环境因素、教育公平与收入差距

——来自去偏差机器学习的新发现

张　晨　李沫霖　张　琦　吴　煜*

内容提要　教育公平是社会公平的基础，是扎实推进共同富裕的必然要求。本文从机会公平的视角关注教育均等性，呈现中国教育机会差距的客观现实，并评估教育机会差距对收入分配的影响。以2010–2021年中国综合社会调查数据为研究样本，本文采用去偏差机器学习方法测算教育机会差距。研究发现，个体特征和父代属性所代表的环境因素解释了超过70%的教育差距，样本期内教育公平性获得显著改善。年龄、户口和性别是影响子代教育最为重要的三项因素，其中年龄背后可能隐含了教育扩张的影响。城乡内部均存在显著的教育机会差距，女性群体面临的教育机会差距显著大于男性。基于可解释机器学习模型的评估结果表明，教育因素对收入差距的贡献度约为13%，其中环境因素占据主导地位，通过教育引致了约10%的收入差距。根据研究发现，本文提出了促进教育机会公平的政策建议。

关 键 词　环境因素　教育机会公平　收入分配　机器学习

一　引言

党的二十大报告提出，“坚持以人民为中心发展教育，加快建设高质量教育体系，发展素质教育，促进教育公平”。教育不仅为个人带来经济回报，也为经济社会高质量

*　张晨，山东财经大学财政税务学院，电子邮箱：20203731@sdufe.edu.cn；李沫霖（通讯作者），山东财经大学财政税务学院，电子邮箱：li_molin@163.com；张琦，中国社会科学院大学应用经济学院，电子邮箱：zhangqi10@ucass.edu.cn；吴煜，厦门国际银行，电子邮箱：2777913876@qq.com。本研究得到山东省社会科学规划研究项目（24CJJJ20）、国家社会科学基金青年项目（21CZZ028）、河北省高等教育教学改革研究与实践项目（2022GJJG179）的资助。

发展提供人力资本和创新源泉。劳动力市场将不公平的教育分配机制及其结果转化为经济差距，并进一步外溢至健康（Lleras-Muney，2005）等其他领域。如果教育的获取受制于个人无法掌控的环境因素（如户口、性别、种族、家庭背景等），那么许多群体将失去通过自身努力改善生活的机会，最终阻碍全社会人力资本的积累进程。

本文关注环境因素在教育获取差异中所扮演的重要角色。诸多文献强调父代经济资源和教育背景的显著影响（周康等，2025）。机会均等化的思想由Fleurbaey（1995）、Van de Gaer（1993）从哲学引入经济学。随后Roemer（1998）提出了机会均等化的强标准：结果分布独立于个人无法掌控的环境因素，付出同等努力的个体将会取得相同的结果[①]。Lefranc et al.（2008）使用随机占优对其进行检验。然而这一方法要求每一环境类型拥有大量样本，因此限制了环境因素的数量。Van de Gaer（1993）提出了一种不需要估计结果条件分布，同时允许纳入大量环境因素的弱标准：通过相同环境类型群体的平均结果差异而非分布差异衡量机会差距。通过抹平环境类型内部差异，从而考察环境类型间的平均水平差异，为机会差距的测算提供了实证思路，该思路至今仍被广泛使用，也是本文所遵循的基本框架。

对于弱标准的评估，少量文献通过统计检验判断机会公平性是否存在（Kanbur & Snell，2019），大量研究则通过划分环境类型量化机会均等化的程度（Almås et al.，2011）。本文同样关注量值测度而非存在性检验。从测算角度来看，对环境和努力因素的区分形成了机会差距的事前和事后估计。事前估计只关注环境因素（Fleurbaey & Peragine，2013），将样本分配至不同的环境类型，结果的组间均值差异即为机会差距。事后估计则关注努力因素（Juárez & Soloaga，2014），机会差距由相同努力程度的组内结果差异进行衡量。由于努力因素通常难以观测和度量，事前估计方法在实证研究中应用更为广泛。从测算方法来看，可通过非参数方法对样本进行分组，但面临维度诅咒问题。参数方法则利用包含环境和努力因素的结果方程进行反事实预测。虽然非参数方法更加符合机会均等化的定义且应用灵活，但受限于样本量和高维问题，更多文献使用参数方法进行实证研究（史新杰等，2022）。

本文仍采用Roemer（1998）提出的“环境—努力”二元分析框架，从事前角度利用非参数方法测算教育机会差距。本文认为，机会差距的测算本质上属于一个基于环境和努力因素对结果进行预测的问题。现有参数方法（例如线性回归模型）由于存在欠拟合问题，在预测任务上相较于非参数方法（例如随机森林、深度学习）具有一定

① 不仅包括工作努力和学习努力等狭义的努力，还包括运气等因素。

劣势。当采用线性模型估计结果方程时，研究者需要自行决定哪些环境因素应被纳入模型。原因在于，普通线性回归模型缺乏自动筛选变量的机制，若将所有环境因素直接纳入，多重共线性和高维等问题将干扰系数估计，导致预测偏误。然而，人为筛选又可能遗漏重要环境变量。Ferreira & Gignoux（2011）指出，可观测的环境因素只是个体无法掌控且对结果具有重要影响的外生因素的子集，基于可观测子集的机会差距测算结果存在向下偏误。实际上，模型设定错误是制约线性模型预测效果的一个主要原因。无论从经济理论还是预测实践角度考虑，都应将环境因素的交互项和高阶项纳入预测模型。例如，在收入决定方程中应考虑年龄或经验的二次项，而不同性别、不同户籍等群体的结果方程应具有显著差异。如果模型不能考虑环境因素的交互影响及非线性效应，则会再次遗漏重要环境因素，导致机会差距进一步低估（Bourguignon et al.，2007）。

非参数方法过去受制于维度诅咒问题，难以被广泛应用。然而，通过引入正则化项，利用算法自动选择有用的环境因素，机器学习甚至可以用于解决高维问题。近年来，利用机器学习方法基于环境因素对结果变量进行预测已成为一种常用研究策略。Hufe et al.（2022）使用最小绝对收缩与选择算子评估了12个新兴经济体的机会差距。万相昱等（2024）、Brunori et al.（2023）在预测阶段使用条件推断森林以解决过拟合问题。然而，Escanciano & Terschuur（2022）指出，在预测阶段使用机器学习可能会对机会差距的测算引入显著偏差。其原因在于，为了提高泛化能力，机器学习需要在预测偏差与方差之间进行权衡，这意味着为了降低预测方差，机器学习允许在第一阶段的预测中出现偏差，而这种偏差将会蔓延至第二阶段的机会差距测算。

本文的边际贡献和增量工作体现在三个方面。第一，采用前沿的机器学习方法获得更具可信度的教育机会差距测算结果。相较于传统参数方法，Escanciano & Terschuur（2022）提出的去偏差机器学习方法具有两点优势：一是允许使用主流机器学习方法弥补传统参数方法的预测缺陷，更加精准地分离出环境因素的真实影响；二是直接应用机器学习进行机会差距测算将会产生统计偏误，去偏差机器学习方法通过纠偏保证机会差距估计值具有优良的统计性质，即渐进正态分布，同时允许对机会差距测算结果的不确定性进行量化评估。而无论采用参数方法还是非参数机器学习方法，现有文献均无法对估计结果（特别是组间差异）进行统计检验。理论上的无偏性与有效性提高了测算结果的可信度。本文的估计结果表明，已有研究显著低估了中国的教育机会差距。第二，对环境因素的偏效应进行估计和检验，为理解环境因素如何影响机会公平提供了新的视角。与已有研究不同，本文发现教育的代际传递并非决定子代教育的最

重要因素。相反，年龄、户口和性别等个人特征成为影响子代教育获取的主要因素。第三，本文为教育机会公平如何影响收入差距提供了量化结果。教育是环境作用于子代收入的重要渠道，但具体效应大小鲜有实证结论。本文将机器学习算法与收入分配效应评估的经典方法沙普利（Shapley）分解相结合，充分发挥非线性模型的预测优势，发现环境因素通过教育大约引致10%的收入差距，为理解教育机会差异的经济效应提供了新的经验证据。

二　机会差距测算方法

令 $W_i = (Y_i, X_i)$ 表示来自分布 F_0 的独立观测。其中，Y_i 为代表结果变量的标量，例如受教育年限，X_i 为环境因素构成的向量，i 为样本标识。$\gamma_0(X_i) = \mathbb{E}_{F_0}[Y_i|X_i]$ 表示结果变量 Y_i 的条件期望函数，例如环境因素决定的平均受教育年限。本文感兴趣的是随机变量 $\gamma_0(X_i)$ 的分布差异即机会差距。此处采用基尼系数作为机会差距的度量指标。

（一）去偏差机会差距

基于条件期望 $\gamma_0(X_i)$ 的基尼系数可表示为：

$$\theta_0(\gamma_0) = \frac{\mathbb{E}\left[\left|\gamma_0(X_i) - \gamma_0(X_j)\right|\right]}{\mathbb{E}\left[\gamma_0(X_i) + \gamma_0(X_j)\right]} \tag{1}$$

左右同时乘以分母并移项可将基尼系数测算转化为一个矩估计问题：

$$\mathbb{E}\left[\theta\left(\gamma_0(X_i) + \gamma_0(X_j)\right) - \left|\gamma_0(X_i) - \gamma_0(X_j)\right|\right] = 0, \text{ iff } \theta = \theta_0(\gamma_0) \tag{2}$$

显然，需要首先使用样本数据预测出 γ_0 即 $\hat{\gamma}$，然后代入式（2）求解 θ。对于此类代入估计量，上述两步法属于常规做法。但是，需要注意的是，机会差距的估计取决于 γ_0，对于预测偏误 $\gamma_0 - \hat{\gamma}$ 比较敏感：$\partial\theta_0(\gamma_0)/\partial\gamma_0 \neq 0$。

无论使用线性模型还是非线性机器学习算法对条件期望 γ_0 进行预测，不可避免将引入各种偏误。对于线性模型，偏误可能来自模型形式误设。对于机器学习算法，虽然通常不对模型形式进行约束，但是由于正则化项的存在，将会引致正则偏误。利用两步法求解式（2）无法保证机会差距估计量关于冗余参数 γ_0 具有局部稳健性，这一性质称之为内曼正交。据此，Escanciano & Terschuur（2022）提出一种新的矩条件：

$$\mathbb{E}\left[\theta\left(Y_i + Y_j\right) - sgn\left(\gamma_0\left(X_i\right) - \gamma_0\left(X_j\right)\right)\left(Y_i - Y_j\right)\right] = 0, \text{ iff } \theta = \theta_0\left(\gamma_0\right) \tag{3}$$

在Chernozhukov et al.（2018）提出的去偏差机器学习框架中，主要通过构建满足内曼正交性质的矩条件规避机器学习的正则偏误。利用U型矩的第一阶段影响函数表示，可证明基于式（3）的机会差距估计满足内曼正交性质，意味着至少能够实现估计量在真实值附近的局部稳健性，不会受到冗余参数γ_0预测偏误的影响。样本矩条件为：

$$\sum_i\sum_j\theta\left(Y_i + Y_j\right) - sgn\left(\hat{\gamma}\left(X_i\right) - \hat{\gamma}\left(X_j\right)\right)\left(Y_i - Y_j\right) = 0 \tag{4}$$

求解可得：

$$\theta = \frac{\sum_i\sum_j sgn\left(\hat{\gamma}\left(X_i\right) - \hat{\gamma}\left(X_j\right)\right)\left(Y_i - Y_j\right)}{\sum_i\sum_j\left(Y_i + Y_j\right)} \tag{5}$$

使用环境因素进行预测和根据式（5）估计机会差距两个步骤不能使用相同样本，原因在于机器学习算法所产生的过度拟合将引致向上偏误，对于估计结果的统计推断产生不良影响（Chernozhukov et al.，2018），因此采用交叉拟合方式进行预测和估计。基本思路分为以下四个步骤。第一，由于差距估计涉及样本间比较，因此需要将样本进行遍历形成配对样本$S_{(i,\ j)}$，其中i，$j \in [1, \cdots, n]$，$i < j$。第二，将配对样本$S_{(i,\ j)}$切分为L份：$I_1, \cdots, I_L$。令$\hat{\gamma}_l$表示使用非I_l子样本训练得到的条件期望函数，然后对I_l子样本进行结果变量预测并计算样本间差距，即$\hat{\gamma}_l\left(X_i\right) - \hat{\gamma}_l\left(X_j\right)$。第三，为了充分利用样本信息，交换样本用途，如此每份配对样本都将被用于进行一次差距估算。第四，对全部配对样本的$sgn\left(\hat{\gamma}\left(X_i\right) - \hat{\gamma}\left(X_j\right)\right)\left(Y_i - Y_j\right)$进行求和，可得去偏差估计量如下：

$$\hat{\theta} = \frac{\sum_{l=1}^{L}\sum_{(i,j)\in I_l} sgn\left(\hat{\gamma}_l\left(X_i\right) - \hat{\gamma}_l\left(X_j\right)\right)\left(Y_i - Y_j\right)}{\sum_i\sum_j\left(Y_i + Y_j\right)} \tag{6}$$

Escanciano & Terschuur（2022）证明去偏差估计量满足渐进正态分布：

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \rightarrow {}_dN\left(0, V\right)$$

其中，$V = \mathbb{E}\left[Y_i\right]^{-2}\mathbb{E}\left(\mathbb{E}\left[\theta_0\left(Y_i + Y_j\right) - sgn\left(\gamma_0\left(X_i\right) - \gamma_0\left(X_j\right)\right)\left(Y_i - Y_j\right)|Y_i,\ X_i\right]^2\right)$。

式（1）为机会差距绝对量，本文还通过计算机会差距占比$\theta_0^R = \theta_0/I$（I表示根据结果变量Y所测算的总体差距）刻画环境因素的重要影响。当得到绝对量估计值后，

相对占比的一致估计量为$\hat{\theta}^R = \hat{\theta}/\hat{I}$。

（二）环境因素的偏效应估计

机会差距测算通常需要纳入诸多环境因素，因此研究者需对环境因素的重要性进行评估。在应用机器学习进行机会差距测算时，通常在预测阶段评估环境因素对预测结果的影响。例如，Brunori et al.（2023）利用条件推断森林提供的变量重要性指标评估环境因素的影响。然而，此类方法并不能将环境因素与最终的机会差距结果直接关联，也就是说，仍然无法明确环境因素对机会公平的重要性。因此，有必要直接评估环境因素与机会差距的相关性。

假设存在M种环境变量$X_i \in R^M$。令$\theta_{0,-m}$表示未纳入第$m \in (1, \cdots, M)$种环境因素时的机会差距。定义环境因素的偏效应如下：

$$k_{0,m} = \theta_0 - \theta_{0,-m} \tag{7}$$

理论上$k_{0,m} \geqslant 0$，意味着包含更多的环境因素将会增加机会差距，除非某些因素对于结果预测不起作用，此时机会均等性将不会改变。同时说明基于部分可观测环境因素的机会差距测算只是真实差距的下界。然而，因为样本噪声的存在，上述分析在经验评估中可能存在出入。偏效应的估计量为：

$$\hat{k}_m = \hat{\theta} - \hat{\theta}_{-m} \tag{8}$$

由于已知$\hat{\theta}$和$\hat{\theta}_{-m}$的统计性质，因此可直接推导出两者之差的标准误和置信区间。为了进行组间比较，将相对偏效应定义如下：

$$k^R_{0,m} = \frac{\theta_0 - \theta_{0,-m}}{\theta_0} \tag{9}$$

一致估计量为：

$$\hat{k}^R_m = \frac{\hat{\theta} - \hat{\theta}_{-m}}{\hat{\theta}} \tag{10}$$

$k^R_{0,m}$和$\hat{k}^R_m$度量了如果排除第m种环境因素，机会差距的变动比例。

（三）组间机会差距差异检验

机会差距估计量的渐进分布可以被用于进行组间差异检验，例如，评估男性和女性群体的机会公平性差距。令$\hat{\theta}_A$和$\hat{\theta}_B$分别表示两组样本的去偏差机会差距估计量，相应的标准误分别为$\widehat{se}\left(\hat{\theta}_A\right)$和$\widehat{se}\left(\hat{\theta}_B\right)$。可以直接计算两组差异$\hat{\theta}_A - \hat{\theta}_B$的标准误：

$$\widehat{se}\left(\hat{\theta}_A - \hat{\theta}_B\right) = \sqrt{\widehat{se}^2\left(\hat{\theta}_A\right) + \widehat{se}^2\left(\hat{\theta}_B\right)} \tag{11}$$

三　教育机会差距的测算与分析

（一）样本选择与变量度量

1. 样本来源与筛选

本文用于实证分析的样本来自中国综合社会调查（Chinese General Social Survey，CGSS）。CGSS始于2003年，是中国最早的全国性、综合性、连续性学术调查项目。CGSS全面收集社会、社区、家庭、个人多层次数据，系统反映社会变迁趋势，已成为研究中国社会问题最为重要的数据来源，被广泛应用于社会学、政治学和经济学等学科问题的研究。相比其他微观调查数据，CGSS拥有最为丰富的父代信息（特别是子代14岁时的生活环境），包括教育、政治面貌、宗教信仰以及工作信息。鉴于环境因素的充分获取是保障机会差距测算可靠性的先决条件，因此CGSS是研究机会差距、代际流动等问题的首选数据来源（李莹、吕光明，2019）。由于早期调查问卷关于父代信息的问题较少，且与近年调查存在出入，在保证环境因素数量的条件下，选取2010-2021年调查样本。主要样本筛选规则如下：要求子代已完成学校教育、限制样本年龄小于等于80周岁①、剔除数据缺失和异常样本。最终获得52634个有效样本。

2. 变量选择与度量

子代教育是机会差距测算中涉及的结果变量，本文根据受教育年限将教育处理为连续型变量。CGSS将受教育程度划分为13个选项：没有受过任何教育、私塾和扫盲班、小学、初中、职业高中、普通高中、中专、技校、大学专科（成人高等教育）、大学专科（正规高等教育）、大学本科（成人高等教育）、大学本科（正规高等教育）和研究生及以上。根据具体选项，将受教育年限分别设置为0~19年。

环境因素的选择有以下两点考虑。第一，由于各年份涉及父代信息的问题并不统一，因此剔除问题较少年份，其余年份取交集。第二，考虑有效样本数量，比如子代14岁时父母“工作单位或公司所有制性质”在多数年份答案缺失比较严重，因此进行剔除。最终获取五种父代环境因素，包括教育（同样赋值为教育年限）、政治面貌（党员和其他）、14岁时父母就业状况（务农、自雇佣、就业和其他）、14岁时父母职务级别（无行政职务和有行政职务）、14岁时父母单位类型（企业、机关事业单位和其他）。

① 遵循罗楚亮和刘晓霞（2018）的做法，剔除高年龄段数量稀少的样本。在完成学业的基础上，80周岁的年龄大约处于95%分位数附近。

另外，纳入子代个体特征，包括年龄、性别、民族和户口（农业和城镇）[①]。变量含义、类型、度量和符号如表1所示。

表1　变量选择与度量

含义	类型	度量	符号
性别	分类	男为1，女为2	*sex*
年龄	连续	调查年份减去出生年份	*age*
户口	分类	农村为1，城镇为2	*hu*
民族	分类	汉族为1，少数民族为2	*nation*
教育年限	连续	未受教育、私塾和扫盲班为0，小学为6，初中为9，职高、普高、中专、技校为12，大学专科为15，大学本科为16，研究生及以上为19	*edu*
父亲教育年限	连续	未受教育、私塾和扫盲班为0，小学为6，初中为9，职高、普高、中专、技校为12，大学专科为15，大学本科为16，研究生及以上为19	*fedu*
父亲政治面貌	分类	党员为1，其他为2	*fparty*
14岁时父亲就业状况	分类	务农为1，自雇佣（个体工商户、老板、在自己家生意或企业工作）为2，就业（受雇佣于他人）为3，其他为4	*fwork*
14岁时父亲职务级别	分类	无行政职务为1，有行政职务为2	*fclass*
14岁时父亲单位类型	分类	企业为1，机关事业单位为2，其他为3	*ftype*
母亲教育年限	连续	未受教育、私塾和扫盲班为0，小学为6，初中为9，职高、普高、中专、技校为12，大学专科为15，大学本科为16，研究生及以上为19	*medu*
母亲政治面貌	分类	党员为1，其他为2	*mparty*
14岁时母亲就业状况	分类	务农为1，自雇佣（个体工商户、老板、在自己家生意或企业工作）为2，就业（受雇佣于他人）为3，其他为4	*mwork*
14岁时母亲职务级别	分类	无行政职务为1，有行政职务为2	*mclass*
14岁时母亲单位类型	分类	企业为1，机关事业单位为2，其他为3	*mtype*

资料来源：根据2010–2021年中国综合社会调查数据计算得到。

3. 描述性统计

如表1所示，大量环境因素只能通过离散型变量进行度量。已有的机会公平性测度中的预测通常使用简单线性模型实现，模型约束将使得这些详尽的环境因素在预测中

① 参照罗楚亮和刘晓霞（2018），将所有“农转非”个体划入农村样本。

的作用大打折扣。当采用线性模型进行预测时，往往需要对模型进行修正。第一，引入环境变量的交互项和高阶项。例如，环境因素对教育的边际回报可能在男性和女性之间存在差异，导致性别间的机会差距存在较大差异。因此，至少需要纳入环境因素的两两交互项以解决异质性问题。第二，剔除冗余环境信息。通过表2可以看到，离散型环境变量的分布并不均衡，如非汉族、党员、机关事业单位、自雇佣等所占比例一般低于10%。样本非均衡问题不可避免地为模型引入冗余信息，导致系数估计和预测结果不稳定。因此，需要手动或依赖算法自动挑选非冗余环境变量，才能保障最终的预测效果，避免将本应归因于环境因素的教育成就差异错误地归结为努力或者运气。遗憾的是，上述两种缓解模型约束的操作并未受到重视，几乎所有采用线性模型的机会差距测算均选择直接引入环境变量。

表3展示了连续型变量的统计特征。伴随教育扩张等因素（罗楚亮、刘晓霞，2018），可以发现子代受教育年限明显高于父代。虽然母亲受教育年限显著低于父亲，但并不能因此断定父亲对子代教育的影响更大，后文关于机会差距的偏效应估计将对此进行解答。年龄因素在机会均等性测算中通常被重点考虑，其背后可能隐藏着不同出生队列（Golley & Kong，2018）由于宏观政策、教育资源年代差异、环境因素及其边际影响的显著不同，最终导致教育机会差距（林文炼、李长洪，2020）。

表2　离散变量分布统计

变量	含义	样本数量	比例	变量	含义	样本数量	比例
sex	男性	25444	0.483	*ftype*	企业	7245	0.138
	女性	27190	0.517		机关事业单位	5944	0.113
hu	农村	35418	0.673		其他	39445	0.749
	城镇	17216	0.327	*mparty*	党员	1375	0.026
nation	汉族	48147	0.915		非党员	51259	0.974
	非汉族	4487	0.085	*mwork*	务农	41519	0.789
fparty	党员	6500	0.123		自雇佣	1653	0.031
	非党员	46134	0.877		就业	9242	0.176
fwork	务农	37042	0.704		其他	220	0.004
	自雇佣	2058	0.039	*mclass*	无行政职务	43373	0.824
	就业	13195	0.251		其他	9261	0.176
	其他	339	0.006	*mtype*	企业	6137	0.117
fclass	无行政职务	41856	0.795		机关事业单位	3334	0.063
	其他	10778	0.205		其他	43163	0.820

资料来源：根据2010–2021年中国综合社会调查数据计算得到。

表3　连续变量统计特征

变量	样本数量	均值	标准差	最小值	25%	50%	75%	最大值
edu	52634	8.585	4.661	0	6	9	12	19
age	52634	48.688	15.079	17	37	48	60	80
fedu	52634	4.375	4.665	0	0	6	9	19
medu	52634	2.986	4.203	0	0	0	6	19

资料来源：根据2010–2021年中国综合社会调查数据计算得到。

（二）教育机会差距测算结果

基于环境因素对子代教育进行预测时，可以采用线性模型，也可以使用非线性的机器学习算法。由于模型种类繁多，本文在进行机会差距估计时，首先从备选模型中选出最优模型，评价标准为样本外预测精度，具体使用均方误差（MSE）指标进行评估和比较。根据模型形式，备选模型集合纳入线性模型：最小绝对收缩和选择算子（least absolute shrinkage and selection operator，Lasso）和岭回归（ridge regression，RR）；非线性模型：随机森林（random forest，RF）、极端梯度提升算法（extreme gradient boosting，XGB）和类别特征梯度提升算法（CatBoost，CB）。

应用线性模型Lasso和RR进行预测时，应该考虑模型形式误设，需要将环境因素的交互项和高阶项一并纳入预测模型，否则将出现欠拟合。首先，将分类型环境变量转换为虚拟变量。其次，设置连续变量的高阶项，加入年龄变量的二次项和三次项。最后，引入包括连续型变量在内的两两交互项。最终每年预测模型包含的环境因素数量为875个。考虑到不同年度数据分布特征的差异，因此逐年选择最优预测模型，结果如表4倒数第二列所示。可以看到，无论是分年度预测还是全样本分析，最优预测模型（即样本外MSE最小的模型）均为CB，因此教育机会差距的测算基于CB完成。根据前文模型介绍，测算机会差距可以使用完全独立的两步法（预测–估计，记为Plugin），也可以使用对预测偏误保持较低敏感性的去偏差方法（记为Debiased）。表4同时汇报了两种估计方法的估计结果。与已有机会差距测算文献最大的不同之处在于，本文根据机会差距估计量的渐进分布对估计结果进行统计检验，补充汇报了标准误和显著性水平，用于评估样本估计结果的不确定性。图1、图2和图3将表4的数值转换为趋势图，用于更为直观地观察测算结果。

粗略观察表4，可以非常清楚地看到，不同估计方法、不同年份样本，甚至不同机会差距指标（绝对量和相对占比）均在1%水平上具有统计显著性，说明环境因素对子代教育的影响非常显著。对比不同估计方法，从图2和图3可以直观看出，如果

只考虑结果数值，Plugin方法明显高估了教育机会差距。对比表4中两种方法所得的标准误，可以发现Plugin方法的标准误始终大于Debiased方法，特别是在机会差距占比方面，通常为Debiased方法标准误的两倍，说明简单插入子代教育预测的测算结果具有较高的不确定性。Terschuur（2023）发现两者标准误相差1.8~16.1倍，显然此处得到了较为一致的实证结果。然而，当考虑测算结果的波动范围时，图2和图3显示两者的95%置信区间（CI）高度重合，因此仅从数值角度判定Plugin方法高估机会差距的结论并不可靠。尽管如此，由于Debiased方法对预测偏误的敏感性较低，即具有无偏性，同时估计量的方差较小，即具有有效性，因此本文仍以Debiased方法的结果为准。

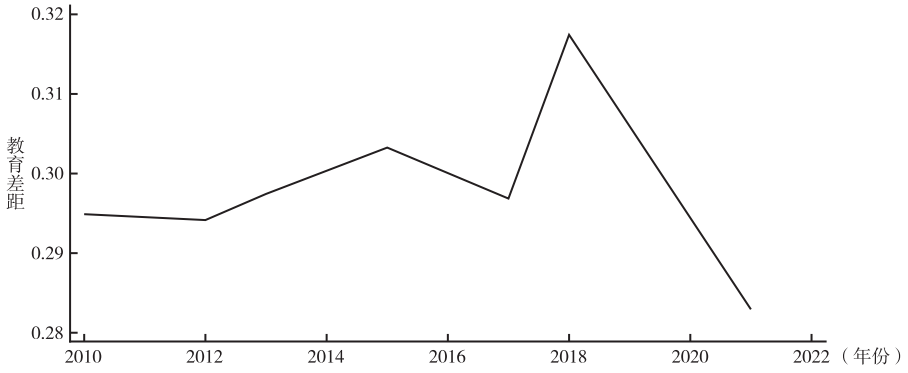

图1　2010–2021年教育差距

资料来源：根据2010–2021年中国综合社会调查数据绘制得到。

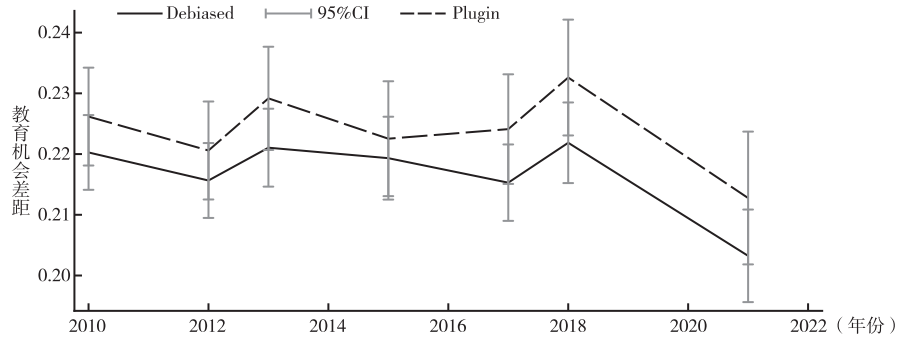

图2　2010–2021年教育机会差距

资料来源：根据2010–2021年中国综合社会调查数据绘制得到。

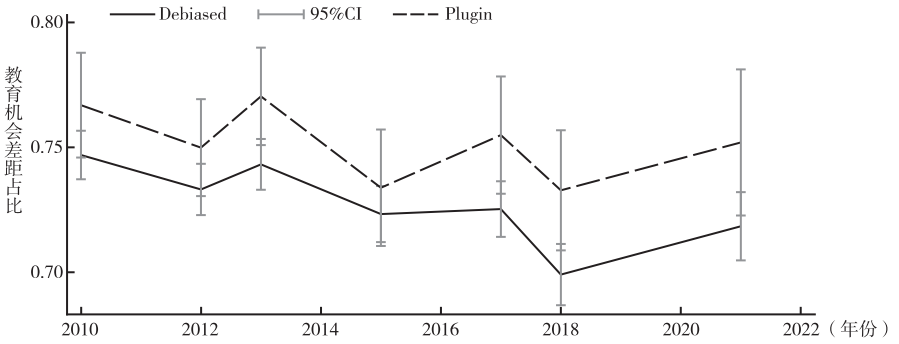

图3　2010–2021年教育机会差距占比

资料来源：根据2010–2021年中国综合社会调查数据绘制得到。

表4　教育机会差距估计结果

年份	机会差距	机会差距占比	总差距	机器学习算法	MSE	估计方法
2010	0.220*** (0.003)	0.747*** (0.005)	0.295	CB	3.185	Debiased
	0.226*** (0.004)	0.767*** (0.011)				Plugin
2012	0.216*** (0.003)	0.733*** (0.005)	0.294	CB	3.217	Debiased
	0.221*** (0.004)	0.750*** (0.010)				Plugin
2013	0.221*** (0.003)	0.743*** (0.005)	0.297	CB	3.176	Debiased
	0.229*** (0.004)	0.770*** (0.010)				Plugin
2015	0.219*** (0.003)	0.723*** (0.006)	0.303	CB	3.265	Debiased
	0.223*** (0.005)	0.734*** (0.012)				Plugin
2017	0.215*** (0.003)	0.725*** (0.006)	0.297	CB	3.359	Debiased
	0.224*** (0.005)	0.755*** (0.012)				Plugin
2018	0.222*** (0.003)	0.699*** (0.006)	0.317	CB	3.540	Debiased
	0.233*** (0.005)	0.733*** (0.012)				Plugin
2021	0.203*** (0.004)	0.718*** (0.007)	0.283	CB	3.322	Debiased
	0.213*** (0.006)	0.752*** (0.015)				Plugin

续表

年份	机会差距	机会差距占比	总差距	机器学习算法	MSE	估计方法
全样本	0.221*** (0.001)	0.739*** (0.002)	0.299	CB	3.262	Debiased
	0.221*** (0.001)	0.741*** (0.003)				Plugin

注：括号内为标准误；*、**和***分别表示10%、5%和1%的显著性水平。

资料来源：根据2010-2021年中国综合社会调查数据计算得到。

接下来考虑教育机会差距大小。基于Debiased的各年份机会差距落在0.203~0.222范围内，机会差距占比约为0.699~0.747，全样本下两者分别为0.221和0.739。显然这一结果显著高于已有研究。表5汇总了赵心慧（2023）基于中国家庭收入调查（CHIP）数据的测算结果，可以总结出两点不同。第一，教育差距的测算结果明显低于本文。考虑相同年份，2013年和2018年的教育基尼系数分别为0.213和0.177，但是表4则给出了0.297和0.317的评估结果。实际上，赵心慧（2023）所列出的四年教育差距都要显著低于本文基于CGSS的测算结果。可能的原因在于教育年限赋值方式，虽然CHIP给出了9种文化程度（未上过学、小学、初中、高中、职高/技校、中专、大专、大学本科、研究生），但是赵心慧（2023）却将前两种文化程度统一赋值为6，而本文对未上过学选项赋值为0（受教育年限为0样本共计7312，占比约为14%）。因此，忽略未上过学样本可能会导致教育差距被低估①。第二，教育机会差距占比明显低于本文。表5中教育机会差距占比的最大值为0.514，但本文普遍在0.700以上，意味着超过70%的教育差距由个体属性、父代特征等环境因素所决定。尽管这一比例（超过70%）显著高于现有文献，但本文认为该结果是稳健且合理的，主要基于以下三方面原因。

首先，由于环境变量选择问题，赵心慧（2023）关于子代教育的预测结果可能存在偏差。赵心慧（2023）将性别、户口、年龄、家庭规模、父亲受教育年限、父亲职业和地区作为环境因素。与本文相比，可以明显看出其环境因素数量偏少，尤其是未将母亲的教育、职业等信息纳入模型。Terschuur（2023）发现，在30个国家中，有17个国家母亲教育是最重要的教育机会差距决定因素，因此遗漏重要环境因素将导致机会差距被低估。此外，由于居住地迁移、工作变动等情况，家庭规模、父亲职业甚至地区都难以成为令人信服的环境因素。CHIP问卷中关于此类问题并不存在时间限制，

① 稳健性评估采用相同的赋值方式，结果发现如此赋值确实导致教育总体差距和机会差距的绝对量减小，但对机会差距的占比影响不大。

更为合理的做法是像CGSS一样，考虑子代成长期（例如14岁）时的家庭规模、父亲职业和所处地区等信息。

其次，过于简单的预测模型可能制约环境因素的预测效果。赵心慧（2023）只是采用基本的线性回归模型，并未对模型误设问题进行考虑。也就是说，即使使用线性模型，也应考虑环境因素间的交互项和高阶项，本质上这也是一种遗漏重要环境因素的错误预测方式。

最后，通过对比收入机会差距和教育机会差距的差异，可以增强本文结果的可信度。本文认为，相较于教育，子代未来所能取得的收入水平明显更容易受到努力、运气等因素的影响。原因在于，教育结果的形成期相对于收入而言较短，同时教育领域受到的市场化冲击较小，更加可能与家庭环境密切相关。因此，相比收入，教育机会差距的占比理应显著更高。借助收入机会差距的相关数据，可以说明本文约70%的测算结果具有合理性。例如，万相昱等（2024）利用条件推断森林（Plugin方法）表明，收入机会差距占比大约为50%。而本文基于Plugin方法的收入机会差距结果为50%~66%（见后文）。因此，通过与收入机会差距进行对比，再次说明教育机会差距的测算结果具备可信度。

表5　已有研究教育机会差距测算结果

内容	指标	2002年	2007年	2013年	2018年
教育差距	基尼系数	0.236	0.211	0.213	0.177
	变异系数平方的一半	0.089	0.074	0.079	0.053
教育机会差距	变异系数平方的一半	0.046	0.037	0.040	0.019
教育机会差距占比	变异系数平方的一半	0.514	0.498	0.501	0.352

资料来源：根据赵心慧（2023）整理得到。

关于机会差距的演变趋势也是普遍关注的焦点问题。根据图1可知教育差距在样本期内呈现先上升后下降的时间趋势，从2010年的0.295变为2021年的0.283，降幅约4%。根据图2，教育机会差距在绝对量上基本与总差距呈现同步的演变趋势，从2010年的0.220降至2021年的0.203，降幅约8%。如果只考虑样本期两端，那么教育差距和机会差距具有下降的时间趋势。将机会差距测算结果的不确定性考虑在内时，2010年机会差距的95%置信区间为［0.214，0.226］，2021年为［0.196，0.211］，两者并不重合。差异性检验的标准误为0.005，获得5%的统计显著性，因此可认为样本期内教育机会差距下降的时间趋势显著存在。对于相对机会差距，教育机会差距占比在样本期内同样存在显著的下降趋势，2010年为0.747，2021年为0.718，前者95%置信区间为［0.737，0.757］，后者为［0.705，0.732］，同样并不重合。差异性检验的标准误为

0.009，获得了5%的统计显著性，因此4%的下降幅度显著存在。

结合上述数值结果，可以看到教育机会差距的缩小是教育差距得以缓解的重要原因。赵心慧（2023）将持续减弱的环境因素影响归结于义务教育和高等教育扩张。实际上，根据下文对于环境因素贡献度的评估和分析，发现教育扩张的确带来了显著的机会差距缩小（罗楚亮、刘晓霞，2018）。基于户口和性别的分组测算结果显示，可能得益于教育扩张的影响①，城镇内部、性别组内的机会差距持续缩小：样本期两端城镇内部机会差距从0.124降至0.108，机会差距占比由0.635降至0.605；男性（女性）机会差距由0.167（0.261）降至0.147（0.239），机会差距占比由0.686（0.762）降至0.639（0.731）。但与之形成鲜明对比的是，农村内部机会差距并无显著变化，相对占比甚至从0.621提高至0.659。这一结果表表明，旨在推动教育公平的普惠性政策，可能并未有效渗透至农村内部并产生预期影响。

根据有效维持不平等假设（effectively maintained inequality，EMI）（Lucas，2001），只有当资源优势群体的教育机会饱和后，教育扩张带来的教育机会才会向资源有限群体扩散。EMI理论能够解释为何城镇教育机会差距持续缩小，而农村机会差距却未得到缓解。至于机会差距占比反而有所增加，可能的原因是农村内部经济能力较强的家庭通过择校、借读等方式部分享受城镇公共教育资源，由此拉大了与物质条件较差的其他农村家庭的机会差距。这部分群体原本就无法“被动”享受教育扩张带来的福利，而“主动”付出经济成本以改善教育条件的可行性也更低，从而长期滞留在低端教育层次。本文还从模型设定、结果变量度量、差距指标和环境因素选择四个方面评估了教育机会差距测算结果的稳健性，数据显示实证发现十分稳健②。

（三）环境因素的偏效应估计

除了机会差距大小，环境因素贡献度也是研究者普遍关注的焦点。关于环境因素对机会差距影响程度的测算遵循同样的思路：包含和未包含特定环境变量下机会差距的差异。表6汇报了全样本下环境因素偏效应的估计结果③。在正式阐述本文实证发现之前，首先回顾一下已有研究关于环境因素贡献度的评估结果。赵心慧（2023）、Golley & Kong（2018）发现户口、年龄和父亲教育是影响子代教育最重要的三类环境因素。关于本文的测算结果，可以看到影响教育机会差距最为重要三个因素分别为年

① 根据匿名评审专家的意见，为便于读者快速把握核心内容，本文简要汇报分样本结果，详细结果留存备索。

② 由于篇幅有限，相关结果留存备索。

③ 只展示偏效应至少获得10%显著性的环境因素。

龄、户口和性别。全样本下，忽视年龄将导致机会差距减少0.015，降幅为7%；忽视户口将导致机会差距减少0.012，降幅为5.4%；忽视性别将导致机会差距减少0.009，降幅为4%。年份虚拟变量显著的偏效应意味着年度间教育机会差距存在显著差异，在全样本中贡献了约1.3%的机会差距。父亲和母亲受教育年限被忽视后将引致机会差距下降约1%和0.6%。子代民族身份和父亲就业状况虽然也取得了至少5%水平的显著性，但是影响力的绝对量比较小①。

表6　环境因素偏效应估计结果

环境变量	绝对偏效应	相对偏效应
age	0.015*** (0.000)	0.070*** (0.002)
hu	0.012*** (0.000)	0.054*** (0.002)
sex	0.009*** (0.000)	0.040*** (0.001)
year	0.003*** (0.000)	0.013*** (0.001)
fedu	0.002*** (0.000)	0.010*** (0.001)
medu	0.001*** (0.000)	0.006*** (0.001)
nation	0.001*** (0.000)	0.002*** (0.000)
fwork	0.000** (0.000)	0.001** (0.000)

注：括号内为标准误；*、**和***分别表示10%、5%和1%的显著性水平。

资料来源：根据2010–2021年中国综合社会调查数据计算得到。

表6同样揭示出年龄和户口的重要影响，但是在性别因素上与现有文献存在差异。已有研究在Shapley分解时使用线性回归模型（未加入交互项和高阶项），本文基于同样方法进行测算，发现前三位重要因素与赵心慧（2023）和Golley & Kong（2018）一致②。因此可知，性别因素的差异很有可能来自模型不同，由此反映出性别对教育的影响显著体现在与其他环境因素的交互作用上，比如不同地区（城镇和农村）、不同出生队列（所处

① 分年度估计结果具有相似规律。根据匿名评审专家的意见，本文简要汇报分样本测算结果，相关数据留存备索。

② 在匿名评审专家的启发下，本文采用赵心慧（2023）和Golley & Kong（2018）相同的分解方法对环境因素的贡献度进行测算（相关结果留存备索）。户口、年龄和父亲教育的影响力度位居前三，与已有研究一致，说明性别因素贡献度的差异来源于模型的不同。

年代）的性别影响具有显著差异，在构建预测模型时应该充分纳入性别与其他因素的交乘项，显然已有文献忽视了这一点。上述发现说明在测度环境因素贡献度时，去偏差机器学习方法能够提供增量发现。接下来本文将逐一分析年龄、户口和性别的重要影响[①]。

1. 教育扩张与教育机会差距

中国的教育扩张始于20世纪80年代。1986年《中华人民共和国义务教育法》正式颁布实施，九年义务教育的推行显著增加了适龄儿童和青少年接受学校教育的可能性。1990年，为适应经济社会发展需要，中国高等教育提高录取比例并扩大招生规模，成为中国教育扩张道路上最为重要的标志性事件。相比九年义务教育的实施，高等教育扩张带来的教育公平问题更受关注。李春玲（2010）发现，大学扩招并没有提高教育机会公平性，反而导致城乡之间的教育差距上升。张建华和万千（2018）提供的证据表明，高校扩招在高等教育资源丰富的省份弱化了教育的代际传递，在高等教育资源匮乏的省份则增加了高学历家庭子女进入大学的可能性。陆雪琴等（2023）使用CGSS的研究表明，高校扩招总体上缩小了群体间的相对教育差距。

综合教育扩张历史及其相关研究成果，有理由怀疑年龄之所以成为影响教育机会差距最大的环境因素，可能是不同出生队列由于受到或者未受到教育扩张政策的影响从而产生不同的教育后果。如果教育扩张真正实现了教育公平，那么受到政策影响的群体内部教育机会差距将会显著小于未受到政策影响的群体。为了验证上述假说，遵循巫锡炜等（2022）根据出生日期将样本划分为两个出生队列，出生日期在1981年之前（记为*old*）和出生日期在1981年及其之后（记为*young*），后者被认为受到教育扩张的影响。表7汇报了两类群体的教育机会差距测算结果，并对组间差异进行了统计检验。

通过群体间横向对比可以揭示年龄因素影响力的来源。首先，根据教育年限（表7最后一列）可以非常直观地看到教育扩张的政策效果，*young*群体的受教育年限明显高于*old*群体。其次，教育扩张显著降低了教育机会差距。从机会差距的绝对量上看，*young*群体基本维持在*old*群体一半水平。全样本下*old*群体的机会差距为0.222，*young*群体为0.112，3位有效数字下相同的标准误意味着两者差异具有统计上的显著性：0.110的差异在1%水平下显著异于零。最后，考虑相对机会差距，机会差距占比的测算结果同样表明*young*群体内部环境因素对于子代教育的影响力不如*old*群体。全样本下*old*群体的机会差距占比为0.695，*young*群体则为0.655，0.040的差异同样在1%水平

① 根据匿名评审专家的意见，本文简要汇报了分组测算结果。在年度维度上进行了分组评估，相关发现与全样本结果一致，具体结果留存备索。

下显著异于零。综合实证结果的分析，有理由认为年龄偏效应的来源可能是教育扩张政策带来的教育机会差距变动。

表7　组间教育机会差距测算结果（全样本）

分组依据	分组	机会差距	差异	机会差距占比	差异	教育年限
出生队列	*old*	0.222*** (0.002)	0.110*** (0.002)	0.695*** (0.003)	0.040*** (0.007)	7.807
	young	0.112*** (0.002)		0.655*** (0.006)		11.740
户籍	农村	0.208*** (0.002)	0.081*** (0.002)	0.653*** (0.003)	−0.005 (0.006)	7.162
	城镇	0.127*** (0.002)		0.658*** (0.005)		11.514
性别	男性	0.167*** (0.001)	−0.097*** (0.002)	0.680*** (0.004)	−0.075*** (0.005)	9.339
	女性	0.264*** (0.002)		0.756*** (0.003)		7.879

注：括号内为标准误；*、**和***分别表示10%、5%和1%的显著性水平。

资料来源：根据2010–2021年中国综合社会调查数据计算得到。

2. 户籍与教育机会差距

虽然已知户籍是决定教育机会公平的重要环境因素，但进一步分析农村和城镇群体内部教育机会均等性的差异仍具有重要意义。根据表7的结果进行横向对比，发现城镇群体拥有较长的受教育年限，同时存在较低的机会差距。尽管农村和城镇的教育机会差距绝对值存在显著差异，但由于城镇的教育差距总量较小，导致农村和城镇间相对机会差距占比并不存在显著差异，统计检验也未获得显著性。上述结果产生的原因可能如下：相较于城镇，农村地区教育资源匮乏，群体受教育程度的上限虽然较低，但下限更低，导致其教育总差距更高。与此同时，在资源受限的条件下，具有环境优势的家庭能够更容易获得相较于其他家庭更为优质的教育资源，因此农村地区的机会差距绝对值更高。然而，农村家庭由于缺乏抗风险（如疾病、意外事故等）能力，运气等非环境因素引致的教育结果离散度也更大，在一定程度上稀释了环境因素的作用，从而导致其与城镇在机会差距占比上并未表现出显著差异。

3. 性别与教育机会差距

观察表7中关于性别的机会差距测算分组结果，可以看到男性群体拥有较长的教育年限，同时具有较低的机会差距绝对量和相对值，所有统计检验均获得了1%水平下的显著性。以上结果说明环境因素对不同性别子代的教育影响力存在差异，大致可区分

为两个原因：环境禀赋差异和环境边际贡献差异。根据数据进行简单统计，男性和女性的户口平均值（农村为1，城镇为2）分别为1.338和1.327，年龄平均值分别为49.069和48.331，父亲平均教育年限分别为4.389和4.362，母亲平均教育年限为2.997和2.976。上述数字说明不同性别群体的环境禀赋差异并不明显。排除禀赋差异后，男性和女性在教育机会公平方面的差距只能归因于环境因素的边际贡献存在差异。这再次说明，在设定预测模型时，应尽可能引入更多环境因素的交互项。

（四）教育机会差距与收入机会差距

收入可能是机会均等性研究中最受关注的结果变量，而教育通常是解释环境因素影响收入的重要渠道。因此，对比收入和教育机会差距具有重要意义，允许观察环境因素对于教育和收入差距的解释力度差距。表8汇报了收入机会差距的测算结果，环境变量及其度量与前文相同。遵循李莹和吕光明（2019）、万相昱等（2024），将样本年龄限制为18~60周岁并且收入大于零。为了剔除通胀因素的干扰，以2009年为基期，使用消费者价格指数对收入数据进行调整①。图4至图6绘制了收入机会差距、收入机会差距占比以及教育机会差距占比的时间走势②。

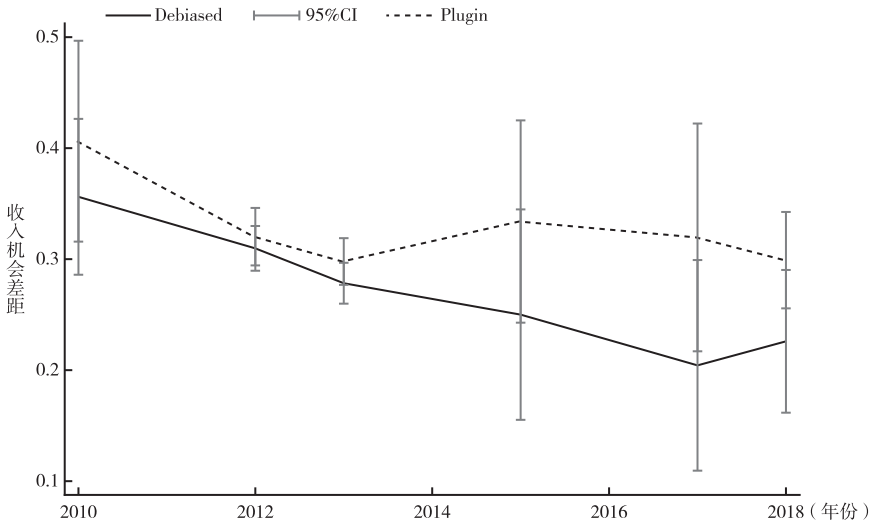

图4　2010–2018年收入机会差距

资料来源：根据2010–2018年中国综合社会调查数据绘制得到。

① 资料来源：国家统计局（https://data.stats.gov.cn/easyquery.htm?cn=C01）。

② 由于2021年有效样本仅有2474个，无法排除因样本量过少而导致的估计问题（万相昱等，2024），因此舍弃该年度。鉴于篇幅有限，图中数值结果留存备索。

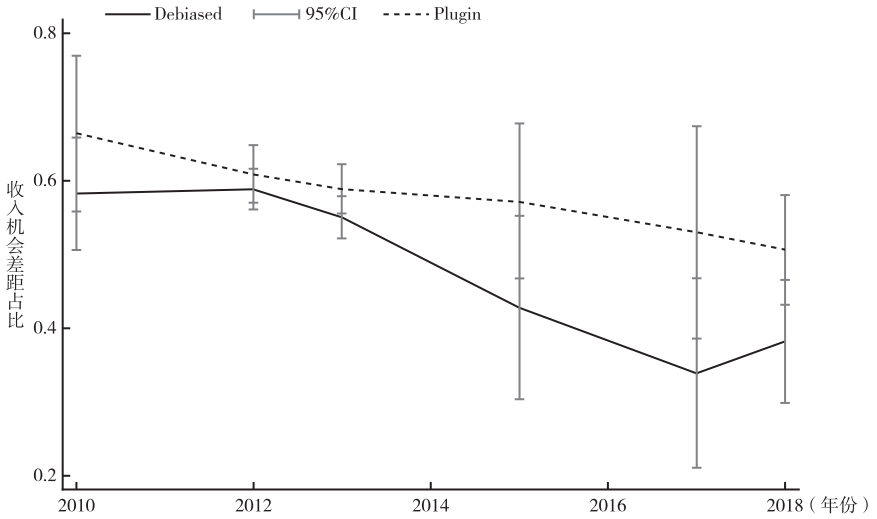

图5　2010–2018年收入机会差距占比

资料来源：根据2010–2018年中国综合社会调查数据绘制得到。

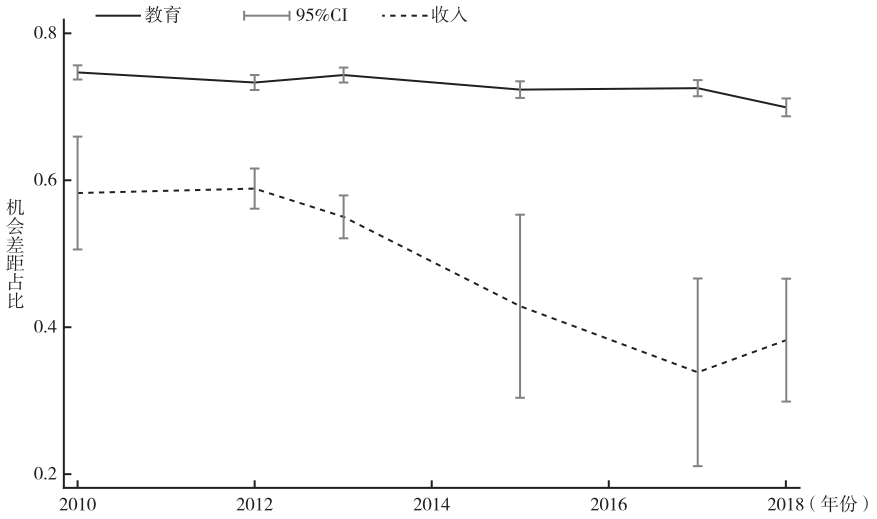

图6　2010–2018年收入和教育机会差距占比

资料来源：根据2010–2018年中国综合社会调查数据绘制得到。

根据测算数据，环境因素对个人收入的影响与教育因素一样非常显著，所有结果均在1%水平下显著大于零。收入机会差距占比的Plugin估计结果落在0.506~0.664范围内，与采用相同样本、类似Plugin方法的万相昱等（2024）测算结果0.381~0.524较为

接近。根据图6，教育机会差距占比显著高于收入机会差距，这意味着环境因素对教育的影响力度显著高于对收入的影响。如前文所述，相较于收入，教育领域受到的市场化冲击较弱，后天努力和运气发挥作用的空间较小，因此个人属性和父代特征显得更为重要。同时，样本期内收入机会差距呈现出比教育机会差距更为明显的下降趋势。具体来看，机会差距的绝对量在2010年和2018年分别为0.356和0.226，下降幅度达到37%。但差异性检验的标准误为0.049，未达到统计显著性。机会差距占比分别为0.582和0.382，下降幅度达到34%。差异性检验的标准误为0.058，在5%水平上显著。与教育机会差距占比4%的下降幅度相比，环境因素对收入的解释力度正在加速减弱，其中的作用机制显然不能完全通过教育进行解释。

四　教育机会差距的收入分配效应评估

教育被广泛认为是环境因素影响子代收入最为重要的渠道，机会公平性研究中也专注于评估教育在其中产生的影响大小（史新杰等，2022）。本文在对教育机会差距进行一系列测算后，同样希望明晰环境因素如何通过教育作用于收入差距。本文借助收入分配效应评估中常用的Shapley分解（Shorrocks，2013）及其改进模型进行实证分析。根据万相昱（2024）和万相昱等（2024），在劳动经济学领域被广泛使用的Shapley分解存在三个缺陷：第一，线性模型约束降低预测精准性；第二，利用取均值方式剔除因素影响不符合Shapley值的基本思想；第三，预测变量选择主观且数量较少，制约预测效果。因此，在评估教育机会差距的收入分配效应时，本文遵循已有研究，采用沙普利加性解释模型（SHAP）。

此处的评估思路如下。首先，基于环境因素预测个体受教育年限代理环境因素所决定的教育水平。同时使用实际教育年限减去预测值获得努力和运气决定的教育水平。其次，将两种教育水平代入收入决定方程，使用非线性的机器学习算法测算两种教育水平对每一样本收入的贡献度。最后，使用实际收入分别减去两种教育水平的贡献度获得反事实收入。对比实际收入和反事实收入下的差距指标可评估教育机会公平的收入分配效应，评估结果如表8所示[①]。

① 评估前进行了最优模型选择，备选模型为Lasso、XGB、CB和轻量级梯度提升学习（light gradient boosting machine，LightGBM），LightGBM的预测效果最佳。因此，此处主要基于LightGBM进行评估。

表8　教育机会差距的收入分配效应评估结果

年份	原始收入	剔除环境因素		剔除其他因素		剔除教育因素		机器学习算法
		反事实Gini	变动	反事实Gini	变动	反事实Gini	变动	
2010	0.612	0.508	−0.170	0.589	−0.038	0.487	−0.204	LightGBM
2012	0.541	0.463	−0.145	0.527	−0.027	0.444	−0.180	
2013	0.528	0.457	−0.133	0.515	−0.024	0.444	−0.159	
2015	0.617	0.602	−0.024	0.597	−0.031	0.561	−0.090	
2017	0.622	0.594	−0.046	0.596	−0.041	0.560	−0.099	
2018	0.618	0.563	−0.089	0.607	−0.017	0.545	−0.118	
2021	0.651	0.599	−0.081	0.629	−0.035	0.573	−0.121	
全样本	0.614	0.559	−0.090	0.595	−0.031	0.533	−0.132	

资料来源：根据2010–2021年中国综合社会调查数据计算得到。

粗略观察表中数据，所有变动均为负值，说明教育作为一个整体因素拉大了收入差距。无论是源于环境因素的教育水平，还是源于努力与运气的教育水平，都扩大了收入差距。当剔除教育水平时，收入差距下降了13.2%，说明教育在样本中起到了拉大收入差距的作用。将教育进一步拆分后，全样本下环境决定的教育水平引致了9%的收入差距。与之形成鲜明对比的是努力和运气因素，其所决定的教育水平使基尼系数（Gini）增加了约3.1%。对比可知，教育所产生的收入分配效应中，个体属性和父代特征占据主导地位。从时间趋势来看，整体教育水平和环境决定的教育水平对收入差距的拉动作用正在逐年减弱，意味着环境因素通过教育影响收入差距的渠道正在发生变化。

五　研究结论与政策建议

鉴于当前关于中国教育机会差距的测算并无权威结果，本文使用去偏差机器学习重新对环境因素所决定的教育差距进行评估。基于CGSS数据尽可能纳入充足的个体特征、父代信息等环境变量，在完成对教育机会差距科学测算的基础上，评估每一环境因素的偏效应，对比分析环境因素在教育和收入分配中的作用差异，最后使用可解释机器学习模型对教育在环境和收入中间扮演的重要作用进行测算。归纳相关结论如下。

第一，环境因素决定了70%~75%的教育差距，显著高于收入机会差距。使用去偏差机器学习方法对中国2010–2021年教育机会差距进行测算后，发现接近74%（全样本）的教育差距可以由个体特征和父代信息所决定，明显高于已有实证结果。在样本期内，教育机会差距略有降低，但远不及收入机会差距的下降幅度，这意味着教育显然并不能成为驱动当前收入分配格局演变的主导力量。

第二，针对环境因素的偏效应估计结果表明，年龄、户口和性别是导致教育机会差距的重要因素，其影响显著高于父代教育水平。进一步分析发现，年龄影响的背后蕴含着教育扩张对不同出生队列人群产生的显著作用，教育扩张显著缩小了教育机会差距。尽管农村与城镇样本在受教育年限上存在明显差距，但机会差距的占比并无显著差异。男性与女性在可观测环境禀赋上的差异，不足以解释两者在教育机会上的显著差距。因此，这种差距更可能源于不同环境因素对男女教育回报的边际贡献存在差异。

第三，教育机会差距对收入分配的影响显著，造成了约10%的收入差距。基于去偏差机器学习预测的教育水平，将实际受教育年限分解为由环境因素决定的教育水平和由个人努力与运气决定的教育水平两部分，采用SHAP方法评估三种教育水平对收入差距的贡献。结果表明，整体教育水平拉大了约13%的收入差距，环境因素决定的教育水平导致了约10%的收入差距，而努力和运气通过教育共同产生了3%的收入分配效应。上述结果充分说明，在教育带来的收入分配效应中，环境因素占据主导作用。

虽然在机会公平性研究中使用机器学习方法屡见不鲜，但去偏差机器学习方法尚未受到广泛重视。该方法能够解决参数方法以及简单应用机器学习方法时存在的一些问题，但在应用过程中仍需注意以下三个方面。首先，预测模型的选择标准尚不明确。去偏差机器学习允许使用主流机器学习方法，但尚不清楚预测模型对目标参数估计的影响程度。其次，去偏差机器学习无法解决机器学习算法中的超参数调整问题。最后，在应用于机会差距测算时，去偏差机器学习无法解决因果识别问题。此外，与参数方法相比，去偏差机器学习还存在模型复杂度高、运行速度慢以及内存空间占用大等不足。

基于研究发现，本文提出如下政策建议。第一，加大对城市地区教育资源的投入力度，优化资源配置，提升地区承载能力。同时，持续放开户籍限制，完善随迁子女入学政策，降低农村学生进城入学门槛，保障其平等接受教育的权利。第二，建议延长义务教育年限，为学生提供更完整的教育基础。同时，完善普职分流机制，尊重学生个体差异与兴趣，将职业教育提升至与普通教育同等重要的地位，加大职业教育投入，优化课程体系，提升教学质量，拓宽职业教育学生的升学与就业渠道。第三，重视女性教育，通过宣传引导、政策扶持等措施，提高社会对女性教育的重视程度，消除性别偏见，为女性创造更公平的教育环境，助力其增加受教育年限，缩小与男性的教育差距，从而实现教育公平。

参考文献：

李春玲（2010），《高等教育扩张与教育机会不平等——高校扩招的平等化效应考查》，《社会学研究》第3期，第82–113页。

李莹、吕光明（2019），《中国机会不平等的生成源泉与作用渠道研究》，《中国工业经济》第9期，第60–78页。

林文炼、李长洪（2020），《“入学年龄规定”会产生教育不平等吗？——来自1986年〈义务教育法〉的证据》，《经济学（季刊）》第3期，第959–976页。

陆雪琴、马汴京、陈慧文（2023），《高等教育扩张与阶层间教育机会不平等》，《中国经济问题》第1期，第180–196页。

罗楚亮、刘晓霞（2018），《教育扩张与教育的代际流动性》，《中国社会科学》第2期，第121–140页。

史新杰、李实、陈天之、方师乐（2022），《机会公平视角的共同富裕——来自低收入群体的实证研究》，《经济研究》第9期，第99–115页。

万相昱（2024），《共享机制能否推动企业高质量发展？——来自机器学习的实证发现》，《劳动经济研究》第6期，第13–46页。

万相昱、张晨、唐亮（2024），《中国居民收入机会不平等再测算——来自机器学习的新发现》，《数量经济技术经济研究》第1期，第192–212页。

巫锡炜、曹增栋、武翰涛（2022），《高等教育扩张与小家庭崛起——来自大学扩招政策的证据》，《社会学研究》第3期，第92–114页。

张建华、万千（2018），《高校扩招与教育代际传递》，《世界经济》第4期，第168–192页。

赵心慧（2023），《教育机会不平等的变化趋势及成因：2002–2018年》，《财经研究》第2期，第79–94页。

周康、张俊森、李琼琼（2025），《就业冲击与教育的代际传递：来自国企改革的证据》，《劳动经济研究》第3期，第43–70页。

Almås, Ingvild, Alexander Cappelen, Jo Thori Lind, Erik Sørensen & Bertil Tungodden (2011). Measuring Unfair (In)equality. *Journal of Public Economics*, 95 (7–8), 488–499.

Bourguignon, François, Francisco Ferreira & Marta Menéndez (2007). Inequality of Opportunity in Brazil. *Review of Income and Wealth*, 53 (4), 585–618.

Brunori, Paolo, Paul Hufe & Daniel Mahler (2023). The Roots of Inequality: Estimating

Inequality of Opportunity from Regression Trees and Forests. *The Scandinavian Journal of Economics*, 125 (4), 900–932.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey & James Robins (2018). Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, 21 (1), C1–C68.

Escanciano, Juan Carlos & Joël Robert Terschuur (2022). Machine Learning Inference on Inequality of Opportunity. *arXiv Working Paper*, No. 2206.05235.

Ferreira, Francisco & Jérémie Gignoux (2011). The Measurement of Inequality of Opportunity: Theory and an Application to Latin America. *Review of Income and Wealth*, 57 (4), 622–657.

Fleurbaey, Marc (1995). Three Solutions for the Compensation Problem. *Journal of Economic Theory*, 65 (2), 505–521.

Fleurbaey, Marc & Vito Peragine (2013). Ex Ante Versus Ex Post Equality of Opportunity. *Economica*, 80 (317), 118–130.

Golley, Jane & Sherry Tao Kong (2018). Inequality of Opportunity in China's Educational Outcomes. *China Economic Review*, 51, 116–128.

Hufe, Paul, Andreas Peichl & Daniel Weishaar (2022). Lower and Upper Bound Estimates of Inequality of Opportunity for Emerging Economies. *Social Choice and Welfare*, 58 (3), 395–427.

Juárez, Florian Wendelspiess Chávez & Isidro Soloaga (2014). iop: Estimating Ex-Ante Inequality of Opportunity. *The Stata Journal*, 14 (4), 830–846.

Kanbur, Ravi & Andy Snell (2019). Inequality Indices as Tests of Fairness. *The Economic Journal*, 129 (621), 2216–2239.

Lefranc, Arnaud, Nicolas Pistolesi & Alain Trannoy (2008). Inequality of Opportunities vs. Inequality of Outcomes: Are Western Societies All Alike? *Review of Income and Wealth*, 54 (4), 513–546.

Lleras-Muney, Adriana (2005). The Relationship Between Education and Adult Mortality in the United States. *Review of Economic Studies*, 72 (1), 189–221.

Lucas, Samuel (2001). Effectively Maintained Inequality: Education Transitions, Track Mobility, and Social Background Effects. *American Journal of Sociology*, 106 (6), 1642–1690.

Roemer, John (1998). *Equality of Opportunity*. Cambridge: Harvard University Press.

Shorrocks, Anthony (2013). Decomposition Procedures for Distributional Analysis: A Unified Framework Based on the Shapley Value. *Journal of Economic Inequality*, 11 (1), 99–126.

Terschuur, Joël (2023). Educational Inequality of Opportunity and Mobility in Europe. *arXiv Working Paper*, No. 2212.02407.

Van de Gaer, Dirk (1993). *Equality of Opportunity and Investment in Human Capital*. Belgium: Catholic University of Louvain.

Environmental Factors, Educational Equity and Income Gap: New Discoveries from Debiased Machine Learning

Zhang Chen[1], Li Molin[1], Zhang Qi[2] & Wu Yu[3]

(School of Public Finance and Taxation, Shandong University of Finance and Economics[1];
Faculty of Applied Economics, University of Chinese Academy of Social Sciences[2];
Xiamen International Bank[3])

Abstract: Educational equity is the foundation of social justice and a prerequisite for achieving common prosperity. Adopting an equity of opportunity perspective, this paper examines the disparities in educational opportunities in China and assesses their impact on income distribution. Using data from the 2010–2021 China General Social Survey (CGSS), the study employs a debiased machine learning method to measure the educational opportunity gap. The findings indicate that environmental factors – encompassing individual characteristics and parental background – account for over 70% of educational inequality. Educational equity showed significant improvement during the sample period. Age, household registration (hukou), and gender emerge as the three most critical factors determining an individual's educational attainment, with the age effect likely reflecting the impact of national educational expansion. Significant educational opportunity gaps persist within both urban and rural areas, and women face a markedly larger opportunity gap than men. Furthermore, results from an interpretable machine learning model show that education contributes to approximately 13% of income inequality. Environmental factors are the primary drivers, channeling through education to explain about 10 percentage points of this income inequality. Based on these findings, the paper puts forward policy recommendations to promote equity of educational opportunity.

Keywords: environmental factors, equity of educational opportunity, income distribution, machine learning

JEL Classification: D63, E24, I24

（责任编辑：西　贝）