户籍歧视导致的收入差距依然存在吗*

——基于机器学习方法的再讨论

江求川[1] 鲁元平[2,3]

摘要：中国劳动力市场供需结构变化促使农民工与城镇职工间收入差距呈现新态势，有关户籍歧视是否依然存在的争论越来越多。本文用双重去偏机器学习方法重新检验农民工与城镇职工收入差距中的户籍歧视现象。经验分析表明：第一，迁移溢价干扰了对户籍歧视的识别，考虑迁移溢价因素后户籍歧视现象更加明显；第二，用双重去偏机器学习方法选择更加符合条件独立性假设要求的模型后，农业户籍对劳动者小时工资收入、全年总收入和全年工资收入均有负面影响，且对小时工资收入的负面影响更为显著；第三，经双重去偏机器学习修正后的 Oaxaca-Blinder 分解结果表明，农民工和城镇职工收入差距中大约有 8%～15%属于户籍歧视；第四，Oster 检验证实双重去偏机器学习的估计相较 OLS 估计更加可靠，不同机器学习算法下的双重去偏机器学习估计与 Lewbel 工具变量估计也表明本文结论是稳健的。

关键词：户籍歧视　机器学习　收入差距　农民工

中图分类号：F241.2　　**文献标识码：**A

一、引言

中华人民共和国成立之初，政府推行重工业优先发展战略，户籍制度等一系列城市偏向型政策和制度应运而生（陈斌开和林毅夫，2013）。虽然大多数城市偏向型政策和制度随着社会经济形势的变化已经退出历史舞台，但户籍制度改革却久攻不克（蔡昉，2023）。户籍制度在一定程度上造成了城乡居民的发展机会不均等。尤其是 20 世纪 80 年代之后，大量农村剩余劳动力流入城市，城乡居民的发展不均衡问题快速凸显。其中，最突出的表现就是城镇职工与农民工之间的收入差距和就业机会差距（Meng and Zhang，2001；邢春冰，2008；章莉等，2014；孙婧芳，2017；Ma，2018；吴彬彬等，

*本文研究得到国家社会科学基金一般项目“共同富裕目标下代际收入流动阻碍因素与提升路径研究”（编号：22BJL074）的资助。本文通讯作者：鲁元平。

2020）。不少学者认为，城镇职工与农民工之间的收入差距有一部分源于劳动力市场针对农民工的户籍歧视（吴珊珊和孟凡强，2019），甚至有研究表明户籍歧视构成了城镇职工与农民工收入差距的主体（谢嗣胜和姚先国，2006；邓曲恒，2007）。然而，也有学者发现，随着刘易斯转折点的到来，农村剩余劳动力的无限供给特征已经消失，针对农民工的户籍歧视正在减弱，城镇职工与农民工的工资收入出现了缓慢的趋同（吕炜等，2019；许岩，2022），甚至出现了针对农业户籍的优待或“反向歧视”（章莉和吴彬彬，2019）。那么，户籍歧视导致的城镇职工与农民工收入差距真的消失了吗？

一直以来，针对城镇职工与农民工在劳动力市场上被差别对待的研究大多聚焦于农民工的农业户籍身份（孙婧芳，2017；吴彬彬等，2020；孟凡强等，2022）。从人口迁移角度看，收入增长是移民选择流动的主要动机之一，对农民工流入城市而言更是如此。经验证据表明，在中国劳动力市场上人口迁移带来的收入增长效应较为明显（孙三百，2015）。农民工的迁移人口和农业户籍双重身份意味着忽略任何一种效应都会影响对另一种效应的精确识别。虽然已有文献注意到农业户籍身份对识别流动人口迁移溢价的潜在干扰（刘金东等，2020；黄静和祝梦迪，2022），但在识别户籍歧视对劳动力收入的影响时，鲜有研究关注迁移溢价的潜在干扰。

识别农业户籍对劳动力收入的影响面临的最大挑战是实证检验时必须考虑农业户籍和迁移人口这两个变量的潜在内生性。以往文献几乎均在条件独立性假设（conditional independence assumption，简称 CIA）成立的前提下估计农业户籍的收入效应，或对城镇职工与农民工的收入均值和分布进行分解（刘金东等，2020；邢春冰等，2021；孟凡强等，2022）。CIA 要求观察到的个体特征信息足够充分，足以包含所有与户籍身份、迁移身份和潜在收入同时相关的信息。但是，以往研究大多数只控制了非常有限的个体特征信息，CIA 是否成立成为识别户籍歧视存在与否的关键。

与现有研究相比，本文可能的贡献有三点。第一，本文提供了潜在遗漏变量问题有可能对户籍歧视识别造成严重影响的证据，对以往研究结论做出了补充。本文的主要控制变量除包含以往研究中的常规变量外，还包含个体职业声望、身体特征、家庭背景等信息。然而，实证结果表明，即便如此丰富的控制变量也无法完全排除遗漏变量造成的影响。第二，本文为回答 2012－2021 年这一阶段的城镇职工与农民工收入差距中是否依然存在户籍歧视问题提供了更加稳健的证据。本文在识别农业户籍对劳动力收入的影响时排除迁移溢价的干扰，并同时考虑这两个关键变量潜在的内生性问题，不仅补充了户籍歧视研究和迁移溢价研究这两类文献，也为城镇职工与农民工收入差距中的户籍歧视识别提供了更稳健的结果。第三，为尽可能识别出农业户籍对收入的影响，本文使用 Chernozhukov et al.（2018）提出的双重去偏机器学习（double debiased machine learning，简称 DML）方法，不仅在研究方法上丰富了户籍歧视研究，也为基于双重机器学习的项目评估理论提供了新的经验支撑。

二、文献回顾

经济学研究一直对劳动力市场分割问题有浓厚兴趣，因为劳动力市场分割不仅导致低效率，还会加剧经济不平等。户籍制度导致的劳动力市场分割无疑是中国劳动力市场最主要的特征之一。20 世纪

80 年代以后，随着城市劳动力市场中农业户籍劳动者的规模不断扩大，农村流动人口与城镇本地职工在就业和收入获得方面的差异引起广泛关注。早期的大量研究发现农民工和城镇职工之间的收入差距中只有较少的部分可以由两个群体间的特征差异解释，超过 50%的差异是不可解释的，这部分差异通常被认为是劳动力市场针对农业户籍的歧视（王美艳，2005；谢嗣胜和姚先国，2006；邓曲恒，2007）。2005 年之后的研究通过对新数据分析发现，中国劳动力市场中各类针对农业户籍的歧视现象出现了新特征。越来越多的证据表明，虽然户籍歧视导致的收入差距仍然不可忽视，但出现了明显的减弱趋势（邢春冰，2008；章莉等，2014；孙婧芳，2017；于潇和孙悦，2017）。究其原因，可能是中国劳动力市场的供需结构变化和制度变化的综合结果（孙婧芳，2017）。

户籍歧视现象的新趋势促使后续研究朝三个方向发展。一是聚焦于户籍歧视的其他表现形式，如就业的户籍歧视（章莉和吴彬彬，2019；吴彬彬等，2020）；二是考察农民工和城镇职工收入的趋同速度（陈珣和徐舒，2014；吕炜等，2019；胡凤霞和叶仁荪，2019；许岩，2022）；三是探究是否存在针对外来农民工的“反向歧视”（陈昊等，2017；章莉和吴彬彬，2019；邢春冰等，2021；孟凡强等，2022）。与早期户籍歧视研究将“流动人口”“农业户籍人口”“农民工”等名词相互替换使用的做法不同，上述研究开始关注“户籍来源地”与“户籍性质”之间的差异。陈昊等（2017）发现外地户籍人口相比本地居民有 58.64%的收入溢价，并将这一现象称为户籍所在地的“反向歧视之谜”。为了避免针对农业户籍的歧视干扰对外地户籍溢价的识别，他们将样本划分为农业户籍和城镇户籍两个子样本，并证实在两个子样本中均存在外地户籍溢价。但是，这种方式本身并不能说明农业户籍歧视是否存在。类似地，徐凤辉和赵忠（2014）、邢春冰等（2021）均发现外地户籍获得了收入溢价。但这些研究均未能充分考虑农业户籍歧视的潜在效应，不能排除户籍歧视存在的可能性。刘金东等（2020）、黄静和祝梦迪（2022）在研究流动人口工资溢价时控制了农业户籍因素，但户籍因素并非核心变量，且未考虑潜在的遗漏变量问题对户籍歧视识别的干扰。刘金东等（2020）的结果表明迁移溢价和户籍歧视均不存在，而黄静和祝梦迪（2022）的结论表明流动人口工资溢价和户籍歧视均存在。与上述研究不同的是，在识别农业户籍对劳动力收入的影响时，鲜有研究关注迁移溢价对结果的干扰。章莉和吴彬彬（2019）、吴彬彬等（2020）在不考虑迁移溢价的情况下发现农业户籍人口在获得收入时其实受到了优待。同样是未考虑迁移溢价，孟凡强等（2022）、于潇等（2022）却发现户籍歧视仍然存在。

综上所述，现有研究对城镇职工与农民工收入差距中户籍歧视的影响进行了极为有益的探索。从研究结论上看，相关文献大体可分三个阶段：户籍歧视在城乡劳动力收入差距中占主导地位阶段、户籍歧视作用下降阶段和户籍歧视存在性存疑阶段。前两个阶段中大量经验研究结论相对统一，但这些研究结论均以 CIA 成立为前提，研究本身均未对 CIA 成立与否进行讨论。在户籍歧视较为严重甚至作为城镇职工和农民工收入差距的主导因素时，CIA 不成立所带来的估计偏差可能不会影响对户籍歧视存在性的判断。但随着户籍歧视的减弱，CIA 不成立可能导致户籍歧视的识别结果在存在歧视、存在“反向歧视”和不存在歧视之间反复。因此，现阶段探讨户籍歧视导致的收入差距是否依然存在，有必要关注 CIA 是否成立，并在尽可能满足 CIA 的基础上识别户籍歧视对劳动力收入的影响。

三、计量方法与数据

（一）计量方法

为将户籍歧视研究的常规方法和 DML 方法置于同一框架，本文先考虑 Rubin 因果模型。用 $rural_i$=1 和 $rural_i$=0 分别表示农业户籍和城镇户籍；y_i^1 和 y_i^0 分别表示农业户籍和城镇户籍状态下的潜在收入，则 $\tau_i = y_i^1 - y_i^0$ 为户籍差异对收入影响的净效应。实践中，只能观察到个体的实际收入 y_i、与收入相关的特征信息 x_i 和户籍状态 $rural_i$。y_i 的表达式为：

$$y_i = y_i^1 \times rural_i + y_i^0 (1 - rural_i) \tag{1}$$

虽然 τ_i 永远无法被直接观测到，但在一定条件下，平均处理效应 $\tau_{ate} = E(\tau_i) = E(y_i^1 - y_i^0)$ 是可以被识别的。其中，CIA 是识别 τ_{ate} 的重要基础。具体而言，CIA 要求在给定可观察的个体特征信息 x 的条件下，户籍身份与个体的潜在收入之间没有任何关联，即：

$$y_i^1, y_i^0 \perp rural_i \mid x_i = x \tag{2}$$

简单来说，在（2）式成立的条件下，条件平均处理效应 $\tau_{ate}(x) = E(\tau_i \mid x_i = x)$ 是可识别的，即：

$$\tau_{ate}(x) = E(y_i \mid x_i = x, rural_i = 1) - E(y_i \mid x_i = x, rural_i = 0) \tag{3}$$

（3）式成立的关键在于 CIA 保证了 $E(y_i^t \mid x_i) = E(y_i \mid x_i, rural_i = t)$，其中 t=0，1。因此，$\tau_{ate}(x)$ 可以用观测到的样本信息（y_i，x_i，$rural_i$）估计。根据迭代期望原理，$\tau_{ate}(x)$ 构成识别平均处理效应 τ_{ate} 的基础。一般而言，$E(y_i^t \mid x_i)$ 可用参数或非参数方法估计得到。大多数经验研究采用参数估计方法（刘金东等，2020；黄静和祝梦迪，2020；邢春冰等，2021）。参照现有研究的做法，本文假定 y_i^t 是 x_i 的线性函数，即 $y_i^0 = x_i\gamma + u_i^0$，$y_i^1 = \beta \times rural_i + x_i\gamma + u_i^1$，且 $E(u_i^t \mid x_i)$=0。其中，γ 和 β 为系数，u_i^0、u_i^1 和 u_i^t 为残差项。根据上述假定可构建如下线性模型：

$$y_i = x_i\gamma + \beta \times rural_i + u_i \tag{4}$$

（4）式中：$u_i = u_i^0 + (u_i^1 - u_i^0) \times rural_i$。$\beta$ 反映的是户籍差异对收入的影响，其估计值是否具有因果解释能力的关键前提之一在于（2）式是否成立。为了让（4）式更好地满足 CIA，最简单的做法是在（4）式中增加控制变量并提升模型灵活性。然而，经济学理论和经验可能会说明哪类变量是重要变量，但不一定会说明哪一个变量是重要变量。因此，在实践问题中，变量和模型形式选择并非易事。

近年来，大量机器学习算法被应用到因果识别问题当中，推动了因果机器学习这一新领域的快速发展。与传统的机器学习研究以提高模型的预测能力为目标不同，因果机器学习的目标是提升因果识别的可靠性。对于本文关心的（4）式而言，传统机器学习方法试图提升模型对 y_i 的预测能力，而因果机器学习试图得到 β 的可靠估计量。理论研究表明，用机器学习算法估计（4）式得到的 β 估计值

不是一致估计量，相应的估计偏差被称为正则化偏差（Belloni et al.，2014；Chernozhukov et al.，2018）。为解决正则化偏差，需要进一步考虑辅助方程：

$$rural_i = x_i\alpha + v_i \tag{5}$$

根据（5）式可以构造（4）式的如下简约型：

$$y_i = x_i\varphi + \omega_i \tag{6}$$

（6）式中：$\varphi = \beta\alpha + \gamma$，$\omega_i = \beta v_i + u_i$。用机器学习算法分别拟合（5）式和（6）式得到系数估计量$\hat{\alpha}$和$\hat{\varphi}$。最终，β的估计量$\hat{\beta}$构造如下：

$$\hat{\beta} = \left(\frac{1}{n}(rural_i - x_i\hat{\alpha}) \times rural_i\right)^{-1} \frac{1}{n}\sum_i (y_i - x_i\hat{\varphi})(rural_i - x_i\hat{\alpha}) \tag{7}$$

（7）式中：$\hat{\beta}$被称为双重去偏机器学习估计量。（7）式虽然解决了正则化偏差，但同时也带来了新问题。由于机器学习算法的目标是更好地进行样本内预测，所以通常存在过度拟合问题。（5）式或（6）式的过度拟合都会导致估计（7）式时出现过度拟合偏差（overfitting bias）。根据 Chernozhukov et al.（2018）的建议，本文使用交叉拟合（cross-fitting）方法来缓解过度拟合偏差。最终的估计过程如下：先将总样本分为k份，再用其中的第j（j=1,⋯,k）份S_j作为测试集，用其余样本S_{-j}作为训练集学习（5）式和（6）式，最后将训练集的学习结果代入测试集用于计算（7）式。重复上述过程S次，用S次的平均值作为最终估计结果。

严格来说，所有有监督的机器学习算法均适用于上述估计过程。本文选择 Lasso 作为基准回归中的学习算法，这是目前 DML 方法中使用最广泛的学习算法（Knaus，2021；Bonaccolto-Töpfer and Briel，2022）。为了让模型能更好地满足 CIA，研究者需要提供丰富的潜在控制变量集合。在这种高维数据情形下，基于 Lasso 算法的 DML 估计同时也发挥了变量筛选功能，这为后续的研究（如 Oaxaca-Blinder 分解）提供了基础。当然，就 DML 估计本身而言，其他机器学习算法同样也可以估计（5）式和（6）式。为此，本文在稳健性分析部分还采用随机森林、岭回归和弹性网络这些常规的机器学习算法进行估计。

（二）数据与变量

1.*数据来源*。本文使用的数据来自中国综合社会调查（Chinese general social survey，简称 CGSS）于 2012 年、2013 年、2015 年、2017 年、2018 年和 2021 年的 6 次抽样调查数据。该调查于 2003 年开始，目前由全国 48 所高校联合组织实施，每年在全国范围内通过多阶分层 PPS 随机抽样方法抽取约 10000 户家庭。本文选择使用 2012 年后的最近 6 次调查数据，主要基于两点考虑：第一，关于中国劳动力市场中户籍歧视研究的分歧主要发生在基于 2005 年以后的数据所得出的结论中，且这些研究认为城镇职工和农民工的收入在缓慢趋同。使用最新的数据更有利于回应以往研究的分歧，也有利于观察缓慢趋同后的最新状态。第二，2012 年后的 CGSS 问卷问题设计相对稳定，能最大限度保持变

量的完整性和变量定义的统一性。

本文将研究对象限定为18～60岁且目前从事非农工作的个体[①]，不包含自雇者和个体商户。本文删除了收入、户籍等关键信息缺失的样本，同时为避免离群值的影响，进一步剔除了收入最高和最低的1%样本。由于单一年份的最终有效样本量较少，为提高估计效率并兼顾变量定义的统一，本文将2012年、2013年和2015年3年的样本合并，共计10044个观测值；将2017年、2018年和2021年3年的样本合并，共计9767个观测值。

2.变量定义。本文的被解释变量为劳动者个体的收入。根据CGSS的问卷设计，本文定义3个反映个体收入的变量，分别是全年收入、全年工资收入和小时工资。其中，全年收入和全年工资收入根据受访者对过去一年总收入和职业收入的回答直接获得。根据CGSS问卷，职业收入为劳动者所有劳动收入。由于本文剔除了自雇、个体商户等无劳动合同信息的个体，故职业收入可作为工资收入的代理变量。小时工资由全年工资收入和每周工作小时数据估算得到。

本文的核心解释变量为受访者是否为农业户籍。CGSS将受访者分为农业户口、城镇户口和居民户口3种主要类型[②]。本文将农业户口和取得居民户口前为农业户口的个体定义为农业户籍人口。同时，为了避免迁移溢价干扰，受访者是否为迁移人口也是本文的核心变量。按照以往研究（邢春冰等，2021；黄静和祝梦迪，2022）的常见做法，本文将户籍所在地不是本县（市、区）的个体定义为迁移人口。

在控制变量方面，根据CGSS问卷设计，综合考虑变量可得性、可比性和信息缺失等因素，本文最终从个体特征、工作特征和家庭背景等信息中选定了24个主要控制变量、26个潜在控制变量，共计50个基础变量，再由这些基础变量的高次项和交互项构成机器学习模型的控制变量集合。

主要控制变量在现有研究的常规控制变量基础上扩展而来。首先，在个体特征变量方面，除了包含性别、年龄、受教育程度、婚姻状态这些以往研究中的常规变量外，还增加了自评健康、党员身份、身高、体重、BMI指数等与个人收入相关的变量。其次，在工作特征变量方面，本文控制了比以往研究更丰富的变量集合，包括工作经验、工作单位所有制、工作单位类型、工作单位规模、工作合同类型、职业ISEI得分和职业SIOPS得分[③]。最后，在家庭背景方面，本文控制了以往户籍歧视研究鲜有

[①]将个体样本年龄限定在60岁及以下是现有研究的常规做法（王美艳，2005；邢春冰，2008；章莉等，2014；黄静和祝梦迪，2022）。本文使用的CGSS样本中，农业户籍人口中大于60岁的人口占比为7.3%，城镇户籍人口中大于60岁的人口占比为10.1%，二者均不高。

[②]居民户口指部分地区取消农业户口和城镇户口划分后的统一户口登记类型。此外，CGSS中的户口类型还包括蓝印户口，但由于2000年后各地蓝印户口逐渐被取消，所以样本中蓝印户口较少，这类样本被直接剔除。

[③]2017年以前的CGSS调查按照国际标准职业分类代码ISCO-88（international standard classification of occupations, revised edition 1988）对受访者职业进行编码，2017年以后的调查使用ISCO-08职业编码。ISCO-88和ISCO-08可直接与国际标准职业社会经济指数（international socio-economic index of occupational status，简称ISEI）、标准国际职业声望量表（Treiman's standard international occupational prestige scale，简称SIOPS）接驳，用于更精准地刻画受访人职业特征。

关注的父（母）亲受教育程度、父（母）亲职业 ISEI 得分、父（母）亲职业 SIOPS 得分[①]。

表 1 是本文主要变量的定义及描述性统计结果。

表 1　　主要变量定义及描述性统计结果

变量名称	变量定义与赋值	2012—2015 年		2017—2021 年	
		均值	标准误	均值	标准误
被解释变量					
全年收入	受访者全年收入（元）	44961.994	43164.437	72303.500	183054.964
全年工资收入	受访者全年工资收入（元）	42577.397	33633.585	70028.436	159939.088
小时工资	受访者小时工资收入（元）	19.453	18.067	35.452	85.434
核心解释变量					
是否农业户籍	受访者户籍类型：农业户籍=1，非农业户籍=0	0.352	0.478	0.503	0.500
是否迁移人口	受访者户籍所在地：本县（市、区）以内=1，本县（市、区）以外=0	0.179	0.384	0.243	0.429
个体特征变量					
性别	受访者性别：男=1，女=0	0.567	0.496	0.526	0.499
年龄	受访者年龄（岁）	39.693	10.373	40.458	10.701
受教育程度	受访者学历：本科及以上，大专，中专或高中，初中，小学及以下				
婚姻状态	受访者婚姻状态：未婚，已婚，其他				
自评健康	受访者自评健康状况：很健康，比较健康，一般，比较不健康，很不健康				
党员身份	受访者是否为中共党员：是=1，否=0	0.145	0.353	0.139	0.346
身高	受访者身高（米）	1.669	0.770	1.666	0.790
体重	受访者体重（千克）	63.882	11.511	63.993	12.627
BMI 指数	受访者 BMI 指数	22.844	3.255	22.965	3.667
工作特征变量					
工作经验	受访者参加工作以来累积工作年数（年）	15.663	10.602	15.152	11.234
工作单位所有制	受访者工作单位所有制类型：国有或集体，私营或民营，其他				
工作单位类型	受访者工作单位类型：机关或事业单位，企业，其他				
工作单位规模	受访者工作单位人数：1000 人及以上，150～999 人，50～149 人，10～49 人，10 人以下，不确定				

[①]本文样本中受访者父母职业信息缺失值较多。对于父母职业信息缺失的情况，受访者父母的 ISCO 被标记为 9999，相应的 ISEI 得分和 SIOPS 得分为所有职业的 ISEI 均值和 SIOPS 均值。为了尽可能减少样本损失，本文保留了这些父母职业信息存在缺失值的样本，并在控制变量中增加了父母职业信息是否缺失的虚拟变量。

表 1（续）

工作合同类型	受访者签订劳动合同情况：无固定期限合同，固定期限合同，未签合同				
职业 ISEI 得分	受访者职业 ISEI 得分	4.275	1.484	4.207	1.435
职业 SIOPS 得分	受访者职业 SIOPS 得分	3.861	1.415	4.170	1.250
家庭背景特征变量					
父亲受教育程度	受访者父亲学历：高中及以上，初中，小学，未上学，私塾及其他				
母亲受教育程度	受访者母亲学历：高中及以上，初中，小学，未上学，私塾及其他				
父亲职业 ISEI 得分	受访者父亲职业 ISEI 得分	3.519	1.635	3.272	1.627
母亲职业 ISEI 得分	受访者母亲职业 ISEI 得分	3.178	1.328	2.946	1.382
父亲职业 SIOPS 得分	受访者父亲职业 SIOPS 得分	4.065	1.084	4.236	0.914
母亲职业 SIOPS 得分	受访者母亲职业 SIOPS 得分	3.915	0.773	4.182	0.692
父亲职业信息缺失	受访者父亲职业信息是否缺失：是=1，否=0	0.120	0.325	0.141	0.348
母亲职业信息缺失	受访者母亲职业信息是否缺失：是=1，否=0	0.244	0.430	0.242	0.429

注：①全年收入、全年工资收入和小时工资均以 2021 年价格计，表中展示的是原值，后文估计中使用的是对数值。②所有分类变量（含有序分类变量和无序分类变量）在估计时均转换为多个二元虚拟变量。表中有序分类变量的均值和标准差无意义，故表中未展示。③连续型控制变量在后文估计时均做标准化处理。④2012－2015 年样本和 2017－2021 年样本分别有 10044 个和 9767 个观测值，但小时工资数据存在缺失，2012－2015 年样本和 2017－2021 年样本中小时工资变量分别只有 8767 个和 8570 个观测值。

除上述主要变量外，为了让机器学习算法可以从复杂的变量集合和灵活的方程形式中选出更加符合 CIA 的模型，本文还加入了一些潜在的影响收入的因素作为控制变量。这些变量包括受访者的生理健康、心理健康、网络使用、社交、政治参与、语言能力、家庭结构、性别观念、父（母）亲社会经济特征等共计 26 个变量。

四、户籍歧视效应的估计与分解

（一）户籍歧视效应估计

1.不考虑迁移因素的 OLS 估计。为便于和已有研究结论对比，本文首先沿用以往研究的常规做法（章莉和吴彬彬，2019；孟凡强等，2022），在不考虑迁移因素且只加入主要控制变量的情形下估计农业户籍对劳动力收入的影响。表 2（1）列结果表明，在不考虑城镇职工与农民工之间的特征差异时，2012－2015 年，两个群体的小时工资收入相差 36%左右。表 2（5）列结果表明，城镇职工与农民工之间未调整的小时工资差距在样本期内是相对稳定的。这种未经任何调整的小时工资差距可能完全来自两个群体间的特征差异。表 2（2）列和（6）列进一步控制了性别、年龄、受教育程度、婚姻状态等个体特征变量。结果显示，两个群体间的个体特征差异确实在很大程度上解释了二者之间的工资差异，虽然两组样本的估计结果仍表明农民工的小时工资更低，但与城镇职工小时工资的差距已经大幅

度下降。但是，个体特征差异并不能完全解释两个群体间的工资差异，农民工的小时工资仍显著低于城镇职工。表 2（3）列和（7）列进一步加入了工作相关特征作为控制变量，主要包括工作经验、工作单位所有制、职业 ISEI 得分、职业 SIOPS 得分等。从 2017－2021 年样本的估计结果来看，农业户籍对小时工资的负面影响相较（6）列进一步减弱，而 2012－2015 年样本的估计结果表明农业户籍没有对小时工资造成显著的负面影响。考虑到城镇职工和农民工可能还在社会资本、个体性格等方面存在差异，表 2（4）列和（8）列进一步加入了父母受教育程度、父母职业 ISEI 得分、父母职业 SIOPS 得分等与个人成长环境和社会资本有关的家庭背景因素。从 2017－2021 年样本的估计结果来看，农业户籍对小时工资收入的负面影响相较表 2（7）列结果所体现的进一步下降，2012－2015 年样本的估计结果则表明农业户籍没有对工资收入造成显著的负面影响。

表 2　　农业户籍对劳动者小时工资收入影响的 OLS 估计结果（不区分是否为迁移人口）

变量	被解释变量：小时工资							
	2012－2015 年				2017－2021 年			
	（1）	（2）	（3）	（4）	（5）	（6）	（7）	（8）
是否农业户籍	−0.356***	−0.081***	−0.038	−0.023	−0.385***	−0.123***	−0.098***	−0.085***
	（0.027）	（0.026）	（0.025）	（0.024）	（0.040）	（0.029）	（0.029）	（0.028）
个体特征	未控制	已控制	已控制	已控制	未控制	已控制	已控制	已控制
工作特征	未控制	未控制	已控制	已控制	未控制	未控制	已控制	已控制
家庭背景	未控制	未控制	未控制	已控制	未控制	未控制	未控制	已控制
年份固定效应	已控制	已控制	已控制	已控制	已控制	已控制	已控制	已控制
省份固定效应	已控制	已控制	已控制	已控制	已控制	已控制	已控制	已控制
观测值	8767	8767	8767	8767	8570	8570	8570	8570
调整后的 R^2	0.227	0.361	0.381	0.383	0.186	0.326	0.349	0.352

注：①括号内为异方差稳健标准误。②***表示 1%的显著性水平。

综合而言，表 2 关于是否存在户籍歧视的结论还不稳健，变量控制方式对估计结果影响较大，既给出了类似于吴彬彬等（2020）得出的户籍歧视不明显的结论，也得到了类似于孟凡强等（2022）给出的户籍歧视未能消除的结论。

通过分析表 2 的结果不难发现，按照常规方法估计并判断户籍歧视是否存在时，模型选择和变量控制需要特别谨慎。一方面，有些模型可能无法很好地满足 CIA 条件，导致结果不稳健，如遗漏重要变量或者包括了所谓的坏控制变量。另一方面，人口迁移带来的收入溢价已经被大量研究证实（于潇等，2022），忽略迁移因素也有可能导致迁移溢价和户籍歧视混淆，从而使估计结果不稳健。

2.*考虑迁移因素的 OLS 估计*。表 3 通过两种方式引入迁移因素：一是假定迁移对工资的影响在农民工和城镇职工之间不存在异质性，估计结果见其中的（1）列、（2）列、（5）列和（6）列；二是允许迁移对工资的影响在两个群体间存在差异，估计结果见其中的（3）列、（4）列、（7）列和（8）列。从表 3（1）列和（5）列结果可发现，农民工和城镇职工小时工资差距相较表 2 中未调整的工资差距均有所上升，说明迁移溢价确实在一定程度上掩盖了农业户籍对农民工小时工资收入的负面影响。

表 3（2）列和（6）列进一步控制了个体特征、工作特征和家庭背景变量。表 3（2）列结果表明，农业户籍对农民工的小时工资仍然有显著负面影响；表 3（6）列结果表明，农业户籍对农民工小时工资的负面影响相较表 2 结果略有上升。

迁移溢价在农民工和城镇职工之间可能有差异，这种差异会干扰对户籍歧视的识别。为此，本文根据劳动力是否为迁移人口和是否为农业户籍进一步将样本划分为本地城镇户籍人口、外地城镇户籍人口、本地农业户籍人口和外地农业户籍人口四类，并以本地城镇户籍人口为基准组重新估计小时工资方程，回归方程中本地农业户籍变量的系数可以反映户籍歧视效应的大小。表 3（4）列和（8）列结果表明，相较本地城镇户籍人口，本地农业户籍对农民工小时工资有负面影响，但结论不稳健。此外，在其他条件相同情况下，如果迁移对城镇职工和农民工产生的溢价相同，那么外地城镇户籍和外地农业户籍的回归系数差异也可以在很大程度上反映户籍歧视效应。针对外地城镇户籍和外地农业户籍回归系数差异的 Wald 检验表明两个系数存在显著差异。上述检验在很大程度上证实了户籍歧视并未完全消失，同时也说明，忽略迁移溢价会干扰对户籍歧视的识别。

表 3　　农业户籍对劳动者小时工资收入影响的 OLS 估计结果（区分是否为迁移人口）

变量	被解释变量：小时工资							
	2012—2015 年				2017—2021 年			
	（1）	（2）	（3）	（4）	（5）	（6）	（7）	（8）
是否农业户籍	−0.385***	−0.041*			−0.400***	−0.095***		
	（0.027）	（0.024）			（0.040）	（0.029）		
是否迁移人口	0.112***	0.090***			0.087**	0.078***		
	（0.027）	（0.024）			（0.032）	（0.024）		
外地城镇户籍			0.151***	0.125***			0.117***	0.075*
			（0.029）	（0.031）			（0.041）	（0.037）
本地农业户籍			−0.369***	−0.027			−0.388***	−0.096***
			（0.030）	（0.024）			（0.045）	（0.032）
外地农业户籍			−0.290***	0.032			−0.322***	−0.016
			（0.042）	（0.038）			（0.050）	（0.028）
个体特征	未控制	已控制	未控制	已控制	未控制	已控制	未控制	已控制
工作特征	未控制	已控制	未控制	已控制	未控制	已控制	未控制	已控制
家庭背景	未控制	已控制	未控制	已控制	未控制	已控制	未控制	已控制
年份固定效应	已控制	已控制	已控制	已控制	已控制	已控制	已控制	已控制
省份固定效应	已控制	已控制	已控制	已控制	已控制	已控制	已控制	已控制
外地城乡户籍			132.328	5.341			162.606	6.332
系数差异检验			[0.000]	[0.028]			[0.000]	[0.018]
观测值	8767	8767	8767	8767	8570	8570	8570	8570
调整后的 R^2	0.229	0.385	0.229	0.385	0.187	0.353	0.187	0.353

注：①圆括号内为异方差稳健标准误。②***、**和*分别表示 1%、5%和 10%的显著性水平。③“外地城乡户籍系数差异检验”指外地城镇户籍估计系数与外地农业户籍估计系数相等的 Wald 检验，[ ]中为 p 值。

3.考虑迁移因素的 DML 估计。为了使模型能更好地满足 CIA，研究者应尽可能以更加灵活的方式控制所有干扰因素。参照现有研究的做法（Chernozhukov et al.，2018；Knaus，2021；Bonaccolto-Töpfer and Briel，2022），本文在纳入表 3 的控制变量的基础上，进一步加入其他潜在的控制变量以及所有控制变量的两两交互项和二次项。除常数项及年份和省份虚拟变量外，共计加入 6897 个控制变量。为避免高度共线性，本文对变量进行如下处理：删除 99%分位数和 1%分位数差值小于1×10^{-6}的变量（2012－2015 年和 2017－2021 年两组样本分别删除 969 个和 1053 个变量）；删除相对标准差[①]最小的 10%变量（2012－2015 年和 2017－2021 年两组样本分别删除 595 个和 587 个变量）；删除相关系数高于 0.99 的两个变量中的任意一个（2012－2015 年和 2017－2021 年两组样本分别删除 240 个和 244 个变量）。最终，2012－2015 年和 2017－2021 年两组样本中用于进行 DML 估计的变量分别为 5093 个和 5013 个。

为提高 DML 估计的有效性，遵循 Chernozhukov et al.（2018）的建议，本文采用交叉拟合方法进行估计。基准回归结果中将总样本进行三等分用于交叉拟合，后文将考虑其他等分方式下估计结果的稳健性。在估计表 4（1）列和（3）列结果时，需要先进行小时工资的 Lasso 估计、是否农业户籍的 Lasso 估计和是否迁移人口的 Lasso 估计。表 4（1）列和（3）列汇报的是 DML 估计的最后一步结果，即三个 Lasso 估计的交叉拟合残差回归，因此模型调整后的 R^2 相较表 3 更小。类似地，在估计表 4（2）列和（4）列结果时，需要进行小时工资的 Lasso 估计、外地城镇户籍的 Lasso 估计、本地农业户籍的 Lasso 估计和外地农业户籍的 Lasso 估计，表 4（2）列和（4）列汇报的依然是 DML 估计的最后一步结果。

表 4（1）列结果表明，农业户籍对农民工小时工资的负面影响不仅在经济显著性上相较表 3（2）列有所上升，统计显著性也更强。表 4（3）列结果进一步证实农业户籍对农民工小时工资造成了显著的负面影响。考虑迁移对工资影响的异质性后，农业户籍对本地农民工小时工资的负面影响依然显著。此外，与表 3 结果相比，表 4 结果更加稳健。例如，表 3（2）列和（6）列中是否农业户籍变量的系数绝对值差异更大，而表 4（1）列和（3）列结果中是否农业户籍变量的系数绝对值差异更小。类似地，表 4（2）列和（4）列结果的差异相较表 3 也更小。2012－2015 年和 2017－2021 年两组样本的时间跨度不大，户籍歧视在两个样本期内发生显著变化的可能性也不大，更加稳健的结果表明 DML 估计结果更加可信。据此，本文认为表 4 的结果更加真实地反映了中国劳动力市场中的户籍歧视状况。

表 4　　农业户籍对劳动者小时工资收入影响的 DML 估计结果

变量	被解释变量：小时工资							
	2012－2015 年				2017－2021 年			
	（1）		（2）		（3）		（4）	
	系数	标准误	系数	标准误	系数	标准误	系数	标准误
是否农业户籍	−0.051**	0.020			−0.070***	0.022		

[①]根据 Belloni et al.（2014）的研究，相对标准差为原始变量的原始标准差除以变量 99%分位数和 1%分位数的差值得到的商。

表 4（续）

是否迁移人口	0.112***	0.020			0.087***	0.021		
外地城镇户籍			0.139***	0.029			0.081**	0.032
本地农业户籍			−0.045**	0.022			−0.065***	0.023
外地农业户籍			0.049*	0.028			0.030	0.029
外地城乡户籍系数			6.102				1.841	
差异检验			[0.014]				[0.178]	
观测值	8767		8767		8570		8570	
调整后的 R^2	0.094		0.094		0.076		0.075	

注：①标准误为异方差稳健标准误。②***、**和*分别表示 1%、5%和 10%的显著性水平。③“外地城乡户籍系数差异检验”指外地城镇户籍估计系数与外地农业户籍估计系数相等的 Wald 检验，[ ]中为 p 值。④本表为 DML 估计的最后一步结果，被解释变量为小时工资 Lasso 估计的交叉拟合残差，每个解释变量均为对应 Lasso 估计的交叉拟合残差。

（二）稳健性讨论

本文从四个方面对研究结论的稳健性进行讨论。一是调整被解释变量，用全年收入和全年工资收入替换小时工资；二是在 DML 估计中采用其他样本分割方式进行交叉拟合；三是用其他机器学习算法替代 Lasso 算法；四是采用 Oster（2019）的系数稳定性检验和 Lewbel（2012）的工具变量估计审查潜在的遗漏变量对基本结论的影响。

1.替换被解释变量。小时工资未必能很好地反映劳动者在劳动力市场中获得的全部回报，例如兼职创业者的创业收入可能不在工资收入中统计。部分研究甚至认为迁移人口的被迫创业行为是导致迁移溢价现象的关键原因之一（刘金东等，2020；黄静和祝梦迪，2022）。这是不是意味着使用全年收入就会发现农业户籍人口被优待的现象？鉴于此，本文将被解释变量替换为全年收入，并进行回归，结果如表 5 所示。表 5 结果显示：（1）列和（2）列中是否农业户籍和本地农业户籍系数为正，但统计上不显著；使用 DML 估计后，是否农业户籍和本地农业户籍对全年收入的影响仍然不显著；而 2017−2021 年样本的估计结果仍然证实农业户籍人口在获得收入时面临一定的户籍歧视。因此，总体而言，本文的基准结论是稳健的。

表 5　　农业户籍对劳动者全年收入影响的估计结果

变量	被解释变量：全年收入							
	2012−2015 年				2017−2021 年			
	OLS	OLS	DML	DML	OLS	OLS	DML	DML
	（1）	（2）	（3）	（4）	（5）	（6）	（7）	（8）
是否农业户籍	0.004		−0.012		−0.053***		−0.027	
	（0.019）		（0.019）		（0.018）		（0.018）	
是否迁移人口	0.132***		0.164***		0.130***		0.139***	
	（0.020）		（0.019）		（0.019）		（0.018）	

表 5（续）

外地城镇户籍		0.152***		0.169***		0.112***		0.119***
		（0.027）		（0.027）		（0.027）		（0.027）
本地农业户籍		0.012		−0.020		−0.060***		−0.043**
		（0.021）		（0.020）		（0.019）		（0.020）
外地农业户籍		0.125***		0.136***		0.082***		0.111***
		（0.026）		（0.029）		（0.026）		（0.025）
观测值	10044	10044	10044	10044	9767	9767	9767	9767
调整后的 R^2	0.366	0.366	0.100	0.100	0.381	0.381	0.082	0.082

注：①括号内为异方差稳健标准误。②***和**分别表示 1%和 5%的显著性水平。③DML 估计结果为 DML 估计的最后一步结果，被解释变量为全年收入 Lasso 估计的交叉拟合残差，每个解释变量均为对应 Lasso 估计的交叉拟合残差。

劳动者的小时工资数据需要利用全年工资收入和每周工作小时数进行估算，这一过程不仅会导致样本量损失，还有可能引入更严重的测量误差。为此，本文用全年工资收入作为被解释变量，重新估计表 4 中对应方程的结果，估计结果如表 6 所示。从 2012－2015 年的 DML 估计结果看：（3）列中是否农业户籍的系数为–0.030，虽然统计上不显著，但其 p 值仅为 0.110；（4）列中本地农业户籍的系数为–0.038，在 10%统计水平上显著。可见，表 6（3）列和（4）列的结果与表 4（1）列和（2）列的结果基本一致。同样，与表 4（3）列和（4）列结果对比，表 6（7）列和（8）列的结果也支持了本文的基本结论。

当然，直接使用全年工资收入进行估计有可能低估户籍歧视效应。例如，农民工为追求更高总收入而采取过度劳动行为（刘涛等，2023；卢文秀和吴方卫，2023），这会掩盖工资收入中的户籍歧视现象。通过对比表 6 与表 3 中是否农业户籍和本地农业户籍的 OLS 估计系数以及表 6 与表 4 中是否农业户籍和本地农业户籍的 DML 估计系数可以发现，使用全年工资收入作为被解释变量确实有可能低估农业户籍对农民工工资的负面影响。这也说明，农民工和城镇职工在就业方面的差距不仅表现在收入层面，也表现在就业质量层面。农民工更有可能通过较长的劳动时间来获得更高的工资收入，这会在一定程度上掩盖全年工资收入中反映出来的户籍歧视程度。因此，实际的户籍歧视效应可能比表 6 结果所展示出来的更大。

表 6　　农业户籍对劳动者全年工资收入影响的估计结果

变量	被解释变量：全年工资收入							
	2012－2015 年				2017－2021 年			
	OLS	OLS	DML	DML	OLS	OLS	DML	DML
	（1）	（2）	（3）	（4）	（5）	（6）	（7）	（8）
是否农业户籍	−0.012		−0.030		−0.062***		−0.037**	
	（0.020）		（0.019）		（0.018）		（0.019）	
是否迁移人口	0.138***		0.173***		0.127***		0.138***	
	（0.020）		（0.019）		（0.019）		（0.019）	

表 6（续）

外地城镇户籍		0.156***		0.171***		0.101***		0.101***
		（0.028）		（0.027）		（0.028）		（0.028）
本地农业户籍		−0.004		−0.038*		−0.072***		−0.055***
		（0.021）		（0.020）		（0.020）		（0.020）
外地农业户籍		0.117***		0.130***		0.074***		0.099***
		（0.026）		（0.026）		（0.026）		（0.026）
观测值	10044	10044	10044	10044	9767	9767	9767	9767
调整后的 R^2	0.364	0.364	0.099	0.099	0.368	0.368	0.079	0.079

注：①括号内为异方差稳健标准误。②***、**和*分别表示 1%、5%和 10%的显著性水平。③DML 估计结果为 DML 估计的最后一步结果，被解释变量为全年工资收入 Lasso 估计的交叉拟合残差，每个解释变量均为对应 Lasso 估计的交叉拟合残差。

2.*调整交叉拟合方法*。为降低过度拟合偏差对 DML 估计的影响，本文遵循 Chernozhukov et al.（2018）的建议，采用交叉拟合估计方法。交叉拟合需要将总样本分割成若干份，为验证本文的基本结论和样本分割份数无关，本文进一步考察其他样本分割份数下的估计结果。由于总样本分割份数越多，估计过程越耗时，本文仅考虑了 5 等分、7 等分和 9 等分这 3 种分割方式[①]。结果表明，无论采取哪种分割方式，是否农业户籍和本地农业户籍对农民工各类收入基本表现出负向影响。而且，是否农业户籍和本地农业户籍对小时工资的负向影响最大，对全年收入的负向影响最小。这些结果均和前文的基本结论一致。

3.*使用其他机器学习算法*。另一个可能影响本文结论的问题是机器学习算法的选择。前文用 Lasso 方法进行 DML 估计，但理论上 DML 的机器学习环节也可以用其他有监督的机器学习算法。为证明本文的基本结论和学习算法选择无关，本部分采用其他常规算法重新估计表 4、表 5 和表 6 中对应方程的结果。表 7 展示了随机森林、岭回归和弹性网络这三种除 Lasso 外常规的机器学习算法用于 DML 估计的结果。可以看出，改变机器学习算法对本文的基本结论没有造成明显的影响。

表 7　　机器学习算法选择对 DML 估计结果影响的检验结果

变量	机器学习算法	2012−2015 年			2017−2021 年		
		小时工资（1）	全年收入（2）	全年工资收入（3）	小时工资（4）	全年收入（5）	全年工资收入（6）
是否农业户籍	随机森林	−0.071***	−0.013	−0.031*	−0.101***	−0.046***	−0.054***
		（0.020）	（0.018）	（0.018）	（0.021）	（0.018）	（0.018）
	岭回归	−0.043**	−0.008	−0.023	−0.061***	−0.020	−0.028
		（0.021）	（0.019）	（0.019）	（0.021）	（0.018）	（0.019）
	弹性网络	−0.039*	−0.006	−0.023	−0.062***	−0.021	−0.029
		（0.021）	（0.019）	（0.019）	（0.021）	（0.018）	（0.018）

[①]实践中通常采用较小的样本分割份数，例如，Chernozhukov et al.（2018）在经验分析环节将总样本划分为 2 份。

表 7（续）

本地农业户籍	随机森林	−0.062***	−0.011	−0.030	−0.106***	−0.058***	−0.068***
		（0.022）	（0.019）	（0.019）	（0.023）	（0.019）	（0.020）
	岭回归	−0.036	−0.007	−0.022	−0.058**	−0.036*	−0.048**
		（0.022）	（0.020）	（0.020）	（0.023）	（0.020）	（0.020）
	弹性网络	−0.036	−0.006	−0.022	−0.060***	−0.032*	−0.044**
		（0.022）	（0.020）	（0.020）	（0.023）	（0.019）	（0.020）

注：①括号内为异方差稳健标准误。②***、**和*分别表示 1%、5%和 10%的显著性水平。

4.审查潜在遗漏变量的影响。遗漏重要变量导致模型无法满足 CIA，是识别户籍歧视的最大挑战。本文采用 DML 方法的主要目的在于借助机器学习算法，从包含大量控制变量的灵活模型中识别出最佳模型，从而准确估计是否农业户籍或本地农业户籍两个核心变量对劳动者收入的影响。换言之，采用 DML 估计应该能更好地满足 CIA。劳动者在不同户籍状态下的潜在收入永远无法完全被观测到，所以 CIA 无法直接验证。但是，可以利用 Oster（2019）提供的系数稳定性检验方法间接判断 DML 估计和传统 OLS 估计哪个更有可能满足 CIA。Oster 检验无法直接应用于 DML 估计结果，因此，本文借鉴 Bonaccolto-Töpfer and Briel（2022）的做法，用 OLS 估计（4）式，但控制变量调整为 DML 估计结果中所有系数不为 0 的变量，以此判断 DML 是否能够达到更好的变量选择效果。

表 8 是 Oster 检验结果，前两行分别对应表 4（2）列和（4）列的 Oster 检验，后两行分别对应表 3（4）列和（8）列的 Oster 检验。从后两行结果可以看出，无论是 Beta 边界检验还是 Delta 检验，都有强烈的证据表明 OLS 结果是不稳健的。以 2012－2015 年样本的 OLS 估计结果为例，Oster 集不仅包含了 0，也没有完全落在 $\hat{\beta}$ 的 99.5%置信区间之内。此外，δ 估计值为 0.118，说明遗漏变量的重要性只要达到观测变量重要性的 0.118 倍，就有可能在真实户籍歧视为 0 的情况下看到表 3 的结果。遗漏变量的重要性相当于观测变量的 0.118 倍这一要求很低，所以对应的 OLS 估计结果很有可能是不可信的。2017－2021 年样本的估计结果中虽然 Oster 集更小并且 δ 估计值更大，但仍然不能排除遗漏变量会影响本文结论的可能性。而在对 DML 估计结果的 Oster 检验中，2017－2021 年样本的估计结果顺利通过了 Oster 检验。δ 估计值为 1.728 说明，遗漏变量的重要性需要达到观测变量重要性的 1.728 倍，才有可能导致在真实户籍歧视为 0 的情况下错误地估计出核心解释变量系数不为 0 的结果。由于本文已经控制了大量个体、工作和家庭背景层面的特征，遗漏变量依然如此重要的可能性并不大。换言之，DML 估计结果因遗漏变量导致结果不可信的可能性不大。2012－2015 年样本的估计结果虽然没有通过 Oster 检验，但估计结果受遗漏变量严重影响的证据并不明显。具体来说，2012－2015 年样本估计结果的 Oster 集在 $\hat{\beta}$ 的 99.5%置信区间内，且 Oster 集的上界只是略微超过 0，δ 估计值也非常接近 1。因此，有理由相信 DML 估计违背 CIA 的可能性更小。

表 8　　Oster 遗漏变量偏误检验结果

估计方法	样本期	Beta 边界检验			Delta 检验	
		Oster 集	$\hat{\beta}$ 的 99.5%置信区间	是否通过	δ	是否通过
DML	2012—2015 年	[−0.051, 0.001]	[−0.114, 0.013]	否	0.992	否
	2017—2021 年	[−0.068, −0.040]	[−0.134, −0.003]	是	1.728	是
OLS	2012—2015 年	[−0.032, 0.283]	[−0.093, 0.029]	否	0.118	否
	2017—2021 年	[−0.093, 0.161]	[−0.158, −0.027]	否	0.428	否

前文分析表明，以往的户籍歧视研究中使用的模型可能无法满足 CIA。这意味着 OLS、PSM 和 Oaxaca-Blinder 分解等基于 CIA 的常规研究方法在识别户籍歧视时均存在一定缺陷。当 CIA 不满足时，工具变量估计是常见的识别策略之一。但由于缺乏有效的工具变量，这一识别策略在户籍歧视问题中鲜有使用。Lewbel（2012）提供了一种基于异方差构造工具变量的思路。为保证所构造的工具变量为有效工具变量，参照 Lewbel（2012）的建议，本文选择家庭背景、性别、年龄等比较外生的控制变量为构造工具变量的基础变量。表 9 是 Lewbel 工具变量估计结果。所有结果中本地农业户籍的系数均小于 0，且大部分结果均在 10%以上统计水平上显著，说明本地农业户籍对劳动力收入仍有显著负向影响，这些结果进一步支持了前面的基本结论。此外，从表 9 各列的估计系数看，本地农业户籍对劳动力收入的负向影响均大于前文中对应模型的估计结果。这一结果说明 OLS 估计可能低估了户籍歧视的效应，这也和 DML 估计得出的结论一致。此外，从表 9 的 Hansen J 统计量看，所有模型均通过了过度识别检验，说明工具变量估计结果是可靠的。

表 9　　Lewbel 工具变量估计结果

变量	2012—2015 年			2017—2021 年		
	小时工资	全年收入	全年工资收入	小时工资	全年收入	全年工资收入
外地城镇户籍	0.098	0.211***	0.212***	0.059	0.102**	0.096**
	（0.068）	（0.061）	（0.062）	（0.053）	（0.044）	（0.045）
本地农业户籍	−0.056	−0.081*	−0.082*	−0.150*	−0.199***	−0.191***
	（0.059）	（0.049）	（0.049）	（0.079）	（0.068）	（0.069）
外地农业户籍	0.086	0.112*	0.122**	−0.003	−0.034	0.007
	（0.072）	（0.059）	（0.059）	（0.099）	（0.085）	（0.087）
观测值	8767	10044	10044	8570	9767	9767
调整后的 R^2	0.384	0.364	0.362	0.352	0.377	0.365
Hansen J 统计量	45.752	45.928	52.211	55.795	52.898	51.448
Hansen J 统计量 p 值	0.441	0.434	0.214	0.130	0.196	0.236

注：①括号内为异方差稳健标准误。②***、**和*分别表示 1%、5%和 10%的显著性水平。

（三）Oaxaca-Blinder 分解

当 CIA 成立时，Oaxaca-Blinder 分解也可以给出对户籍歧视效应的无偏估计。与前文 Oster 检验中的做法相同，为让 Oaxaca-Blinder 分解更好地满足 CIA，本文参照现有研究（Bonaccolto-Töpfer and Briel，

2022；Bach et al.，2024）的做法，将（6）式的 Lasso 估计中系数不为 0 的解释变量和（5）式的 Lasso 估计中系数不为 0 的解释变量合并，作为 Oaxaca-Blinder 分解中的解释变量。表 10 展示了 Oaxaca-Blinder 分解结果。表 10 结果表明，2012－2015 年农民工和城镇职工小时工资、全年收入和全年工资收入差距中不可解释部分占比分别为 15.851%、3.585%和 8.843%，全年收入差距中的不可解释部分占比较低，而小时工资差距中的不可解释部分占比较高。2017－2021 年样本的估计结果也呈现类似的结论。总体而言，农民工和城镇职工的收入差距中大约有 8%～15%的部分不可由个体特征差异解释。这说明，户籍歧视导致的城镇职工与农民工收入差距仍然存在，更没有证据表明农业户籍人口受到优待。和以往研究结论相比，现阶段城镇职工与农民工收入差距中的户籍歧视成分有所下降，且主要表现在小时工资层面。

表 10　　Oaxaca-Blinder 分解结果

组成部分	2012－2015 年			2017－2021 年		
	小时工资	全年收入	全年工资收入	小时工资	全年收入	全年工资收入
总差异	−0.443	−0.311	−0.318	−0.449	−0.328	−0.325
可解释部分	−0.373	−0.300	−0.290	−0.381	−0.300	−0.291
不可解释部分	−0.070	−0.011	−0.028	−0.068	−0.028	−0.035
不可解释部分占比（%）	15.851	3.585	8.843	15.239	8.514	10.616

五、结论与启示

随着中国劳动力市场供需结构的变化，农民工在城市劳动力市场中获得的收入快速上升，城镇职工和农民工的收入差距中无法用劳动者特征差异解释的部分逐渐下降。换言之，劳动力市场针对农业户籍劳动者的歧视现象正在消失。然而，最近的研究对于中国劳动力市场中是否仍然存在户籍歧视、农业户籍劳动者是否确实受到了优待等问题还存在较大争议。本文利用因果机器学习理论的最新识别策略和 CGSS 的最新 6 轮调查数据，重新对城镇职工和农民工收入差距中的户籍歧视现象进行了检验。

经验分析发现，以往研究关于城镇职工和农民工收入差距中是否存在户籍歧视的结论存在争议的主要诱因有两点：一是未充分考虑迁移溢价对户籍歧视识别的潜在干扰；二是实证研究中采用了基于 CIA 的识别方法，但数据不能很好地满足 CIA，导致对户籍歧视的识别存在偏误。针对上述问题，本文采取了两个措施：一是将迁移因素纳入模型，二是采用因果机器学习体系下的 DML 估计策略。本文研究发现，迁移因素在一定程度上掩盖了户籍歧视效应。DML 估计结果表明，中国劳动力市场中仍然存在比较显著的户籍歧视。农民工和城镇职工的全年总收入差距中户籍歧视成分较低，统计上的显著性也较低；在小时工资差距上，户籍歧视表现得更加强烈，统计上也更显著。本文用 Oster 检验证实了 DML 识别策略确实可以让模型违背 CIA 的可能性更低，Lewbel 工具变量估计结果也进一步证实了针对农业户籍劳动者的户籍歧视依然存在。本文在 DML 估计基础上利用 Oaxaca-Blinder 分解方法测算了户籍歧视因素在农民工和城镇职工收入差距中的解释能力，结果表明，农民工和城镇职工收入差距之中仍然有 8%～15%的部分是由户籍歧视导致的。

从政策层面看，本文的结论进一步支持了户籍制度改革尚未完成这一观点。这也意味着，实质性推动户籍制度改革，从根本上破除城乡二元结构，仍是未来一段时期内的重点任务。首先，要加强制度建设，建立和完善覆盖全民的社会福利体系，促进基本公共服务均等化。其次，要进一步规范劳动力市场运行机制，破除劳动力市场分割的因素，保护劳动者权益。最后，要加强对农民工的职业技能培训，最大程度防止“前市场歧视”传导为最终的户籍歧视。

本文仍存在可完善和可推进的空间，主要体现在以下三个方面：第一，本文只考察了平均收入水平上的户籍歧视，没有进一步延伸到整个收入分布的其他位置，未来的研究可将 DML 方法拓展到收入分布不同位置上的户籍歧视识别。第二，本文只探讨了收入层面的户籍歧视问题，没有考虑就业、职业选择等其他方面的户籍歧视，未来的研究可以将 DML 方法拓展到劳动力市场的多个方面或更一般的就业层面的户籍歧视。第三，因果机器学习理论的快速发展为大量因果识别问题提供了新方向，本文是因果机器学习方法在实践中运用的初步尝试，未来的研究可以进一步扩大该方法的运用范围。

参考文献

1.蔡昉，2023：《户籍制度改革的效应、方向和路径》，《经济研究》第 10 期，第 4-14 页。

2.陈斌开、林毅夫，2013：《发展战略、城市化与中国城乡收入差距》，《中国社会科学》第 4 期，第 81-102 页。

3.陈昊、赵春明、杨立强，2017：《户籍所在地“反向歧视之谜”：基于收入补偿的一个解释》，《世界经济》第 5 期，第 173-192 页。

4.陈珣、徐舒，2014：《农民工与城镇职工的工资差距及动态同化》，《经济研究》第 10 期，第 74-88 页。

5.邓曲恒，2007：《城镇居民与流动人口的收入差异——基于 Oaxaca-Blinder 和 Quantile 方法的分解》，《中国人口科学》第 2 期，第 8-16 页。

6.胡凤霞、叶仁荪，2019：《农民工与城镇职工的工资差距及其趋同——基于 CHIP 数据的实证分析》，《人口与经济》第 1 期，第 31-41 页。

7.黄静、祝梦迪，2022：《外来劳动力的工资溢价研究》，《经济学动态》第 4 期，第 67-82 页。

8.刘金东、秦子洋、孔培嘉，2020：《流动人口享受工资溢价了吗？——对户籍来源地“反向歧视之谜”的再检验》，《经济学动态》第 12 期，第 92-105 页。

9.刘涛、秦志龙、伍骏骞，2023：《农民工过度劳动行为的同群效应研究》，《中国农村经济》第 9 期，第 101-121 页。

10.卢文秀、吴方卫，2023：《患寡亦患不均：双轨制基本养老保险与农民工过度劳动》，《中国农村经济》第 7 期，第 100-123 页。

11.吕炜、杨沫、朱东明，2019：《农民工能实现与城镇职工的工资同化吗？》，《财经研究》第 2 期，第 86-99 页。

12.孟凡强、刘志辉、彭志勇，2022：《政府推动的就近城镇化能够消除工资歧视吗》，《南方经济》第 11 期，第 92-108 页。

13.孙婧芳，2017：《城市劳动力市场中户籍歧视的变化：农民工的就业与工资》，《经济研究》第 8 期，第 171-186 页。

14.孙三百，2015：《城市移民的收入增长效应有多大——兼论新型城镇化与户籍制度改革》，《财贸经济》第 9 期，第 135-147 页。

15.王美艳，2005：《城市劳动力市场上的就业机会与工资差异——外来劳动力就业与报酬研究》，《中国社会科学》第 5 期，第 36-46 页。

16.吴彬彬、章莉、孟凡强，2020：《就业机会户籍歧视对收入差距的影响》，《中国人口科学》第 6 期，第 100-111 页。

17.吴珊珊、孟凡强，2019：《农民工歧视与反歧视问题研究进展》，《经济学动态》第 4 期，第 99-111 页。

18.谢嗣胜、姚先国，2006：《农民工工资歧视的计量分析》，《中国农村经济》第 4 期，第 49-55 页。

19.邢春冰，2008：《农民工与城镇职工的收入差距》，《管理世界》第 5 期，第 55-64 页。

20.邢春冰、屈小博、杨鹏，2021：《农民工与城镇职工工资差距演变及原因分析》，《经济学动态》第 5 期，第 64-78 页。

21.徐凤辉、赵忠，2014：《户籍制度和企业特征对工资收入差距的影响研究》，《中国人民大学学报》第 3 期，第 19-28 页。

22.许岩，2022：《市民化与农业转移人口的共同富裕——对“农转非”居民工资同化过程的分析》，《人口与经济》第 3 期，第 130-148 页。

23.于潇、陈筱乐、解瑯卓，2022：《流动效应与户籍歧视效应对流动人口工资收入的影响——基于双边随机前沿模型的分析》，《人口研究》第 2 期，第 61-74 页。

24.于潇、孙悦，2017：《城镇与农村流动人口的收入差异——基于 2015 年全国流动人口动态监测数据的分位数回归分析》，《人口研究》第 1 期，第 84-97 页。

25.章莉、李实、William A.Darity Jr.、Rhonda Vonshay Sharpe，2014：《中国劳动力市场上工资收入的户籍歧视》，《管理世界》第 11 期，第 35-46 页。

26.章莉、吴彬彬，2019：《就业户籍歧视的变化及其对收入差距的影响：2002－2013 年》，《劳动经济研究》第 3 期，第 84-99 页。

27.Bach, P., V. Chernozhukov, and M. Spindler, 2024, “Heterogeneity in the US Gender Wage Gap”, *Journal of the Royal Statistical Society Series A: Statistics in Society*, 87(1): 209-230.

28.Belloni, A., V. Chernozhukov, and C. Hansen, 2014, “Inference on Treatment Effects After Selection Among High-Dimensional Controls”, *The Review of Economic Studies*, 81(2): 608-650.

29.Bonaccolto-Töpfer, M., and S. Briel, 2022, “The Gender Pay Gap Revisited: Does Machine Learning Offer New Insights?”, *Labour Economics*, Vol. 78, 102223.

30.Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, 2018, “Double/Debiased Machine Learning for Treatment and Structural Parameters”, *The Econometrics Journal*, 21(1): C1-C68.

31.Knaus, M. C., 2021, “A Double Machine Learning Approach to Estimate the Effects of Musical Practice on Student’s Skills”, *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(1): 282-300.

32.Lewbel, A., 2012, “Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models”, *Journal of Business & Economic Statistics*, 30(1): 67-80.

33.Ma, X., 2018, “Labor Market Segmentation by Industry Sectors and Wage Gaps Between Migrants and Local Urban Residents in Urban China”, *China Economic Review*, Vol. 47: 96-115.

34.Meng, X., and J. Zhang, 2001, "The Two-Tier Labor Market in Urban China: Occupational Segregation and Wage Differentials Between Urban Residents and Rural Migrants in Shanghai", *Journal of Comparative Economics*, 29(3): 485-504.

35.Oster, E., 2019, "Unobservable Selection and Coefficient Stability: Theory and Evidence", *Journal of Business & Economic Statistics*, 37(2): 187-204.

（作者单位：[1]郑州大学商学院；
[2]中南财经政法大学财政税务学院；
[3]中南财经政法大学收入分配与现代财政学科创新引智基地）

（责任编辑：胡　祎）

Does the Income Gap Caused by Hukou Discrimination Still Exist? A Re-discussion Based on Machine Learning Methods

JIANG Qiuchuan　LU Yuanping

Abstract: Changes in the supply-demand structure of China's labor market have led to a new trend in the income gap between migrant workers and urban employees, sparking increasing debates on whether Hukou-based discrimination still exists. This paper re-examines the phenomenon of Hukou-based discrimination in the income gap between migrant workers and urban employees using a doubly debiased machine learning approach. The empirical analysis reveals the following findings. (1) Migration premium interferes with the identification of Hukou-based discrimination, and the phenomenon of Hukou-based discrimination becomes more apparent after accounting for the factor of migration premium. (2) After applying the doubly debiased machine learning method to select models that better meet the conditional independence assumption, the agricultural household registration has a negative impact on the laborers' hourly wage income, the annual total income and the annual wage income, with a more significant negative effect on the hourly wage income. (3) The Oaxaca-Blinder decomposition, corrected by doubly debiased machine learning, indicates that approximately 8% to 15% of the income gap between migrant workers and urban employees can be attributed to Hukou-based discrimination. (4) The Oster test confirms that the estimation of doubly debiased machine learning is more reliable than the OLS estimation, and the doubly debiased machine learning estimation and Lewbel's instrumental variable estimation under different machine learning algorithms also demonstrate the robustness of the conclusions drawn in this paper.

Keywords: Hukou Discrimination; Machine Learning; Income Gap; Migrant Workers