2024 届本科生学士学位论文　　学校代码：10269

華東師範大學

East China Normal University

本科生毕业论文

基于世界范围内男女性劳动力市场表现数据的性别不平等问题研究

The study on gender inequality based on data on the labour market performance of women and men worldwide

姓　　名：王若溪

学　　号：10205000460

学　　院：统计学院

专　　业：统计学

指导教师：李丹萍

职　　称：教授

2024 年 4 月

华东师范大学学位论文诚信承诺

本毕业论文是本人在导师指导下独立完成的，内容真实、可靠。本人在撰写毕业论文过程中不存在请人代写、抄袭或者剽窃他人作品、伪造或者篡改数据以及其他学位论文作假行为。

本人清楚知道学位论文作假行为将会导致行为人受到不授予/撤销学位、开除学籍等处理（处分）决定。本人如果被查证在撰写本毕业论文过程中存在学位论文作假行为，愿意接受学校依法作出的处理（处分）决定。

承诺人签名：王若溪　　日期： 2024 年 4 月 21 日

华东师范大学学位论文使用授权说明

本论文的研究成果归华东师范大学所有，本论文的研究内容不得以其它单位的名义发表。本学位论文作者和指导教师完全了解华东师范大学有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅；本人授权华东师范大学可以将论文的全部或部分内容编入有关数据库进行检索、交流，可以采用影印、缩印或其他复制手段保存论文和汇编本学位论文。

保密的毕业论文（设计）在解密后应遵守此规定。

作者签名：王若溪　导师签名：[illegible]　日期： 2024 年 4 月 21 日

目录

基于世界范围内男女性劳动力市场表现数据的性别不平等问题研究

摘要：

本文旨在探讨世界范围内男女性在劳动力市场上面临的性别不平等问题，并研究导致这种不平等的潜在因素。研究分为三个主要部分：首先，通过性别不平等指数、性别发展指数和历史性别平等指数等指标，描绘了全球范围内性别不平等的直观表现，通过性别薪酬差距、劳动参与率、低薪岗位人数分布和顶端收入人数分布等数据特征刻画劳动力市场的性别不平等问题。其次，重点关注教育、婚育、政治发言权和社会潜在共识四大因素，分析这些因素与性别发展指数、性别不平等指数和性别薪酬差距之间的关系。最后构建以性别不平等指数为响应变量的回归方程，采用多元线性回归模型进行回归分析，包括残差分析、多重共线性检验、异方差检验和回归系数显著性检验等，解决了内生性问题，引入了交互作用构建合理的限制模型，以解释性别不平等。

研究结果表明，生育率、政治权利、劳动参与率、预期寿命和受教育程度这些因素对性别不平等具有显著影响，这些因素自身作用和交互作用构建的回归模型能够很好的解释性别不平等指数，对深入理解劳动力市场性别不平等问题做出了助推。

关键词：性别不平等，劳动力市场，性别薪酬差距，多元线性回归模型，交互作用

The study on gender inequality based on data on the labour market performance of women and men worldwide

Abstract:

The purpose of this paper is to explore the gender inequality faced by men and women in the labour market worldwide and to examine the underlying factors that lead to such inequality. The study is divided into four main parts: Firstly, it depicts the visual manifestation of gender inequality globally through indicators such as the Gender Inequality Index (GII), the Gender Development Index (GDI) and the Historical Gender Equality Index (HGEI), and portrays the gender inequality problem in the labour market through the characteristics of the data, such as the gender wage gap, the labour force participation rate, the distribution of the number of people in low-paying jobs, and the distribution of the number of people in the top-end of the income scale. Secondly, it focuses on four major factors, namely education, marriage and parenthood, political voice and potential social consensus, and analyses the relationship between these factors and the GDI, GII and Gender Wage Gap. Finally, the regression equation with GII as the response variable is constructed, and the multiple linear regression model is used for regression analysis, including residual analysis, multiple covariance test, heteroskedasticity test, and regression coefficient significance test, etc., we also solve the problem of endogeneity, and the interaction is introduced to construct a reasonable restriction model in order to explain gender inequality.

The results of the study show that the factors of fertility, political rights, labour force participation, life expectancy and educational attainment have a significant effect on gender inequality, and the regression model constructed by the effects of these factors themselves and their interactions can explain the gender inequality index very well, which contributes to the in-depth understanding of the issue of gender inequality in the labour market.

Keywords: Gender Inequality, Labour Market, Gender Pay Gap, Multiple Linear Regression Model, Interactions

1、绪论

1.1 背景

在传统观念中，“家庭主妇”一词被用来统称那些在家务工作上花费大部分时间的女性。类似地，“全职妈妈”这个词则是指那些完全投身于家庭而放弃工作的母亲。这两个词在我们的社会中早已司空见惯，以至于我们甚至没有注意到为什么很少有人提到“家庭主夫”或者“全职父亲”。在过去的几十年里，性别角色在社会中的分工一直是一个重要的议题。传统观念认为，男性应该是家庭的经济支柱，承担着养家糊口的责任，而女性则被期望在家中负责照料儿女、料理家务。长期以来，我们似乎默认男性在职场上扮演更重要的角色，而女性则应该回归家庭。即使这种观念并没有被明确提出，但或多或少已经潜移默化地渗透到每个人的成长经历中，形成了潜在的底层意识。这种性别角色的划分不仅影响了个人的职业选择和发展，也对整个社会结构产生了深远的影响[1]。

然而，随着时间的推移和社会变革的发展，这种传统观念逐渐受到了质疑和挑战。越来越多的女性追求教育和事业发展，展现出了与男性同样的能力和雄心。同时，也有一些男性选择在家庭中扮演更积极的角色，愿意放弃一部分职业发展的机会，全身心地投入到家庭事务中[2]。这种对传统性别角色的重新审视和重新定义，为我们通往两性平权的路上起到了助推力量。但从更广阔的地域维度和更长远的时间跨度来看，尽管人们在理论上倾向与赞同男女平等的观念，但在真实生活里，性别不平等仍广泛存在于劳动力市场中。许多研究表明，女性在职业发展、晋升机会等方面仍然面临着种种障碍和歧视，男性的普遍收入往往高于女性，女性在高薪职位上的比例相对较低，而在低薪工作中的比例则相对较高。多年以来，男女在劳动力市场上表现出的各种数据差距反映了一系列性别不平等问题，"职场性别歧视"也成为了许多平权倡导者呼吁解决的焦点议题。

鉴此，本文旨在深入分析全球范围内男女在劳动力市场上所面临的性别不平等问题。对性别不平等指数、性别发展指数、性别薪酬差距、两性劳动参与率等指标进行综合分析。同时，将从教育、婚姻与育儿、政治发言权和社会共识四个方面分析造成这些不平等或差距的原因，并利用影响因子和响应指标寻找最优回归模型。

通过这项研究，我们希望能够实现对性别角色更全面的认知和理解，打破传统观念的束缚，为男女平等的实现所必须的更公正、包容的劳动力市场的构建尽一份力，为实现性别平等的目标提供有益的建议和方向。

只有当我们消除性别不平等，给予每个人平等的机会和尊重，我们才能真正实现社会的进步和繁荣。

1.2 文献回顾与创新点

1.2.1 现有文献对劳动力市场性别不平等表现的研究

性别平等和女性权益对于社会和谐发展至关重要。虽然中国政府自建国以来采取了一系列措施来推动女性权益的历史性变革，但性别不平等问题仍然普遍存在。在社会经济领域，女性在劳动参与率和职业地位和薪酬方面一直落后于男性,劳动力市场上的性别差距严重阻碍了女性的发展，这不符合现代文明发展的基本要求，也不利于中国人口和经济的长期繁荣[3]。

从劳动参与率来说，在经济体制改革后，我国两性的劳动参与率均出现下降趋势，但女性的劳动参与率始终低于男性，而且下降的幅度更大、更明显[4]。姚先国和谭岚（2005）发现，我国城镇女性劳动参与率在 1988-2002 年从 91.37%下降至 83.33%[5]。沈可等（2012）发现从 1990 年至 2010 年间，女性劳动参与率下降的幅度显著超过男性[6]。根据北京大学进行的中国家庭追踪调查（CFPS 2010）显示，2010 年成年女性的劳动参与率比成年男性低了 8.74％[4]，正处于黄金劳动年龄的女性劳动参与率持续下降，由 1990 年的 91%、降至 2000 年的 87.6%、再降至 2010 年的 83.2%；而男性劳动参与率基本保持恒定，20 年间劳动力参与率仅仅下降 2 %[6]。

从女性职业地位来说，在社会结构和体制变革过程中，获得职业地位方面的性别差异明显，女性的职业地位获得明显不如男性，且存在扩大趋势。主要表现为劳动力市场上存在的职业性别隔离，横向隔离将男性和女性区分为“男性职业”和“女性职业”，前者职业地位高、工资收入高、社会福利好，后者职业地位低、工资收入低、社会福利差[7]。在纵向隔离研究中，“玻璃天花板效应”揭示了女性职业向上流动的隐性障碍[8]。同时，随着社会流动性和社会开放程度的提高，男女性在获得职业地位方面的差异也在扩大[9] 。

从女性和男性的性别薪酬差距来说，经济体制的变革使劳动力市场原有的就业结构和收入分配关系发生了巨大变化，市场化进程的加快使女性在劳动力市场中面临多重不利因素，男女工资收入差距逐渐凸显并进一步扩大。根据北京大学进行的中国家庭追踪调查（CFPS 2010）显示，2010 年女性从业人员的平均收入仅为男性的 64％[4]。Gustafsson 和 Li（2000）根据 1988 年和 1995 年中国家庭收入调查（CHIP）的数据发

现，中国城镇劳动力市场的性别收入差距相对较小，然而，这种差距正在逐渐扩大[10]。李春玲和李实（2008）研究了改革开放后 20 多年劳动力市场中的性别收入差距问题，也发现性别收入差距继续稳步扩大[11]。

就现有文献研究而言，针对劳动力市场性别不平等表现的绝大部分研究都是基于对中国本土或国内区域的分析，鲜少有站在世界总体维度，对世界不同地理片区或不同经济发展程度国家的研究。此外，很少有研究直接分析反映性别不平等的指数，而是通过研究两性参与率、职业地位、薪酬等方面来侧面反映不平等程度。就数据新鲜度来看，大部分论文使用的数据时间都较早，鲜少有涉及到 2020 年及以后的数据。

因此基于以上问题，本文会站在世界总体维度，加入对性别不平等指数的直观图表刻画和文字分析，使用数据的最新程度也会更新到 2021 年。

1.2.2 现有文献对劳动力市场性别不平等背后原因的分析

家庭照料也会限制女性的就业选择，然而，随着两性工资差距的缩小和子女数量的减少，女性对就业参与的意愿可能会增加，因为教育人力资本的提高增加了女性选择家庭生产劳动的机会成本。Becker（2003）在新家庭经济理论中引入了人力资本要素，解释了家庭劳动和市场劳动的分配机制。他认为，家庭劳动分工不仅由男女之间的生理差异和社会身份决定，也受到男女性在人力资本方面的差异影响，进而影响女性的劳动供给。此外，教育人力资本水平的提升不仅会增加女性放弃进入劳动市场的机会成本，还会逐步提升女性的独立意识，进而影响她们参与就业的决策[12]。

社会建构理论认为社会文化与个体行动存在复杂联系，其中性别差异不仅涉及天然的生理差异，还包括社会属性差异，这构成了男性和女性之间的不平等分工和差异[7]。社会文化通过塑造女性对性别不平等观念的认知，影响着女性的就业参与。传统观念中的性别歧视和固有两性分工使得家庭责任被视为女性的天职。实际上，性别不平等观念导致的就业歧视仍然存在，通常表现为部分女性独立意识和竞争意识的削弱，选择依附于家庭或男性，并以家庭主妇的身份呈现，从而减少了就业参与[8]。一些观念强调家庭责任，如照料和家务劳动，是非生产性活动，不仅掩盖了女性作为生产者的价值，还加强了就业市场中男性和女性之间的不平等格局，导致一些女性长期负责家庭生产和照料。研究发现，“男主外，女主内”的性别分工认同普遍存在，甚至对女性的就业产生了扭曲影响[4]。

就现有文献研究而言，针对劳动力市场性别不平等背后原因的分析，大部分论文

仅涉及到了生育（子女数量）、教育（人力资本）、社会文化（性别歧视和固有两性分工），但对于具体的出生率，教育年限，两性用于家务劳动的时长等细分指标的深入分析较少。其次，鲜少有论文涉及到政治发言权这个板块，女性在议会上所持席位，在公司中担任高管的比例也很有可能影响到性别平等程度。最后，大部分研究都是基于人文社科类背景，着重于讨论不平等背后的政策措施，社会表象，较少有论文会做丰富全面的可视化来直观用图表说明数据背后蕴藏的规律。

因此基于以上问题，本文会加入政治发言权这一板块做原因分析，并对每个板块的原因挑选细分指标，做完善的可视化分析来从数据的角度出发，解释这些原因对劳动力市场性别不平等的作用。

1.2.3 现有文献对回归模型的建立和分析

袁旭宏等（2022）对性别不平等观念对男女分工、能力、婚姻和家庭角色等多个方面的影响进行了研究，发现这些观念是影响女性就业和劳动绩效的根源性因素。研究结果表明，性别不平等观念对女性的就业参与和收入产生负面影响的结论仍然稳固，特别是在农村地区，对女性的非农就业和收入产生更普遍的负面影响[3]。李虎（2023）利用中国综合社会调查（CGSS）的数据，从性别角色观念的角度入手，重点研究了这些观念对劳动参与、职业地位、工资收入以及性别差异的影响。研究旨在深化和拓展劳动力市场中性别差异的解释框架，并为理解和解析性别差异的成因以及其持续存在提供新的视角[1]。张川川和王靖雯（2020）使用 2010 年中国综合社会调查数据，实证检验了性别角色对女性劳动力市场表现的影响。研究结果显示，性别角色观念越传统的地区，女性从事受雇佣工作的概率越低，从业女性的工资收入也越低[4]。

就现有文献研究而言，绝大多数针对回归模型的建立和分析都是将性别不平等观念或性别角色观念作为自变量，讨论其对与女性就业、收入、职业地位等劳动力市场表现的影响，鲜少有论文将性别不平等作为因变量，讨论不同指标对于性别不平等的贡献度大小。

因此基于以上问题，本文会在回归建模部分将性别不平等作为响应变量，将劳动力市场表现的各类指标和原因分析部分提出的指标作为解释变量，分析影响性别不平等的因素指标。

1.2.4 论文创新点

基于对现有文献的分析和探究，本文提出创新点如下：

数据可视化部分：

（1）在地域上扩宽多样化视角的对比，站在全球范围的视角，比较不同发展程度国家、不同大洲国家以及不同圈层组织国家（如 G7、G20、欧盟成员国）之间的情况。

（2）在时间上具有新鲜度和时效性，本文使用的绝大部分数据已更新到以 2021 年为最新年份的时间点或时间段。

回归分析部分：

（1）在多元线性回归模型中考虑了教育、婚育、政治发言权、社会共识四大维度，并以细分指标反映每一维度的信息，如以预期受教育年限反应教育优劣，以 15-19 岁女性的生育率反应婚育早晚等。

（2）将女性在议会席位上所持比例纳入多元回归模型中进行分析，在回归变量的选择中引入与女性政治发言权相关的指标。

（3）将性别不平等指数作为响应变量构建回归模型，探究影响性别不平等的因素指标。

（4）在回归分析部分引入可能的交互效应，运用不同类型的模型评价指标，力求深入分析构建最优模型。

1.3 论文组织架构

本文主要内容分为三部分：劳动力市场性别差异的直观数据分析、性别不平等背后的影响因素探究、考虑性别不平等指数的回归模型构建与分析。前两部分主要通过综合利用各类图表进行数据可视化和文字分析，第三部分主要涉及回归模型的建立和调试优化模型。具体如下：

第一部分着重于世界范围内两性之间性别不平等程度的表现，包括对性别不平等指数、性别发展指数、历史性别平等指数的直观刻画，对劳动力市场上的性别薪酬差距、劳动参与率、低薪岗位人数分布、顶端收入人数分布等的数据特征体现。利用散点图、折现图、核密度图、箱线图等图表清晰地刻画每组数据在两性间的差异，并探究地理位置和时间变化对劳动力市场性别不平等的影响，以全面了解以 2021 年为最新年份的全球范围内劳动力市场的性别不平等现状。

第二部分侧重探究造成这种不平等的潜在影响因素。重点关注教育、婚育、政治

发言权和社会潜在共识等四个方面，利用图表展现这些因素与性别发展指数（GDI）、性别不平等指数（GII）以及性别薪酬差距之间的关系。通过直方图加拟合曲线、散点图加拟合曲线等可视化方式，深入全面地认识影响性别不平等的主要因素。

第三部分重点是基于 GII 作为响应变量构建回归方程进行回归分析。包含残差分析、多重共线性检验、异方差检验、回归系数显著性检验等，还将针对数据本身的特点考虑解决内生性问题，考虑交互作用，构建合理的最优多元线性回归模型。

2、劳动力市场性别差异的直观数据分析

2.1 总览性别不平等指数（GII）、性别发展指数（GDI）和历史性别平等指数（HGEI）

性别不平等指数（Gender Inequality Index）从三个方面衡量性别不平等现象，包括生殖健康（基于孕产妇死亡率和青少年出生率）、赋权（基于女性在议会席位中所占的比例和 25 岁及以上至少受过中等教育的成年女性的比例）以及经济状况（基于 15 岁及以上女性和男性人口的劳动力市场参与率）。

性别发展指数（Gender Development Index）同样从三个基本方面衡量两性不平等，包括健康（以女性和男性出生时的预期寿命来衡量）、教育程度（以男女儿童的预期受教育年限和 25 岁及以上成年人的女性和男性平均受教育年限来衡量）以及对经济资源的控制（以女性和男性的估计收入来衡量）。

尽管 GDI 和 GII 都是用来衡量性别平等程度的指标，反映了男女在教育、健康、经济等方面的平等程度，它们在侧重点和计算方法上仍存在许多差异。GDI 主要关注性别发展的整体平等，而 GII 则更侧重于揭示男女在多个领域中的不平等情况，包括健康、教育、经济、政治等方面。此外，GDI 通常使用人类发展指数（HDI）作为基础进行修正，而 GII 是单独计算得出的指数。GDI 的取值范围是 0 到 1 之间，1 表示完全性别平等，而 GII 的取值范围也是 0 到 1 之间，1 表示最大程度的性别不平等[14]。总结而言，GDI 和 GII 综合考虑了包含教育水平、健康状况、经济参与率等在内的多个领域数据，在评估不同国家或地区性别平等状况时提供了重要的信息，具体指标构成的区别如图 2-1 所示[16]。

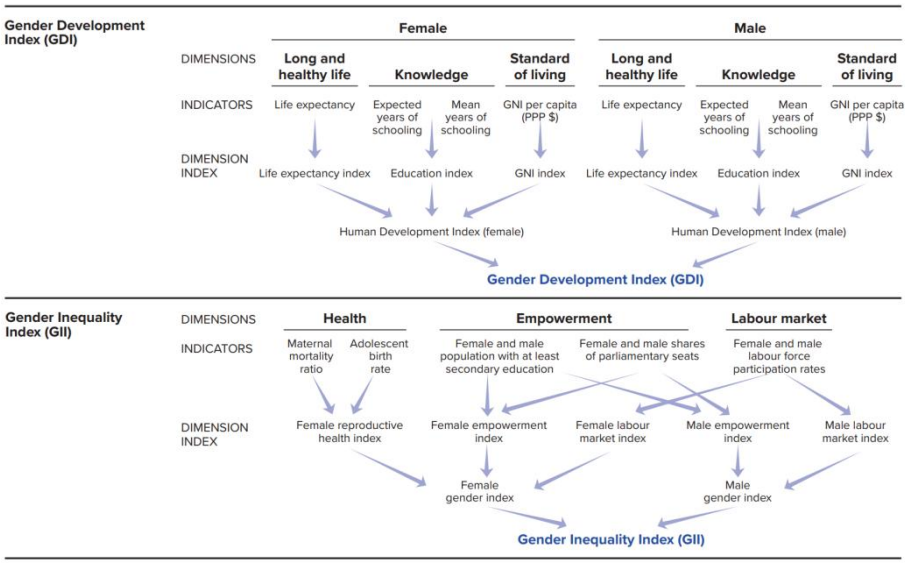

图 2－1 GDI 与 GII 指标构成差异

Figure 2-1 Differences in the composition of GDI and GII indicators

我们从联合国官方数据库收集到了 1990 至 2021 年（最新年份更新至 2021 年）的世界范围内不同国家地区的 GDI 和 GII 数据，并利用其进行可视化。如图 2-2 和 2-3 所示。

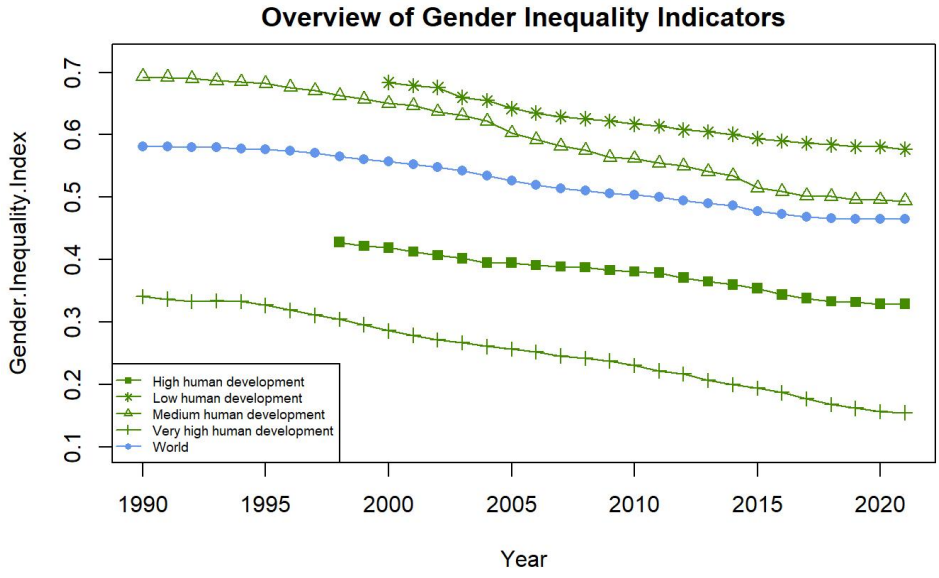

图 2－2 总览 1990 至 2021 年的性别不平等指数

Figure 2-2 Overview of Gender Inequality Indicators from 1990 to 2021

图 2-2 的折线图反映了 1990 年到 2021 年，世界四种不同发展水平的国家平均 GII 随年份变化情况，和世界总体水平的 GII 变动情况。可以看出，总体而言世界范围内 GII 都在随着时间推移而下降，性别不平等程度在逐步削弱。其次，越高水平人类发

展的地区，GII 的值越低，世界的平均值介于中等人类发展水平国家和高等人类发展水平国家之间，从 1990 年的 0.58 下降至 2021 年的 0.47。

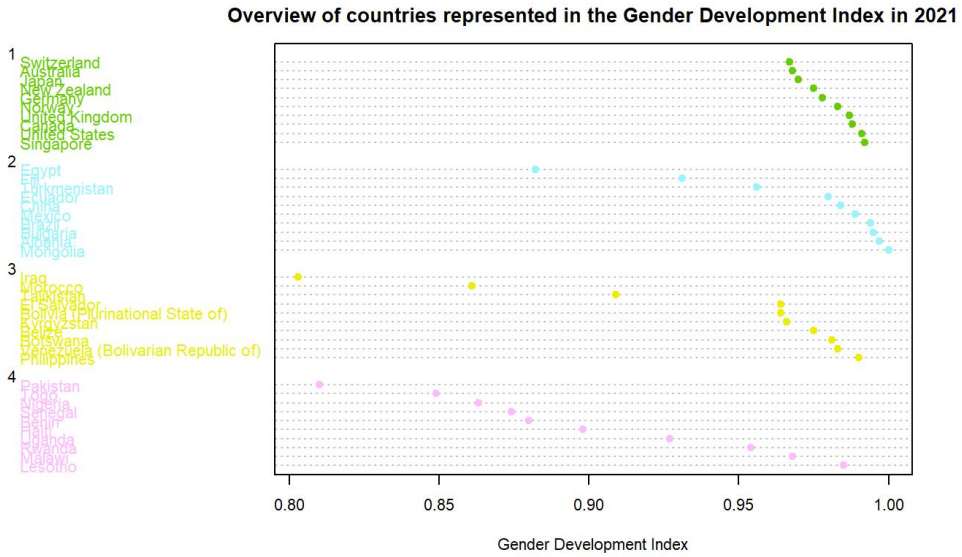

图 2－3 总览 2021 年的性别发展指数

Figure 2-3 Overview of countries represented in the Gender Development Index in 2021

从四个不同发展程度（Very high human development、High human development、Medium human development、Low human development）的国家中，以每种发展程度 10 个国家随机挑选了 40 个地区，做出分层散点图如图 2-3 所示。可以看出在 2021 年，位于极高人类发展的 10 个国家的 GDI 全部落在 0.95 到 1.00 的区间内，非常接近高度平等化，随着发展程度降低，国家之间性别平等的程度逐渐出现了差异化越来越大的现象，落在 0.95 到 1.00 区间内的国家数也越来越少。

下面对历史性别平等指数（Historical Gender Equality Index）进行研究，由于该指数更注重一个历史段内的变化，以便进行长期的综合比较，因此数据区间为 1950 至 2000 年。它涵盖四个维度：（1）健康，以预期寿命中的性别配比来衡量;（2）社会经济资源，以平均受教育年限和劳动力参与率的性别比率来衡量;（3）家庭中的性别差距，以结婚年龄的性别比例来衡量;（4）政治中的性别差距，以议会席位的性别比例来衡量。总体分数越高表示不平等越少。

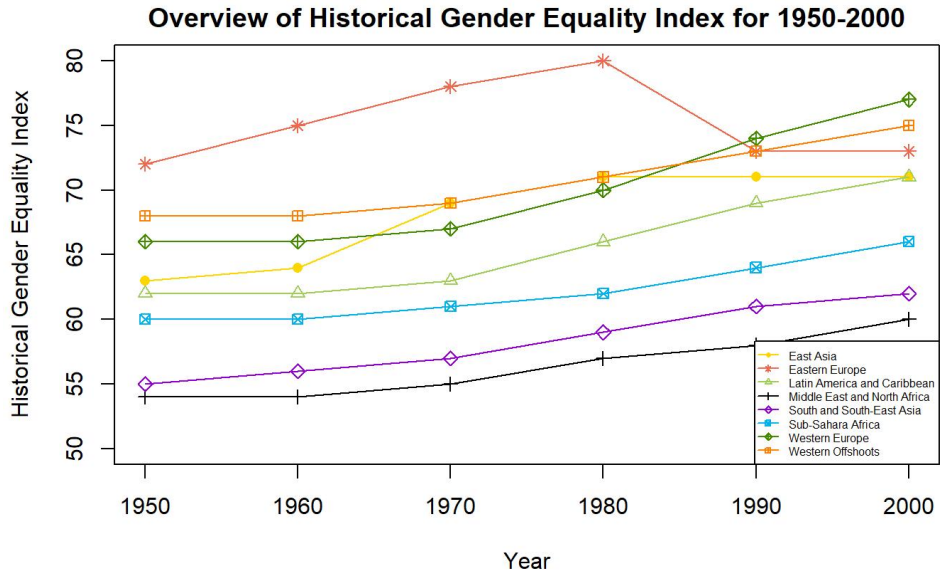

图 2－4 总览 1950 至 2000 年的历史性别平等指数

Figure 2-4 Overview of Historical Gender Equality Index for 1950-2000

图 2-4 绘制了 1950 年至 2000 年，从联合国收集到的 HGEI 变化情况。其中按照不同国家所在的地理片区划分了八个区域，分别是东亚、东欧、拉丁美洲和加勒比地区、中东和北非、南亚和东南亚、撒哈拉以南非洲、西欧、西方国家分支（澳大利亚、加拿大、新西兰和美国等原本是欧洲殖民国家的国家）。可以看出在 1980 年及以前，世界范围内性别平等程度最高的地区是东欧，在 1980 年时历史性别平等指数达到 80，1980 到 2000 年间西欧和西方国家分支逐渐超过了东欧成为世界范围内性别平等程度最高的两个区域。在数据记录期间内，中东和北非一直是性别平等程度最低的，且在这 50 年间历史性别平等指数仅上升了 6 个单位。总体而言，各个区域的指标都在随着时间推移而上升，性别平等情况越来越好。

2.2 总览性别薪酬差距数据

国际上有两个官方组织分别对性别薪酬差距的数据进行了收录，其一是联合国国际劳工组织（ILO），其二是经济合作与发展组织（OECD）[16]。其中 ILO 记录的数据以妇女劳动总收入与男子劳动总收入的比率反映两性工作收入中的性别差异，指标小于 100 表示女性收入低于男性，等于 100 表示男女平等；能收集到的最新年份为 2021 年，数据包含 82 种地域组合和 4 种对世界总体范围划分不同经济收入程度等级的值。OECD 记录的数据被定义为男性平均收入与女性平均收入之间的差额，以男性平均收入的百分比表示，数据仅考虑了全职雇员和自雇人士，不包括因兼职和全职工

人的小时工资差异而产生的差异；能收集到的最新年份为 2021 年，数据包含经合组织注册成员国和欧盟 27 国总计 39 个国家的值。

对比之下，从针对单个国家的细节维度来看，ILO 在数据质量上逊于 OECD，但从覆盖地理范围的广度和囊括统计人群的全面性来看，ILO 在数据质量上优于 OECD。因此为了站在更宏观广泛的维度总览性别薪酬数据的特征，下面将先根据 ILO 收集到的 2011 至 2021 年的性别薪酬差距数据进行可视化。

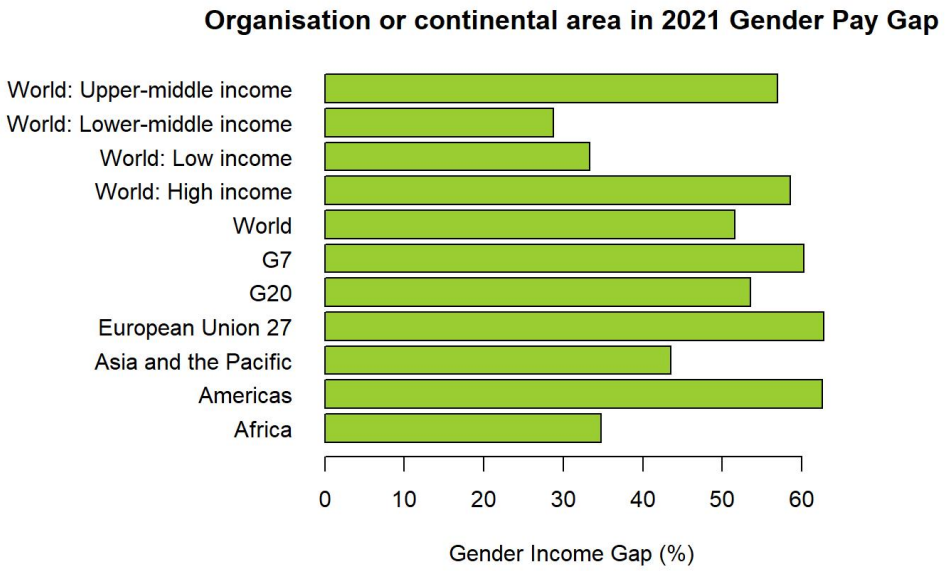

图 2－5　2021 年不同组织或大陆地区的性别薪酬差距

Figure 2-5 Organization or continental area in 2021 Gender Pay Gap

图 2-5 反映了 2021 年世界四种经济收入程度（High income、Upper-middle income、Lower-middle income、Low income）、G20 国家群、G7 国家群、欧盟 27 国、亚洲和太平洋地区、美洲地区、非洲地区、世界总体的性别薪酬差距情况。可以看出美洲和欧盟 27 国的性别薪酬差距最小，妇女总收入能占到男性总收入的 60%以上，非洲和亚洲太平洋均低于世界平均水平，而从经济收入程度来看，收入越高的地区对应的性别薪酬差距越小，中下收入地区妇女总收入仅占男性总收入不到 30%，甚至低于最低收入地区。

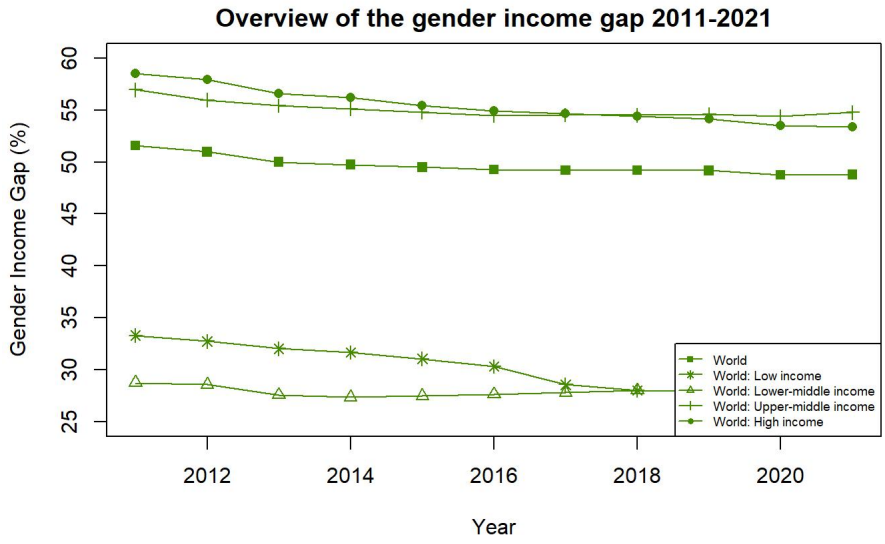

图 2－6 2011-2021 年性别薪酬差距概览

Figure 2-6 Overview of the gender income gap 2011-2021

在更长的时间跨度上来单独分析世界不同经济收入地区的性别薪酬差距，从图 2-6 可以看出世界平均水平在这 11 年内始终低于中上收入和高等收入地区，同时始终高于中下收入和低收入地区，维持在 50%左右，即从世界整体水平来看，妇女总收入在这 10 年内始终只占男子总收入的一半。从 2018 年开始，中下收入地区的差距逐渐缩短，小于了低收入地区，同时中上收入差距小于了高收入地区，处于中间位置的国家在缩小性别薪酬差距上有了较为正向的发展趋势，但极高和极低收入国家的差距仍有进一步扩大的趋势。

2.3 性别薪酬差距与各国人均 GDP 的关系（按经济发展程度划分）

为了探究性别薪酬差距和人均 GDP 的关系，我们从世界银行（World Bank）的世界发展指标集（WDI）中提取出了世界各国人均 GDP 的数值。此时为了匹配人均 GDP 的单个国家收集维度，选择使用 OECD 的性别薪酬数据与之关联。

2021 年同时在人均 GDP 和性别薪酬数据中都有记录的国家共有 28 个，按照不同的经济发展程度可以将这些国家区分为三类（高收入国家，中上收入国家，中下收入国家），将每个国家在 2021 年的性别薪酬差距和人均 GDP 刻画在散点图中，如图 2-7 所示：

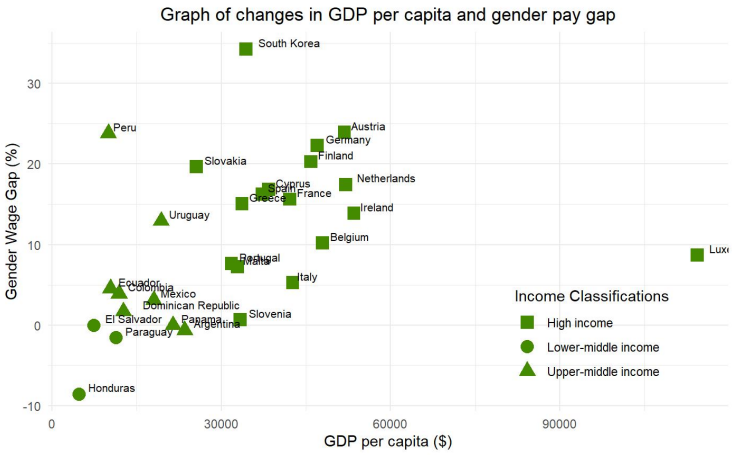

图 2－7 人均国内生产总值和性别薪酬差距关系变化图

Figure 2-7 Graph of changes in GDP per capita and gender pay gap

图 2-7 的纵轴显示未经调整的男女工资差距，计算方法是男性平均收入与女性平均收入之间的差额，以男性平均收入的百分比表示；横轴显示经过通货膨胀和各国生活成本差异调整后的人均 GDP 数额，统一以美元作为单位。可以看出三种经济发展程度的国家（高收入，中上收入，中下收入）都呈现出了随人均 GDP 升高，男女性之间的薪酬差距也进一步扩大的趋势；高收入国家普遍是性别薪酬差距最大的一批国家，以南韩为例，男性的平均收入已经比女性多出了自己工资的 34.27%；而在中下收入国家，女性收入会反过来超过男性，中上收入国家相较而言是最接近性别薪酬平等的一类地区。

2.4 性别薪酬差距与各国人均 GDP 的关系（按大洲划分）

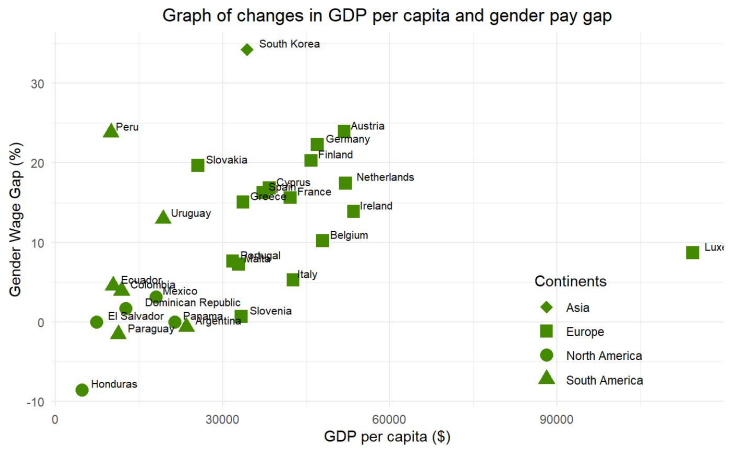

图 2－8 人均国内生产总值和性别薪酬差距关系变化图

Figure 2-8 Graph of changes in GDP per capita and gender pay gap

类似地，将划分指标从三种经济收入程度换成四种地理位置（亚洲、欧洲、北美洲、南美洲）之后，如图 2-8 所示可以看出 2021 年人均 GDP 较高且性别薪酬差距也较大的一批国家位于欧洲，南美洲和北美洲的国家较为接近男女薪酬平等。

从图 2-7 和图 2-8 的对比来看，绝大部分欧洲国家也就代表了高经济收入国家，绝大部分北美洲和南美洲的国家也就代表了中等收入国家。而我们从两张图都能发现同样的规律——中等收入国家的薪酬差距往往最小。

中等收入国家的性别工资差距很小，这在很大程度上是妇女就业选择的结果。有学者对此的解释如下：“如果就业妇女往往具有相对较高的工资特征，那么低女性就业率可能与低性别工资差距相一致，原因很简单，因为低工资妇女不会出现在观察到的工资分布中”[15]。有研究表明，这种模式在数据中是成立的：各国未经调整的性别工资差距往往与性别就业差距呈负相关，也就是说，在女性劳动力相对较少的地方，性别工资差距往往较小，因此，在一些国家观察到的较低的工资差距并不反映出更大的平等，而是可能表明只有具有某些特征的妇女，例如没有丈夫或子女的妇女才会加入劳动力大军中[17]。

由此引出了下一个研究指标——男女性劳动参与率。

2.5 两性劳动力参与率

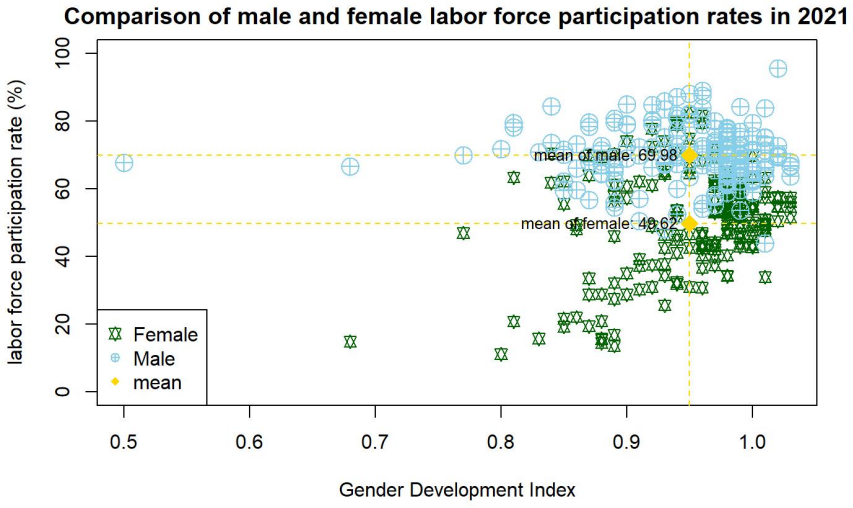

图 2－9 2021 年男女劳动力参与率比较

Figure 2-9 Comparison of male and female labor force participation rates in 2021

图 2-9 的纵轴是 15 岁及以上的男女性的劳动参与率，横轴是性别发展指数，记录了 2021 年 183 个国家的情况。从图中可以看出，在相同性别发展指数下，男性的劳动参与率明显普遍高于女性，达到平均 69.98%，而女性仅有 49.62%。这 183 个国家的平均 GDI 水平是 0.95，可以看出约接近于两性平等的国家，男女性之间的劳动参与率差距也在缩小，反之差距越大。

2.6 两性平均月收入

分析六个国家——巴西、意大利、印度、墨西哥、秘鲁、亚美尼亚在 2017 年到 2021 的五年间，男女性平均月收入大小。六个国家分别来自南美，北美，欧洲，西亚，南亚，在地理位置上相距很远[15]。以性别和年份两个分类指标对月平均收入数据进行划分，由于月收入的计量单位是当地货币，因此在没经过换算的情况下，分开对每个国家画出一幅箱线图，粉色代表女性，蓝色代表男性，箱线图如图 2-10 所示：

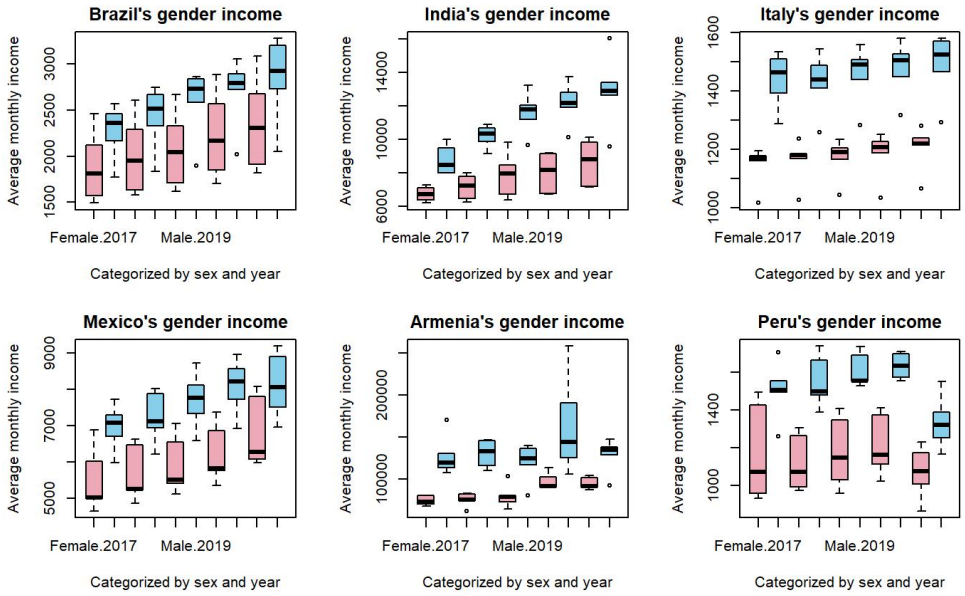

图 2－10 六个国家男女平均月收入对比图

Figure 2-10 Comparison of average monthly income of men and women in six countries

从图 2-10 可以看出，这六个国家无一例外的，女性月平均收入在这五年来都低于男性月平均收入。除意大利和亚美尼亚两个国家之外，其余四国女性月平均收入的箱图跨度更大，意味着女性的收入分布范围更广，更多样且不稳定不集中。而这四国女性收入的平均值都没有到达男性收入的下四分位值，两性收入差距非常大。

2.7 女性在低薪岗位人数占比

从国际劳工组织中收集到 2010 年到 2021 年女性在低收入岗位中的人数占比数据，共计 213 条数据。

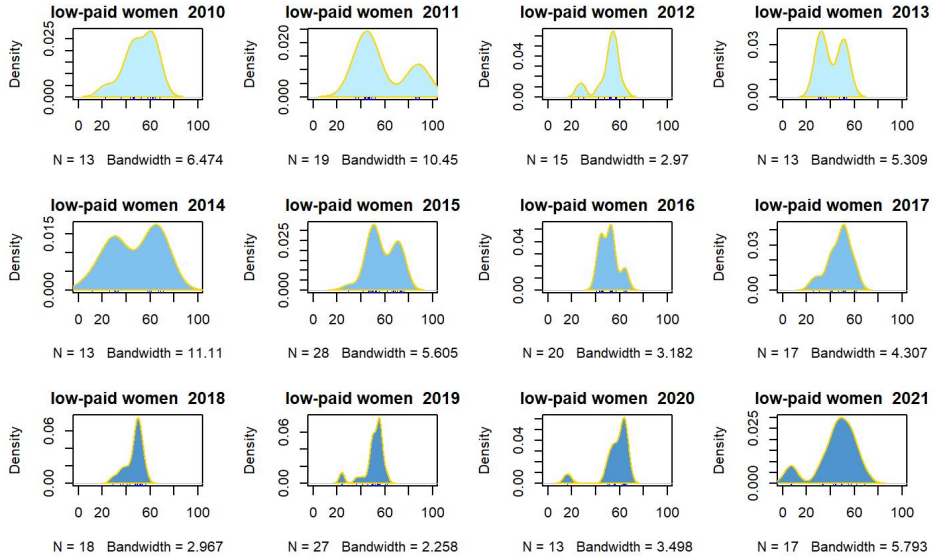

图 2－11 2010 到 2021 年女性在低薪岗位中的平均人数占比分布

Figure 2-11 Distribution of the average share of women in low-paid jobs, 2010-2021

图 2-11 中的 12 张核密度图分别刻画了 2010 到 2021 共 12 年间，每一年世界女性在低薪岗位中的平均人数占比分布。总体来说，分布的范围随着时间往后推移有一个集中缩窄的趋势，峰值所在位置也随着年份推后逐步缓慢左移。从 2016 年以后，峰值基本固定在 40%到 50%之间，这意味着世界绝大地区女性在低薪工作中的人数占比已经下降到一半以下。整体来说，女性的工薪待遇正在逐步得到重视和改善，呈现出一个好的趋势。

2.8 女性在收入分配顶端的比例

从各个国家的统计局或税务局等官方数据库收集到了澳大利亚、加拿大、丹麦、意大利、新西兰、西班牙、英国七个国家在 2000 年、2005 年、2010 年、2015 年这四年处于前 1%和 10%收入的女性比例数据，这七个国家都属于高收入国家，它们仅以个人为基础而不是以夫妻身份征税，因此不会影响对女性顶端收入的比例的计算[18]，根据四个年份做了四张柱形图：

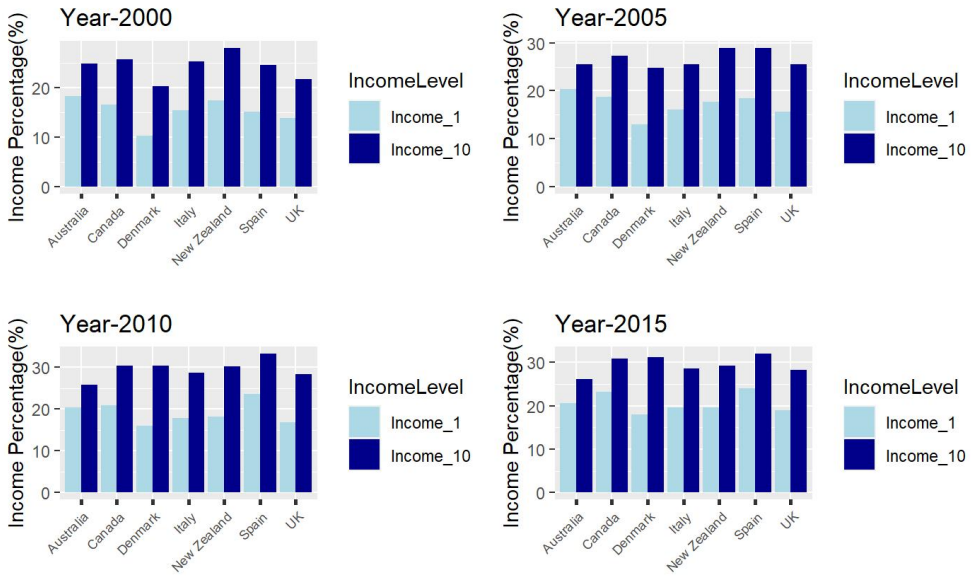

图 2－12 处于前 1%和 10%收入的女性比例数据

Figure 2-12 Data on the proportion of women in the top 1 and 10 per cent of incomes

如图 2-12 所示，随着时间推移，这些国家的女性在顶端收入群体中所占比例有细微的上升，但女性在高收入群体中的代表性仍然严重不足——在这些国家中，女性所占比例远低于 50%，在前 1%的人口中，女性约占 20%，而且各国之间的差异很小；在收入最高的 10%中，每三个收入者中就有一名妇女，总体而言，尽管最近取得了进展，但我们仍然很少能看到有女性能够在今天的收入分配中名列前茅。

2.9 小结

总结以上分析可以发现，男女两性劳动力市场表现在世界范围上呈现出的共性是，女性普遍比男性从事更低薪的工作，获得更少的薪酬，参与劳动的人数更少，处于顶端收入的人数明显不足。这种不平等差距随国家的整体文明发展程度提高而随之降低，也随时间推移而逐渐减少，但不平等现象依然存在且在低发展国家显得尤为明显[13]。缩短两性经济不平等，给世界女性带去更公平的权益和更平等的发展机会，依然是目前社会应考虑的重大议题。

因此第三章将讨论影响男女性在劳动力市场上表现出的性别不平等背后的因素。

3、性别不平等背后的影响因素探究

3.1 教育

从历史发展和社会现状的维度来看，女性和男性之间确实存在不同程度的受教育差异。教育程度在一定因素上影响了掌握技能的大小和接触工作的种类。这里对世界47个国家2021年男女性的受教育程度数据进行分析，其中选取的数据代表至少受过中等教育及以上的人数在其国家当年所占的人口比例，按照性别划分分别作出频率分布直方图和频数分布直方图，如图3-1所示。

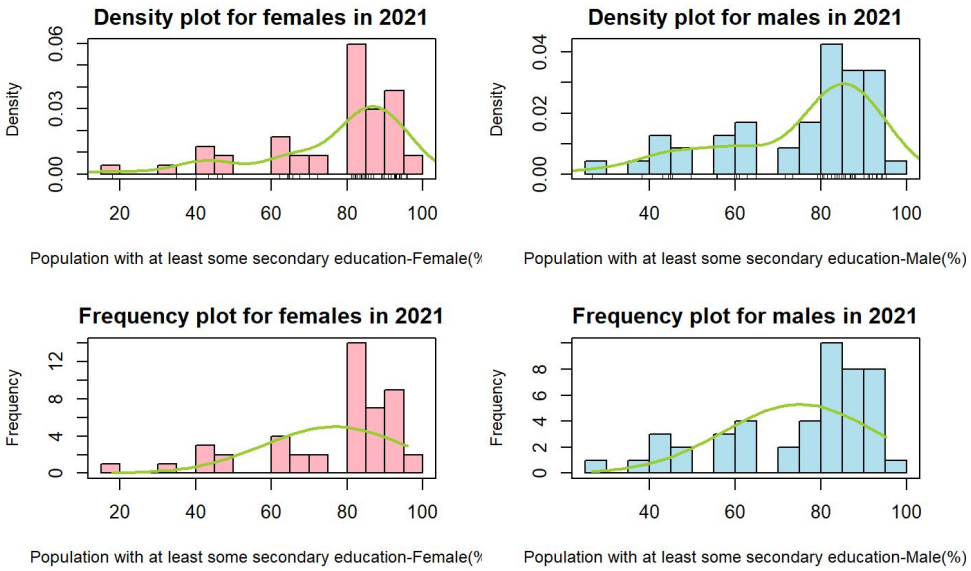

图3-1 2021年男女性的受教育程度

Figure 3-1 Educational attainment of women and men in 2021

第一行的两幅图是2021年男女性受教育程度的频率分布直方图，并加上了拟合曲线。可以看出女性出现峰值极端化，两个波峰分别出现在约43%和87%的位置，说明这47个国家中，女性受教育程度接近90%和不足50%的国家占了大多数。相比之下，男性仅有一个较为明显的波峰，出现在85%的位置，且50%以上的分布比女性更为密集，可看出虽然两性趋势类似，但男性普遍受过中等教育及以上的人数依然高于女性。第二行的两幅图是2021年男女性受教育程度的频数分布直方图，并加上了正态拟合曲线。男女性的波峰都出现在约75%的位置，总体呈现出男性每段区间的分布不如女性峰值如此明显，但整体分布更均匀，尤其50%以上的国家数更多，男性普遍受中等教育及以上的情况还是比女性稍好一些。

下面按照四个年龄段（25-34岁、35-44岁、45-54岁、55-64岁）划分人群，做出

这 47 个国家不同年龄段的至少受过中等教育及以上的人群在总人群中占比的箱线图，如图 3-2 所示：

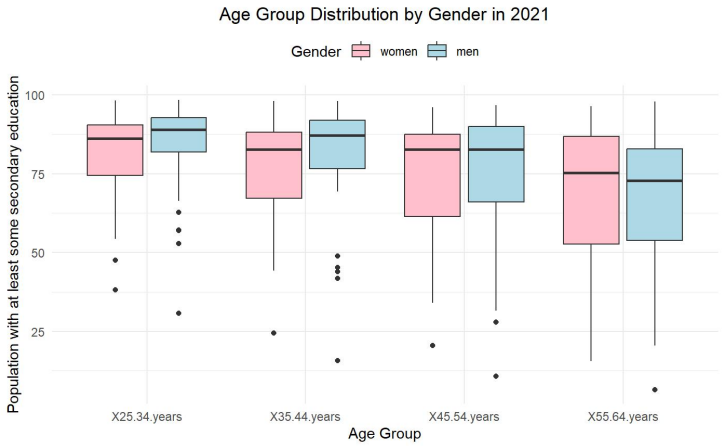

图 3-2 2021 年按性别分列的年龄组分布情况

Figure 3-2 Age Group Distribution by Gender in 2021

可以看出年龄层较小时（处于 25-44 岁内），女性至少受过中等教育及以上的人群在总女性中的占比平均值低于男性，而随着年龄增长，均值逐渐超过男性，但对于两性而言，整体来看随着年龄增长，受中等教育级以上的比例仍然是逐渐下跌的。此外，不论在哪个年龄段，女性的比例分布区间仍然比男性大很多，更多样且不稳定不集中，这说明男性整体受教育比例较高，受教育情况比女性更好。

下面按照四种受教育程度（高中以下学历、高等教育、高中或大专以上非高等教育、以上所有教育类别的综合）划分人群，做出这 47 个国家教育层级的女性相对男性收入（规定男性为 100 基准值）的箱线图，如图 3-3 所示：

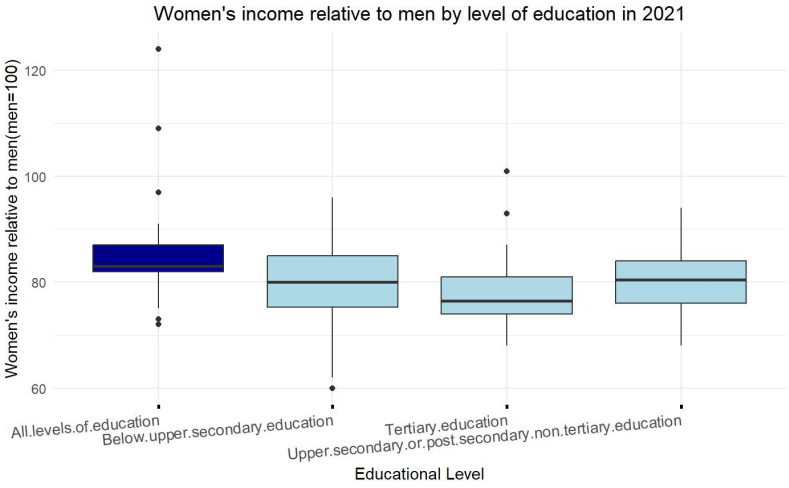

图 3-3 2021 年按教育水平分列的女性相对于男性的收入情况

Figure 3-3 Women's income relative to men by level of education in 2021

可以看出不论什么样的教育层级，女性相对男性的收入绝大部分始终没有超过100，意味着各教育阶层的女性的收入都不如同阶层的男性。具体来看，高中以下学历、高等教育、高中或大专以上非高等教育三种类别的均值都在 77 左右，但随着教育层级提高，箱型图逐渐变窄，女性的收入更加集中稳定。综合所有教育层级来看，女性相对男性收入的均值是 85.38。

3.2 婚育

女性的过早结婚是使得她们在年级很小的时候就脱离社会和正常就业市场的一大原因。大部分早婚女性在没有习得足够就业技能前就步入家庭，未来的生活很大程度上需要仰仗丈夫，这种因素影响到了女性未来的经济收入。此处在世界经合组织的性别发展门户下收集到了全世界 180 个国家 2021 年内 15 到 19 岁结婚的女性占总女性人口数的比例。下面以地域划分为四块（美洲，亚洲，非洲，欧洲），以家庭收入情况分为四类（高收入，中上收入，中下收入，低收入），共 16 类情况，以直方图的形式表达了每个种类下女性早婚分布情况。

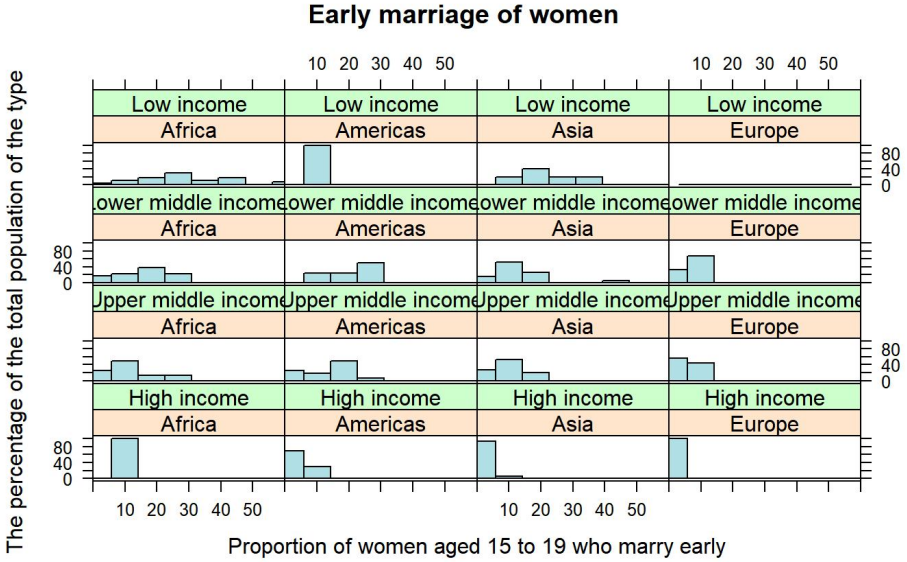

图 3-4 2021 年 15 到 19 岁女性早婚情况

Figure 3-4 Early marriages among women aged 15 to 19, 2021

从图 3-4 可以看出，随着家庭整体收入的增加，女性的早婚比例也在逐渐下降。以非洲为例，低收入群体的早婚比例从 0 到 60%均有分布，而到高收入群体时，早婚比例仅分布在 10%附近。从地域上看，非洲是早婚最严重的的地区，欧洲是早婚最少的地区。文明和经济发展程度一定程度影响了人们的婚姻观念和性别平等观念。

下面收集 2021 年全世界 197 个国家对应的女性早婚比例和性别不平等指数，筛选出两种指标都有数据记录的国家共 163 个，并对它们画散点图，如图 3-5 所示：

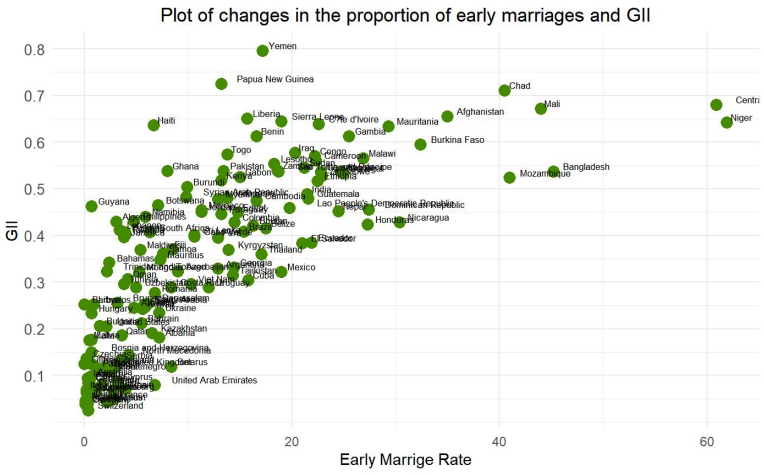

图 3-5 早婚比例和 GII 变化图

Figure 3-5 Plot of changes in the proportion of early marriages and GII

可以看出这 163 个国家的数据分布大体呈现了早婚比例越高，性别不平等程度越高的趋势，具体来看，当 GII 处于 0 到 0.3 区间内时，女性早婚比例基本不超过 10%，随着 GII 逐渐增大，同一水平的 GII 下对应的女性早婚比例跨度也在逐渐加大，最高的从 10%到 60%都有分布。

下面收集到了 7 个国家（匈牙利、丹麦、荷兰、挪威、瑞典、冰岛、芬兰）从 1973 到 2021 年的平均结婚年龄，并作出散点图，如图 3-6 所示：

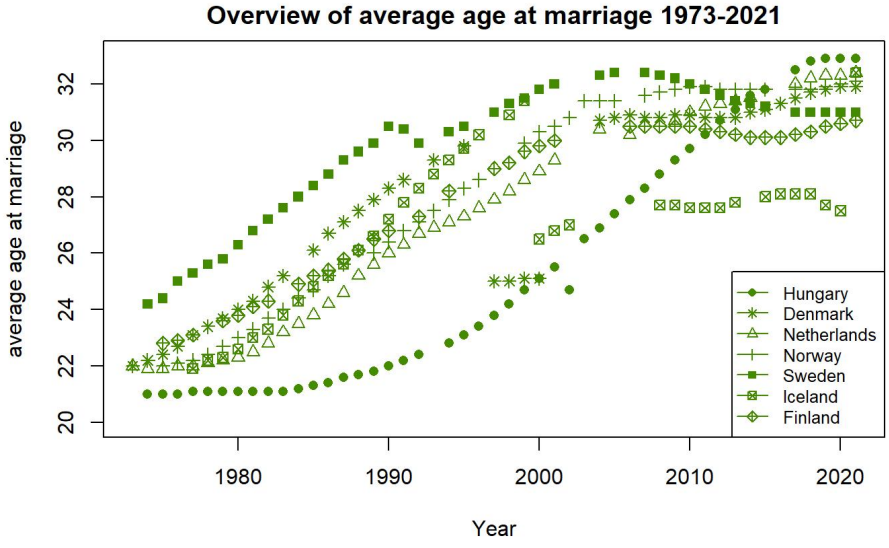

图 3-6 1973-2021 年平均结婚年龄概览

Figure 3-6 Overview of average age at marriage 1973-2021

可以看出这七国的普遍趋势都是随时间推移，女性平均结婚年龄一直在上升，以匈牙利变化最为显著，从 21.0 岁上升到了 32.9 岁。在 2003 年以前它们的上升趋势都很明显，2003 年之后逐步趋于平缓。总体而言，到了 2013 年之后，这七个国家的女性平均结婚年龄都到了 26 岁及以上，越来越趋近于正常生育或晚育。

3.3 政治发言权

女性在政治类岗位上的参与比例和在公司高等职位的在职比例会一定程度上影响到当地女性的收入状况和公司女性员工的薪水。女性执政者针对当地情况作出的扶持女性就业工作的政策可能会影响到女性的收入，同时，随之带来的性别通道也给了更多女性从事高等政治类岗位的机会。虽然直觉上这种影响是微弱且间接的，且与每个国家政体有很大关联。但依然可以通过分析直观看出这种影响在不同地域的强弱。

在世界经合组织的治理门户下收集了 2021 年亚洲、美洲、非洲、欧洲四大地区共 28 个国家的女性参政比例和当地女性的收入状况。收入已由国际货币换算机制统一换算成美元为单位的数据。

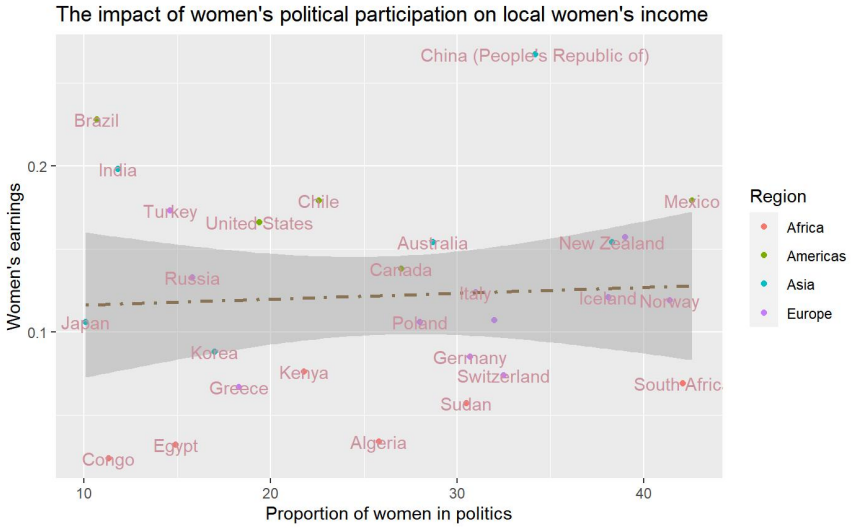

图 3-7 妇女参政对当地妇女收入的影响

Figure 3-7 The impact of women's political participation on local women's income

从图 3-7 可以看出，2021 年这 28 个国家女性的参政比例大概分布在 10%到 40%之间，女性的总体收入大小粗略排名从小到大分别是非洲，欧洲，亚洲，美洲。由拟合曲线可以看出，站在世界总体维度上考虑，女性参政比例对女性收入有较微弱的正相关性，随着女性参政比例增加，当地女性收入有一定程度上的增长。

利用收集到的 2021 年世界 197 个国家担任管理职务的女性的比例和性别不平等指数，筛选出同时在这两个指标都有数据记载的有效国家共 77 个，并根据它们所在的大洲进行分类，绘制散点图如图 3-8 所示：

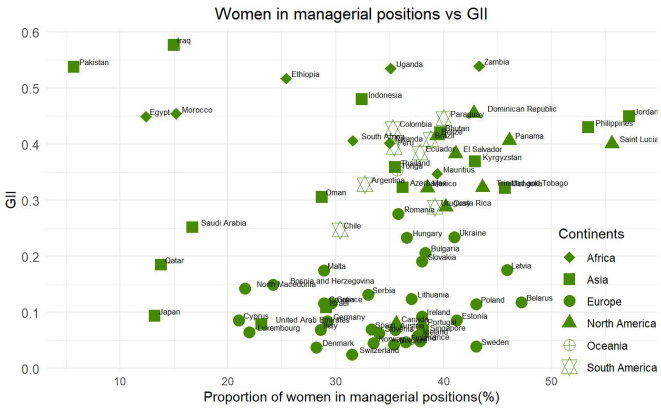

图 3-8 担任管理职务的女性与 GII 关系变化

Figure 3-8 Women in managerial positions vs GII

从图 3-8 可以看出，亚洲国家是在 GII 和女性担任管理职务比例两个指标上跨度都最大的地区，同时存在女性高管比例很少但两性较为平等和女性高管比例较高但两性较不平等的情况；非洲国家集中呈现出不论女性高管的比例处在 10%到 40%区间的哪一位置，GII 始终高于 0.35，两性平等化较弱的现象；欧洲国家女性高管比例集中在 20%到 40%区间内，同时 GII 集中于 0.2 以下，两性平等化较高；南美洲和北美洲国家女性高管比例集中于 30%到 45%区间，GII 集中于 0.3 到 0.45。不同国家在女性高管比例和 GII 的关联性上呈现出了较强的地域化特征，随地理大洲位置的改变有较大差别。但总体来说，不论哪个地区哪种性别平等程度，2021 年鲜少有国家的女性高管比例超过 50%的，世界各地的女性在高等职位工作中的代表性不足，而这些工作往往报酬更高。

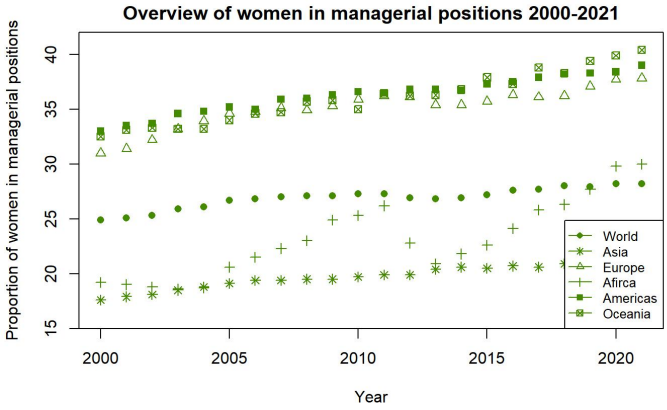

图 3-9 2000-2021 年担任管理职务的妇女概况

Figure 3-9 Overview of women in managerial positions 2000-2021

如图 3-9 所示，在更长的时间维度上观察不同大洲女性高管比例随年份的变化情况，可以看出所有地区随年份增长，女性高管比例都出现了不同幅度的上升，其中欧洲、美洲、大洋洲始终高于世界平均水平，亚洲始终远低于世界平均水平，非洲是变化幅度最大的一个地区，从 2000 年远低于世界平均的 19.2%上升至 2021 年已超过世界平均值的 30%。

下面引入一个新的指标——女性政治赋权指数，它是基于 V-Dem 数据集的专家评估指数，反映了妇女享有公民自由并在政治上有代表的程度，范围从 0 到 1（1 代表政治赋权最高）。这里收集了 2021 年世界 197 个国家女性政治赋权指数和性别发展指数，筛选出同时在这两个指标都有数据记载的有效国家共 154 个，绘制散点图如图 3-10 所示：

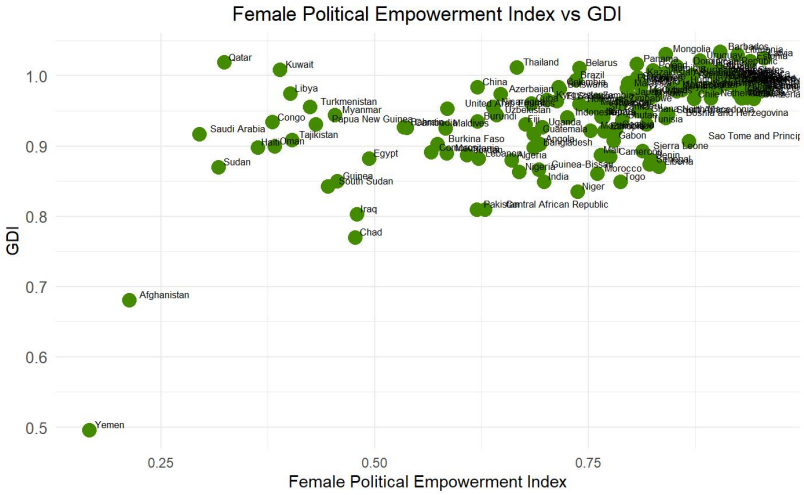

图 3-10 女性政治赋权指数和 GDI 关系变化

Figure 3-10 Female Political Empowerment Index vs GDI

可以看出，整体呈现出随女性政治权利的增大，两性平等也在逐渐增大的趋势，有女性政治赋权指数逐渐向 0.75 以上靠拢，GDI 逐渐往 0.9 以上靠拢的趋势。

3.4 社会潜在共识

长久以来，社会普遍存在一些有关两性的底层意识和默认行为习惯，如家务活天生属于一个家庭中的妻子，而银行账户和金融资产的所有权应该属于丈夫名下，在劳动力市场上，同样一份工作面对素质几乎一致的候选人时，女性有可能不被优先考虑，同样地，当付出同等水平的劳动之后，女性的薪水有可能低于男性[3]。这些潜在的社会共识直接或间接地影响了女性参与劳动力市场的能力和取得报酬的水平，进而影响

了性别平等。

首先考虑同工同酬的影响。同工同酬是指在同等工作条件下，不论性别、种族、宗教等因素，工资待遇应该是一样的，这意味着同样的工作应该获得同样的报酬，而不受到个人属性的影响。收集了 2021 年全世界 197 个国家的法律是否规定男女同工同酬（1=是；0=否）和性别不平等指数，筛选出同时在这两个指标都有记载的有效国家共 166 个，绘制直方图如图 3-11 所示：

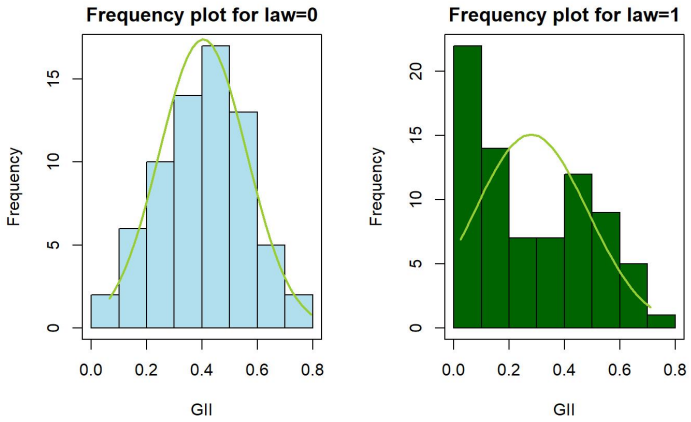

图 3-11 2021 年全世界 197 个国家的法律是否规定男女同工同酬

Figure 3-11 Whether the laws of 197 countries worldwide provide for equal pay for men and women for work of equal value in 2021

可以看出没有法律规定两性同工同酬的国家里，直方图正态拟合曲线的峰值出现在了 0.4 左右，且集中于分布在 0.2 到 0.6 的 GII 范围内；在法律有规定两性同工同酬的国家里，正态拟合线峰值出现在 0.28 左右，且集中分布于 0.0 到 0.2 及 0.4 到 0.6 的 GII 范围内，体现出了更强的性别平等性。

其次收集了 2021 年 69 个国家女性从事无偿护理工作的时间相对男性的比例，无偿护理工作是指在家庭内为家庭成员提供的所有无酬服务，包括照护人员、家务劳动和志愿社区工作，如图 3-12 所示：

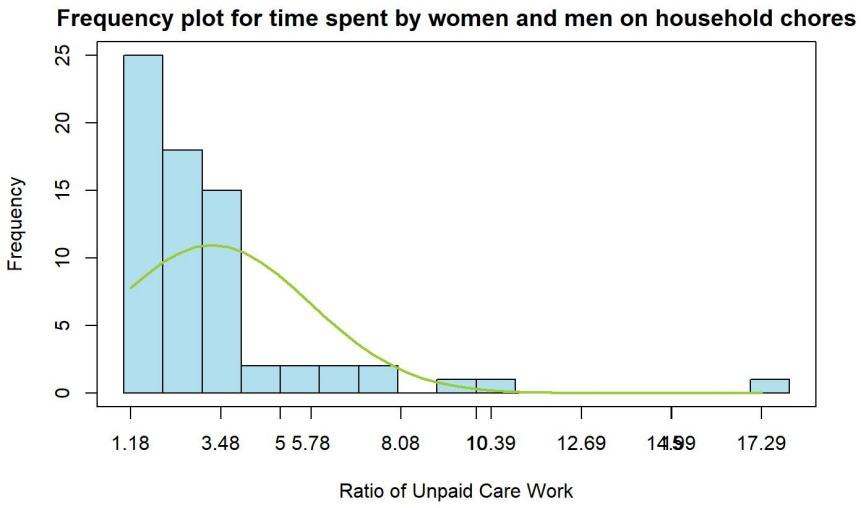

图 3-12 男女用于家务劳动的时间频数图

Figure 3-12 Frequency plot for time spent by women and men on household chores

可以看出 69 个国家记录的比例全都超过了 1，说明女性普遍都会比男性花更多的时间在无偿护理工作上，正态拟合曲线的峰值出现在 3.0 附近，并且呈很长的拖尾分布，可以看出 69 个国家中大部分女性都需要花费男性的 3 倍左右时间在家务工作上，且这个比例可以一直延伸到接近 18 倍之多，男女性在对无偿护理工作的时间付出上非常不对等。

最后收集了 2021 年男女性获得生产性资产或金融资产的能力，即能够获得银行账户所有权的比例，绘制箱线图如图 3-13 所示：

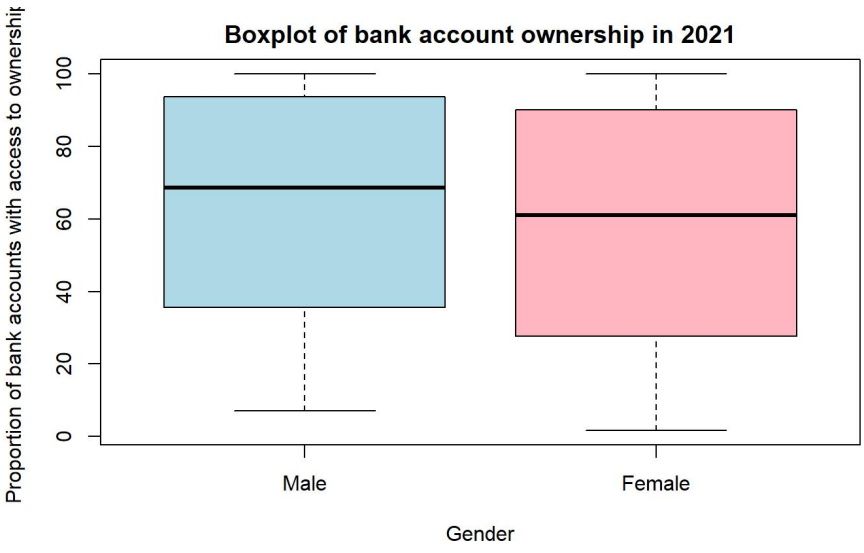

图 3-13 男女性 2021 年银行账户所有权箱型图

Figure 3-13 Boxplot of bank account ownership in 2021

可以看出，男性能够获得银行账户所有权的平均比例为 62.55%，女性为 56.92%，男性的上四分位数和下四分位数都高于女性，体现出普遍有更多男性能够拥有银行账户所有权，获得生产性资产或金融资产的能力上，女性处于劣势。

3.5 小结

总结而言，教育、婚育、政治发言权、社会潜在共识这四方面都在一定程度上对性别不平等造成了影响，以下将基于这些影响因子的分析，选取更为合适精细的指标，构建回归模型做进一步分析。

4、考虑性别不平等指数的回归模型构建与分析

4.1 响应变量和解释变量的选择

作为直观性表现每个国家性别不平等程度大小的指标，GII 和 GDI 无疑是响应变量的首选，它们分别从不同角度衡量性别不平等：GII 从生殖健康、赋权和经济状况三个方面评估，包括孕产妇死亡率、青少年出生率、女性议会席位比例和受过中等教育的成年女性比例等。GDI 则从健康、教育程度和对经济资源的控制三个方面评估，包括预期寿命、儿童受教育年限以及成年人的平均受教育年限和估计收入等。它们都在一定程度上反映了一个国家的性别不平等状况，但它们有不同的性质和重点。

GII 主要关注的是性别不平等的各个方面，包括女性在生活、经济和政治等领域的地位与机会。它考虑了女性参与就业、政治决策和健康状况等因素，并将这些因素综合起来计算一个综合指数，侧重于衡量性别在不同领域中的不平等情况，以及女性面临的挑战和机会的差异。GDI 则更关注性别在人类发展的各个方面的平等程度，它将 HDI 分解为男性和女性的子指数，用于比较两性在教育、健康和收入等领域的平等程度。GDI 强调了性别在人类发展中的平等性，将重点放在教育和健康等基本权利的平等保障上。GDI 主要关注性别发展的整体平等，而 GII 则更侧重于揭示男女在多个领域中的不平等情况，包括健康、教育、经济、政治等方面。

相比基本权利方面的平等保障，本文重点落在以劳动力市场为主，影响性别不平等的各个方面，包括就业、政治参与和健康等，且根据收集到的 2021 年世界各国 GII、GDI 数据，GII 范围在 0.01 到 0.82，GDI 范围在 0.5 到 1.03（基于 HDI 的修正会让 GDI 偶尔出现略微超过 1 的情况），从充分利用取值范围 0 到 1 的条件角度看，GII 也能更加有效地展示不同国家间的性别不平等差异。

综合以上因素考量，选取 GII 作为回归模型的响应变量。

在解释变量的选择方面，首先根据联合国出具的与 GII 测算最直接相关的方面选取了 7 个指标，分别是：孕产妇死亡率（每 10 万例活产中孕产妇的死亡人数）、青少年生育率（每千名 15-19 岁女性的生育率）、议会席位比例（女性持有议会席位的百分比）、至少受过中等教育的女性人口占 25 岁及以上女性的百分比、至少受过中等教育的男性人口占 25 岁及以上男性的百分比、15 岁及以上女性的劳动力参与率、15 岁及以上男性的劳动力参与率。

此外，根据 GDI 的测算定义，又挑选了 8 个和之前不重复的指标作为解释变量补充进来，分别是：女性出生时的预期寿命、男性出生时的预期寿命、女性预期受教育年限、男性预期受教育年限、女性平均受教育年限、男性平均受教育年限、女性人均国民总收入（以 2017 年购买力平价美元作为单位）、男性人均国民总收入（以 2017 年购买力平价美元作为单位）。

以上指标已经涵盖了第二章分析的四大影响因素——教育、婚育、政治发言权、社会潜在共识。最后再根据每个国家所处的发展程度赋予了它们一列定性变量，1 代表极高人类发展水平，2 代表高人类发展水平，3 代表中等人类发展水平，4 代表低等人类发展水平。将这一列作为虚拟变量加入回归模型中。

全模型共计包含一个响应变量，15 个解释变量，一个虚拟变量。

4.2 回归模型的构建与分析

4.2.1 数据预处理

从联合国收集到了 2021 年全球 191 个国家的 GII 数据和 15 个解释变量的数据，并根据联合国出具的对这些国家发展程度的划分为它们赋予了虚拟变量的值，以国家名作为行名称，将这 15 个解释变量，一个虚拟变量，一个响应变量，总计 17 列整合到一起，作为基础分析的数据集。

经过筛查，191 个国家中共有 11 行带有缺失值，且都缺失了关键值 GII，考虑到 GII 在每个国家中都有综合反映了这个国家政治经济健康等方面的两性不平等的性质，任何补全缺失值的方式都会对真实情况造成较大的偏颇，加之很多缺失 GII 的国家对应的解释变量数据也不尽完整，故在带有缺失值行数不算多的情况下，考虑将这 11 行全部删除，保留剩下的 180 个国家数据。

图 4-1 展示预处理过后的数据集的前六行的前四列：

Description: df [6 × 17]

	degree.of.development <int>	GII <chr>	Maternal.mortality.ratio <chr>	Adolescent.birth.rate <dbl>
Switzerland	1	0.02	5	2.2
Norway	1	0.02	2	2.3
Iceland	1	0.04	4	5.4
Australia	1	0.07	6	8.1
Denmark	1	0.01	4	1.9
Sweden	1	0.02	4	3.3

6 rows | 1-5 of 17 columns

图 4-1 预处理过后的数据集的部分展示

Figure 4-1 Partial presentation of the preprocessed dataset

针对响应变量 GII 是否需要根据帕累托原则做对数变换进行分析，画出原本 GII 的直方图和取对数之后的直方图，如图 4-2 所示：

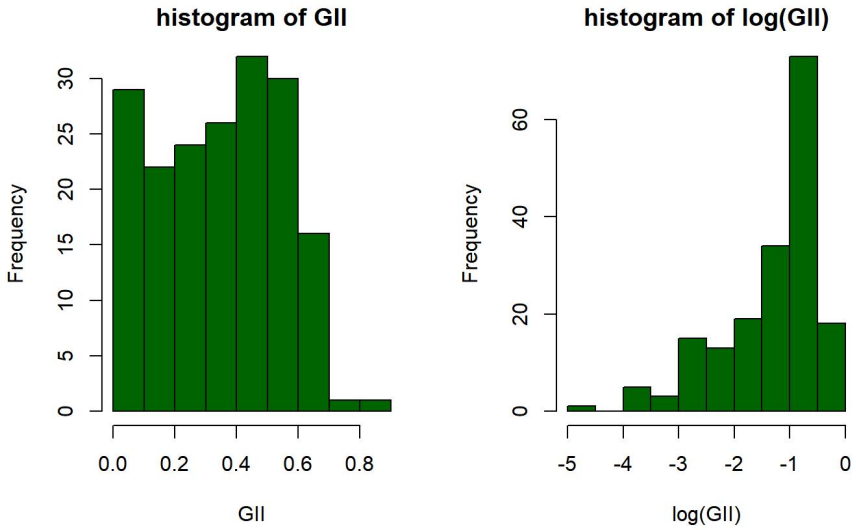

图 4-2 GII 的直方图和 GII 取对数之后的直方图

Figure 4-2 Histogram of GII and histogram after GII taking logarithms

可以看出原始数据的分布是稍微偏斜的，而对数变换后的数据分布成了长尾分布，对数变换后的分布并没有改善数据的偏斜性，反而出现了长尾分布，因此认为对 GII 进行对数变换非必要。故保留原始 GII 进行后续分析。

4.2.2 描述性统计

将虚拟变量转化为因子型变量，其余变量转化为数值型变量，进行描述性统计，部分结果如表 4-1 所示：

	GII	Maternal mortality ratio	Adolescent birth rate
Min.	0.01	2.00	1.90
1st Qu.	0.18	12.00	11.50
Median	0.37	53.00	34.70
Mean	0.35	158.60	44.78
3rd Qu.	0.51	185.00	64.90
Max.	0.82	1150.00	170.50

表 4-1 描述性统计部分结果展示

Table 4-1 Presentation of the results of the descriptive statistics section

这里展示了数据集里前三个指标的描述性统计结果，可以看出 180 个国家里 GII 的取值范围为 0.01 到 0.82，平均值和中位数都在 0.35 附近；每 10 万例活产中孕产妇的死亡人数取值范围为 2 到 1150，平均值为 159，中位数为 53；每千名 15-19 岁女性的生育率的取值范围为 1.90 到 170.50，平均值为 44.78，中位数为 34.70。可以看出同一指标在不同国家间存在明显差异性，数据范围跨度都较大。

4.2.3 构建全模型

首先使用 model.matrix 函数将 degree.of.development 这个分类变量转换为虚拟变量矩阵。对于这个有 4 个不同水平的分类变量，R 会为每个水平创建一个虚拟变量（0 或 1），并且将这些虚拟变量按照列的方式组合成一个矩阵。这样就可以在回归分析中使用这个虚拟变量矩阵来代替原始的分类变量，从而能够更好地处理分类变量对回归分析的影响。

之后构建响应变量 GII 和 15 个解释变量及一个虚拟变量矩阵的回归全模型，并对该模型使用 summary 函数输出回归全模型的摘要信息，如图 4-3 所示：

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	0.972	0.124	7.839	0.0000	***
data$degree.of.development1	-0.074	0.023	-3.208	0.0016	**
data$degree.of.development2	-0.026	0.019	-1.351	0.1785	
data$degree.of.development3	-0.010	0.019	-0.498	0.6191	
data$degree.of.development4	-0.022	0.024	-0.889	0.3755	
data$Maternal.mortality.ratio	0.000	0.000	1.143	0.2547	
data$Adolescent.birth.rate	0.001	0.000	5.984	0.0000	***
data$Share.of.seats.in.parliament	-0.004	0.000	-9.167	0.0000	***
data$Population.with.at.least.some.secondary.education.Female	-0.001	0.001	-0.744	0.4577	
data$Labour.force.participation.rate..Female	-0.003	0.000	-5.855	0.0000	***
data$Life.expectancy.at.birth.Female	-0.007	0.003	-2.233	0.0270	*
data$Expected.years.of.schooling.Female	0.011	0.006	1.910	0.0579	.
data$Mean.years.of.schooling.Female	-0.006	0.011	-0.559	0.5767	
data$Estimated..gross.national.income.per.capita.Female	0.000	0.000	1.185	0.2377	
data$Population.with.at.least.some.secondary.education.Male	-0.001	0.001	-0.606	0.5452	
data$Labour.force.participation.rate..Male	0.002	0.001	2.921	0.0040	**
data$Life.expectancy.at.birth.Male	-0.000	0.003	-0.102	0.9190	
data$Expected.years.of.schooling.Male	-0.008	0.006	-1.481	0.1406	
data$Mean.years.of.schooling.Male	0.009	0.011	0.803	0.4231	
data$Estimated..gross.national.income.per.capita.Male	-0.000	0.000	-2.251	0.0258	*

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 0.05331 on 160 degrees of freedom

Multiple R-squared: 0.9334, Adjusted R-squared: 0.9255

F-statistic: 118 on 160 and 19 DF, p-value: 0.0000

图 4-3 全模型的摘要信息

Figure 4-3　Summary information for the full model

可以看出，在 0.05 显著性水平下，显著的变量有：极高人类发展的发展程度、青少年生育率、女性持有议会席位比例、女性劳动参与率、女性出生时的预期寿命、男性劳动参与率、男性人均国民总收入。R-squared 为 0.9334 考虑了模型自由度的 Adjusted R-squared 为 0.926，因变量对响应变量的解释程度达到了 92%以上。

对模型做残差分析如图 4-4 所示：

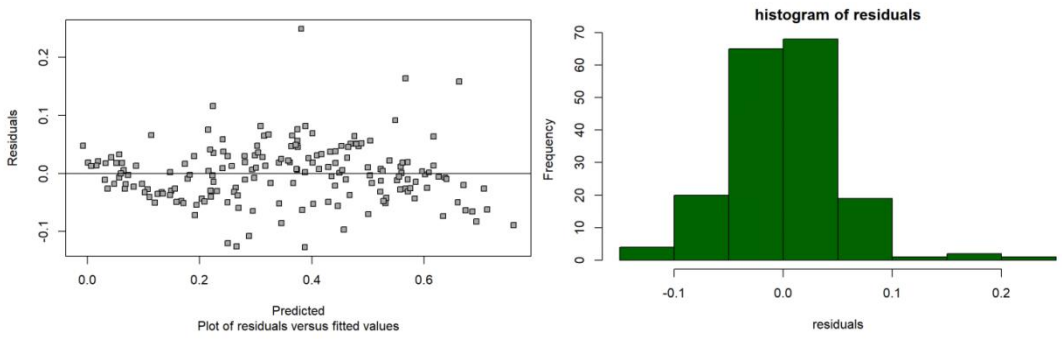

图 4-4 全模型残差分析

Figure 4-4　Full model residual analysis

散点图展示了残差与拟合值之间的关系，残差在 0 附近均匀分布，没有明显的模式或趋势，说明残差的正态性较好；直方图展示了残差的分布情况，残差呈现近似正态分布的钟形曲线，说明残差符合正态性假设。

4.2.4 构建限制模型

通过刚才全模型的分析可以知道，在实际构建回归方程时，部分变量对于解释响应变量更为重要，因此考虑只用这七个全模型中显著的解释变量来拟合回归模型，这实际上是一个限制模型，其他变量被限制为 0。

再一次使用 summary 函数输出限制模型的摘要信息如图 4-5 所示：

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	0.969	0.099	9.776	0.0000	***
data$degree.of.development1	-0.053	0.017	-3.138	0.0020	**
data$Adolescent.birth.rate	0.002	0.000	7.358	0.0000	***
data$Share.of.seats.in.parliament	-0.003	0.000	-8.168	0.0000	***
data$Labour.force.participation.rate..Female	-0.002	0.000	-6.372	0.0000	***
data$Life.expectancy.at.birth.Female	-0.009	0.001	-6.814	0.0000	***
data$Labour.force.participation.rate..Male	0.003	0.001	4.220	0.0000	***
data$Estimated..gross.national.income.per.capita.Male	-0.000	0.000	-3.479	0.0006	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 0.06055 on 172 degrees of freedom

Multiple R-squared: 0.9076, Adjusted R-squared: 0.9038

F-statistic: 241.3 on 172 and 7 DF, p-value: 0.0000

图 4-5 限制模型摘要信息

Figure 4-5 Summary information for the restricted model

在这个限制性模型中，所有解释变量全部高度显著，残差标准误差为 0.061，表明实际数据点与预测值的平均偏差为 0.061，略高于全模型；由于模型中包含的变量减少了，多重相关系数 R-squared 为 0.908，Adjusted R-squared 为 0.904，比全模型小，这表明自变量在回归模型中对因变量变异性的整体解释能力相比全模型略有下降，但由于解释程度仍然在 90%以上，故限制模型的效果依旧是能够被肯定的。经过计算，限制模型残差平方和 RSS 为 0.631，大于完整模型中的 RSS（0.455），这与我们的推测一致，但在该模型中，F 统计量为 241.3，大于全模型中的 118，表示在这一指标评判上限制模型的拟合程度相对较好，并且回归平方和相对于残差平方和的贡献较大。总结而言，限制模型将因变量的个数从 19 减少到 7，大大减小了模型复杂度，且在这样的情况下因变量对响应变量的解释程度仍然到达了 90%以上，且根据 F 检验在统计上是高度显著的。在牺牲了部分残差标准误和残差平方和的条件下，这样的结果可以被接受。

4.2.5 最优子集筛选模型调优

最优子集选择是一种变量选择方法，用于在给定的自变量集合中找到最佳的子集，以构建一个具有最佳拟合度和预测性能的回归模型，运用 R 中的 leaps 包，以 BIC 作为评价指标的最优子集筛选结果，如图 4-6 所示：

```
BIC
BICq equivalent for q in (0.484285553478017, 0.782338955494874)
Best Model:
                                              Estimate   Std. Error    t value     Pr(>|t|)
(Intercept)                                 0.8878504403 1.085024e-01   8.182776 5.935138e-14
Maternal.mortality.ratio                    0.0002939284 9.601845e-05   3.061166 2.558871e-03
Labour.force.participation.rate..Female     0.0022122729 9.217745e-04   2.400015 1.746210e-02
Expected.years.of.schooling.Female         -0.0250945529 8.341414e-03  -3.008429 3.020463e-03
Mean.years.of.schooling.Female             -0.0316872661 7.825244e-03  -4.049365 7.755404e-05
degree.of.development1                     -0.2063011355 6.626982e-02  -3.113048 2.169082e-03
degree.of.development2                     -0.3334645858 5.541281e-02  -6.017825 1.036976e-08
degree.of.development3                     -0.5462872580 4.418690e-02 -12.363104 1.569025e-25
```

图 4-6 最优子集筛选结果

Figure 4-6 Optimal subset filtering results

可以看出使用 R 包筛选的最优模型和 2.4 中的限制模型在解释变量筛选上存在很大差异，故对该 best model 做一次回归摘要信息的输出，如图 4-7 所示：

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	0.747	0.051	14.767	0.0000	***
data$degree.of.development1	-0.109	0.031	-3.537	0.0005	***
data$degree.of.development2	-0.006	0.026	-0.242	0.8094	
data$degree.of.development3	0.030	0.021	1.458	0.1466	
data$Maternal.mortality.ratio	0.000	0.000	4.562	0.0000	***
data$Labour.force.participation.rate..Female	-0.002	0.000	-3.744	0.0002	***
data$Expected.years.of.schooling.Female	-0.012	0.004	-2.990	0.0032	**
data$Mean.years.of.schooling.Female	-0.018	0.004	-4.957	0.0000	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 0.08036 on 172 degrees of freedom

Multiple R-squared: 0.8372, Adjusted R-squared: 0.8306

F-statistic: 126.4 on 172 and 7 DF, p-value: 0.0000

图 4-7 best model 的回归摘要信息

Figure 4-7 Regression summary information for best model

从比限制模型更大的残差标准误和更小的多重相关系数 R-squared、Adjusted R-squared、F 统计量可以看出，使用 leaps 包自动挑选出的最优模型效果并不好，但该最优模型中的确挑出了三个限制模型中没有考虑到的解释变量，分别是：孕产妇死亡率、女性预期受教育年限、女性平均受教育年限。

故下面将这三个解释变量加入限制模型中，再一次输出回归模型摘要信息，如图 4-8 所示：

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	0.878	0.114	7.688	0.0000	***
data$degree.of.development1	-0.048	0.016	-2.986	0.0032	**
data$Maternal.mortality.ratio	0.000	0.000	1.399	0.1636	
data$Adolescent.birth.rate	0.001	0.000	6.170	0.0000	***
data$Share.of.seats.in.parliament	-0.003	0.000	-8.592	0.0000	***
data$Labour.force.participation.rate..Female	-0.002	0.000	-6.634	0.0000	***
data$Life.expectancy.at.birth.Female	-0.006	0.001	-4.285	0.0000	***
data$Labour.force.participation.rate..Male	0.002	0.001	3.727	0.0003	***
data$Estimated..gross.national.income.per.capita.Male	-0.000	0.000	-3.191	0.0017	**
data$Expected.years.of.schooling.Female	0.004	0.003	1.356	0.1770	
data$Mean.years.of.schooling.Female	-0.013	0.002	-5.461	0.0000	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 0.05489 on 169 degrees of freedom

Multiple R-squared: 0.9254, Adjusted R-squared: 0.921

F-statistic: 209.6 on 169 and 10 DF, p-value: 0.0000

图 4-8 限制模型的回归摘要信息

Figure 4-8 Regression summary information for restricted model

根据结果显示，将这三个解释变量放入限制模型后，孕产妇死亡率和女性预期受教育年限这两个指标再次不显著，故删除它们，仅加入女性平均受教育年限这一个解释变量后，再次输出回归模型摘要信息和 ANOVA 表，如图 4-9 和图 4-10 所示：

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	0.976	0.090	10.819	0.0000	***
data$degree.of.development1	-0.041	0.016	-2.636	0.0092	**
data$Adolescent.birth.rate	0.001	0.000	6.367	0.0000	***
data$Share.of.seats.in.parliament	-0.003	0.000	-8.461	0.0000	***
data$Labour.force.participation.rate..Female	-0.002	0.000	-6.362	0.0000	***
data$Life.expectancy.at.birth.Female	-0.007	0.001	-5.595	0.0000	***
data$Labour.force.participation.rate..Male	0.002	0.001	3.387	0.0009	***
data$Estimated..gross.national.income.per.capita.Male	-0.000	0.000	-2.943	0.0037	**
data$Mean.years.of.schooling.Female	-0.013	0.002	-6.058	0.0000	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 0.0551 on 171 degrees of freedom

Multiple R-squared: 0.9239, Adjusted R-squared: 0.9204

F-statistic: 259.6 on 171 and 8 DF, p-value: 0.0000

图 4-9 回归模型摘要信息

Figure 4-9 Regression model summary information

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
data$degree.of.development1	1	3.882	3.882	1278.67	< 2e-16	***
data$Adolescent.birth.rate	1	1.328	1.328	437.40	< 2e-16	***
data$Share.of.seats.in.parliament	1	0.591	0.591	194.63	< 2e-16	***
data$Labour.force.participation.rate..Female	1	0.096	0.096	31.62	7.50e-08	***
data$Life.expectancy.at.birth.Female	1	0.199	0.199	65.44	1.06e-13	***
data$Labour.force.participation.rate..Male	1	0.053	0.053	17.51	4.56e-05	***
data$Estimated..gross.national.income.per.capita.Male	1	0.044	0.044	14.61	0.000185	***
data$Mean.years.of.schooling.Female	1	0.111	0.111	36.70	8.52e-09	***
Residuals	171	0.519	0.003			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

图 4-10 方差分析表

Figure 4-10 ANOVA Table

此时所有解释变量全部显著，并且相较最初的限制模型而言，残差标准误更小，为 0.055；多重相关系数 R-squared 更大，为 0.924；Adjusted R-squared 更大，为 0.920；F 统计量更大，为 259.6，从以上指标来看现在的模型都优于最初的限制模型。故将该模型确立为目前的最优模型，在此模型基础上做后续分析。该模型表达式如下：

GII=0.976-0.041*是否是极高人类发展程度国家+0.001*青少年出生率-0.003*女性持有议会席位比例-0.002*女性劳动参与率-0.007*女性出生时的预期寿命+0.002*男性劳动参与率-0.000001*男性人均国民总收入-0.013*女性平均受教育年限

4.2.6 置信区间和预测区间

一般来说，假设α'是一个 p 维向量，β是线性回归模型的系数向量，则：

$$\mathrm{CI}(\alpha'\beta)=[\alpha'\hat{\beta}_{LS}\pm\sqrt{\alpha'(X'X)^{-1}\alpha}\hat{\sigma}t_{n-p}(\frac{\alpha}{2})],$$

$$\mathrm{PI}(\alpha'\beta)=[\alpha'\hat{\beta}_{LS}\pm\sqrt{1+\alpha'(X'X)^{-1}\alpha}\hat{\sigma}t_{n-p}(\frac{\alpha}{2})].$$

回归均值的置信区间：

$$[\hat{Y}_i\pm t_{1-\alpha/2}(n-2)\hat{\sigma}\Big/\sqrt{1/\mathrm{n}+(X_i-\overline{X})^2/l_{xx}}].$$

响应变量的预测区间：

$$[\hat{Y}_{\mathrm{h}}\pm t_{1-\alpha/2}(n-2)\hat{\sigma}\Big/\sqrt{1+1/\mathrm{n}+(X_h-\overline{X})^2/l_{xx}}].$$

在置信水平α为 0.05 下，回归参数的置信区间如图 4-11 所示：

	2.5 %	97.5 %
(Intercept)	7.977579e-01	1.153810e+00
data$degree.of.development1	-7.180661e-02	-1.030819e-02
data$Adolescent.birth.rate	9.542070e-04	1.811690e-03
data$Share.of.seats.in.parliament	-3.944316e-03	-2.452102e-03
data$Labour.force.participation.rate..Female	-2.792922e-03	-1.470265e-03
data$Life.expectancy.at.birth.Female	-8.917620e-03	-4.266469e-03
data$Labour.force.participation.rate..Male	8.165120e-04	3.096915e-03
data$Estimated..gross.national.income.per.capita.Male	-1.692301e-06	-3.334670e-07
data$Mean.years.of.schooling.Female	-1.718652e-02	-8.739129e-03

图 4-11 回归参数的置信区间

Figure 4-11 Confidence intervals for regression parameters

回归均值的置信区间如图 4-12 所示（此处仅展示前五行结果）：

	fit	lwr	upr
Switzerland	-0.0005152644	-0.024023690	0.02299316
Norway	0.0014545448	-0.021136535	0.02404562
Iceland	-0.0036732677	-0.024440936	0.01709440
Australia	0.0389537466	0.021151133	0.05675636
Denmark	0.0258033120	0.005098171	0.04650845

图 4-12 回归均值的置信区间

Figure 4-12 Confidence intervals for regression to the mean

响应变量的预测区间如图 4-13 所示（此处仅展示前五行结果）：

	fit	lwr	upr
Switzerland	-0.0005152644	-0.11179375	0.1107632
Norway	0.0014545448	-0.10963377	0.1125429
Iceland	-0.0036732677	-0.11440516	0.1070586
Australia	0.0389537466	-0.07126054	0.1491680
Denmark	0.0258033120	-0.08491687	0.1365235

图 4-13 响应变量的预测区间

Figure 4-13 Prediction intervals for response variables

预测区间反映的是单个值的不确定性，而置信区间反映的是平均预测值的不确定性。因此，对于相同的值，预测区间通常比置信区间宽得多，而图 4-14 直观地显示了这一点（阴影区域为 CI，绿色边线为 PI 边界）：

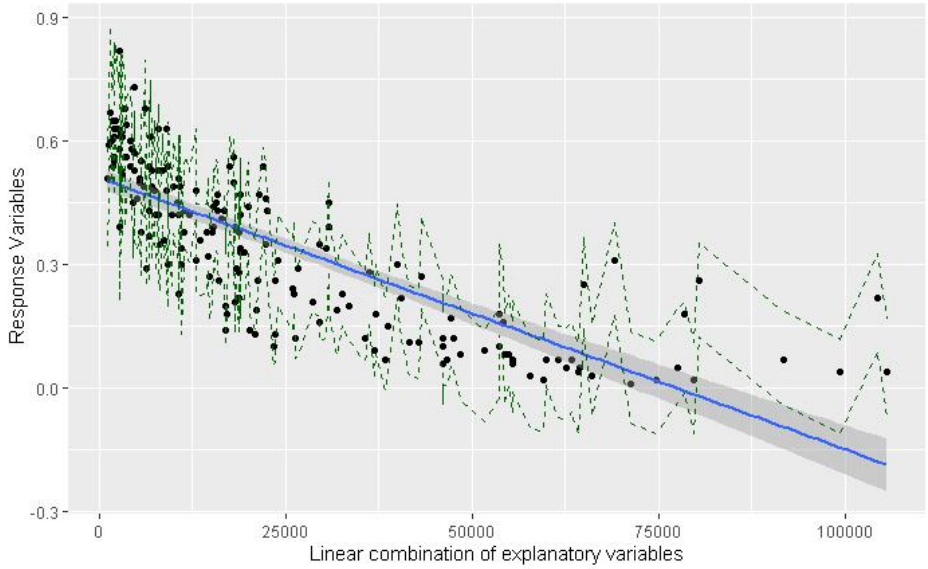

图 4-14 置信区间和预测区间的关系图

Figure 4-14 Plot of confidence intervals against prediction intervals

4.2.7 多重共线性检验

多重共线性是指在多元回归模型中，两个或多个解释变量之间存在高度相关性的情况，这种情况下，解释变量之间的相关性会引入如估计系数不稳定之类的问题。下面将分别用三种方法检测回归模型是否存在多重共线性[2]。

方差膨胀因子法： $VIF = 1/(1-R^2)$

```
> vif(Restrict.lm1)
                                    data$degree.of.development1
                                                       3.223132
                                data$Share.of.seats.in.parliament
                                                       1.222295
                              data$Life.expectancy.at.birth.Female
                                                       4.988104
data$Estimated..gross.national.income.per.capita.Male
                                                       3.826214
                                    data$Adolescent.birth.rate
                                                       4.137081
                  data$Labour.force.participation.rate..Female
                                                       1.547966
                   data$Labour.force.participation.rate..Male
                                                       1.341313
                         data$Mean.years.of.schooling.Female
                                                       3.234681
```

图 4-15 方差膨胀因子的检测

Figure 4-15 Detection of variance inflation factors

一般认为当 0<VIF<10 时，不存在多重共线性；当 10≤VIF<100 时，存在较强的多重共线性，当 VIF≥100，多重共线性非常严重。如图 4-15 所示，我们的回归模型所有解释变量的 VIF 都在 10 以下，不存在多重共线性。

Kappa 系数法：

一般认为 kappa<100 时，设计矩阵多重共线性的程度很小；$100 \le kappa \le 1000$ 时，设计矩阵存在较强的多重共线性；kappa>1000 时，存在严重的多重共线性。经由计算，我们的回归模型 kappa 系数为 28.81，多重共线性程度很小。

简单相关系数法：

相关系数热力图如图 4-16 所示（此处为了图表美观，解释变量名称暂时用数字替代，1-8 分别代表：是否是极高发展水平国家、青少年出生率、女性持有议会席位比例、女性劳动参与率、女性出生时的预期寿命、男性劳动参与率、男性人均国民收入、女性平均受教育年限）：

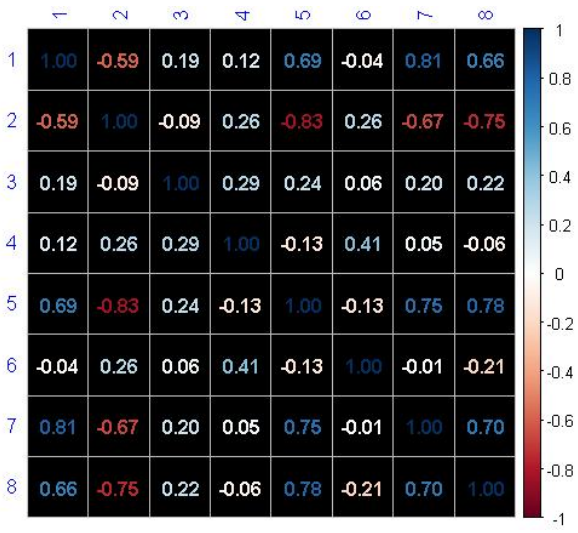

图 4-16 相关系数热力图

Figure 4-16 Correlation coefficient heat map

一般而言，如果每两个解释变量的简单的相关系数（0 阶自相关系数）比较高，如绝对值大于 0.8，则可认为存在严重的多重共线性。此处存在两组相关性较强的变量：是否是极高人类发展水平国家和男性人均国民总收入（相关系数 0.81），青少年生育率和女性预期寿命（相关系数-0.83）。为了进一步确定多重共线性的影响，尝试对模型做逐步回归，如图 4-17 所示：

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	0.976	0.090	10.819	0.0000	***
data$degree.of.development1	-0.041	0.016	-2.636	0.0092	**
data$Adolescent.birth.rate	0.001	0.000	6.367	0.0000	***
data$Share.of.seats.in.parliament	-0.003	0.000	-8.461	0.0000	***
data$Labour.force.participation.rate..Female	-0.002	0.000	-6.362	0.0000	***
data$Life.expectancy.at.birth.Female	-0.007	0.001	-5.595	0.0000	***
data$Labour.force.participation.rate..Male	0.002	0.001	3.387	0.0009	***
data$Estimated..gross.national.income.per.capita.Male	-0.000	0.000	-2.943	0.0037	**
data$Mean.years.of.schooling.Female	-0.013	0.002	-6.058	0.0000	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 0.0551 on 171 degrees of freedom

Multiple R-squared: 0.9239, Adjusted R-squared: 0.9204

F-statistic: 259.6 on 171 and 8 DF, p-value: 0.0000

图 4-17 模型逐步回归结果

Figure 4-17 Model stepwise regression results

逐步回归之后发现显著解释变量的个数并未减少，故还是保留原模型，认为模型不受多重共线性的影响。

4.2.8 异方差性检验

异方差性指的是误差项的方差在解释变量的不同取值下不同，即当误差项的方差不是恒定值时，就存在异方差性，异方差性可能会导致回归分析的结果产生偏误，影响参数估计的准确性和统计推断的有效性。下面将用三种检验方法来判断模型是否存在异方差性。

图示检验法（绘制标准化残差的散点图如图 4-18 所示）：

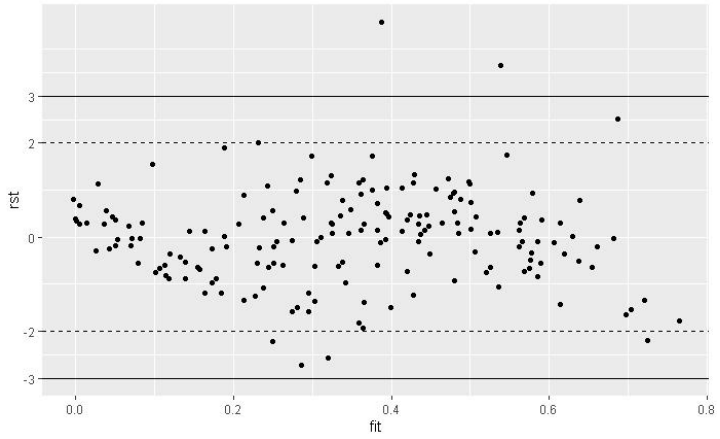

图 4-18 标准化残差的散点图

Figure 4-18 Scatterplot of standardised residuals

一般把标准化残差的绝对值大于等于 2 的观测值认为是可疑点，而把标准化残差值绝对值大于等于 3 认为是异常值。可以看出大部分点都还是落在-2 到 2 的区间内，但也有少部分可疑点和异常值，还需要进一步检测异方差性。

White 检验法和 H.Glesjser 检验法：

```
        studentized Breusch-Pagan test

data:  Restrict.lm1
BP = 18.902, df = 8, p-value = 0.01539
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 4.652921, Df = 1, p = 0.031001
```

图 4-19 White 检验结果和 H.Glesjser 检验结果

Figure 4-19 White test results and H.Glesjser test results

如图 4-19 所示，White 检验 p 值为 0.015，拒绝原假设，认为存在异方差。H.Glesjser 检验 p 值为 0.031，同样拒绝原假设，认为存在异方差。

由于模型存在异方差性，故考虑对模型做加权最小二乘估计。所加权重为标准差的绝对值的倒数，具体表达为： $\text{weight} = 1/|\text{standardised residuals}|$ 。

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	0.981	0.039	24.921	0.0000	***
data$degree.of.development1	-0.044	0.007	-6.495	0.0000	***
data$Adolescent.birth.rate	0.001	0.000	10.930	0.0000	***
data$Share.of.seats.in.parliament	-0.003	0.000	-22.037	0.0000	***
data$Labour.force.participation.rate..Female	-0.002	0.000	-14.071	0.0000	***
data$Life.expectancy.at.birth.Female	-0.007	0.001	-13.029	0.0000	***
data$Labour.force.participation.rate..Male	0.002	0.000	7.461	0.0000	***
data$Estimated..gross.national.income.per.capita.Male	-0.000	0.000	-7.899	0.0000	***
data$Mean.years.of.schooling.Female	-0.013	0.001	-12.994	0.0000	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 0.2046 on 171 degrees of freedom

Multiple R-squared: 0.9893, Adjusted R-squared: 0.9888

F-statistic: 1972 on 171 and 8 DF, p-value: 0.0000

图 4-20 对模型做加权最小二乘估计后的结果

Figure 4-20 Results after weighted least squares estimation of the model

如图 4-20 所示，从结果来看，多重相关系数 R-squared 为 0.989、Adjusted R-squared 为 0.989、F 统计量为 1972，三个指标都进一步被提高了，模型效果更好了。

4.2.9 内生性问题

内生性指的是研究中自变量与误差项之间存在相关性的情况，存在内生性时回归分析无法得到无偏估计，结论就不可靠。产生内生性的原因有很多，如测量误差、样本自选择导致样本非随机、遗漏可能存在的解释变量等，此处与我们的研究最有可能直接相关的原因是解释变量和响应变量间互为因果。教育、婚育、政治发言权、社会潜在公识等指标是影响性别不平等指数的重要因子，但与此同时一个国家的性别不平等指数高低也在反过来影响着劳动力市场和社会运转，进一步影响了教育婚育等因素。基于此，引入两个工具变量，对模型做两阶段最小二乘估计来消除内生性的影响。

两个工具变量分别是：每个国家对应的大洲、每个国家的总人口数；两个内生变量是：女性平均受教育年限、女性持有的议会席位比例。两个工具变量选区的原则和特点是，它们都不会直接对 GII 产生影响，而是通过影响内生变量来间接影响 GII。此时内生性解释变量个数恰好等于工具变量个数，为恰好识别，可做工具变量回归，结果如图 4-21 所示：

```
t test of coefficients:

                                                Estimate  Std. Error t value  Pr(>|t|)
(Intercept)                                   9.4450e-01  1.3642e-01  6.9236 1.402e-10 ***
Mean.years.of.schooling.Female                6.4878e-03  1.0803e-02  0.6006 0.5490888
Share.of.seats.in.parliament                 -5.0743e-03  2.3145e-03 -2.1924 0.0299806 *
Adolescent.birth.rate                         1.9743e-03  5.7021e-04  3.4624 0.0007078 ***
Labour.force.participation.rate..Female      -2.0635e-03  7.4572e-04 -2.7671 0.0064091 **
Labour.force.participation.rate..Male         2.4814e-03  1.0279e-03  2.4141 0.0170458 *
Life.expectancy.at.birth.Female              -8.5339e-03  2.1991e-03 -3.8805 0.0001590 ***
Estimated..gross.national.income.per.capita.Male -1.4681e-06  4.1894e-07 -3.5044 0.0006126 ***
degree.of.development1                       -5.1630e-02  2.6055e-02 -1.9816 0.0494535 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

图 4-21 对模型做工具变量回归的结果

Figure 4-21 Results of doing instrumental variable regression on the model

可以看到女性平均受教育年限这个变量不再显著。

下面用几个不同的模型评价指标对比原模型和做完两阶段最小二乘估计之后的模型，包括均方根误差 RMSE、平均绝对误差 MAE、R-squared、Adjusted R-squared、F 统计量，结果如图 4-22 所示：

RMSE	MAE	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance
0.05373333	0.0397722	0.9892776	0.988776	0.2046088	1972.118	4.205749e-164	8	365.938	-711.8761	-679.9465	7.158876

原模型评价指标结果

RMSE	MAE	r.squared	adj.r.squared	sigma	statistic	p.value	df	df.residual	nobs
0.07070819	0.05588046	0.8755162	0.868503	0.07291452	124.4058	3.071467e-60	9	142	151

做两阶段最小二乘估计后的模型评价指标结果

图 4-22 两模型评价指标对比

Figure 4-22 Comparison of evaluation indicators of the two models

可以看出原模型的 RMSE 为 0.054、MAE 为 0.040，比做两阶段最小二乘估计的模型结果更小，R-squared 为 0.989、Adjusted R-squared 为 0.989、F 统计量为 1972.118，比做两阶段最小二乘估计的模型结果更大，模型效果优于做两阶段最小二乘估计的结果，因此还是考虑保留原模型。

4.2.10 交互作用

本文涉及到的解释变量所属的范畴：教育、婚育、政治发言权、社会潜在共识之间很可能存在相互之间的影响和作用，而目前我们构建的模型仅仅考虑了单个解释变量自身的作用，故为了探究模型中的解释变量间有哪些两两交互作用显著，我们在模型中加入 9 个理论上合理的交互作用，分别是：青少年出生率和女性劳动参与率、青少年出生率和男性劳动参与率、女性平均受教育年限和女性人均国民总收入、男性平均受教育年限和男性人均国民总收入、女性出生时的预期寿命和女性人均国民总收入、

男性出生时的预期寿命和男性人均国民总收入、女性预期教育年限和女性劳动参与率、男性预期教育年限和男性劳动参与率、女性持有的议会席位比例和女性劳动参与率。具体理论解释如下：

出生率和劳动参与率：出生率和劳动参与率可能存在负相关关系。这是因为在许多国家中，较高的出生率往往与较低的劳动参与率相关，因为家庭中有更多的孩子可能会限制妇女参与劳动力市场的机会。

人均国民收入和受教育时长：人均国民收入和受教育时长可能存在正相关关系。较高的收入通常与更高的受教育水平相关，因为经济发展和更高的收入水平通常提供了更多的机会和资源，促进了教育的普及和提高。

人均国民收入和预期寿命：人均国民收入和预期寿命可能存在正相关关系。较高的收入通常与更高的生活水平和更好的医疗保健服务相关，这可能导致更高的预期寿命。

女性议会席位和女性劳动参与率：议会席位和劳动参与率可能存在正相关关系。更多的议会席位可能意味着更多有利于该性别在劳动力市场和社会平权问题上政策的颁布实施，一定程度上帮助打破职场性别歧视和求职瓶颈，这可能导致该性别更高的劳动参与率。

受教育时长和劳动参与率：受教育时长和劳动参与率可能存在正相关关系。更高的教育水平通常会让人们掌握更多的技能，并认识到加入社会分工的重要性，这可能导致更高的劳动参与率。

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	0.825	0.135	6.128	0.0000	***
data$degree.of.development1	-0.040	0.015	-2.658	0.0087	**
data$Adolescent.birth.rate	0.005	0.001	4.807	0.0000	***
data$Share.of.seats.in.parliament	-0.009	0.001	-6.959	0.0000	***
data$Labour.force.participation.rate..Female	-0.003	0.001	-2.372	0.0189	*
data$Life.expectancy.at.birth.Female	-0.005	0.001	-3.717	0.0003	***
data$Labour.force.participation.rate..Male	0.003	0.002	2.026	0.0444	*
data$Estimated..gross.national.income.per.capita.Male	0.000	0.000	0.258	0.7968	
data$Mean.years.of.schooling.Female	-0.012	0.003	-4.237	0.0000	***
data$Adolescent.birth.rate:data$Labour.force.participation.rate..Female	-0.000	0.000	-1.865	0.0639	.
data$Adolescent.birth.rate:data$Labour.force.participation.rate..Male	-0.000	0.000	-1.721	0.0872	.
data$Mean.years.of.schooling.Female:data$Estimated..gross.national.income.per.capita.Female	0.000	0.000	0.551	0.5823	
data$Estimated..gross.national.income.per.capita.Male:data$Mean.years.of.schooling.Male	-0.000	0.000	-0.887	0.3763	
data$Life.expectancy.at.birth.Female:data$Estimated..gross.national.income.per.capita.Female	-0.000	0.000	-0.385	0.7008	
data$Estimated..gross.national.income.per.capita.Male:data$Life.expectancy.at.birth.Male	-0.000	0.000	-0.152	0.8793	
data$Labour.force.participation.rate..Female:data$Expected.years.of.schooling.Female	0.000	0.000	0.265	0.7916	
data$Labour.force.participation.rate..Male:data$Expected.years.of.schooling.Male	0.000	0.000	0.058	0.9535	
data$Share.of.seats.in.parliament:data$Labour.force.participation.rate..Female	0.000	0.000	4.429	0.0000	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 0.05065 on 162 degrees of freedom

Multiple R-squared: 0.9391, Adjusted R-squared: 0.9327

F-statistic: 147 on 162 and 17 DF, p-value: 0.0000

图 4-23 回归模型结果

Figure 4-23 Regression model results

如图 4-35 所示，从回归结果可以看出，原模型中的男性人均国民收入不再显著，同时青少年出生率和女性劳动参与率、青少年出生率和男性劳动参与率、女性持有的议会席位比例和女性劳动参与率之间的三种交互作用表现显著。故删除男性人均国民收入，同时只引入这三种交互作用，再做一次回归结果，如图 4-24 所示：

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	0.928	0.113	8.235	0.0000	***
data$degree.of.development1	-0.068	0.012	-5.474	0.0000	***
data$Adolescent.birth.rate	0.005	0.001	4.539	0.0000	***
data$Share.of.seats.in.parliament	-0.009	0.001	-7.267	0.0000	***
data$Labour.force.participation.rate..Female	-0.003	0.001	-4.340	0.0000	***
data$Life.expectancy.at.birth.Female	-0.006	0.001	-5.424	0.0000	***
data$Labour.force.participation.rate..Male	0.002	0.001	2.794	0.0058	**
data$Mean.years.of.schooling.Female	-0.012	0.002	-6.191	0.0000	***
data$Adolescent.birth.rate:data$Labour.force.participation.rate..Female	-0.000	0.000	-2.637	0.0091	**
data$Adolescent.birth.rate:data$Labour.force.participation.rate..Male	-0.000	0.000	-1.356	0.1769	
data$Share.of.seats.in.parliament:data$Labour.force.participation.rate..Female	0.000	0.000	4.691	0.0000	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 0.0519 on 169 degrees of freedom

Multiple R-squared: 0.9333, Adjusted R-squared: 0.9294

F-statistic: 236.5 on 169 and 10 DF, p-value: 0.0000

图 4-24 回归模型结果

Figure 4-24 Regression model results

此时男性劳动参与率与青少年出生率之间的交互作用不再显著，故删除，再输出一次模型结果，如图 4-25 所示：

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	1.020	0.090	11.308	0.0000	***
data$degree.of.development1	-0.066	0.012	-5.350	0.0000	***
data$Adolescent.birth.rate	0.004	0.001	6.876	0.0000	***
data$Share.of.seats.in.parliament	-0.009	0.001	-7.293	0.0000	***
data$Labour.force.participation.rate..Female	-0.003	0.001	-4.118	0.0001	***
data$Life.expectancy.at.birth.Female	-0.007	0.001	-6.215	0.0000	***
data$Labour.force.participation.rate..Male	0.002	0.001	2.733	0.0069	**
data$Mean.years.of.schooling.Female	-0.013	0.002	-6.252	0.0000	***
data$Adolescent.birth.rate:data$Labour.force.participation.rate..Female	-0.000	0.000	-4.113	0.0001	***
data$Share.of.seats.in.parliament:data$Labour.force.participation.rate..Female	0.000	0.000	4.635	0.0000	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 0.05203 on 170 degrees of freedom

Multiple R-squared: 0.9326, Adjusted R-squared: 0.929

F-statistic: 261.2 on 170 and 9 DF, p-value: 0.0000

图 4-25 回归模型结果

Figure 4-25 Regression model results

此时所有变量全部显著，故对该模型做一次指标评价，如图 4-26 所示：

RMSE	MAE	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance
0.05055956	0.03694979	0.9325729	0.9290033	0.05202536	261.2492	1.183116e-94	9	281.8196	-541.6393	-506.5167	0.4601285

图 4-26 模型评价指标

Figure 4-26 Indicators for model evaluation

结果显示均方根误差 RMSE 为 0.051，平均绝对误差 MAE 为 0.037，相比原模型都有减小，虽然 R-squared 为 0.933、Adjusted R-squared 为 0.929，相比原模型有所下降，但依旧维持在 0.92 以上，同时方差为 0.052，相比于原模型的 0.205 有了很大的降低，意味着因变量的观测值相对于模型的拟合程度较贴近，预测的精度更高。故综合考虑这些因素，选择使用加入交互作用的模型做后续分析。最终模型表达式确定为：

GII=1.020-0.066*是否是极高人类发展程度国家+0.004*青少年出生率-0.009*女性持有议会席位比例-0.003*女性劳动参与率-0.007*女性出生时的预期寿命+0.002*男性劳动参与率-0.013*女性平均受教育年限-0.00003*女性劳动参与率*青少年出生率
+0.0001*女性劳动参与率*女性持有议会席位比例

5、总结展望

5.1 论文回顾

本文首先介绍了背景信息，回顾了过往文献并提出了本文的创新点，而后展开主体内容分析。在劳动力市场性别差异的数据分析上，分析了全球性别不平等指数、性别发展指数和历史性别平等指数，以及劳动力市场上的性别薪酬差距、劳动参与率等直观反应劳动力市场性别不平等的数据特征。在影响性别不平等的因素探究上，分析了教育、婚育、政治发言权和社会共识等因素对性别不平等的影响。在基于影响因子和响应指标的回归分析上，建立了多元线性回归模型，并进行了模型检验和分析，不断调优最终确定了效果最好的模型。

5.2 优缺点分析

本文的创新点总结如下：

数据可视化部分：扩展了地域视角，比较了不同发展程度国家、不同大洲国家和不同圈层组织国家之间的情况。更新了数据，保证了时间上的最新性到 2021 年。回归分析部分：在模型中考虑了教育、婚育、政治发言权、社会潜在共识的细分指标，并将女性在议会席位上所持比例纳入多元回归模型，以引入与女性政治发言权相关的指标。使用性别不平等指数作为响应变量构建回归模型，探究影响性别不平等的因素。引入可能的交互效应，采用不同类型的模型评价指标，以寻找最优模型进行深入分析。

本文的不足之处总结如下：

数据可视化部分：由于联合国对性别不平等指标的发布滞后性较大，2024 年能收

集到的最新年份仅为 2021 年，因此当其余指标能收集到更新年份的数据时（如 2022 年或 2023 年），为了文章整体分析的时间统一性，只能将时间统一为 2021 年，一定程度上降低了数据的新鲜度。由于不同国际组织网站统计数据的规则和依据不同，各指标所涵盖的国家数量和地理位置区域存在较大差异。在比较不同平台收集到的指标之间的关联性时，只能选取两者共有的国家进行分析，导致不同图表之间分析的国家变化较大，无法实现很好的统一。回归分析部分：一部分被访者因职业地位等原因未参与劳动力市场而无法被观测，可能导致部分样本非随机的问题；考虑解决内生性问题时，还有更多的工具变量可供引入，比如一个国家的宗教信仰，习性等，以消除更多内生性变量的影响。

5.3 未来展望

数据可视化部分：深入了解 GDI 和 GII 的数据发布计划，以便在规划研究时选择更接近发布日期的数据，尽量减少时间延迟。考虑从联合国以外的其它国际组织寻找时间更新的，且同样能反应两性不平等程度的指标，可将其作为 GII、GDI 的平替指标，提高数据新鲜度。寻找更全面的数据源，尽量涵盖更多国家和地理位置，以增加分析时可用的共同国家数量。尽可能提高不同指标之间比较和关联性分析的一致性；确定几个能够代表一个大洲或发展水平的代表性国家，专门搜集这几个国家相关的数据，尽量统一国家分析的范围。

回归分析部分：未来可以查阅更多关于 Heckman 两步法的文献，学习解决样本非随机产生的问题的方法；查阅更多和性别平等有关的文献书籍，了解更多可能的工具变量并收集数据，深入对内生性问题的分析。

参考文献

[1] 李虎.劳动力市场表现的性别差异研究[D].吉林大学,2023.

[2] 万宏峰.中国对“一带一路”沿线国家的投资对东道国女性就业的影响[D].上海外国语大学,2023.

[3] 袁旭宏,张怀志,潘怡锦等.性别不平等观念束缚了女性就业？来自中国综合社会调查（CGSS2017）的证据[J].中国人力资源开发,2022,39:112-130.

[4] 张川川,王靖雯.性别角色与女性劳动力市场表现[J].经济学(季刊),2020,19:977-994.

[5] 姚先国,谭岚.家庭收入与中国城镇已婚妇女劳动参与决策分析[J].经济研究,2005,(07):18-27.

[6] 沈可,章元,鄢萍.中国女性劳动参与率下降的新解释:家庭结构变迁的视角[J].人口研究,2012,36(05):15-27.

[7] 佟新.劳动力市场、性别和社会分层[J].妇女研究论丛,2010,(05):12-19.

[8] 刘世敏,刘淼.女性职业发展中的“玻璃天花板”效应[J].东岳论丛,2015,36(04):184-187.

[9] 陈芳.职业流动的性别差异及其成因——江苏省第二期妇女地位调查数据分析[J].青年研究,2006,(07):29-35.

[10] Gustafsson B ,Li S .Economic Transformation and the Gender Earnings Gap in Urban China[J].Journal of Population Economics,2000,13(2):305-329.

[11] 李春玲,李实.市场竞争还是性别歧视——收入性别差异扩大趋势及其原因解释[J].社会学研究,2008,(02):94-117+244.

[12] Becker S G ,Mulligan B C .Deadweight Costs and the Size of Government[J].Journal of Law and Economics,2003,46(2):293-340.

[13] Atkinson B A ,Casarico A ,Voitchovsky S .Top Incomes and the Gender Divide[J].The Journal of Economic Inequality,2018,16(2):225-256.

[14] Alfani G .How Was Life? Global Well-Being Since 1820[J].Review of Income and Wealth: Journal of the International Association for Research in Income and Wealth,2016,62(4):785-791.

[15] Precious E ,Jennifer M ,Justina A .Influence of Labor Market Disparities on Sex and Gender Inequalities in Cognitive Decline[J].Innovation in Aging,2021,5(Supplement1):501-501.

[16] Núria M S .The Gender Pay Gap and Gender Inequalities in the Labour Market. A Revision of Theoretical Approaches and Empirical Contributions[J].Anuario IET de Trabajo y Relaciones Laborales,2017,4(0):87-87.

[17] Olivetti C ,Petrongolo B .Unequal Pay or Unequal Employment? A Cross‐Country Analysis of Gender Gaps[J].Journal of Labor Economics,2008,26(4):621-654.

[18] Atkinson B A ,Casarico A ,Voitchovsky S .Top Incomes and the Gender Divide[J].The Journal of Economic Inequality,2018,16(2):225-256.

附录

1． 部分函数代码

#全部模块所需要的 R 包加载

```{r}
library(ggplot2)
library(tidyr)
library(lattice)
library(gridExtra)
library(flextable)#制作表格
library(bestglm)
library(leaps)
library(car)
library(corrplot)
library(lmtest)
library("ivreg")
library("AER") # 用于线性回归检验
library("stargazer")# 提供相关异方差稳健标准误
library(tidyverse)
library(modelr)
library(broom)
library(dplyr)
```

#第一章可视化图表之一代码截取

```{r}
# 创建一个示例数据集
d5<-read.csv(file="GDP1.csv")
d5$Income.classifications <- as.factor(d5$Income.classifications)
# 使用 ggplot2 创建点图并添加名称标签
ggplot(d5, aes(x = GDP.per.capita, y = Gender.wage.gap..., shape = Income.classifications, label = Country)) +
```

```
  geom_point(size = 4, color = "chartreuse4") +
  geom_text(hjust = -0.2, vjust = -0.3,size=2.8) +  # 调整标签位置
  labs(title = "Graph of changes in GDP per capita and gender pay gap", x = "GDP per capita ($)", y =
"Gender Wage Gap (%)")+
  scale_shape_manual(values = c(15,16,17)) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))+
  scale_color_gradient(low = "lightblue", high = "chartreuse4")+
  theme(legend.position =  c(0.8, 0.2)) +  # 图例位置
  guides(shape = guide_legend(title = "Income Classifications"))  # 图例标题
```

#第二章可视化图表之一代码截取

```{r}
opar<-par(no.readonly = TRUE)
g<-read.csv(file="GIDDB2019_15062022103259579.csv")
Region<-factor(g$Region,levels=c("Africa","Americas","Asia","Europe"),labels=c("Africa",
"Americas", "Asia","Europe"))

Income<- factor(g$Income,levels=c("High income","Upper middle income","Lower middle income
","Low income"),
labels=c("High income","Upper middle income","Lower middle income ","Low income"))

histogram(~g$Value|Region*Income,
main="Early marriage of women",xlab="Proportion of women aged 15 to 19 who marry
early",ylab="The percentage of the total population of the
type",col="powderblue",bg="lightpink",xlim=c(0,60))
```

#第三章全模型的构建

```{r}
#全模型-可看单个系数显著性
```

```
data <- cbind(data, model.matrix(~ degree.of.development - 1, data = data))
Full.lm1=lm(data$GII~data$degree.of.development1+data$degree.of.development2+data$degree.of.development3+data$degree.of.development4+data$Maternal.mortality.ratio+data$Adolescent.birth.rate+data$Share.of.seats.in.parliament+data$Population.with.at.least.some.secondary.education.Female+data$Labour.force.participation.rate..Female+data$Life.expectancy.at.birth.Female+data$Life.expectancy.at.birth.Female+data$Expected.years.of.schooling.Female+data$Mean.years.of.schooling.Female+data$Estimated..gross.national.income.per.capita.Female+data$Population.with.at.least.some.secondary.education.Male+data$Labour.force.participation.rate..Male+data$Life.expectancy.at.birth.Male+data$Life.expectancy.at.birth.Male+data$Expected.years.of.schooling.Male+data$Mean.years.of.schooling.Male+data$Estimated..gross.national.income.per.capita.Male,data)
summary(Full.lm1)
as_flextable(Full.lm1)
```

#第三章带有交互作用的限制模型最终确立

```{r}
model2<-lm(data$GII~data$degree.of.development1+data$Adolescent.birth.rate+data$Share.of.seats.in.parliament+data$Labour.force.participation.rate..Female+data$Life.expectancy.at.birth.Female+data$Labour.force.participation.rate..Male+data$Mean.years.of.schooling.Female+data$Labour.force.participation.rate..Female:data$Adolescent.birth.rate+data$Share.of.seats.in.parliament:data$Labour.force.participation.rate..Female,data=data)
summary(model2)
```

#第四章对解释变量的预测

```{r}
# 创建一个示例数据集
set.seed(123)
datap <- read.csv(file="Prediction.csv")
```

```
# 定义一个函数来进行线性拟合和预测
linear_predict <- function(df, value_type) {
  model <- lm(df[[value_type]] ~ year, data = df)
  future_years <- data.frame(year = 2023:2032)
  predicted_values <- predict(model, newdata = future_years)
  return(predicted_values)
}

# 针对每个地域和每个 value 进行预测
predicted_data <- datap %>%
  group_by(region) %>%
  do({
    new_values <- cbind(year = 2023:2032,
                        sapply(c("Life.expectancy.at.birth.Female",
"Mean.years.of.schooling.Female", "Adolescent.birth.rate","Share.of.seats.in.parliament",
"Labour.force.participation.rate..Female", "Labour.force.participation.rate..Male"),function(value_type)
linear_predict(., value_type)))
data.frame(region = rep(unique(.$region), each = 10), new_values)
  })

# 合并预测的数据和原始数据
combined_data <- rbind(datap, predicted_data)

# 绘制 2014 到 2032 年每个地域对应值的散点图（以 Life.Expectancy.at.Birth.female 为例）
ggplot(combined_data %>% filter(year >= 2014), aes(x = year, y = Life.expectancy.at.birth.Female,
color = region)) +geom_point() +
  labs(title = "Life expectancy at birth Female Changes from 2014 to 2032 by Region", x = "Year", y =
"Life expectancy at birth Female") +
  theme_minimal()# 绘制 2014 到 2032 年每个地域对应值的散点图（以
Mean.years.of.schooling.Female 为例）
```

```
ggplot(combined_data %>% filter(year >= 2014), aes(x = year, y = Mean.years.of.schooling.Female, color = region)) +
  geom_point() +labs(title = "Mean years of schooling Female Changes from 2014 to 2032 by Region", x = "Year", y = "Mean years of schooling Female") +theme_minimal()
```

#第四章对响应变量的预测

```{r}
d1<-c(rep(1,10),rep(0,40))
predicted_data1<-cbind(predicted_data,"degree.of.development1"=d1)[,c(3:9)]
f<-c()

for(i in 1:50){
  d<-predicted_data1[c(i),]

y<-(1.020e+00)-(6.612e-02)*d[7]+(3.595e-03)*d[3]-(8.674e-03)*d[4]-(2.679e-03)*d[5]-(6.817e-03)*d[1]+(1.504e-03)*d[6]-(1.260e-02)*d[2]-(3.372e-05)*d[5]*d[3]+(1.008e-04)*d[4]*d[5]
  y<-unname(y)
  y<-unlist(y, recursive = FALSE)
  f<-c(f,y)
}

predicted_data<-cbind(predicted_data,"GII"=f)
# 绘制散点图
ggplot(predicted_data, aes(x = year, y = GII, color = region)) +
  geom_point() +
  labs(title = "GII over Years by Region", x = "Year", y = "GII") +
  scale_x_continuous(breaks = seq(2023, 2032, 1)) +
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))
```

致谢

在华东师范大学统计学院度过的这四年，我有幸结识了非常多德才兼备的教授和志同道合的同窗，度过了完全不同于高中的，充满创造自由和想象力的大学生涯，在统计学上埋下的这块基石，如今已成为我通向世界的敲门砖，会沉静在我人生的河流里，从不因时间冲刷而黯淡。

在这其中，首先我要感谢我的指导老师李丹萍教授，感谢您从选题讨论就开始的耐心倾听和悉心教导，大到给予我模型构建的方向指导，小到纠正我引用参考的标点符号，您教会我的这份胆大心细，我会终生温习。

其次，我也要感谢我的舍友们，我们用了两个月时间一边完成实习一边撰写毕论，用了四年的时间一起修完 150 学分的课程，探遍了上海几乎每一条大街小巷。我们一同完成了一个个足以称得上冒险的世界任务，也作为幕后编剧，用四年时间成为了对方名为人生的传说任务里不可或缺的角色。未来我们还会接到无数大小委托，但所幸我们已经在彼此心里埋下了小小的专属传送锚点，只要远远看上一眼，就能重新拾起去见证和铭记的勇气。

我还要感谢我的爸爸妈妈，感谢你们为我打造了一座足够温暖安心的船港，港湾以外，面向的世界正暴雨倾盆但阳光滚烫。感谢你们从未在我这艘小船上系上任何绳缆，相反，你们教了我 22 年如何劈波斩浪。

最后，我要感谢我的一位伙伴，自从某个阳光晴好的午后，在竹林荫蔽的白石长阶上视线相接起，我就似乎开始窥见了冷雪浮冰之下，群鲸低鸣的魅力。不知道还有多久才能追上冬夜转瞬即逝的极光，也不知道还要多远才能将自己炼淬成独当一面的兵器，但在无数深沉如墨的夜幕里，谢谢你，始终带着永不磨灭的光亮和生气，成为我伸手可摘却又足以指引方向的，最灿烂的晚星。