley-Blackwell,2012.
[9] Cohan C L,Bradbury T N.Cohabitation and marital dissolution:The significance of marriage expectations[J].Journal of Marriage and Family,2011,73(2),298-317.
[10] Chandrashekar G,Sahin F.A survey on feature selection methods[J].Computers & Electrical Engineering,2014,40(1):16-28.
[11] Do Van T,Nguyen G C,Thi H D,et al.Classification and variable selection using the mining of positive and negative association rules[J].Information Sciences,2023, 631:218-240.
[12] Drentea P,Reynolds J R.Neither a borrower nor a lender be:The relative importance of debt and SES for mental health among older adults[J].Journal of Aging and Health,2012,24(4), 673-695.
[13] Dehejia R H,Wahba S.Propensity score-matching methods for nonexperimental causal studies[J].Review of Economics and statistics,2002,84(1):151-161.
[14] Draper N R,Smith H.Applied regression analysis[M].New York:Wiley,1998.
[15] Fan J,Lv J.A selective overview of variable selection in high dimensional feature space[J].Statistica Sinica,2010,20(1):101.
[16] Gary K,Richard N.Why Propensity Scores Should Not Be Used for Matching[J].Political Analysis,2019,27(4):435-454.
[17] Glynn A N, Quinn K M.An Introduction to the Augmented Inverse Propensity Weighted Estimator[J].Political Analysis,2010,18(1):36-56.
[18] Guyon I,Elisseeff A.An introduction to variable and feature selection[J].Journal

of machine learning research,2003,3(Mar):1157-1182.
[19] Heckman J J,Ichimura H,Todd P.Matching as an Econometric Evaluation Estimator[J].The Review of Economic Studies,1998,65(2):261-294.
[20] Ho D E,Imai K,King G,et al.Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference[J].Political analysis,2007,15(3):199-236.
[21] Hastie T,Tibshirani R,Friedman J H,et al. The elements of statistical learning:data mining,inference,and prediction[M].New York:Springer,2009.
[22] Imbens G W,Rubin D B.Causal inference in statistics,social,and biomedical sciences[M].Cambridge:Cambridge University Press,2015.
[23] Imbens G W,Wooldridge J M.Recent developments in the econometrics of program evaluation[J].Journal of economic literature,2009,47(1):5-86.
[24] James G,Witten D,Hastie T,et al.An introduction to statistical learning[M].New York:springer,2013.
[25] Cheng J,Sun J,Yao K,et al.A variable selection method based on mutual information and variance inflation factor[J].Spectrochimica Acta Part A:Molecular and Biomolecular Spectroscopy,2022, 268:120652.
[26] Liu Y.Adaptive lasso variable selection method for semiparametric spatial autoregressive panel data model with random effects[J].Communications in Statistics-Theory and Methods,2024,53(6):2122-2140.
[27] Luchman J N,Gonzalez-Morales M G.Demands, control, and support: A meta-analytic review of work characteristics interrelationships[J].Journal of Occupational Health Psychology,2013,18(1), 37-52.
[28] Luellen J K,Shadish W R,Clark M H.Propensity scores:An introduction and experimental test[J].Evaluation review,2005,29(6):530-558.
[29] Mroczek D K, Almeida D M.The effect of daily stress,personality,and age on daily negative affect[J].Journal of Personality,2004,72(2),355-378.
[30] Pearl J.Probabilistic reasoning in intelligent systems: networks of plausible inference[M].San Francisco:Morgan Kaufmann,1988.
[31] Pearl J,Mackenzie D.The book of why:the new science of cause and effect[M].New York:Basic Books,2018.
[32] Pearl J.Causal inference in statistics:An overview[M].New York:Wiley2009.
[33] Peres-Neto P R,Jackson D A.How well do multivariate data sets match?The advantages of a Procrustean superimposition approach over the Mantel test[J].Oecologia,2001,129:169-178.
[34] Robert T.Regression Shrinkage and Selection via the Lasso[J].Journal of the Royal Statistical Society.1996,58(1):269-271.
[35] Rosenbaum P R,Rubin D B.Constructing a control group using multivariate matched sampling methods that incorporate the propensity score.TheAmerican Statistician.1985;39(1):33-38.
[36] Stuart E A.Matching methods for causal inference:A review and a look forward[J].Statistical science:a review journal of the Institute of Mathematical Statistics,2010,25(1):1-21.
[37] Shen G,Ru X,Wang K,et al.Influence of labor migration on rural household food

waste in China:Application of propensity score matching (PSM)[J]. Journal of Environmental Management,2024,351:119840.
[38] Spirtes P,Glymour C N,Scheines R.Causation,prediction,and search[M].Cambridge:MIT Press,2000.
[39] Williams K J,Umberson D.Marital status,marital transitions,and health:A gendered life course perspective[J].Journal of Health and Social Behavior,2004,45(1),81-98.
[40] Xiao D,Yu F,Guo C.The impact of China's pilot carbon ETS on the labor income share:Based on an empirical method of combining PSM with staggered DID[J].Energy Economics,2023,124:106770.
[41] Zubizarreta J R.Stable weights that balance covariates for estimation with incomplete outcome data[J].Journal of the American Statistical Association,2015,110(511):910-922.
[42] Zou H.The Adaptive Lasso and Its Oracle Properties[J].Journal of the American Statistical Association,2006,101(476):1418-1419.
[43] Zou H,Hastie T.Regularization and variable selection via the elastic net[J].Journal of the Royal Statistical Society Series B: Statistical Methodology,2005,67(2):301-320.
[44] Zhang L,Zhang L,Du B.Deep learning for remote sensing data:A technical tutorial on the state of the art[J].IEEE Geoscience and remote sensing magazine,2016,4(2):22-40.
[45] 程蒙.工作时间对劳动者职业幸福感的影响研究[D].南京财经大学, 2021.
[46] 陈春明,孔灵芝.中国成人超重和肥胖症预防控制指南[M].北京:人民卫生出版社,2006.
[47] 陈淑云,杨建坤.住房是否影响了居民健康?—来自中国综合社会调查(2015)的实证分析[J].华中师范大学学报(人文社会科学版),2018,57(05):55-64.
[48] 邓小惠,向燕辉.婚姻是幸福的坟墓吗?—基于中国家庭追踪调查的双重差分倾向得分匹配法的估计[J].心理科学,2023,46(3):635-643.
[49] 黄寰,何广,肖义.低碳城市试点政策的碳减排效应[J].资源科学,2023,45(5):1044-1058.
[50] 蒋青嬗.因果推断中基于能源距离的协变量分布平衡[J].统计研究,2023,40(5):144-151.
[51] 康娜.生育二孩对于女性工资的异质性影响—基于倾向得分的双重差分分析[J].统计学与应用,2022,11(6):1329-1337.
[52] 李金鑫.工作压力对员工职业健康的影响[D].长沙:湖南师范大学,2017.
[53] 廖君瑶.工作压力对员工幸福感的影响[D].广州:广州大学,2021.
[54] 马红梅,代亭亭.工作时间长度对劳动者健康的影响—基于 CFPS(2020)数据的实证研究[J].西北人口, 2022,43(06):99-112.
[55] 任国强,姚舜禹.工作时间对城乡就业者健康的影响[J].中国劳动关系学院学报, 2023, 37(02):80-93.
[56] 司健敏.新生代员工工作时间与其工作绩效的关系研究[D].山西财经大学,2018.
[57] 王玉荣,段玉婷,卓苏凡.工业互联网对企业数字创新的影响—基于倾向得分匹配的双重差分验证[J].科技进步与对策,2022,39(8):89-98.

[58]魏祥迁,马红宇,张明亮.员工家庭压力与工作压力对健康的影响[J].心理研究, 2012, 5(04): 41-45.
[59]徐小兵,李迪,孙扬等.基于倾向得分匹配的农村中老年人慢性病共病对失能的影响研究[J].中国全科医学,2023,26(04):434-439.
[60]谢申祥.传统 PSM-DID 模型的改进与应用[J].统计研究, 2021, 38(2):146-160.
[61]杨青,吕赟.浅析体育锻炼对健康的促进意义[J].冰雪体育创新研究,2020(01):99-100.
[62]俞林伟.居住条件对流动人口健康的影响研究[D].福州:福建师范大学,2019.
[63]张云桥.婚姻与主观幸福感—基于 CGSS2013 数据的实证分析[J].文化创新比较研究,2017,1(16):88-91.
[64]赵梦阳,刘同山.促进还是抑制:农民合作社对农村集体经济影响的计量分析[J].云南农业大学学报 (社会科学),2023,17(4):27-36.

攻读硕士学位期间参与的科研任务及主要成果

科研项目：

甘肃省中央引导地方科技发展资金项目《城市计算方法体系构建及甘肃智慧城市应用》（项目编号：YDZX20216200001876）

比赛获奖：

2022 年第五届全国应用统计专业学位研究生案例大赛二等奖

2023 年第六届全国应用统计专业学位研究生案例大赛三等奖

致谢

论文写到致谢，也就意味着即将向我的硕士研究生生涯说再见了，回首这过去三年的学习经历，想想完成这篇论文的一年光阴，心中感慨良多。

从拿到硕士研究生录取通知书的那一刻，我就明白了我已经站到了自己梦寐以求的平台上，那时的我充满激情、梦想、抱负和无尽的志向。三年的岁月，在求知、求学、积累、历练中如斯逝去，这段岁月让我逐渐褪去了本科刚出校门时的稚气和冲动，多了几分成熟与冷静，依然不变的是同样的激情和梦想。它是我人生的一个重要的缓冲和积累的过程，使我的心灵得到了锤炼，学业也得到了升华，让我可以充满自信，身怀所学，以不减的激情投入到新的人生阶段，将这宝贵的三年财富转化为生活的动力。

在写这篇论文的一年中，我的身心也经历了巨大的考验。刚开始为选题困惑，后来又为文章结构所累，再就是对论文所要解决的问题提出合理的参考性观点时又遇到了思路瓶颈，同时在这个过程中发现了自己知识沉淀的不足和对理论以及文字的驾驭力的欠缺。不过最终论文还是成型，我深感欣慰，因为这其中也的确投入了自己的心血和付出，只是当前的知识素养只能支撑自己写出这种水平和高度，希望在今后不断努力的过程中，能有更大的提高和更深的见解来使自己的观点更加完善。

我满怀一颗感恩的心面对身边所有的老师、同学、朋友和家人，不知道“感谢”二字是否可以表达心底那份最深的情意。感谢我的导师牛成英教授，一个才华横溢的学者，一个宽厚仁慈的长辈，一个可以谈心的朋友，我从您身上学到的东西虽九牛一毛，但足够我受益终身。在我论文的每一次修改过程中，牛老师帮助我开拓研究思路，精心点拨、热忱鼓励，并为我指点迷津，才使得我在面对各种问题的时候得以豁然开朗。牛老师不仅授我以文，而且教我做人，虽历时三载，

却给以终生受益无穷之道，我对牛老师的感激之情无法用言语来表达！

感谢我师门的同学，这个集体团结友爱，志向高远，我能成为这个大家庭中的一员感到非常荣幸。正是因为有了这样一批兄弟姐妹，才使我在求学的路上感到充满力量。感谢我的室友，在像家一样温暖的宿舍里，我们彼此关爱，同喜同悲，一同渡过了生命中非常重要的时光，希望未来的日子里大家都能幸福！

我要由衷地感谢我的父母，他们是我人生中最伟大的支持者和榜样。他们无私地付出了无数的努力和牺牲，给予我无尽的爱和鼓励。没有他们的支持和信任，我无法完成这篇论文。他们是我坚强的后盾和智慧的指导，时刻鼓励我追求知识。无论我遇到多大的困难，他们总是给予我坚定的支持和无私的建议。同时他们也是我生活中最温柔和最关怀的存在。他们在我成长的每一个阶段都给予我无微不至的照顾和无限的爱。他们的鼓励和支持使我有勇气面对挑战，并不断超越自己。他们为我提供了良好的家庭环境和优质的教育资源，为我打下了坚实的知识基础。他们的辛勤工作和牺牲精神是我不断努力和追求卓越的动力。我想对他们表达我深深的感激之情，他们是我人生中最重要的人，我将永远珍惜他们对我的支持和爱，没有他们，我不可能取得今天的成就，在今后的生活中我将继续努力。

最后，我想对自己说，硕士研究生生涯的结束意味着新的人生阶段的开始，不负他们的期望。在今后的岁月里，我不论做任何事情，都会认真、努力，不断成就自己的梦想和更加精彩的人生!

仅以此文献给所有在我人生大道上曾经支持、鼓励、帮助过我的人，谢谢你们!