

## 定量研究方法课程作业（R语言操作2）

1. 利用上次获得的数据集cgssincome，构建线性回归模型，因变量为收入（income）的对数，自变量为教育年限（eduear）。根据模型的参数估计结果，写出回归方程，并解释教育年限的回归系数的含义，该系数在0.05的显著性水平上显著吗？截距的估计值是多少，其含义是什么？（建议：构建回归方程前，剔除掉收入为零的观测）
2. 将回归模型所代表的直线绘制在其散点图上。回归系数的95%的置信区间是多少（可以使用confint函数直接计算）？教育年限为12年的人的收入预测值为多少？预测值的95%的置信区间是多少？并且（较难可选做）绘制回归直线的95%置信区间。
3. 根据下面给定的关系构建线性回归模型，
$$\log(\text{income}) = \beta_0 + \beta_1 \text{eduear} + \beta_2 \text{workexp} + \beta_3 \text{workexp}^2 + \beta_4 \text{party} + \beta_5 \text{sex} + \epsilon$$
，其中因变量为收入（income）的对数，自变量为教育年限（eduear）、工作年限（workexp）、是否党员（party）、性别（sex）。解释得到的各个回归系数。与前面的简单回归相比，教育年限的系数发生了什么变化，可能是什么原因导致的变化？
4. 用plot命令绘制回归诊断图，并解释其结果。对模型的多重共线性进行检验。检验两个回归模型之间是否存在统计意义上的差异，应该选择哪个模型？
5. （较难可选做）将两个回归模型所代表的直线绘制在其散点图上比较其差异；通过分性别制图对比教育对收入的影响在性别间的差异；绘制工作年限与收入对数之间的关系的拟合曲线。
6. 在回归模型中添加性别与工作年限的交互项，解释交互项的系数。通过模型比较，有必要增加该交互项吗？（可选做）将省份变量（province）转化为若干个虚拟变量，将虚拟变量引入回归模型，解释省份虚拟变量的回归系数。