附录 1 遗漏变量导致 OLS 估计低估原理

由于遗漏变量偏误的表达式为 $E\left(\beta_1\right)-\beta_1=\gamma\beta_3$，其中 β_3 为村/居经济文化因素等遗漏变量对个体初育年龄的影响，预期 $\beta_3<0$，即村/居越落后个体的初育年龄越提前。γ 为遗漏变量对个体是否留守的影响，由于落后地区的农民越倾向于外出务工以增加收入，个体被留守的概率也越大，因此预期 $\gamma>0$。由此可推断 $\gamma\beta_3<0$，$E\left(\beta_1\right)>\beta_1$，村/居层面的遗漏变量会导致 OLS 估计低估。

附录 2 工具变量“个体 0-12 岁期间出生地春季是否降水异常”的构造方法

首先，利用中国国家气象中心（NMIC）1980-2010 年的日度降水数据，将 2000 多个地面气象观测站的数据使用反距离加权法（IDW）插值成 $0.1^{\circ}\times0.1^{\circ}$ 分辨率的网格数据，再将网格数据转换成栅格数据，接着按区域汇总平均得到各区县历年的春季平均降水量。然后，参考魏东霞和陆铭（2021）采用的降水距平百分率指标，来衡量个体 0-12 岁期间出生地春季的降水情况：

$$H_{ij}=\frac{\left(\bar{x}_{ij}-\bar{x}_j\right)}{\bar{x}_j}\times100\%$$

其中，$\bar{x}_{ij}$ 为个体 i 0-12 岁期间出生地所在区县 j 的春季平均降水量，$\bar{x}_j$ 为区县 j 在 1980-2006 年的春季平均降水量；考虑到本文样本出生年份在 1980-1994 年，所定义的留守经历为其 0-12 岁期间的经历，因此 1980-2006 年覆盖了全部样本的童年阶段。以上式计算得到的降水距平百分率 H_{ij} 即衡量了个体 0-12 岁出生地的春季平均降水量相对于整个样本期内平均降水量的偏离情况；该指标考虑了地区常年降水量特征，具有很好的时空对比性。最后，我们借鉴鞠笑生等（1997）按照降水距平百分率的取值划分 7 个旱涝等级的做法，将个体 0-12 岁出生地春季的降水情况等分为 7 个等级，分别为降水非常少、降水很少、降水较少、降水正常、降水较多、降水很多和降水非常多，并构造是否降水异常的虚拟变量，当降水正常时取值为 0，否则取值为 1。

附表 1 2010CFPS 问卷中用于测度大五人格的问题

大五人格	描述	2010CFPS 问卷对应问题
严谨性	具有胜任、公正、条理、尽职、成就、自律、谨慎、克制等特点	1.受访者的衣着整洁程度*
		2.有成就感的重要程度
		3.多大程度上赞同“努力工作能得回报”
		4.受访者对调查的疑虑*
外向性	表现出热情、社交、果断、活跃、冒险、乐观等特质	1.受访者的待人接物水平*
		2.不孤单的重要程度
		3.生活有乐趣的重要程度
顺同性	具有利他、直率、依从、谦虚、移情等特质	1.不被人讨厌的重要程度
		2.在与人相处方面能大几分

		3.受访者对调查的配合程度*
开放性	具有想象、审美、情感丰富、求异、创造、智能等特质	1.受访者对调查的兴趣* 2.传宗接代的重要程度（反向）
情绪稳定性	能较好地平衡焦虑、敌对、压抑、自我意识、冲动、脆弱等情绪的特质，即具有保持情绪稳定的能力	1.感到神经紧张的频率（反向） 2.感到坐卧不安，难以保持平静的频率（反向） 3.感到情绪沮丧、郁闷、不振奋的频率（反向） 4.感到未来没有希望的频率（反向） 5.做任何事都感到困难的频率（反向） 6.认为生活没有意义的频率（反向）

注：标*的问题为调查人员回答的问题，取值范围从 1-7 分别代表非常差/低-非常好/高；其他问题为受访者自行回答的问题，取值范围从 1-5 分别代表非常不重要/不赞同/低-非常重要/赞同/高；标记为反向的问题即意味着对其进行了反向赋值，使之能正确反映所衡量的特质。

附表 2　　变量描述性统计

变量	样本量 （1）	均值 （2）	标准差 （3）	最小值 （4）	最大值 （5）
初育年龄	1027	22.820	2.486	15	30
0~12岁是否留守（是=1）	1027	0.090	0.286	0	1
0~3岁是否留守（是=1）	1027	0.041	0.198	0	1
4~12岁是否留守（是=1）	1027	0.078	0.268	0	1
是否短期留守（是=1）	1027	0.020	0.142	0	1
是否较长留守（是=1）	1027	0.042	0.200	0	1
是否长期留守（是=1）	1027	0.027	0.163	0	1
性别（男性=1）	1027	0.414	0.493	0	1
年龄	1027	26.490	2.742	16	31
年龄的平方	1027	709.400	141.600	256	961
民族（非汉族=1）	1027	0.154	0.361	0	1
城乡（居住地在城镇=1）	1027	0.287	0.453	0	1
健康状况	1027	1.378	0.610	1	5
受教育年限	1027	7.792	3.479	0	19
父亲学历	1027	2.187	0.998	1	6
母亲学历	1027	1.683	0.861	1	4
父亲政治面貌	1027	3.715	0.844	1	4
母亲政治面貌	1027	3.927	0.387	1	4
家庭年收入	1027	10.140	0.914	5.707	13.61
家庭房产	1027	1.993	0.997	0	5.303

注：受教育年限为文盲/半文盲=0，小学=6，初中=9，高中=12，大专=15，本科=16，研究生=19；健康状况为健康=1，一般=2，比较不健康=3，不健康=4，非常不健康=5；父母学历为文盲/半文盲=1，小学=2，初中=3，高中=4，大专=5，本科=6；父母政治面貌为共产党员=1，共青团员=3，群众=4。

附表 3　　Heckman 两步法回归结果

变量	(1) 第一阶段 是否生育	(2) 第二阶段 初育年龄
0~12岁是否留守		0.612**
		(0.251)
是否结婚（是=1）	3.149***	
	(0.259)	
个体特征	控制	控制
家庭特征	控制	控制
区县效应	控制	控制
样本量	3163	
Wald chi^2	60.67***	
ρ	0.8441***	

注：*、**和***分别表示在 10%、5%和 1%的显著性水平上显著；括号内是聚类到区县的稳健标准误，下表同。

附表 4　　控制可能遗漏变量的估计结果

变量	(1) 初育年龄	(2) 初育年龄	(3) 初育年龄
是否留守	0.719**	0.621*	0.691**
	(0.303)	(0.335)	(0.347)
农业劳动力占比	-0.011**		-0.015***
	(0.005)		(0.005)
计划生育超生最低处罚金额对数		0.134	0.069
		(0.149)	(0.134)
个体特征	控制	控制	控制
家庭特征	控制	控制	控制
区县固定效应	控制	控制	控制
R^2	0.468	0.459	0.472
样本量	831	714	702

注：样本量相对于基准回归的变化是因为样本在村/居层面的变量存在不同程度的缺失，并且剔除了 12 岁之前和被调查时居住地发生了变化的样本。

附表 5　　工具变量的估计结果

变量	(1) 第一阶段 是否留守	(2) 第二阶段 初育年龄	(3) 第一阶段 留守时长	(4) 第二阶段 初育年龄	(5) 第二阶段 初育年龄	(6) 第二阶段 初育年龄
0~12 岁是否留守		1.530*		1.683**	1.521*	1.674**

		(0.830)		(0.676)	(0.804)	(0.708)
0~12 岁期间出生地春季是否降水异常（是=1）	0.045*		0.057*			
	(0.025)		(0.032)			
出生地留守儿童占比	1.008***		1.134***			
	(0.141)		(0.144)			
个体特征	控制	控制	控制	控制	控制	控制
家庭特征	控制	控制	控制	控制	控制	控制
区县固定效应	控制	控制	控制	控制	控制	控制
第一阶段F统计量	25.40		31.35		——	——
Hansen J 统计量p值	0.143		0.113		——	——
样本量	986		705		986	705

注：列（1）-（2）为工具变量的 2SLS 估计结果，（3）-（4）为剔除样本所在村/居为自然灾害频发区的样本后的工具变量 2SLS 回归结果，（5）-（6）为使用含内生变量的 Tobit 模型进行回归的结果。其中 Tobit 模型的估计系数均已转换为边际效应，样本量相对于基准回归的减少是因为剔除了 3 岁及 12 岁的居住地和出生地发生了变化的样本。

附表 6　　替换变量的稳健性检验

变量	(1) 初育年龄	(2) 是否晚育	(3) 初育年龄（2020）	(4) 初育年龄（2022）
0~12岁是否留守		0.122**	0.429*	0.515**
		（0.053）	（0.218）	（0.242）
留守时长	0.114*			
	(0.001)			
个体特征	控制	控制	控制	控制
家庭特征	控制	控制	控制	控制
区县固定效应	控制	控制	控制	控制
R^2	0.470	0.402	0.528	0.522
样本量	1027	1027	1709	1587

附表 7　　缓解潜在回忆偏差的稳健性检验

变量	(1) 初育年龄	(2) 初育年龄
0~12岁是否留守	0.536**	0.463*
	（0.268）	（0.277）
留守时长		
个体特征	控制	控制
家庭特征	控制	控制

区县固定效应	控制	控制
R^2	0.487	0.514
样本量	914	796

附表 8　　评估不可观测因素影响

有限集控制变量	全集控制变量	*Ratio*
无控制变量	所有控制变量	2.26
仅控制区县固定效应	所有控制变量	3.20
仅控制个体和家庭特征、无固定效应	所有控制变量	6.84
无控制变量	所有控制变量+重要遗漏变量	2.13
仅控制区县固定效应	所有控制变量+重要遗漏变量	2.89
仅控制个体和家庭特征、无固定效应	所有控制变量+重要遗漏变量	5.31
		3.77

附表 9　　不同群体的影响异质性

	(1)	(2)	（3）	（4）
	高先天人力资本	低先天人力资本	男性	女性
变量	初育年龄	初育年龄	初育年龄	初育年龄
0~12 岁是否留守	1.073	1.877**	0.654	1.915**
	(2.495)	(0.896)	(1.151)	(0.901)
个体特征	控制	控制	控制	控制
家庭特征	控制	控制	控制	控制
区县固定效应	控制	控制	控制	控制
样本量	117	869	409	577
	(5)	(6)	(7)	(8)
	到县城时间短	到县城时间长	现居住在农村	现居住在城镇
变量	初育年龄	初育年龄	初育年龄	初育年龄
0~12 岁是否留守	1.737	2.528*	2.145*	0.933
	(1.501)	(1.437)	(1.118)	（1.115）
个体特征	控制	控制	控制	控制
家庭特征	控制	控制	控制	控制
区县固定效应	控制	控制	控制	控制
样本量	411	430	704	282

注：到县城时间长/短的划分依据为村/居到县城需要花费时间高/低于所有样本对应时间的中位数；上表回归均采用工具变量 2SLS 方法缓解内生性问题。