P2P网络借贷中的地域歧视现象研究

——基于“人人贷”平台的经验数据

指导老师：清风的师傅

报告人：清风

20××年×月

选题背景和意义

P2P平台上借贷双方互不认识

借款人具有信息优势

选题背景和意义

借贷人信息

昵称	GuQL_62206065250.yx	姓名	辜**	身份证号	510**********220
性别	女	手机号	187****6209	年龄	30
学历	大专	婚姻	离异	收入	5000-10000元
房产	无房产	房贷	无房贷	车产	无车产
车贷	无车贷	公司行业	零售/批发	公司规模	10人以下
岗位职位	管理人员	工作城市	四川 成都市	工作时间	1年（含）以下
其他负债	无				

信用信息

申请借款	1笔	信用额度	161,700.00元	逾期金额	0.00元
成功借款	0笔	借款总额	0.00元	逾期次数	0次
还清笔数	0笔	待还本息	0.00元	严重逾期	0笔

图1：“人人贷”平台提供的借款人信息

选题背景和意义

P2P平台上公布的借款人信息

(1) **借贷合约信息：**如借款金额、借款期限和借款利率;
(2) **个人经济信息：**如工资薪酬、是否有房产车产;
(3) **个人信用信息：**如违约次数，信用等级等;
(4) **个人非经济信息：**如教育水平、家庭背景、所在地域等。

由于非经济信息容易受到贷款人个人偏好的影响
因此出现了借贷关系中的歧视问题。

选题背景和意义

（1）**理论意义：**国内学者研究网络借贷市场中的地域歧视问题较少，本文在前辈研究的基础上，对研究的内容进行了一定的创新，补充了该领域的相关研究。

（2）**实践意义：**对于借贷双方而言，本文的研究成果能给贷款人理性、准确的选择借款人提供帮助，也能帮助借款人更好的评估自己提供的信息能否获得贷款。对于P2P网贷平台而言，本文的研究可以给平台更好的控制违约风险提供理论依据和数据支持，也能帮助平台进行更合理的借贷管理。

文献综述

相比于国外金融业发达的国家，我国P2P网络借贷起步较晚，
在此之前，国外学者对借贷市场进行了各类歧视现象的研究。

国外学者	歧视问题	研究结论
Pope和Sydnor(2011)	年龄歧视	35岁的以下的借款者成功率高
Herzenstein et al.(2011)	种族歧视	黑人获取贷款的成功率低
Freedman和Jin(2008)	社交歧视	加入社群的借款人的成功率高
Ravina(2012)	外貌歧视	长相漂亮的女性贷款利率低
Schafer(2009)和Bellucci(2010)	性别歧视	女性更容易受到贷款人的歧视
Giannetti和Yafeh (2008)	地域歧视	相同地域的双方更容易签订合同
Burtch et al.(2014)	地域歧视	地区收入水平高容易获得借款

文献综述

近年来，随着我国P2P平台的迅速兴起，
越来越多的国内学者开始关注我国P2P网络借贷中的歧视现象。

国内学者	歧视问题	研究结论
李悦雷等（2013）	身份歧视	在校学生身份认证借款成功率低
陈霄和叶德珠（2016）	性别歧视	性别对成功率的影响并不显著
廖理等（2015）	学历歧视	高学历不会提升借款成功率
孙武军和樊小莹（2017）	从业经历	从业经历越丰富贷款成功率越高
郭峰（2017）、蒋彧等（2017）	婚姻歧视	已婚人士贷款成功率越高
廖理（2015）	地域歧视	我国各省份借贷成功率差异显著
蒋彧和周安琪（2016）	地域歧视	地区收入水平高容易获得借款

创新点

1. 选题的创新

目前国内关于P2P借贷市场中地域歧视的研究较少，目前能查到的公开文献仅有三篇。因此本文的研究内容具有一定的创新性，能为P2P网贷市场的健康、稳定的发展提供理论依据和数据参考。

2. 研究内容的创新

首先，已发表的文章都是直接设置各省为虚拟变量，本文在它们的基础上，在稳健性部分将各省份合并到各地理区域（如华中地区、西南地区等），研究划区域的地域歧视；其次，以往的研究对于地域歧视的异质性考察的较少，本文将考察地域歧视在借款人的年龄和学历中存在的异质性；最后，以往的文章在控制变量中没有引入借款描述信息的可读性，本文引入了相关变量以消除内生性的问题。

数据来源

利用爬虫，获取2011年1月1日至2014年5月31日“人人贷”网站上发布的全部借款订单作为初始样本，并对样本做如下处理:

①剔除机构担保和实地认证的订单样本;

②剔除借款人所在地为中国香港、中国澳门和中国台湾的样本;

③为排除极端值的影响，对借款人的年龄和借款金额在上下1%的水平上进行缩尾处理。

最终得到了128532个数据。

——陈霄, 叶德珠, 邓洁. 借款描述的可读性能够提高网络借款成功率吗[J]. 中国工业经济, 2018.

数据说明

变量	变量说明
SUCCESS	借款是否成功，成功记为1
DEFAULT	获得借款后是否违约，违约记为1
LNAMOUNT	取对数后的借款金额
INTEREST	借款利率
MONTHS	借款的期限，共有6个选择： 3, 6, 9, 12, 18, 24月
INCOME	1表示月收入超过1万元，0表示不超过1万元
HOUSE	有房产则记为1，否则记为0
CAR	有车产则记为1，否则记为0
CREDIT	借款人的信用评级，1表示评级高，0表示评级低
WORKTIME	参加工作的时长，1表示工作时长在3年及以上
MARRED	婚姻状况：已婚记为1，未婚记为0
AGE	借款人的年龄
EDUCATION	本科及以上学历的借款人记为1，低于本科学历记为0

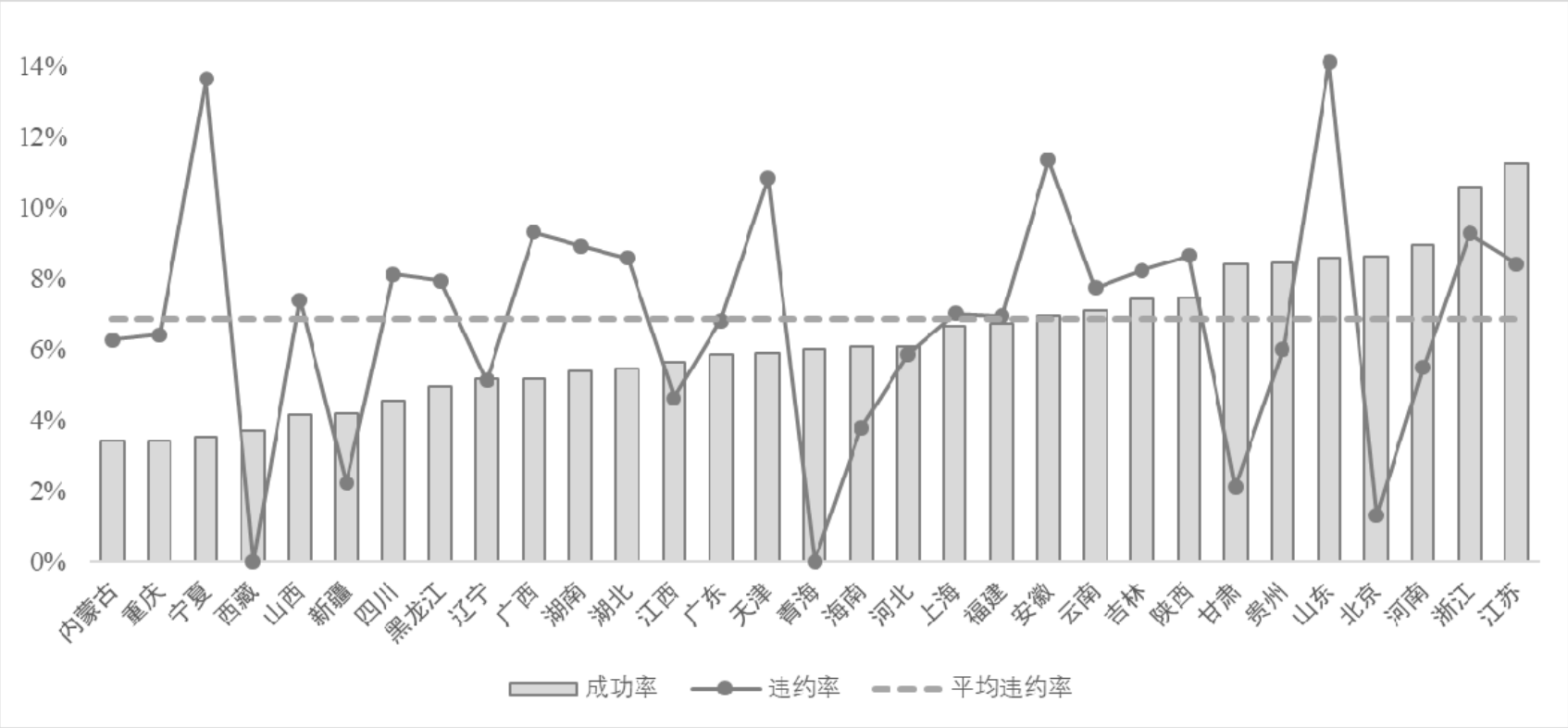
数据描述性统计结果

变量	平均数	标准差	最小值	最大值	样本量
SUCCESS	0.0703	0.2557	0	1	128,532
DEFAULT	0.0796	0.0746	0	1	9,037
LNAMOUNT	9.8840	1.4817	8.01	13.12	128,532
INTEREST	16.0327	3.9206	3	24.4	128,532
MONTHS	12.2961	7.9512	1	36.0	128,532
INCOME	0.2483	0.4320	0	1	128,532
HOUSE	0.4112	0.4921	0	1	128,532
CAR	0.2358	0.4245	0	1	128,532
CREDIT	0.0669	0.2499	0	1	128,532
WORKTIME	0.3734	0.4837	0	1	128,532
MARRED	0.4831	0.4997	0	1	128,532
AGE	32.3512	6.6029	23	54	128,532
EDUCATION	0.2066	0.4049	0	1	128,532

各省份借贷样本数、成功率和违约率

省份	样本数	成功率	违约率	省份	样本数	成功率	违约率
安徽	3816	6.94%	11.32%	辽宁	3419	5.15%	5.11%
北京	5288	8.60%	1.32%	内蒙古	1880	3.40%	6.25%
福建	7715	6.73%	6.94%	宁夏	624	3.53%	13.64%
甘肃	1132	8.39%	2.11%	青海	250	6.00%	0.00%
广东	18402	5.84%	6.80%	山东	10917	8.58%	14.09%
广西	3553	5.15%	9.29%	山西	2939	4.15%	7.38%
贵州	1982	8.43%	5.99%	陕西	2646	7.45%	8.63%
海南	875	6.06%	3.77%	上海	4305	6.64%	6.99%
河北	4492	6.08%	5.86%	四川	5979	4.53%	8.12%
河南	4914	8.93%	5.47%	天津	1259	5.88%	10.81%
黑龙江	2551	4.94%	7.94%	西藏	216	3.70%	0.00%
湖北	4512	5.43%	8.57%	新疆	1075	4.19%	2.22%
湖南	4602	5.37%	8.91%	云南	2560	7.07%	7.73%
吉林	1973	7.40%	8.22%	浙江	10410	10.59%	9.26%
江苏	8896	11.23%	8.41%	重庆	2276	3.43%	6.41%
江西	3074	5.63%	4.62%				

各省份借款成功率、违约率和平均违约率



内蒙古的借款成功率最低，仅有3.4%，江苏的成功率最高，达到了11.2%；从违约率的角度来看，内蒙古的违约率略低于平均违约率，而江苏的违约率却略高于违约率。

检验地域歧视是否存在的模型（1）

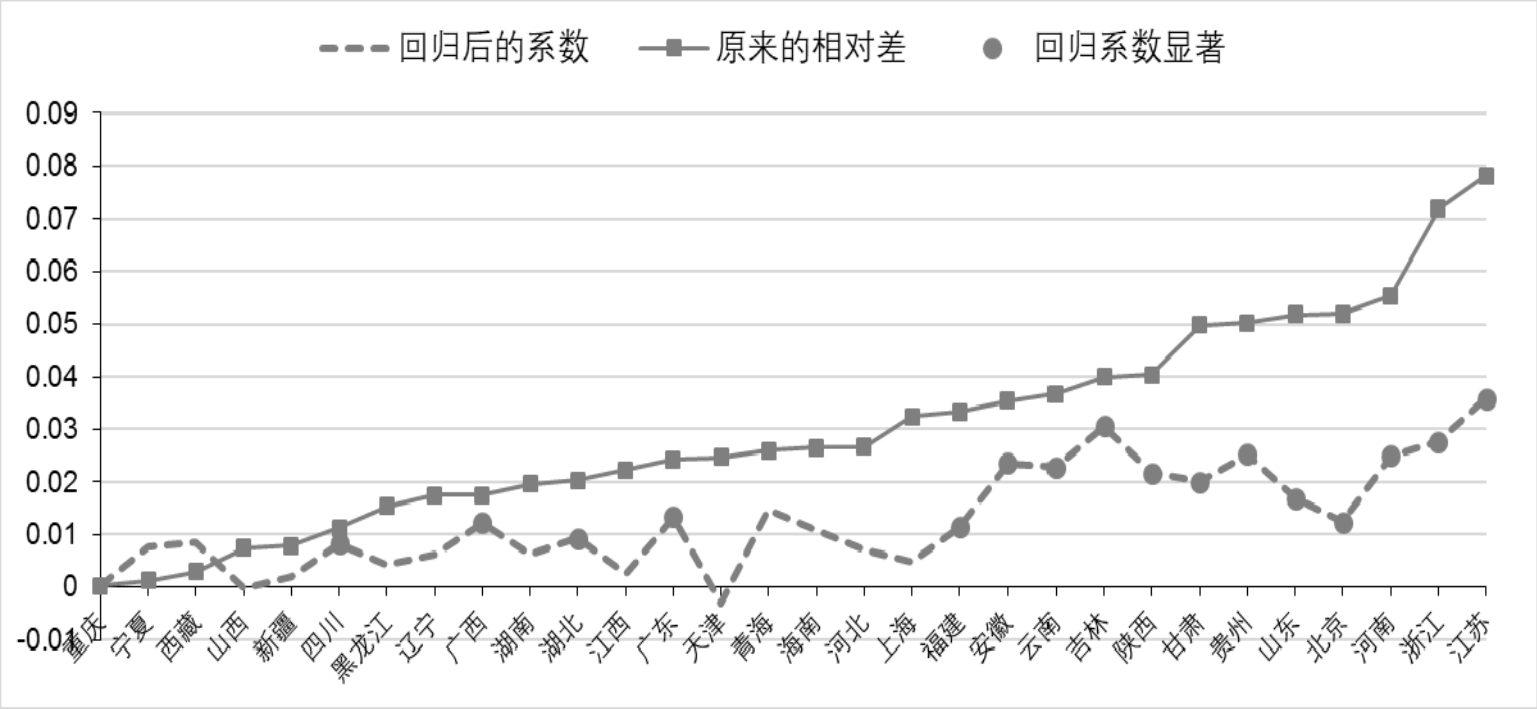
$$SUCCESS_i = \alpha + \sum \beta_n \times Province_n + \lambda \times Controls_i + \varepsilon_i$$

这里$SUCCESS_i$表示第i个借款人是否成功获得贷款，$Province_n$是省份的虚拟变量，剔除了港澳台三地的数据后，还有31个省份，因此这里设置内蒙古为对照组后，还剩下三十个虚拟变量，我们只需要检验$\beta_1 = \beta_2 = \cdots = \beta_{30}$是否联合显著，就可以说明存在地域歧视。

各省份系数的回归结果

省份	回归系数	标准误	省份	回归系数	标准误
上海	0.00471	0.005354	河北	0.00706	0.0050115
云南	0.0228***	0.0060837	河南	0.0249***	0.0052523
北京	0.0123**	0.0052505	浙江	0.0277***	0.0048218
吉林	0.0307***	0.0064521	海南	0.0109	0.0076571
四川	0.00825*	0.0047659	湖北	0.00925*	0.0050054
天津	-0.00312	0.0073158	湖南	0.00624	0.0050286
宁夏	0.0078	0.0082028	甘肃	0.0200**	0.0078838
安徽	0.0235***	0.0054286	福建	0.0114**	0.0047222
山东	0.0170***	0.004691	西藏	0.00844	0.008708
山西	-0.00028	0.0051386	贵州	0.0252***	0.0058424
广东	0.0134***	0.0044345	辽宁	0.00623	0.0051384
广西	0.0122**	0.0052037	重庆	0.000213	0.0054139
新疆	0.00201	0.0068627	陕西	0.0218***	0.0058938
江苏	0.0359***	0.004904	青海	0.0147	0.0129193
江西	0.00259	0.005292	黑龙江	0.0044	0.0054267

回归结果的可视化



上方的实线是通过统计性描述计算出来的各省份与内蒙古的成功率的相对差。
下方的虚线是回归系数，其实际意义是控制了其他变量（包括借贷合约信息、个人信用、经济与非经济信息）后的相对差。
圆形的实心标记表示该省份的回归系数显著（置信水平为90%）。

各省份系数的回归结果

本文对30个省份的系数进行了联合显著性检验，得到的F统计量：F(30,128489)=9.09，对应的p值为0.00，这说明在1%的显著性水平上，P2P借贷市场中存在着地域歧视。

附件1：模型（1）中控制变量的回归系数

被解释变量：SUCCESS			
INTEREST	-0.00355***	LNAMOUNT	-0.0139***
	(0.000136)		(0.000544)
MONTHS	-0.000818***	INCOME	0.0386***
	(7.48e-05)		(0.00181)
HOUSE	0.000955	CREDIT	0.504***
	(0.00142)		(0.00539)
CAR	0.0253***	WORKTIME	0.0211***
	(0.00185)		(0.00139)
Year	-0.00952***	MARRY	0.0104***
	(0.000719)		(0.00135)
AGE	0.00204***	EDUCATION	0.00935***
	(0.000114)		(0.00167)
Constant	19.29***	Observations	128,532
	(1.444)	R-squared	0.334

研究地域歧视的异质性的模型（2）

模型（2）从地域歧视的异质性角度来探究。首先，利用借款人的某个特征将样本进行分组，用分组后的数据分别进行回归，并比较结果的差异。本文采用的分组特征为借款人的学历和工作时间。

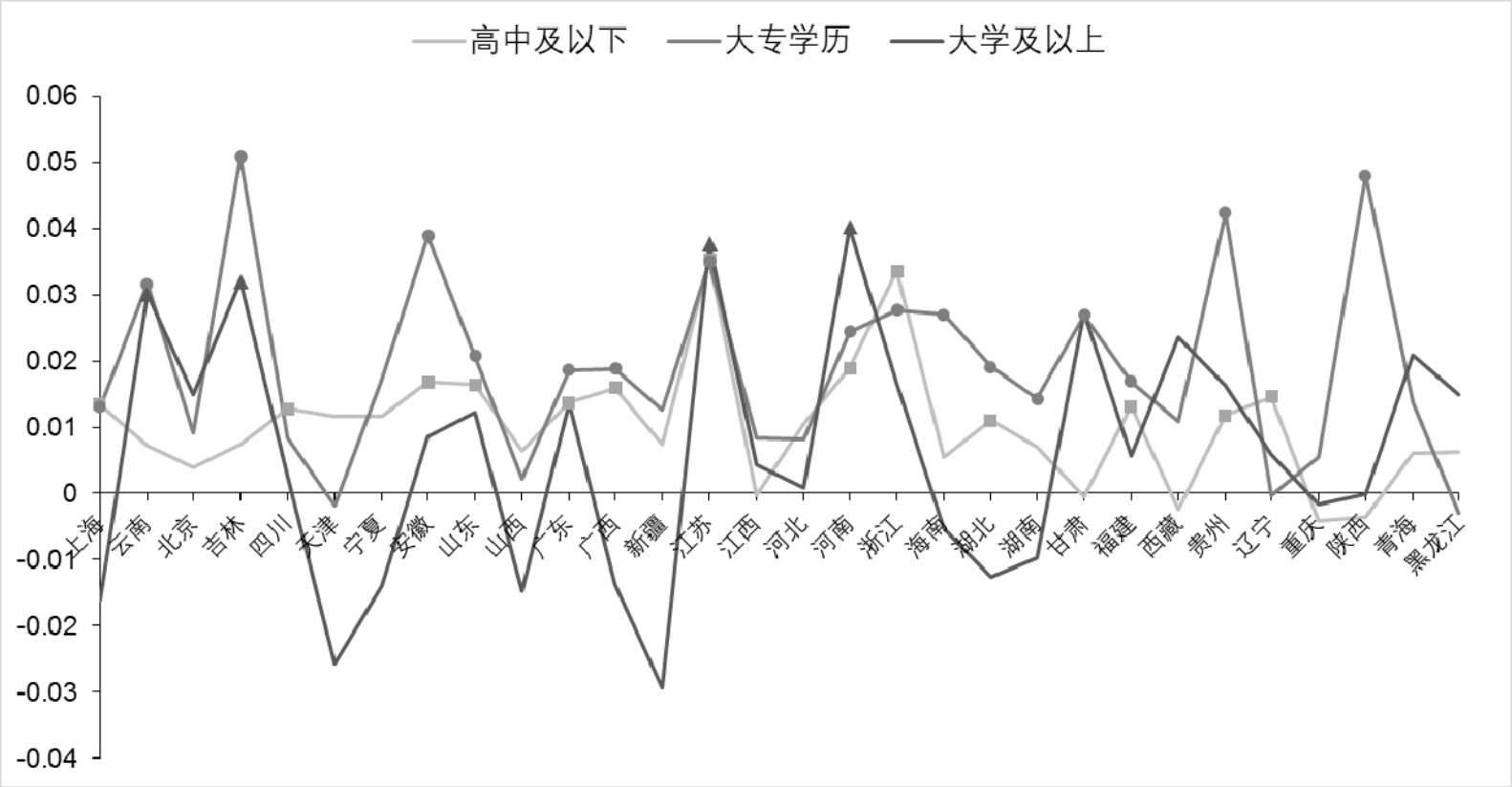
假设根据某个特征可将样本分为三个组，那么模型（2）可用下面的式子表示：

$$SUCCESS_i^{(1)} = \alpha^{(1)} + \sum \beta_n^{(1)} \times Province_n^{(1)} + \lambda^{(1)} \times Controls_i^{(1)} + \varepsilon_i^{(1)}$$

$$SUCCESS_i^{(2)} = \alpha^{(2)} + \sum \beta_n^{(2)} \times Province_n^{(2)} + \lambda^{(2)} \times Controls_i^{(2)} + \varepsilon_i^{(2)}$$

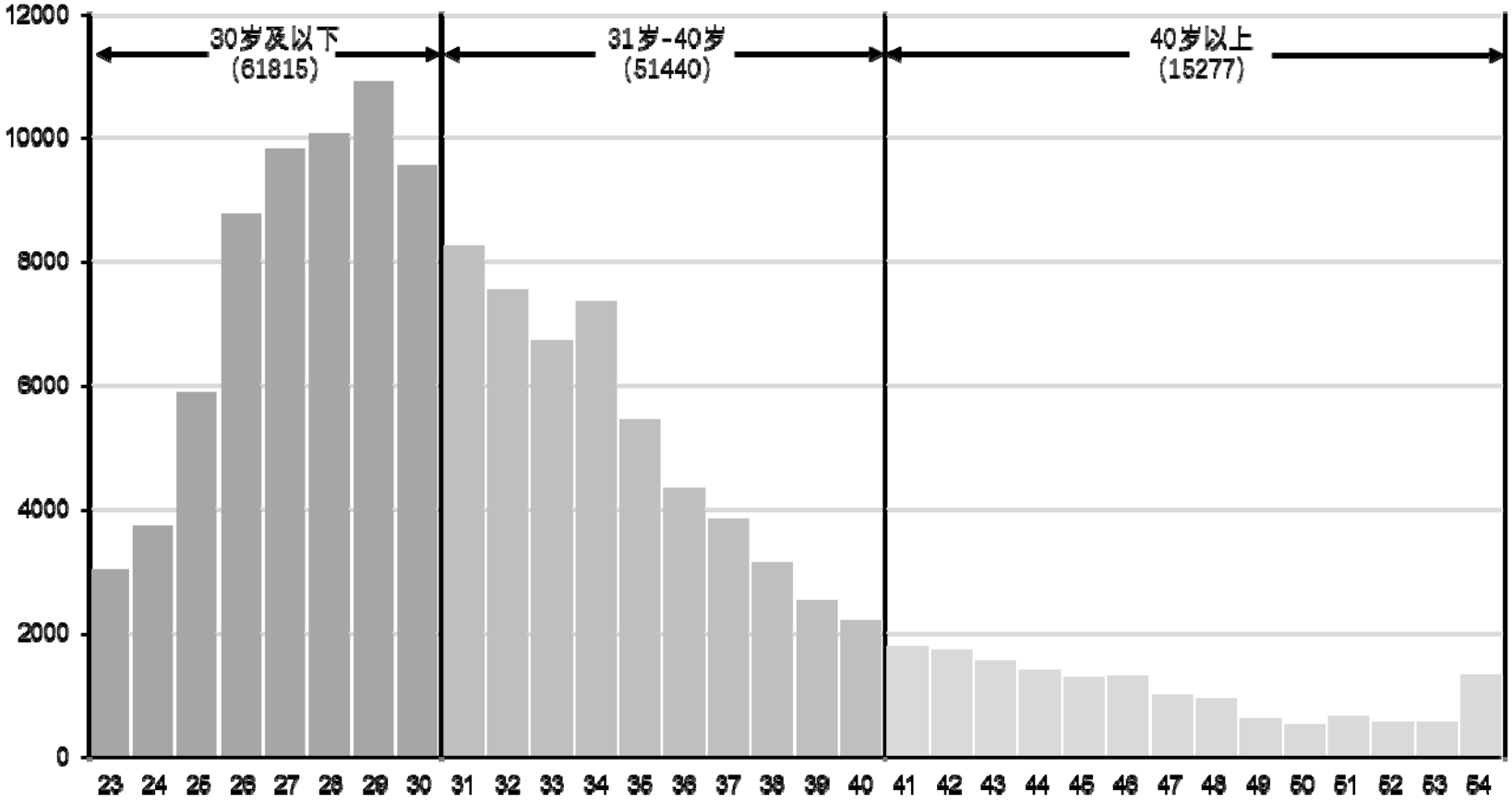
$$SUCCESS_i^{(3)} = \alpha^{(3)} + \sum \beta_n^{(3)} \times Province_n^{(3)} + \lambda^{(3)} \times Controls_i^{(3)} + \varepsilon_i^{(3)}$$

地域歧视对学历的异质性

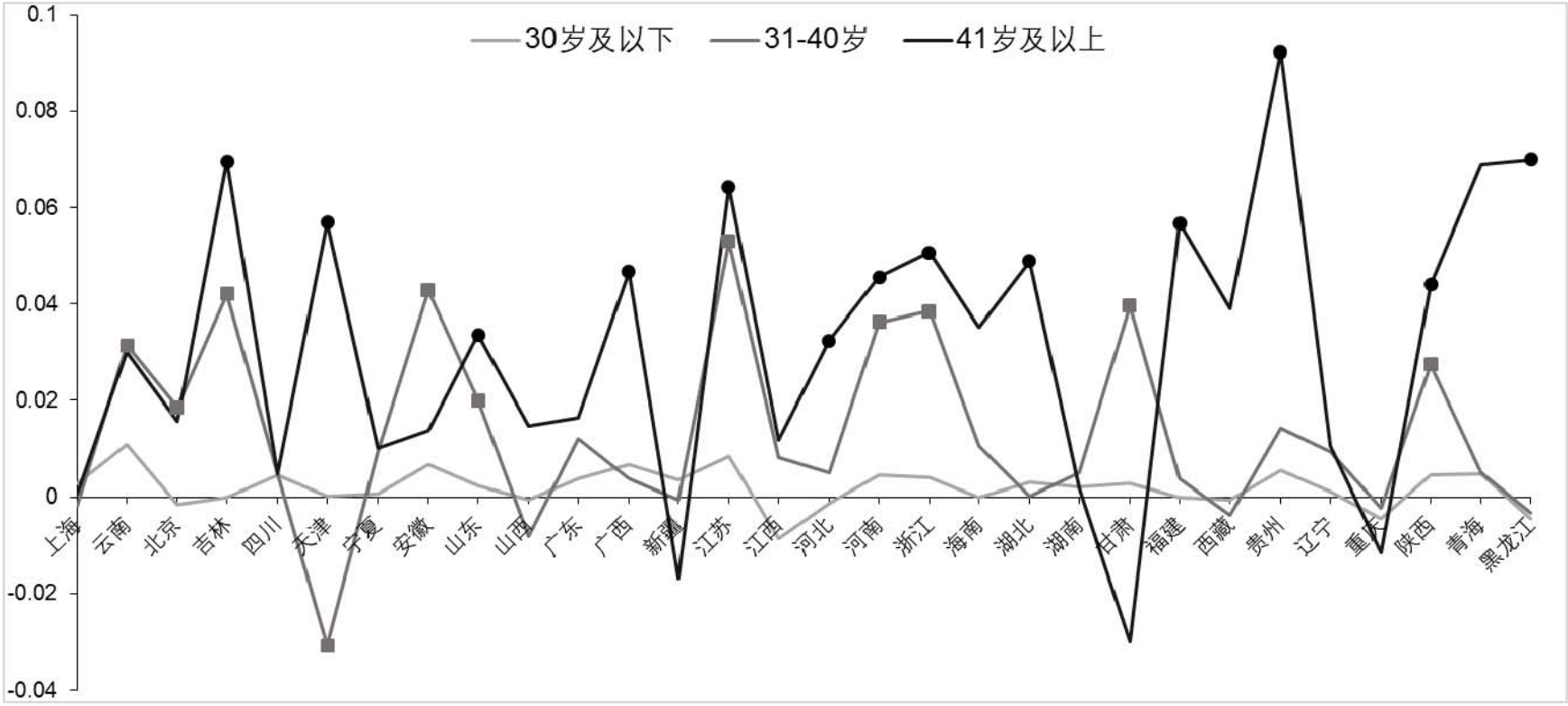


从图中可以发现，相比于另外两个分组，大学及以上学历这组中，回归系数围绕横轴上下波动，且系数显著的省份的个数远远低于另外两组，这表明：**借款人的学历越高，越不容易受到地域歧视的影响。**

借款人年龄的频数分布直方图



地域歧视对年龄的异质性



分组为30岁及以下的借款人的各省份系数相对于其他两组较小，且并没有出现显著的省份，对其进行联合显著性检验对应的p值为0.1099，这说明在30岁及以下的借款人在借款时没有受到地域歧视的影响。

结论：随着借款人年龄的增加，P2P市场上地域歧视的现象也更明显。

地域歧视是否理性的理论

早期学者根据歧视产生的原因是否具有经济解释，将歧视分为**有效统计歧视**和**非有效偏好歧视**（Becker 1957, Pheleps 1972, Arrow 1973）。

具体到我们要研究的网络借贷这一话题，**如果贷款人的歧视是由于借款人本身具有较高的违约率而引起的，那么这种歧视就是有效统计歧视（理性的）。**

若歧视完全由贷款人的个人偏好决定的，即背后并不存在合理的经济原因，则这种歧视为非有效偏好歧视。

判断地域歧视是否理性的模型（3）

$$DEFAULT_i = \alpha' + \sum \beta'_n \times province_n + \lambda' \times Controls_i + \varepsilon_i$$

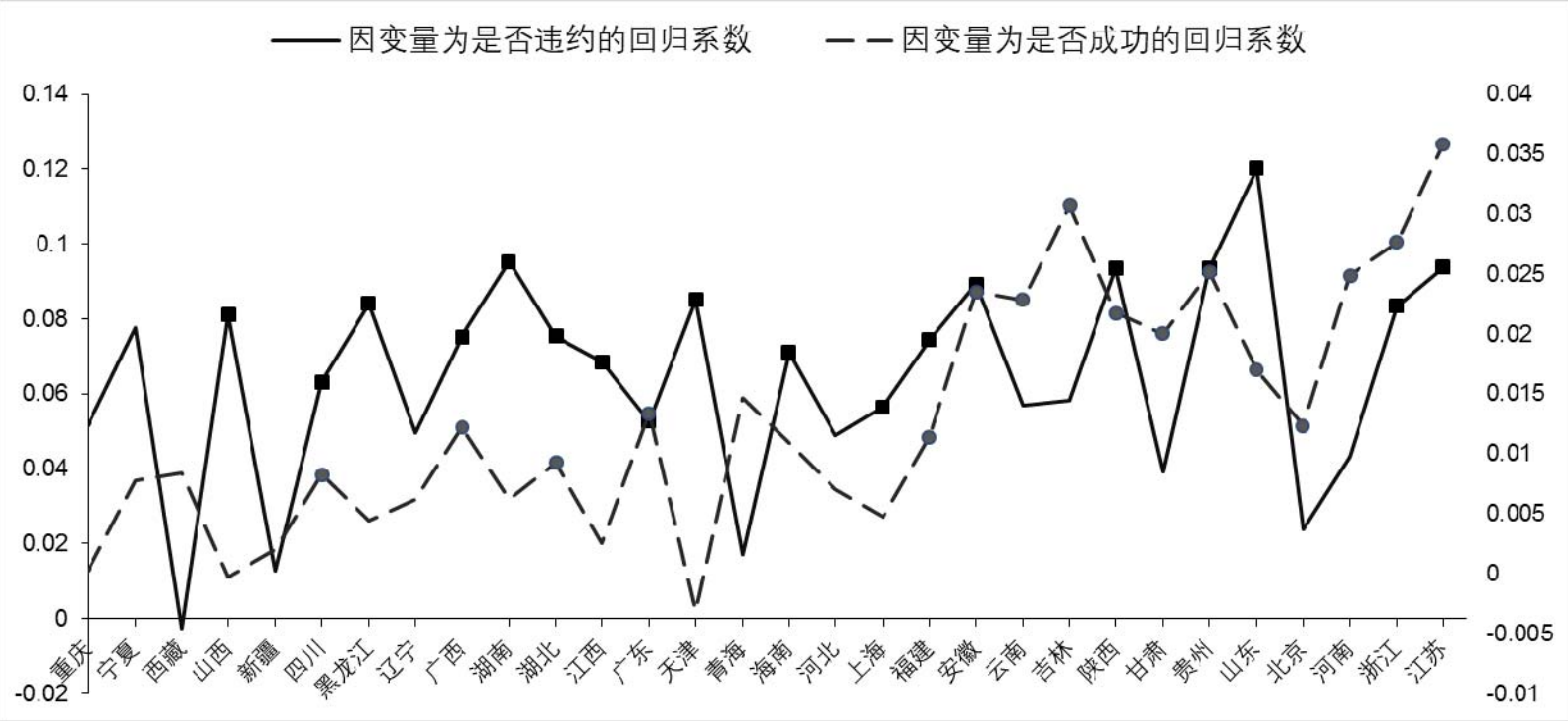
这里样本为获得借款的订单，$DEFAULT_i$表示第i个借款人是否成功获得贷款。

我们可以估计**模型（3）**，其因变量为是否违约，样本为所有成功获得借款的借款人，使用OLS可估计出每个省份的回归系数，它们反映了**各省份违约率与对照组违约率的相对差额。**

而我们在**模型（1）**中也同样估计了各个省份的回归系数，其因变量为借款是否成功，因此其回归系数表示的含义是**各省份借款成功率与对照组借款成功率的相对差额。**

如果这两次回归得到的系数具有显著的负相关关系，则表明地域歧视是理性的；若相关关系不显著甚至具有显著的正相关关系则表明地域歧视是非理性的。

模型（1）和模型（3）各省份系数的回归结果



我们可以计算两组数据的样本相关系数r,并对其进行费希尔的t显著性检验，结果显示两组数据的相关系数为0.1936，对应的t统计量的p值为0.3054，这表明两者没有显著的相关关系。

(除掉不显著的数据后，两组回归系数仍然没有显著的相关关系)

结论： P2P借贷市场中的地域歧视是非理性的，即借款人的违约率高低不是引起贷款人决定投资的主要原因。

稳健性检验：更改地域歧视的研究对象

前文的研究对象是各省份，而各省份的样本数目有较大差异，如广东样本数为18402，而西藏的样本数仅为216，两者相差85倍，样本数据的不均匀可能会导致回归结果存在一定的误差。

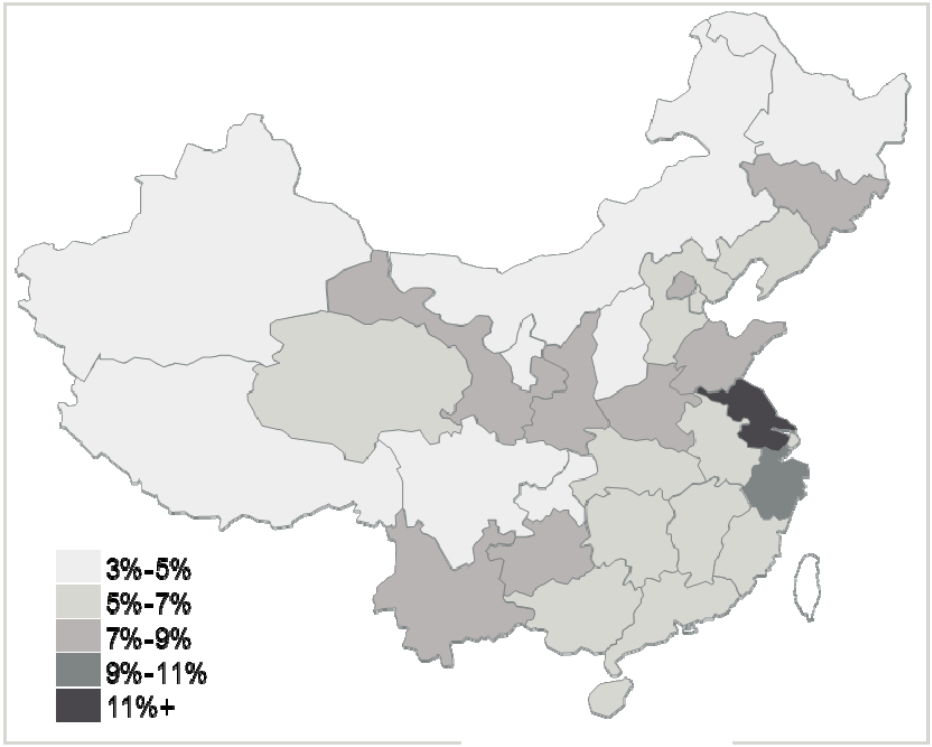


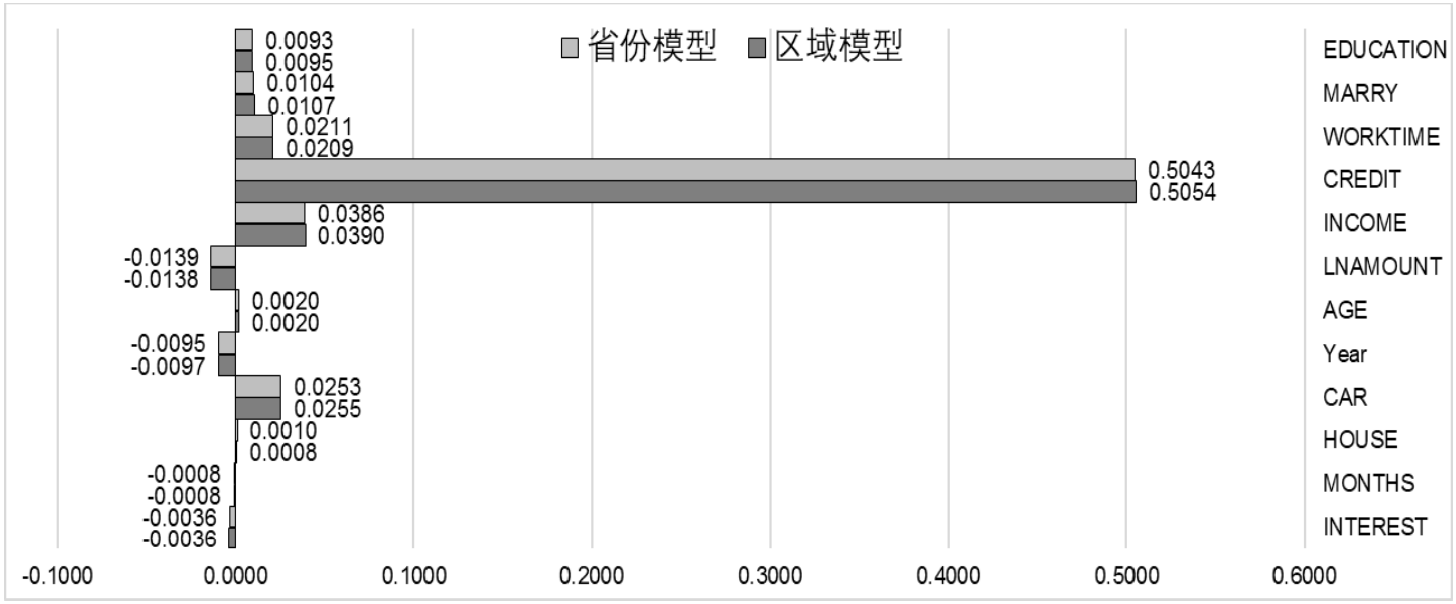
图1：各省份借款成功率的热力图

图2：按七大区域划分的中国地图

稳健性检验：更改地域歧视的研究对象

我们从下面两个角度来说明模型的稳健性：

角度一：控制变量的回归系数



角度二：区域回归系数的联合显著性

我们对模型中六个区域的回归系数进行联合显著性检验，得到的F统计量为11.25，对应的p值为0.00，这说明在1%的显著性水平上，P2P借贷市场中存在着地域歧视，与研究对象为省份时保持了一致。

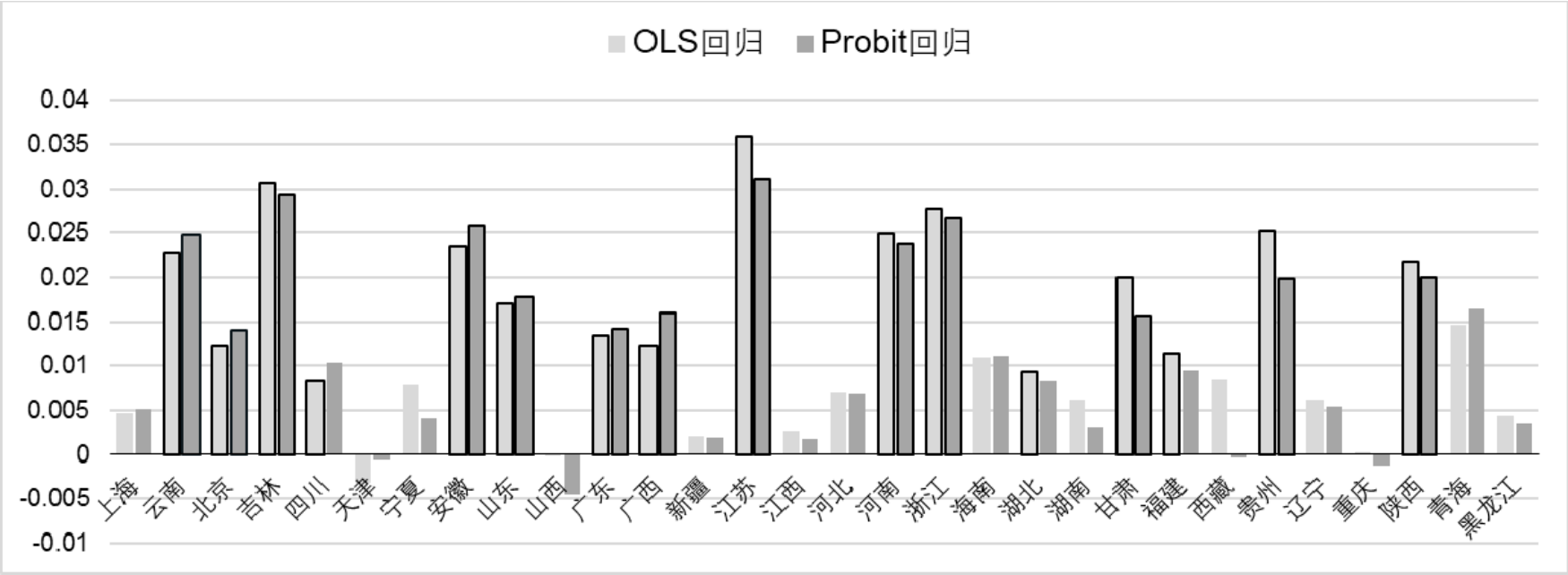
稳健性检验：更改计量方法

因为本文并不关注借款人所在的省份对于其借款成功率的具体影响大小，所以前文的实证模型采用的均为OLS估计。

但是，订单的借款成功率为二值变量，为了排除计量方法选择导致的误差，本文重新使用Probit模型进行估计。

由于 *Probit* 是非线性模型，估计量 $\hat{\beta}_{MLE}$ 表示的并不是自变量的边际效应，注意到: $\frac{\partial P(y=1|\boldsymbol{x})}{\partial x_i} = \frac{\partial P(y=1|\boldsymbol{x})}{\partial(\boldsymbol{x}'\boldsymbol{\beta})}\frac{\partial(\boldsymbol{x}'\boldsymbol{\beta})}{\partial x_i} = \frac{\partial \Phi(\boldsymbol{x}'\boldsymbol{\beta})}{\partial(\boldsymbol{x}'\boldsymbol{\beta})}\beta_i = \phi(\boldsymbol{x}'\boldsymbol{\beta})\beta_i$，所以其边际效用会随着解释变量的变化而变化。为了和OLS模型的系数进行对比，我们需要计算平均边际效应，即先计算在每个样本的边际效应，然后再计算其简单算术平均数。

稳健性检验：更改计量方法



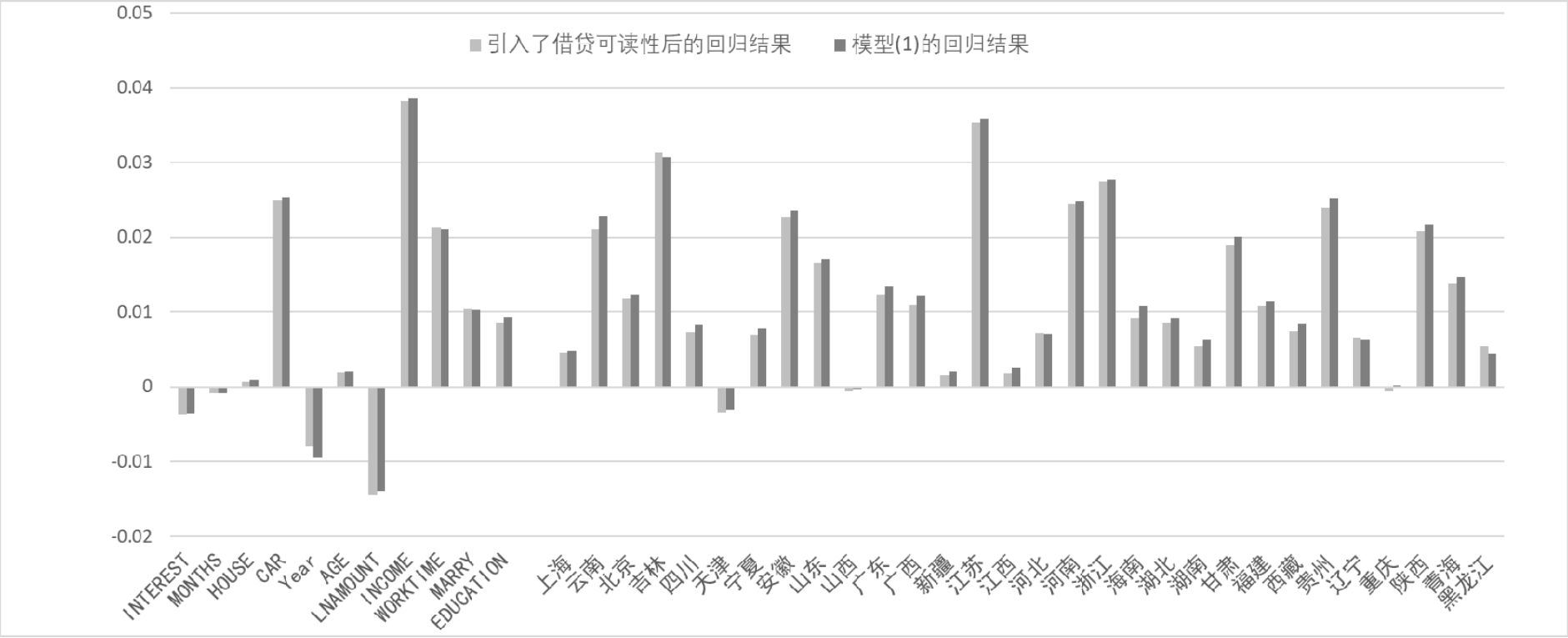
1. 从图中可以看出，两个模型各省份系数的估计量没有较大差异。
2. 除了3个省份的显著性不一样，其他27个省份的显著性完全一致。
3. 论文附件中也给出了控制变量的系数对比图，两个模型的差异并不明显。
4. 对各省份回归系数进行联合显著性检验，得到的p值为0，这说明了地域歧视现象的存在。

稳健性检验：加入其他控制变量

前文的回归模型中引入了12个控制变量以消除内生性的影响，但我们忽略了借款人的借款描述信息。事实上，李焰（2014）、陈霄（2018）等人的研究均发现：借款人在借款时描述的可读性能显著提高借款的成功率。因此我们在模型中引入衡量借款人描述可读性的控制变量，对模型（1）重新进行回归。

具体的回归结果见论文，与不增加这一控制变量的模型对比，各自变量回归系数的大小和显著性均无明显变化，另外对各省份系数进行联合显著性检验，得到F统计量：$F(30,128485) = 9.01$，其对应的p值为0，这说明地域歧视现象仍然存在，因此我们的模型是稳健的。

稳健性检验：加入其他控制变量



是否加入借贷信息可读性的回归系数对比图

本文研究结论

本文利用人人贷平台上借贷交易的数据，研究了我国P2P市场上地域歧视的现象，主要得到了以下三点结论：

1. 在控制了借款人的信息和借贷信息等变量后，回归结果表明各省份的借款成功率存在显著的差异，且其系数联合显著不为0，这说明**我国P2P借贷市场上存在着地域歧视现象**；此外，对省份按地域重新分组后进行回归的结果也得到了相同的结论。
2. 我国P2P市场上的地域歧视现象对于借款人的学历和年龄存在异质性。一方面，**借款人的学历越高，越不容易受到地域歧视的影响**；另一方面，**随着借款人年龄的增加，P2P市场上地域歧视的现象也更明显**。
3. P2P市场上的地域歧视并不是理性的，借款成功率低的省份的借款者并没有更高的违约率，这说明**我国P2P平台上存在着“非有效的偏好地域歧视”**。

本文的建议

首先，作为在P2P网络借贷中提供资金的**贷款人**，在选择借款订单进行投资时，应摒弃固有的地域歧视观念，更多关注借款人自身的收入水平、信用等级、学历水平等个人信息以及借款订单相关信息，充分运用P2P网站提供的各类信息综合判断借款人的偿债能力与信用水平。

其次，对**借款人**而言，应注重自身信用等级的培养，合理设计借款订单，提供更多的有效个人信息，以消除地域歧视的影响，进而提高借款成功率。

最后，对**P2P网络借贷平台**而言，应利用平台优势，借助大数据分析技术，进一步设计和完善个人信用评价体系，发挥信用评级在P2P借贷中的重要作用；在订单信息公布中，突出能够反映借款人偿还能力的个人信息以及反映违约风险的订单信息，进而弱化地域歧视对借贷双方的影响，提高P2P网络借贷中的效率；应敦促借款人完善个人信息，为借款订单设计更为详尽的信息提供方式，尽量避免地域歧视产生的影响。

不足之处与后续展望

（1）样本在各省份的分布不够均匀。本文使用的数据来自人人贷官网，共计128532条（2011年1月到2014年6月的订单），由于各省份的数据量存在较大差异，因此回归中可能会出现偏误。

（2）本文对于地域歧视的原因的探讨还不够深入，在未来的研究中可以引入区域经济发展水平、文化差异，从这些角度去进一步探讨，进而为缓解“地域歧视”现象提供依据。

（3）未来可以控制更多变量以尽可能消除内生性的影响。

请各位老师批评指正！