实践作业: 评审定量研究论文

姓名: 方嘉聪 学号: 2200017849 得分: 95

1 党员的生活是否更加幸福?

1.1 研究整体评价

该文章研究问题为 中国共产党的党员身份是否能够显著影响人们的生活满意度? 作者使用了中国家庭最终调查 (CFPS) 数据, 分别构建了一元线性回归模型与多元线性回归模型, 来论证中国党员身份可以显著提升人们的生活满意度. 整体来说, 该文章的研究问题具有一定的现实意义, 但在定量研究设计细节、统计方法使用和解读方面还有诸多缺陷, 下面将详细分析:

1.2 研究设计及变量选取

该文章选取 CFPS 调查中的生活满意度指数作为因变量, 是否为中共党员为核心自变量, 这一选取是相对符合研究问题的.[1]

(1) **控制变量选取**: 作者额外选取了教育水平、婚姻状况和省份作为控制变量. 注意这里婚姻状况和省份对是否为中共党员无关, 选择控制变量需要同时满足对因变量和核心自变量相关.
   - 该文章遗漏了一些其他可能会同时影响生活满意度和党员身份的变量, 如收入水平、家庭背景、城乡户籍等.
   - 只考虑文中有限的控制变量, 而未考虑上一点提到的其他可能会导致遗漏变量偏误, 从而影响到结果的可靠性.

(2) **变量统计性描述**: 在该文章中, 生活满意度可以视作定量变量, 求均值和方差是合理的.
   - 但是否为党员为二分变量, 通常应当将值设置为 0/1, 而不是 0/2. 且在后续回归分析中, 应当引入虚拟变量进行处理.
   - 身份及教育水平为定序变量, 在这里计算均值和方差是不合适的, 应当使用频数/频率表进行描述.
   - 此外在表 1 中没有具体指出统计的样本为 CFPS 全体, 还是进行了进一步筛选.

此外在分析变量分布时, 依据对教育水平 (定序变量) 的没有实际意义的均值得出党员在教育水平上也略微高于非中共党员, 是不可靠的. 可以考虑使用受教育教育年限作为教育水平的度量或是使用频数表进行描述.

1.3 统计方法使用

该文章使用了一元线性回归模型与多元线性回归模型, 但在具体的统计方法使用上还有一些问题:

(1) 对于党员身份、教育水平、婚姻状况等定类或定序变量, 应当先将其转化为虚拟变量, 再进行回

[1] 不确定是否有更好的指标来衡量生活满意度

归分析. 从提供的回归表 (没有指明参照组) 可以发现该文章直接将这些变量作为定量变量纳入回归方程, 这样得到的回归系数是没有论证意义的.

(2) 回归方程的细节上, 作者并未将残差项纳入其中, 在公式表示上不够规范.

1.4 统计结果解读

文章认为两个模型表明在 95% 的置信区间内党员身份对生活满意度具有显著的正向影响. 进而得到了结论: 中国共产党的党员身份能够显著提升人们的生活满意度. 存在以下问题:

(1) 在回归分析中一般使用 R^2 来衡量模型的拟合的显著性, 文章中构建的模型 $R^2 = 0.0039, 0.0062$, 这说明文中构建的两个模型对因变量的解释力很弱. (这里可以考虑纳入其他的控制变量)

(2) 置信区间一般分析的是回归系数的显著性. 本文中得到的党员身份回归系数在 0.01 的显著性水平上统计显著, 上这里在分析时更为规范的表述为"在置信度为 99% 的情况下, 控制其他变量的影响后, 成为党员可以提高生活满意度若干个单位".

1.5 总结

由于文章将所有的定序变量都当作定量变量进行回归处理, 得到得回归模型 R^2 很小, 模型的解释力弱, 且在解释上存在问题, 得到的结论是不可靠的. 建议应当如上重新考虑其他的控制变量, 并使用虚拟变量进行处理, 设计更为合理的回归模型.

2 成长环境会影响生育意愿吗?

2.1 研究整体评价

该文章试图研究 城乡二元结构是否会对人的生育意愿产生影响, 并以此为基础探究城乡差异之下收入、受教育水平等变量对生育意愿的效用的交互效应. 利用 2017 年中国社会综合调查 (Chinese General Social Survey, CGSS) 的数据进行多元线性回归分析. 下面将详细分析该文章的研究设计、统计方法使用和统计结果解读:

2.2 定量研究设计

该文章选择个人生育意愿作为因变量 (通过 CGSS 调查中的相关问题得到), 选择成长环境和受教育程度为核心自变量, 选择了年龄、性别、健康程度、个人全年总收入作为控制变量.

(1) **核心自变量选取**: 该文章以"您 14 周岁的常居地属于?"来衡量成长环境, 这一问题不能完全代表成长环境. 此外, 文章预期想要探究的是城乡二元结构对生育意愿的影响, 仅用成长环境这一变量无法完全符合预设的研究问题. 更贴合研究问题的设计可以按照不同年代的不同城镇化水平进行划分或者考虑使用如户籍等其他变量, 再控制其他变量进行分析.

(2) **变量统计描述**: 对于定类 (如性别) 和定序变量 (如成长环境和健康程度) 的均值和方差的计算是没有意义的, 应当使用频数表或百分比表进行描述.

(3) **控制变量选取**: 可以考虑其他的控制变量, 如家庭人口规模、民族、政治面貌等.

2.3 统计方法使用

该文章使用了多元线性回归模型, 构造了嵌套模型, 但在具体的统计方法使用上还有一些问题:

(1) **制表规范**: 没有在回归表中注明虚拟变量 (如性别) 的参照组.

(2) 作者将成长环境 (定序变量) 直接作为定量变量纳入回归方程. 可以采用的改进方式是按成长环境处理为分组数据, 来探究对于不同的成长环境, 是否存在差异.

(3) **回归模型设计**: 文章关注的核心自变量为成长环境和受教育程度, 但文章中选择的模型 I 只纳入了控制变量. 建议可以考虑将模型 I 中只将核心自变量 (或其一) 纳入, 在模型 II,III,IV 或其他模型中加入控制变量或交互项.

(4) 注意这里对于计数型的变量使用 OLS 是不合适的

2.4 统计结果解读

在统计结果解读上, 有以下问题:

(1) **表述不规范**: 在分析教育年限和成长环境的主效应时, 应当强调“控制其他变量的影响”, 而不是直接得出结论.

(2) 此外, 对于主效应而言, 使用模型 III 的结果来分析更为合适, 而模型 IV 中教育年限和成长年限的部分影响包含在了交互项中. 类似的比较标准化系数也更倾向于使用模型 III 的结果.

2.5 总结

文章最后的总结中认为“随着城镇化的不断推行，在城市中成长的个体越来越多，由人口流动带来城乡差异经历，进一步改变个体生育意愿的情况将得到缓解，从而有可能使生育意愿的下降趋于平缓”. 但从文章的回归分析结果来看, 说明的是教育对生育意愿和成长环境对生育意愿的影响时相互削弱的, 但整体还是会存在着负向影响. 文章的这一推论很古怪.

因而, 该文章存在如上所述的问题, 应当考虑对核心自变量重新设计, 并合理对变量进行处理 (区分好变量层次), 以得到更为可靠的结论.

最后编译时间: 2024 年 12 月 24 日